Impacts of missing data in risk management

Patricia dos Santos

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Université Paris I Panthéon-Sorbonne
École de Management Panthéon-Sorbonne - UFR 06
Laboratoire de rattachement : PRISM Sorbonne

THÈSE
Pour l'obtention du titre de Docteur en Sciences de Gestion
Présentée et soutenue publiquement
le 20 Décembre 2021 par

Patricia Dos Santos

Impacts of Missing Data in Risk Management

Membres du jury

Michel CROUHY, Co-responsable de la thèse CIFRE en entreprise, Natixis
Jean-David FERMANIAN, Professeur, ENSAE Paris, Rapporteur
Julie JOSSE, Advanced Researcher (HdR), INRIA
Jean-Paul LAURENT, Professeur, Université Paris I Panthéon-Sorbonne, Directeur de thèse
Yannick MALEVERGNE, Professeur, Université Paris I Panthéon-Sorbonne, Président du Jury
Adil REGHAI, Responsable de la thèse CIFRE en entreprise, Natixis
Christian-Yann ROBERT, Professeur, ENSAE Paris, Rapporteur

Ce travail a été réalisé dans le cadre du laboratoire d'excellence ReFi, portant la référence ANR-10-LABX-0095. Ce travail a bénéficié d'une aide de l'Etat gérée par l'Agence Nationale de la recherche au titre du projet Investissements d'Avenir Paris Nouveaux Mondes portant la référence n° ANR-11-IDEX-0006-02.

Remerciements

Tout d'abord, je tiens à remercier mon directeur de thèse, le Professeur Jean-Paul Laurent, pour son encadrement, ses multiples conseils, sa disponibilité et la confiance qu'il m'a accordée en acceptant d'encadrer cette thèse. Je remercie également Adil Reghai de m'avoir donné la chance de faire une thèse CIFRE au sein de Natixis, sans qui ce travail doctoral n'aurait sans doute jamais vu le jour. De plus, je remercie les rapporteurs de cette thèse, le Professeur Jean-David Fermanian et le Professeur Christian-Yann Robert, pour leur disponibilité et leur fine lecture qui m'a permis d'améliorer, de façon certaine, cette thèse. Je remercie Julie Josse, Michel Crouhy et Yannick Malevergne d'avoir accepté de faire partie du jury de cette thèse. Je voudrais également remercier les professeurs du pôle finance des laboratoires PRISM Sorbonne et LABEX ReFi pour leur soutien. Bien entendu, je remercie mes collègues et amis de Natixis, mais également tous les doctorants du PRISM pour leur conseil, leur aide et surtout leur bonne humeur. Je tiens aussi à remercier Abida Saidyassine, la responsable administrative et financière du PRISM, pour son implication et son aide si précieuse. Pour finir, je remercie ma famille, mon compagnon et mes amis pour avoir su m'apporter tout le soutien et l'amour dont j'avais besoin pour aller jusqu'au bout de mes objectifs.

Contents

List of Abbreviations

AUC Area under the curve

ARMSPE Average of the root mean squared prediction errors

BCBS Basel Committee on Banking Supervision

BICS Bloomberg industry classification standard

BRL Brazil

CDS Credit default swap

CHN China

CRR Capital Requirement Regulation

CVA Credit valuation adjustment

DAC Distributional accuracy for classification

DDC DetectDeviatingCells

DINEOF Data interpolation with empirical orthogonal functions

DTCC Depository Trust & Clearing Corporation

EBA European Banking Authority

ECB European Central Bank

EIA Energy Information Administration

EM Expectation-maximization

EMB Expectation-maximization with bootstrapping

EMIR European market infrastructure regulation

EONIA Euro overnight index average

ES Expected shortfall

ESMA European Securities Market Authority

ETF Exchange-Traded Fund

FRTB Fundamental review of the trading book

G-SIB Global systematically important banks

GAIN Generative adversarial imputation nets

GAN Generative adversarial networks

GICS Global industry classification standard

HEOM Heterogeneous Euclidean overlap metric

ICPCA Iterative classical principal component analysis

IFC Irving Fisher Committee

IMA Internal model approach

IND India

ISDA International swaps and derivatives association

JBF Journal of Banking and Finance

JF Journal of Finance

JFE Journal of Financial Economics

JFQA Journal of Financial and Quantitative Analysis

K***-NN*** K-nearest neighbors

LGD Loss given default

LOCF Last observation carried forward

MAE Mean absolute error

MAR Missing at random

MAAR Missing always at random

MACAR Missing always completely at random

MCAR Missing completely at random

MI Multiple imputation

MissPALasso Missingness pattern alternating imputation and l_1-penalty

MCD Minimum covariance determinant

MLP Multi-layer perception

MNAAR Missing not always at random

MNAR Missing not at random

MRD Mean relative deviation

MSCI Morgan Stanley capital international

MSE Mean squared error

MSSA Multivariate (or multichannel) singular spectral analysis

MICE Multiple imputation by chained equations

M-REM Multivariate regularized expectation maximization

NA Not available

NOCB Next observation carried backward

NRMSE Normalized root mean squared error

OLS Ordinary least square

OTC Over the counter

PAC Predictive accuracy in classification

PCA Principal component analysis

PCC Percent correctly classified

PFC Proportion of falsely classified

PMM Predictive mean matching

PP Projection pursuit

RAND South Africa

RFS Review of financial studies

RMSE Root mean squared error

ROBPCA Robust principal component analysis

RUB Russia

SFTR Securities financing transactions regulation

SIB Systemically important banks

SOM Self organization maps

S&P Standard & Poor's 500

SSA Singular spectral analysis

SVD Singular value decomposition

TRF Total return futures

TRIM Targeted review of internal models

TRS Total return swap

VaR Value-at-risk

WTI West Texas Intermediate

List of Figures

List of Tables

List of Empirical Research of the Thesis

Synthèse

Les problématiques liées aux données manquantes concernent aussi bien la sphère académique que la sphère professionnelle. En effet, de nombreuses études scientifiques traitent des données manquantes, de leur cadre théorique, de leurs impacts, ainsi que de leur gestion, et de plus en plus d'articles les mettent en lien avec des problématiques très pratiques. Les données manquantes sont présentes partout, en commençant par les jours de non-cotation. French [85] a montré en 1980 l'existence d'un effet week-end, et notamment que des rendements négatifs étaient attendus le lundi. Ainsi, un rendement entre un vendredi et le lundi suivant ne devrait pas être comparé à n'importe quel autre rendement de la semaine. De plus, Liu et An [147] montrent que les périodes de non-cotation (et notamment les week-ends) ont un impact sur les mesures de risque. C'est pourquoi certains font le choix de retirer des données de leurs analyses afin que l'effet week-end ne biaise pas leurs résultats (Giot et Laurent [93]). Concernant les banques, les données manquantes sont devenues un véritable enjeu réglementaire suite à une multitude de nouvelles réglementations traitant, directement ou indirectement, de la qualité de la donnée. C'est donc sur cette base que les banques se sont intéressées aux méthodes de complétion. Mais encore faut-il être en mesure de savoir quelle méthode utiliser, parmi de nombreux algorithmes de complétion pouvant aller de l'approche heuristique à l'algorithme mathématique. C'est d'ailleurs la problématique que pose Kahneman [62], où il oppose les intuitions d'experts aux formules mathématiques. Il explique que ces dernières sont plus fiables car moins sujettes à des stimuli extérieurs. Klein [132] quant à lui contredit cette approche, car selon lui, Kahneman [62] base ses conclusions sur de "faux experts", alors que les vrais experts sont fiables et leurs intuitions font qu'ils le sont encore plus que les formules mathématiques. Il faut alors être capable de différencier le vrai du faux expert. De la même façon, certains algorithmes de complétion peuvent avoir recours à une approche heuristique, alors que d'autres sont basés sur des formules mathématiques bien précises. Or, ces dernières apparaissent comme plus fiables aux yeux du régulateur, et sont donc préférées par les banques.

Cette thèse consiste donc à présenter le contexte et les problématiques de données manquantes en finance, avant d'introduire leur cadre théorique, ainsi qu'un certain nombre de méthodes de complétion. Enfin, ces dernières seront appliquées sur des échantillons de données simulées et historiques, afin de les comparer et de discuter des résultats.

Dans la littérature, les problématiques de données manquantes sont souvent traitées conjointement à celles des outliers. Ceci est dû au fait, qu'en réalité, les données manquantes ne représentent qu'un aspect d'une thématique bien plus vaste : la qualité

de la donnée. En effet, la qualité de la donnée concerne non seulement les données manquantes, mais également la source (pouvant être multiple) de la donnée, son accessibilité ou encore la présence de valeurs aberrantes. Naturellement, chacun de ces aspects implique des problématiques qui lui sont propres. Or, cette thèse n'a pas pour vocation d'étudier les impacts liés à la qualité de la donnée d'une façon générale, mais plutôt de mettre l'accent sur ceux liés à la présence de données manquantes dans des données financières.

Bien entendu, les problématiques de données manquantes ne sont pas propres aux données financières, mais concernent tous les domaines de recherche dès lors qu'une dimension quantitative est utilisée. C'est pourquoi cette thèse est construite autour de nombreux articles provenant de domaines de recherche très différents : océanologie, épidémiologie, cancérologie, imagerie, etc. D'où l'intérêt d'introduire, dans un premier temps, les impacts liés aux données manquantes d'une façon générale, avant d'orienter l'analyse vers les données financières. Ainsi, les impacts relatifs aux données manquantes sont présentés à travers, entre autres, l'étude de Verma et Goddale [207] qui ont montré que les données manquantes avaient un impact négatif sur la puissance statistique (via la taille de l'échantillon), mais également à travers l'étude de Roth, Switzer et Switzer [172] qui ont, quant à eux, mis en avant le fait qu'elles pouvaient biaiser les estimateurs et donc les analyses.

Ces effets négatifs concernent bien évidemment la Gestion et de la Finance. Tsikriktsis [201] explique, en 2005, que 33% des articles en Gestion recensés dans son étude (103 au total) font mention de données manquantes, avec une proportion moyenne de 13%. Le même type d'étude a été fait par Kofman et Sharpe [133] en 2003, sur une base de 946 articles publiés exclusivement dans des revues financières. Parmi ces articles, 27% d'entre eux font mention de données manquantes dont la proportion moyenne est de 23.3%. Ces études montrent donc une présence indéniable de données manquantes dans les articles de recherche, mais la sphère académique n'est pas la seule impactée.

Les répercussions des données manquantes peuvent, en effet, aller bien au-delà des analyses et études statistiques, et remettre en cause la stabilité du système financier dans son ensemble. Les données manquantes peuvent conduire les banques à, volontairement ou involontairement, sous-estimer l'évaluation de leurs expositions aux risques, ce qui pourrait avoir des conséquences désastreuses en cas de chocs importants sur les marchés. Dès lors, les données manquantes deviennent un véritable enjeu réglementaire, c'est pourquoi, depuis plusieurs années maintenant, de nombreuses réglementations ont été mises en place afin de garantir, directement ou indirectement, une bonne évaluation du risque.

Si certaines réglementations ne traitent de la qualité de la donnée que dans quelques principes ou paragraphes, d'autres y sont, au contraire, totalement dédiées. C'est notamment le cas de la réglementation BCBS 239 [24] qui, à travers ses quatorze principes, vise à renforcer la gestion et la transparence des données de risque. Les onze premiers

principes, destinés aux banques, consistent à renforcer la gouvernance et la gestion des données de risque, mais incitent également les banques à mettre en place les moyens informatiques suffisants afin de garantir la précision et l'intégrité des données. Quant aux trois derniers principes, adressés directement aux superviseurs, ils mettent l'accent sur le contrôle mais également sur la mise en place de mesures correctives et prudentielles en cas d'effort insuffisant de la part des banques. Cette réglementation exige ainsi que les banques mettent en œuvre les moyens suffisants afin de garantir des données de risque de qualité. Mais, avant même d'énumérer l'ensemble de ces principes, la réglementation BCBS 239 [24] mentionne explicitement les données manquantes, en indiquant qu'un avis d'expert est impératif pour leur gestion, bien que cela doive rester exceptionnel. Ainsi, le régulateur ne nie pas l'existence de données manquantes, et tolère leur gestion, sans pour autant donner davantage d'information sur la méthode à employer (imputation, interpolation, suppression, etc.).

Depuis sa publication, la date de mise en application de la réglementation BCBS 239 [24] n'a cessé d'être repoussée. Celle-ci était initialement prévue pour janvier 2016, mais les banques ont connu beaucoup de difficultés pour être conformes aux principes, et il aura fallu attendre 2020 pour que la majorité d'entre elles les respecte pleinement. Ainsi, ces difficultés de mise en application sont bien la preuve de réelles lacunes en matière de qualité de la donnée dans les banques.

En parallèle, les autorités réglementaires ont mis en place des référentiels centraux afin de lutter contre l'opacité des marchés mais également de prévenir et détecter leurs abus. Il s'agit d'entités visant à collecter et sauvegarder de façon centralisée des données de transaction et notamment celles issues des marchés de gré à gré. Petit à petit, les référentiels centraux sont devenus des outils de plus en plus présents dans les réglementations, afin d'assurer la transparence des marchés. Elles sont, en effet, utilisées dans de nombreux textes réglementaires, dont les principes pour les infrastructures de marchés financiers [58] en 2012, le règlement sur l'infrastructure du marché européen (EMIR) [168] publié la même année, le Dodd-Frank Act [203] en 2010, ou encore le règlement sur les opérations de financement sur titres (SFTR) [169] en 2015.

Ainsi, ces référentiels centraux sont devenus des acteurs essentiels aux réglementations actuelles, mais ont également permis de mettre en avant les nombreuses failles liées à la qualité de la donnée du secteur bancaire. En effet, l'Irving Fisher Committee (IFC) a mené une étude [121], publiée en 2018, révélant que plus de 40% des banques centrales interrogées considèrent la qualité de la donnée, transitant par ces référentiels centraux, comme étant moyenne voire médiocre. De plus, cette étude révèle que les banques centrales font face à des données manquantes, étant donné qu'au moins 25% d'entre elles déclarent corriger des observations incomplètes, mais également qu'au moins 20% d'entre elles les suppriment.

Les référentiels centraux constituent un pas supplémentaire vers la transparence des marchés, néanmoins ils révèlent de sérieuses failles concernant la qualité de la donnée,

et notamment la présence de données manquantes. De plus, ils révèlent également le recours à des méthodes et algorithmes de complétion de données manquantes au sein même des banques centrales.

Bien qu'aucune réglementation ne précise comment imputer les données manquantes, il existe néanmoins des problématiques comparables dans la réglementation relative aux exigences de fonds propres (CRR) [167]. Cette réglementation explicite le calcul de l'ajustement de valeur de crédit (CVA) à partir d'un certain nombre de paramètres et autorise notamment leur approximation à l'aide d'un proxy, dans le cas où ils seraient non observables. En effet, d'après le rapport sur la CVA de l'autorité bancaire européenne (EBA) [79], 10 banques sur 11 ont plus de 75% de leurs contreparties concernées par l'utilisation de données de remplacement. Ici, les données sont manquantes car elles ne sont pas observables, ce qui ne signifie pas qu'elles n'ont pas été enregistrées mais plutôt qu'elles n'existent pas. Certains produits sont, en effet, peu liquides et ne cotent pas nécessairement de façon quotidienne. Ainsi, le régulateur propose une méthodologie afin de pallier ces données manquantes, mais autorise le recours à des méthodologies alternatives. Or, le rapport de l'EBA [79] montre bien que ces méthodologies (propres à chaque banque) peuvent conduire à des résultats sous-optimaux, ce qui peut entrainer, par ailleurs, une mauvaise estimation des mesures de risque.

Ainsi, la réglementation autour du calcul de la CVA illustre bien la difficulté pour les banques de trouver la bonne donnée mais également de la modéliser. Le choix de cette dernière est fondamental étant donné qu'elle impacte directement la CVA mais également les mesures de risque qui en sont déduites.

La revue fondamentale du portefeuille de négociation (FRTB) [12] est une des nouvelles réglementations impliquant les plus gros changements concernant l'évaluation du risque. Elle est issue de nombreux documents consultatifs publiés entre 2012 et 2015 [8] [9] [10], et dont la première version a été publiée en 2016 [11], pour aboutir à une version finale en 2019 [12]. Cette nouvelle réglementation a pour but d'établir une frontière claire entre le "banking book" et le "trading book", d'améliorer le modèle interne mais également l'approche standard en la rendant notamment plus crédible.

L'un des enjeux de la réglementation FRTB [12] est l'éligibilité de la modélisation des facteurs de risque. En effet, pour qu'un facteur de risque soit modélisable il doit contenir suffisamment d'observations, sinon il est considéré comme non modélisable (NMRF) et est soumis à de lourdes charges en capital. L'existence même de ce type de test d'éligibilité est bien la preuve de la présence de données manquantes dans les séries financières. De plus, les difficultés pour rendre les facteurs de risque modélisables sont telles que les critères d'éligibilité ont dû être assouplis entre la première [11] et la dernière version [12] du FRTB. Malgré cela, un rapport de l'EBA [80] révèle qu'en moyenne 37% des facteurs de risque, des 8 banques interrogées, sont non modélisables et qu'au moins une de ces banques déclare la totalité de ses facteurs de risque comme étant non modélisables.

Le second enjeu en lien avec les données manquantes réside dans le calcul des mesures de risque stressées. Selon le FRTB [12], ces mesures de risque stressées doivent être calculées sur la base des 12 mois les plus sévères déterminés à partir d'un historique remontant à 2007. Ainsi, pour calculer de telles mesures de risque, la banque doit posséder des données historiques depuis 2007, ce qui rend sa mission encore plus complexe.

Enfin, en 2017 la Banque Centrale Européenne a lancé le projet intitulé "targeted review of internal model" (TRIM) [81], qui consiste à assurer la transparence dans l'interprétation des réglementations et leur application. Concernant la qualité de la donnée, ce guide met l'accent sur l'importance d'une documentation claire relative aux interventions manuelles qui peuvent être faites, mais également sur des processus de vérification de données et particulièrement si elles sont utilisées dans le modèle interne. La banque est ainsi rendue responsable de sa donnée mais également de sa qualité. Il est même stipulé qu'aucune donnée manquante ne doit être présente lors de la calibration des modèles. Le projet TRIM [81] permet donc d'éclaircir de nombreux éléments obscurs dans l'application pratique des réglementations sans pour autant donner plus de détail sur les méthodologies d'imputation acceptables du point de vue du régulateur.

Si toutes ces récentes réglementations s'adressent principalement aux banques, elles concernent également d'autres acteurs. Le premier acteur est naturellement le régulateur lui-même qui, mis à part l'élaboration et la mise à jour de ces nouvelles réglementations, se doit de contrôler les banques. C'est d'ailleurs l'objectif des trois derniers principes de BCBS 239 [24] qui s'adressent directement aux autorités de contrôle, ou encore à travers le FRTB [12] et TRIM [81], qui évoquent la surveillance et le rôle du régulateur. Ainsi, le rôle du régulateur consiste à répondre aux interrogations des banques à travers notamment de nombreux Q&A et FAQ, mais également à s'informer sur l'évolution de l'application de la réglementation et sa faisabilité, à l'aide d'exercices, sondages, études, etc. L'ensemble de ces documents publiés permet non seulement au régulateur de garder un oeil sur les banques, mais permet également aux banques de se tenir informées sur ce qui se fait dans les autres banques. Concernant les données manquantes, le régulateur a pour objectif de s'assurer que la banque utilise une méthode d'imputation qui reproduit les données les plus représentatives possibles, et non une méthode qui aurait tendance à sous-estimer son risque et donc à réduire ses charges en capital.

De plus, l'émergence de ces nouvelles réglementations a également un intérêt pour les fournisseurs de données. Bien qu'ils ne soient pas directement concernés par celles-ci, ces réglementations représentent un enjeu commercial pour eux. En effet, l'ensemble de ces réglementations met l'accent sur l'importance de la qualité de la donnée, ce qui est l'occasion rêvée pour les fournisseurs de données de vendre leurs historiques de données, mais également les mesures de risques associés ou encore l'imputation de bases de données. D'ailleurs, le fait que les fournisseurs de données vendent des méthodes

d'imputation interroge sur des possibles retraitements des données qu'ils vendent. Pour autant, aucune information publique n'est donnée à ce sujet.

Si les données manquantes ne représentent qu'une partie de la qualité de la donnée, elles soulèvent de nombreuses questions dans tous les domaines de recherche dès lors que des données historiques sont utilisées. Bien entendu, le monde de la finance est concerné par les problématiques de données manquantes, à tel point qu'elles sont devenues un véritable enjeu réglementaire pour les banques, mais également pour le régulateur lui-même et les fournisseurs de données.

Même si les données manquantes sont présentes dans de nombreux domaines de recherche, il faudra attendre la fin des années 80 pour qu'un véritable cadre théorique leur soit dédié. Il existe différentes classifications de données manquantes. Celles de Little et Rubin [145] ainsi que celle de Schafer et Graham [182] en sont les deux principales. Ils distinguent trois types de catégories de données manquantes : les données manquantes de façon complètement aléatoire (MCAR), les données manquantes de façon aléatoire (MAR) et les données manquantes de façon non-aléatoire (MNAR). D'après Little et Rubin [145] : la probabilité qu'une donnée soit MCAR ne dépend ni des données observables ni des données manquantes ; celle qu'une donnée soit MAR dépend des données observables ; enfin, la probabilité d'une donnée soit MNAR dépend des données observables et des données manquantes elles-mêmes. La catégorisation de Schafer et Graham [182] est proche de celle de Little et Rubin [145], mais intègre une variable extérieure, n'ayant aucun lien avec les données observables et les données manquantes, mais pouvant expliquer les données manquantes. D'autres auteurs, tels que Gelman et Hill [90] nuancent la catégorisation de Little et Rubin [143], néanmoins cette dernière reste la référence dans la littérature.

Suite à la catégorisation des données manquantes, des tests permettant de détecter si les données manquantes sont MCAR ou non ont été développés. C'est notamment le cas du test de Little [142] et celui de Jamshidian et Jalal [123]. Le test de Little [142] consiste en un test de ratio de vraisemblance avec l'hypothèse nulle que les données sont MCAR, alors que le test de Jamshidian et Jalal [123] combine deux tests (un test de Hawkin amélioré avec un test non-paramétrique) afin de tester l'hypothèse nulle que les données sont MCAR. Ces deux tests sont fréquemment utilisés dans la littérature, c'est pourquoi ils seront appliqués sur des échantillons de données simulées et de données historiques afin de commenter leurs résultats.

Little et Rubin [145] définissent également un certain nombre de distributions des données manquantes. Selon eux, elles peuvent, en effet, être distribuées de façon univariée ou, au contraire, multivariée, mais aussi de façon monotone, ou hasardeuse, etc.

Une fois le cadre théorique présenté, il semble pertinent de se recentrer sur un cadre financier relatif aux données manquantes. Ces différentes catégories de données

manquantes sont représentées dans des séries financières, or il est notamment important d'identifier les causes des données manquantes sur des séries financières. Ainsi, des données manquantes peuvent être dues : au calendrier boursier (week-ends et jours fériés), à des différences de calendrier d'une place boursière à une autre, des problèmes informatiques, à des migrations informatiques, à des suppressions volontaires (outliers), à des processus tardifs d'historisation des données, à des oublis de sauvegarde manuelle, au fait qu'elles soient inaccessibles ou invalides, à une entrée (ou sortie) en bourse, à une suspension de cotation, à l'illiquidité de certains marchés, ou encore à l'absence d'accord entre le prix de l'acheteur de celui du vendeur. Il est, en effet, important d'identifier la nature des données manquantes, car elles peuvent informer l'expert sur la façon dont elles doivent être complétées.

Traditionnellement, les données manquantes sont gérées par listwise deletion ou pairwise deletion. La listwise deletion consiste à supprimer toutes les observations (lignes) contenant au moins une donnée manquante, alors que la pairwise deletion consiste à ne supprimer que les observations inutilisables et d'utiliser toutes les autres observations possibles (méthode particulièrement utilisée lors de calculs de covariance par exemple). Kim et Curry [130] montrent d'ailleurs qu'utiliser une pairwise deletion est plus efficace qu'utiliser une listwise deletion (entre autres), dans un cadre de matrice de covariance et matrice de corrélation. Par ailleurs, les matrices pairwise ne sont pas toujours semi-définies positives, et certaines méthodologies (Rousseeuw et Molenberghs [173], Higham [110] ou encore Ledoit et Wolf [139]) peuvent être utilisées afin de la régulariser. Ces méthodes de régularisation sont, par ailleurs, fréquemment utilisées dans un cadre de gestion de portefeuilles.

Avant de présenter l'ensemble des méthodes qui seront utilisées pour l'analyse comparative, il est important de faire un point sur l'encodage des données, et notamment les nombreux choix qu'il implique. Le choix de la méthode de complétion est tout aussi important que le choix de l'échantillon lui-même. Il va dépendre du type de données, de la longueur de l'historique, des autres variables utilisées ou encore du format des données. Certaines méthodes de complétion se prêtent plus à une application sur des séries de prix, alors que d'autres se destinent davantage à des taux par exemple. Ainsi, distinguer le type de données à imputer influe nécessairement sur le choix de la méthode de complétion. Il en est de même pour la longueur de l'échantillon : un échantillon trop long peut noyer les algorithmes sous trop de données, alors qu'un échantillon trop court pourrait ne pas révéler assez d'information pour assurer leur bon fonctionnement. Les performances des méthodes de complétion exploitant la dimension transversale des données peuvent être particulièrement impactées par le choix des séries incluses dans l'échantillon. Il est possible d'effectuer cette sélection à partir de nombreux critères, pouvant être qualitatif, quantitatif ou les deux. Enfin, si pour certaines méthodes le format des données est intuitif, ce n'est pas toujours le cas. Certaines méthodes de complétion peuvent s'avérer plus efficaces sur des prix bruts, sur des rendements ou

encore sur des données standardisées.

C'est donc sur cette base que les différentes méthodes de complétion utilisées lors des analyses comparatives seront présentées. Tout d'abord, des méthodes très simples mais fréquemment utilisées en pratique par les banques seront intégrées dans l'analyse comparative. C'est le cas de l'interpolation linéaire ainsi que la méthode LOCF (last observation carried foward). L'interpolation linéaire consiste à relier linéairement la dernière observation à la prochaine, alors que la méthode LOCF consiste à figer la dernière observation disponible. Ces méthodes seront utilisées comme les benchmarks lors des analyses comparatives. Une autre méthode d'interpolation sera utilisée, il s'agit du pont Brownien, qui consiste à compléter les données manquantes en reliant des prix à l'aide d'un mouvement Brownien. Ces trois méthodes ont la caractéristique d'être unidimensionnelle, c'est-à-dire qu'elles n'utilisent pas la dimension transversale des données, mais simplement la série elle-même. En revanche, ce n'est pas le cas des autres méthodes de complétion. Sept autres méthodes, utilisant toutes la dimension transversale seront utilisées.

La première d'entre elles est la méthode des plus proches voisins, aussi appelée K-NN. Cette méthode d'imputation consiste à remplacer les données manquantes par une moyenne pondérée (inversement proportionnelle à la distance) des observations des séries les plus proches. Cette méthode est combinée avec du bagging (bootstrap averaging) afin d'améliorer les performances du modèle : l'échantillon initial est bootstrappé (tirage aléatoire avec remise) afin d'appliquer les K-NN sur chaque échantillon bootstrappé, avant de moyenner les imputations. Cette méthode nécessite un certain nombre de paramètres à fixer : la distance, le nombre de plus proches voisins et l'ordre d'imputation. Dans cette thèse, la méthode des K-NN utilise une distance Euclidienne, le nombre de plus proches voisins est défini par cross-validation (méthode des k-folds) et l'ordre d'imputation des séries est défini par ordre croissant de données manquantes (la première colonne à imputer est celle qui contient le moins de données manquantes).

Une autre méthode de l'analyse comparative est le MSSA (multivariate singular spectrum analysis). Cette méthode consiste à décomposer les séries afin d'isoler la tendance du bruit. La décomposition ne se fait pas directement sur la matrice des données, mais sur la matrice de trajectoires qui correspond à une matrice où chaque série est laggée de façon à obtenir une matrice d'Hankel. Par la suite, une décomposition en valeurs singulières est faite sur cette matrice de trajectoires, puis les données sont reconstruites en veillant à imputer les données manquantes. Cette méthode est fréquemment utilisée sur des données financières et est d'ailleurs la méthode d'imputation vendue par Bloomberg L.P. [65]. Il était donc important de l'intégrer à l'analyse comparative. De plus, cette méthode nécessite de fixer deux paramètres : la fenêtre (qui définit le lag de matrice de trajectoire) ainsi que le rang (le nombre de composantes principales utilisé pour la reconstruction). Ces paramètres sont fixés en fonction des recommandations de l'équipe de recherche de Bloomberg L.P. [65].

L'analyse comparative intègre également des imputations par forêts aléatoires. Cette méthode consiste à combiner des arbres de décision avec du bagging. Ainsi, l'échantillon initial est bootstrappé (tirage aléatoire avec remise) pour que des arbres de décision soient construits à partir de chacun de ces échantillons bootstrappés, avant de faire la moyenne des imputations obtenues. Cette méthode est très fréquemment utilisée sur des données médicales mais ses applications en finance sont plus rares. C'est pourquoi il est intéressant de l'intégrer dans cette thèse. En revanche, cette méthode implique des problèmes de non-réplicabilité des résultats dus à la composante aléatoire introduite par le bootstrap, qui pourrait ne pas être acceptable du point de vue du régulateur. En effet, cette méthode donne des résultats différents d'une exécution à l'autre, et la banque pourrait être tentée d'utiliser une exécution qui serait en sa faveur (lui permettant de minimiser ses charges en capital) c'est pourquoi le régulateur peut être méfiant et réticent à son égard.

Une version améliorée de l'algorithme EM (expectation-maximization) est également incluse dans l'analyse comparative. Il s'agit de l'algorithme Amelia, qui combine l'EM avec de l'imputation multiple. L'imputation multiple consiste à bootstrapper (tirage aléatoire avec remise) l'échantillon initial, d'appliquer sur chacun d'eux une méthode de complétion (ici l'algorithme EM), pour ensuite utiliser l'ensemble de ces échantillons imputés pour les analyses. Les analyses sont, en effet, calculées à partir de chacun des échantillons imputés, et peuvent ensuite être moyennées. Contrairement au bagging où les analyses sont faites sur la moyenne des imputations, ici les analyses sont faites sur chacun des échantillons imputés avant d'être moyennés. La moyennisation a donc lieu un temps plus tard. L'algorithme EM est très souvent appliqué à des données financières dans la littérature, d'où l'intérêt d'utiliser Amelia dans cette thèse. De plus, Amelia ne nécessite pas de paramétrage spécifique, sauf l'initialisation des paramètres de l'algorithme EM. Little et Rubin [145] traitent d'ailleurs de ce sujet dans leur ouvrage.

Une autre méthode d'imputation multiple est utilisée dans cette thèse, et il s'agit de l'algorithme MICE (multivariate imputation by chained equation). Même si cette méthode n'utilise pas de bootstrap, elle donne lieu à plusieurs échantillons imputés. L'algorithme MICE consiste, à partir d'un même échantillon de départ, à imputer les données manquantes par PMM (predictive mean matching), qui intègre plusieurs composantes aléatoires. L'idée de cette méthode est d'estimer la donnée manquante par régression linéaire, et de l'imputer en tirant aléatoirement parmi un certain nombre de valeurs observés proches (donneur) de l'estimation qui a été faite. Cette méthode implique donc faire un choix concernant le nombre de donneurs à utiliser : il ne doit pas être trop grand au risque de s'éloigner trop de l'estimation qui a été faite, mais ne doit pas être trop faible non plus, au risque d'imputer trop souvent la même valeur. De plus, cette procédure est itérative et utilise un échantillonnage de Gibbs pour l'imputation. Cette méthode est souvent utilisée dans la littérature, mais plus rarement sur des données financières.

Enfin, les deux dernières méthodes utilisées sont proches l'un de l'autre étant donné

que l'une est une version améliorée de l'autre. Il s'agit de l'IPCA (iterative principal component analysis) et de sa version en imputation multiple. L'IPCA consiste à décomposer l'échantillon initial à partir d'une analyse en composantes principales afin de le reconstruire en utilisant un nombre de composantes principales bien défini, et de remplacer les données manquantes par leur prédiction. De plus, cette procédure est réitérée plusieurs jusqu'à convergence des prédictions. Cette méthode a, par ailleurs, été implémentée au sein d'un algorithme visant à détecter les outliers et imputer les données manquantes (MacroPCA [115]). De plus, Josse et Husson [126] ont amélioré l'IPCA sous deux aspects : régularisation pour éviter les erreurs de sur-apprentissage et intégration dans un processus d'imputation multiple. Comme pour le MSSA, l'IPCA et le MIPCA nécessite de définir le nombre de composantes principales à utiliser pour la reconstruction. Ici, il est défini par une formule généralisée d'une méthode de cross-validation (la méthode one-leave-out), établie par Josse et Husson [126].

Ces méthodes ne sont qu'une sélection de ce qui se fait dans la littérature. Il en existe bien d'autres, comme des méthodes d'imputation ayant recours à un modèle autorégressif qui permettent d'exploiter la dimension temporelle des données [146][218], ou encore, des méthodes basées sur les réseaux antagonistes génératifs (GAN), où la recherche est particulièrement active ces dernières années [96][215][216][213]. Ces dernières sont d'ailleurs de plus en plus utilisées sur des données financières [107][134][49][50][72].

Après avoir présenté l'ensemble des méthodes qui seront utilisées, il est temps de les mettre à l'épreuve et de les confronter les unes aux autres. Ces méthodes de complétion seront utilisées dans un premier temps sur un échantillon simulé auquel des données auront été supprimées en fonction de différents mécanismes de données manquantes, avant de les appliquer sur des échantillons historiques avec les mécanismes de données manquantes les plus pertinents. Ce type de procédure va permettre la comparaison de toutes ces méthodes de complétion à partir de différents critères d'analyse. Les comparaisons seront faites en termes de moments statistiques, de mesures de proximité (MAE et RMSE), de différence de matrice de covariance et de mesure de risque (VaR et ES).

Ainsi, la première analyse consiste à comparer les méthodes de complétion sur un échantillon simulée contenant des données qui auront été supprimées de l'échantillon de façon complètement aléatoire (MCAR), c'est-à-dire en fonction d'une loi uniforme. Les données sont supprimées dans la première colonne de l'échantillon simulé, et avec des proportions allant de 5% à 70%.

La deuxième analyse est très similaire à la première, étant donné que le même mécanisme de données manquantes est utilisé (MCAR) mais cette fois sur la quasi-totalité de la matrice (toutes les colonnes sauf la dernière). Les proportions de données manquantes par colonnes vont de 5% à 70%. Ce scénario est plus réaliste que le précédent,

car en pratique les données sont manquantes dans plusieurs colonnes de l'échantillon et rarement dans une seule.

L'analyse suivante consiste à appliquer les méthodes d'imputation sur un échantillon hétéroscédastique, c'est-à-dire sur un échantillon à volatilité non-constante. Ainsi, l'analyse est, cette fois, faite à partir de plusieurs échantillons simulés contenant une période de crise de plus en plus violente (avec une volatilité de plus en plus forte), afin de voir ou non la dégradation de la qualité d'imputation des algorithmes en fonction de la déformation de la série. Un mécanisme de données MCAR (avec une proportion fixée à 30%) est alors appliqué dans la première colonne de chacun des échantillons, pour rendre la comparaison possible.

La quatrième analyse vise à observer l'impact de la présence de sauts dans la série sur la qualité d'imputation. Ainsi, à partir du même tirage de données, plusieurs échantillons contenant de plus en plus de saut ont été simulés. Des données MCAR ont été supprimées (avec une proportion fixée à 30%) dans la première colonne de chacun de ces échantillons, avant d'être imputées par l'ensemble des méthodes. Ceci a pour intérêt de mettre en évidence les méthodes sensibles à la présence de sauts dans la série.

La cinquième analyse comparative reprend le même échantillon simulé que pour les deux premières analyses, mais supprime les données de façon à suivre un mécanisme MAR. Pour cela, les données sont supprimées en fonction des valeurs de la dernière colonne de la matrice. Plus précisément, les valeurs extrêmes de la dernière colonne sont identifiées, et les observations correspondantes sont supprimées dans la première colonne. Les proportions de ce mécanisme de données manquantes vont de 5% à 70%.

Un autre mécanisme de données manquantes est appliqué à l'échantillon simulé et consiste à supprimer des données successivement au milieu de la première colonne. Les proportions de données manquantes associées à ce mécanisme sont les mêmes que précédemment, à savoir 5% à 70%. Ceci permet de comparer les méthodes face à un grand nombre de données manquantes successivement.

Le mécanisme suivant est très proche du précédent, étant donné qu'il consiste à supprimer les données successivement également, mais cette fois à la fin de la série (toujours avec des proportions allant de 5% à 70%). Ainsi, aucune autre observation n'est disponible après.

Enfin, le dernier mécanisme appliqué à l'échantillon simulé consiste à supprimer les valeurs extrêmes de la série, avec des proportions allant de 5% à 70%. Ce mécanisme s'attaque directement à la distribution de la série, ce qui risque de mettre en difficulté un certain nombre de méthodes de complétion.

Après avoir effectué toutes ces études comparatives sur des échantillons simulés, il convient d'en mener quelques-unes sur des données historiques. Pour cela les données de l'Euro Stoxx 300 sont utilisées afin de former deux échantillons historiques sur la période de Janvier 2020 à Février 2021. Un premier échantillon basé sur une approche heuristique (actions d'un même secteur et fortement corrélées) est formé, puis un se-

cond basé sur une approche de graphical Lasso [86]. Des données MCAR (basées sur une loi uniforme) ont été supprimées de la première colonne de ces deux échantillons historiques, puis un second mécanisme de données manquantes est appliqué, consistant à supprimer successivement des données à la fin de la première colonne. Le premier mécanisme a été appliqué car il s'agit d'un cas fréquemment étudié dans la littérature, et le second car il est fréquemment observable sur des données financières (c'est typiquement le cas avant l'introduction d'une action en bourse).

Ainsi, après avoir présenté et analysé l'ensemble des résultats obtenus pour chaque analyse comparative, quelques conclusions sont discutées. Tout d'abord, les résultats peuvent être différents d'une analyse comparative à une autre, et il est clair qu'ils dépendent de l'échantillon auquel ils sont appliqués.

Ensuite, certaines méthodes impliquent une non réplicabilité des résultats, ce qui peut être très problématique du point de vue du régulateur. La banque peut, en effet, être tentée d'utiliser un scénario qui lui est favorable afin de réduire ses charges en capital. Il serait possible de fixer une "seed" pour résoudre ce type de problème, mais encore faut-il qu'elle soit représentative. Une autre conclusion est que le choix de la méthode d'imputation doit être fait en fonction de la nature des données, mais surtout de leur utilisation future. Il est certain qu'imputer de façon à obtenir une valeur proche de la valeur manquante sera efficace d'un point de vue des mesures de proximité, mais sans doute beaucoup moins en termes de mesures de risque. Si la méthode des forêts aléatoires apparait comme efficace pour tous les critères et toutes les proportions de données manquantes, ce n'est pas le cas des autres méthodes. L'algorithme des K-NN est efficace en termes de mesures de proximité, mais pas en termes de mesures de risque. Les méthodes Amelia et MIPCA obtiennent également de très bons résultats pour de nombreux critères. En revanche, Amelia n'est pas capable de gérer une trop forte proportion de données manquantes, et ça, peu importe le critère d'analyse.

Un des enseignements de cette thèse est également de relever qu'une méthode n'est pas un algorithme et que parler d'une méthode seule n'a pas grande valeur si toutes les étapes de pré-processing, et autres, ne sont pas explicitement présentées. C'est notamment le cas des forêts aléatoires, dont deux implémentations sont couramment utilisées dans la littérature, mais qui ne donnent pas les mêmes résultats.

Ceci met d'ailleurs l'accent sur le point suivant : le risque opérationnel engendré par les méthodes d'imputation. Une méthode d'imputation peut, dans certains cas, ne pas fonctionner (pas assez de données observables), ce qui peut être catastrophique pour une banque. De plus, les différences d'interprétation des algorithmes peuvent conduire à des implémentations différentes, d'où l'importance d'une documentation claire, transparente et précise les concernant.

Concernant les méthodes en particulier, il a été observer que l'algorithme Amelia était particulièrement sensible à la proportion de données manquantes. Au-delà de 50% de données manquantes, la méthode d'imputation a tendance à perdre en qualité d'im-

putation. Ceci peut notamment s'expliquer par l'initialisation des paramètres qui est basée sur les données observables. Or, lorsque la proportion de données manquantes devient trop importante, ces paramètres peuvent s'éloigner fortement de ceux de la série complète.

De plus, certains résultats paradoxaux sont observé : les données imputées dans la première colonne sont de meilleure qualité lorsque toutes les colonnes contiennent des données manquantes par rapport au cas où le reste de l'échantillon est complet. Ceci est particulièrement visible en termes de mesures de proximité, pour les méthodes IPCA, MIPCA, MICE et Amelia. Après investigation, ceci n'est pas dû à une meilleure estimation en loi, et s'observe dès deux colonnes contenant des données manquantes. Ces méthodes sont toutes itératives, et il semblerait que leur procédure itérative passe par un chemin de convergence différent (lorsque plusieurs colonnes contiennent des données manquantes) qui les conduisent, pour une raison inconnue, à de meilleurs résultats.

Enfin, Amelia et les forêts aléatoires apparaissent comme les deux bons élèves de l'ensemble de ces analyses comparatives. Pour autant, les forêts aléatoires ont tendance à être plus performantes qu'Amelia. Ceci s'explique en partie par le bagging et notamment la moyennisation des imputations des forêts aléatoires, alors qu'Amelia implique de calculer chaque critère sur chaque échantillon imputé, avant de moyenner ces critères. Si les échantillons imputés sont moyennés avant d'y appliquer les critères d'analyse, les performances d'Amelia sont généralement améliorées. De plus, une autre partie de ces différences peut s'expliquer par la taille de l'échantillon. Modarresi et Diner [156] ont effectivement montré que les performances des algorithmes étaient variables en fonction de la taille de l'échantillon. Dans leur étude, ils montrent notamment qu'Amelia est plus efficace que les forêts aléatoires pour une certaine taille d'échantillon, et inversement pour une autre taille. Ce qui renvoie au fait que les performances des algorithmes dépendent des données elles-mêmes.

Ainsi, après avoir présenté le contexte et les différents enjeux liés aux données manquantes en finance, de nombreuses méthodes de complétion ont été appliquées sur des échantillons (simulés et historiques) afin de les comparer. Les analyses comparatives ont permis de mettre en avant un certain nombre de points et notamment des performances remarquables pour l'algorithme des forêts aléatoires, mais également pour Amelia. D'autres méthodes, comme MICE, obtiennent des résultats mitigés et nécessitent d'être appliquées sur un plus grand nombre d'échantillons avant de pouvoir tirer des conclusions. Enfin, certaines méthodes apparaissent comme des mauvais élèves pour la quasi-totalité des analyses, comme la méthode LOCF. Pour autant, cette méthode est intéressante du point de vue du régulateur car elle est la plus conservatrice.

Des pistes d'amélioration sont évoquées tout au long de la thèse, et notamment celle visant à refaire l'analyse (à partir des méthodes les plus pertinentes) sur un grand nombre d'échantillons de données afin d'être en mesure de généraliser les résultats.

Extended Summary

Issues related to missing data affect both the academic and professional spheres. Indeed, many scientific studies deal with the question of missing data, including theoretical frameworks, impacts and management strategies, and an increasing number of articles link it to very practical issues. Missing data is everywhere, starting with non-trading days. French [85] showed in 1980 the existence of a "weekend effect", specifically that negative yields were expected on Mondays. Thus, the yield between a Friday and the following Monday should not be compared to any other yield for the week. In addition, Liu and An [147] show that non-trading periods (and particularly weekends) have an impact on risk measures. This is why some people choose to remove data from their analyses so that the weekend effect does not skew their results (Giot and Laurent [93]). With regard to banks, missing data has become a real regulatory issue following a multitude of new regulations which directly or indirectly address data quality. This has led banks to take an interest in completion methods. The issue remains of deciding which method to use, from among many approaches to completion that can range from the heuristic to the mathematical. This is the issue raised by Kahneman [62], when he contrasts the intuition of experts with mathematical formulas. He explains that the latter are more reliable by virtue of being less susceptible to external stimuli. Klein [132], on the other hand, contradicts this approach, because according to him, Kahneman [62] bases his conclusions on "fake experts", whereas real experts can be trusted and their intuition makes them even more reliable than mathematical formulas. This raises the question of differentiating real experts from fake experts. Likewise, some completion algorithms may use a heuristic approach, while others are based on precise mathematical formulas. However, the latter are considered more reliable in the eyes of the regulator, and are therefore preferred by banks.

This PhD thesis therefore consists of presenting the context of and problems related to missing data in finance, before introducing their theoretical framework and describing certain completion methods. Finally, the latter will be applied to simulated and historical data samples, in order to compare them and discuss the results.

In the literature, problems related to missing data are often dealt with in conjunction with those related to outliers. This is because, in reality, missing data is just one aspect of a much broader theme: data quality. Data quality is determined not only by any missing data, but also by the source (which may be multiple) of the data, its accessibility, and the presence of any outliers. Naturally, each of these aspects creates its own set of problems. However, this PhD thesis does not seek to study the impact of

data quality in general, but rather to focus on issues linked specifically to missing data in financial data sets.

Of course, issues caused by missing data are not unique to financial data sets, and in fact concern all areas of research with a quantitative dimension. As such, this PhD thesis is built around numerous articles from diverse fields of research: oceanography, epidemiology, oncology, imaging, etc. The impacts of missing data are introduced in a general way, before focusing the analysis on financial data. The impacts of missing data are presented through, among others, a study by Verma and Goddale [207] which showed that missing data had a negative impact on statistical power (via the sample size), as well as through a study by Roth, Switzer and Switzer [172] which highlighted the fact that missing data could skew the estimators and therefore the analyses. These negative effects obviously concern Management and Finance. Tsikriktsis [201] explains, in a 2005 paper, that 33% of the management articles identified in his study (103 in total) mention missing data, on average amounting to 13% of the data set. The same type of study was done by Kofman and Sharpe [133] in 2003, based on 946 articles published exclusively in financial journals. Of these articles, 27% mention missing data, amounting to an average proportion of 23.3%. These studies show the undeniable presence of missing data in research papers, but its effects also exist outside the academic sphere.

Indeed, the impact of missing data can go far beyond statistical analyses and studies and undermine the stability of the financial system as a whole. Missing data may lead banks to intentionally or unintentionally underestimate the assessment of their risk exposures, which could have disastrous consequences in the event of significant market shocks. Missing data thus becomes a real regulatory issue, which is why, for several years now, many regulations have been put in place to guarantee, directly or indirectly, accurate risk assessments.

While some regulations deal with data quality only in a few principles or paragraphs, others, on the contrary, are entirely dedicated to it. This is specifically the case for regulation BCBS 239 [24] which, through its fourteen principles, aims to strengthen the management and transparency of at-risk data. The first eleven principles, intended for banks, consist of strengthening the governance and management of at-risk data, and also encourage banks to put in place sufficient IT resources to guarantee the accuracy and integrity of data. As for the last three principles, addressed directly to supervisors, they focus on monitoring as well as on the implementation of corrective and prudential measures in the event of insufficient efforts on the part of banks. This regulation thus requires that banks implement sufficient means to guarantee the quality of at-risk data. However, even before listing all of these principles, BCBS 239 [24] explicitly mentions missing data, stating that expert advice is required for its management, although this should remain an exceptional circumstance. The regulator thus does not deny the existence of missing data, and tolerates its management, without, however, providing further information on the method to be used (imputation, interpolation, deletion, etc.).

Since its publication, the date of application of BCBS 239 [24] has been repeatedly postponed. It was initially planned for January 2016, but banks have experienced many difficulties in complying with the principles, and it took until 2020 for the majority of them to fully comply. Indeed, these implementation difficulties are evidence of real data quality deficiencies at banks.

At the same time, the regulatory authorities have set up central repositories to combat the opacity of the markets, and to prevent and detect abuses. These are entities designed to collect and centrally back up transaction data, specifically data from over-the-counter markets. Gradually, the central repositories have become increasingly present in the regulations, in order to ensure market transparency. They have been used in many regulatory texts, including the principles for financial market infrastructure [58] in 2012, the European Market Infrastructure Regulation (EMIR) [168] published the same year, the Dodd-Frank Act [203] in 2010, and the Securities Financing Transactions Regulation (SFTR) [169] in 2015.

Thus, these central repositories have become essential players in the current regulations, and have also made it possible to highlight the numerous issues linked to the quality of banking sector data. The Irving Fisher Committee (IFC) conducted a study [121], published in 2018, revealing that more than 40% of the central banks surveyed consider the quality of the data passing through these central repositories to be average or even mediocre. Moreover, this study reveals that central banks are dealing with missing data, given that at least 25% of them report correcting incomplete observations, but also that at least 20% of them have deleted data.

The central repositories are another step towards market transparency, but they reveal serious flaws in data quality, notably the presence of missing data. Moreover, they also reveal the use of completion methods and algorithms to deal with missing data within the central banks themselves.

Although no regulation specifies how to allocate missing data, there exist comparable issues in the regulation around requirements for cash reserve ratios [167]. This regulation explains the calculation of the credit value adjustment (CVA) based on certain parameters and specifically authorizes their approximation using a proxy, in the event that they are unobservable. Indeed, according to the report on the CVA of the European Banking Authority (EBA) [79], at 10 out of 11 banks more than 75% of counterparties are affected by the use of replacement data. Here, the data are missing because they are not observable, which does not mean that they have not been recorded but rather that they do not exist. Some products are not very liquid and are not necessarily costed on a daily basis. As such, the regulator suggests a methodology to mitigate these missing data, but authorizes the use of alternative methodologies. However, the EBA report [79] clearly shows that these methodologies (specific to each bank) can lead to sub-optimal results, which may also lead to a poor estimates of risk measures. The regulations around the calculation of the CVA clearly illustrate the difficulty banks face

in both finding the right data and modelling it. The choice of the latter is fundamental, given that it directly impacts the CVA as well as the risk measures that are deduced from it.

The fundamental review of the trading book (FRTB) [12] is one of the new regulations involving the biggest changes to risk assessment. It is the result of numerous advisory documents published between 2012 and 2015 [8] [9] [10], the first version of which was published in 2016 [11], leading to a final version in 2019 [12]. The aim of this new regulation is to establish a clear boundary between the banking book and the trading book, to improve both the internal model and the standard approach by making it more credible.

One of the challenges of the FRTB regulation [12] is the eligibility of risk factor modelling. For a risk factor to be modelled it must contain sufficient observations, otherwise it is considered non-modelled (NMRF) and is subject to heavy capital charges. The very existence of this type of eligibility test is proof of the presence of missing data in financial series. Furthermore, the difficulties involved in making risk factors modellable are such that the eligibility criteria had to be relaxed between the first [11] and the latest version [12] of the FRTB. Despite this, an EBA report [80] reveals that on average 37% of the risk factors at the 8 banks surveyed are non-modellable and that at least one of these banks declares all its risk factors as non-modellable.

The second issue with missing data is the calculation of stressed risk measures. According to the FRTB [12], these stressed risk measures must be calculated on the basis of the most severe 12 months, chosen from a history dating back to 2007. Thus, to calculate such risk measures, the bank must have historical data from at least 2007, making the mission even more complex.

Finally, in 2017 the European Central Bank launched a project entitled "targeted review of internal model" (TRIM) [81], which consists of ensuring transparency in the interpretation of regulations and their application. With regard to data quality, this guide emphasizes the importance of clear documentation relating to possible manual interventions, as well as data verification processes, particularly if these are used in the internal model. The bank is thus made responsible for both the data and its quality. It is even stipulated that no missing data should be present during calibration of the model. The TRIM project [81] therefore makes it possible to clarify many obscure elements in the practical application of the regulations, without giving more detail on the imputation methodologies which are acceptable from the point of view of the regulator.

While these recent regulations are mainly aimed at banks, they also concern other players. The first player is, of course, the regulator itself, which, apart from drawing up and updating these new regulations, must monitor the banks. This is the aim of the last three principles of BCBS 239 [24] which are addressed directly to the supervisory

authorities, or through the FRTB [12] and TRIM [81], which refer to supervision and the role of the regulator. The role of the regulator is thus to respond to banks' questions through numerous Q&As and FAQs, and also to learn about changes to the application of the regulations and their feasibility, using exercises, surveys, studies, etc. All these published documents not only allow the regulator to keep an eye on the banks, but also allow banks to stay informed about what is being done at other banks. With regard to missing data, the regulator aims to ensure that the bank uses an imputation method that reproduces the most representative data possible, and not a method that would tend to underestimate its risk and therefore reduce its capital costs.

Moreover, the emergence of these new regulations is also of interest for data providers. Although the latter are not directly concerned by them, in their eyes these regulations do represent a commercial challenge. All of these regulations emphasize the importance of data quality, which is the perfect opportunity for data providers to sell their data histories, as well as the the associated risk measures and the imputation of databases. Moreover, the fact that data providers sell imputation methods raises questions about possible re-processing of the data they sell. However, no public information is given on this subject.

Although missing data is only one element of data quality, it raises many questions in all research areas where historical data is used. Of course, the financial world is affected by the problems of missing data, to the point that it has become a real regulatory issue for banks, as well as for the financial regulator and data providers.

Despite the fact that missing data are present in many areas of research, it was not until the end of the 1980s that a real theoretical framework was dedicated to them. There are different classification systems for missing data. The most commonly used are those of Little and Rubin [145] and Schafer and Graham [182]. They distinguish three types of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). According to Little and Rubin [145] the probability of a data point being MCAR does not depend on observed data or unobserved data, that of a data being MAR depends on observed data, and finally, the probability of a point data being MNAR depends on both observed data and the missing data itself. Schafer and Graham's categorization [182] is similar to that of Little and Rubin [145], but includes an external variable, unrelated to the observed data and missing data, but which can explain missing data. Other authors, such as Gelman and Hill [90] add nuance to the categorization of Little and Rubin [145], but the latter remains the standard in the literature.

Following the categorization of missing data, tests to detect whether the missing data is MCAR or not have been developed. These include notably the Little test [142] and that of Jamshidian and Jalal [123]. Little's test [142] consists of a likelihood ratio test with the null hypothesis that the data are MCAR, while the Jamshidian and Jalal's

test [123] combines two tests (an improved Hawkin test with a non-parametric test) in order to test the null hypothesis that the data are MCAR. These two tests are frequently used in the literature, which is why they will be applied to the simulated data samples and historical data to comment on their results.

Little and Rubin [145] also define a number of distributions of missing data. According to them, these can be distributed in a univariate or multivariate manner, but also in a monotonous or random manner, etc.

Once the theoretical framework has been presented, it is appropriate to refocus on a financial context related to missing data. These different categories of missing data are represented in financial series, but it is particularly important to identify the causes of missing data in financial series. Missing data may be due to: the stock market calendar (weekends and public holidays), differences in the calendar from one stock market to another, IT problems, IT migrations, voluntary deletions (outliers), late data logging processes, manual backup omissions, the fact that they are inaccessible or invalid, an entry (or exit) on the stock market, a suspension of listing, the illiquidity of certain markets, or the lack of agreement between the price of the buyer and that of the seller. It is indeed important to identify the nature of the missing data, as this can inform the expert as to how they should be completed.

Traditionally, missing data is managed by listwise deletion or pairwise deletion. Listwise deletion consists in deleting all observations (lines) containing at least one piece of missing data, whereas pairwise deletion consists in deleting only unusable observations and using all other possible observations (a method used specifically in covariance calculations, for example). Kim and Curry [130] also show that using pairwise deletion is more effective than using listwise deletion (among others), within a framework of covariance matrix and correlation matrix. Furthermore, pairwise matrices are not always positive semidefinite, and certain methodologies (Rousseeuw and Molenberghs [173], Higham [110] or Ledoit and Wolf [139]) can be used to regularize them. These regularization methods are also frequently used in portfolio management.

Before presenting all the methods that will be used for the comparative analysis, it is important to look at data encoding, and in particular the many choices that it involves. The choice of a completion method is just as important as the choice of the sample itself. This will depend on the type of data, the length of the history, the other variables used as well as the format of the data. Some completion methods are more suitable for application to price series, while others are more suitable for rates, for example. Thus, distinguishing the type of data to be attributed necessarily influences the choice of the completion method. The same applies to the length of the sample: a sample that is too long can drown the algorithms in too much data, while a sample that is too short may not reveal enough information to ensure their proper functioning. The application of completion methods using the cross-functional dimension of the data can be particularly impacted by the choice of series included in the sample. It is possible to

make this selection based on numerous criteria, which may be qualitative, quantitative or both. Finally, while for some methods the data format is intuitive, this is not always the case. Certain completion methods can be more effective for gross prices, yields or standardized data.

It is therefore on this basis that the different completion methods used in the comparative analyses will be presented. Firstly, very simple methods that are frequently used in practice by banks will be included in the comparative analysis. This is the case for linear interpolation as well as the LOCF (last observation carried forward) method. Linear interpolation consists in linearly linking the last observation to the next one, whereas the LOCF method consists in freezing the last available observation. These methods will be used as benchmarks during comparative analyses. Another method of interpolation will be used, namely the Brownian bridge, which consists of completing missing data by linking prices using Brownian motion. These three methods have the characteristic of being one-dimensional, i.e. they do not use the transverse dimension of the data, but simply the series itself. However, this is not the case for other completion methods. Seven other methods, all using the transversal dimension, will be used.

The first is the "closest neighbor method" also called K-NN. This imputation method consists in replacing the missing data with a weighted average (inversely proportional to the distance) of the observations of the closest series. This method is combined with bagging (bootstrap averaging) in order to improve model performance: the initial sample is bootstrapped (random draw with replacement) in order to apply the K-NN to each bootstrapped sample, before averaging the imputations. This method requires a certain number of parameters to be set: distance, number of closest neighbors and order of imputation. In this PhD thesis, the K-NN method uses a Euclidean distance, the number of closest neighbors is defined by cross-validation (k-folds method) and the order of imputation of the series is defined in increasing order of missing data (the first column to be imputed is that which contains the least missing data).

Another method of comparative analysis is MSSA (multivariate singular spectrum analysis). This method consists in breaking down the series in order to isolate the noise trend. The breakdown is not done directly on the data matrix, but on the trajectory matrix which corresponds to a matrix where each series is lagged so as to obtain a Hankel matrix. Subsequently, a breakdown into singular values is completed on this trajectory matrix, then the data is reconstructed while making sure to attribute the missing data. This method is frequently used for financial data and is also the allocation method sold by Bloomberg L.P. [65]. It was therefore important to include it in the comparative analysis. In addition, this method requires setting two parameters: the window (which defines the trajectory matrix lag) as well as the rank (the number of main components used for reconstruction). These parameters are set according to the recommendations of the Bloomberg L.P. research team. [65].

The comparative analysis also includes allocations using random forests. This

method involves combining decision trees with bagging. As such, the initial sample is bootstrapped (random draw with replacement) so that decision trees can be built from each of these bootstrapped samples, before taking the average of the imputations obtained. This method is very frequently used on medical data, but its application in finance is rarer. That is why it is interesting to include it in this PhD thesis. However, this method involves issues of non-replicability of the results due to the random component introduced by the bootstrap, which may not be acceptable from the regulator's point of view. Indeed, this method gives different results from one execution to another, and the bank could be tempted to use an execution that would be in its favor (allowing it to minimize its capital costs), which is why the regulator may be suspicious and reticent regarding its use.

An improved version of the EM (expectation-maximization) algorithm is also included in the benchmarking analysis. This is the Amelia algorithm, which combines EM with multiple imputation. Multiple imputation consists of bootstrapping (random draw with replacement) the initial sample, applying a completion method to each of them (here the EM algorithm), and then using all these imputed samples for analyses. The analyses are calculated from each of the imputed samples, and can then be averaged. Unlike bagging, where the analyses are applied to the average of the imputations, here the analyses are applied to each of the imputed samples before being averaged. Averaging takes place later. The EM algorithm is frequently applied to financial data in the literature, hence the appeal of using Amelia in this PhD thesis. In addition, Amelia does not require any specific settings, beyond initialization of the EM algorithm settings. Little and Rubin [145] also deal with this subject in their work. Another multiple imputation method is used in this PhD thesis: the MICE (multivariate imputation by chained equation) algorithm. Although this method does not use bootstrap, it results in several imputed samples. The MICE algorithm consists, from the same starting sample, in imputing the missing data by PMM (predictive mean matching), which integrates several random components. The idea of this method is to estimate the missing data by linear regression, and to impute it by randomly drawing from a certain number of observed values (donors) that are close to the estimate made. This method therefore involves making a choice regarding the number of donors to use: this must not be too large, or it risks moving too far away from the estimate made, but it must not be too small either, or it risks imputing the same value too often. Furthermore, this procedure is iterative and uses Gibbs sampling for imputation. This method is often used in the literature, but more rarely on financial data. Finally, the last two methods used are similar, since one is an improved version of the other: IPCA (iterative principal component analysis) and the multiple imputation version of this method. IPCA consists of breaking down the initial sample from an analysis into main components in order to reconstruct it using a well-defined number of main components, and replacing the missing data with their prediction. This procedure is repeated several times until the predictions converge. This method has also been implemented within

an algorithm aimed at detecting outliers and imputing missing data (MacroPCA [115]). In addition, Josse and Husson [126] improved the IPCA in two respects: regularization to avoid overfitting errors and integration into a multiple imputation process. As with MSSA, IPCA and MIPCA require defining the number of main components to be used for reconstruction. Here, it is defined by a generalized formula from a cross-validation method (the leave-one-out method), established by Josse and Husson [126].

These methods represent only a selection of what is done in the literature. There are many others, such as imputation methods using a self-regulatory model that make it possible to exploit the temporal dimension of data [146][218], or methods based on generative antagonistic networks (GAN), where research has been particularly active in recent years [96][215][216][213]. The latter are increasingly used for financial data [107][134][49][50][72].

After presenting all the methods that will be used, it is time to put them to the test and to compare them to each other. These completion methods will initially be used on a simulated sample from which data will have been deleted based on different missing data mechanisms, before applying them to historical samples with the most relevant missing data mechanisms. This type of procedure will allow for the comparison of all these completion methods based on different analytical criteria. Comparisons will be made in terms of statistical moments, proximity measurements (MAE and RMSE), covariance matrix difference and risk measurement (VaR and ES). Thus, the first analysis consists of comparing the completion methods on a simulated sample containing data that will have been deleted from the sample completely at random (MCAR), i.e. according to a uniform law. The data are deleted in the first column of the simulated sample, and in proportions ranging from 5% to 70%.

The second analysis is very similar to the first, as the same missing data mechanism is used (MCAR) but this time on almost the entire matrix (all columns except the last one). The proportions of missing data per column range from 5% to 70%. This scenario is more realistic than the previous one, because in practice data are usually missing in several columns of the sample and rarely in just one.

The following analysis consists of applying the imputation methods to a heteroskedastic sample, i.e. a sample with non-constant volatility. Thus, the analysis is, on this occasion, based on several simulated samples containing an increasingly volatile period of crisis, in order to see whether or not the deterioration of the algorithms' quality of imputation depends on the deformation of the series. An MCAR data mechanism (with a proportion set at 30%) is then applied to the first column of each sample, to make comparison possible.

The fourth analysis aims to observe the impact of the presence of jumps in the series on the quality of allocation. Thus, based on the same data draw, several samples containing an increasing number of jumps were simulated. MCAR data were deleted

(to a proportion set at 30%) in the first column of each of these samples, before being imputed by all methods. This reveals which methods are sensitive to the presence of jumps in the series.

The fifth comparative analysis uses the same simulated sample as the first two analyses, but deletes the data in such a way as to follow an MAR mechanism. To do this, the data are deleted based on the values of the last column of the matrix. More precisely, the extreme values from the last column are identified, and the corresponding observations are deleted in the first column. The proportions of this missing data mechanism range from 5% to 70%.

Another missing data mechanism is applied to the simulated sample and consists in deleting data successively in the middle of the first column. The proportions of missing data associated with this mechanism are the same as previously, i.e. 5% to 70%. This makes it possible to compare the methods when handling a large number of missing data points in succession.

The following mechanism is very similar to the previous one, as it also consists of deleting the data successively, but this time at the end of the series (always in proportions ranging from 5% to 70%). Thus, no further observations are available afterwards.

Finally, the last mechanism applied to the simulated sample consists of removing the extreme values of the series, in proportions ranging from 5% to 70%. This mechanism directly addresses the distribution of the series, which may pose issues for a number of completion methods.

After conducting all these comparative studies on simulated samples, some of them should be conducted on historical data. For this purpose, the Euro Stoxx 300 data is used to form two historical samples over the period from January 2020 to February 2021. A first sample based on a heuristic approach (shares from the same sector and which are strongly correlated) is formed, then a second sample based on a graphical Lasso approach [86]. MCAR data (based on a uniform law) has been deleted from the first column of these two historical samples, then a second missing data mechanism is applied, consisting of successively deleting data at the end of the first column. The first mechanism was applied because it is a case frequently studied in the literature, and the second because it is frequently observable in financial data (this is typically the case before an IPO).

After presenting and analyzing all the results obtained for each comparative analysis, some conclusions are discussed. Firstly, the results may differ from one benchmarking analysis to another, and it is clear that they depend on the sample to which they are applied.

Secondly, some methods result in the non-replicability of the results, which can be very problematic from the regulator's point of view. The bank may be tempted to use a scenario that is favorable to it in order to reduce its capital costs. It would be possible

to set a "seed" to solve this type of problem, but this still needs to be representative. Another conclusion is that the choice of the imputation method must be made according to the nature of the data, and especially their future use. Imputing so as to obtain a value close to the missing value will certainly be effective from the point of view of proximity measures, but probably much less so in terms of risk measures. While the random forests method appears to be effective for all criteria and proportions of missing data, this is not the case for other methods. The KNN algorithm is effective in terms of proximity measurements, but not in terms of risk measurements. The Amelia and MIPCA methods also achieve very good results for many criteria. On the other hand, Amelia is not able to manage an excessive proportion of missing data, regardless of the analysis criterion. One of the lessons of this PhD thesis is also to note that a method is not an algorithm and that talking about a method alone is not very valuable if all the pre-processing and other steps are not explicitly presented. This is particularly the case for random forests, two implementations of which are both commonly used in the literature, but which do not give the same results.

This emphasizes the importance of the the operational risk generated by the imputation methods. An imputation method may, in some cases, not work (not enough observable data), which can be catastrophic for a bank. In addition, differences in the interpretation of algorithms can lead to different implementations, hence the importance of clear, transparent and precise documentation concerning them.

Regarding the methods in particular, it was observed that the Amelia algorithm was particularly sensitive to the proportion of missing data. Beyond 50% missing data, this imputation method tends to lose quality. This can be explained notably by the initialization of parameters based on observable data. When the proportion of missing data becomes too large, these parameters may deviate significantly from those of the complete series.

In addition, some paradoxical results are observed: the data imputed in the first column are of better quality when all the columns contain missing data compared to when the rest of the sample is complete. This is particularly evident in terms of proximity measurements, for the IPCA, MIPCA, MICE and Amelia methods. Upon further investigation, this is not due to a better estimate in law, and can be seen beginning from two columns containing missing data. These methods are all iterative, and it would appear that their iterative procedure goes through a different convergence path (when several columns contain missing data) that leads them, for an unknown reason, to better results.

Finally, Amelia and random forests appear to be the two best performers of all these comparative analyses. However, random forests tend to perform better than Amelia. This is partly explained by bagging and in particular the averaging of random forests imputations, whereas Amelia involves calculating each criterion on each imputed sample, before averaging these criteria. If the imputed samples are averaged before applying the analysis criteria, Amelia's performance is generally improved. In addition,

another element of these differences can be explained by the sample size. Modarresi and Diner [156] in fact showed that the performance of the algorithms was variable depending on the sample size. In their study, they show that Amelia is more effective than random forests for a certain sample size, and vice versa for another size. This once again reflects the fact that the performance of algorithms depends on the data itself.

In sum, after presenting the context and the various issues related to missing data in finance, a variety of completion methods were applied to samples (simulated and historical) in order to compare them. The comparative analyses made it possible to highlight certain points, and in particular the remarkable performance of the random forests algorithm, as well as of Amelia. Other methods, such as MICE, have mixed results and must be applied to more samples before conclusions can be drawn. Finally, some methods appear to be poor performers in almost all analyses, such as the LOCF method. However, this method is appealing from the regulator's point of view because it is the most conservative.

Avenues for improvement are discussed throughout the PhD thesis, in particular that aimed at repeating the analysis (using the most relevant methods) on a large number of data samples in order to be able to generalize the results.

Introduction

All empirical researchers, whatever their field of study, must address the issue of missing data. Whether quantitative or qualitative, all analyses that are based on real data are potentially sensitive to missing data. Throughout the scientific world, missing data and their management have been a point of difficulty for decades.

Missing data have probably always been a concern, but it was not until the late 1980s that a theoretical framework dedicated to their management was constructed. The pioneers in the field were Little and Rubin [145]. In 1987, they proposed not only a definition of missing data, but also a categorization, a means of identifying their distribution and solutions that enable statistical analyses to be executed when data are missing.

According to Little and Rubin [145] "missing data are unobserved values that would be meaningful for the analysis if observed; in other words, a missing value hides a meaningful value". Therefore, they saw the information that the missing data contain as being of equal importance to that in observed data. Thus, relying exclusively on observed data would necessarily introduce bias into the statistical analysis. For this reason, they distinguished between different missing data patterns and between different mechanisms, namely "missing completely at random" (MCAR), "missing at random" (MAR) and "missing not at random" (MNAR). The idea behind these mechanisms is to understand why the data are missing, especially if they are missing in relation to other values. The analysis of the data will, in fact, be different when the missing data are entirely independent of the data in the sample (MCAR) and when they are related to them (MAR or MNAR). This is why some authors have even designed statistical tests that attempt to address these issues (for example, Little [142] in 1988 and Jamshidian and Jalal [123] in 2010).

Thus, Little and Rubin [145] introduced new concepts and solutions to the subject. Accordingly, they became the leading authorities on missing data. Later, other researchers tried to refine their theoretical framework. The leading papers are that by Schafer and Graham [182] from 2002 or even the one published by Gelman and Hill [90] in 2006. Those contributions still draw on Little and Rubin's theory [145]. After its publication, that theory became a starting point for many researchers.

Missing data issues occur in management and finance, as well as in other fields where real data is used. Indeed, according to Tsikriktsis [201] and Kofman and Sharpe [133], approximately a third of research articles in management and finance, respectively, mention missing data and their management. This figure does not mean that only a

third of researchers are confronted with missing data but rather that only a third have mentioned its presence and management in their publications. The actual proportion is probably much higher. However, many authors do not write about the problem because of a lack of interest or even a fear that their study would be considered less credible.

In fact, many studies have shown the negative effect that missing data can exert on statistical studies. In 1977, Kim and Curry [130] demonstrated that even if only 2% of the data in a sample are missing, deleting the observations that contain them entails reducing the size of the sample by more than 18%. This reduction in sample size involves a loss of statistical power (among other things). Therefore, revealing that a significant proportion of a sample had to be deleted due to missing data is liable to discredit an article.

Moreover, the challenges that missing data pose extend far beyond the academic sphere; they are the source of very practical problems. Banks (as well as asset managers and insurers) must also confront missing data. Missing data can interfere with the proper execution of many quotidian calculations and therefore require automated management.

Banks must calculate their profit and loss (P&L) on a daily basis. For a few years now, financial regulations have required banks to calculate various risk measures daily in order to compute the amount of capital charges that would ensure their stability in a crisis. To fulfill these obligations, banks purchase much of their data from external providers, who must manage missing data actively. The banks also save data in order to reduce their expenditure.

Developments such as these mean that the management of missing data has become essential for banks, especially since regulators understood that data quality is essential to financial stability. If a bank possesses data of good quality, their risk measures gauge exposure accurately. If the data are poor, then the same measures can over- or underestimate risk exposure, resulting in vast capital charge expenditures or, more dramatically, in bankruptcy. Thus, data quality, and therefore missing data, do have an impact on the stability of individual banks and the entire financial system.

Data may be missing from financial series for many reasons, but the most obvious one is probably market closure. The stock markets are not open every day, and they follow specific calendars. Stock markets all over the world typically close on weekends, but other closing dates vary between countries. For example, European stock exchanges (and others) close on Good Friday and on Easter Monday, but Asian stock exchanges do not. Accordingly, it is normal for financial series to lack data for weekends and for exchange-specific holidays.

In practice, data from weekends are not really considered to be missing. In databases or even in downloadable data, weekends very often do not form part of the time series,

and they are automatically removed from the record. However, in some cases, it may be wise to fill in the missing data even if they do not correspond to quotation dates. In the case of daily returns, removing the missing data instead of filling them in implies that the distribution of returns is identical for all days of the week and in particular, that returns between Friday and Monday are identical to those observed in other periods.

This is problematic because the day of the week, and weekends and holidays in particular, do affect returns. Indeed, in 1980, French [85] showed that between 1953 and 1977, the daily returns from Standard & Poor's (S&P) composite portfolio were on average significantly negative between Friday and Monday, whereas those from other days of the week were positive. French [85] went even further and compared the returns that follow holidays to returns from periods without holidays. He found that the negative returns were indeed attributable to a weekend effect and not simply to market closure. Moreover, his results show that the weekend effect persists over long weekends (negative returns for Tuesdays when Monday is a holiday). In 1981, Gibbons and Hess [91] confirmed the existence of a weekend effect on asset returns in a study based on data from the S&P and the CRSP, where they also observed negative returns that persisted after Mondays.

In 1988, Miller [155] explained this phenomenon by "a slight tendency for self-initiated sell trades to exceed self-initiated buy trades over the weekend, while broker-initiated buy trades cause a slight surplus of buying during the rest of the week." In other words, investors tend to sell more over the weekend. Accordingly, brokers must buy on Monday and sell during the week. In 1994, Abraham and Ikenberry [1] added that investors are more active on Mondays, especially following the circulation of a negative announcement or publication over the weekend. Therefore, investors are also sensitive to bad news.

For these reasons, particularly negative returns are expected after weekends and long weekends. This tendency may have repercussions for risk measures. Indeed, in 2014, Liu and An [147] showed that the information accumulated during non-trading hours impacts risk measures. To establish that relationship, they used daily data (opening and closing) from copper, rubber and soybean futures from the Chinese market over the period between 1993 and 2012. Then, they modeled value-at-risk (VaR) copulas and expected shortfall (ES) copulas [1] that allowed them to account for trading and non-trading information. Table 0.0-1 overviews the parts of the risk measures that are related to trading and non-trading returns and their proportional contributions.

[1] Liu and An [147] use copula functions to obtain a joint distribution of trading and non-trading returns before computing risk measures.

Tab. 0.0-1: Component VaR and ES and risk contributions of trading and non-trading returns (Source: Liu and An, 2014 [147])

Market	Risk Measures	95% confidence level		99% confidence level	
		Trading returns	Non-trading returns	Trading returns	Non-trading returns
Copper	Component VaR	0.8475	1.0187	1.2409	1.9056
	Risk contribution	45.41%	54.59%	39.44%	60.56%
	Component ES	1.2325	1.7928	1.6054	2.5939
	Risk contribution	40.74%	59.26%	38.23%	61.77%
Copper	Component VaR	1.3631	0.9541	2.2482	1.8724
	Risk contribution	58.83%	41.17%	54.56%	45.44%
	Component ES	1.8100	1.8234	2.9334	3.8886
	Risk contribution	49.82%	50.18%	43.00%	57%
Copper	Component VaR	0.9204	0.6503	1.4991	1.3250
	Risk contribution	58.60%	41.40%	53.08%	46.92%
	Component ES	1.3670	1.4083	2.4008	2.7610
	Risk contribution	49.26%	50.74%	46.51%	53.49%

This table reports the component VaRs and ESs at the 95% and 99% confidence levels, as well as the risk contributions of trading and non-trading hours in copper, rubber, and soybean futures markets.
The non-trading returns have a higher contribution to the overall risk for all markets. For example, for copper, the non-trading returns contribute 54.59% and 59.26% to the VaR and the ES, respectively, at the 95% confidence level.

Liu and An [147] pointed out that on average, the contribution of non-trading returns is larger than that of trading returns, as far as total risk measures are concerned. If all markets are aggregated, non-trading hours impact at least 40% of VaR. This threshold increases to more than 50% for ES. Therefore, it represents a non-negligible proportion of total risk measures.

In order to understand the origins of the risk introduced by non-trading information better, they analyzed the relationship between risk measures and the duration of non-trading periods. They distinguished between weeknights, weekends and holidays and recalculated the risk measures for non-trading periods accordingly. Their results are presented in Table 0.0-2.

Tab. 0.0-2: VaRs and ESs based on weeknight, weekend and holiday returns (Source: Liu and An, 2014 [147])

Market	**Non-trading periods**	**95% confidence level**		**99% confidence level**	
		VaR	**ES**	**VaR**	**ES**
Copper	Weeknights	1.8271	3.0560	3.0140	5.0766
	Weekends	1.9430	3.1311	3.8736	5.1878
	Holidays	2.1651	3.9443	4.3171	9.9120
Copper	Weeknights	1.9287	3.4845	4.2380	6.611
	Weekends	2.2556	4.4134	5.4663	8.8830
	Holidays	2.0905	2.8475	3.3896	3.9654
Copper	Weeknights	1.4275	2.1541	2.6553	3.4319
	Weekends	1.5934	2.8289	3.4112	4.2237
	Holidays	1.7078	4.8849	4.7819	7.8437

This table reports the estimated VaRs and ESs at the 95% and 99% confidence levels based on weeknight, weekend, and holiday returns for copper, rubber, and soybean futures markets.
The VaRs and ESs of weeknights are lower than the VaRs of weekends or holidays at both the 95% and 99% confidence levels. For example, for copper, the VaR of weeknights is 1.8271 whereas that of weekends is 1.9430 and that of holidays is 2.1651, at the 95% confidence level.

They found a positive relationship between risk measures and the duration of the non-trading periods. In other words, risk measures tend to be higher (and therefore associated with greater losses) when calculated over weekends or holidays than when calculated over weeknights. The authors therefore concluded that the longer a non-trading period, the greater the associated risk.

Given that non-listing periods can impact risk measures significantly, the weekend (and holiday) effect also impacts risk measures. For example, in the case of the Easter weekend, when there no trading occurs on Friday or Monday in Euronext countries, a comparison between a return calculated between Thursday and Tuesday and any other one-day return would be unrepresentative. The return associated with the Easter weekend may be abnormally high. The abnormality would not be attributable to high volatility but to the passage of several days since the last close. Concerns of this kind have prompted some authors to remove certain data to prevent bias from intruding into their analyses. This is notably the case for Giot and Laurent [93], who sought to model VaR from realized volatility and the ARCH model [77]. They voluntarily removed some data to forestall the weekend effect.

As a general matter, when an expert uses data from different stock markets that follow different calendars, weekends are removed because no observations are available, regardless of the series. There is a weekend effect on daily returns, but it is common to all series, and it is therefore less problematic. Conversely, some dates may be missing for some series and not for others due to holidays. In such instances, the expert must choose between removing the dates in question and losing information, which could bias their analysis considerably, and imputing the missing data in order to make full use of the available information. Of course, the choice of imputation method is essential,

but it is necessary to determine which completion method is the most appropriate. It is obvious that imputing a long weekend naively by using the last observed quote, as some banks do by default, is tantamount to doing nothing and impacts VaR far more than a simple linear interpolation.

In the light of these considerations and the difficulty of implementing certain regulatory principles, regulators have noted that data quality, especially missing data, is a fundamental issue for banks. As a result, regulators have introduced additional measures for several years in order to impel banks to become more transparent and considerate in their treatment of data. Often, these measures are included in regulations with a more general objective, as in the case of the fundamental review of the trading book (FRTB) [12], which aims to reform financial regulation in its entirety to strengthen the financial system. However, regulators have also been able to establish specific principles on data quality, such as the Basel Committee Standard 239 (also known as BCBS 239) [24], and on transparency, such as the Principles for Financial Market Infrastructures [58].

Therefore, data pose important challenges to banks, which is why recent regulations insist that data of a certain quality be used for regulatory calculations while encouraging banks to use completion methods in order to fill in omissions. However, the regulations do not explicitly state how banks can or should manage missing data. The regulator therefore leaves the selection of suitable completion methods to banks, as long as the methods that are chosen are documented carefully. Consequently, banks have found themselves forced to question the different completion methods that are available and their efficiency in financial series. Numerous completion methods exist in the literature. Some are intuitive; others are more complex and sophisticated.

Banks already use some such methods in a context that is quite different from what has been discussed here. These methods can be used by a trader who wants to fix the price of a financial product. The trader will then set their price by using their knowledge of the market (for example, their knowledge of the price of other products) and their experience in the field. The completion method is thus based on a heuristic approach that draws on the trader's market expertise. This is particularly obvious when illiquid products, which do not necessarily have a daily quotation, must be traded. Evidently, the data are not always clean market data, and they are probably more fabricated than expected. This heuristic method could also be used by banks to fill in the missing data in their databases.

The following question emerges: is an expertise-based methodology sufficient to estimate missing data? Is it able to outperform a specific algorithm? This problematic can be related to the work of Kahneman [62] and his 2001 book "*Thinking, Fast and Slow*", where he pits the intuitions of experts against formulae. He contrasts System

1, which corresponds to an intuitive, approximate, fast and often effective strategy (heuristic approach), to System 2, which is based on logical, analytical reasoning that is cognitively more costly and slower but more reliable. His argument is based on the 20 studies by Meehl [153], published in 1954, which reveal that clinical predictions that are based on the subjective impressions of experts are often less effective than predictions made through the use of a precise statistical method. Kahneman [62] explains that since the publication of Meehl's studies [153], more than 200 such studies have been published, and in 60% of the cases, they showed that algorithms were more accurate than individuals. Moreover, among the remaining cases, statistical methods performed as well as the experts, never worse. Thus, in prediction, it is preferable to use a simple analytical method rather than a complex combination that works only occasionally. Kahneman adds that the experts are even less efficient than statistical formulae if they have access to the information obtained by the algorithm. They then enter into a competition of sorts. On average, the additional information causes the experts to predict the wrong result.

Kahneman [62] adds that humans are inferior to formulae because of their inconstancy when confronted with complex information. They can base different answers on the same information. In a study by Hoffman, Slovic and Rorer [111], experienced radiologists were presented with the same radiographies on two different occasions. It emerged that 20% of the diagnoses were contradictory. Brown [48]reports similar results for internal auditing. Kahneman [62] explains that human judgment is influenced by imperceptible external stimuli. This affords an advantage to algorithms, which cannot be influenced and always infer the same answer from the same data.

Thus, Kahneman [62] suggests that "to maximize predictive accuracy, final decisions should be left to formulas, especially in low-validity environments." Overall, it appears that algorithms are more reliable than expert opinions because they are not sensitive to external events and because they cannot be influenced. In finance, the data that is produced by traders and which is based on their own expertise can be a poor estimator of market prices. As noted earlier, traders are liable to incorporate all kinds of external influences, such as those of poor investment choices made during the day or their P&L goal or even the day of the week, into their prices. It would therefore be preferable for traders to use a specific algorithm or a statistical rule to set market prices because those techniques are impervious to external stimuli.

Nevertheless, Kahneman [62] refined these statements, in particular following his collaboration with Klein, a great defender of expert intuition. In 1999, Klein [132] related the story of a team of firemen during an intervention. Their commander ordered an immediate evacuation without any particular reason. The commander's sixth sense saved the entire team: the floor collapsed after the evacuation. The commander had perceived something abnormal about the fire without knowing precisely what. It

transpired that the main source of the fire was in the basement, under the feet of the firemen. Driven by this anecdote, Klein [132] elaborates his theory of decision-making, which is called the recognition-primed decision model. It involves both System 1 and System 2. The first step of a plan is conceived through the automatic functioning of associative memory (System 1). Then, the plan is simulated mentally in order to verify its correctness (System 2). Therefore, the commander's decision originated from a plausible mental simulation of the situation that was based on years of experience. Expert intuition can be underpinned by a much more complex and more reliable process.

Kahneman and Klein realized that their disagreement about intuition was partly due to differences in their respective definitions of expertise. While Klein had worked extensively with fire commanders, nurses and other professionals of undoubted expertise who were aware of the limits of their knowledge, this was not necessarily the case for Kahneman, who had worked with clinicians, brokers and pseudo-experts who had no idea that they did not know what they were doing. This is reminiscent of the Dunning-Kruger effect [135], whereby the least qualified people tend to overestimate their competence. In their studies, Dunning and Kruger [135] show that the underqualified individuals fail to recognize their incompetence and to evaluate their capacities. Conversely, qualified individuals tend to underestimate their competence, hence the existence of pseudo-experts. It is thus important to differentiate between true experts and pseudo-experts, which is why Kahneman and Klein agree that "the confidence that people have in their intuitions is not a reliable guide to their validity."

According to Kahneman [62], algorithms and statistical methods are better adapted to problems of missing data. Some heuristic methods can be efficient, but it is still necessary to ascertain whether the expert is reliable or not. However, there is not enough data on the behavior of traders to gauge their reliability. Therefore, banks tend to prefer algorithms for the management of missing data, especially since they are more easily validated by the regulator. A method that tends to be, directly or indirectly, sensitive to external stimuli, in the sense in which Kahneman uses the term [62], is very likely to produce inaccurate estimates and to be rejected by the regulator. The regulator insists that banks select transparent completion methods. However, the use of an algorithm does not necessarily imply regulatory acceptance. Some complex and non-transparent methods, often referred to as "black box," do not curry favor with regulators.

Considerations of this kind drive the choice of completion method. Many such models are available in the literature. Therefore, the identification of the ones that are most relevant to financial series is a matter of interest. This PhD thesis will also aim to compare the imputations made by different methods according to different criteria, such as statistical moments, proximity measures, covariance matrices and risk measures.

To this end, 10 very different completion methods are examined. Some are very

simple and known to be default solutions, while others are much more complex. The simplest methods include linear interpolation and the last-observation-carried-forward method. These methods are often used by default in the automatic processes of banks because they are simple and quick to implement. For this reason, they will be employed as benchmarks in the comparative analysis.

Other methods that are frequently used for the imputation of financial data will also be analyzed. One of them, the Brownian bridge, entails interpolating prices by using Brownian motion. This imputation method is simple. Like the previous ones, it is one-dimensional, in the sense that it requires no information other than the series itself. This is not the case for other completion methods, which rely on the cross-sectional dimension of the data.

Multidimensional algorithms draw on information from other series in order to capture market dynamics. In this way, they predict the missing data better. For this reason, a large proportion of the algorithms that this PhD thesis covers are multidimensional. One notable example is the EM algorithm, which is widely used on financial data because it is based on a Gaussian framework. The distribution of the missing data is estimated from the distribution of the observed data. This is one of the methods that Little and Rubin [145] promoted. The EM algorithm can be combined with bootstrapping to improve performance. The version implemented by Honaker, King and Scheve [112] in 2002, which is called Amelia, is used in this PhD thesis. Its analysis enables the differences between a method and an algorithm to be highlighted (thus integrating one or several methods) and the importance of precise and detailed documentation to be highlighted alongside the attendant operational risks.

The principal component analysis (PCA) method is used very frequently in financial data analysis, but not necessarily for data imputation. Like the EM algorithm, PCA been integrated into much more sophisticated imputation algorithms. The literature circles on two main PCA-based algorithms: MacroPCA by Hubert, Rouwweeuw and Van den Bossche [115] and MIPCA by Josse, Pagès and Husson [127], which is an improved version of the iterative PCA. Given that PCA has already been proven to solve factorial analysis problems in finance, it is not surprising that these methods are also efficient for data imputation.

The comparison also accounts for the very famous K-NN method that was introduced by Fix and Hodges [83] in 1951. Its use for imputing missing data is widely accepted in many research fields. It involves imputing the missing data using observations (prices or returns) of series that the model has considered as close (according to a specific distance). This method is often considered to be one of the first machine learning methods. Unlike others, it is simple and understandable.

The random forests method, which is more complex but also more recent, will also be analyzed. It was created by Breiman [42] in 2001. A non-parametric algorithm

estimates the missing data by using decision trees built from bootstrapped samples. Breiman [41] also integrated bagging into random forests, which seems to be one of the reasons for the success of the algorithm. The PhD thesis discusses it extensively.

The MICE algorithm is also employed often in the literature. It was introduced by Van Buuren [54] in 1999, and it combines predictive mean matching and regression. Its operation is longitudinal: it imputes missing data by drawing on the values of other observations that meet a certain number of criteria. Despite its popularity, no study has applied MICE to financial data, a further reason for its inclusion in this analysis.

Finally, banks are not the only organizations that are interested in completion methods; they concern data providers, too. Since data providers are increasingly subject to regulations on data quality, their interest is mainly commercial. Banks' outlays on data are considerable. Accordingly, they use completion methods, especially if they already own data. However, it is not always easy to implement an efficient completion method and have it validated by the regulator. The bank has to allocate a budget to research, to choose a method and to document it precisely to secure regulatory approval. Data providers offer to lessen that burden by imputing data through a proprietary completion method that the regulator approves. Articles on the completion of data by data providers have been published, notably on the MSSA algorithm that was implemented in 2016 by Dash and Zhang [65], researchers at Bloomberg L.P. It is based on the spectral decomposition of data and since it is sold by data providers, it belongs in this PhD thesis. Of course, there are many other completion methods. The selection in this PhD thesis reflects considerations of relevance.

Evidently, missing data problems abound in finance. They can skew risk measures significantly and thus mislead both banks and regulators. Banks, encouraged by regulation, are therefore interested in completion methods. However, as mentioned, the literature offers many methods. Banks must study them to identify the one that is most appropriate to their needs. Accordingly, this PhD thesis aims to present the general problems of missing data before focusing on finance and regulation. Thereafter, it turns to the relevant imputation methods and their efficiency.

It is necessary to begin with a review of the academic and practical issues that surround missing data. The review underscores the proposition that the treatment of missing data forms part of a much larger topic, that of data quality. The latter notion subsumes many issues that are as important as missing data. They are also related to it. The general problems associated with missing data will also be presented. Given that missing data affect all areas of research, the literature on the subject is vast. For this reason, the meaning and impact of missing data will be discussed in the abstract before they are related to financial data. Finally, as noted previously, the use

of completion methods by banks is partly motivated by financial regulations. Therefore, the regulations on missing data will be described.

Once the context and the motivations of the research are set out, it will be possible to present the theoretical framework of missing data in financial series, as well as some alternatives. Little and Rubin's framework [145] will be presented, as will the approaches of other authors. Thereafter, the theoretical framework will be applied, and an attempt will be made to explain missing data. Before the operation of the different completion methods that are used in this PhD thesis is presented in full, the emphasis will be placed on data encoding, whose importance is no lesser than that of completion methods.

Finally, a comparative study of the completion methods will be performed using both simulated and historical data. Different missing data mechanisms (MCAR, MAR and MNAR) will be applied to simulated data. Various simulated series will be used in order to observe the efficiency of the methods on heteroskedastic series and on series that contain jumps. Then, historical data from Europe that contains missing observations will be presented. Two sub-samples will be constructed, and MCAR data will be introduced artificially to compare the performance of the algorithms.

Chapter 1:
Issues around missing data

Missing data are only an element of a much larger topic, data quality. The issues that missing data raises are connected to data quality, and the two are often managed together. Nevertheless, this PhD thesis focuses specifically on missing data. In research, especially in quantitative domains, they are a significant problem. Despite their ubiquity in finance, missing data have only begun to attract regulatory attention in the last decade. As a result, financial data pose numerous challenges for banks and other participants in the financial markets.

Contents

1.1 Missing data is only one aspect of data quality

Missing data is only one of many data-quality issues. Indeed, components of data quality also include ensuring the completeness of the data, as well as its validity, accuracy,

consistency and availability. This section presents three main data-quality issues in a financial series framework. First of all, there may be many data sources, in which case it is necessary to be able to determine which is the best source. Conversely, it may be difficult to obtain certain data, leading to long investigations to find a valid source. Finally, the data may be composed of false or unrepresentative market data, presented in the literature as outliers.

1.1.1 Multitudes of data sources: the new enemy of consistent data

The first step in a quantitative analysis is data collection and the decision of which data provider to turn to. In finance, various data sources can be relevant, either free of charge or with a fee for access, providing access to data series. A wide range of different data providers, paid or free, can be used, including Bloomberg L.P., FactSet, Thomson Reuters, Markit, or Yahoo Finance, Investing.com, Swissquote and so forth. These sources provide market data for many types of financial instruments (e.g., shares, bonds, funds, options, futures and yield curves), but also regulatory data or data from brokers and dealer desks. This high volume of much data allows efficient exchange on the markets. Is data from one provider really the same as data from another provider?

In the case of financial data, there are many different sources of data and it would be logical to assume that for a specific series, the data would be the same regardless of the source. In 2001, Elton, Gruber and Blake [75] compared financial databases, specifically the CRSP and Morningstar databases. The CRSP built a mutual fund database to challenge that of Morningstar as the source of fund data for academic research. Therefore, the authors explain that the purpose of their paper is to examine potential errors in the CRSP database by comparing the performance of CRSP data with that of Morningstar data.

Their sample comprises common stock funds from the Growth and Income group in the CRSP database, which had over $15 million in total net assets in 1998 and, that also had complete sets of monthly returns in both the CRSP and Morningstar databases for a 20-year period (from January 1979 through December 1998). They obtained the following results (see Table 1.1-1). They use the same model as Elton, Gruber and Blake in 1996 [76], a four-index model regressing the excess return of each fund against the excess return of the S&P 500 index, the return on the small stocks minus large stocks, the return on growth stocks minus the return on value stocks and the return on long-term bonds minus the T-bill rate. The difference in alphas are the intercepts of this four-index model applied to the CRSP and Morningstar data for each of the four five-year samples presented in Table 1.1-1.

Tab. 1.1-1: Differences in monthly alphas estimated from four-index model using CRSP and Morningstar monthly return data (in basis points) (Source: Elton, Gruber and Blake, 1996[76])

Sample Period	Number of funds with differences	Average differences	Average absolute differences	Number of differences greater than or equal to		
				10 basis point	5 basis point	1 basis point
Large Funds						
1979-1983	25	-1.34	4.84	3	6	11
1984-1988	25	0.09	0.60	0	0	3
1989-1993	25	-0.12	0.21	0	0	1
1994-1998	24	-0.14	0.26	0	0	3
Overall	99	-0.378	1.48	3	6	18
Small Funds						
1979-1983	25	-5.08	7.71	5	7	14
1984-1988	25	-0.11	2.32	2	3	9
1989-1993	25	-0.81	1.20	0	2	7
1994-1998	24	-0.26	1.29	2	2	4
Overall	99	-1.56	3.13	9	14	34

All CRSP fund data obtained from CRSP Mutual Fund Databases, version 1. 1998.
All Morningstar fund data obtained from Principia Pro January 1999,CD.
Differences measured as alpha using CRSP data minus alpha using Moringstar data.
For example, when examining the small funds, 34 out of the 99 observations have differences in alpha greater than 1 basis points per month, 14 have differences greater than 5 basis points per month, and 9 have differences greater than 10 basis points per month.

First, they notice that the largest difference in alphas is the one the use of the oldest data. Indeed, the average of the absolute difference of large funds is 4.84 (and for small funds, 7.1) basis points for the period 1979-1983, compared to 0.26 (and, respectively, 1.29) basis points for the period 1994-1998. The older the data is, the greater the gap will be. This gap may not seem like much, but the data used are monthly data, so if they are input on an annual basis, the differences take on another dimension. The difference is then, for the oldest period, 16 basis points per year for large funds and 61 basis points per year for small fund. At the same time, the authors note that this difference is greater for the small funds than for the large funds.

Moreover, the authors find (on the right-hand side of Table 1.1-1) a number of large differences between the alphas of individual funds. Among the 99 small funds studied, 34 have alpha differences greater than 12 basis points per year, 14 have differences greater than 60 basis points per year and 9 have differences greater than 120 basis points per year. The five largest differences are 6.8%, 4.9%, 4.9%, 3.1% and 2.5% per year. Fewer large differences exist for large funds than for small funds (18 have alpha differences greater than 12 basis points per year, 6 have differences greater than 60 basis points per year and 3 have differences greater than 120 basis points per year); the difference are also smaller (4.9%, 3.1%, 1.9%, 1.6% and 1.4% per year).

As such, more large differences exist for small funds than do for large funds and these differences are larger. On the other hand, the authors note that the five largest differences are concentrated in the first period (1979-1983) for large funds, whereas they are scattered over the entire period (1979-1998) for small funds.

These differences in alphas are important and can significantly impact conclusions made on individual funds, especially based on old historical data. Elton, Gruber and Blake [75] thus recommend extreme caution when manipulating these numbers. They state that the differences observe in alphas are caused by differences in the returns reported by CRSP and Morningstar. Hence, the authors repeated this same analysis, but this time using the CRSP and Morningstar returns (see Figure 1.1-2).

Tab. 1.1-2: Differences in monthly total returns using CRSP and Morningstar monthly return data (in percent)(Source: Elton, Gruber and Blake, 1996[76])

Sample Period	Number of months with differences	Average differences	Average absolute differences	Number of differences greater than or equal to		
				5.0%	1.0%	0.5%
Large Funds						
1979-1983	532	0.001	0.151	14	44	59
1984-1988	421	-0.002	0.030	0	8	19
1989-1993	297	0.000	0.015	0	2	6
1994-1998	185	0.000	0.009	0	4	7
Overall	1,435	0.000	0.052	14	58	91
Small Funds						
1979-1983	639	-0.030	0.280	20	91	145
1984-1988	473	-0.004	0.122	8	31	57
1989-1993	231	-0.007	0.029	0	13	28
1994-1998	190	0.002	0.013	1	3	7
Overall	1,533	-0.010	0.111	29	138	237

All CRSP fund data obtained from CRSP Mutual Fund Databases, version 1. 1998.
All Morningstar fund data obtained from Principia Pro January 1999,CD.
Differences measured as return using CRSP data minus alpha using Moringstar data.
For example, when examining the small funds, 237 out of the 1,533 observations have differences in alpha greater than 0.5% per month, 138 have differences greater than 1% per month, and 29 have differences greater than 5% per month.

This follow-up study allowed them to logically draw the same conclusions, namely that the differences on small fund data are larger than on large fund data and the older the data, the larger the differences. For example, Table 1.1-2 clearly shows the differences between the CRSP and Morningstar databases. Finally, the authors add that, among the 52 cases in which the difference in monthly performance is more than 0.5% between the two databases (18 for large funds and 34 for small funds), 51 cases contain a difference of 12 basis points per year or more on the alpha difference. They therefore consider this rule to be useful in identifying problem cases.

With this article, Elton, Gruber and Blake [75] demonstrate differences between the CRSP and Morningstar databases. This difference may exist also in other financial databases and in fact, their comparison is far from an isolated case.

WTI price comparison based on four sources

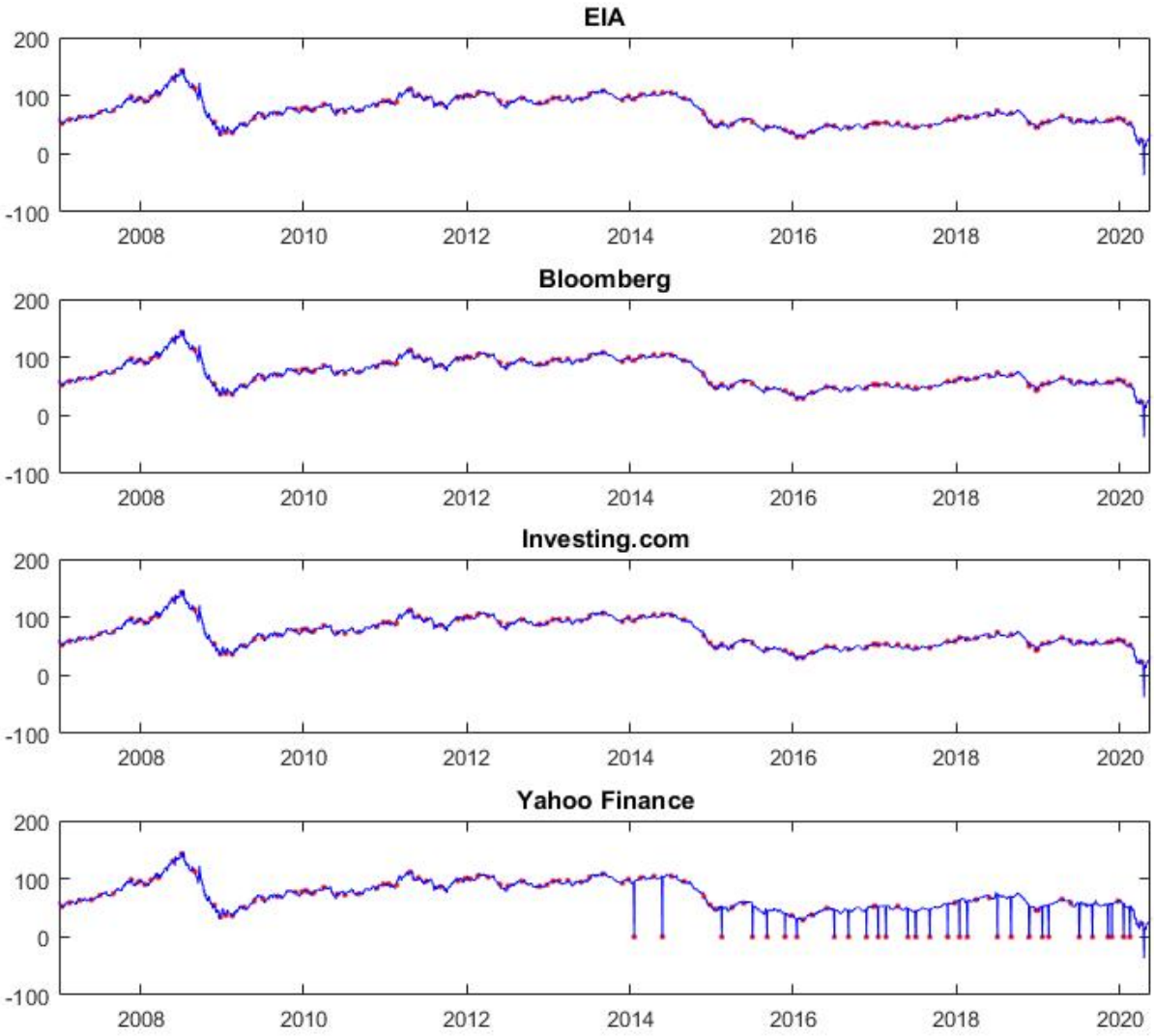
Indeed, it is not very difficult to find other cases of this kind. For example, data for U.S. oil company West Texas Intermediate (WTI) should be the same regardless of the source, but it is not. One obvious example is the price of a barrel of American oil, reports of which differ by source. In the following case study, the spot price of a barrel of oil from 01/01/2007-01/05/2020 is presented according to four different sources: U.S. Energy Information Administration (EIA; to be used as a reference), Bloomberg L.P., Investing.com and Yahoo Finance. The downloading prices correspond to the closing prices (i.e. the raw prices before the market officially closes). In this study, only weekends and January 1st were removed. As a result, each series can accommodate less than 3480 values (due to stock market closing dates not having been removed).

Even before analysis of the differences between the series, it may be preferable to use as a reference the data from the EIA, as the independent statistical agency attached to the U.S. Department of Energy, whose mission is to provide data and forecasts independent of political power. Thus, since the EIA is an independent institution, one might think it likely to yield different results than do other sources.

The first observation that can be made when comparing these four different sources is based on the missing data. As mentioned earlier, the series was downloaded without considering the stock market calendar and in particular the market closing dates. Thus, it is normal for the series to have missing data. Moreover, being a widely used series, they should all have the same number of data (the one due to market close), but they do not.

Figure 1.1-1 represents the spot price per barrel of WTI, from 2007 to 2021, from the 4 difference sources discussed: EAI, Bloomberg L.P., Investing.com and Yahoo Finance. It can be seen graphically (Figure 1.1-1) that from 2014 onwards, the series from Yahoo Finance has 32 values equal to zero. In many missing data management processes, when the information is missing, the data is reset to zero. These data are plausibly considered missing data, since leaving them at zero could bias further analysis.

Fig. 1.1-1: Spot price per barrel of oil WTI (in blue) and their missing data (in red) for four different sources: EIA, Bloomberg L.P., Investing.com and Yahoo Finance



Thus, 115 missing data points are observed in the EIA series, compared to 109 for Bloomberg L.P., 60 for Investing.com and 115 (i.e.,83 true missing data points and 32 zeroes considered as missing data too) for Yahoo Finance over the period studied. Rather surprisingly, most missing data is found in the series of the EIA, one of whose

missions is to collect data for the U.S. Department of Energy.

In addition, the data from Yahoo Finance also has 115 missing data points and upon closer examination, the EIA and Yahoo Finance series have exactly the same missing data. Yahoo Finance data might be sourced from the EIA source (the reverse would be surprising for a supposedly independent institution), or both sources might be sourced from the same common third source. This remains to be verified by comparing the values of the two series.

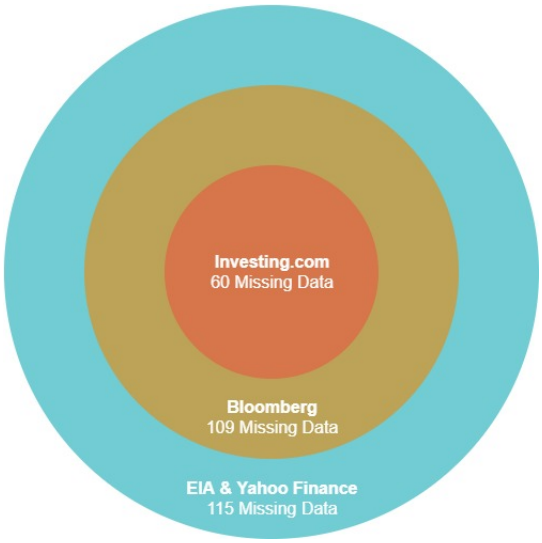
Moreover, for 60 dates, data are missing from all four series at the same time. In other words, the 60 dates with missing data for Investing.com are missing for the entire sample. On these dates, markets were likely closed. Furthermore, these 60 dates do correspond to data that were initially missing from the Yahoo Finance series (and not zeros that would have been considered missing), which supports the idea that no data were searched for that day because no quotes were expected.

Thus, apart from these 60 missing observations for all series, the missing data remaining to be analyzed are as follows: 55 missing data points for the EIA and thus for Yahoo Finance (as they have the same data); 49 missing data points for Bloomberg L.P.; and none for Investing.com.

It remains to be seen which of these remaining missing values are in common and which of the 6 missing data points present in EIA and Yahoo Finance and not in Bloomberg L.P. or Investing.com. After verification, the 49 missing data points in Bloomberg L.P. are also missing in the EIA (and thus Yahoo Finance) series. After verification, the 49 missing data points in Bloomberg L.P. are also missing in the EIA (and thus Yahoo Finance) series, but present in the Investing.com series.

Thus, in the sample studied, the missing data from Investing.com is also missing from Bloomberg L.P., which is missing from both the EIA and Yahoo Finance data. These relationships are visualized in Figure 1.1-2.

Fig. 1.1-2: Missing Data on spot price per barrel of oil WTI for EIA, Bloomberg L.P., Investing.com and Yahoo Finance series



Thus, to conclude, the sample of 3,480 observations studied here is composed of

- only 3,365 observations where the values of the four different series are available, because 115 observations have only missing data (for the four sources at the same time);
- no observation with only 1 missing data point and 3 observed data points, because if a data is missing it is necessarily missing in EIA and Yahoo Finance at the same time (the missing data included in Bloomberg L.P. and Investing.com are necessarily in EIA and Yahoo Finance);
- 6 observations with 2 missing data points (EIA and Yahoo Finance) and two available data (Bloomberg L.P. and Investing.com); and
- 49 observations with 3 missing data points and only 1 data point is available (from Investing.com).

A comparison of the series in values remains to be done. Given the great similarities obtained from the study of missing data, exactly the EIA source may contain exactly the same data series as does the Yahoo Finance source. For other sources, however, it is harder to make assumptions.

Above all, the mean of each series, not considering the missing data (sum divided by the number of observations) and the standard deviation may be of interest. The results are as presented in Table 1.1-3.

Tab. 1.1-3: Mean and standard deviation of four different sources of spot price per barrel of oil WTI

	EIA	Bloomberg L.P.	Investing.com	Yahoo Finance
Mean	72.7156	72.7443	72.5922	72.7156
Std dev	23.5831	23.5108	23.4919	23.5831

The means and standard deviations of EIA and Yahoo Finance may be the same, leading us to believe that the series are the same, but this will assessed later in this study. Moreover, Bloomberg L.P. and Investing.com have different means and standard deviations, giving the impression that there are three different series.

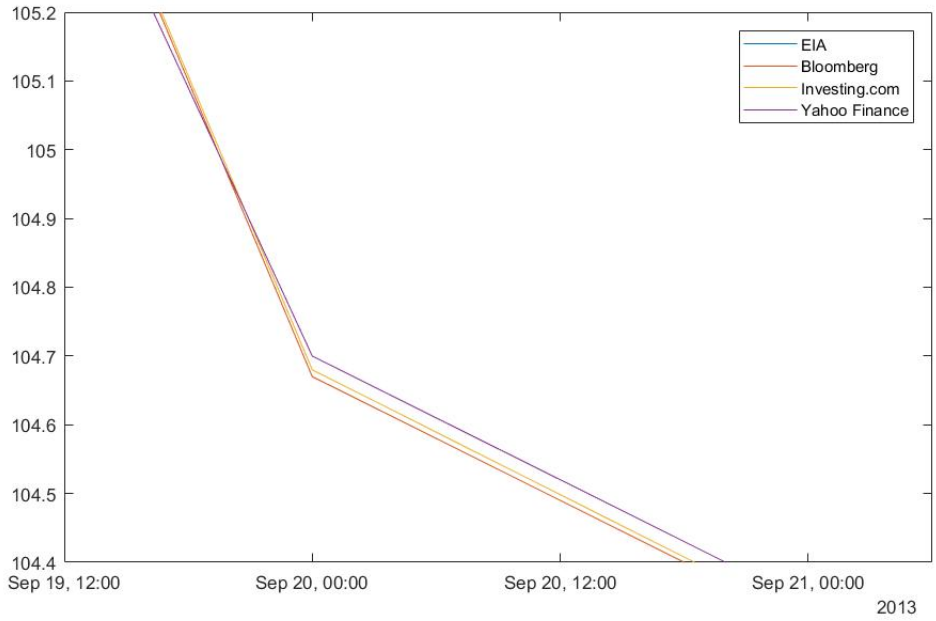
Thus, the first thing that is obvious is that the series from EIA and Yahoo Finance are identical. Not only did these two series have exactly the same missing data, but also the sum of the absolute differences between these two series amounts to zero. In other words, these two series are indeed the same.

Conversely, the sum of the absolute differences between the EIA series and the Bloomberg L.P. series or between the EIA series and the Investing.com series does not yield a result of zero (639.65 and 639.64, respectively). The sum of the absolute differences is calculated from the available data; since EIA is priced as a reference, 115 observations are not counted in this sum. These two results remain close, suggesting that the data between the Bloomberg L.P. series and the Investing.com series are similar.

Among the 3,365 available observations without missing data, in only 399 observations are the four series exactly identical to each other, representing only a little more than 10%.

Logically, the sum of the absolute differences between Bloomberg L.P. and Investing.com is calculated from 3,371 observations (109 fewer observations because the Bloomberg L.P. series contains 109 missing data points). Moreover, all of Bloomberg L.P.'s values, apart from the 109 missing observations, are identical to the Investing.com data, except for one: On September 20, 2013, Bloomberg L.P. posted a price of \$104.67 versus \$104.68 for Investing.com (and \$104.7 for EIA and Yahoo Finance). The gap is small but appreciable. Therefore, for one and the same date, it is possible to find three different quotations, without knowing which is correct (see Figure 1.1-3).

Fig. 1.1-3: Four sources and three different values for the spot price per barrel of oil WTI for the September 20, 2013



Thus, among the 3,480 observations, 3,365 observations are observed by all four series at once (115 observations have at least 1 missing data point). Among these, 3,365 observations available; the EIA and Yahoo Finance series are identical, as are the Bloomberg L.P. and Investing.com series, except for one observation where the Bloomberg L.P. and Investing.com data differ (i.e., on September 20, 2013). In addition, 60 observations are missing for the four series at a time, so no data are available for these 60 dates, suggesting that the market was closed on these dates. At present, the 55 observations with at least 1 missing data point and 1 observed value at a time (the 115 with at least one missing data points minus the 60 with no observed value for the four series) have yet to be analyzed.

Of these 55 observations, no observed values from EIA and Yahoo Finance are available, 49 have only 1 observed value (that of Investing.com) and 6 have 2 observed values (that of Bloomberg L.P. and Investing.com).

First of all, concerning the 6 cases composed 2 two missing data points (EIA and Yahoo Finance) and 2 observed data points (Bloomberg L.P. and Investing.com), after verification, the 2 observed data points generally differ from the previous day's data. This difference suggests that the data were not frozen from one day to the next and that there were many quotes on that date. For this reason, it seems odd that they are missing for the EIA and Investing.com.

Finally, for the 49 observations with 3 missing data points and 1 observed data point

(that of Investing.com), only 1 of the 49 observations was taken from the previous day: On April 3, 2015, the only observable data is that of Investing.com at $49.14 and the data from April 2, 2015, is at exactly the same. The data would have or be repeated from one day to the next, as some data recovery processes resume the previous day's data when the information is missing or unavailable. If this procedure had been in place, other cases like this one would have been present and this is the only case in which the data is the same as that of the previous day. Therefore, either the market has a real quotation, which the other sources would not have been able to record, or Investing.com has itself resorted to data-completion methods to avoid missing data in its history.

Thus, this example shows that it is not so complicated to find a data series available from several different non-homogeneous sources. Indeed, the case of the spot price of the WTI oil barrel clearly shows that for the same date, prices differ by source (up to three different prices for the same date in the previous example) — some are even missing quotes whereas others are available.

Finally, this example raises several questions: How does one explain differences between databases? Is one specific quotation not supposed to be identical? How should the source be chosen in this case? Do the sources use completion methods to reduce their rate of missing data?

Given these questions, it is not always easy to choose a data source and even if one source is preferable, there is nothing that says that a source unused today will not be used tomorrow. Thus, as a preventive measure, banks today prefer to store all the historical data they come across. This data could eventually be used for production, comparison, or verification purposes.

This is especially the case for Natixis, which has created an application that provides access to databases storing a multitude of different sources for many types of data. For several years now, the bank has invested in numerous databases that have been centralized in one application, allowing all bank employees to access them. The application hides behind it many historical databases of various lengths and is updated daily. Thanks to this kind of application, it is possible to access many different sources for spot prices, interest rates, repo rates, dividends, volatility surfaces and so forth.

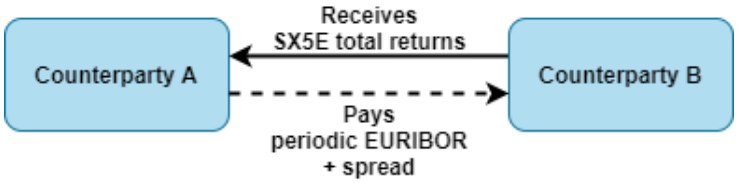
More specifically, regulatory requirements have led Natixis to rely on historical index repo rates. Two sources have been identified: the repo rate used in index forward contracts and the repo rate used in index total return swaps (TRSs). According to Eurex [78], an index forward contract is a contract between two counterparties, where one party agrees to buy the underlying equity index from the other party for a given price at a given future date. The forward price of the equity index S, under risk-neutral probability $\mathbb{Q}$, starting at time 0 and with maturity T is given as follows:

$$F_{0,T} = \mathbb{E}^{\mathbb{Q}}(S_T) = S_0 e^{(i-r-d)T}, \tag{1.1-1}$$

where S_0 is the index spot at time 0, i is the funding rate, r is the repo rate and d is the dividend rate.

An index TRS is where the buyer and the seller of the contract agree to exchange, at a periodic dates, two cash flows based on a notional amount of the reference index. An equity index TRS can be decomposed into two legs: on one side, there is the equity amount that is the total return performance of the index (including dividends) and on the other side, the floating rate amount that is quoted as a floating interest rate plus a fixed spread (see Figure 1.1-4).

Fig. 1.1-4: Total Return Swap on SX5E (Source: Eurex)



with *spread* corresponds to the TRS fixed spread. Mathematically, an index TRS can be written as as follows:

$$\mathbb{E}^{\mathbb{Q}}\left(\frac{S_T}{S_0}\right)e^{dT} = e^{(i+spread)T} \tag{1.1-2}$$

The left side of the Equation 1.1-2 corresponds to the equity amount and the right side, to the floating rate amount.

These two historical datasets were saved by Natixis but of mediocre quality. They were composed of missing data and outliers that made them non-modellable according to the regulator. The repo rate series from forward contracts are saved in order to match the price of the forward contract itself. Since the only unknown in the Equation 1.1-1 is the repo rate, it is possible to implicitly solve the equation, but this method implies that if there is strong movement by one or more other variables, the repo rate will be directly affected and non-representative jumps will be observed. For this reason, another source for the repo rate was sought.

However, it turns out that it is possible to approximate a TRS contract by a forward contract, starting with Equation 1.1-2, as follows:

$$\begin{aligned} & \mathbb{E}^{\mathbb{Q}}\left(\frac{S_T}{S_0}\right)e^{dT} = e^{(i+spread)T} \\ \leftrightarrow \quad & \frac{\mathbb{E}^{\mathbb{Q}}(S_T)}{S_0}e^{dT} = e^{(i+spread)T} \\ \leftrightarrow \quad & \mathbb{E}^{\mathbb{Q}}(S_T) = S_0 e^{(i+spread-d)T} \end{aligned} \tag{1.1-3}$$

By equalizing Equation 1.1-1 and le result of Equation 1.1-3, it appears clearly that the TRS spread is nothing more than the opposite of the repo rate. This is true if

the rate of the forward contract and the rate of the TRS is the same, generally it is the 3-month Euribor. TRS spreads can take both negative and positive values as the repo rate can be positive or negative. A negative repo rate means that the buyer pays an interest to the seller who is borrowing cash. According to Billio and Varotto [34], average repo rates for the euro are negative, and have been since mid-2014.

This relationship between the repo rate and the TRS spread is very convenient, as the TRS spread is quoted in the markets. Brokers provide traders with the TRS spread when a trade will take place. It is also for this reason that the repo rate deducted from the TRS spread is considered to be a market data, as it is quoted in the market independently of the rest. Thus, using a TRS contract to deduce the repo rate allows one to recover the true market value and not a value that would be impacted by shocks to other variables.

Thus, for the same dataset (index repo rates here), it was possible to find two different sources, one of which is real market data. Thus, this section clearly demonstrates the data-quality issues around the chosen source. Elton, Gruber and Blake [75] who show in their article the undeniable presence of differences between the CRSP and Morningstar databases. The case study comparing EIA, Bloomberg L.P., Investing.com and Yahoo Finance spot price per barrel of oil WTI shows differences in terms of missing data and value. Finally, there is the case of Natixis bank, which has implemented an application that allows access to a multitude of sources and which, for one data item, searches different sources in order to find the right market data.

Many different data sources can thus provide the same data. However, as has been said, very often these bases differ significantly, leading sooner or later to the need for a decision to be made. Precisely within this multi-source framework has Markit set up the Markit Totem consensus data to provide a global view of the market. Indeed, Markit offers consensus prices, more precisely data based on the consensus over-the-counter (OTC) market prices of the main active market participants for each product. In other words, Totem consensus data is an average of the prices of the largest market participants. In particular, it provides an idea of the prices that apply on OTC markets and thus ensures a true reflection of the market. The primary purpose of these Totem consensus data is to provide accurate market prices (validated after rigorous controls according to Markit).

The very existence of this consensus data shows, first of all, that there is actually a market, because the reporting of market prices highlights the existence of the market, even if it is not very liquid. These consensus prices are not meant to be considered as real prices, but only serve to show that a market exists. Secondly, the consensus prices, calculated as an average, clearly show that there is no single price. On the other hand, no information is given about the dispersion of the collected prices, and thus the standard deviation around the consensus prices.

Finally, consensus data is a new type of data source, which provides a global view of

the market and Markit claims that this same consensus source is used both by auditors and by major regulators in their oversight of OTC derivative markets. Finally, this type of data could be the perfect benchmark to help impute missing data from a series or to generate proxy data, in case no data are available.

Moreover, the regulators are well aware of the existence of these multi-sources. The standard BCBS 239 [24] (detailed in Section 1.3.1) is a regulation dedicated to the data management and the data quality of banks. The Basel Committee states, in Principle 3, that the bank must reconcile data with the different sources it uses and that it should, to the extent possible, use only one authoritative source. Thus, banks must choose their sources, but also document and argue them in order to inform the regulator.

1.1.2 Data accessibility

Another aspect of data quality is data accessibility. While the previous section highlighted the existence of many sources for the same data, finding certain data can sometimes be complicated. In finance, it is probably easier to find data than in other areas of research, as there are quite a lot of data provider: Bloomberg L.P., Thomson Reuters, Six Financial Information, Markit, FactSet and so forth. These suppliers are generally very expensive, but they have the advantage of offering features and options (quotes from the live market for example) that free providers do not offer. When the use does not require one to have live data, but simply to have access to historical data, there are less expensive platforms, or even totally free ones: Swissquote, Yahoo Finance, Investing.com, Boursorama, Investing and so forth.

The data is not always accessible or easy to access, however. In the previous case, where Natixis use TRS contracts to approximate the repo rate, the data is transmitted directly from the broker to the trader, via the Bloomberg L.P. chat room. These data are transmitted directly in conversations between broker and trader, so it is difficult to set up an automatic data recovery system. If the data arrived by e-mail, it would be possible to consider an automatic process to save the data without any problems. On the other hand, since the data is transmitted in conversations, setting up an automatic process is nearly impossible for several reasons. There is no guarantee that the information comes from the same broker, the time will not always be the same, or the format of the data can change and so forth.

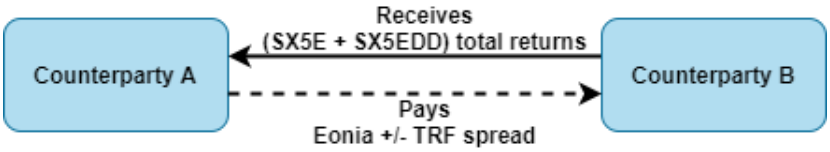
As it is impossible to automate the backup of TRS spread data, traders are obliged to backup the data manually. The TRS spread series is thereby exposed to the risk of missing data: the trader in charge of saving TRS spreads manually may forget to do so. Even before they had regulatory motivations (such as risk measures calculations), Natixis traders were saving these data for their personal use. Then, these data were saved in multi-source databases to ensure that they were not lost and that they could be consulted and used by everyone. In the case where no trader would have taken the

time to back up these data on a daily basis, however, historical TRS spreads would have been very complicated or even impossible to find.

Therefore, there would have been no alternative but to use the repo rates implied by the levels of forward contracts, which react to movements in other variables and are not true market data. In the case where TRS spreads are used to approximate repo rates, an implicit repo rate can still be used in the forward contract formula, as seen previously. Even if this implicit method makes the repo rate sensitive to other variables, it is still better than having no data. On the other hand, if the intention is really to have historical data of TRS spreads, it will be impossible to have true market data. TRS spread data cannot be bought or accessed. The only solution will be to compute these TRS spreads with Equation 1.1-2, using the historical data of the other variables, which may themselves contain some bias. The bank must therefore ensure (and document) that the TRS spreads are not biased and that they are representative of the market. Otherwise, it would be preferable to use the original source.

Unlike the TRS, Eurex's website makes available (and keeps available for two weeks) all the data necessary to calculate total return futures (TRF), including TRF spreads. The TRF is a new financial instrument introduced on the markets in December 2016 for the Euro Stoxx Index, which is very similar to the TRS. Figure 1.1-5 shows that the equity amount is the same for a TRF as for a TRS (Euro Stoxx 50 Distributions Point Index [SX5EDD] allows for the accounting of dividends). Only the floating rate amount differs. Unlike the TRS, which uses the 3-month Euribor rate as a reference funding rate and would exchange net cash flows quarterly, the TRF uses the euro overnight index average (EONIA) as a reference funding rate that is reset daily. Moreover, in a TRF the total returns are incorporated into the daily settlement price and the daily P&L is paid as a daily cash flow via the regular variation margin process.

Fig. 1.1-5: Total Return Futures on SX5E (Source: Eurex)



These TRF spreads can be copied by downloading the previous day's data (available for 2 weeks) from the Eurex website. Otherwise, there remains the possibility to buy these data from Eurex, even if it is extremely expensive. It is possible to deduce the repo rate in the same way as above, but by incorporating an adjustment factor due to the differences between the TRF and the TRS.

Thus, in the case of the repo rate, different methods of calculation are viable, but which is best remains uncertain. It thus becomes possible to infer the repo rate through TRS or TRF.

Of course, the case of repo rates is an example that works because it is possible to retrieve information by various means, but some data are very complicated to retrieve if no data could be recorded in the past and this is especially the case for very old data.

Moreover and quite obviously, it goes without saying that data from the primary market are much more readily available than data from the secondary market. As soon as a product is traded on the OTC market, not all the information is publicly available to all market participants. It then becomes possible to pay for these data, although financial data is traded at very high prices. Moreover, the problem becomes more complex when the data are no longer simply historical prices of financial products but historical risk factors. These are the issues that reporting obligations via trade repositories aim to resolve (see Section 1.3.2). Many regulations - such as the Principles of Financial Market Infrastructure [58] in 2012, the European Market Infrastructure Regulation [168] in 2012, or even the Securities Financing Stability Board [169] in 2015 - have included in their principles the obligation to report all derivative transactions via trade repositories. This obligation is supposed to encourage market participants to report all these activities, particularly on OTC markets, in order to reduce market opacity and increase the transparency of the risks to which each market player is exposed.

Finally, the data may not be accessible because it simply does not exist, resulting in missing data in the time series. The inaccessibility of the data may simply be due to a lack of supply or demand or to the illiquidity of the products, meaning that the market does not exist (as explained in Section 2.2.1).

1.1.3 Treatment of outliers

Finally, another aspect of data quality is the validity of the data, in other words, ensuring that the data are consistent and that they do not include unrepresentative data points, also called outliers. These outliers can be the result of shocks observed in markets or of processing errors. An analysis done in the presence of these data and without reprocessing will necessarily lead to biased results and if they are numerous, they can lead to a completely erroneous study.

The issues of outliers concern any quantitative studies, whether financial or other research. All are affected by the presence of outliers in their sample that could potentially bias their results. In the article by Adams, Hayunga, Mansi, Reeb and Verardi [2] in 2019, presenting multivariate outlier detection, they submit a preliminary study, showing a real awareness of the data quality and especially the outliers. These authors have managed to identify, among a multitude of recent financial articles, those that have identified outliers in their sample and how they are treated. This study was based on articles published each year in the following financial journals: *Journal of*

Finance, *Journal of Financial Economics*, *Review of Financial Studies* and the *Journal of Financial and Quantitative Analysis*.

Adams, Hayunga, Mansi, Reeb and Verardi [2] show that over the past 30 years, articles have increasingly mentioned outliers in their research. Indeed, in the early 1990s, less than 10% of papers mentioned outliers, compared to about 30% since 2010.

Overall, they show that the papers increasingly refer to outliers and that one-third of the papers mention outliers and the ordinary least square (OLS). The outliers therefore seem to be linked by the literature with the OLS framework.

Tab. 1.1-4: The incidences of articles with outlier mention, with OLS mention and with OLS and outlier mentions from 1988 to 2017 (Source: Adams and al. 2019 [2])

Year	All papers JF, JFE, RFS, JFQA	% All papers mentioning outliers	All papers utilizing OLS	% All papers utilizing OLS	% OLS papers mentioning outlier
1988	194	7%	64	33%	17%
1989	209	5%	71	34%	13%
1990	228	7%	89	39%	15%
1991	195	5%	58	30%	10%
1992	201	6%	66	33%	15%
1993	207	7%	77	37%	9%
1994	182	13%	68	37%	21%
1995	201	12%	58	29%	28%
1996	199	14%	86	43%	26%
1997	223	11%	106	48%	19%
1998	201	14%	85	42%	28%
1999	208	8%	96	46%	23%
2000	216	9%	90	42%	23%
2001	220	16%	102	46%	26%
2002	243	13%	110	45%	25%
2003	241	18%	118	49%	34%
2004	250	15%	105	42%	28%
2005	248	23%	137	55%	35%
2006	259	24%	163	63%	32%
2007	288	23%	185	64%	30%
2008	298	27%	200	67%	37%
2009	379	25%	252	66%	35%
2010	381	24%	234	61%	38%
2011	400	26%	279	70%	34%
2012	365	27%	231	63%	27%
2013	364	30%	178	49%	44%
2014	316	28%	152	48%	33%
2015	328	30%	137	42%	36%
2016	355	31%	163	46%	36%
2017	386	32%	195	51%	37%

For example, in 1988, 7% of papers mention outliers, 33% used an OLS and 17% of the papers that used OLS mention outliers.

Table 1.1-5 presents the methods used to deal with outliers, among the articles collected between 2008 and 2017.

- *Winsorizing*: replacing extreme values with the nearest value not considered an outlier (according to a quantile);
- *Trimming*: removing all outliers without replacement as the pairwise deletion (see Section 2.2.2); and
- *Dropping*: dropping all the observations containing outliers as the listwise deletion (see Section 2.2.2).

These three methods identify and treat outliers in a univariate setting. The percentages total more than 100% are due to multiple treatments in some papers.

Tab. 1.1-5: Outliers mitigation methods used in the financial journal from 2008 to 2017 (Source: Adams and al. 2019 [2])

Year	% Winsorize	% Trim	% Drop	% Winsorize, trim and/or drop	% All other treatments
2008	35	11	31	75	38
2009	46	14	21	80	25
2010	41	24	10	75	29
2011	53	12	12	78	29
2012	64	20	6	91	21
2013	54	15	39	109	11
2014	35	13	30	78	35
2015	69	7	17	93	14
2016	56	26	21	103	12
2017	72	10	8	90	5
Average	52	16	17	85	24

On average, 52% of papers winsorize, 16% trim, 17% drop missing data, and 85% used at least two of these methods. 24% of papers used other treatment.

They find that, over these 10 years of study, the *winsorization* method (used by more than half of the articles) tends to be applied more often than are *trimming* and *dropping*. On the other hand, almost a quarter of the articles use other ways to deal with outliers.

Their preliminary study shows that they are well aware of the quality of the data and especially the outliers in this case. More and more articles dealing with a financial theme mention outliers and hence, use a method to get around them.

Furthermore, outliers are common in samples and can be easily found in a review of the sample of the spot price of the WTI oil barrel used previously in Section 1.1.1. In this example, the price series from Yahoo Finance contained 32 zeros. These are

very basic outliers, but they still need time to be treated. These were obviously missing data that had been initialized to zero by default and thus, created artificial spikes in the series. Figure 1.1-1 reveals anomalies in the price series.

Thus, if an analysis is carried out on these data, without an identification of outliers in the series or without attempts to correct or remove them, the results obtained may be highly biased and the analysis will certainly be discredited.

There are different ways to detect outliers, such as Tukey's criterion (by Tukey in 1986 [202]), Dixon's Q test (by Dixon in 1950 [69]), Grubbs' test (by Grubbs in 1950 [100]), Tietjen-Moore's test (by Tietjen and Moore in 1972 [199]), Student's generalized test of extreme deviation (by Rosner in 1983 [171]), Thompson's modified tau test (by Thompson in 1935 [197]), Peirce's criterion (by Peirce in 1877 [165]) and many more.

It would be interesting to conduct one or more of these tests on the sample used, to determine which data can be considered as outliers and which data should be removed or replaced in order not to bias the analysis. This would take time, though and it should not be forgotten that this PhD thesis deals with missing data and not outliers.

Notably, on the other hand, issues with outliers and issues with missing data remain closely linked. Once a value has been considered an outlier by a given test, the analyst has two choices: remove it from the sample, or replace it with a value from a chosen completion method. The problems of outlier management are thus followed by those of missing data.

Finally, missing data represent only one consideration of data quality. Data quality raises many more questions than those pertaining only to management of missing data, encompassing issues of data accessibility, data consistency, data validity, data frequency, and so forth. Missing data are therefore part of a much broader research theme and are coming back to the center of attention, particularly because of new regulations that raise the question of the sustainability of banks.

The BCBS 239 [24] standard is a regulation entirely dedicated to data and aimed at solving problems of poor data quality. This standard is composed of 11 principles aimed at pushing banks to arm themselves with the best possible data quality, as well as solid governance, architecture and infrastructure, in order to be able to respond to reporting requests concerning the management of their risk factors. Principle 3 (cited above) emphasizes the accuracy and integrity of risk data and the same data should be used to produce accurate and precise risk management reports to represent the risks incurred by the bank as accurately as possible. This issue is highly topical for banks, because despite an initial implementation date of January 2016, the BCBS 239 [24] standard has still not, at the present time, been fully implemented by the banks concerned, all systemically important banking institutions worldwide. This new regulation implies huge changes in production methods, which requires a significant mobilization of resources on all fronts, which is why the regulator has granted additional time to banks.

As data quality is a vast subject, this PhD thesis will focus precisely on missing data and the problems they create, which will be presented in the following section.

1.2 Missing data issues across research areas

Missing data are an integral part of data quality, and the problems they create are not limited to a particular field of study. Missing data are relevant to any quantitative study, be it in finance (as in this PhD thesis) or epidemiology, cancerology, oceanology, economics, signal or image processing, and others. The literature in this field is large, so it will be interesting to highlight the multi-domain issues of missing data and then focus on the problems they create in a financial framework.

1.2.1 Data are essential in all quantitative domains

Missing data are present in any area. As soon as a database is used, it is likely to contain missing data, so missing data is therefore a vast subject that brings together various fields of expertise.

Missing data and in particular, the reconstruction of missing data are subjects examined in various fields, both academic and corporate. This is the case of a report written in 2006 by the Société de Calcul Mathématique SA with the scientific participation of the University of Donetsk in Ukraine [190]. This report addresses the company Veolia, which commercializes its services in water management, waste management and recovery and energy management.

This report presents Veolia's need to reconstruct missing data. The data to be reconstructed are those derived from time series of the flows of 19 rivers in the Vendée, a French department. Veolia called upon the Société de Calcul Mathématique SA to complete the missing data present in this database. The importance of these reconstructions is that these data are used on a daily basis to know whether a dam should be closed or not or if withdrawals can be authorized, but also, in the longer term, whether certain groundwater can be accessed or new constructions planned. Veolia's decisions depend on the historical data in their possession: the flow of the 19 rivers of the Vendée from October 1967 to January 2006, namely more than 30 years of daily data.

The missing data are due to some rivers being correctly measured and others not. The data can therefore be highly irregular from one river to another: some series do not start until 2001 and others have large gaps. The percentage of missing data in the series ranges from 1% to 89%. Obviously, the older the data, the greater the probability of missing data. They use a conditional probability table based on the most correlated series to reconstruct missing data. Thus, the completion of missing data is applied here to potamology (study of rivers) data.

There are also many papers dealing with marine data that use completion methods to optimize the use of their database. This is notably the case of Lagona and Picone in 2011 [137], who work with wave height and direction data in different wind conditions. Their database is composed of missing data, which makes their study complicated to set up. Their database is composed of series with a maximum of 16.3% missing. Faced with these missing data, conducting an unbiased study can become complicated. Because searches on marine data often contain missing data, the authors propose in this article to use an expectation-maximization (EM) algorithm (see Section 2.4.6) calibrated on marine data to impute them. They show that their model can explain most of the variability in the data and re-impute artificially suppressed values with reasonable precision. The needs between research areas do not substantively differ, because an improved version of the same algorithm (see Section 2.4.6 for more details) was used in this PhD thesis to be applied to financial data.

There are, of course, also many examples in medical research. In 2010, Jerez, Molina, García-Laencina, Alba, Ribelles, Martín and Franco [124] compared different completion algorithms applied to breast cancer data-set. Their data were collected from the "El Álamo" project, one of the largest databases on breast cancer in Spain and is used in 2004 by Martín, Llombart-Cussac, Lluch, Alba, Munárriz, Tusquets, Barnadas, Balil, Dorta y Picó [151]. The data-set includes demographic, therapeutic and recurrence-survival information from 3,679 women with operable invasive breast cancer diagnosed in 32 hospitals belonging to the Spanish Breast Cancer Research Group between 1990 and 1993. The missing data represent 5.61% of the overall data-set (considering every attribute for every patient case as data). A share of 45.61% of cases had at least one missing data point, meaning that about half of the observations would not have been used in the case of listwise deletion (presented in Section 2.2.2). The data is concentrated among five variables (out of nine in total) with proportions ranging from 0.19% to 42.84%. The authors use several different algorithms to fill in these missing data, including the K-NN method, presented in Section 2.4.3, which yields some of the best results (see Section 2.4.3 for details on the results). The K-NN method was also used in this PhD thesis to fill in missing data on financial series.

Some studies have dealt with the completion of missing data applied to epidemiological data. Indeed, in 2014, Shah, Bartlett, Carpenter, Nicholas and Hemingway [187] have worked on the calibration of the multivariate imputation by chained equations (MICE) model (see Section 2.4.7) on epidemiological data. For this work, they created samples from the cardiovascular disease research using linked bespoke studies and electronic records (CALIBER) database, randomly removing data in order to fill them. This database comprises linked electronic health records collected regularly in England, including data from primary care, hospital admissions, the national register of acute coronary syndrome and the national death register. This corresponds to a sample of about 53,000 observations, where only just over 10,000 observations are complete.

They therefore worked from a sample of more than 13,000 observations, containing data originally missing but not in the variables of interest and they randomly added missing data to the complete observations according to a specific pattern.

They then used the multivariate chain equation imputation model (MICE) to fill in the missing values for these epidemiological data. The same algorithm is applied in this PhD thesis to see how effective it is for financial data.

Finally, in a purely computational world, in 2010 Pan and Li [163] worked on the imputation of missing data on the data of wireless sensor networks. They explain that missing sensor data is inevitable due to the inherent characteristics of wireless sensor networks, creating difficulties in various applications. Solutions to this problem demand that missing data be imputed as accurately as possible. In their study, Pan and Li use the improved K-NN method (see Section 2.4.3), considering spatial more than temporal correlation of sensor data. Their experimental study conducted on two different data sets showed stable and reliable performance.

Many further examples of the application of missing-data completion methods exist in various fields of research: neuroscience (see Rubin, Witkiewitz and and Reilly [178] in 2007), image reconstruction (see Dass and Nair [66] in 2011), speech recognition (see Josifovski, Cooke, Green and Vizinho [125] in 1999) and so forth.

1.2.2 General impact of missing data

The superabundance of articles regarding the completing missing data, from across research fields, indicates that the problem concerns all empirical studies. It was presented at the beginning of this chapter that it was complicated to create a sample without any missing data, hence the interest in the subject.

As mentioned earlier, the presence of missing data can bias statistical analyses and thus reduce statistical power. The statistical power of a test is the probability of rejecting the null hypothesis knowing that it is false. According to Verma and Goodale [207] in 1995, the power of a statistical test depends on three criteria: significance level, size effect and sample size. Only the latter is used to control power, however. The significance level α corresponds to the probability of a Type I error (the null hypothesis is rejected when it is true) and this probability is set arbitrarily. Generally, in the literature, this level of significance is set at 5%. The size effect is an index measuring the strength of the association between the populations of interest. An important effect may not be detected because of a lack of power. Different measures can be applied to evaluate this size effect, but the most well-known is the omega-square introduced by Keppel [128] in 1991. As with significance level, effect size can be assumed to be fixed because, generally, it is impossible to change the effect of a particular phenomenon. Verma and Goodale [207] explain than the sample size is then the only parameter than

can be used to design empirical studies with high statistical power. In general, the larger the sample size, the more powerful the statistic will be and relatively small expected effect sizes require substantial sample sizes to achieve reasonable power. Dallal [61] in 1986 demonstrated the relationship between statistical power, size effect and sample size, for two different levels of significance ($\alpha = 1\%$ and $\alpha = 5\%$). The results of his study shows that for a given effect size, the sample size must be increased larger to maintain a power level. Moreover, the same is true for the level of significance: for a smaller α and for a given effect size and power, the sample size must be lager. In order to maintain high power with a small effect size, the sample must be larger.

Hence, if the sample size has as much impact on the power of a test, one should try to have as little missing data as possible or at least find a way to complete them without loss of power. Often, when one or more data points are missing in an observation, the whole observation is removed from the sample, this method is called listwise deletion (see Section 2.2.2) and leads directly to a drastic reduction of the sample size in case of many missing data points. In 1977, Kim and Curry [130] introduced the subject by pointing out that with only 2% of missing data in their sample of 10 variables, listwise deletion leads to the deletion of 18.2% of the observations: almost one-fifth of their sample. In an even more extreme case, in 2001, Allison [5] explains that for a data sample of $1,000$ observations and 20 columns, if only 5% of the data are missing in each of these columns, then it is expected to have only 360 fully observed observations, which means 64% of the sample partially missing. Thus, missing data directly impacts analyses based on the dataset, especially if they are ignored or treated incorrectly. In 1999, Quinen and Raaijmakers [166] claimed that listwise deletion leads to an excessive loss of statistical power from 35% for the smaller scales with 10% missing data, to 98% for the larger scales, with 30% missing data.

In addition to the power loss, missing data will also result in biased estimates. In 1999, Roth, Switzer and Switzer [172] showed, based on a simulated sample, that missing data could lead to biases. To do so, they simulated samples with different pattern deletion, each time with 20% missing data. Subsequently, they applied listwise deletion (see Section 2.2.2), as well as four different completion methods (mean substitution across individuals, mean substitution across items, hot-deck using a Euclidean distance function, and regression imputation with no error term). In their study, Roth, Switzer and Switzer [172] find that the root mean square error (RMSE) obtained from listwise deletion is generally three to four times higher than that obtained by imputation methods. This difference arises from the loss of information involved in this method. Their study is based on a sample size of 400 observations with 20% of removed data and they show that listwise deletion leads to consider, on average (out of 100 simulations), only 266.67 observation. In other words, for 20% of missing data, more than a half of the sample becomes unused. Furthermore, in general, the authors find that imputation methods recreate estimates better than does listwise deletion. This is one of the conclusions that Kim and Curry's [130] had in 1977, where a simple average imputation

gave the best result for listwise deletion.

Based on these previous works, it turns out that missing data have two major impacts: reduction of statistical power by reducing sample size and biased estimators resulting from these data. More generally, it does not matter whether the data are used for classical data analysis or applied to statistical tests, regressions, or even more-sophisticated models. For the results to be credible, the data must also be credible. Reliable results cannot be obtained if the data themselves are not reliable. This insight is even more important when the survival of a financial institution depends on it and that is what will be presented in the rest of this chapter.

1.2.3 Financial data affected by missing data

The topic of missing data is gradually affecting the world of management research and thus, the world of finance research. As Adams, Hayunga, Mansi, Reeb and Verardi [2] did in 2019, examining outliers (presented in Section 1.1.3), so in 2005 did Tsikriktsis [201] study the treatment of missing data in the literature based on 103 articles published in the *Journal of Operations Management* between 1993 and 2001. The handling of missing data in his study is assessed on the basis of two raters who independently evaluated the articles. In case of disagreement, the raters discussed the differences and then agreed. The results of that study are presented in Table 1.2-1.

Tab. 1.2-1: Use of missing data techniques (MDT) in *Journal of Operations Management* (Source: Tsikriktsis, 2005 [201])

	Articles
Item non-response discussed	
Yes	34 (33%)
No	69 (67%)
Agreement between raters (N=103)	93.2%
Method for arriving at MDT judgment	
MDT stated in article	4 (8.9%)
MDT inferred	41 (91.1%)
Agreement between raters (N=145)	95.5%
Missing data technique	
Listewise deletion	45 (100%)
Other	0 (0%)
Agreement between raters (N=45)	93.3%
Sample size	
Average sample size	263.22
Average number of missing data	34.32
% of missing data	13.04%

Among the 103 articles, 34 articles mention the presence of missing data. Among 45 articles mentioning missing data treatment all use listwise deletion. On average, the missingness proportion is 13.04%.

First of all, 67% of the articles in this study do not mention the presence of missing data or, if so, how they are treated. According to Tsikriktsis [201], researchers do not necessarily see the value in talking to readers about the missing data in detail and some simply do not deal with the missing data in their data analysis.

Moreover, he finds very few articles that explicitly discuss the method of treatment of missing data used. Among the 45 articles in which the reviewers agreed, only four articles discuss explicitly the treatment of missing data, but all the 45 refer to listwise deletion (see Section 2.2.2). Otherwise, in general, it can be inferred that missing data have been processed on the basis, for example, of the sample size or the degree of freedom used.

In almost half of the articles in this study (45 out of 103), the authors used particular expressions that let raters interpret their way of dealing with the missing data: for example, "the analysis is based on 145 completed questionnaires" or, "160 usable questionnaires were returned". Knowing that it is almost impossible to have no missing data in a large-scale survey, the authors necessarily used listwise deletion to obtain these so-called "complete" questionnaires. Globally, Tsikriktsis [201] explains that researchers do not always mention their technique for dealing with missing data, presumably because they do not see the need for it, as they are likely to use simple techniques such as listwise deletion or pairwise deletion (see Section 2.2.2), or because they want to avoid potential reviewer comments in the publication process.

Finally, he finds that 13% of the sample data was missing on average. As mentioned earlier, such a large proportion of missing data necessarily and negatively impacts statistical power (because listwise deletion seem to be the preferred technique) on statistical power. Tsikriktsis [201] adds that in one of the 103 studies, 301 observations out of 576 contained missing data, leading the authors to conduct their study on the basis of less than half of their sample.

Through its study, Tsikriktsis [201] shows that a large number of articles in the *Journal of Operations Management* do not mention missing data, or in the constrained case, the treatment remains rudimentary (listwise deletion). Although this sample covers only a portion of management articles, there is every reason to believe that the trend would be the same in other management journals.

Just as Tsikriktsis [201] did for management, Kofman and Sharpe's [133] 2003 study aims to highlight the management of missing data in the field of applied finance. To do so, they examine 946 empirical articles published between 1995 and 1999 in the five largest banking and financial journals: the *Journal of Banking and Finance* (JBF), the *Journal of Finance* (JF), the *Journal of Financial Economics* (JFE), the *Journal of Financial and Quantitative Analysis* (JFQA) and the *Review of Financial Studies* (RFS). Their results are presented on Table 1.2-2 (in parenthesis, the total number of articles and in brackets, the number of empirical articles appearing in the journal).

Tab. 1.2-2: Literature survey of incomplete data in finance (Source: Kofman and Sharpe, 2003 [133])

Journal	Articles acknowledging missing values	Missing values occur in			Missing values occur in a	
		Independent variable	Dependent variable	Both	Cross-sectional analysis	Time-series analysis
JBF	67 (397) [270]	34	14	19	50 (215)	40 (220)
JF	98 (365) [292]	69	16	13	85 (246)	37 (252)
JFE	53 (222) [201]	34	8	11	51 (189)	20 (178)
JFQA	20 (135) [95]	13	3	4	14 (77)	9 (84)
RFS	19 (176) [88]	10	5	4	13 (70)	17 (79)
All	257 (1295) [946]	160	46	51	213 (797)	123 (813)

JBF: *Journal of Banking and Finance*, JF: *Journal of Finance*, JFE: *Journal of Financial Economics*, JFQA: *Journal of Financial and Quantitative Analysis* and RFS: *Review of Financial Studies*.
The total number of articles is in parenthesis.
The number of empirical articles appearing in the journal is in brackets.
Among the 1,296 articles of the study, 257 mentioned missing data, 160 with independant variable, 46 with dependent variable, 51 with both, 213 in cross-sectional analysis and 123 in time-series analysis.

Of all these articles, only 257 explicitly acknowledge the presence of missing data in their database, representing only 27%, a little less than what Tsikriktsis [201] had found in his study dealing with operational management articles (respectively 33%).

The authors explain that in this investigation, being very complex due to non-standard and sometimes incomplete descriptions, they were able to identify the proportion of missing data in only 72 articles (among 257), representing only 28%. Still, among those 28%, the proportion of missing data varied from 0.1% to 81.1%, with a mean of 23.3% and a median of 15.7%. On average, the proportion of missing data in applied finance is very high and can sometimes reach extreme levels that would undermine the reliability of an empirical study.

Kofman and Sharpe [133] also show that missing data are more often reported in studies dealing with independent variables (over 62%), probably because it would seem more complicated to explain the presence of missing data within dependent data. In addition, for 27% (i.e. 213 articles out of 797) of the articles dealing with cross-sectional analysis, issues of missing data appear more often addressed in cross-sectional studies than in time-series studies (respectively 15%). Deleting missing data in cross-sectional analyses results in the deletion of an entire row of data, even the observed ones (listwise deletion). This results in a larger reduction in sample size.

Following this first expertise, the authors complete their analysis by referencing which methodology was applied, to compare these missing data (see Table 1.2-3).

Tab. 1.2-3: Treatment of missing data in Finance (Source: Kofman and Sharpe, 2003 [133])

Journal	Articles acknowledging missing values	Missing values occur treatment			
		Listwise deletion	Regression imputation	Ad hoc imputation	Proxy imputation
JBF	67	56	5	5	3
JF	98	77	6	9	7
JFE	53	44	2	3	5
JFQA	20	18	0	1	2
RFS	19	10	3	7	1
All	257	205	16	25	18

JBF: *Journal of Banking and Finance*, JF: *Journal of Finance*, JFE: *Journal of Financial Economics*, JFQA: *Journal of Financial and Quantitative Analysis* and RFS: *Review of Financial Studies*.

Among all the 257 articles mentioning missing data, 205 used listwise deletion, 16 used regression imputation, 25 used ad-hoc imputation and 18 used proxy imputation.

Table 1.2-2 and Table 1.2-3 show that problems of missing data are relatively frequent in the literature of applied finance and more specifically, for cross-sectional studies. In addition, the most common method researchers use to manage missing data is listwise deletion, which aims primarily to avoid missing data by reducing the sample size. This method could reduce study's statistical power and bias results, even making them totally inaccurate.

The work of Tsikriktsis [201] and of Kofman and Sharpe [133] yields to the following conclusion: missing data occurs commonly in Operational management and applied

finance research. Furthermore, in both studies, listwise deletion is used to deal with missing data in most cases. However, this method of managing missing data is not more optimal because, as mentioned earlier, it leads to a (sometimes drastic) reduction in sample size and can bias analyses and diminish their statistical power.

1.3 Data: a new regulatory challenge

Many new banking and financial regulations have been enacted in recent years. This development was prompted by the first two Basel Accords (published in 1996 [14] and 2004 [15]). The Cooke ratio and the implementation of standard and internal methods were insufficient to avoid the losses and bankruptcies of banks during one of the greatest financial crises of the century. In the aftermath of the crisis of 2007, the authorities became aware of the shortcomings of the regulations and intent on limiting the damage, they hastily signed a new set of agreements called "Basel 2.5"[16], which would serve as a basis for the Basel III agreements [18][17], in 2010. Notably, Basel 2.5 saw the emergence of new quantitative standards, such as the leverage ratio, the liquidity coverage ratio and the net stable funding ratio as well as a countercyclical capital buffer. However, regulation took a completely different turn with Basel 3.5 [12], a much more thorough and binding regulation for banks. Its aim is to account for all known financial risks. The authorities have also circulated several recommendations and principles that aim to implement the agreements of the Basel Committee effectively. For banks, these new regulations have created problems that concern implementation, modeling and of course, databases. The regulators found that banks were not always aware of their risk exposure because they lacked the necessary data. Databases (if they existed) were often of poor quality, difficult to access and incomplete. This discovery led to the promulgation of regulations on data quality, database design and systematic data storage. Regulators realized that banks must save all data automatically and that they must maintain their databases with the uttermost care to gauge their exposure to risk as accurately as possible. This section will therefore present a list of the main modern regulations that have prompted banks, be it directly or indirectly, to improve their databases.

1.3.1 BCBS 239: the new data regulation

Standard number 239 of the Basel Committee, also known as BCBS 239 (BCBS stands for "Basel Committee on Banking Supervision") [24] includes 14 principles that aim to maintain the stability of the global financial system. After the 2007 crisis, the authorities realized that many banks were unable to provide balance sheets of their risk exposures. The Basel Committee decided to enact a new regulation, and BCBS 239 [24] was published in January 2013. The goal was to have all systematically important

banks (SIB) apply it from January 2016. Its main purpose is to strengthen the capacity of banks to aggregate risk-related data and to improve internal risk reporting practices, which should improve risk management and decision-making.

According to BCBS 239 [24], the adoption of the 14 principles would enable the targeted banks to aspire to a better quality of banking management by improving their infrastructure for reporting important information, their decision-making processes and the processing of information within the different legal entities as well as by reducing the probability and magnitude of losses, accelerating the dissemination of information and enhancing the quality of strategic plans and the management of risks in new products and services.

The 14 principles are organized along four closely related axes:

- Overarching governance and infrastructure
- Risk data aggregation capabilities
- Risk reporting practices
- Supervisory review, tools and cooperation

The first axis, "overarching governance and infrastructure", consists of two principles. Their aim is to force banks to implement a strong governance framework, an IT architecture and an infrastructure for risk data. The principles emphasize the involvement of various banking departments in improving communications and thus the quality of regulatory reporting and the decision-making process. These improvements must be based on a data architecture and an IT infrastructure that facilitates automation and ensures the reliability of data aggregation and reporting.

The second axis includes four principles (Principles 3 to 6). They aim to encourage banks to develop "robust risk data aggregation [1] capabilities" in order to report their risk exposure accurately. These principles involve improving the accuracy and integrity of all risk data. The banks are also required to have updated aggregated risk data at their disposal in order to be able to respond quickly to specific requests.

The third axis, is the one with the most principles. The "risk reporting practices" axis is based on five principles (Principles 7 to 11) that are related to the implementation of effective risk management, which should be based on accurate, comprehensive and recent data. Banks must be ready to provide precise and accurate risk management reports on different types of risks. These reports must be clear, concise, relevant and suitable for the needs of the recipient. A production frequency for these reports must

[1] The Basel committee [24] define the the risk data aggregation by "defining, gathering and processing risk data according to the bank's risk reporting requirements to enable the bank to measure its performance against its risk tolerance/appetite. This includes sorting, merging or breaking down sets of data".

be defined. The aim is to distribute them to the relevant parties while preserving the confidentiality of the risk data.

Finally, while the first three axes target banks, the last, "supervisory review, tools and cooperation," addresses regulators. The last three principles require regulators to ensure that compliance with the principles is monitored regularly and to adopt effective measures if banks fall short of the standards. Finally, authorities must cooperate with each other. All of the principles purport to improve risk monitoring for both regulators and banks.

Regulators noted that banks use data of poor quality and lack data infrastructures. Many manual and unsophisticated processes were employed, making aggregation laborious and risk management reports unreliable. For this reason, the first challenge that BCBS 239 [24] standard tackles is the automation of processes. Automation also causes banks to become more responsive to requests from regulators and to address risk more effectively in times of crisis.

There is a growing trend towards reporting. Regulators are increasingly requesting information in order to increase transparency. Therefore, banks must produce more reports. As noted, data was often aggregated manually, often resulting in non-standardized reports. Some contained duplicate data, and reconciliation and validation could only be accomplished manually. These difficulties made the production of reports laborious, tedious and inefficient. The reports were often plagued by inaccuracies. The same was true of the resultant risk management decisions. Thus, BCBS 239 [24] attempted to make banks more responsive and efficient in producing reports.

The Basel Committee addresses the infrastructures related to risk management and the quality of the data on risk factors (as stated earlier in Section 1.1)directly through the principles of BCBS 239 [24]. In addition, the paragraph that precedes the principles refers to missing data (paragraph 25 of BCBS 239 [24]). If data is incomplete, ad hoc expert judgments may be used to facilitate aggregation. However, their use should remain exceptional, and it must not interfere with compliance [24]. Of course, when expert judgments are used, the authorities should be able to observe a clearly documented and transparent procedure that can be subjected to independent review.

According to the regulators, data may be missing but such states of affairs should remain exceptional. In such situations, the regulator will tolerate expert judgments, but no guidance is provided about the methods to be adopted (imputation, interpolation, deletion, and such like) or the control criteria that the regulator will use to regulate them. The banks are thus left in the dark and may invest time and money in a solution that is eventually rejected.

Following the January 2013 publication of the BCBS 239 principles [24], banks were forced to undertake a huge project whose results would be explored fully from January 2016. The Basel Committee called on banks to institute strong governance arrangements, efficient IT architectures and infrastructures, strong aggregation capabilities and

effective risk data management systems in a short period of three years. To monitor progress, since December 2013, the Basel Committee has been publishing reports called *Progress in adopting the principles for effective risk data aggregation and risk reporting.* Usually, the reports are published on annual basis.

The Basel Committee's Working Group on SIB Supervision, aided by national supervisors, has set up a self-assessment questionnaire for global systematically important banks (G-SIB). It consists of 87 questions (or requirements) that concern the 11 principles that are addressed to banks. The ratings are based on a scale that ranges from 4 (best) to 1 (worst).

Table 1.3-1 shows the results from the first report, which reflects the answers of 30 banks. It was published in December 2013 [19].

Tab. 1.3-1: Self-assessment ratings from December 2013 report: number of banks reporting compliance with each principle (Source: Basel Committee, 2013 [19])

	Governance/ infrastructure		**Risk data aggregation capabilities**				**Risk reporting practices**				
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**
Fully compliant	0	0	0	0	2	1	0	8	3	2	7
Largely compliant	25	14	18	22	17	15	21	20	26	21	23
Materially non-compliant	5	16	12	8	11	14	9	2	1	7	0
Non-compliant	0	0	0	0	0	0	0	0	0	0	0
Average rating	2.8	2.5	2.6	2.7	2.7	2.6	2.7	3.2	3.1	2.8	3.2

Among the banks surveyed, none consider themselves as fully compliant with Principle 1, 25 as largely compliant, 5 as materially non-compliant and none as non-compliant.

The average score for all 11 principles is 2.8, indicating that, on average, banks report themselves to be between "largely compliant" and "materially compliant." In addition, a large proportion of banks report low compliance with Principle 2 (architecture and IT infrastructure), Principle 6 (adaptability) and Principle 3 (accuracy and integrity). Many banks are struggling to implement sound governance, architecture and data aggregation processes. Unsurprisingly, as mentioned earlier, banks typically resort to manual workarounds that the regulator rejects because their use could hinder the aggregation of risk data. Conversely, banks reported high compliance with Principle 11 (report distribution), Principle 8 (comprehensiveness) and Principle 9 (clarity and usefulness). Overall, the scores for the principles that pertain to the "risk reporting practices" axis are much better than those for "governance and infrastructure' and "risk data aggregation capabilities."

Two similar reports were published two years later, just before the implementation deadline for the BCBS 239 principles [24]. One was published in January 2015 and another in December 2015). The January 2015 [20] report, which was based on 2014 data, showed that banks had not processed sufficiently to apply all the principles of BCBS 239 [24] from January 2016. Therefore, the Basel Committee decided to postpone implementation until 2018. In the December 2015 [21] report (see Table 1.3-2), banks

tended to rate their performance in "risk reporting practices" and "risk data aggregation capabilities" highly, while they performed poorly in "governance and infrastructure."

Tab. 1.3-2: Self-assessment ratings from December 2015 report: number of banks reporting compliance with each principle (Source: Basel Committee, 2015 [21])

	Governance/ infrastructure		**Risk data aggregation capabilities**				**Risk reporting practices**				
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**
Fully compliant	2	0	0	2	1	1	0	9	4	3	10
Largely compliant	21	13	15	22	20	16	20	20	25	23	20
Materially non-compliant	7	17	15	6	9	13	10	1	1	4	0
Non-compliant	0	0	0	0	0	0	0	0	0	0	0
2014 average rating	2.83	2.43	2.50	2.87	2.73	2.60	2.67	3.27	3.10	2.97	3.33
2013 average rating	2.83	2.47	2.60	2.73	2.70	2.57	2.70	3.20	3.07	2.83	3.23

Among the banks surveyed, 2 consider themselves as fully compliant with Principle 1, 21 as largely compliant, 7 as materially non-compliant and none as non-compliant.

More banks considered themselves to be "fully compliant" in this report than in the December 2013 one (see Table 1.3-1). In addition, the average scores for 2014 is often higher than the average scores for 2013. The date of application of the principles of BCBS 239 [24] (originally 2016) would have been postponed anyway, as the December 2014 report showed that banks were far from being "fully compliant."

The June 2018 report covers 2016 and 2017 (see Table 1.3-3). Evidently, five years after the publication of the BCBS 239 principles [24], there was no principle with which all banks complied fully. Progress between 2016 and 2017 was very marginal. Table 1.3-3 shows that the change did not exceed 0.13 compliance score points.

As far as particular axes are concerned, "governance and infrastructure," which had not seen positive developments in the previous report (see Table 1.3-2), had caught up, at least to some extent. The veracity of these reports is doubtful because Principle 1 and Principle 2 are prerequisites for a sound governance framework as well as for an adequate risk data architecture and infrastructure.

Tab. 1.3-3: Self-assessment ratings from June 2018 report: number of banks reporting compliance with each principle (Source: Basel Committee, 2018 [23])

	Governance/ infrastructure		**Risk data aggregation capabilities**				**Risk reporting practices**				
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**
2016											
Fully compliant	6	3	6	5	3	6	6	7	8	7	12
Largely compliant	14	12	8	18	17	15	11	16	17	16	17
Materially non-compliant	9	15	14	7	9	9	13	7	5	6	1
Non-compliant	1	0	2	1	0	0	0	0	0	1	0
2016 average rating	2.90	2.73	2.60	2.90	2.87	2.90	2.73	3.03	3.03	2.97	3.33
2017											
Fully compliant	7	5	6	5	6	8	5	8	7	7	13
Largely compliant	14	12	8	17	15	11	12	15	17	16	14
Materially non-compliant	8	13	14	8	8	11	13	7	6	6	3
Non-compliant	1	0	2	1	0	0	0	0	0	1	0
2017 average rating	2.83	2.60	2.60	2.93	2.73	2.90	2.77	3.00	3.10	2.97	3.37

Among the banks surveyed in 2016, 6 consider themselves as fully compliant with Principle 1, 14 as largely compliant, 9 as materially non-compliant and 1 as non-compliant.

Moreover, unlike Table 1.3-2 from the December 2015 report, the June 2018 report shows that at least some banks declared themselves to be "fully compliant" with each of the principles. This clearly shows that banks were gradually implementing the BCBS 239 principles [24]. Regulators noted that only three banks were "fully compliant" with all 11 principles, compared to the 11 that had been expected in the 2016 report.

The most recent report is from April 2020 [25]. It compares the results of 34 G-SIB questionnaires, four more than in the previous reports. Table 1.3-4 shows that compliance has improved significantly. In 2019, no banks described themselves as "non-compliant," a first since the reports began. In addition, the number of banks reporting material non-compliance has decreased systematically for each of the principles, meaning that "largely compliant' or even "fully compliant" are now more likely responses. In addition, it can be seen that efforts to implement Principle 1 and Principle 2 have been redoubled since 2019.

Tab. 1.3-4: Self-assessment ratings from April 2020 report: number of banks reporting compliance with each principle (Source: Basel Committee, 2020 [25])

	Governance/ infrastructure		**Risk data aggregation capabilities**				**Risk reporting practices**				
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**
Fully compliant	9	5	4	9	6	9	7	13	16	7	18
Largely compliant	21	19	22	20	22	17	20	17	16	23	14
Materially non-compliant	1	9	7	4	5	7	6	3	1	3	1
Non-compliant	0	0	0	0	0	0	0	0	0	0	0

Among the banks surveyed, 9 consider themselves as fully compliant with Principle 1, 21 as largely compliant, 1 as materially non-compliant and none as non-compliant.

This having been said, the report points out that no bank is "fully compliant" with all 11 principles. In the June 2018 report, three banks had reported full compliance. The

ratings had been downgraded partly due to increased awareness. The Basel Committee noted, however, that most banks expected to be fully or largely compliant by the end of 2020.

The BCBS 239 standard [24] was introduced to raise awareness of the risks to which banks are exposed. This is a significant task for banks, which are obliged to institute solid governance and risk data architectures as well as sound risk data aggregation capabilities and effective risk management procedures. The 11 principles of BCBS 239 [24] are difficult to implement, and regulators underestimated the time that would be needed. This failure may be attributable to the vagueness of the text of the regulation, especially in respect of the methods that must be used, and the difficulty of introducing the changes. BCBS 239 is barely 30 pages long. Each principle is described in brief, without clear indications of the expectations of regulators. The principles should have been implemented by January 2016, but the April 2020 report [25], which is the most recent, indicates that no bank had succeeded in implementing all of them in 2019. Supervisors are aware of the difficulties that inhere in the implementation of the principles, and they are lenient with institutions that show goodwill. Others must provide a recovery plan.

The BCBS 239 [24] standard is one of the reasons why banks have had to review the quality of their data since 2013. This is the banking regulation that is designed specifically to address data quality issues, yet it overlooks missing data, treating them as "an exception" in a single paragraph. The various studies that were overviewed in Section 1.2.3 show clearly that missing data are common in financial series.

1.3.2 Trade repository: transparency of data quality

The supervisory authorities have set up trade repositories to mitigate systemic risk, to reduce the opacity of markets (particularly that of the over-the-counter markets) and to detect and prevent market abuse. In 2012, the principles for financial market infrastructures that were published by the Committee on Payment and Settlement Systems and the Technical Committee of the International Organization of Securities Commissions [58] defined a trade repository as "an entity that maintains a centralized electronic record (database) of transaction data". In 2018, the Irving Fisher Committee (IFC) [121], which is composed of central bank economists and statisticians, defined trade repositories as "legal entities tasked with centrally collecting the transactions information and maintaining the related records." Trade repositories have emerged as a new type of financial market infrastructure that is designed to provide information. That information enables central banks to gauge risk. Prudential supervisors can monitor the exposures of banks, and market regulators can detect instances of possible market manipulation. Given the large number of stakeholders who use trade repository data, it is essential that they remain available, reliable and accurate.

The first trade repositories appeared upon the emergence of credit default swaps (CDS) in the late 1990s and early 2000s. At that time, trades were conducted orally, which caused many errors. In addition, the confirmation of transactions was often very lengthy, leading to the accumulation of unidentified or unreconciled risks. It is in this context that several national authorities recommended the development of an electronic reconciliation and processing service for CDS transactions. In 2003, the Depository Trust & Clearing Corporation (DTCC) set up the automated platform Deriv/SERV to reconcile and confirm CDS trades. Then, in 2006, a second platform, called Trade Information Warehouse, was set up by DTCC in order to save all modifications and amendments to CDS contracts, which can be resold or transferred several times before they mature. Thus, DTCC created the first trade repositories. In the light of this success, regulators decided to allocate a more important role to trade repositories, extending their use to other asset classes. Consequently, trade repositories have become essential to ensuring the transparency and security of market operations. The global crisis of 2008 led to the G20 commitment, made in Pittsburgh in September 2009, whose aim is to improve the transparency and security of financial markets.

The 24 principles for Financial Market Infrastructures [58] were published in April 2012. Their purpose is to harmonize and strengthen international standards for systemically important payment systems, central securities depositories, securities settlement systems and central counterparties, that is, for the entire financial market infrastructure. These principles emphasize the use of different types of infrastructure to increase market transparency and transaction security. The last of the principles identifies trade repositories as the source of "timely and accurate data to the relevant authorities and the public in line with their respective needs."

Trade repositories also feature in the European Market Infrastructure Regulation (EMIR) [168] of August 2012, which aims to reduce systemic risk, to improve transparency in the over-the-counter market and to preserve the stability of the financial sector. This regulation is based on four principles.

- A central clearing requirement for all over-the-counter derivatives deemed sufficiently liquid and standardized by European Securities Market Authority (ESMA). As a result, counterparty risk is transferred in full to the clearing houses.
- The harmonization of the European legal framework to ensure that clearing houses comply with robust capital, organizational and conduct-of-business requirements.
- The use of a set of operational and counterparty risk mitigation techniques for uncleared contracts
- An obligation to report all derivatives transactions to trade repositories.

Thus, since February 2014 it is mandatory for both parties to report on all derivative transactions.

Accordingly, since February 2014, it has been mandatory for both parties to a derivative transaction to report it. In the US, the G20 commitments are reflected in the Dodd-Frank [203] of 2010, which requires only one of the two parties to report a derivatives transaction to a swap data repository (the US equivalent of a trade repository).

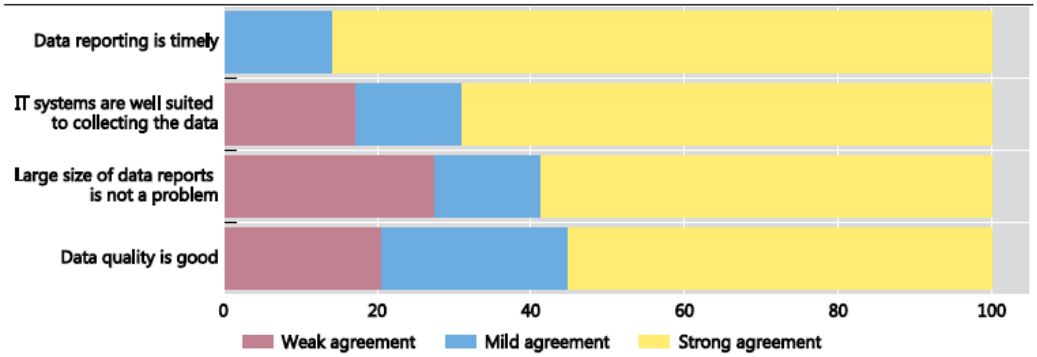
The use of trade repositories seems to be in its developmental stages. In November 2015, following the recommendations of the Financial Stability Board about the regulation of shadow banking [82], a new regulation, titled the Securities Financing Transactions Regulation (SFTR) [169], was adopted. Its goal is to make securities financing transactions in the EU more transparent. It focuses on securities lending and borrowing and on repo transactions and it imposes an obligation to report securities transactions to trade repositories from 12 January 2016.

There are many trade repositories around the world. ESMA publishes a list of registered and approved European trade repositories every year. According to the list published by ESMA on 31 March 2020, eight trade repositories were approved this year: DTCC Derivatives Repository Plc, Krajowy Depozyt Papierów Wartosciowych S.A., Regis-TR, UnaVista Limited, CME Trade Repository Ltd, ICE Trade Vault Europe Ltd. NEX Abide Trade Repository AB and UnaVista TRADEcho B.V. NEX Abide Trade Repository AB and UnaVista TRADEcho B.V. The EMIR regulation also authorizes ESMA to recognize trade repositories from non-EU countries, subject to conditions, such as an equivalent supervision regime, the completion of cooperation agreements between regulators and such like.

Trade repositories are an essential element of contemporary regulation. They emerged when transparency became essential to financial stability. They allowed domestic authorities, as well as various foreign supervisory authorities, market participants and the general public in the surveyed jurisdictions (define by the report of the committee on payment and settlement systems and the international organization of securities commissions [57]), to become fully aware of the risks to which they were exposed. However, even if these data are now available to central banks, they cannot be applied effectively because they are not sufficiently comprehensive. If data are missing from reports of transactions, risk cannot be measured precisely.

In 2018, the IFC [121], drawing on a survey of respondents from 50 countries, showed that data processing at trade repositories raises no major issues for central banks. However, these data must be of sufficient quality. The survey reveals some deficiencies in data quality. As is evident from Figure 1.3-1, more than 80% of central bankers were of the view that trade repositories are processing data in a timely manner. Timeliness is important because it allows the central banks to access real-time information and thus to make better decisions.

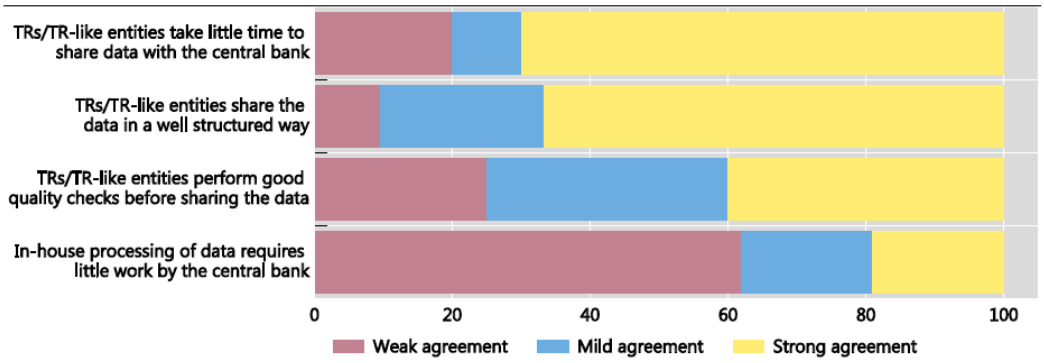
Fig. 1.3-1: Trade repositories data processing as described by central banks (in per cent) (Source: Irving Fisher Committee, 2018 [121])



Less than 60% of surveyed banks weakly agree that the data is of good quality, while more than 20% mildly agree and about 20% strongly agree.

Figure 1.3-2 shows that, according to central bankers, the connection between central banks and trade repositories is fast and structured well. However, expedience appears to have come at the cost of data quality. More than 40% of respondents considered the data to be of average or even poor quality (see Figure 1.3-1). Furthermore, Figure 1.3-2 shows that only 40% of central bankers were satisfied with the data quality checks that trade repositories implemented, that is, 60% believed that trade repositories should perform more data quality checks before transmitting data to them.

Fig. 1.3-2: Quality of derivatives trade repositories data as described by central banks (in per cent) (Source: Irving Fisher Committee, 2018 [121])



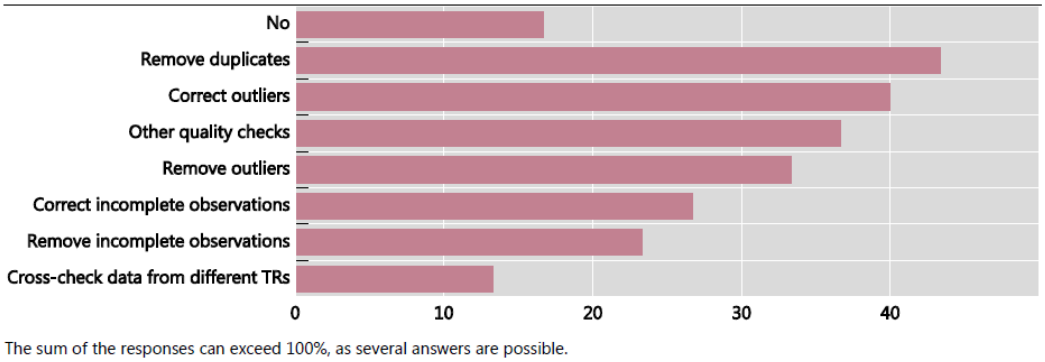
Less than 10% of surveyed banks weakly agree that good quality checks are performed before sharing the data, while about 20% mildly agree and more than 60% strongly agree.

In addition, more than 60% of central banks seem to reprocess the data provided by trade repositories in-house to a significant degree, and another 20% reported engaging in medium-level reprocessing. Less than 20% of central banks did not adjust the data

(see Figure 1.3-2). This suggests that the data transmitted by trade repositories may be unrepresentative or incomplete or that it includes outliers.

Figure 1.3-3 shows the types of reprocessing that central banks perform: more than 40% removed duplicate data, 40% corrected outliers, more than 30% deleted them, more than 25% corrected missing data, more than 20% deleted observations with missing data, and more than 35% performed other quality checks. In addition, more than 10% of the respondents reported cross-checking the data with that provided by other trade repositories. Clearly, the quality of trade repositories data leaves something to be desired.

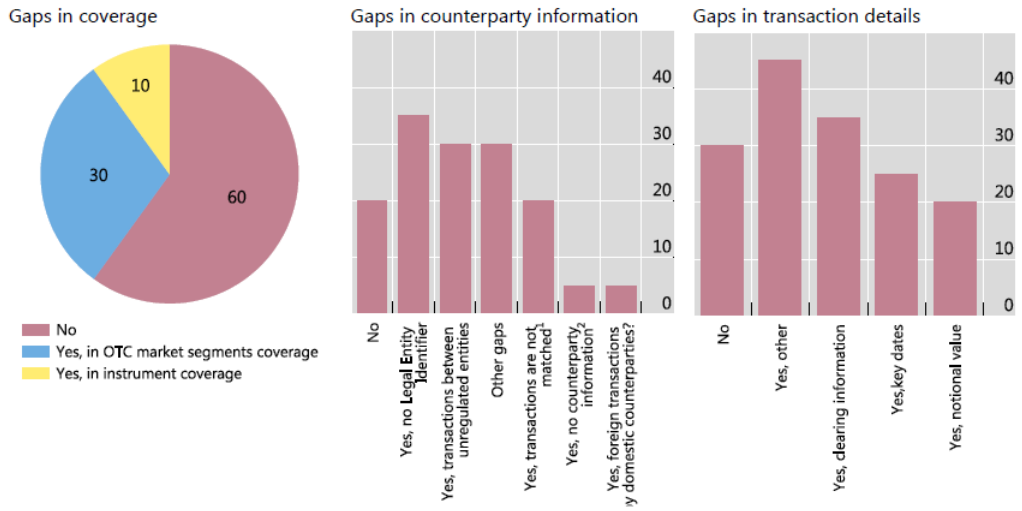
Fig. 1.3-3: Quality checks on trade repository data conducted by central (in per cent) (Source: Irving Fisher Committee, 2018 [121])



More than 25% of surveyed banks correct incomplete observations while, less than 25% remove them.

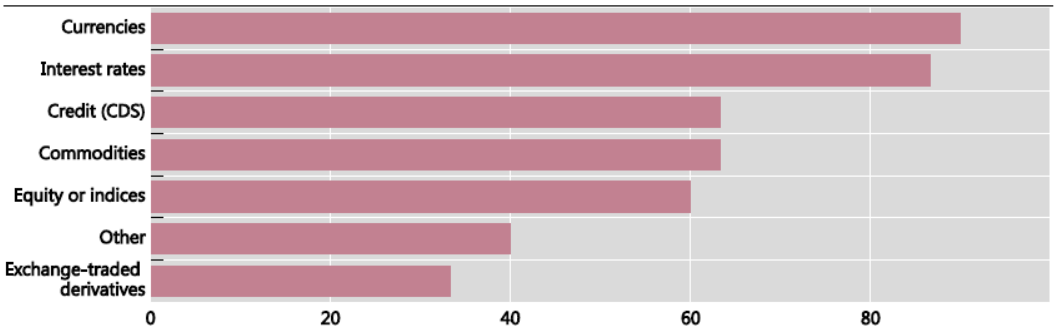
The IFC [121] also show that 72.4% of central banks encountered difficulties with gaps when they analyzed trade repository data. They add that there are three categories of gaps: incomplete coverage of market segments and/or instruments, omitted counterparty information and missing details on critical elements of derivatives transactions. Among the respondents, 60% thought that there were no gaps in coverage, 30% thought that there were gaps in the coverage of some important OTC segments, and 10% thought that there were gaps in some important instruments (see the left-hand side of Figure 1.3-4).

Fig. 1.3-4: Gaps in the trade repository data that central banks access (in per cent) (Source: Irving Fisher Committee, 2018 [121])



As far as individual market segments are concerned, Figure 1.3-5 shows that about 90% of central banks reported having access to transaction data for currency and interest rate derivatives and that about 60% reported that they could access transaction data for CDS, commodities, equities and indices. More than 60% identified gaps in exchange-traded derivatives trade repositories data, and 60% identified gaps in the "other" category.

Fig. 1.3-5: Coverage of derivatives transactions in trade repository data that central banks access (in per cent) (Source: Irving Fisher Committee, 2018 [121])



About 90% of surveyed banks reported having access to transaction data for currency and interest rate derivatives.

The IFC found that the main gaps in trade repository data concern counterparty information and transaction details. As can be seen in the center of Figure 1.3-4 only

20% of respondents reported that there were no gaps in counterparty information. In 35% of the cases, the Legal Entity Identifier was missing, and the gaps were in data for transactions organized by unregulated entities in 30% of the cases.

The right-hand side of Figure 1.3-4 shows that only 30% of central banks did not encounter gaps in data on transaction details. Conversely, 35% reported gaps in clearing information, 25% in key transaction dates and 20% in notional values. Various "other" gaps were identified in the survey, including information on collateral and market value. The gaps are partly attributable to ambiguous reporting requirements. Missing information may prevent the correct identification of novated and compressed trades. To solve this problem, the IFC [121] proposes that information be collected to track events in the lifecycles of derivative contracts. For instance, data on previous Unique Transaction Identifiers may be collected when evaluating the impact of transaction compression. All in all, the survey shows that despite the imposition of reporting obligations on trade repositories, there are holes in the data. Those holes can bias the risk assessments of central banks.

Trade repositories increasingly feature in regulations that purport to combat market opacity. Their use can improve the quality of market data. The regulations oblige banks to report data on derivatives transactions and allow regulators to monitor their activity. However, the effective exploitation of data reporting remains complicated because the ambiguity of some reports leads to missing or aberrant data. Indeed, as shown in Figure 1.3-3, more than 25% of respondents reported correcting missing data. This shows once more that missing data are a concern in financial series. Moreover, so far, nothing has been written about imputation methods. Central banks must reprocess trade repository data before exploiting it. According to the central banks, trade repositories should apply more quality checks to improve data quality.

1.3.3 Proxy spread methodology for the Capital Requirement Regulation

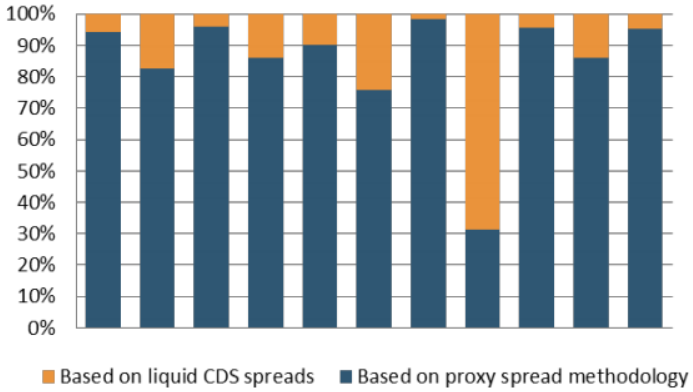
Although there are no specific regulations on the imputation of missing data, similar difficulties seem to have been identified in EU law. This is notably the case in Regulation 575/2013, also known as the Capital Requirement Regulation (CRR) [167], which was published in June 2013 by the European Banking Authority (EBA). This regulation covers the calculation of credit valuation adjustment (CVA), which corresponds to the price that an investor would have to pay to hedge the counterparty risk associated with a derivative instrument.

According to Article 383-1 of the CRR, for CVA to be calculated, a credit spread and a loss given default (LGD) must be identified for every counterparty. Their values must be deduced from observable market data, particularly CDS. When these parameters are

unobservable, the regulations allow approximation. In addition, Article 383-7 stipulates that the EBA must implement "draft regulatory technical standards (RTS) to specify in greater detail how a proxy spread is to be determined."

However, according to the EBA report on CVA [79], CDS spreads are available for few counterparties. According to the report, among the 11 banks that responded to the survey, 10 had more than 75% of their counterparties affected by the proxy spread methodology (see Figure 1.3-6). Only one bank had liquid CDSs whose spreads did not need to be approximated.

Fig. 1.3-6: Number of counterparties subject to proxy spread for surveyed (Source: European Banking Authority, 2015 [79])



Among the 11 banks (11 bars plotted), 10 had more than 75% of their counterparties affected by the proxy spread methodology

This EBA report therefore reveals the unobservability of the data. If values must be approximated for so many counterparties, then it must be true that very few CDS are finally liquid and therefore observable. Data are not necessarily missing because they are not recorded or because they are erroneously deleted. It is possible that they simply do not exist. In this case, illiquidity leads to unobservability, and unobservability leads to missing data. This makes the enforcement of data quality regulations such as the BCBS 239 standards (see Section 1.3.1) even more legitimate. The same is true of the intensive use of trade repositories (see Section 1.3.2) presented earlier. Allowing banks to access CDS data from trade repositories would, indeed, allow them to more easily deduce the missing information.

As mentioned by the CRR, an RTS has been published by the EBA to define acceptable spreads and LGD approximations. These are presented in Policy Recommendation 7 and Policy Recommendation 8 of the report on CVA [79]. Policy Recommendation 7 is germane to the present argument. It contains a proxy spread methodology that is based on three attributes: rating, region and industry. However, the EBA is acutely aware

of the modeling difficulties and encourages institutions to use alternative approaches. Those approaches should be based on a more fundamental analysis of credit risk when no time series of credit spreads (or their counterparts) are available. Consequently, institutions are free to consider additional attributes.

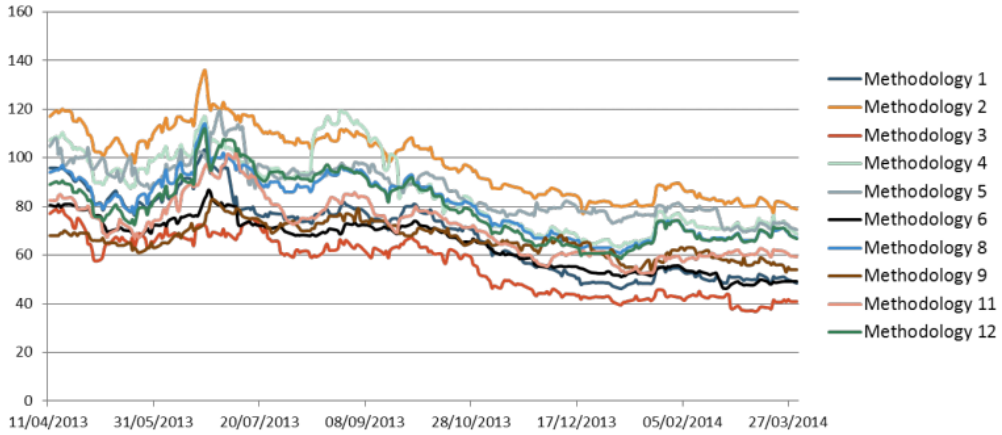
In practice, institutions do not use additional attributes systematically. The CVA report shows that among the 12 institutions surveyed, 5 use additional attributes. When they did, they would use currency, seniority or even country or sovereign ownership.

Of course, the methodology that is chosen must always be documented in such a way that it can be overseen by the authorities. The regulation is therefore aimed at both banks and supervisory authorities: banks must implement a methodology in order to be able to calculate their CVA as accurately as possible, and the regulator is responsible for monitoring it.

Good documentation and control are all the more important because the credit spreads that are obtained differ with the choice of methodology. The EBA asked the 12 institutions surveyed to calculate proxy spreads over a period of one year for different counterparties. The CVA report [79]describes the methodologies. It seems that the vast majority used a bucketing approach, while two banks used regressions and one employed a decision-tree approach. Globally, banks use the median of buckets, as well as single-name CDS spreads, while others also included bond spreads and even CDS index spreads.

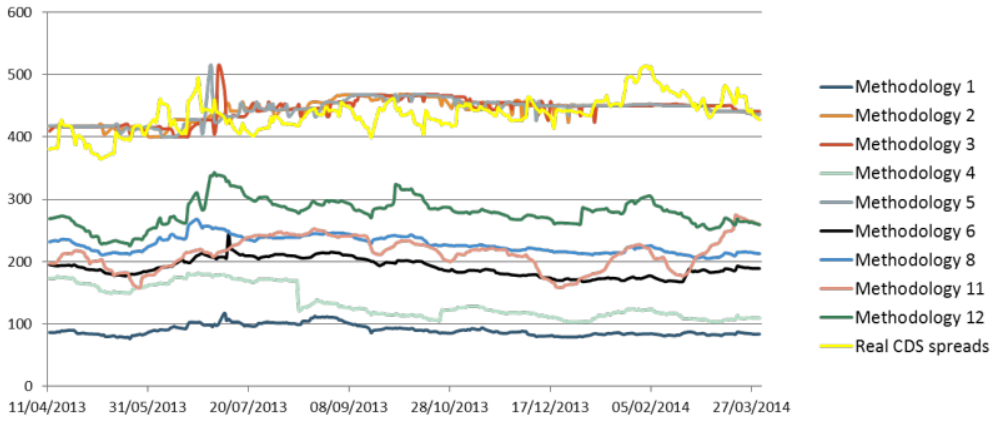
Figure 1.3-7 and Figure 1.3-8 present the one-year-maturity proxy spreads of Tata Motors Ltd and the AA-rated UK insurer counterparty for the period between April 2013 and April 2014 and for each methodology used by the 12 institutions. The real CDS spreads that the authorities expected are depicted by the yellow line.

Fig. 1.3-7: 1-year history of five-year proxy spread benchmarking of the AA-rated real UK insurer counterparty (Source: European Banking Authority, 2015 [79])



The one-year-maturity proxy spreads of Tata Motors Ltd computed by each methodology used by the 12 institutions.

Fig. 1.3-8: 1-year history of five-year proxy spread benchmarking of the real counterparty Tata Motors Ltd (Source: European Banking Authority, 2015 [79])



The one-year-maturity proxy spreads of AA-rated UK insurer counterparty computed by each methodology used by the 12 institutions, and the observed CDS spread (in yellow).

It is clear from Figure 1.3-7 that the 12 methodologies yielded proxy spreads of approximately the same volatility and level. At the same time, Figure 1.3-8 presents very heterogeneous results for the same methodologies. The real CDS spread seems quite variable over the time, while the methodologies used by institutions to obtain proxy spreads are much more stable, that is, less volatile. In addition, half of the

proxy spreads presented here differ substantially from the real CDS spreads (some are close to four times lower). This disparity demonstrates that the choice of proxy spread methodology can change results. In its report, the EBA notes [79] that the contrast between methodologies is even more important when it comes to transactions with non-financial counterparties, where few CDSs are quoted at the end. It adds that some methods are not capable of maintaining the appropriate level of volatility due to discontinuities and abnormal jumps.

The significant variability between methodologies certainly impacts risk measures. Notably, EBA [79] highlights this concern by calculating a 1-day, one-sided 99% non-stressed CVA VaR. The results for Tata Motors Ltd are presented in Table 1.3-5.

Tab. 1.3-5: Computation of 1-day regulatory CVA VaR 99% of Tata Motors Ltd (Source: European Banking Authority, 2015 [79])

	CVA 31/03/14	**CVA VaR 99%**	**CVA VaR 99% (in % CVA 31/03/14)**	**CVA 31/03/14 (in % Mean CVA 31/03/14)**	**CVA VaR 99% (in % Mean CVA VaR 99%)**
Methodology 1	95,404	4,672	4.9%	-52.6%	-52.8%
Methodology 2	290,439	19,064	6.6%	44.4%	92.6%
Methodology 3	292,744	19,993	6.8%	45.6%	101.9%
Methodology 4	115,249	4,501	3.9%	-42.7%	-54.5%
Methodology 5	298,930	21,381	7.2%	48.7%	116%
Methodology 6	171,762	5,291	3.1%	-14.6%	-46.6%
Methodology 8	195,657	3,406	1.7%	-2.7%	-65.6%
Methodology 11	203,820	6,238	3.1%	1.4%	-37%
Methodology 12	243,451	16,043	6.6%	21.1%	62%
Mean	201,066	9,901	4.9%		
Stdev	75,933	7,698	2%	37.8%	77.7%

The first methodology provides an amount of CVA as of 31/03/24 equal to 95,404, which leads to an amount of CVA VaR (at 99% confidence level) equal to 4,672, that is 4.9%. This methodology leads to an amount of CVA 52.6% lower than the average of CVA of all methodology, and to a VaR 52.8% lower than the average CVA VaR of all methodology.

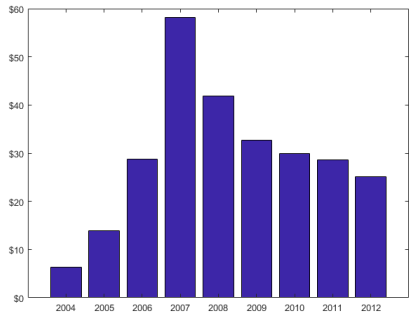
Unsurprisingly, the amount of the CVA as of 31/04/2014 could vary threefold depending on the chosen method (from 95,404 to 298,930). The average CVA obtained by the different methodologies at that date was 201,066, and the results vary by 50% around it, with standard deviation close to 40%.

These different methodologies also impact the VaR calculation. VaR is between 3,406 and 21,381 (for Methodology 8 and Methodology 5, respectively), the latter amount more than six times higher than the smallest VaR. These correspond to 1.7% and 7.2%, respectively, of the amount of the CVA as of 31 April 2014, giving a CVA VaR that can vary between -65% and $+116\%$ around the average VaR of the different methods—variability is close to 80%. If the choice of methodology impacts CVA, it impacts the VaR that is calculated from it even more.

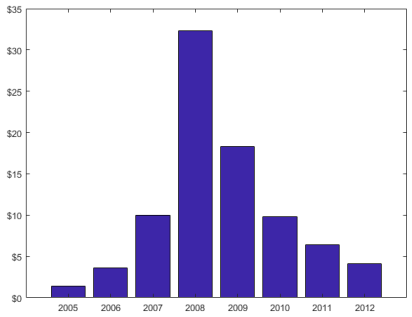
Finally, EBA seems to acknowledge that these differences are partly attributable to data quality. A report by the International Swaps and Derivatives Association (ISDA)

[119] that was published in October 2013 shows that the notional amount of CDSs has been declining steadily since 2007 (based on the Bank for International Settlement data; see Figure 1.3-9a), but it seems to suggest that the decline has been driven by the increasing compression of portfolios [2] (see Figure 1.3-9b). However, the DTCC's market risk transaction activity clearly shows a reverse trend, with the number of new CDS trades increasing since 2011 (see Figure 1.3-9c) and the associated notional amount rising by 15% between 2012 and 2013 (see Figure 1.3-9d). It appears that the increase has been fomented by index CDS trading.

Fig. 1.3-9: CDS market between 2004 and 2013 (Source: International Swaps and Derivatives Association, 2013 [119])

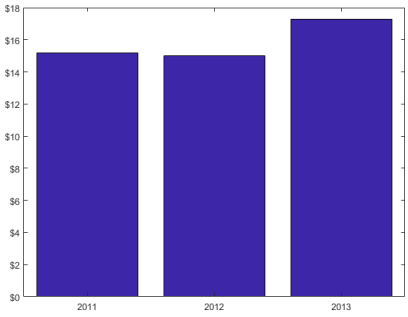


(a) Annual CDS notionals outstanding (in $tn)

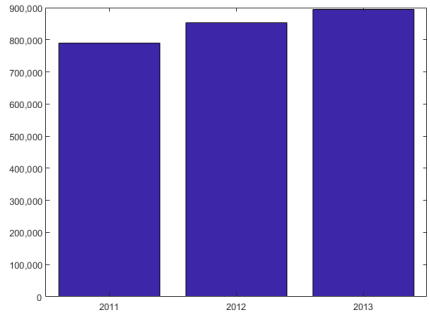


(b) Annual CDS portfolio compression (in $tn)

[2] According to the BIS [7], portfolio compression "is a process that enables early termination of economically redundant derivatives trades without changing the net position of each participant". This is done by terminating existing transactions and replacing them with a smaller number of new transactions with significantly smaller notional amounts that have the same risk profile and cash flows as the original portfolio. Portfolio compression thus reduces the overall notional size and number of outstanding contracts in the derivatives portfolios, improving derivatives risk management.



(c) Number of new CDS trades from February to July



(d) Notional amounts of new CDS from February to July (in $tn)

However, despite all the new contracts and the efforts to standardize CDS agreements, securing high-quality data seems to be a substantial challenge. In 2015, it was found that more than 75% of the counterparties had to be modeled by a proxy spread, while the ISDA report is from 2013. New regulations, such as the implementation of the incremental risk charge and the regulation of short selling and of leverage ratios, have not supported the CDS market and its liquidity.

Finally, the problem raised by the CRR is close to that of missing data: in a context where data is not necessarily available, banks must use methodologies that allow them to reproduce their risk exposure as accurately as possible. As far as CVA calculation is concerned, the supervisory authorities understand that the market may be illiquid and that data may be missing. For this reason, they encourage banks to implement methodologies to determine the best possible proxy spreads.

The case of CDS, presented in the EBA report [79], highlights the difficulty of finding and modelling the right data. The choice of methodology affects the amount of the CVA and therefore the set of risk measures that are applied to it. VaR can mushroom between methods, and it is impossible to know which method is the most reliable.

1.3.4 Fundamental review of the trading book

The fundamental review of the trading book (FRTB) is one of the recent financial regulations enacted by the Basel Committee. It involves a significant change in risk assessment. It is a successor to the Basel III [18][17] agreements, and it purports to respond to the failings identified after the 2007 crisis. Specifically, it addresses the evaluation of market risk in a prudential framework with the aim of making banks more resilient. An initial draft was circulated in January 2016 [11] (after three consultative documents were published in 2012 [8], 2013 [9] and 2014 [10]). The final version of the

FTRB was released three years later, in January 2019 [12]. The new regulation focuses on three major issues:

- defining a stricter boundary between the banking book and the trading book,
- improving the internal model approach (IMA) and
- reviewing the standardized approach (SA) comprehensively in order to make it more credible.

Before the FRTB [12], the boundary between the banking book and the trading book was set in accordance with subjective criteria. Banks often used the failure to move positions from the banking book to the trading book to reduce the amount of their capital charge. In order to overcome this problem, the FRTB defined a list of instruments that must be included in trading books. Deviations from the definition require explicit regulatory approval.

The improvements to the IMA account for liquidity risk and tail risk better. More granularity has been introduced into liquidity horizons, and VaR, the main risk measure, has been replaced with expected shortfall (ES). The validation processes of these internal models have also been strengthened through the implementation of backtesting and the P&L attribution test. Finally, risk must be modelled by reference to a rigorous selection of risk factors, and a special treatment is afforded to non-modellable risk factors (NMRF).

Concerning the SA, the authorities plan to set up a credible alternative to the internal models by employing a sensitivity approach to the calculation of capital charges, but also charges based on default risk and residual risk (for example, gap risk, correlation risk, behavioral risk and so on). Thus, the new SA is much more sophisticated and elaborate, and all banks can calculate it. In fact, it must become a standard because the regulators demand and enforce its implementation. Banks that are authorized to use their IMA must always declare their capital charges according to the SA.

For many years, financial regulation was premised on the calculation of a few ratios, often simple. Modelling and implementation were unproblematic. The publication of the FRTB [12] in 2016 has completely overhauled the regulatory practices of banks by requiring them to be ready to provide numerous risk measures, backtests, stress tests and such like. The FRTB is undoubtedly one of the biggest regulatory projects that banks have undertaken. So much is evident from the constant postponement of the implementation date. Initially, the regulation was intended to come into force in January 2019; the current deadline is January 2022.

The implementation of new regulatory requirements involves a significant amount of data. As noted in Section 1.3.1, there are some deficiencies in banks' treatment of data. The progress report on the application of the BCBS 239 principles [24], which

was published by the Basel Committee in April 2020 [25], shows that none of the surveyed bank were fully compliant with all of the principles in 2019. Delays in the implementation of BCBS 239 [24] inevitably affect the implementation of the FRTB. Banks must implement many interconnected regulations, a complex endeavor.

The FRTB regulation [12] introduces new demands of implementing, modelling and calculating the risk measures that requires excellent databases. Although data quality is not the central concern of FTRB (BCBS 239 [24] is already dedicated to that issue), some of its sections are dedicated to it, be it implicitly or explicitly.

Non-modellable risk factors

One of the new features of the FRTB [12] is the adaptation of capital charges to data quality through the concept of non-modellable risk factors. Contrary to the Basel 2.5 agreements [16], where all the risks in the trading book were deemed modellable, the FRTB [12] accounts for the issues that stem from missing data. The uncertainty of risk modeling when there is little historical data is not negligible. Thus, risk factors for which there is insufficient observed data are considered non-modellable, or NMRF, and they are subject to higher capital charges.

Naturally, the Basel Committee has set conditions that determine whether a risk is modellable or not. Thus, the risk factor eligibility test (defined in paragraph 31.13 in Chapter MAR31, which is dedicated to the model requirement of the IMA) of the FRTB [12]) allows banks to classify a risk factor as modellable if it meets one of the following two criteria over a quarter:

1. at least 24 real price observations over the previous 12 months and a minimum of four real price observations in a 90-day period or

2. at least 100 real price observations over the previous 12 months.

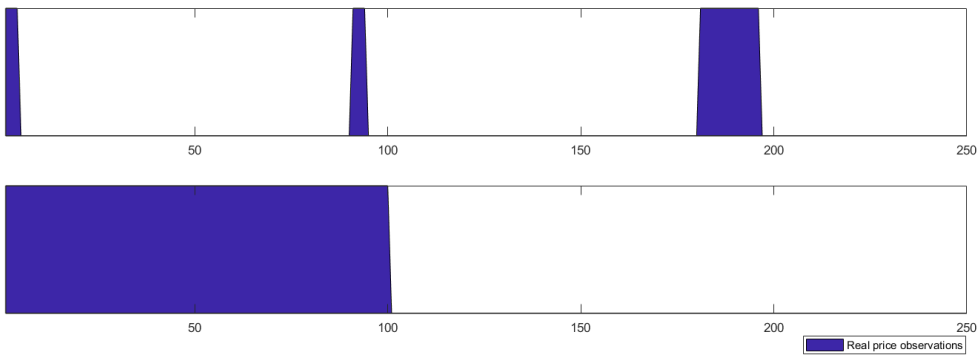
If one of the two conditions is met, then the risk factor is deemed modellable on a regular basis and is therefore subject to an ES calculation. If the risk factor does not meet either of the conditions, then it is considered an NMRF and is therefore subject to its own capital charge, determined by a stressed ES. Consequently, the capital charges associated with NMRF are higher than those associated with modellable risk factors. The challenge for banks is to have the most modellable risk factors, and they are encouraged to institute daily data backup processes.

This definition seems to be relatively unrestrictive, as series containing a large proportion of missing data, especially a large proportion of successive missing data, can be modellable. Two examples of modellable time series are presented in Figure 1.3-10, where real price observations are shown in blue and missing data in white. The first

example, at the top of Figure 1.3-10,corresponds to a time series of 250 data points (that is, one year) that is made up of two periods of four successive observed values followed by 86 successive missing data and then a period of 16 successive observed values followed by 54 successive missing data. This first example contains a total of only 24 observations, which are divided into three historical clusters (two contain four observations and one contains 16 observations); all the rest are missing. In other words, 90% of the historical data are missing. They are split into three groups of successive missing data that represent 34%, 34% and 22% of the total data, respectively. This time series, 90% of which comprises missing data, is indeed composed of 24 observations with at least four real price observations over a period of 90 days. It meets the first criterion from the definition of modellable risk factors.

Similarly, in the second example, which is shown at the bottom of Figure 1.3-10, a time series that consists of 100 successive observed values (a little short of five months) followed by 150 successive missing data (a little over six months) is also considered modellable, in line with the second criterion of the test. In this case, 60% of the sample is made up of successive missing data.

Fig. 1.3-10: Two examples of modellable time series over a period of one year



Therefore, modellability is attainable when missing data account for 90% of the series or when 60% of the series comprises successive missing data. This definition may not seem very restrictive, and it is much laxer than that of the first version of the FRTB [11]. The first definition required "at least 24 observable real prices per year with a maximum period of one month between two consecutive observations." The Basel Committee adjudged this first definition too demanding, even though only two real observed prices per month would have sufficed to pass the test. The relaxation of the criterion shows clearly how difficult it is for banks to gather good-quality historical data. The voluntary data collection exercise on NMRFs, which was conducted by EBA in 2019, lends further credence to this proposition. The results were published in December 2020 [80]. Eight institutions participated. One was German, four were French, two were Italian, and one was from the United Kingdom. Table 1.3-6 shows

the total number of risk factors and NMRFs collected in the exercise. They are sorted by risk category. Table 1.3-7 presents the distribution of the number of risk factors and NMRFs by bank and by risk category.

Tab. 1.3-6: Total number of risk factors and non-modellable risk factors sorted by risk category (Source: European Banking Authority, 2020 [80])

Risk factor category	Total number of risk factors	Of which: time series provided	Total number of NMRFs	Of which: time series provided	Average share of NMRFs
Commodity	2,921	2,921	211	211	7%
Credit spread	11,510	11,448	4,943	4,913	43%
Equity	16,016	15,686	4,485	4,482	28%
Foreign exchange	3,389	3,276	685	685	20%
Interest rate	14,449	10,737	5,222	4,516	36%
Total	**48,285**	**44,068**	**15,546**	**14,807**	**32%**

On average, 48,285 risk factor are reported of which 44,068 provide from time series. On average, 15,546 NMRF are reported, representing 32% of risk factors, and of which 14,807 provide from time series.

Tab. 1.3-7: Distribution of number of risk factors and non-modellable risk factors sorted by bank and by risk category (Source: European Banking Authority, 2020 [80])

	Number of banks	Min.	Q1	Median	Q3	Max.	Average
All RFs of which	**8**	**58**	**370**	**825**	**2,321**	**40,603**	**6,036**
Commodity	8	0	0	8	101	2,214	417
Credit spread	8	0	30	72.5	124	11,058	1,439
Equity	8	0	5	53	192	15,489	2,002
Foreign exchange	8	0	3	40	72	3„155	424
Interest rate	8	27	184	508	1,909	8,290	1,806
All NMRFs of which	**8**	**0**	**81**	**357**	**818**	**12,367**	**1,943**
Commodity	8	0	0	0	10	187	30
Credit spread	8	0	0	3	113	4,645	618
Equity	8	0	0	35	96	4,193	561
Foreign exchange	8	0	0	0	65	553	86
Interest rate	8	0	47	271	777	2,789	653
Share of NMRFs in total Rfs	**8**	**0%**	**15%**	**27%**	**55%**	**100%**	**37%**

On average, 6,036 risk factor and 1,943 NMRF are reported. The average share of NMRF among the total risk factors, for the 8 surveyed banks, is 37%.

In the data collected by the EBA, 48,285 risk factors were represented. Of those, participants considered 15,546 non-modellable (see Table 1.3-6). Thus, almost one-third of the risk factors that were collected were non-modellable, a very significant proportion. Evidently, the number of risk factors provided by the banks differed considerably: one bank provided 58 risk factors, while another provided 40,603 (see Table 1.3-7). This means that there is a wide dispersion in the number of risk factors reported by the banks. This is particularly visible in the gap observed between the mean and the median.

Turning to the distribution of risk factors, 87% concerned equity risk, interest rate risk and credit spread risk. Similarly, 94% of the NMRFs were included in the same

categories. Commodity risk and foreign exchange risk had very low shares. Proportionally, the NMRFs represented between 20% and 40% of all risk factors. Commodity risk is an exception. In that category, NMRFs accounted for only 7% of the total number of risk factors.

On average, 37% of the risk factors reported by the eight banks were NMRFs. The median is 27%. The difference between the average and the median is probably due to the fact that (at least) one bank took the view that none of the risk factors could be modeled. Conversely, (at least) one bank reported that it had been able to model all of its risk factors. Thus, both extremes are represented in this small sample (last line of Table 1.3-7).

These results should be treated cautiously because the sample is exceedingly small (only eight banks) and because data collection began in 2019. The FRTB regulation [12] was published in January of that year, leaving little time for the banks to consider and implement the modifications. Nevertheless, the analysis is very recent and shows that some banks still struggle to make risk factors modellable. Few banks were able to respond to the survey. It is difficult to believe that one of them categorized all risk factors as non-modellable. The implementation deadline of January 2022, one year after the publication of the report, highlights the urgency with which banks must act. In addition, the study provides an indication of the share of NMRFs in all risk factors, between a quarter and a third. This is a significant proportion that can lead to sizable capital charges, making the issue all the more exigent.

In addition to defining NMRFs, the Basel Committee has listed seven principles (paragraph 31.26 of Chapter MAR31) that banks must apply when modelling a risk factor. These principles enumerate modeling possibilities and condition them: it is possible to use combinations of modellable risk factors, the data must accurately reflect the volatility and correlation of the risk positions, the use of proxies is allowed but limited, and so on. The principles stipulate that banks must demonstrate compliance. Otherwise, their risk factors could be categorized as non-modellable.

The fifth principle is of particular interest because it concerns the updating of data. It states that data must be updated with sufficient frequency, ideally every day and at least once a month. Similarly, when the bank uses a particular model, each parameter that touches on the risk factor should be re-estimated regularly. The Basel Committee, therefore, encourages banks to establish their own data historization processes in order to be able to reproduce a particular scenario if necessary.

Notably, the last sentence of the fifth principle implies that banks may backfill or gap-fill their missing data when appropriate. The FRTB [12] is over 150 pages long, yet this is the only explicit mention of missing data and of gap-filling. The Basel Committee therefore authorizes banks to use completion methods to improve the quality of their data, provided that they are guided by clear policies. However, no indication is given

of the method to be used or the conditions under which a method would be accepted by the regulator. Since the principle is relatively binding, banks are obliged to toil in the dark. If a bank attempts to make a risk factor modellable, it must shoulder both the higher capital charges and the sunk cost of efforts to ensure modellability.

Historical data from 2007

Chapter MAR33 of the FRTB [12] is dedicated to the capital requirement calculation of the IMA. It begins by addressing the calculation of the ES (instead of the VaR). It is stated that banks must use the 97.5th percentile, a one-tailed confidence level and a liquidity horizon of 10 days to compute an ES for the last 250 days. Over a year, the ES that is calculated with these parameters gives the average of the 10-day returns behind the VaR (calculated with a 10-day horizon and a 97.5% confidence level). The measure must be calculated daily, implying that daily data must be available.

The calculation of the ES becomes more complex because it requires adjustments to be made on the basis of liquidity horizons. The riskier the risk factor, the longer the horizon that is applied. The possible horizons are 10, 20, 40, 60 and 120 days. Thus, the ES becomes the root of a quadratic sum of a 10-day ES with scaling applied to its horizon. Moreover, the regulation states that the ES must be calibrated for a stress period. The idea is to then calculate the ES of a portfolio for the case in which the relevant risk factors are undergoing stress. In this way, the set of risk factors is circumscribed, with only the ones that are most relevant and for which there is sufficient historical data considered. This reduced set of risk factors must explain at least 75

Thus, in paragraph 33.6 of the FRTB [12], the Basel Committee defined ES for market-risk capital purposes as

$$ES = ES_{R,S} \times \min\left\{\frac{ES_{F,C}}{ES_{R,C}}, 1\right\}, \tag{1.3-1}$$

where

1. $ES_{F,C}$ is the ES measure that uses the full set of risk factors, which is based on the current period (observations over the past 12 months),

2. $ES_{R,C}$ is the ES measure that uses the reduced set of risk factors, which is also based on the current period, and

3. $ES_{R,C}$ is the ES measure that uses the reduced set of risk factor, which is based on the most severe 12-month period of stress that is available over the observation horizon.

It would appear that $ES_{R,S}$ is scaled by the ratio (floored at 1) of the current ES, which is calculated from the full set of risk factors, to the current ES, which is calculated from the reduced set of risk factors.

The next paragraph (33.7) of the FRTB [12] defines the most severe 12-month period of stress, which is required for the calculation of $ES_{R,C}$. Banks are asked to identify the 12-month period during which the portfolio suffered the gravest losses. This period of stress must be determined from historical data that go back to at least 2007, and it must be updated at least quarterly.

These few lines of the FRTB regulation came as a shock to many banks: the historical data requirement demands a large amount of data, which are difficult to unearth for many risk factors. Of course, if the historical data do not meet the conditions for modellability that were quoted above between 2007 and the present date, the risk factors are categorized as non-modellable and are subjected to high capital charges. This complexity could explain why a bank might report that 100% of risk factors were categorized as non-modellable (see Table 1.3-7 from the EBA report [80]).

The FRTB regulation [12] differs significantly from all of its predecessors. It aims to address their weaknesses by defining the boundaries between the banking book and the trading book clearly. It also ensures the homogenization of banking regulations by imposing a standard approach on all banks, and it improves risk assessment by introducing more appropriate risk measures. Nevertheless, the regulation can only be successful if it is applied to high-quality data, especially modellable historical data that goes back to 2007. Therefore, before they can confront all the issues that modeling and implementing the FRTB [12] entails, banks must be able to build their own databases. Those databases should be complete and of as high a quality as possible.

Banks may have the most sophisticated models, but if they apply poor-quality data to them, the result will be unsatisfactory. The idea is captured well by the well-known adage, “garbage in, garbage out”.

1.3.5 The targeted review of internal models

Since banks are allowed to use internal models to calculate risks, the European Central Bank (ECB) has published a guide on internal models to ensure transparency in the interpretation of regulations and their application. The project is called a targeted review of internal models (TRIM). It was initiated by the ECB, and its first version was published in 2017. It has been updated on several occasions. Its latest version is called the ECB *Guide to Internal Models* [81], and it was published at the end of 2019.

The project aims to assess the conformity of the internal models (according to articles 143, 283 and 363 of Regulation (EU) No 575/2013 [167]) and to verify the reliability and comparability of their results. One of TRIM’s [81] main objectives is to reduce inconsistencies and the unwarranted variability of the models. Another is to ensure that supervisory practices are consistent through proactive monitoring and to guarantee the proper use of internal models. Of course, the ECB cannot monitor the

internal models of all European banks. It focuses on approximately 65 large banks, and it covers credit, market and counterparty risk (excluding operational risk).

The TRIM project [81] refers to the regulations presented in this chapter (and many others) and provides some additional details about the data and their quality. In the first chapter of TRIM [81], which is dedicated to credit risk, it is stipulated that any manual intervention into the data must be clearly formalized in order to ensure traceability and control. This point is particularly salient to paragraph 25 of the BCBS 239 standards [24], which indicates that the imputation of missing data is possible (although it must remain exceptional), provided that good-quality documentation is maintained and that all steps used are recorded for expert evaluation.

The ECB also requires banks to establish data verification processes and therefore data quality standards that govern the completeness, accuracy, consistency, timeliness, uniqueness, validity, availability (or accessibility) and the traceability of the data. A data quality check process, which is additional to verification, is also expected and must be applied throughout the lifecycle of the data. The quality check process must include procedures for identifying and correcting deficiencies in data quality.

All these stipulations substantiate the principles of the second axis of BCBS 239 [24], which is related to risk "data aggregation capabilities." The BCBS 239 principles can be vague. Through the TRIM project [81], the ECB indicates that banks are responsible for their data and their quality and that they are obliged to adopt methods for improving it, such as outlier detection and of course, data completion.

However, the ECB is silent on the completion methods that ought to be used and on the outlier detection tests that are permissible. Each of the changes applied to a time series must be documented because the regulator must read the documentation and verify the changes. Regulators may even develop their own models and compare their results with those of the banks to validate or reject them. Obviously, the banks would be interested in discovering the methods used by the authorities (if they use any) in order to benchmark their internal methods. Extending this argument further, it would be interesting to establish a standard approach to data quality methods and to authorize the internal approaches after validation, mirroring the FRTB regulation [12].

Whatever the methodology used, it must be documented clearly so that it can be inspected not only by the regulator but also by all of the relevant departments of the bank. Boosting transparency to improve decision-making processes throughout banks is one of the challenges of BCBS 239 [24]. Documentation is important for ensuring methodological transparency.

Although the regulation is silent on the matter, black-box methods such as machine learning do not seem to be suitable. Applying a machine learning algorithm in order to impute missing data could meet with rejection, even if it is documented well. The black-box aspects of some of those methods encounter a certain hostility at banks because

they remain theoretically unclear. Similarly, banks seek methodologies that reduce their capital charges by avoiding the categorization of a risk factor as non-modellable. Banks might even be prompted to choose the methodology that allows them to minimize capital charges. Ideally, the regulator should have provided a list that gives some indication of the methods that are acceptable according to their criteria.

The second chapter of TRIM [81] concerns market risk and includes an entire section on data quality. It specifies that the data used in an internal model must meet several standards and that compliance should be monitored on a quarterly basis. According to the ECB, the purpose of these standards is to ensure the observability of the true volatility of a position or portfolio. The ECB therefore asks the banks to set up quality controls that identify

- the number of daily data points that were missing initially and were then completed,
- the number of daily data points that were available initially and were then replaced,
- the number of daily data points without daily changes,
- the maximum number of consecutive days without daily changes.

In adopting all these standards, the ECB again recognizes the presence of missing data and outliers but also that of constant data, which tend to affect volatility, in financial series. As mentioned previously, banks have the duty to use the series that reflect true volatility best, hence the use of the outlier completion and detection method. When reporting, banks must also be able to provide documentation related to the latter or, more generally, to describe any changes in the time series in sufficient details.

In addition, the ECB recommends that banks should generate documents that are built around certain minimum standards on data quality and that they should be able to justify the use of time series that contain large amounts of successive missing data or successive data without daily changes. When few data are available, the bank must also explain why the available data are sufficient to reflect the true volatility of the series correctly or why the modifications to the series do not affect volatility.

The TRIM project [81] therefore appears to help banks to apply several regulations correctly. It provides practical answers to many practical queries about data collection, data quality and even missing data. This said, it does not define the expectations of the ECB clearly. Admittedly, it is difficult to define "good-quality data." However, the project seems to be a first attempt to arrive at a definition.

Maximum thresholds for the standards should have been set. The definition of a NMRF from FRTB [12] refers to precise numbers of missing daily data points, but it is

otherwise unclear. Data that are available initially but are then replaced because they are not representative should not be price observations under the NMRF criteria, but the regulation does not state so explicitly. In the same way, the supervisory authorities could identify a maximum number of permissible successive days without daily changes. This would give banks a more precise indication of the modifications that they need to introduce to their time series.

Finally, the ECB insists on the preservation of the volatility of time series when using the completion method. That volatility must be calculated. However, when a time series contains missing data, calculating volatility solely by reference to observed data can skew results. If only the time series is used, then it must be assumed that the volatility of the missing data is the same as that of the observed data, which is not necessarily true. It may be possible to use a benchmark to determine the volatility of the series on the assumption that the high correlation between the volatilities of the series and its benchmark remains consistent over time when there are missing data. The ECB expresses its interest in preserving volatility, but it does not identify the methods that ought to be used to attain this goal. For example, the TRIM guide states that "there should be no missing data points for the final time series of shocks used to calibrate the model." Meeting this requirement is much more complicated than it seems. No additional information is given on an acceptable way of handling missing data from a regulator's point of view. However, a standard approach based on data imputation would be welcome. The TRIM project [81] contains clues about its content, but it remains unclear.

1.3.6 Regulators and data providers

So far, this chapter has highlighted the impact of new data quality regulations on banks. Obviously, these regulations mostly affect the banks to which they are addressed. However, they also pose challenges to two other actors: regulators and data providers.

Regulator

The role of regulators is paramount. They must develop new regulations and update existing ones, but they must also strive to ensure that banks understand and apply those regulations properly. The regulatory innovations of the last few years have prompted many questions. Regulators must engage with banking institutions continuously in order to answer those questions and to understand the difficulties that they encounter. When necessary, regulations must also be modified.

This relationship manifests in the numerous questions-and-answers (Q&A) and frequently-asked-questions (FAQ) documents that are circulated regularly to clarify the text of the regulations. In fact, some articles of the CRR regulation [167] that were presented earlier in this section incorporate Q&As. Likewise, the text of the FRTB

citefrtb2019 includes FAQs. Several data quality questions are included, for example in the NMRF section. The Basel Committee even shared a FAQ on market-risk capital requirements [22], between the publication of the first and the last version of the FRTB [11] [12] in order to encourage banks and thus avoid delays. The text of the FAQs [22] about the FRTB was published in January 2017, updated in March 2018, and it was finally integrated into the FRTB [12] in January 2019.

In addition to FAQs and Q&A, the supervisory authorities regularly communicate with banks through surveys, exercises, quantitative impact studies (also known as QIS) and progress reports. These collaborations enable the participants to grasp the regulations that they must deal with. In fact, these very reports and surveys have been used to illustrate regulations throughout this chapter. In particular, the overview drew on the annual reports on the adoption of the BCBS 239 principles [19][20][21][23][25], the IFC survey on the quality of data at trade repositories [121], the report on CVA [79] that reflects the difficulties of using proxies of credit spreads and the data collection exercise [80] that reflects the difficulties of using proxies of credit spreads and the data collection exercise

The documents have the advantage of maintaining dialogue between the banks and the regulator: the banks ensure that they understand the regulations (mainly through FAQs and Q&A), and the regulator monitors the difficulties that the banks encounter. The documents also allow banks to help each other. Not all banks face the same problems at the same time. The questions that are asked by some banks are sure to be asked by others later, hence the appeal of the publications. The reports overview the application of the regulations, highlighting problems and deficiencies. Supervisory authorities, then, produce many reports and surveys. These publications enable them to monitor the implementation of the regulations, to clarify them and to discuss specific cases.

Incidentally, this is also the purpose of the last three principles of BCBS 239 [24]. The reader might recall that the regulation in question comprises 14 principles. The first 11 are addressed to banks, and the last three are directed at authorities. Notably, the latter must ensure effective supervision, compliance and enforcement. Similarly, regulatory texts such as the FRTB [12] and the TRIM project [81] implicitly target banks by asking them to control and validate IMA. The FTRB stipulates that banks must verify compliance with modellability criteria at least once every quarter. The implication is that regulators may have to monitor compliance with these criteria four times a year. As noted earlier, reports can highlight failures and regulatory weaknesses. As a result, regulations are updated, redesigned or amended.

Returning to missing data, the BCBS 239 regulation [24], the FRTB [12] and the TRIM [81] mention that banks may use methods to complete missing data. Their use must be documented comprehensively to ensure transparency and to enable regulatory validation. The regulators act as validators in the same way as in the IMA. The

authorities should therefore ensure that they understand the methodologies used by banks to fill in missing data or to detect outliers. As mentioned earlier, they may choose methodologies that allow them to minimize capital charges. The regulator must ensure that the bank has tried to reconstruct its data history as faithfully as possible and without seeking to reduce volatility.

It follows that banks are not the only ones affected by missing data. The regulator cannot bypass them, either. Regulations must be adapted to these problems. Provisions on data quality may have to be introduced into existing texts. It may even be necessary to dedicate entire regulations to data quality. As soon as regulators authorize banks to complete historical data for regulatory purposes, they must ensure that banks do minimize their risk exposure by selecting a methodology strategically. Therefore, data quality problems compound the complexity of regulation.

Data providers

Data providers are also affected by data quality. If banks are confronted with incomplete historical data, the same is likely true of data providers. This is one of the points that Section 1.1.1 highlighted through the case study of price data on WTI crude oil. Recall that data for this stock had been downloaded from four different sources: EIA, Bloomberg L.P., Investing.com and Yahoo Finance. The data from the EIA and Yahoo Finance were identical and had the most missing data. The Bloomberg series contained slightly fewer missing data, and Investing.com was the source with the most available observations. The study sample also revealed differences in the spot prices cited by the sources. The reasons for these differences and the absence of quotations were not explained by the data providers. Therefore, it is sensible to inquire whether data providers use specific methodologies to improve data quality.

A data provider sells data. To navigate the market, they must offer customers data of the best possible quality. A real market for data has developed in the digital age. It is impossible to imagine a career as a trader that does not involve investing several tens of thousands of euros into a license to access to valuable market data.

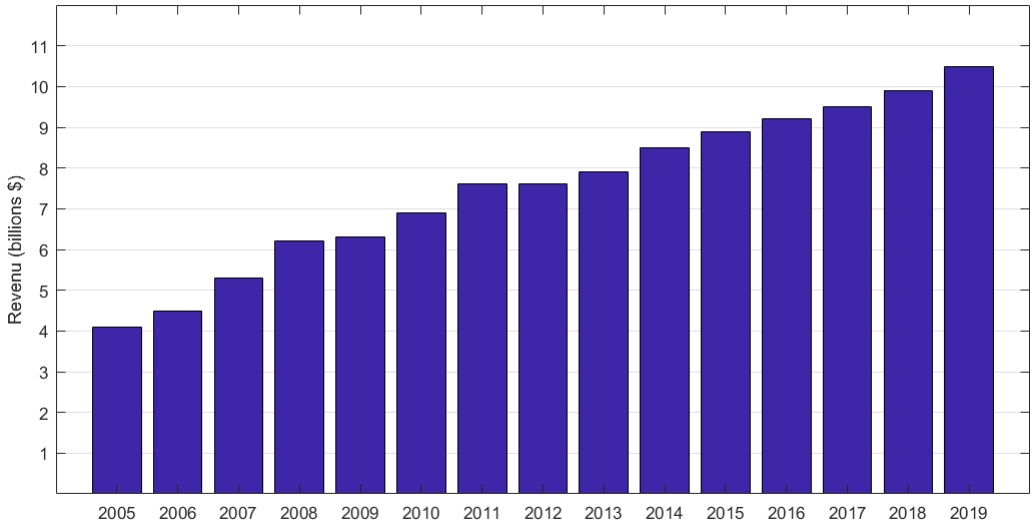
Bloomberg L.P. is one of the leaders in the data market. It emerged in the 1980s. According to Douglas B. Taylor [195], the managing director of data research and consulting firm Burton-Taylor International Consulting, who spent years working for Thomson Reuters (Bloomberg's primary competitor), its success can be attributed to the evolution of its services in line with the market and its needs.

Despite a very expensive annual license fee, which ranges $\$20,000$ to $\$24,000$ per year, Bloomberg L.P. had more than 330,000 subscribers in 2018. The Bloomberg terminal contains an impressive array of market data, financial instruments, stock exchanges from around the world and company information. It also provides tools for transactions, communication, financial calculations, financial analysis and investment comparisons.

Evidently, the business of data providers involves not only database maintenance but also the provisions of tools and services.

The revenues of Bloomberg L.P. (see Figure 1.3-11) have been increasing steadily since 2005, which indicates that the data market is expanding. There was no decline during the subprime crisis: traders need data even when markets are collapsing.

Fig. 1.3-11: Annual revenues of Bloomberg L.P. since 2005 (Source: Burton-Taylor International Consulting, a TP ICAP company [195])



Between 2005 and 2019, bloomberg's revenues have grown from about $4 billions to about $10.5 billions

Speaking to the Vox newspaper in 2019 [195], Taylor said that since the beginning "it [Bloomberg L.P.] looks the same, but it actually does more and it has evolved with the markets and its needs." More data has been added to the Bloomberg database. Its content reflects both novel financial products and new regulatory requirements. Modern Bloomberg terminals are increasingly integrating risk measure calculation tools, such as the possibility to calculate VaR or tracking errors.

Of course, the introduction of new regulatory requirements has benefitted data providers. The modellability criteria for risk factors and the requirement to produce historical data that goes back to 2007 have undoubtedly been profitable for Bloomberg L.P. (see Figure 1.3-11). The FRTB regulation [12] requires that these historical data be furnished for risk measures to be applied. Many banks used market data on an ad hoc basis, without necessarily recording it. Some of that data can now be bought from data providers.

Section 1.1.1 described a problem of this kind that Natixis encountered in 2017. The FRTB regulation [12] requires data on historical repo rates. The data that Natixis had

in its possession covered historical repo rates that were implied by forward contracts. They were of average quality. When the repo rate is implied, it is subject to instabilities that are not specific to it but which are due to all the other parameters of the forward contract. The repo rate, then, becomes the adjustment variable for the correct price of the forward contract. Thus, the historical data that Natixis had could have been affected by periods of high volatility that did not necessarily correspond to market fluctuations. For this reason, the bank decided to deduce the repo rates from the TRS by taking the opposite of the TRS spread (see Section 1.1.1 for further details). This methodology yielded much more stable historical data by recovering real market information from discussions between traders and brokers. Of course, this type of data is not included in the annual license. If Natixis had been unable to discover its own historical repo rates, the bank would have been forced to incur considerable expense to purchase them from a data provider to avoid having the risk factor categorized as non-modellable.

Bloomberg L.P. also offers services outside of their terminals, including a missing data completion service. Their completion methodology is based on the singular spectrum analysis algorithm (presented in Section 2.4.4). Bloomberg L.P. probably decided to commercialize their own completion algorithm because they saw banks struggling to unearth the high-quality historical data that regulators require. The commercial proposal that Bloomberg L.P. made to Natixis indicates that this is likely to have been the case. Once the latter had collected all the data that it needed to build its repo rate history, it discovered missing data. Bloomberg L.P. offered to solve the problem. Natixis preferred to use their own missing data completion algorithm, which will not be presented for reasons of confidentiality. It produced satisfactory results, and it enabled savings to be realized.

The example shows that a bank can reconstruct its data history without the help of data providers. There are many risk factors. Many banks have missing data issues. This state of affairs poses a real challenge to data providers. The regulations offer data providers the opportunity to increase their profits through the sale of historical data (which can be deep and is therefore expensive) or the completion of banks' missing data. The completion models that the data providers offer must also be documented well if they are used for regulatory purposes. Therefore, data providers must comply with regulations in order to sell their services: even if their methodologies are not subject to direct oversight, they are reviewed once purchased.

Finally, that data providers possess completion methods raises doubts about the authenticity of the data that are available directly from terminals. The historical data on a Bloomberg terminal (or that of another company) may not be true market data. It is possible that it may have been corrected or imputed in the course of a data quality check. No information on this subject is available on the terminals or on the Bloomberg website. This is a matter that regulators might need to investigate.

This chapter showed that missing data, despite being only one of the many elements of data quality, are at the root of many problems in financial modeling and analysis. The example of WTI crude oil prices that was presented at the beginning of the chapter (see Section 1.1.1) shows that missing data are commonplace. Numerous studies have identified the problem: missing data are present in 33% of management research articles, according to Tsikriktsis [201], and Kofman and Sharpe [133] report that they affect 27% of financial research articles.

Missing data bias results and reduce statistical power. They pose many challenges to banks. Many of the financial regulations that have been passed account for missing data, at least to some degree. Attempts are made to remedy missing data problems. Banks have been forced to create databases of the best possible quality. Otherwise, they risk significant capital charges. As their field of action is expanding, regulators are also increasingly affected by missing data. Finally, missing data are important to data providers, for whom new regulations represent an opportunity to increase profits.

Chapter 2:
Missing data and alternatives

This chapter presents the theory of missing data. It covers both their categorization and their distribution, and it discusses them in the context of financial series. The first step in completing data is to encode it, that is, to construct a sample. Encoding involves many choices that impact completion directly. Once this step has been executed, the completion method can be applied. However, it is still necessary to approach its choice with caution. There are many completion methods. In this chapter, they are described. In the next, they are compared. Each method relies on a number of parameters that influence its effectiveness significantly. The methodologies and their parameters will be presented and illustrated with applications from the literature. Not all of the applications originate from finance research. Once all this has been explained, it will become possible to intuit the results of the application of the algorithms to financial data, which form the subject matter of the next chapter.

Contents

2.1 Theoretical presentation of missing data

Since missing data is a vast and interdisciplinary subject, several researchers have succeeded in establishing a theoretical framework to best respond to the problems it raises. Two major categorizations of missing data are discussed in this section. Following this discussion, the statistical tests developed to specify the category of missing data are presented. Finally, the last section describes the different distributions of observable missing data.

2.1.1 Categorization of missing data

Before presenting the different categorizations of missing data, Graham [98] discusses, in his book, what is missing data and in particular the two kinds of missing data present in the literature. According to him, the missing data from item nonresponse and those from wave nonresponse.

Non-response items occur when part of the survey is completed but not all of it, for example, because a question was forgotten or because the survey is too long. On the other hand, missing data due to wave non-response are specific to longitudinal research studies, i.e. studies where the same individual is surveyed several times (waves). Thus, wave nonresponse is the case when the respondent forgets to answer the survey or is absent during a wave of the longitudinal study.

Item nonresponse is not really a problem according to Graham [98] because there are many cross-sectional solutions in the literature to overcome it. On the other hand, the case of wave nonresponse becomes extremely problematic when the individual does not respond until the end of the longitudinal study (if the individual responds to the last wave, an interpolation is possible).

The context described by Graham seems to be specific to survey-based studies, but it can be applied to a financial data context. Missing data due to a nonresponse item would correspond, for example, to data that would not have a quote on a specific date or to the trader forgetting to save a data manually, while the wave nonresponse could correspond to days when the market is closed (for example, after a weekend, when no information is available since Friday), or to a deletion of historical data by mistake.

Graham [98] therefore presents the missing data in a context specific to survey-based studies, but not at all generalized. This is why he then presents the two major categorizations of missing data in the literature: that of Little and Rubin [145] in 1987 and more recently, that of Schafer and Graham [182] in 2002.

Little and Rubin's categorization

The theory of missing data was initiated by Little and Rubin in 1987 in *Statistical analysis with missing data* and remains the reference book today [145]. They define missing data as unobserved values that would be meaningful for analysis if observed; in other words, a missing value hides a meaningful value. This definition fully justifies the existence of imputation methods.

Little and Rubin [145] then focus on various patterns of missing data. They explain that the missingness mechanisms are crucial because the properties of missing data methods depends very strongly on the nature of the dependencies in these mechanisms. Let $Y = (y_{ij})$ be a data matrix and $M = (m_{ij})$ be the missingness indicator matrix such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is observed. Assuming that the rows (y_i, m_i) are independent and identically distributed over i, the missingness mechanism is characterized by the conditional distribution of M given Y, say $f(M|Y, \phi)$, where ϕ denotes unknown parameters.

The data are called **missing completely at random** (MCAR), if missingness does not depend on the values of the data, missing or observed, which means,

$$f(M|Y, \phi) = f(M|\phi) \tag{2.1-1}$$

In this case, each observation in a variable has the same probability of missingness.

They describe another mechanism, called **missing at random** (MAR). It is the case, the probability of missingness depends only on available information or, according to the notations previously used, the distribution of M depends on the missing components of Y.

Let Y^{obs} denote the components of Y that are observed and Y^{miss} denote the components of Y that are missing. A less restrictive assumption than MCAR is that missingness depends on Y only through the observed components Y^{obs}, that is

$$f(M|Y, \phi) = f(M|Y^{obs}, \phi) \tag{2.1-2}$$

However, this notion is not new since it was first introduced in 1976 by Rubin [174], where he explains that it is appropriate to ignore the missing data process if they are MAR but also that the observed data are "observed at random".

Finally, they present the mechanism called **missing not at random** (MNAR), corresponding to the data that are neither MCAR nor MAR, that is when the distribution of M depends on Y^{obs}, the observed component of Y and Y^{miss} the missing components of Y:

$$f(M|Y,\phi) = f(M|Y^{obs}, Y^{miss}, \phi) \tag{2.1-3}$$

Little and Rubin [145] explain then that MAR is a sufficient condition for pure likelihood and Bayesian inferences to be the valid without modeling the missingness mechanism.

Moreover, in 2015, Mealli and Rubin [152] introduced the term *always* on the classification of Little and Rubin [145]. Indeed, data are called missing always at random (MAAR) when the missing data mechanism always produces data that are MAR. Hence, the data are MAAR if any missing data of Y depend on Y^{obs}; in other words, they are all MAR and all respect the Equation 2.1-2. If a missingness mechanism is MAAR, then any realization from it is MAR.

In the same way, they introduce the notions of missing always completely at random (MACAR) and missing not always at random (MNAAR). In the case of MACAR, each of the missing data of Y are MCAR and therefore, satisfy the Equation 2.1-1. In the same way, the data qualify as MNAAR if each of the missing data of Y are MNAR and therefore, satisfy the Equation 2.1-3. Due to its very specific nature, this type of missing data remains rarely discussed in the literature.

Schafer and Graham's categorization

In 2002, Schafer and Graham [182] presented their point of view on the missing data framework. In particular, they note that the definitions given by Little and Rubin [145] essentially derive relationships between observed and missing data, not cause and effect relationships. Hence, they add that the three types of missing data can be represented as in Figure 2.1-1, where M is the missingness pattern and where Z denotes the component of cause that is unrelated to X (completely observed) and Y (partly missing). MCAR means that the data are missing with probability unrelated to any variables in the system; in other words, it requires that the causes of missingness are entirely contained within the unrelated part Z. Schafer and Graham [182] include in Z all reasons for missingness, such as age or health status, that could explain why participants did not show up, cognitive functioning that would help understand why individuals tend to answer, "I don't know" or privacy concerns that would explain an outright refusal to respond. Thus, in case of MCAR, these causes are directly

responsible for the missingness M, without being related to X and Y (see Figure 2.1-1 [a]). Mathematically, the result is as follows:

$$f(M, Y|X, Z) = f(M|Z)f(Y|X) \tag{2.1-4}$$

In a financial context, data that would have been, for example, manually deleted would be considered as MCAR according to Schafer and Graham [182]. The missing data mechanism M depends on a variable Z which corresponds to a history listing all the modification dates of the database. In this case, Z has no relationship with Y (and X). Unlike MCAR, MAR allows some causes of Z to be indirectly related to X (see Figure 2.1-1 [b]) resulting in

$$f(M, Y|X, Z) = f(M|X, Z)f(Y|X) \tag{2.1-5}$$

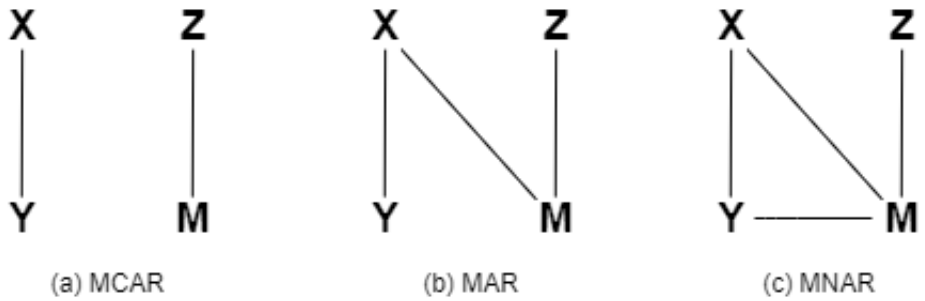
In this case, if the missing data are due to a suspension of quotation, then there is a variable X, corresponding to the dates of quotation of Y, related to the series Y and thus allowing to explain the mechanism of missing data M. Z can correspond to the same variable as in the previous case (MCAR), thus confirming that no manual deletion has taken place.

Moreover, MNAR requires some causes to be residually related to Y after relationships between X and M are taken into account (see Figure 2.1-1 [c]) which leads to

$$f(M, Y|X, Z) = f(M|X, Z)f(Y|X, M) \tag{2.1-6}$$

This missing data mechanism appears, for example, when the extreme data of a series are removed volontarly. In this case, the missingness mecanism M is explained by: the values of Y which no longer contains extreme values; X which corresponds to a series strongly correlated to Y allowing to understand that the missing data of Y are its extreme values; and finally Z corresponding to the history of the dates of modifications of Y in the database.

Fig. 2.1-1: Representation of the three types of missing data according to Schafer and Graham (Source: Schafer and Graham, 2002 [182])



Gelman and Hill's categorization

Gelman and Hill [90] (2006) defined four different default mechanisms: missingness completely at random, missingness at random, missingness that depends on unobserved predictors and missingness that depends on the variable itself. The first two types of missingness are common to the definitions of Little and Rubin [145] and the next two are more specific than the notion of MAR.

They present the missingness that depends on unobserved predictors, which depends on information that has not been recorded and that predicts the missing values. They present the example of people with a college degree, who are less likely to reveal their incomes, i.e. having a college degree is predictive of incomes. The income level are not MAR. Let Y be the partially observed component, M the missingness matrix and V the unobserved component. The data are qualified as missing depending on unobserved predictors if $f(M|Y,V) = f(M|V)$.

Finally, they complete their presentation by introducing the missingness that depends on the missing value itself, which means that the probability of missingness depends on the (potentially missing) variable itself. For example, people with higher incomes are less likely to reveal them. In their example, all persons earning more than $\$100,000$ refuse to respond. More formally, let Y^{miss} be the missing part of Y, in their example, $Y^{miss} = \mathbb{1}_{(Y>\$100,000)}$; then, missing data depends on the missing value itself if $f(M|Y) = f(M|Y^{miss})$.

Although there is some disagreement on the categorization of missing data, there is general consensus that there are several types of missing data that are more or less randomly characterized. Moreover, Little and Rubin's [145] and Schafer and Graham's [182] categorization are the most widely used in the literature.

These two categorizations seem to be close, yet in practice the category of missing data depends on the definition chosen. A mechanism can be MAR in the sense of Schafer and Graham [182] if the missing data of a series depend on the observations of another series, but if this same series is absent from the data matrix, then it means that the data are MCAR in the sense of Little and Rubin [145]. Similarly, data could be considered as MCAR in the sense of Little and Rubin [145] although the probability of having missing data is not the same for each observation. These discrepancies are discussed in more detail in a practical case, in the next chapter (see Section 3.3.2).

2.1.2 Testing MCAR data: Little's test & Jamshidian and Jalal's tests

Once the categorization of the missing data was completed, two main statistical tests were developed to determine whether or not the data were MCAR in nature.

Little's test

When historical data are used, it is not always clear whether the data are MCAR, MAR, or MNAR. For this reason, shortly after the classification of missing data with Rubin [145], Little provides a likelihood ratio test of the assumption of MCAR [142]. It is in 1988 that he presents his test of MCAR, also called "Little's test" in the literature.

Little [142] presents his test as follow. Consider the data matrix $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})^T$ to be an i.i.d. (independent and identically distributed) sequence of p-dimension, with $(i = 1, 2, \ldots, n)$ where n is the sample size and where $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T$ is the $(n \times p)$ data matrix. He supposes that there are J missing value patterns among all y_i. Fully observe cases, if present, count as a pattern. Moreover, let $\boldsymbol{\mu}_j^{obs}$ and $\boldsymbol{\Sigma}_{.j}^{obs}$ be respectively, the mean vector of dimension p_j and covariance matrix of dimension p_j of only the observed components for j-th missing pattern and $\bar{\boldsymbol{y}}_j^{obs}$ the observed sample average for the same j-th missing pattern. Finally, consider $\boldsymbol{I}_j \subseteq \{1, 2, \ldots, n\}$ to be the index set of pattern j in the sample and $n_j = |\boldsymbol{I}_j|$, then $\sum_{j=1}^{J} n_j = n$.

The Little's test statistic for MCAR is as follows:

$$d_0^2 = \sum_{j=1}^{J} n_j (\bar{\boldsymbol{y}}_j^{obs} - \boldsymbol{\mu}_j^{obs*})^T \boldsymbol{\Sigma}_j^{obs\,-1} (\bar{\boldsymbol{y}}_j^{obs} - \boldsymbol{\mu}_j^{obs*}). \tag{2.1-7}$$

where $\boldsymbol{\mu}_j^{obs*}$ denote the maximum likelihood estimate of $\boldsymbol{\mu}_j^{obs}$. This test statistic follows a χ^2 distribution with a degree of freedom equal to $\sum_{j=1}^{J}(p_j) - p$

If the data are MCAR, then conditional on the missing indicator $\boldsymbol{m}_i$ (with $\boldsymbol{m}_i = (m_{i1}, \ldots, m_{ip})^T$ denote the indicator of whether each component in vector $\boldsymbol{y}_i$ is observed; that is, $m_{ik} = 1$ if y_{ik} is observed and $m_{ik} = 0$ if y_{ik} is missing, for $i = 1, \ldots, n$ and $k = 1, \ldots, p$), the following null hypothesis is,

$$H_0 : \boldsymbol{y}_i^{obs} | \boldsymbol{m}_i \sim \mathcal{N}(\boldsymbol{\mu}_j^{obs}, \boldsymbol{\Sigma}_j^{obs}) \qquad \text{if } i \in \boldsymbol{I}_j, 1 \leq j \leq J, \tag{2.1-8}$$

where $\boldsymbol{\mu}_j^{obs}$ is a sub-vector of the mean vector $\boldsymbol{\mu}$.

If H_0 is rejected, then the alternative hypothesis is supposed true:

$$H_1 : \boldsymbol{y}_i^{obs} | \boldsymbol{m}_i \sim \mathcal{N}(\boldsymbol{\nu}_j^{obs}, \boldsymbol{\Sigma}_j^{obs}) \qquad \text{if } i \in \boldsymbol{I}_j, 1 \leq j \leq J, \tag{2.1-9}$$

where $\boldsymbol{\nu}_j^{obs}$, $j = 1, \ldots, J$ are mean vectors of each pattern j and can be distinct.

Little [142] specifies that his test still works even if the normality of the data is not satisfied, because of the asymptotic sense for quantitative random vectors (but it is not suitable for categorical variables). If the data under each y_i are not multivariate normal, but has the same $\boldsymbol{\mu}$ and the same $\boldsymbol{\Sigma}$, then d_0^2 follows asymptotically the same χ^2 distribution, thanks to the multivariate central limit theorem.

Because of the missing values, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown, Little [142] proposes to replace them with the unbiased estimator $\hat{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}} = n\hat{\boldsymbol{\Sigma}}/(n-1)$, where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$

are the maximum likelihood estimators based on H_0 and can be obtained with a EM algorithm (see Section 2.4.6) using $\boldsymbol{Y}^{obs}$ the observed value of $\boldsymbol{Y}$.

In the same way, Σ_j^{obs} in Equation 2.1-7 is replaced by the sub-matrix $\tilde{\Sigma}_j^{obs}$ from $\tilde{\Sigma}$. Thus, the null hypothesis (Equation 2.1-8) given above, becomes

$$d^2 = \sum_{j=1}^{J} n_j (\bar{\boldsymbol{y}}_j^{obs} - \hat{\boldsymbol{\mu}}_j^{obs})^T \tilde{\boldsymbol{\Sigma}}_j^{obs\,-1} (\bar{\boldsymbol{y}}_j^{obs} - \hat{\boldsymbol{\mu}}_j^{obs}). \tag{2.1-10}$$

As d_0^2, d^2 follows the same χ^2 distribution with $\sum_{j=1}^{J}(p_j) - p$ degree of freedom. Thus, H_0 is rejected of $d^2 > \chi^2_{\sum_{j=1}^{J}(p_j - p)}(1-\alpha)$ where α is the significance level.

Little [142] finally applied his test on simulated data to check the consistency of the results. To do this, he generated three different samples of data that followed a multivariate normal, skewed (log-normal) and a long-tailed (χ_3^2) distribution from which he removed 60% of the data by an MCAR mechanism. According to this procedure, 1,000 incomplete data-sets were generated (the same sample with different missing values) before Little's test was applied.

The results are presented in Table 2.1-1. For example, the first value of the table means that the null hypothesis of MCAR (Equation 2.1-8) was rejected in 202 cases out of 1,000 with a confidence level of 20%. In addition, the superscripts a and b show a significant difference between the empirical size and the nominal level at 1% and 5%, for sample size equal to 40 and 20. For the biggest sample size, there is no significant difference from nominal level. The test seems to be more conservative for the small sample, particularly with the lower nominal level.

He adds that there is a relatively small impact of non-normality on empirical sizes, suggesting a fairly high degree of robustness for the method. This statement reinforces the fact that, asymptotically, this test does not require normality.

Tab. 2.1-1: Percent empirical sizes for a test of the MCAR assumption, from N=1000 simulated data sets (Source: Little, 1988 [142])

Sample size	Distribution	Nominal level of test			
		20%	**10%**	**5%**	**1%**
80	Normal	20.2	10.9	4.9	0.5
	Lognormal	18.9	8.8	3.7	0.7
	t on 3 df	21.2	11.2	5.5	1.0
40	Normal	21.2	9.5	2.5^a	0.2^b
	Lognormal	18.8	8.9	3.2^b	0.3^b
	t on 3 df	20.8	9.6	4.1	0.8
20	Normal	20.3	6.8^a	2.8^a	0.3^b
	Lognormal	21.1	8.3	3.5^b	0.5
	t on 3 df	21.6	8.1^b	2.0^a	0.3^b
Standard errors		1.27	0.95	0.69	0.315

[a] 1% level of significance.
[b] 5% level of significance.
The null hypothesis of MCAR was rejected in 20.2% of the cases with a confidence level of 20%, and with a normal distribution.

Nevertheless, he concludes by pointing out some of the limitations of his test and suggesting alternatives. The first limitation is that Little's test appears to be more suitable for quantitative rather than categorical variables. Another limitation of the test and not the least, is that d^2, even in the alternative hypothesis, allows missing data to affect the means but constrains the variances and covariances to be the same for all patterns. Furthermore, in the case of a large number of missing values, this assumption may not be satisfied. Thus Little [142] proposes to relax this limitation on covariance matrices and replace H_1 (in Equation 2.1-9) by

$$H_1^* : \boldsymbol{y}_i^{obs} | \boldsymbol{m}_i \sim \mathcal{N}(\boldsymbol{\nu}_j^{obs}, \boldsymbol{\Gamma}_j^{obs}) \qquad \text{if } i \in \boldsymbol{I}_j, 1 \leq j \leq J, \tag{2.1-11}$$

where $\boldsymbol{\Gamma}_j^{obs}$ is the new covariance matrix that contains, as $\boldsymbol{\nu}_j^{obs}$, distinct parameters for each missing patterns j. Thus, to test H_0 against H_1^*, Little [142] gives the following augmented likelihood ratio statistic :

$$d_{aug}^2 = d^2 + \sum_{j=1}^{J} n_j [tr(\boldsymbol{S}_j^{obs} (\hat{\boldsymbol{\Sigma}}_j^{obs})^{-1}) - p_j - ln|\boldsymbol{S}_j^{obs^{-1}}| + ln|(\hat{\boldsymbol{\Sigma}}_j^{obs})^{-1}|] \tag{2.1-12}$$

where $\boldsymbol{S}_j^{obs}$ is the sample covariance matrix of the observed values of pattern j and where $tr(.)$ is the trace operator.

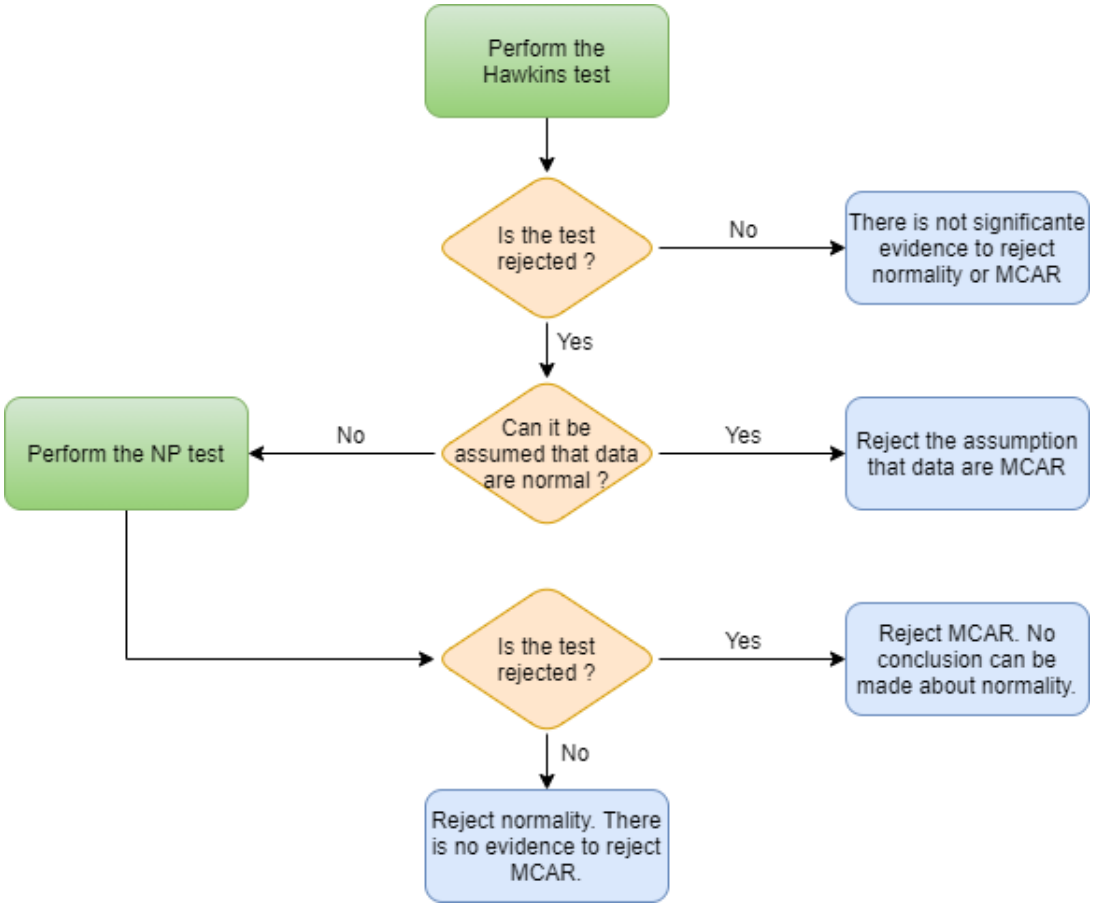
Under H_0, d_{aug}^2 follows a χ^2 distribution with $\sum_{j=1}^{J} p_j(p_j+3)/2 - p(p+3)/2$ degree of freedom. This degree of freedom can be large (42 in his study), leading to doubts about the effectiveness of the test. Moreover, d_{aug}^2 may not be applicable if some patterns

have too-small sample sizes, such that $n_j \leq p_j$ since $\boldsymbol{S}_j^{obs}$ will be singular, so $ln|\boldsymbol{S}_j^{obs}|$ cannot be computed.

Jamshidian and Jalal's tests

Since Little's test, other tests have been developed and one of the latest is that of Jamshidian and Jalal (2010) [123]. They use an improvement on Hawkins' test (in 1981) [106] in conjunction with a non-parametric test to determine whether the data verify the MCAR hypothesis. Hawkins' test is initially a multivariate normality and homoscedasticity test, which works specifically on small and complete samples. It has been improved for application to samples with missing data and for testing of whether these are MCAR. In addition, Jamshidian and Jalal [123] combine it with a non-parametric k-sample test, a homoscedasticity test, to determine whether the data are MCAR. If the null hypothesis of the improved Hawkins' test is rejected, then this means that the data are non-normal and/or homoscedastic (i.e. not MCAR) and if the null hypothesis of the non-parametric test is accepted, then the data are definitely homoscedastic. It can then be concluded that the data are MCAR (and non-normal). The whole mechanism of Jamshidian and Jalal's tests has been summarized in Figure 2.1-2.

Fig. 2.1-2: Flowchart of Jamshidian and Jalal's test (Source: Jamshidian and Jalal, 2010 [123])



First of all, they present the Hawkins' test as follows. Let $\boldsymbol{Y}_i$ denote the matrix, of dimension $(n_i \times p)$, of values corresponding to the i-th missing data pattern group in $\boldsymbol{Y}$, with $\boldsymbol{Y}_i^{obs}$ and $\boldsymbol{Y}_i^{miss}$ respectively denoting the observed and the missing part of $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_{ij} = (\boldsymbol{Y}_{ij}^{obs}, \boldsymbol{Y}_{ij}^{miss})$ denoting the j-th case in $\boldsymbol{Y}_i$. Let $\boldsymbol{m}_{ij}$ denote a vector of dimension p composed by indicator variables with elements 1 and 0, if the value in $\boldsymbol{Y}_{ij}$ is respectively observed and missing. They assume that given $\boldsymbol{m}_{ij}$, $\boldsymbol{Y}_{ij}$ has the density $f(\boldsymbol{Y}_{ij}; \boldsymbol{\Sigma}_i, \boldsymbol{\theta})$ with $\boldsymbol{\Sigma}_i = cov(\boldsymbol{Y}_{ij})$ the covariance matrix depending on the missing data pattern i and $\boldsymbol{\theta}$ including homogeneous parameters.

Their further analysis is based on the fact that, under this setting, testing the homogeneity of covariances is like testing whether the missing data are MCAR, because homogeneity of covariances implies MCAR. Then, Jamshidian and Jalal [123] propose testing MCAR using Hawkins' test of homoscedasticity and normality. This test assumes that a set of complete data $\boldsymbol{X}$ of dimension $(n \times p)$ from g groups is

available (g is the number of missing data patterns), with $\boldsymbol{X}_{ij}$ denoting the j-th case from the i-th group (with $j = 1, \ldots, n_i$ and $i = 1, \ldots, g$). Hawkins [106] assumes that $\boldsymbol{X}_{ij} \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ (where $\mathcal{N}_p$ is a p-variate normal distribution).

The null hypothesis of this test can be written as follows:

$$H_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_g \equiv \boldsymbol{\Sigma} \tag{2.1-13}$$

Additionally, to test this hypothesis, he uses the statistic F_{ij} associated with case j in group i, defined by

$$F_{ij} = \frac{(n-g-p)n_i V_{ij}}{p[(n_i-1)(n-g) - n_i V_{ij}]}, \quad \text{where } V_{ij} = (\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i)^T S^{-1} (\boldsymbol{X}_{ij} - \bar{\boldsymbol{X}}_i), \tag{2.1-14}$$

with $\bar{\boldsymbol{X}}_i$ and S_i respectively denoting group i sample mean and the overall pooled sample covariance. The statistic F_{ij} follows a Snedecor's F distribution with p and $(n-g-p)$ degrees of freedom. Then, Hawkins [106] presents the statistic $A_{ij} = \mathbb{P}(\mathcal{F} > F_{ij})$, which is the probability that an $\mathcal{F}$-distributed random variable with degrees of freedom p and $(n-g-p)$ exceeds F_{ij} and if the model of homoscedastic normal distribution holds, then A_{ij} is distributed as a uniform random variate over the range $(0, 1)$. For this reason, he proposed testing A_{ij} for uniformity as a test of homoscedasticity. If this statistic is not uniform on $(0, 1)$, then the null hypothesis (Equation 2.1-13) is rejected.

Originally, the Hawkins' test requires a complete data sample to be applied and Jamshidian and Jalal [123] improved it to use it with incomplete data-set, by imputing the missing values before applying the Hawkins' test.

For this test, they assume that

$$\boldsymbol{Y}_{ij} = \begin{pmatrix} \boldsymbol{Y}_{ij}^{obs} \\ \boldsymbol{Y}_{ij}^{miss} \end{pmatrix} \sim \mathcal{N}_p(\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \equiv \mathcal{N}_p \left[\begin{pmatrix} \boldsymbol{\mu}_i^{obs} \\ \boldsymbol{\mu}_i^{miss} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_i^{obs,obs} & \boldsymbol{\Sigma}_i^{obs,miss} \\ \boldsymbol{\Sigma}_i^{miss,obs} & \boldsymbol{\Sigma}_i^{miss,miss} \end{pmatrix} \right], \tag{2.1-15}$$

with $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ being partitioned according to their missing and observed values. Then the conditional distribution of $\boldsymbol{Y}_{ij}^{miss}$ given $\boldsymbol{Y}_{ij}^{obs}$,$\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ is

$$\begin{aligned} \boldsymbol{Y}_{ij}^{miss} | \boldsymbol{Y}_{ij}^{obs}, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \sim \mathcal{N}_{p-pi}(&\boldsymbol{\mu}_i^{miss} + \boldsymbol{\Sigma}_i^{miss,obs} \boldsymbol{\Sigma}_i^{obs,obs\,-1} (\boldsymbol{Y}_{ij}^{obs} - \boldsymbol{\mu}_i^{obs}), \\ &\boldsymbol{\Sigma}_i^{miss,miss} - \boldsymbol{\Sigma}_i^{miss,obs} \boldsymbol{\Sigma}_i^{obs,obs\,-1} \boldsymbol{\Sigma}_i^{obs,miss}) \end{aligned} \tag{2.1-16}$$

They propose to impute missing values using random draws from this conditional distribution, but $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are not known, so to test the null hypothesis (Equation 2.1-13), they assume $\boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_g = \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$ and estimate the common mean $\boldsymbol{\mu}$ and the common $\boldsymbol{\Sigma}$ using the method of maximum likelihood. Those

estimates are used in Equation 2.1-16 to impute the $\boldsymbol{Y}_{ij}^{miss}$ for all i by randomly drawing in this law.

After this imputation is made, the statistics A_{ij} can be computed and tested for uniformity, as explained previously. Then, the classical Hawkins' test can be computed to test whether the data are normal and homoscedastic.

Jamshidian and Jalal [123] combine this improved Hawkins' test with a non-parametric test that assumes that data come from a density of the form $f(\boldsymbol{Y}_{ij}; \boldsymbol{\Sigma}, \boldsymbol{\theta})$ and tests equality of the covariances $\boldsymbol{\Sigma}_i$. They show that if the data have a density of the form $f(\boldsymbol{Y}_{ij}; \boldsymbol{\Sigma}, \boldsymbol{\theta})$ and the n_is are equal or large, then under the null hypothesis (Equation 2.1-13), the distribution of F_{ij} (from the Hawkins' test) for all g groups must be the same, hence the interest in testing the homoscedasticity. To compute F_{ij}, an imputed data-set is required with an imputation method that makes no assumptions about the distribution and a k-sample test must be employed to test equality of distribution of F_{ij} in g groups.

Then, as for the Hawkins' test, data need to be imputed using the Srivastava and Dolatabadi's (2009) [191] approach. This imputation method only assumes the independence of observations from case to case and the continuity of their cumulative distribution function.

Then, they compute F_{ij} as in the Hawkins' test, before applying the Anderson-Darling k-sample test introduced by Scholz and Stephens in 1987 [184]. This test uses a rank statistic of the form $T = \frac{1}{N}\sum_{i=1}^{g} T_i$ with

$$T_i = \frac{1}{n_i}\sum_{j=1}^{N-1}\frac{(NM_{ij} - jn_i)^2}{j(N-j)} \tag{2.1-17}$$

where $N = \sum_{i=1}^{g} n_i$ is the size of the pooled sample of F_{ij}s and M_{ij} is the number of the observations in the i-th sample that are not greater than the j-th order statistic in the pooled sample of F_{ij}s. If this test is rejected, the covariances are non-homogeneous and data are not MCAR.

They developed a R-package called *MissMech* that computes all those tests (the improved Hawkins' test and the k-sample test).

Then, Jamshidian and Jalal [123] applied their tests to simulated data to see their power. First, they perform power studies of the improved Hawkins' test applying it to a simulated sample draw from normal and correlated normal distributions. They show the results in Table 2.1-2, which is divided into four parts: the rows labeled N/MAR and Corr-N/MAR show the power of a sample drawn respectively from a normal distribution $\mathcal{N}_p(0, I)$ and a correlated normal distribution $\mathcal{N}_p(0, \Sigma)$ with missing data of the MAR type; the rows N/Corr-N (denote the case where they generate n cases according to $\mathcal{N}_p(0, I)$ then add MCAR missing data before replacing the missing data from the group with the largest number of cases with data generated from $\mathcal{N}_p(0, \Sigma)$ and Corr-N/N (the

same as before, but by drawing a sample from $\mathcal{N}_p(0, \Sigma)$ and replacing the missing data by drawing from $\mathcal{N}_p(0, I)$ shows the power of the test under the hypothesis of non-homogeneity of the covariates. The missingness proportion is denoted by q.

First, they note that the larger the sample size, the more Hawkins' test rejects his null hypothesis. Moreover, the authors remove the patterns with $n_i \leq 6$, which sometimes leads to insufficient data remaining to calculate the test statistic (reported in the table by the NED mention - "not enough data"). Overall, they find that the results for the power of the Hawkins' test are reasonable for $n = 500$ and respectable for $n = 1000$.

Tab. 2.1-2: Rejection rates of the Hawkins' test under MAR alternative and non-homogeneity of covariances alternative (Source: Jamshidian and Jalal, 2010 [123])

		n=200			**n=500**			**n=1000**		
Dist	**q**	**p=4**	**p=7**	**p=10**	**p=4**	**p=7**	**p=10**	**p=4**	**p=7**	**p=10**
N/MAR	0.1	18.0	8.0	5.7	50.2	19.6	11.4	90.5	58.5	27.0
	0.2	16.7	8.5	NED	45.5	23.6	NED	86.9	50.7	27.5
	0.3	19.7	NED	NED	42.9	29.1	NED	82.8	53.6	NED
Corr-N/MAR	0.1	17.2	9.6	5.6	49.5	20.3	10.6	89.6	58.5	26.6
	0.2	18.7	9.0	NED	43.5	23.4	NED	83.9	46.1	27.3
	0.3	18.6	NED	NED	39.5	27.5	NED	76.7	49.5	NED
N/Corr-N	0.1	31.7	39.2	45.2	66.2	85.0	84.4	83.9	97.3	99.8
	0.2	23	22.6	NED	42.8	50.0	NED	80.9	84.1	84.3
	0.3	17.9	NED	NED	40.5	35.1	NED	79	70.6	NED
Corr-N/N	0.1	12.4	11.9	13.5	28.1	25.8	24.8	38.3	37.3	40.3
	0.2	12.5	8.3	NED	19.8	14.2	NED	42.4	23.3	25.2
	0.3	9.3	NED	NED	18.5	13.5	NED	41.9	20.1	NED

NED: not enough data

For a sample of 200 (n) observations and 4 (p) columns, with data drawn from $\mathcal{N}_p(0, I)$ containing MAR data (N/MAR), and 10% of missing data ($q = 0.1$), the rejection rate of the Hawkin's test is equal to 18%.

In Table 2.1-3, they present the rejection rate for the non-parametric test when the data are generated by standard multivariate normal $\mathcal{N}_p(0, I)$ (denoted by N); a correlated multivariate normal $\mathcal{N}_p(0, \Sigma)$ (denoted by Corr-N); a multivariate Student distribution with mean 0, covariance I and 4 degrees of freedom (denoted by t); a multivariate Student distribution with mean 0, covariance Σ and 4 degrees of freedom (denoted by Corr-t); a multivariate Uniform distribution with independent $(0, 1)$ random variates (denoted by U); a multivariate Uniform distribution with independent $(0, 1)$ random variates and multiplying by $\Sigma^{1/2}$ to have a covariance Σ (denoted by Corr-U); a random variate, $W = N + 0.1N^3$ where N is the standard multivariate normal (denoted by W); and finally a multivariate Weibull distribution with scale 1 and shape parameter 2 (denoted by Weibull).

They find that, as with the Hawkins' test, the larger the sample size, the greater

the power of the non-parametric test. Overall, the results are quite satisfactory for sample sizes $n = 500$ and 1000. Concerning the distributions, they note that the power is particularly good for the distributions with thick tails (Uniform and Correlated Uniform); on the other hand, for the W and Weibull distributions, the power of the test deteriorates considerably. The results obtained for the sample from a normal multivariate distribution are relatively similar to those obtained for a correlated normal multivariate. However, in Table 2.1-4, they note that data from a distribution denoted N/Corr-N (detailed above) score higher on the non-parametric test than data from a Corr-N/N distribution. In this Table 2.1-4, they remark again that the power of the test increases as the sample size increases and is quite good for the light-tailed uniform distribution.

Tab. 2.1-3: Rejection rates of the non-parametric test under MAR alternative (Source: Jamshidian and Jalal, 2010 [123])

		n=200			**n=500**			**n=1000**		
Dist	**q**	**p=4**	**p=7**	**p=10**	**p=4**	**p=7**	**p=10**	**p=4**	**p=7**	**p=10**
N	0.1	39.8	19.1	12.5	78.2	40.8	24.4	98.8	83.3	54.5
	0.2	28.4	12.6	NED	56.2	33.4	NED	88.2	63.5	35
	0.3	18.2	NED	NED	20.2	29.7	NED	29.6	42.5	NED
Corr-N	0.1	39.2	19.6	12.4	78.2	41.6	24.4	99	83.7	53.5
	0.2	28.4	12.6	NED	56.2	33.4	NED	88.2	63.5	35
	0.3	17.7	NED	NED	20.4	27.3	NED	27.8	36.4	NED
t	0.1	30.5	51.2	49	64.8	91.2	91.3	86.8	100	100
	0.2	29.9	31.4	NED	46.1	83.7	NED	59.7	99.1	97.8
	0.3	30	NED	NED	36.2	84.2	NED	43.1	98.6	NED
Corr-t	0.1	31.1	48.9	47.8	62.5	91.5	91.2	87	100	100
	0.2	30.4	31.2	NED	46.6	83.1	NED	62.5	99.4	98
	0.3	29.7	NED	NED	36.2	82.7	NED	45.1	97.8	NED
U	0.1	88.9	48.5	32.2	100	95.9	62.2	100	100	98.7
	0.2	91.7	37.3	NED	99.9	94.6	NED	100	100	90.3
	0.3	38.8	NED	NED	90.2	59.9	NED	100	98.7	NED
Corr-U	0.1	89.5	48.5	34.3	100	95.9	63.2	100	100	99.2
	0.2	90.4	36.5	NED	99.9	94.3	NED	100	100	89.7
	0.3	38.9	NED	ED 8	7.6	56.9	NED	99.9	97.1	NED
W	0.1	14.4	12.2	11.7	20.8	12.8	13.4	28.4	25.2	19.5
	0.2	20.5	15.9	NED	21.5	27.4	NED	27.6	35.4	34.7
	0.3	26.5	NED	NED	25	48.5	NED	26.5	61.7	NED
Weibull	0.1	23.9	17.6	13.1	51.9	29.3	17.8	81.8	61	38.2
	0.2	14.4	11	NED	16.5	20.1	NED	20.4	26	23.6
	0.3	24.1	NED	NED	41.2	30.9	NED	71.3	47	NED

NED: not enough data

For a sample of 200 (n) observations and 4 (p) columns, with data drawn from $\mathcal{N}_p(0, I)$ containing MAR data, and 10% of missing data ($q = 0.1$), the rejection rate of the non-parametric test is equal to 39.8%.

Tab. 2.1-4: Rejection rates of the non-parametric test under non-homogeneity of covariances alternative (Source: Jamshidian and Jalal, 2010 [123])

		n=200			n=500			n=1000		
Dist	**q**	**p=4**	**p=7**	**p=10**	**p=4**	**p=7**	**p=10**	**p=4**	**p=7**	**p=10**
N/Corr-N	0.1	42	50.1	55.1	79.9	92	91.3	92.4	99.1	100
	0.2	34.5	30.3	NED	56.5	59.8	NED	90.7	89.9	91.4
	0.3	27	NED	NED	51.4	44.6	NED	87.5	79.3	NED
Corr-N/N	0.1	24.4	26	26.6	54.1	50	43.1	67.1	63.8	65.7
	0.2	25.5	14.1	NED	37.3	25.6	NED	71.8	43.7	45
	0.3	16.6	NED	NED	32	22.3	NED	64.9	36.9	NED
t/Corr-t	0.1	31	28.8	29.2	67.4	56.9	45	79.7	83.6	79.6
	0.2	24	19.8	NED	37.9	31.9	NED	70.5	58.4	59.3
	0.3	20	NED	NED	34.5	29.8	NED	58.7	59.2	NED
Corr-t/t	0.1	12.1	13.1	13.6	16.5	17.4	18.2	29.1	19.2	25.8
	0.2	13.2	13	NED	22.9	18.3	NED	42.4	29.4	34.4
	0.3	15	NED	NED	28.1	26	NED	44	43.7	NED
U/Corr-U	0.1	80.2	81.3	81.2	99.5	100	99.9	100	100	100
	0.2	65.2	47.2	NED	93	90.6	NED	99.8	100	99.8
	0.3	46.7	NED	NED	87.8	64.9	NED	99.8	97.8	NED
Corr-U/U	0.1	57.8	46.5	46.5	94.1	88.4	81.1	98.7	97.7	98.2
	0.2	52.6	22.2	NED	77	52.7	NED	99.2	82.3	75
	0.3	31	NED	NED	66.4	34.9	NED	97	65.5	NED

NED: not enough data
For a sample of 200 (n) observations and 4 (p) columns, with data drawn from a correlated multivariate normal distribution $\mathcal{N}_p(0, \Sigma)$, and 10% of missing data ($q = 0.1$), the rejection rate of the non-parametric test is equal to 18%.

After seeing, through Little's test and the Jamshidian and Jalal's test, how it was possible to determine whether missing data are missing completely randomly or not, it may be interesting to apply them to historical data. In this PhD thesis, these tests are applied to simulated data but especially on empirical data, offering some additional information on the nature of the missing data.

2.1.3 Distribution of missing data

Little and Rubin [145] also found it useful to distinguish between missingness patterns. Let $Y = (y_{ij})$ denote an $(n \times p)$ rectangular dataset without missing values, that is, a complete dataset, with i-th row $y_i = (y_{i1}, \ldots, y_{i,p})$, where y_{ij} is the value of variable Y_j for unit i. With missing data, define the missingness indicator matrix $M = (m_{ij})$, such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is observed. The matrix M then defines the pattern of missing data.

They also consider it useful to allow different missing data codes that indicate different types of missing data; for example $m_{ij} = 1$ for unit non-response due to non-contact, $m_{ij} = 2$ for unit non-response due to refusal to participate and $m_{ij} = 3$ for

refusal to answer a particular item. The goal here is to use the information contained in these codes in the statistical analysis. Little and Rubin [145] also present different missingness patterns in Figure 2.1-3:

(a) Univariate non-response: where the missingness is confined into a single variable. For example, Y_p is incomplete and $Y_1, \ldots, Y_{p-1}$ are fully observed (here $p = 5$).

(b) Multivariate with two patterns: the same case as previously, but Y_p is replaced by a set of single incomplete variables $Y_{J+1}, \ldots, Y_p$ that are all observed or missing on the same set of units (here $p = 5$ and $J = 2$). An example here is the unit non-response in sample surveys, where a questionnaire is administered and a subset of sampled individuals does not complete the questionnaire because of non-contact, refusal, or some other reason.

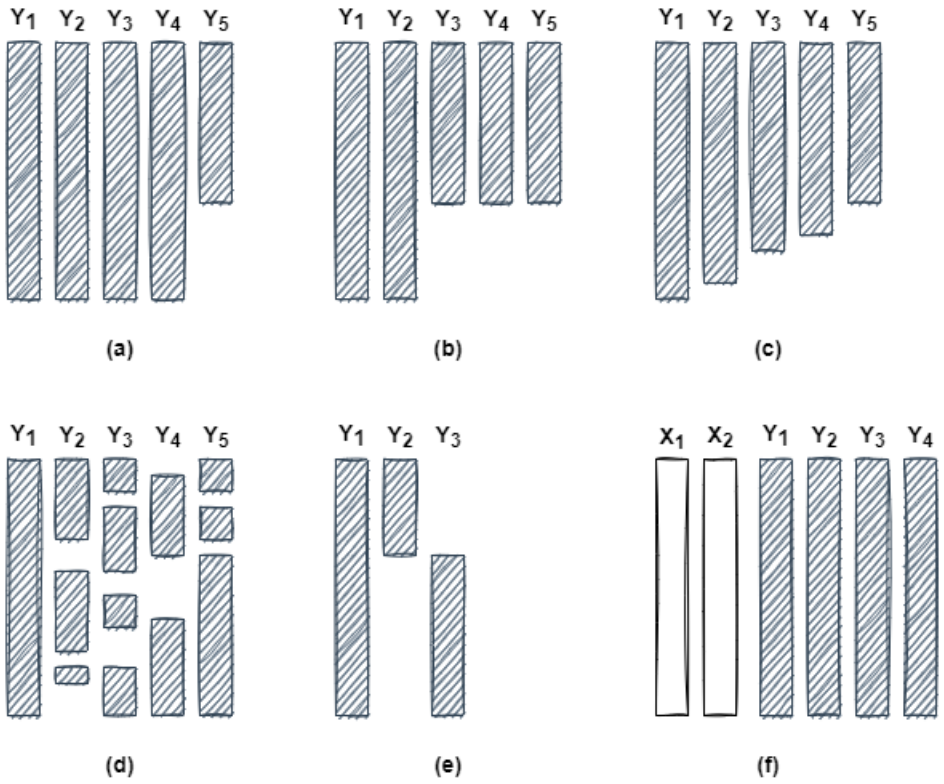
(c) Monotone: for example, attrition in longitudinal studies (collecting information on a set of units repeatedly overtime),where units drop out prior to the end of the study and do not return - in other words, where the variables can be arranged so that all $Y_{j+1}, \ldots, Y_p$ are missing for all units where Y_j is missing, for all $j = 1, \ldots, p-1$ (here $p = 5$).

(d) General: where missing values typically have a haphazard pattern.

(e) File matching: the case of the file-matching problem with two sets of variables never jointly observed - for example, in this pattern where Y_1 represents a set of variables common to both data sources and fully observed; Y_2, a set of variables observed from the first data source but not the second; and Y_3, a set of variables observed from the second data source but not the first. There is no information in this data pattern about the partial associations of Y_2 and Y_3 given Y_1.

(f) Factor analysis: the case of patterns with latent variables that are never observed. Consider $X = (X_1, X_2)$ represents two latent variables that are completely missing and $Y = (Y_1, Y_2, Y_3, Y_4)$ is a set of variables that are fully observed. Factor analysis can be viewed as an analysis of the multivariate regression of Y on X for this pattern. As none of the regressor variables are observed, some assumptions are needed

Fig. 2.1-3: Examples of missingness patterns (rows correspond to units and columns to variables) (Source: Little and Rubin, 2019 [145])



These examples show that the distribution of missing data can take several forms, three of which are often cited in the literature: univariate and multivariate missing data (Figure 2.1-3 [a] and [b]), monotonous (Figure 2.1-3 [c]) and non-monotonous (Figure 2.1-3 [d]).

The distributions presented above are also valid for longitudinal data where, for example, a monotonic distribution corresponds to a censoring to the right of the data.

Only after this complete theoretical framework of missing data has been established can the question of filling in the missing data arise. Some methods are better suited to certain categories of missing data, or to certain distributions. Before presenting a selection of completion methods, however, it is necessary to go back to a financial framework and point out the specificities of the missing data of the financial series.

2.2 Specificities of missing data in finance

Missing-data theory was designed to be generalist, so it is legitimate to ask what characterizes missing data in financial series. Thus, in this section we will present what types of missing data are present in financial series, the reasons for their presence and how are they usually managed.

2.2.1 Where do the missing data in finance come from?

A financial series is considered complete when it includes all observations outside weekends (Saturday and Sunday) and New Year's Day. In this framework, any unavailable information is considered missing. Therefore, it is entirely possible to retrieve the three types of missing data described by Little and Rubin [145] in a financial data framework.

First of all, there are several explanations for the MNAR phenomenon on financial series. This type of missing data is also known as "non-ignorable missingness" because this mechanism is directly due to the missing data itself. For example, missing data may tend to appear when a stock price exceeds a certain threshold. In this case, the data are not randomly missing, because they depend directly on the stock price.

In the case of missing data of type MAR, the missing data are related to observations of other variables. Contrary to the MNAR data where the missing data depends only on the variable itself, MAR data can depend on the observe value of the variable but also on other related variables. For example, missing stock price data may occur randomly at first glance, but further investigation may reveal a link to other stock prices. More concretely, missing stock price data can occur when the price of one or more other stocks in the portfolio exceeds a certain threshold.

Finally, in the case of MCAR missing data, they are not linked to their values or even to the other variables. If the data are categorized as MCAR, then they are such due to a purely random process. This is the case, for example, with stock prices, where missing data appear randomly throughout the period without any explanation of their presence.

It is indeed possible to find all types of missing data categorized by Little and Rubin [145], in the financial series. The question now regards the reasons for the appearance of missing data in finance. Data can be missing from financial series for a number of reasons:

- The first reason for missing data in a financial data context is the **market calendar**. The financial markets are not, indeed, open every day. It is normal to find missing data on weekends, but also on days during market holidays (January 1^{st}, Christmas, Easter, etc.). Generally, if no observations are available on a date, there is a good chance that it coincides with a market closing date.

- Some missing data may, of course, be due to **calendar differences** from one country to another. Indeed, market holidays are not always the same from one country to another country and it may be the cause of missing data in a financial series. For example, December 25 is closed for many stock exchanges, but not in Japan. It will be possible to quote on December 25 for Japanese exchanges but not for U.S. exchanges.

- It is possible that a simple **computer problem** could be the cause of a missing data. An error in an automatic batch, a server problem, a broken internet connection, or the like may be the cause of random missing data in the series.

- Missing data may also come from **IT migration**, as a bank or insurance company may need to review the architecture of its databases in order to improve data exploitation, or for regulatory purposes (e.g., the case of the BCBS 239 standard; see Section 1.3.1). These computer migrations may result in an involuntary loss of data creating successive missing data in a series.

- It may also happen that data, especially outliers, are **deleted voluntarily** for data-quality reasons. Some data-quality controls set up by the banks warn of the presence of outliers, which are (automated or not) deleted from the databases in order not to bias the analyses made afterwards. Thus, outliers can be removed, resulting in missing data.

- Some banks or insurance companies have taken some time to set up a **data historization process**, especially for their own data (e.g., volatility, smile, or calibration parameters). Before the advent of numerous regulations, data backup was not a primary preoccupation for banks and insurance companies. It is notably with the FRTB [12], which has required good quality data histories since 2007, that many data storage processes have been put in place. This can be the reason of many successive missing data at the beginning of the time series.

- Some missing data are due to **non-computer data entry**, namely manually recorded daily in the databases, as no automatic saving process is possible. If the person in charge of saving the data forgets to save the data, this oversight is sufficient to create a missing data in the series. Since the data is forgotten by the trader, this leads to random missing data in the series, but if he has forgotten for several days, this can lead to successive missing data.

- Sometimes data is simply missing due to **inaccessibility and unavailability**. As presented earlier, in Section 1.1.2, some financial data are not easily accessible, such as the Totem Markit consensus data that are available to contributors but may no longer be available once the contribution in exchange stops or becomes of poor quality. This could be the reason of successive missing data.

- Usually, successively missing data at the very beginning of a series are simply due to a date of **introduction on the markets** during the observation period studied. To make a study that goes back several years, it is possible to find the case of actions introduced on the markets during the desired observation period. The whole beginning of the associated series will be composed of successive missing data.

- Data may also be missing due to a **trading halt**. In fact, the market operator may decide to suspend the price of a financial asset, but in practice a price suspension often follows an important event in the life of the company (e.g., restructuring operation, leveraged buy-out, tender offer, financial press announcement, or planned financial operation). The reasons for these suspensions are often responsible for large price variations, so these missing data may actually correspond to extreme values in the series.

- Some **illiquid financial assets** due to a lack of investors, which cannot be bought or sold easily. This lack of trading on the markets generally results in non-daily quotes. Potentially, for some specific assets, the associated series may not be observed on a daily basis. Although, some data in the series may then appear to be randomly missing, in reality they are simply not observed.

- Furthermore, the missing data may be the result of **no price agreement between buyer and seller**. In some cases, sellers and buyers are unable to agree on an exchange price and therefore they create no market results, despite the existence of supply and demand. Moreover, this kind of event can lead to random missing data in the series.

This list of arguments that can explain the presence of missing data in financial series, whether obvious or not, shows that they can be explained by human or computer error, a regulatory motivation manifesting itself late, or market particularity. In addition, regulatory motivations, among others, have caused banks and insurance companies to investigate, implement and use missing-data completion algorithms.

In addition, it is important to identify the nature of the missing data, as it can also provide information to the expert on how to complete it. It is obvious that data missing due to a calendar differences will not be imputed in the same way as data missing due to an illiquid market. In the first case, a price adjustment may be sufficient whereas in the second case there is no data and many variables can be considered for the estimation. It is therefore important for the expert to identify the nature of the missing data.

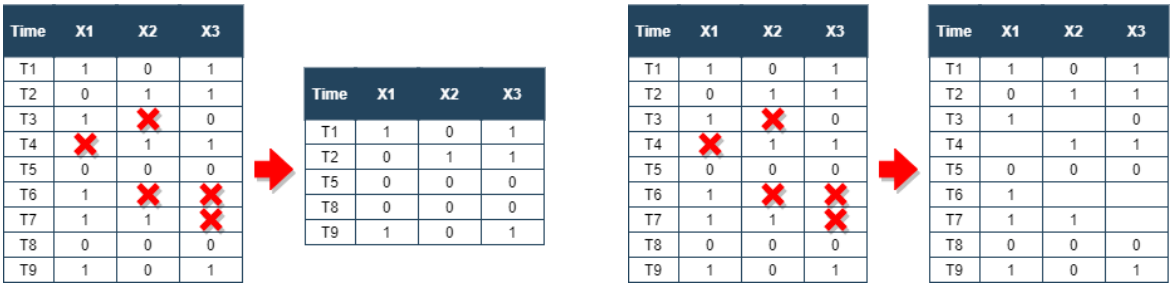
2.2.2 Traditional management of missing data

Listwise and pairwise deletion

Even before talking about filling in the missing data, other methods can be applied to get around the missing data. Indeed, some deletion data methods are often applied for statistical study purposes.

There are two main deleting methods: listwise deletion and pairwise deletion. The listwise deletion method is the process of deleting an observation when one or more variables in this observations are unobserved (Figure 2.2-1a). Furthermore, the pairwise deletion method removes only the specific missing values from the analysis and not the entire case (generally used for correlation calculations). The best way to understand pairwise deletion is in the case of a correlation matrix. It will be computed according to each pair of variables for which data is available. In other words, all pairs of available data are included (Figure 2.2-1b).

Fig. 2.2-1: Listwise and pairwise deletion



(a) Listwise deletion

(b) Pairwise deletion

The regulatory environment pushes banks, asset management and insurance companies to have good quality and complete data histories. Rather than trying to complete the data at all costs, however, in some cases it may be more cost-effective to use a proxy or to simply use the observed data.

In 1977, Kim and Curry [130] show in their study that replacing missing data does not lead to better results. First, they show that pairwise deletion is more efficient (in terms of mean deviation from the model) than in listwise deletion, based on variance-covariance and correlation matrices of sociological data, where they add 10% of artificial missing values on each variable.

Then, they compare the results obtained thanks to the five following samples:

- Mean imputation: missing values of the variable completed by the mean of the observed values of the same variable;
- Listwise deletion: deletion of all rows containing missing value;

- Pairwise deletion: each pair of variables for which data is available are taken into account;
- Regression estimation of random component instead of missing values; and
- Complete sample with no missing data.

The findings of their study show that pairwise deletion offers the best results of regression imputation. This means that replacing missing data does not necessarily offer a better result than does simply considering only the available data, using pairwise deletion. In addition, the estimated parameters obtained from the pairwise deletion data give results very close to those of the original data.

The article by Kim and Curry [130] is not very recent; however, it is one of the rare articles to question the utility of the data completion. In some cases, it will be preferable to use as much of the data as possible, without necessarily trying to complete the missing data. Pairwise deletion is still widely used in the literature, suggesting that data completion is not always the best option.

On the other hand, when pairwise deletion is applied, the resulting covariance matrix is not necessarily positive semidefinite. Thus, they are not Gram matrices (or Gramian matrix; see Jain and Gupta [122] in 1970). A Gram matrix is a square and symmetrical matrix in which each term corresponds to a scalar product. A Gram matrix is always at least positive semidefinite. However, in the case of asset management, for example, this matrix cannot be used because it is not necessarily positive semidefinite; in other words, it could contain negative eigenvalues. Thus, one should be able to transform these matrices into Gram matrices. In their case, Kim and Curry [130] may not need this property but in a financial framework, the application of pairwise deletion alone may not be sufficient. Thus, these covariance matrices can be regularized so that they become Gram matrices and are, thus, positive semidefinite.

A methodology to transform a non-positive semidefinite matrix into a positive semidefinite matrix was proposed by Rousseeuw and Molenberghs [173] in 1993. This method depends on the use of the covariance matrix at the end. Their covariance matrix method consist of carrying out a transformation on their own values. They explain that a positive semidefinite matrix $\mathbf{R}$ can be decomposed as follows:

$$\mathbf{R} = \mathbf{PDP}^T, \tag{2.2-1}$$

where $\mathbf{D}$ is a diagonal matrix containing the (non-negative) eigenvalues of $\mathbf{R}$ and where $\mathbf{P}$ is the matrix of corresponding eigenvectors.

This decomposition still holds when $\mathbf{R}$ is a symmetric but not positive semidefinite matrix. In this case, $\mathbf{D}$ contains some negative eigenvalues. Their approach consist of computing $\tilde{\mathbf{D}}$ by replacing the negative eigenvalues of $\mathbf{D}$ with zeroes and computing

$\tilde{\mathbf{R}} = \mathbf{P}\tilde{\mathbf{D}}\mathbf{P}^T$. $\tilde{\mathbf{R}}$ is the nearest symmetric positive semidefinite matrix in the Frobenius norm; in other words, $\tilde{\mathbf{R}}$ is the matrix that minimizes $\|\tilde{\mathbf{R}} - \mathbf{R}\|_F$ with

$$\|X\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p}|x_{ij}|^2} = \sqrt{tr(\mathbf{X}\mathbf{X}^T)}, \tag{2.2-2}$$

where $tr()$ is the notation of the trace of a square matrix. This transformation proposed by Rousseeuw and Molenberghs [173] makes it possible to have the closest positive semidefinite covariance matrix and avoid getting stuck with negative eigenvalues. Additionally, this method will be reused in the continuation of this PhD thesis in order to compare the performances of the chosen imputation methods to this one when the objective is to impute covariance matrices.

However, there are other methods of making a positive semidefinite matrix. Higham [110], in 2002, introduced a methodology, similar to the previous one, that allows one to find the nearest positive semidefinite matrix to a pairwise matrix. Apart from the fact that these matrices are not positive semidefinite, according to him, pairwise correlation matrices are approximate, due to the use of only the available peers. His methodology allows one to find the nearest correlation matrix in the Frobenius norm. His method is described for correlation matrices but it can also be applied to covariance matrices.

Another way to repair invalid covariance and correlation matrices is through shrinkage. One of the most famous articles about the shrinkage method is that of Ledoit and Wolf [139], published in 2003. The shrinking method is based on a convex combination of the empirical estimator S (the covariance matrix) with some suitable chosen target F (a highly structured estimator), such as $\delta F + (1-\delta)S$, with $\delta \in [0, 1]$.

Thus, in the case of a covariance matrix, sophisticated imputation methods can be applied to data before computing the covariance matrix, or it is possible to compute a Pairwise covariance matrix and then regularize it by the Rousseeuw and Molenbergh's method [173] (or other). Thus, it will be interesting to compare the results of these two techniques to see which is most efficient when the interest is to use the covariance matrix.

Multiplying completion methods on data management

In a context of regulatory calculations, deletion methods cannot be used because all the data is required. Data-completion methods are therefore at the heart of regulatory concerns. Above all, a distinction must be made between imputation methods and interpolation methods. According to The Oxford Dictionary of Statistical Terms [70], an imputation method is "the process of replacing missing data in a large-scale survey" whereas an interpolation method is defined "as the use of a formula to estimate an intermediate data value". In other words, imputation is the act of replacing missing data with an estimated or observed value, while interpolation is the act of creating new data points between other points.

Traditionally, to manage missing data in a daily process (but also for historical process), banks and insurance companies use imputation methods, including the method of retrieving the last available observation (see Section 2.4.1). Very often in the daily backup processes of data, if there is no data to be recovered, the method will search for the data of the previous day in order not to block all the calculations that are requested afterwards. The same method can also be applied for the purpose of revising historical data. Often, though, the linear interpolation method is applied, as the starting and end points are known. The use of more sophisticated algorithms is clearly not the primary solution for completing missing data. In particular, they are used for specific projects where good quality data is required, such as, recently, to meet regulatory requirements. The current trend is to have data of the best possible quality. Hence, for specific data, where they are highly exposed to risks, banks and insurance companies no longer hesitate to invest in research and computing capacity in order to provide both historical and daily data of the highest possible quality; hence the interest in looking at the different completion methods in place and their impacts in terms of risk measures.

The regulatory environment pushes banks, asset management companies and insurance companies to have high quality and complete data histories.

2.3 Data encoding

Data encoding is a fundamental step in any algorithmic procedure. It is the way in which the data are formatted, organized and, reprocessed before being used in a statistical model. This is a preliminary step that should not be neglected because the choices that will be made regarding the encoding of data can directly impact the performance of an algorithm. Thus, this step is as important as the choice of the algorithm used. The importance of data encoding was pointed by Cerda, Varoquaux and Kégl [56] in 2018, while they were working on the encoding of categorical data, seeking to clean up data as well as possible in order to give algorithms the clearest possible information.

Thus, this section raises the main issues related to the application of a database on a given algorithm and how the database should be built in order to optimize the results of that algorithm. Before applying an algorithm, many choices have to be made in terms of data encoding in order to optimize the algorithm's application, including the nature of the data, its historical length, the other variables that make it up and the format of the data.

2.3.1 Model selection according to data type

First of all, the choice of the method or algorithm used will depend directly on the type of data. First of all and in an obvious way, the efficiency of the algorithms may vary depending on the type of data applied to them. An algorithm may be particularly

effective for a certain type of data, but not necessarily for another. For this reason, the results should be put into perspective. The conclusions obtained by applying the models to spot prices will probably not be the same as if it had been used on yield curves, volatility surfaces, correlation, or variance-covariance matrices, smiles, model parameters and so forth.

Among the completion methods that will be presented in the rest of this Chapter 2, some may be very effective for one type of data and others less effective, but the same ranking is absolutely not guaranteed if the experiment is repeated for another type of data. Some algorithms make the assumption of data normality (for example *Amelia* presented in Section 2.4.6), which could be advantageous for them if the data used are price returns. The prices from financial time series are log-normally distributed, so assuming that the data are Gaussian would make sense. On the other hand, in dealing with volatility surfaces, covariance matrices, or interest rate curves, using a model based on normality no longer makes sense.

Finally, the very purpose of completion may cast doubt upon the choice of method. For example, if the objective is to reconstruct a missing maturity (that has never been observed) in a term structure of interest rates, then an interpolation method should be preferred instead of an imputation method. Of course, not all methods are not usable and they must be well chosen according to the nature of the data. Applying a Brownian Bridge (presented in Section 2.4.2) may not be the most appropriate method to preserve the shape of a rate curve. On the other hand, if some data are observed for this maturity, it is possible to isolate this maturity in order to consider it a time series containing missing data. In this case, a possible completion method would consist in first using an imputation method on this time series, in order to be consistent with the observed values from this maturity and then to check whether the rate curve is respected, using rate models such as the Nelson, Siegel and Svensson model [161] [196].

In addition, the financial data generally manipulated is continuous data, yet each type of then has its own particularities and constraints. It is therefore necessary to choose a completion method that can impute continuous data, but also by respecting their constraints. Data encoding may be the solution to this last problem. For example, with few exceptions and special cases, spot price series, volatility surfaces and rate curves (before the subprimes crisis) are data series that are always positive. Thus, the models through which they pass must respect this condition or be transformed in such a way as to preserve this specificity. One of the methods used by Marshall, Altman, Royston and Holder [150] in their paper (see Section 2.4.7) was to apply an imputation method to transformed medical data. In their case, the value 1 was added to the data used before the application of a logarithmic function and before application of the imputation method. Then an exponential function is used on the imputed dataset to return to the levels of the original series to ensure that these were only positive values. The same procedure could be applied to a financial framework, in order to

ensure positivity (with some exceptions) of spot prices, areas of volatilities, or interest rates.

However, in some cases financial data may be categorical, meaning that the imputation method used must be able to handle the categorical data. Applying a Brownian bridge or EM algorithm to categorical data makes no sense. On the other hand, the methods of K-NN, random forests and others can be used on categorical data. This PhD thesis will focus on continuous data imputations, as they are mostly present in a financial data framework.

The choice of model depends directly on the type of data to be processed. However, in the framework of this PhD thesis, the type of data used will mainly concern spot price series, so the completion methods selected will be methods that could be relevant when applied to spot price series. Nevertheless, it would be relevant, in future works, to carry out similar studies concerning volatility surfaces, interest rate curves and covariance matrices to compare the results.

2.3.2 Historical length

Once the type of data to be completed is known, one can start creating the sample. However, one of the first questions that may arise is the length of the sample to be used. This subsection deals with the longitudinal aspect of the sample.

First of all, determining the correct historical length of a database is fundamental when the method used involves the entire database to fill in missing data. The one-dimensional methods that will be presented in this chapter (i.e., methods that require only one series and not other methods to predict the dynamics of missing data) consider the data only locally, not as a whole. These kind of methods will not be impacted by the length of the chosen history. On the other hand, the length of the variables will indeed have an impact on the two-dimensional completion methods (i.e., methods that use the dynamics of other series to predict the dynamics of missing data). Thus, the methods that aim to seek to estimate the distribution of the sample, those that estimate the proximity between series, or those that proceed to a spectral decomposition, use the full set of data in their procedure and will be impacted by the length of the history studied.

Obviously, using a sample that is too short may lead to results not representative of the series. Any statistical study based on a sample that is too short is subject to disagreement. Similarly, in the case of data completion, a sample that is too small in length will not be credible. Models to estimate the distribution of series will need a certain amount of data to provide realistic results. The same is true for models that attempt to decompose the principal components and those based on proximity criteria.

On the other hand, if the sample is too long, then the algorithm may be drowned under too much information and yield results unrepresentative of the period in which the missing data is located. Imputations could be smoothed in some cases. With a sample that is too long, several sub-periods can be grouped, with their own characteristics. For example, in the case of financial data, it is highly probable that volatility will vary over time. However, if a sample that is too long is used by a completion algorithm that aims to estimate the distribution of the observed data, then it will consider only one volatility and will therefore lose valuable information for completion. Completing data in times of bull markets is not the same as completing missing data in the middle of a financial crisis. For this reason, depending on the models used, one should not always try to give as much information as possible, naively thinking that one cannot have too much, but select one's sample by correctly defining its time period.

In addition, the choice of length may depend on the proportion of missing data present in the history. For example, the history may be long or if the proportion of missing data is too large, the history may not be long enough. To operate more or less properly, the completion method will need a certain level of observed data. On the other hand, if complete observed data are too far back in time, they may have no relation to the period with missing data and may mislead the completion method. Thus, the choice of length can be even more problematic in the case of heteroskedasticity. As said previously the variance may change over time and then some methods could be misled by too much information.

In this case, a possible solution would be to divide the original sample into several sub-samples of smaller length, in order to apply the completion methods to sub-samples with a volatility that is almost constant. Alternatively, it would be possible to use a rolling window procedure in order to apply the completion method to the entire sample, considering this variation in volatility.

Thus, obviously, the length of the history impacts the performance of the imputation methods' results. For this reason, the length of the sample must be well chosen before the application of a completion method, or in general any method of statistical analysis.

2.3.3 Data bucketing

Multi-dimensional methods always work, by definition, from several time series. As mentioned above, these methods use other time series to reproduce the dynamics of missing data. In other words, the data to be completed will depend on several or all of the available sample data. For this reason, the selection of the variables that make up the sample under study should not be neglected. In the case of data completion, choosing a sample composed of data selected at random, which may have nothing in common with each other, means taking the risk of having bad missing data prediction.

This is one of the reasons why, when setting up a sample, it is as important to choose the length of the history as it is to choose the variables that make it up. Many criteria could be used to constitute a sample, based on both quantitative and qualitative criteria. One of the most commonly used classifications on financial data is the sector classification, which is practical to use because it is easily accessible by data providers. For example, Bloomberg offers these users the option of organizing the data into different non-quantitative classifications. The two main ones are the Global Industry Classification Standard (GICS) and Bloomberg Industry Classification Standard (BICS).

According to the Morgan Stanley capital international (MSCI) website [159], the GICS was developed by MSCI and S&P in 1999 to provide an effective investment tool that captures the breadth, depth and evolution of inductive sectors. It is based on four levels of hierarchical industrial classification: 11 sectors, 24 industry groups, 69 industries and 158 sub-industries. The level of hierarchy chosen will depend on the desired granularity. That is, if the study requires a broad classification, the first level should be sufficient (the data will be divided among the 11 sectors); on the other hand, if the study requires one to know in detail the activity of the company, then it will be more appropriate to choose the last level of classification (then the data will then be divided among 158 sub-industries).

The BICS is another hierarchical classification according to the general business activities of companies. It is based on eight hierarchical levels. The first level contains 9 macro-sectors and is the broadest classification of general business activities of companies and the last level has up to 2,294 sectors. With its BICS, Bloomberg offers a much more detailed classification of individual companies, in contrast to the GICS, which can only divide into 158 sub-sectors. Bloomberg thus offers twice as many hierarchical levels as the GICS and a much more detailed classification of individual companies, unlike the GICS, which can only divide into 158 sub-sectors.

On the other hand, data bucketing by sector can be sub-optimal, as it can still bring together industries that differ considerably. Unless a very detailed classification is used, divided into many very specific sub-sectors, this way of constituting a sample might bring together series that differ considerably. The use of a categorical classifier does not seem to be the best option for the construction of a sample when the aim is to carry out quantitative analyses. Sector classifications are often used to constitute samples in the literature, although there are many other methods of constituting a sample.

In the same way as for the sectors and in a more general way, it is possible to form series buckets from other non-quantitative criteria, for example their type (e.g., shares with shares, bonds with bonds, or rates with rates), their rating, their nationality, their currency and so forth. but can be just as sub-optimal (or worse) as using categorical classifications. The variables can also be grouped according to the risk factor criteria set out in the FRTB regulations [12]. This regulation leaves the possibility for the

bank to use their own bucketing approach, but also proposes a standardized bucketing approach, in order to compute the sensitivities of financial instruments (delta, vega and curvature risk capital requirements) based on these buckets. These buckets group together risk factors according to their common characteristics. In this regulatory framework, a bucket corresponds to a set of instruments with the same risk class, sharing the same characteristics and therefore the same risk profile. The standardized bucketing approach is detailed in the chapter MAR21 of the last version of the FRTB [12]. Thus, interest rate buckets correspond to groups organized by currencies and exchange rate buckets are also organized by currencies. Commodity buckets correspond to categories of raw materials (e.g., energy, metals, or gaseous). Buckets associated to credit instruments (unrelated to securitization and securitization within the correlation trading portfolio) are based on a two-level classification, by credit quality (investment grade and high yield) and then by economic sector (e.g., sovereigns, consumer goods and services and health). The classification made for credit buckets concerning non-correlation trading portfolios is close from the previous credit bucket, because buckets also correspond to two levels of classification: the first level is the credit quality (senior investment grade, non-senior investment grade and high yield) and the second level is the seniority sector (senior investment grade, non-senior investment grade and high yield) and asset family (residential mortgage-backed security, commercial mortgage-backed securities, asset-backed securities and collateralized loan obligation). Finally, equity buckets are organized into three levels of classification: market capitalization (large and small), economy (emerging market and advanced) and sector (e.g., consumer goods and services, telecommunications and financials). The bucketing standardized approach from the FRTB regulation [12] is overwhelmingly based on non-quantitative criteria (e.g., currency, economic sector and securitization sector).

Quantitative criteria can also be used in the constitution of a sample. For example, a simple correlation analysis can be performed to determine which series are most likely to have the same dynamics as the series with missing data. In the case of financial series, it is entirely possible to form a bucket of series highly correlated with each other by having, as presented above, a well-defined study period. It is possible to make buckets based on many quantitative criteria: market capitalization, volume, volatility, price-to-book ratio, price earnings ratio, betas (in the case of asset management) and so forth.

It is even possible to apply somewhat more sophisticated methods than simple analytic criteria to form groups. For example, it is possible to call upon classification methods, typically a principal component analysis (PCA; introduced in 1901 by Pearson [164]) is widely used with financial data. The objective of a PCA is to reduce the size of the data used by distorting reality as little as possible (with the least loss of information). A PCA allows to graphically identify groups of individuals within the factorial designs. Methods such as K-means or hierarchical clustering can then be used.

Another method used aims to infer dependence relationships between variables by covariance matrix. Indeed, it is possible to estimate a covariance matrix from a multivariate data sample by maximizing its likelihood, while penalizing the covariance so that its graph is sparse. This problem is notably known as covariance selection initiated by Dempster [67] in 1972. He decides to focus on the zeros of the inverse of the covariance matrix, which correspond to the conditionally independent variables. In particular, Dempster [67] explains that a covariance matrix derived from a multivariate normal data sample can be simplified by imposing elements of the inverse covariance matrix to be zero. This is a way of showing the dependency between variables and thus in the present framework, of constructing a sample of dependent variables. The idea is to set to 0 the elements of the inverse covariance matrix that correspond to the weakest dependencies. Then, it is necessary to find a method that determines how to penalize the model, in other words in what order and on what criteria to set the dependencies to 0. This can be seen as a model selection problem. Traditionally, model selection is based on the use of the Akaike information criterion (by Akaike [4] in 1973) or Bayesian information criterion (by Schwartz [185] in 1978), but it is also possible to use techniques such as that of Lasso (Tibshirani [198] in 1996), which is a penalization technique based on a l_1 norm. More recently, in 2008, d'Aspremont, Banerjee and El Ghaoui [60] proposed a solution using a sequence of eigenvalue decomposition.

Finally, once the question of the choice of variables that make up the sample under study has been answered, it is sometimes necessary to set a number of variables to establish the sample. If, despite the variable section criteria, the sample remains too large, it might be conceivable to use only a certain number of variables that best meet the criteria. This would mean, for example, using the K variables that are most correlated with the series to be completed, so as to use only information that is relevant to the data imputation. There is little chance of forming a sample of several dozen variables, each of which has a positive effect on completion. Finally, in the same way as for the number of observations, using too many different variables could drown the algorithm in too much information without being able to extract the essential information, which could lead smooth the predictions. Conversely a sample consisting of too few variables might not contain enough information to predict missing data.

Thus, the quality of imputation may be directly related to the number of variables used in the sample, but this is not necessarily the case. Some models do not use the entire sample provided, but only the part defined as the closest (on distance criteria for example), such as the K-NN method, which will be presented in Section 2.4.3. The imputed data from K-NN algorithm will not be impacted by a sample containing too many variables, because it does not use all of these variables to impute the missing data but only some of them. This algorithm is a kind of exception, however, because generally, algorithms use the information available in all variables.

Moreover, the optimal choice of the number of variables depends wholly on the

nature of the data used and of course, the imputation algorithm applied, making this problem even more complex. This is why there is no perfect solution in the selection of the number of variables.

More generally, buckets can be formed from many criteria, whether quantitative or not, simple or more complex. It is even possible to combine several of these criteria. In any case, for the construction of a credible sample, it is necessary to use adapted and above all, justified criteria and methods. Finally, the most important thing is undoubtedly justifying and arguing the construction of the sample.

2.3.4 Raw, return, or normalized data?

Once the sample has been constituted, it is possible to apply the completion method to it. However, depending on the method used, data may or may not need to be transformed.

Some methods require raw data as input, such as interpolation methods that aim to fill in missing data using the last observation and the next observation surrounding the missing data. Thus, this kind of method requires knowledge of the data level in order to function properly.

Apart these usual methods, the question of the format of the input data may arise. Moreover, applying certain algorithms to raw data does not always make sense. It can be more or less efficient to apply such and such a completion method on raw data, on returns, or on standardized data. For example, some algorithms involve estimating the distribution of observed data in order to predict missing data. However, estimating the distribution of spot prices does not make sense, since price series are not stationary and the goal is to estimate the distributions of price returns. On the other hand, some other methods do not assume stationarity of the data, which means that they can be applied to spot prices.

Finally, some methods are more efficient when applied to series where autocorrelation is strong. However, it is obvious that the autocorrelation on a series of prices will be higher than on a series of returns (generally associated by a random walk).

Thus, the choice of inputs must be made on a case-by-case basis according to the algorithm used. Sometimes, this choice is made logically, but in some cases the logic may not be so obvious and should be discussed.

Furthermore, the data can be transformed before the application of a specific completion method. Actually, the very calculation of returns is a transformation procedure applied to spot prices that allows one to compare prices on a common basis; in other words, it is a way to standardize the data.

Thus, if the input choice is price returns, then it is necessary to know what type of return to calculate. Different types of returns are commonly used: absolute returns,

which is the simple difference $S_t - S_{t-1}$;, relative returns, which is the ratio $\frac{S_t}{S_{t-1}} - 1$; or log returns, $ln(\frac{S_t}{S_{t-1}})$, that are an approximation of relative returns. As with sample selection, the use of a particular data format must first and foremost be justified in order to be credible. Generally, the most commonly used returns are relative returns or log-returns, but they cannot to handle negative values. Negative prices are rare but they do exist. This phenomenon was observed on WTI crude oil prices in 2020. However, if an automatic process uses relative or log-normal returns, this method can lead to erroneous results, which can be catastrophic for a bank in a VaR calculation framework. Absolute returns, on the other hand, are not impacted by negative prices and could be a solution.

Now, if the choice of returns as input has been made, it is also possible to make the choice to normalize them, which means dividing them by their standard deviation so as not to be sensitive to volatility. Problems of heteroskedasticity are frequent in financial series, as mentioned above. Thus, if volatility varies over time, returns may have a more or less significant amplitude over time. Since some imputation models are unable to take this heteroskedasticity into account, it is up to the statistician to transform these data, if they do not want the imputation to be impacted. It would then be appropriate to divide these returns by their standard deviation, but this is not such an obvious procedure to implement, even on a complete dataset.

For this, it is necessary to be able to determine a method for obtaining a daily volatility, which would vary over time to take heteroskedasticity into account. This whole process of standardization of financial series constitutes a full field of research in its own right. It is possible to compute daily volatility in several ways, for example by using historical results through a rolling window, or to model volatility by a generalized autoregressive conditional heteroskedasticity (GARCH) model (conditional variance modeling by Bollerslev [37] in 1986) or even by stochastic modeling (for example, the Heston modeling proposed by Heston [109] in 1993, where the variance is written as a stochastic differential equation). Each of these methods has its own advantages and disadvantages, which can directly impact the quality of imputation depending on the method chosen.

Data standardization could therefore be a solution to improve the results of imputation methods that cannot consider the heteroskedasticity of the data. However, it implies additional choices on sample formation that can make the procedure and calculation time more cumbersome.

On the other hand, the transformation of a series of spot prices into a series of price returns in a context of missing data is not without repercussions on data completion. When the returns are calculated from a series containing missing data, then the proportion of missing data is necessarily higher. Since the calculation of a return is based on two observations, if at least one observation is missing, then the return will

be automatically missing. Hence, if the missing data are all grouped together in the price series, their proportion will be almost the same in the price return series (just one missing value is added), but if the missing data are scattered throughout the price series, then the proportion of missing data in the price return series will be much higher and may even be up to twice as high. However, the larger the proportion of missing data is, the more uncertain the imputation will be. For the same database, therefore, an algorithm using a price series could have an advantage over an algorithm requiring a price return series, because the latter could have to impute much more data.

Finally, the construction of the sample itself will have a real impact on the quality of the data completion, whether it is obviously the very nature of the data used, or the choice of the historical data length, or even the variables that make up the sample, or finally the choice of their format and their transformation. All of these choices will have a direct impact on the completion results. For this reason, whatever choices are made, they must first and foremost be thought out and debated. Obviously, one cannot apply a completion method without having asked these preliminary questions.

2.4 Reviews of the completion methods

To deal with missing data issues, many different completion methods are possible, from the simplest to the most sophisticated. Indeed, there are many methods of data completion and many of them have already been proven in the literature. Often, these methods are applied to medical data (e.g., epidemiology, clinical studies, or cancer research) and less often to financial data. This part of this study presents the completion methods retained in this PhD thesis, through their theories and their applications.

2.4.1 Usual methods

Banks have now all set up daily data-backup processes, on a growing set of data. Once these data are saved in the databases, they can be instantly called up by a multitude of internal programs to provide daily calculations (e.g., profit and loss calculation, value-at-risk and ES). However, if one or more of these data are missing among this ocean of new daily information, this can lead to a multitude of calculation errors later on. Thus, in order not to deal with this wave of computer errors to be processed or absurd results, banks use default missing data management methods.

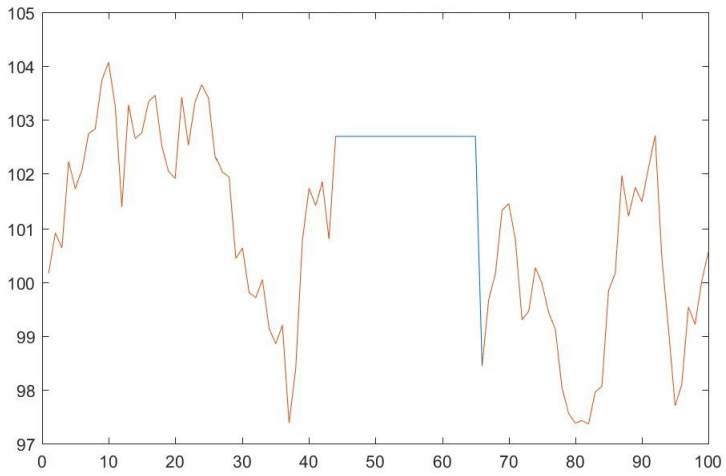
Last observation carried forward

Generally the default solution implemented for a daily process is to replace the missing data by retrieving data from the previous day. This method is called last observation

carried forward (LOCF). In other words, if the value at time t (noted S_t) is missing, then this methodology will consider that $S_t = S_{t-1}$. This method is based on the assumption that the price series is a martingale, where the present price is the conditional expectation (knowing past prices) of prices at a future date. The reverse method also exists, which consists of retrieving the next observation to complete the missing data preceding that same observation. This is the next observation carried backward (NOCB) method. In fact, it is common to combine these two methods, first using the LOCF and then the NOCB.

However, this method of management must remain exceptional and temporary. If the frequency of missing data increases, then the quality of the series will necessarily be impacted. If a large number of data are missing successively in the series, then this methodology will show a plateau due to the lack of variation of the data. In addition, in the case of a significant difference between the value preceding the missing data and the value following the missing data, this will reveal a significant shock to returns, which is not necessarily representative of the series and can be costly in terms of capital charges, as shown in Figure 2.4-1.

Fig. 2.4-1: Interpolation by the last observation



As already discussed in the introduction, this method is equivalent to doing nothing since it leads to returns that have quotation dates that can be more or less distant. The LOCF method involves, indeed, the same problems as making a return between a Friday and a Monday, without considering the weekend effect as described by French [85] in 1980.

Linear interpolation

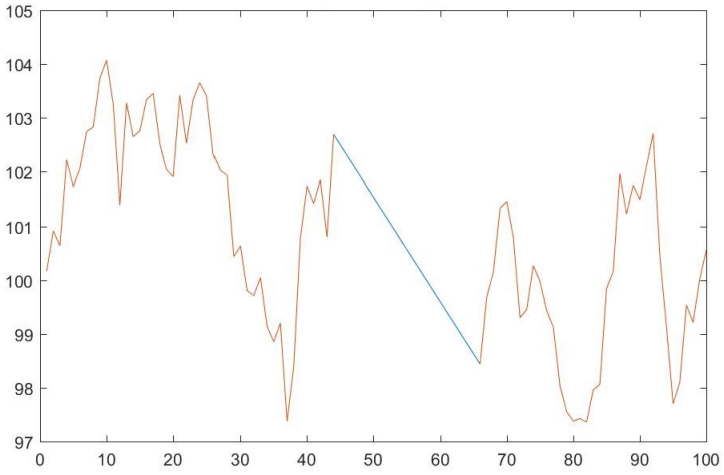
Another obvious way of dealing with missing data is through linear interpolation. The simplest way to fill in missing data, without suffering a large jump, is to connect the values preceding and following the missing data with a straight line.

Let Y be the time series containing missing data; y_t, the last observation before the hole (one or more missing data); and y_{t+l+1}, the first observation after it, with l missing values to fill. The straight line connecting each y_i with $i \in [t, \cdots, t+l+1]$ is given by

$$y_i = y_t + (y_{t+l+1} - y_t)\left(\frac{i-t}{l+1}\right). \tag{2.4-1}$$

In this way, the completed data will give smooth returns and will not cause any shock, as could be the case with the previous method (Figure 2.4-2).

Fig. 2.4-2: Linear Interpolation



Still, this can happen only if the starting point and especially the end point are known. In the case of daily missing-data management, the end point is unknown and a linear extrapolation can be used. On the other hand, if the value following the missing data does not go in the same direction as the slope of the interpolation, then this methodology may inject a non-representative jump.

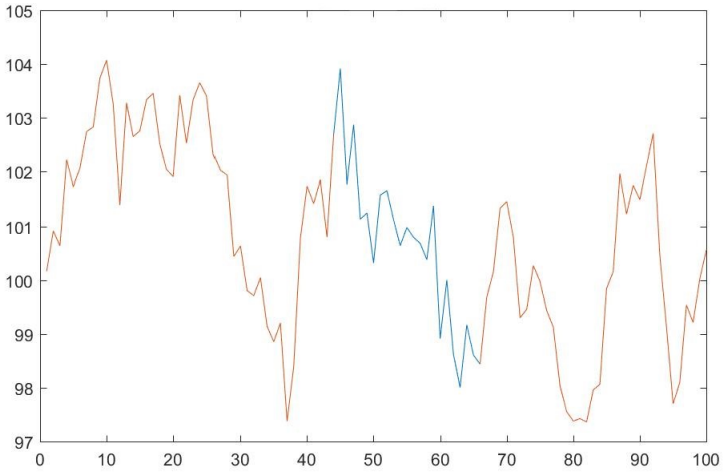
Usually these two methods are applied directly to spot prices as they are interpolation methods. A price adjustment (in order to find the right prices observed) is needed if these methods are applied to returns. Moreover, they have the advantage of being one-dimensional; namely, they do not use any other series as benchmarks or proxies, which makes their implementation even simpler.

In some IT procedures, missing data are automatically replaced by zero. For obvious reasons, this method is undoubtedly the least optimal. It does not represent the market as well as possible because it does not consider the level of the market or even its variations, but in addition it will create outliers in the series. For this reason, this solution sometimes used by default in computer systems is to be rejected and it will be useless to study it further.

2.4.2 Brownian bridge

There is another slightly more sophisticated one-dimensional method based on stochastic calculus. The Brownian bridge is an interpolation method that uses a Brownian motion [6] at each moment where there are missing returns, in order to replace them (Figure 2.4-3).

Fig. 2.4-3: Brownian bridge interpolation



Unlike the previous interpolation methods, this one introduces a random variable. This means that the interpolation varies according to the draw. This implies a problem of non-reproducibility of the results (which will be discussed in more detail in Section 2.4.5). The user may, in fact, be tempted to use the interpolation result from a favorable scenario.

Brownian bridge definition

According to Borodin and Salminen [39] a Brownian motion initiated at a on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is defined as a stochastic process W with the following properties:

(a) $W_0 = a$ almost surely,

(b) $s \mapsto W_s$ is continuous almost surely,

(c) for all $0 = t_0 < t_1 < \ldots < t_n$ the increments

$$W_{t_n} - W_{t_{n-1}}, W_{t_{n-1}} - W_{t_{n-2}}, \ldots, W_{t_1} - W_{t_0}$$

are independent and normally distributed with

$$\begin{aligned} \mathbb{E}(W_{t_i} - W_{t_{i-1}}) &= 0 \\ \mathbb{E}(W_{t_i} - W_{t_{i-1}})^2 &= t_i - t_{i-1}. \end{aligned} \tag{2.4-2}$$

Further on, the authors define a Brownian bridge from a to b of length l. Let $a, b \in \mathbb{R}$ and $l > 0$ be given. A continuous Gaussian process $Y = Y_t : 0 \leq t \leq l$, $Y_0 = a$, such that

$$\mathbb{E}(Y_t) = a + (b-a)\tfrac{t}{l}, \qquad \boldsymbol{Cov}(Y_t, Y_s) = (s \wedge t) - \tfrac{st}{l}$$

Hence, $\mathbb{E}(Y_l) = b$ and $\boldsymbol{Cov}(Y_t, Y_s) = 0$ if $s = l$ or $t = l$. Therefore, almost surely $Y_l = b$. Because of this property, Y is called a Brownian bridge.

Furthermore, a Brownian bridge from a to b can also be characterized as the unique solution of the stochastic differential equation (Ikeda and Watanabe [118]):

$$\begin{cases} dY_t = \frac{b-Y_t}{l-t}\, dt + \, dWt, & \leq t < l, \\ Y_0 = a, \end{cases} \tag{2.4-3}$$

with W a standard Brownian motion started at 0.

The solution is

$$Y_t = a + (b-a)\frac{t}{l} + (l-t)\int_0^t \frac{dW_s}{l-s} \tag{2.4-4}$$

Thus, as mentioned earlier, the Brownian bridge is suitable for Gaussian returns, but its application to non-Gaussian data does not make sense. This method is not universal, contrary to linear interpolation and the retrieval of the previous day's data, as it requires prior verification of the normality of the series. However, the Brownian bridge is theoretically suitable for spot price series, as their performance can be assimilated to a log-normal distribution. If the price of a quotation changes from S_0 to S_T, between time 0 to time T, then the continuously compounded return $r_{0,T}$ is defined as

$$r_{0,T} = \ln\left(\frac{S_T}{S_0}\right) \iff \frac{S_T}{S_0} = \exp\left(r_{0,T}\right) \tag{2.4-5}$$

with $r_{0,T} \sim \mathcal{N}(\mu T, \sigma^2 T)$, which means that $\ln(\frac{S_T}{S_0}) \sim \mathcal{N}(\mu T, \sigma^2 T)$; therefore by definition $\frac{S_T}{S_0}$ is a log-normal random variable.

Since it is reasonable to assume that financial price series follow log-normal dynamics, the application of a Brownian bridge makes sense. However, using it on any financial series is not so simple. Indeed, nothing guarantees the normality of volatility series, repo rates, correlation and so forth, which is not always the case. Moreover, it is quite possible that the returns of a price series are not normal, for example in the case of a series with high volatility or in the presence of multiple jumps. It is thus important to check beforehand whether the series is normal.

2.4.3 K-nearest neighbor

Since univariate methods are relatively limited in their effectiveness, it is possible to use other series in order to fill in missing data. These kinds of algorithms are called multidimensional algorithms, because they no longer work only from the series itself, but consider the dynamics of other given series. Thus, the missing data of a series are filled using the dynamics of well-chosen benchmarks.

K-nearest neighbor presentation

The first multidimensional algorithm presented here is the K-nearest neighbors (K-NN) algorithm, introduced by Fix and Hodges in 1951 [83]. It is one of the first machine-learning algorithms, generally used for classification.

This clustering algorithm is a non-parametric method based on a simple principle: find the K-samples closest to the sample to be complete (with $K \in \mathbb{N}$). The proximity between samples is calculated from distance measurement, usually a Euclidean measurement, to find the closest K-samples that minimize this distance. Thus, the K-NN-algorithm needs only a positive integer K, a sample space and a proximity metric to work.

This method can be adapted to impute missing values of a time series using the K closest time series. To fill in the missing values, it is assumed that one value can be approximated by the closest other values. In this case, suppose $\boldsymbol{Y}$ is the data matrix of size $(n \times p)$, composed of p time series $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_p$. Let $\boldsymbol{Y}_1$ be the time series with missing data and $\boldsymbol{Y}_2, \dots, \boldsymbol{Y}_p$ the $p-1$ complete time series used to reconstruct the missing data of $\boldsymbol{Y}_1$. If a Euclidean measure is used into the K-NN algorithm, the distance between $\boldsymbol{Y}_1$ and the other $\boldsymbol{Y}_j$ with $j \in [2, \dots, p]$ is defined by

$$\begin{aligned} dist(\boldsymbol{Y}_1, \boldsymbol{Y}_j) &= ||\boldsymbol{Y}_1 - \boldsymbol{Y}_j|| \\ &= \sqrt{\sum_{i=1}^{n^{obs}} (\boldsymbol{Y}_{i,1} - \boldsymbol{Y}_{i,j})^2}. \end{aligned} \tag{2.4-6}$$

where n^{obs} is the set of indices where $\boldsymbol{Y}_1$ is observed, and $\boldsymbol{Y}_{i,j}$ corresponds to the i^{th} observation (with $i \in [1, \dots, n^{obs}]$) of the j^{th} column (with $j \in [2, \dots, p]$) of $\boldsymbol{Y}$.

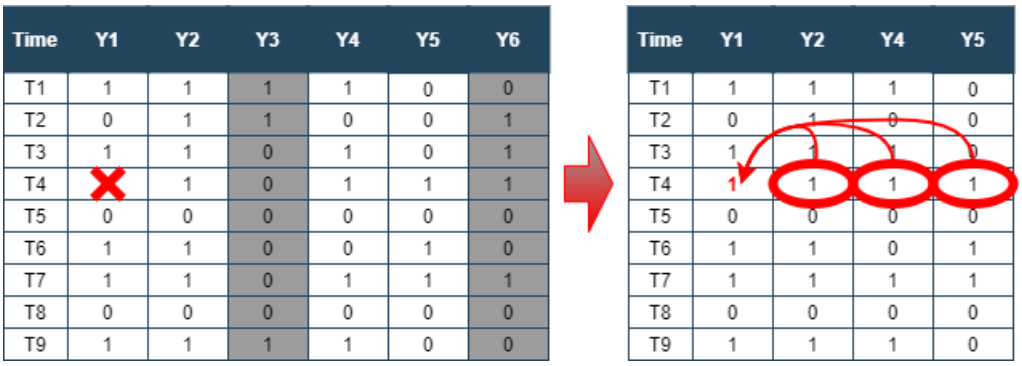
Further, the closest neighbors are the K-samples (from the p time series) that minimize this distance. These K neighbors are then used to reconstruct missing values of $\boldsymbol{Y}_1$ using weights inversely proportional to distances. Thus, the missing values of $\boldsymbol{Y}_1$ are imputed as

$$\boldsymbol{Y}_{i,1} = \sum_{k=1}^{K} \left(\frac{dist(\boldsymbol{Y}_1, \boldsymbol{Y}_k)^{-1}}{\sum_{k=1}^{K} (dist(\boldsymbol{Y}_1, \boldsymbol{Y}_k)^{-1})} \boldsymbol{Y}_{i,k} \right), \tag{2.4-7}$$

with K the number of nearest neighbors and $i \in [1, \ldots, n^{miss}]$ (where n^{miss} is the set of indices where $\boldsymbol{Y}_1$ is missing).

The operation of the K-NN is shown in Figure 2.4-4, where the $\boldsymbol{Y}_1$ series has one missing data point to complete. If the operator wishes to use the three closest, series to complete the missing data, then it turns out that the $\boldsymbol{Y}_2$, $\boldsymbol{Y}_4$ and $\boldsymbol{Y}_5$ series will be the closest according to a Euclidean measurement ($\boldsymbol{Y}_3$ and $\boldsymbol{Y}_6$ are grayed out because they are the furthest apart series). Thus, data from $\boldsymbol{Y}_2$, $\boldsymbol{Y}_4$ and $\boldsymbol{Y}_5$ at time T_4 will be used to reconstruct the missing value of $\boldsymbol{Y}_1$.

Fig. 2.4-4: Example of K-NN algorithm



Time	Y1	Y2	Y3	Y4	Y5	Y6
T1	1	1	1	1	0	0
T2	0	1	1	0	0	1
T3	1	1	0	1	0	1
T4	✖	1	0	1	1	1
T5	0	0	0	0	0	0
T6	1	1	0	0	1	0
T7	1	1	0	1	1	1
T8	0	0	0	0	0	0
T9	1	1	1	1	0	0

Time	Y1	Y2	Y4	Y5
T1	1	1	1	0
T2	0	1	0	0
T3	1	1	1	0
T4	1	1	1	1
T5	0	0	0	0
T6	1	1	0	1
T7	1	1	1	1
T8	0	0	0	0
T9	1	1	1	0

In the case of financial time series, the K-NN algorithm must take returns as input in order to compare the dynamics of each time series and thus, choose the K most similar time series.

The K-NN algorithm gives good results if applied to a huge database; indeed, the more time series are used in the algorithm, the more it will learn about neighbors to help fill in missing values. Of course, for the algorithm to be effective, a sufficiently large number of complete observations is required. The proximities between series must not be calculated from too few observations. One disadvantage of this algorithm is its computation time if it is used with very large database, because it must scan all the data before performing any completion.

K-NN imputation widely used on medical data

The K-NN method is very popular in medicine. García-Laencina, Sancho-Gómez and Figueiras-Vidal [89], in 2009, classified incomplete data using four different estimation techniques:

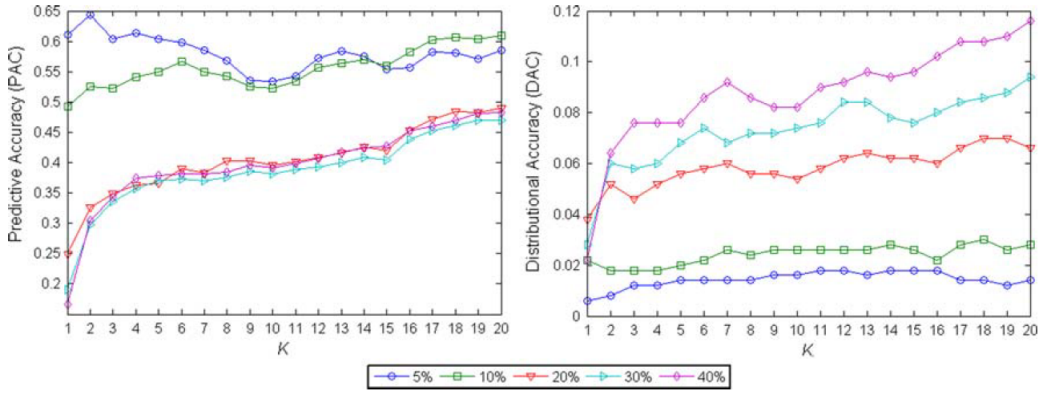
- K-NN: selects the K closest cases that are not missing values in the attributes to be imputed, such that they minimize the heterogeneous euclidean overlap metric (HEOM) distance [1] (see Section 2.4.3);
- Self-organization maps (SOM): a neural network model made out of a set of nodes organized on a grid and fully connected to the input layer and trained to estimate missing values;
- Multi-layer perceptron (MLP): estimates missing data by training multiple layers of computational units interconnected in a feed-forward way, to learn incomplete features, using the remaining complete features as inputs; and
- EM: computes the log likelihood of the data and then find the parameters that maximize this likelihood (see Section 2.4.6).

First, they used standardized simulated data where 5 to 40% of missing data were randomly inserted. Then they repeated the experiment using data on thyroid disease composed of five Gaussian components.

The study is first conducted using simulated data. The authors use two comparison criteria: predictive accuracy in classification (PAC) and distributional accuracy for classification (DAC). The PAC aims to analyzed the proximity between imputed and true data. This analysis criterion is based on the Pearson correlation between imputed and true data, in other words, a good imputation method will have a PAC close to 1. On the other hand, the DAC aims to measure de preservation of the distribution of the true values, by measuring the Kolmogorov-Smirnov distance (which represents the difference in distribution between imputed data and true data). Thus, a good imputation method will have a small distance value, i.e. a small DAC. The authors show in the lest part of Figure 2.4-5, that PAC highlights a positive relationship between the accuracy of the prediction and the number of K-neighbors chosen. In fact the greater the number of K-neighbors used for imputation, the more accurate the prediction will be (as a PAC close to 1 means a good imputation quality). Conversely, the right part of Figure 2.4-5 reveals a lower DAC for smaller K. This means that the distribution of imputed data is closer to the one of true data when K is low. Moreover, the smaller the amount of missing data, the greater the forecast accuracy and the better the distribution will be respected. The blue curve (corresponding to 5% of missing data) has the PCA closer to 1 and the minimum DAC.

[1] A distance measure that incorporates continuous and categorical attributes. [210]

Fig. 2.4-5: PAC and DAC obtained by K-NN imputation using different values for K, with missing data artificially inserted (Source: García-Laencina, Sancho-Gómez and Figueiras-Vidal, 2010 [89])



A good imputation method in terms of preservation of the true values will have a PAC close to 1. A good imputation method in terms of preservation of the distribution of the truce values will have a DAC close to 0.

The researchers then complete the artificial missing data using the four methods presented earlier (K-NN, MLP, SOM and EM) and obtain the following classification performed by an artificial neural network with six hidden neurons (Table 2.4-1). Moreover, the choice of the K parameter for the K-NN method is made on the basis of the results given in Figure 2.4-5: the K chosen is the one that gives the most precise classification. The EM algorithm outperforms other imputation methods by a wide margin, regardless of the proportion of missing data. They also find that when there are few missing data, the K-NN method is a better classification method than is the MLP method, but for a higher proportion of missing data the reverse is true.

Tab. 2.4-1: Misclassification error rate (mean ± standard deviation from 20 simulations) according to each methods, using a neural network with 6 hidden neurons (Source: García-Laencina, Sancho-Gómez and Figueiras-Vidal, 2010 [89])

	Missing data imputation			
Missing data in x_1 (%)	**K-NN**	**MLP**	**SOM**	**EM**
5	9.21 ± 0.56	9.97 ± 0.48	9.28 ± 0.84	8.29 ± 0.24
10	10.85 ± 1.06	10.86 ± 0.79	9.38 ± 0.52	9.27 ± 0.54
20	11.88 ± 1.01	11.42 ± 0.44	10.63 ± 0.54	10.78 ± 0.59
30	13.50 ± 0.81	12.82 ± 0.51	13.88 ± 0.67	12.69 ± 0.57
40	14.89 ± 0.49	13.72 ± 0.37	15.55 ± 0.66	13.31 ± 0.56

For 5% of missing data in x_1, the missclassification error rate for the K-NN méthod is equal to 9.21 ± 0.56 (standard deviation based on 20 simulations).

Table 2.4-2 presents the classification task trained by a neural network with 18

hidden neurons. The K-NN and MLP methods provide the best classifications among the four methods and these results are even more visible with a large proportion of missing data. The authors then decide to highlight the K-NN method because it is a simple method (unlike MLP) that leads to the best classification. The K-NN method consists only in a simple calculation of distance and the choice of K, while the MLP method requires one to train several neural networks (which requires a certain cost in terms of computing time).

Tab. 2.4-2: Misclassification error rate (mean ± standard deviation from 20 simulations) according to each methods, using a neural network with 18 hidden neurons (Source: García-Laencina, Sancho-Gómez and Figueiras-Vidal, 2010 [89])

	Missing data imputation			
Missing data in x_2 (%)	**K-NN**	**MLP**	**SOM**	**EM**
5	15.92 ± 1.26	15.84 ± 1.13	16.32 ± 1.13	16.19 ± 0.99
10	16.88 ± 1.16	16.87 ± 1.16	16.97 ± 1.18	16.85 ± 1.03
20	18.78 ± 1.29	19.09 ± 1.29	19.30 ± 1.23	19.23 ± 1.12
30	20.58 ± 1.31	20.76 ± 1.34	22.04 ± 1.01	21.22 ± 1.12
40	22.61 ± 1.30	22.76 ± 1.23	24.06 ± 1.29	23.11 ± 1.37

For 5% of missing data in x_2, the missclassification error rate for the K-NN méthod is equal to 15.92 ± 1.26 (standard deviation based on 20 simulations).

Finally, García-Laencina, Sancho-Gómez and Figueiras-Vidal [89] apply all four methods to empirical data from sick-thyroid disease, composed of 2,800 training instances and 972 test instances. Table 2.4-3 shows the percentage of missing data for the five selected variables.

Tab. 2.4-3: Missing data percentage in the five selected variables of thyroid dataset (Source: García-Laencina, Sancho-Gómez and Figueiras-Vidal, 2010 [89])

	Input feature				
Thyroid dataset	x_1 (%)	x_2 (%)	x_3 (%)	x_4 (%)	x_5 (%)
Training	10.14	20.89	6.57	10.61	10.54
Test	8.74	18.93	4.84	9.26	9.26

The proportion of missing data for the training set of x_1 is equal to 10.14% and that of the test set is equal to 8.74%.

The target class variable is modeled by an artificial neural network composed of 20 hidden neurons. Once missing data are imputed, the authors find lower misclassification rates (over 20 simulations) for the K-NN method (Table 2.4-4), while the EM algorithm provides the worst results.

Tab. 2.4-4: Misclassification error rate (mean ± standard deviation from 200 simulations (Source: García-Laencina, Sancho-Gómez and Figueiras-Vidal, 2010 [89])

	Missing data imputation			
	***K*-NN**	**MLP**	**SOM**	**EM**
Misclassification error rate (%)	3.01 ± 0.33	3.23 ± 0.31	3.49 ± 0.35	3.60 ± 0.31

The misclassification error rate in sick-thyroid dataset for the K-NN model is 3.01 ± 0.33 (standard deviation based on 20 simulations).

In 2010, Jerez, Molina, García-Laencina, Alba, Ribelles, Martín and Franco [124] worked on the completion of missing data in the medical field by applying the K-NN method as an imputation method. The group of Spanish researchers studied a sample of empirical breast cancer data to apply eight different completion algorithms and then compared the results with those of listwise deletion (see Section 2.2.2), using the following methods:

- Mean imputation: completes the missing data by the mean of the non-missing data;
- Hot-deck imputation: uses observations from the sample according to a similarity criterion (in other words, a K-NN with $K = 1$);
- SAS: multiple imputation (five imputations) using the function *PROC MI* from SAS software, which assumes a multivariate normal distribution ;
- Amelia: multiple Imputation based on an EM algorithm on bootstrapped data (see Section 2.4.6) with five imputations;
- Multiple imputation by chained equations (MICE): multiple imputation with a linear mixed model based on predictive mean matching and regression methods (see Section 2.4.7) with five imputations combined by averaging;
- Multi-layer perception (MLP): estimates missing data by training a multiple layers of computational units interconnected in a feed-forward way, to learn incomplete features, using the remaining complete features as inputs;
- K-NN: selects the K closest cases that are not missing values in the attributes to be imputed such that they minimize the HEOM distance [2] (see Section 2.4.3); and
- SOM: a neural network model made out of a set of nodes organized on a 2D grid, fully connected to the input layer and trained to estimate missing values.

[2] A distance measure that incorporates continuous and categorical attributes. [210]

They apply all of these algorithms to data on breast cancer in Spain, involving 3,679 women with operable invasive breast cancer diagnosed in 32 hospitals belonging to the Spanish Breast Cancer Research Group between 1990 and 1993. This database consists of both continuous and categorical data. Missing data represents 5.61% of the total data. They counted 1,678 observations with at least one associated missing data point, representing 45.61% of the patients. Hence, the listwise deletion would allow one to exploit only a little more than half of the database.

In order to best compare each methodology, missing data were first removed from the sample using the listwise deletion. As shown in Table 2.4-5 and Figure 2.4-6, all imputation methods except the hot-deck method lead to an improvement in the accuracy of prediction, as measured by the area under the curve (AUC). In addition, statistical tests were conducted (Friedmans'test and Wilcoxon signed-rank test) to determine whether the differences observed were significant; only the results obtained using the three machine-learning algorithms (MLP, K-NN and SOM) differed significantly from those in which records containing missing values are eliminated (listwise deletion method). The best predictions were obtained using K-NN, in which the AUC was 0.7345 ± 0.0289 (mean plus or minus the standard deviation calculated using 10-fold cross-validation); this represents a 2.71% improvement over the listwise deletion. Moreover, methods based on machine learning outperform statistical methods and that the K-NN method obtains the best results.

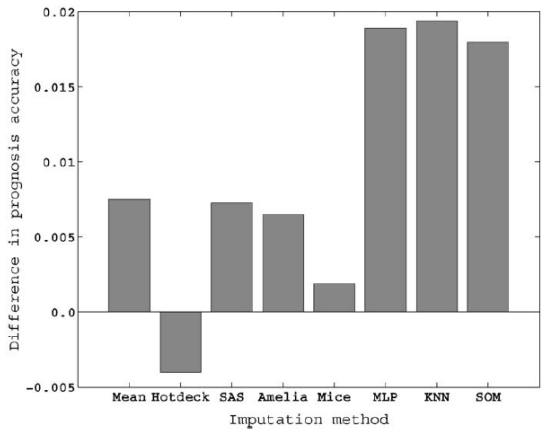
Thus, for this study, the K-NN method was found to be the best approach to fill in missing data because it leads to an improvement in prediction accuracy.

Tab. 2.4-5: Mean, standard deviation and MSE values for the AUC values computed for the control model and for each of the eight imputation methods considered (Source: Jerez and al., 2010 [124])

AUC	**LD**	**Mean**	**Hot-deck**	**SAS**	**Amelia**	**Mice**	**MLP**	***K*-NN**	**SOM**
Mean	0.7151	0.7226	0.7111	0.7216	0.7169	0.7250	0.7340	0.7345	0.7331
Std dev.	0.0387	0.0399	0.0456	0.0296	0.0297	0.0301	0.0305	0.0289	0.0296
MSE	0.0358	0.0235	0.0324	0.0254	0.1119	0.1119	0.0240	0.0195	0.0204

The mean of the area under the curve obtain by listwise deletion is equal to 0.7151, its standard deviation is equal to 0.0387 and its MSE is equal to 0.0358.

Fig. 2.4-6: Difference in AUC means between the reference model and each imputed data (Source: Jerez and al., 2010 [124])



Mean imputation lead to an improvement of accuracy in terms of area under the curve (around 0.008), contrary to hotdeck imputation that lead to a degradation of accuracy (around -0.004).

This section shows that the K-NN method is advantageous due to its simplicity of understanding and implementation - but it is no less efficient for all that. García-Laencina, Sancho-Gómez and Figueiras-Vidal [89] in 2009 and Jerez, Molina, García-Laencina, Alba, Ribelles, Martín and Franco [124], in 2010, show in their studies that this method can be very effective in filling in missing data, even competing with other, more-sophisticated models.

Optimal number of K neighbors

Obviously, the performance of this method remains very sensitive to the sample used to complete missing data but also to the number of neighbors and the methodology to choose them. For example, if the column to be imputed has exactly the same returns as another column in the sample, choosing a $K \geq 1$ will be useless since the use of any additional columns will have a negative impact on the imputation quality. Although this is a very particular scenario, it raises the need to choose the number of neighbors in relation to the sample used. Thus, using a fixed number of nearest neighbors without considering the sample clearly appears to be sub-optimal, that is why, a dynamic procedure for determining the number of neighbors to use must be implemented.

In a general way, if the K parameter chosen is too small, it will lead to underfitting, because the few variables used will have a important impact on the result and the method is likely to produce unstable results. In contrast, if K is too big, then this will lead to overfitting, because the distinction made between classes will lose efficiency,

which means here lower variance but increased bias. Hence, the challenge here is to find the K parameter that allows neither underfitting nor overfitting. However, in the literature, there is no predefined statistical method for determining this K parameter and this determination is even more complicated with missing data.

However, it is possible to find several ways to choose the K parameter. The goal here is to find the K that does neither under-fits nor overfits. Generally, the rule of thumb commonly used is that K must be equal to $\sqrt{p}$ [71]. This is clearly sub-optimal. Some people even prefer to test all possible Ks, to deduce the one that allows one be the most accurate, before continuing the analyses. This is notably the case of the article by García-Laencina, Sancho-Gómez and Figueiras-Vidal [89] presented previously; they start from a complete sample where they determine the optimal K, then take this level to apply it on the same sample containing this time missing data. However, even from a complete sample, the choice of K is not obvious because they determine the optimal K by calculating the classification accuracy for different levels of K (ranging from 1 to 20). Furthermore, the K selected here are those that maximize this accuracy.

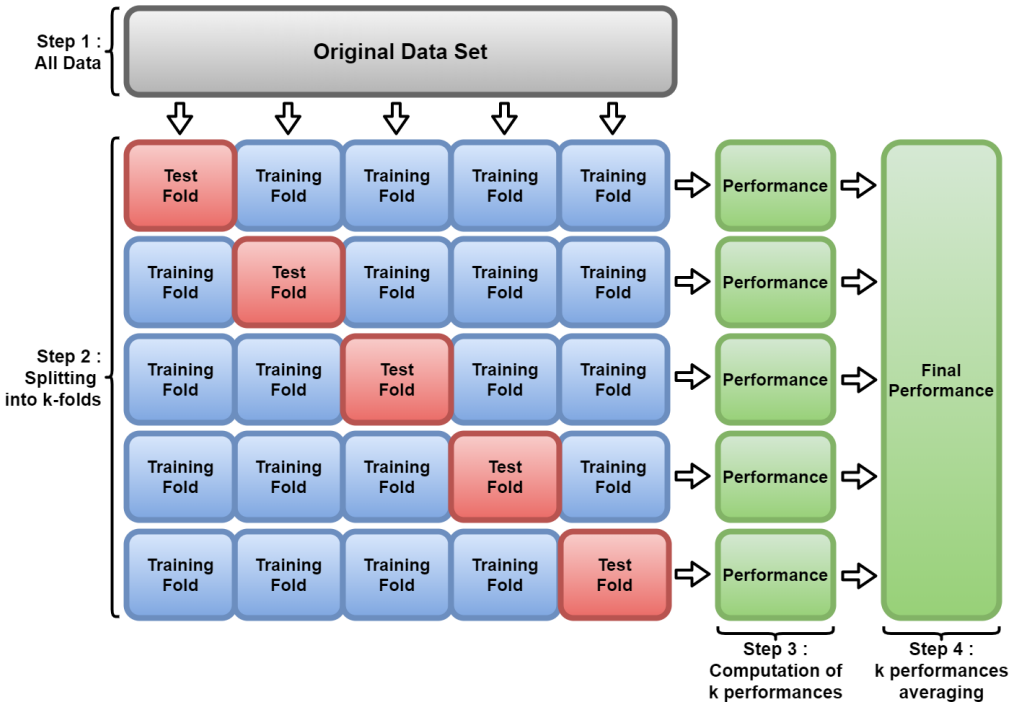
However, one of the most satisfactory way to determine K still remains by cross-validation. In the case of Jerez, Molina, García-Laencina, Alba, Ribelles, Martín and Franco [124], they select the K parameter by cross validation. This methodology is widely used to choose the number of neighbors to used in the K-NN algorithm, but also for other algorithm that need to specify some parameters. An improved cross-validation procedure is used by Josse and Husson [126], in Section 2.4.8, in order to be used with iterative PCA.

Cross validation was introduced in 1951 by Mosier [160], who sought to evaluate the predictive validity of linear regression equations used to predict a performance criterion. This technique allows one to measure the performance of a model and in this case, to find out for which K the K-NN method is the most efficient to give the best prediction. This is a re-sampling method that randomly divides the data of approximately equal size in order to measure the performance of the model. The idea here is then to measure the performance of the model for several possible K and to select the K that will maximize the performance of the K-NN algorithm.

The intuition of the cross-validation technique is to split the dataset in two parts, where one part will be dedicated to training and the other to testing. Thus, the model will be evaluated from the training data and the performance of the model will be computed from the test data. More precisely, the dataset is divided into q-folds, where one of them is the test fold and the other are the training folds. Usually the method is called k-folds, but here the notation q will be used instead of k to avoid any confusion with the K value corresponding to the number of neighbors used. This procedure is repeated q times, so that each fold was once the test fold. As a result of these q iterations, an average performance is obtained. In this way, the performance of the model is not deduced from a single data split, but from the average performance of q

data splits calculated from q iterations. An example with $q = 5$ is presented in Figure 2.4-7.

Fig. 2.4-7: q-fold cross validation procedure with $q = 5$



Thus, all data is split into q-folds (usually 5 or 10) and for each iteration this dataset is divided in two parts: the training set composed of $(q-1)$-folds and the test set composed of one fold. Usually the training set represents 90%, 80%, or 70% and respectively, the test set 30%, 20%, or 10%.

In this case, the goal is to find the best value of K from the K-NN imputation method by q-fold cross validation. It is important to not confuse here q and K: q is the number of folds used in the cross validation and K is the number of nearest neighbors used to impute missing data. The cross validation for K-NN imputation consist in finding for which K the K-NN algorithm gives the best performance on imputed data. Hence, the procedure needs to be repeated for several K.

Thus, for a given K, the dataset with missing data has to be imputed using the K-NN algorithm with parameter K. Then, data is divided into q folds in order to split the data into a training set and a test set, as explained previously. When $q = 1$, the process is called a hold-out validation. This method is dependent on just one train-test split, making the hold-out method performance dependent on how the data is split into training and test set. On the other hand, when $q > 1$, a K-NN predictive model is applied to the training set for each of the q iterations. Forecasting with K-NN works

in the same was as K-NN imputation: the predictions are computed by a weighted average of the K nearest neighbors observations. Hence, the training set allows one to find which of the variables are the K-nearest neighbors and what weights are allocated to them. Once this is done, these same parameters are applied to the test set to calculate the performance of the model. The test set allows one calculate a mean squared error (MSE) between the forecasts made and the real observations. After the q-iterations are done, q-MSE has been calculated. From then on, the performance of the model is given by the average of the performances, which means by the average of the MSEs. Thus, an average MSE is calculated for each K and the K used for imputation is that which minimizes the average MSE.

The advantage of cross-validation is that it allows the totality of the data to be used. In this PhD thesis, q is fixed at 10, so the model is trained on 90% of the data. Moreover, this method allows one to see whether the method is stable or not, by comparing the MSEs to the others. If the MSEs calculated in each iteration differ considerably from each other, then the performance of the model depends strongly on the dataset it is trained on and therefore the model is unstable.

The K-NN method is rather fast when the K parameter is given, on the other hand, it becomes much more expensive in computation time when this parameter is to be defined. It is necessary to first impute the data and then calculate the performance for a every possible K.

Thus, cross-validation can be used to determine the number of neighbors, but also the distance to use. In the context of a financial series, the K-NN method will be applied to price returns. Financial returns are usually Gaussian, applying a Euclidean distance makes sense. In order not to further complicate the algorithm, the K-NN method will use only the Euclidean distance in these estimates.

On the other hand, choosing neighbors based on a metric distance can be sub-optimal. For example, if one of the series in the sample has returns twice as high as those of the series to be imputed: since the two are perfectly correlated, this single series contains all the information necessary for imputation, but it may not be the closest in terms of Euclidean distance. In this case, the discrimination by Euclidean distance, could lead not to use this series which would result in the best prediction. The data standardization could help, but it consists in centering and reducing the data from the mean and standard deviation of the observed data only. The presence of missing data will bias these parameters. The correlation of the observed standardized data still be 1 but the covariance can be very different which will lead to distort the observed data. This example is not the only one. If, this time, a series is imputed using, for example, the 2 closest series in terms of Euclidean distance, it is not impossible that the use of a third more distant series will improve the imputation quality by adding noise to the imputation. There are many examples revealing the limits of the choice of the number of neighbors per Euclidean distance.

To some extent, K-NN imputation is quite similar to regression imputation, except that the selection of columns to be used for regression imputation relies on methods that can be more sophisticated and optimal than a simple distance metric. The regression imputation is based on column discrimination methodologies such as the stepwise regression [40] or the Lasso [198], allowing to use the right columns in order to maximize the R^2. On the contrary, the K-NN method uses only the distance to choose its columns (the nearest only) and does not focus on determining the best combination of columns to improve the imputation quality. Thus, as already mentioned, it is possible to a column to be far from the column to be imputed, but to bring an additional information (a noise for example) improving the imputation quality. It is clear that the choice of nearest neighbors is sub-optimal compared to the variable selection methodology used in the case of regression imputation, nevertheless, the K-NN method is still widely used in the literature.

K-NN improved by bagging

The K-NN algorithm is a very simple and generalized imputation technique, which can be combined with other methods to refine the model and improve the quality of imputation. Previously a cross-validation method was presented, which can be added to the K-NN in order to improve the results. Another possibility to improve the results can be to add bootstrap. It is possible to use B samples drawn randomly with replacements in the original sample before imputing missing data using the K-NN method. Thus, once the B samples have been imputed, they can be aggregated, in order to obtain the final result. This is a methodology called bagged K-NN. The term "bagged" refers here to the bagging technique, introduced by Breiman [41] in 1996, which consists in creating bootstrap samples, then applying a basic rule to them (here the K-NN method) before aggregating all the results obtained. The bagging procedure is explained detailed in Section 2.4.5, because it is an integral part of the random forests algorithm. Globally, the combination between the bagging and the K-NN allow one to stabilize the results obtained by the K-NN method, giving less weight to outliers that would negatively affect imputation. Using bagging in conjunction with K-NN is relevant since financial returns are supposed to be independent and identically distributed.

Combining the K-NN method with bagging is notably what Biau and Devroye [32] and also Biau, Cérou and Devroye [31] did in 2010 where they talked about "bagged nearest neighbors". In the first paper, Biau and Devroye [32] show the consistency of the bagged K-NN model in a regression framework: the bagged K-NN estimator converges well to the true regression function when the number of observations is large enough. In the second paper mentioned, Biau, Cérou and Devroye [31] present bagging as a procedure to make the K-NN method much more efficient: the K-NN method is quite poor in the sense that it uses only the closest variables and integrating bootstrap allows one to take into account different information in variables that are farther from

the closest neighbors but still close.

In the same way as cross-validation, adding bagging to an algorithm such as K-NN (or to any other algorithm) will contribute to significantly increase computation time. The imputation by K-NN will have to be done on each of the bootstrapped samples. The computation time will be much longer as the number of bootstrap samples will be important. In addition, this computation time will be even longer if a cross validation step is necessary to find the number of K to use, especially if it is done on the bootstrap samples.

Thus, the algorithm that will be used in the comparative analysis of this PhD thesis will consist of four main steps, for each K:

1. Creation of B bootstrap samples drawn randomly with replacement from the original sample;

2. Completion of each bootstrapped sample using the K-NN method by imputing variable by variable in ascending order of missing data (explained in the next sub-section);

3. Deduction of the average imputed sample by averaging the B previously imputed sample; and

4. Calculation of a performance measure (mean absolute error) by cross-validation applied to the mean imputed sample.

Thus, each K is associated to a single imputed sample and to a performance measure. The K retained and thus the imputed sample chosen as the final imputation result, is therefore the one with the best performance measure (here the lowest mean absolute error).

Finally, integrating bagging into the K-NN method and especially bootstrapping, adds a random component that was not originally present in the algorithm. The random draw with bootstrapped samples makes the results of this methodology non replicable. However, this can be a constraint for the regulator because, at first glance, there is no guarantee that the result is not the result of an optimist scenario. The problem of non-replicability of results is discussed in more detail at the end of Section 2.4.5.

Imputation ordering

Finally, the last problem raised here is a problem common to all multidimensional imputation methods that impute missing data variable by variable, in other words, where the result of the imputation is based on observations of other variables as K-NN.

In the literature, variables are generally treated by ascending of missing data (from the one with the least missing data to the one with the most)[114][179][217]. This

is notably the case of Huang, Keung, Sarro, Li, Yu, Chan and Sun [114] who are working on an empirical study of the cross validation of a K-NN model on software quality datasets. Ordered imputation can potentially impact the quality of the final imputation. The algorithm naively imputes the first column from the others, then the second column from the other columns, including the first that was imputed in the previous step and so on. This means that the imputed values depend on all the observed values but also on all the imputed values before it. Therefore, if imputation first treats the variables with the least amount of missing data, this will lead to an increase in the weight on the imputation of the observed values compared to the previously imputed values. This is, indeed, what Sahri, Yusof and Watada [179] show in their article in 2014. They use a K-NN ordering imputation and explain that this allows more accuracy imputation and to preserve the correlation between each variables.

2.4.4 Multivariate singular spectrum analysis

According to Borio [38], the financial cycle "denote self-reinforcing interactions between perceptions of value and risk, attitudes towards risk and financing constraints, which translate into booms followed by busts. These interactions can amplify economic fluctuations and possibly lead to serious s financial distress and economic dislocations". In other words, economic actors are optimistic when the financial market is rising, which then leads to financial imbalances that can result in a brutal correction in the event of a market slowdown, followed by a limitation in the credit supply from banks, accentuating the scale and duration of the recession. Some methods initially developed to reproduce marine cycles can be used on financial data, given their cyclicality. Singular spectrum analysis (SSA; introduced by Broomhead in 1986 [46][47]) has been used in various fields, including biomedical signal processing, image processing, earth sciences, hydrology, climatology, meteorology, marine science and recently, in finance and economics. SSA is particularly useful for analyzing data with complex seasonal patterns and non-stationary trends. It is a non-parametric method that does not make assumptions about the data. An extension for multivariate time series has been developed. It is called multivariate (or multichannel) singular spectral analysis (MSSA).

Singular spectrum analysis

The SSA decomposes chronological series into three components, namely trend, harmony and noise. The series is then reconstructed using the trend and harmony estimators. It is in this reconstruction phase that the missing data are imputed. The SSA algorithm is a non-parametric technique, and it is used often to model non-linear, non-stationary and noisy time series [103][45].

The SSA algorithm depends on two parameters that are set: the size of the window L and the number of eigenvalues r. The model is very sensitive to these parameters.

The choice of L influences decomposition and thus the accuracy of the forecast. If r is not large enough, the algorithm suffers a loss of information. As a result, the reconstructed series are less accurate. However, if r is too large, the reconstruction is affected by noise. The optimal choice is not simple. It becomes even more complex in the case of MSSA, where orthogonality and the similarity of series are important.

In 2013, Hassani and Mahmoudvand [104] explained that understanding MSSA requires a firm grasp of its univariate version, where it is often assumed that the series is noisy. Suppose that there is a univariate stochastic process $\{y_t\}_{t\in\mathcal{Z}}$ of size n such that $Y_n = \{y_1, y_2, \ldots, y_n\}$. In a first step, the noisy series Y_n is decomposed.. In the second step, once reduced by its noise (which is used for prediction purposes), the series is reconstructed. In the decomposition step, the "embedding" procedure entails transforming the one-dimensional series into a multidimensional one by choosing an observation vector of size L and moving it into the observed series Y_n. This produces a trajectory matrix $\boldsymbol{H}$ whose dimensions are L by $K = n - L + 1$ defined as follows:

$$\boldsymbol{H} = [H_1, H_2, \ldots, H_K] = (h_{i,j})_{i,j=1}^{L,K} = \begin{pmatrix} y_1 & y_2 & \cdots & y_K \\ y_2 & y_3 & \cdots & y_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & \cdots & y_n \end{pmatrix}. \tag{2.4-8}$$

The trajectory matrix $\boldsymbol{H}$ is a Hankel matrix, which means that all the elements along the "anti-diagonals" are equal. This step is called the embedding step, and and it transforms the data that is contained in a vector into a matrix. This procedure is reminiscent of the block bootstrap method that was introduced in 1989 by Künsch [136], in which a matrix is created from bootstrapped data.

The singular value decomposition (SVD) of $\boldsymbol{H}$ is another step that is based on the spectral decomposition of the lag-covariance matrix $\boldsymbol{H}\boldsymbol{H}^T \in \mathbb{R}^{L\times L}$, which is symmetric and positive semidefinite. Therefore, it has a complete set of eigenvectors and can be diagonalized as $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T$, where $\boldsymbol{\Sigma}$ is the diagonal matrix of dimensions $(L \times L)$ of eigenvalues $\lambda_1 \geq \ldots \geq \lambda_L \geq 0$ and $\boldsymbol{U} = (U_1, \ldots, U_L)$ is an orthogonal matrix of the eigenvectors of the matrix $\boldsymbol{H}\boldsymbol{H}^T$. The singular value decomposition of the trajectory matrix $\boldsymbol{H}$ allows the grouping step to proceed and can be written as follows:

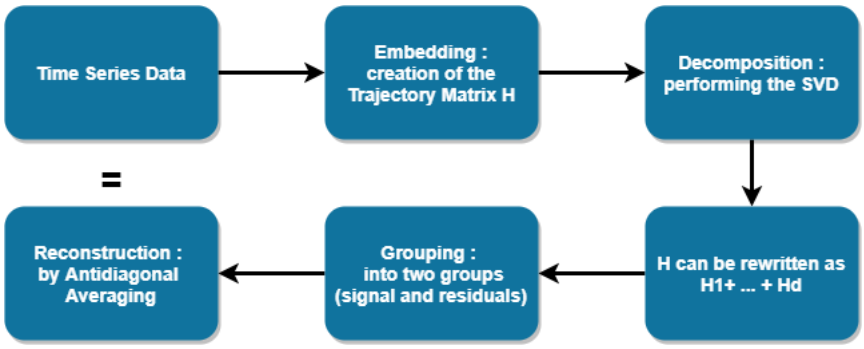
$$\boldsymbol{H} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T = \boldsymbol{H}_1 + \ldots + \boldsymbol{H}_d = \sum_{i\in I} \boldsymbol{H}_i + \sum_{i\notin I} \boldsymbol{H}_i, \tag{2.4-9}$$

where d is the rank of $\boldsymbol{H}$ and $d = max\{i; i = 1, \ldots, L|\lambda_i > 0\}$, $V_i = \boldsymbol{H}^T U_i/\sqrt{\lambda_i}$ for $i = 1, \ldots, d$ and $I \subset \{1, \ldots, d\}$. The noise-reduced series is reconstructed by $\boldsymbol{H}_I = \sum_{i\in I} \boldsymbol{H}_i$ by selecting a set of indices I.

In general, $\boldsymbol{H}_I$ is not a Hankel matrix, and the problem is bypassed by using the "anti-diagonals" mean. It is by the application of the $\boldsymbol{H}_I$ matrix that the signal s_t can be reconstructed, and it provides the decomposition of the original series y_t such that $y_t = s_t + \varepsilon_t$ with ε_t the residual series after the extraction of the signal and $t = (1, 2, \ldots, n)$. This is the reconstruction step. Then, the extracted signal is denoted by $S = (s_1, \ldots, s_n)$ and used to forecast the new data point y_{n+1}.

The SSA procedure is schematized in Figure 2.4-8, which presents the four fundamental steps of the algorithm, namely embedding, decomposition, grouping and reconstruction.

Fig. 2.4-8: Singular spectrum analysis algorithm



This method bears semblance to PCA. Indeed, the two techniques make it possible to reduce a large dataset to a smaller number of dimensions while retaining important information. They both allow signal to be separated from noise.

Example of SSA application

Before switching to the multivariate version, it is important to ensure that the SSA has been understood correctly. The example that follows is useful. Let A be a time series defined as $\boldsymbol{A} = [1, 2, 3, \ldots, 7]$. To decompose $\boldsymbol{A}$, the embedding step is performed first. Thus, suppose that $L = 3$, the embedding step is performed first. Thus, suppose that $\boldsymbol{H}$ can be written as

$$\boldsymbol{H} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 5 & 6 & 7 \end{pmatrix}. \tag{2.4-10}$$

The second step consists in the singular value decomposition of the trajectory matrix $\boldsymbol{H}$ (rounded to two decimal points) as $\boldsymbol{H} = \boldsymbol{U}\sqrt{\boldsymbol{\Sigma}}\boldsymbol{V}^t$, with

$$\boldsymbol{U} = \begin{pmatrix} -0.44 & 0.80 & 0.41 \\ -0.57 & 0.10 & -0.82 \\ -0.69 & -0.59 & 0.41 \end{pmatrix}$$
$$\boldsymbol{\Sigma} = \begin{pmatrix} 278.92 & 0 & 0 \\ 0 & 1.08 & 0 \\ 0 & 0 & 4.58.10^{-14} \end{pmatrix} \tag{2.4-11}$$
$$\boldsymbol{V} = \begin{pmatrix} -0.22 & -0.74 & 0.47 \\ -0.32 & -0.44 & -0.72 \\ -0.42 & -0.14 & 0.29 \\ -0.53 & 0.16 & -0.31 \\ -0.63 & 0.45 & 0.27 \end{pmatrix},$$

where the three eigenvalues are $\lambda_1 = 278.92$, $\lambda_2 = 1.08$ and $\lambda_3 = 4.58.10^{-14}$. Thus, the trajectory matrix $\boldsymbol{H}$ can be rewritten as $\boldsymbol{H} = \boldsymbol{H}_1 + \boldsymbol{H}_2 + \boldsymbol{H}_3$, where $\boldsymbol{H}_i = \sqrt{\lambda_i} U_i V_i^T$. From this, it is possible to deduce the following results (rounded to two decimal points):

$$\boldsymbol{H}_1 = \begin{pmatrix} 1.62 & 2.37 & 3.12 & 3.87 & 4.62 \\ 2.08 & 3.05 & 4.02 & 4.98 & 5.95 \\ 2.54 & 3.73 & 4.91 & 6.10 & 7.28 \end{pmatrix}$$
$$\boldsymbol{H}_2 = \begin{pmatrix} -0.62 & -0.37 & -0.12 & 0.13 & 0.38 \\ -0.08 & -0.05 & -0.02 & 0.02 & 0.05 \\ 0.46 & 0.27 & 0.09 & -0.10 & -0.28 \end{pmatrix}$$
$$\boldsymbol{H}_3 = \begin{pmatrix} -5.71.10^{-15} & -3.17.10^{-15} & -7.25.10^{-16} & 1.81.10^{-15} & 4.53.10^{-15} \\ 1.14.10^{-14} & 6.35.10^{-15} & 1.45.10^{-15} & -3.63.10^{-15} & -9.06.10^{-15} \\ -5.71.10^{-15} & -3.17.10^{-15} & -7.25.10^{-16} & 1.81.10^{-15} & 4.53.10^{-15} \end{pmatrix}. \tag{2.4-12}$$

The grouping step is next. Consider that $\boldsymbol{H}_1$ is one group and $\boldsymbol{H}_2$ and $\boldsymbol{H}_3$ comprise another. This means that $r = 1$. Therefore,

$$\boldsymbol{H}_{I_1} = \boldsymbol{H}_1 = \begin{pmatrix} 1.62 & 2.37 & 3.12 & 3.87 & 4.62 \\ 2.08 & 3.05 & 4.02 & 4.98 & 5.95 \\ 2.54 & 3.73 & 4.91 & 6.10 & 7.28 \end{pmatrix}, \tag{2.4-13}$$

and

$$\boldsymbol{H}_{I_2} = \boldsymbol{H}_2 + \boldsymbol{H}_3 = \begin{pmatrix} -0.62 & -0.37 & -0.12 & 0.13 & 0.38 \\ -0.08 & -0.05 & -0.02 & 0.02 & 0.05 \\ 0.46 & 0.27 & 0.09 & -0.10 & -0.28 \end{pmatrix}. \tag{2.4-14}$$

The last step is the antidiagonal averaging. Let $\boldsymbol{S}_A$ and $\boldsymbol{\varepsilon}_A$ be the result of the antidiagonal averaging of $\boldsymbol{H}_{I_1}$ and $\boldsymbol{H}_{I_2}$ respectively.

$$\begin{aligned}\boldsymbol{S}_A &= \begin{pmatrix} 1.62 & 2.22 & 2.90 & 3.87 & 4.84 & 6.02 & 7.28 \end{pmatrix} \\ \boldsymbol{\varepsilon}_A &= \begin{pmatrix} -0.62 & -0.22 & 0.10 & 0.13 & 0.16 & -0.02 & -0.28 \end{pmatrix}\end{aligned} \tag{2.4-15}$$

Finally, by summing these two antidiagonal averages, it is possible to reconstruct and retrieve the signal initially studied as follows:

$$\boldsymbol{A} = \boldsymbol{S}_A + \boldsymbol{\varepsilon}_A = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \end{pmatrix}. \tag{2.4-16}$$

Now that the univariate case has been presented, it is possible to extend the model to multivariate data.

The multivariate version

Only the organization of the trajectory matrix is different in the multivariate case. Hassani and Mahmoudvand [104] present the MSSA organized vertically (which will be presented below) or horizontally. To explain MSSA, let p be a time series of a different size n_i: $Y_{n_i}^{(i)} = \left(y_1^{(i)}, \ldots, y_{n_i}^{(i)}\right)$ with $i = 1, \ldots, p$.

MSSA consists of the same four steps:

1. **Embedding**: This step involves transforming a one-dimensional time series $Y_{n_i}^{(i)} = \left(y_1^{(i)}, \ldots, y_{n_i}^{(i)}\right)$ into a multidimensional matrix $[H_1^{(i)}, \ldots, H_{K_i}^{(i)}]$ with vectors $H_j^{(i)} = \left(y_j^{(i)}, \ldots, y_{j+L_i+1}^{(i)}\right)^T \in \mathbb{R}^{L_i}$, where $L_i\,(2 \leq L_i \leq n_i)$ is the window length for each series with length n_i and $K_i = n_i - L_i + 1$. It gives the trajectory matrix, which is defined as $\boldsymbol{H}^{(i)} = [H_1^{(i)}, \ldots, H_{K_i}^{(i)}]$. Consequently, each p series will provide a new trajectory matrix (Hankel matrix) $\boldsymbol{V}^{(i)}$ of dimensions $(L_i \times K_i)$, with $i = 1, \ldots, p$.

   To form a new (vertical) block Hankel matrix, it is necessary that $K_1 = \ldots = K_p = K$. MSSA allows various windows lengths L_i with different series lengths n_i, but similar K_i for all series. If these conditions are respected, it is possible to write the following block Hankel trajectory matrix:

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}^{(1)} \\ \vdots \\ \boldsymbol{H}^{(p)} \end{bmatrix} = \begin{pmatrix} y_1^{(1)} & y_2^{(1)} & \cdots & y_K^{(1)} \\ y_2^{(1)} & y_3^{(1)} & \cdots & y_{K+1}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ y_{L_1}^{(1)} & y_{L_1+1}^{(1)} & \cdots & y_{n_1}^{(1)} \\ \vdots & \vdots & & \vdots \\ y_1^{(p)} & y_2^{(p)} & \cdots & y_K^{(p)} \\ y_2^{(p)} & y_3^{(p)} & \cdots & y_{K+1}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ y_{L_p}^{(p)} & y_{L_p+1}^{(p)} & \cdots & y_{n_p}^{(p)} \end{pmatrix}. \tag{2.4-17}$$

2. **Decomposition**: Then the singular value decomposition is applied to $\boldsymbol{H}$. Consider $\lambda_1, \ldots, \lambda_{L_{sum}}$ the eigenvalues of $\boldsymbol{H}\boldsymbol{H}^T$ (detailed in Equation 2.4-18), organized by descending order ($\lambda_1 \geq \ldots \geq \lambda_{L_{sum}} \geq 0$) and the corresponding eigenvectors $U_1, \ldots, U_{L_{sum}}$, where $L_{sum} = \sum_{i=1}^{p} L_i$.

$$\boldsymbol{H}\boldsymbol{H}^T = \begin{pmatrix} \boldsymbol{H}^{(1)}\boldsymbol{H}^{(1)T} & \boldsymbol{H}^{(1)}\boldsymbol{H}^{(2)T} & \cdots & \boldsymbol{H}^{(1)}\boldsymbol{H}^{(p)T} \\ \boldsymbol{H}^{(2)}\boldsymbol{H}^{(1)T} & \boldsymbol{H}^{(2)}\boldsymbol{H}^{(2)T} & \cdots & \boldsymbol{H}^{(2)}\boldsymbol{H}^{(p)T} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{H}^{(p)}\boldsymbol{H}^{(1)T} & \boldsymbol{H}^{(p)}\boldsymbol{H}^{(2)T} & \cdots & \boldsymbol{H}^{(p)}\boldsymbol{H}^{(p)T} \end{pmatrix} \tag{2.4-18}$$

   The SVD of $\boldsymbol{H}$ gives $\boldsymbol{H} = \boldsymbol{H}_1 + \ldots + \boldsymbol{H}_{L_{sum}}$, where $\boldsymbol{H}_i = \sqrt{\lambda_i} U_i V_i^T$ and $V_i = \boldsymbol{H}^T U_i \sqrt{\lambda_i}$ (with $\boldsymbol{H}_i = 0$ if $\lambda_i = 0$).

3. **Grouping**: The grouping step consists of dividing the matrices $\boldsymbol{H}_1, \ldots, \boldsymbol{H}_{L_{sum}}$ into several disjoint groups and summing the matrices within each. The distribution of the set of indices $\{1, \ldots, L_{sum}\}$ into disjoint subsets $I_1, \ldots, I_m$ corresponds to the representation $\boldsymbol{H} = \boldsymbol{H}_{I_1} + \ldots + \boldsymbol{H}_{I_m}$. In the simple case, where there are only signal and noise components, two groups of indices are identified, $I_1 = \{1, \ldots, r\}$ and $I_2 = \{r+1, \ldots, L_{sum}\}$. The group $I = I_1$ is associated with the signal component, and group I_2 is associated with the noise.

4. **Reconstruction**: Here, the process aims at the transformation of the reconstructed matrix $\hat{\boldsymbol{X}}_i$ into a Hankel matrix so that it can be converted into time series. The reconstruction is then obtained through the antidiagonal averaging of each $\boldsymbol{H}_{I_1}, \ldots, \boldsymbol{H}_{I_m}$. The sum of all this antidiagonal averaging gives the approximation $\tilde{Y}_{n_i}^{(i)} = \left(\tilde{y}_1^{(i)}, \ldots, \tilde{y}_j^{(i)}, \ldots, \tilde{y}_{n_i}^{(i)}\right)$.

Forecasting and data imputation

With the functioning of SSA in the multivariate framework explained, it is important to explain how this algorithm can be used for completion before returning to the problem of missing data. In the literature, MSSA is often employed for forecasts. The basic condition for making a forecast from the SSA is that the series satisfy a linear recurrent formula. First of all, the forecasting method of Hassani and Mahmoudvand [104] is presented, and thereafter the case of missing data imputation is discussed.

The following procedure is the forecasting method based on MSSA, presented by Hassani and Mahmoudvand [104]. Therefore, a series $Y_n = (y_1, \ldots, y_n)$ checks a recurrent linear formula of order $L-1$ if $Y_n = (y_1, \ldots, y_n)$ checks a recurrent linear formula of order $L-1$ if

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_{L-1} y_{t-L+1}, \qquad t = L+1, \ldots, n. \tag{2.4-19}$$

Let U_j^{∇} be the vector of the first $L-1$ components of the eigenvector U_j, and let π_j be the last U_j component (with $j = 1, \ldots, r$). Then the recurrent SSA forecasting method allows $\hat{y}_{n+1}, \ldots, \hat{y}_{n+h}$ to be obtained through the following expression:

$$\hat{y}_i = \begin{cases} \tilde{y}_i, & i = 1, \ldots, n \\ \dfrac{\sum_{j=1}^{r}\left(\pi_j U_j^{\nabla T}\right)}{1-\sum_{j=1}^{r}\pi_j^2} \begin{bmatrix} \hat{y}_{i-L+1} \\ \vdots \\ \hat{y}_{i-1} \end{bmatrix}, & j = n+1, \ldots, n+h. \end{cases} \tag{2.4-20}$$

Example of SSA forecasting

Thus, to estimate the eighth and ninth values of series $A = [1, 2, \ldots, 7]$ in the previous example would yield the following result (rounded to two decimal points), with $L = 3$ and $r = 1$:

$$U_1^{\nabla} = \begin{bmatrix} -0,44 \\ -0,57 \end{bmatrix}, \qquad \text{and} \qquad \pi_1 = -0.69. \tag{2.4-21}$$

Thus, the Equation 2.4-20 becomes as follows:

$$\hat{y}_i = \begin{cases} \tilde{y}_i, & i = 1, \ldots, n \\ \begin{bmatrix} 0.59 & 0.76 \end{bmatrix} \begin{bmatrix} \hat{y}_{i-L+1} \\ \vdots \\ \hat{y}_{i-1} \end{bmatrix}, & j = n+1, \ldots, n+h. \end{cases} \tag{2.4-22}$$

Therefore, for $h = 1$, which is $\hat{y}_8$, the SSA gives:

$$\hat{y}_8 = \begin{bmatrix} 0.59 & 0.76 \end{bmatrix} \begin{bmatrix} \hat{y}_6 \\ \hat{y}_7 \end{bmatrix} = \begin{bmatrix} 0.59 & 0.76 \end{bmatrix} \begin{bmatrix} 6.02 \\ 7.28 \end{bmatrix} = 9.12 \tag{2.4-23}$$

In the same way, for $\hat{y}_9$ (where $h = 2$):

$$\hat{y}_9 = [0.59 \quad 0.76] \begin{bmatrix} \hat{y}_7 \\ \hat{y}_8 \end{bmatrix} = [0.59 \quad 0.76] \begin{bmatrix} 7.28 \\ 9.12 \end{bmatrix} = 11.27 \tag{2.4-24}$$

It follows that the series and its predictions are as follows:

$$A = [1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 9.12 \quad 11.27]. \tag{2.4-25}$$

A naive but obvious prediction would be that the eighth and ninth values are 8 and 9, but the prediction made by the SSA is higher. The model seems to overestimate these values. In the previous example, it was clear that A is not stationary. Thus, the procedure is repeated with strong white noise, which is stationary by definition. Let B be a strong white noise (rounded to two decimal points) as follows:

$$B = [0.59 \quad 0.71 \quad -0.11 \quad -0.45 \quad 0.61 \quad -1.82 \quad 0.63]. \tag{2.4-26}$$

The SSA algorithm is applied in the same way and with the same parameters as before (i.e., $L = 3$ and $r = 1$), giving the following eigenvalues: $\lambda_1 = 7.16$, $\lambda_1 = 1.82$ and $\lambda_1 = 1.13$, and the following decomposition $\boldsymbol{B} = \boldsymbol{S}_B + \boldsymbol{\varepsilon}_B$, with

$$\begin{aligned} \boldsymbol{S}_B &= \begin{pmatrix} -0.09 & 0.15 & -0.03 & -0.31 & 0.70 & -1.24 & 1.24 \end{pmatrix} \\ \boldsymbol{\varepsilon}_B &= \begin{pmatrix} 0.68 & 0.56 & -0.08 & -0.15 & -0.09 & -0.58 & -0.62 \end{pmatrix}. \end{aligned} \tag{2.4-27}$$

In the previous example, the main information was contained in $\boldsymbol{S}_B$, and $\boldsymbol{\varepsilon}_B$ was used as an adjustment variable. In this example, however, $\boldsymbol{\varepsilon}_B$ may have more weight than $\boldsymbol{S}_B$. For the first value of the vector B, $\boldsymbol{\varepsilon}_B$ is much closer to the value of the associated B, and $\boldsymbol{S}_B$ seems to be its fitting variable. Then, concerning the forecasts for $h = 2$, the SSA method gives

$$B = [0.59 \quad 0.71 \quad -0.11 \quad -0.45 \quad 0.61 \quad -1.82 \quad 0.63 \quad \textbf{-1.44} \quad \textbf{1.61}], \tag{2.4-28}$$

which seems to be consistent with a random walk.

What if, this time, MSSA was applied to perfectly recurring series, such as the following:

$$C = [1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1], \tag{2.4-29}$$

or even

$$D = [1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0]. \tag{2.4-30}$$

The next two forecasts for the Series C and Series D should obviously be 0 and 1. However, the MSSA method gives

$$C = [1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 0 \quad 1 \quad \mathbf{0} \quad \mathbf{1}], \tag{2.4-31}$$

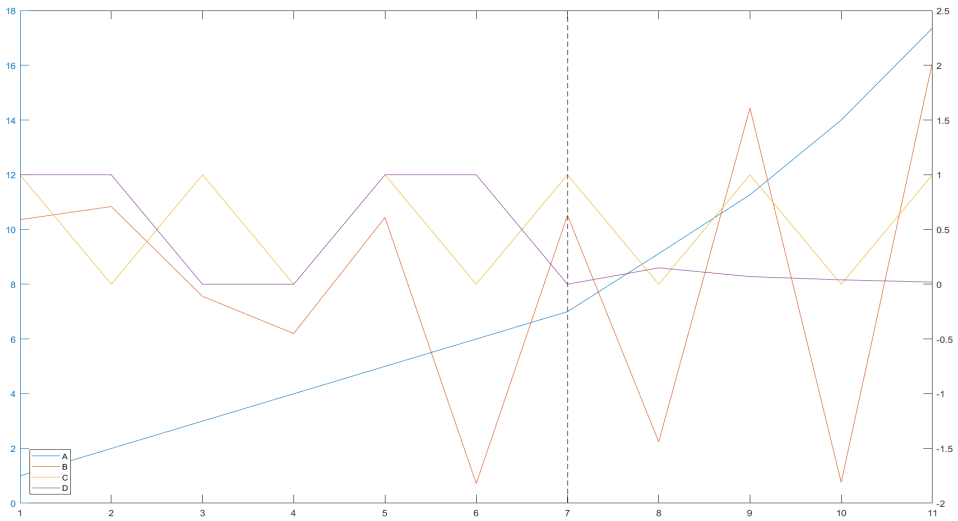
and

$$D = [1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad \mathbf{0.15} \quad \mathbf{0.07}]. \tag{2.4-32}$$

Series C is perfectly predicted, unlike Series D. The predictions for the latter are approaching 0 without reverting back to 1. The results of the forecast for $h = 4$ from the four previous examples (to the right of the dotted line) are presented in Figure 2.4-9.

This example set is akin to a Turing test, in that MSSA is tested to see if it can mimic a series. Yet it becomes obvious immediately that for the first example (Series A), there is a break between observed and forecast data: the slope of the forecasts is steeper than that of the observed data, and it is growing increasingly steeper. Conversely, in the second example (Series B), the series still resembles a random walk when the forecast data are considered. Moreover, the MSSA method estimates the future values of Series C, in which 1 and 0 alternate, perfectly. However, as soon as the pattern becomes complicated (Series D), the first forecast is relatively close to the expected value, which is 0. The following forecasts tend towards 0 and cannot return to 1 to reproduce the initial scheme.

Fig. 2.4-9: Forecasts of a stationary series (B) and a non-stationary series (A)



In the last example, the observed Series D data stops in the middle of a scheme, between two zeros, which could cause the forecast to become erroneous. However, when the scheme ends correctly, that is, when an eighth value is added to Series D and it

is 0, the forecast is not performed correctly. The forecasts for a series of the type [1 1 0 0 1 1 0 0] also tend towards 0.

Hassani and Mahmoudvand [104] explain the procedure for the h-step-ahead forecast for multivariate series. As before, for a p time series such as $Y_{n_i}^{(i)} = \left(y_{n_i}^{(i)}, \ldots, y_{n_i}^{(i)}\right)$ with a corresponding window length L_i, where $1 < L_i < n_i$ and $i = 1, \ldots, p$, the procedure is as follows: for a fixed value of K, the trajectory matrix $\boldsymbol{H}$ is written as Equation 2.4-17. Let $\boldsymbol{U}_j = \left(\boldsymbol{U}_j^{(1)}, \cdots, \boldsymbol{U}_j^{(p)}\right)$ be the j-th eigenvector of $\boldsymbol{H}\boldsymbol{H}^T$, where $\boldsymbol{U}_j^{(i)}$ with length L_i, corresponding to the series $\boldsymbol{Y}_{n_i}^{(i)}$ $(i = 1, \cdots, p)$. Moreover, consider that $\hat{\boldsymbol{H}} = [\hat{H}_1 : \ldots : \hat{H}_K] = \sum_{i=1}^{r} UiU_i^T \boldsymbol{H}$ is the reconstructed matrix that r eigentriples produce, as follows:

$$\hat{\boldsymbol{H}} = \begin{bmatrix} \hat{\boldsymbol{H}}^{(1)} \\ \vdots \\ \hat{\boldsymbol{H}}^{(p)} \end{bmatrix}. \tag{2.4-33}$$

Now, consider that $\tilde{\boldsymbol{H}}^{(i)} = \mathcal{H}\hat{\boldsymbol{H}}^{(i)}$ $(i = 1, \ldots, p)$ is the result of the Hankelization procedure of the matrix $\hat{\boldsymbol{X}}^{(i)}$ (with $\mathcal{H}$ being the Hankel operator). Then $U_j^{(i)\nabla}$ denotes the vector of the first $L_i - 1$ components of the vector $U_j^{(i)}$, and $\pi_j^{(i)}$ is the last component of the vector $U_j^{(i)}$ $(i = 1, \ldots, p)$. Thus, $\boldsymbol{U}^{\nabla p} = \left(\boldsymbol{U}_1^{\nabla p}, \ldots, \boldsymbol{U}_r^{\nabla p}\right)$, with r the number of eigentriples for the reconstruction step, with $\boldsymbol{U}_j^{\nabla p}$ is defined as follows:

$$\boldsymbol{U}_j^{\nabla p} = \begin{bmatrix} \boldsymbol{U}_j^{(1)\nabla} \\ \vdots \\ \boldsymbol{U}_j^{(M)\nabla} \end{bmatrix}. \tag{2.4-34}$$

The matrix $\boldsymbol{W}$ (of dimensions $(p \times r)$ is defined as follows:

$$\boldsymbol{W} = \begin{pmatrix} \pi_1^{(1)} & \pi_2^{(1)} & \cdots & \pi_r^{(1)} \\ \pi_1^{(2)} & \pi_2^{(2)} & \cdots & \pi_r^{(2)} \\ \vdots & \vdots & & \vdots \\ \pi_1^{(p)} & \pi_2^{(p)} & \cdots & \pi_r^{(p)} \end{pmatrix}. \tag{2.4-35}$$

If the matrix $\left(\boldsymbol{I}_{p\times p} - \boldsymbol{W}\boldsymbol{W}^T\right)^{-1}$ exists and $r \leq L_{sum} - p$, then the h-step-ahead forecast exists, and it is given by the following formula:

$$[\hat{y}_{j_1}^{(1)}, \ldots, \hat{y}_{j_p}^{(p)}]^T = \begin{cases} [\tilde{y}_{j_1}^{(1)}, \ldots, \tilde{y}_{j_p}^{(p)}], & j_i = 1, \ldots, n_i \\ \left(\boldsymbol{I}_{p\times p} - \boldsymbol{W}\boldsymbol{W}^T\right)^{-1} \boldsymbol{W}\boldsymbol{U}^{\nabla p T}\boldsymbol{Z}_h, & j_i = n_i + 1, \ldots, n_i + h, \end{cases} \tag{2.4-36}$$

where $\boldsymbol{Z}_h = [Z_h^{(1)}, \ldots, Z_h^{(p)}]^T$ and where $Z_h^{(1)} = [\hat{y}^{(i)}_{n_i - L_i + h + 1}, \ldots, \hat{y}^{(i)}_{n_i + h - 1} h]^T$ $(i = 1, \ldots, p)$. It should be noted that Equation 2.4-36 indicates that the h-step-ahead forecasts for the refined series $\hat{Y}^{(i)}_{n_i}$ are obtained by a multidimensional linear recurrent formula. For the multivariate case, there is only a one-dimensional linear recurrent formula.

Having seen how the SSA works and how it can be used to make forecasts, it is now time to present another way of using it to fill in missing data.

If a series contains missing data, the first step is to pre-impute them by using linear interpolation 2.4.1. Then, the operation is exactly the same as with a series without missing data: the four steps (embedding, decomposition, grouping and reconstruction) are executed in the normal fashion, and the missing data are obtained from the result of the fourth step. The missing data are replaced by the value obtained from the SSA. This process can be iterated in order to improve the results. In the following iterations, the SSA is applied to series in which the missing data are replaced by the results obtained in the reconstruction step from the previous iteration.

Example of missing data imputation by SSA

The first example used above can be recycled to explain how the imputation of missing data works. Let

$$A' = [1 \quad 2 \quad 3 \quad NA \quad 5 \quad 6 \quad 7], \tag{2.4-37}$$

where NA is the missing data in the example.

As explained earlier, the first step is to impute the missing data by simple interpolation. This imputation gives exactly the A series that were studied in the first example and simplifies the calculations. The four steps are then executed as in the first example, and the reconstruction step gives

$$\boldsymbol{S}_A = \begin{pmatrix} 1.62 & 2.22 & 2.90 & 3.87 & 4.84 & 6.02 & 7.28 \end{pmatrix}. \tag{2.4-38}$$

Thus, the SSA method estimates the missing data for the A' series at 3.87. This gives the following result

$$A^* = [1 \quad 2 \quad 3 \quad \mathbf{3.87} \quad 5 \quad 6 \quad 7] \tag{2.4-39}$$

A^* is therefore the result of using the SSA method to complete missing data when only one iteration is necessary. If a second iteration is needed, then the SSA method is applied to A^* and it gives:

$$\boldsymbol{S}_A = \begin{pmatrix} 1.61 & 2.20 & 2.87 & 3.83 & 4.81 & 6.01 & 7.29 \end{pmatrix} \tag{2.4-40}$$

The final result after two iterations is

$$A^{**} = [1 \quad 2 \quad 3 \quad \mathbf{3.83} \quad 5 \quad 6 \quad 7]. \tag{2.4-41}$$

Ultimately, for a large number of iterations, the results of the estimation of the missing data converge to 3.80. There is no theoretical result to ensure the convergence of the results after a large number of iterations. It is therefore necessary to see what the authors recommend or to implement a convergence test. Since the version used in this PhD thesis is that of Dash and Zhang [65], their recommendation in terms of convergence will be presented later in this section.

The MSSA method is often used for forecasting purposes in the literature. The problems of forecasting and data completion are in fact quite similar. What is sought in both cases is the best estimate of an unknown value. The imputation of missing data can be seen as a forecast of unknown data from the past. Therefore, it would be possible to use the data prediction technique to complete the data, provided that only the data preceding the gap is taken into account. Conversely, the completion technique can also be deployed to predict future values by extrapolating future data (instead of interpolating them linearly) and then applying the usual procedure.

The examples that were used previously will be repeated to compare the two methods. Thus, if the 100-iteration completion method is used for forecasting, the results for Series A, B, C and D (with two missing data points added at the end of each series) are as follows:

$$\begin{aligned}
A &= \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & \mathbf{8.81} & \mathbf{10.83} \end{pmatrix} \\
B &= \begin{pmatrix} 0.59 & 0.71 & -0.11 & -0.45 & 0.61 & -1.82 & 0.63 & \mathbf{-1.16} & \mathbf{1.12} \end{pmatrix} \\
C &= \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & \mathbf{0} & \mathbf{1} \end{pmatrix} \\
D &= \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 1 & 0 & \mathbf{0.09} & \mathbf{0.01} \end{pmatrix}
\end{aligned} \tag{2.4-42}$$

Using the imputation method to predict future data seems to yield comparable results to the forecasting method. As noted earlier, for Series A, the slope of the forecasts is higher than the slope of the observed data, but lower than when the forecasting model is used. For this series, a linear regression provides better results. The two forecasts in Series B seem to be consistent with a random walk, Series C is perfectly predictable, and Series D tends towards 0. The results appear to be similar, and perhaps a little better in view of Series A.

MSSA on financial data

SSA has been applied to empirical data in the literature, including in a paper by Rodrigues and Mahmoudvand from 2017 [170]. They compare the SSA and MSSA to

model the daily currency exchange rate data of the BRICS countries, that is, Brazil (BRL), Russia (RUB), India (IND), China (CHN) and South Africa (RAND). However, the data for Russia were impossible to obtain, and the authors were forced to exclude them from the sample. The database that they studied covers 14 years of data (from September 2001 to September 2015), with 3,516 observations for each time series. Rodrigues and Mahmoudvand use the univariate and the multivariate version of the prevision algorithm (and their modeling variants) to predict the last 35 values in the series and compare them to observed values by using the RMSE.

One of their many results is presented in Table 2.4-6. They show an RMSE that is based on 35 forecasts with a window of 60 values.

Tab. 2.4-6: RMSE based on 35 currency forecasts and L=60 for each combination of forecasting method, time series and number of steps ahead (1, 5 and 10; Source: Rodrigues and Mahmoudvand, 2017 [170])

	BRL			IND			CHN			RAND		
Method	**1**	**5**	**10**	**1**	**5**	**10**	**1**	**5**	**10**	**1**	**5**	**10**
VMSSA-V	0.1265	0.1457	0.1708	**0.3439**	**0.6248**	1.0882	0.0845	0.0898	0.0946	**0.1021**	**0.1592**	0.2856
VMSSA-R	0.1418	0.1633	0.1932	0.3812	0.6272	1.0387	0.0881	0.0917	**0.0944**	0.1293	0.1871	0.2986
HMSSA-V	0.0462	0.1067	0.1553	0.3593	0.7327	1.0345	**0.0278**	0.0703	0.0958	0.1115	0.2198	0.2923
HMSSA-R	**0.0453**	**0.1022**	0.1534	0.3571	0.6998	**1.0258**	0.0283	**0.0687**	0.0954	0.1093	0.1999	0.2834
VSSA	0.0562	0.1202	0.1719	0.4113	0.7469	1.1611	0.0284	0.0739	0.1003	0.1216	0.2511	0.2986
RSSA	0.0581	0.1102	0.1574	0.4206	0.6874	1.0676	0.0291	0.0732	0.1001	0.1153	0.1832	**0.2522**

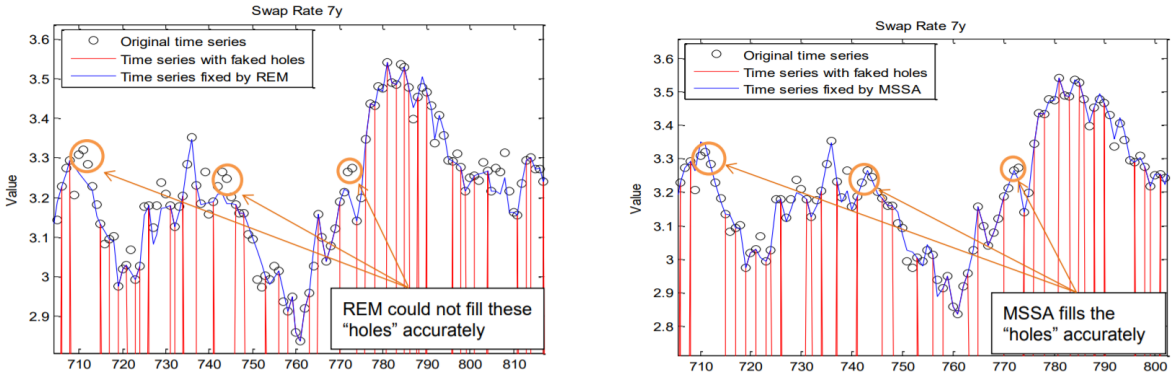
Horizontal recurrent MSSA a RMSE of Brezilian currency for 1-step ahead equal to 0.0453, that is lower than that of recurrent SSA, which is equal to 0.0581.

The results are representative of the entire analysis. They note that multivariate methods (vertical, horizontal, vector and recurrent) performed better than univariate methods (vector or recursive), except for the exchange rate of the rand. They conclude that MSSA allows co-integration between time series to be accounted for and that SSA does not. MSSA is therefore even more efficient when there is dependency between the time series.

Bloomberg's reference imputation model

Dash and Zhang [65], who are in the employ of Bloomberg, worked on the imputation of missing data in financial time series through the MSSA algorithm in 2016. In particular, they compared the MSSA imputation model with multivariate regularized EM (M-REM). Like the classical EM algorithm (see Section 2.4.6), the M-REM method is effective when the normalcy assumption is satisfied, and it relies on the iterations of linear regressions of variables with missing values on variables with available values. It uses a regularized regression method in which a regularization parameter controls the filtering of the noise in the data. Their analysis was based on USD swap rate data; 90% missing data were added artificially. Figure 2.4-10a and Figure 2.4-10b show the results from the M-REM and MSSA completions of the seven-year USD swap rate.

Fig. 2.4-10: Data reconstruction for USD swap rate 7Y (90% artificial data gaps; Source: Dash and Zhang, 2016 [65])



(a) M-REM data reconstruction

(b) MSSA data reconstruction

Figure 2.4-10a shows that M-REM completes missing data poorly, with many missing data. The quality of the MSSA reconstruction is much higher (Figure 2.4-10b). Dash and Zhang's [65] explanation is that the swap rate data was not normally distributed. In Table 2.4-7, they compare the two methods with the original series by reference to the bases of mean average error (MAE), mean daily volatility, mean daily correlation, mean 95% distribution tail and mean 5% distribution tail. The performance of MSSA is significantly better than that of the M-REM in all of the comparisons in Table 2.4-7. MSSA reconstructs the time series while maintaining volatilities and tails. It performs much better than the M-REM.

Tab. 2.4-7: M-REM vs MSSA data gap filling on USD swap rate with 90% gaps (Source: Dash and Zhang, 2016 [65])

Metrics	**M-REM algorithm**	**MSSA algorithm**	**M-REM vs MSSA relative difference**
Average MAE (bp)	1.75	1.07	64%
Average daily vol diff (%)	0.96	0.36	167%
Average daily correlation diff	0.03	0.02	50%
Average daily right tail diff (%)	1.59	0.62	156%
Average daily left tail diff (%)	1.36	0.48	183%

Multivariate regularized EM obtain MAE equal to 1.75 basis points while that of MSSA is equal to 1.07 basis points, in other words, the MAE from Multivariate regularized EM is 64% higher than that from MSSA.

They conclude that MSSA is more efficient than M-REM because it is a non-parametric method that does not require a normalcy assumption. Empirically, it provides better results in filling in missing data because it preserves not only the dynamics of a series but also its volatility, correlation and distribution. Since Bloomberg use

MSSA data imputation for their entire client base, the remainder of this PhD thesis is premised on it rather than on the forecasting method.

Choice of inputs: raw prices or price returns

Dash and Zhang [65] apply the MSSA algorithm to raw prices rather than price returns. The authors' explanation is that applying MSSA to prices works better than applying it to returns because the rates have a memory that is captured by the time lag. The authors also point to their experience, but do not provide additional justifications. Moreover, while presenting their model in 2016, Dash and Zhang [65] spoke about a "lag-covariance matrix" decomposition. The MSSA algorithm is based on the decomposition of the trajectory matrix and the computation of the eigenvectors of this matrix in particular. However, decomposing the trajectory matrix is the same as decomposing the variance-covariance matrix of the trajectory matrix. Returning to the choice of inputs, the correlations and the autocorrelations are generally much higher in series of raw prices than in series of price returns. This explains the superior performance of MSSA when it is applied to raw prices rather than returns. During the discussions between Bloomberg and Natixis staff, a data standardization option was discussed with a view to optimizing the performance of the algorithm. According to Bloomberg, the algorithm must be applied to standardized raw prices. Therefore, MSSA was tested on crude oil prices and then on crude oil price returns in order to identify the differences.

MSSA imputation based on WTI and Brent data: price versus return

The sample that is used here covers WTI crude oil daily prices and Brent crude oil daily prices over a period of five years, from August 30, 2015, to July 31, 2020. These data originate from the EIA, the same source that was used in Section 1.1.1. The EIA is an independent statistical agency founded in 1977. It is part of the U.S. Department of Energy, and its objective is to provide independent data, forecasts and analyses to enable governments to make appropriate policy decisions. The spot price data used by the EIA, which can be downloaded from their website, is from Thomson Reuters.

Brent is extracted from the North Sea in Europe, while WTI oil is extracted in North America. In terms of quality, WTI is slightly better than Brent (even though both are of very good quality), which explains the slightly higher price of Brent in the past. However, since 2017, the price of Brent has often exceeded that of WTI oil due to the high levels of U.S. oil inventories that put pressure on the price of WTI.

Figure 2.4-11a and Figure 2.4-11b show the price and returns series for WTI and Brent, respectively. It is evident from Figure 2.4-11a that the price series for Brent and WTI remained very close during the first two periods and diverged gradually over the third and the fourth period before converging again during the last one. In general, the dynamics of the two series are similar. The correlations between them for the first four

periods are over 72%, 69%, 50% and 70%. The correlation in the last period is only 42%.

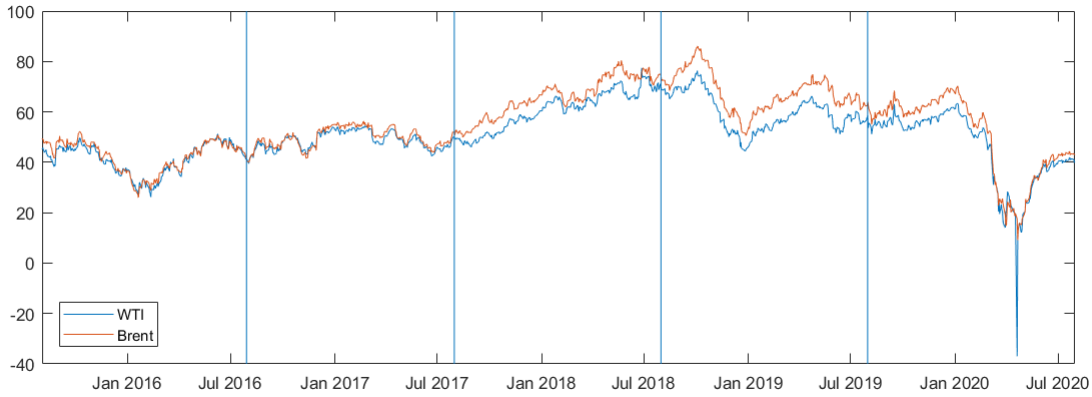
One of the characteristics of crude oil prices over the fifth period, and of WTI prices in particular, is that they became negative for one day. This can be explained by the coincidence of three events: in March 2020, Saudi Arabia and Russia decided to increase their oil production in order to reduce prices. Then, a drop in global demand due to the confinement of the population during the Covid-19 crisis led to a further drop. Finally, prices fell due to storage capacity becoming overstretched, forcing producers to pay to dispose of their oil on April 20, 2020.

On April 17, WTI was trading at \$18.31. On April 20 it dropped to −\$36.98, and the next day, April 21, it went back up to \$8.91. This represents a decrease of \$55.29 between April 17 and 20. However, as soon as prices turn negative, the use of relative returns no longer makes much sense. Here, the relative return between April 17 and April 20, when positive prices became negative, was −302%, while between April 20 and April 21, when negative prices became positive, the relative return was −124%. Between the latter dates, prices increased. Logically, returns should have been positive. However, because of the negative price, they are negative. Applying an automatic process to relative returns from WTI prices over this period would have a catastrophic impact in risk measure terms. The process would identify two periods of very strong decline. In reality, only one such period occurred.

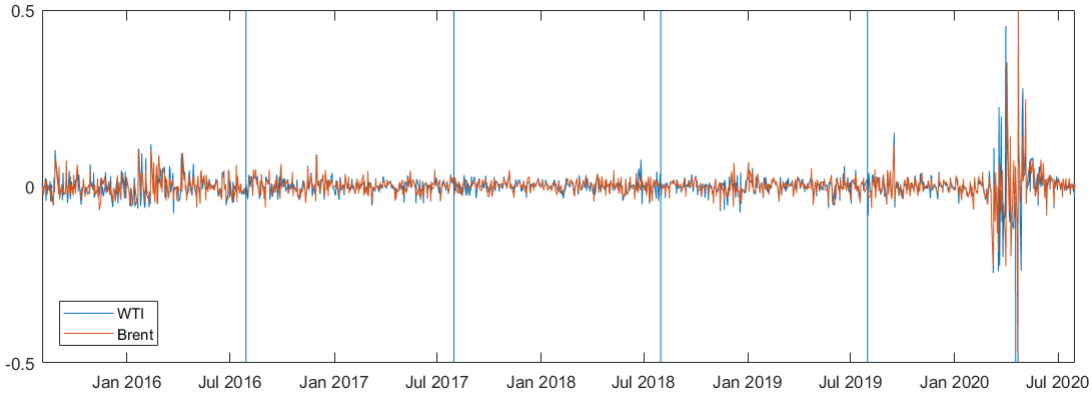
Generally, when negative prices occur, absolute returns can replace relative returns (corresponding to the price variation between the two dates) to circumvent this kind of problem. Here, the price variation between April 17 and 20 was $\$ - 55.29$, and the price variation between April 20 and 21 was \$45.89, a positive figure. However, if the relative returns method has always been applied because prices have never been negative, it is very laborious to repeat the entire calculation and to review the analyses, risk measures, decisions and so forth that result from it. Many hours of overtime would have to be expended on the task. Since negative prices remain exceptional, special rules are in place to compensate for such market anomalies. Compensation can be actuated by setting a minimum price threshold of 0 (or a value very close to 0) or by adjusting relative returns to be positive.

Here the relative returns are left unchanged because they concern only one value. It can be assumed that this value would have little impact on the analysis. After verification, the effect of the descriptive statistics would be negligible if the false negative return is replaced by an absolute value. For this reason, the rest of the analysis is executed without any exception rule for the negative value.

Fig. 2.4-11: WTI and Brent crude oil spot prices and returns between August 2015 and July 2020



(a) Spot prices



(b) Price relative returns

Unsurprisingly, returns are largely contained within a range of -5% to $+5\%$, except for the last period, when relative returns exploded, ranging from -301% to $+53\%$. In the main, this development was observed for WTI oil (see Figure 2.4-11b). Table 2.4-8 shows clearly that the standard deviation and thus the volatility of the WTI series is comparable to that of the Brent series over the first four periods, but this is not the case for the last period, in which the volatility of the WTI series is almost four times higher than that of Brent.

Moreover, when the analysis proceeds to the third and fourth moment, it transpires that the skewness of the WTI series is generally lower than that of the Brent series (except for the third period). Unsurprisingly, the skewness of the WTI series in the

second period is much lower than that of the Brent series and that in the other periods, reflecting a left-spreading tail distribution, which is attributable to numerous negative returns. The kurtosis coefficients reveal an overall mesokurtic distribution, that is, a close-to-normal one (because it is close to 3), except for the last period. The distributions of the two series in the last period can be qualified as leptokurtic, meaning that their extremities are larger than normal, that is, that extreme values are more frequent. This tendency is more pronounced in the WTI series than in the Brent series.

Tab. 2.4-8: First four moments of WTI and Brent crude oil returns

Sample Period	**Mean**		**Standard Error**		**Skewness**		**Kurtosis**	
	WTI	**Brent**	**WTI**	**Brent**	**WTI**	**Brent**	**WTI**	**Brent**
08.15 - 07.16	0.000	-0.000	0.033	0.030	0.699	0.0748	4.067	4.098
08.16 - 07.17	0.001	0.001	0.020	0.021	0.203	0.457	4.584	4.709
08.17 - 07.18	0.002	0.002	0.017	0.016	0.062	-0.174	4.822	3.268
08.18 - 07.19	-0.001	-0.000	0.020	0.020	-0.635	-0.193	4.546	4.547
08.19 - 07.20	-0.013	0.001	0.220	0.070	-10.953	0.606	146.427	25.087

Now that the sample has been described, it is appropriate to use MSSA to determine whether it is preferable to use raw prices or price returns. First of all, a listwise deletion (see Section 2.2.2) is performed on the data to remove observations where at least one of the two quotations is missing (see Figure 2.4-11a). Thereafter, the sample is divided into five subsamples so that one-year periods can be analyzed.

In order to observe the impact of the inputs on the MSSA model, 30% of missing data were introduced at random in each sub-sample of the WTI spot price series. Moreover, 100 scenarios containing 30% missing prices (uniformly distributed) were created in order to draw conclusions from multiple results. The data sample is the same, only the missing data location is different from one scenario to another. On average, for each sub-sample, 75 missing data points (prices) were added to the WTI price series among the 100 scenarios. These missing prices correspond, on average, to 127 missing returns (varying between 115 and 135, depending on the missingness scenario). An imputation method using price returns imply handling a higher proportion of missing data. This gives an advantage to methods that impute price series (such as interpolation methods), as they have to impute less missing data. The aim of the example is to demonstrate the superiority of the MSSA algorithm when it is applied to a series of raw prices rather than price returns. To do so, the comparison is made between the returns of the series where the prices were imputed (MSSA applied to price series) and those of the series where the returns were imputed (MSSA applied to return series). For the purposes of the comparison, the MSSA method will be called *PriceMSSA* when it is applied to prices and *ReturnMSSA*when it is applied to returns. In both cases, the data are not standardized so that only the impact of the input can be observed. The MSSA applied

here uses a length window of 60 and a rank of 60. The choice of parameters remains arbitrary. The problem is discussed in more detail at the end of this section. Finally, the two methods are iterated 10 times before the final results are obtained. No scaling procedure is used.

The methods are compared by using the mean absolute error (MAE) measure, which is computed for each of the 100 scenarios. The MAE enables a comparison of the difference between the returns from the original series and the returns from the completed series. Accordingly, MAE_{Price} corresponds to the average of the absolute differences between the returns from the original series and the returns from the imputed prices that are obtained by the use of the *PriceMSSA* method. In the same way, the MAE_{Return} measure corresponds to the average of the absolute differences between the returns from the original series and the imputed returns obtained by the use of the *ReturnMSSA* method.

The results obtained for the MAE criterion are presented in Table 2.4-9. $\overline{MAE}_{Price}$ corresponds to the mean of the 100 MAE_{Price} and $\overline{MAE}_{Return}$ corresponds to the mean of the 100 MAE_{Return}.

Tab. 2.4-9: Average MAE of returns from *PriceMSSA* and *ReturnMSSA* imputations

Sample Period	$\overline{\boldsymbol{MAE}}_{Price}$	$\overline{\boldsymbol{MAE}}_{Return}$
08.2015 - 07.2016	0.96%	1.37%
08.2016 - 07.2017	0.50%	0.79%
08.2017 - 07.2018	0.54%	0.76%
08.2018 - 07.2019	0.60%	0.83%
08.2019 - 07.2020	4.46%	4.22%

Cursory scrutiny reveals the *PriceMSSA* method is more suitable than the *ReturnMSSA* method for reconstructing missing data. The averaged MAE that is based on the *PriceMSSA* results is lower than the one that is based on the *ReturnMSSA* results, which means that the *PriceMSSA*reconstruction is further removed from the original series than the *ReturnMSSA* one. Over the first four periods (08.2015 to 07.2016, 08.2016 to 07.2017, 08.2017 to 07.2018 and 08.2018 to 07.2019), MAE_{Price} is always lower than the MAE_{Return} and 100% of the scenarios give lower MAEs for MAE_{Price}. Here, using returns as inputs leads to an MAE that is higher by around 40% (43%, 58%, 41% and 38% for the first four periods, respectively).

In the last period (08.2019 to 07.2020), MAE_{Return} is almost equal to MAE_{Price} on average. It is even a little lower. The *PriceMSSA* method does not always give the best results: in the last period, only 46% of the scenarios give lower MAEs for the method that is applied to spot prices. Therefore, the MSSA that is applied to returns

yields better imputations in 54% of cases. It follows that applying the MSSA algorithm systematically to series of spot prices is not always the best solution.

The same logic was reproduced, with the RMSE criterion, a quadratic measure, in order to observe the impact of extreme values and outliers. $RMSE_{Price}$ corresponds to the root of the mean squared differences between the returns from the original series and the returns from the imputed prices that were obtained by the *PriceMSSA* method. Likewise, the $RMSE_{Return}$ measure corresponds to the root of the mean squared differences between the returns from the original series and the imputed returns obtained through the use of the *ReturnMSSA*.

The results for the RMSE criterion, are presented in Table 2.4-10. $\overline{RMSE}_{Price}$ corresponds to the mean of the 100 $RMSE_{Price}$ and $\overline{RMSE}_{Return}$ corresponds to the mean of the 100 $RMSE_{Return}$.

Tab. 2.4-10: Average RMSE of returns from *PriceMSSA* and *ReturnMSSA* imputations

Sample Period	$\overline{\boldsymbol{RMSE}}_{Price}$	$\overline{\boldsymbol{RMSE}}_{Return}$
08.2015 - 07.2016	1.74%	2.66%
08.2016 - 07.2017	0.88%	1.57%
08.2017 - 07.2018	1.01%	1.60%
08.2018 - 07.2019	1.10%	1.58%
08.2019 - 07.2020	21.61%	17.75%

The conclusion for the RMSE criterion is the same: for the first four periods, the average RMSE obtained with the *PriceMSSA* method is systematically lower than that obtained with the *ReturnMSSA* method. The results obtained with the *ReturnMSSA* method give results that are about 50% higher than those obtained with the *PriceMSSA* method, except for the second period, when the results are approximately five times higher. The results for the first four periods with the *PriceMSSA* method are 53%,78%,58% and 44% higher, respectively, than those from the application of the *ReturnMSSA* method. Moreover, 100% of the scenarios yield a lower RMSE for *PriceMSSA* than for *ReturnMSSA*, except in the third period, when one scenario gives a better result for *ReturnMSSA* method. The results for the first four periods therefore indicate that using spot prices as inputs is always the optimal solution.

In the last period, it is possible to observe the same pattern as when the MFA criterion was used, that is, the *PriceMSSA* method commands a small advantage. The spread between the average RMSE for both methods in the last period is smaller than in the first four periods, and 51% of the scenarios yield a better RMSE for *PriceMSSA* than for *ReturnMSSA*. Therefore, applying the MSSA algorithm to spot prices does not

seem to always be the best methodology to follow, especially when volatility is high. The presence of extreme values makes the algorithm less efficient in general, and it has a slight disadvantage when applied to series of spot prices.

In order to ensure that the negative price observed in the WTI series has no impact on the results, the same calculations were made by replacing the false return with its absolute value (−124% is replaced by 124% as may have been done earlier). The results obtained with both methods are very close to the previous ones. The average MAE of the MSSA that is applied to price returns is a little higher but still lower than that of the MSSA that is applied to spot prices (in the last period, MAE_{Return} increases to 4.29%). The same pattern is observed in respect of the RMSE: the *ReturnMSSA* method obtains a higher average RMSE than before (18.45% upon correcting the false return). This means that the false return caused by the negative prices had very little impact on the imputation results.

That MSSA runs into difficulties in the last period is undoubtedly due to the strong movements of the market. The last period includes the Covid-19 crisis, a period of high volatility that included negative WTI prices on April 20, 2020 (see Figure 2.4-11a). As noted previously, the annual volatilities in the first four periods are between 1.6% and 3.3% (for both series). In the last period, they are 22% and 7% for WTI and Brent, respectively (see Table 2.4-8). In this period, there were some scenarios in which MSSA fit the original series better when it was applied to a returns series than when it was applied to raw prices. Finally, the most important observation that emerges from this example is that an automatic process can have negative consequences for data quality. Oil prices had never been negative before April 20, 2020, so it is very likely that automatic processes had never been confronted with negative prices. As a result, data quality may suffer from the illogical use of relative returns. Moreover, there is no reason to believe that *PriceMSSA* is always a better solution than *ReturnMSSA*. In a specific sample (or a specific scenario), it is perfectly possible that applying MSSA to returns is better than applying it to raw prices. The advantages of the application of MSSA to prices become less pronounced when a series is volatile, when the period of observation is long and when it includes jumps. In the last analysis, the choice of prices as an input for MSSA is not obvious. A preliminary study is likely necessary. Since Bloomberg apply MSSA to prices to fill in missing data for their clients, however, it is used in the rest of the PhD thesis.

Optimal choice of length window and rank

MSSA is complex and difficult to understand. Its performance is also extremely sensitive to the choice of its two parameters: the length of the window L and the number of eigenvalues used for reconstruction r.

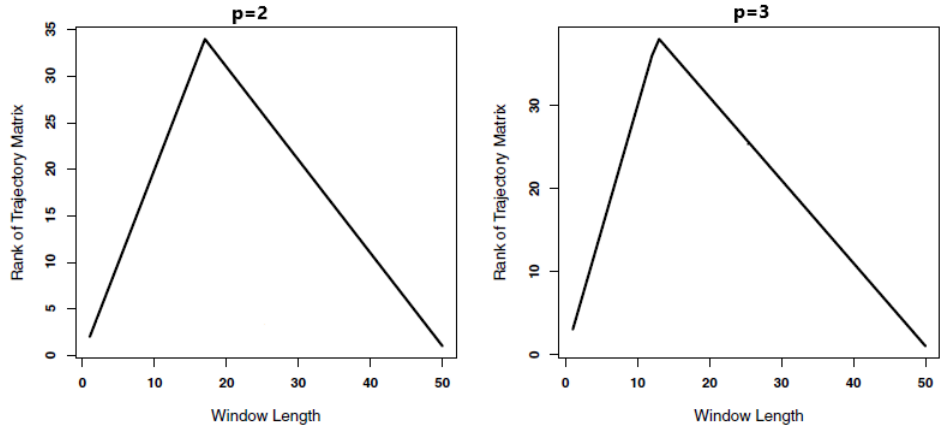
First of all, the usual bound for L, is $2 \leq L \leq n-1$, which enables the construction of a trajectory matrix. In 1996, Elsner and Tsonis [74] found that, in practice, setting $L = n/4$ is relatively common. That a method is commonly applied in a certain way does not indicate optimality. In their 2001 book, Golyandina, Nekrutkin and Zhigljavsky [94] recommended an L that is large enough but no more than $n/2$. Once again, this was only a recommendation, and it has no theoretical basis. It was not until 2011 that Hassani, Mahmoudvand and Zokaei [105] showed that the optimal value of the parameter, at least for the reconstruction step, is $(n+1)/2$. This result is applicable only to the univariate model. Rodrigues and Mahmoudvand [170] explain that applying SSA to multivariate series is much more complicated than using MSSA directly because one univariate model is applied for each time series. Therefore, a pair of parameters must be chosen for each univariate model. In the case of MSSA, only one L and one r must be chosen.

Regarding the multivariate algorithm, Rodrigues and Mahmoudvand [170] point out that finding the optimal L in the case of MSSA is much more complicated, even more so when the series are not of the same length. However, in 2013, Hassani and Mahmoudvand [104] presented, what they saw as the optimal value of L in a multivariate framework. In their view, the optimal value of L is the one that produces the trajectory matrix with maximum rank. They assume that all the p series have the length n and show that the trajectory matrix, which has the dimensions $(pL \times K)$, is of maximum rank when $L = \frac{n+1}{p+1}$. Let d denote the rank of $\boldsymbol{H}$. Then, the maximum value of d is obtained as follows:

$$\begin{aligned} max(d) &= \max_{L \in \{2,\ldots,n-1\}} min(pL, n-L-1) \\ &= \max_{L \in \{2,\ldots,n-1\}} \frac{L(p-1)+n+1-|L(p+1)-(n+1)|}{2} \end{aligned} \tag{2.4-43}$$

The equation shows that the maximum rank of $\boldsymbol{H}$ is reached when $|L(p+1)-(n+1)|$ is minimized. Hence $L = \frac{n+1}{p+1}$. However, working from a maximum rank trajectory matrix allows one to retain as much detail as possible during the decomposition step. Hassani and Mahmoudvand [104] illustrates their result by the following example: if $n = 50$, then the value of L that maximizes the rank of the trajectory matrix is $L = \frac{50+1}{2+1} = 17$ for $p = 2$ and $L = \frac{50+1}{3+1} \approx 13$ for $p = 3$. These results are presented in Figure 2.4-12 and show that Hassani and Mahmoudvand's optimal L does maximize rank.

Fig. 2.4-12: The rank of the trajectory matrix with respect to different values of L and p ($p = 2$ on the left-hand side and $p = 3$ on the right-hand side) for two datasets of lengths $n = 50$ (Source: Hassani and Mahmoudvand, 2013 [104])



For a sample with 50 observations and 2 columns ($n = 50$ and $p = 2$), the value of L that maximizes the rank of the trajectory matrix is 17.

Finally, in their presentation of the application of MSSA to financial data, Dash and Zhang [64] described window size as the longest periodicity of the time series that the model captures. They also advise an $L = 60$ for a daily time series, equivalent to a quarter, in order to account for the trend of seasonality as well as possible.

The choice of r follows that of SSA and MSSA because the r parameter only affects the number of eigenvalues involved in the calculation. In the univariate case, if $r = L$, all eigenvalues are used, and the reconstruction is equivalent to the original series. However, if $r < L$, then the reconstruction is likely to be less noisy because the small eigenvalues typically contain the noise. Similarly to the L parameter, there is no definite theory about the choice of r, only a few lines of research. Golyandina and Zhigljavsky [95] make general recommendations for separating noise from the signal appropriately. Notably, they suggest generating one-dimensional graphs of eigenvalues as well as two-dimensional plots of successive eigenvectors. However, they proffer no definitive procedure for deducing the number of values that should be used for the reconstruction of the data. Dash and Zhang [65], for their part, are of the view that the value to be retained for their reconstruction must be at least equal to p. For this reason, $r = p$ will be used hereafter. Other methods could be used to determine the rank. In particular, it would be possible to analyze a scree plot (graph of the eigenvalues of the principal component). Thus, by a scree test it is possible to deduce the number of principal component (i.e. the rank) by detecting an elbow. Cross-validation could also be used here to decide the number of principal component. Josse and Husson [126]

present a generalized leave-one-out cross-validation method used to determine the number of principal component of the multiple imputation with PCA. This cross-validation procedure is presented in more detail in Section 2.4.8.

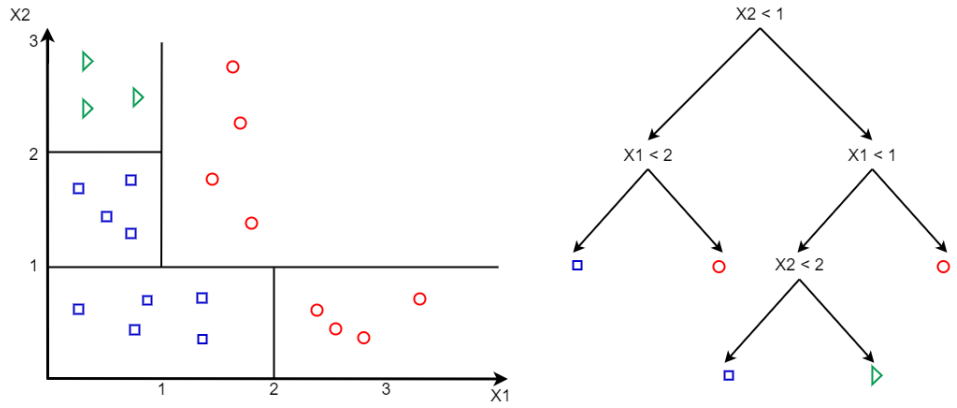
2.4.5 Random forests

The multidimensional random forests algorithm, developed by Breiman [42] in 2001, can also be used. It is a non-parametric method and a machine learning algorithm that is based on regression trees. It is particularly useful because it can be applied to both continuous and categorical data. Random forests entail the joint application of decision trees and bagging (bootstrap aggregating). Breiman [41] developed the bagging concept in 1996. He then applied decision trees, thus creating the random forests algorithm.[42]

Decision Tree Construction

To understand how random forests work, one must first grasp decision trees. A decision tree, used for the first time by Morgan and Sonquist [158],], is based, as its name would imply, on the image of a tree that is composed of a root and nodes (where branches divide) as well as branches and leaves. A decision starts at the root node, moves from node to node through the branches and ends in a leaf. It works as a cutting recursive algorithm where each slicing is parallel to the axes. In 1984, Breiman [44] introduced the classification and regression trees (CART) statistical method, constructing tree-based predictors for both regression and classification. Figure 2.4-13 shows schematically how a decision tree works with two predictor variables $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. The first split is of $\boldsymbol{X}_2$, which may be less than 1 or not. If the answer is affirmative, the left split of $\boldsymbol{X}_1$ is greater than 2. If the answer is negative, the right split of $\boldsymbol{X}_1$ is greater than 1. It is possible to proceed down the tree in this manner, a top-down approach. For example, it is possible to see that if $\boldsymbol{X}_2 > 1$ (right branch of the origin node), then $\boldsymbol{X}_1 < 1$ (left branch from the second node) and that if $y > 2$ (right branch of the third node), then the data are in the green triangle class (terminal leaf).

Fig. 2.4-13: Example of simple decision tree



The general principle of CART is to partition space recursively into two in order to find a sub-partition that is optimal for prediction. To this end, it is necessary to first build a maximal tree. That tree must be pruned to achieve optimality. In the case of imputation by random forests, the pruning phase is not necessary. For this reason, the exposition focuses only on the maximal tree construction phase.

Formally, let $(\boldsymbol{X}, \boldsymbol{Y})$ be a couple of random variables, with $\boldsymbol{X}$ a matrix with dimensions $(n \times p)$ also called an input matrix, and $\boldsymbol{Y}$ a vector with n observations, which is called the output vector. Let $\mathcal{L}_n = \{(\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)\}$ be a sample of n independent and identically distributed observations, also called the learning set. A decision tree result in a classifier (predictor) $h(\boldsymbol{X}, \mathcal{L}_n)$.

In general, the construction of a maximal tree starts from a root node (Level 0), here noted as g, which contains the set of indices $i \in \{1, \ldots, n\}$. It is divided into two new nodes g_L (the left node) and g_R (the right node) in such a way that $g_L \cup g_R = g$. Therefore, Level 1 consists of two nodes. Subsequently, each new nodes is divided into two. At Level 2, there are four new nodes (the two nodes of Level 1 gives rise to two nodes each at level 2). The process continues until the stop criterion is fulfilled. If the stop criterion of the tree is reached at level $Q \in \mathbb{N}$, then that level will contain 2^q nodes and the maximum tree is composed of $\sum_{q=0}^{Q} 2^q$ nodes. Their union is equal to $\{1, \ldots, n\}$. As said previously, the tree construction continue until reaching the maximal tree (which means one observation in each terminal leaves).

What remains to be described is the division of the nodes. The explanation that follows focuses on the split of the root node g into two new nodes g_L and g_R, but it can be generalized and applied to all nodes in a tree. Beginning with the root node g, the splitting method used in maximal trees aims to find the (k, d) pair that minimizes the intra-group variance of $\boldsymbol{Y}$, where k is the k-th column of $\boldsymbol{X}$ and $d \in \mathbb{R}$ such that $g_L = \{i \in g | \boldsymbol{X}_{i,k} \leq d\}$ and $g_R = \{i \in g | \boldsymbol{X}_{i,k} > d\}$.

Moreover, if the variance of a node g is defined as the variance of $\boldsymbol{Y}_i$, for all i present in node g

$$\mathbb{V}_g(\boldsymbol{Y}) = \frac{1}{\#g}\sum_{i\in g}(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}}_g)^2, \tag{2.4-44}$$

then the intra-group variance to be minimized is defined as follows:

$$\begin{aligned}\mathbb{V}_g^{intra}(\boldsymbol{Y}) &= \frac{1}{n}\sum_{i\in g_L}(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}}_{g_L})^2 + \frac{1}{n}\sum_{i\in g_R}(\boldsymbol{Y}_i - \bar{\boldsymbol{Y}}_{g_R})^2 \\ &= \frac{\#g_L}{n}\mathbb{V}_{g_L}(\boldsymbol{Y}) + \frac{\#g_R}{n}\mathbb{V}_{g_R}(\boldsymbol{Y}),\end{aligned} \tag{2.4-45}$$

where $\bar{\boldsymbol{Y}}_{g_L}$ and $\bar{\boldsymbol{Y}}_{g_R}$ are the means of the $\boldsymbol{Y}_i$ observations for all $i \in g_L$ and all $i \in g_R$, respectively. Lastly, the splitting method consists in finding k and d such that $\mathbb{V}_g^{intra}(\boldsymbol{Y})$ is minimized.

As noted previously, once the root node has been divided, the same operation can be repeated on each of the new nodes. The tree is thus extended until the stopping criterion is met. In the case of continuous variables, that criterion is tied to the number of observations that are left in each terminal node. The intra-group variance minimization step is similar to the ascending hierarchical classification method. In that method, the number of classes retained at the end of the process is determined by minimizing intra-group variance or by maximizing inter-group dispersion. However, the similarity between the two methods is superficial. The construction of a decision tree begins with the whole set of data, which is divided into increasingly smaller groups. The ascending hierarchical classification method proceeds from individual pieces of data and gradually groups them into larger groups. In addition, the intra-group variance minimization step is a partition criterion in decision trees. In ascending hierarchical classification, it is only a simple indicator of partition quality. In the ascending hierarchical classification method, partitions are based exclusively on a distance criterion.

Random forests construction

Random trees involves fitting trees to bootstrapped samples and combining them through averaging. In other words, the random forests model entails taking B bootstrap samples from the original data, growing a tree on each of these datasets and averaging all the results. The random forests algorithm is akin to the bean machine, also known as the Galton board, that is used to demonstrate the central limit theorem. Each of the values in the database passes through several decision trees before reaching its terminal node, and the distribution of the data is deduced by averaging the obtained results. The difference is that the data circulate randomly in the Galton board; in random forests, they respect an intra-group variance minimization criterion.

Let $\mathcal{L}_n$ be the learning set (introduced previously, in the decision tree presentation), B the number of tree in the forest, and k the number of variables (columns) used for the tree construction, where $1 \leq k \leq p$ (p is the number of columns of $\boldsymbol{X}$). According to Breiman [42], a random forests is a classifier consisting of a collection of tree-structured classifiers $\left\{\hat{h}_1(\boldsymbol{X}, \hat{\boldsymbol{X}}_1, \Theta_1), \ldots, \hat{h}_B(\boldsymbol{X}, \hat{\boldsymbol{X}}_B, \Theta_B)\right\}$ where $\hat{\boldsymbol{X}}_b$ (for each $b = 1, \ldots, B$) is the learning set from $\mathcal{L}_n$ chosen by the law Θ_b. The $\{\Theta_b\}$, for all $b = 1, \ldots, B$, are independent identically distributed random vectors.

More formally, for each $b = 1, \ldots, B$, the random forests algorithm works in three steps:

1. Drawing an independent bootstrap samples $\hat{\boldsymbol{X}}_b$ from $\mathcal{L}_n$ chosen by the law Θ_b;

2. Constructing a decision tree $h_b(\boldsymbol{X}, \hat{\boldsymbol{X}}_b, \Theta_b)$. For each splitting node, k variables are randomly selected in order to calculate the best split (as presented previously). The tree is constructed until there is only one observation left in the terminal leaves (maximal tree).

3. Adding the classifier to the collection $\left\{\hat{h}_1(\boldsymbol{X}, \hat{\boldsymbol{X}}_1, \Theta_1), \ldots, \hat{h}_B(\boldsymbol{X}, \hat{\boldsymbol{X}}_B, \Theta_B)\right\}$.

Thus, this algorithm includes two levels of randomness: the first one is due to the random vector Θ_b (for all $b = 1, \ldots, B$) and the other one if k the number of variable used into the tree construction.

Then, a random forests consists in the aggregation of the classifiers from each tree. Its predictor is obtained by averaging these classifiers.

Once the step of building a decision tree is understood, the algorithm is relatively simple to conceptualize, visualize and interpret. A decision tree is a white-box method because it is simple to comprehend. However, in the case of random forests, many trees are taken into consideration. They may comprise hundreds or even thousands of nodes, which can make the results less easy to interpret. For this reason, among others, random forests is still considered a black box models. Another disadvantage of random forests is that each tree must be developed to its maximum (until only one observation remains in each leaf), which can prolong computation. Calculation time also increases with the number of trees that are chosen for the forest.

The lack of theoretical results remains the biggest disadvantage of the method. Breiman [42] shows that the practical results of bagging are satisfactory, but they are not based on theory, and there is no expected upper bound of the generalization error of the forests in terms of correlation or of the strength of individual trees.

Biau, Devroye and Lugosi [33] work has made random forests consistent, on Breiman's [42] definition. It was only in 2015 that Scornet, Biau and Vert [186] demonstrated the l_2 consistency of random forests, the first theoretical guarantee of its efficiency.

The bagging procedure and the general aspects of random forests make the method relatively unclear, similarly to other machine learning methods. In banking and financial regulation, machine learning methods (such as neural network methods) are generally not received well because they are often associated with the notion of a black box. That notion does not suit the goal of modelling and explaining every risk.

Existent theoretical work does not make random forests completely transparent and interpretable. This motivated Bénard, Biau, Da Veiga and Scornet [30]to develop the SIRUS algorithm in 2021. SIRUS is an interpretable version of random forests that is based on a simple structure, stable behavior and results that are just as accurate as those of the other versions of random forests. SIRUS enables the use of random forests algorithms whose theoretical framework is fully understood. SIRUS also uses bagging. However, by aggregating the forests structure rather than predictions, it stabilizes the prediction rule. This algorithm, being very recent, could not be applied in this PhD thesis. Nevertheless, it would be interesting to adapt it to data imputation in order to compare its results with those obtained in the next chapter. Its authors assert that it is transparent and interpretable, which corresponds to the goals of regulators.

Bagging

Now that the random forests algorithm is presented, it is interesting to discuss about the bagging procedure. As noted previously, the random forests algorithm is based on a method called bagging (short for bootstrap aggregating), that was introduced by Breiman [41] in 1996. Bagging was developed before random forests: the latter is a combination of the bagging technique and decision trees. The bagging method entails applying a prediction method to independently drawn bootstrapped samples with replacement to obtain predictors. Then, the results are aggregated. The technique was originally created for decision trees, but it can be used with other prediction methods. Notably, Biau, Cérou and Guyader [31] and Biau and Devroye [32] have applied the bagging technique to K-NN, and they even write about bagged nearest neighbors.

Bagging is particularly attractive because it allows a single prediction to be based on many. In this way, one can build a varied collection of predictors; without bagging, only one predictor, which would be based on the whole sample, would be deduced from the chosen prediction method. The intuition is that the collection of predictors obtained through bagging enables the analyst to explore a set of possible solutions and prediction rules. Even if one of them is wrong, its impact is small because the other predictions are of better quality. The idea, simply put, is that there is strength in numbers.

To be more precise, Breiman [41] shows that the main advantage of bagging is variance reduction. He explains that the use of a decision-tree method for prediction can lead to high variance and that the use of bagging reduces this variance by averaging results from multiple decision trees. Since each bootstrapped sample is identically distributed, the bias introduced by a bagged tree is the same as that introduced by an

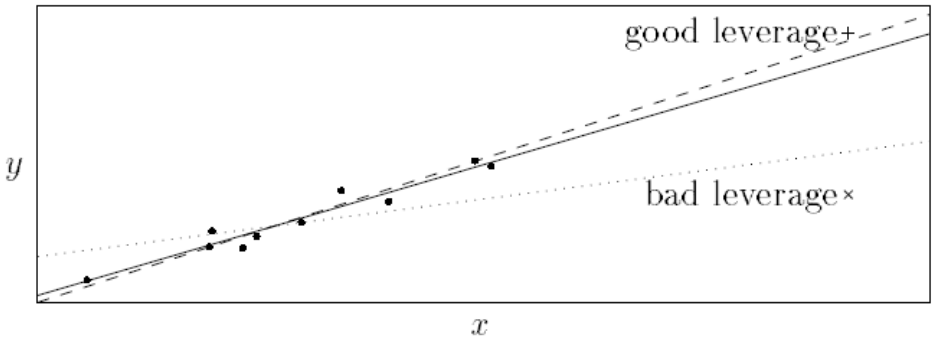
individual decision tree. However, because of the averaging step of the bagging method, the final predictor leads to an improvement because variance is reduced. Therefore, the variance of a bagged estimator is less than or equal to the variance of an individual estimator. In his paper, Breiman [41] applies bagging to several datasets (socio-economic, meteorological and simulated data) and shows that mean squared error is reduced by between 21% and 46% when bagging is applied to regression trees. He adds that variance reduction can be drastic if the individual predictor is "unstable." Bagging reduces mean squared error considerably if the predictors are "unstable," but it can remain constant if they are "stable." According to Breiman [41], the term "unstable" is relatively vague. "Unstable" predictors, on his definition, are just predictors that are far from each other or, in other words, if small changes in the data can cause large changes in the prediction. However, he provides no formal definition. Predictors can be described as unstable when they are derived from weakly positively correlated bootstrap samples. In random forests, predictor instability comes in one or two forms: the instability may be attributable to the bootstrap (the decision trees are constructed from bootstrap samples that differ from each other because they are randomly selected) or from the number of random variables used to compose the bootstrap samples (which can further uncorrelate them). Decision trees built from bootstrapped samples that differ significantly from each other produce unstable predictors. Owing to bagging, they enable gains in estimation accuracy and drastic reductions in variance. As far as bagging is concerned, the instability of the classification and regression trees (CART) model is an advantage.

The bagging method is not fully understood, and it remains the subject of intensive academic investigations. In 2000, Buhlmann and Yu [52] formalized the notion of instability and adumbrated a theory that may explain the variance reduction effect. In 2005, Hall and Samworth [102] showed that the application of bagging to the nearest neighbor method has no impact if the sampling fraction is large. In 2006, Friedman and Hall [88] presented a model in which stochastic perturbations produce excessive nonlinear variation in the resulting estimators. They explained that bagging helps to reduce these nonlinear variations. In 2006, Buja and Stuetzle [53] investigated a simplified method that applies bagging to U-statistics. They showed that bagged U-statistics (a family of estimates that generalize the concept of an average) reduce variance often but not always. However, they always increase bias. In 2008, Biau, Devroye and Lugosi [33] delivered a series of theoretical answers to conundrums about the universality of bagging's consistency. They showed that bagging is consistent whenever the base classifiers (prediction rules) are consistent and that bagging may convert inconsistent rules (like the nearest neighbor one in their paper) into consistent ones. Conversely, if a rule is inconsistent, bagging is inconsistent too.

These authors are all in agreement that bagging is a very effective technique for improving instability estimations. The non-exhaustive list in the preceding paragraph

is intended solely to highlight that research on the causes of bagging's effectiveness remains vigorous. Although many researchers agree that bagging causes a reduction in variance, some, like Buja and Stuetzle [53] explain that bagging does not necessarily have this effect. They even write about situations in which bagging could have the opposite result. Grandvalet [99] elaborated on this idea in a 2004 article. He highlighted what he calls "leverage points," extreme values or outliers that have a significant impact on the estimation of predictors. Generally, these values affect the estimate negatively and cause the predictor to deteriorate; he called them "bad leverage points." However, there may be cases where the values have a beneficial effect on the accuracy of the estimate, that is, there are "good leverage points." Grandvalet [99] illustrated this possibility with Figure 2.4-14, in which he showed that an outlier can have a negative effect on an estimate (here, a classical linear regression by the least squares; the negative effect is represented by the dotted line) as well as improving it (as shown by the dashed line).

Fig. 2.4-14: Least squares linear regressions without outlier (solid line), with a good leverage point (dashed line) and with a bad leverage point (dotted line; Source: Grandvalet, 2004 [99])



A bad leverage point can lead to a strong deviation from the original regression while a good leverage point can improve the precision of the estimate.

This example allowed Grandvalet [99] to draw an analogy with bagging. In most situations, extreme values have a negative impact on predictors. These are "bad leverage points." Bagging reduces their weight in the final predictor because they are not present in all of the bootstrapped samples and because it reduces predictor variance. In fact, it is due to the presence of "bad leverage points" that the predictors are qualified as unstable. The opposite effect occurs in the case of "good leverage points," and bagging leads to the deterioration of the final predictor.

Grandvalet [99] considered a mean estimation problem from data that were generated from a mixture distribution (a distribution that is contaminated by a widespread component that is centered on the same location) of size $n = 20$. He computed four

estimates of the mean: the sample average, the sample median, the bagged average and the bagged median. The bagged estimates were obtained over 100 bootstrap replications, and the experiment was repeated 1,000 times on independent samples. The results are presented in Table 2.4-11. All estimators are unbiased, so their expected squared errors are equal to their variances. Moreover, as the unbagged and the bagged averages are identical, their variance is the same. For this reason, Table 2.4-11 presents only the results for the average, but not those for the bagged average.

Tab. 2.4-11: Variance of the unbiased mean estimates according to the contamination proportion P (Source: Grandvalet, 2004 [99])

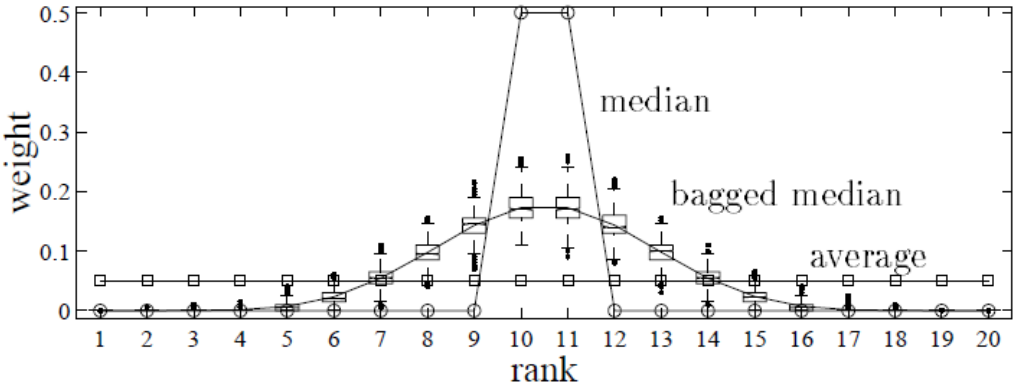
P	Average	Median	Bagged median
0	0.050 ± 0.002	0.069 ± 0.003	0.061 ± 0.003
0.05	0.321 ± 0.020	0.077 ± 0.003	0.068 ± 0.003
0.2	1.062 ± 0.050	0.109 ± 0.005	0.102 ± 0.005
0.7	3.565 ± 0.148	1.121 ± 0.084	1.544 ± 0.091
1	5.020 ± 0.216	6.933 ± 0.296	6.143 ± 0.261

For no contamination ($P = 0$), the average is equal to 0.050 ± 0.002, the median is equal to 0.069 ± 0.003 and the bagged median is equal to 0.061 ± 0.003. On the other hand, with 5% of contaminated data, the average is equal to 0.321 ± 0.020 (much more than without), the median is equal to 0.077 ± 0.003 and the bagged median is equal to 0.068 ± 0.003.

It was noted previously that, like all unbiased estimators, the bagged average gives the same results as the unbagged average. Therefore, in this instance, bagging does not reduce variance. If the median and the bagged median results are considered, bagging reduces the variance on some occasions but not on others. In Table 2.4-11, bagging increases variance when the contamination proportion is equal to 0.7. Then Grandvalet [99] explained that studies of bagging generally concern the "macroscopic" effect on the estimate, whereas his focus was on "atomic" effects, that is, the weight attached to a single observation.

Figure 2.4-15 shows the distribution of the weights allocated to values of ordered x_i, where $x_1 \leq x_2 \leq \ldots \leq x_{20}$, over 1,000 scenarios. Unsurprisingly, all points contribute equally to the computation of the average, bagging has no effect on it, and the median yields a weight of 1/2 for x_{10} and x_{11} and 0 for the other values in the sample. In the case of the bagged median, the median of the bootstrap sample b is defined as $(x^b_{10} + x^b_{11})/2$ with x^b_{10} and x^b_{11}, which may not correspond to x_{10} and x_{11}.

Fig. 2.4-15: Boxplot of the weight allocated to the examples x_i according to the rank of x_i for original and bagged mean estimates (Source: Grandvalet, 2004 [99])



Grandvalet [99] explained that bagging is a method for distributing weights more evenly over an entire sample. He concluded by voicing his support for Breiman's idea that bagging increases the accuracy of the predictor when the prediction method is unstable. However, Grandvalet [99] insists that instability is not related to the intrinsic variability of the predictor. Instead, it is connected to the presence of influential points. In many cases, these influential points are outliers ("bad leverage points"). Bagging reduces their weight, which, in turn, reduces the variance of the predictor. However, the use of bagging can be detrimental when the estimates are made more precise by "good leverage points." Therefore, according to Grandvalet [99], bagging is a method that improves robustness to outliers.

In 2012, Bühlmann [51] showed the efficiency of bagging for solving instability problems through a toy example that is based on an indicator function. Nevertheless, he eventually wrote that that "...bagging is a smoothing operation. The amount of smoothing is determined 'automatically' and turns out to be very reasonable (we are not claiming any optimality here). The effect of smoothing is that bagging reduces variance due to a soft-instead of a hard-thresholding operation." Thus, bagging is an effective method because of the smoothing that it involves, but it is not optimal. Despite the theoretical contributions made by Bühlmann and Yu [52] or those of Biau, Devroye and Lugosi [33], the difficulty of interpreting it is probably its greatest disadvantage.

Bagging is based on simple and recent methods. It is valuable because it can be used to improve unstable methods, such as some machine-learning methods that are currently in vogue, considerably. However, the reasons for its good functioning remain unclear due to the lack of indisputable mathematical proof.

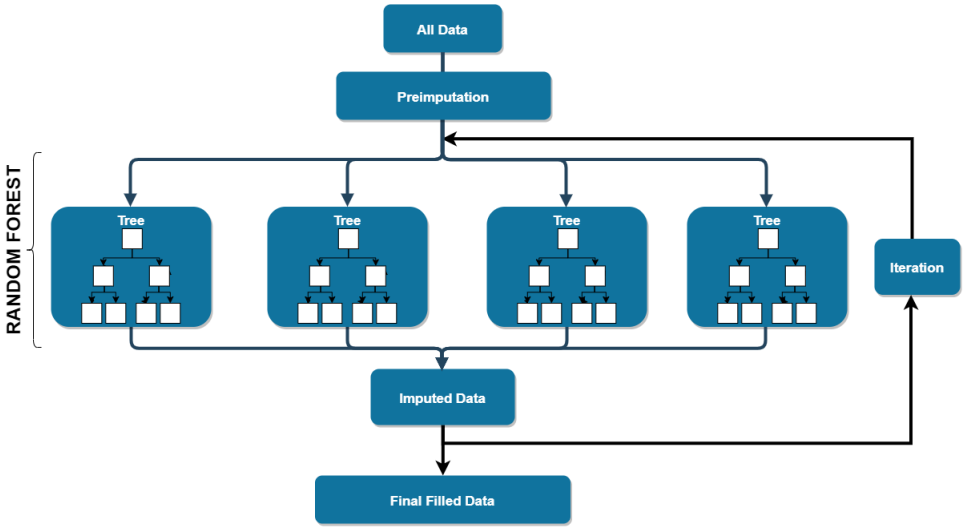
Random forests for data imputation

Random forests can be used to fill in missing data because forecasting is very similar to imputation. Two important implementations of random forests can be found in the literature (both in the R software): rfImpute (written by Breiman [43] in 2003) and missForest (written by Stekhoven and Bühlmann[194] in 2011).

The *rfImpute* function is part of the *randomForest* package, that Breiman [43]developed to impute missing data. It uses a proximity matrix to estimate missing data. It starts with a naïve (median) imputation. Random forests are then generated from the pre-completed sample in order to calculate a proximity matrix. The intuition of the proximity matrix is that similar observations should be in the same terminal nodes. Then, the imputed value is the weighted average of the non-missing observations. The weights are the proximities (from the proximity matrix). The proximity matrix shows how similar the observations are to each other by counting how often two data points end in the same terminal node. The proximity matrix corresponds to a $(n \times n)$ matrix. The bootstrapped observations are run down each tree once it is grow, and if two cases i and j happen to be in the same terminal node, their proximity measure is increased by one (the i^{th} element of the j^{th} column, and conversely, of the proximity matrix). At the end of the forests construction, these proximities are normalized. In other words, each element of the proximity matrix corresponds to the percentage of time each pair of observations ends in the same terminal node. The higher the proximity measure, the more similar the pair of observations. The drawback of this method is that $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ must be fully observed (without any missing data).

It is for this reason that Stekhoven and Bühlmann [194] have implemented another version of random forests. The function is *missForest* was implemented by Stekhoven and Bühlmann [194] in 2011, and it is distributed in an R package with the same name. It is expected to be particularly effective for mixed-type data, including complex interactions and non-linear relationships. Their algorithm works a little differently from that of Breiman's (*rfimpute* function) as it uses an iterative scheme. It consists in permuting each variable as an explanatory variable, and building a forests only for the non-missing values, then iterating. To do so, first, missing data are pre-completed through mean imputation (or another method). Thereafter, the variables are sorted by amount of missing data (in ascending order) in order to use them (in this order) as explanatory variable of random forests. Then, the random forests imputation is used iteratively, using the sorted variables as explanatory variables for each iteration. In other words, the first iteration consist in applying random forests algorithm on all available observations from the data matrix, using the column with the most missing data as explanatory variable, to impute missing data. Then, the second iteration consists in applying random forests algorithm using the second column with the most missing data, and so forth until a stop criterion is met. The Stekhoven and Bühlmann [194] algorithm is presented in Figure 2.4-16.

Fig. 2.4-16: Random forests algorithm for data imputation



The stop criteria are defined so that the imputation process terminates as soon as the difference between two successive imputed data matrices increases for the first time. The formal definition is as follows:

$$\Delta = \frac{\sum(\boldsymbol{Y}^{old} - \boldsymbol{Y}^{new})^2}{\sum(\boldsymbol{Y}^{new})^2}. \tag{2.4-46}$$

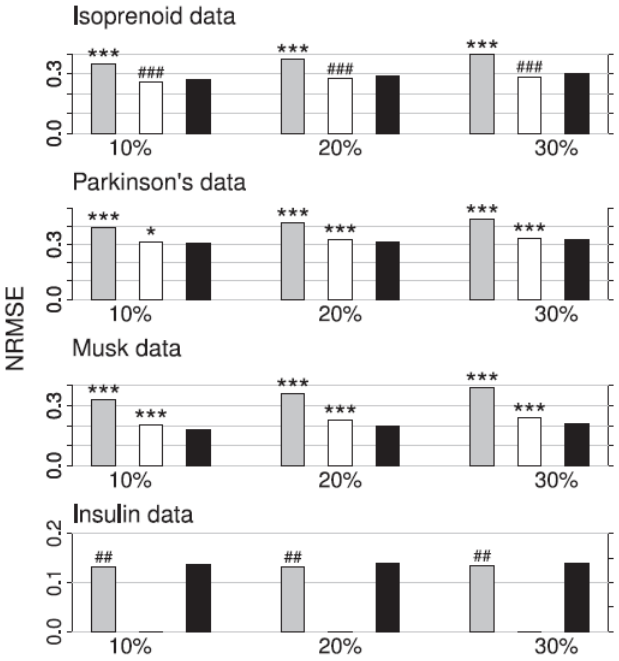
As mentioned above, these two functions see the most use in the literature. The three papers presented below motivate the choice of the function that is used in the rest of this PhD thesis. First of all, Stekhoven and Bühlmann [194] show the efficency of the *missForest* function by comparing it with different samples and to three alternative imputation methods.

- K-NN: selects the K-closest cases that are not missing values in the attributes that are to be imputed so that Euclidean distance is minimized (see Section 2.4.3)
- Missingness pattern alternating imputation and l_1-penalty (*MissPALasso*) [192]: an EM algorithm in which the missing variables are regressed on the observed ones using the Lasso penalty by Tibshirani [198]
- Multiple imputation by chained equations (MICE): a multiple imputation with a linear mixed model that is based on predictive mean matching and regression methods (see Section 2.4.7), with five imputations combined by averaging.

In each experiment, the authors performed 50 independent simulations on samples where 10%, 20% and 30% of the data had been removed in a completely randomized

manner (MCAR). Observing a database composed exclusively of continuous variables (Figure 2.4-17), they noted that the random forests method performs well, with normalized root mean squared error (NRMSE) falling by 25% relative to K-NN. In the case of the musk data, the reduction is larger than 50%. In the case of the insulin data, the *MissPALasso* method performs better due to the large size of the sample, which complicates the calculation.

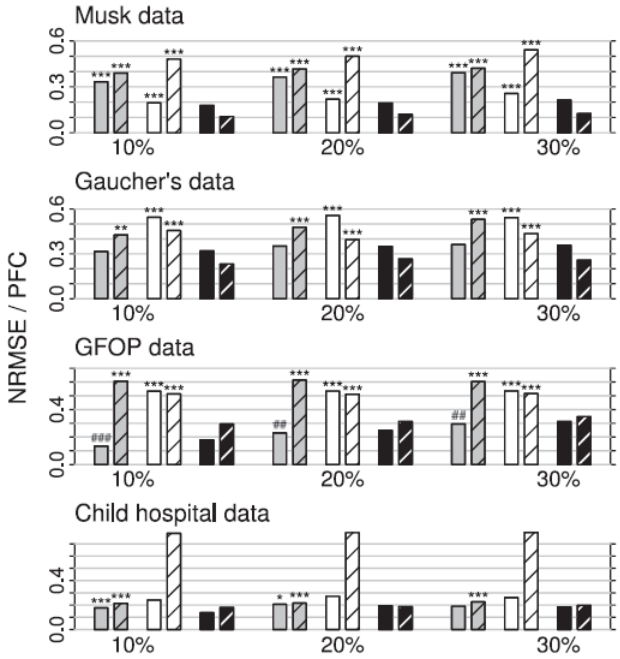
Fig. 2.4-17: Average NRMSE for K-NN (grey), *MissPALasso* (white) and random forests (black) on four continuous datasets and three different amounts of missing data. (Source: Stekhoven and Bühlmann, 2011 [194])



Standard errors are in the order of magnitude of 10^{-4}. Significance levels for the paired Wilcoxon tests in favor of random forests are encoded as $,*,(<0.05)$, $,**,(<0.01)$, and $,***,(<0.001)$. If the average error of a method is smaller than that of *missForest* [193], the significance level is encoded by a hash (#) instead of an asterisk. In the lowermost dataset, results for *MissPALasso* are missing due to the implementations limited capability with regard to accommodate high dimensions.

The authors also experiment with mixed data (both continuous and categorical) in four new samples(see Figure 2.4-18). They calculated the NRMSE for continuous variables again as well as the proportion of false classification (PFC) for categorical variables. They found that random forests generally produce better results than the other two methods, reducing imputation error by more than 50% in several cases. In the case of GFOP data, they noted that the K-NN method yields a lower NRMSE than random forests, but that its PCA is at least twice as high.

Fig. 2.4-18: Average NRMSE (left bar) and PFC (right bar) for K-NN (grey), MICE (white) and random forests (black) on four mixed-type datasets and three different amounts of missing data. (Source: Stekhoven and Bühlmann, 2011 [194])



Standard errors are in the order of magnitude of 10^{-3}. Significance levels for the paired Wilcoxon tests in favor of random forests are encoded as $,*,(<0.05)$, $,**,(<0.01)$, and $,***,(<0.001)$. If the average error of a method is smaller than that of *missForest* [193], the significance level is encoded by a hash (#) instead of an asterisk.

Finally, Stekhoven and Bühlmann [194] compared the calculation times of the algorithms (see Table 2.4-12). They found that the K-NN method is the fastest. However, the random forests method is still relatively fast, especially when compared to *MissPALasso* and MICE.

Tab. 2.4-12: Average computation time (in seconds) for imputing the analyzed datasets (Source: Stekhoven and Bühlmann, 2011 [194])

Dataset	n	p	K-NN	MissPALasso	MICE	missForest
Isoprenoid	118	39	0.8	170	-	5.8
Parkinson's	195	22	0.7	120	-	6.1
Musk (continue)	476	166	13	1,400	-	250
Insulin	110	12,626	1,800	NA	-	6,200
Mussk (mixed)	476	167	27	-	2,800	500
Gaucher's	40	590	1.3	-	1,400	40
GFOP	595	18	2.7	-	1,400	40
Children	55	124	2.7	-	4,000	110

n: number of rows
p: number of columns
NA: not available
For isopenoid data (of dimension (188×39) the K-NN is executed in 0.8 seconds, MissPALasso in 170 seconds and missForest in 5.8 seconds.

In 2016, Jamal [180] applied different algorithms to eight databases composed of either continuous or categorical financial data (no mixing). In addition to random forests, represented in her study by the *missForest* function, she compares three other methods.

- LOCF: the missing value is replaced with the last known data for the variable (see Section 2.4.1)
- K-NN: selects the K closest cases that are not missing values for the attributes to be imputed so that distance is minimized (see Section 2.4.3)
- Multiple imputation by chained equations (MICE): multiple imputation with a linear mixed model that is based on predictive mean matching and regression methods (see Section 2.4.7), with five imputations combined by averaging

Each of these methods was applied to data from to Natixis Assurance savings contracts that where subject to surrender between 2005 and 2010 and in 2013. First, a listwise deletion (see Section 2.2.2) was executed to obtain a complete sample. Then, missing data were generated randomly and progressively (seven databases are generated, with missing data ranging from 10% to 80% with a step of 10%). Once the data had been completed in this fashion, the results were compared to the initial data by reference to Hamming distance (the sum of the absolute differences).

Table 2.4-13 shows the results for one database (contract seniority data), but the conclusions are the same for the other seven. Jamal [180] noted that the random forests method is the one that minimizes the imputation error rate (Hamming distance) best. The error rate is very low for the quasi-totality of the variables. Unsurprisingly, LOCF is the method with the highest error rate because it is based on a non-existent

neighborhood relationship. She also found that the MICE imputation method is not suitable for the dataset under observation, despite the missing data being of the MCAR type.

Tab. 2.4-13: Percentage imputation errors for missing data in contract seniority data (Source: Jamal, 2016 [180])

	Missing data proportion							
	10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**
LOCF	1.93	1.32	1.84	2.03	1.77	1.38	1.62	1.27
K-NN	1.35	1.42	0.86	1.31	0.92	0.83	0.9	0.84
MICE	1.17	0.88	0.88	1.4	0.95	0.81	0.85	0.84
Random forests	0.93	0.82	0.69	1.09	0.53	0.57	0.63	0.52

For a missingness proportion at 10%, the imputation error is equal to 1.93% for the LOCF method while it is 0.93% for the random forests.

Then, she compared the NRMSE (for continuous variables) and the PCA (proportion of falsely classified entries; for discrete variables) of the random forests method, the K-NN and MICE. She found that the imputation of missing data through the random forests method minimizes quadratic error (see Table 2.4-14).

Tab. 2.4-14: Comparison between error rates of imputation methods (Source: Jamal, 2016 [180])

	K-NN	**MICE**	**Random forests**
NRMSE	27.1%	32.5%	19.3%
PFC	36.7%	29.0%	13.7%

The normalized RMSE is equal to 27.1% for the K-NN method while it is 19.3% for the random forests. The proportion of falsely classified entries is equal to 36.7% for the K-NN method while it is 13.7% for the random forests.

Lastly, in 2017, Young [214] compared Breiman's *rfimpute* function to three other methodologies.

- New/Rough: using median imputation to fill in missing values before running a random forests algorithm (the same procedure as the *missForest* function).
- New/Rand: using a random value from the sample to fill in missing values before running a random forests algorithm.
- Rough: using median imputation.
- Prox: filling in missing values with the help of the proximity matrix of the random forests (the same procedure as the *rfimpute* function).

- Mice/For: MICE (see Section 2.4.7) with random forests
- Mice/Pmm: MICE (see Section 2.4.7) with predictive mean matching.

These six methods were applied to several medial datasets (from the UC Irvine repository [141]), whereby missing at random data (10% to 60% in steps of 10%) were entered and compared by reference to the proportion of correct classifications.

Some of the results are presented in Table 2.4-15 Young [214] reported that the most accurate method for these data is MICE with PMM (Mice/Pmm). New/Rough (median imputation before running the random forests algorithm) came second. It should also be noted that the proximity matrix completion method appears to be the least efficient on these data.

Tab. 2.4-15: MAR datasets percent correctly classified (Source: Young, 2017 [214])

	New/Rough	**New/Rand**	**Rough**	**Prox**	**Mice/For**	**Mice/Pmm**
Soybean	91.44	91.09	91.36	90.59	91.24	92.09
Hepatitis	86.19	84.46	84.83	84.76	85.92	86.21
Adult	79.90	79.87	79.68	79.69	79.79	79.80
Horse	76.69	75.93	77.06	74.63	77.58	78.55
Bridge	70.33	69.68	69.67	69.61	70.31	71.74

For soybean data, using median imputation to fill in missing values before running a random forests algorithm (New/Rough) lead to correctly classification in 91.44% of cases, while using a random value from the sample to fill in missing values before running a random forests algorithm (New/Rand) lead to correctly classification in 91.09% of cases.

Young [214] concluded that New/Rough is the most accurate method. It is also known as the *missForest* function by Stekhoven and Bühlmann [194]. It seems to be more efficient than the one that uses the procedure with a proximity matrix (like in the *rfimpute* function), which was proposed by Breiman. In their articles, Stekhoven and Bühlmann [194] and Jamal [180] show the relative efficiency of random forests and the *missForest* function in particular. It is for this reason that the *rfimpute* function is not used in the remainder of this PhD thesis. The empirical studies use the R package by Stekhoven and Bühlmann [194] (the *missForest* package [193]).

The effectiveness of random forests is still sensitive to the choice of parameters, particularly the number of trees and the number of iterations. It is certainly possible to choose high parameters to maximize the accuracy of the results but doing so would have a direct impact on computation time.

Non-replicability of results

One of the problems for regulators is that the results of random forests are not reproducible. Because of their random component, it is not possible to obtain the same

output from a given set of inputs. If random forests are applied to a sample twice, then, holding all other things constant, the two results obtained will be close (if the algorithm is robust) but different. This problem affects all algorithms that have a random component. It is the bootstrap step of random forests that makes it random. As noted earlier, each predictor is derived from a decision tree that has been fitted from bootstrap samples that are drawn from the original sample at random. The final predictor is the average of the collection of predictors that are obtained from the bootstrap samples. Therefore, if the algorithm is applied several times in a row, it is unlikely that the bootstrap samples in each application would coincide.

Many other algorithms include a random component. One notable example is the Brownian bridge method (see Section 2.4.2), which uses a Gaussian random variable in its process. The Amelia and MICE algorithms, which are presented in the following sections (see Section 2.4.6 and Section 2.4.7), are multiple imputation methods that integrate a bootstrap step and a Monte Carlo step, respectively.

Non-replicability is a problem for regulators in risk management because it can enable a bank to reduce its capital charges. Regardless of methodology, there is no guarantee that the results provided by the bank are not generated from a scenario that it favors, which would lead to an underestimation of risk. In the case of a VaR calculation, which is simply a quantile of the distribution, the effect can be significant. If the data used for the calculation of this risk measure reflect an optimistic scenario, this could lead to a higher VaR. Conversely, if the data reflect a pessimistic scenario, then the bank would be overestimating its level of risk by decreasing the level of VaR, which would have a direct impact on the capital charges that it must pay. Manipulation through a small number of bootstrap samples is also a possibility. Finally, each of the parameters that make up the model must be inspected in order to validate the resultant risk measures. The task of the regulator becomes more difficult if they have to understand not only the risk measures of the bank and the model but also all of the parameters that comprise it. What the regulator wants to avoid above all is moral hazard, that is, banks manipulating data to reduce their capital charges. In fact, banks could be tempted to present results to the regulators that understate risks in order to minimize charges.

A possible solution to the non-replicability of the results would be to fix a seed in the random bootstrap generator. This would allow banks and regulators to retest the model by changing parameters and observing the impact of the changes without being disturbed by the random effect generated by the bootstrap. The use of a seed would allow banks to be more transparent about the completion methods that they use because each effect of the algorithm would be identifiable, enabling them to understand their completion methods. On the other hand, setting a seed is only relevant if the code used is fully transparent, which refers to the importance of documentation and its transparency. The regulator could also repeat the calculations in order to obtain a

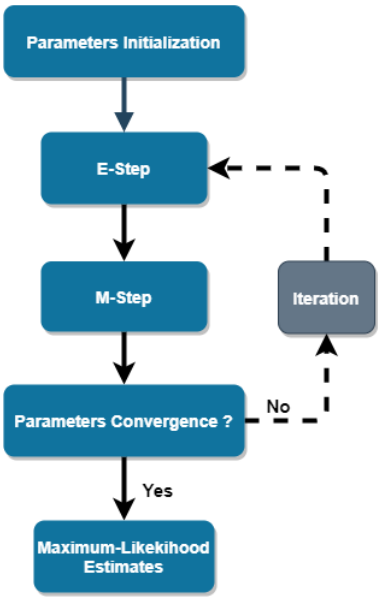
range of imputed values or a range of risk measures obtained. This methodology would make it possible to differentiate between favorable and unfavorable scenarios, and to select a scenario that would provide conservative risk measures.

2.4.6 Amelia: improved expectation-maximization algorithm

EM algorithm

The methods presented so far are adapted to samples with missing data. However, there is one method that has been developed specifically for incomplete samples. In 1977, Dempster, Laird and Rubin introduced the EM algorithm [68], that is a parametric estimation method within the general framework of maximum likelihood. This method consists in repeating two steps at each iteration: the E step (expectation step) which consists in estimating the missing values knowing the observed values and the parameters determined at the previous iteration; and the M step (maximization step) which consists in maximizing the likelihood using the estimates of the completed values of the previous E step, and in updating the parameters for the next iteration. These steps are iterated until the parameters converge (see Figure 2.4-19).

Fig. 2.4-19: EM algorithm



Schafer [181] presents the EM algorithm by separating the observed data from the missing data. Let $\boldsymbol{Y}$ be the data matrix, let $\boldsymbol{Y}^{obs}$ be the observed part of $\boldsymbol{Y}$, let $\boldsymbol{Y}^{miss}$ be the missing part of $\boldsymbol{Y}$, and let θ be the unknown parameters. Let $f(\boldsymbol{Y}|\theta)$ be the probability density function of $\boldsymbol{Y}$ for all θ. The EM algorithm assumes that $\boldsymbol{Y}^{miss}$ are

MAR, which means that missing data can be marginalized. Then, the log-likelihood function of the marginalized data (observed data with the missing data ignored) can be written as follows:

$$l(\boldsymbol{Y}^{obs};\theta) = lnf(\boldsymbol{Y}^{obs};\theta) = ln\int f(\boldsymbol{Y}^{obs},\boldsymbol{Y}^{miss};\theta)d\boldsymbol{Y}^{miss} \tag{2.4-47}$$

However, this log-likelihood function cannot be computed, which is why θ is estimated by iteratively maximizing the log-likelihood function of the complete data:

$$l(\boldsymbol{Y};\theta) = lnf(\boldsymbol{Y}^{obs},\boldsymbol{Y}^{miss};\theta) \tag{2.4-48}$$

Consider that $\theta^{(0)}$ are the initial parameters and that $\theta^{(t)}$ are the estimated parameters at the t-th iteration. The EM-algorithm starts with the parameters $\theta^{(0)}$ and repeats the following two steps:

- The E-step finds the conditional expectation of the log-likelihood of complete data given the observed data and the current estimated parameters.

$$Q(\theta,\theta^{(t)}) = \mathbb{E}\left[l(\boldsymbol{Y};\theta)|\boldsymbol{Y}^{obs};\theta^{(t)}\right] \tag{2.4-49}$$

- The M-step maximizes the Q function in order to determine $\theta^{(t+1)}$.

$$\begin{aligned}\theta^{(t+1)} &= \underset{\theta}{argmax}\left\{Q(\theta,\theta^{(t)})\right\}\\ &= \underset{\theta}{argmax}\left\{\mathbb{E}\left[l(\boldsymbol{Y};\theta)|\boldsymbol{Y}^{obs};\theta^{(t)}\right]\right\}\end{aligned} \tag{2.4-50}$$

This algorithm is based on the following property (introduced by Wu [212] in 1983):

$$l(\boldsymbol{Y}^{obs};\theta) - l(\boldsymbol{Y}^{obs};\theta^{(t)}) \geq Q(\theta,\theta^{(t)}) - Q(\theta^{(t)},\theta^{(t)}), \qquad \text{for all } \theta,\theta^{(t)} \tag{2.4-51}$$

Accordingly, at each M-step, $\theta^{(t)}$is known, and the algorithm aims to find the best θ that increases the log-likelihood so that $(\boldsymbol{Y}^{obs};\theta) - l(\boldsymbol{Y}^{obs};\theta^{(t)}) \geq 0$. As noted previously, $l(\boldsymbol{Y}^{obs};\theta)$ cannot be computed, so the solution is to find a function that can be maximized, such as $Q(\theta,\theta^{(t)})$, to then find the θ value that maximizes $Q(\theta,\theta^{(t)}) - Q(\theta^{(t)},\theta^{(t)})$. This results in an iterative sequence of estimated parameter $(\theta^{(t)})_{t\geq 0}$.

Notably, this is possible because of Jensen's inequality, which posits that if f is a convex function defined by the interval I and if $x_1,\ldots,x_n \in I$ and $\lambda_1,\ldots,\lambda_n \geq 0$ such that $\sum_{i=1}^{n}\lambda_i = 1$, then

$$f(\sum_{i=1}^{n}\lambda_i x_i) \leq \sum_{i=1}^{n}\lambda_i f(x_i). \tag{2.4-52}$$

Therefore, in this case, f is a logarithmic function. Equation 2.4-51 can be deduced from it. Wu [212] focused on this problem in more detail in his 1983 article.

Application to multivariate normal data

It has emerged that the EM-algorithm can be applied to multivariate normal data. Suppose that $\boldsymbol{Y}$ is a matrix with dimensions $(n \times p)$ such that $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p)$ have a p-variable normal distribution with a mean of $\mu = (\mu_1, \ldots, \mu_p)$ and a covariance matrix $\Sigma = (\sigma_{ij})$. Let $\boldsymbol{Y}^{obs}$ denote the observed part of $\boldsymbol{Y}$, and let $\boldsymbol{Y}^{miss}$ denote its missing part. In addition, let y_i^{obs} be the set of variables for the i-th observation with $i = 1, \ldots, n$. The log-likelihood function that is based on the observed data is given by

$$\begin{aligned} l(\boldsymbol{Y}^{obs}|\mu, \Sigma) = &- nln(2\pi) - \frac{1}{2}\sum_{i=1}^{n} ln|\Sigma_i^{obs}| \\ &- \frac{1}{2}\sum_{i=1}^{n}(y_i^{obs} - \mu^{obs})^T {\Sigma_i^{obs}}^{-1}(y_i^{obs} - \mu^{obs}), \end{aligned} \tag{2.4-53}$$

where μ^{obs} and Σ_i^{obs} are the mean and the covariance matrix of the observed components of the i-th observations of $\boldsymbol{Y}$, respectively. Here, the constant $-nln(2\pi)$ can be ignored because it has no impact on the maximization process. Then, at the t-th iteration, the algorithm estimates the parameters $\mu^{(t)}$ and $\Sigma^{(t)}$. At iteration $t+1$, the EM algorithm consists in computing these two expectations conditionally on the observed values:

$$\mathbb{E}\left(\sum_{i=1}^{n} y_{ij}|\boldsymbol{Y}^{obs}, \mu^{(t)}, \Sigma^{(t)}\right) = \sum_{i=1}^{n} y_{ij}^{t+1}, \quad j = 1, \ldots, p, \tag{2.4-54}$$

and

$$\mathbb{E}\left(\sum_{i=1}^{n} y_{ij}y_{ik}|\boldsymbol{Y}^{obs}, \mu^{(t)}, \Sigma^{(t)}\right) = \sum_{i=1}^{n}\left(y_{ij}^{t+1}y_{ik}^{t+1} + c_{ijk}^{(t+1)}\right), \quad j, k = 1, \ldots, p, \tag{2.4-55}$$

with

$$y_{ij}^{(t+1)} = \begin{cases} y_{ij}, & \text{if } y_{ij} \text{ is observed} \\ \mathbb{E}(y_{ij}|y_i^{obs}, \mu^{(t)}, \Sigma^{(t)}), & \text{if } y_{ij} \text{ is missing} \end{cases}, \tag{2.4-56}$$

and

$$c_{ijk}^{(t+1)} = \begin{cases} 0, & \text{if } y_{ij} \text{ or } y_{ik} \text{ is observed} \\ Cov(y_{ij}, y_{ik}|y_i^{obs}, \mu^{(t)}, \Sigma^{(t)}), & \text{if } y_{ij} \text{ and } y_{ik} \text{ are missing} \end{cases}, \tag{2.4-57}$$

The missing values y_{ij} are replaced by $\mathbb{E}(y_{ij}|y_i^{obs}, \mu^{(t)}, \Sigma^{(t)})$, which is the conditional mean of y_{ij} given the set of observed values of the i-observation y_i^{obs} and the current parameters $\mu^{(t)}$ and $\Sigma^{(t)}$. These conditional means and the nonzero conditional covariances are easily found from the current parameter estimates by sweeping the augmented covariance matrix so that the variables y_i^{obs} are predictors in the regression equation and the remaining variables y_i^{miss} are outcome variables (see Beaton's article in 1964 for more details about the sweep operator [28]). The adjustments c_{ijk} that are included in Equation 2.4-55 are needed to correct for biases in the resulting estimated covariance matrix by imputing conditional means for the missing values.

Then, the M-step consists in computing the new estimates $\mu^{(t+1)}$ and $\Sigma^{(t+1)}$ from the estimated complete data $y_{ij}^{(t+1)}$, in the following manner:

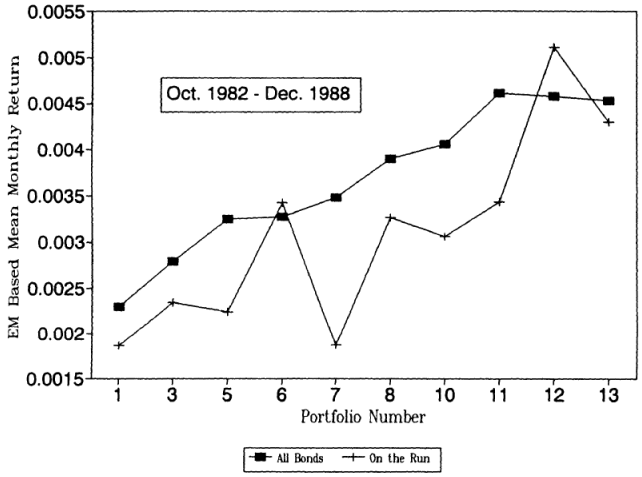
$$\begin{aligned}
\mu_j^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n} y_{ij}^{(t+1)}, \quad j = 1, \dots, p \\
\sigma_{jk}^{(t+1)} &= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} y_{ij}y_{ik} | \boldsymbol{Y}^{obs}, \mu^{(t)}, \Sigma^{(t)}\right) - \mu_j^{(t+1)}\mu_k^{(t+1)} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\left(y_{ij}^{(t+1)} - \mu_j^{(t+1)}\right)\left(y_{ik}^{(t+1)} - \mu_k^{(t+1)}\right) + c_{ijk}^{(t+1)}\right], \quad j, k = 1, \dots, p
\end{aligned} \tag{2.4-58}$$

These steps are repeated until the parameters converge.

Imputation by EM on financial data

The application of the EM algorithm to financial data is not novel. In 1992, Warga [208] studied the holding period returns of constant-duration portfolios of U.S. Government notes and bonds and measured the return premium generated by liquidity differences between bonds. The comparison is between constant-duration portfolios constructed in two ways: one type of portfolio contains only the bonds issued at the last Treasury auction (called "on the run"), and the other type of portfolio contains all other bonds (called "off the run"). The comparisons between the series are made meaningful by the choice of narrow maturity ranges in the portfolio construction process. To execute the analysis, he Warga [208] used 13 constant-duration bond portfolios with four-month duration ranges. He faced a problem of missing data in his time series and decided to use the EM algorithm to estimate the parameters of the return distributions. In particular, the EM algorithm enabled him to estimate the mean returns of the excess portfolios, which are presented in Figure 2.4-20.

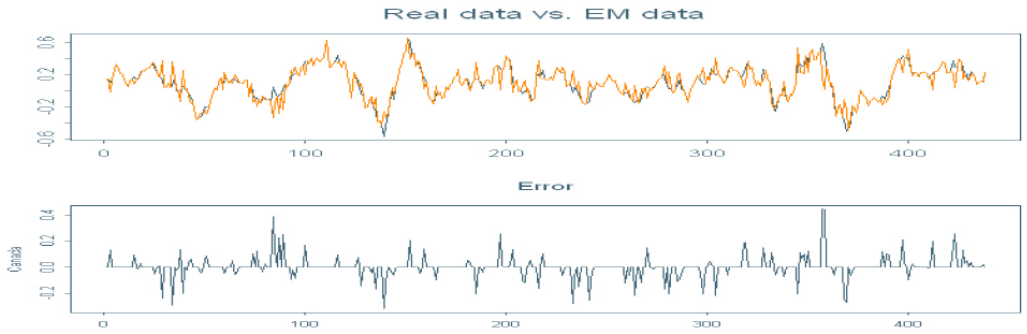
Fig. 2.4-20: Excess mean monthly returns from constant-duration bond portfolios (Source: Warga, 1992 [208])



These estimators allowed him to continue his analysis and showed that during the period between October 1982 and December 1988, the behavior of the mean return of the off-the-run portfolio differed significantly from that of the on-the-run portfolio. Therefore, there were differences in bond liquidity.

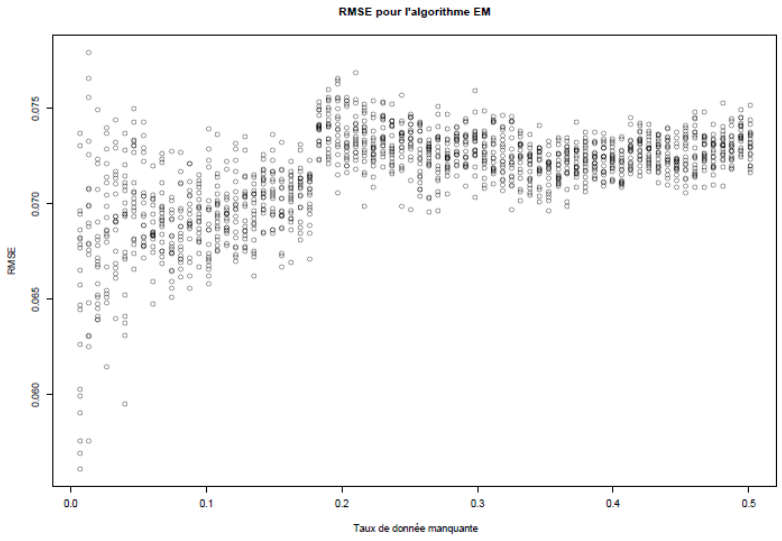
In 2007, Urli used the EM algorithm with financial time series to impute missing data [204]. The sample that he studied comprises 18 Morgan Stanley Capital International (MSCI) stock market indices for the period between 1970 and 2006 for which all data are available. He used annual rates for the indices to facilitate his analysis. He began by removing data in a completely randomized manner (MCAR), regardless of variable or time. He analyzed the best result from 10 EM-algorithm imputations. He found that this imputation method allowed him to report the behavior of the series correctly, albeit with a few exceptions. The gap around the 130th return was larger due to a drop in Canadian rates that the algorithm could not predict because it was much larger in Canada than elsewhere in the world (see Figure 2.4-21).

Fig. 2.4-21: Canadian annual rate (original series in blue and EM imputation 25% of missing values in orange) at top. Error between observed and imputed data at bottom (Source: Urli, 2007 [204])



As far as imputation accuracy (in RMSE terms) is concerned, Urli [204] pointed out that the size of the mean error increases with the rate of missing data due to a decrease in observable data(Figure 2.4-22). The less observable data one has, the poorer one's estimate of θ. He also noted that average error stabilizes and remains the same at between approximately 20% and 50% of missing data. He also observed a fairly large scatter of points when there was little missing data. The explanation is that the imputation is randomly drawn from the estimated distribution and that the distance between these imputations and the original can vary, with a small divisor in RMSE terms.

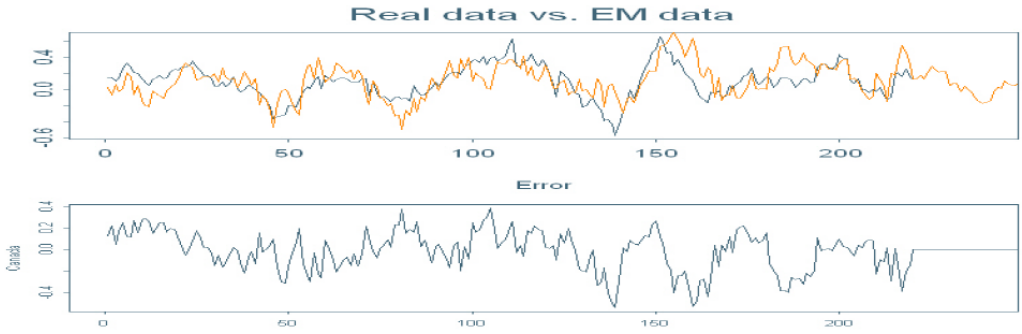
Fig. 2.4-22: RMSE for different levels of missing data (Source: Urli, 2007 [204])



In the second step, Urli [204] analyzed the completion results for another missing data pattern. In that pattern, many successive missing data occur in a particular series

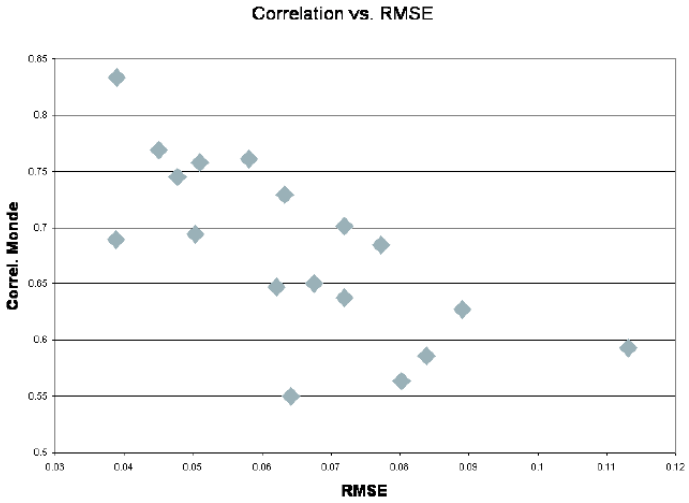
(MAR). He then removed 50% of the original data from the Canadian annual rates and applied the EM algorithm to them (see Figure 2.4-23). The completed data seemed to follow a similar dynamic to the original data, but the correspondence was not perfect. The RMSE of the imputation is 0.1928, which is relatively high for annual rates. That result may be attributable to the use of the trend for all MSCI countries (θ estimated from all available data) to replicate Canadian behavior, generating somewhat rough results. He noted a close relationship between the efficiency of the EM algorithm and the correlations between variables.

Fig. 2.4-23: Canadian annual rate (original series in blue and EM imputation of 50% successive missing values in orange) at top. Error between observed and imputed data at bottom (Source: Urli, 2007 [204])



Urli [204] decided to test this relationship by comparing the RMSE of the series that were correlated to the portfolio under observation. He found that the most correlated variables were also those on which the EM algorithm made the smallest errors (see Figure 2.4-24). In general, the algorithm succeeds in reproducing important trends in a dataset, but it struggles to capture the local particularity of each variable. This tendency is, as expected, even more pronounced in the case of variables that are poorly correlated with the others. For example, the Netherlands has a higher correlation with the dataset than Austria (0.83 and 0.55, respectively), and it has a lower RMSE for a missing data rate of 30% (0.037 versus 0.063, respectively).

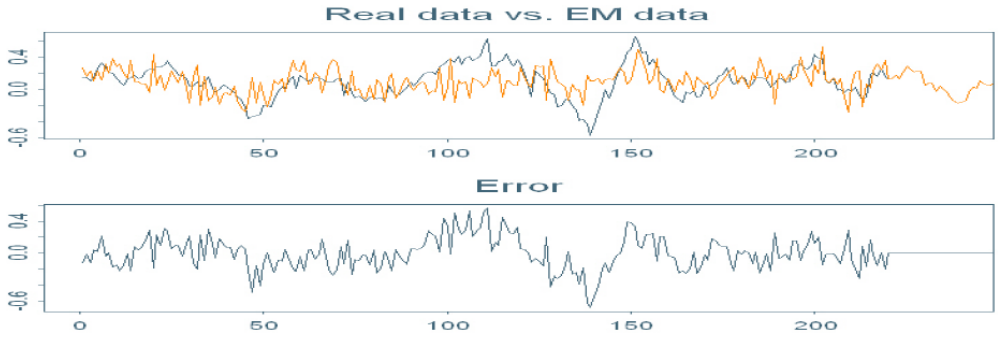
Fig. 2.4-24: Relationship between correlation and RMSE with 50% missing data (Source: Urli, 2007 [204])



Urli [204] completed his analysis by addressing missing data by period, that is, cases in which data are missing at the very beginning of the series for several variables (MNAR). He removed 50% of the first observations for all variables except those that covered the United States. The imputation results were of poor quality. The RMSE is 0.31, compared to 0.20 for missing data of the MCAR type. The scarcity of the information that was available to the algorithm for the estimation of the distribution parameters explains this outcome.

It is possible to see from the Canadian case presented in Figure 2.4-25 that imputation fails to replicate even the largest trends. Indeed, in this case, imputation resembles a straight line with noise.

Fig. 2.4-25: Canadian annual rate (original series in blue and EM imputation of 50% data MNAR in orange) at top. Error between observed and imputed data at bottom (Source: Urli, 2007 [204])



Amelia algorithm

The Amelia method, created in 2002 by Honaker, Joseph, King, Scheve and Sigh [112], is a multiple imputation method (introduced by Rubin [177]) that aims to impute B values for each missing data in the sample by creating B complete samples (rows are sampled with replacement). Like the EM algorithm, it assumes that all data (observed and missing) follow a multivariate Gaussian distribution. For a dataset $\boldsymbol{Y}$ of dimensions $(n \times p)$ in which $\boldsymbol{Y}^{obs}$ is the observed part and $\boldsymbol{Y}^{miss}$ is the missing part,

$$\boldsymbol{Y} \sim \mathcal{N}_p(\mu, \Sigma) \tag{2.4-59}$$

which means that $\boldsymbol{Y}$ follows a multivariate normal distribution with a mean μ and covariance matrix Σ. Like most imputation methods, Amelia assumes that the data are randomly missing (MAR). Therefore, the model depends only on observed data $\boldsymbol{Y}^{obs}$ and not on unobserved data $\boldsymbol{Y}^{miss}$.

The Amelia algorithm, also called Expectation-Maximization with Bootstrapping (EMB) combines the classical EM algorithm with a bootstrap approach. The EM algorithm allows the estimation of the distribution. More precisely, for each draw, the data are bootstrapped to simulate estimation uncertainty. Then, the EM algorithm is applied to estimate the distribution of these bootstrapped data.

Honaker, Joseph, King, Scheve and Sigh [112] explained that in order to combine the results from the B datasets, it is necessary to first decide on the quantity of interest q (mean, regression coefficient, predicted probability or first difference) that is to be computed. The complete data analysis model is run on each of the B-imputed datasets. The authors explained that the easiest way to combine the results from these models is to draw $1/B$ simulations of q from each of the B models and combine them into one set of B simulations (represented by Equation 2.4-62). Then, the standard simulation-based methods for the interpretation of single datasets can be used. It is possible to combine the model results directly and to use them as the multiple imputation estimate of this parameter, $\bar{q}$ and the average for B separate estimates, $q_b (b = 1, \ldots, B)$:

$$\bar{q} = \frac{1}{B} \sum_{b=1}^{B} q_b \tag{2.4-60}$$

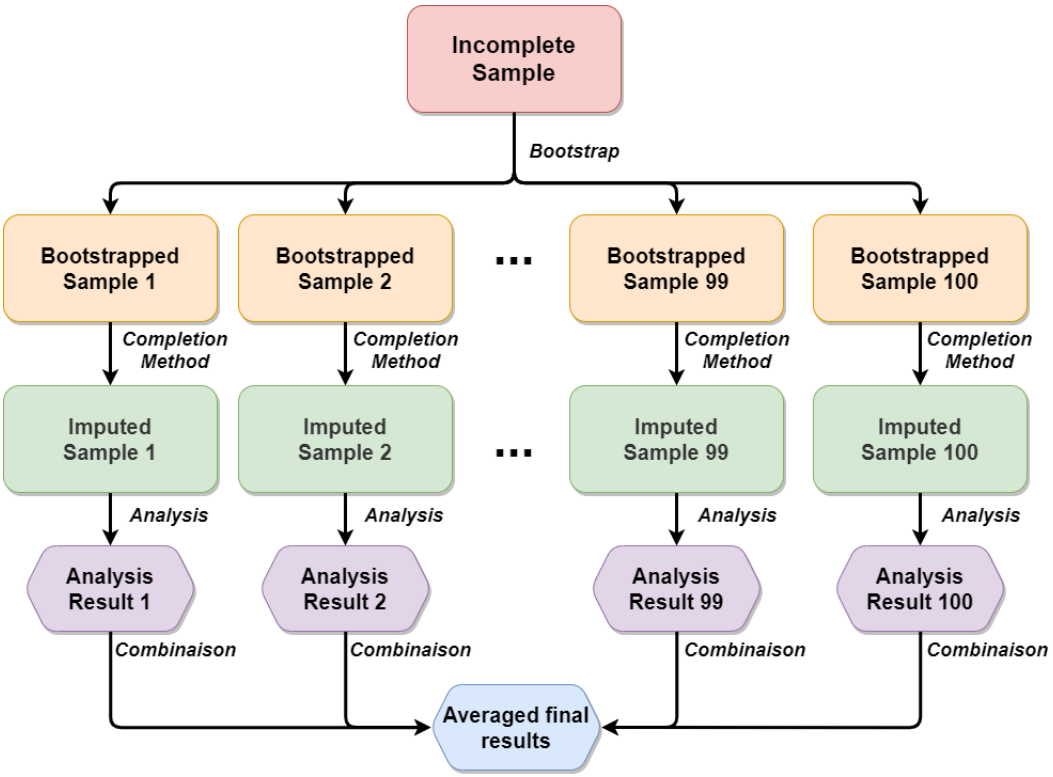
The variance of the point estimate is the sum of the averages of the estimated variance from within each completed dataset and the sample variance in the point estimates across the datasets (multiplied by a factor that corrects for bias because $B < \infty$). Let $\tilde{\sigma}^2_{q_b}$ denote the estimated variance (squared standard error) of q_b from the dataset b, and let σ^2_q be the sample variance across the B point estimates. The

standard error of the multiple imputation final estimate is the square root of

$$\begin{aligned}\tilde{\sigma}_q^2 &= \frac{1}{B}\sum_{b=1}^{B}\tilde{\sigma}_{q_b}^2 + (1+\frac{1}{B})\sigma_q^2 \\ &= \frac{1}{B}\sum_{b=1}^{B}\tilde{\sigma}_{q_b}^2 + \frac{(1+\frac{1}{B})}{(B-1)}\sum_{b=1}^{B}(q_b-\bar{q})^2\end{aligned} \tag{2.4-61}$$

Figure 2.4-26 represents the operation of the Amelia algorithm schematically, where the completion method used is the EM algorithm.

Fig. 2.4-26: Multiple imputation with 100 bootstrap samples. In the case of Amelia, the completion method is the EM algorithm.



Multiple imputation and bootstrap

Before proceeding, it is important to highlight the advantages of multiple imputation and bootstrap.Before proceeding, it is important to highlight the advantages of multiple imputation and bootstrap. The methodologies presented so far are simple imputation methods. In other words, they assign a single value to a single missing value. Linear interpolation, the Brownian bridge, K-NN, MSSA, iterative PCA models (presented in

Section 2.4.8) and even random forests entail attaching a unique value to each missing data. Simple imputation methods are advantageous because they are easier to execute and integrate into analyses. They provide complete databases that can be reused and offer the possibility of accessing more information than what is available initially. However, according to Little and Rubin [144], these methods treat the imputed data as if they were real and overstate their accuracy. Using a sample with 30% missing data, they showed that multiple imputation (with three imputation sets) gives the correct estimate in 95% of cases with a confidence level of 95%; simple imputation only gives the correct estimate in 85% of cases. In addition, a simple imputation cannot reflect the variability of a model that is due to missing data or the uncertainty of their imputation. Indeed, imputed data are considered observed values and do not account for the uncertainty of the prediction, which means that the standard deviation of the parameters that are calculated from these imputed data is underestimated. It is for this reason that multiple imputation is of interest here.

In 1978, Rubin [175] developed multiple imputation to solve the problems posed by simple imputation. Multiple imputation no longer aims at a single answer. It assumes that missing data are MAR and involves producing B complete datasets by imputing missing values B times through an appropriate model that incorporates random variation. The desired analysis can be performed on each dataset through complete data methods to obtain a collection of B individual estimates q_b with $(b = 1, \ldots, B)$. Finally, a single estimate $\bar{q}$ can be obtained by averaging the values of the parameter estimates over the B samples (as in Equation 2.4-60 of the Amelia algorithm). Thus, the standard error is computed as before (see Equation 2.4-62 or Equation 2.4-61). This standard error is a sum of two components. The first is the average of the within-imputation variance obtained from each of the completed datasets, and the second is the between-imputation variance of the complete-data parameter estimates. In other words, the within-variance corresponds to the variance of the q parameter estimate that is due to the model, while the between-variance corresponds to the variance from one imputation to another. Therefore, total variance accounts for both the variance that is due to the analysis model and the variance that is attributable to the imputation method.

$$\tilde{\sigma}_q^2 = \frac{1}{B}\sum_{b=1}^{B}\tilde{\sigma}_{q_b}^2 + \frac{(1+\frac{1}{B})}{(B-1)}\sum_{b=1}^{B}(q_b - \bar{q})^2 \tag{2.4-62}$$

It follows that multiple imputation makes the integration of imputation risk possible. It allows one to retain the advantages of simple imputation while correcting its disadvantages. Missing data can be imputed by integrating an error term. Moreover, the repetition of the imputation yields a better standard error of the estimators. Finally, multiple imputation is a very general procedure. It can be applied to any type of data and in any type of analysis. Little and Rubin [145] explain that multiple imputation has the obvious disadvantage of demanding more labor and storage space. However,

they point out that storage issues are trivial at the present time and that the analysis remains modest because it consists of repeating the same procedure for each imputed dataset.

Multiple imputation is analogous to the bagging procedure that was presented in the previous section (see Section 2.4.5). Both involve imputing missing data several times and aggregating the results. There are two main differences. The first is at the level of the samples to which the imputation method is applied. According to Rubin [177], multiple imputation consists of imputing the initial sample B times, bagging applies the imputation method to B samples from a bootstrap of the initial data. The strict definition given by Rubin [177] has evolved over time because multiple imputation has become a procedure to obtain B imputed datasets from the same original data or from a bootstrap. If the broad definition of multiple imputation is retained, there is no real difference between it and bagging, at least on this dimension.

The second difference concerns the aggregation step and the averaging of the data in particular. In the case of bagging, averaging is applied to all imputed data; in multiple imputation, it is the analysis estimators that are averaged. In other words, bagging averages B imputed datasets to obtain a single dataset and then calculates an analysis estimator, while multiple imputation generates an analysis estimator for each of the B imputed samples and averages them. For example, in the case of a VaR calculation, random forests provide a single imputed dataset, and it is possible to use it to obtain a single VaR. In contrast, multiple imputation like Amelia entails calculating a VaR for each of the imputed datasets before deducing an average. Obviously, a method integrating bagging is much more time-efficient than a multiple imputation method.

Another solution would be to construct a distribution of the data by considering all possible imputations (calculated by the multiple imputation method), weighting them so that the sum of the weights of all imputations for a missing data is equivalent to the weight of an observed data. In other words, if a series has n observations, then each data (observed or missing) would have a weight of $\frac{1}{n}$ in the distribution. In the context of missing data that is imputed by multiple imputation using B bootstrapped samples, each imputation would have a weight of $\frac{1}{B \times n}$ in the distribution. This method would have no impact on the computation of linear measures (typically the mean). However, it may impact non-linear measurements such as VaR, which is simply a quantile of a distribution.

In the case of the Amelia algorithm, a bootstrap step is integrated into multiple imputation. That step consists of a multiple random reproduction of data from an original sample whereby the lines are sampled with replacement. In 1994, Efron [73], a pioneer in the field, used bootstrap with missing data to perform multiple imputations that were based on the EM algorithm. Let $Y = (Y_1, \ldots, Y_n)$ be the data matrix with n observations and let Y_i contain the p-values of i-th observation of Y. Thus, each observation of a bootstrap sample $Y^{(b)}$ is obtained by drawing randomly from observed

values of Y with replacement.

This method has the advantage of simplicity, and it can be combined with others. It allows the integration of error terms into estimates and confidence intervals without making assumptions about the distribution of the data. Like the bagging technique presented earlier (see Section 2.4.5), it also allows results to be stabilized. As far as missing data are concerned, no assumptions about missingness patterns are necessary. Indeed, its application to uncertainly distributed data is very interesting. However, the simulations of multiple imputation and bootstrapping involve a random component in their simulation, making the imputation results non-replicable. Non-replicability is a concern for the Amelia algorithm and for multiple imputation methods in general. There is no means of ensuring that the figures presented to regulators do not reflect optimistic scenarios and that they do not understate the risk that a bank has assumed (see the end of Section 2.4.5 for more details). It is therefore necessary to check that a large number of bootstrapped samples have been used, which is not so obvious in the literature since Rubin [177] considers that $B = 10$ is sufficient.

EM algorithm versus Amelia

To determine whether the bootstrap has a genuinely positive impact on the imputation of missing data through the EM algorithm, it is possible to compare the results of the EM algorithm when it is used in isolation to those that obtain when it is used in tandem with a bootstrap (the Amelia model). The sample used in Section 2.4.4, which covered WTI and Brent crude prices between August 2015 and July 2020, was employed to this end. The two algorithms that were applied to it are the EM algorithm (from the Amelia function in the R package *Amelia*, which allows the bootstrap step to be removed) and Amelia.

As in Section 2.4.4, 100 samples were created from the two series. Then, 30% of the data (crude oil prices) were removed from the WTI series at random in each sub-period in order to be imputed by the EM algorithm and by the Amelia algorithm. The samples were exactly the same as those used in Section 2.4.4, which means that each sample has approximately 250 values (daily data) and that, on average, 127 values were randomly deleted (ranging from 115 to 135 deleted data). Comparisons were made on the basis of MAE and RMSE and then on the basis of a 1-day 99% VaR.

Since the Amelia algorithm is a multiple imputation method, it was decided that 100 samples of results would be calculated and that the final imputation would be the average of the 100 results. In the same way as in Section 2.4.4, the MAE (Table 2.4-16) and the RMSE (Table 2.4-17) between the returns of the original series and the returns from the series that were imputed (by EM and by the Amelia algorithm) were calculated for each of the 100 scenarios in order to observe the impact on extreme values. As far as MAE is concerned, it is clear that the EM algorithm always produces

an imputation that is further from the original series than the Amelia algorithm. It is possible to see from Table 2.4-16 that for all periods and for all 100 scenarios, all the MAEs calculated from the results of the Amelia algorithm are lower than those calculated from the results of the EM algorithm.

Tab. 2.4-16: Average MAE of returns between original series and imputed series (imputed by EM and Amelia)

Sample Period	$\overline{\boldsymbol{MAE}}_{EM}$	$\overline{\boldsymbol{MAE}}_{\boldsymbol{Amelia}}$
08.2015 - 07.2016	1.30%	0.88%
08.2016 - 07.2017	0.82%	0.57%
08.2017 - 07.2018	0.79%	0.55%
08.2018 - 07.2019	0.82%	0.58%
08.2019 - 07.2020	8.32%	3.03%

This table above presents the average over the 100 scenarios of the MAE between the returns of the original data and the returns of the imputed data, denoted by $\overline{MAE}_{EM}$ and $\overline{MAE}_{Amelia}$. For all periods, $\overline{MAE}_{Amelia}$ is lower than $\overline{MAE}_{EM}$. When a verification is performed on all scenarios, it emerges that returns imputed through the Amelia algorithm always provide a lower MAE than returns imputed through EM.

If MAE amplitudes are compared, it transpires that the Amelia method allows for imputations that are more precise by at least 40%. Still according to the MAE criteria, the Amelia algorithm gives more precise imputations by 47%, 44%, 44%, 41% and 175% in each of the periods.

Finally, there is a significant gap between $\overline{MAE}_{Amelia}$ and $\overline{MAE}_{EM}$ in the last period. The last period is 10 times as volatile as the previous period (due to the Covid-19 crisis). Therefore, the EM algorithm seems to run into considerable difficulties when it must impute the missing data because the MAE is 10 times higher in this period than in the previous period. However, owing to bagging, the Amelia algorithm enables the results to be stabilized and an MAE that is five times higher than that for the previous period to be obtained. Amelia, then, manages to follow the dynamics of the original series much better than EM even in the presence of extreme values. This said, the results show that both methods perform worse in more volatile periods.

Turning to the quadratic measurement (here, RMSE), Table 2.4-17 shows that the Amelia method always gives the best results.

Tab. 2.4-17: Average RMSE of returns between original series and imputed series (imputed by EM and Amelia)

Sample Period	$\overline{\boldsymbol{RMSE}}_{EM}$	$\overline{\boldsymbol{RMSE}}_{\boldsymbol{Amelia}}$
08.2015 - 07.2016	2.29%	1.64%
08.2016 - 07.2017	1.45%	1.03%
08.2017 - 07.2018	1.44%	1.05%
08.2018 - 07.2019	1.45%	1.05%
08.2019 - 07.2020	20.47%%	13.45%

As before, the average RMSE is around 40% higher when returns imputed from the EM algorithm are used (40%, 41%, 37%, 38% and 52%, respectively). Therefore, Amelia handles outliers and extreme values better than EM. In addition, calculating the imputed returns from Amelia rather than EM yields a lower RMSE in all of the scenarios. The sole exception is the last period, in which four scenarios provides a lower RMSE when EM returns are used. This said, the average RMSE results in the last period are at least about 13 times higher than those in the previous one for both EM and Amelia. The two RMSEs fluctuate in line with volatility. It would appear that neither method manages outliers well when volatility is high. Finally, the last period departs from the Gaussian framework on which the EM algorithm is based, so it is logical that the results are worse.

It is interesting to note that the Amelia algorithm performs better than the EM algorithm when the sample is less correlated. In the first four periods, the Amelia algorithm produces imputations that are between 41% and 47% more accurate than those of the EM algorithm (in terms of MAE). These four periods are the ones for which the series are the most correlated (the correlations are 72%, 69%, 50% and 70%). The last period is the one in which Amelia performs considerably better than EM (the MAE here is 175% lower) and also the one in which the series are the least correlated (the correlation is 42%). It would thus seem that, relative to the imputation accuracy of EM, that of Amelia increases when correlation decreases.

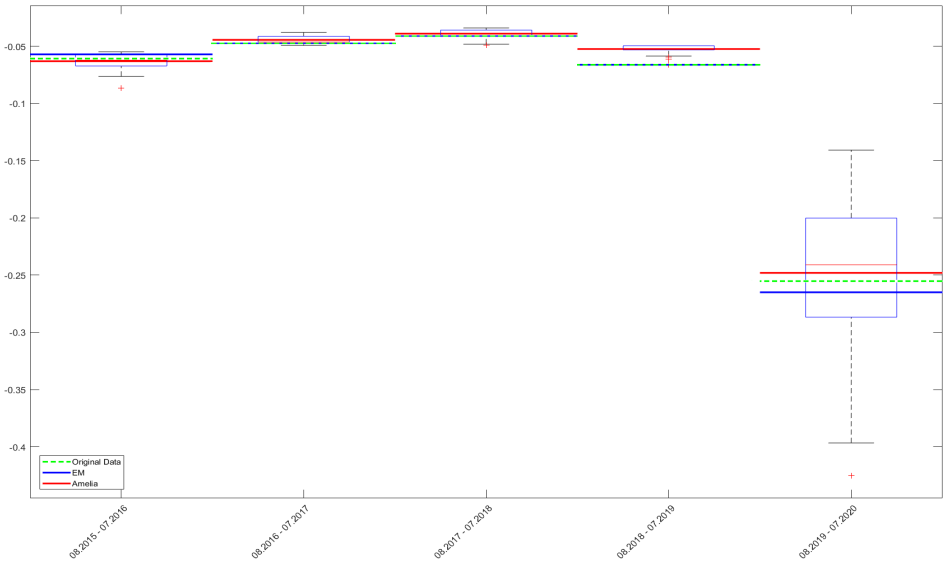
The Amelia algorithm outperforms the EM algorithm the most when the sample is uncorrelated. The EM algorithm attempts to estimate the distribution of the whole sample. The Amelia algorithm estimates the distribution of 100 bootstrap samples that are smaller, which increases precision because the estimation is less sensitive to extreme values and outliers.

It is important to remember that multiple imputation methods involve applying the method of analysis to each of the imputed samples before the results are aggregated. It is for this reason that the relationship between the choice of method and VaR is

of interest. The following results are those of the 1-day 99% VaR for one of the 100 scenarios that were used in the previous results, with 30% missing data in the WTI series imputed by the EM algorithm and by Amelia.

Figure 2.4-27 presents a comparison between the results for the VaR of the original series (green line), the VaR obtained from the data imputed by the EM algorithm (blue line) and the results for the Amelia-imputed data (the box plot represents the set of results and the red line the average VaR). The results cover all of the periods under observation.

Fig. 2.4-27: The 1-day 99% VaR of the fully observed crude oil WTI (green line) price returns, the 1-day 99% VaR computed from price returns imputed by the EM algorithm (blue line) and the 1-day 99% VaR computed from price returns imputed by the Amelia algorithm (the boxplot represents each imputation of VaR and the red line represents the average VaR) for all periods

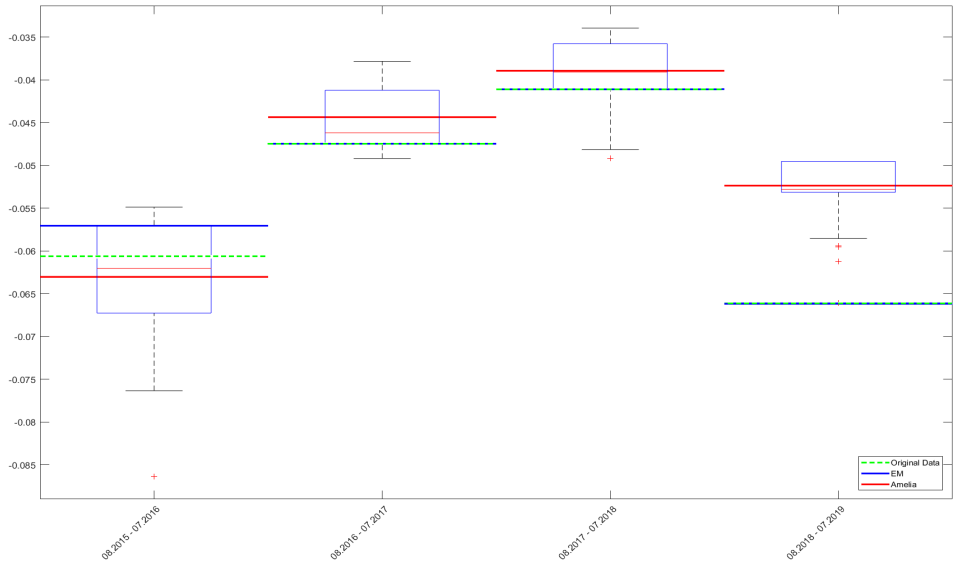


Since the last period is considerably more volatile than the others, the results for the first four are hardly observable in Figure 2.4-27. However, it is possible to see that even if the three VaRs for the last period are very close, the one obtained by the Amelia algorithm (which corresponds to the average of the VaR calculated from the imputed samples) comes closest to the true VaR. The true VaR is equal to -0.2552, and the average Amelia VaR is equal to -0.248. The EM VaR is equal to -0.2650, that is, the spread between the EM VaR and the true VaR is almost twice as high as that between the Amelia VaR and the true VaR. Furthermore, the boxplot shows that some imputed samples produce VaRs that are much more negative than the true one, hence

the importance of averaging the results.

Figure 2.4-28 represents the same results as Figure 2.4-27, but the last period is excluded.

Fig. 2.4-28: The 1-day 99% VaR of the fully observed crude oil WTI (green line) price returns, the 1-day 99% VaR computed from price returns imputed by the EM algorithm (blue line) and the 1-day 99% VaR computed from price returns imputed by the Amelia algorithm (the boxplot represent each imputation of VaR and the red line represents the average VaR), for the first four periods



In the first period, the true VaR lies between the VaR obtained by the EM imputations and the average of VaRs that are based on Amelia imputations. However, the Amelia algorithm (−0.0630) gives an average VaR that is closer to the true VaR (−0.0606) than the VaR calculated from the EM imputations (−0.0571). The latter seem to reflect a less pessimistic scenario. In the second and the third period (i.e., 08.2016-07.2017 and 08.2017-0.7.2018) the true VaRs correspond exactly to the EM imputations (−0.0393 and −0.0411, respectively), but the average Amelia-imputed VaRs remain close (−0.0444 and −0.0389, respectively). The EM algorithm was able to produce exactly the right VaR for the two periods, but it must not be forgotten that it is calculated from a single scenario (which seems to have been optimistic, in VaR terms) and that the results can differ considerably depending on the random draw that is used.

Finally, in the fourth period (08.2018-0.7.2019), the difference between the VaR from the Amelia-imputed data and the true VaR is greater. This can be explained by the much more negative skewness coefficient for this period, relative to the one in the Brent

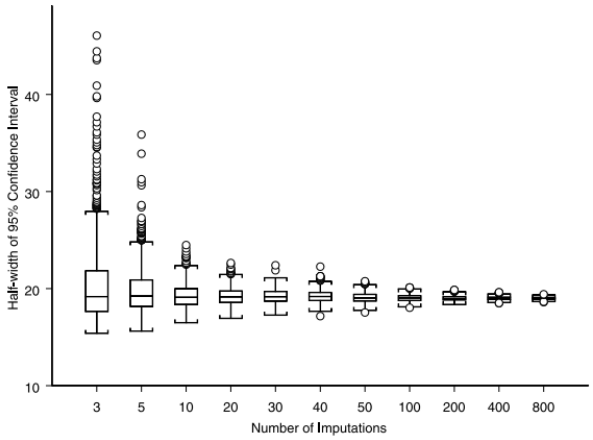
series (as seen in the presentation of the sample in Section 2.4.4). In other words, the WTI crude oil series for the fourth period is characterized by a left-tailed distribution, and it has a larger number of negative extreme values than the Brent series. The presence of missing data causes these extreme values to become less visible, leading to an estimated distribution that has a higher skewness than the original. The Amelia imputation gives a VaR of −0.0524. This result is very far from the true VaR, which is −0.0661. The EM algorithm provides a scenario which yields a VaR of −0.0662. However, the caveat from the preceding paragraph applies to this finding, too.

The repetition of the analysis does not allow the two methods to be ranked definitively in VaR terms. Everything depends on the trajectory of the EM imputation. It is not at all unusual for the imputation to give better results for some scenarios than for others. This case illustrates the problem of non-replicability that was presented in Section 2.4.5. This analysis could nevertheless be repeated to see, on average, the performance of the two methods. Given the results presented in the Table 2.4-16 and the Table 2.4-17, Amelia should have an advantage over the EM.

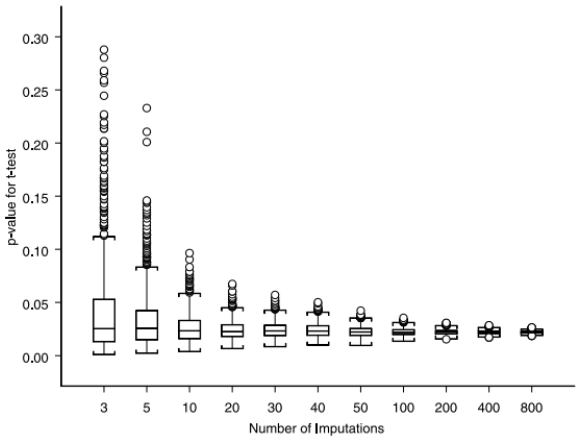
The use of multiple imputation entails setting the number of imputations and thus the number of complete samples that are obtained at the end of the imputation. Multiple imputations make it possible to account for the uncertainty around missing data. This is not true of simple imputation, where only a single estimate is available. It is normal to ask how many imputations are necessary to achieve sufficient quality. It may therefore be possible to impute missing data not through a single possibility but through an interval that includes all possible imputations, similarly to Monte Carlo forecasting simulations. However, a large number of Monte Carlo simulations is typically expected. The literature does not necessarily indicate that the same is true of multiple imputations. According to Rubin [177], between two and 10 imputations are necessary for "a modest fraction of missing information." Schafer recommends a larger number of imputations if there are more missing data. However, the literature does not indicate a specific number. Moreover, algorithm documentations are often presented with relatively low imputation numbers, suggesting that the methods are performing well even with low imputation. For this reason, Bodner [36] having reconsidered the matter, suggests using a much larger number of imputations than Rubin's recommended range [177].

Bodner [36] shows that the number of imputations can affect quality. He uses cholesterol data (the same as the ones Schafer [181] uses in his book) in order to show that the number of imputations affects the confidence interval half-widths of the null hypothesis significance test, their p values and the fraction of missing information. According to Allison [5],the last criterion refers to "how much information is lost about each coefficient because of missing data." Figure 2.4-29 shows the fluctuation of the three criteria according to the number of imputations used over 1,000 replications.

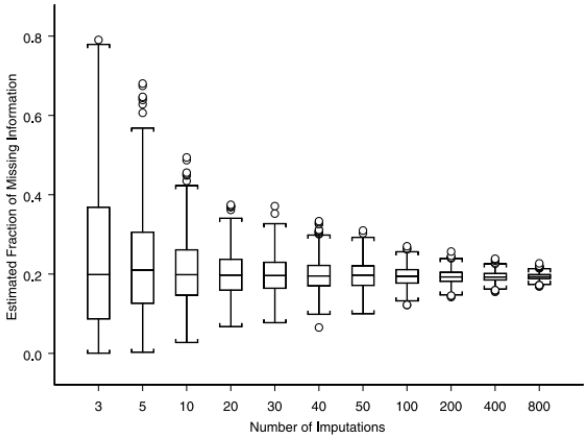
Fig. 2.4-29: Boxplot of 95% confidence interval half-widths, null hypothesis significance test p values, and estimated fractions of missing information as a function of the number of imputations (Source: Bodner, 2008 [36])



(a) 95% confidence interval half-widths



(b) Null hypothesis significance test p values



(c) Estimated fraction of missing information

The confidence interval half-widths of the null hypothesis significance test, their p values and the fraction of missing information are much more variable when the number of imputations is low than when it is high.

Bodner [36] shows that the median of the three criteria remains relatively stable whatever the number of imputations used. On the other hand, the three criteria are much more variable when the number of imputations is low than when it is high. Indeed, he finds that the half-amplitude of the confidence interval can vary threefold from one replication to another, that the maximum p-value is 280 times higher than the minimum

and that this tendency is even more pronounced for the fraction of missing information, whose maximum value is 1,975 times higher than the minimum..

Thus, Bodner [36] shows that using more than the 10 imputations recommended by Rubin[177] is better. He adds that the computation time is not necessarily longer because the imputation is usually based on a random draw within a specific law. Thus, in the same way as a Monte Carlo simulation, a large number of imputations are necessary to stabilize the results. Using a low number of imputations can induce high variability in the results because they seem to stabilize with 100 imputations. Therefore, the number of imputations will be fixed at 100 for all of the multiple imputation methods that are used in this PhD thesis.

Parameter initialization

The last question that remains to be answered concerns the initialization of parameters. The principal disadvantage of the EM algorithm is its sensitivity to starting parameters. In this regard, Little and Rubin [145] present four straightforward possibilities. Honaker, Joseph, King, Scheve and Sigh [112] implemented the first solution as a default option in the Amelia algorithm.

1. It is possible to use a complete-case solution, that is, to only use observations in which all the variables are present (like the listwise deletion presented in Section 2.2.2). This method is the simplest because it can be applied without modifications. However, its employment leads to important information being lost (precision diminishes and bias increases, as seen in Section 1.2.2). The complete-case solution provides consistent parameter estimates if the data are MCAR and if there are more complete observations than variables.

2. It is possible to use one of the available-case solutions, that is, to use all available observations of a variable. This method allows all available data to be used, but it involves univariate analysis because the available sample changes with each variable. The use of this procedure leads to incomparable variables. It can provide an estimated covariance matrix that is not definite and positive, which can be a problem in the first iteration.

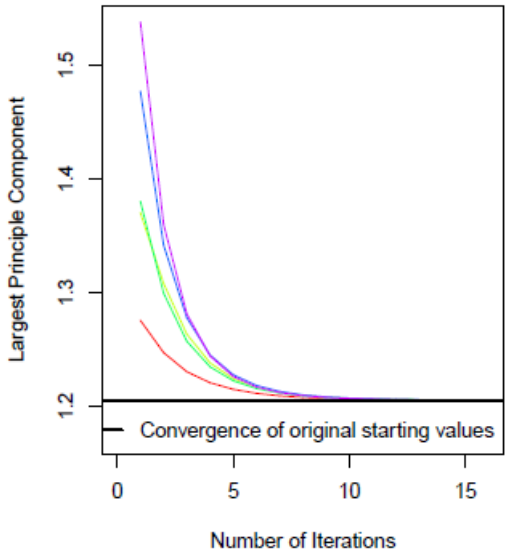
3. It is possible to form the sample mean and the covariance matrix of the data that are filled in through one of the single-imputation methods. This method entails treating the variables individually and applying a univariate imputation, such as a mean imputation, a linear interpolation, a regression imputation, a Brownian bridge and so forth, to them. In general, this method leads to inconsistent estimates of the covariance matrix, but semidefinite positive estimates are generally usable as starting values. For example, a valid solution would be to estimate the

pairwise variance-covariance matrix and then regularize it, as shown in Section 2.2.2, in order to make it semidefinite positive.

4. It is possible to form means and variances from the observed values of each variable and to set all starting correlations to 0. This methodology also leads to an inconsistent covariance matrix, but definite positive estimates are workable initial parameters.

None of these solutions account for the fact that the EM algorithm can converge to a local maximum but not necessarily to the global maximum. If the initial parameters are set close to a local maximum, the EM algorithm is highly likely to converge to it, despite the existence of a global maximum farther away. The problem may be overcome by testing the algorithm with several different and scattered starting values in order to verify their convergence. Scenarios that, for the most part, converge to the same value confirm that the convergence value probably corresponds to the global maximum. However, if this test has to be carried out for each bootstrap sample, computation time is likely to mushroom. For this reason, Honaker, Joseph, King, Scheve and Sigh [112]leave open the possibility of performing these steps as an index test, whereby the user can check the convergence of the parameters if they want to. This approach enables the convergence of different starting values to be tested for a given sample. The convergences of the parameters for five different scenarios are represented in Figure 2.4-30. Evidently, the starting values converge to the same maximum (colored curves). The values in question are obtained through listwise deletion (the first possibility of Little and Rubin [145]).

Fig. 2.4-30: Overdispersion diagnostic where all scenarios with different starting values are converging to the same maximum (Source: Honaker, Joseph, King, Scheve and Sigh, 2002 [112])



For these five different scenarios, starting values converge to the same maximum with 10 iterations.

Amelia imputation applied to financial data

In 2013, Bauer, Angelini and Denev [27] applied different imputation methods, including the Amelia algorithm, to CDS data that include missing values. They performed Little's test on these data and concluded that the MCAR hypothesis was rejected in most cases, with a very low p value. The multivariate normal distribution assumption of Little's test (Section 2.1.2) provides a partial explanation. For a sufficiently large sample, the distribution is close to an multivariate normal distribution, and the test is almost always significant. The authors therefore decided to further investigate the causes of the missing data. Illiquidity, that is, insufficient market trading data, made it impossible to produce a reliable price quotation. They concluded that the data are not MNAR, so they must be MAR or MCAR. Therefore, assuming that the data are MCAR appears to be sound.

They then classified the data into five clusters:

1. relatively small fraction of missing data;

2. missing values, mainly for long maturities and with relatively short and alternating stretches of consecutive missing values;

3. missing values in long streaks for longer maturities;

4. patterns with considerable amount of missing data and substantial variation;

5. patterns with large amount of missing data with uniform long stretches, often covering all maturities.

They decide to use only the first three clusters in their data, which are also the ones with the least amount of missing data. They found that the overwhelming majority of the patterns are in the first two clusters. For the first cluster, the authors stated that the missing data are MCAR (due to a very low proportion of missing data), but they could not make the same observation about the second and the third cluster. Before applying different imputation methods, the missing data of these three clusters are imputed by linear interpolation, in order to work on complete samples without introducing a particular bias. More precisely, they decided to create 200 samples of tickers containing less than 1% missing data and linearly interpolate them. Now that their samples are complete, they can remove data. Contrary to the standard approach in the literature, they did not remove data from the sample at random. Instead, they reproduced a realistic missing data pattern observed on CDS data. Then, the following imputation methods were applied:

- Amelia multiple imputation based on an EM algorithm and bootstrapped data (see the beginning of Section 2.4.6) with 5 imputations;
- data interpolation with empirical orthogonal functions (DINEOF), which allows the detection of the number of statistically significant EOFs by a cross-validation procedure for complete and incomplete datasets [29];
- Multiple imputation by chained equations (MICE), a multiple imputation based on predictive mean matching and regression methods (see Section 2.4.7) with a window length of 10 for the first cluster and a window length of 40 for the second and the third cluster;
- Random forests, which is based on multiple regression trees, and the MissForest implementation (see Section 2.4.5);
- Multivariate Singular Spectrum Analysis (MSSA), which entails decomposing the chronological series into three components, trend, harmony and noise, and reconstructing it (see Section 2.4.4) with 5 imputations.

Each algorithm was applied to 200 missing data scenarios and the mean relative deviation (MRD) was computed as follows:

$$MRD = \frac{1}{n_M} \sum_{j=1}^{p} \sum_{i \in M_j} \left| \frac{y_{ij}^* - y_{ij}}{y_{ij}^*} \right| \tag{2.4-63}$$

with M_j the set of missing observations for component i, $n_M = \sum_p |M_j|$, and y^*_{ij} as the true value and y_{ij} as the imputed value. Table 2.4-18 shows the summary statistics for the MRD (also known as mean absolute percentage error) results, for each clusters:

Tab. 2.4-18: Summary statistics for MRD metrics for each cluster (Source: Bauer, Angelini and Denev, 2017 [27])

	Amelia	**DINEOF**	**Mice**	**Random forests**	**MSSA**
Cluster 1					
Mean	0.017	0.024	0.031	0.019	0.016
Std Dev	0.010	0.019	0.032	0.000	0.001
Min	0.002	0.001	0.002	0.000	0.001
Max	0.057	0.141	0.374	0.077	0.102
Cluster 2					
Mean	0.035	0.064	0.052	0.046	0.048
Std Dev	0.035	0.053	0.056	0.057	0.056
Min	0.005	0.011	0.009	0.002	0.005
Max	0.328	0.384	0.497	0.483	0.492
Cluster 3					
Mean	0.093	0.141	0.111	0.098	0.128
Std Dev	0.135	0.121	0.158	0.103	0.125
Min	0.009	0.012	0.010	0.014	0.008
Max	0.980	0.728	1.522	0.650	0.739

For cluster 2, the mean absolute percentage error from Amelia imputation is equal to 0.035, while that from random forests is equal to 0.046.

The results obtained for the first cluster seem to be accurate (generally between 1% and 3%). The best results are obtained by Amelia and MSSA, even if the other algorithms produce similar results, because there are relatively few missing data in the first cluster (1.5% on average) and because successive data are rare. Concerning Cluster 2, which has 13% missing data on average, the results are less accurate (MRD between 2% and 7%). Amelia gives the best results, followed by random forests and MSSA. Finally, the third cluster is composed of 19% missing data on average, a good part of which are successive. Figure 3 shows that, on average, the MRD is between 9% and 14%. Amelia performs best again, followed by random forests. For all three clusters, the DINEOF method produced the worst results, followed by MICE.

In the second and the third cluster, data are sometimes missing for all maturities in an observation (20 and 23 patterns, respectively), that is, a row may only contain missing data. Imputation is particularly difficult in these cases. The authors recomputed their statistics to exclude them, which may have biased the results (see Table 2.4-19). The results in Table 2.4-19 show the relative effectiveness of the Amelia method.

Tab. 2.4-19: Summary statistics for second and third cluster where patterns were filtered out if entire rows were missing (Source: Bauer, Angelini and Denev, 2017 [27])

	Amelia	**DINEOF**	**Mice**	**Random forests**	**MSSA**
Cluster 2					
Mean	0.028	0.064	0.046	0.037	0.041
Std Dev	0.015	0.054	0.052	0.032	0.041
Min	0.005	0.011	0.009	0.002	0.005
Max	0.104	0.384	0.497	0.256	0.342
Cluster 3					
Mean	0.061	0.135	0.095	0.092	0.126
Std Dev	0.084	0.124	0.155	0.104	0.129
Min	0.009	0.012	0.010	0.014	0.008
Max	0.705	0.728	1.522	0.650	0.739

For cluster 2, the mean absolute percentage error from Amelia imputation is equal to 0.028, while that from random forests is equal to 0.037.

Bauer, Angelini and Denev [27] concluded by explaining that the Amelia algorithm is the most suitable method to complete the missing data in their sample, despite the data not being normally distributed, a fundamental assumption of the Amelia framework. Amelia is used widely, but it is based on the assumption of data normality. Sometimes, that assumption is not met. It is not at all obvious that its application to a non-Gaussian sample would yield comprehensible results.

2.4.7 Multivariate imputation by chained equations

Just like the Amelia algorithm, multivariate imputation by chained equations (MICE) is a multiple imputation method. In 1987, Rubin [177] explained that each missing value should be completed by B values, creating B imputed datasets. Then, each imputed dataset is analyzed through standard complete-data procedures that ignore the distinction between real and imputed values. Finally, the resultant B analyses are combined. The MICE imputation algorithm was created in 1999 by Van Buuren [206], and it assumes that the user specifies a conditional distribution for the missing data on the basis of the observed data for each missing data.

MICE algorithm

Van Buuren [54] defined the MICE algorithm as follows: let $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p)$ be a set of p incomplete variables. The observed and missing parts of $\boldsymbol{Y}_j$ (with $j = 1, \ldots, p$) are denoted by $\boldsymbol{Y}_j^{obs}$ and $\boldsymbol{Y}_j^{miss}$. Therefore, it is possible to split the observed and missing data of $\boldsymbol{Y}$ as $\boldsymbol{Y}_j^{obs} = (\boldsymbol{Y}_1^{obs}, \ldots, \boldsymbol{Y}_p^{obs})$ and $\boldsymbol{Y}_j^{miss} = (\boldsymbol{Y}_1^{miss}, \ldots, \boldsymbol{Y}_p^{miss})$. The number of imputations is equal to $B \geq 1$, and $\boldsymbol{Y}^{(b)}$ is the b-th imputed dataset, where $b = 1, \ldots, B$. Unlike the previous algorithms, B does not correspond to the number of bootstrapped samples, but simply to the number of imputed samples. The MICE algorithm (with

PMM) does not include a bootstrap step. Let $\boldsymbol{Y}_{-j} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{j-1}, \boldsymbol{Y}_{j-1}, \ldots, \boldsymbol{Y}_p)$ be the collection of $\boldsymbol{Y}$ with $\boldsymbol{Y}_j$ excluded. Finally, let Q be the quantity of scientific interest (for example, a regression coefficient), generally represented as a multivariate vector.

The analysis starts from an observed and an incomplete dataset. It is not possible to estimate Q without making unrealistic assumptions about the missing data. The general framework of multiple imputation allows missing data to be replaced with several plausible values that are drawn from a distribution modeled specifically for each data that must be completed. This process is the same as the one shown in Figure 2.4-26 (in Section 2.4.6), except that the completion method is Mice with PMM. The multiple imputed samples are based on the same observed data.

The second step is to estimate Q for each imputed dataset. The estimates $\hat{Q}^{(1)}, \ldots, \hat{Q}^{(B)}$ differ from each other because of the B different imputed datasets. The last step is to pool the B estimates $\hat{Q}^{(1)}, \ldots, \hat{Q}^{(B)}$ into a single one, which is denoted by $\bar{Q}$, and to estimate its variance.

The procedure is therefore the same as that of the Amelia algorithm (see Section 2.4.6), as both are multiple imputation methods. The main differences are that Amelia includes a bootstrap step and the EM algorithm, while MICE allows the user to choose among many imputation methods, but no bootstrap step is performed with PMM. It is even possible to specify a different imputation method for each variable of the data matrix.

According to Van Buuren [54], the chained equations technique is used in order to address the problems posed by the real complexity of the data. Therefore, it is necessary to specify the imputation model separately for each column. Let the hypothetically complete data $\boldsymbol{Y}$ be a partially observed random sample from the p-variable multivariate distribution $f(\boldsymbol{Y}|\theta)$, where the multivariate distribution of $\boldsymbol{Y}$ is completely specified by θ (a vector of unknown parameters). The posterior distribution of θ can be obtained by using chained equations as follows:

$$\begin{aligned} & f(\boldsymbol{Y}_1|\boldsymbol{Y}_{-1}, \theta_1) \\ & \vdots \\ & f(\boldsymbol{Y}_p|\boldsymbol{Y}_{-p}, \theta_p) \end{aligned} \tag{2.4-64}$$

Van Buuren [54] remarked that the parameters $\theta_1, \ldots, \theta_p$ are specific to the corresponding conditional densities and that they are not necessarily the product of the factorization of the true Bayesian joint distribution $f(\boldsymbol{Y}|\theta)$. The MICE algorithm therefore includes the Gibbs sampler algorithm, which is a Bayesian simulation technique that samples from conditional distributions in order to obtain samples from the joint distribution. The Gibbs sampler can be written as follows:

$$
\begin{aligned}
\theta_1^{*(t)} &\sim f(\theta_1|\boldsymbol{Y}_1^{obs}, \boldsymbol{Y}_2^{(t-1)}, \ldots, \boldsymbol{Y}_p^{(t-1)}) \\
\boldsymbol{Y}_1^{*(t)} &\sim f(\boldsymbol{Y}_1|\boldsymbol{Y}_1^{obs}, \boldsymbol{Y}_2^{(t-1)}, \ldots, \boldsymbol{Y}_p^{(t-1)}, \theta_1^{*(t)}) \\
&\vdots \\
\theta_p^{*(t)} &\sim f(\theta_p|\boldsymbol{Y}_p^{obs}, \boldsymbol{Y}_1^{(t)}, \ldots, \boldsymbol{Y}_{p-1}^{(t)}) \\
\boldsymbol{Y}_p^{*(t)} &\sim f(\boldsymbol{Y}_p|\boldsymbol{Y}_p^{obs}, \boldsymbol{Y}_1^{(t)}, \ldots, \boldsymbol{Y}_p^{(t)}, \theta_p^{*(t)}),
\end{aligned}
\tag{2.4-65}
$$

where $\boldsymbol{Y}_j^{(t)} = (\boldsymbol{Y}_j^{obs}, \boldsymbol{Y}_j^{*(t)})$ is the j-th imputed variable at iteration t.

The term "chained equations" refers to the possibility of implementing the Gibbs sampler as a concatenation of univariate procedures to fill out the missing data. Van Buuren [54] noted that the convergence can be rapid, unlike in many Markov chain Monte Carlo (MCMC) methods, and that it can be achieved with relatively few iterations. It is a semiparametric method that involves imputing missing data by randomly drawing an observed value from a group of donors that comprise complete observations (non-parametric part). The group of donors is based on a proximity criterion that is calculated from the errors of a linear regression model (parametric part).

Predictive mean matching

There are 24 imputation techniques in the MICE algorithm, of which 18 are applicable to continuous data. It is possible to impute missing data by simple methods, such as mean imputation or linear regression, but also by more sophisticated ones, such as predictive mean matching or random forests. The method that is used in this PhD thesis, predictive mean matching (PMM), is implemented by default for continuous data. This method was first proposed in 1986 by Rubin [176]and then developed in the context of missing data in 1988 by Little [143].

Let $\boldsymbol{Y} = (\boldsymbol{Y}^{obs}, \boldsymbol{Y}^{miss})$ be an incomplete sample of n observations, and let $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ be a set of p variables that are observed fully, where $\boldsymbol{X}^{obs}$ and $\boldsymbol{X}^{miss}$ correspond to the observations with the same partition of $\boldsymbol{Y}^{obs}$ and $\boldsymbol{Y}^{miss}$, respectively. In addition, n^{obs} is the set of indices where $\boldsymbol{Y}$ is observed, and n^{miss} is the set of indices where $\boldsymbol{Y}$ is missing. The PMM algorithm involves six steps:

1. Estimation of $\hat{\beta}$, $\hat{\sigma}$ and $\hat{\varepsilon}$ by a linear regression of $\boldsymbol{Y}^{obs}$ given $\boldsymbol{X}^{obs}$ by means of ordinary least squares: $\boldsymbol{Y}^{obs} = \boldsymbol{X}^{obs^T}\beta + \varepsilon$;
2. Computation of σ^{*2} as $\sigma^{*2} = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{A}$ by drawing A from $\chi^2_{\#n^{obs}-p}$;
3. Drawing β^* from $\mathcal{N}(\hat{\beta}, \sigma^{*2}(\boldsymbol{X}^{obs^T}\boldsymbol{X}^{obs})^{-1})$;
4. Computing $\hat{\boldsymbol{Y}}^{obs} = \boldsymbol{X}^{obs^T}\hat{\beta}$ and $\hat{\boldsymbol{Y}}^{miss} = \boldsymbol{X}^{miss^T}\beta^*$;

5. Computing a matrix Δ such that $\Delta(i,j) = |\hat{\boldsymbol{Y}}_i^{obs} - \hat{\boldsymbol{Y}}_j^{miss}|$ with $i \in n^{obs}$ and $j \in n^{miss}$;

6. Randomly drawing one of the d donors for each value of $\boldsymbol{Y}^{miss}$, whereby each observation j (with $j \in n^{miss}$) is associated with d donors that correspond to d observations of $\boldsymbol{Y}^{obs}$ whose indices are given by the same indices as the d smallest values in the vector $\Delta(i,j)$, where $i \in n^{obs}$.

The multiple imputation is obtained by repeating this algorithm B times. As already said, the MICE algorithm with PMM do no include bootstrap. The B different imputed samples are due to the random components from PMM. The last step of the PMM algorithm integrates a random draw, which makes the method non-replicable. It faces the same problems of non-replicablility as random forests (see Section 2.4.5).

Choice of proximity criterion and number of donors

The intuition of this method is close to that of K-NN (see Section 2.4.3). In both algorithms, missing data is imputed from nearby data. However, the techniques are different. The main differences are the orientation of the imputation, the proximity criterion and the final imputation. The logic of K-NN is latitudinal because information about the missing data is sought in other variables; in PMM, it is longitudinal because the imputation information sourced from other observations. The proximity criterion for the K-NN method is a distance measure, whereas the PMM method compares the differences between the forecasts that result from a linear regression. In addition, the final imputation of the K-NN model is an inverse distance-weighted mean of the nearest k neighbors. The PMM method consists of a random draw from the observations of the donor d. However, in both cases, it is necessary to choose the number of neighbors or donors to be used for imputation. It would be possible to use cross validation (see Section 2.4.3), like in the K-NN method, but computation time would be strongly affected.

Due to its longitudinal operation (drawing from donors), the PMM via MICE method is less sensitive to heteroskedasticity, to deviations from normality and to non-linear relationships between the data. The interest of the PMM method lies in the fact that the imputed data are observed in the sample. Therefore, the value of d should not be too small. Otherwise, the imputations would be too close to each other because of the repetition of values. At the same time, d should not be too large: a bias might result from the increase in the probability of a mismatch. Schenker and Taylor [183] compare results with $d = 5$, $d = 10$ and an adaptive method (d varies depending on the data) and find that they are similar. Van Buuren [205] set the default number of donors in the *MICE* function (in R software) at 5. This value is based on the results from Table 2.4-20. In a small study, he compared the impact of the number of donors

on the performance of the PMM method and the properties of the β_1 parameter in particular, for two different sample sizes.

Tab. 2.4-20: Properties of β_1 under multiple imputation by PMM with $B = 5$, 50% MCAR data and for sample sizes of 50 and 1,000 (Source: Van Buuren, 2018 [205])

Method	d	Bias	% Bias	Coverage	CI Width	RMSE
Missing y, $n = 50$						
PMM	1	0.016	5.4	0.884	0.252	0.071
PMM	3	0.028	9.7	0.890	0.242	0.070
PMM	5	0.039	13.6	0.876	0.241	0.075
PMM	10	0.065	22.4	0.804	0.245	0.089
Missing x, $n = 50$						
PMM	1	-0.002	0.8	0.916	0.223	0.063
PMM	3	0.002	0.9	0.931	0.228	0.061
PMM	5	0.008	2.8	0.938	0.237	0.062
PMM	10	0.028	9.6	0.946	0.261	0.067
Listwise deletion		0.000	0.0	0.946	0.251	0.063
Missing y, $n = 1000$						
PMM	1	0.001	0.2	0.929	0.056	0.014
PMM	3	0.001	0.4	0.950	0.056	0.013
PMM	5	0.002	0.6	0.951	0.055	0.013
PMM	10	0.003	1.2	0.932	0.054	0.013
Missing x, $n = 1000$						
PMM	1	0.000	0.2	0.926	0.041	0.011
PMM	3	0.000	0.1	0.933	0.041	0.011
PMM	5	0.000	0.1	0.937	0.042	0.011
PMM	10	0.000	0.1	0.928	0.042	0.011
Listwise deletion		0.000	0.1	0.955	0.050	0.012

CI: confidence interval

For a sample size at 50, where y is missing and with a PMM method with 1 donor, the bias is equal to 0.016 (5.4%), the coverage is equal to 0.884, the confidence interval width is equal to 0.252 and the RMSE is 0.071.

Van Buuren [205] noted that bias increases with the number of donors. The strongest bias is observed for high values of d. This relationship is even more pronounced when the sample is small. He also found that the bias is smaller when the missing data are in $\boldsymbol{X}$ than when they are in $\boldsymbol{Y}$. He added that when the sample is large, $d = 5$ seems to be a suitable parameter, even though slight undercoverage can be observed when the missing data is in $\boldsymbol{X}$.

The main flaw of the PMM method is that there is a risk that donor values might be duplicated. Missing data can end up with the same donors on multiple occasions, so it is likely that the final imputation will be the same. The problem manifests more clearly when the number of missing data is large in comparison to the number of observed data. The fewer the observed data, the less choice there is between donors.

Application of MICE imputation in the literature

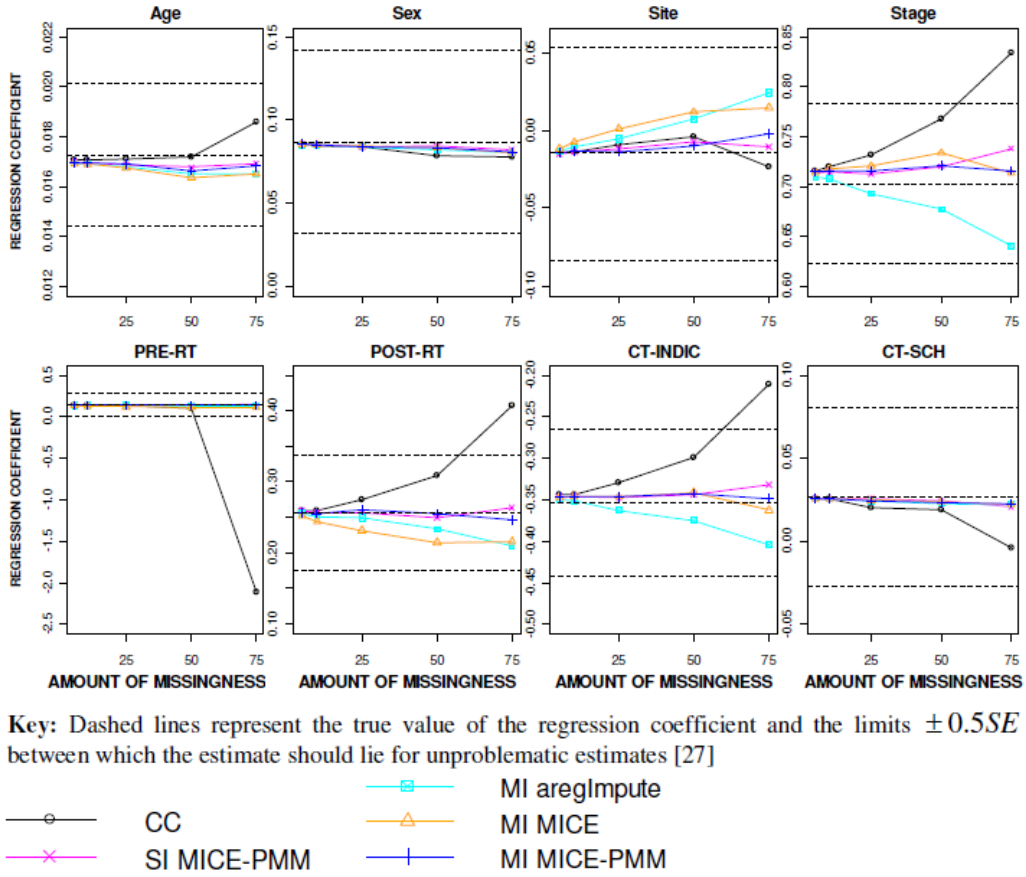
In 2010, Marshall, Altman and Holder [149] used the MICE algorithm with PMM and other imputation methods to investigate the effects of different missing data methods on the performance of a medical model. They used empirical colorectal cancer data. The sample is large because it includes complete data from May 1994 to September 2003 for 7,507 patients who received chemotherapy. Eight characteristics (e.g., age, sex and stage) are listed. A re-sampling step was executed. In that step, each new sample was made up of 100 randomly drawn cases from the initial study (with replacement). They used 500 replications. In each, missing data were injected artificially through a MAR mechanism and in proportions ranging from 5% to 75%. Marshall, Altman and Holder's [149] comparative study involved applying a medical model (the Cox proportional hazards model) to data obtained from the following methods:

- CC: only complete cases, specifically based on listwise deletion data (see Section 2.2.2);
- SI: single imputation with predictive mean matching (from the library *mice* in R);
- MI-aregImpute: multiple imputation that fits separate flexible additive imputation models to each incomplete variable, with predictive mean matching (from the library *Hmisc* in R). The difference from the MICE algorithm is that I-aregImpute takes random selections of missing elements instead of proceeding across columns;
- MI-MICE: multiple imputation with linear or logistic regression according, in line with the data (from the package *mice* in R);
- MI-MICE-PMM: multiple imputation with predictive mean matching (from the package *mice* in R).

Once the data were imputed, Marshall, Altman and Holder [149] compare the methods by reference to five criteria: regression coefficient estimates from the medical model, the standard error of the regression coefficient estimates, coverage, the significance of variables in the prognostic model and model performance measures.

Concerning the regression coefficient estimates, the CC method unsurprisingly produces the most unstable results. Figure 2.4-31 shows the regression coefficient estimates for each of the eight variables. Evidently, the methods that produce the lowest biased estimates are SI and MI-MICE-PMM. Apart from the CC method, the MI-MICE method produces the highest bias.

Fig. 2.4-31: Regression coefficient estimates after the application of each missing data method to increasing percentages of MAR missingness (Source: Marshall, Altman and Holder, 2010 [149])



The results closest to those of the original series should be close to the center dashed line, and remain within the confidence interval (dashed lines above and below).

As far as standard error is concerned, the CC method also gives the worst results. The results of the multiple imputation methods are similar. The MI-MICE-PMM method, which gives standard errors estimates close to those of the complete dataset, is an exception. The SI method has the worst coverage, at about 70%. Coverage is not below 90% for any of the other methods. Very few variables remain significant, even when there are few missing data points. For the CC method, only two out of eight remain. The last criterion concerns performance measures that are specific to the medical field. The authors compared the methods by reference to Negelkerke's R^2 (an indicator of prediction quality), similar prognostic separation value and predicted survival probabilities. For all of these criteria, the methods obtain very comparable and above all constant results, despite an increase in the proportion of missing data. The

CC method is an exception. It is the least satisfactory method, even if its Negelkerke R^2 increases with the proportion of missing data).

Marshall, Altman and Holder [149] concluded that the CC method (based on listwise deletion) is satisfactory up to a maximum of 10% missing data. Beyond that threshold, results are affected strongly. Likewise, they did not recommend the use of SI when the 10% threshold is exceeded. Multiple imputation emerged as the most satisfactory solution, even when 50% of the data are missing. According to the authors, MI-MICE-PMM outperforms all other multiple imputation methods.

In the same year, Marshall, Altman, Royston and Holder [150] conducted a similar study. They used a different data sample and added other imputation methods to their comparative analysis. That study uses German breast cancer data, which is composed of eight variables (e.g., age, tumor size and tumor stage). The study is based on 1,000 simulations. Each simulation comprises 1,000 observations. The medical model from the earlier article, the Cox proportional hazards model, was applied, and the results obtained from the original data were compared to the results that were obtained after missing data (ranging from 5% to 75%) had been injected for four variables. Continuous variables were truncated by using only their upper observed limits in order to produce realistic values. Incomplete continuous variables were transformed through a logarithmic function. The methods that were compared were as follows:

- CC: only complete cases, based on listwise deletion data (see Section 2.2.2)
- SI: single imputation with predictive mean matching (from the library *mice* in R)
- MI-aregImpute: multiple imputation that fits separate flexible additive imputation models to each incomplete variable, with predictive mean matching (from the library Hmisc in R). The difference from the MICE algorithm is that MI-aregImpute takes random selections of missing elements instead of proceeding across columns.
- MI-MICE: multiple imputation with linear or logistic regression, in line with the data (from the package *mice* in R)
- MI-MICE-PMM: multiple imputation with predictive mean matching (from the package *mice* in R)
- MI-MICE-PMM-no transformation: same as MI-MICE-PMM but without using a logarithmic transformation on continuous variables (from the library *mice* in R)
- MI-NORM: data augmentation under a normal-inverted Wishart prior (from the library *norm* in R)

- MI-MIX: Markov chain Monte Carlo method for generating posterior draws of the parameters of the unrestricted general location model given a matrix of incomplete mixed data (from the library *mix* in R)

- MI-MIX-no truncating: same as MI-MIX but using non-truncated continuous data (from the library *mix* in R)

The results presented below use the same comparison criteria as those used in the article by Marshall, Altman and Holder [149]. When the missing data follow a MAR scheme, the regression coefficient estimates seem to be reproduced well for all the variables with a simple method such as the CC method, which gives results close to those from the original series (within the confidence interval). For SI and most multiple imputation methods, the regression coefficient estimates are outside of the confidence interval for three of the eight variables. Among the multiple imputation methods, the MI-MICE-PMM method without data transformation obtains the best results, with 50% missing data (as well as for two variables with 75% missing data). In terms of standard error, the CC method produces the worst results. Conversely, the SI method is the most efficient and the most stable. The multiple imputation methods produce similar results, as do CC and SI. However, SI has the worst regression coefficient estimates coverage. For all the other methods, the coverage level is adequate as long as the proportion of missing data remains low. However, coverage deteriorates strongly for three of the eight variables. Nevertheless, the MI-MICE-PMM method without transformation allows one to preserve a coverage of 93% for a variable that contains 75% missing data. As far as the significance of the variables in the prognostic model is concerned, the more missing data there is in the sample, the more it is damaged. In all cases, the CC method leads to loss of significance. Naturally, this loss is due to the drastic reduction of the sample size. The other imputation methods do not seem to affect the significance of the estimators at all. Globally, the significance of the estimators from the imputed data is the same as that of the estimators from the original data. Finally, the examination of the performance measures of the model reveals that the likelihood ratio test was significant for all imputation methods. Among the multiple imputation methods, the MI-MICE-PMM method without transformation has one of the lowest p values. The same method gives the best results in terms of Negelkerke R^2 and the prognostic separation statistic. The probability of survival is not impacted by the imputation of the missing data.

Marshall, Altman, Royston and Holder [150] replicated the analysis by injecting MCAR missing data and found no difference. However, when the missing data are of the MNAR type, it is possible to observe some discrepancies, particularly in the estimates of the regression coefficients of certain variables. However, the MI-MICE-MICE method without transformation is among those that perform best in this regard, too.

Marshall, Altman, Royston and Holder [150] concluded with the idea that the use of a multiple imputation method is not justified when the proportion of missing data is below 10%. If that proportion is exceeded, multiple imputation produces much better results than listwise deletion. According to the authors, when the proportion of missing data is significant but below 50% and the missing data are not MNAR, MICE with PMM should be the preferred multiple imputation method.

The application of MICE with PMM to financial data is rare. In fact, there appear to be no comparative studies that are based on a sample of financial time series. On the other hand, in 2015, Grabka and Westermeier from the German Socio-Economic Panel edited a research report that explains how wealth data can be prepared from a social scientific panel study [97]. Some data were missing, and the authors made suggestions about filling the gaps. They described two completion methods: a basic method and a fallback method.

The basic method is the row-and-column imputation technique, which is a univariate method that combines data that are available for the entire duration of the panel and for every unit (row) and cross-sectional trend information (column). It entails adding a residual that is derived from nearest-neighbor matching, thereby introducing a stochastic component into an otherwise deterministic approach. The fallback method is the MICE algorithm with PMM, which was presented in this section. The basic method is used for observations with missing values where information from other waves is available. The MICE algorithm is applied when there are some missing observations for which only cross-sectional information is available.

2.4.8 Iterative Principal Component Analysis

More recently, in 2018, Hubert, Rousseeuw and Van den Bossche [115] developed an imputation and outlier detection method called *MacroPCA*. It can deal with cellwise outliers (suspicious cells that can occur anywhere in the data matrix) and row-wise outliers (entire observations which do not fit the model). *MacroPCA* is based on a principal component analysis that covers missingness and cellwise and rowwise outliers. It has been implemented in the R software, and it is available in the package *cellWise* from the *MacroPCA* function.

Theoretical presentation

Let $\boldsymbol{Y}$ be a data matrix composed of n rows and p columns. According to the authors, if there are no outliers or missing data, is possible to rewrite $\boldsymbol{Y}$ so that

$$\boldsymbol{Y} = \mathbf{1}_n \boldsymbol{\mu}^T + \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{\varepsilon}, \tag{2.4-66}$$

where $\mathbf{1}_n$ is the column vector of value 1 and of dimension n, $\boldsymbol{\mu}$ is the column vector of dimension p containing the means of each column of $\boldsymbol{Y}$, $\boldsymbol{T}$ is the $(n \times K)$ score matrix (with K the chosen number of principal components), $\boldsymbol{P}$ is the $(p \times K)$ loading matrix, and $\boldsymbol{\varepsilon}$ is the error matrix with dimensions $(n \times p)$. Moreover, K should be low, even if it can vary from 1 to p.

The theory of PCA consists in decomposing a matrix $\boldsymbol{Y}$, to which the mean has been adjusted, into two orthogonal matrices, here denoted by $\boldsymbol{T}$ and $\boldsymbol{P}$. The loading matrix $\boldsymbol{P}$ contains the eigenvectors of the $\boldsymbol{Y}$-adjusted covariance matrix. Finally, the score matrix $\boldsymbol{T}$ is the matrix that is orthogonal to the loading one. It is obtained by projecting the $\boldsymbol{Y}$-adjusted matrix into the loading matrix.

When data are missing, the variables $\boldsymbol{\mu}$, $\boldsymbol{T}$ and $\boldsymbol{P}$ are unknown. For that reason, Hubert, Rousseeuw and Van den Bossche [115] proposed the *MacroPCA* algorithm that considers that data y_{ij} may be missing. Data are supposed to be MAR in their framework. They also assume, however, the possible presence of row-wise outliers (the entire row is considered an outlier) and the possibility that cellwise outliers (incorrect, inaccurate or unusual single data) may also be present. In that case, the variables $\boldsymbol{\mu}$, $\boldsymbol{T}$ and $\boldsymbol{P}$ are unknown.

The algorithm used to fill in missing data in *MacroPCA* is the iterative classical PCA (*ICPCA*) algorithm developed by Nelson, Taylor, MacGregor [162] and Kiers [129]. It works like an EM algorithm (see Section 2.4.6) by replacing the missing values with initial estimates (mean imputation or zero imputation). Then, it fits a classical PCA iteratively. It consists of four main steps.

1. Initializing the missing elements with their estimates (calculated as the mean of the corresponding column means)
2. Performing a PCA in order to deduce the loading and score matrices $\boldsymbol{P}$ and $\boldsymbol{T}$.
3. Reconstructing them with the predefined number of factors.
4. Replacing the missing data with the predicted values and repeating the last three steps until convergence.

This process is close to the one presented for the MSSA method in Section 2.4.4, the difference being that the MSSA algorithm reshapes the data through a trajectory matrix (lagged matrix) before decomposing it. In iterative PCA, the decomposition is performed directly on the data matrix.

Van Den Bossche used the Matlab methodology of Folch-Fortuny, Arteaga and Ferrer [84] to implement the iterative PCA function. More formally, $\boldsymbol{Y}$ is the data matrix of dimensions $(n \times p)$ that contains missing data and $\boldsymbol{M}$ is the missingness matrix, where $m_{ij} = 1$ if y_{ij} is missing and 0 otherwise (with $i = 1, \ldots, n$ and $j = 1, \ldots, p$).

Thus, the preliminary step of iterative PCA consists in writing the matrix $\boldsymbol{Z}$ so that it is equal to $\boldsymbol{Y}$, but the missing data is replaced by 0, which means that $\boldsymbol{Z} = \bar{\boldsymbol{M}} \circ \boldsymbol{Y}$, where $\bar{\boldsymbol{M}}$ is the complement matrix of $\boldsymbol{M}$, and $\circ$ is the Hadamard product operator.

The iterative process begins at this point. For each iteration t,

1. the data matrix is centered so that $\boldsymbol{X}^{(t)} = \boldsymbol{Y}^{(t)} - \mathbf{1}_n(\boldsymbol{\mu}^{(t)})^T$, where $\boldsymbol{Y}^{(0)}$ is initialized to $\boldsymbol{Z}$ when $t = 0$, $\mathbf{1}_n$ is a column vector of dimension n with values 1, and $\boldsymbol{\mu}^{(t)}$ is the column vector of dimension p that contains the means of each column of $\boldsymbol{Y}^{(t)}$ (for iteration t),

2. a PCA is performed on the matrix $\boldsymbol{X}^{(t)}$ using K principal components so as to obtain $\boldsymbol{X}^{(t)} = \boldsymbol{T}^{(t)}(\boldsymbol{P}^{(t)})^T + \boldsymbol{\varepsilon}^{(t)}$, where $\boldsymbol{T}^{(t)}$ is the score matrix, $\boldsymbol{P}^{(t)}$ is the loading matrix, and $\boldsymbol{\varepsilon}^{(t)}$ is the error matrix (for iteration t), and

3. the new matrix $\boldsymbol{Y}^{(t+1)}$ is calculated so that $\boldsymbol{Y}^{(t+1)} = \bar{\boldsymbol{M}} \circ \boldsymbol{Y}^{(t)} + \boldsymbol{M} \circ (\boldsymbol{T}^{(t)}(\boldsymbol{P}^{(t)})^T + \mathbf{1}_n(\boldsymbol{\mu}^{(t)})^T)$. This new matrix is equal to the original matrix $\boldsymbol{Y}$, where missing data are replaced with their predicted values.

The first three steps are iterated for so long as the tolerance threshold is not reached (the difference between consecutive imputations $\boldsymbol{Y}^{(t)}$ and $\boldsymbol{Y}^{(t+1)}$ is below the specified threshold, set by default to 0.005) and the maximum number of iterations is not completed (set by default to 20).

In 2005, Hubert, Rousseeuw and Vanden Branden [116] explaind that the *MacroPCA* algorithm uses a robust version of the iterative PCA (ROBPCA or MROBPCA in the case of missing data) in the sense that it accounts row-wise outliers. However, in order to compare like with like, the method that will be used in this PhD thesis is the iterative PCA technique that Hubert, Rousseeuw and Van den Bossche [115] used to impute missing data in their *MacroPCA* algorithm. The *MacroPCA* could be compared to other methods that combine outlier detection and data imputation. However, since this PhD thesis only concerns imputation methods, it is preferable to refer to the imputation method used in *MacroPCA*.

Empirical application of iterative PCA

Hubert, Rousseeuw and Van den Branden [116] then compared the performance of *MacroPCA* to those of *ICPCA* and *MROBPCA* (*ROBPCA* for missing values) models with simulated data. As a reminder, the *ICPCA* method only aims to impute missing data, the *MROBPCA* method aims to impute missing data and row-wise outliers, and the *MacroPCA* method aims to impute missing data as well as row-wise and cellwise outliers.

A clean data sample $\boldsymbol{Y}$ is generated from a multivariate Gaussian distribution with $n = 100$, $p = 200$, and mean $\mu = 0$ that is associated with two types of covariance matrix Σ of dimension p. The first covariance matrix used, called A09,

is a structured correlation matrix where each value except those on the diagonal is $\rho_{ij} = -0.9^{|i-j|}$. The second one, called ALYZ, is based on the random correlation matrices of Agostinelli, Leung, Yohai and Zamar [3]. In both cases, the correlation matrices are converted into covariance matrices with other eigenvalues, which means that the elements of the diagonal of the matrix $\boldsymbol{L}$ of dimension p (from the spectral decomposition that gives $\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{L}^T$ are replaced by the desired values, namely $\boldsymbol{L} = diag(30, 25, 20, 25, 10, 5, 0.098, \ldots, 0.002, 0.0015)$. As the sum of the first six eigenvalues represents 91.5% of all 200 eigenvalues, Hubert, Rousseeuw and Van den Branden [116] use $K = 6$. They also note that the *MacroPCA* algorithm takes less than one second to run for $n = 100$ and $p = 200$.

Hubert, Rousseeuw and Van den Branden [116] proffered four specific samples. In the first one, they took $\boldsymbol{Y}$ and randomly integrated $5\%, 10\%, \ldots, 30\%$ missing data cells. The second sample incorporates 20% missing data and 20% outlying cells with values ranging from $\gamma\sigma_j$, where σ_j^2 is the j-th elements of the diagonal of Σ, and γ, ranging from 0 to 20. The third sample contains 20% missing data (introduced at random) and 20% outlying rows generated from $\mathcal{N}(\gamma\boldsymbol{\nu}_{K+1}, \boldsymbol{\Sigma})$, with γ varying from 0 to 50, and $\boldsymbol{\nu}_{K+1}$ corresponds to the $(K+1)$-th eigenvector of $\boldsymbol{\Sigma}$. Finally, the last sample accounts for the three types of anomalies, which consist of 20% missing data, 10% cellwise outliers and 10% row-wise outliers. The methodology is the same as that for the two previous samples.

In order to compare the effectiveness of the *MacroPCA* algorithm to that of *ICPCA* and *MROBPCA*, the authors measured mean squared error (MSE), quantifying the difference between the clean data $\boldsymbol{Y}$ and the results from the application of each the algorithms to the created samples. The MSE is then averaged over 100 replications.

Figure 2.4-32 presents the results for the sample in which missing data rang from 5% to 30%. The algorithm with the best results is *ICPCA*, which is due to the absence of outliers. The results obtained by *MROBPCA* and *MacroPCA* are similar and slightly higher than those of *ICPCA*. However, referring to the scale, which is very tight, the authors concluded that the three methods yield similar results. The next sample also includes cellwise outliers. Figure 2.4-33 shows that the MSE grows very rapidly with the magnitude of the outliers (γ parameter). Therefore, the methods are not robust in the presence of outlier data. For the *MacroPCA*, the results remain within a finite interval. Specifically, when γ is lower than 5, MSE increases; when γ is higher than 5, it decreases. The explanation that the authors gave is that for a low γ, the outlying cells are still close to their original values and are not easily detected by the algorithm.

Figure 2.4-34 shows the results for the sample with 20% missing data and 20% row-wise outliers. The *ICPCA* algorithm, being unable to handle outliers at all, sees its MSE mushroom. *MROBPCA* and *MacroPCA* give adequate results.

The last sample has 10% missing data, 10% cellwise outliers and 10% row-wise outliers. Figure 2.4-35 confirms the efficiency of authors' *MacroPCA*. MSE is very high

for the *ICPCA* and *MROBPCA*, while *MacroPCA* produces reasonable results.

Fig. 2.4-32: Average MSE as a function of the fraction of missing data. The data were generated using the A09 (left) and the ALYZ (right) correlation matrix (Source: Hubert, Rousseeuw and Van den Bossche, 2019 [115])

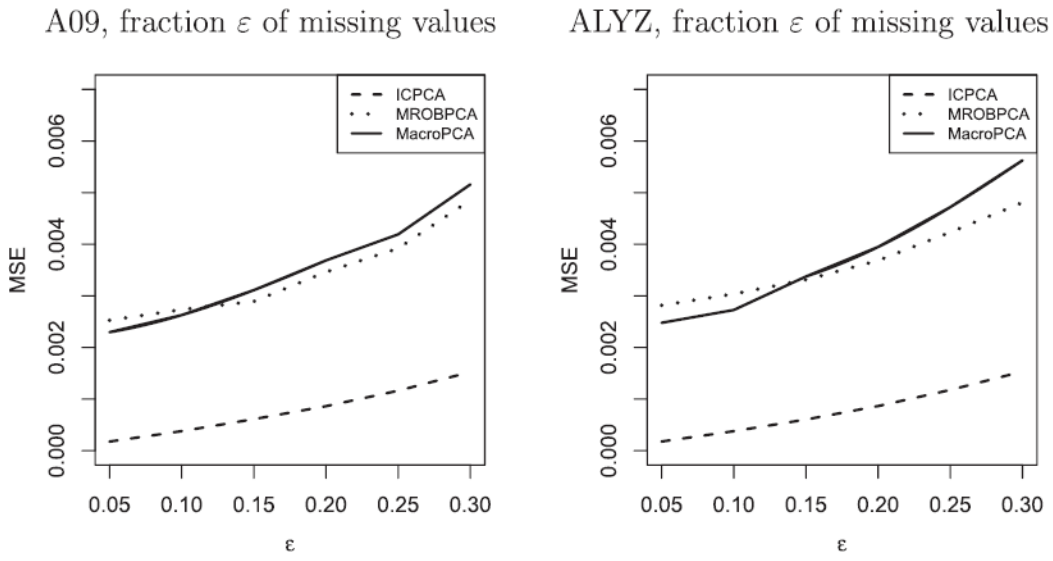


Fig. 2.4-33: Average MSE for sample with 20% missing data and 20% cellwise outliers, as a function of the distance of the cellwise outliers (Source: Hubert, Rousseeuw and Van den Bossche, 2019 [115])

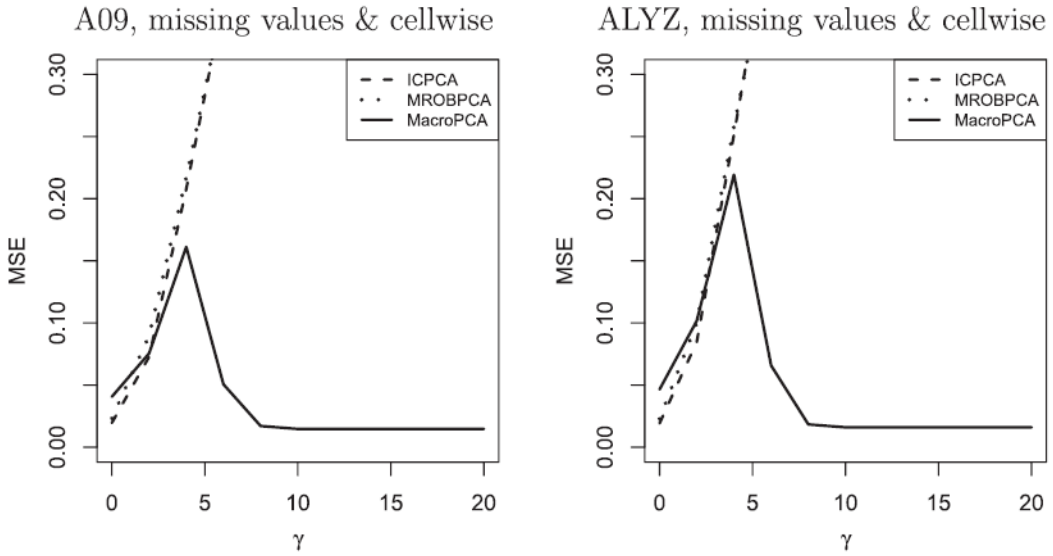


Fig. 2.4-34: Average MSE for sample with 20% missing data and 20% row-wise outliers as a function of the distance of the row-wise outliers (Source: Hubert, Rousseeuw and Van den Bossche, 2019 [115])

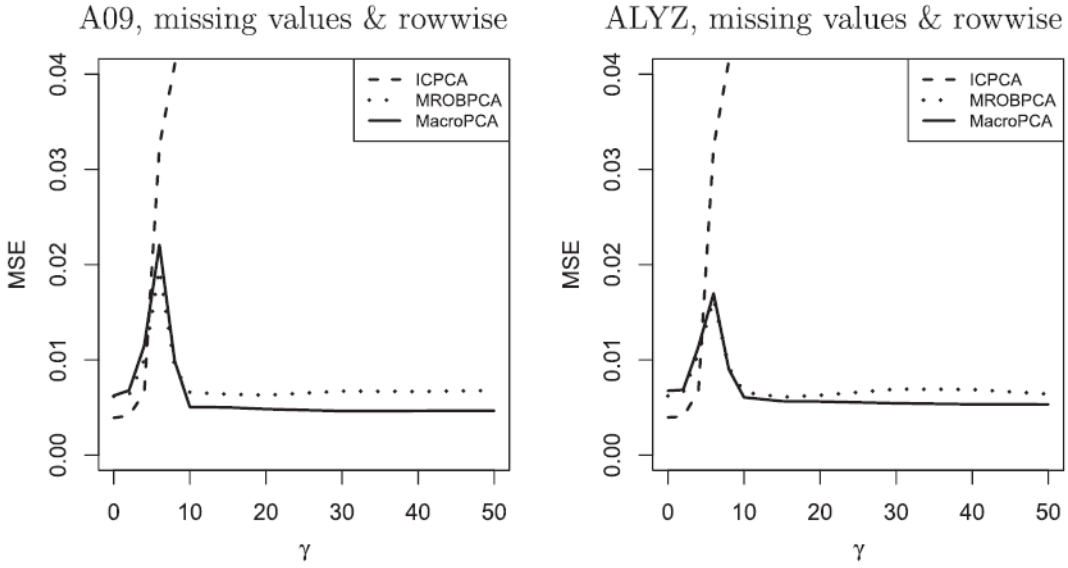
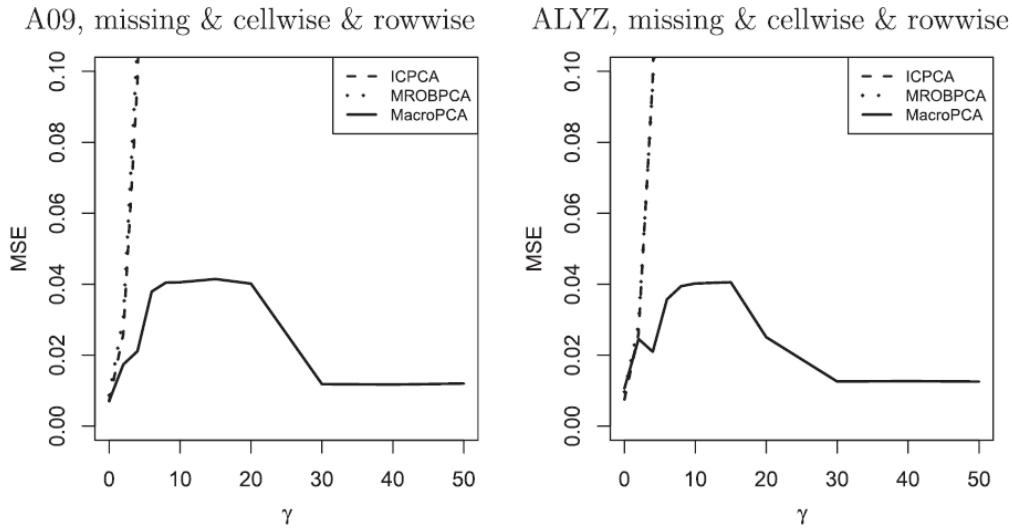


Fig. 2.4-35: Average MSE for sample with 10% missing data, 10% cellwise outliers and 10% of row-wise outliers as a function of the distance of the cellwise and the row-wise outliers (Source: Hubert, Rousseeuw and Van den Bossche, 2019 [115])



Finally, Hubert, Rousseeuw and Van den Branden [116] showed the superiority of the *MacroPCA* for missing data, outlier cells and outlier lines that occur simultane-

ously. However, when missing data are the only problem, *ICPCA* seems to be the most effective.

The *MacroPCA* method being very recent, no other authors have written comparative studies. This said, the iterative PCA method is older and has already been used to impute missing data. Notably, Josse and Husson [126] have presenting the PCA iterative method, identified its defects and suggests means to remedy them.

Regularized iterative PCA

The first flaw of the PCA iterative method concerns the overfitting problem. It is characterized by good estimates of observed data, which means that $\boldsymbol{\varepsilon}$ errors are low. At the same time, prediction quality for missing data is dissatisfactory due to the poor estimation of the axes and the components. This problem is especially pronounced when the proportion of missing data becomes too large and too many dimensions must be retained.

To overcome this first problem, Josse and Husson [126] propose a regularized iterative PCA method. It consists in repeating the same process as the iterative PCA method (presented previously) but using a shrunk imputation step. To that end, it is assumed that the first dimensions contain all the information and noise and that the other dimensions contain only noise. The idea of regularization is to remove the noise from the latter in order to reduce the instability of the prediction. Therefore, the regularized method consists of recalibrating the impact of the eigenvalues used for the reconstruction. Instead of using $\boldsymbol{T}^{(t)}(\boldsymbol{P}^{(t)})^T$ directly in the third step of the original algorithm, $\boldsymbol{Y}^{(t+1)}$ is computed from $\hat{\boldsymbol{X}}^{(t)}$ so that

$$\hat{\boldsymbol{X}}_{ij}^{(t)} = \sum_{k=1}^{K} \frac{\lambda_k - \sigma^2}{\lambda_k} \boldsymbol{T}_{ik}^{(t)} \boldsymbol{P}_{jk}^{(t)}, \tag{2.4-67}$$

where λ_k denotes the eigenvalues of the k-th principal component, $\sigma^2 = \frac{1}{p-K}\sum_{k=K+1}^{p} \lambda_k$ is the average of the unused eigenvalues, and $\hat{\boldsymbol{X}}_{ij}^{(t)}$, $\boldsymbol{T}_{ij}^{(t)}$ and $\boldsymbol{P}_{ij}^{(t)}$ are the value of the i-th row and the j-th column of $\hat{\boldsymbol{X}}^{(t)}$, $\boldsymbol{T}^{(t)}$ and $\boldsymbol{P}^{(t)}$, respectively.

The regularized model matrix is given by $\hat{\boldsymbol{Y}}^{(t+1)} = \boldsymbol{Z} + \boldsymbol{M} \circ (\hat{\boldsymbol{X}}^{(t)} + \mathbf{1}_n(\boldsymbol{\mu}^{(t)})^T)$. Josse and Husson [126] provide an example with a sample of simulated data. Its two-dimensional configuration is given in Figure 2.4-36 (left). Then, they remove 50% of the data at random. By imputing the missing data through the iterative PCA method (with two principal components), they obtain the results presented in Figure 2.4-36 (right). It is possible to see that the imputed data are much closer to each other than the original data. This so because they are imputed from the same number of dimensions. Moreover, the classical method reinforces projection quality because the inertia percentage that is calculated from the completed data and associated with the dimensions that are used is much higher than that calculated from the original data (94% versus 82% in this example). The shrinkage process of the regularized method

(see Figure 2.4-37) allows this inertia percentage to be reduced, which is why it is of interest to the present ends.

Fig. 2.4-36: Illustration of the overfitting problem: original data (on the left) and same dataset with 50% data removed and imputed by iterative PCA (Source: Josse and Husson, 2012 [126])

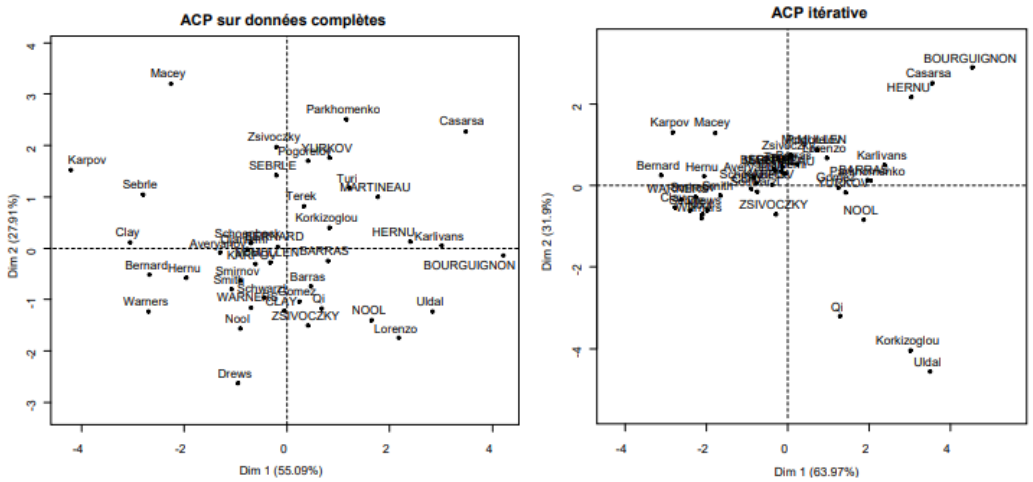
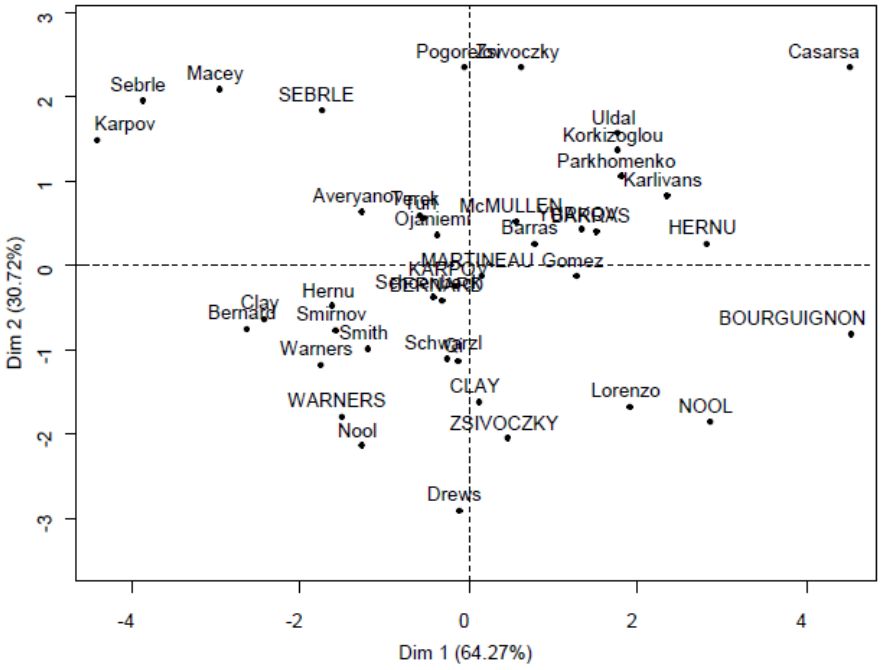


Fig. 2.4-37: Illustration of the overfitting problem: same dataset with 50% data removed and imputed by regularized iterative PCA (Source: Josse and Husson, 2012 [126])



Multiple imputation with (regularized) iterative PCA

The second main flaw of the iterative PCA is that it is a simple imputation method. As seen in the presentation of K-NN in Section 2.4.3, simple imputation methods do not account for the variability of missing data. An analysis of data imputed by a simple imputation method treats it in exactly the same way as observed data. Currently, the solution is to combine this method with multiple imputation, as explained in Section 2.4.6. A bootstrap step can be added to obtain a collection of possible imputations.

However, the bootstrap step that is integrated here is different from the one in K-NN, random forests, Amelia or MICE. The idea here is not to impute B bootstrapped sample (from which the observations would have been drawn with replacement from the original sample) but rather to impute the original sample as it is through the (regularized) iterative PCA method before replicating it B times and integrating the desired variability in each of the resultant samples. Josse, Pagès and Husson [127] propose a new multiple imputation method that aims to account for both the variability of the estimated values of the PCA parameters and the variability that is due to noise.

Accordingly, B bootstrap samples $\boldsymbol{Y}^b$ (with $b = 1, \ldots, B$) are created by adding B new matrices of residuals obtained by bootstrapping the current residuals matrix $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ (bootstrap among the observed residuals) to the estimator $\hat{\boldsymbol{Y}}$ (in the case

of a regularized iterative PCA). Then, a regularized iterative PCA algorithm is applied to each $\boldsymbol{Y}^b$ (adjusted by its means) in order to arrive at B new estimators $\hat{\boldsymbol{Y}}^1, \ldots, \hat{\boldsymbol{Y}}^B$ that represent the variability of the PCA parameters. Finally, the B imputed samples are obtained by adding noise to each imputed value of $\hat{\boldsymbol{Y}}^b$ (with $b = 1, \ldots, B$), which is drawn from a Gaussian distribution of means equal to 0 and with variance equal to the variance of the residuals $\hat{\boldsymbol{\varepsilon}}$. Since this method incorporates bootstrap and thus a random component, the results run into the non-replicability problems that were presented in Section 2.4.5.

Choice of the number of principal components

As in K-NN, choosing the number of main components to be used is not trivial in the case of incomplete data, especially since it has a direct impact on the quality of the imputation. If the number of dimensions that are used is too small, then too much relevant information may be lost. Conversely, if this number is too large, overfitting is more likely.

Josse and Husson [126] explain the literature offers no solution to the problem of determining the optimal number of dimensions to use in the presence of missing data. Cross-validation is one candidate, as shown in Section 2.4.3, which discussed the determination of the K nearest neighbor. The cross-validation method that was presented in the case of K-NN is the k-folds method, which consists of separating the sample into k sub-samples in order to train the prediction model on $k-1$ sub-samples and validating it on the last one for different parameter levels. Finally, the chosen model parameter is the one that minimizes the difference between the predicted value of the validation sample and its observed value.

However, Josse and Husson [126] use a different method. They start with leave-one-out cross-validation, which is even simpler. It involves removing a single observation from the sample, predicting it and repeating this procedure for each of the observations in the sample, in order to compute an average prediction error. This procedure is done for each parameter levels (here, the number of principal components K) to determine for which one the average prediction error is minimized. In a fully observed database, it is possible to determine the parameter K that minimizes the mean square error of prediction (MSEP). It is given by

$$MSEP(K) = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} (\hat{\boldsymbol{Y}}_{ij}^K - \boldsymbol{Y}_{ij})^2, \tag{2.4-68}$$

with $\hat{\boldsymbol{Y}}_{ij}^K$ the prediction value of the i-th observation of the j-th column of $\boldsymbol{Y}$ based on the first K principal components.

Obviously, the main flaw of this method is its computation time: the larger the matrix, the longer the procedure. In order to remedy this problem, Josse and Husson

[126] propose using an approximation of the method. They define the generalized cross-validation (GCV) criterion as follows:

$$GCV(K) = \frac{np \times (\sum_{i=1}^{n} \sum_{j=1}^{p} (\hat{\boldsymbol{Y}}_{ij}^{K} - \boldsymbol{Y}_{ij})^2)}{(np - p - nK + K^2 + K)^2}. \tag{2.4-69}$$

Josse and Husson [126] propose extending the generalized cross-validation criterion to missing data as follows:

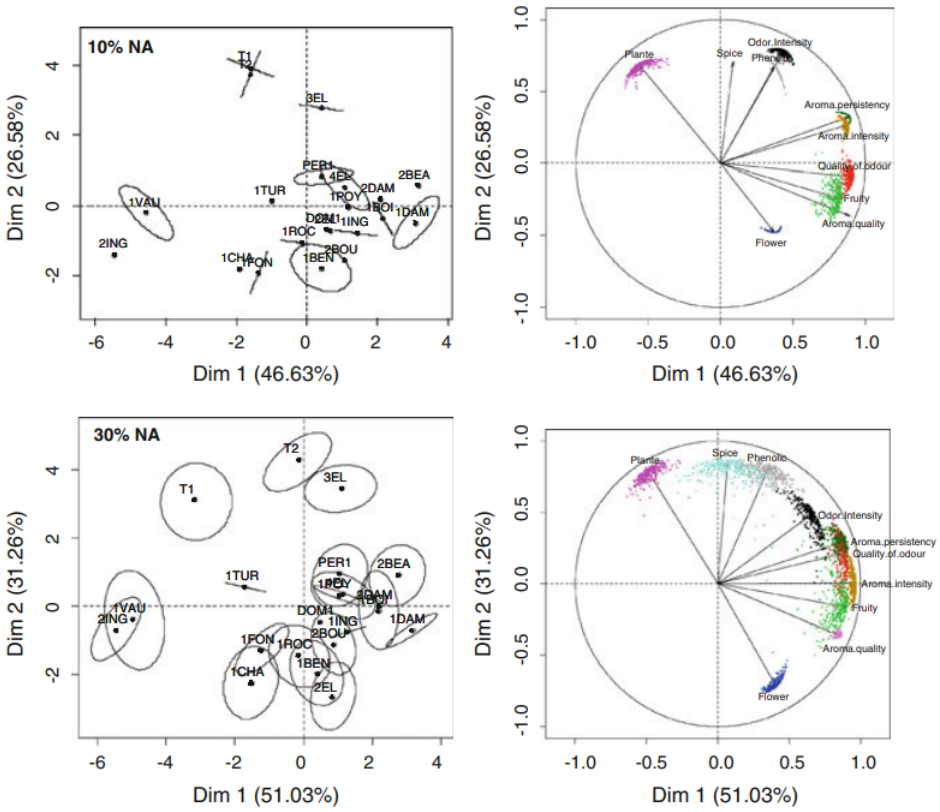
$$GCV(K) = \frac{(np - nbmiss) \times (\sum_{i=1}^{n} \sum_{j=1}^{p} (\bar{\boldsymbol{M}}_{ij} (\hat{\boldsymbol{Y}}_{ij}^{K} - \boldsymbol{Y}_{ij}))^2)}{(np - nbmiss - p - nK + K^2 + K)^2}, \tag{2.4-70}$$

where $nbmiss$ is the number of missing cells, $\bar{\boldsymbol{M}}$ is the complement missingness matrix ($\bar{\boldsymbol{M}}_{ij}$ is equal to 0 if y_{ij} is missing; otherwise, it is 1), and $\hat{\boldsymbol{Y}}$ is the prediction matrix that is based on K dimensions. The details of the calculations that lead to this result are presented in their article [126].

Application of multiple imputation with iterative PCA

Josse, Pagès and Husson [127] present a simulation study of a fully observed database that comprises 21 individuals and 10 variables. The first two dimensions of the complete database explain more than 70% of the variability. For testing purposes, the authors removed data (10% then 30%) at random before imputing them. They explain that it is not possible to say with certainty whether the imputation results obtained by the regularized iterative PCA method are plausible or even interpretable. Therefore, applying the method in a multiple imputation context is a matter of interest. Figure 2.4-38 presents the results obtained for an imputation of 10% of missing data (top) and 30% of missing data (bottom).

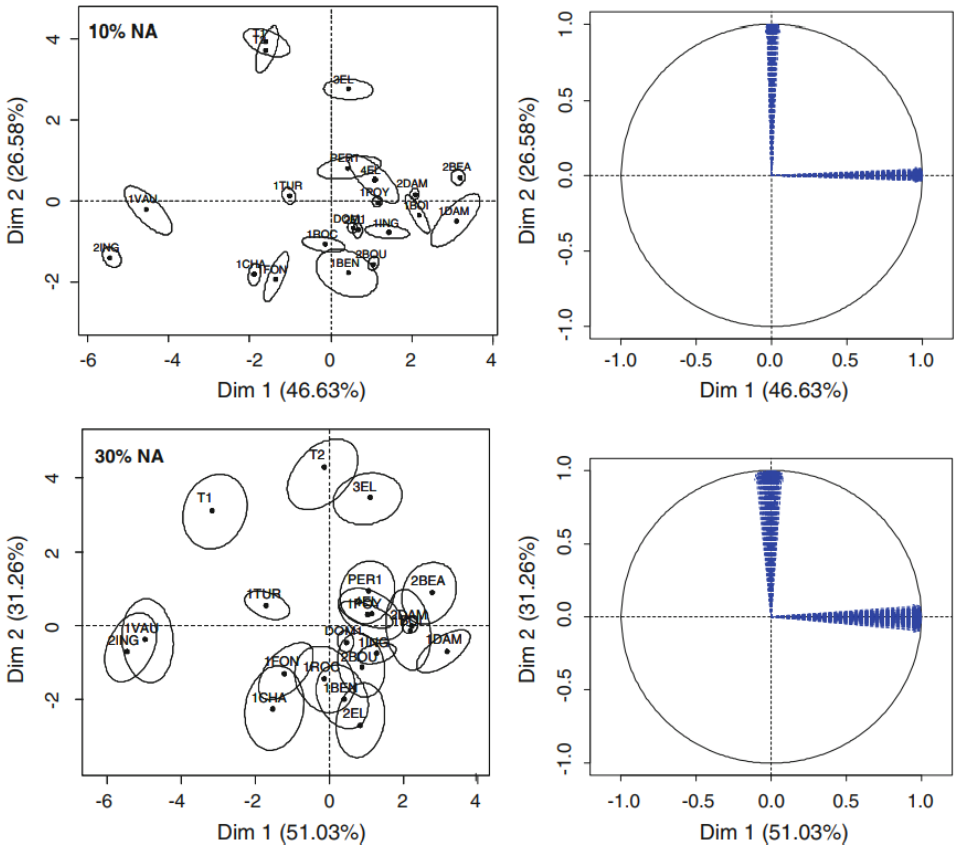
Fig. 2.4-38: Visualization of the uncertainty of individuals (on the left) and of variables (on the right) for two level of missing data (Source: Josse, Pagès and Husson [127])



Multiple imputation was performed on 500 bootstrapped samples, allowing the prediction uncertainty of missing data imputed by the PCA method to be taken into account. In Figure 2.4-38, this uncertainty is represented by ellipses (left) and point clouds (right) at a confidence level of 95%. The authors find that the size of the ellipses and the point clouds increases with the proportion of missing data. However, their size does not mushroom. Therefore, the proportion of missing data does not impact the estimate here.

Finally, in order to determine the impact of data imputation on parameter uncertainty, a regularized PCA is performed on each of the 500 samples imputed previously in order to compare them to the reference configuration. The results are presented in Figure 2.4-39.

Fig. 2.4-39: Visualization of uncertainty due to missing data (for two amounts of missing data): of individuals (on the left) and of two-first dimension representations based on 500 datasets (on the right; Source: Josse, Pagès and Husson [127])



For the left plots, they find that the different imputations imply slightly different axes and components, meaning that the imputations affect the positions of all individuals, even if they do not contain missing data. Conversely, the more missing data an individual contains, the greater the uncertainty around it. The authors note that their results are mostly stable, even with 30% missing data, as illustrated by the graphs on the right that correspond to the projections of the first two components that are obtained from the imputed data on the reference configuration. They believe that the interpretation of the dimensions of the reference configuration is accurate.

In this PhD thesis, the *ICPCA* method (implemented by Van Den Bossche [115]) and the *MacroPCA* were compared to the regularized iterative PCA version that is integrated in Josse and Husson's multiple imputation [126].These methods were compared to determine their effectiveness on financial data and to solve regulatory problems, among other things. The performance of one-dimensional techniques (linear interpo-

lation, last observations and Brownian bridge) was compared to that of much more sophisticated methods (K-NN, random forests, MSSA, Amelia, MICE, *ICPCA* and Regularized iterative PCA).

2.5 Other completion methods and expected results

This chapter aims to present a selection of completion methods that are used in the following chapter. Other completion methods are available in the literature, and some of them are presented in this section and may be implemented in future work. Moreover, all completion methods presented previously may provide an indication of the likely results of the empirical analysis. Therefore, this last section also provides a summary of the completion methods that were presented in this chapter, and the anticipated outcomes of their application.

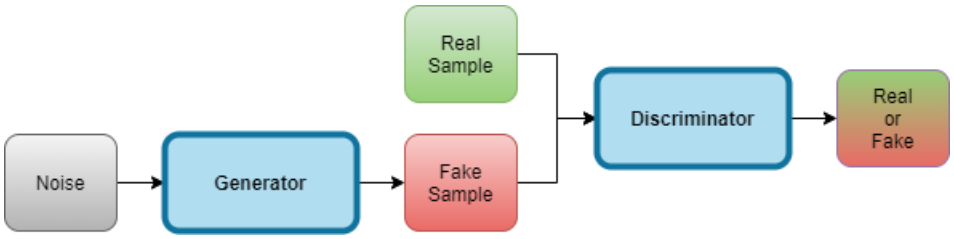
2.5.1 Other completion algorithms

Many other methods can be used to impute missing data, especially missing data from time series. This chapter has highlighted the methods that will be applied to simulated and historical data, but these are only a selection of them. The literature related to missing data imputation is composed of many other methods, notably methods specifically designed for imputing missing data from time series.

A natural approach is to impute missing data using an autoregressive model, which is commonly used to analyze time series data. Moreover, since financial data are not always Gaussian, an autoregressive model with more general heavy-tailed innovations is preferred. In 2019, Liu, Kumar and Palomar [146] proposed an efficient framework for parameter estimation from incomplete heavy-tailed time series based on a stochastic approximation EM coupled with a Markov chain Monte Carlo procedure. To do so, they formulated a maximum likelihood estimation problem based on an autoregressive process and developed an efficient approach to obtain the estimates based on the stochastic EM. A stochastic EM consists of replacing the E-step of a classical EM by a stochastic approximation procedure, which approximates the expectation by combining new simulations with the previous simulations. Moreover, Liu, Kumar and Palomar [146] showed the convergence of their algorithm, and that it provides reliable estimates from an incomplete time series for different missingness proportions. This algorithm is based on a univariate time series, which is why Zhou, Liu, Kumar and Palomar [218] extended it to a multivariate time series (using vector autoregression).

In addition, many machine learning and deep learning methods have been developed over the last few years, and so they increasingly present in the literature. The most

Fig. 2.5-1: GAN architecture: the generator generates a fake sample based on noise, and the discriminator distinguishes whether this sample is fake or real



popular models are based on generative adversarial networks (GANs). They are a class of unsupervised learning algorithms introduced by Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville and Bengio [96] in 2014. GANs are deep learning methods that consist of training generative models through an adversarial process between, generally, two networks. According to their authors, a GAN can be considered, from a game theory point of view, as a two-player minimax game between a generating network and a discriminating network. In other words, the generator generates a fake sample from noise, while the discriminator tries to detect if this sample is real, or if it is a fake from the generator (see Figure 2.5-1).

GAN methods were later adapted to fill in missing data. Iizuka, Simo-Serra and Ishikawa [117] were the first to use GAN methods for missing data imputation and, in particular, image completion. Their GAN consists of a completion network generator and a global and a local discriminator. Both the global and local discriminators are trained to distinguish between real and fake images. Iizuka, Simo-Serra and Ishikawa [117] explain that this type of GAN can be used to fill in missing areas and show very satisfactory results. On the other hand, it can produce blurred and distorted images that are not coherent if the missing area is very different from the nonmissing area. Based on these limitations, Yu, Lin, Yang, Shen, Lu and Huang [215] introduced a new GAN based on two generators and two discriminators. This last version of GAN is efficient in imputing large rectangular forms of missing data but not free forms. Hence, an improved version, by the same authors [216], enables to take into account free forms of missing data.

Finally, Yoon, Jordon and van der Schaar [213] present generative adversarial imputation nets (GAIN) models specific to missing data imputation. These GAINs are composed of one generator and one discriminator. They differ from GANs on several points. The generator receives noise as well as the missing data pattern as input. Moreover, the discriminator receives the data matrix imputed by the generator, as well as an information matrix as inputs. This information matrix reveals partial information about the missing data pattern of the initial sample. Finally, the discriminator tries to determine which elements of the sample are fake or true and does not analyze the

whole sample as in the case of GAN. In their paper, Yoon, Jordon and van der Schaar [213] even compared the performance of GAIN with random forests, MICE and the EM algorithm on five different samples. Their application reveals an out-performance of GAIN compared to the other methods in terms of RMSE and prediction accuracy. Subsequently, Camino, Hammerschmidt and State [55] proposed making the structure of GAIN models more complex, with multiple inputs and multiple outputs for the generator and for the discriminator.

In recent years, banks have shown a strong interest in GAN models. Indeed, many articles in the literature apply GANs to financial data. This is notably the case of Henry-Labordere [107], who uses a GAN model to calibrate the parameters of the SABR model (a stochastic volatility model [101]); Kondratyev and Schwarz [134] use a GAN model able to learn complex multivariate distributions from a limited number of samples and apply it to currency data; Buelher, Horvath, Lyons Perez Arribas and Wood [49][50] present a GAN model that is reliable even with a small amount of training data (with a numerical experiment on S&P historical data); Eckerli and Osterrieder [72] show that GANs can generate consistent data for various purposes, including retail banking and market data, and others.

The imputation methods presented in this chapter already constitute a large selection of frequently used methods for missing data imputation. However, there are many other imputation methods, such as those presented above based on autoregressive models or GANs, which would be interesting to include in future comparative analyses. Moreover, it must be kept in mind that the use of GAN methods is very recent, and so they can be misunderstood and are not very transparent, which can be a constraint for the regulator.

2.5.2 The expected results through literature

Referring to the completion methods selected for the comparative analysis performed in the next chapter, the literature provides some indications of expected results. First of all, certain models have never been applied to financial data; others have only seen sporadic application. The literature on the K-NN method very often includes applications to medical data of the kind described in Section 2.4.3. Thus, García-Laencina, Sancho-Gómez and Figueiras-Vidal [89] use data on sick thyroid disease, and Jerez, Molina, García-Laencina, Alba, Ribelles, Martín and Franco [124] use breast cancer data. The application of this completion method to financial data and monitoring its effectiveness may be promising.

Other methods, such as random forests and MICE, are applied in different fields. Random forests is applied to medical data, as in Young's article [214], which uses a large sample of clinical data (presented in Section 2.4.5) and to insurance data, as in Jamal's article [180] presented in Section 2.4.5. Similarly, Section 2.4.7 presented two

applications of MICE on medical data from 2010, one by Marshall, Altman and Holder [149] and another by Marshall, Altman, Royston and Holder [150] have also applied it to socio-economic data. So far, these methods have never been applied to financial market data. Their effectiveness merits further investigation.

In contrast, MSSA and *Amelia* have been applied to financial data, or at least to Gaussian data. The applications of MSSA to financial data are presented in Section 2.4.4. Two notable examples are the paper on the currency exchange rates of BRICS countries by Rodrigues and Mahmoudvand [170] and Dash and Zhang's [65] study of USD swap rate data. Finally, Section 2.4.6 presented several illustrative applications of an EM algorithm to financial data, namely Warga's [208] work on U.S. bond data, Urli's [204] research on MSCI stock market indices and Bauer, Angelini and Denev [27] study of CDS. These articles proved the usefulness of the Brownian bridge and the EM algorithm for financial data. Therefore, it is expected that those methods will yield good results.

Overall, half of the completion methods presented in this chapter have been applied to financial data. The selected articles prove their effectiveness. Further work, however, is necessary to compare them to methods that are common in other fields of research.

2.5.3 Advantages and disadvantages of each algorithm

Table 2.5-1 summarizes the hypotheses of each of the models that were presented. It also adumbrates their advantages and disadvantages, which were discussed at length in this chapter .

Tab. 2.5-1: Summary table of imputation methods used (part 1)

Model	Hypothesis	Advantages	Drawbacks
Last Observation Carry Forward	- non-parametric method	- easy to understand - immediate result	- create non-realistic zero returns - can form jumps - data distribution distortion
Linear Interpolation	- non-parametric method	- easy to understand - immediate result	- constant returns due to smoothing - data distribution distortion - underestimation of variance
Brownian Bridge	- Gaussian data distribution	- easy to understand - short calculation time - no distribution distortion	- unsuitable for punctual jumps and crises - non-replicability problem
K-NN	- non-parametric method	- easy to understand - considers the interaction between the variables - results stabilized by bagging	- choice of the number of neighbors and distance - requirement of certain number of complete observations - long computation time with cross-validation and bootstrap - non-replicability problem
MSSA	- non-parametric method	- separates signal from noise	- choice of window length
Random forests	- non-parametric method	- few parameters - stable results (bagging) - low sensitivity to outliers	- black-box algorithm (hard to understand and interpret) - long computation time - non-replicability problem

Tab. 2.5-1: Summary table of imputation methods used (part 2)

Model	Hypothesis	Advantages	Drawbacks
Amelia	- Gaussian data distribution	- uses all available data - bias reduction due to multiple imputation - consideration of the uncertainty of missing data	- choice of initial parameters - convergence to a local maximum - requirement of large computer storage capacity due to multiple imputation - long computation time - non-replicability problem
MICE with PMM	- semi-parametric method	- insensitivity to heteroskedasticity - bias reduction due to multiple imputation - consideration of the uncertainty of missing data	- choice of the number of donors - repetition of imputed data - requirement of large computer storage capacity due to multiple imputation
ICPCA	- non-parametric method	- separates signal from noise	- choice of the number of principal components - overfitting problem - sensitivity to outliers
MI-Regularized *IPCA*	- semi-parametric method	- separates signal from noise - bias reduction due to multiple imputation - handling of overfitting problem	- requirement of large computer storage capacity due to multiple imputation - choice of the number of principal components - sensitivity to outliers

This chapter showed that, as a matter of theory, some models suit missing data in financial series better than others. First of all, in financial data, particularly in returns series, assuming normality is not absurd. Thus, applying Brownian bridge (see Section 2.4.2) and improved EM algorithm (Amelia; see Section 2.4.6) make sense on price returns. Therefore, good results on VAR and ES may be expected when the proportion of missing data is not too high. Out of the two, Amelia should perform better with missing data because of its multidimensional operation and its bootstrap step.

Satisfactory results are also expected from the MSSA method (see Section 2.4.4). That Bloomberg have chosen it to complete missing data and to address general data quality issues attests to its efficiency. Being multivariate in nature, it allows signal to be separated from noise by accounting for both time dependence and cross-sectional dependence. If the results of the MSSA algorithm when pitted against an EM-based method is considered (as Dash and Zhang[65] did in 2016), it is even possible that MSSA outperforms Amelia.

Moreover, some sophisticated methods have been used little or not at all with financial data, but their effectiveness has been proven in other research domains. These include random forests, MICE with PMM and iterative PCA methods. These are non-parametric or semi-parametric and can be applied to financial data easily. The random forests method (see Section 2.4.5) is often employed to solve problems in forecasting or even missing data. It could be efficient when applied to financial data, but its black-box aspects make comprehension and interpretation difficult, meaning that the possibility of regulatory approval is remote. In the same way, MICE with PMM (see Section 2.4.7) could be effective on a long record of financial data that is characterized by heteroskedasticity. MacroPCA (see Section 2.4.8), especially its iterative classical PCA part that concerns imputation, is close to MSSA. Therefore, it would be interesting to apply it to financial data and to compare the results. Finally, simpler methods, such as K-NN or the Brownian bridge (see Section 2.4.3 and Section 2.4.2), and trivial ones, such as linear interpolation or LOCF (see Section 2.4.1), are useful benchmarks. If a sophisticated algorithm yields results that are close to those that are obtained through simpler methods, it could be discredited. Alternatively, such a finding might indicate that imputing missing data is a serious challenge.

As far as the literature is concerned, the theory of missing data originated from Little and Rubin. They remain the foremost authorities, even though others have tried to refine their analysis. Listwise deletion (see Section 2.2.2) is probably the most common methodology for dealing with missing data. However, many studies, such as those by Kim and Curry [130] and Dallal [61], have shown that listwise deletion is inappropriate and that it biases results. For this reason, other methodologies should be preferred, including data imputation. There are many methods of imputation, of

which this chapter presented only a selection. Methods can be parametric or non-parametric, with single or multiple imputation, unidimensional and multidimensional, simple or sophisticated, famous or obscure, and so on. Each method requires the analyst to make certain choices, either about the construction of the sample (data encoding) or in setting parameters. Each has advantages and disadvantages that merit testing through application on financial data. It is with this problem that the next chapter is concerned.

Chapter 3:
Empirical studies of simulated and historical data

The literature proposes many approaches to missing data, but some of them have not been applied to financial data. For this reason, an empirical study may be of interest. First, it is important to compare the application of these approaches to simulated data that possess specific characteristics (Gaussian distribution, heteroskedasticity, jumps). Different missing data schemes should also be compared. As seen in the previous chapter, there are many reasons why a financial series may include missing data. Of course, the patterns that they follow depend on these reasons. Filling in missing data about an IPO is not the same as filling in missing data because a trader forgot to save them manually. Once the performance of the algorithms has been compared with simulated data, they are pitted against real historical data. The purpose of this chapter is therefore to ascertain the performance of the completion methods in finance and to discover their practical limitations. The latter are important because they may cause banks to incur heavy capital charges or even operational risks.

Contents

3.1 Simulated sample, algorithm configurations, comparative tools and process

Therefore, the first part of this chapter consists of applying different missing data proportions and mechanisms to simulated data to compare the completion methods. The use of a simulated sample allows the expert to analyze the impact of the completion methods according to the specificities of the sample (volatility and jumps, in this PhD thesis). The second part is dedicated to the same exercise on historical samples.

While many studies on historical data are available in the literature (Jerez, Molina, García-Laencina, Alba, Ribelles, Martín and Franco [124] in 2010; Marshall, Altman and Holder [149] in 2010; Marshall, Altman, Royston and Holder [150] in 2010; Stekhoven and Bühlmann [194] in 2011; Dash and Zhang [65] in 2016; Jamal [180] in 2016; Bauer, Angelini and Denev [27] in 2017; Young [214] in 2017), those on simulated samples are also frequently used. This is notably the case of Hubert, Rousseeuw and Van den Branden [116] in 2005, who used their algorithm (MacroPCA) on simulated

Gaussian and correlated data; García-Laencina, Sancho-Gomez and Figueiras-Vidal [89] in 2010, who injected different proportions of missing data on simulated data before extending their analysis to real data; Abrahantes, Sotto, Molenberghs, Vromman and Bierinckx [59] in 2011, who injected MCAR and MAR data on simulated samples; and Josse and Husson [126] in 2012, who imputed data in a completely random way in 50% of a simulated sample.

This chapter therefore uses analysis procedures that are already well established in the literature.

The purpose of this section is not to review each algorithm in detail, because that was already done in Chapter 2, but to present the implementation choices as well as the parameter choices that have been made for each of the methods and the comparative tools used to confront them. The different types of graphs used in the chapter will also be presented in a general framework to facilitate their reading later. The idea is therefore to understand how the comparative study will be conducted in the rest of the chapter.

3.1.1 Presentation of simulated sample

The first element to be presented in this section is the simulated sample. To fit with a financial framework, the data were simulated using a log-normal distribution. Indeed, as mentioned previously in this PhD thesis, stock prices are often modeled by a log-normal distribution. Thus, for S_{t-1} the price of a stock at time $t-1$ and S_t, the price of the same stock at time t, the log return between $t-1$ and t corresponds to an approximation of the relative return between $t-1$ and t. Thus:

$$ln\left(\frac{S_t}{S_{t-1}}\right) \approx \frac{S_t}{S_{t-1}} - 1 \tag{3.1-1}$$

In addition, it is possible to model the price dynamics of a stock price S_t, using a geometric Brownian motion, as:

$$dS_t = \mu S_t dt + \sigma S_t dW_t, \tag{3.1-2}$$

where μ and σ are constant drift and diffusion coefficients, respectively, and $dW_t = \varepsilon\sqrt{dt}$ with $\varepsilon \sim \mathcal{N}(0,1)$.

If price returns are supposed to follow a log-normal distribution, similar to Black and Scholes [35] in their model, then:

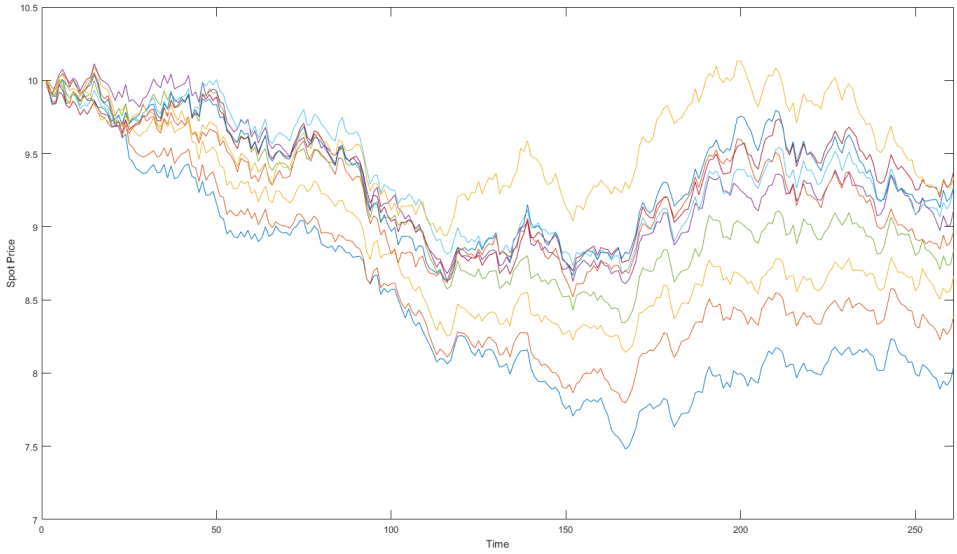
$$S_t = S_{t-1} \exp\left((\mu - \frac{\sigma^2}{2})dt + \sigma dW_t\right). \tag{3.1-3}$$

To apply each of the methods presented in Chapter 2, there is a need in this simulation randomly generate a daily stock price sample of 10 series with a length of 261 observations each (or slightly more than a year, in order to be able to compute 10-day VaR and ES measures), respecting the log-normal dynamics presented above (see Equation 3.1-3). Thus, the drift μ is set as 0, the volatility σ is arbitrarily set at 10%, ε is drawn from the law $\mathcal{N}(0,1)$ and finally dt is set equal to 1/252 (daily frequency). Thus, this simulation gives 10 time series (columns) of 261 observations (rows), where each time series has an annualized volatility of 10%. In addition, the original price S_0 has been set at 10 by default for all the series.

As discussed in Section 2.3, sample selection is important in an empirical study. For example, choosing variables that are relatively close to each other could have a positive impact on the quality of imputation. Thus, the sample generated here should be more or less correlated so that each of the variables has a dynamic relatively close to each other. To achieve this, a correlation matrix was generated such that the correlation between columns i and j is given by $\rho_{i,j} = 0.95^{|i-j|}$. Thus, for 10 time series, this gives correlations ranging from 95% to 63% (0.95^9). Thus, depending on the draw used to generate the simulated sample, the correlations between the simulated variables should be approximately between 95% and 63%. This correlation procedure is the same as that used by Hubert, Rousseeuw and Van den Branden [116] in 2005 to see the efficiency of the iterative PCA method, and especially of MacroPCA in the presence of outliers. Thus, all imputation methods will be based on a single sample composed of a specific correlation matrix. Nevertheless, it is not always easy to form correlated samples with historical data.

Thus, for a MATLAB seed set at 1, this procedure gives a simulated series, as presented in Figure 3.1-1.

Fig. 3.1-1: Simulated sample composed of 10 time series with 261 observations and an annualized volatility of 10%



This single sample is of course complete, which is why some data are removed voluntarily to apply the different imputation algorithms presented in the previous chapter and to be able to analyze the results. Thus, all imputations and comparative analyses are based on a single simulated sample.

The results obtained in the following sections, using this single simulated data sample, depend of course on the sample itself. Although it has been simulated from log-normal dynamics, well known in the literature, it is still highly correlated. The results and performances of the algorithms can therefore vary depending on the correlation matrix of the sample used. Thus, the results presented in the rest of this PhD thesis depend on the correlation matrix defined above.

Finally, to see the impact of the correlation matrix and, in particular, the importance of the choice of the sample on the performance of the algorithms, two historical data samples will be used in the last part of this chapter. These two historical samples are presented in Section 3.5.2 and Section 3.5.3.

3.1.2 Parametrization of the algorithms

In this chapter, different missing data mechanisms will be applied to the simulated sample presented above. These mechanisms will then give rise to one or more missing data scenarios that will be imputed by the set of methods presented in the previous

chapter. Hence, it is now important to present (or remember) all the functions used in this chapter as well as the choice of their parameters.

Many implementations of the same algorithm can exist, so that using one without saying precisely under which software, with which function, with which parameters, etc. is meaningless. In practice, the documentation of a completion method must be as transparent and clear as possible, and the choice of software, functions and their parameters must necessarily be part of it. The previous chapter has often emphasized the importance of the replicability of the results, yet it is notably the transparency concerning these parameters that allows the good replicability of the results for the regulator.

Before going into the details of the parameters of each methodology, it is appropriate to review the material used. The analysis related to this PhD thesis and the implementations have been done in MATLAB, using the MATLAB R2017a version, but many algorithms implemented under R software by their authors have been called (using the R-4.0.3 version). The aim is to directly use directly the available methods to compare their results on both simulated and historical data.

All these software programs were on a 64bit computer with a 3.30 GHz processor, and 32 Gb RAM memory. It is with this configuration that the calculation times presented in this chapter were obtained. The computation times can be optimized using a parallelization technique but this has not been used here to compare all the algorithms on the same basis.

Last observation carried forward

This method is one of the simplest to implement, but being very frequently used, a MATLAB function was already available and usable. Thus, this method was applied through the *fillmissing* function. This function completes missing data by different basic methods and notably by the last observation carried forward method by specifying the *previous* option. Thus, missing data are replaced by the previous observed value.

In addition, in the case where the missing data are at the very beginning of the series, there is no last observed value, meaning that the LOCF method will not fill in these missing data. To ensure that no data are missing after the imputation method application, it is possible to take the next available observation. Thus, this kind of missing data value is replaced by the first observed value of the series. It is possible to manage this particular case by using the function *fillmissing* with the *next* option so that no missing data remain at the end of the completion.

Finally, as presented in Section 2.4.1, this method is applied to raw prices directly and not to price returns like any interpolation method. Using the LOCF method on returns leads to a modification of the price series after the missing data. An adjustment

must be made to recover the original price series, which is finally equivalent to using another method than LOCF. In addition, if the missing data are present in more than one column, the method works independently, column by column. The interpolation of one column has no impact on the others and vice versa.

Linear interpolation

The *fillmissing* function can also be used to interpolate linearly missing data. In the previous paragraph it is used to apply the LOCF method, but it can be reused here to apply a linear interpolation by specifying the *linear* option. Thus, missing data located between two observed data points are completed by a simple linear interpolation, as presented in Section 2.4.1.

Missing data not surrounded by observed data, in other words, missing data located at the very beginning or at the very end of the series, are completed by default by a linear extrapolation. The extrapolation implemented in MATLAB consists of reproducing the closest observed returns to reconstruct the missing data. This extrapolation, therefore, leads to a linear evolution of the series.

For the last observation carried forward method, the linear interpolation here applies to raw prices directly, and not to price returns, to avoid modifying the observed price series. Moreover, in the case where missing data are present in several columns, linear interpolation is performed column by column and independently of the other columns (one-dimensional method).

Brownian Bridge

Contrary to the two previous methods, no method has been implemented to interpolate the data by a Brownian Bridge in MATLAB. Thus, this method has been implemented for this PhD thesis based on its theory presented in Section 2.4.2. It thus consists of replacing the missing data surrounded between two observations by forming a Brownian Bridge, and in extrapolating the missing data that would be located at the very beginning or at the very end of the series (missing data that are not surrounded by observed data) by injecting a the Brownian motion. The parameters of the Brownian motion are based on the available observed price returns.

This interpolation method, like all interpolation methods, is applied to gross prices and completes the missing data column by column, independent of the other results.

K-nearest neighbors

The K-NN method has been integrated into an improved algorithm using bagging with a cross-validation step to determine the right number of neighbors to use, and is organized in such a way as to treat the columns in ascending order of missing data.

To implement this improved bagged K-NN algorithm, the functions *knnimpute*, *bootstrp* and *crossval* from the MATLAB software are used. Moreover, this algorithm requires the setting of several parameters:

- The number of bootstrapped samples for the bagging step is fixed at 100. Each of the observations from these bootstrapped samples is randomly drawn with replacement from the original sample.
- The number of folds used in the cross-validation procedure is set at 10 (the default number),
- The metric used in the K-NN method is the Euclidean distance.

The improved K-NN algorithm is used to impute the original sample for every possible K (between 1 and $p-1$). Finally, the average imputed sample retained is the one among the K, which minimizes the prediction error calculated in the cross-validation step.

Moreover, concerning the choice of the number of neighbors by cross-validation, this method is not very discriminating in practice. Afterward, a sample of simulated data will be presented and used to compare the completion methods. However, the KNN cross-validation step almost systematically leads to imputing the missing data of a column from all the other available columns. More precisely, the simulated sample presented hereafter is made of 10 columns, and the cross-validation concludes that the use of 9 is optimal for 98.9% of the missingness scenarios tested on the simulated data (across all missingness mechanisms). There are some rare scenarios where cross-validation leads to the use of 8 or 7 columns but this remains an exception.

Finally, this algorithm is applied to price returns, making the hypothesis that the missing information depends on the dynamics of the other series (and not on their price level). Moreover, as already mentioned, when missing data are on more than one column, the algorithm imputes each column in ascending order of missing data.

Multivariate singular spectrum analysis

The MSSA algorithm, being commercialized by the data provider Bloomberg, is not available for free. It was therefore necessary to implement it entirely to carry out this comparative study based on all the theory presented in Section 2.4.4.

The algorithm has therefore been implemented in MATLAB software and requires many parameters to be set:

- The window size L used to form the trajectory matrix. As discussed earlier in Section 2.4.4, the Bloomberg research teams recommend in their paper [64] to set

this parameter to 60 for daily data, which represents approximately one quarter of the total.

- The number of eigenvalues K used for the reconstruction step. All eigenvalues are used to predict missing data. Thus, the number of eigenvalues used is exactly equal to the number of lines in the trajectory matrix, that is, $p \times L$, with p being the number of variables in the sample.
- The number of iterations to improve the imputations. Dash and Zhang [64] explain that, for example, imputation converges after 5 to 10 iterations. The maximum number of iterations used here is set at 10.
- The scaling option adds a step to scale the data to reduce noise, as proposed by Dash, Yang, Stein and Bondioli [63] in 2016. Scaling the data leads to optimizing the algorithm's performance according to them, which is why this option has also been implemented and is used in the applications below.

Finally, the MSSA algorithm is applied to a series of raw prices (rather than returns) for the reasons discussed in Section 2.4.4. Thus, the input of the algorithm is the price matrix directly.

Random forests

Regarding random forests, the function that has been put forward in Section 2.4.5 is *missForest* (from the R package with the same name) implemented under R software by Stekhoven and Bühlmann [194]. It consists of imputing missing data, column by column, in ascending order of missing data (like the K-NN method).

Thus, like the other algorithms, random forests require setting some parameters:

- The number of decision trees in the random forests, which correspond to the number of bootstrapped samples where observations are randomly drawn with replacement from the original sample, is set at 100 to obtain an average imputed sample based on a large number of imputations.
- The number of variables (columns) randomly sampled at each split which is set (by default) to $\sqrt{p}$, which corresponds to three variables for the next simulated sample (but also for the historical samples).
- The maximum number of iterations to perform, in case the stopping criterion is not reached, is arbitrarily set at 10 because, according to Stekhoven and Bühlmann [194], the stopping criterion is often reached below 10 iterations (5 iterations are enough in their study).

The random forests algorithm is applied to price returns, rather than raw prices, to be consistent with the dynamics of the series and to avoid jumps in spot prices.

Improved EM algorithm: Amelia

The EM algorithm applied to bootstrapped samples, presented in Section 2.4.6, was implemented in R software by Honaker, Joseph, King, Scheve and Sigh [112] in 2002 to solve data imputation problems. This is the function *amelia*, which is available in the package *Amelia*.

Similar to the other methods using bootstrapping, the number of bootstrapped samples (where each observation is randomly drawn with replacement among the observations of the original series) is chosen by the user. To remain consistent with the study of Bodner [36] about the number of imputations for multiple imputation methods (presented in Section 2.4.6) and with the other methods, this number is set at 100.

It is also possible to specify the initial parameters of the EM algorithm (the mean and the covariance matrix). By default, the algorithm calculates these parameters from the total available observations, in other words, using data from a listwise deletion. This is also the method recommended by Little and Rubin [145] in their book to initialize the parameters, because it provides consistent parameter estimates if the data are MCAR, and if there are strictly more complete observations than variables. Thus, this is the method used for the rest of this PhD thesis.

The function *amelia*, whose purpose is to estimate the distribution of the data, is applied to a price return matrix. There is no need to standardize returns because it is already implemented in the *amelia* algorithm.

Multivariate Imputation by Chained Equations

The MICE algorithm with the PMM method was implemented by Van Buuren [54] through the *mice* function under the R software package available with the same name(*MICE*).

Thus, the parameters chosen for the application of this algorithm are as follows:

- the number of imputed samples is, as in the other multiple imputation methods, set at 100;
- the option *PMM* is specified impute missing values by PMM (even if it is not specified, PMM is used by default in the case of continuous data);
- the number of donors, specific to the PMM model, is set at 5 as recommended by Van Burren [54]; and
- the algorithm can be iterated to improve the results, and the number of iterations is set at 10.

It is also possible to use an option to not remove a variable that is collinear to another. Van Buuren [54] explains that the MICE algorithm is implemented with safety measures, to not consider, for example, collinear variables, but it is also to bypass this safety measure by specifying it in the *mice* function.

Finally, the MICE with PMM method is applied to price returns, as applying it to spot prices would, logically, make it less efficient. Spot prices can be highly variable over time; therefore, the PMM method would tend to look for information close in time, which would probably lead to fewer realistic possibilities of response if the time series contains many missing data. On the other hand, if the method is applied to price returns, then it would be possible to find similar observations both near and far in time. The financial series are characterized by heteroskedasticity, and this method makes it possible to take into account variations in volatility over time.

Iterative PCA

As presented in Chapter 2, two different iterative PCA methods are applied in this chapter. The first consists of using the iterative classical PCA method implemented by Van Den Bossche in the *MacroPCA* algorithm through the *ICPCA* function from the R package *cellWise*. The second method used in this comparative analysis, is the regularized iterative PCA associated with multiple imputation, using the *MIPCA* function from the R package *missMDA* implemented by Husson and Josse [127].

The iterative classical method requires:

- the number of principal components to be used. This number is computed according to the generalized cross-validation method implementing the function *estim_ncpPCA* (available in the *missMDA* R package) by specifying the use of the iterative classical PCA method;
- the maximum number of iterations is set at 20 by default and is not changed; and
- its tolerance threshold is set at 0.005 by default and is not changed.

The MIPCA requires:

- the number of principal components; this number is computed in the same way as for the iterative PCA but by specifying the use of the regularized method; and
- the number of imputed samples is set at 100 like other multiple imputation methods.

In addition, the MIPCA algorithm scales the data by default, to give the same weight to each variable.

These two algorithms are applied to price returns to classify the dynamics of the financial data.

Now that the implementation and parameterization choices of each model have been presented, it is appropriate to present the tools that will be used for the comparative analysis.

3.1.3 Comparison tools

To compare the effectiveness of each of the completion methods, different comparison tools are used in this chapter. First, it is necessary to understand how these tools are calculated by the algorithms, as they do not all generate the same number of outputs depending on whether a single or multiple imputation method is used. Then, the different tools used to compare these algorithms will be presented.

Calculation of comparison tools

Thus, for each sample and missing data mechanism used, comparisons are made to highlight the performances, strengths and weaknesses of the methods to better understand their potential. However, knowing that not all algorithms generate the same number of outputs, it is necessary to explain how to calculate the comparison tools by defining the three main categories of imputation methods:

- single imputation methods without random component,
- single imputation methods with random component and,
- multiple imputation methods.

First, the simplest case is that of **simple imputation methods without a random component**; these can be define as methods that always obtain the same output (imputed sample), or in other words, methods where the output is replicable. These are methods such as:

- linear interpolation,
- LOCF,
- MSSA and
- IPCA.

Thus, for a sample containing missing data, these imputation methods always lead to the same imputed sample, and thus to the same comparison tools.

The second case concerns **simple imputation methods with random components**. These are methods that result in a single output (imputed sample) but are not uniqu, in the sense that they are not replicable due to a random component. In other words, these methods obtain an imputed sample that may be different from one run to another. These methods are:

- Brownian bridge,
- K-NN and,
- random forests.

Although the random forests and K-NN methods use bootstrapping, they still result in a single imputed sample as output due to the bagging process. These methods consist of imputing several bootstrapped samples, and the output corresponds to the average of all imputed samples.

Since the imputed samples resulting from this method category (simple imputation methods with random components) may vary from one run to another, they were repeated 100 times to analyze the results on average, and not from a single run. Thus, these imputation methods result in 100 imputed samples on which each comparison tool is applied before being averaged.

Finally, the last case concerns **multiple imputation methods**, which are methods that obtain several outputs (imputed samples). The multiple imputation methods used here are:

- Amelia,
- MICE and,
- MIPCA

The purpose of a multiple imputation method is to calculate comparative tools from each imputed sample before averaging these tools. As Honaker, King and Blackwell [113] explain, in practice, the analysis tools are calculated from each of the imputed samples before being averaged. In this PhD thesis, multiple imputation methods will result in 100 imputed samples, allowing 100 comparison tools to be calculated before being averaged.

Thus, the categorization of each algorithm is presented in Table 3.1-1. Single imputation methods without random components are run once and provide a unique imputed sample and thus a unique set of comparison tools. Simple imputation methods with a random component are run 100 times to generate 100 imputed samples. The comparative tools are computed on each of these imputed samples to obtain a set of average

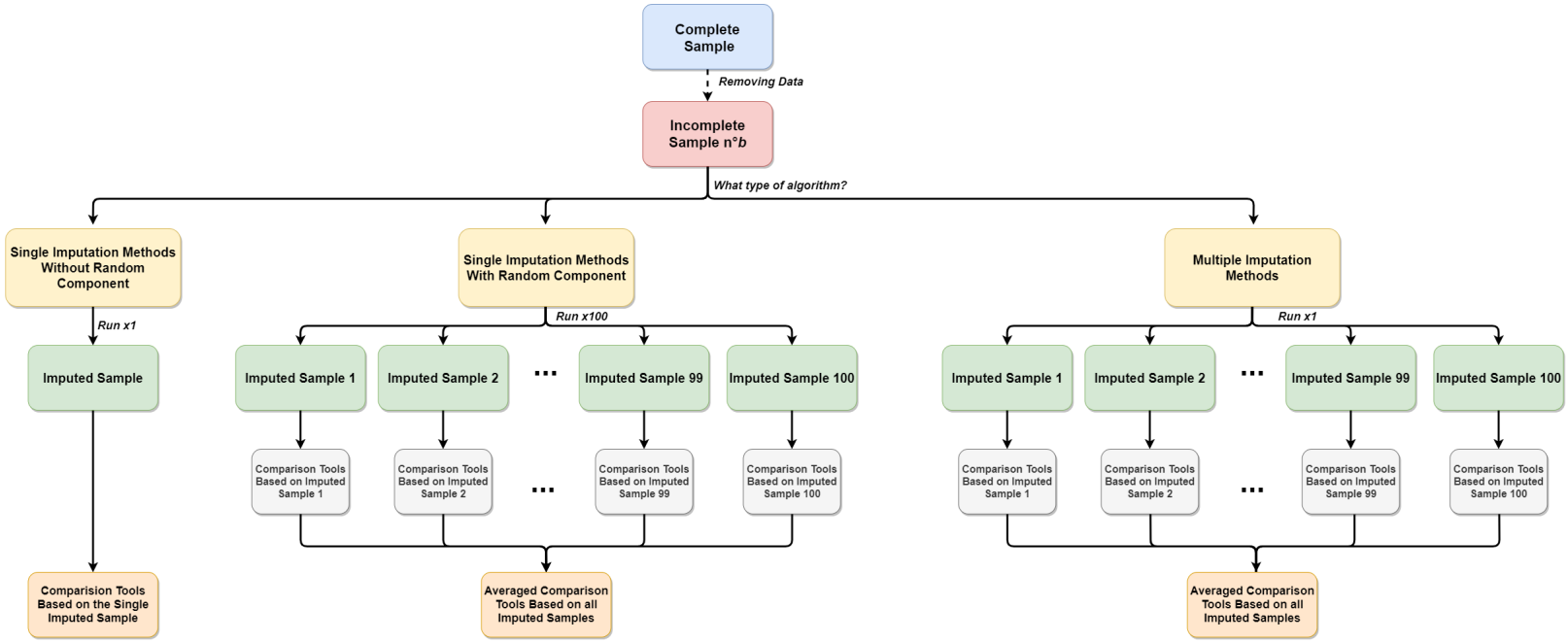
comparative tools. Finally, multiple imputation methods generate, by construction, 100 imputed samples. Then, the comparative tools are computed on each of these imputed samples to obtain a set of average comparative tools.

Tab. 3.1-1: Completion method categorization, number of outputs per run and number of runs needed in this PhD thesis

Category	**Imputation Methods**	**Number of output per run**	**Number of runs**
Single Imputation Methods without random component	Linear interpolation LOCF MSSA IPCA	1	1
Single Imputation Methods with random component	Brownian bridge $K-$-NN Random forests	1	100
Multiple Imputation Methods	Amelia MICE MIPCA	100	1

Knowing this, the process of data completion and the comparative tool calculations are represented in Figure 3.1-2.

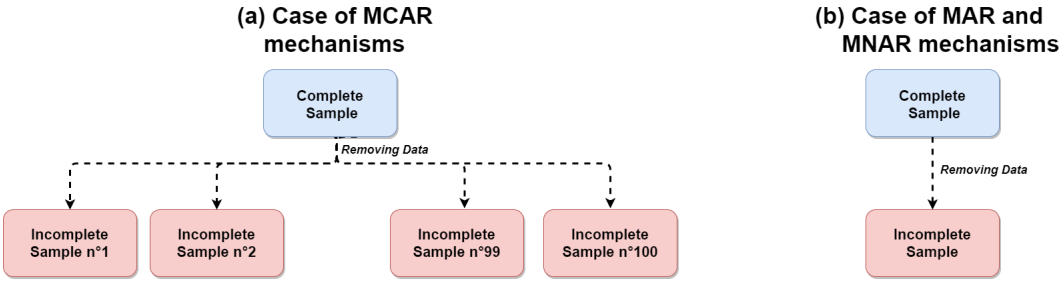
Fig. 3.1-2: Process of missing data completion and calculation of comparison tools according to the imputation algorithm category. This process is shown for the b^{th} missingness scenario.



Thanks to this process, each imputation method leads to a unique set of comparison tools for each sample containing a specific scenario of missing data. When the completion algorithm includes a random component, then the final comparison tools is the average of the comparison tool of each imputed sample. It is on this basis that the algorithms can then be compared.

The process presented in Figure 3.1-2 considers that a single sample containing missing data is obtained from a complete sample. This depends, of course, on the missing data mechanism used. Some mechanisms imply, by construction, a single missingness scenario, while others allow the generation of a large number of missingness scenarios. As presented in Figure 3.1-3, in the context of this PhD thesis, the MAR and MNAR mechanisms used will generate a single missingness scenario per data sample, while the MCAR mechanism, integrating a random component, will generate several (100 to be precise).

Fig. 3.1-3: The MCAR mechanism results in 100 missingness scenarios while the MAR and MNAR mechanisms result in a single missingness scenario.



For MAR and MNAR mechanisms, the b^{th} incomplete sample is the only one incomplete scenario ($b = 1$ in Figure 3.1-2) whereas, for the MCAR mechanism, 100 incomplete scenario are generated, then Figure 3.1-2 is performed for each $b = 1, \ldots, 100$.

Now that the way to calculate the comparative tools has been clarified, it is time to present the comparative tools themselves.

Mean absolute error and root mean square error

Different measures are used in this chapter to highlight the imputation error of each of the methods to compare them. These are proximity measures that aim to compare the imputed series with the fully observed original series. Thus, two proximity measures that have already been used in this PhD thesis (see Section 2.4.4 and Section 2.4.6) are also used in this chapter: the mean absolute error (MAE) and the root mean square error (RMSE). As they have already been used, these two measures are briefly presented below.

The purpose of these two measures is to quantify the proximity between the returns of the original series $\boldsymbol{Y}$ (fully observed) and those of the imputed series $\boldsymbol{Y}^*$. More precisely, $\boldsymbol{Y}^*$ corresponds to the $\boldsymbol{Y}$ series to which some complete data have been removed and then imputed. The proximity measures involve comparing y_i and y_i^* corresponding to observations of, respectively, $\boldsymbol{Y}$ and $\boldsymbol{Y}^*$. The proximity measures only take into account the observations that have been imputed in $\boldsymbol{Y}^*$. Thus, for $i = 1, \ldots, n^{obs}$, $y_i = y_i^*$ whereas this equality is not (necessarily) true for $i = 1, \ldots, n^{miss}$. In other words, the MAE and RMSE only compute the distance between the original data and imputed data.

In a missing data framework based on financial time series, the MAE and the RMSE are computed from price return series; in other words, $\boldsymbol{Y}$ and $\boldsymbol{Y}^*$ contain the original and imputed price returns, respectively.

Thus, the MAE measures the average of the differences between the original and imputed observations without taking into account their direction. Its computation consists of the mean of the absolute differences between y_i and the y_i^* for all $i \in n^{miss}$. More formally, the MAE is computed as follows:

$$MAE = \frac{1}{n^{miss}} \sum_{i=1}^{n^{miss}} |y_i - y_i^*|. \tag{3.1-4}$$

The differences between the observed values y_i and the imputed values y_i^* could take on both positive and negative values, hence the presence of the absolute value, so that these differences do not cancel each other out.

The RMSE is a quadratic measure that aims to compute the square root of the average of squared differences between y_i and y_i^* as follows:

$$RMSE = \sqrt{\frac{1}{n^{miss}} \sum_{i=1}^{n^{miss}} (y_i - y_i^*)^2}. \tag{3.1-5}$$

The differences are squared here, whereas previously they were converted into absolute values. As in the case of the MAE, the RMSE measure allows positive and negative deviations to be taken into account without offsetting each other. In both cases, the lower the value given by the measure, the more accurate the imputation is.

On the other hand, in the case of RMSE, the fact that the differences are squared before being averaged allows one to focus on large differences, especially extreme values. Due to its definition, the RMSE gives high weight to large deviations, which completes the information given by the MAE. In the case where the RMSE is much higher than the MAE, it means that the errors are of different magnitudes.

Covariance matrix differences according to the Frobenius norm

In a financial framework, it is also interesting to compare imputation methods in terms of the covariance matricesx. For this, the Frobenius norm is usually used, which corresponds to the matrix norm. It consists of the root of the sum of all the elements of the matrix in absolute value and is squared.

Thus, a good imputation method minimizes the differences between the covariance matrix obtained from the returns of imputed data and the covariance matrix obtained from the returns of original data. Thus, let $\boldsymbol{C}$ be the covariance matrix computed from the returns of original data $\boldsymbol{Y}$, and $\boldsymbol{C}^*$ the covariance matrix obtained from the returns of the imputed data $\boldsymbol{Y}^*$. Contrary to the proximity measures previously presented, the covariances are computed on the whole set of observations of $\boldsymbol{Y}$ and $\boldsymbol{Y}^*$, i.e., for $i = 1 \ldots, n$ (and not only $i = 1 \ldots, n^{miss}$). These covariance matrices are of dimension $(p \times p)$ since $\boldsymbol{Y}$ and $\boldsymbol{Y}^*$ contain p columns. Thus, the differences between both covariance matrices, with respect to the Frobenius norm, are calculated as follows:

$$\|\boldsymbol{C} - \boldsymbol{C}^*\|_F = \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{p} |c_{ij} - c_{ij}^*|^2} \tag{3.1-6}$$

where c_{ij} and c_{ij}^* are the elements of the i-th row and the j-th column of $\boldsymbol{C}$ and $\boldsymbol{C}^*$, respectively.

Some problems in certain fields, such as portfolio management, do not directly exploit the series data itself but rather their covariance and correlation with other series, hence the interest in comparing the imputation methods by comparing the results obtained from the covariance matrix.

Value-at-risk and expected shortfall

A completion method can also be the source of a significant impact on risk measures and therefore on the amount of capital charges paid by banks. If a completion method is used, the banks must, therefore, find the one that allows them to complete the missing data as accurately as possible to reflect, as well as possible, the risks incurred by the bank.

A completion method that leads to overestimation of the risk taken by the bank would force it to pay out huge capital charges. Conversely, if the completion method underestimates the risk, regulatory authorities may reject it, leading the bank to pay unexpected additional charges. Moreover, even if the regulatory authorities validate a completion method that underestimates risk, then the bank will not have sufficient reserves to deal with a crisis, which could lead to bankruptcy.

Hence, adding value-at-risk (VaR) and expected shortfall (ES) to this comparative analysis is indispensable.

VaR is a risk measure that has already been used in Section 2.4.6 to compare the performance of the Amelia algorithm compared to the classical EM algorithm in terms of VaR. This risk measure is frequently used as a risk measure in recent regulations. The VaR corresponds to an amount of losses that should only be exceeded with a given probability α, over a given time horizon t. More formally, the VaR_α^t is nothing more than a quantile of the distribution of realized profits and losses. Hence, the VaR of confidence level α and time horizon t is as follows:

$$VaR_\alpha^t = F_{\boldsymbol{Y}}^{-1}(\alpha), \tag{3.1-7}$$

where $F_{\boldsymbol{Y}}^{-1}$ corresponds to the quantile distribution function associated with $\boldsymbol{Y}$ the profit and loss distributions (spot price returns).

The Basel 2.5 reforms recommend the calculation of a $VaR_{99\%}^{1-day}$ to use it for regulatory backtesting, which corresponds to the amount given by a historical VaR with a probability set at 99% and with a 1-day horizon. There are different methodologies for calculating VaR, but the methodology recommended by regulators is the historical method, which simply consists of using historical data. The 1-day horizon aims to use a distribution of daily returns (in other words, returns based on 1 day), and generally, this distribution includes the last 250 returns. Thus, this risk measure tells the bank that there is a 99% chance that the loss observed on one day is less than the amount given by the $VaR_{99\%}^{1-day}$.

The Basel Committee also recommends the daily computation of a 99% VaR with a 10-day horizon for the capital metrics. The regulation specifies that the $VaR_{99\%}^{10-day}$ cannot be deduced by the $VaR_{99\%}^{1-day}$, using a square root to expand the daily horizon. Thus, in this PhD thesis, $VaR_{99\%}^{10-day}$ is computed using the overlapping 10-day price returns from the past year, which means the 250 10-day price returns where each price return is defined as $ln\left(\frac{S_t}{S_{t-10}}\right)$, with S_t being the stock price at time t.

The other risk measure that is frequently used in addition to the VaR is the ES. It consists of observing losses beyond the VaR through their average. One of the reasons why VaR is often criticized is that the risk is only represented by a single value, without giving any information about losses beyond the quantile defined by VaR. Thus, the ES offers a more complete analysis than the one made by the VaR through the averages of the worst losses beyond the VaR.

Thus, similar to VaR, the ES requires two parameters: probability α and time horizon t. The ES is, therefore, defined as follows:

$$ES_\alpha^t = \frac{1}{\alpha}\int_0^\alpha F_{\boldsymbol{Y}}^{-1}(p)dp \tag{3.1-8}$$

For the VaR, Basel 2.5 recommends conducting regulatory backtesting based on an ES of probability 97.5%, and with a 1-day horizon and capital metrics based on a

horizon of 10 days. Thus, $ES_{97.5\%}^{1-day}$ and $ES_{97.5\%}^{10-day}$ are used in this PhD thesis to compare the completion methods on these risk measures.

3.1.4 Graphical comparison process

Now that all imputation processes are understood and all the comparison tools are calculated, it is necessary to present the different types of graphs that will be presented recurrently in the rest of the chapter. The analysis and comparison process is divided into six parts:

- preliminary results (absolute differences for one scenario),
- statistical moments,
- proximity measures,
- covariance matrices comparison,
- risk measures and,
- computation time.

In addition, this process uses graphics with the following abbreviations to designate their model:

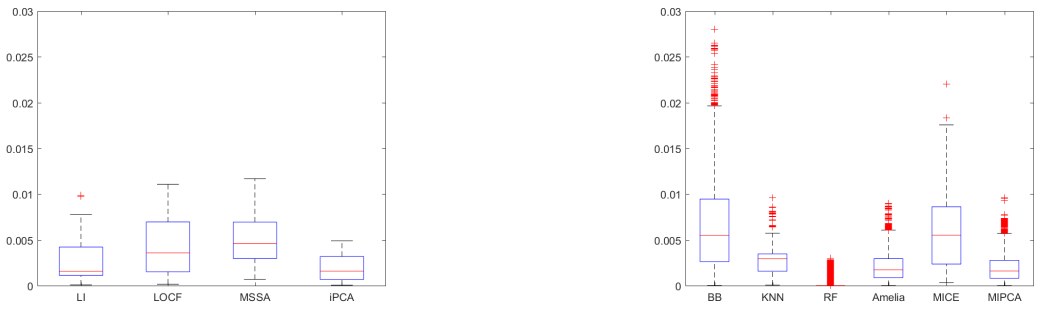
- *original* for original data,
- LI for linear interpolation,
- $LOCF$ for last observation carried forward,
- BB for Brownian bridge,
- KNN for K-nearest neighbors,
- $MSSA$ for multivariate singular spectrum analysis,
- RF for random forests,
- *Amelia* for the EM algorithm with bootstrapping implemented through the Amelia algorithm,
- $MICE$ for multiple imputation by chained equations,
- $IPCA$ for iterative principal component analysis and finally
- $MIPCA$ for the IPCA with multiple imputations.

The rest of this section is not intended to comment on the results of the graphs, but simply to present them, how they are calculated and what they highlight.

Graphical presentation: preliminary results

Before comparing the algorithms using the different tools presented above, a preliminary analysis is conducted, consisting of observing the absolute differences between the returns of the first column from the complete original series and those from the imputed series. These absolute differences are represented by whisker boxes and are based on a single missing data scenario (the first one simulated in the case of MCAR data and the only scenario in the case of MAR and MNAR). On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The red crosses represent the outliers. For m% of missing data (10% and 30% in practice), these absolute deviations are represented by Figure 3.1-4.

Fig. 3.1-4: Distribution of absolute return differences between the original series and imputed series for a single missingness scenario containing $m\%$ missing data



(a) Methods without a random component for $m\%$ of missingness

(b) Methods with a random component for $m\%$ of missingness

The y-axis corresponds to the absolute deviations between the returns of the original and the imputed data. For example, in the case of linear interpolation, it is possible to see that the imputations give returns that are at most 0.01 (or 1%) away from their original return. As this section is dedicated to the presentation of the graphs, their interpretation will be done in the following sections.

The simple imputation methods without random components are represented in a graph (see Figure 3.1-4a), and the (simple or multiple) imputation methods with a random component are presented in another graph (see Figure 3.1-4b). This is because in the first case, the absolute deviations are those of a single imputed sample, while in the second case, they are those of 100 imputed samples. In other words, the complete sample (represented in blue in Figure 3.1-2) is compared with all imputed samples (represented in green in Figure 3.1-2). Thus, Figure 3.1-4b contains 100 times more data than Figure 3.1-4a. Moreover, this explains the reason why there are many more

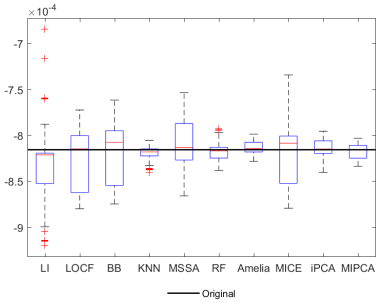
outliers in Figure 3.1-4b than in Figure 3.1-4a.

Graphical presentation: the first four statistical moments

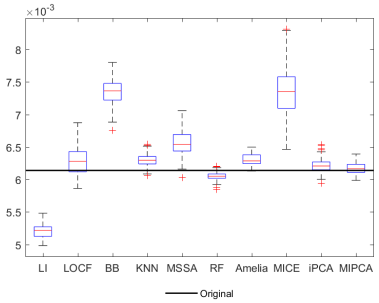
Now, the comparison tools can be used to interrogate the algorithms. The first comparative analysis concerns the results obtained in terms of statistical moments. For this, the first four statistical moments (mean, standard deviation, skewness and kurtosis) were calculated, following the process presented in Figure 3.1-2, to analyze and compare the results. These moments are computed from the returns of the first column of the imputed data and those of the first column of the original data.

When the missingness mechanism generates 100 missingness scenarios (MCAR mechanism in this PhD thesis), a specific focus is made. The missingness proportion is set at 30% to compare the results of the statistical moments of each of the 100 imputed samples (based on the 100 missingness scenarios generated). This consists of analyzing the distribution of comparison tools (in orange in Figure 3.1-2) obtained for each of the 100 incomplete samples. Thus, Figure 3.1-5a, Figure 3.1-5b, Figure 3.1-5c and Figure 3.1-5d represent the means, standard deviations, skewness and kurtosis, respectively, for each of these 100 imputed missing data scenarios. As previously described, the central mark of the whisker boxes indicates the median, and the bottom and top edges indicate the 25th and 75th percentiles, respectively. The red crosses represents the outliers.

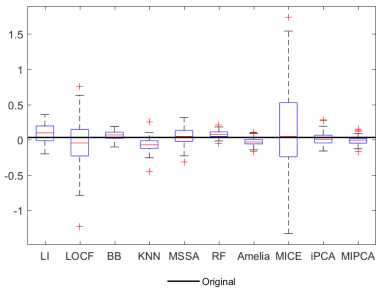
Fig. 3.1-5: Distribution of the first four statistical moments obtained from 100 missingness scenarios with 30% MCAR data



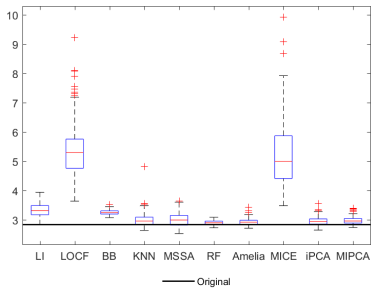
(a) Mean



(b) Standard Deviation



(c) Skewness



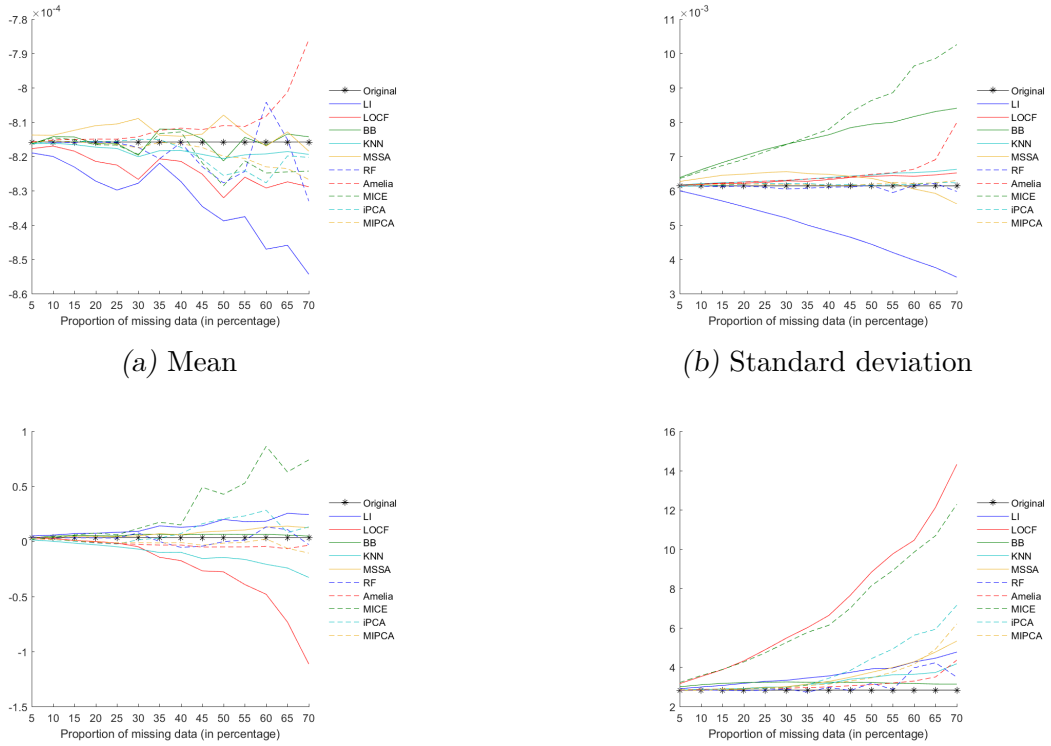
(d) Kurtosis

Thus, the levels of these statistical moments are represented on the y-axis. The statistical moments of the complete original series (the first column of the data matrix) are represented by the black line as the reference.

The second type of graph concerns all the missing data mechanisms (MCAR, MAR and MNAR) and shows the evolution of each statistical moment according to the proportion of missing data (calculated from the returns of the first column of the imputed matrix).

Thus, in the case of MCAR data, Figure 3.1-6a, Figure 3.1-6b, 3.1-6c and Figure 3.1-6d are, respectively, the mean, standard deviation, skewness and kurtosis averaged across the 100 missingness scenarios for each proportion of missing data presented. For each method, this consists of averaging the results obtained in orange in Figure 3.1-2. In the case where the data are MAR or MNAR, these same figures represent the statistical moments based on a single imputed sample for each missingness proportion. As already mentioned, these missing data mechanisms result in a single missingness scenario for each proportion tested, a single set of comparison tools is provided (in orange in Figure 3.1-2), so no averaging is needed. It is for these missing data mechanisms that the evolution of statistical moments (but also of other comparison tools) is the most unstable. This is because they are based on a single incomplete sample, and sampling errors are possible.

Fig. 3.1-6: The first four statistical moments of the returns of the imputed data according to the missingness probability



(a) Mean

(b) Standard deviation

(c) Skewness

(d) Kurtosis

As previously described, the y-axis represents the levels of these statistical moments. The moments of the complete original series (for the first column) are designated by the black starred lines. For example, the volatility of the original series is approximately 0.06% (i.e., 10% annualized), while the volatility of the linearly interpolated series decreases to approximately 0.035% when the series contains 70% missing data.

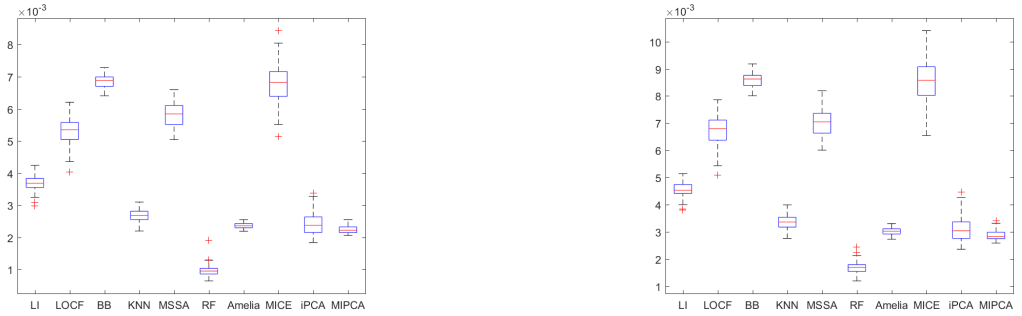
Graphical presentation: proximity measures

The completion methods will also be compared in terms of proximity measures and, in particular, the MAE and RMSE, presented previously. These measures are computed between the first column of the complete original data matrix and the first column of the imputed matrix. Thus, for the statistical moments, two types of graphs will be used in the case of MCAR data, against one for MAR and MNAR data.

First, in the case where the data are MCAR, 100 missingness scenarios are imputed. Then, it is possible to analyze the distribution of proximity measures computed from each of the 100 scenarios of missing data generated. Thus, as before, the analysis focuses on the 100 scenarios containing 30% missing data and, in particular, on the resulting

MAE and RMSE. This consists of analyzing the distribution of comparison tools (in orange in Figure 3.1-2) obtained for each of the 100 incomplete samples. Obviously, these proximity measures are calculated between the returns of the first column of the complete original matrix and those of the imputed matrix. These results are presented by graphs such as Figure 3.1-7a and Figure 3.1-7b. On each whisker box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The red crosses represent the outliers.

Fig. 3.1-7: Distribution of the MAE and RMSE computed from the 100 scenarios containing 30% of MCAR data



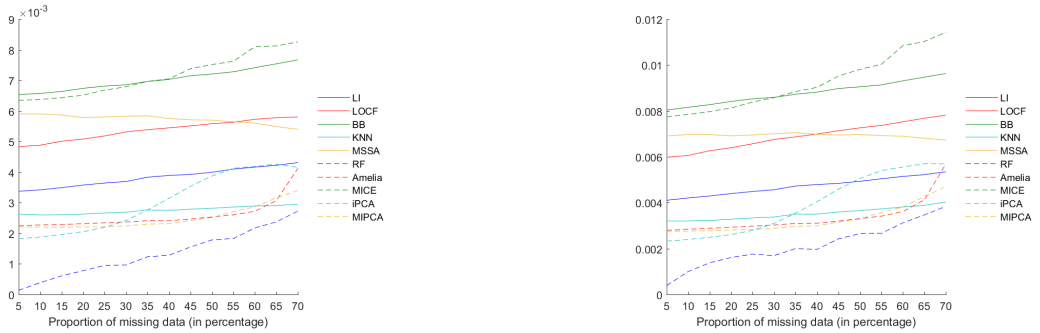
(a) MAE *(b)* RMSE

The y-axis represents the MAE and RMSE values. The closer they are to zero, the closer the imputed series is to the original series. Thus, for example, the 100 scenarios containing 30% missing data imputed by linear interpolation obtain MAEs between approximately 0.3% and 0.44%.

The algorithms can also be compared in terms of MAE and RMSE for different proportions of missing data. This allows for observation of the behavior of the algorithms regarding the proximity measures, in front of an increasingly empty sample.

For statistical moments, when the missingness mechanism is MCAR, the proximity measures shown for the graphs are the average of the MAEs and RMSEs. These averages are calculated from the 100 proximity measures obtained from the 100 missingness scenarios. For each method, this consists of averaging the results obtained in orange in Figure 3.1-2. Conversely, when the missingness mechanism is MAR or MNAR, a single missingness scenario is generated, which means that the reported values are the proximity measures of a single imputed missingness scenario (in orange in Figure 3.1-2). In both cases, these types of comparisons are made using Figure 3.1-8a and Figure 3.1-8b.

Fig. 3.1-8: MAE and RMSE from matrices containing missing data, according to the missingness probability



(a) MAE *(b)* RMSE

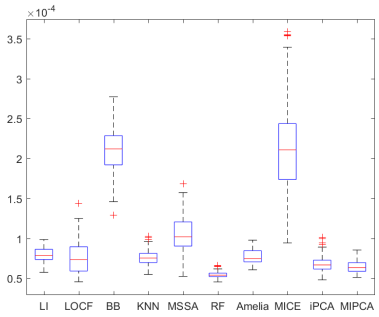
The y-axis represents the levels of MAE and RMSE obtained for each proportion of missing data. For example, when the series is linearly interpolated for 70% missing data, the MAE is approximately 0.4% (on average among 100 missingness scenarios if the data are MCAR).

Graphical presentation: covariance matrix comparison

Comparisons of covariance matrices are also made to illustrate the impacts in terms of portfolio management. Thus, the covariance matrices of the imputed data are compared to the original matrices, according to a Frobenius norm, as detailed above. While previously, the analyses focused on the first column of the matrix, here it is all the columns of the matrix that are used (thanks to the Frobenius norm).

As before, a focus is made for the 100 samples containing 30% MCAR data to see the distribution of covariance differences from one missing data scenario to another. This consists of analyzing the distribution of comparison tools (in orange in Figure 3.1-2) obtained for each of the 100 incomplete samples. This analysis is, thus, represented by Figure 3.1-9. On each whisker box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The red crosses represent the outliers.

Fig. 3.1-9: Covariance matrix differences, according to the Frobenius norm based matrices containing 30% MCAR data

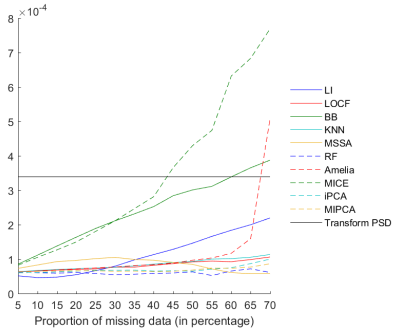


The y-axis thus represents the value of the covariance matrix differences according to a Frobenius norm. For example, the 100 covariance errors obtained for imputation by linear interpolation are between 0.00005 and 0.0001.

Following the same logic as the previous analysis criteria, it is possible to follow the evolution of the covariance differences with respect to an increasing proportion of missing data. This type of analysis is possible for any missing data mechanism.

This is represented by a graph like Figure 3.1-10. Thus, in the case of MCAR data, the average of the covariance differences (among the 100 missingness scenarios) are listed for each proportion of missing data. For each method, this consists of averaging the results obtained in orange in Figure 3.1-2. On the other hand, in the case of MAR or MNAR data, the single covariance difference obtained for each missingness proportion is reported (since only one missingness scenario is available).

Fig. 3.1-10: Covariance matrix differences, according to the Frobenius norm, from matrices containing missing data, according to the missingness probability



As before, the y-axis corresponds to the value of the covariance differences according to a Frobenius norm. The black line designated by $TranformPSD$ corresponds

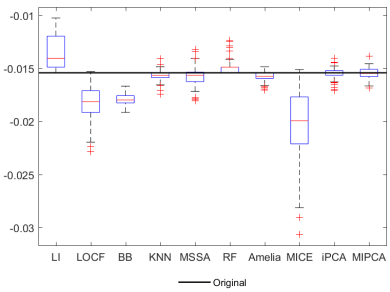
to the pairwise covariance matrix that has been transformed to be positive semidefinite, according to Rousseeuw and Molenberghs [173] (presented in Section 2.2.2). This method, which is traditionally used to make covariance matrices quickly manipulable, is a good benchmark here.

Graphical presentation: value-at-risk and expected shortfall

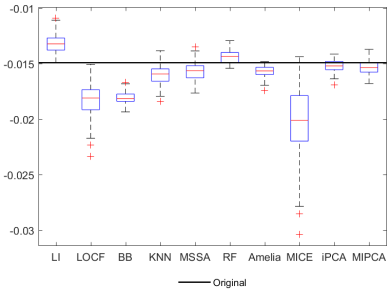
Finally, comparisons will also be made in terms of VaR and ES for 1-day and 10-day horizons. In this section, the horizon is noted d to simplify the explanation. The choice of horizons and confidence levels has been detailed previously. Thus, the risk measures of the first column of the return matrix are calculated, analyzed and compared to those of the return of the first column of the complete original matrix.

As before, a focus on the 100 scenarios containing 30% of the MCAR data will be analyzed to observe the variations in risk measures from one missingness scenario to another. This consists of analyzing the distribution of comparison tools (in orange in Figure 3.1-2) obtained for each of the 100 incomplete samples. The VaR of each of the missingness scenarios is thus represented by a graph such as Figure 3.1-11a, and similarly for the ESs with Figure 3.1-11b. On each whisker box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The red crosses represent the outliers.

Fig. 3.1-11: Distribution of the d-day risk measures computed from the 100 scenarios containing 30% MCAR data



(a) d-days VaR at 99%



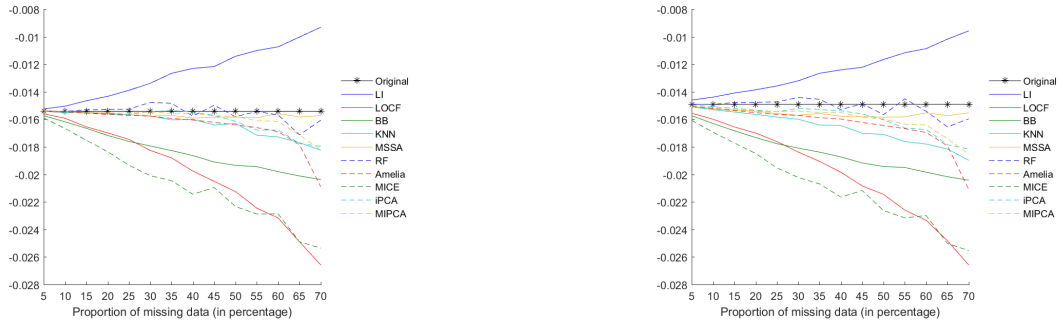
(b) d-days ES at 97.5%

The y-axis, therefore, represents the negative return corresponding to the risk measure and the black line represents the risk measure of the original series. For example, the VaRs obtained for each of the 100 scenarios containing MCAR data are above the level of the original VaR, and approximately between -0.01 and -0.016; that is, a loss between 1.0% and 1.6%, while the original VaR is 1.6%.

Finally, in a more general way, it is possible to analyze the impacts of the proportion of missing data on the risk measures for all missing data mechanisms.

As before, Figure 3.1-12a and Figure 3.1-12b represent, in the case of MCAR data, the average of the 100 VaRs and 100 ESs obtained among the 100 missingness scenarios for each proportion of missing data. For each method, this consists of averaging the results obtained in orange in Figure 3.1-2. In the case of MAR and MNAR data, the VaRs and ESs obtained for each proportion of missing data (from a single scenario).

Fig. 3.1-12: The d-day risk measures, computed from matrices containing missing data, according to the missingness probability



(a) d-days VaR at 99%

(b) d-days ES at 97.5%

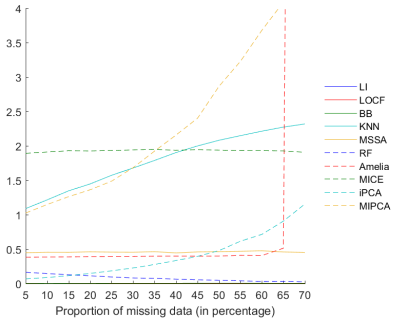
Thus, as in the previous figures (Figure 3.1-11a and Figure 3.1-11b), the y-axis corresponds to the return associated with the risk measure, and that of the original series is represented by the black starred line.

Graphical presentation: computation time

Finally, the last comparison of algorithms concerns the computation time necessary to execute each method, according to the proportion of missing data to be processed. Thus, these calculation times (in seconds) are represented by graphs such as Figure 3.1-13.

In the case of MCAR data, this computation time is averaged among the 100 runs made for each proportion of missing data. On the other hand, the results represented for MAR or MNAR data are those of a single run.

Fig. 3.1-13: Computation time of the missing data imputation according to the missingness probability



The y-axis corresponds to the computation time necessary, in seconds, to execute the algorithm. For example, the MSSA method requires a computation time of half a second, regardless of the proportion of missing data in the sample.

3.2 Imputation of data: MCAR on simulated Gaussian sample

To compare the efficiency of each algorithm presented in Chapter 2, all the methods are applied to simulated data (presented in Section 3.1.1), to which missing data are artificially added. Indeed, this procedure makes it possible to compare the original values artificially removed from the sample with the data imputed by the different algorithms, and then to compare the algorithms between them.

As presented in Chapter 2, there are different types of missing data, and this section is dedicated to MCAR data. Data are MCAR, according to Little and Rubin [145], if the missing data do not depend on the observed or missing data of the data matrix, in other words, if the probability of missingness is the same for each data. This definition is close to that of Schafer and Graham [182] because the MCAR data depend only on external reasons and are not related to the data matrix. Giving an explanation for all the missing data is not always obvious in a financial context. As mentioned in Section 2.2.1, some data can be deleted in a totally random way by operational staff, and others can be randomly missing due to failing IT processes, or even due to the trader forgetting to save the data manually. It is also possible to find this type of data missing when the financial product is illiquid, or when no price agreement has been reached between the buyer and the seller. These cases may result in randomly distributed missing data in the database, which could be categorized as MCAR by Little and Rubin [145].

3.2.1 Impact of MCAR data on the first column

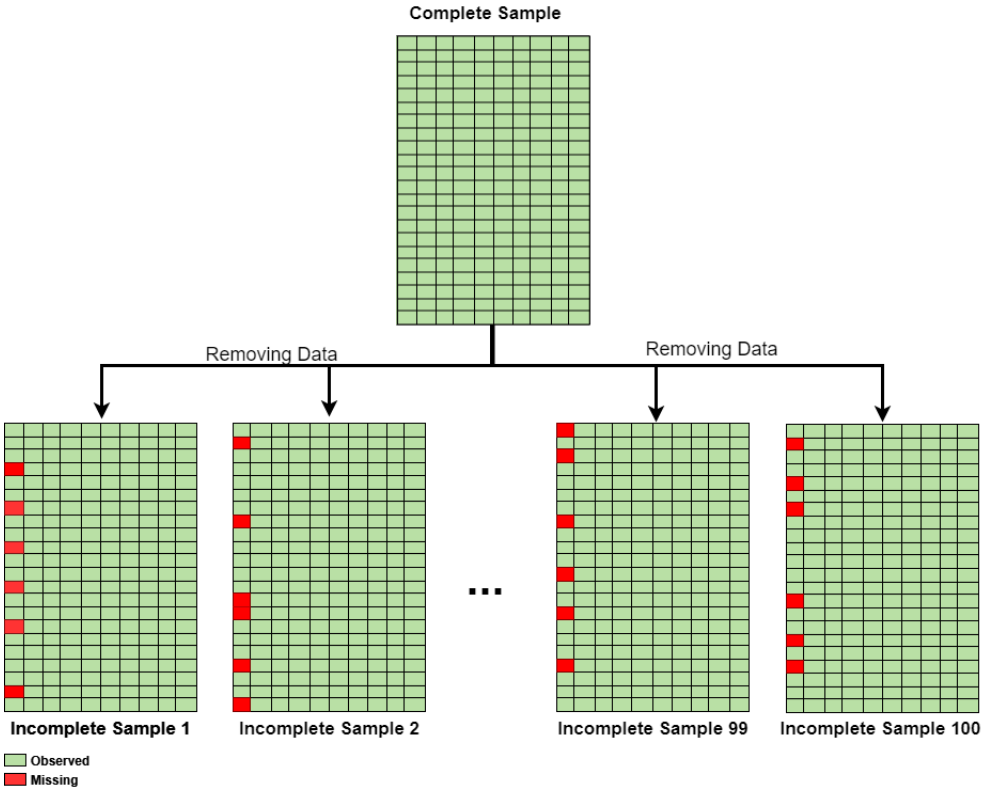
Thus, data are considered MCAR according to Little and Rubin [145] if each data point has the same probability of being missing. To remove data completely at random in the simulated sample presented in Section 3.1.1, a uniformly distributed random variable is used. This means that each value in the sample has the same probability of being missing. Depending on the result of this random variable, the data were removed or not. This mechanism is in agreement with the definition of Little and Rubin [145] since it does not depend on the data matrix, and the probability that data are missing is the same for each observation; it is also in agreement with the definition of Schafer and Graham [182] because the missing data mechanism depends on a uniform law totally independent of the data matrix.

Two main MCAR mechanisms are tested in this chapter. The first consists of removing the data in a MCAR way, based on a uniform distribution, in the first column of the price matrix. The second consists of removing data in a MCAR way in all columns, as presented in Section 3.2.2.

MCAR mechanism

As mentioned, the first MCAR mechanism tested consists of removing data only in the first column of the data matrix. Given that the data are missing from random draws, the results obtained depend on each of these draws. Thus, to avoid analyzing the results of a single missingness scenario, 100 missingness scenarios are drawn based on the same initial full sample to observe the results on average from this MCAR mechanism. This mechanism is represented in Figure 3.2-1.

Fig. 3.2-1: 100 missingness scenarios following a MCAR mechanism: randomly removing data in the first column of the data matrix



Finally, several levels of missingness probability are used to analyze the impact of increasing amounts of missing data. More precisely, this proportion is between 5% and 70% missing data per sample, by steps of 5% (which represents 14 proportions of missing data tested).

The procedure aims to compare the algorithms according to the proportion of missing data. Therefore, the idea here is, for each missingness proportion, to impute the missing data from each of the 100 missingness scenarios before analyzing the (averaged) results. Each comparison tools are computed as defined in Figure 3.1-2.

The missing data mechanism is applied here to the first column of the price matrix, but the algorithms are applied to prices or returns depending on the method. As already mentioned, the proportion of missing returns will depend on how much price data is missing in the sample. If the missing data are all pooled together, then the proportion of missing returns will be equivalent to the proportion of missing data. On the other hand, in the opposite case, this can introduce a large difference in the proportion of missing data and missing returns. Since the proportion of missing data on prices is not equivalent to the proportion on returns for this missingness mechanism, Table 3.2-1 highlights the proportion (and quantity) of missing returns associated with

the proportion of MCAR prices injected in the first column of the simulated sample.

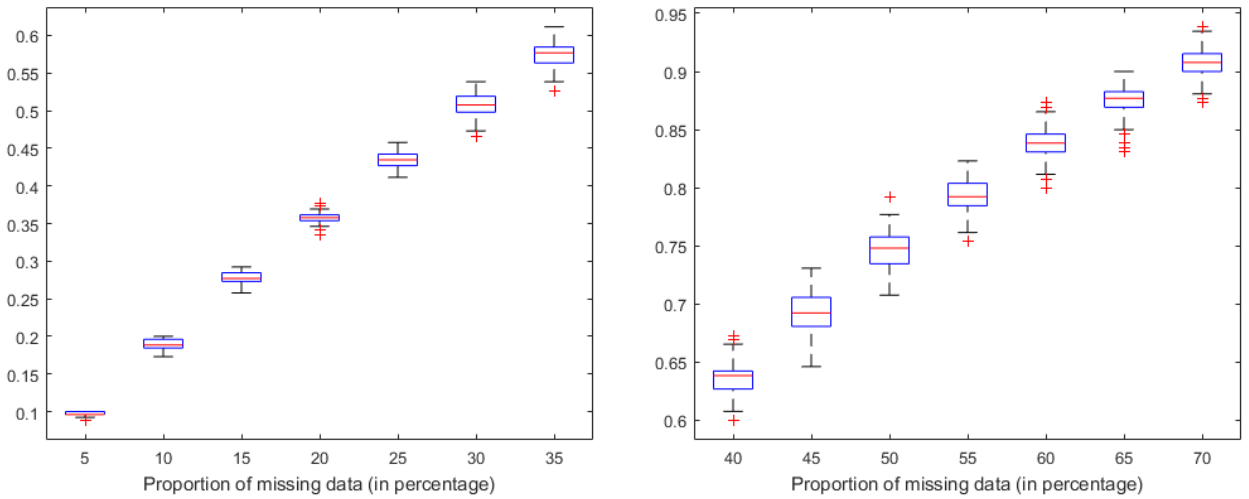
Tab. 3.2-1: Average proportion (number) of missing returns (among the 100 missingness scenarios) associated with the proportion of MCAR raw data injected into the first column of the simulated sample of length 261 (260 for return sample)

	Proportion (and number) of missing returns associated with missing data													
Data	**5%** (13)	**10%** (26)	**15%** (39)	**20%** (52)	**25%** (65)	**30%** (78)	**35%** (91)	**40%** (104)	**45%** (117)	**50%** (130)	**55%** (143)	**60%** (156)	**65%** (169)	**70%** (182)
Return	10% (25)	19% (49)	28% (72)	36% (93)	44% (113)	51% (132)	57% (149)	64% (165)	69% (180)	75% (194)	79% (207)	84% (218)	87% (227)	91% (236)

Thus, with this type of random missing data mechanism, algorithms applied to returns must impute between approximately 25% to 50% more data than those applied to prices directly. In an extreme scenario where 70% of the prices are missing, an algorithm using returns will have to impute on average 91% of the first column of the sample. Such levels of missing data are excessive but realistic when put in conjunction with the NMRF eligibility tests presented in Section 1.3.4, where 24 real price observations are required.

In addition, Figure 3.2-2 represents the distribution, among the 100 missingness scenarios, of the proportions of missing returns that the algorithms using returns are dealing with.

Fig. 3.2-2: Distribution of the proportion of missing returns according to the proportion of missing data (prices) injected in the first column of the simulated sample



From one missing data scenario to another, the proportion of missing returns can vary significantly. For example, in the case where 25% of the prices were deleted, then the samples may contain a proportion of missing returns ranging from approximately 52% to 61%. This can introduce some variability in the results.

Finally, the templates of all the graphs presented in this section have been presented and detailed in Section 3.1.4 (what they represent and how they were obtained).

MCAR tests

Before imputing the MCAR data, MCAR tests are applied to observe whether the tests are able to detect the missing data as MCAR or not. In this section, data are artificially removed from the simulated sample following a MCAR mechanism; hence, the fact that the data are MCAR is known. On the other hand, in the case of an empirical study based on a historical sample, the mechanism of missing data is unknown. Thus, the application of MCAR tests on these missingness scenarios will reflect the reliability and robustness of these tests. These tests are often applied in the literature to justify the application of a particular method; however, there is no study comparing Little's test [142] and Jamshidian and Jalal's test [123] based on many missingness scenarios. This is why the application of these tests, on this simulated sample but also on historical ones, and with different missing data mechanisms, are interesting.

First, it is necessary to test whether the missing data, injected into these 100 missingness scenarios, are indeed MCAR according to Little's test [142] and to Jamshidian and Jalal's test [123] (presented in Section 2.1.2). Thus, these statistical tests were applied to each of the 100 missingness scenarios for each proportion of missing data.

To conduct this analysis, Little's test [142] was implemented for this PhD thesis, and Jamshidian and Jalal's test [123] were computed using the *MissMech* package (from R-software) to perform both the improved Hawkins test and the nonparametric test.

As a reminder, the data are considered MCAR according to the Little test [142] in the case where the null hypothesis is accepted. For Jamshidian and Jalal's test [123], the data are considered MCAR if the null hypothesis of the improved Hawkins test is accepted or the null hypothesis of the nonparametric test is accepted in the case where the null hypothesis of the improved Hawkins test is rejected.

Thus, both tests were applied to each of the 100 missingness scenarios and each missingness proportion. Little's test is always calculable, by construction, but this is not the case of Jamshidian and Jalal's test [123]. This test may fail because it does not consider missing data patterns (based on a raw data) that appear less than 6 times in the whole data matrix. This is a missing data pattern that appears less than 6 times, and the rows are deleted from the data matrix, leading to a matrix with no more missing data pattern. This parameter can be modified by the user, but 6 is set

by default and recommended by the authors. This is why, among the 100 scenarios of missing data generated for each missingness proportion, this function was not always able to reach a solution. The fact that this test may not be calculable for any matrix, may constitute an operational risk for the bank and implies that these tests cannot be used in an automatic data quality control process.

Table 3.2-2 represents, for both tests applied to return matrices, the probability of not rejecting H_0 when H_0 is true, among the tests that are calculable, with a 5% significance level. As the Jamshidian and Jalal's test can be incalculable, the details of the number of scenarios that is calculable is available in Appendix A.1. Appendix A.1 reveals that Jamshidian and Jalal's tests were performed without difficulty for all missingness scenarios for each of the fourteen proportions of missing data applied. Then, the results from Table 3.2-2 are based on all the missingness scenarios.

Tab. 3.2-2: Confidence level (probability of not rejecting H_0 when H_0 is true) for both MCAR tests applied to price return matrices containing MCAR on the first column of the matrix, for a 5% significance level

	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Little's test	97%	96%	97%	93%	93%	94%	99%	91%	96%	94%	93%	97%	97%	96%
J&J's test	95%	93%	94%	93%	89%	97%	94%	94%	99%	93%	96%	94%	91%	92%

From these results, both tests obtain almost the same confidence level for all missingness proportions. Moreover, these tests are able to detect that data are MCAR without being impacted by an increasing proportion of missing data. Little's test [142] obtains confidence levels between 91% and 99%, and Jamshidian and Jalal's test [123] obtains confidence levels between 89% and 99% of cases.

Thus, through this example, it appears that both tests are able to detect MCAR data, even when the proportion of missing data becomes too large. Based on these results, an expert could use these two tests to detect the MCAR nature of the data. MCAR

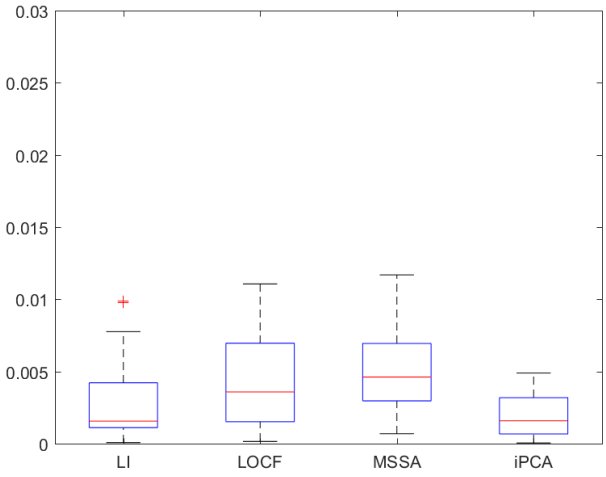
Preliminary results

The MCAR data contained in the first column of the matrix are now imputed to compare each completion method.

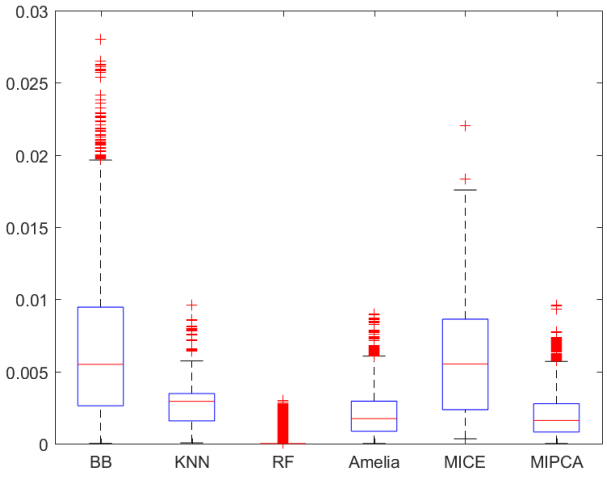
Hence, Figure 3.2-3 represents the distribution of the absolute return differences between the original data and imputed data from the first missingness scenario for 10% (at the top) and 30% (at the bottom) missing data. There are three times more absolute differences represented in the figures at the bottom than in those at the top.

As explained in Section 3.1.4, the distribution from simple imputation methods (linear interpolation, LOCF, MSSA and IPCA), represented on the left, are based (by definition) on a single imputation, whereas the distribution from methods containing a random component, on the right, is based on 100 imputations (all represented here).

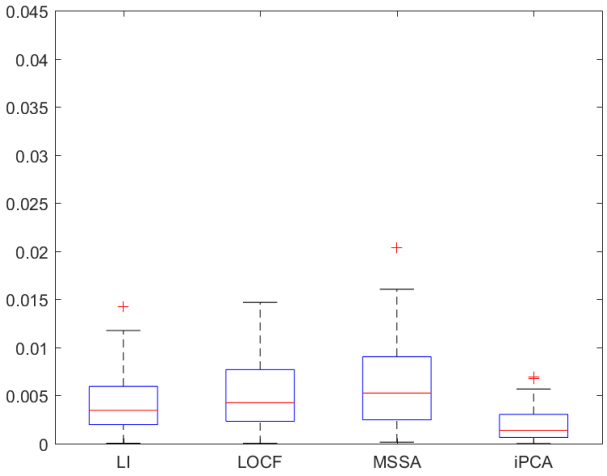
Fig. 3.2-3: Distribution of absolute return differences between the original series and imputed series for a single missingness scenario containing 10% (at the top) and 30% (at the bottom) MCAR data in the first column



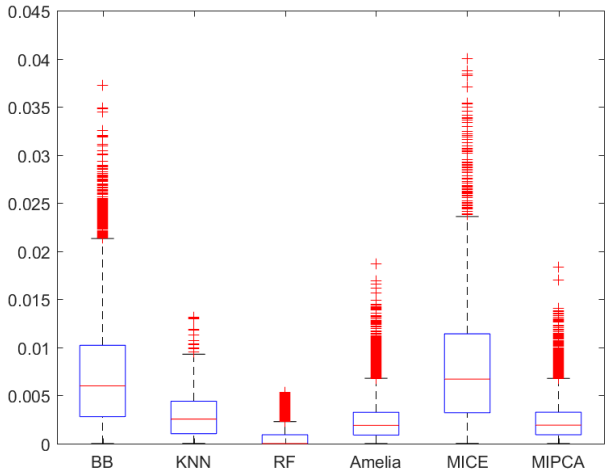
(a) Methods without a random component for 10% missingness



(b) Methods with a random component for 10% missingness



(c) Methods with a random component for 30% missingness



(d) Methods with a random component for 30% missingness

These figures report the behavior of the algorithms for a specific proportion of missing data. By comparing Figure 3.2-3a and Figure 3.2-3b with Figure 3.2-3c and Figure 3.2-3d, it can be seen that the more missing data are present in the sample, the more imputations tend to deviate from the original series. In fact, there are larger absolute differences when the missingness proportion is 30%: the distributions are more spread out and contain more outliers (represented by red crosses).

In addition, imputation methods with a random component (Figure 3.2-3b and Figure 3.2-3d) have more outliers in their results compared to the single imputation method (Figure 3.2-3a and Figure 3.2-3c). This is because there are 100 times more data from a random component in these distributions and therefore a higher probability of having an outlier. Moreover, the proportion of outliers increases with the amount of missing data, reflecting the difficulty for the algorithms to correctly impute data when the proportion of observed values decreases.

Looking now, algorithm by algorithm, at these figures, the random forests method is the one that obtains the smallest absolute deviations from the original series (the smallest distribution). These outliers are very concentrated around the distribution, such that all imputed values have absolute deviations of less than 0.5% from the original series (with both levels of missing data). In this case, the stability of random forests due to the bagging procedure, as Breiman [41] talked about, is verified.

The IPCA method also obtains small deviations from the original series but as its improved version (MIPCA) and the Amelia algorithm. Without considering outliers, these three completion methods have similar distributions. Moreover, MIPCA and Amelia appear to obtain very close results based on these figures (including extreme

values), with absolute differences under 1% with 10% missing data and under 2% with 30% missing data. Moreover, the K-NN method obtains results also comparable to the MIPCA and Amelia algorithms.

Some of these results were expected, while others were not. Since the sample is composed of Gaussian data, it is not surprising that Amelia is among the best performing algorithms. This algorithm is, indeed, based on the assumption that the data are Gaussian and seems to be perfectly adapted since it aims at estimating the parameters of a multivariate Gaussian distribution. Good results were also expected for the IPCA and MIPCA methods, which integrate a PCA, very frequently applied to financial data. PCA is indeed a method close to data analysis and factor analysis that is very frequently used in finance. PCA is still very much used to solve portfolio management problems and, in particular, portfolio selection, as Ioan [120] did in 2020, on crisis data and in particular on the COVID-19 crisis. On the other hand, the fact that random forests appear to perform even better than Amelia, IPCA or even MIPCA is rather surprising. If random forests have often been applied and compared to other models on medical data [194], applications on financial data, or just on Gaussian samples, are almost nonexistent. It was therefore difficult to anticipate such good performances for this algorithm. This algorithm was expected to be particularly effective for mixed-type data (continuous and categorical data), including complex interactions and nonlinear relationships, but the fact that it does even better than algorithms such as Amelia and MIPCA was not expected. In any case, the results presented above correspond to those of a single missingness scenario (for 10% and 30% missing data); hence, it remains to be seen if these good performances persist for the other scenarios.

If outliers are set aside, single imputation methods obtain more spread out distributions of absolute deviation than these sophisticated methods (namely, K-NN, random forests, Amelia and MIPCA). Single imputation methods lead to larger deviations from the original series (or, in the best case, equivalent for the IPCA). Overall, considering outliers, these sophisticated methods fill missing values with spreads of approximately less than 1%, with 10% missing data and less than 2%, with 30% missing data. Without considering outliers, the distributions of the linear interpolation, LOCF and MSSA methods are indeed twice as large as those of K-NN, random forests, Amelia, IPCA and MIPCA.

However, this trend is not true for all imputation methods including a random component: the Brownian bridge method and the MICE algorithm indeed have distributions three to four times larger than the others, even without considering outliers. The absolute differences are particularly important for Brownian bridges because the more missing data there are in the sample, the more Gaussian parameters from the original data are harder to estimate. This method uses only the information from the series itself (one-dimensional method); and the more missing data it contains, the more

biased the parameters of the normal distribution are, since they are calculated from observed data only. This reinforces, in fact, the differences with the original series.

Large differences are also observed for the MICE algorithm because it is a longitudinal algorithm, meaning that the imputations are made from other observations (rows) of the same series. The simulated series, having a constant volatility, leads the algorithm to hardly find suitable potential donors, as all are suitable. Thus, algorithms with latitudinal functioning (such as K-NN) appear to be much more efficient in comparison.

The results presented in Figure 3.2-3 are those of a single missingness scenario (the first scenario contains 10% and 30% missing data); and to bring consistency to the results, other missingness scenarios are generated and imputed. Then, to compare all these scenarios, the following analyzed results are based on the average of each missingness scenario for a given proportion of missing data, and then on the average of all scenarios based on each missing data proportion.

Statistical moments

To observe, with more detail, the performance of these algorithms, the following figures focus on imputations made from the 100 scenarios containing 30% missing data. This proportion of missing data is used very frequently in this chapter. It may seem important, but its choice is not meaningless. Three main arguments explain this choice.

First, as mentioned in Chapter 1, after analyzing 946 empirical articles from the largest banking and financial journals that were published between 1995 and 1999, Kofman and Sharpe [133] showed that in practice, the proportion of missing data observed was on average 23.3% and could vary from 0.1% to 81.1%. Thus, the choice of analyzing scenarios containing 30% missing data would only be slightly more conservative than what they observed on average. Of course, their analysis is based on a large selection of articles, which do not necessarily deal with market data. It would not be surprising to observe high proportions of missing data for surveys, which would probably impact their average upwards. Nevertheless, the first idea here is to compare results on stressed scenarios.

However, the FRTB [12] regulation stipulates that a risk factor is modellable if it contains at least 24 real price observations over the last 12 months, with at least 4 real price observations every 90 days; or if it contains at least 100 real price observations over the last 12 months. Thus, as discussed in the section on FRTB [12], the Basel Committee considers risk factors that may contain up to 90% missing data to be modellable according to the first criterion. Moreover, the eligibility criteria for modellable risk factors have been relaxed compared to the first version of the FRTB [8], which required at least 24 real price observations per year with a maximum period of one month between two consecutive observations. This relaxation leads one to believe that the banks were faced with too many non-modellable risk factors, which led the Basel Committee to

revise their criterion, so as not to penalize them too much. Thus, the fact that missing data are found in a significant proportion of risk factor series must not be an exception, given that the banks declare that 32% of risk factors are non-modellable on average.

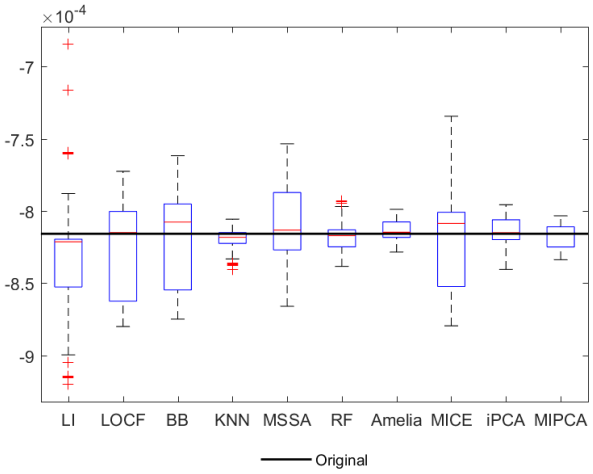
The third reason for choosing 30% missing data is that financial data are not simply stock or bond prices. In empirical studies, stock price series are very frequently used because they are accessible and of good quality. These data rarely contain more than 20% missing data (approximately 10-15% for the historical data applications in this PhD thesis). Nevertheless, this type of data represents only one part of financial data. As already mentioned, some data are missing because the market is illiquid for some special financial products, leading to the absence of transactions. Thus, it is not exceptional to find missing data in large proportions, which can be much higher than 30%, in single-name CDS or in repo rate historical data. These data are difficult to access and often confidential, which explains the absence of empirical studies on them, especially about missing data issues.

Finally, the last argument asks about the hidden part of the iceberg. Some series may appear to be complete, but in reality, they may have been completed by the data providers or even by the traders. Some illiquid data may be complete due to the intervention of traders and not due to the presence of transactions in the markets. Financial data are probably more fabricated than people think. Usually, when there are no transactions, traders estimate the price. Thus, when concluding a transaction, the trader has a kind of algorithm in mind to be able to set his price. The trader will probably look at the price that has been realized for other underlying assets to estimate the price of his transaction (latitudinal functioning as for the K-NN). In addition, this kind of procedure will necessarily have an impact on the correlation. Thus, the existence of these prices set by traders is undeniable, which is why it would be interesting to know what proportion of the data on the OTC markets comes from the front office without being based on transaction prices. This phenomenon clearly appears to be difficult to quantify.

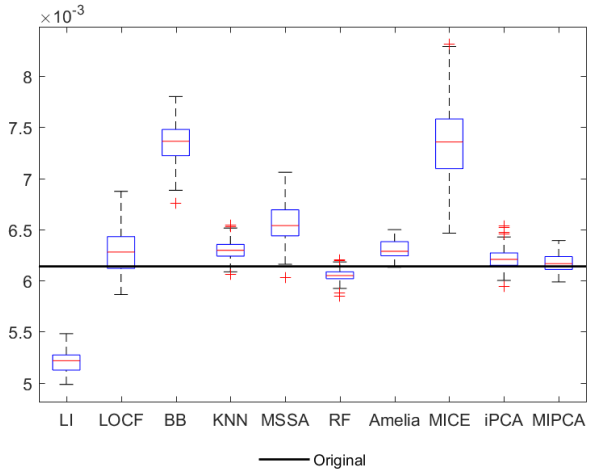
Considering these arguments, comparing completion methods from data histories containing 30% missing data is no longer appears to be excessive.

Thus, Figure 3.2-4, represents the distribution of the 100 means, standard deviations, skewnesses and kurtosises, for each of the methods used compared to those of the original series (represented by the black line).

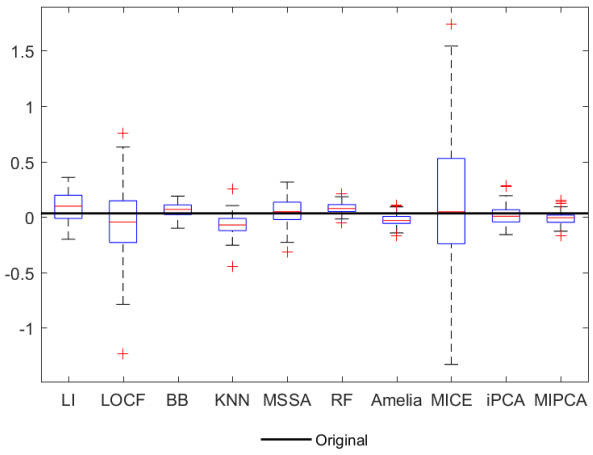
Fig. 3.2-4: Distribution of the first four statistical moments obtained from 100 scenarios with 30% MCAR data in the first column



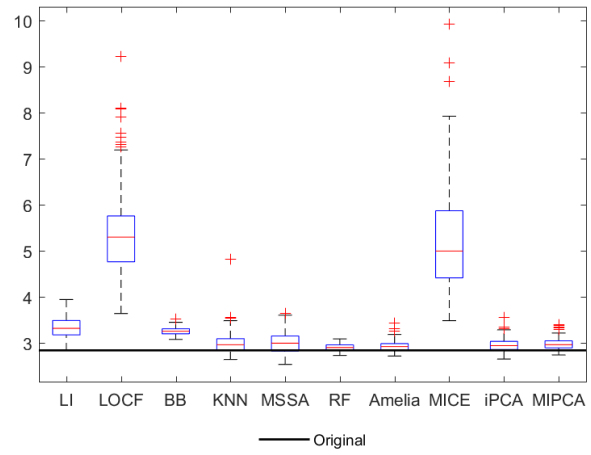
(a) Mean



(b) Standard Deviation



(c) Skewness



(d) Kurtosis

Among the 100 missingness scenarios analyzed, it appears that it is simpler for the methods to reproduce the mean and skewness than the standard deviation and kurtosis. The mean (Figure 3.2-4a) and skewness (Figure 3.2-4c) of the original series crosses the distributions of all methods. In other words, each of the methods results in at least one imputation sample, resulting in a mean and a skewness coefficient very close to that of the original series.

However, this is not the case for the standard deviation (Figure 3.2-4b) and kurtosis (Figure 3.2-4d). Some methods give values far from those of the original series. This is

notably the case for Brownian bridges and the MICE algorithm, which tend to overestimate the standard deviation and kurtosis. These two methods already gave the largest distributions in Figure 3.2-3; hence, it is not surprising that this affects the statistical moments. The Brownian bridge is calibrated from the observed data, distorting the original volatility. The MICE algorithm has serious difficulties finding credible donors because the series follows the same trend. Thus, its potential donors can be any returns from the column, which is why its kurtosis tends to be greater than 3.

On the other hand, linear interpolation is the only method that underestimates the standard deviation. This is not surprising since the imputations made do not add any noise to the series and thus slightly overestimate the kurtosis for almost all scenarios. Moreover, although the LOCF method preserves the volatility of the series (even if the distribution is one of the most spread out), it tends to strongly overestimate the kurtosis due to the 0-returns imputed.

Imputation methods that incorporate a random component, and therefore whose results are averaged, obtain smaller distributions than others, except for the Brownian bridge and MICE methods. The K-NN, random forests, Amelia and MIPCA methods are able to preserve the first four statistical moments of the original series, or at least not to deviate too far from them. Moreover, the Amelia algorithm seems to be much more efficient than MICE and MSSA while obtaining results close to those of random forests, such as Bauer, Angelini and Denev [27] in their paper.

In addition, the IPCA method also appears to be an acceptable solution, but its improved version (MIPCA) allows the results to be further stabilized by presenting tighter distributions.

Establishing a ranking of these methods based on these results is not obvious, but the methods of random forests, MIPCA, Amelia and K-NN seem to be at the top of the list. Conversely, the MICE method and Brownian bridge appear to perform the worst. Referring back to the preliminary results previously observed, where the random forests appeared to be more efficient than Amelia, here this conclusion is not obvious. Both methods obtain results close to the original series, but Amelia appears to be more conservative than the random forests, as it slightly overestimates the volatility, while the random forests underestimate it. This gives the advantage to Amelia.

Nevertheless, these results concern samples with 30% missing data and may be quite different for other proportions of missing data. The variability of the statistical moments for each proportion of missing data can be analyzed through the standard deviation of the moments obtained for each of the missingness scenarios. These results are presented in Appendix A.2 and provide an idea of the variability of these moments from one scenario to another and for a given proportion of missing data. These tables reveal the high variability of the MICE algorithm compared to other completion methods. This algorithm obtains among the largest standard deviations for each of the four statistical

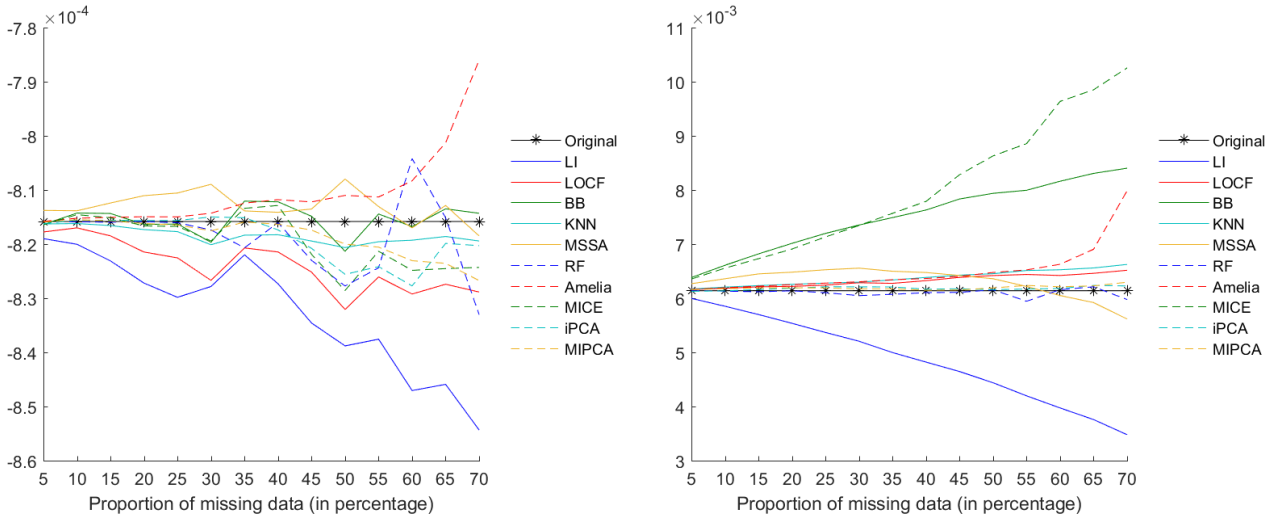
moments regardless of the proportion of missing data. On the other hand, the random forests appear to be the most stable by systematically presenting among the lowest standard deviations.

Thus, these results and those obtained for other proportions of missing data should be averaged to compare the algorithms with each other with respect to the proportion of missingness. These tables in Appendix A.2 give an idea of how far each algorithm deviates from their average.

For this purpose, the same calculation was performed for several proportions of missing data before being averaged to obtain the average of the first four statistical moments presented in Figure 3.2-5. More precisely, they are the averaged mean (Figure 3.2-5a), standard deviation (Figure 3.2-5b), skewness (Figure 3.2-5c) and Kustosis (Figure 3.2-5d) of returns from imputed data based on 100 missingness scenarios (for a given missingness proportion).

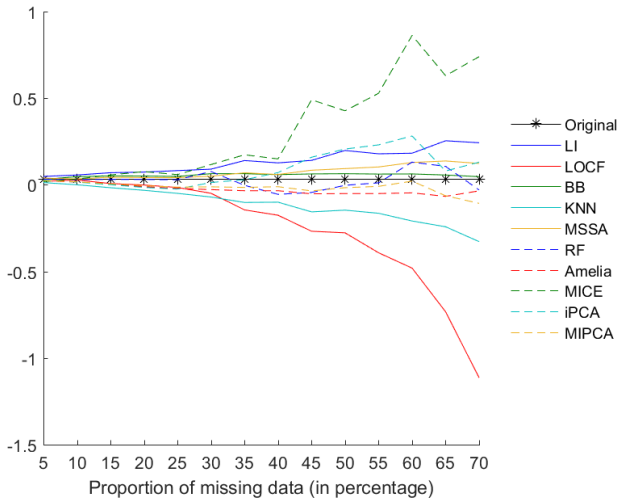
Therefore, Figure 3.2-5 represents the average results of the first four statistical moments obtained for the returns of the imputed series and can be compared with those of the completely observed original series (represented by the black starred line).

Fig. 3.2-5: Average of the first four statistical moments of the returns of the imputed data based on a matrix containing MCAR data (only in the first column) according to the missingness probability

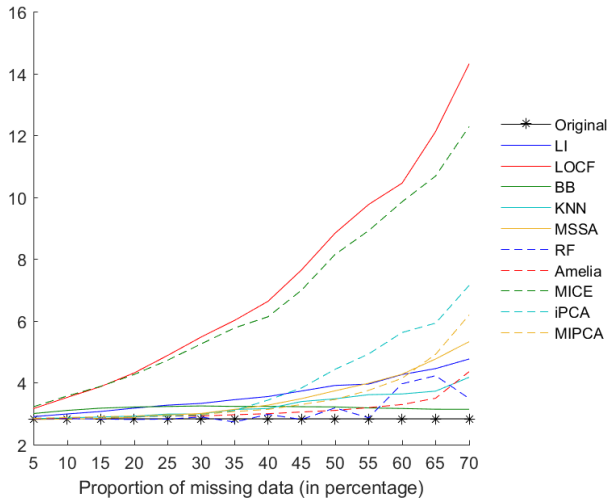


(a) Mean

(b) Standard deviation



(c) Skewness



(d) Kurtosis

In general, as the proportion of missing data increases, the number of imputation complications affecting models also increases. This is not surprising, as shown by Figure 3.2-5, where the greater the proportion of missing data, the greater the deviation of the statistical moments obtained from imputed data from those of the original series. On the other hand, some methods perform better than others, and this is notably what is presented below.

Considering the scale (at 10^{-4}), the methods manage to keep the average of the original series, with differences below a basis point (see Figure 3.2-5a). On the other hand, many methods tend to significantly impact the volatility of the series as the proportion of missing data increases (see Figure 3.2-5b). Completion methods also have an effect on the skewness and kurtosis;: as the proportion of missing data increases, the distribution is less symmetrical (see Figure 3.2-5c) and tends to be leptokurtic (see Figure 3.2-5d).

One of the first observations concerns the usual methods, such as linear interpolation or the LOCF method. These methodologies, usually applied by default in bank systems, lead to a bad estimation of at least two of the four statistical moments presented in Figure 3.2-5 as the proportion of missing data increases.

First, linear interpolation gives some of the worst results in terms of the mean and standard error. As the proportion of missing data increases, the mean and standard deviation decrease, which is clearly because this method tends to smooth returns without injecting additional noise into the series. On the other hand, the results of this usual method, in terms of skewness and kurtosis, appear good relative to other methodologies, although they tend to increase slightly with a high proportion of missing data.

The LOCF method is among the worst performers in terms of mean, skewness and kurtosis. Even if this method obtains better results than those of linear interpolation in

terms of the mean, it is still among the worst performers compared to the other methods. Furthermore, the LOCF is the method that has the greatest impact on skewness and kurtosis. This was expected, given that the missing returns are replaced by 0 followed by a correction return (the one calculated from the next available observation), which can be very negative or very positive if the difference between the last observed value and the next one is very large. Thus, the more missing data there are, the greater the probability of having successive missing data. This implies a higher proportion of zero returns, hence a leptokurtic distribution (kurtosis higher than 3) and a higher probability of having extreme values in the series, leading to an asymmetric distribution. A negative skewness was expected for this method, given that the trend of the simulated series was downward (see Figure 3.1-1), which implies strongly negative correction returns.

The Brownian bridge methodology does not totally distort the distribution; in contrast, this method correctly preserves the original symmetry and flatness of the series, which was expected because of the Gaussian assumption. On the other hand, it is less effective in preserving the level of volatility of the series. This is because Brownian Bridge is a one-dimensional methodology that uses only the available data of the series, as said previously. Thus, the more missing data the series contains, the greater the difference of the standard deviation of the observed values from that of the complete original series, leading to a Brownian motion generated with the wrong parameters.

Among the worst performing algorithms, it is also interesting to highlight the results obtained by the MICE algorithm. Imputations made by this method tend, in fact, to increase the volatility of the series but also to distort its distribution by making it asymmetric and leptokurtic. Thus, the average results obtained for these first four moments show that this relatively sophisticated method is not suitable to impute missing data from a Gaussian sample. Then, Figure 3.2-4 shows the poor performance of the MICE algorithm, and in view of these latest results, it gets much worse with a high proportion of missing data.

As mentioned previously, this algorithm is also very unstable from one missingness scenario to another. Appendix A.2 shows that this method has among the highest standard deviations for the four statistical moments for all missingness proportions. This method obtains scenarios where the mean, variance, skewness and kurtosis are 7.0710^{-5}, 2.1710^{-3}, 1.86 and 3.57 around their mean, respectively. This method leads to a mean of approximately 0.539 and 3.63 around their mean, which is huge considering the scales of Figure 3.2-5.

On the other hand, it is also important to note the most efficient algorithm according to these statistical moments. Three models appear to be able to impute the missing data of the series in a particularly satisfactory way by preserving their distribution, namely, the MIPCA, random forests and Amelia algorithms. These three methods obtain slightly different results in terms of the averaged mean but nevertheless satisfactory

(given the scale at 10^{-4}) and very good results in terms of standard deviation, skewness and kurtosis.

Among these methods, the MIPCA algorithm provides good imputation that preserves the mean, standard deviation and skewness. On the other hand, this method is not the best to reproduce the true kurtosis beyond a certain proportion of missing data. The greater the missingness proportion is, the greater the tendency of the distribution to be leptokurtic. Moreover, the IPCA version (thus without multiple imputation and based on the nonregularized model) also provides a good mean and standard deviation, but it is less efficient with respect to the asymmetry coefficient and especially the kurtosis of the distribution. The performance of this method is also negatively impacted by the proportion of missing data. This is partly because the number of main components used, determined by the function *estim_ncpPCA()* and presented in Appendix A.3, is constant for IPCA (equal to 4 on average) for any missingness proportion, whereas that for MIPCA gradually decreases (from 4 to 3 on average).

The random forests method is also one of the best performers, even though it is the method with the greatest variation in terms of averaged mean from one missingness proportion to another. For the other statistical moments, the algorithm gives very good results and even better than the performances of MIPCA regarding kurtosis. Moreover, Appendix A.2 reveals that random forests is one of the most stable methods (especially when the proportion of missing data are lower than 50%). This algorithm, indeed, obtains one of the lowest standard deviations, which shows that from one missingness scenario to another, the results obtained for each of the statistical moments are never too far from the results presented in Figure 3.2-5.

Finally, the Amelia algorithm is also one of the best performing models, even if the quality of imputation seems to be negatively impacted when the proportion of missing data exceeds 55% of the series. The averaged mean, standard deviation and kurtosis have, in fact, an increasing tendency beyond this threshold. The Amelia algorithm has, indeed, some difficulties converging correctly when the proportion of missing data becomes too large. At the end of this section, the average computation time of each algorithm is represented (Figure 3.2-14), and it is possible to notice that the computation time of the Amelia algorithm explodes when the proportion of missing data is very high, which reflects its difficulties of convergence. On the other hand, given the Gaussian sample, the Amelia model was expected to perform best. In reality, random forests are as good as Amelia or even better when the proportion of missing data becomes significant. Thus, this confirms the previously observed trend where random forests appear to predict Gaussian data better than an improved EM algorithm. It would not have been surprising if random forests were better than Amelia at capturing nonlinear relationships in a sample, but on this sample with zero mean and constant volatility, this is unexpected.

The MSSA algorithm, used by the Bloomberg data provider as the imputation

method, and the K-NN algorithm also lead to reasonable results. The MSSA algorithm allows imputations to reproduce the distribution properly and without too much impact on its volatility. On the other hand, it is far from being the best option, given the results of the other algorithms. The K-NN method also gives good results in terms of the mean, standard deviation and kurtosis, but to a lesser extent regarding to the skewness. Moreover, looking at Appendix A.2, the K-NN method is also one of the most stable, especially for the estimation of mean, standard deviation and skewness when the proportion of missing data is high.

One-dimensional imputation methods (linear interpolation, LOCF and Brownian bridge) are probably the simplest and least constraining methods to understand and implement, but they are also the least likely to preserve the original distribution. Thus, the results discussed here suggest that these methods are likely to provide suboptimal results in future trials as well.

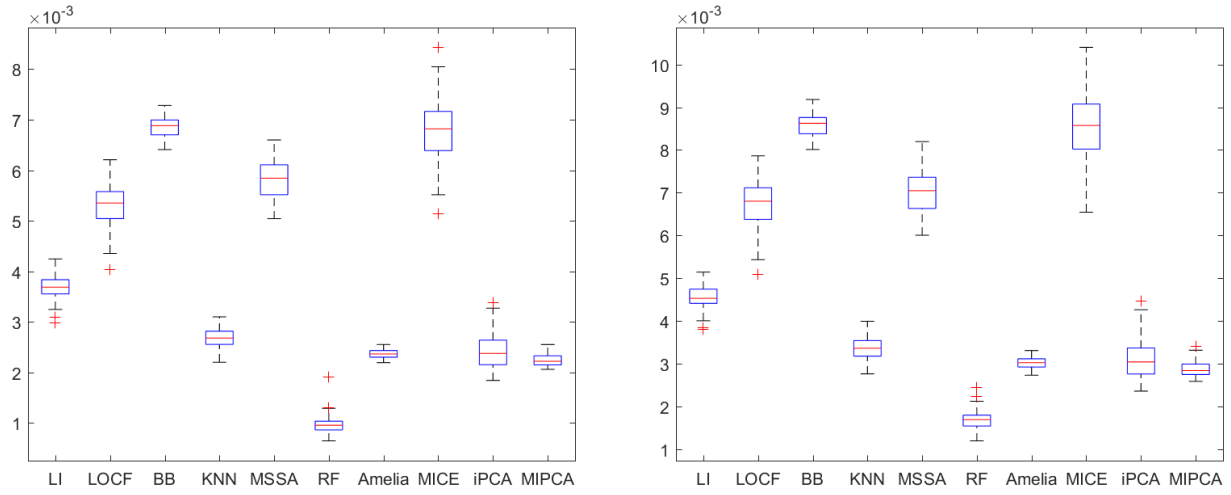
On the other hand, more complex multiple imputation methods (such as MIPCA, Amelia and random forests) are among the most efficient. Nevertheless, the MICE algorithm, combining MCMC with PMM, gives very disappointing results and is likely to remain unsatisfactory in the future.

Proximity metrics

Once the distribution of the imputed data has been observed, it is interesting to analyze the proximity between the original series and the imputed series. Thus, the two proximity measures presented earlier are used here: the MAE and the RMSE.

Figure 3.2-6a and Figure 3.2-6b represent the distributions of the 100 MAE and RMSE, respectively, between the returns of the original series and those of the imputed series obtained from the 100 scenarios containing 30% missing data.

Fig. 3.2-6: Distribution of the MAE and RMSE computed from the 100 scenarios containing 30% MCAR data in the first column



(a) MAE *(b)* RMSE

The first remark is that Figure 3.2-6a and Figure 3.2-6b look very similar. Even if the scales of the two figures are slightly different, each of the distributions associated with the MAE seems to be comparable to those of the RMSE, which means that all the models react in approximately the same way, not revealing too large differences with the original series.

For a missingness proportion of 30%, the random forests method minimizes the two proximity measures for almost all scenarios and thus fits as close as possible to the original series. The results of this model are very stable, as has already been the case previously; the MAE and RMSE from one scenario to another are very close, which confirms that Breiman's method [42] is very stable and efficient here. Moreover, these results clearly show the superiority of random forests over other imputation methods when the missingness proportion is 30%. The results are in agreement with those of Stekhoven and Bülhmann [194] or those of Jamal [180], who concluded that random forests were more efficient, in particular on proximity measures, than the MICE algorithm or the K-NN algorithm.

The Amelia and MIPCA algorithms again give good and comparable results; the MAE and RMSE distributions obtained are very close for these two methods and stable from one scenario to another (both distributions are very tight). As previously mentioned, good performances were expected for the Amelia algorithm since it relies on a Gaussian framework (EM algorithm) for data imputation. Moreover, the MIPCA algorithm gives results comparable to those of Amelia, which is not surprising given that PCAs are very frequently used on financial data (and therefore Gaussian). The

IPCA method obtains almost the same median as MIPCA; on the other hand, it gives more variable results among the missingness scenarios due to the absence of multiple imputation. Nevertheless, it is the most efficient simple imputation method here.

Finally, the K-NN method also leads to a MAE and RMSE very close to those of Amelia and MIPCA, although slightly higher and with a slightly wider distribution. The fact that the sample is relatively highly correlated helps this method work well. Thus, when it is possible to construct a highly correlated sample, the K-NN method is just as effective as Amelia or MIPCA when the missingness proportion is 30%. Thus, among the sophisticated methods, random forests are the most efficient in terms of proximity measures, followed very closely by Amelia, MIPCA and K-NN.

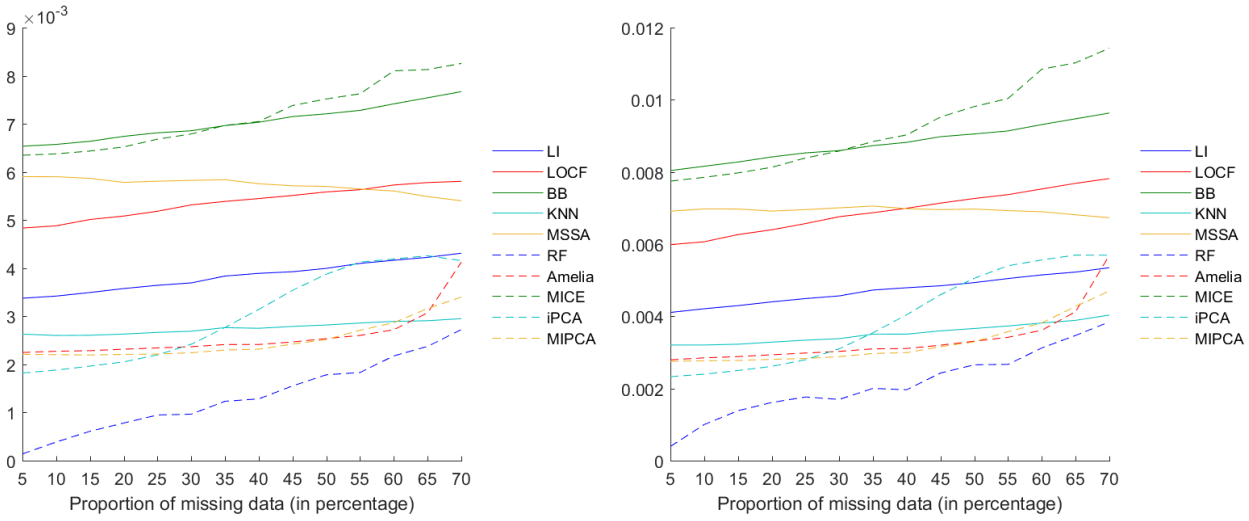
Concerning the least performing imputation methods, regarding MAE and RMSE, the MICE algorithm leads to much higher proximity measures and is very variable from one missingness scenario to another, considering the distributions. In fact, these distributions are the most spread out in Figure 3.2-6 and have the highest proximity measures. This is also totally consistent with the previous results.

Even if the Brownian bridge obtains proximity measures that are relatively stable from one scenario to another, these measures are also among the largest ones (close to MICE proximity measures). Moreover, the fact that the proximity measures are calculated from each imputed sample (see Figure 3.1-2 in Section 3.1.4) and not on the average of the imputed samples, gives results less satisfactory than a simple linear interpolation. If the 100 samples imputed by Brownian bridge had been averaged before applying the proximity measures, then results close to the linear interpolation would have been expected. Here, the fact that the proximity measures are averages of MAE and RMSE among the 100 imputed samples (for a single missingness scenario) considerably degrades the performance of this method. This shows that, independent of the method, the choice of implementation can significantly affect the quality of the imputation.

Thus, it is preferable to impute missing data using the usual methods, such as linear interpolation or LOCF, rather than using the MICE algorithm or the Brownian bridge. Moreover, these two usual methods of completion also seem to be a better option than using the MSSA algorithm. The latter being one of the bad performers.

To observe the robustness of the models in terms of proximity measures, it is therefore appropriate to average these results, as well as those for other levels of missingness. Thus, Figure 3.2-7 represents the MAE and RMSE between the returns of the original series and the returns of the imputed series, according to the missingness proportion. Moreover, the variability of each method, according to the missingness proportion, is available in Appendix A.4.

Fig. 3.2-7: Average MAE and RMSE from matrices containing MCAR data in the first column, according to the missingness probability



(a) MAE *(b)* RMSE

The first observation here concerns, as previously described, the similarity between the two graphs, but considering their scales, RMSE is approximately 20% to 40% higher than MAE for most algorithms. This means that the more missing data there are, the more likely it is that the imputation methods will impute values far from the original series but in a relatively moderate way. In addition, as the proportion of missing data increases, the RMSE tends to move further away from the MAE, which means that the imputation contains a greater number of large differences compared to the original series when the proportion of missing data is high.

Comparing the methods for all missingness proportions, the best performer is the random forests algorithm, followed by MIPCA, IPCA, Amelia and K-NN. The trend observed for 30% missing data is almost the same for any other missingness proportion. The IPCA algorithm achieves even better proximity measures than MIPCA when the proportion of missing data is less than 25%; however, its imputation quality deteriorates beyond this proportion to reach the same proximity measures as a linear interpolation, whereas MIPCA obtains almost stable proximity measures. This is because the IPCA used almost the same number of principal components (4 on average), whereas the MIPCA tended to adjust it (from 4 to 3 on average), as the missingness proportion increased (see Appendix A.3).

On the other hand, even with a low proportion of missing data, the Brownian bridge, MICE and MSSA algorithms are the least efficient. These three methods obtain even less satisfactory MAE and RMSE than the usual methods (linear interpolation and

LOCF). Moreover, the MICE algorithm is the one with the least stable results from one missingness scenario to another (see Appendix A.4).

Regardless of the proportion of data, the algorithm that minimizes both MAE and RMSE is the random forests algorithm. While the random forests obtain proximity measures comparable to the previously mentioned methods (MIPCA, IPCA, Amelia and K-NN) when the proportion of missing data becomes significant, they are well below when the proportion of missing data is low. When the proportion of missing data is 5%, the random forests obtain almost zero MAE and RMSE, revealing a quasi-perfect imputation, while Amelia, MIPCA, IPCA and K-NN obtain MAE and RMSE approximately 15 times and 7 times larger, respectively. This reveals the clear superiority of random forests in imputing highly correlated Gaussian data. In addition, random forests appear to be one of the most stable algorithms, from one missingness scenario to another (see Appendix A.4).

The second-best algorithm is IPCA when the proportion of missingness is lower than 25%. Beyond this threshold, the MIPCA, Amelia and K-NN lead to better proximity measures. The MIPCA algorithm has the advantage until the proportion of missing data exceeds 45%, and the Amelia algorithm then becomes the most efficient until this proportion reaches 65%. Beyond 65%, the K-NN algorithm becomes the best performing of the three and obtains results comparable to those of random forests. Moreover, the MIPCA, Amelia and K-NN algorithms provide among the most stable MAE and RMSE (see Appendix A.4). While MIPCA and Amelia lose stability as the proportion of missing data increases, the opposite is observed for K-NN. This is because the more data are missing, the more K-NN uses the same weights for weighted mean imputation, whereas MIPCA and Amelia have to estimate their parameter from less data, which strongly biases the imputation.

While all other methods display an increasing evolution of their MAE and RMSE as data are missing from the sample, the MSSA algorithm shows the opposite. This is counterintuitive because it means that the more data missing from the sample, the closer the algorithm moves to the original series. This would mean that the MSSA algorithm is able to more efficiently reconstruct an entire period rather than a few isolated data points of the sample, or at least that the deviations from the original series are on average larger when there are few missing data. Despite this, the MSSA results are less efficient than those of random forests, Amelia, MIPCA, or even a simple linear interpolation.

Covariance matrices comparison

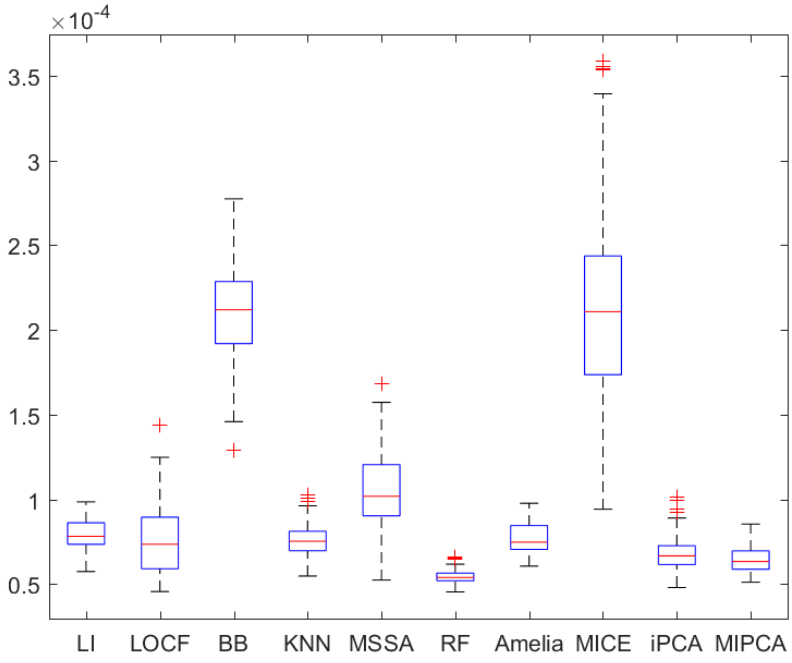
Here, a comparison is made based on the covariance matrices. The missing data are concentrated only in the first column of the data matrix, and only one part of the covariance matrix is impacted, that is, the one relating the first column to the others.

Nevertheless, it remains interesting to study whether completion methods are able to preserve the covariance matrix.

In the same way as previously conducted, the analysis is, first, done among the set of scenarios containing 30% missing data before analyzing the average results among different levels of missingness.

Thus, Figure 3.2-8 represents, for each of the methods, the differences between the covariance matrix (according to the Frobenius norm) from the original data and that resulting from the imputed data, where each of the 100 scenarios contains 30% MCAR data.

Fig. 3.2-8: Covariance matrix differences, according to the Frobenius norm, based matrices containing 30% MCAR data in the first column



Considering the previous results, it is not surprising that the covariance matrix obtained by MICE or Brownian bridge imputations is far from the original matrix. These methods are the least efficient in preserving the original distribution, especially the volatility (see Figure 3.2-5b) but also, the most distant from the original series, which leads to distancing in terms of the covariance matrix.

Moreover, the differences between the original covariance and the one obtained from the MICE method are highly variable from one scenario to another, ranging from simple to triple. This is the method that generates the most spread-out distribution of

covariance differences. The median of the deviations from the Brownian bridge method is close to that of MICE, but its distribution is less spread out (by half).

The MSSA method gives a covariance deviation distribution as wide as that of the Brownian bridge, but with values half as high. Nevertheless, this rather complex method obtains fewer good results than the usual classical methods (linear interpolation and LOCF).

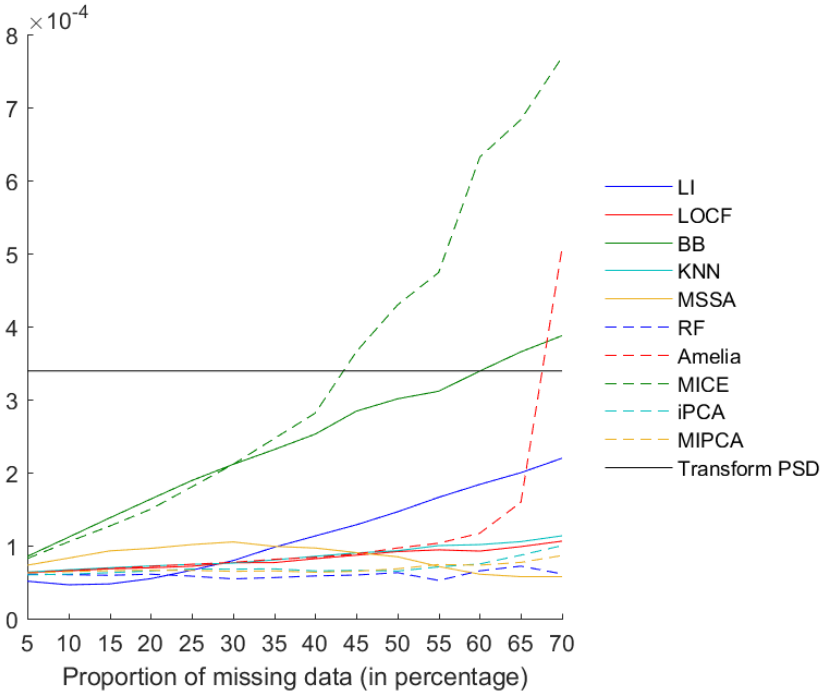
It is also unsurprising to see that, considering the previous results, the random forests method has the smallest differences and a tighter distribution.

Since the beginning of this comparative analysis, the MIPCA and Amelia methods give very similar results; however, here, the MIPCA algorithm offers a slight advantage with a distribution slightly closer to zero than Amelia. The IPCA, on the other hand, gives a distribution comparable to that of MIPCA but slightly more spread out, thus revealing slightly larger deviations in terms of covariance for this method.

Finally, linear interpolation gives a distribution of covariance differences similar to that of K-NN but also to that of Amelia. Then, for covariance matrix issues, linear interpolation challenges more complex and sophisticated algorithms when the missingness proportion is 30%.

The previous results are now averaged for different levels of missingness to obtain a global idea of the behavior of the algorithms according to the proportion of missing data. Thus, Figure 3.2-9 represents the average differences (among all the generated scenarios) between the covariance matrix of the returns of the original series and that of the imputed series, according to the Frobenius norm. These results are completed with the standard deviation of the covariance differences, presented in Appendix A.5.

Fig. 3.2-9: Average covariance matrix differences, according to the Frobenius norm, from matrices containing MCAR data in the first column, according to the missingness probability



$TranformPSD$, which is traditionally used to make covariance matrices quickly manipulable, obtains much higher results than imputation methods. This methodology distorts the covariance matrix even more than the majority of the other algorithms. Moreover, using a linear interpolation results in a covariance matrix closer to the original matrix than the one obtained by the Rousseeuw and Molenberghs [173] transformation for any missingness proportion.

Unsurprisingly, some models fail to preserve the covariance matrix when the proportion of missing data is too high, which is particularly the case for the MICE algorithm, as well as for the Brownian bridge. These two methods have already shown their failures in the previous graph. This is due to the longitudinal functioning of the MICE algorithm, which imputes missing data based on the closest available observations and not on the closest variables (unlike K-NN). Similarly, a Brownian bridge is a unidimensional algorithm that uses only the information from the series itself and therefore has no link with the rest of the sample. Moreover, these two methods are also the least stable in terms of covariance differences from one scenario to another and especially MICE (see Appendix A.5).

However, linear interpolation, which is also a one-dimensional method, obtains the

best results when the proportion of missing data is below 20%, but above this threshold, it tends to distort the covariance matrix, and then it becomes one of the worst performing models. Thus, linearly interpolating missing data, when the proportion of missing data is below 20%, seems to have little impact on the covariance matrix. Thus, when few MCAR data are present on a single Gaussian series, linear interpolation appears to be an excellent solution.

Finally, the results obtained for the other methods are relatively comparable, except for the Amelia algorithm, which, here again, has difficulty managing a large proportion of missing data (over 55%). The results of this method are also highly variable from one scenario to another when the missingness proportion is high (see Appendix A.5). Moreover, the method that obtains the smallest differences (and the most stable results according to Appendix A.5) is the random forests method, followed closely by MIPCA and IPCA, which both give similar results.

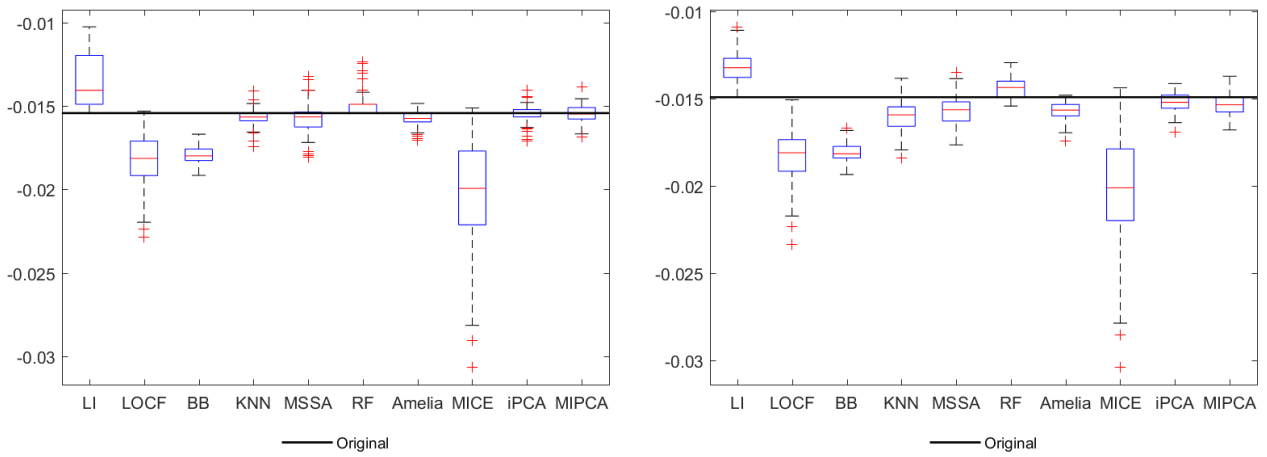
The MSSA algorithm provides stable results from one missingness proportion to another, with a slightly decreasing trend as the proportion of missing data increases. However, these results coincide with the previous results, pointing out that the higher the proportion of missingness, the closer the series imputed by MSSA is to the original series in terms of MAE and RMSE (see Figure 3.2-7). These results must of course be put into perspective, given that they were obtained from a single sample of data, and that they may vary for another sample of data.

Value-at-risk and expected shortfall

Each of the completion methods imputes missing data by more or less distorting the original distribution, deviating more or less from the original series, with impacts on the covariance matrices; therefore, this is not without consequences on risk measures. As already mentioned in Section 1.3, the imputation methods are authorized by the regulator, to allow banks to model their risk factor to calculate VaR and ES.

Thus, the results presented in Figure 3.2-10a and Figure 3.2-10b show that for each of the methods, the 1-day VaR and ES with confidence levels of 99% and 97.5%, were calculated from the imputed samples, for the 100 scenarios with a missingness proportion of 30% to be compared with those calculated from the original series (black line).

Fig. 3.2-10: Distribution of the 1-day risk measures computed from the 100 scenarios containing 30% MCAR data in the first column



(a) 1-day VaR at 99%

(b) 1-day ES at 97.5%

Since the beginning of this section, the MICE and Brownian bridge methods distort the distribution the most compared to the original series. These two methods are unsatisfactory for many criteria, and this is also the case here. First, the MICE method gives the most unstable VaR and ES from one scenario to another. This is not the case for the Brownian bridge method, which gives almost stable risk measures, but each of these imputations give rise to VaR and ES below their true level. In fact, both methods tend to have VaR and ES below their true level in almost all scenarios. This coincides with the fact that these methods tend to asymmetrize the distribution with positive skewness (see Figure 3.2-5c). Thus, these two methods would lead banks to pay much higher capital charges than they should.

The usual imputation methods are also among the least stable methods in terms of VaR and ES. Moreover, while the LOCF method tends to have a more conservative VaR and ES (which means below those of the original series), the linear interpolation method leads to much more optimistic VaR and ES (above those of the original series). These results were expected, given that the LOCF method can add more extreme returns in the series due to the correction returns, and that the linear interpolation method imputes without injecting any noise to the series.

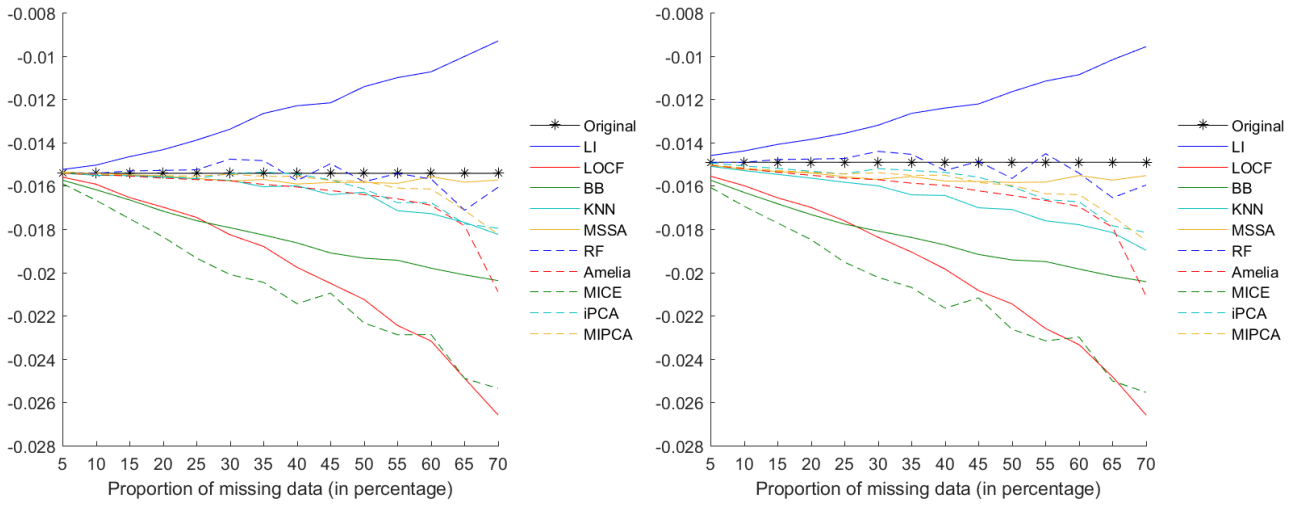
Some of the other methods obtain much more stable VaR and ES. The random forests method tends to slightly overestimate these risk measures, unlike the other completion methods. This could lead the regulator not to consider this algorithm as a good imputation technique, even though it has been satisfactory on the many criteria previously observed. It would be possible to integrate a penalty to the risk measures calculated by the random forests to make them valid for the regulator. Indeed, a

confidence interval could be calculated using the standard deviation obtained from one scenario to another, which would obtain the most conservative VaR and ES. Moreover, this methodology can be used in a more general way for all methods (and all analysis criteria) to ensure the validation of a methodology by the regulator.

The remaining methods tend to underestimate the levels of VaR and ES, which will be in accordance with the expectations of the regulators, especially if they include a penalty, as explained for the random forests. The IPCA and MIPCA algorithms give very similar results here. In addition, the Amelia algorithm provides comparable results and is one of the most stable for these two risk measures.

Now, the average results for each proportion of missing data are represented in Figure 3.2-11a and Figure 3.2-11b. As previously described, the variability of the risk measures is available in Appendix A.6 (first table).

Fig. 3.2-11: Average 1-day risk measures computed from matrices containing MCAR data in the first column, according to the missingness probability



(a) 1-day VaR at 99%

(b) 1-day ES at 97.5%

The trend previously observed in Figure 3.2-10 for 30% missing data is also observed in Figure 3.2-11 for all other missingness proportions. The linear interpolation method tends to result in higher risk measures than those from the original series. In contrast, other methods tend to be more conservative than the original risk measures.

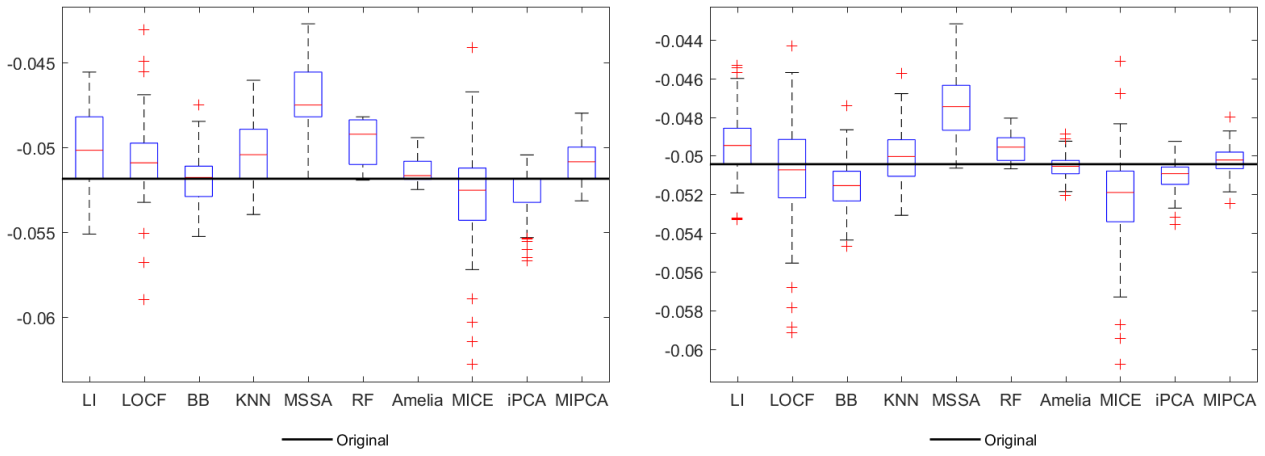
Linear interpolation is therefore one of the methods that cannot faithfully reproduce the risk measures of the original series. This is also the case for the LOCF, MICE and Brownian bridge methods, which become increasingly far apart as missing data appear

in the series. As in the previous analysis, Appendix A.6 shows a high variability of the MICE algorithm but also for the LOCF. Moreover, these were the two methods with the highest kurtosis in Figure 3.2-5d.

Moreover, the Amelia method, previously presented as one of the best methods for estimating VaR and ES with 30% missing data, has risk measures that degrade very strongly when the proportion of missing data exceeds 60%. As already mentioned, this algorithm appears to have some difficulties converging correctly when the proportion of missing data is high. The variability is also high when the proportion of missing data is equal to 70% (see Appendix A.6). Moreover, it is possible to observe the same phenomenon for the MIPCA method but to a lesser extent. Finally, the method that offers the closest average results from the original risk measures with a 1-day horizon, regardless of the proportion of missing data, is the MSSA algorithm. Its average evolution remains close to the VaR and ES of the original series, compared to the other methods, for any missingness proportion. The risk measures of this method do not appear to be impacted by the proportion of missing data, and their estimations are very stable from one missingness proportion to another, which is not the case for all methods.

As presented in Section 1.3.4, the FRTB regulation aims to replace the 1-day VaR at 99%, by a 10-day ES with a confidence level of 97.5%. Thus, Figure 3.2-12a and Figure 3.2-12b represent the 10-day VaR and ES at a 10-day horizon with confidence levels of 99% and 97.5%, respectively, for all scenarios associated with 30% missing data.

Fig. 3.2-12: Distribution of the 10-day risk measures computed from the 100 scenarios containing 30% MCAR data in the first column



(a) 10-day VaR at 99%

(b) 10-day ES at 97.5%

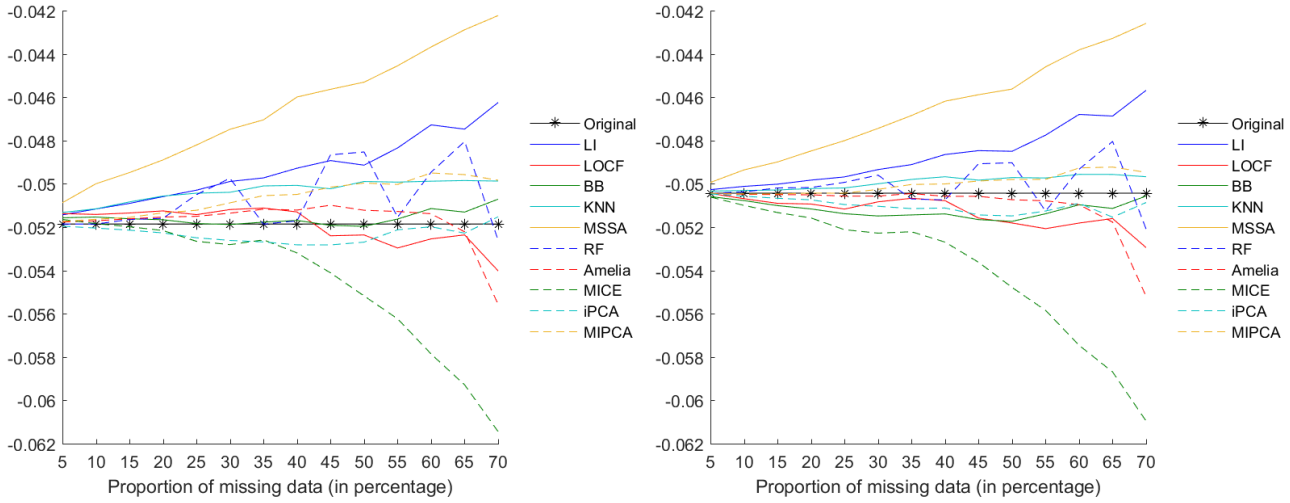
Looking from a 1-day VaR at 99% to a 10-day ES at 97.5% radically changes the results. While almost all of the methods previously tended to give more conservative risk measures than those in the original series, here, the results have completely changed. The only method that would tend to be more conservative (that is, below the level of true risk measures) is the MICE algorithm; this is true for a majority of scenarios but not for all. Nevertheless, this method leads to imputation with the most conservative risk measures.

Overall, the results obtained by the completion methods revolve around the true risk measures but with more spread-out distributions than when the horizon was set at 1 day (see Figure 3.2-10). The interval of these risk measures has indeed increased, thus leading the methods to a loss of accuracy.

The largest distribution is, as usual, that of the MICE algorithm, followed by that of LOCF, and this is the case for all other missingness proportions, as shown in Appendix A.6. The other distributions are relatively comparable from one method to another. The methods that would give the most stable risk measures among all the scenarios would be the random forests, MIPCA and Amelia algorithms.

Figure 3.2-13a and Figure 3.2-13b represent the averaged evolution of 10-day VaR and ES for confidence levels of 99% and 97.5%, respectively, for each missingness proportion. In addition, the standard deviation of each risk measure is available in Appendix A.6 (second table).

Fig. 3.2-13: Average 10-day risk measures computed from a matrix containing MCAR data in the first column, according to the missingness probability



(a) 10-day VaR at 99%

(b) 10-day ES at 97.5%

As before (see Figure 3.2-11), the MICE algorithm and the linear interpolation are among the least suitable methods to replicate the risk measures from the original series.

On the other hand, the MSSA algorithm, which appears to be satisfactory for a horizon set at 1 day, gives risk measures that are significantly above their expected level here. Thus, the MSSA method appears to correctly reproduce risk measures at the 1-day horizon but is totally unable to do so when this horizon is extended to 10 day. Moreover, the variability of the risk measure, from one scenario to another, is part of the highest when the missingness proportion is low (see Appendix A.6).

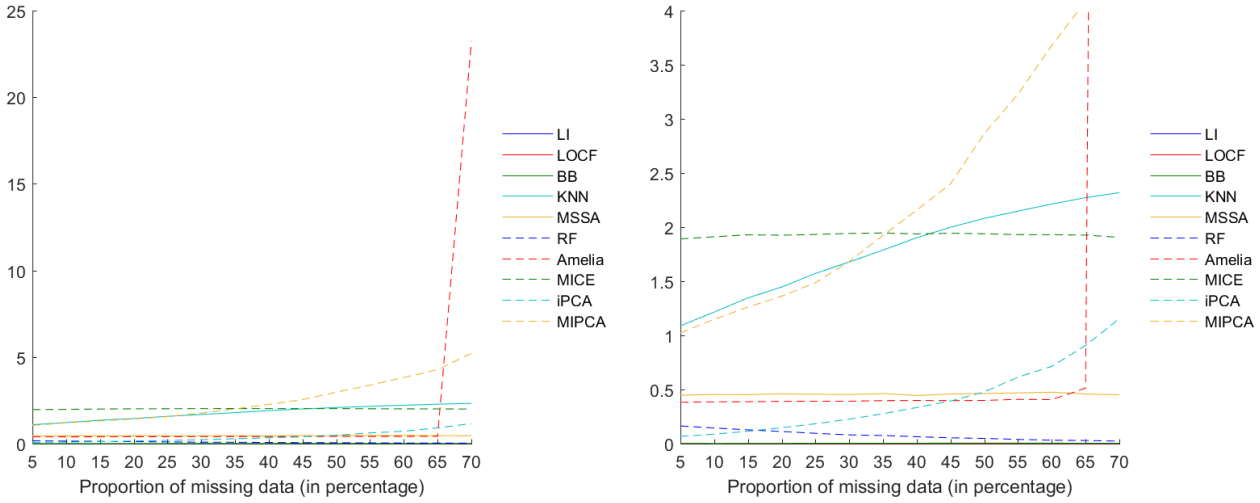
The most satisfactory method here, when the proportion of missing data is below 60%, is undoubtedly the Amelia algorithm, although above this threshold the algorithm is no longer able to correctly impute the data, which has an impact on the risk measure estimation. Again, the fact that Amelia's imputation is of poor quality for a proportion of 70% may be due to the sample used here. Nevertheless, this anomaly reveals that the algorithm may not handle a high proportion of missing data. In addition, it obtains very stable results from one missingness scenario to another (see Appendix A.6).

The IPCA and Brownian bridge methods also provide risk measures close to the original series. On the other hand, the random forests algorithm seems to give totally random result, but this is probably due to the sample used. The same calculation should be done for several simulated sample to make these results more robust.

Computation time

Finally, the last element of comparison here concerns the computing time. This variable is not negligible given that imputation techniques can be applied to very large or very numerous samples, and that computation time can become a real issue for banks. Figure 3.2-14 represents the evolution of the average computation time of each algorithm (in seconds) to impute a scenario according to the proportion of missing data to complete. As a reminder, all the algorithms were done on the same computer and without parallelization of the calculations.

Fig. 3.2-14: Average computation time of the imputation of MCAR data in the first column (with two different scales) according to the missingness probability



The first obvious observation is the explosion in computing time for the Amelia algorithm when the proportion of missing data exceeds 65% (left figure). For a sample containing 70% missing data, the algorithm takes an average of 24 seconds to impute one sample, while for a lower proportion, the algorithm takes less than half a second to find a solution. This highlights the complications that the EM algorithm may have to converge, which clearly explains the poor performance of the algorithm when the proportion of missing data becomes too high.

The MIPCA algorithm also has a computation time that increases as the proportion of missing data is large, but in a much more progressive way than for Amelia. It takes approximately 1 second to complete 5% of the sample, compared to approximately 5 seconds to complete 70%.

The K-NN algorithm needs twice as much computation time to impute 70% of missing data as it does to compute 5% of missing data. The slope of this evolution is comparable to that of the IPCA. Moreover, the IPCA algorithm actually has a much smaller slope than the MIPCA algorithm, which is certainly due to the bootstrap that the latter integrates.

Finally, the MICE algorithm is one of the slowest algorithms, but its computation time is relatively stable in relation to the missingness proportion managed. The MSSA algorithm is also stable in terms of computation time but is four times faster than the MICE algorithm.

Random forests have a computation time that decreases as the sample is emptied. This is because each tree is based on fewer observations and is therefore faster to create.

Finally, the usual methods, as well as the Brownian bridge, have an almost instantaneous computation time.

Thus, these first results provide an idea of the performance of the algorithms in an easy case, that is, a Gaussian sample containing MCAR data in only one column of the matrix. The usual methods are obviously inefficient on many criteria, but they are not always the worst performers. It appears clear that the MICE algorithm, in addition to being totally unstable from one missingness scenario to another, tends to deviate strongly from the original series for all the mentioned criteria. Thus, this algorithm, which operates longitudinally, appears to be inefficient on this sample. Moreover, the Brownian bridge tends to increase the volatility of the series, which has an impact on the other analyzed criteria. Finally, despite the ability of the MSSA algorithm to correctly replicate the 1-day risk measures, this algorithm remains mediocre for the other criteria, sometimes not even outperforming the linear interpolation. This method, based on a spectral analysis of the data, is close to the one used in signal processing, but intuition would suggest that it is particularly efficient on data integrating mixtures of periodicity and seasonality. The application of Gaussian data without regime variation leads the method to obtain suboptimal results compared to the usual methods. However, it is possible that this method is more relevant on other types of data. This method being the one used by Bloomberg, raises questions about the documentation on which the implementation is based. Dash and Zhang [65] may have deliberately omitted some details of the implementation, to not allow the potential client to reproduce this algorithm himself.

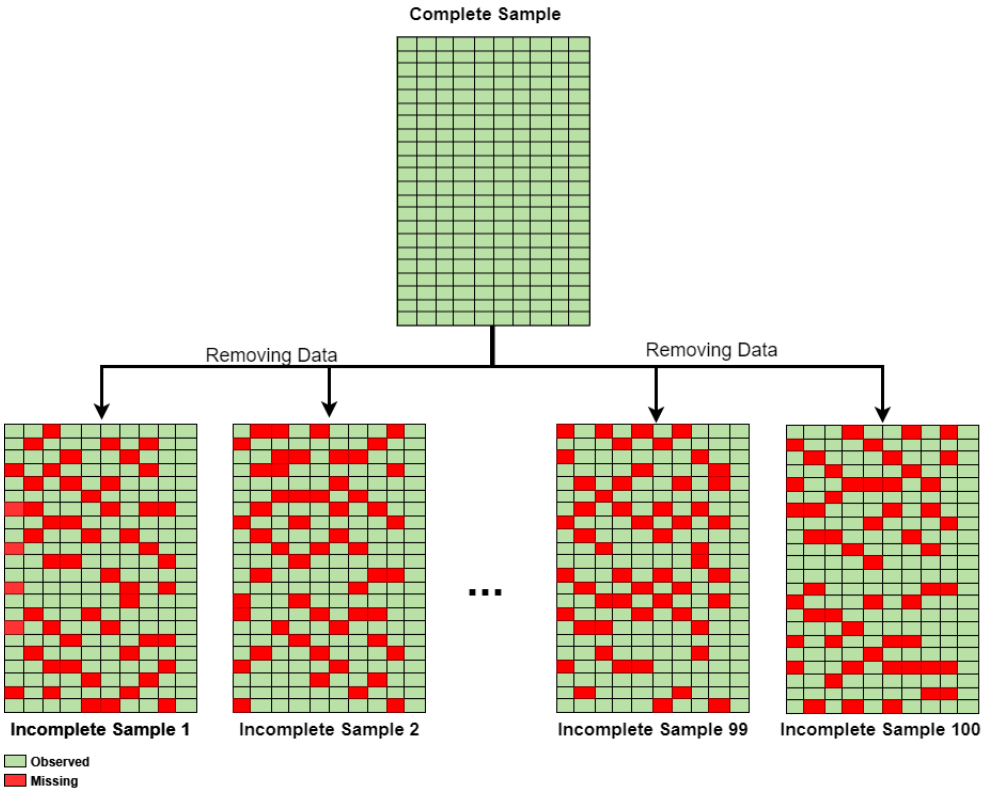
On the other hand, and fortunately, some other methodologies perform better than the usual methods. This is the case for random forests, Amelia, MIPCA IPCA and K-NN. Random forests perform well on many criteria and are very stable from one scenario to another, which can make them very credible to the regulator. The Amelia and MIPCA algorithms obtain very satisfactory results and are comparable to one another on many criteria of analysis, but the performances of the Amelia algorithm deteriorate very strongly when the proportion of missing data becomes too significant, leading to MIPCA being preferred. Moreover, MIPCA also appears to perform better than its simplified version, IPCA. Finally, despite its satisfactory results, the K-NN does not significantly outperform these other methods.

The fact that these methods work well is not very surprising; on the other hand, as has already been said, the fact that Amelia is not the most efficient method is not obvious. The Amelia algorithm is based on the assumption that the data are Gaussian and should have no problem in reproducing the missing data in this case. However, random forests allow better quality imputations. Some explanations and hints will be discussed at the end of this chapter to understand the magic behind this algorithm, often described as a “black box”.

3.2.2 Impact of MCAR data in the whole sample

The set of results presented previously in Section 3.2.1, compares the algorithms when the missing data are concentrated only on one series. This is, of course, a very special scenario because when data are missing for one series, it is highly probable that they are missing for the others too. Hence, it is important to see if the performance of the algorithms is strongly impacted by the presence of missing data for the whole sample studied. Thus, as this missingness mechanism includes a random component (by drawing randomly from a uniform law), the results of this section are based on 100 missingness scenarios of MCAR data in each column of the data matrix, as presented in Figure 3.2-15. The goal of this section is to see the degradation of imputation quality with respect to the results from Section 3.2.1, where the missingness was only in the first column.

Fig. 3.2-15: 100 missingness scenarios following a MCAR mechanism: randomly removing data in each column of the data matrix (except the last column)



In this second type of missingness pattern, the deletion of data concerns only the first 9 columns of the matrix to guarantee the correct application of the completion methods. This means that the last column is completely observed, so that there is

no observation (row) containing only missing data. Having the last column complete ensures that there is at least one observed value on each row.

The fact that data are missing only in the first column (as in the previous section) is, of course, a simplified pattern of missingness. In practice, if there are missing data in a sample, it is rarely in a single column but throughout the entire sample. Nevertheless, this brings up a practical problem; should the imputation methods be applied to the original data only (which may contain missing data) or should they be applied to already imputed data? If, for example, an expert has already imputed his databases, he can impute the missing data of a new series using other imputed series. Using data that have already been imputed to impute another series will necessarily have an impact on these new imputations. Moreover, this impact leads to less transparent imputation, which is also less easily replicable for the regulator.

In this PhD thesis, the missing data mechanism is often applied to the first column of the data matrix, which is useful to observe the performances of the algorithms without being too expensive in computation time. Nevertheless, it is interesting to see the impact of missing data on the whole matrix, especially how much it degrades the imputation quality.

Moreover, when the missing data are distributed over the whole sample, the question of the column imputation order arises for algorithms imputing column by column. One dimensional imputation methods, such as linear interpolation, LOCF, and Brownian bridges independently impute each column. On the other hand, some methods iteratively impute the missing data of a column using the available dataset of the other columns, without making any distinction between the originally observed data and the imputed data. This is notably the case for the K-NN, random forests and MICE algorithms. Thus, when the imputation method works iteratively, it is obvious that the imputation order of the columns may have an impact on the imputation quality. In the literature, these imputation methods often treat the columns in ascending order of missing data, but this methodology is not necessarily the most optimal one and the imputation order becomes an issue in the same way as the choice of any parameter of the chosen model.

As explained in the previous chapter (see Section 2.4.3), the K-NN method was implemented based on the results of Sahri, Yusof and Watada [179], who showed that imputing columns in ascending order of missing data improved the imputation accuracy and preserved the correlation between the columns. Furthermore, Stekhoven and Bühlmann's implementation of random forests [194] follows the same logic, and imputes columns in increasing order of missing data by default. Finally, in the case of MICE, the algorithm is implemented to process the columns naively, one after the other, in the reading direction (from left to right), but it is possible to specify an option to process them in ascending order of missing data (which was not used here).

Since the missing data in this section will be injected in equal proportions for all columns (which is explained in detail later), the order of imputation is not very important for this study. On the other hand, in the case of historical data, the order of imputation can represent a real issue for the imputation quality.

Thus, in the same way as before, data were removed from the sample in a completely random way but this time from (almost) the entire sample, and not just the first column of the data matrix. The sample is exactly the same as presented in Section 3.2.1, following a simulation process as presented in Section 3.1.1, and the data deletion process remains the same as before, that is, based on a uniform law. This time it is not only the first column of the data matrix that is affected by these deletions.

To allow all the algorithms to work properly, the missing data have been deleted among the first 9 columns of the data matrix (9 among 10). Thus, the same missing data proportion as before is removed in each of the first 9 columns. Thus, the proportions of missing returns associated with the proportions of missing data presented in Table 3.2-1 are the same as in the previous section but this time they involve all the columns of the matrix (except the last one), not only the first. Thus, in the most extreme scenarios, an algorithm applied to price returns will have to impute approximately 90% of the data (that are returns) of 9 columns out of 10.

The fact that the deletion concerns only the first 9 columns, and not the last one, ensures the presence of at least one available observation on each row (as in Figure 3.2-15). Moreover, this allows the number of cases where no observation is observable for all the series to be limited, at least for the small proportion of missing data, to have at least one row completely observed by all.

If the deletion was made for all columns of the data matrix, then the K-NN and MICE algorithms would not work at all, even for low missingness proportions, and should be removed from the comparative analysis.

Thus, the fact that the missing data are found over almost the entire sample makes it possible to compare the performances of all the algorithms, without making any concession on the objective of this procedure, that is, to confront the algorithms with missing data distributed over the whole (or almost the whole) sample. The results obtained, on samples where the missing data are distributed over 9 or 10 columns should, indeed, be relatively similar. Thus, the purpose of this section is to impute samples with missing data over the entire (or almost the entire) sample, to compare the results obtained with those of Section 3.2.1.

Despite these precautions, the K-NN method is not able to work properly when the proportion of missing data exceeds 15%. This is because, to work properly, it requires at least one completely observed row (i.e., without any missing data) to calculate the weights associated with each of the columns. However, the missing data were injected into almost the entire sample, and the higher the proportion of missing data, the lower

the probability of having complete rows. The way the data were removed from the sample, and considering that the K-NN uses return data, corresponds to just over five fully observable rows for 15% missing data, just over two rows for 20% missing data, and less than one for 25% missing data (the probability of having complete rows is given by $(1-(2p-p^2))^9$). Moreover, the chosen methodology integrates a bootstrap step, which makes the algorithm even less workable, given that each of the bootstrapped samples must contain at least two completely observable rows (to have at least one row of returns); otherwise, the algorithm stops. No steps were implemented to search for a usable bootstrapped sample. Thus, the results obtained for the K-NN method, with a proportion of missing data at 15%, are totally inconsistent with those obtained for 5% or 10%. This is because the weights are calculated from too few available observations. For this reason, these results are not presented afterward, so as not to distort the scales of the following graphs.

Moreover, the K-NN algorithm is not the only algorithm that is not calculable. It is also the case for Amelia, which is no longer computable after 30% missing data. The reasons for this are exactly the same as for K-NN, because it requires at least one completely observed line to make possible the initialization of the parameters after a listwise deletion. On the other hand, the algorithm integrates a step that allow to find a bootstrapped sample exploitable by the EM algorithm. Thus, Amelia manages to handle a slightly larger proportion of missing data than the K-NN. This is why the Amelia results in this section will be presented for proportions less than or equal to 30% missing data.

Finally, the MIPCA algorithm is also not computable for at least one of the scenarios containing 70% missing data, because there is too much missing data in the sample to be able to compute the eigenvalues necessary for the PCA imputation. This algorithm is indeed applied to the returns, which means that for 70% missing prices, less than 10% of the returns are observable in the first 9 columns. Thus, the results for this algorithm will be presented for all but the highest proportion, i.e., 70%.

As mentioned earlier, missing data on the whole matrix are not exceptional. Generally, data are missing for several series and not only for one of them, hence the interest of removing data from the whole matrix in this section. Thus, this missingness mechanism highlights a significant limitation; missing data distributed over the whole sample (even for a small proportion) can make the K-NN and Amelia algorithms unusable. In practice, a bank may use this type of algorithm in a data quality verification process before calculating risk measures or just P&L calculations. However, if more than 10% (30%) missing data are present in the total sample, then the use of K-NN (Amelia) as a completion method becomes noncalculable, automatically generating an error, and leading to an operational risk that can have heavy repercussions for the bank.

Thus, the completion methods can constitute a risk for the banks if they are incorrectly chosen or not adapted to the data to be imputed, but also if they become

noncalculable. In this PhD thesis, a certain number of parameterizations have been implemented for each of the methods to avoid missing any imputation after passage. For example, in the case of linear interpolation, an extrapolation was parameterized in the case of missing data at the extremities of the series. Despite these precautions, the competition methods may not be able to impute missing data, as is the case here with K-NN and Amelia. Moreover, the MCAR tests used in this chapter are not always calculable, which may refer to an operational risk in the same way as the completion methods.

The aim of this section is to impute all the missing data that are in the whole sample to compare the results with those of the previous section. Thus, the same analyses as before were done based on the imputed data from the first column: statistical measures, proximity measures and risk measures. On the other hand, the covariance matrix comparisons are based logically on the whole sample. No impact is expected on the unidimensional methods (linear interpolation, LOCF and Brownian bridge); on the other hand, the other methods could see their performance deteriorate. However, before imputing the missing data, the two MCAR tests previously used will be applied to samples containing missing data over almost the entire sample to see if they are able to detect them as MCAR.

Finally, each comparison tools are computed as defined in Figure 3.1-2, and the templates of all the graphs presented in this section have been presented and detailed in Section 3.1.4 (what they represent and how they were obtained).

MCAR tests

Thus, Little's test [142] and Jamshidian and Jalal's test [123] were applied to the same simulated data as in the previous section, but this time with missing data in almost all columns (except the last one). The results of these tests, applied to return matrices, are presented in Table 3.2-3. This table highlights the proportion of tests that do not reject the null hypothesis of data being MCAR with a 5% significance level.

This proportion was calculated among the calculable tests. With missing data present in almost the entire sample, Jamshidian and Jalal's test [123] may have more difficulty obtaining a result. Appendix B.1 presents the number of scenarios (among 100) that are correctly calculable. Little's test [142] is, of course, always calculable. On the other hand, Jamshidian and Jalal's test [123] was successfully conducted for all missingness scenarios only when the proportion of missing data was less than or equal to 10% or greater than or equal to 50%. For proportions of 15% and 45%, a large part of the tests are calculable but not all (97 and 80, respectively), while for proportions of 20% and 40%, only a few of them are calculable (6 and 4, respectively). Finally, when the proportion of missing data is between 25% and 35%, no scenario (among the 100)

obtains results for this test. This is because the test does not take into account missing data patterns that occur less than 7 times in the entire sample. In other words, all missing patterns appearing less than 7 times are removed from the sample. The test may then not be computable, if, once all these patterns are deleted, none remains (no pattern occurring more than 6 times). This is why Jamshidian and Jalal's test [123] may fail.

The simple fact that this test is not always calculable for samples containing missing data on the whole matrix is problematic. Missing data distributed over the whole data matrix are far from an exceptional phenomenon. However, the impossibility of calculating these MCAR tests can lead experts to make the wrong choice. Jamshidian and Jalal's test [123] may not be calculable as soon as 15% of data are missing, yet higher proportions are frequently observed in practice. As presented in Section 1.2.3, after analyzing 946 published empirical studies, Kofman and Sharpe [133] concluded that the average proportion of missing data in finance was 23.3%, and could vary from 0.1% to 81.1%. Then, the results from Jamshidian and Jalal's test [123] are rather problematic, because it would not be applicable for a large part of financial samples. This point gives an advantage to Little's test [142].

Now that the number of calculable tests (among the 100) is known, the probability of not rejecting the null hypothesis when this null hypothesis is true can be analyzed and is presented in Table 3.2-3.

Tab. 3.2-3: Confidence level (probability of not rejecting H_0 when H_0 is true) for both MCAR tests applied to price return matrices containing MCAR on almost the whole data matrix, for a 5% significance level

	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Little's test	95%	96%	96%	100%	100%	100%	100%	100%	100%	100%	100%	97%	98%	100%
J&J's test	89%	90%	88%	83%	-	-	-	100%	91%	93%	99%	99%	97%	97%

Little's test [142] generally concludes that the data are MCAR. The null hypothesis is, in fact, npt rejected in almost all cases for all proportions of missing data. The confidence levels are between 95% and 100%. Thus, Little's test [142] is efficient, even when the missing data are scattered on the whole sample.

Jamshidian and Jalal's test [123] also obtains good results (when it is calculable). Excluding those obtained on samples containing between 20% and 45% missing data (corresponding to the proportions with few or no calculable test), Jamshidian and Jalal's test [123] very often concludes that the data are MCAR. The confidence levels of this test are between 89% and 99%.

Thus, among the two tests presented here, both obtain comparable results, but as Little's test [142] is always calculable, it is probably the most effective test. When missing data are found throughout the whole sample, Jamshidian and Jalal's test [123] is indeed less calculable. The threshold of the minimum number of patterns to consider is changeable in the function from *MissMech*, but it has been left at 6 to respect the recommendations of their authors. Thus, Little's test [142] is always calculable and able to detect MCAR data when they are actually MCAR.

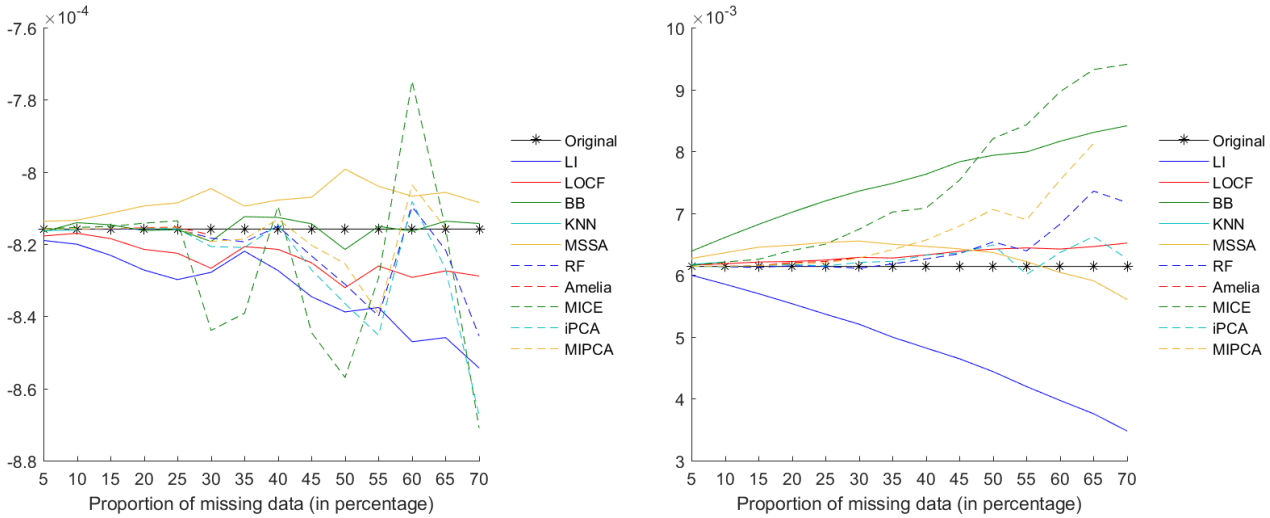
Thus, considering the results obtained here, as well as those of the previous section, both tests obtain good confidence levels as they are able to detect MCAR data. On the other hand, Jamshidian and Jalal's test [123] is not always calculable, which makes it complicated to use in practice. Little's test [142] is then preferable.

Statistical moments

The results obtained for the distribution of the first four statistical moments of the first column of the data matrix over the set of scenarios containing 30% missing data are presented in Appendix B.2 and show very little difference from those already observed in Figure 3.2-4 of the previous section. The fact that the missing data are distributed over the entire sample does not seem to have a significant impact on the mean, standard deviation, skewness and kurtosis of the first series. The only difference would concern the MICE algorithm, which gives slightly more stable imputations but is still one of the most variable methods from one scenario to another, which is not surprising, given that the method sees its universe of potential donors reduced when the missing data concerns the whole sample.

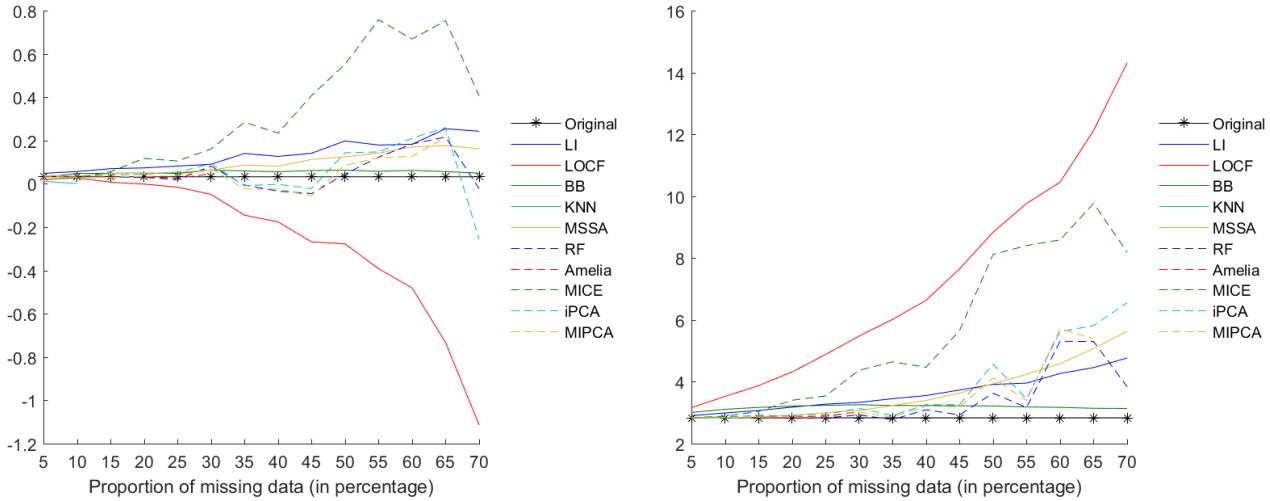
However, the performance of the MICE algorithm deteriorates strongly as the proportion of missing data increases in the sample. Figure 3.2-16a, Figure 3.2-16b, Figure 3.2-16c and Figure 3.2-16d represent the average of the means, standard deviation, skewness and kurtosis of the first time series, respectively, among the 100 scenarios of each missingness proportion for each completion method.

Fig. 3.2-16: Average of the first four statistical moments of the returns of the imputed data based on a matrix containing MCAR data in the whole sample, according to the missingness probability



(a) Mean

(b) Standard deviation



(c) Skewness

(d) Kurtosis

Thus, among all the imputation methods, the MICE algorithm is clearly one of the worst methods to adopt when the sample contains missing data over the entire data matrix (which was already the case when the missing data were concentrated in the first series only). This algorithm gives totally different means from one missingness proportion to another, but again, this may be due to the single sample used. In addition, it

very significantly overestimates the standard deviation, skewness and kurtosis. Finally, as in the previous section, this method gives unstable results from one missingness scenario to another, especially for the standard deviation and the skewness (according to Appendix B.3).

As mentioned earlier, the K-NN method does not work beyond 10% missing data, which is why its evolution is not comparable. On the other hand, the linear interpolation, LOCF, MSSA, Brownian bridge and IPCA methods provide the same (or almost the same) results as in Section 3.2.1, as if they were not impacted by the fact that the entire sample now contains missing data. This is completely normal for the one-dimensional methods (linear interpolation, LOCF and Brownian bridge), which do not use, by definition, the transversal dimension. On the other hand, the IPCA and MSSA use the transversal dimension, but their results are unchanged. This may be due to an overfitting problem (presented in Section 2.4.8). The columns of the sample are highly correlated, and it is possible that the methods trust the relationship between variables too much. These comparisons can be made thanks to Table 3.2-4, which compare the results from the previous section and from this section for several proportions of missing data. For example, the standard deviation of the original returns is 0. 061%, while the standard deviation of the linearly interpolated returns is 0.035% when 70% missing data are in the first column or in all columns.

Tab. 3.2-4: Average of the first four statistical moments for 10%, 30%, 50% and 70% missing data in the first columns versus missing data in the whole matrix

		Missingness proportion in the first column*				**Missingness proportion in all columns**			
		10%	**30%**	**50%**	**70%**	**10%**	**30%**	**50%**	**70%**
Mean (10^{-4})	**Original**	-8.16				-8.16			
	LI	-8.2	-8.28	-8.39	-8.54	-8.2	-8.28	-8.39	-8.54
	LOCF	-8.17	-8.27	-8.32	-8.29	-8.17	-8.27	-8.32	-8.29
	MSSA	-8.14	-8.09	-8.08	-8.18	-8.13	-8.05	-7.99	-8.08
	BB	-8.14	-8.2	-8.21	-8.14	-8.14	-8.19	-8.21	-8.14
	IPCA	-8.16	-8.15	-8.26	-8.2	-8.16	-8.21	-8.37	-8.68
	KNN	-8.16	-8.2	-8.21	-8.19	-8.16	-	-	-
	RF	-8.16	-8.17	-8.28	-8.33	-8.16	-8.18	-8.31	-8.45
	MICE	-8.15	-8.19	-8.29	-8.24	-8.15	-8.44	-8.57	-8.71
	MIPCA	-8.15	-8.18	-8.2	-8.27	-8.16	-8.19	-8.25	-
	Amelia	-8.15	-8.14	-8.11	-7.86	-8.16	-8.18	-	-
Standard Deviation (10^{-3})	**Original**	0.61				0.61			
	LI	0.59	0.52	0.44	0.35	0.59	0.52	0.44	0.35
	LOCF	0.62	0.63	0.64	0.65	0.62	0.63	0.64	0.65
	MSSA	0.64	0.66	0.64	0.56	0.64	0.65	0.64	0.56
	BB	0.66	0.74	0.79	0.84	0.66	0.74	0.79	0.84
	IPCA	0.61	0.62	0.62	0.62	0.61	0.62	0.65	0.63
	KNN	0.62	0.63	0.65	0.66	0.62	-	-	-
	RF	0.61	0.60	0.62	0.60	0.61	0.61	0.65	0.72
	MICE	0.66	0.73	0.86	1.03	0.62	0.67	0.82	0.94
	MIPCA	0.62	0.62	0.62	0.63	0.61	0.63	0.71	-
	Amelia	0.62	0.63	0.65	0.80	0.61	0.63	-	-

* Results from Section 3.2.1

		Missingness proportion in the first column*				Missingness proportion in all columns			
		10%	**30%**	**50%**	**70%**	**10%**	**30%**	**50%**	**70%**
Skewness	**Original**	0.03				0.03			
	LI	0.06	0.09	0.20	0.24	0.06	0.09	0.20	0.24
	LOCF	0.03	-0.05	-0.28	-1.11	0.03	-0.05	-0.28	-1.11
	MSSA	0.03	0.05	0.09	0.12	0.04	0.06	0.12	0.16
	BB	0.05	0.06	0.06	0.05	0.05	0.06	0.06	0.05
	IPCA	0.02	0.01	0.21	0.13	0.03	0.09	0.14	-0.26
	KNN	0.00	-0.07	-0.15	-0.33	0.00	-	-	-
	RF	0.03	0.08	0.00	-0.03	0.03	0.08	0.05	-0.02
	MICE	0.03	0.12	0.43	0.74	0.03	0.16	0.55	0.41
	MIPCA	0.02	-0.01	-0.02	-0.11	0.03	0.07	0.09	-
	Amelia	0.02	-0.03	-0.05	-0.03	0.03	0.05	-	-
Kurtosis	**Original**	2.83				2.83			
	LI	2.99	3.33	3.91	4.77	2.99	3.33	3.91	4.77
	LOCF	3.52	5.47	8.84	14.31	3.52	5.47	8.84	14.31
	MSSA	2.82	3.01	3.74	5.33	2.83	3.07	3.93	5.63
	BB	3.10	3.25	3.22	3.14	3.11	3.26	3.22	3.13
	IPCA	2.87	2.95	4.43	7.15	2.86	3.14	4.57	6.56
	KNN	2.86	2.99	3.48	4.18	2.86	-	-	-
	RF	2.84	2.89	3.19	3.49	2.83	2.93	3.63	3.83
	MICE	3.57	5.25	8.15	12.28	2.88	4.37	8.12	8.19
	MIPCA	2.87	2.97	3.46	6.19	2.84	3.02	4.12	-
	Amelia	2.86	2.92	3.10	4.36	2.84	3.03	-	-

* Results from Section 3.2.1

While some methods do not seem to be impacted by the presence of missing data in the entire sample, this is not the case for all the methods. The random forests method, which previously gave very satisfactory results (see Figure 3.2-5), tends to have increasing difficulty in reproducing the original series as the proportion of missing data increases throughout the sample. The algorithm tends to increase the volatility of the series as well as its kurtosis, especially for missingness proportion. Thus, this algorithm appears to lose its efficiency when the proportion of missingness exceeds 55%. On the other hand, Appendix B.3 shows that this algorithm is one of the most stable from one scenario to another (with MIPCA and Amelia), particularly when the proportion of missing data is less than 50%.

The MIPCA algorithm tends to be less efficient than the other methods when the entire sample contains missing data, compared to missing data are in the first column.

Apart from the fact that the mean is totally unstable from one proportion of missing data to another (but this is probably due to sampling effects), the volatility of the series tends to increase progressively as the missing data are added to the sample. In the case where the missing data only concerned the first series, the volatility of the data imputed by MIPCA was among the closest to that of the original series (see Figure 3.2-5b), whereas here, the volatility is almost at the same level as that imputed by Brownian bridge for a proportion of missingness at 65% (see Figure 3.2-16b). Moreover, the volatility obtained by IPCA is much lower than that obtained by MIPCA, which leads one to believe that the improved algorithm has no positive effect when the missing data concern the entire sample. In addition, both methods use a decreasing number of principal components as the missingness proportion increases, as presented in Appendix B.4. For skewness and kurtosis, the trends of the MIPCA and IPCA are constant until a missingness proportion of 40%. Beyond this threshold, it is possible to observe that the MIPCA method reproduces kurtosis much better when missing data are present in the whole sample (Table 3.2-4). Moreover, Appendix B.3 shows that this method is also very stable under this same threshold and less stable beyond it.

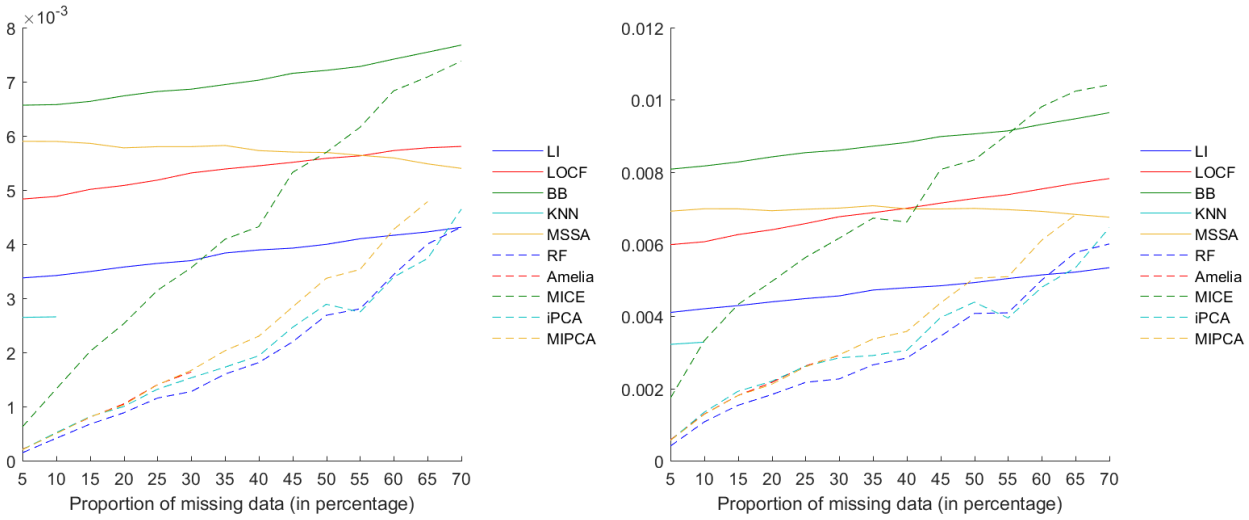
Finally, the Amelia algorithm seems (at least for a missing data proportion less than or equal to 30%) to impute missing data in the whole sample with the same efficiency as when they were only in one column of the sample.

Proximity measures

With respect to the proximity measures, the results obtained for a proportion of missing data of 30% over the entire sample (see Appendix B.5) are similar to those presented previously when missing data were only in the first column (see Figure 3.2-7). The only difference concerns the MICE algorithm, which, for the first four statistical moments, gives lower and more stable results.

On the other hand, this trend is not true for all proportions of missing data. The averaged MAE and RMSE per missing data proportion are presented in Figure 3.2-17a and Figure 3.2-17b. These proximity measures are based on the first column of the data matrix to compare the returns of the first column of the original sample with those of the first column of the imputed sample.

Fig. 3.2-17: Average MAE and RMSE between the return of the imputed data from a matrix containing MCAR data in the whole sample and the original data matrix, according to the missingness probability



(a) MAE *(b)* RMSE

As previously mentioned, Table 3.2-5 represents the MAE and RMSE results from the previous section and from this section for several missingness proportions to observe the impact of missing data in the whole sample. For example, the MAE of the linearly interpolated returns is equal to 0.431% when 70% missing data are in the first column or in all columns.

Tab. 3.2-5: Average MAE and RMSE for 10%, 30%, 50% and 70% missing data in the first columns versus missing data in the whole matrix

		Missingness proportion in the first column*				**Missingness proportion in all columns**			
		10%	**30%**	**50%**	**70%**	**10%**	**30%**	**50%**	**70%**
MAE (10^{-3})	**LI**	3.42	3.69	4.00	4.31	3.42	3.69	4.00	4.31
	LOCF	4.88	5.31	5.59	5.80	4.88	5.31	5.59	5.80
	MSSA	5.90	5.82	5.70	5.40	5.90	5.80	5.69	5.40
	BB	6.57	6.86	7.21	7.67	6.58	6.86	7.21	7.68
	IPCA	1.88	2.42	3.87	4.16	0.52	1.53	2.89	4.65
	KNN	2.60	2.69	2.82	2.95	2.66	-	-	-
	RF	0.39	0.96	1.78	2.72	0.41	1.28	2.68	4.32
	MICE	6.38	6.79	7.52	8.26	1.32	3.56	5.70	7.38
	MIPCA	2.20	2.24	2.52	3.40	0.50	1.67	3.37	-
	Amelia	2.27	2.37	2.54	4.13	0.50	1.64	-	-
RMSE (10^{-3})	**LI**	4.21	4.56	4.93	5.35	4.21	4.56	4.93	5.35
	LOCF	6.06	6.76	7.26	7.82	6.06	6.76	7.26	7.82
	MSSA	6.97	7.01	6.97	6.73	6.98	7.00	6.99	6.75
	BB	8.16	8.59	9.05	9.63	8.17	8.60	9.05	9.64
	IPCA	2.39	3.10	5.05	5.69	1.34	2.85	4.39	6.45
	KNN	3.20	3.38	3.66	4.03	3.28	-	-	-
	RF	1.00	1.69	2.65	3.84	1.07	2.26	4.08	6.00
	MICE	7.85	8.58	9.81	11.43	3.32	6.17	8.33	10.41
	MIPCA	2.77	2.88	3.29	4.70	1.28	2.93	5.06	-
	Amelia	2.85	3.03	3.31	5.69	1.29	2.92	-	-

* Results from Section 3.2.1

The first major difference with Figure 3.2-7, as previously observed, is indeed about the MICE algorithm. It appears to be very efficient when the proportion of missing data distributed over the entire sample is low (see Table 3.2-5). In fact, it is much more efficient than when the missing data are distributed over only one variable (approximately 5 times less for the MAE, and 2 times less for the RMSE for 10% missing data). When this proportion increases, however, the differences with the original series also increase, and even tend to reach the same level as when the missing data are only in one series. Thus, the MICE method is much more efficient at imputing 10% missing data, when all variables also contain 10% missing data than when the rest of the sample is complete, and remains more efficient as long as the proportion of missing data remains below 70%. On the other hand, completing 70% of missing data by MICE gives almost the same results when the rest of the sample is complete, or contains 70% missing data.

This algorithm has the highest slope of the evolution of MAE and RMSE.

Moreover, the MAE of linear interpolation is lower than that of MICE when the proportion of missingness exceeds 30%, and this threshold decreases to 15% for the RMSE. In addition, there are several methods that, regardless of the proportion of missing data, give MAE and RMSE lower than that of MICE. This is notably the case for the random forests, Amelia (when calculable), IPCA and MIPCA algorithms. Although MICE performs well for a small proportion of missing data, it is not the best performer and fails to maintain its performance as the proportion of missing data increases.

The reason why the MICE algorithm performs better when the entire sample contains missing data (compared to when the missing data are only in the first column) is not obvious. Indeed, it is rather counterintuitive to think that a sample containing more information gives poorer results; this will be discussed later, but the computation time of the MICE algorithm is 6 times higher than in the previous case (comparing Figure 3.2-14 and Figure 3.2-20). In addition, Appendix B.6 shows that the MICE algorithm is still the less stable method from one scenario to another, which makes it not very reliable.

The same trend can be observed for the IPCA, MIPCA and Amelia algorithms up to a certain proportion of missing data; their proximity measures are lower here than with missing data in only the first column when the proportion of missing data is low (or at least, not too high), but they become greater when the proportion of missing data is significant (except for Amelia, which is not calculable for a high missingness proportion). Moreover, the evolution of the proximity measures of the IPCA is very similar to that of random forests and MIPCA. In the previous section (see Figure 3.2-7 or Table 3.2-5), where the missing data were only in the first time series, these three models had different proximity measures. However, when the missing data concern the whole sample, these three models (but also Amelia when it is calculable) seem to behave in a similar way.

For IPCA and MIPCA, the results are close to each other for any proportion of missing data (see Figure 3.2-17), whereas previously, the two models diverged when the proportion of missing data was over 30% (see Figure 3.2-7).

When the missing data are in the first column or across the entire sample, the MAE and RMSE levels from random forests are similar for a low proportion of missing data (see Table 3.2-5). On the other hand, the proximity measures from this model become higher when a high missing data proportion is present in all columns (see Figure 3.2-17), because decision trees are constructed from fewer observations and therefore lose precision.

The results of the MSSA are similar to the previous case (see Table 3.2-5). The fact that the missing data are spread over a single column or over the entire sample does

not change the proximity measures of the first column. This seems to indicate that the method makes little use of the transversal dimension. Thus, the performance of this method in terms of proximity measures remains poor.

The Amelia algorithm is the only algorithm that significantly improves its proximity measures, at least for the proportions where it is calculable. The MAE and RMSE, obtained from samples with missing data over the whole sample, are always lower than those obtained from samples with missing data only in the first column (comparing Figure 3.2-7 and Figure 3.2-17). Thus, it seems that the algorithm imputes missing data from a column better when the rest of the matrix also contains missing data rather than being whole, which appears to bee completely paradoxical. For MICE, this point will be discussed at the end of the chapter in Section 3.7.

Even if the random forests model is the most efficient in terms of proximity measures, the results for Amelia remain very close (see Figure 3.2-17). Moreover, these two methods appear to be equally stable from one scenario to another (see Appendix B.6).

Covariance matrices comparison

Thus, based on these results, some completion methods appear to more accurately impute missing data when the sample contains missing data that are distributed homogeneously throughout the entire sample and not just in one column. These results suggest that these same methods would be better at preserving the original covariance matrix when the missing data are distributed throughout the entire sample.

As a reminder, the previous case, presented in Figure 3.2-8 and Figure 3.2-9, shows the differences between the covariance matrices of the original series and those of the imputed series. In the first case, only the first row (and thus the first column) of the matrix was impacted. Here, given that the entire (or almost the entire) matrix contains missing data, the entire matrix is impacted.

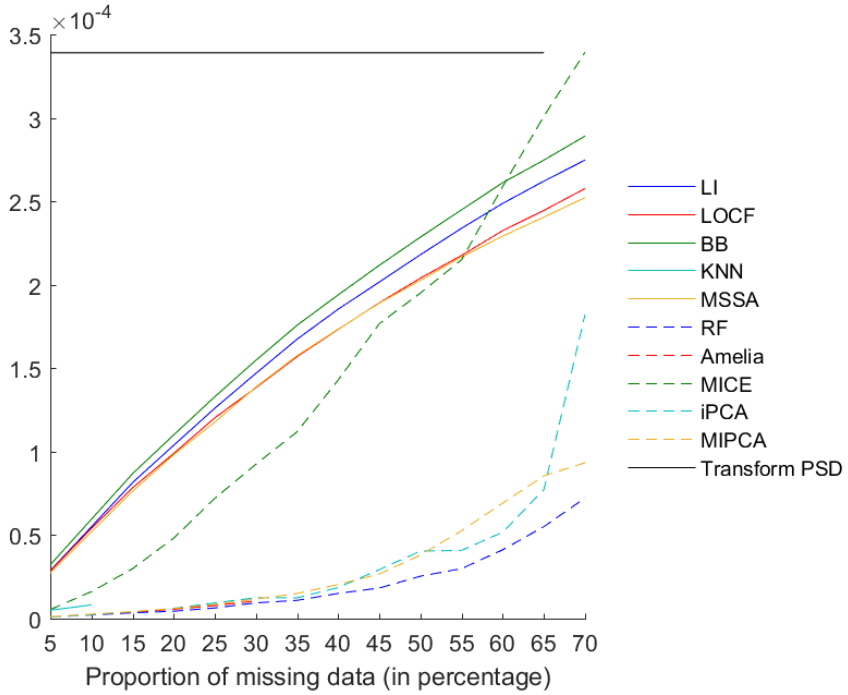
The results obtained for the 100 scenarios, containing 30% missing data are presented in Appendix B.7. When compared to the previous methods, presented in Figure 3.2-8, it can be seen that the usual imputation methods (linear interpolation and LOCF) have higher results. This is normal because, as said, the differences are in the whole covariance matrix and not only in one part.

Whereas previously these usual methods obtained covariance matrix differences relatively comparable to other methods (see Figure 3.2-8), here, the random forests, Amelia, IPCA and MIPCA methods give differences that are approximately 7 times smaller. This is consistent with previous results, where these same methods obtained much lower proximity measures than when the missing data were concentrated only in the first series. Moreover, the same phenomenon is observable for the MICE algorithm (more stable results) but to a lesser extent. Thus, as before, this trend is probably

observable for the proportion of low missingness but may be progressively less as it increases.

Figure 3.2-18 thus represents the average result (among the 100 scenarios) of the covariance matrix deviations, according to the missingness proportion.

Fig. 3.2-18: Average covariance matrix differences, according to the Frobenius norm, based on original returns and the imputed returns from a matrix containing MCAR data in the whole sample, according to the missingness probability



The first remark is that the usual methods (linear interpolation and LOCF) give higher results than when the missing data were only in the first column of the data matrix (see Table 3.2-6). However, this is not a surprise, as the entire covariance matrix was impacted.

However, these usual methods remain preferable to the pairwise matrix transformation as Rousseeuw and Molenberghs [173] define it (see Section 2.2.2). The differences between the pairwise transformed matrix (to have a positive semidefinite matrix) and the original covariance matrix are even greater than with the usual methods. It is, therefore, better to use a covariance matrix based on linearly interpolated data than to transform the pairwise matrix. In addition, when the proportion of missing data becomes too large, the pairwise matrix may not be calculable. When the proportion

of missingness is greater than or equal to 60%, some scenarios do not even make it possible to calculate a pairwise matrix. The results presented here are all based on 100 scenarios in which the proportion of missing data is less than 60%. However, for a proportion of 60%, only 95 pairwise matrices could be computed, 59 pairwise matrices for a proportion at 65%, and none for a proportion at 70% (hence a shortened curve).

Tab. 3.2-6: Average covariance differences (10^{-4}), according to the Frobenius norm, for 10%, 30%, 50%, and 70% missing data in the first columns versus missing in the whole matrix

	Missingness proportion in the first column*				**Missingness proportion in all columns**			
	10%	**30%**	**50%**	**70%**	**10%**	**30%**	**50%**	**70%**
Transform PSD	3.39	3.39	3.39	-	3.39	3.39	3.39	-
LI	0.46	0.79	1.46	2.20	0.55	1.47	2.18	2.75
LOCF	0.65	0.76	0.92	1.06	0.54	1.38	2.04	2.58
MSSA	0.83	1.05	0.85	0.57	0.52	1.39	2.03	2.52
BB	1.11	2.11	3.01	3.87	0.60	1.55	2.29	2.89
IPCA	0.61	0.68	0.65	1.00	0.02	0.12	0.40	1.82
KNN	0.67	0.76	0.93	1.13	0.08	-	-	-
RF	0.60	0.54	0.63	0.61	0.02	0.09	0.25	0.72
MICE	1.05	2.11	4.29	7.69	0.16	0.93	1.95	3.39
MIPCA	0.64	0.65	0.68	0.86	0.03	0.12	0.38	-
Amelia	0.66	0.77	0.97	5.07	0.02	0.11	-	-

* Results from Section 3.2.1

By contrast, the behavior of the Brownian bridge and the MSSA is similar to that of usual methods. In other words, as the missing data multiply in the whole sample, the quality of the covariance matrix deteriorates. This was already the case of the Brownian bridge when the missing data were only in the first column but not for the MSSA. The MSSA algorithm was one of the methods that made it possible to keep the covariance matrix, even when the missingness proportion in only the first column was very important (see Table 3.2-6), and here it is comparable to usual methods.

Once again, the MICE algorithm is one of the worst methods to preserve the original covariance matrix and also the least stable (see Appendix B.8). Nevertheless, the results here are weaker than in the previous section (see Table 3.2-6), while the entire covariance matrix is impacted here. This means the MICE algorithm is better able to preserve the covariance matrix when the missing data are distributed throughout the sample than when it is only in one column. This phenomenon was already observed with the previous comparison tools.

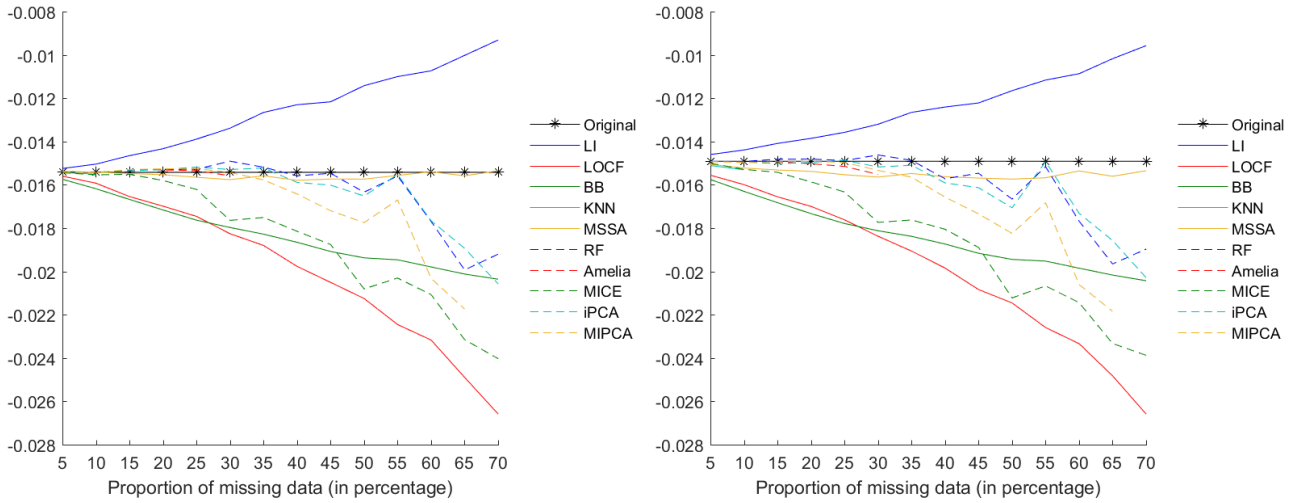
Finally, the random forests, IPCA, MIPCA, and Amelia (when it is calculable) methods are the most efficient for reproducing the covariance matrix when the proportion of missing data are below 40%. Beyond this threshold, the random forests, IPCA, and MIPCA tend to distort the covariance matrix as the missing data proportion increases through the sample. However, these algorithms obtain a better result than when the data were missing only from the first column (see Table 3.2-6), except for a high missingness proportion. Thus, the fact that these methods impute data throughout the matrix allows for less impact on the covariance matrix.

Value-at-risk and expected shortfall

Finally, the presence of missing data across the entire sample makes little difference in terms of VaR and ES when missing data are in only one column of the sample. The risk measures were recalculated from a sample initially containing 30% missing data in all the samples, and the results of the 100 scenarios are presented in Appendix B.9. These results are very similar to those already presented in Figure 3.2-10 and Figure 3.2-12, where the missing data are concentrated on the first series. But this is not necessarily true for higher proportions of missing data.

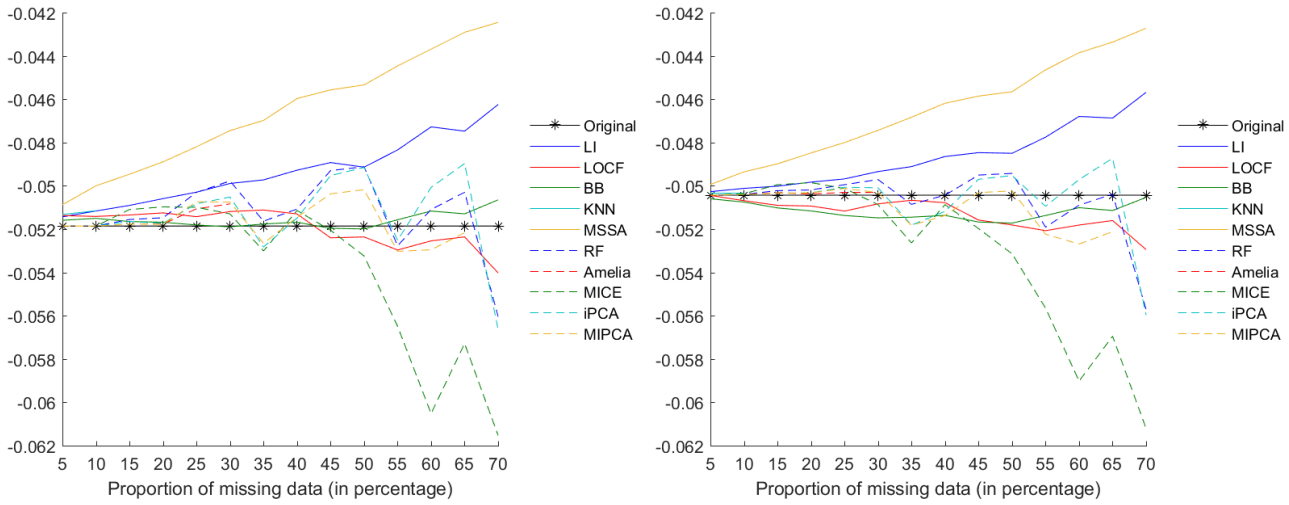
Thus, in Figure 3.2-19, the same calculation was performed for each missingness proportion and averaged to compare these results with those of Figure 3.2-11 and Figure 3.2-13, where the missing data were concentrated in the first column of the sample.

Fig. 3.2-19: Average 1-day and 10-day risk measures computed from a data matrix containing MCAR data in the whole sample, according to the missingness probability



(a) 1-day VaR at 99%

(b) 1-day ES at 97.5%



(c) 10-day VaR at 99%

(d) 10-day ES at 97.5%

Again, the results are comparable to those of Figure 3.2-10(see Table 3.7-5) presented in the previous section. For example, the 1-day VaR from original data is equal to -1.54%, whereas that of the linearly interpolated returns is equal to -0.93% when 70% of missing data are in the first column or all the columns.

The MSSA algorithm remains the most efficient to reproduce risk measures when the horizon is 1 day but remains one of the worst when the horizon increases to 10 days. Conversely, the Brownian bridge is optimal for a 10-day horizon but not for a 1-day horizon.

All methods lead to conservative 1-day risk measures (except linear interpolation, of course), but this is not the case for 10-day risk measures.

Tab. 3.2-7: Average 1-day and 10-day risk measures for 10%, 30%, 50%, and 70% missing data in the first columns versus missing in the whole matrix

		Missingness proportion in the first column*				**Missingness proportion in all columns**			
		10%	**30**%	**50**%	**70**%	**10**%	**30**%	**50**%	**70**%
$VaR_{99\%}^{1-day}$ (10^{-2})	**Original**	-1.54				-1.54			
	LI	-1.5	-1.34	-1.14	-0.93	-1.5	-1.34	-1.14	-0.93
	LOCF	-1.59	-1.83	-2.13	-2.66	-1.59	-1.83	-2.13	-2.66
	MSSA	-1.54	-1.58	-1.58	-1.57	-1.54	-1.58	-1.57	-1.54
	BB	-1.62	-1.79	-1.93	-2.04	-1.62	-1.8	-1.94	-2.04
	IPCA	-1.55	-1.54	-1.61	-1.8	-1.54	-1.53	-1.65	-2.06
	KNN	-1.55	-1.57	-1.63	-1.82	-1.55	-	-	-
	RF	-1.54	-1.48	-1.58	-1.61	-1.54	-1.49	-1.63	-1.92
	MICE	-1.67	-2.01	-2.23	-2.54	-1.55	-1.76	-2.08	-2.4
	MIPCA	-1.55	-1.55	-1.58	-1.82	-1.54	-1.54	-1.77	-
	Amelia	-1.55	-1.58	-1.64	-2.09	-1.54	-1.56	-	-
$ES_{97.5\%}^{1-day}$ (10^{-2})	**Original**	-1.49				-1.49			
	LI	-1.44	-1.32	-1.17	-0.96	-1.44	-1.32	-1.17	-0.96
	LOCF	-1.6	-1.84	-2.15	-2.66	-1.6	-1.84	-2.15	-2.66
	MSSA	-1.52	-1.57	-1.58	-1.55	-1.52	-1.56	-1.57	-1.53
	BB	-1.63	-1.81	-1.94	-2.04	-1.63	-1.81	-1.94	-2.04
	IPCA	-1.51	-1.52	-1.6	-1.82	-1.49	-1.52	-1.71	-2.03
	KNN	-1.53	-1.6	-1.71	-1.9	-1.53	-	-	-
	RF	-1.49	-1.44	-1.57	-1.6	-1.49	-1.46	-1.67	-1.9
	MICE	-1.69	-2.02	-2.26	-2.55	-1.53	-1.77	-2.12	-2.39
	MIPCA	-1.52	-1.54	-1.6	-1.85	-1.5	-1.53	-1.82	-
	Amelia	-1.52	-1.57	-1.64	-2.11	-1.5	-1.55	-	-

* Results from Section 3.2.1

		Missingness proportion in the first column*				Missingness proportion in all columns			
		10%	**30%**	**50%**	**70%**	**10%**	**30%**	**50%**	**70%**
$VaR_{99\%}^{10-day}$ (10^{-2})	**Original**	-5.19				-5.19			
	LI	-5.12	-4.99	-4.91	-4.62	-5.12	-4.99	-4.91	-4.62
	LOCF	-5.14	-5.12	-5.24	-5.4	-5.14	-5.12	-5.24	-5.4
	MSSA	-500	-4.75	-4.53	-4.22	-500	-4.74	-4.53	-4.24
	BB	-5.15	-5.19	-5.2	-5.07	-5.15	-5.19	-5.2	-5.06
	IPCA	-5.2	-5.26	-5.27	-5.15	-5.19	-5.05	-4.92	-5.66
	KNN	-5.12	-5.04	-4.99	-4.99	-5.12	-	-	-
	RF	-5.19	-4.97	-4.85	-5.26	-5.19	-4.98	-4.91	-5.6
	MICE	-5.18	-5.28	-5.52	-6.14	-5.19	-5.13	-5.32	-6.15
	MIPCA	-5.16	-5.09	-500	-4.98	-5.19	-5.08	-5.02	-
	Amelia	-5.17	-5.14	-5.12	-5.56	-5.19	-5.08	-	-
$ES_{97.5\%}^{10-day}$ (10^{-2})	**Original**	-5.04				-5.04			
	LI	-5.01	-4.93	-4.85	-4.57	-5.01	-4.93	-4.85	-4.57
	LOCF	-5.07	-5.08	-5.18	-5.29	-5.07	-5.08	-5.18	-5.29
	MSSA	-4.94	-4.74	-4.56	-4.26	-4.94	-4.74	-4.57	-4.27
	BB	-5.08	-5.15	-5.17	-5.06	-5.08	-5.15	-5.17	-5.05
	IPCA	-5.06	-5.1	-5.15	-5.09	-5.04	-5.01	-4.95	-5.6
	KNN	-5.03	-500	-4.97	-4.97	-5.03	-	-	-
	RF	-5.04	-4.96	-4.9	-5.21	-5.04	-4.97	-4.94	-5.57
	MICE	-5.1	-5.23	-5.48	-6.1	-5.04	-5.09	-5.31	-6.12
	MIPCA	-5.05	-5.03	-4.98	-4.95	-5.04	-5.03	-5.02	-
	Amelia	-5.05	-5.06	-5.07	-5.52	-5.04	-5.03	-	-

* Results from Section 3.2.1

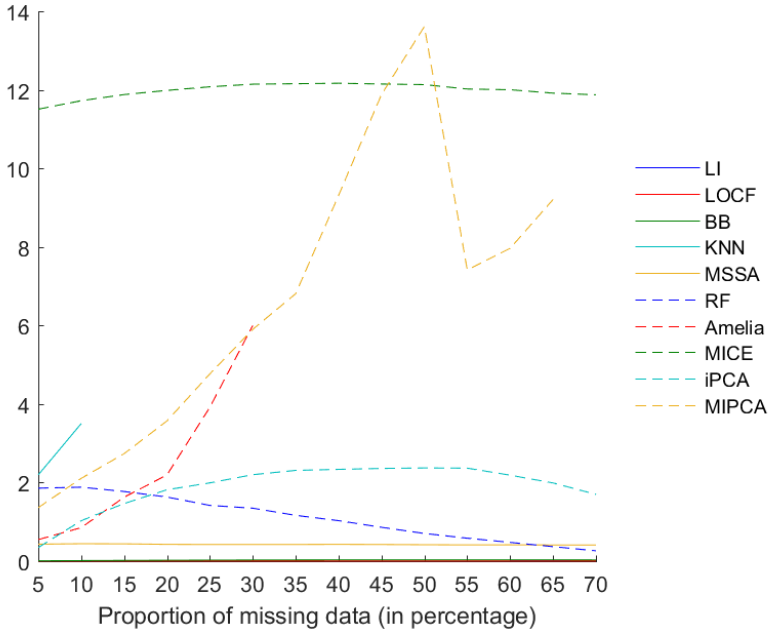
Thus, while some algorithms seem to impute that missing data (e.g. proximity measures) are improved, this is not enough to improve the risk measures. The fact that missing data are found across the entire sample has relatively little impact on the risk measures. None of the completion methods appear to be significantly optimal for either 1-day or 10-day risk measures. In addition, the fact that missing data are found throughout the entire sample tends to make the risk measures slightly more unstable from one missingness proportion to another.

Computation time

Finally, the very last comparison criterion concerns the computation time. Given that the proportion of missing data in the whole sample is higher, the computation time

should naturally be higher as well. The computation time (in seconds) is presented in Figure 3.2-20.

Fig. 3.2-20: Average computation time of each algorithm imputing a matrix that contains MCAR data in the whole sample, according to the missingness probability



Overall, all completion methods impute missing data more slowly than before (see Figure 3.2-14). As mentioned earlier, the computation time of the MICE algorithm is six times longer than before. The MIPCA algorithm has an unstable computation time that varies from approximately 2 to 14 seconds to impute a sample, compared with 1 to 4 seconds previously, depending on the proportion of missing data. The Amelia algorithm also has a computation time that increases from 0.5 to 6 seconds for proportions of missing data ranging from 5% to 30%. The IPCA has a stable calculation time (between 1 and 2 seconds), whereas in the previous case (from Figure 3.2-14), it was increasing but was still less than 1 second. Moreover, these methods (MICE, MIPCA, Amelia, MIPCA, and IPCA) are among those whose performance is improved when the missing data are distributed over the whole matrix. Thus, if the computation time is more important here, it is because there is more data to impute but also because the algorithm increases its number of iterations to reach the convergence of the parameters or the imputed values.

As before, random forests have a computation time that decreases with the proportion of missing data; and yet, it remains higher than before (albeit among the lowest).

Only the computation time of the MSSA algorithm remains unchanged (approximately 0.5 seconds), but this is logical given that the dimension of the data matrix is also unchanged.

In this section, it appears that the MCAR data are much more complex to impute when spread over the whole sample than when it is concentrated in the first column as in the previous section (section 3.2.1).

Overall, the MICE, Brownian bridge, and MSSA methods struggle to be much more efficient than at least one of the two usual methods. They are either too unstable, too volatile, or simply not much more efficient than a simple linear interpolation. This phenomenon was already observed in Section 3.2.1, albeit to a lesser extent. Moreover, the K-NN, which was among the best-performing algorithms in the previous section, cannot even impute the missing data without generating an error because the missing data are spread over all the columns.

By contrast, the random forests, IPCA, MIPCA, and Amelia algorithms appear to be much more efficient than the usual methods. These methods obtain very comparable results when the proportion of missing data is below 40% (30% for Amelia), which makes it difficult to establish a ranking. Beyond this threshold, all the methods' performances start to deteriorate.

However, MICE, Amelia, IPCA, and MIPCA seem to obtain even more satisfactory results when the proportion of missing data is not too high, which seems to be counter-intuitive.

3.2.3 Impact of heteroskedasticity

In the previous section, volatility was assumed to be constant throughout the period. The annualized volatility is set at 10% for the entire sample, which means the volatility is the same every day. In practice, daily volatility is far from constant. Financial series are generally characterized by a succession of periods of low and high variance, representing quiet moments and phases of high volatility, respectively. These periods of high volatility can last a few minutes, a few days, or even a few months following, for example, a tweet, a public offer, or a bank failure. Thus, volatility clusters are common in financial series.

In the previous chapter, the case of the WTI and Brent crude oil prices were used, and volatility was not constant over the period (see Figure 2.4-11b) – hence, the interest in analyzing the behavior of each algorithm concerning heteroskedastic time series.

Simulated sample with non-constant volatility

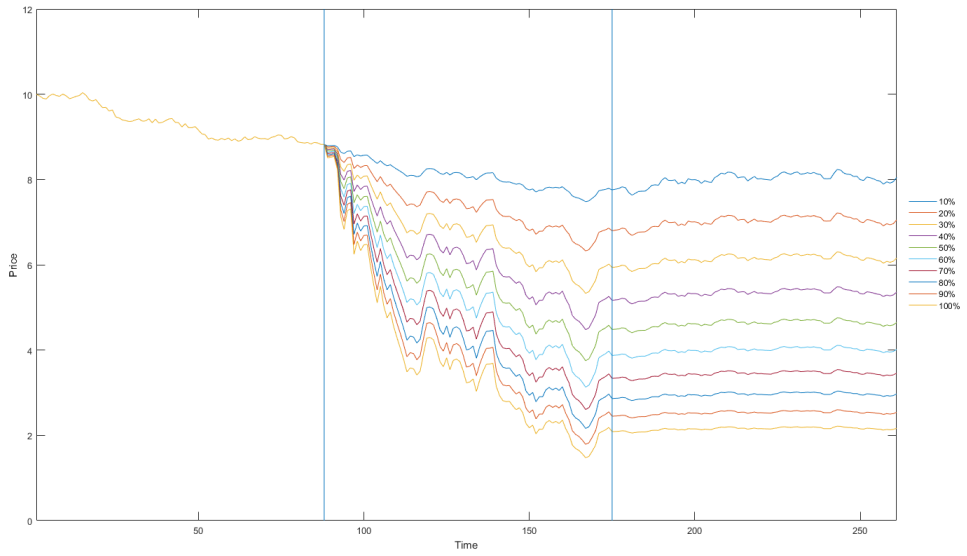
The sample used here is the same as the one used in the previous section (3.1.1), which is a sample of 10 price series, each containing 261 observations, where the price dynamics

follow a log-normal process given by Equation 3.1-3. The only difference is the volatility parameter σ, which now takes two different values depending on the period.

The idea of approach used in this section is to integrate a period of financial crisis: a sub-period of high volatility between two calm periods. To this end, the sample was divided into three parts so that the first and final thirds had an annualized volatility of 10% (as in the previous sections), but the second third had a period of higher volatility.

Thus, the analyses in this section are based on 10 simulated samples containing a crisis period for all columns, with increasing annualized volatility. For each sample, this annualized volatility of the calm period is set at 10%, and that of the crisis period takes a value from 10% to 100%, in increments of 10%. For example, the first sample contains an annualized volatility of 10% for the entire sample (normal periods and crisis period), the second sample is composed of a crisis period of volatility at 20% (the normal period is still the same, that is, with an annualized volatility at 10%), the third of a crisis period of volatility at 30%, and so forth, as presented in Figure 3.2-30. Given that the sample consists of 261 days, each of these periods counts as 87 days.

Fig. 3.2-21: First column of the data matrix with heteroskedasticity: normal period with an annualized volatility of 10% and crisis period with an annualized volatility ranging from 10% to 100%



All the columns of the data matrix follow this volatility pattern, which means the volatility of each column from a data matrix is the same.

The first of these samples (with a constant annualized volatility at 10%) is the same as in Section 3.2.1 and Section 3.2.2 and is used as a benchmark. The others are used

to analyze the behavior of completion algorithms faced with a sub-period of increasing volatility.

Finally, data are removed in a completely random way from each of the first columns of these price samples (as in Figure 3.2-1), following the same mechanism presented in Section 3.2.1 (drawing from a uniformly distributed random variable). The proportion of missing data injected into all the first columns is set at 30% for all the samples in this section, which corresponds to 51% missing returns on average for the algorithms applied to the yields (see Table 3.2-1). In order not to observe results based on a unique MCAR scenario, 100 scenarios containing 30% missing data are used.

The application of MCAR tests or completion algorithms on this type of data remains debatable. As the data are simulated with a crisis period, it is not stationary across the whole period, which calls into question the application of tests and completion methods in the whole period. Indeed, historical data are not necessarily stationary. It is clear that applying an EM algorithm on non-stationary data is problematic and will necessarily lead to a deterioration of the performance of the Amelia algorithm. Thus, it could have disastrous impacts in an automatic process context where no volatility check is used. To overcome this problem, it is possible to apply these tests and completion methods on well-chosen subparts, but it is still necessary to find at which moments the changes of scheme took place. This is simple with simulated data but becomes a real problem with historical data.

MCAR tests and completion methods will be applied over the whole period, even if this may appear suboptimal for some methods. Moreover, some algorithms can perform better on non-stationary data, such as the MICE, MSSA, and random forests algorithms.

The rest of this section consists of comparing the algorithms with each other and examining the results previously obtained to gauge the impact of heteroskedasticity on the algorithms. The first step is to apply the MCAR tests to these samples to see whether they give different results differently with heteroskedasticity.

Finally, each comparison tools is computed as defined in Figure 3.1-2, and the templates of all the graphs presented in this section have already been presented and detailed (i.e., what they represent and how they were obtained) in Section 3.1.4.

MCAR tests

As mentioned, the application of these tests on non-stationary data is not optimal, but it is presented for illustrative purposes. As in Section 3.2.1 and Section 3.2.2, Little's test [142] and Jamshidian and Jalal's test [123] are applied to the return matrices to see whether they could determine that the data are MCAR.

Appendix C.1 shows that Jamshidian and Jalal's test [123] has no difficulty to be calculated for a proportion of missing data of 30%, regardless of the volatility of the crisis period.

Table 3.2-8 represents the proportion of tests not rejecting the null hypothesis that the data are MCAR.

Tab. 3.2-8: Confidence level (probability of not rejecting H_0 when H_0 is true) for both MCAR tests applied to heteroskedastic price return matrices containing MCAR on the first column of the matrix, for a 5% significance level

	Volatility of the crisis period									
	10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Little's test	94%	93%	91%	90%	90%	90%	89%	90%	91%	91%
J&J's test	97%	86%	82%	77%	69%	67%	61%	58%	50%	46%

The observation is clear: The more violent the crisis period is, the less Jamshidian and Jalal's test [123] can recognize data as MCAR, while Little's test [142] is not affected. Jamshidian and Jalal's test [123] is effective at correctly detecting MCAR data when the volatility of the crisis period is low, but its efficiency decreases as the volatility of the crisis period increases. The confidence levels of Jamshidian and Jalal's test [123] decrease from 97% to 46%. By contrast, confidence levels from Little's test [142] are stable: between 89% and 94%.

Given these results, Little's test [142] appears to be efficient at detecting MCAR data when the latter are MCAR, even with heteroskedasticity. This is not the case with Jamshidian and Jalal's test [123], which more and more rejects the null hypothesis that the data are MCAR when the series is heteroskedastic.

The MCAR tests were applied here to the entire period, even though the data were heteroskedastic, but another approach would be to apply the tests over a sliding period to observe the behavior of the test on heteroskedastic data in order to analyze the results of all possible subparts (the choice of the sliding window is still to be determined).

Statistical moments

The impact of the crisis period's increasing volatility on the first four statistical moments is analyzed below. The results presented in Appendix C.2 represent the averaged statistical moments obtained for each of the completion methods associated with the entire heteroskedastic series (first column of the data matrix).

For the sample used in this section, each moment will vary as the volatility of the crisis period increases. The mean will decrease (due to the downward trend observed in

Figure 3.2-30), the standard deviation will increase (this is the goal of the exercise), the skewness will drop a little below zero for this sample, and the kurtosis will rise above 3 (specific to Gaussian mixtures).

These results give an average answer over the total sample used; however, they do not give any information about the behavior of the algorithms specific to the crisis and non-crisis periods.

Since the sample is divided into three periods, the following analysis is also split into three periods:

1. before the crisis
2. during the crisis, and
3. after the crisis.

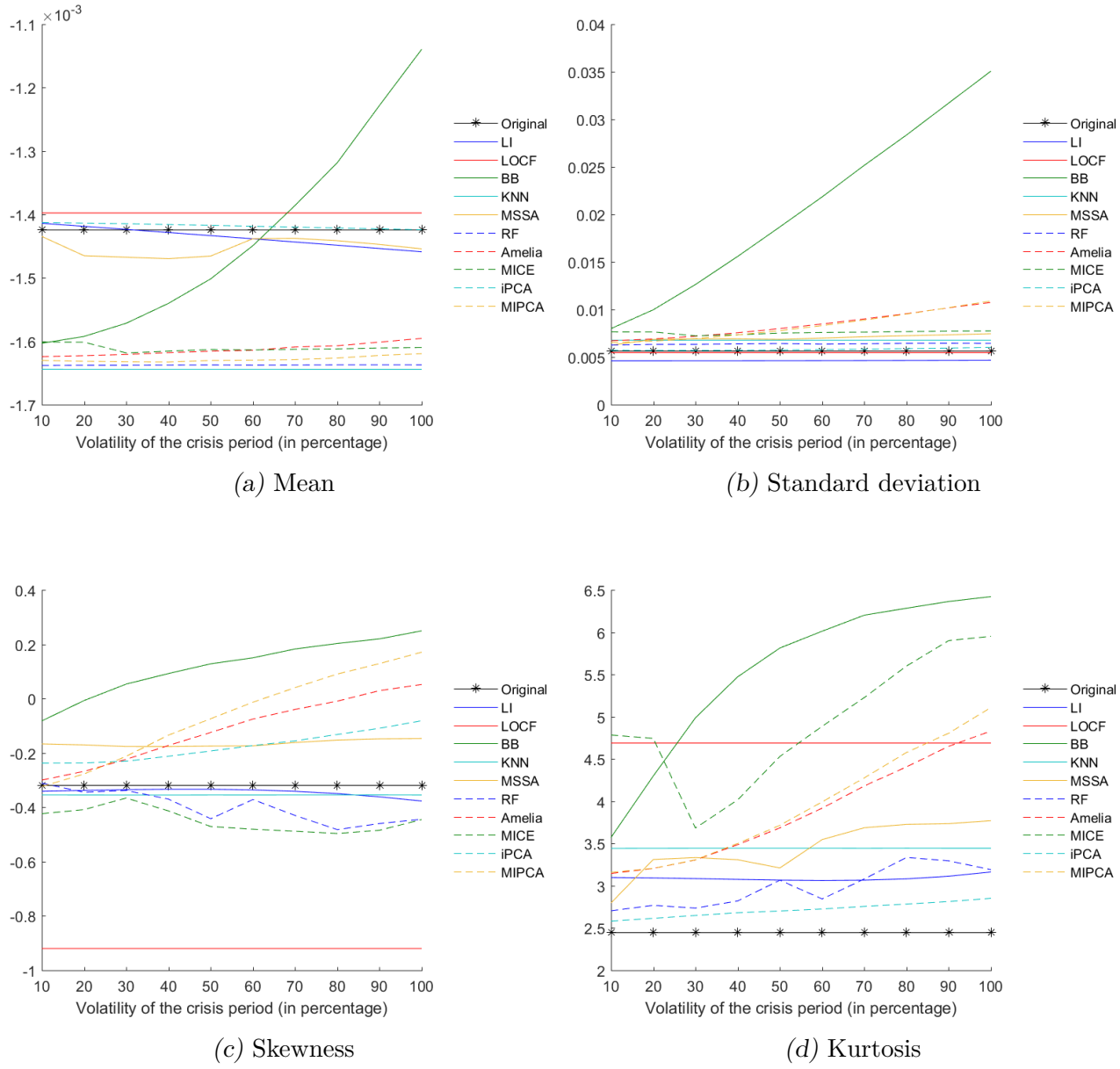
Rather than analyzing the average results of the whole sample, the results are analyzed period by period to see whether the algorithms can take into account the heteroskedasticity of the series. Each of these periods consists of approximately 87 days.

Thus, the moments calculated for each period are those of the analyzed period; that is why the statistical moments of the original series (represented by the black starred line) are different for the three periods. In the case of the standard deviation, the x-axis represents the (annualized) volatility of the crisis period, and the y-axis represents the average observed volatility of the analysis period (over 87 days). For the other graphs, the y-axis represents the average mean, skewness, and kurtosis of the returns for the period.

Before the crisis

Figure 3.2-22 depicts the evolution of the first four statistical moments (based on the returns of the first column) for the first period (calm period).

Fig. 3.2-22: Average of the first four statistical moments of the returns of the imputed data based on a matrix containing MCAR data for the first period (calm period) according to the volatility of the crisis period



(a) Mean

(b) Standard deviation

(c) Skewness

(d) Kurtosis

This first period corresponds to a calm period since its annualized volatility is 10%, as in the previous studies. Over this period, the parameters of the Gaussian law used to make this simulation are constant for all samples (because volatility increases in the next period); therefore, the statistical moments of the original data associated with this period are, logically, constant for each sample used.

Here, many of the methods tend to underestimate the mean because the mean of the series over the crisis period is decreasing. Moreover, many methods overestimate kurtosis because the kurtosis of the total series is higher than 3 and increases with the volatility of the crisis period. Nevertheless, the standard deviation evolution is relatively constant for the majority of methods, while the standard deviation is impacted upward in the next period.

The Brownian bridge method is the furthest from the real moments of the series. Besides increasingly overestimating the volatility of the first period as the volatility of the second period increases, it completely overestimates skewness and kurtosis. These results are because the Brownian bridge sets these parameters from the available set of observations of all series without distinguishing the periods of high and low volatility. Thus, the greater the volatility of the crisis period, the greater the impact on calm periods. This method, as implemented, is unable to take into account the heteroskedasticity of the series. The solution is to implement a Brownian bridge with a non-constant volatility parameter (as in the GARCH model [37]) or to impute the missing data by dividing the sample into sub-samples.

The usual methods obtain (almost) constant results given that the proportion of missing data is the same for all levels of volatility tested. Moreover, since these methods use local information, they are not (or little) impacted by the high volatility of the second period. A slight slope is observed for linear interpolation, due to missing data located at the end of the first period and, thus, data from the crisis period being used to perform the interpolation. Between these two methods, linear interpolation is much more satisfactory and is one of the methods that produce moments closest to those of the original period.

Among the more sophisticated methods, the IPCA method can adequately reproduce the statistical moments of the first period and obtains much better results than the MIPCA algorithm. Yet, both methods use about the same number of principal components (four on average, according to Appendix C.4). Moreover, IPCA appears to be one of the least stable methods from one missingness scenario to another, according to Appendix C.3. Here again, the MIPCA algorithm obtains results comparable to those of Amelia. Both methods have moments that become increasingly distant from those of the original period, as the volatility of the second period becomes more and more important. It is not surprising for the Amelia algorithm that, despite the preliminary stage of data standardization of the whole matrix, reproduces the global distribution of the matrix without distinguishing between crisis and non-crisis periods.

Working longitudinally, the MICE algorithm could consider the heteroskedasticity of the series. This algorithm, which imputes based on the closest observations, could pool low-volatile and high-volatile returns. The results of the MICE algorithm are better than those previously observed but not exceptional; however, they are comparable to

those of Amelia and MIPCA, which proved to be good performers in the previous sections. Moreover, in the previous results presented in Section 3.2.1 and Section 3.2.2, MICE was always among the least performing methods, whereas here, the results are more consistent with those of many other algorithms.

It should nevertheless be noted that a break is observable for each of the moments from MICE – between 20% and 30% of missing data, which suggests sampling errors. The sample that is used can cause this algorithm to react abnormally when the volatility is 10% or 20%. Thus, the results observed for this algorithm in the previous sections (i.e., a constant annualized volatility of 10%) would not be representative of this method's performance. For verification, the same analysis should be done on a large number of samples, but this will be the subject of future work. In this PhD PhD thesis , two historical samples will be used and will allow a comparison of the results.

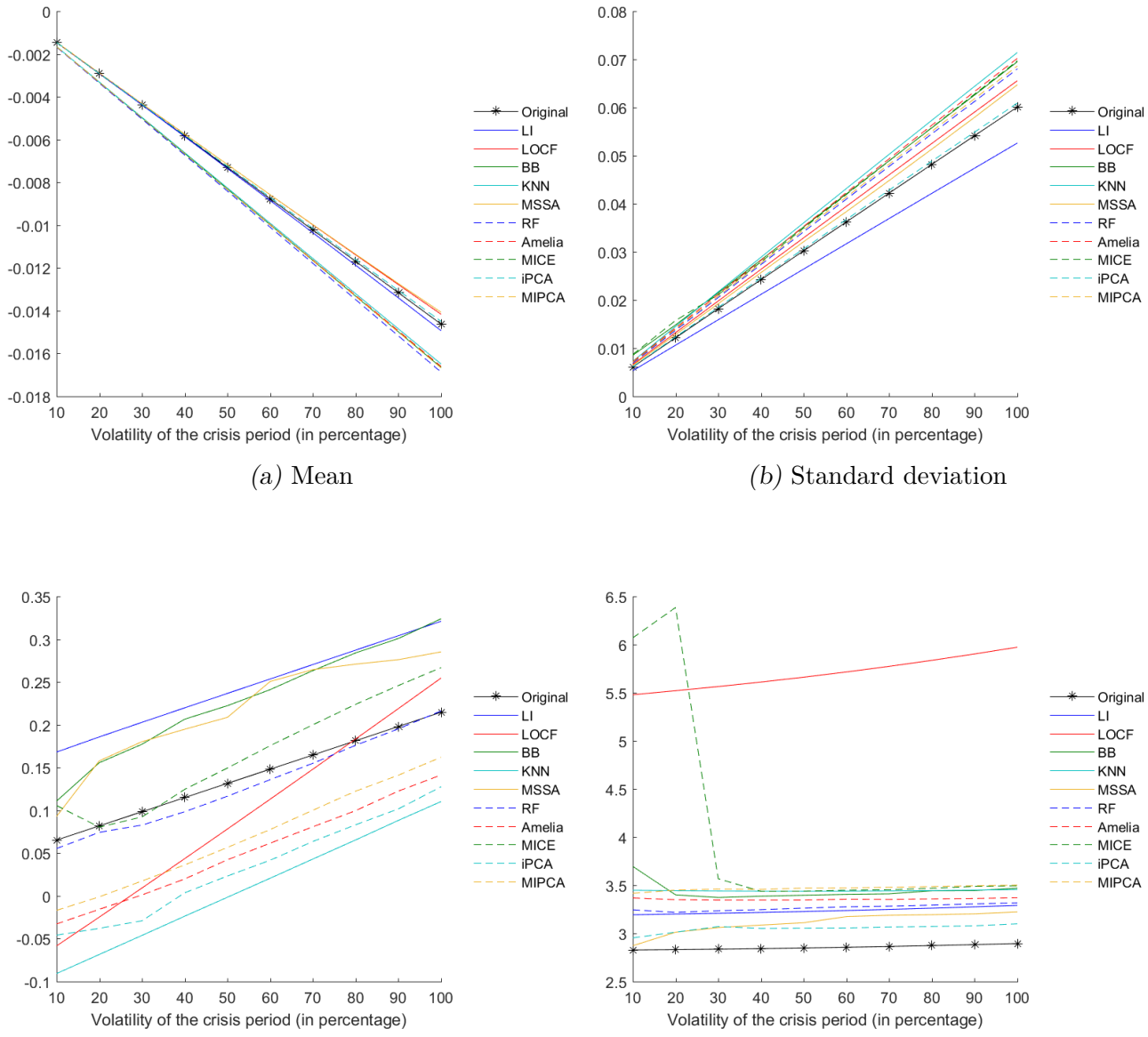
The MSSA algorithm obtains better results than the Amelia and MIPCA algorithms when the volatility of the crisis period increases. This method is even preferable, considering the statistical moments of this first period when the volatility of the crisis period exceeds 40%. Moreover, Appendix C.3 shows that the MSSA is even less stable than the IPCA presented previously.

The random forests algorithm, which worked rather well in the previous sections, gives rather satisfying and among the most stable (see Appendix C.3) results here (except for the mean), but this remains to be confirmed at a later stages.

During the crisis

Now, the aim is to highlight the average performance of the algorithms during the crisis period, with the help of the first four statistical moments. These results are presented in Figure 3.2-23 below. This second part of the sample was simulated using increasing annualized volatility (represented by the x-axis).

Fig. 3.2-23: Average of the first four statistical moments of the returns of the imputed data based on a matrix containing MCAR data for the second period (crisis period) according to the volatility of the crisis period



(a) Mean

(b) Standard deviation

(c) Skewness

(d) Kurtosis

Contrary to Figure 3.2-22, the imputation methods follow the same trend as the original series, as the volatility of the crisis period increases. Thus, the trends are downward for the mean, upward for standard deviation and skewness, and constant for kurtosis, and the algorithms follow the same dynamics. As a reminder, the samples

used for each volatility level are from a single random draw. Thus, the trends observed here are those obtained for a particular seed. If the seed changes, so will the trends. Moreover, if the experiment had been done on a large number of samples, the mean would have had a stable mean trend during the crisis period.

Moreover, Figure 3.2-23b shows that all the methods (except linear interpolation) impute missing data in a way that increases the level of volatility, in other words, the methods tend to overreact in periods of high volatility. Therefore, this is good news for the regulator because it is possible to use a method that would be conservative for all missing data scenarios.

Concerning the Brownian bridge algorithm, its results are more efficient in a crisis period than in a non-crisis period. Previously, the results were very bad for this method; here, they are quite suitable. The results from the third period should logically be as bad as in the first period because they were taking into account the volatility of the second period.

Usual methods obtain among the worst results: The linear interpolation underestimates the standard deviation (as expected) but also strongly overestimates skewness, whereas the LOCF method completely overestimates kurtosis because, in periods of high volatility, the correction return can take very large values.

The IPCA algorithm, which had good results in the first period (see Figure 3.2-22), obtains very good results here as well: The algorithm obtains means, standard deviations, and kurtosis of this period that follow very closely the evolution of those of the original data. Moreover, this method is one of the most stable in the estimation of these statistical moments (see Appendix C.3). Nevertheless, the IPCA tends to underestimate the skewness of this period.

Concerning the MIPCA algorithm, it remains less efficient than the IPCA, except for skewness. In addition, the performance of MIPCA is still close to that of Amelia. Even if these results are very stable from one missingness scenario to another (see Appendix C.3), these methods are unable to correctly manage heteroskedasticity, as they make no difference between periods of low and high volatility.

Also, the MSSA algorithm still manages to keep mean, standard deviation, and kurtosis relatively close to those results of the original crisis period, but not the skewness (as the IPCA).

The statistical moments obtained from the MICE imputations are not particularly satisfactory, and when attention is focused on kurtosis, they are even totally unstable, according to the volatility of the crisis period, especially with a crisis period volatility between 20% and 30%. As already mentioned, the reason could be the sample.

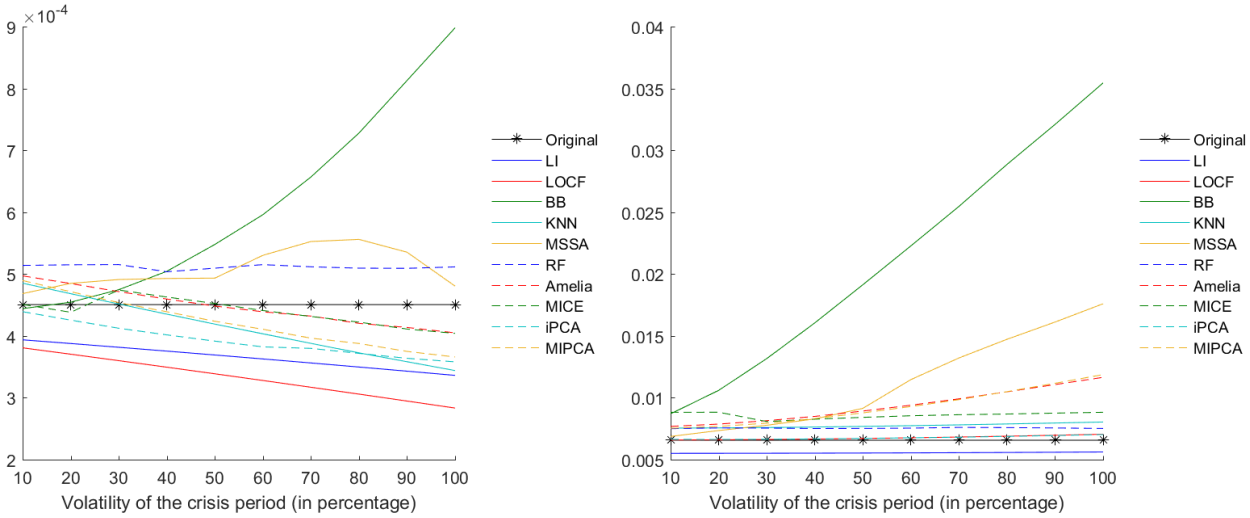
Finally, although the random forests algorithm is the only one to correctly conserve the skewness of the crisis period, this algorithm tends to slightly overestimate its volatility and kurtosis and underestimate its mean more than the other methods.

Nevertheless, the results obtained are very stable, even when the volatility of the period becomes important.

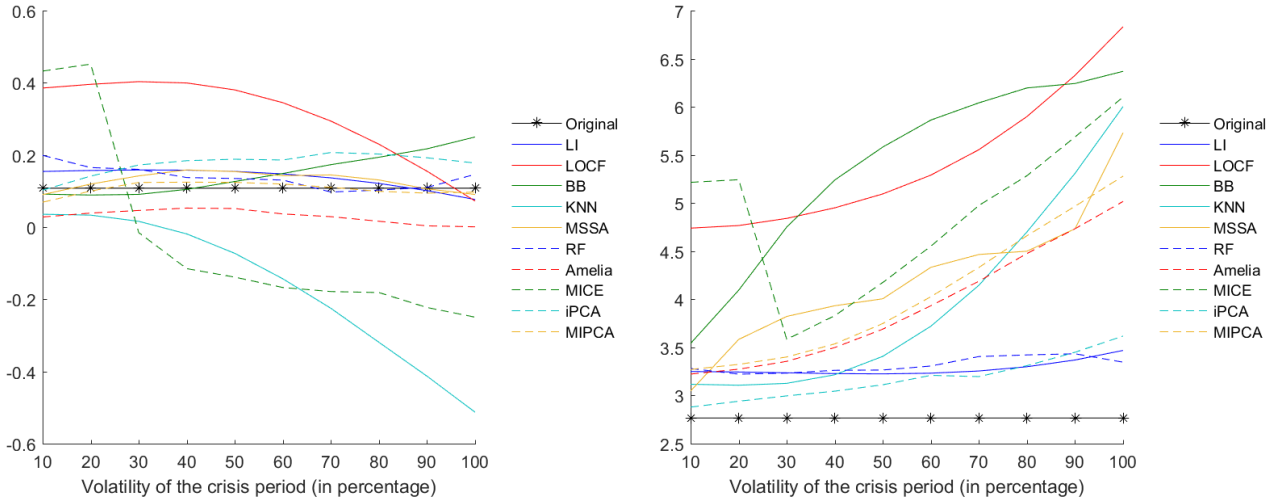
After the crisis

Finally, Figure 3.2-24 represents the first four statistical moments of the last 87 days of the series, corresponding to the post-crisis period.

Fig. 3.2-24: Average of the first four statistical moments of the returns of the imputed data based on a matrix containing MCAR data, for the third period (calm period) according to the volatility of the crisis period



(a) Mean *(b)* Standard deviation



(c) Skewness *(d)* Kurtosis

The results obtained for standard deviation and kurtosis in the latter period are comparable to those observed in the first period (see Figure 3.2-22) since the true standard deviation and kurtosis were close to those of the first period.

Over this last period, the Brownian bridge is very far from the original period, as expected.

Random forests obtain statistical moments close to those of the original period and follow the same evolution, which makes it a very satisfactory method to impute this heteroskedastic series.

The MICE algorithm gives satisfactory means and standard deviations, but skewness and kurtosis are completely unstable depending on the volatility of the crisis period. These instability seem to be sampling errors. Nevertheless, this method appears to be much more efficient when imputing this sample with heteroskedasticity than when there is constant volatility.

Here, the IPCA algorithm, which was close to the original series over the first two periods, stays relatively close to the moments of the original period. Thus, in this example, this method makes it possible to take into account the heteroskedasticity of the series in a satisfactory way. By contrast, the MIPCA algorithm offers a slight advantage over the IPCA algorithm in terms of mean and skewness. Finally, the results of the Amelia algorithm remain close to those of MIPCA but are slightly closer to those of the original series.

The standard deviation and kurtosis from the MSSA algorithm tend to increase sharply at the end of the period when volatility increases during a crisis, which makes it less efficient than linear interpolation.

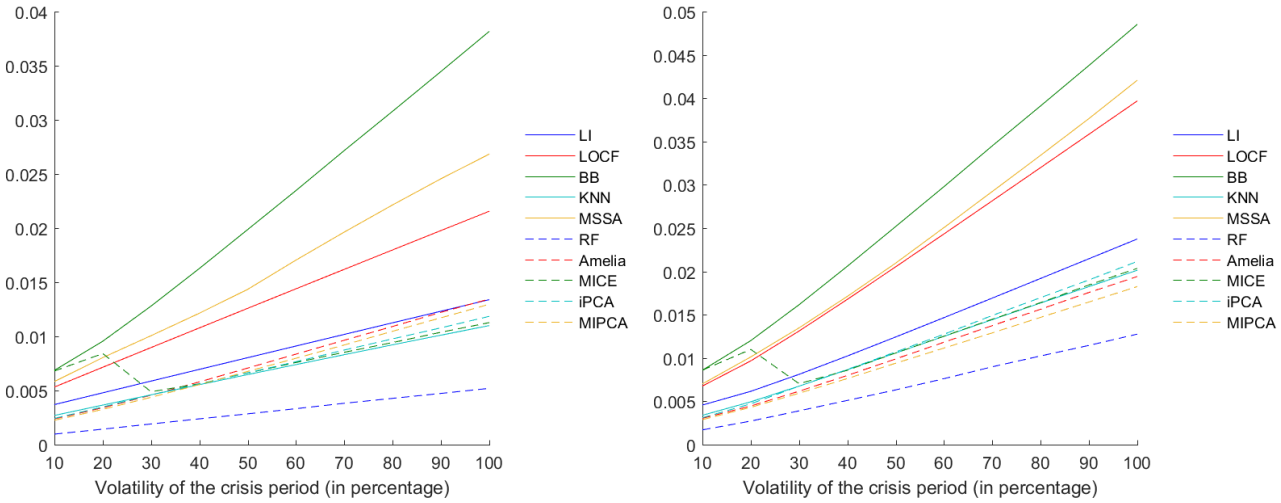
If the algorithms were difficult to compare with the results of Appendix C.2, it is more obvious here to understand that random forests and IPCA algorithms are among the methods that offer the most satisfactory results for properly preserving the moments of the original series. The fact that random forests can adapt to these regime changes is not surprising given that it is designed to take into account the complex relationships in the data. However, the performance of the IPCA, compared with MIPCA and Amelia, remains difficult to explain. Again, these results are illustrative, as they are based on a single sample of data. But the fact that the MICE results are better than those in the previous sections seems to reveal sampling errors.

Historical data samples with non-constant volatility will be used further in this chapter and will allow confirming or rejecting these results.

Proximity metrics

Figure 3.2-25 represents the average of the MAE and RMSE between the returns of the original series and the returns of the imputed data over the whole period. These results correspond to the average of the proximity measures among the 100 scenarios (each containing 30% of missing data) according to the volatility of the crisis period. Moreover, Appendix C.5 shows the proximity measures obtained for each period.

Fig. 3.2-25: Average MAE and RMSE between the original returns and the returns of the imputed data based on a matrix containing MCAR data, according to the volatility of the crisis period



(a) MAE *(b)* RMSE

Thus, over the entire period, we can see the method that minimizes the gaps with the original series the most is the random forests method , with an MAE and an RMSE

approximately equal to 0.5% and 1.2%, respectively, when the annualized volatility of the crisis period is 100%. Appendix C.5 shows the random forests algorithm allows minimizing these proximity measures, regardless of the period (crisis period or not) and the volatility of the crisis period. Moreover, the increase in proximity measures is due to the crisis period, because their evolution is constant during calm periods.

The K-NN, MICE, IPCA, MIPCA, and Amelia methods obtain MAE and RMSE very close to each other and higher than those of random forests. The MICE algorithm is close to the other methods when the volatility of the crisis period is higher than 20%. As previously, this method behaves very differently for lower volatility. While their proximity measures may be among the lowest, their results are approximately twice as high as those obtained with the random forests method. Moreover, the ranking of these methods is not the same between MAE and RMSE. The K-NN, MICE, and IPCA methods perform better in terms of MAE, while the MIPCA and Amelia methods present lower results than the other methods in terms of RMSE. This means that the K-NN, MICE, and IPCA methods tend to impute some missing data with much larger deviations than the MIPCA and Amelia methods; in other words, some imputations are very far from the original series.

These results are due to a better performance of the MIPCA and Amelia algorithms during the crisis period. The results obtained for the calm periods (first and last periods) are completely comparable, but those of the second period are quite different (Appendix C.5). While the K-NN, MICE, and IPCA methods perform better than MIPCA and Amelia in non-crisis periods, this is not the case in crisis periods. Moreover, this observation is valid for the MAE but also the RMSE. The evolution of the two proximity measures of the K-NN, MICE, and IPCA methods is constant in non-crisis periods, while those of MIPCA and Amelia tend to increase as the volatility of the crisis period becomes higher. However, all these methods increase their MAE and RMSE with the level of volatility of the crisis period, particularly the K-NN, MICE, and IPCA methods.

Thus, the K-NN, MICE and IPCA methods are more efficient when the analysis is done on the entire sample (see Figure 3.2-25), which is because the crisis period represents only one third of the sample. It is important to keep in mind that the results presented in Figure 3.2-25 are average results across the entire sample and that the results would have been different if the crisis period represented two-thirds of the sample or even with another sample. Finally, among all this group of methods, the K-NN, MICE, and IPCA algorithms perform better in calm periods than in crisis periods, whereas the opposite is true for the MIPCA and Amelia algorithms. By contrast, none of these methods is preferable for all the periods.

In addition, MIPCA and Amelia appear to be the two most stable methods in the estimations of these proximity measures from one missingness scenario to another, regardless of the volatility of the crisis period (see Appendix C.6).

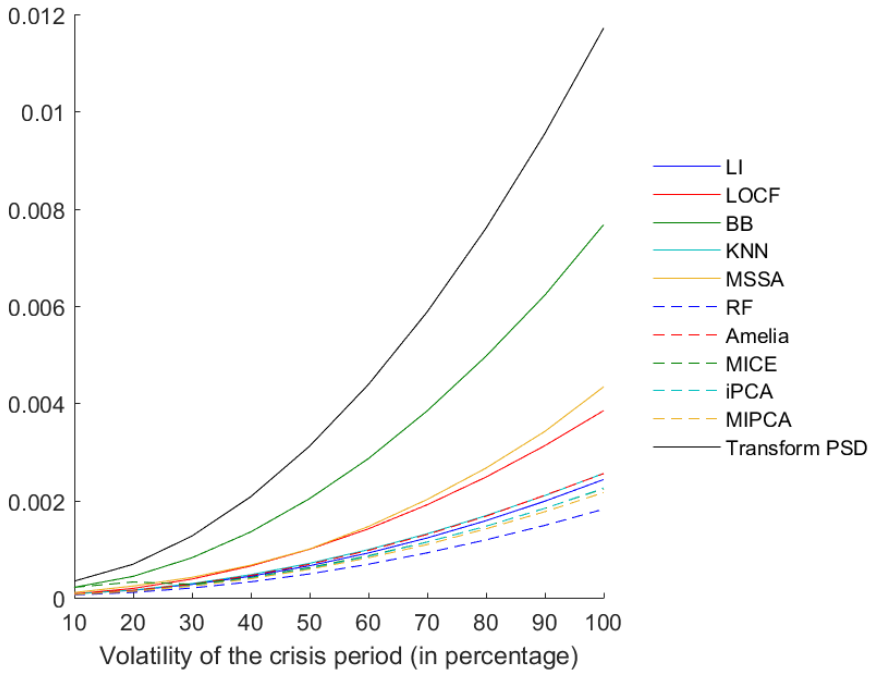
Nevertheless, the linear interpolation method obtains proximity measures that are a little less satisfying but still relatively close to this group of methods (MIPCA, Amelia, K-NN, MICE, and IPCA). It even obtains a MAE similar to the Amelia algorithm, when the volatility of the crisis period is at its maximum. This reflects a decrease in the imputation quality of these methods.

Finally, the Brownian bridge, MSSA, and LOCF methods obtain proximity measures reflecting a growing deviation from the original series as the crisis period is subject to high volatility. Appendix C.5 reveals that the Brownian bridge method moves away from the original series for each of the three periods, whereas the proximity measures of the MSSA and LOCF methods only increase sharply during the crisis periods (they remain constant during the non-crisis periods).

Covariance matrices comparison

Proximities in terms of the covariance matrix (according to the Frobenius norm) are represented in Figure 3.2-26. As before, this corresponds to the average among the 100 scenarios (each containing 30% of MCAR data) for different levels of volatility of the crisis period. Since only the first column of the sample contains MCAR data, it follows the same logic as in Section 3.2.1, where the differences in the covariance matrices concern only one part of the matrix and not the whole matrix.

Fig. 3.2-26: Average covariance matrix differences, according to the Frobenius norm, based on original returns and the imputed returns from a matrix containing MCAR data on all the samples, according to the volatility of the crisis period



Thus, the period of crisis with increasing volatility in the series tends to increase the differences between the original covariance matrix and that of the imputed data. The gaps with constant volatility over the period are at least 20 times smaller than when there is a crisis period with an annualized volatility of 100%.

Moreover, all the methods used here observe approximately the same curvature in their covariance matrix deviation evolution, except for the MICE algorithm, which, as from the beginning, reveals a slight instability (as already observed) and only joins the general trend when the volatility of the crisis period reaches 30%.

As before, the random forests method is the one that minimizes the gaps with the original covariance matrix, even though they remain much larger than when the series is homoscedastic. According to Appendix C.7, this method is also the most stable from one missingness scenario to another, which makes it very reliable for covariance estimation.

The MIPCA and IPCA algorithms obtain results very close to each other and slightly superior to those of the random forests algorithm. Moreover, here the MICE algorithm shows behavior very similar to that of the IPCA algorithm when the volatility of the

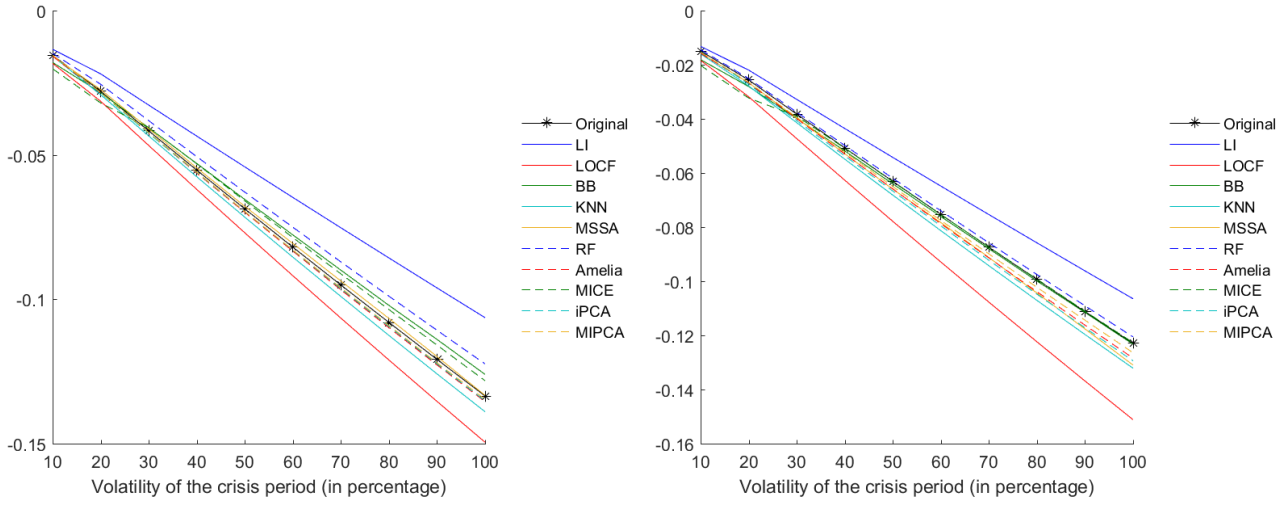
crisis period is higher than 30% (the two curves are almost merged). Thus, MICE performs much better in the presence of heteroskedasticity, although it cannot beat the random forests. Again, MICE is probably subject to sampling errors, so these results should be taken with caution.

Finally, all the other methods obtain larger covariance matrix deviations than a simple linear interpolation. Thus, the Amelia and K-NN algorithms, which obtain almost similar results, are slightly less efficient than linear interpolation at preserving the covariance matrix, and the MSSA algorithm is even less efficient than the LOCF method. The Brownian bridge method is the least performing of the methods, except for the pairwise transformed matrix ; this was expected because of the statistical moment results and the proximity measures.

Value-at-risk and expected shortfall

In terms of risk measures, the fact that a period of crisis is included in the VaR and ES calculation window would lead to a reduction in their level. Thus, the greater the volatility observed during the crisis period, the lower the level of VaR and ES would be. It is this trend that can be observed in Figure 3.2-27 and Figure 3.2-28 for a 1-day and a 10-day horizon, respectively. The risk measures decrease with the volatility of the crisis period.

Fig. 3.2-27: Average 1-day risk measures computed from a data matrix containing MCAR data on all the samples, according to the volatility of the crisis period



(a) 1-day VaR at 99%

(b) 1-day ES at 97.5%

Considering the risk measures with a 1-day horizon and as in Figure 3.2-11, llinear interpolation excessively overestimates VaR and ES, while the LOCF method underestimates them. This phenomenon is the same as observed when the data were homoscedastic for any proportions of missing data (see Section 3.2.1 and Section 3.2.2). The other methods are surrounded by the usual ones.

The MSSA method is the most efficient for VaR but not for ES. The Brownian bridge and MICE methods are, in fact, the best-performing methods for ES (and not for VaR). These methods did not perform particularly well in previous analyses, but they appear to be the closest to the original set for 1-day risk measures.

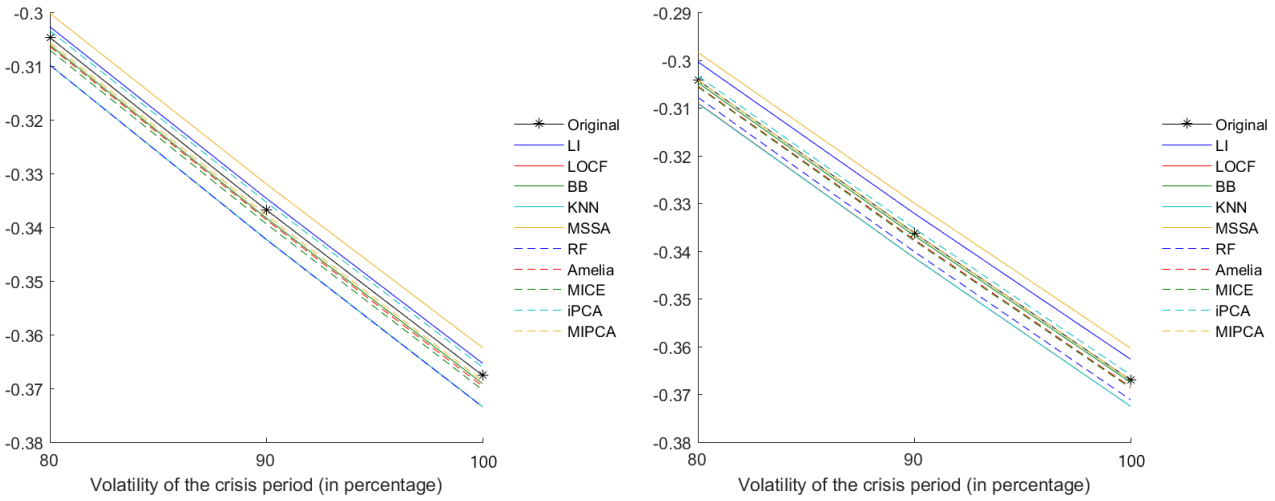
Globally, the results obtained for the IPCA, MIPCA, and Amelia algorithms remain close to each other, conservative, among the closest to the original series, and also more efficient than those of K-NN.

Random forests, which were one of the imputation methods that gave the closest results to the original series in all previous analyses, do not obtain the closest risk measures here. Again, this may be due to the sample; however, it should be noted that one method may be particularly effective for some parameters of analysis but not for others.

Moreover, while MICE is the most conservative method for a crisis period volatility of 10% or 20%, it obtains similar behavior from other completion methods above this threshold. It is likely that, when the crisis period volatility is below 30%, the MICE algorithm uses among its potential donors one (or more) extreme value, which would recur in the imputation and would be the cause of a distortion of the distribution. This would explain the poor results of MICE for this sample and, in particular, the risk measures that are even more conservative than LOCF.

In the case of 10-day risk measures, the VaR and ES decrease very sharply as the volatility of the crisis period increases, which makes the graphs difficult to analyze (see Appendix C.8). For this reason, Figure 3.2-28 (below) highlights only the results for the three series with the most violent crisis periods, as the results with constant volatility have already been presented in Section 3.2.1.

Fig. 3.2-28: Average 10-day risk measures computed from a data matrix containing MCAR data on all the samples, according to the volatility of the crisis period



(a) 10-day VaR at 99% *(b)* 10-day ES at 97.5%

Here the results reveal the efficiency of the MIPCA algorithm. This algorithm is indeed the closest to both VaR and ES of the original series, when the crisis period is very volatile. The MIPCA is also one of the most stable methods from one missingness scenario to another, as it has low standard deviations in Appendix C.9. The Brownian bridge method also reproduces risk measures well over a 10-day horizon (comparable to those of MIPCA).

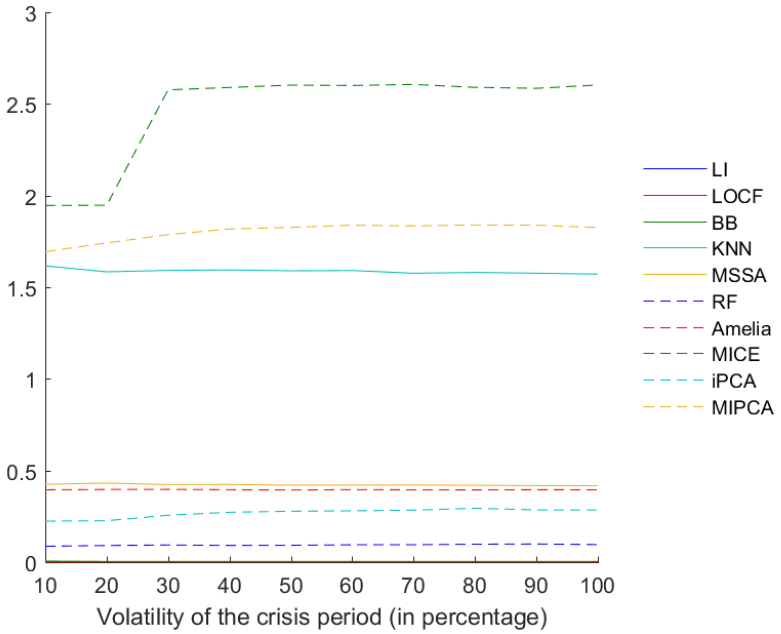
The K-NN algorithm results in the same risk measures as LOCF (the curves are overlayed), which makes it useless if the objective of the imputation is to compute the risk measures. The random forests algorithm also obtains results similar to the LOCF in the case of 10-day VaR and improves a bit in the case of 10-day ES. This algorithm was the most efficient according to a large number of criteria, but here, as for 1-day risk measures, it is not as satisfactory as expected.

Finally, the MSSA method, which was previously the most effective at reproducing VaR at a 1-day horizon (see Figure 3.2-27), overestimates 10-day risk measures even more than the linear interpolation method.

Computation time

Finally, the last analysis criterion is calculation time. Figure 3.2-29 represents the average computation time (in seconds), required by each algorithm to impute a sample of missing data.

Fig. 3.2-29: Average computation time of the imputation of MCAR data on only the first series (with two different scales) according to the volatility of the crisis period



Contrary to Section 3.2.1, the calculation times are constant, regardless of the level of volatility of the crisis period. Computation times are not affected by the violence of crises, except for the MICE algorithm, which sees its computation time increase by approximately 25% when the volatility of the crisis period exceeds 20%. This reflects a greater number of iterations to impute missing data beyond 20% volatility, which would be at the root of better imputation quality, given the results.

As in Section 3.2.1: the MICE, MIPCA, and K-NN algorithms are among the algorithms with the longest computation times (more than 1.5 seconds). The MSSA and Amelia algorithms have a similar computation time (approximately 0.5 seconds), and the random forests algorithm is the fastest of the sophisticated algorithms.

This section has been dedicated to the impact of a heteroskedastic sample on the different imputation methods used. Non-constant volatility is a well-known phenomenon in financial series and, according to the results of this section, can influence the choice of the imputation technique to use. Once again, the results obtained in this section remain debatable since they are specific to the sample used. However, they clearly illustrate the problems that an expert may encounter when selecting an imputation method.

One of the major differences observed in this section is the performance of the MICE algorithm. This algorithm was among the least successful imputation methods; nevertheless, the variable volatility of the series reveals this algorithm is as efficient as other sophisticated methods. The results of this section, therefore, challenge those of the previous two sections, where the method tended to deviate significantly from the original series. The algorithm probably had to use an extreme value as a potential donor, which distorted the series. The MICE algorithm finally manages to do better than the usual methods; however, it is still not a serious candidate compared with other methods such as random forests.

Moreover, the presence of heteroskedasticity in the series also gives an advantage to IPCA over MIPCA and Amelia. The latter methods use bootstrapping and treat all the observations in the series in the same way without distinguishing between periods of high or low volatility.

As in the previous sections, the random forests algorithm appears to be the most efficient when measured against many criteria. It reproduces the statistical moments of the original series and remains the closest model in terms of proximity measures but also in terms of the covariance matrix. Nevertheless, other methods can reproduce the risk measures of the original series more faithfully than the random forests.

By contrast, the MSSA and the Brownian bridge techniques are not as efficient and often do not reach better results than a simple linear interpolation.

The results of this particular comparative analysis are based on this sample. However, the sample has been constructed in such a way that two-thirds of it is similar to the one used in Section 3.2.1 and Section 3.2.2, thus making a comparison possible. However, this sample remains non-stationary, which is why it would be relevant for future researchers to redo the experiment on a sample simulated with a GARCH process [37] (for example) to compare the results.

3.2.4 Impact of jumps

Financial series are also characterized by the presence of jumps. Financial markets are sensitive to a large number of parameters that can cause high returns for many stocks at the same time. Jumps in a financial time series can occur for many reasons: Tender offers may result in an upward jump (arbitrage opportunity); official announcements by the Central Bank (e.g., since the 2007 crisis, Central Bank announcements about key interest rates have been the cause of numerous downward jumps that may have had repercussions on all interest rates); social media, such as the tweets published by former U.S. President Donald Trump, also generated jumps on financial series; or simply the payment of dividends to shareholders when the series used is the total return.

This section aims to show the effectiveness of imputation techniques for MCAR data when the series contains jumps.

Simulated sample with jumps

In 1976, Merton [154] proposed integrating these jumps into the Black and Scholes [35] processes. In the Merton jump-diffusion model, the jump times and the jump amplitude are random and based on a compound Poisson process. Hence, let $\{Q_i\}_{i>1}$ be a sequence of independent and identically distributed random variables and $\{N_t\}_{t\geq 0}$ be a Poisson process with intensity λ (which corresponds to the expected number of jumps). The compound Poisson process $\{J_t\}_{t\geq 0}$ is defined as follows:

$$J_t = \sum_{i=1}^{N_t} Q_i, \tag{3.2-1}$$

with, $Q_i \sim \mathcal{N}(\mu^J, (\sigma^J)^2)$ and $J_t = \sum_{i=1}^{0} Q_i = 0$ when $N_t = 0$.

Then, the stock price under the Merton jump-diffusion model is as follows:

$$S_t = S_0 \exp\left((\mu - \frac{\sigma^2}{2})t + \sigma W_t + \sum_{i=1}^{N_t} Q_i\right) \tag{3.2-2}$$

The Merton jump-diffusion model uses the same log-normal dynamics presented in Section 3.1.4 through Equation 3.1-3, including $\left\{\sum_{i=1}^{N_t} Q_i\right\}_{t\geq 0}$ corresponding to a compound Poisson process with normally distributed jumps $\mathcal{N}(\mu^J, (\sigma^J)^2)$ and intensity λ. In the case of this simulated data, the intensity parameter corresponds to the number of jumps per year - in other words, the daily probability that a jump occurs corresponds to λdt.

To be consistent with the previous simulation, a sample size of 10 times series of 261 observations are simulated with a drift $\mu = 0$ and an annualized volatility $\sigma = 10\%$. For each column, the jumps occur at the same dates and with the same amplitude before being distorted by the correlation matrix. Each series is correlated in the same way as presented in Section 3.1.1. This illustrates the market jumps (i.e., events that make the whole market react) that can occur on financial time series. The idea here is the same as before: The sample has been built in series relatively close to each other and, therefore, with jumps of more or less similar amplitudes to optimize the performance of the algorithms.

Regarding the parameters of the jump process, it is possible to find studies in the literature that attempt to estimate the parameters of the Merton jump-diffusion model on empirical data. Bates is one of the pioneers in this field. In 1991, he published an article [26] in which he sought to estimate the implied mean and standard deviation of jumps based on the transactions prices of S&P 500 futures options between 1985 and 1987.

In 2019, Lau, Goh and Lai [138] published a study to estimate these same parameters on more recent data and from several indices. They used data from the Dow Jones

(GJI), the NASDAQ 100, the FTSE 100 Index, S&P 500, and the NYSE ARCA Oil & Gas Index over a period from January 2005 to December 2014. The parameters of the Merton jump-diffusion model are estimated from the Gibbs sampler, which consists of using Bayesian inference from the MCMC with the Metropolis-Hasting model (an iterating process that updates initialized values toward the distribution with an accept-reject method). After dividing their sample into two periods, they obtained the following results:

Tab. 3.2-9: Jump parameter of the Merton jump-diffusion model for different indices between 2005 and 2010 (Source: Lau, Goh and Lai, 2019 [138])

Parameters	**DJI**	**S&P 500**	**NASDAQ 100**	**FTSE 100**	**OilGas**
μ^J	-0.00407	-0.00519	-0.00388	-0.00475	-0.01131
σ^J	0.04524	0.04711	0.05277	0.04554	0.06655

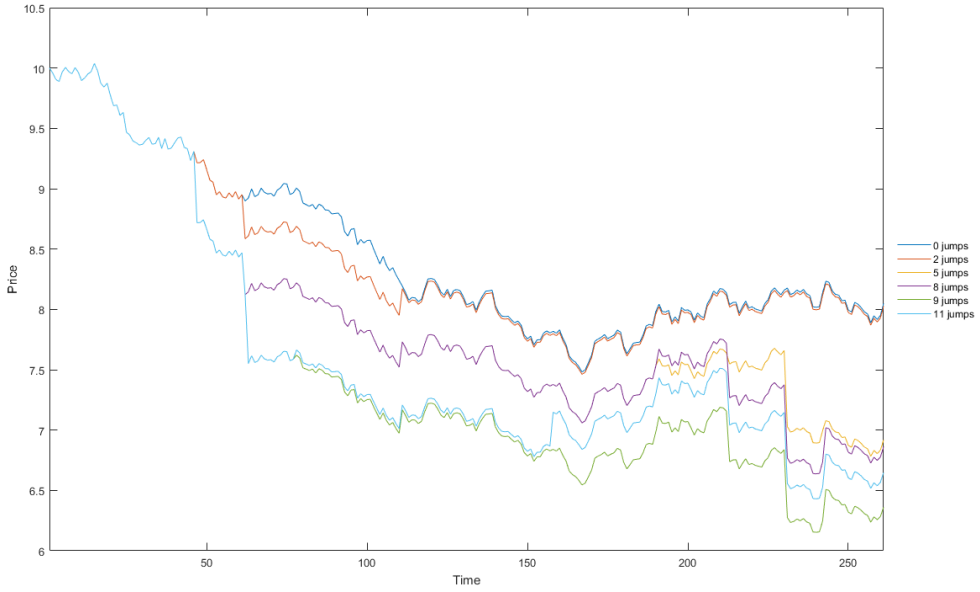
Tab. 3.2-10: Jump parameter of the Merton jump-diffusion model for different indices between 2010 and 2015 (Source: Lau, Goh and Lai, 2019 [138])

Parameters	**DJI**	**S&P 500**	**NASDAQ 100**	**FTSE 100**	**OilGas**
μ^J	-0.00592	-0.00478	-0.00568	-0.00446	-0.00615
σ^J	0.05644	0.05294	0.05739	0.05095	0.04405

Although the first period includes the 2007 crisis, the results are on average quite similar. What differs between the two periods is the occurrence of jumps that can be up to five times higher in the first period. On average, $\mu^J = -0.005$, which means the jumps tend to be more downward than upward; and $\sigma^J = 5\%$. These are the parameters that will be used in this PhD PhD thesis to agree with the market.

In addition, Lau, Goh, and Lai [138] observe between 14 and 20 jumps in the first period, then between six and seven in the second period (OilGas sees approximately 13 jumps, but it is the only one). Rather than using these values, the number of jumps will arbitrarily chosen. The goal of this section is to test different levels of λ to see the impact of more and more jumps on the completion methods. Thus, the intensity of the Poisson λ process will take values 0, 2, 3, 4, 5, and 6, which leads to eight samples containing an increasing number of jumps from one sample to another, as follows: 0 (which correspond to the same sample as in Section 3.2.1), 2, 5, 8, 9 and 11 jumps. This allows analyzing whether the completion methods are negatively impacted by the appearance of one or several new jumps in the series. The first columns of each simulated data matrices, with the different levels of λ, are shown in Figure 3.2-30.

Fig. 3.2-30: First time series (among 10) of a simulated sample with 261 observations according to the number of jumps applied



An increasing number of jumps is applied to the same matrix of simulated data. Moreover, to make the samples comparable between them, the jumps of a matrix based on a high λ are also present in the matrices based on a lower λ.

Thus, it is possible to compare the behavior of different algorithms facing series that contain an increasing proportion of jumps. The first sample, simulated with a λ parameter equal to zero, is the same as in Section 3.2.1 and Section 3.2.2 (and presented in Section 3.1.1) and will be used as a benchmark here.

As in Section 3.2.1, missing data are drawn completely randomly from the first column of the sample (as in Figure 3.2-1) from a uniformly distributed random variable, with a proportion of missing data set at 30%, which corresponds to 51% missing returns on average for the algorithms applied to the returns (see Table 3.2-1). Finally, the results obtained in this section are based on 100 missingness scenarios containing 30% of MCAR data. In addition and as before, the first step is to apply the MCAR tests to evaluate whether they conclude that the data are MCAR when the series contains jumps.

Finally, each comparison tools is computed as defined in Figure 3.1-2, and the templates of (almost) all the graphs presented in this section have already been presented and detailed (i.e., what they represent and how they were obtained) in Section 3.1.4.

MCAR tests

Thus, Little's test [142] and Jamshidian and Jalal's test [123] were applied to the price returns matrices containing different numbers of jumps to see whether these jumps have an impact on the efficiency of these tests. First of all, Appendix D.1 represents the number of tests (among the 100) that are calculable (run without error). The same observation as before can be made: Little's test [142] has more difficulty performing without error when the data contain jumps, whereas Jamshidian and Jalal's test [123] is always calculable here.

Based on the calculable tests, the probabilities not to reject the null hypothesis that the data are MCAR, depending on the volatility of the crisis period, are presented in Table 3.2-11.

Tab. 3.2-11: Confidence level (probability of not rejecting H_0 when H_0 is true) for both MCAR tests applied to price return matrices containing jumps and MCAR on the first column of the matrix, for a 5% significance level

	Number of jumps in the series					
	0	**2**	**5**	**8**	**9**	**11**
Little's test	94%	93%	95%	94%	92%	94%
J&J's test	97%	95%	95%	96%	95%	95%

Both tests tend to conclude that the data are MCAR. Little's test [142] obtains confidence levels between 92% and 95%; for Jamshidian and Jalal's test [123], these levels are between 95% and 97%. The tests do not appear to be impacted by jumps in the series. The results obtained are almost comparable for each sample used.

Overall, the MCAR tests are not sensitive to the presence of jumps in the series. Once again, the results presented here are from a single initial sample, but it would be appropriate to repeat the experiment on a large number of samples to ensure their robustness.

Jump imputation

Before considering the performance of algorithms when exposed to a series containing jumps, it is natural to ask how these methods impute the jump itself. The data to be imputed may be a jump and, in this case, it is interesting to see whether the completion methods can recognize it as a jump and to impute it as such (or not).

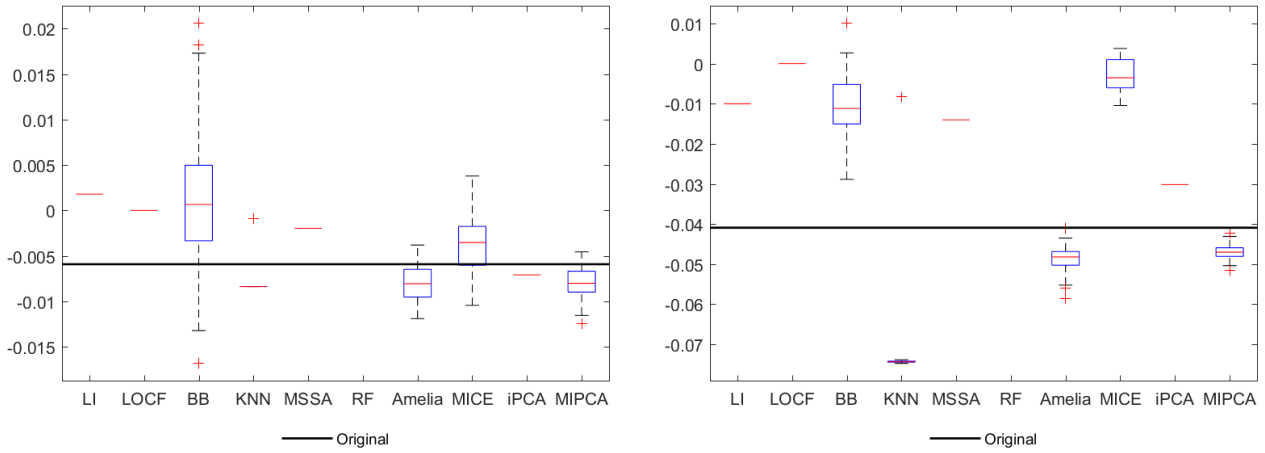
Thus, among the 100 scenarios used in this study, there is at least one scenario where a jump is missing. The first jump (for a λ set at 2) is observed between the 61^{th} and the 62^{th} stock price, which corresponds to the 61^{th} return of the series. Moreover,

this jump is present in all the series where the λ parameter is greater than or equal to 2 and has the same jump amplitude because the same seed is used. Thus, the 61^{th} return of all the series where $\lambda \geq 2$ corresponds to a jump.

This 61^{th} return is part of the missing data for the fifth missingness scenario. The goal is, therefore, to look at the imputations obtained for this 61^{th} return of the series from the second missingness scenario for each λ parameter used.

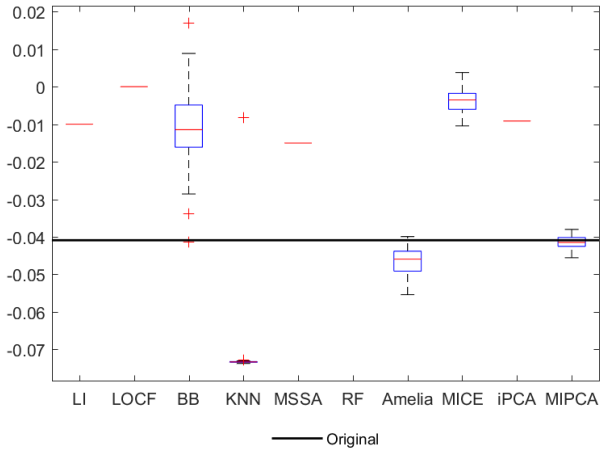
Figure 3.2-31 represents the distribution among the 100 imputations for this 61^{th} imputed returns (from the fifth missing data scenario) for each completion method and each λ parameter used.

Fig. 3.2-31: Distribution of the 61^{th} return (which corresponds to a jump) obtained by completion methods, according to the number of jumps in the series

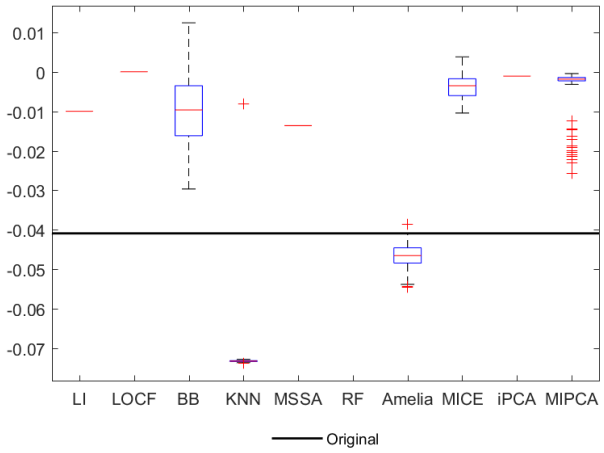


(a) Series with no jump

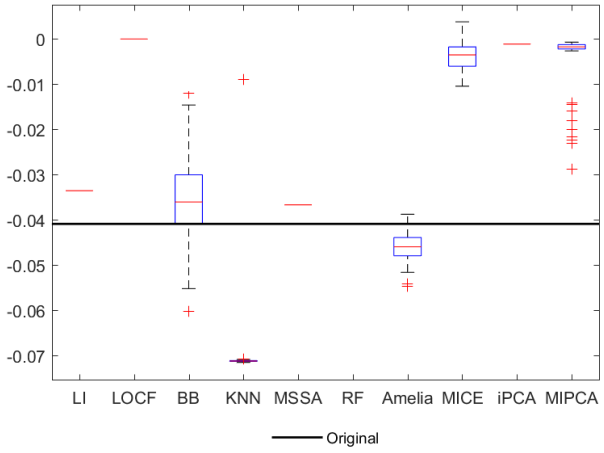
(b) Series with two jumps



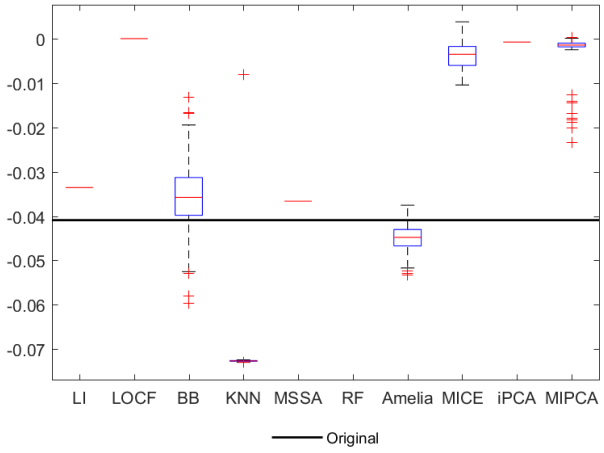
(c) Series with five jumps



(d) Series with height jumps



(e) Series with nine jumps



(f) Series with 11 jumps

When the series does not contain any jump (see Figure 3.2-31a), the 61^{th} price return is a return like any other in the series, which is why this return from the original data is different from the other plots. This original return is perfectly estimated by the random forests method. No matter the number of jumps in the series, this specific jump is always perfectly imputed by the random forests method. This original return is also very close to the imputations made by IPCA and within the confidence interval of the Amelia and MIPCA algorithms. Finally, the Brownian bridge method has, as usual, the imputations with the widest distribution, which can deviate up to 2.5% from their original value.

Now that the distributions and performance of each method have been put forward

for this 61^{th} return, which was just an ordinary return, it is time to see the performance of the algorithms if this same return becomes a jump (see Figure 3.2-31b, Figure 3.2-31c, Figure 3.2-31d, 3.2-31e and Figure 3.2-31f).

The random forests algorithm manages to perfectly impute this jump of any other series containing between two and 11 jumps for this missingness scenario. Since this jump occurs simultaneously and with almost the same amplitude (slightly distorted after making the columns correlated) for all the columns of the data matrix, the algorithm manages to predict exactly the value of this missing jump.

The original return corresponding to this jump is still within the confidence interval of Amelia's algorithm, meaning that among the 100 samples imputed by Amelia, at least one is close to the original return. Furthermore, this method tends to impute this return with an even more negative return, since the median of the distribution is lower than the original return.

The IPCA and MIPCA algorithms tend to impute this jump with a value closer and closer to zero as the series contains more jumps. Both methods use on average the same number of principal components (which is 4 as presented in Appendix D.3). If the MIPCA method imputes this jump correctly when the series contains a total of five jumps, it tends to impute it slightly more negatively when the series contains only two jumps but imputes it as a regular return when the series contains eight jumps or more (imputations are very close to zero in Figure 3.2-31d, Figure 3.2-31e and Figure 3.2-31f).

If IPCA and MIPCA can recognize the presence of jumps when their proportion is low, this is not the case here for MICE, which imputes them like any other value (return close to zero). However, the more jumps there are in the sample, the more the MICE algorithm should be able to use them as potential donors at that moment. These results are from a single scenario and may not be representative of the outcome on average.

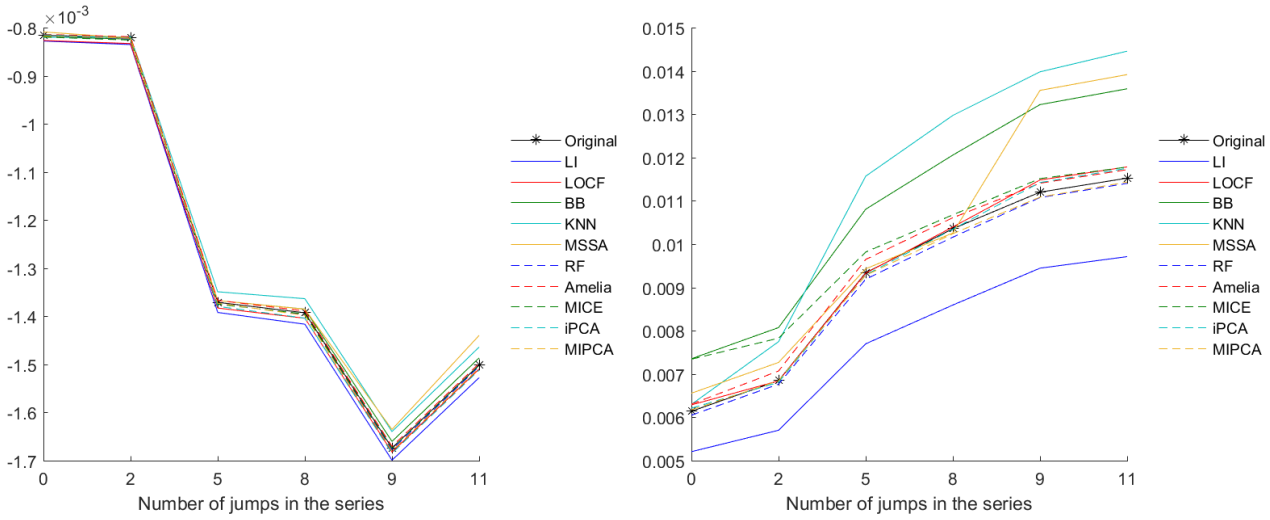
Despite its latitudinal functioning, the K-NN method fails to reproduce the level of the original series at all – in fact, it creates an even larger jump for each of the series used here. This is because the jumps in the other columns are even more violent (because of the process of correlating the columns), which pushes down the imputation.

Not surprisingly, the worst method is the LOCF, which will always give a zero return. By contrast, it will lead to a jump at the next point in time (if only the jump is missing), which may be even more negative than this one. This method will not reproduce the jump at the right time, but there will be one the next time. However, linear interpolation is no longer efficient at all. Moreover, the linear interpolation gives even more negative imputations, which appear to be more satisfactory, between the series containing eight to nine jumps (compare Figure 3.2-31d and 3.2-31e) because there is a jump corresponding to the 62^{th} price return for the series containing nine jumps (or more) - i.e., two successive jumps.

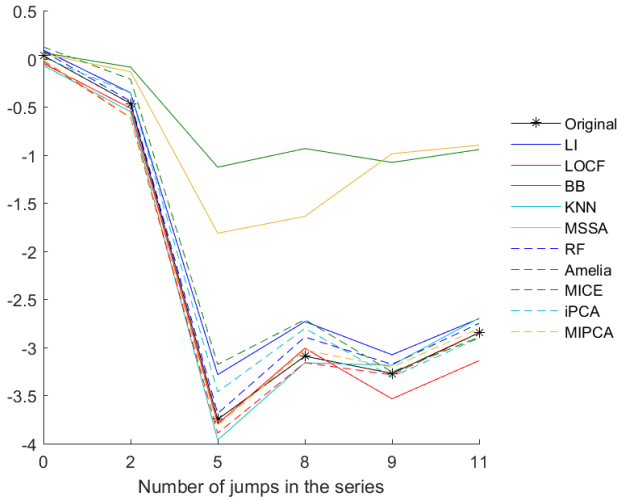
Statistical moments

After this focus on imputing a missing jump for a specific scenario, it is interesting to see how these algorithms behave on the whole series containing jumps. After imputing the 30% missing data from the sample, the average of the first four moments of price returns was compared with the original series. The results, for these first four statistical moments, are presented in Figure 3.2-32, representing the average of these moments obtained among the 100 missingness scenarios according to the intensity of the jump process.

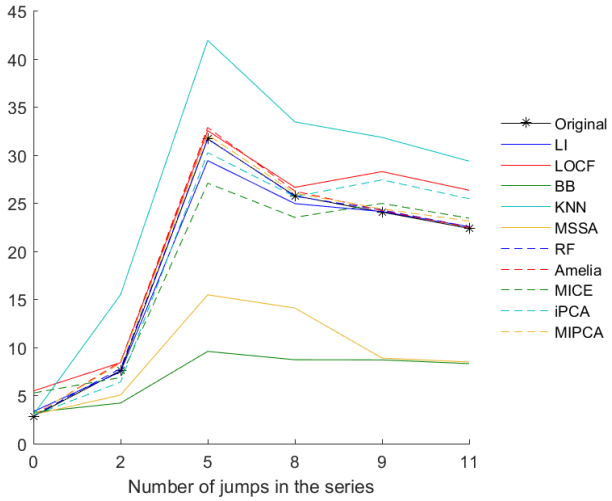
Fig. 3.2-32: Average of the first four statistical moments of the returns of the imputed data based on a matrix containing MCAR data, according to the number of jumps in the series



(a) Mean *(b)* Standard deviation



(c) Skewness



(d) Kurtosis

Since the jumps are more likely to be downward than upward (the mean of the jump process was chosen to be negative), it would be normal for the series mean to tend to decrease with the number of jumps. Given the evolution of the data presented in Figure 3.2-30, the jumps are indeed downward, so the decreasing evolution observed in Figure 3.2-32a is logical. The presence of these jumps also implies higher volatility (see Figure 3.2-32b), as well as also a negative skewness (see Figure 3.2-32c), since the tail of the distribution is spread to the left. Finally, the kurtosis tends to be greater than 3 since these jumps generate fatter distribution tails (see Figure 3.2-32d).

Three methods appear to be suboptimal for the imputation of these samples: Brownian bridge, K-NN, and MSSA. Brownian bridge and MSSA tend to overestimate the level of skewness and underestimate the level of kurtosis. Regarding the K-NN, it is one of the methods that increase the volatility the most (like the Brownian bridge) but also the kurtosis.

These three methods are less effective than the LOCF method when the series contains jumps, which calls into question the use of more sophisticated methods. Moreover, linear interpolation tends to reduce the volatility of the imputed series, as has been observed before.

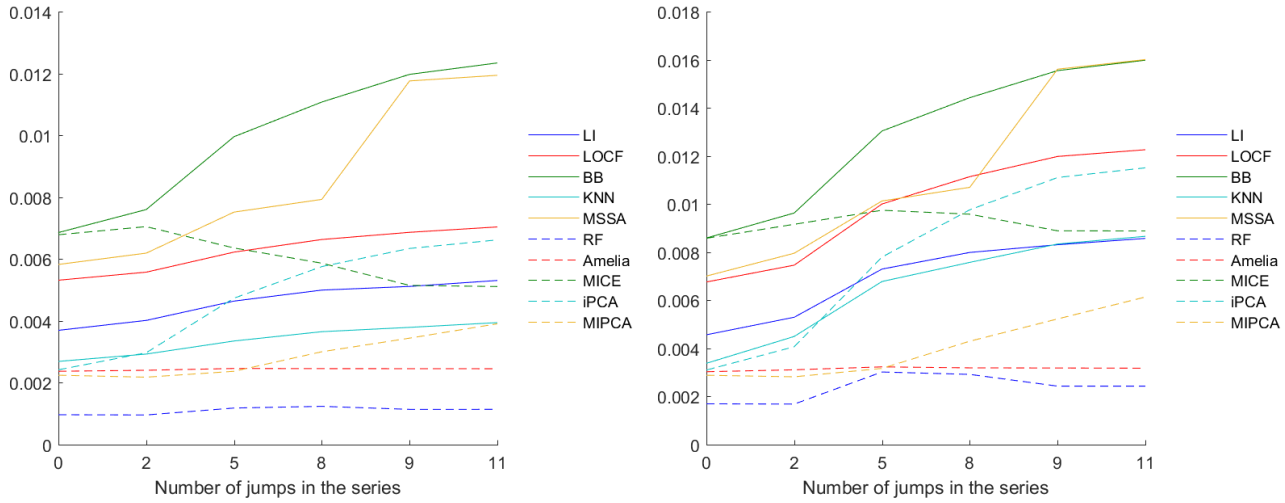
Apart from these methods, the other algorithms tend to follow, on average, the same evolution as the moments of the original series. Moreover, the MICE algorithm is more efficient in the presence of jumps than in their absence. This would not be surprising, given that missing jumps could use the data from other available jumps to be imputed. But, it is also possible that this algorithm is inefficient on the sample without jumps, as has been noted before.

The algorithms of random forests, Amelia, IPCA, and MIPCA obtain results that are relatively similar to each other and satisfactory to preserve the moments of the original series. In addition, Appendix D.2 shows very stable results for these four methods, but especially for the random forests, Amelia, and MIPCA.

Proximity metrics

Figure 3.2-33 presents the average MAE and RMSE (among the 100 MCAR scenarios), representing the proximity between the returns of the original series and the returns of the imputed series.

Fig. 3.2-33: Average MAE and RMSE between the original returns and the returns of the imputed data based on a matrix containing MCAR data, according to the number of jumps in the series



(a) MAE *(b)* RMSE

Across all the methods, the proximity measures tend to increase, on average, with the number of jumps in the series.

The method that minimizes the proximity measures is the random forests algorithm, with an average MAE of approximately 1% and an average RMSE of approximately 2%. This algorithm does not appear to be impacted by an increase in the presence of jumps. The proximity measures are the same whether there is no jump, one jump, or several jumps. Moreover, their estimations are very stable from one missingness scenario to another, according to the results of Appendix D.4. The standard deviations of these proximity measures increase with the number of jumps in the series, revealing a slight imputation instability.

The Amelia algorithm obtains averaged proximity measures slightly higher than those of random forests. In addition, the difference between MAE and RMSE is relatively small, indicating that the algorithm never deviates too far from the original series. This algorithm also has the advantage of not being impacted at all by the presence of jumps in the series, since the results obtained when the series contains no jumps are the same as those obtained when it contains one or more jumps. Moreover, Amelia gets the smallest standard deviations of these proximity measures, which means this method is very stable for the 100 missingness scenarios tested (see Appendix D.4).

Moreover, this method obtains proximity measures similar to the MIPCA algorithm when the series contains eight jumps or fewer. Beyond this threshold, the proximity measures of the MIPCA algorithm increases, and it is one of the least stable methods from one missingness scenario to another (see Appendix D.4). Regarding the IPCA algorithm, it is impacted much more by the presence of jumps in the series than its improved version. While the results between IPCA and MIPCA are the same when there are no jumps in the series, the MAE and RMSE of the IPCA method increase progressively as the jumps appear in the series while the methods use the same number of principal components on average (see Appendix D.3). The proximity measures of the IPCA are even higher than those of the linear interpolation when the series contains more than five jumps.

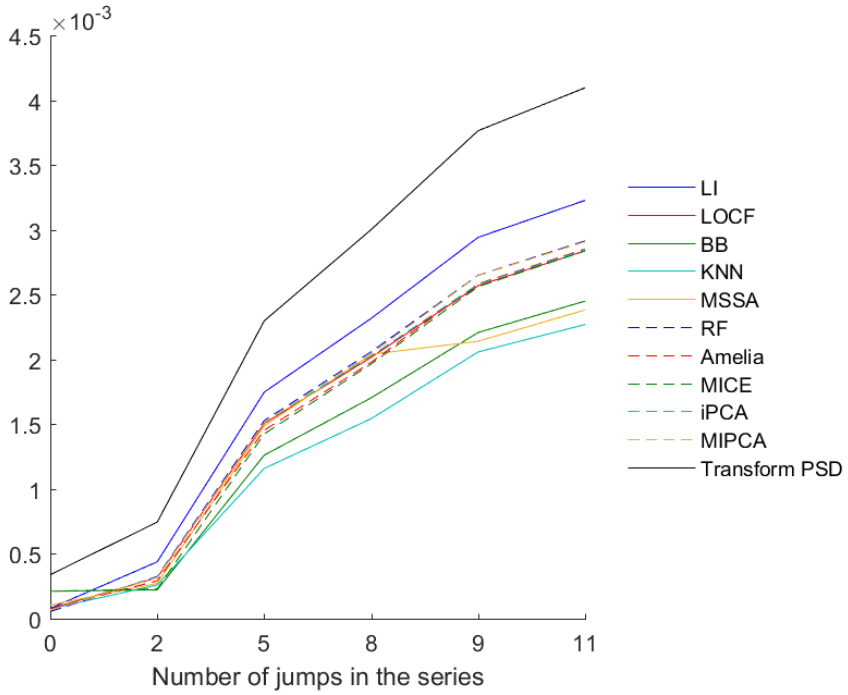
The K-NN algorithm obtains a slightly higher averaged MAE than Amelia, while the averaged RMSE levels tend to increase as the series contains jumps, leading to the same RMSE as linear interpolation. This indicates that some imputations made by this method may be far removed from the original series.

Finally, the remaining methods (i.e., Brownian bridge, MICE, MSSA, and LOCF) obtain the highest proximity measures. They were already the worst-performing methods when the series did not contain jumps (see Figure 3.2-7), and this remains the case when jumps are present. Nevertheless, the MICE algorithm is the only one that has decreasing proximity measures with the number of jumps in the series and this is because the imputed jumps use the data of other jumps, but this could also be due to a sampling effect.

Covariance matrices comparison

Figure 3.2-34 represents the differences between the covariance matrices of the original series and those from the imputed data, according to the Frobenius norm. As in Section 3.2.1, the missing data is distributed only in the first column of the data matrix, which implies that only the first column (and row) of the covariance matrix is impacted by the imputations.

Fig. 3.2-34: Average covariance matrix differences, according to the Frobenius norm, based on original returns and the imputed returns from a matrix containing MCAR data on all the samples, according to the number of jumps in the series



As for proximity measures, the impact on covariance matrices increases with the number of jumps in the series. It appears that the methods have more difficulty correctly reproducing the covariance matrix when there are jumps than when there are no jumps. Each method seems to face the same difficulties in preserving the covariance matrices, given a general trend followed by all of them.

Nevertheless, the Brownian bridge and K-NN methods appear to perform best here. However, these same methods differed significantly from the original series based on proximity measures. Moreover, while the Brownian bridge is the least efficient method when the series contains no jump, it has the smallest covariance differences.

The random forests algorithm, which has often been among the best-performing on many criteria since the beginning, obtains some of the most distorted covariance matrices. These results are close to those of the Amelia, MIPCA, and IPCA algorithms but also of the LOCF method. Nevertheless, even if their results are very stable from one missingness scenario to another (see Appendix D.5), these methods always remain closer to the original matrix than the pairwise transformed matrix.

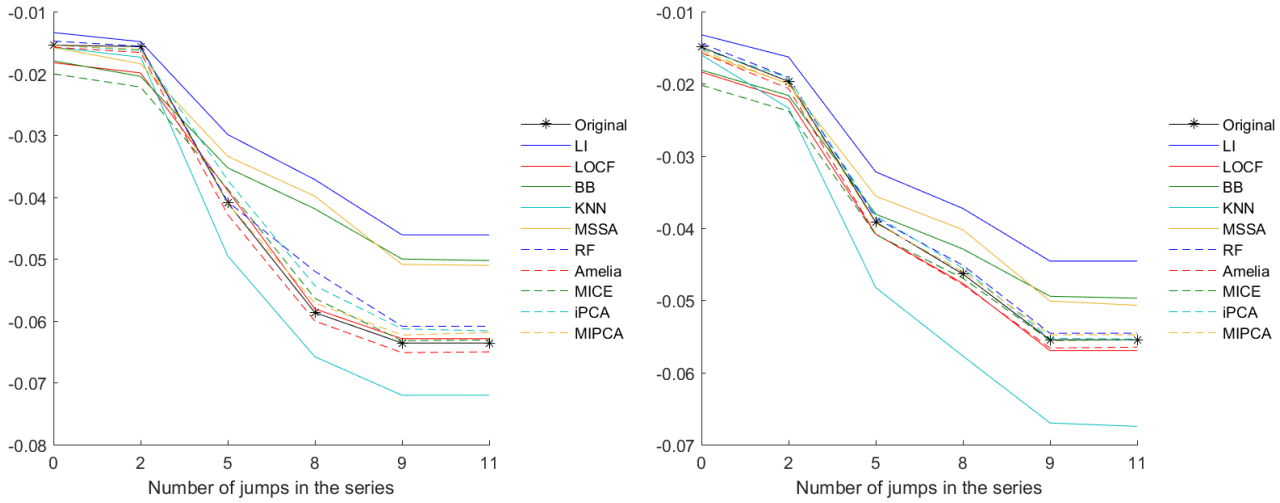
Overall, the results obtained in terms of covariance proximity highlight methods like

the Brownian bridge and K-NN, which were not at all among the best-performing in the previous results. Again, these results are based on a single sample and should be repeated on a larger number of samples to ensure the robustness of the results. But this will be the subject of future work.

Value-at-risk and expected shortfall

The performance of each model concerning the 1-day risk measures is presented in Figure 3.2-35. These figures represent the 1-day VaR and ES with a confidence level set at 99% and 97.5%, respectively, for series containing an increasing number of jumps.

Fig. 3.2-35: Average 1-day risk measures computed from a data matrix containing MCAR data on all the samples, according to the number of jumps in the series



(a) 1-day VaR at 99%

(b) 1-day ES at 97.5%

Thus, the presence of jumps in the series tends to decrease the level of risk measures. As the sample is constructed, the risk measures from one sample are the same or more negative than those of the previous sample (with fewer jumps).

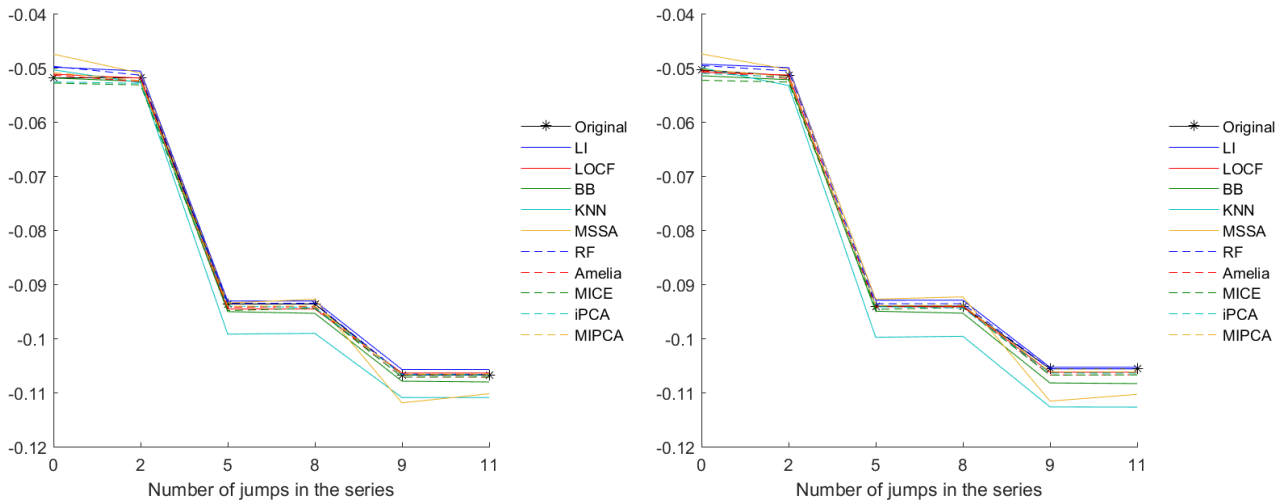
Not all of the methods used here can reproduce the VaR level, especially when the series contains more than two jumps. The Brownian bridge, MSSA, and linear interpolation methods are unable to preserve the VaR for this series; in other words, they failed to properly impute the missing jumps. In the calculation of the VaR and ES, the extreme values are emphasized, and these three methods overestimate the level of these values.

Nevertheless, the other methods follow the true VaR correctly, which reflects good management of the extreme values of the series. If the random forests algorithm tends to overestimate the level of VaR and ES, the Amelia algorithm tends to underestimate their level, which would be more acceptable to the regulator.

Moreover, the K-NN algorithm obtains the more conservative VaR and ES but becomes more and more conservative with the number of jumps. However, this method delivers the least stable results from one missingness scenario to another (see Appendix D.6)

Figure 3.2-36 represent the same VaR and ES as before, but here, with a horizon of 10 days and a confidence level of 99% and 97.5%, respectively „ for the same data containing an increasing number of jumps.

Fig. 3.2-36: Average 10-day risk measures computed from a data matrix containing MCAR data on all the samples, according to the number of jumps in the series



(a) 10-day VaR at 99%

(b) 10-day ES at 97.5%

As the horizon increases to 10 days, the risk measures tend to decrease as the number of jumps in the series increases. In contrast with the horizon set at 1 day, all the methods used here can approximate the true risk measures with deviations of less than 1%.

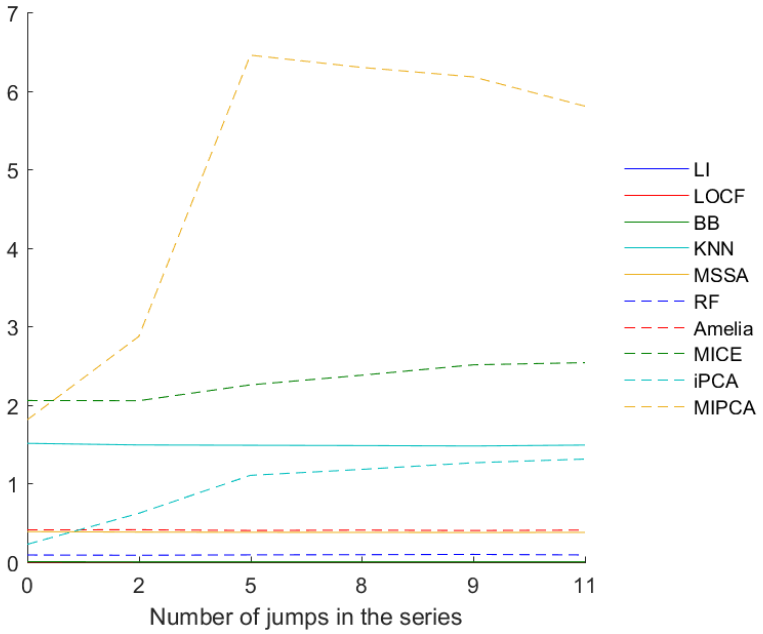
Nevertheless, similar behaviors can be observed. As before, the K-NN algorithm is the most conservative completion method and becomes even more conservative as the number of jumps in the sample increases.

The random forests algorithm tends to stay (slightly) above the level of the risk measures, while the Amelia algorithm is one of the conservative methods closest to the original VaR and ES and also the most stable, as presented in Appendix D.6.

Computation time

Finally, Figure 3.2-37 represents the average computation time (in seconds) required by each algorithm to impute the missing data of the first column of the data matrix when the series contains jumps.

Fig. 3.2-37: Average computation time of the imputation of MCAR data on only the first series (with two different scales) according to the number of jumps in the series



Computation times are constant for all the methods used, except for the IPCA and MIPCA algorithms. These two methods have a higher computation time to perform the PCA, as the jumps are important in the sample.

The MICE algorithm also has a computation time that increases slightly when the series contains more than two jumps. The other completion methods impute samples with jumps as quickly as without jumps.

This section highlights the impact of jumps on completion methods. While these anomalies may complicate and mislead imputation for some methods, they may also allow others to make progress.

Although the MICE algorithm is not the best-performing in this section, these results are more satisfactory in the presence of jumps than without jumps. Thus, the MICE algorithm appears to be more efficient at imputing a heteroskedastic series or a series with jumps than without heteroskedasticity or jumps.

Finally, the random forests and Amelia algorithms are the most efficient and sometimes appear not to be impacted by the presence of jumps in the series (this is particularly visible with the proximity measures). The MIPCA algorithm, generally comparable to Amelia, is a little less efficient here but is nevertheless among the most efficient.

These results should be put into perspective, as they are based on a single sample. Still, they are illustrative of the results that can be obtained with historical data. In future research, this same type of analysis should be repeated on a large number of samples to generalize the results.

3.3 Imputation of data: MAR on simulated Gaussian sample

The missing data included in the sample have, until now, been drawn completely at random, but as presented in Section 2.1.3, there are other mechanisms of missing data. The interest of this section lies in incorporating MAR data in line with the definition of Little and Rubin [145] or Schafer and Graham [182]. The data are removed, following a MAR mechanism, from the first column of the data matrix to compare the results with those of Section 3.2.1, where the missing data are MCAR.

Even if MCAR data represent the most common theoretical framework, it is not obvious that in a financial context the data are predominantly MCAR. The case of MAR is often present in survey data where, for example, an individual with a high salary will tend not to answer questions related to it.

In this section, three types of MAR mechanisms often observed on financial data are used:

1. missing data that depend on extreme values from another series (Section 3.3.1)
2. missing successive data in the middle of the columns (Section 3.3.2)
3. missing successive data at the end of the columns (Section 3.3.3)

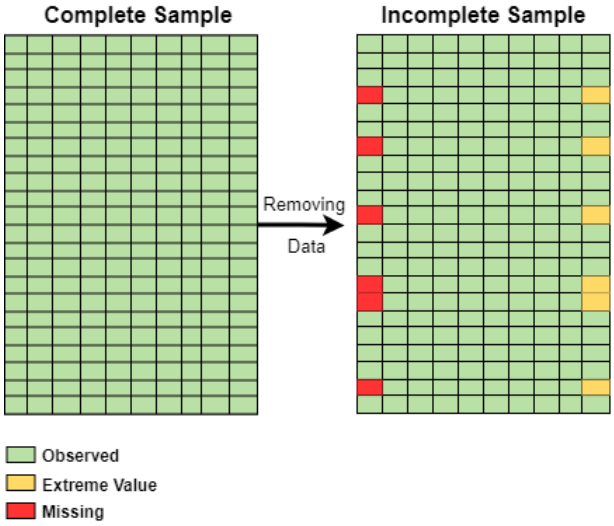
3.3.1 Impact of missing values depending on extreme values of another series

The data of this section are removed according to the extreme values of another column of that data matrix. Knowing the missing data mechanism influences the choice of which imputation method to adopt. If the missing data from one variable depend on the observed data from one or more other variables, it would make sense to use certain imputation methods rather than others. By definition, one-dimensional (unlike two-dimensional) methods cannot use the information of other variables. Hence, linear interpolation, LOCF, and Brownian bridge are not expected to provide good imputations for this kind of missingness. By contrast, the K-NN method is expected to impute missing data correctly since it uses a linear combination of the available observed data.

Missing data mechanism depending on extremes values of another series

As in previous sections, missing data are injected into the first column of the simulated sample (the same sample presented in Section 3.1.1). The data of the first column are missing according to a condition applied to the values of the last (10^{th}) column of the data matrix in order to obtain MAR data. More precisely, the process used here removes data from the first column according to extreme values of the last column: If an observation in the last column is part of its extreme values, for a given confidence level, then the same observation (i.e., in the same row) is missing in the first column. Thus, the procedure works in two steps: determine which observations contain extreme values (positive and negative) in the last column and delete the corresponding observations in the first column. This deletion process is illustrated in Figure 3.3-1.

Fig. 3.3-1: Data MAR depending on the last column of the data matrix



Because the sample comprises columns that are correlated with each other (the first and last columns correlate approximately 63%), the extremes values of the last column may also match the extremes values of the first column.

This MAR mechanism is in line with the definition of Little and Rubin [145], as the missing data depend on the observed data of the matrix and, specifically, the observed data of one of the columns of the matrix. Moreover, it is also in line with the definition of Schafer and Graham [182] as the missing data do not depend directly on the column containing this missing data but on another column that relates to the partially missing column.

This missing data mechanism can appear in financial data during a suspension of trading. Indeed, it is highly probable that the suspension of the quotation of a company's shares gives rise to a strong variation in the price of these shares but also of its suppliers and/or competitors. The collateral effects of a suspension of listing represent a research problem in its own right, which is not easy to analyze due to the data not always being accessible. Nevertheless, this missing data mechanism illustrates the extent to which completion methods can impute missing data that occur at the same time as a trading halt or some strong variations.

Thus, this missingness mechanism is applied to the same sample as in Section 3.2.1 (which is a sample of 10 columns and 261 rows) to compare the results between MCAR and this MAR mechanism.

Finally, to test the impact of an increasing proportion of MAR data on completion methods, the same proportions used in Section 3.2.1 are used here, that is, from 5% to 70%. Based on this MAR mechanism, the data that are missing in the sample with 5% are also missing in the sample with a higher missingness proportion (and so forth with any other missingness proportion). One of the main differences with the previous sections, is that the deletion condition here is based on price returns and not on the prices themselves (previously, the deletion process was performed directly on the price data). For example, for a proportion of 5%, the 2.5% of the highest returns and the 2.5% of the lowest returns in the last column will be identified and the corresponding prices will be deleted in the first column. For an extreme return observed between time $t-1$ and t in the last column, the deleted price is the one observed at time t for the first column.

Table 3.3-1 presents the proportion (and quantity) of missing returns associated with the proportion of missing prices injected in the first column of the simulated sample, following this MAR mechanism.

Tab. 3.3-1: Proportion (number) of missing returns (among the 100 missingness scenarios) associated with the proportion of raw MAR data injected into the first column of the simulated sample of length 261 (260 for return sample) and depending on extreme values of another column

	Proportion (and number) of missing returns associated with missing data													
Data	**5%** (12)	**10%** (26)	**15%** (39)	**20%** (52)	**25%** (64)	**30%** (78)	**35%** (90)	**40%** (104)	**45%** (116)	**50%** (130)	**55%** (142)	**60%** (156)	**65%** (168)	**70%** (182)
Return	9% (24)	20% (51)	28% (73)	37% (95)	43% (113)	52% (134)	58% (150)	63% (164)	70% (181)	76% (198)	81% (211)	84% (217)	87% (225)	90% (233)

Of course, the more the missingness proportion increases, the less the deletion will involve extreme data, but this will still highlight the impact of an increase in the proportion of MAR data.

Finally, each comparison tools is computed as defined in Figure 3.1-2, and the templates of all the graphs presented in this section have already been presented and detailed (i.e., what they represent and how they were obtained) in Section 3.1.4.

MCAR tests

The first analysis involves applying the MCAR tests on these 14 samples containing data MAR. Given their construction, these samples initially contain the same data, but their difference consists of their (proportion of) missing data. As explained, the first sample contains 5% of missing data, and the last 70%. This MAR data mechanism follows a specific mechanism without any random component, leading to a unique missingness scenario for each proportion of missing data (and not 100 as with the MCAR mechanism). This results in the application of a unique Little's test [142] and Jamshidian and Jalal's test [123] on each of these 14 missingness scenarios.

When data were removed completely at random (as in Section 3.2.1 and Section 3.2.2), both Little's test [142] and Jamshidian and Jalal's test [123] were efficient at detecting MCAR data on the simulated sample. Based on the previous results, Little's test [142] appears to be even more efficient than Jamshidian and Jalal's test [123] as it is always calculable and not sensitive to heteroskedasticity. The missing data mechanism of this section is MAR (in keeping with the definition of Little and Rubin [145]); hence, the purpose is to see whether the tests can also reject the null hypothesis that the data are MCAR because it is MAR now. In other words, the Type II error is analyzed, which is the probability to accept the null hypothesis when the alternative hypothesis is true.

Thus, the tests are successfully calculated for each of the 14 samples containing the return of the series. This kind of MAR data, whatever its proportion, does not affect the computation of these tests – even for Little's test, which did not always support a high proportion of missing data in the previous sections. This is the case for this missing data mechanism but not necessarily for others. The results of the two tests, with a 5% significance level, are available for each of the 14 missingness scenarios to which they have been applied.

First, Little's test [142] concludes that the data are MCAR for any proportion of missing data tested, while it is MAR. Thus, the test fails to properly assess that there is a link between the missing data in the first column and the extreme returns in the last column. By contrast, Jamshidian and Jalal's test [123] rejects the null hypothesis that the data are MCAR when the missingness proportion is lower than 60%, but above this threshold, the test no longer rejects it. As already observed, this test appears to be more efficient than Little's test [142] because it recognizes that the data are not MCAR when the amount of missing data is lower than 60%. However, when the missingness proportion is too important, the test can no longer see the relationship between the missing data and the observed data.

As these results are based on a unique sample, further investigation of a large number of samples (with other simulated samples) to which the same MAR mechanism is applied would clarify the conclusions of these tests. But since the purpose of this PhD thesis is to impute missing data, these investigations will be reserved for future study.

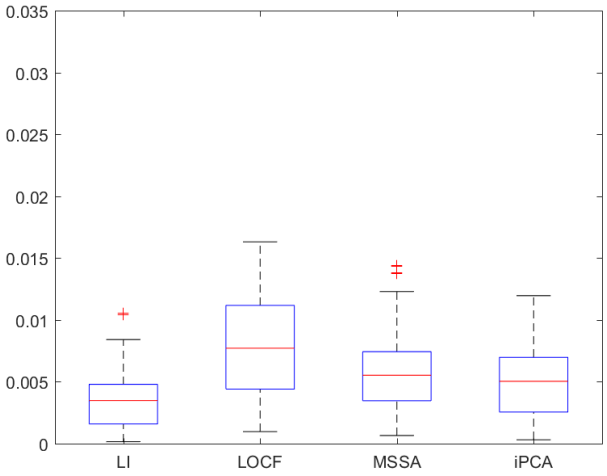
Preliminary results

Before analyzing in detail the performance according to the different criteria already used in the previous sections, it is interesting to note how each algorithm reacts to the MAR data for a specific proportion of missing data. The idea here is to compare the distribution of absolute deviations between the original performances and those of the imputed series for a specific proportion of missingness.

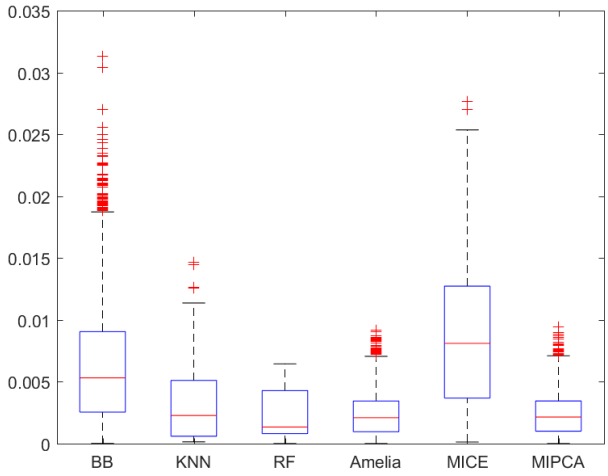
The same analysis was done in Section 3.2.1, where the absolute deviations of a scenario were analyzed in Figure 3.2-3 for proportions of MCAR data set at 10% and 30%. This analysis aims to show here whether the absolute differences between the returns of the original and imputed series are larger or not when the missing data of the first column correspond to the extreme values of the last column of the sample and not to MCAR data.

The study was repeated, and the results obtained are shown in Figure 3.3-2, with proportions of missing data at 10% (at the top) and 30% (at the bottom).

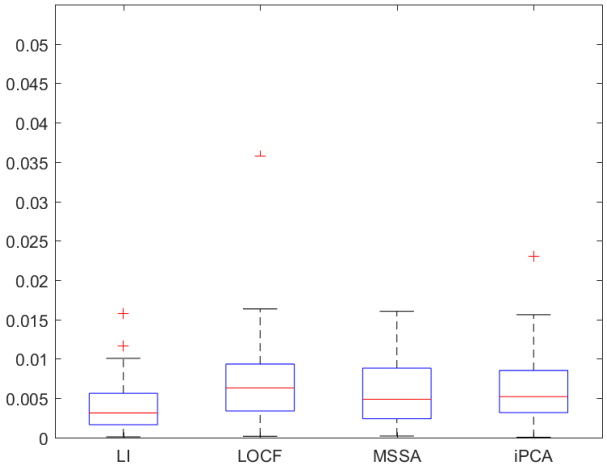
Fig. 3.3-2: Distribution of absolute return differences between the imputed series and original series for a sample containing 10% (at the top) and 30% (at the bottom) MAR data (only in the first column and depending on extreme values of another column)



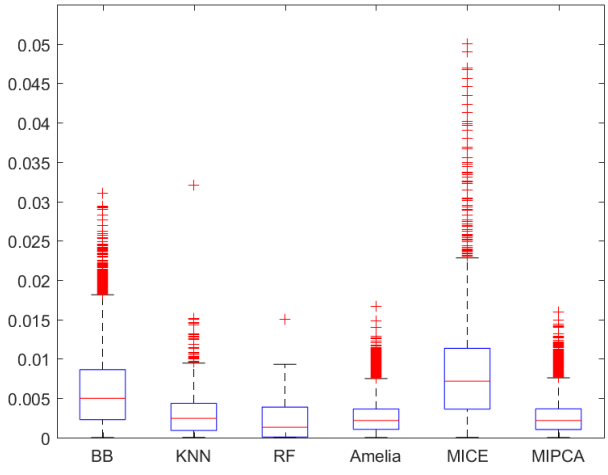
(a) Methods without a random component for 10% missingness



(b) Methods with a random component for 10% missingness



(c) Methods without a random component for 30% missingness



(d) Methods with a random component for 30% missingness

As already noted, the missing data are in the first column of the data matrix and correspond to the same observations of extreme values of the last column. The extreme

values of the first and last columns may be the same as they are positively correlated (around 63%). Thus, if the completion methods fail to detect that the missing data are part of the extreme data, the absolute deviations could be even greater than those observed in Figure 3.2-3 in Section 3.2.1.

Figure 3.3-2 shows a lot of similarity with the results of Section 3.2.1.. The linear interpolation, LOCF, MSSA, Brownian bridge, Amelia, and MIPCA methods obtain quasi-similar results when the data to impute are MCAR. Thus, the absolute differences observed when imputing MAR data are very similar to those observed in Section 3.2.1 when imputing MCAR data. Given that the missing data are the extreme values (especially when the proportion of missing data is 10%), similar distributions mean the methods manage to impute this MAR data with equivalent (not larger) deviations than when the data are MCAR.

The IPCA, K-NN, MICE, and random forests algorithms are impacted by this missingness mechanism because they do not impute MAR and MCAR data in the same way. These three methods have more difficulty with MAR data given that the absolute differences are larger than in Section 3.2.1.

First of all, the absolute differences in the results of the IPCA algorithm are approximately twice as large as those in the case where the data were MCAR. However, these results should be put into perspective since they are based on a single scenario. Moreover, in the case of MCAR data, the distribution of absolute deviations of the IPCA was comparable to that of MIPCA (excluding outliers), whereas here, the IPCA algorithm tends to have much larger absolute deviations than the MIPCA algorithm (including outliers).

The K-NN algorithm also results in larger differences when the data are MAR and not MCAR for a proportion of missing data at 10%. By contrast, when the proportion of missing data increases to 30%, the absolute deviations are not greater than when the proportion is 10% (as when the data were MCAR) – they even tend to be slightly lower if the outliers are excluded.

The MICE algorithm leads to a larger distribution with MAR data than with MCAR data when the missingness proportion is equal to 10% but not with a proportion at 30

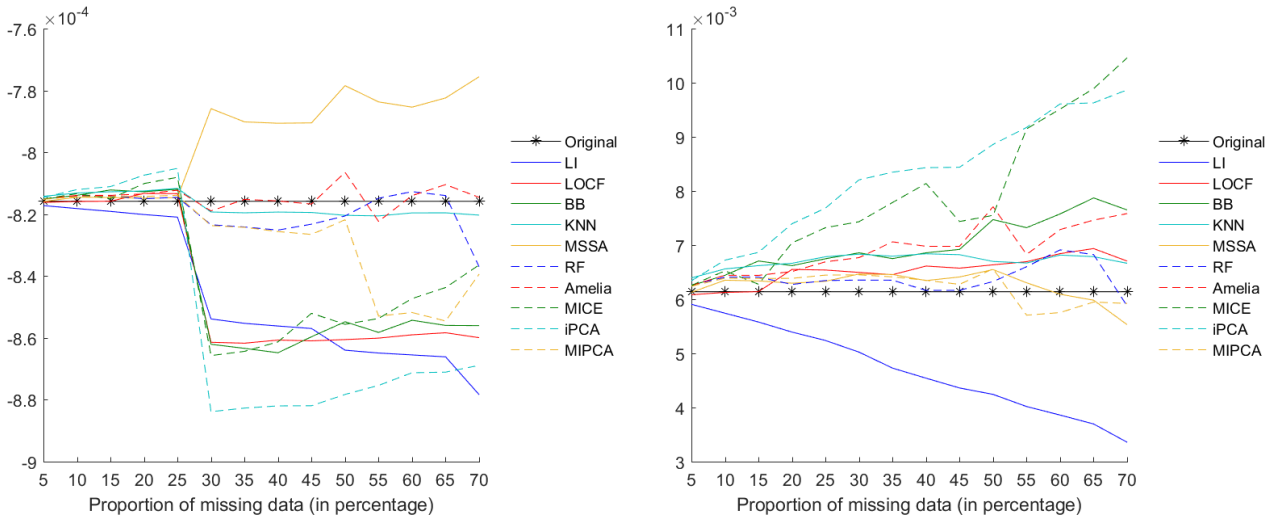
Finally, the random forests algorithm, which obtained very stable results with absolute deviations very close to zero for the two proportions of MCAR data (see Figure 3.2-3), obtain much larger deviations here. Whereas the observed distribution was previously very close to zero, reflecting small deviations from the original series, the algorithm is less efficient when dealing with this kind of MAR data. The distribution obtained here is comparable to that of Amelia and MIPCA algorithms (excluding outliers). Moreover, absolute deviations tend to be larger and larger as the series contains missing data. Despite this decreasing performance, the random forests algorithm remains the most efficient at imputing the MAR data here.

The results presented above, as well as those presented in Section 3.2.1, are from a single scenario and for a given proportion of missing data, so they should be put into perspective. For this reason, in the rest of this section, the analysis will focus on all the proportions of missing data used.

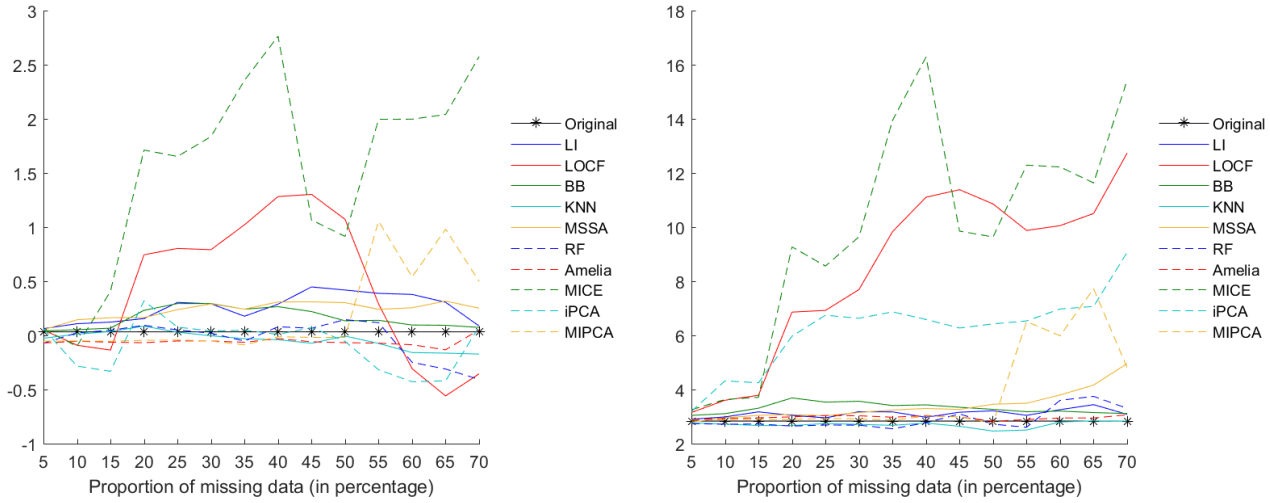
Statistical moments

The results below in Figure 3.3-3 highlight the ability of each algorithm to preserve the first four statistical moments of the sample when MAR data are removed from the sample, as previously explained.

Fig. 3.3-3: The first four statistical moments of the returns of the imputed data based on a matrix containing MAR data (depending on extreme values of another column) according to the missingness proportion



(a) Mean *(b)* Standard deviation



(c) Skewness *(d)* Kurtosis

As these graphs are associated with a single sample and a single scenario of missing data for each proportion, they should be put into perspective. Overall, the methods tend to impute missing data in a way that overestimates the level of volatility of the series, as also observed in the case of MCAR data (see Section 3.2.1). However, the skewness and kurtosis coefficients tend to be preserved for most of the methods.

The usual methods obtain large deviations for at least half of the statistical moments: Linear interpolation fails to reproduce the mean and standard deviation of the original series, and the LOCF method is among the worst performers as it fails to estimate the mean, skewness, and kurtosis.

However, the algorithm that deviates the most from the original series in terms of statistical moments is the MICE algorithm. This method leads to the highest level of skewness and kurtosis here, starting at 15% of MAR data while adding more volatility to the series and underestimating its mean. This algorithm was already one of the least efficient with MCAR data, and it is not surprising that it falls behind here. Nevertheless, these results have to be moderated, given that in Section 3.2.3 and Section 3.2.4, the errors seemed to relate to the sample used.

While the IPCA algorithm was one of the relatively satisfactory methods in Section 3.2.1, it is among the least performing here in terms of mean, standard deviation, and kurtosis. Indeed, this algorithm obtains moments that deviate completely from the original series as the proportion of MAR data increase. Here, it is considerably less efficient than its improved version, MIPCA, which deviates from the original series only when the proportion of missing data exceeds 50%. Below this proportion, the algorithm is still able to preserve the moments of the series, but above 50% of the

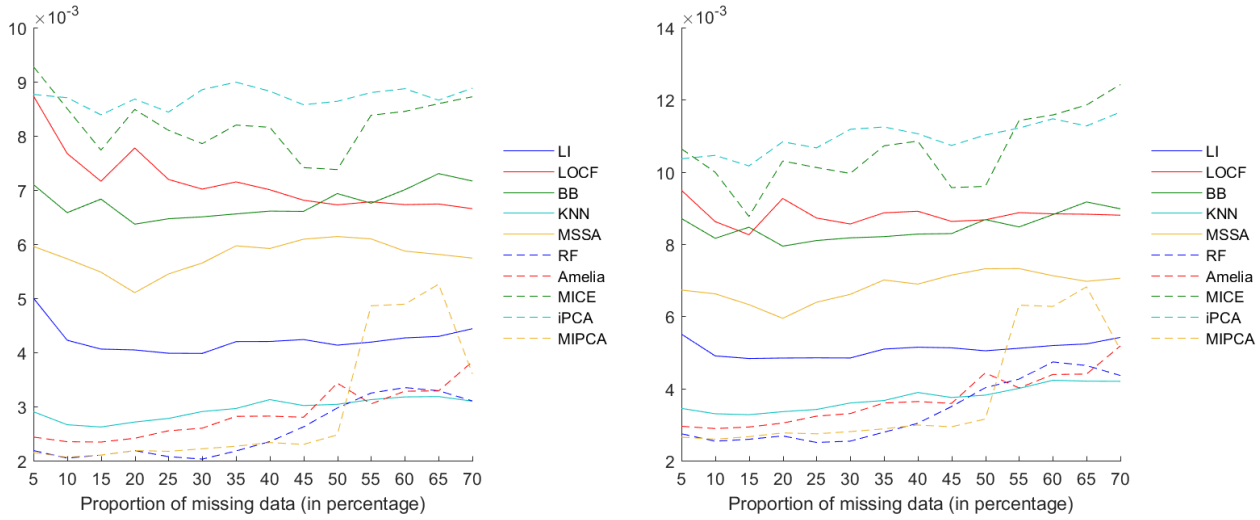
average, skewness, and kurtosis suddenly move away from the original series. This is in part because the IPCA and MIPCA algorithms do not use the same number of principal components (three and two, respectively), as presented in Appendix E.1. However, it is necessary to keep in mind that these results are from a single missingness scenario for each proportion of missing data and not 100 as in Section 3.2.1. The abnormal behavior of MIPCA is probably the result of sampling effects.

Some of the most stable results, regardless of the proportion of MAR data present in the sample, are those of the K-NN algorithm, which obtains quite satisfactory results. Lastly, Amelia and random forests algorithms also obtain good results for this specific sample with this specific missingness mechanism.

Proximity metrics

Figure 3.3-4 represents the MAE and RMSE obtained between the returns of the original series and those of the imputed series for the different proportions of MAR data present in the column.

Fig. 3.3-4: MAE and RMSE between the return of the imputed data from a matrix containing MAR data (only in the first column and depending on extreme values of another column) and the original data matrix, according to the missingness probability



(a) MAE *(b)* RMSE

First of all, the results presented here are higher than the average proximity measures obtained when the data were MCAR, as presented in Figure 3.2-7 from Section 3.2.1. Moreover, if the random forests method was the closest to the original series, according to the MAE and RMSE, here the results are less clear. Contrary to what Young [214]

found in his article, the results of the random forests are unlike those obtained by the MICE algorithm. And yet, the general conclusion of his article is in line with the results above: The random forests method seems to be more precise for the imputation of MAR data.

As these results are based on a single sample, it is not possible to draw general conclusions; however, some groups of methods do manage better than others: Random forests, MIPCA, Amelia, and K-NN are some of the imputation methods that make it possible to minimize the proximity measures. These methods always perform better than linear interpolation.

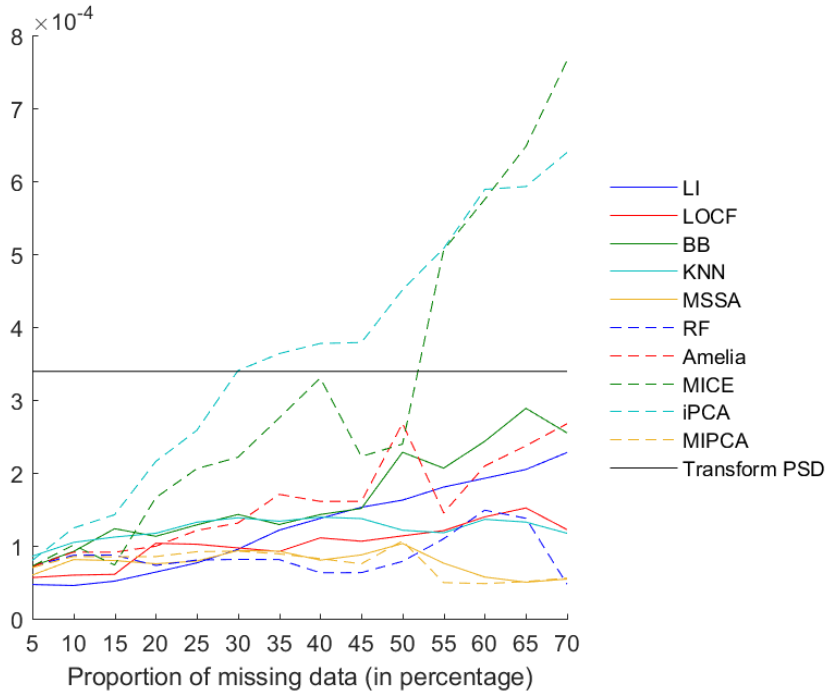
This analysis calls into question the use of other, often more complex methods. This is the case for the Brownian bridge, MSSA, and MICE methods, which obtain average deviations from the original series that are sometimes four times greater than other methods.

The poor performance of the IPCA method is not cited here, as it does not seem to be representative of the method; it has been one of the worst-performing methods since the beginning of the chapter. Nevertheless, this shows how much it can differ from its multiple imputation version.

Covariance matrices comparison

Thus, the algorithms can also be compared in terms of the covariance matrix. For this, Figure 3.3-5 represents the differences between the covariance matrix of the original series and the one obtained from the imputed data, based on the Frobenius norm, for the different levels of missingness proportion. Since the MAR data are only present in the first column, the impact on the covariance matrix is limited to its first column (and its first row).

Fig. 3.3-5: Covariance matrix differences, according to the Frobenius norm, based on original returns and the imputed returns from a matrix containing MAR data (only in the first column and depending on extreme values of another column) according to the missingness probability



It appears that this MAR mechanism leads to more difficulties when imputing missing data than with MCAR data (see Figure 3.2-9 from Section 3.2.1), since the transformation to make the matrix positive semidefinite achieves results much closer to those of the other algorithms.

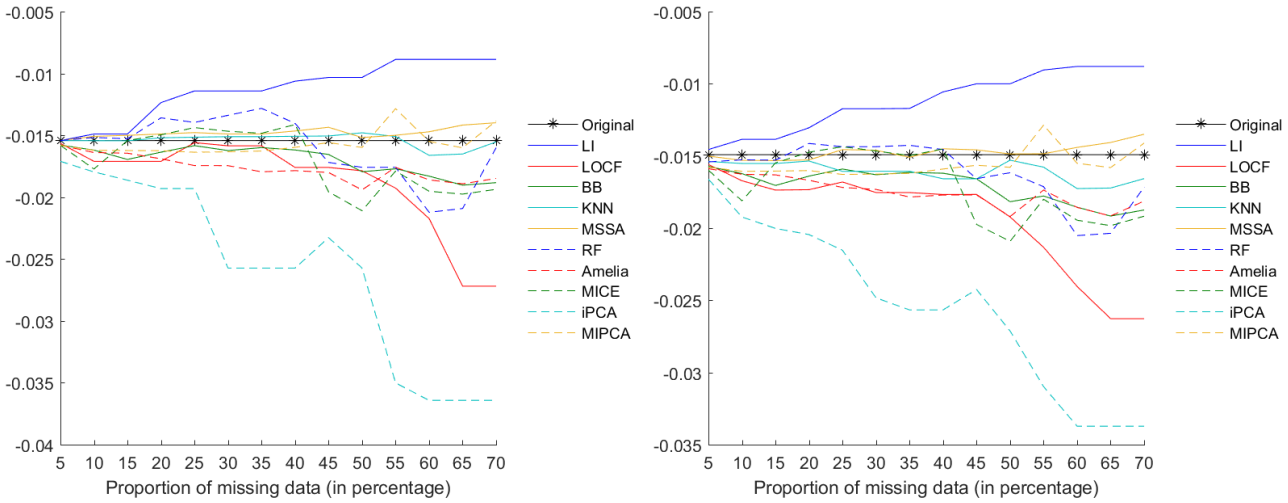
Regarding the differences in terms of covariance, some similarities with MCAR data should be highlighted. The MICE algorithm tends to increase considerably the gaps between covariance matrices, regardless of whether the data are MAR or MCAR. Linear interpolation obtains the same results in both cases (MAR and MCAR) and is most efficient when the proportion of missing data is less than 20%.

Otherwise, the methods tend to distort the covariance more and more as there are missing data in the sample. Finally, random forests and MIPCA seem to be the methods that minimize these differences. By contrast, using IPCA and MIPCA on this sample with this missingness mechanism leads to drastically different results.

Value-at-risk and expected shortfall

Finally, completion methods are compared in terms of risk measures. Figure 3.3-6 represents VaR and ES with a 1-day horizon for a confidence level of 99% and 97.5%, respectively, derived from imputed data according to the different proportions of missing data.

Fig. 3.3-6: The 1-day risk measures computed from a matrix containing MAR data (only in the first column and depending on extreme values of another column) according to the missingness probability



(a) 1-day VaR at 99%

(b) 1-day ES at 97.5%

Overall, the imputation methods tend to be conservative in terms of risk measures (i.e., below the level of the original series).

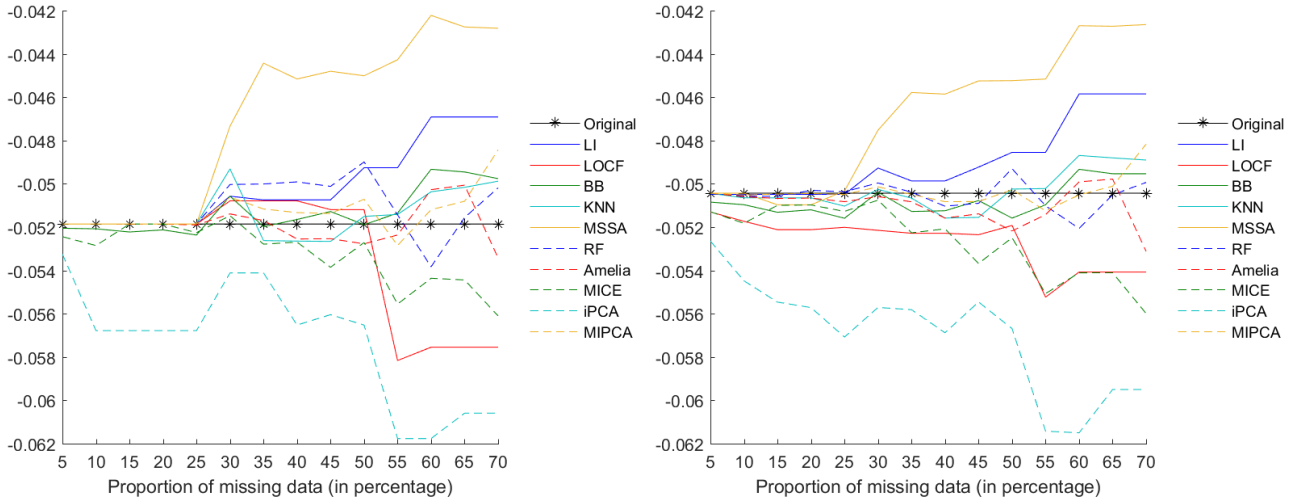
The poor performance of the IPCA is particularly visible and further highlights the effectiveness of its improved version, the MIPCA algorithm.

However, the MSSA algorithm appears to be one of the closest methods of original risk measures with a 1-day horizon. This method was already the one that most faithfully reproduced the 1-day risk measures when the data were MCAR (see Figure 3.2-11 from Section 3.2.1). Moreover, these results are close to those obtained by the MIPCA and K-NN methods. In fact, these three methods are the most effective in terms of 1-day risk measures when the data are MAR.

The random forests algorithm remains relatively unstable from one missingness proportion to another but probably because of the sample. Lastly, the Amelia algorithm obtains risk measures comparable to those obtained from MCAR data.

In addition, Figure 3.3-7 represents the VaR and ES for a horizon of 10 days and a confidence level of, 97.5% and 97.5%, respectively.

Fig. 3.3-7: The 10-day risk measures computed from a matrix containing MAR data (only in the first column and depending on extreme values of another column) according to the missingness probability



(a) 10-day VaR at 99%

(b) 10-day ES at 97.5%

For the 10-day risk measures, the methods are relatively close to each other when the proportion of missing data is less than 30% (except for IPCA), but beyond this threshold, the methods distort the risk measures.

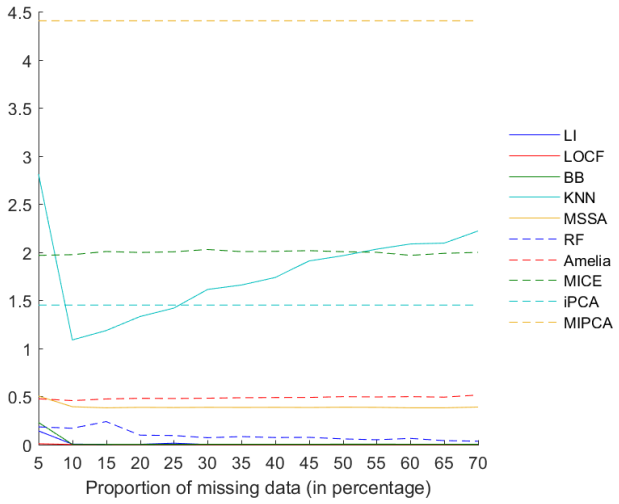
Thus, for the 10-day risk measures, the MSSA algorithm and linear interpolation overestimate their levels, as when the data were MCAR. However, the IPCA algorithm, which was very efficient with MCAR data, here becomes the least capable of reproducing the risk measures. Opting for such a completion method would have a considerable impact in terms of capital charge for a bank.

The K-NN, random forests, Amelia, and MIPCA methods obtain risk measures that oscillate around those of the original series. Again, this lack of stability (not observed in Section 3.2.1) is explained, in particular, by the fact that a single scenario of MAR data is used here, as opposed to 100 scenarios of MCAR data. Thus, these four methods provide 10-day risk measures that are relatively comparable to those of the original series.

Computation time

Finally, Figure 3.3-8 represents the computation time required by each algorithm to impute the samples, according to the proportion of missing data.

Fig. 3.3-8: Computation time of the imputation of MAR data (depending on extreme values of another column) on only the first series according to the missingness probability



The computation time for all the completion methods is relatively constant for all proportions of missing data, except for the K-NN algorithm, which requires a lot of computation time for the first sample (approximately 3 seconds) before decreasing sharply (by approximately 2 seconds) and then to gradually increase. Once again, these results are those of a single scenario and, thus, should be considered in this context. Moreover, this algorithm seems to require a launch time. Nevertheless, apart from the launch of the algorithm, the evolution of the computation time is the same as when the data were MCAR (see Figure 3.2-14 from Section 3.2.1).

In addition, the computation times of the other algorithms are also similar to those obtained when processing MCAR data, except for two methods: The MIPCA algorithm requires 4.5 times more computation time to process MAR data than MCAR data, and the IPCA algorithm can take more than a second longer.

This section demonstrates the differences between the imputation of MCAR data (done in Section 3.2.1) and the imputation done on MAR data while respecting a well-defined mechanism. This mechanism is based on the extreme values of another column of the data matrix. The results obtained in this section are, therefore, specific to this

mechanism and will vary as soon as the MAR data mechanism is modified (or the sample). For this reason, this MAR mechanism should be applied to a large number of simulated samples before drawing conclusions.

Nevertheless, for this type of MAR mechanism on the simulated sample, it is clear that neither the usual methods, nor the Brownian bridge, nor even the MICE algorithm are well-adapted. Even more surprisingly, the IPCA algorithm seems to be unsuitable for this missing data scheme.

On the contrary, random forests and Amelia appear to be the most efficient at dealing with these missing data, but this time, the random forests face imputation difficulties when the proportion of missing data becomes too important (whereas it was the case for Amelia in Section 3.2.1). The K-NN is also among the best-performing methods, as is the MIPCA, although the latter's performance degrades strongly beyond 50% of missing data. This is probably due to sampling effects.

While this MAR mechanism remains a special case, it has the benefit of showing the difficulties of imputation in the case of an empirical study. The application of MCAR tests for these kinds of MAR data gives poor information on the nature of the missing data (depending on the proportion of missing data). Thus, if the imputation of missing data is done using IPCA, which until now has given satisfactory results, the analyses done afterward would not be representative of the sample given the results obtained here.

3.3.2 Impact of successive missing data in the middle of the series

The next MAR mechanism studied here involves successive missing data. This kind of missing data mechanism is frequently observed in financial time series: For example, data can be erroneously deleted over a certain period (e.g., during an IT migration), or that a computer process fails for several days, or that data requiring a manual backup are forgotten over a longer or shorter period, or that some illiquid financial products are not quoted on the markets for several days because of a lack of buyers, or even that a security is in a long-term suspension of listing.

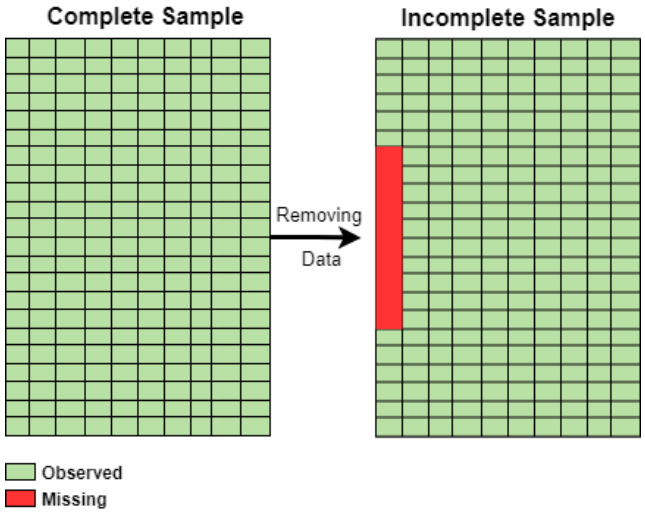
These events can explain the presence of successive missing data in a sample.

This mechanism is different from the previous one because the missing data depend not on another column of the data matrix but on another variable: time. The mechanism highlights the differences between successive data imputations and scattered data imputations over the whole sample.

Successive missing data mechanism

This section analyzes the impact of an increasing proportion of successive missing data. These missing data depend not on the series but on time. This mechanism is also applied to the first column of the data matrix, and successive data will be removed from the middle of this column, as represented in Figure 3.3-9.

Fig. 3.3-9: Successive missing data in the middle of the first column



The categorization of missing data for this mechanism remains debatable. Here, the data are successively missing, which means the probability for a data point to be missing depends on the observability at the previous time. To be more specific, the probability of a missing data point knowing that the previous data point is missing will be 1 until the wanted proportion of missing data is reached. Thus, the probability that a data point is missing is not the same each time. However, according to Little and Rubin [145], this mechanism is categorized as MCAR because the missingness does not depend on either observed or missing data. If there were a variable that listed quotation days and non-quotation days, this mechanism would be categorized as MAR according to Little and Rubin [145], but such a variable does not exist in practice.

According to Schafer and Graham [182], this missing data mechanism gives rise to MAR data since the missing data depend on another variable: time.

As a reminder, according to Schafer and Graham [182], data are MCAR as soon as they are explained by another variable related to the one containing the missing data. This missing data mechanism could illustrate, for example, the case of a suspension of quotation through a partially observed column representing the prices quoted on the stock exchange and a binary variable representing its quotation dates (taking the value 1 if the stock is quoted and 0 otherwise). Thus, this mechanism would represent MAR data in the sense of Schafer and Graham [182], since the missing data do not

depend directly on the stock prices: They depend on its quotation schedule. However, as previously noted, the very existence of such a binary variable is not obvious. In practice, if data are available, it is because the price was quoted on the market; but in the opposite case, no process is put in place to justify that there was no quotation. This makes it difficult to differentiate between data that are missing by accident and data that are missing because they were not quoted.

Moreover, if such a binary variable existed and was integrated into the data matrix, the data would be MAR in the sense of Little and Rubin [145], which would make it possible to converge to the same category as Schafer and Graham [182]. However, integrating such a variable into the data matrix will not make sense for the completion algorithms.

Thus, this missingness mechanism can give rise to missing data categorized as MCAR if Little and Rubin's definition of MCAR is strictly considered. By contrast, if the definition chosen is that of Schafer and Graham [182], this mechanism results in MAR data. Since the definitions do not converge and are debatable, this mechanism will be considered as MAR in this PhD thesis, since it is a much broader category.

As in previous sections, an increasing proportion of data (from 5% to 70%, in increments of 5%) is removed from the first column to see the impact of the completion methods. This MAR mechanism implies that the data that are missing for a specific proportion of missing data are also missing for a higher proportion of missing data.

These deletions are made directly on price series. Since this MAR mechanism does not depend on a random component, a unique missingness scenario is given for each proportion of missing data.

For this missing data mechanism, the proportion of missing data injected into the price series is very close to the proportion of missing returns. This is because the missing prices are missing successively. The correspondences for this MAR data mechanism are presented in Table 3.3-2.

Tab. 3.3-2: Average proportion (number) of missing returns (among the 100 missingness scenarios) associated with the proportion of MAR raw data (based on successive missing data in the middle of the series) injected into the first column of the simulated sample of length 261 (260 for return sample)

	Proportion (and number) of missing returns associated with missing data													
Data	**5%** (13)	**10%** (27)	**15%** (39)	**20%** (53)	**25%** (65)	**30%** (79)	**35%** (91)	**40%** (105)	**45%** (117)	**50%** (131)	**55%** (143)	**60%** (157)	**65%** (169)	**70%** (183)
Return	5% (14)	11% (28)	15% (40)	21% (54)	25% (66)	31% (80)	35% (92)	41% (106)	45% (118)	51% (132)	55% (144)	61% (158)	65% (170)	71% (184)

The same data matrix used previously in Section 3.2.1, Section 3.2.2, and Section 3.3.1 is used here to compare the results and the impact of this kind of MAR data.

Since the deletion process does not depend on the distribution of the data, and since the returns of the simulated sample are independent and identically distributed, the distribution of missing data is approximately the same as the distribution of observable data. This means that algorithms that use the law of observable data should perform well here.

Finally, each comparison tool is computed as defined in Figure 3.1-2, the templates of all the graphs presented in this section have already been presented and detailed (i.e., what they represent and how they were obtained) in Section 3.1.4.

MCAR tests

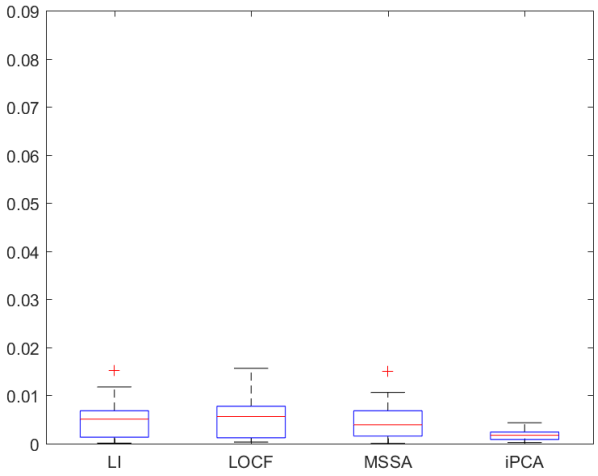
The MCAR tests are based on the categorization of Little and Rubin [145]. As previously discussed, this missingness mechanism (successive missing data in the middle of the series) is categorized as MCAR according to Little and Rubin's [145] definition. Thus, the MCAR tests are efficient if the null hypothesis that data are MCAR is accepted. Little's test [142] and Jamshidian and Jalal's test [123] were applied to the price return matrices containing successive missing data on the first column. No errors were encountered to run these tests; that is, all the tests are calculable here, so the results are obtained for each of the 14 samples (containing an increasing proportion of missing data).

The results obtained are clear: Neither test rejects the null hypothesis that the data are MCAR, regardless of the missingness proportion. Thus, these two tests can categorize data as MCAR following the classification of Little and Rubin [145].

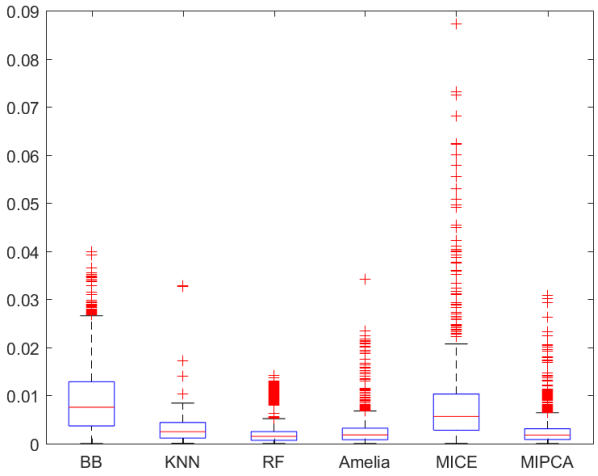
Preliminary results

First, the results presented in Figure 3.3-10 highlight the absolute differences between the returns of original and imputed series for each of the methods used for proportions of missing data set at 10% (top) and 30% (bottom).

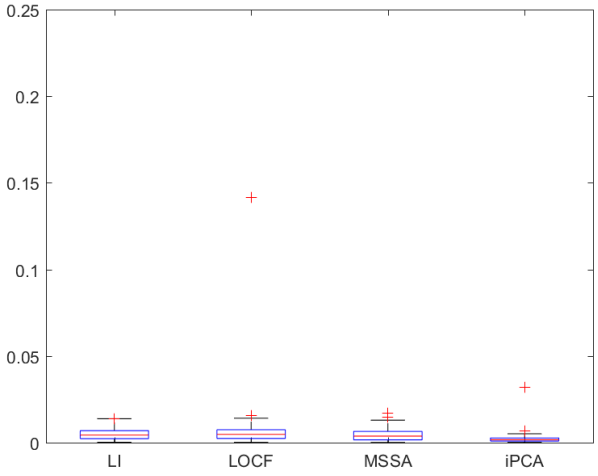
Fig. 3.3-10: Distribution of absolute return differences between the imputed series and original series for a sample containing 10% (at the top) and 30% (at the bottom) MAR data (successive missing data in the middle of the first series)



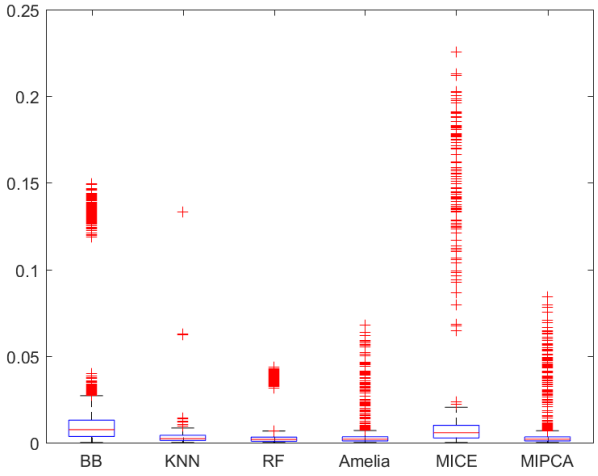
(a) Methods without a random component for 10% missingness



(b) Methods with a random component for 10% missingness



(c) Methods without a random component for 30% missingness



(d) Methods with a random component for 30% missingness

Completion methods face greater difficulty imputing successive missing data than MCAR or even the first MAR mechanism (presented in Section 3.2.1 and Section 3.3.1,

respectively). Absolute deviations from the original series are represented here by scales approximately three and five times larger for 10% and 30%, respectively, of successive missing data than of MCAR data (see Figure 3.2-3 from Section 3.2.1). But these differences are mainly due to outliers of distributions (i.e., red crosses).

If the analysis focuses on distributions, excluding outliers, then the linear interpolation methods, MSSA, K-NN, Amelia, and MIPCA obtain distributions comparable to those previously obtained for MCAR data (see Figure 3.2-3 from Section 3.2.1) or even for the previous MAR mechanism (see Figure 3.3-2 from Section 3.3.1). The random forests, Amelia, and MICE algorithms, which were among the best-performing in those sections, could be effective in this section as well.

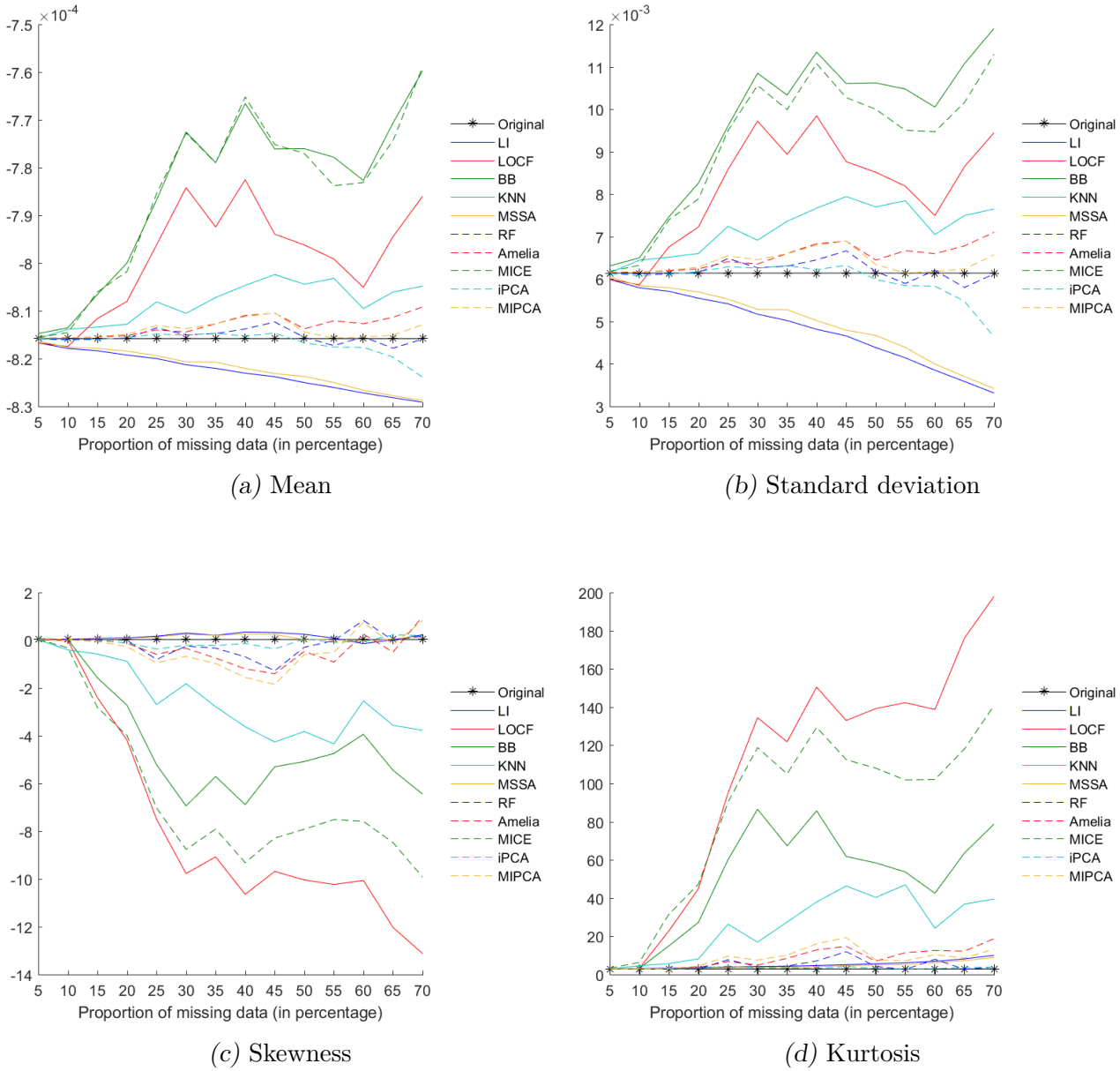
The Brownian bridge and MICE methods are the ones with the most spread-out distributions (including outliers), and this has always been the case, regardless of the type of missing data observed so far. The Brownian bridge method leads to larger differences than the other types of missing data previously tested, even without outliers. This means that this method is likely to remain among the worst performers in this missing data mechanism, even if this result was expected.

Finally, the Amelia and MIPCA algorithms also obtain distributions comparable to each other here, suggesting that these methods will be comparable to each other throughout this section as well. K-NN obtains good results too, but the random forests method provides the smallest absolute differences.

Statistical moments

To compare the impact of an increasing proportion of successive missing data in the series, the first four statistical moments of the series were calculated after imputation for a comparison with the original moments. The results of these statistical moments are presented in Figure 3.3-11.

Fig. 3.3-11: The first four statistical moments of the returns of the imputed data based on a matrix containing MAR data (successive missing data in the middle of the first series) according to the missingness proportion



(a) Mean

(b) Standard deviation

(c) Skewness

(d) Kurtosis

The results presented show that the LOCF, Brownian bridge, and MICE methods are far from the original series for at least three of the four statistical moments. These methods distort the distribution when imputing successive missing data. These results are not surprising given that the LOCF and Brownian bridge methods are one-dimensional interpolation methods that do not use information from other columns.

The kurtosis obtained from the LOCF data is increasingly large – even huge – due to imputations creating zero returns (except the last one). Furthermore, the MICE algorithm was already one of the worst performers on this simulated Gaussian sample, regardless of the missing data mechanism used; hence, it is not surprising to have it among the worst performers here too.

Moreover, the MSSA algorithm here obtains results very similar to linear interpolation in the case of successive missing data. Thus, the model does not appear to be more satisfactory than a simple linear interpolation.

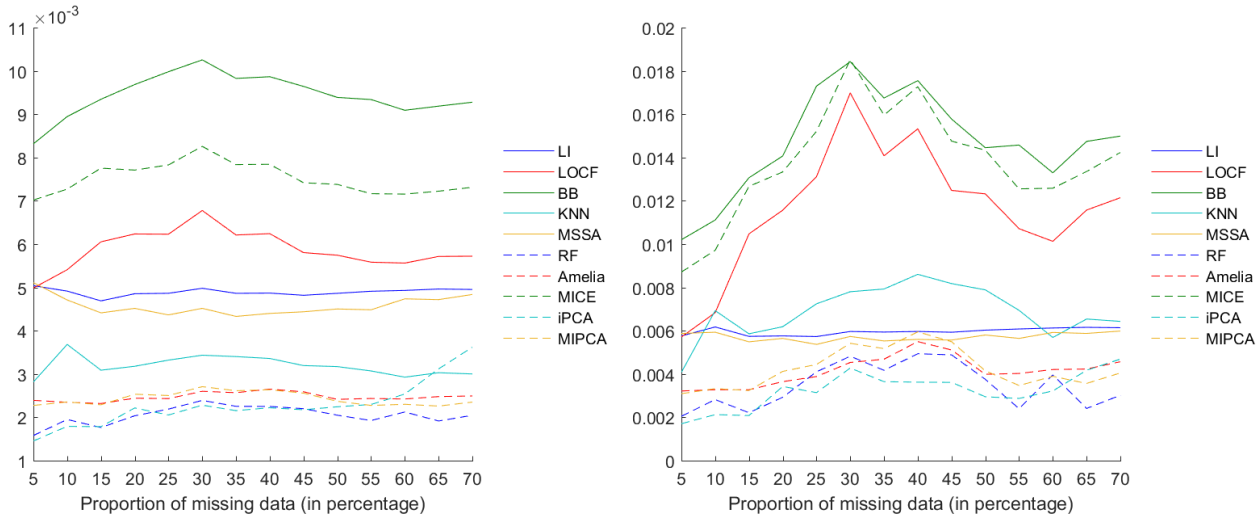
Random forests, Amelia, IPCA, and MIPCA are among the best methods to preserve the moments of the original series. These four algorithms can impute successive data in such a way as to reproduce the moments of the original series relatively accurately. The data that are missing follow, a priori, the same distribution as the observable data, which is why these algorithms obtain results close to the MCAR data (see Section 3.2.1) or at least a little higher. Among these methods, the IPCA algorithm is particularly efficient when the proportion of missing data is less than 50%. Moreover, this method uses the same number of principal components as MIPCA (see Appendix F.1). The random forests algorithm also obtains moments very similar to those of the original series for any missingness proportion.

Even though these results are from a single sample of data, they still allow us to distinguish the good performers from the bad ones.

Proximity metrics

As previously, MAE and RMSE were calculated between the returns of the original series and those of imputed series to see the impact of an increasing proportion of missing data on the imputation quality. These results are presented in Figure 3.3-12.

Fig. 3.3-12: MAE and RMSE between the return of the imputed data from a matrix containing MAR data (successive missing data in the middle of the first series) and the original data matrix, according to the missingness proportion



(a) MAE *(b)* RMSE

The good performance of the random forests, Amelia, IPCA, and MICA algorithms, previously observed in terms of statistical moments, is also observed in terms of proximity measures. These same algorithms minimize the MAE and RMSE for any proportion of missing data. As these results are on a single sample, it is not possible to define a precise ranking of the methods; however, since these methods have already been among the best-performing in the previous sections. Moreover, while previously in the case of MCAR data, random forests appeared to be the best-performing, here the four methods get similar results, which is why a multi-sample analysis for this missing data mechanism will be relevant for future study.

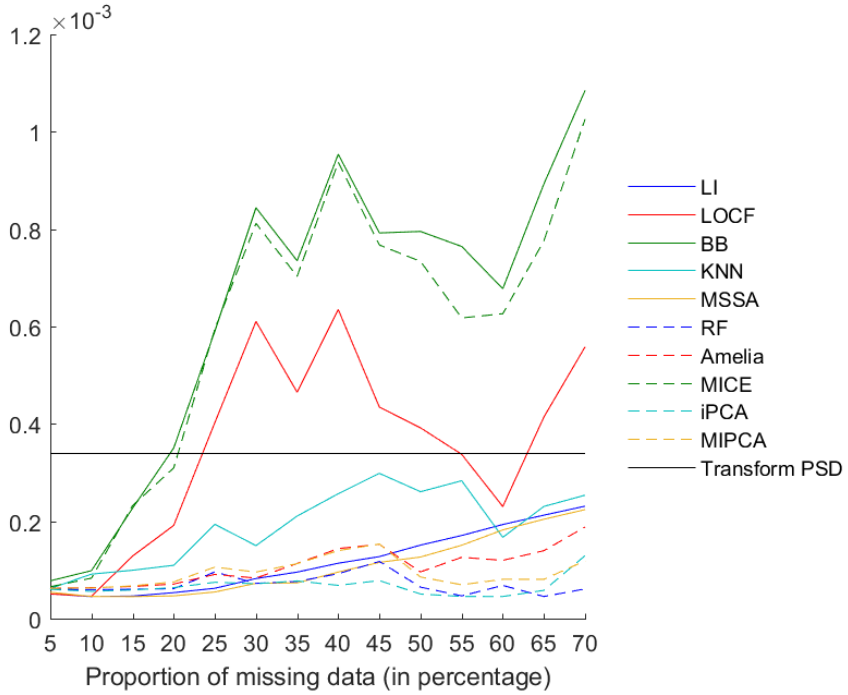
These four algorithms are closely followed by the K-NN algorithm. Although this method performs better than a simple linear interpolation in terms of MAE, it tends to impute some data far from the original series, given its RMSE exceeding that of the linear interpolation. Moreover, as with the statistical moments, the MSSA algorithm obtains results comparable to those of linear interpolation.

Finally, the LOCF, Brownian bridge, and MICE algorithms obtain results farthest from those of the original series, as well as an RMSE that tends to be higher than the MAE, which implies that these methods impute some data with particularly large deviations from the original series.

Covariance matrices comparison

Figure 3.3-13 represents the proximity of each of the imputation methods in terms of the covariance matrix. For this, the differences between the original covariance matrix and the one obtained from the imputed series, according to a Frobenius norm, have been computed here.

Fig. 3.3-13: Covariance matrix differences, according to the Frobenius norm, based on original returns and the imputed returns from a matrix containing MAR data (successive missing data in the middle of the first series) according to the missingness proportion



The transformation of the pairwise matrix is preferable to the Brownian bridge, MICE, and LOCF methods as soon as 20% of missing data are in the column. The differences between the original covariance matrix and those of these imputation methods increase greatly as soon as 15% of the data in the column are missing. K-NN is also one of the least efficient algorithms at preserving the covariance matrix for this type of MAR data, leading to differences close to the positive semidefinite pairwise matrix. These methods were already the least efficient in terms of statistical moments and proximity measurements, so these results were expected.

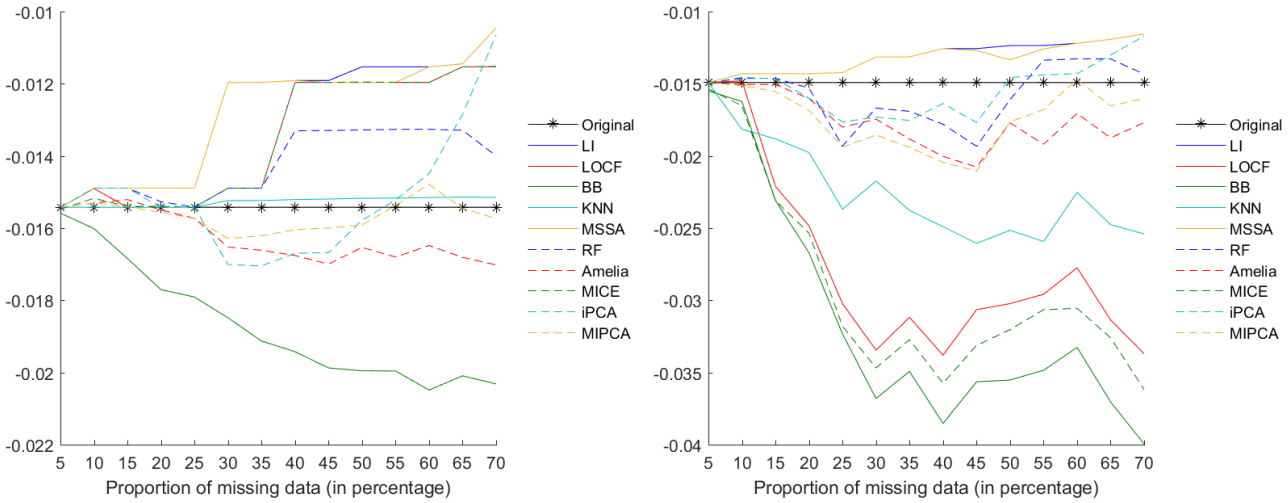
On the other hand, the methods that were previously the most efficient on the other analysis criteria are also the most efficient here. This is the case for the IPCA and

random forests algorithms, which obtain the lowest matrix deviations for this simulated sample.

Value-at-risk and expected shortfall

To compare completion methods in terms of risk measures, VaR and ES with a 1-day horizon and a confidence level of 99% and 97.5%, respectively, have been calculated and are presented in Figure 3.3-14.

Fig. 3.3-14: The 1-day risk measures computed from a matrix containing MAR data (successive missing data in the middle of the first series) according to the missingness proportion



(a) 1-day VaR at 99% *(b)* 1-day ES at 97.5%

These analyses of risk measures are based on a single sample; hence, they should be put into perspective. Efficient methods like Amelia and MIPCA manage to preserve the risk measures for a proportion of missing data inferior or equal to 15% (even 20% in the case of VaR). Beyond this threshold, the risk measures are impacted.

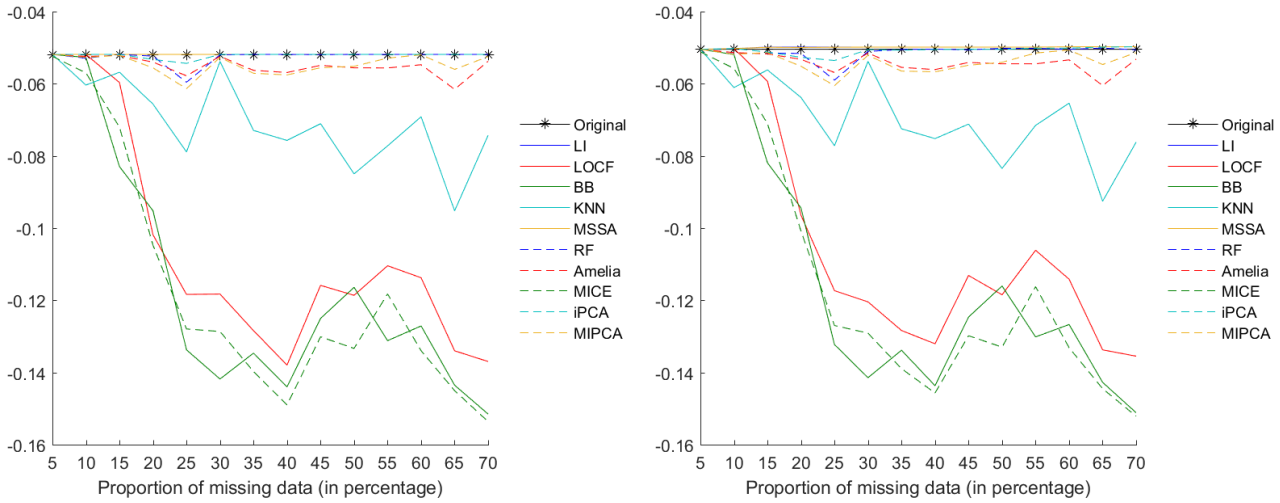
Despite the good performance of the random forests and IPCA algorithms in the previous results, these methods are unable to reproduce in a stable way (from one missingness proportion to another) the original VaR and ES level, for a 1-day horizon. This reveals that both algorithms are unable to faithfully reproduce the missing extreme returns. However, they are not part of the algorithms that strongly overestimate or underestimate these risk measures.

No method performs for both VaR and ES for any proportion of missing data. By contrast, the K-NN method can reproduce the VaR level of the original series quite

faithfully for any proportion of missingness. The K-NN method strongly underestimates the level of the ES, which may lead the bank to pay high additional capital charges.

Similarly, VaR and ES with a 10-day horizon and a confidence level of 99% and 97.5%, respectively, have been calculated and are presented in Figure 3.3-15.

Fig. 3.3-15: The 10-day risk measures computed from a matrix containing MAR data (successive missing data in the middle of the first series) according to the missingness proportion



(a) 10-day VaR at 99%

(b) 10-day ES at 97.5%

In the case of a 10-day horizon, risk measures are particularly well reproduced by a simple linear interpolation. This leads one to believe that the highest 10-day returns of the original series are at the beginning and the end of the first column of the data matrix because linear interpolation does not impute by high returns (extreme values). This is true for both VaR and ES. Given that the results of MSSA have been comparable to those of the linear interpolation since the beginning of this section, it is not surprising that this algorithm also performs well for these 10-day risk measures.

The IPCA and random forests algorithms are also very efficient at reproducing true VaR and ES, except for the sample containing 25% missing data where the risk measures are not the same as the true ones, albeit still close. Moreover, the Amelia and MIPCA algorithms obtain risk measures that are relatively faithful to the true risk measures but less than the methods mentioned above.

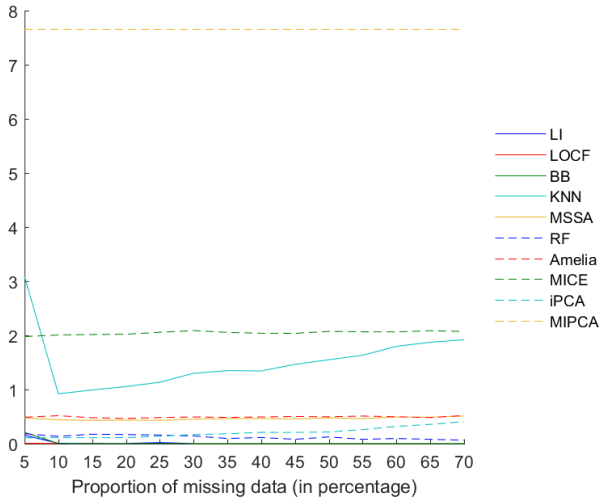
The K-NN, MICE, and Brownian bridge algorithms, as well as the LOCF method, obtain risk measures far below the target levels, but this is consistent with their poor

results in this section. These methods can be very costly for a bank if one of them is chosen to handle missing data.

Computation time

Finally, the computation times (in seconds) required for each algorithm to impute a sample are presented in Figure 3.3-16.

Fig. 3.3-16: Computation time of the imputation of MAR data (successive missing data in the middle of the first series) according to the missingness proportion



The computation times observed here are the same as those presented for the first MAR mechanism (see Section 3.3.1), except in the case of MIPCA. The computation time of MIPCA is twice as high here as in the first MAR mechanism tested, and this is the highest computation time observed for this algorithm so far.

The missing data in this section follow a mechanism that consist in missing data concentrated in only a part of the sample (successive missing data), to see how the algorithms impute the data.

The algorithms that were inefficient in the previous sections are still inefficient here. However, the K-NN, which could obtain good performances in the previous sections, obtain bad ones as it does not manage to do better than a simple linear interpolation. This may be due to the simulated sample used. Moreover, although sophisticated, the MSSA algorithm appears to be as efficient as a linear interpolation because these two methods obtain very close results for each comparative analysis.

Even if the performances observed here are not as good as those observed in Section 3.2.1, where the data were MCAR and distributed over the whole column, the random forests, IPCA, Amelia, and MIPCA algorithms are also the most efficient here.

3.3.3 Impact of successive missing data at the end of the series

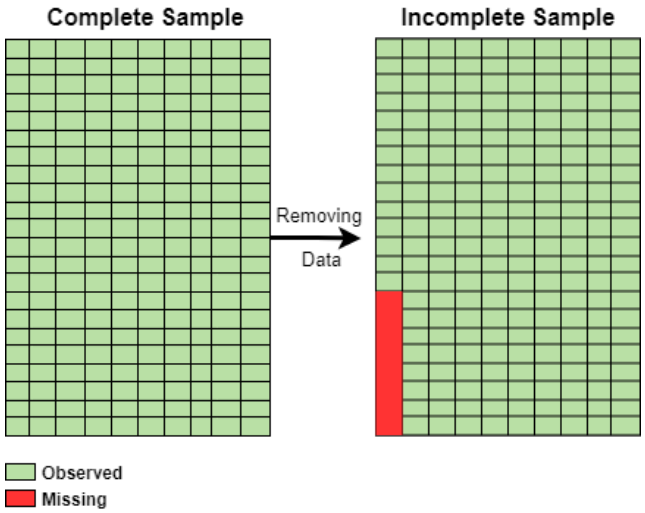
Finally, the last kind of MAR data tested in this chapter is very similar to the one tested in the previous section (see Section 3.3.2), since it consists of successive missing data but on an extremity of the series. The idea here is based on the fact that the series may not have data available before its IPO, nor after its buyback or exit from the market, which would cause missing data at the beginning or the end of the data history. This missing data mechanism can also be due to the inaccessibility or the unavailability of the data (this is the case with data from Totem Markit consensus, which are available only for its contributors). Finally, this pattern can also take place following the implementation of a data historization process, where the data are saved every new day but the recovery of past data depends on past databases that do not always exist. This is very common in financial databases and is a pattern widely observed in the historical sample used in the following section (section 3.5).

Missing data mechanism

Successive data were removed from the first column of the sample. As a reminder, the sample here is the same as the one used in the previous sections (namely Section 3.2.1, Section 3.2.2, Section 3.3.1, and Section 3.3.2). As before, it is a sample made up of 10 columns and 261 observations.

To integrate MAR data corresponding to successive data at the end of the series, the data were successively deleted, starting from the end of the first column of the data matrix. Thus, proportions of data ranging from 5% to 70%, in increments of 5%, were removed from the sample, successively and starting from the end of the series, as shown in Figure 3.3-17.

Fig. 3.3-17: Successive missing data at the end of the first column



Here, there could be the same discussion as in the previous section (see Section 3.3.2) regarding the categorization of the missingness mechanism. Depending on which definition of missingness is used, the missing data from this mechanism could be MCAR according to Little and Rubin [145] or MAR according to Schafer and Graham [182]. The fact that the missing data mechanism is for the middle or the end of the series does not change this argument. Thus, as in the previous section, missing data will be categorized as MAR because it is a much broader category than MCAR.

Since this MAR mechanism does not depend on a random component, a unique missingness scenario is given for each proportion of missing data. In addition, the data that are missing for a missingness scenario set at 5% are also missing for higher missingness proportions.

As before, the deletions were made directly on the spot price matrix, but the missingness proportion based on prices is exactly the same as the one based on returns, as presented in Table 3.3-3.

Tab. 3.3-3: Average proportion (number) of missing returns (among the 100 missingness scenarios) associated with the proportion of MAR raw data (based on successive missing data at the end of the series) injected into the first column of the simulated sample of length 261 (260 for return sample)

	Proportion (and number) of missing returns associated with missing data													
Data	**5%** (13)	**10%** (27)	**15%** (39)	**20%** (53)	**25%** (65)	**30%** (79)	**35%** (91)	**40%** (105)	**45%** (117)	**50%** (131)	**55%** (143)	**60%** (157)	**65%** (169)	**70%** (183)
Return	5% (13)	10% (27)	15% (39)	20% (53)	25% (65)	30% (79)	35% (91)	40% (105)	45% (117)	50% (131)	55% (143)	60% (157)	65% (169)	70% (183)

Finally, each comparison tool is computed as defined in Figure 3.1-2. The templates of all the graphs presented in this section have already been presented and detailed (i.e., what they represent and how they were obtained) in Section 3.1.4.

MCAR tests

The expected results of the MCAR tests for this type of MAR data are the same as those of Section 3.3.2, given the similarity of these two missing data mechanisms. These tests are based on the categorization of Little and Rubin [145]; hence, the purpose here is to accept the null hypothesis that data are MCAR.

As previously, the tests applied to the price returns run without error, and the results are very close to those of Section 3.3.2: Jamshidian and Jalal's test [123] does not reject the null hypothesis that the data are MCAR in all cases, and the same is true for Little's test [142] in almost all cases, except for two missingness proportions (15% and 45%).

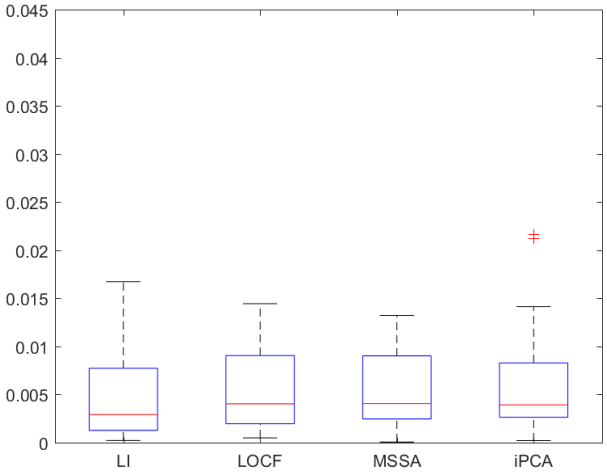
Generally, the tests do not reject the null hypothesis that the data are MCAR when the data are actually removed from the sample based on a uniform law (see Section 3.2.1 and Section 3.2.2). For this kind of MCAR mechanism (according to Little and Rubin [145]), the tests also efficiently accept the null hypothesis, especially Jamshidian and Jalal's test [123].

Since this PhD thesis focuses on the imputation of missing data, further study of these tests will be done in the future. These tests are the two main ones, but other tests also exist (e.g., Kim and Bentler's test [131]), and they can be applied to other simulated samples to compare the results for this MAR data mechanism.

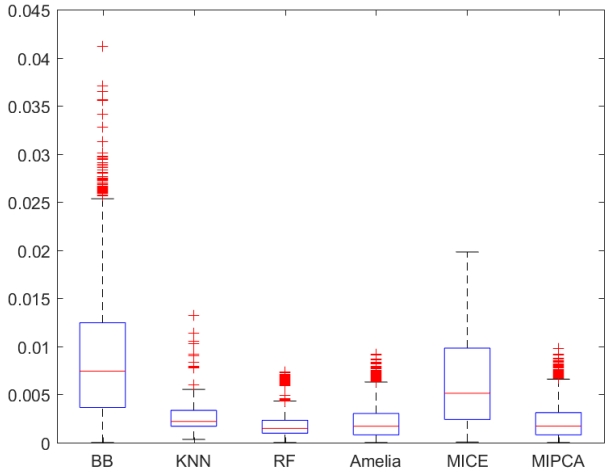
Preliminary results

To begin this analysis, Figure 3.3-18 represents the absolute deviations for series containing 10% (top) and 30% (bottom) of successive missing data at the end of the series. Absolute differences are realized between the returns of the original series and those of the imputed data to have an idea of the distribution of these differences for given proportions of missing data, depending on the method used.

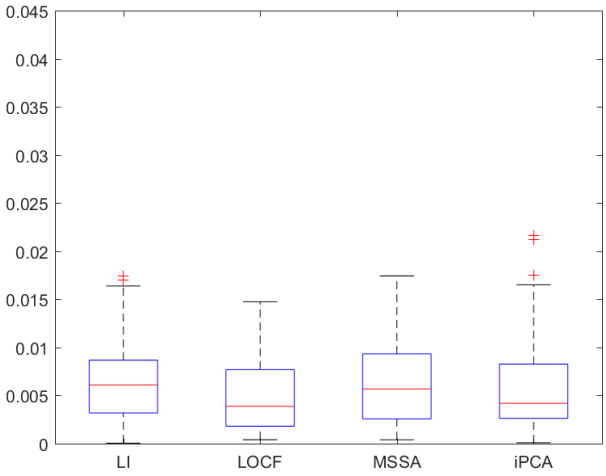
Fig. 3.3-18: Distribution of absolute return differences between the imputed series and original series for a sample containing 10% (at the top) and 30% (at the bottom) MAR data (successive missing data at the end of the first series)



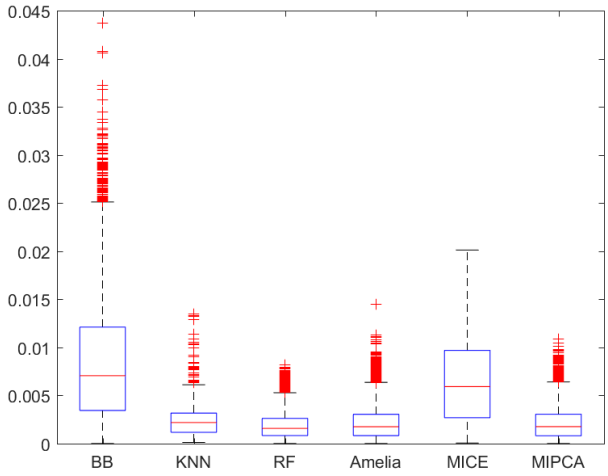
(a) Methods without a random component for 10% missingness



(b) Methods with a random component for 10% missingness



(c) Methods without a random component for 30% missingness



(d) Methods with a random component for 30% missingness

Although the outliers (red crosses) are not considered here, the results are comparable to those obtained with successive missing data in the middle of the series presented in the previous section. The fact that the successive data are located in the middle or at the end of the series only has an impact on the outliers of these absolute differences distributions.

This is true for all the methods applied here, except for the IPCA algorithm, which obtains a distribution of absolute deviations three times larger in the case where the successive data are located at the end of the series. In Section 3.3.2, the absolute deviations were less than 50 basis points, whereas here they can be as high as 150 basis points. Thus, this method could result in a significant deterioration of these results compared with those in the previous section.

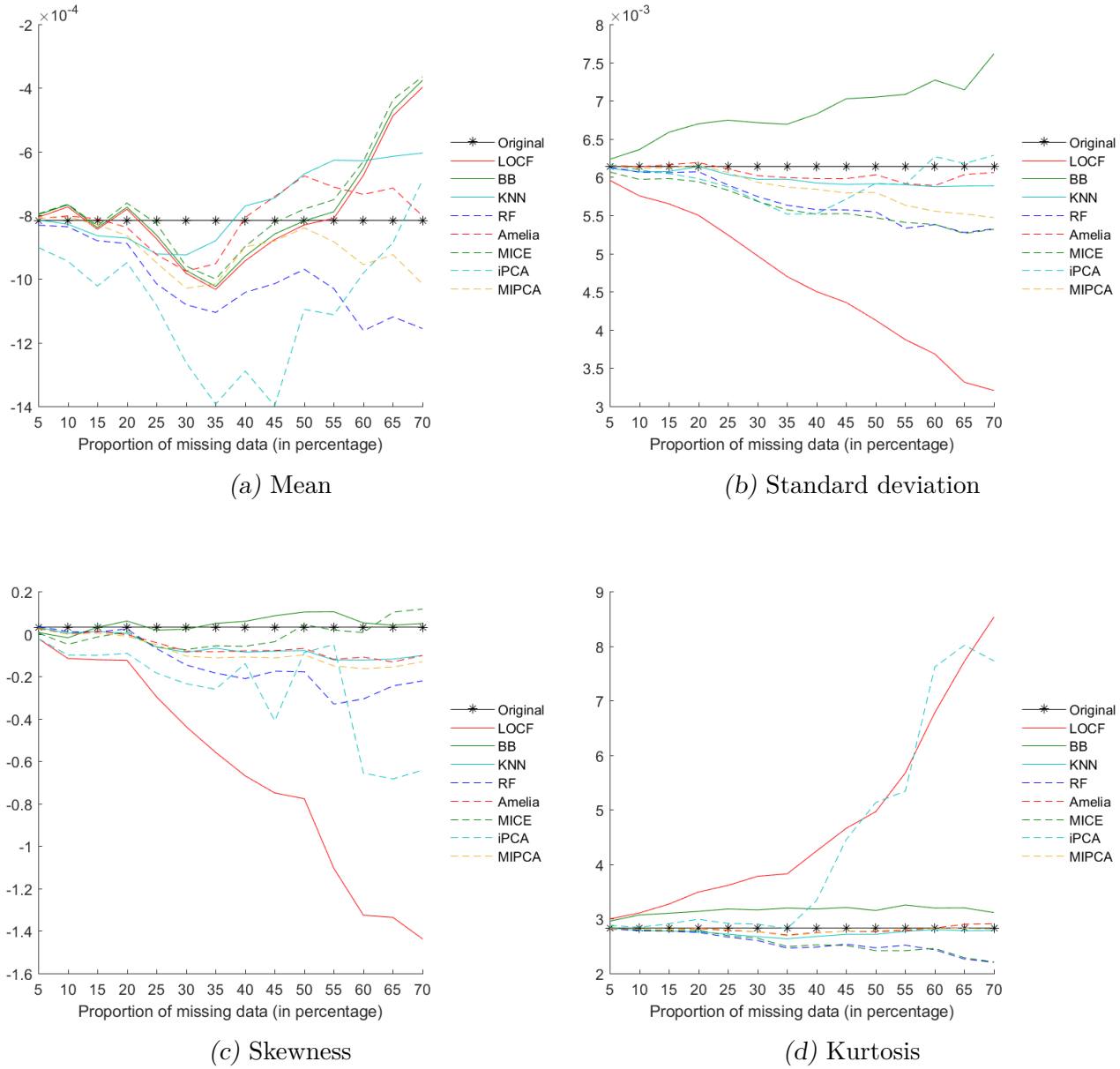
Statistical moments

As in the previous sections, the first four statistical moments have been calculated here from the return of imputed data for each missingness proportion. These results are presented in Figure 3.3-19 for all except two methods: linear interpolation and MSSA.

These two methods have been removed from the analysis in this section because beyond 40% of MAR data distributed successively at the end of the series, they lead to extreme results, making an analysis of the other algorithms (graphically) impossible. Bad results were expected for linear interpolation since a linear extrapolation is used to fill in missing data at the end (but also at the beginning) of the time series. The higher the proportion of missing data is, the more dangerous the imputation by linear extrapolation will be. This missing data mechanism is frequently observed in practice,

and the use of linear extrapolation, in this case, can have catastrophic consequences for the bank. Moreover, linear interpolation and MSSA already obtained results very close to each other in Section 3.3.2, so it is not surprising that this is the case here as well. Nevertheless, Appendix G.1 contains the complete figures, with the results of the linear interpolation and MSSA methods.

Fig. 3.3-19: The first four statistical moments of the returns of the imputed data based on a matrix containing MAR data (successive missing data at the end of the first series) according to the missingness proportion



(a) Mean

(b) Standard deviation

(c) Skewness

(d) Kurtosis

While the results of the previous MAR mechanism were very similar to those of the previous sections, the results here are totally different. In this instance, it appears that the methods impute by reducing the volatility of the series; previously, the opposite had been observed.

In the previous section (see Section 3.3.2), the IPCA algorithm was one of the best-performing algorithms when it came to preserving the statistical moments of the series. Here, its performance has strongly degraded, as expected from the previous results (see Figure 3.3-18). The IPCA underestimates the mean and the standard deviation of the original series, skews the distribution, and explodes its kurtosis for a proportion of missing data greater than 35%. As this analysis is performed on a single sample, the reason could be the sample, which is why it would be interesting to repeat the same procedure on other simulated data.

Moreover, the random forests algorithm was among the most efficient at preserving the moments of the original series, whereas here, the algorithm leads to an underestimate of the set of statistical moments.

By contrast, the Amelia, MIPCA, and K-NN algorithms obtain relatively comparable statistical moments here. These methods are the ones that remain closest to the original series in terms of statistical moments, even if the results obtained are less satisfactory than those of the previous section.

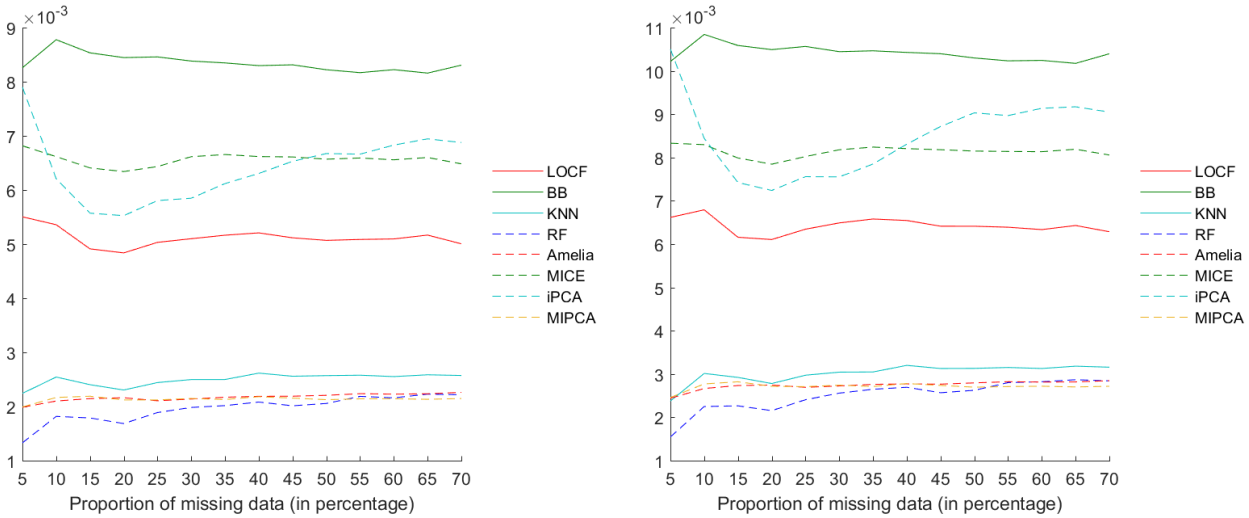
Furthermore, the MICE algorithm and the Brownian bridge method deviate much more from the skewness and kurtosis of the original series when successive missing data are positioned in the middle of the sample than at the end of it.

Finally, the LOCF method remains among the methods that distort the series the most in terms of statistical moments. The linear interpolation and MSSA methods (which can be seen in Appendix G.1) obtain even less satisfactory results than LOCF, even for a small proportion of missing data.

Proximity metrics

Concerning the proximity of the imputed series with the original one, the MAE and RMSE have been computed. These proximity measures are presented in Figure 3.3-20 for all methods, except for linear interpolation and MSSA (the original figures are available in Appendix G.3). Proximity measures are computed here by using the returns of the original series and those of the imputed series.

Fig. 3.3-20: MAE and RMSE between the return of the imputed data from a matrix containing MAR data (successive missing data at the end of the first series) and the original data matrix according to the missingness proportion



(a) MAE *(b)* RMSE

The random forests, Amelia, MIPCA, and K-NN algorithms obtain proximity measures that are close to each other and also very satisfactory. These four methods obtain the best performance in terms of proximity measures, as in the previous case, where successive data were removed from the middle of the sample. Among these methods, the best one remains the random forests algorithm that minimizes both proximity measures for any missingness proportion, but the same analysis should be repeated on other samples to confirm this result.

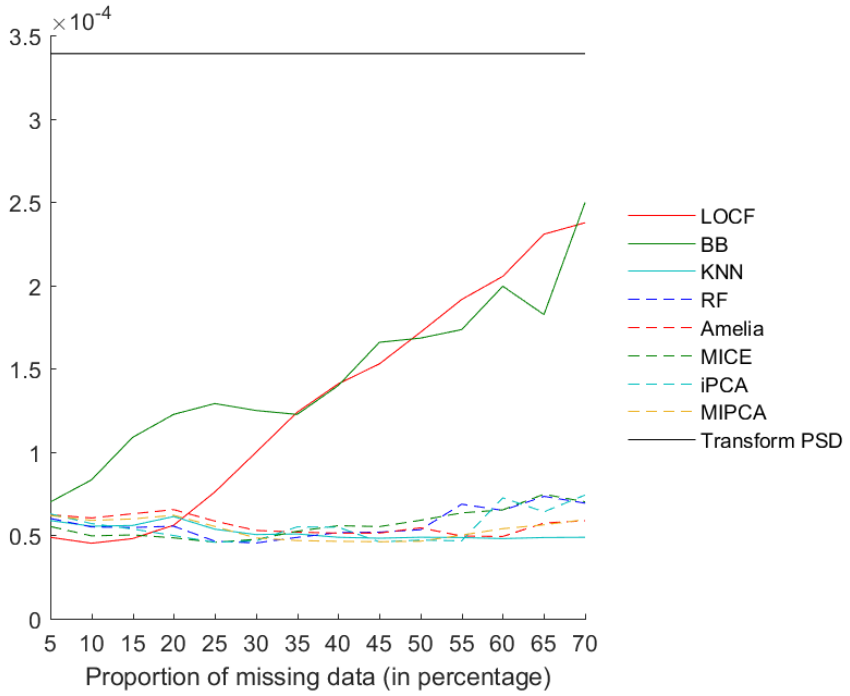
By contrast, in the previous section, the IPCA method obtained proximity measures comparable to random forests, whereas here, the results have clearly deteriorated. This method uses the same number of principal components as MIPCA (see Appendix G.2) but obtains proximity measures comparable to those of the MICE algorithm, which itself is far from a satisfactory result. Moreover, the IPCA, MICE, and Brownian bridge methods obtain less satisfactory results than the usual LOCF method.

As before, the linear interpolation and MSSA methods obtain less satisfactory results than the LOCF method when the proportion of missing data is less than 50% and explodes beyond this threshold (see Appendix G.3).

Covariance matrices comparison

Figure 3.3-21 represents the differences between the original covariance matrix and those from imputed data, based on a Frobenius norm. Thus, it allows us to see the impact of completion methods on covariance matrices. As previously, linear interpolation and MSSA were excluded from these results, but the original figure remains available in Appendix G.4.

Fig. 3.3-21: Covariance matrix differences, according to the Frobenius norm, based on original returns and the imputed returns from a matrix containing MAR data (successive missing data at the end of the first series) according to the missingness proportion



All the methods used here obtain covariance matrices closer to the original series than those from the transformed pairwise matrix (to make it positive semidefinite). Nevertheless, the LOCF and Brownian bridge methods are still the least effective at preserving the original covariance matrix. Their deviation from the original covariance matrix becomes more and more important as the missingness proportion increases.

By contrast, the other methods obtain relatively similar results. Imputing missing data by using K-NN, random forests, Amelia, or MIPCA (or even MICE or IPCA) leads to comparable covariance differences. The results presented here are more stable,

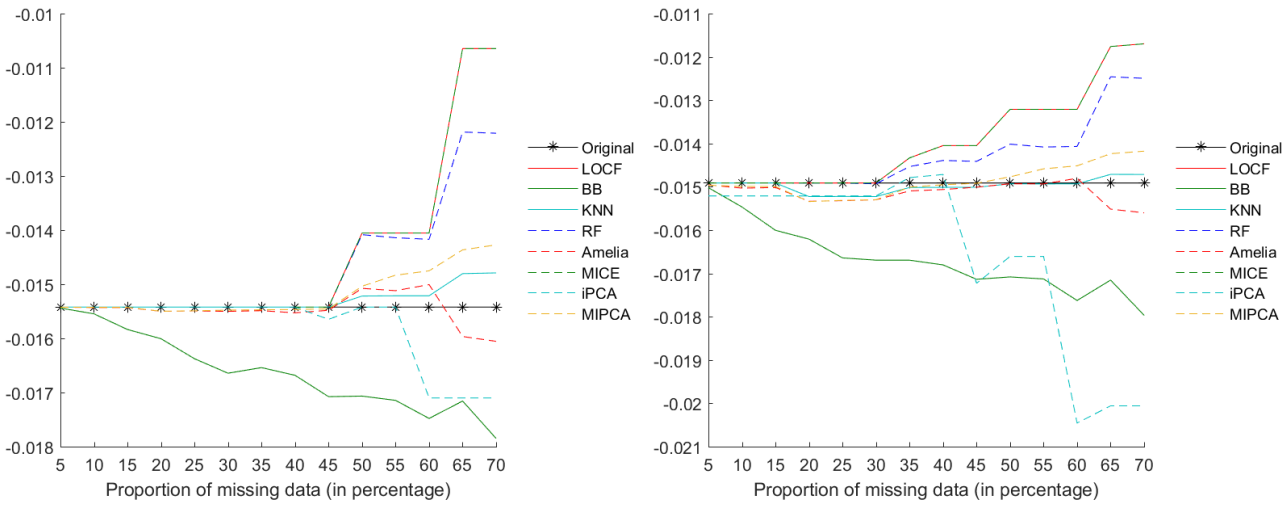
from one missingness proportion to another, than when the missing data were located in the middle of the series.

Although the results obtained by the linear interpolation or MSSA methods are not readable (see Appendix G.4), their results are comparable to previous methods when the proportion of missing data is below 35%.

Value-at-risk and expected shortfall

This type of MAR data can also have a strong impact on risk measures. Figure 3.3-22 represent the VaR and ES levels at a 1-day horizon and for a confidence level of 99% and 97.5%, respectively, calculated from the imputed data. The risk measures of linear interpolation and MSSA have been excluded here but remain available in Appendix G.5.

Fig. 3.3-22: The 1-day risk measures computed from a matrix containing MAR data (successive missing data at the end of the first series) according to the missingness proportion



(a) 1-day VaR at 99% *(b)* 1-day ES at 99%

The Brownian bridge method obtains risk measures that deviate very quickly from those of the original series, and this was already observed in this chapter for the 1-day risk measures. Thus, the behavior obtained here is comparable to the results obtained in the previous Section 3.3.2 or even with MCAR data in Section 3.2.1.

All other completion methods (even linear interpolation and MSSA in Appendix G.5) result in VaRs that are very close to each other when the proportion of missing

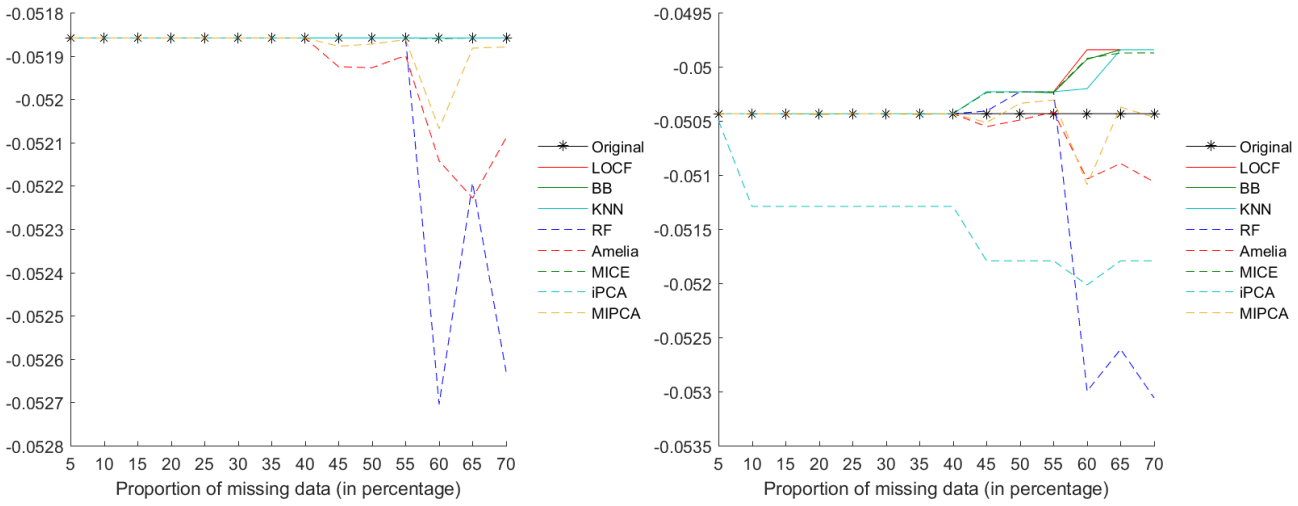
data is below 45%. Beyond this threshold, all the methods tend to deviate from the true VaR level. The K-NN and Amelia algorithms are the ones that stay around the true VaR level when the proportion of missing data becomes too high.

Regarding, the ES, it appears that all the methods (except the Brownian bridge but including the linear interpolation and the MSSA presented in Appendix G.5) are able to obtain an ES close to that of the original series for a missingness proportion below 35%. Beyond this proportion, only the MIPCA, Amelia, and K-NN algorithms report an ES close to its true level; this is particularly true for the K-NN algorithm (which was, however, unsatisfactory in the previous section).

These results depend on the sample used. It seems that, here, the value of the VaR corresponds to the observed data, and the methods' imputation of extreme values will distort this risk measure when the proportion of data is large. The same logic is true for ES. The results for another sample will likely be very different from these, so future research will examine this matter.

Finally, the same risk measures were calculated for a 10-day horizon. These results are presented in Figure 3.3-23 for all the completion methods except for linear interpolation and MSSA, which remain observable in Appendix G.5.

Fig. 3.3-23: The 10-day risk measures, computed from a matrix containing MAR data (successive missing data at the end of the first series) according to the missingness proportion



(a) 10-day VaR at 99% *(b)* 10-day ES at 97.5%

The same type of behavior observed for a 1-day horizon can also be observed for a 10-day horizon: The VaRs of all completion methods are close to each other when

the proportion of missing data is below 45%, and so is the ES (except for the IPCA method) for a threshold of 45% as well.

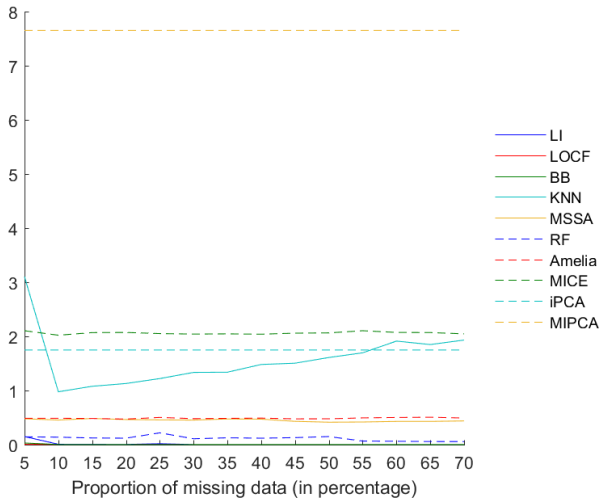
Beyond this threshold, the Amelia, MIPCA, and random forests algorithms can strongly underestimate the 10-day risk measures. The other methods (i.e., LOCF, BB, K-NN, and MICE) are close to the true level of the VaR but slightly overestimate (by almost 5 basis points) the level of ES for the highest missingness proportion, at least for this sample.

Again, this study must be repeated on several samples of data to draw valid conclusions.

Computation time

Finally, the computation time (in seconds) required for each completion method is presented in Figure 3.3-24.

Fig. 3.3-24: Computation time of the imputation of MAR data (successive missing data at the end of the first series) according to the missingness proportion



The computation times presented here are comparable to those presented in Section 3.3.2 in the case where missing data were pooled in the middle of the series.

However, there is a difference: The IPCA requires much more computation time here than before. Here, the algorithm requires almost 2 seconds to complete a sample against approximately 0.2 seconds in Section 3.3.2. The same algorithm already has a calculation time of this order when the sample contained MCAR data throughout the sample for a relatively high proportion of missing data, or in the case of MAR

data depending on extreme values of another column (where the algorithm requires approximately 1.5 seconds).

Thus, having the missing data at the very end of the column leads to relatively different results than in the previous section, where the missing data were in the middle of the column.

Linear interpolation, which extrapolates, clearly appears to be a poor option when dealing with missing data, especially above a certain proportion of missing data. Moreover, as in the previous section, MSSA obtains results comparable to those of linear interpolation, which makes choosing this method pointless.

The Brownian bridge, as well as MICE, obtains unsatisfactory results – sometimes even less than the LOCF method. Nevertheless, these methods have been among the least efficient since the beginning of the chapter. By contrast, while IPCA obtained rather satisfactory results in the previous section, here it tends to deviate from the original series, sometimes even with a small proportion of missing data. These results still need to be confirmed through a study on several simulated samples.

Conversely, the K-NN method, which in the previous section performed poorly when imputing successive missing data in the middle of the column, imputes them more efficiently when they are at the end of the column – hence, the interest of a study based on several simulated samples. Finally, as in previous sections, random forests, Amelia, and MIPCA remain among the best performers.

This missing data mechanism is interesting to observe and study because it highlights the proximity between missing data issues and prediction issues. Although the missing data in this section correspond to data from the past, they could also correspond to data from the future as predicted. Thus, the imputation of missing data or prediction of future data face the same challenges, and it is clear that the Amelia, MIPCA, random forests, and K-NN methods are the best adapted to a simulated sample such as this one.

3.4 Imputation of data: MNAR on simulated Gaussian sample

The last missing data category still to be tested is MNAR data. This section will be dedicated to the impact of missing data following an MNAR mechanism. As a reminder, missing data are considered MNAR according to Little and Rubin [145] if they are neither MCAR nor MAR or if the missing data depend on the observed and missing values of the series. This definition is different from that of Schafer and Graham

[182], who consider MNAR data to be dependent on both the partially missing column and other columns.

As mentioned in the previous section, knowing the mechanism of the missing data necessarily influences the choice of imputation method to be used. However, it is not always easy to understand that the missing data are not randomly distributed in the case of a historical sample, given that there is missing information. This is why it can be interesting to analyze the performance of the algorithms when faced with this type of missing data.

The missing data mechanism used in this section is close to the one used in the previous section, as it consists of removing extreme values for a given proportion.

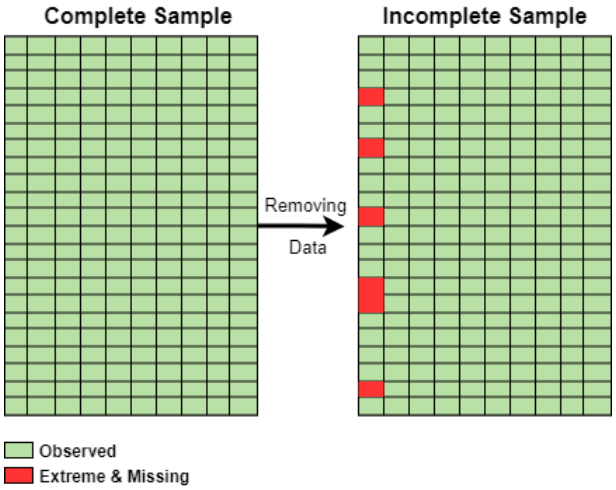
Extreme values mechanism

The data analyzed and suppressed in this section will be those of the first column of the data matrix, contrary to the previous section (Section 3.3.1), where it concerned the last and first column of the sample, respectively. Thus, the procedure aims to determine the extreme values in the first column of the sample before removing them, as presented in Figure 3.4-1.

In a financial framework, missing data can coincide with large variations, as in the case of a trading halt due to a tender offer or a request from authorities to protect investors from potential price fluctuations. Trading halts can hide strong price variations, which would correspond exactly to this missing data mechanism.

This missingness mechanism can also be observed during the bank's data quality checks. Tests are applied to the financial data to ensure that no outliers or unrepresentative values are present in the databases. This can have a serious impact on risk measures or the P&L. Therefore, the banks frequently proceed to data quality controls to remove all outliers or unrepresentative values. This means that an extreme value can be voluntarily removed by the expert and replaced by a more reliable value.

Fig. 3.4-1: Missing data are extreme values of the first column



The missing data of this mechanism are in line with the MNAR definition of Little and Rubin [145], since the whole series is needed to understand that the missing data correspond to the extreme values of the column. Data are MNAR according to Little and Rubin [145] if the missingness depends on observed and missing data. Here, observed and missing values are needed to understand that missing data are the extreme values of the column.

This mechanism is also classified as MNAR according to Schafer and Graham [182] as it depends on a column partially observed and another column that is completely observed and relates to the first one. Considering that the first column is highly correlated (95%) with the second column of the data matrix, the extreme values of the first column may correspond to the extreme values of the second one. Hence, the missingness depends on the first column having partially missing data but also on the second column (and maybe more), which is highly correlated with the first one.

As in previous sections, different proportions of missing data will be removed from price returns to observe the impact of an increasing proportion of MNAR data. Since this MNAR mechanism does not depend on a random component, a unique missingness scenario is given for each proportion of missing data. The value of this proportion will be between 5% and 70%, in increments of 5%, giving 14 samples to analyze. For example, for a probability of missing data set at 5%, the 2.5% most negative returns and the 2.5% most positive returns make it possible to identify the 5% of prices to be removed. As in Section 3.3.1, the data that are missing when the sample contains 5% missing data are also missing for any other sample with more than 5% missing data, and so forth.

Thus, the proportion of missing returns corresponding to the proportion of missing prices is, here, exactly the same as in Section 3.3.1 and are presented in Table 3.4-1.

Tab. 3.4-1: Proportion (number) of missing returns (among the 100 missingness scenarios) associated with the proportion of MNAR raw data (based on extreme returns) injected into the first column of the simulated sample of length 261 (260 for return sample)

	Proportion (and number) of missing returns associated with missing data													
Data	**5%** (12)	**10%** (26)	**15%** (38)	**20%** (52)	**25%** (64)	**30%** (78)	**35%** (90)	**40%** (104)	**45%** (116)	**50%** (130)	**55%** (142)	**60%** (156)	**65%** (168)	**70%** (182)
Return	9% (23)	19% (49)	26% (68)	35% (91)	42% (110)	50% (130)	55% (144)	63% (163)	68% (177)	76% (198)	80% (210)	84% (217)	88% (228)	92% (240)

By removing 70% of the prices at the origin of the extreme readings, 92% of the returns disappear from the sample. Such levels of missing data are excessive but realistic knowing the NMRF eligibility tests, presented in Section 1.3.4, where 24 real price observations are required.

The sample used here is the same as the one used in Section 3.2.1, Section 3.2.2 and Section 3.3 (and presented in Section 3.1.1). The data matrix is composed of 10 columns and 261 observations, where the columns are more or less correlated with each other.

Finally, each comparison tool is computed as defined in Figure 3.1-2. The templates of all the graphs presented in this section have already been presented and detailed (i.e., what they represent and how they were obtained) in Section 3.1.4.

MCAR tests

Before imputing the missing data, MCAR tests are applied to see whether they can reveal that the data are not MCAR.

Little's [142] and Jamshidian and Jalal's [123] tests are applied to the return matrices to see and compare their conclusions. All the tests were applied to the 14 return matrices containing missing data ranging from 5% to 70%. These tests did not have any problems being calculated with this missingness mechanism.

Regarding the results, Little's test [142] does not reject the null hypothesis that the data are MCAR for each missingness proportion. Thus, regardless of the proportion of missing data, this test is unable to detect that missing data correspond to the extreme returns of the series. The results obtained with Jamshidian and Jalal's test [123] are ambiguous: The null hypothesis that the data are MCAR is rejected when the missingness proportion is between 10% and 35%; by contrast, when this proportion is 5%

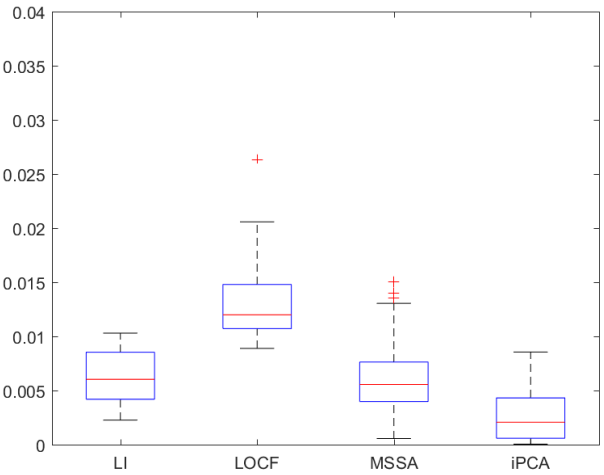
or greater than or equal to 40%, the test does not reject it. This test is partly able to recognize that the data are not MCAR (in fewer than half of the missingness scenarios). The application on a larger number of samples would undoubtedly make it possible to see more clearly; therefore, this issue will be studied in more detail in the future.

Even if Jamshidian and Jalal's test [123] accepts the null hypothesis for some missingness proportions, it is the most efficient test here because Little's test [142] always accepts it. The results need to be put into perspective because they are based on a unique simulated sample.

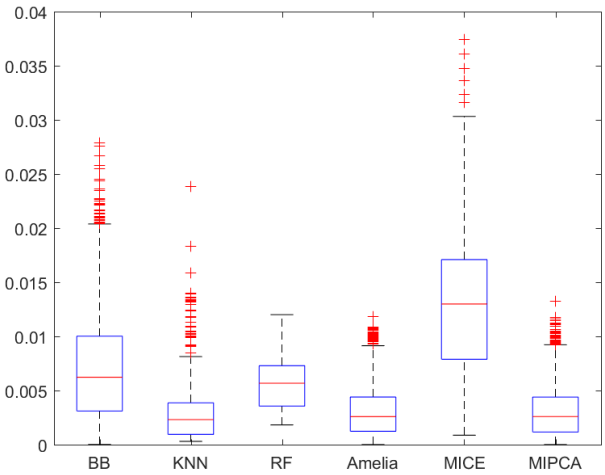
Preliminary results

First, Figure 3.4-2 represents the absolute differences between the returns of the original series and those of the imputed series for each of the methods for scenarios where 10% (at the top) and 30% (at the bottom) of the extreme values have been removed from the series.

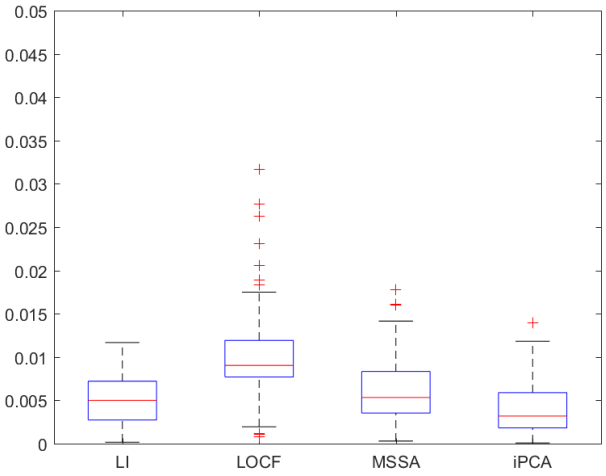
Fig. 3.4-2: Distribution of absolute return differences between the imputed series and original series for a sample containing 10% (at the top) and 30% (at the bottom) MNAR data (extreme values of the first series)



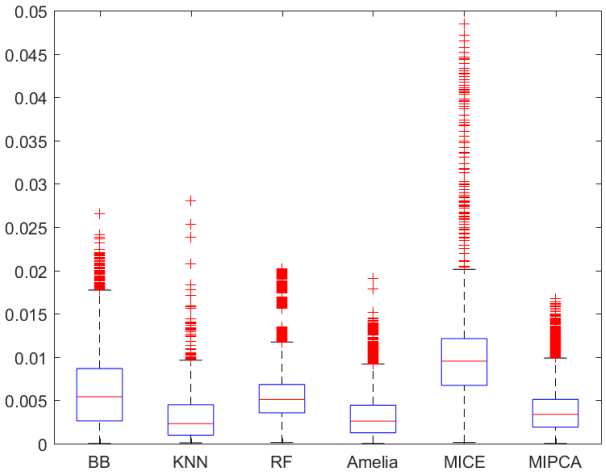
(a) Methods without a random component for 10% missingness



(b) Methods with a random component for 10% missingness



(c) Methods without a random component for 30% missingness



(d) Methods with a random component for 30% missingness

The results obtained by some of the methods above are quite different from those obtained in Section 3.2.1 where the algorithms were imputing MCAR data or those obtained in Section 3.3 with MAR data.

The LOCF method obtains larger absolute deviations when the data are MNAR, compared with results obtained for MCAR or MAR data. This was expected since the missing data are the extreme values of the distribution. The method replaces these extreme returns with zero returns, which implies high absolute return differences. In addition, the distribution for the sample with 30% missing data starts at zero, unlike the one with 10%, which means that not only extreme values are removed from this sample.

In the case of MCAR data, the random forests method was the most stable with absolute deviations close to zero. Now, the MNAR data specifically concerns the extreme values of the distribution, which considerably degrades the performance of this algorithm. The performance had already been significantly degraded in the case where the data were MAR depending on the extreme values of another column (see Section 3.3.1), but the data being MNAR further intensifies this phenomenon.

The K-NN method produces results comparable (excluding outliers of the distribution represented by red crosses) to those obtained with MCAR data. Nevertheless, there is more imputation far from the original series when the data are MNAR than with MCAR data, which reveals some extreme values were imputed incorrectly.

The other methods generally obtain absolute differences comparable to those observed in the case of MCAR data. This is the case of the Amelia and MIPCA algorithms, which obtain distributions comparable to each other (including outliers) and to the distribution of K-NN (excluding outliers).

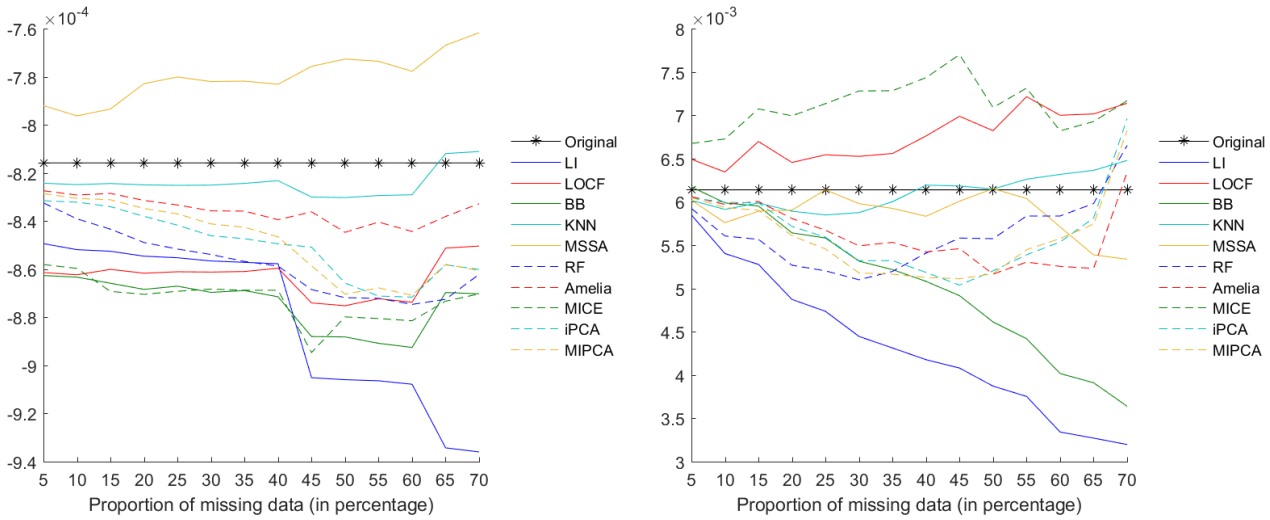
The MICE and Brownian bridge methods obtain the distributions with the largest differences from the original series. These results are not surprising given that the Brownian bridge uses the available data from a truncated distribution with no information on the tails, and all of the realistic potential donors for the MICE algorithm are missing.

Based on these preliminary results, it appears that the random forests algorithm is less efficient here, leading to K-NN, Amelia, or even MIPCA to perform better. Therefore, the average results for all proportions of missing data need to be analyzed to see the impact of this kind of MNAR data for any proportion of missing data.

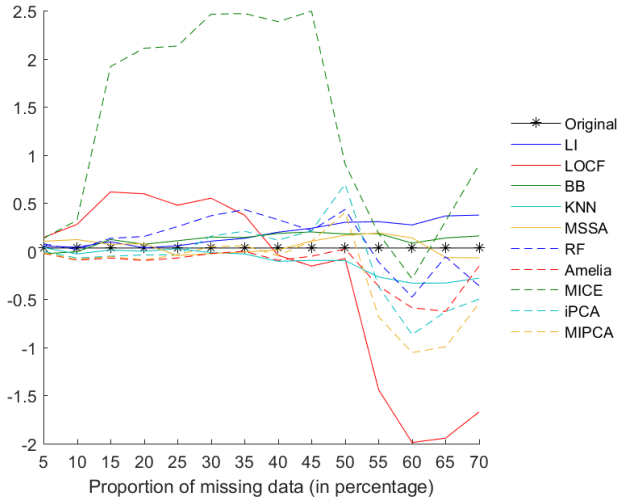
Statistical moments

Figure 3.4-3 represents the first four statistical moments of the imputed series to see the impact of removing the tails of the distribution on the imputation quality.

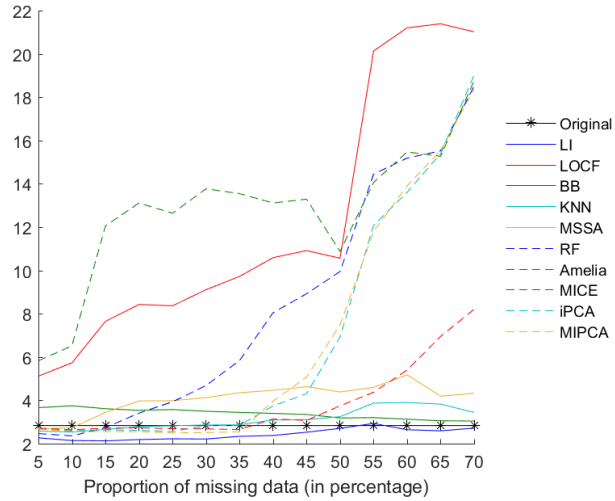
Fig. 3.4-3: The first four statistical moments of the returns of the imputed data based on a matrix containing MNAR data (extreme values of the first series) according to the missingness proportion



(a) Mean *(b)* Standard deviation



(c) Skewness



(d) Kurtosis

First of all, since the missing data are extreme values, they have a direct impact on the mean of the series, even with a low proportion of missingness. In the previous sections, the mean of the imputed data progressively deviated from that of the original series, as the missing data appeared in the sample, whereas here, the means are far from that of the original series as soon as a low proportion of missing data appears (see Figure 3.2-5a). Many imputation methods are far from the original mean starting at 5% missing data and tend to deviate further and further away as the missingness proportion increases. This is related to the mean's lack of robustness. This would probably not be the case with the median. Nevertheless, this phenomenon reveals the difficulties that completion methods can have in imputing extreme values. To some extent, this same phenomenon of the mean can be observed for the standard deviation and the kurtosis.

Some methods appear to be very unsuccessful in preserving the statistical moments of the original series. This is notably the case of the MICE algorithm which deviates very strongly from the original series for each of the four statistical moments. This is not surprising, since the suppression of extreme values is done by proportion and not one by one, which leaves fewer meaningful donors for the MICE algorithm (working longitudinally). For example, for 5% missing data, the MNAR mechanism will remove the 6 most negative returns as well as the 6 most positive returns, thus missing data correspond actually to close original data.

The usual methods are also very unsatisfactory here. Linear interpolation considerably reduces the volatility of the series (as always), as well as its mean. The LOCF method, by contrast, increases its volatility, reduces its mean, and leads to a leptokurtic distribution when the proportion of missing data becomes too large.

Although the Brownian bridge can preserve the original skewness and kurtosis, it underestimates the mean and volatility of the series, which is logical since the tails have been deleted.

Contrary to the results observed in the case of MCAR data, the random forests algorithm distorts the distribution of the series by imputing this kind of MNAR data. This algorithm leads to exploding the kurtosis of the series while underestimating its mean and standard deviation. This is probably related to the decision trees, which do not make any assumption of linearity or normality. It is very likely that if the missing data did not concern all the extreme values but only some of them, the random forests would be as efficient as when using the other missing data mechanism previously presented.

Moreover, its results are comparable to (and sometimes worse than) those obtained by the IPCA and MIPCA methods. Nevertheless, despite their poor results for a high proportion of missing data, IPCA and MIPCA are relatively efficient in terms of skewness and kurtosis, when the proportion of missing data is less than 40%.

Although the mean from MSSA is above that of the original series, this method is quite effective at preserving the level of volatility (up to 50% of MNAR data), as well as the skewness and kurtosis of the series.

The Amelia algorithm slightly underestimates the mean and volatility of the series, but it is very effective at preserving the skewness and kurtosis of the original series when the proportion of missing data does not exceed 45%. However, the kurtosis obtained by Amelia for higher proportions of missing data tends to increase. This could be explained by the fact that the bootstrapped samples are usually very different from each other when the proportion of missing data is important, which leads the algorithm to estimate a Gaussian mixture distribution rather than a simple Gaussian distribution. This can, indeed, cause bias on the kurtosis, which is generally higher than 3 for Gaussian mixtures.

Finally, the most stable (from one missingness proportion to another) and closest method to the original series, regardless of the proportion of missing data in the sample, is the K-NN method. The four statistical moments from the data imputed by K-NN remain close to those of the original series, no matter how much missing data are present in the sample. The way the data are simulated matters a lot in the construction of the weights needed for K-NN imputation, and removing the tails of distributions disturbs it very little compared with other methods. The weights are defined from the observed data, which correspond to the center of the distribution and then used to impute the tails of the distributions, which works nicely considering how the sample was simulated.

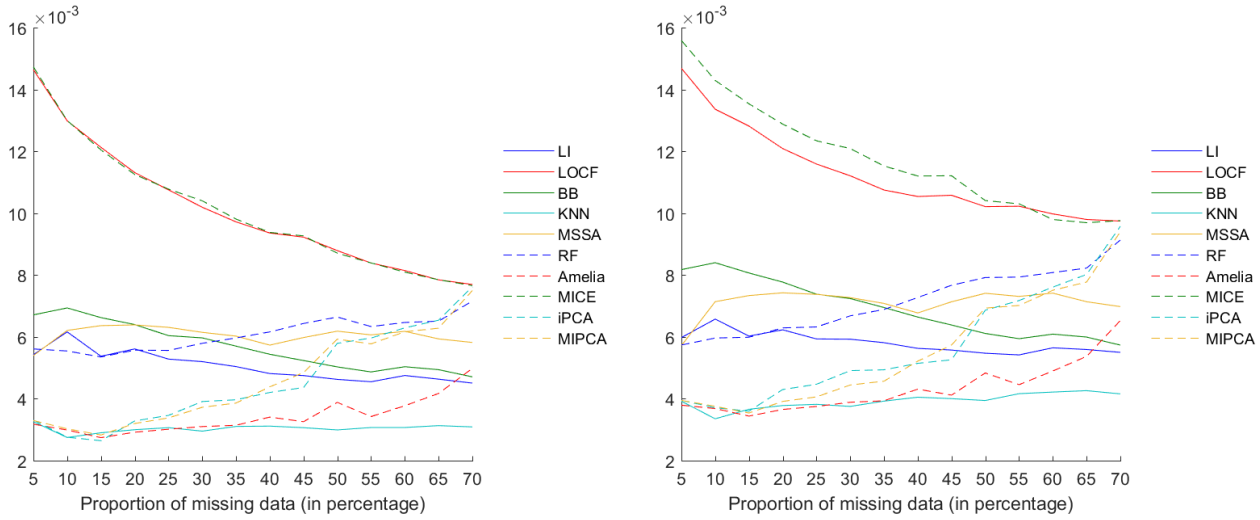
Overall, it is clear that imputation methods have much more difficulty imputing missing data as closely as possible to the original series. This type of MNAR data, which distorts the distribution of returns in particular, makes it difficult to preserve

the true statistical moments. Of course, these results are based on a single sample and must be put into perspective. Nevertheless, they illustrate well the difficulties of imputation that this type of missing data mechanism can cause. An in-depth study using a large number of simulated samples will be carried out in future research.

Proximity metrics

Another way to observe the performance of completion methods is by using proximity measures. Figure 3.4-4 represents the MAE and RMSE to highlight the proximity between the returns of the original series and those of the imputed series for these kinds of MNAR data (by removing the extreme values).

Fig. 3.4-4: MAE and RMSE between the return of the imputed data from a matrix containing MNAR data (extreme values of the first series) according to the missingness proportion



(a) MAE *(b)* RMSE

The first remark concerns the decreasing trend of proximity measures for some methods. The MICE and LOCF methods obtain smaller and smaller proximity measures as the sample contains missing data, whereas, in the previous sections, the opposite behavior was represented. This is because when few missing data are present in the sample, the proximity measures are high as the few imputations made are very far from the original series, but if more data are missing in the sample, these extreme imputations have less weight in the MAE and RMSE. As a reminder, the data that are missing when the missingness proportion is set at 5% are also missing for any other proportion.

In addition, the proximity measures obtained for MNAR data are up three times larger than when the data were MCAR.

Moreover, since the beginning of this chapter, the random forests method appears to be one of the methods that minimize these proximity measures. However, with this kind of MNAR data, this algorithm obtains poor results and is less efficient than a simple linear interpolation. Thus, with these results and those of the statistical moments, the random forests algorithm appears not to be as suitable for imputing the tails of the distribution, at least for this simulated sample.

The IPCA and MIPCA algorithms also obtain very close results here, even though they do not use the same number of principal components (three for IPCA and two for MIPCA, as presented in Appendix H.1). Moreover, these methods are as efficient as the K-NN and Amelia algorithms at minimizing the proximity measures for a missingness proportion below 20%. By contrast, when the sample contains 20% or more of missing data, the proximity measures of IPCA and MIPCA reveala a greater and greater distance from the original series, even leading them to be less efficient than a simple linear interpolation when the proportion of missing data is higher than 45%.

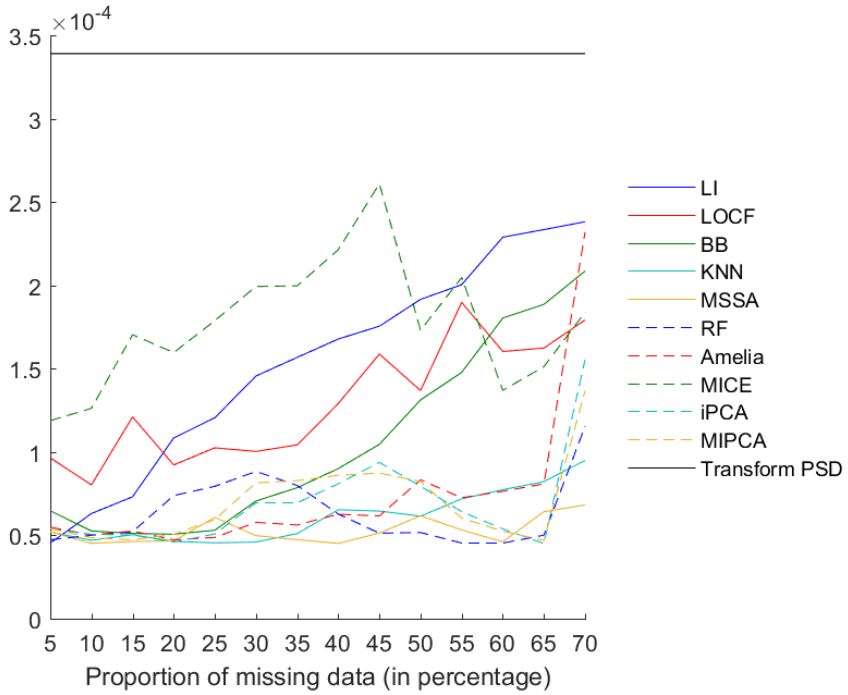
Finally, the K-NN algorithm appears to be the most stable from one missingness proportion to another and imputes the missing data efficiently, whether they are in large numbers or not. These results are confounded with those of the Amelia algorithm when the proportion of missing data is below 35%, but above this threshold, Amelia obtains more proximity measures than the K-NN algorithm.

As always, these results are based on a single missingness scenario and are only representative of this particular sample (used since the beginning of the chapter).

Covariance matrices comparison

Figure 3.4-5 represents the differences between the covariance matrix of the original series and the one derived from the imputed data, according to the Frobenius norm, for each of the respective methods. As the missing data were only injected into the first column of the data matrix, the covariance matrix from the imputed sample is impacted only in the first column (and row).

Fig. 3.4-5: Covariance matrix differences, according to the Frobenius norm, based on original returns and the imputed returns from a matrix containing MNAR data (extreme values of the first series) according to the missingness proportion



First of all, the set of methods used here always results in a closer covariance matrix (according to the Frobenius norm) than the positive semidefinite pairwise matrix, which was not the case when the missing data were MCAR (see Figure 3.2-9 from Section 3.2.1) or MAR (see Figure 3.3-5 from Section 3.3.1, Figure 3.3-13 from Section 3.3.2, and Figure 3.3-21 from Section 3.3.3).

The usual methods (linear interpolation and LOCF) are among the methods that distort the covariance matrix the most. However, linear interpolation was effective when dealing with MCAR data with a proportion of less than 20% (see Figure 3.2-9 from section 3.2.1), but not here.

The MICE algorithm and Brownian bridge method also distort the covariance matrix the most because they tend to seriously distort the covariance matrix, which results in differences comparable to usual methods for the highest proportions of missing data.

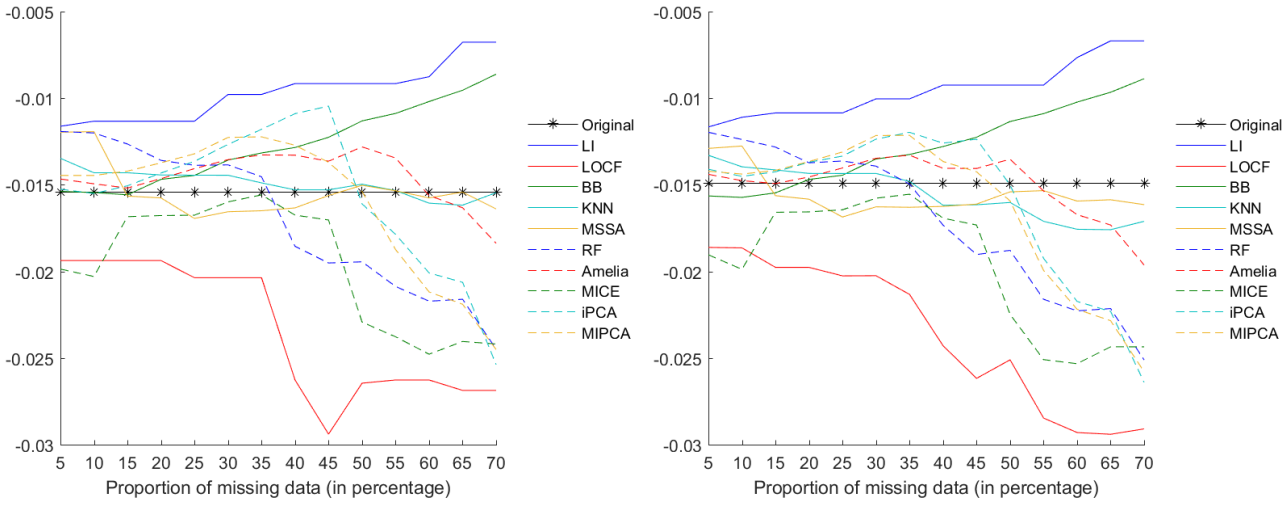
For the other algorithms, namely random forests, IPCA, MIPCA, MSSA, K-NN, and Amelia, they seem to follow a similar trend, with covariance differences between 0.00005 and 0.0001. Moreover, random forests, IPCA, MIPCA, and Amelia tend to

explode in terms of covariance deviation for a proportion of missing data of 70%. However, it is difficult to compare these methods in detail since these results are based on a single sample.

Value-at-risk and expected shortfall

Figure 3.4-6 represents VaR and ES for a 1-day horizon with a confidence level of 99% and 97.5%, respectively, obtained from the data imputed by the completion methods.

Fig. 3.4-6: The 1-day risk measures computed from a matrix containing MNAR data (extreme values of the first series) according to the missingness proportion



(a) 1-day VaR at 99%

(b) 1-day ES at 97.5%

As this missing data mechanism directly targets the extreme values, the deviations on the risk measures are visible from the smallest proportions of missing data. However, it is also possible to see that these risk measures tend to be below their original level as the proportion of missing data increases.

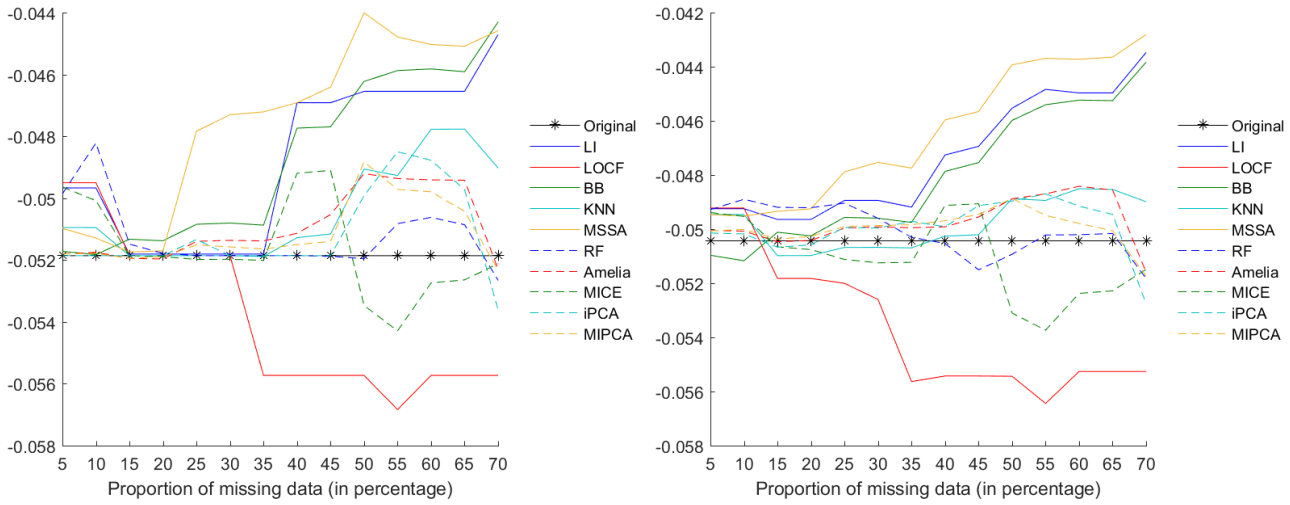
As in the case of MCAR data (see Figure 3.2-11 from Section 3.2.1), the usual methods are the least likely to reproduce the risk measures of the original series, but this is also the case for the MICE and Brownian bridge methods. These methods have been the least efficient since the beginning of the chapter, so these results are not surprising.

As for the other methods, it is not easy to draw general conclusions because the results are obtained from a single sample. However, a downward trend can be observed for the risk measures, which move further and further away from the original series.

The algorithms closest to the risk measures of the original series are K-NN and MSSA. Moreover, the MSSA method was already the most satisfactory method according to these criteria when the data were MCAR (see Figure 3.2-11 from Section 3.2.1).

Finally, Figure 3.4-7 represents the VaR and ES for a 10-day horizon and a confidence level of 99% and 97.5%, respectively.

Fig. 3.4-7: The 10-day risk measures computed from a matrix containing MNAR data (extreme values of the first series) according to the missingness proportion



(a) 10-day VaR at 99% *(b)* 10-day ES at 97.5%

For the 10-day risk measures, the MSSA algorithm loses all its effectiveness (compared with a 1-day horizon), just as it did when the data were MCAR (see Figure 3.2-13 from Section 3.2.1).

Here, the least efficient methods for a 1-day horizon (usual methods, Brownian bridge, and MICE) continue to perform poorly.

The Amelia, IPCA, MIPCA, and K-NN algorithms are quite effective at reproducing the risk measures of the original series when the proportion of missing data is below 45%, but above this threshold, they deviate from the level of the original series. Furthermore, the random forests method is the one that leads to the best performance above this threshold.

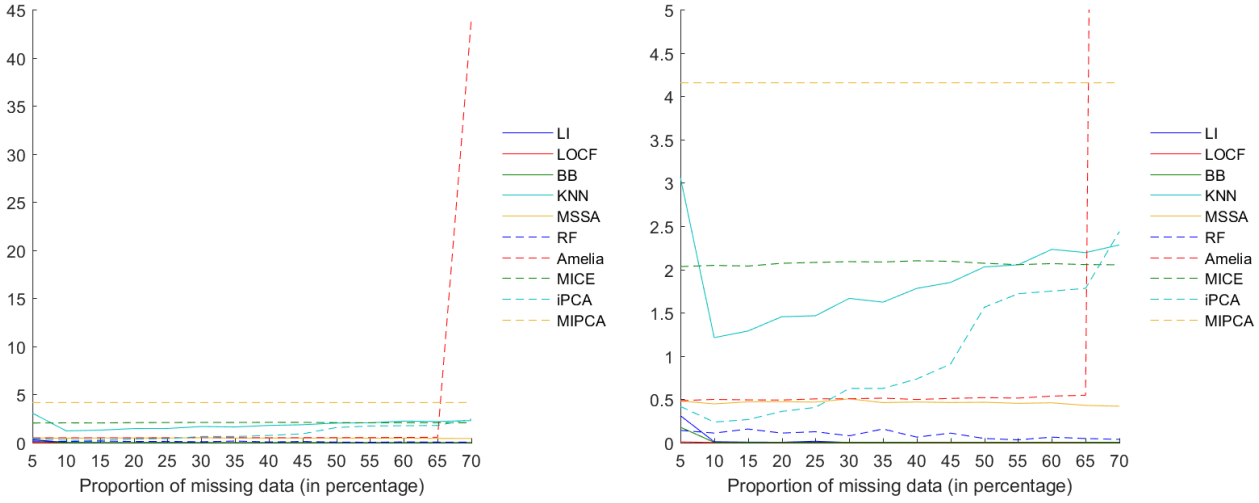
More generally, it can be seen that there is no (or little) impact on the 10-day (but also the 1-day) risk measures for a large number of methods when the proportion of

missing data is below 20%. This observation gives an idea of the validity of the methods and, in particular, of the Amelia algorithm, which is one of the best imputation methods.

Computation time

To complete this study, Figure 3.4-8 represents the computation time (in seconds) required by each algorithm to impute a sample according to its proportion of missing data.

Fig. 3.4-8: Computation time of the imputation of MNAR data (extreme values of the first series) according to the missingness proportion



As in the case of MCAR (see Figure 3.2-14 from section 3.2.1), the calculation time of the Amelia algorithm explodes for the last missingness proportion to be completed, hence the interest of the second graph with a different scale. For the other samples, the Amelia calculation time is similar to the one needed for MCAR data.

The MICE algorithm, random forests, and MSSA have a computation time similar to the one when the data were MCAR (see Figure 3.2-14 from section 3.2.1). However, this is not the case for the MIPCA algorithm, which requires a constant computation time here, corresponding to the highest level observed in the case of MCAR data. The K-NN algorithm also observes a higher computation time for the first imputation (as in Figure 3.3-8 from section 3.3.1), but beyond this first sample, the computation time is the same as in the case of MCAR data (due to the launching of the algorithm).

The purpose of this section is to highlight the impact of MNAR data on the performance of the completion algorithms used. The type of MNAR data used here directly

targets data distribution by truncating it in each of its extremities. The algorithms are misled as the distribution of the missing data is different from that of the observable data.

The methods that were already not performing well in the case of MCAR data are naturally no better in this case of MNAR data. But some algorithms that had previously been efficient are among the least efficient here. The usual methods, Brownian bridge, and MICE, which were already performing poorly, continue to perform poorly in the case of these MNAR data; by contrast, the random forests algorithm, which was one of the best-performing, sees its imputation quality seriously deteriorate.

Here, the K-NN algorithm appears to be the most efficient across many criteria at reproducing the tails of missing distributions. The Amelia, IPCA, and MIPCA algorithms are also satisfactory, especially when the proportion of missing data does not exceed 40%. Finally, the MSSA algorithm obtains rather mitigated results, which may or may not be satisfactory, depending on the analysis criteria observed.

As always, these results must be put into perspective, but since the same sample has been used since the beginning of this chapter, these results illustrate the impact of different missing data mechanisms on the completion methods.

3.5 Imputation of data: MCAR on historical data

This chapter ends with an application of the algorithms to historical data; the goal is to observe their efficiency in practice and to compare them. This last section presents the chosen historical sample, as well as the missing data it contains. Two sub-samples will be created – one based on a heuristic approach, and the other on a graphical Lasso approach – to remove data from an MCAR mechanism and finally apply the set of completion methods.

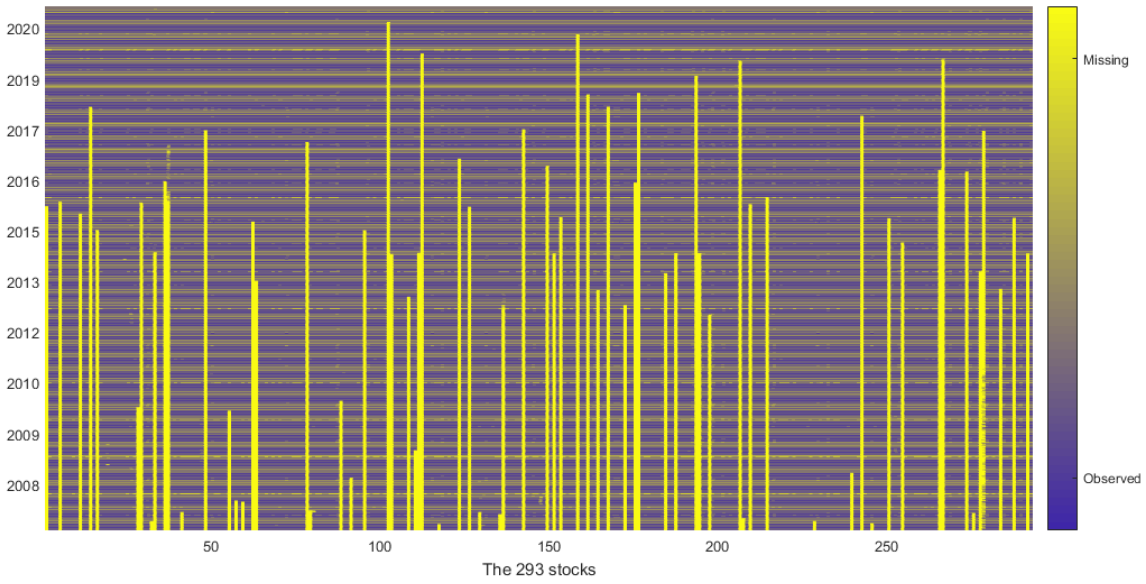
3.5.1 Data presentation

To observe a large sample, the data used here are those of the Euro Stoxx 300, Lyxor's exchange-traded fund (ETF), whose purpose is to replicate the Dow Jones EURO STOXX Total Return Index. This fund was chosen because it reflects the performance of the global equity market in the Eurozone. It has a variable number of components (approximately 315) and includes large, medium, and small caps from 12 countries of the Eurozone.

The data used in this section are those of the component stocks of the Euro Stoxx 300 as of 5 March 2021. On this date, the ETF was composed of exactly 293 shares. The daily historical data of these 293 series from 1 January 2007 to 1 March 2021 were downloaded via Bloomberg L.P. to see the pattern of missing data over a long period.

These downloaded data correspond to a matrix of 5,174 rows, which are all the dates between 1 January 2007 and 1 March 2021 and of 293 columns, which are the 293 stocks constituting the Euro Stoxx 300 as of 5 March 2021. Figure 3.5-1 represents the matrix of the historical prices of the 293 stocks (columns) over the period (rows); the observable data are reported in blue, and the missing data are in yellow.

Fig. 3.5-1: Missing data pattern of the Euro Stoxx 300 from 1 January 2007 to 1 March 2021



The missing data here represent 39% of the total sample, but among these missing data, some are due to market holidays. They are easily recognizable because, for each of these market closing dates, no stocks are quoted: The weekends (Saturdays and Sundays), as well as 25 and 26 December and 1 January, are systematically missing from all the columns. These days lead to a raw with no available information and must be removed because the European stock market is closed and no shares are listed on those dates.

More precisely, among the 5,174 days between 1 January 2007 and 1 March 2021, there are 1,544 days when none of the stocks were listed on the market (empty rows). Of these:

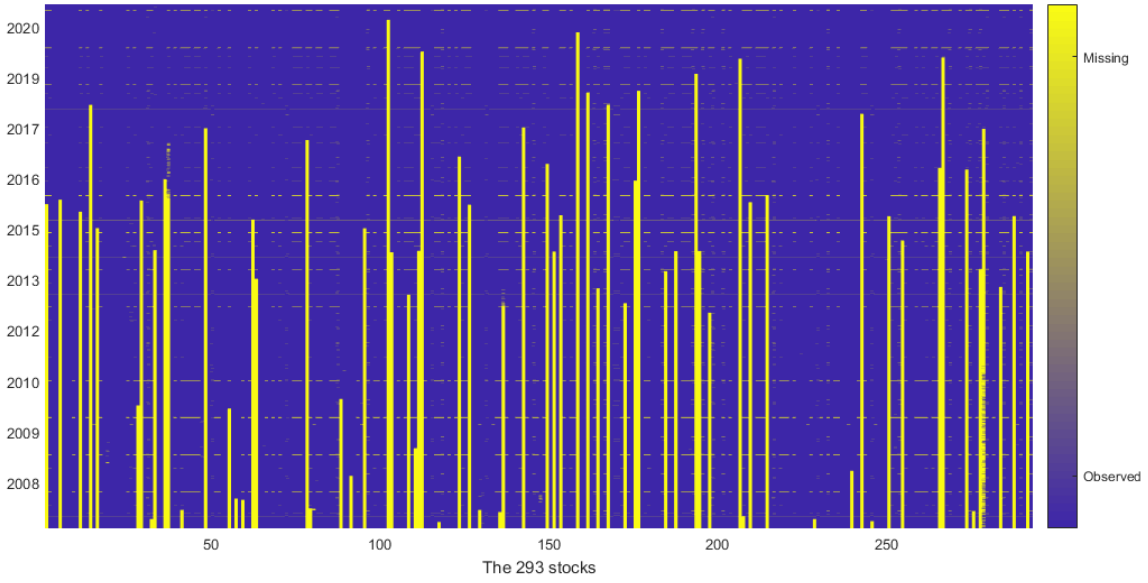
- $1,478$ fall on weekends
- 15 are 1 January
- 28 correspond to the Christmas period (25 and 26 December)

- 14 are 1 May and
- 28 fall on Good Friday or Easter Monday

The missing data do not correspond to the sum of each of the above events, since two of them may concern the same date (for example, 1 January may also be a weekend day). Thus, only one of these events has to happen for the row of this matrix to be completely empty.

Once these 1,544 dates are removed from the sample, there are still 3,630 dates containing at least one missing stock price. The data matrix without the 1,544 dates is presented in Figure 3.5-2, and the missing data are still very present since they represent 13% of this sample.

Fig. 3.5-2: Missing data pattern, excluding stock market holidays, of the Euro Stoxx 300 from 1 January 2007 to 1 March 2021



In this sample, it is possible to observe two types of missing data.

First, for some shares, no observation is available before a certain date. This can be explained by shares that have been included in the index (or have changed their ticker code) during the period. Some stocks have a very large number of missing data at the end of the series (the end corresponding here to the oldest part in the history) like the missing data mechanism analyzed in Section 3.3.3. For example, the share of the German company Siemens Energy has 97% missing data over the period, as it was only introduced on the stock index on 28 September 2020 2020 [189].

It is also possible to observe, through a latitudinal analysis of Figure 3.5-2, that some dates lead to missing data for many stocks. Among these dates are 24 and 30 December, which are market closing eves. Thus, these dates correspond to the closing of the stock market for some countries, like Germany and Italy. This shows that the closing dates can vary from one country to another, which can cause missing data in the sample.

There is a significant amount of missing data in a historical sample such as the Euro Stoxx 300 for data going back to 2007. Without any reprocessing, the sample contains almost 40% missing data; even after the days with no available quotation (corresponding to market holidays) are deleted, some 13% of the sample is still unobserved. This may be due, for example, to country-specific market holidays or to the fact that the shares were not yet listed on the stock exchange. However, for regulatory purposes, the bank must be able to have complete historical data, regardless of the reason for the missing data.

3.5.2 Impact on a sample based on a heuristic approach

Now that the full sample has been analyzed over a long period, the history must be shortened and the stock sample reduced to obtain a well-chosen final sample so that missing data can be injected into it for later completion.

Presentation of the sample based on a heuristic approach

To use the algorithms on a recent period, the final sample is built based on the full historical data of the 293 stocks composing the Euro Stoxx 300 (composition of 5 March 2021) from 1 January 2020 to 1 February 2021, which corresponds to exactly 398 dates (with market holidays). The period is chosen so that it corresponds to a little more than one year, which will make it possible to calculate at least 250 price returns of 10 days for risk measure calculations. This is why the period chosen is a little more than one year.

The first sub-sample from historical data used is based on a heuristic approach, in the same way as an expert might do, to impute missing data from a sample containing series close to each other and with an economic meaning. This heuristic approach is based on the following two steps:

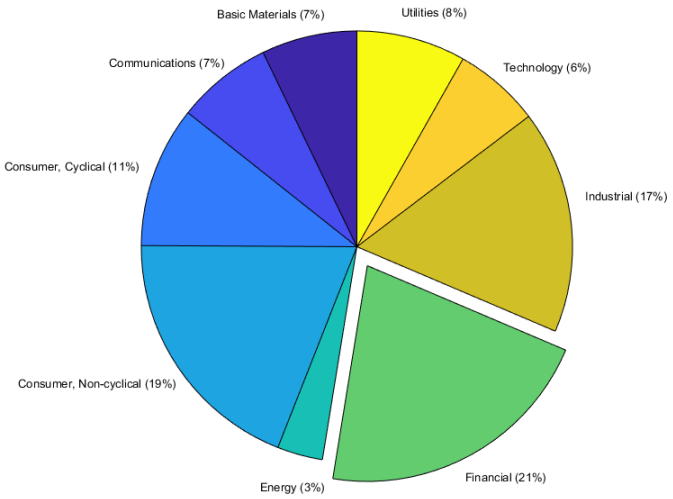
1. select stocks from the same sector
2. select the most correlated stocks

This approach can be related to Kahneman's System 1 [62], discussed in the introduction, since it is intuitive and fast, and it often works (although not always).

This bucket selection is the one used to select this chapter's first historical subsample. It was chosen arbitrarily to have a well-chosen sample. In practice, the stock with missing data to be filled in is known, but here it will be defined ex-post. However, other bucketing processes can also be used, as seen in Section 2.3.

The first selection step concerns the sector categorization of stocks. To see the composition of the ETF by sector, the first level of BICS categorization (presented in Section 2.3 is used here. Thus, the sector composition of the stocks from Euro Stoxx 300, as of 5 March 2021, is represented in Figure 3.5-3.

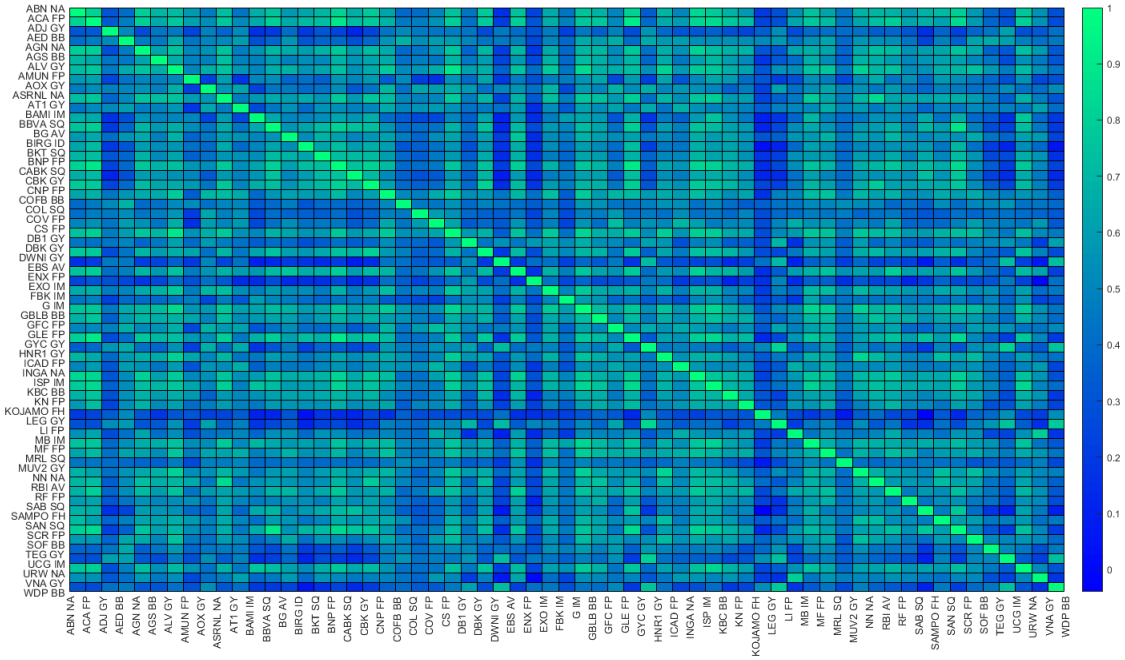
Fig. 3.5-3: Sectors of the Euro Stoxx 300 components



The majority of stocks belong to the "financial", "consumer, non-cyclical", "industrial" and "consumer, cyclical" sectors, with proportions of 21%, 19%, 17% and 11%, respectively. Thus, the most represented sector is the "financial" one, with a total of exactly 62 shares (among 293) contained in the Euro Stoxx 300. Therefore, the stocks belonging to this sector are retained to form a preliminary sample.

The second step is to analyze the pairwise correlations of these 62 stocks. The correlation matrix is represented in Figure 3.5-4.

Fig. 3.5-4: Correlation of stock returns of the 62 stocks from the Euro Stoxx 300's financial sector



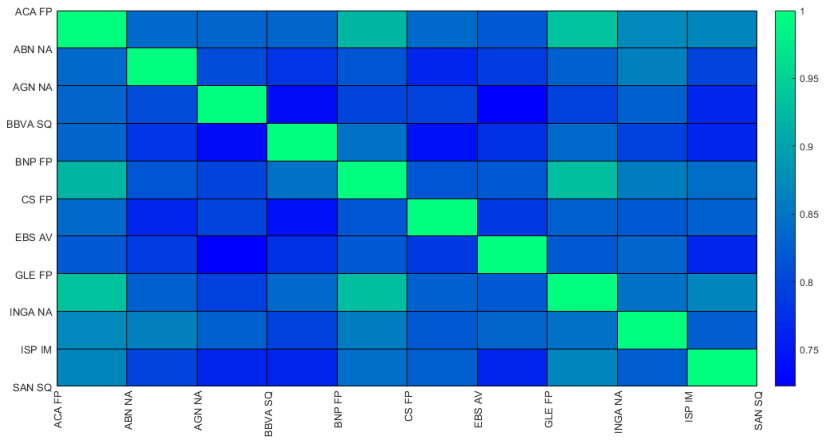
By studying this correlation matrix, it is possible to identify a group of highly correlated actions. This double selection leads to a group of 11 stocks that are from the same sector and correlated with each other at least 70%. The stocks in this historical sample are (the columns of the data matrix are in this specific order):

1. Crédit Agricole (ACA FP)
2. ABN AMRO Bank N.V. (ABN NA)
3. Aegon N.V. (AGN NA)
4. BBVA S.A. (BBVA SQ - Banco Bilbao Vizcaya Argentaria)
5. BNP Paribas S.A. (BNP FP)
6. AXA S.A. (CS FP)
7. Erste Group Bank (EBS AV)
8. Société Générale Bank (GLE FP)
9. ING Groep N.V. (INGA NA)

10. Intesa SanPaolo (ISP IM)
11. Banco Santander S.A. (SAN SQ)

Their correlation matrix is represented as follows (Figure 3.5-5):

Fig. 3.5-5: Correlation of the historical sample based on a heuristic approach

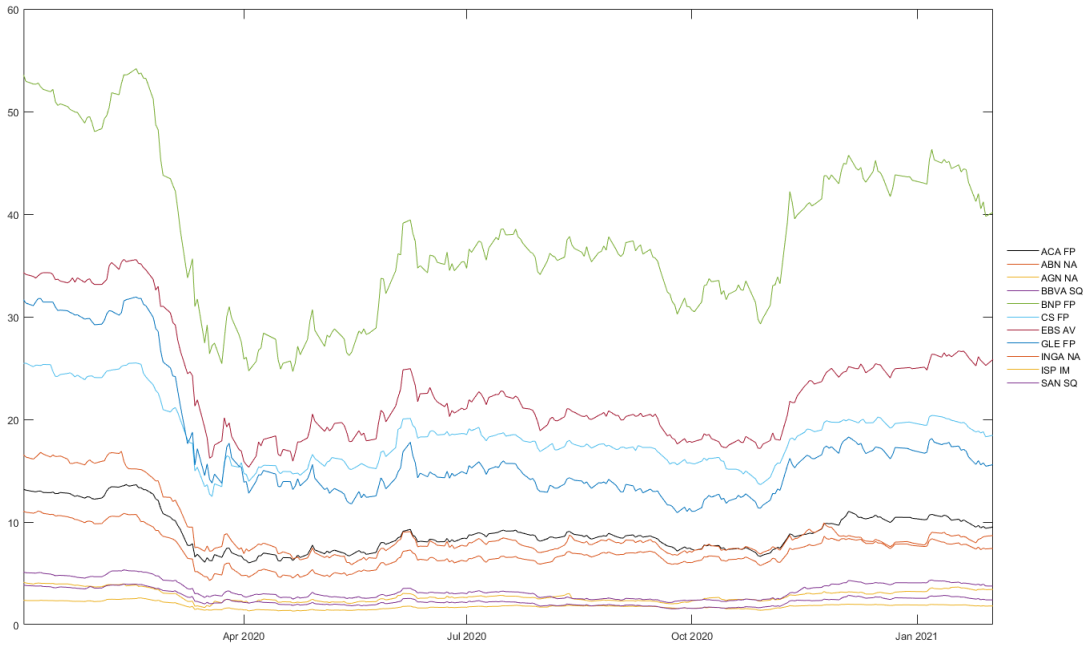


All these shares were introduced on the stock exchange before 1 January 2020, which means none of them have successive missing data since the beginning of the period.

OAfter these stocks have been selected, the data matrix from this sample contains 398 rows, corresponding to the dates between 1 January 2020 and 1 February 2021 and 11 columns associated with the 11 stocks enumerated above. Among these 398 observations, market holidays must be removed, as well as all the dates containing at least one missing data point in order to have a fully observed sample. Before removing the missing data, the MCAR tests of Little [142] and the Jamshidian and Jalal [123] were applied to this matrix. Little's test [142] concludes that missing data from this historical sample are MCAR. However, Jamshidian and Jalal's test [123] is not calculable because, after removing the cases with all variables missing, not enough missing data patterns remained. As previously discussed in Section 3.2.2, the fact that this test generates errors when the distributed missing data are spread over the whole matrix is very problematic, as it is a pattern frequently observed in practice.

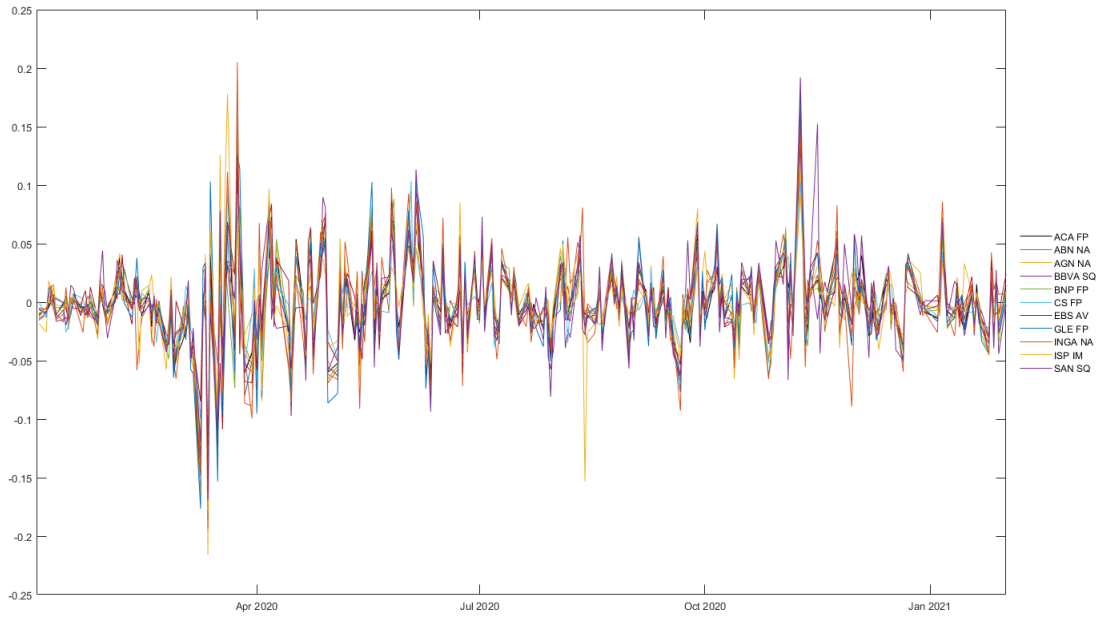
Thus, a listwise deletion 2.4.1 is applied to this sample so that no missing data is present in the sample. Following the listwise deletion, this historical sample is left with 274 dates (rows) and 11 stocks (columns). The evolution of these stocks during the period studied is presented in Figure 3.5-6.

Fig. 3.5-6: Final sample based on a heuristic approach: 11 financial stocks from 1 January 2020 to 1 February 2021



This period is characterized by an initial period of low volatility (until March 2020), followed by a period of higher volatility. This type of heteroskedastic pattern is reminiscent of the one tested in Section 3.2.3; in addition, some series seem to contain jumps, as in Section 3.2.4. This can be seen in Figure 3.5-7.

Fig. 3.5-7: Price returns of the historical sample based on a heuristic approach



Missing data can now be injected into the first column of this historical sample to complete them and compare the performances of the algorithms.

The sample is constructed in such a way that the missing data are injected in the first column, which corresponds to the Crédit Agricole stock prices, and then completed using (or not, depending on the method used) the prices of stocks from the same sector (financial) and highly correlated.

Thus, this historical sample defined above will be used to apply the same missing data mechanism as the one presented in Section 3.2.1. As already mentioned in this chapter, knowing the kind of missing data provides information about the algorithm to use or not. However, while Jamshidian and Jalal's test [123] can detect MCAR data when the data are actually MCAR (see Section 3.2.1 and Section 3.2.2), it does not appear to be as effective at detecting that the data are not MCAR (future studies have yet to confirm this). In other words, it is not easy to know the categorization of the missing data present in a historical sample. That is why, to avoid using a missing data mechanism in this historical sample that is too specific, data will be removed following the same MCAR mechanism presented in Section 3.2.1.

Thus, MCAR data will be injected into the first column of the matrix (corresponding to the Crédit Agricole share prices) to apply the algorithms. The missing data will be injected in different proportions ranging from 5% to 70%, in increments of 5%, and for

each of these proportions, 100 missingness scenarios – each time containing a different MCAR data pattern – will be tested. Thus, the proportion of missing returns associated with missing prices is presented in Table 3.5-1. The orders of magnitude are comparable to those observed in Section 3.2.1, where the MCAR data were injected into a simulated series.

Tab. 3.5-1: Average proportion (number) of missing returns (among the 100 missingness scenarios) associated with the proportion of MCAR raw data injected into the first column of the historical sample based on a heuristic approach of length 274 (273 for return sample)

	Proportion (and number) of missing returns associated with missing data													
Data	**5%** (13)	**10%** (27)	**15%** (41)	**20%** (54)	**25%** (68)	**30%** (82)	**35%** (95)	**40%** (109)	**45%** (123)	**50%** (136)	**55%** (150)	**60%** (164)	**65%** (178)	**70%** (191)
Return	10% (25)	20% (52)	29% (76)	37% (97)	46% (119)	53% (138)	60% (156)	67% (174)	73% (189)	78% (204)	83% (217)	88% (229)	92% (239)	95% (248)

Finally, each comparison tool is computed as defined in Figure 3.1-2. The templates of all the graphs presented in this section have already been presented and detailed (i.e., what they represent and how they were obtained) in Section 3.1.4.

MCAR tests

As in the previous sections, Little's test [142] and Jamshidian and Jalal's test [123] are applied to this first historical sample where MCAR data were artificially added. The proportion of tests that do not reject the null hypothesis that the data are MCAR is presented in Table 3.5-2 for each missingness proportion. These results are based on the tests that were calculable (presented in Appendix I.1).

Tab. 3.5-2: Confidence level (probability of not rejecting H_0 when H_0 is true) for both MCAR tests applied to price return matrices containing MCAR on the first column of the historical matrix based on a heuristic approach for a 5% significance level

	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Little's test	91%	94%	94%	90%	95%	97%	97%	100%	96%	94%	92%	94%	91%	92%
J&J's test	86%	86%	88%	85%	84%	86%	87%	89%	91%	89%	86%	97%	94%	93%

The missing data injected here follow the same missingness mechanism as those

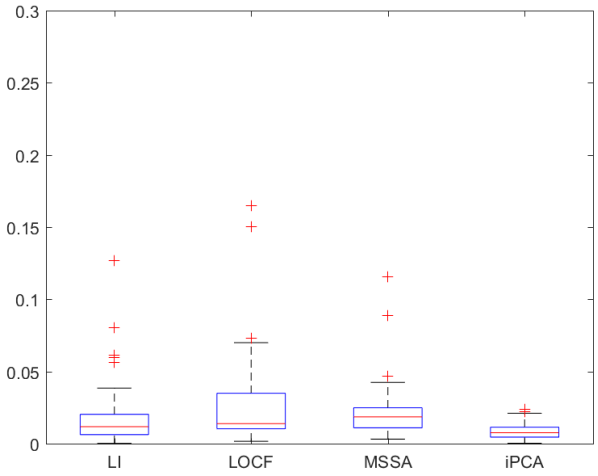
injected in Section 3.2.1; the results are close to those observed in this section. Little's test [142] tends not to reject the null hypothesis that the data are MCAR, with confidence levels between 91% and 100%. Regarding Jamshidian and Jalal's test [123], the results obtained on historical data (between 84% and 97%) are slightly weaker than those obtained on simulated data (between 89% and 99%) but still comparable. However, in Section 3.2.3, it was also shown that the heteroskedasticity of the series had a downward impact on the performance of Jamshidian and Jalal's test [123]. Thus, the differences between the historical sample and the simulated sample can be fully explained by the heteroskedasticity of the historical sample.

Both tests are able to detect the presence of MCAR data. Once the missing data are analyzed, they can finally be imputed to see the impact of each completion method on a sample of historical data.

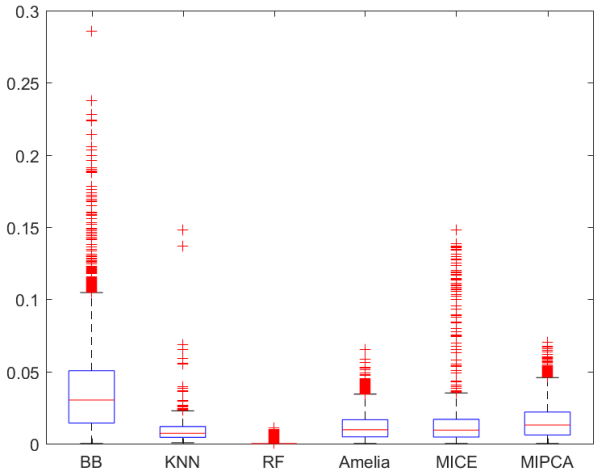
Preliminary results

To have an idea of the stability of each algorithm, all the absolute deviations between the original data and the first scenario containing 10% missing data (at the top) and 30% missing data (at the bottom) are plotted in Figure 3.5-8. These figures show the absolute differences between each imputed return and the original return for the first missingness scenario used. The figures at the bottom are based on three times more absolute differences than those at the top.

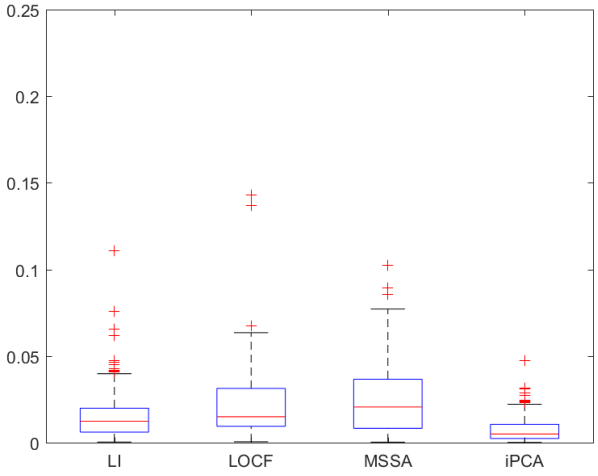
Fig. 3.5-8: Distribution of absolute return differences between the imputed historical series and original historical series (based on a heuristic approach) for a single scenario containing 10% (at the top) and 30% (at the bottom) MCAR data (only in the first column)



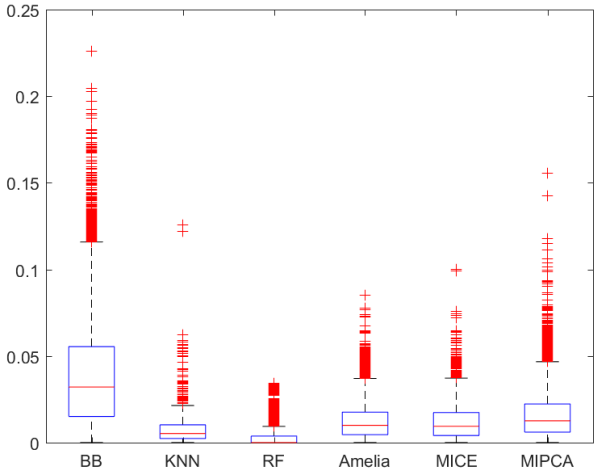
(a) Methods without a random component for 10% missingness



(b) Methods with a random component for 10% missingness



(c) Methods without a random component for 30% missingness



(d) Methods with a random component for 30% missingness

First of all, if the results below are compared with those obtained in the case of simulated data (see Figure 3.2-3 from Section 3.2.1), it is possible to see that the

absolute differences are much larger when the algorithms deal with these historical data. The scales of these figures are, in fact, five to 10 times larger than those used in the case of simulated data. Thus, the historical sample is very different from the simulated one, but it is clear that the algorithms have more difficulty correctly imputing these historical data than is the case with results based on simulated samples since they contain heteroskedasticity and jumps.

These preliminary results point out the performances of the random forests algorithm, which obtains absolute deviation distributions closest to 0. This was already the case in Section 3.2.1. Thus, this algorithm also appears to be efficient when dealing with these historical data (at least for the first missingness scenario).

Here, the Amelia, IPCA, MIPCA, and K-NN methods obtain results comparable to each other and are among the lowest. These results were expected, given these methods' results in the previous sections. However, one of the big differences here is the good performance of MICE which, at least for this missingness scenario, obtains results comparable to those of the previously mentioned methods. The good performance of the MICE algorithm suggests that this algorithm would be more suitable if applied to this historical sample than to the simulated sample used in the previous sections. The sections where the MICE algorithm was the most effective were those where the sample had non-constant volatility (see Section 3.2.3) or jumps (see Section 3.2.4) – in other words, the sections where the samples were modified. Thus, the bad performance of the MICE algorithm may be due to the particular sample that was used, but this remains to be confirmed by later analysis. The MICE algorithm appears to be as efficient as Amelia's algorithm because the (historical) sample has non-constant volatility and jumps.

Brownian bridge, MSSA, and LOCF are the methods with the largest absolute deviations in this missingness scenario.

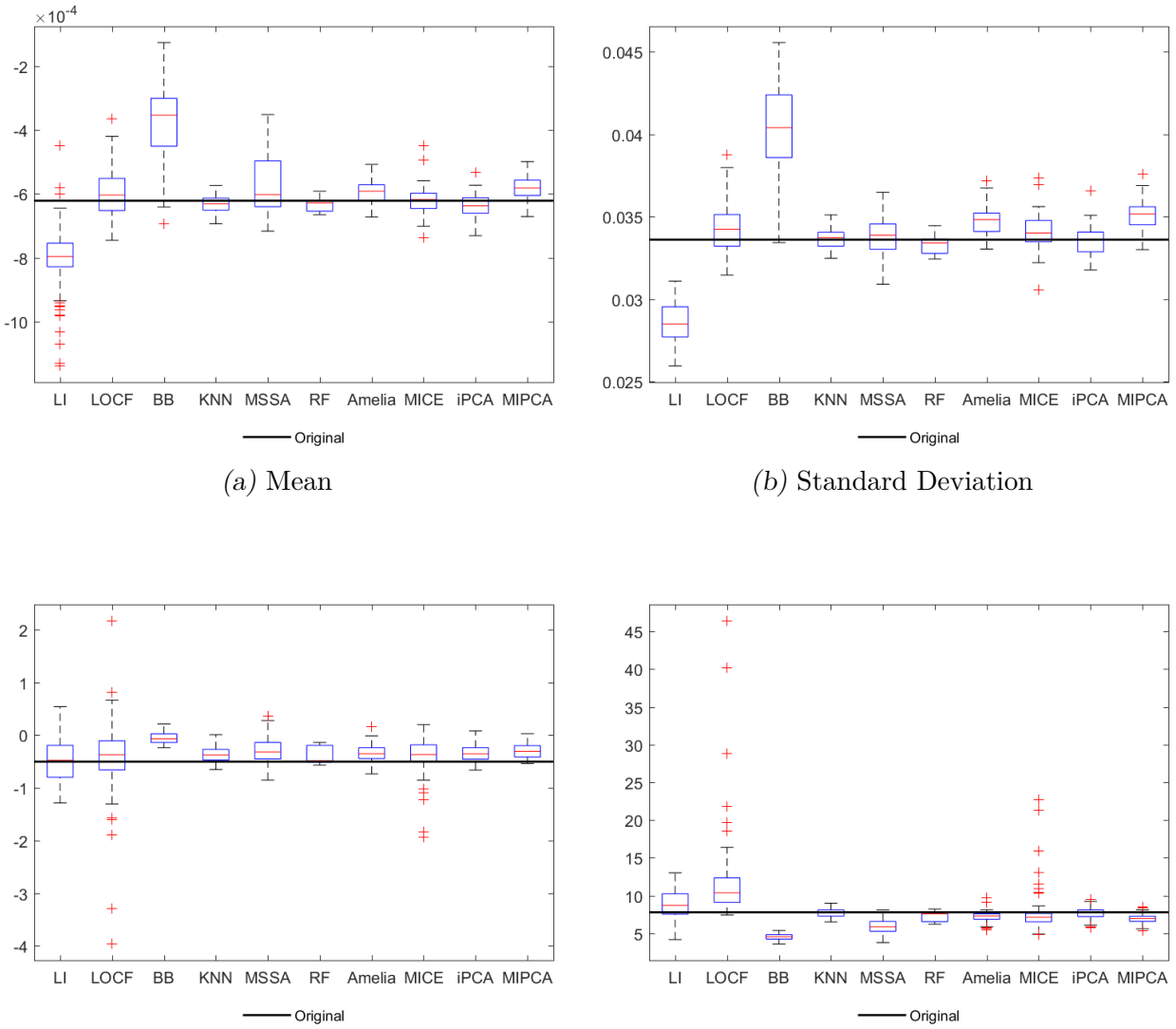
The above results are associated with a single scenario (the first one) for a given missingness proportion. It is now necessary to analyze the results by exploiting all the scenarios used and all the proportions of missing data tested. Thus, as in some previous sections, a comparative analysis of all the scenarios used will be put forward for a fixed level of missing data (at 30%); then, the results obtained among these 100 scenarios will be averaged to compare different levels of missingness.

Statistical moments

Figure 3.5-9 represents the first four statistical moments obtained by each of the completion methods for the 100 missingness scenarios containing 30% missing data. These figures give an idea of the distribution of statistical moments obtained from one miss-

ingness scenario to another. The standard deviations of these distributions are available for all missingness proportions in Appendix I.2.

Fig. 3.5-9: Distribution of the first four statistical moments obtained for the 100 scenarios based on historical sample based on a heuristic approach with 30% of MCAR data (only in the first column)



(a) Mean

(b) Standard Deviation

(c) Skewness

(d) Kurtosis

These figures reveal two categories of algorithms: simple imputation methods without random components and methods with a random component. Linear interpolation, LOCF, Brownian bridge, and MSSA obtain wider distributions than the other algorithms and can deviate significantly from the moments of the original series, especially

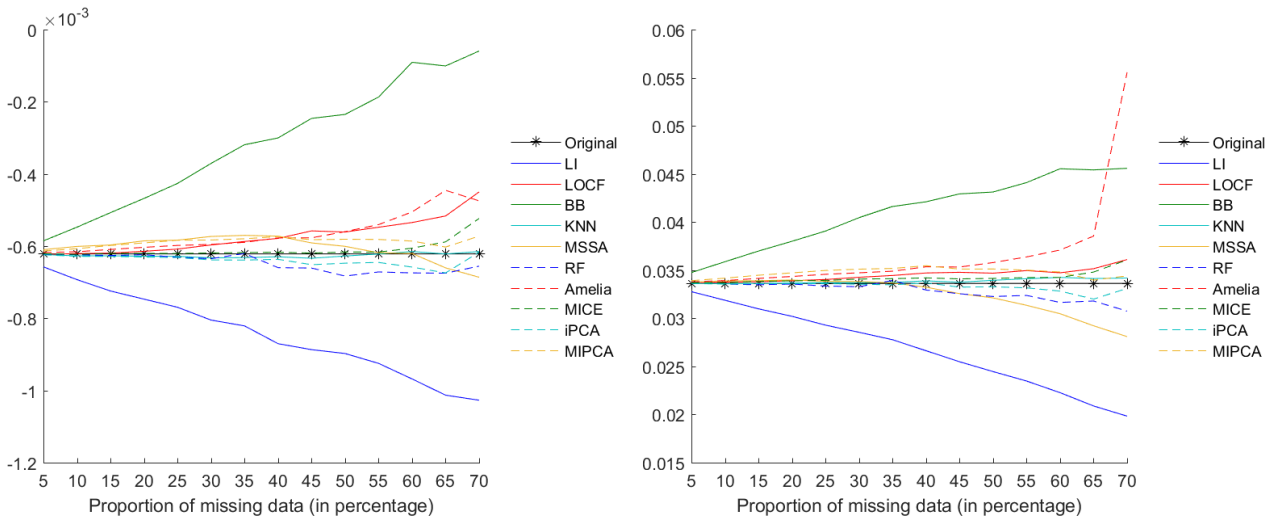
for the mean and standard deviation. Conversely, the other completion methods, which contain a random component, obtain moments relatively close to those of the original series for about all scenarios.

It should be noted that the distributions presented here are much more spread out than those obtained in the case of simulated data, presented in Figure 3.2-4.

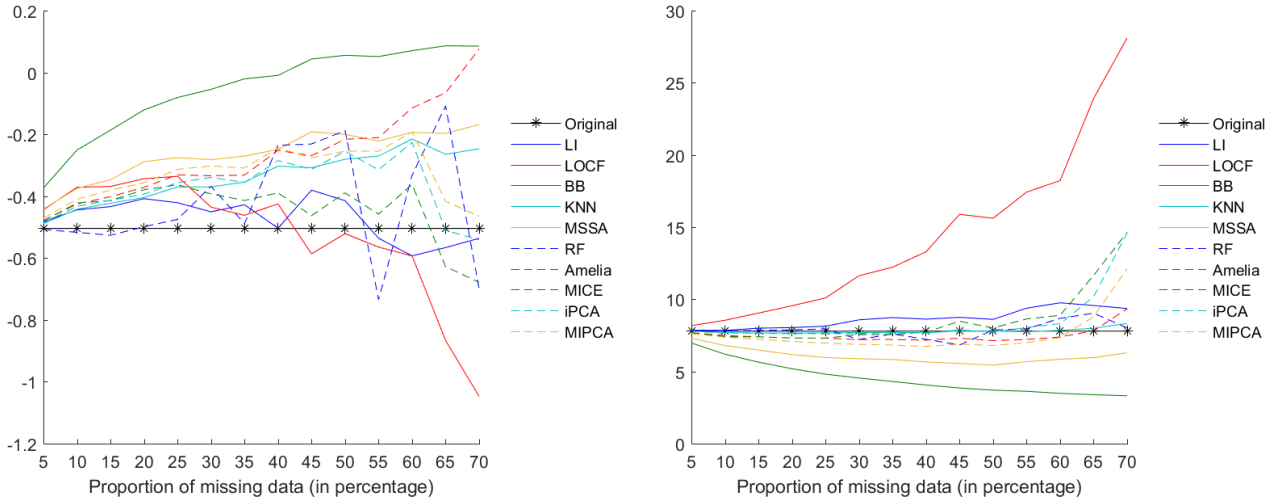
Moreover, the difference with the previous sections also concerns the MICE algorithm, which appears to be much more stable and closer to the original series. Nevertheless, considering extreme values, this method still obtains certain missingness scenarios that do not make it possible to correctly reproduce the moments of the original series.

All the scenarios are now averaged to compare the average performance of each algorithm with respect to the missing proportion. The average results for the first four statistical moments are presented in Figure 3.5-10.

Fig. 3.5-10: Average of the first four statistical moments of the returns of the historical imputed data matrix based on a heuristic approach containing MCAR data (only in the first column) according to the missingness probability



(a) Mean *(b)* Standard Deviation



(c) Skewness *(d)* Kurtosis

It is clear that the linear interpolation, LOCF and Brownian bridge methods tend to further distort the distribution of the series as it contains missing data.

Apart from these methods, it is hard to distinguish between the algorithms. Besides the results for skewness, which are very variable from one proportion of missing data to another (see Figure 3.5-10c), the algorithms succeed in reproducing relatively the statistical moments of the original series faithfully but less and less precisely as the proportion of missing data becomes too big.

In the previous sections, the Amelia algorithm was often in trouble when faced with too high a proportion of missing data; this is also the case here. This loss of efficiency for the particularly high proportion of missing data is also observed here for the MSSA, MICE, IPCA, and MIPCA methods, which deviate more or less significantly from their trend.

The K-NN algorithm appears to be one of the most stable and efficient methods for imputing MCAR data in a way that correctly reproduces the moments of the original data. It does not seem to be impacted by an increase in the proportion of missing data. The historical sample series are correlated more than the simulated sample series, which logically helps the K-NN algorithm.

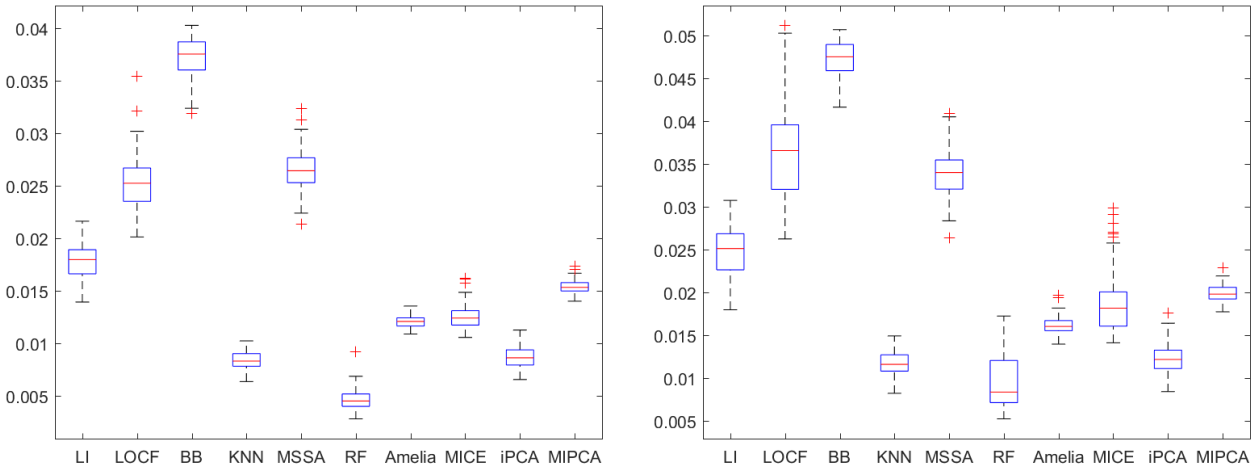
The random forests method, which was one of the best performers in the previous sections, also obtains satisfactory and stable results (excluding skewness, where there appear to be sampling effects), albeit less than the K-NN.

Proximity metrics

The results should now be analyzed in terms of proximity measures. Figure 3.5-11 represents the MAE and RMSE obtained for the 100 scenarios containing 30% missing

data for each of the completion methods. These proximity measures are calculated based on the returns of the original series and those of the imputed series.

Fig. 3.5-11: Distribution of the MAE and RMSE computed from the 100 scenarios containing 30% of MCAR data (only in the first column) on the historical sample based on a heuristic approach



(a) MAE *(b)* RMSE

The results obtained from historical data are represented on a much larger scale (five times larger) than those obtained from simulated data (see Figure 3.2-6). The methods struggled much more when imputing this historical sample. The simulated sample is highly correlated and with constant volatility, which was highly favorable to the imputation. However, this is not the case with the historical sample, which has a strong impact on the quality of the imputation.

Random forests appear to be the method that minimizes the MAE for almost all scenarios; however, this is not the case for its RMSE. This method imputes some data far from the original data, which leads to an increase in the RMSE. Thus, the K-NN and IPCA algorithms obtain for some missingness scenarios, lower RMSE than random forests, and the second-best one in terms of MAE. These methods obtain a proximity measure whose magnitude is relatively similar for both measures, which suggests that the deviations between the original series and the imputed series are relatively similar for each imputation.

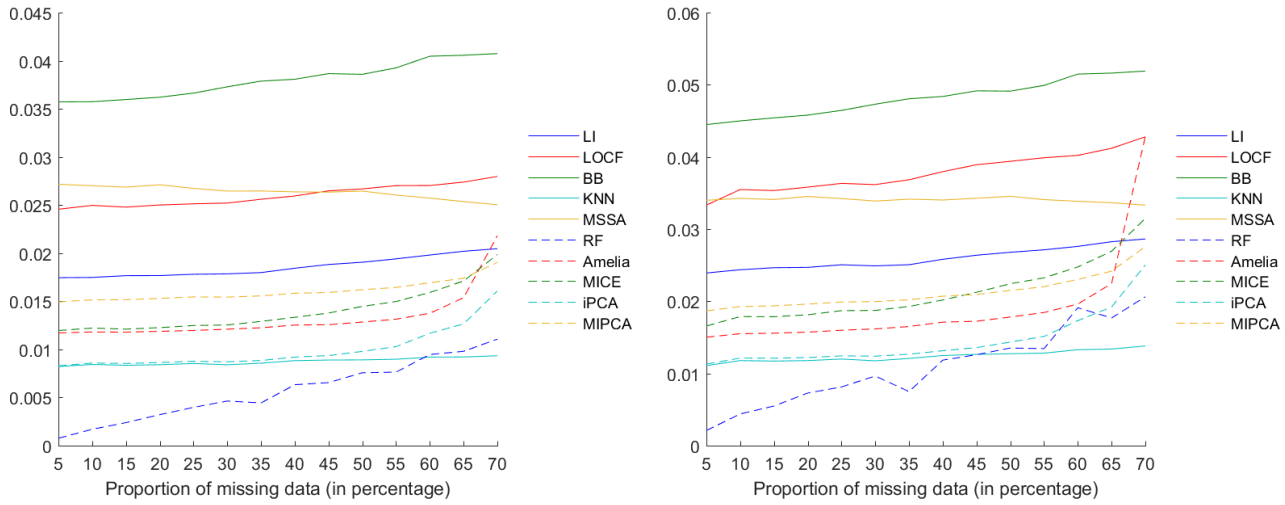
The Amelia and MICE algorithms obtain similar results in terms of MAE and perform better than the MIPCA. Amelia and MIPCA seem to perform worse here than

K-NN and IPCA, whereas the opposite behavior was observed in the previous sections. Nevertheless, the MICE algorithm tends to obtains much higher RMSE than Amelia.

As in the previous sections, linear interpolation, LOCF, Brownian bridge, and MSSA appear to be the worst-performing methods in terms of proximity measures.

Figure 3.5-12 represents the average MAE and RMSE obtained for each missingness proportion and each completion method.

Fig. 3.5-12: Average MAE and RMSE between the return of the imputed data from the historical sample containing MCAR data (only in the first column) and the original historical sample based on a heuristic approach, according to the missingness probability



(a) MAE *(b)* RMSE

The first point concerns the scale of the measures. These proximity measures are much larger on this historical sample than on the simulated data (see Section 3.2.1). In the case of the simulated data, the MAE and RMSE were at most 0.8% and 1.2%, respectively (for the highest proportion of missing data), whereas here they reach 4% and 5%, respectively. Given that these results are on daily data, this means that, on average, the imputations can be within 4% of the original value, which is significant. Finally, only the random forests method makes it possible to impute missing data with average deviations lower than 1% (even lower than 0.5% for less than 30% missing data).

When the proportion of missing data is below 45%, random forests minimize the average proximity measures. However, beyond this proportion, the K-NN method is

preferred. K-NN also obtains very stable proximity measures from one proportion of missing data to another. Furthermore, for a proportion of missing data lower than 45%, the IPCA leads to proximity measures very close to the K-NN, and it is more efficient than its improved version, MIPCA. It is also possible to observe that IPCA uses, on average, two principal components for imputation, while MIPCA uses only one (see Appendix I.3).

The Amelia and MICE algorithms obtain comparable proximity measures that increase with the missingness proportion, which has been the case before in this section. MIPCA appears even less efficient than these two methods when the proportion of missing data is below 45%.

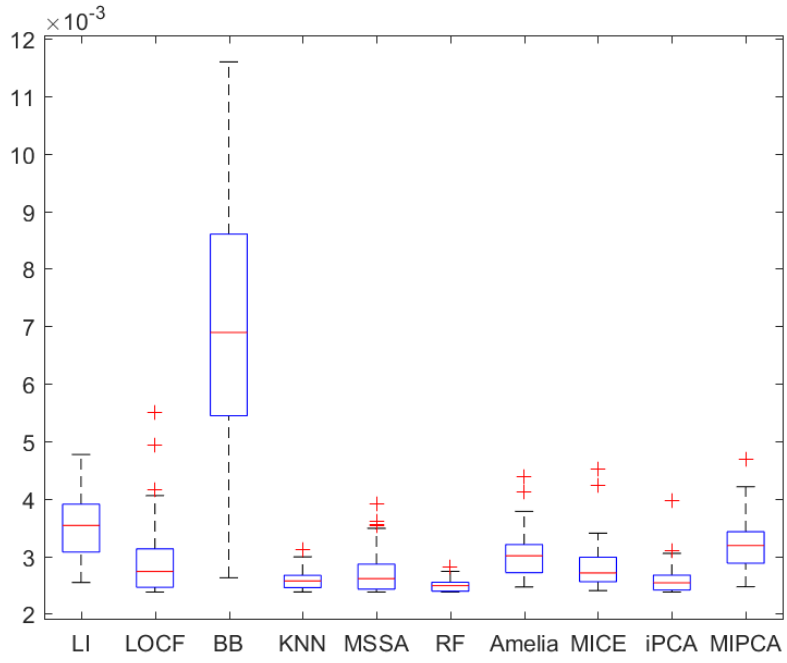
Globally, the best-performing methods remain the same as in the previous sections: random forests, Amelia, MIPCA, IPCA, and K-NN. The main difference is the appearance of MICE in this group of good performers.

Finally, the LOCF, Brownian bridge, and MSSA methods obtain much higher proximity measures than a simple linear interpolation, which makes them unattractive.

Covariance matrices comparison

Figure 3.5-13 represents the differences between the covariance matrix of the imputed series and that of the original series, based on a Frobenius norm, among the 100 scenarios containing a proportion of missing data of 30%. Since the missing data is injected only into the first column of the data matrix, the covariance matrix has been modified only in its first column (and row).

Fig. 3.5-13: Covariance matrix differences, according to the Frobenius norm, based on original historical returns and the imputed returns from historical data based on a heuristic approach containing 30% MCAR data (only in the first column)

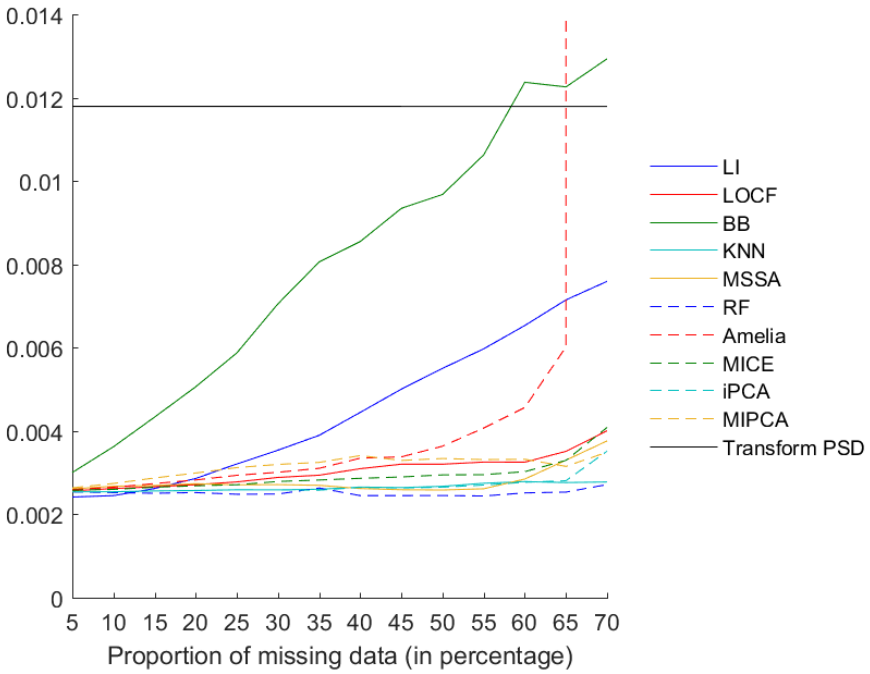


The smallest covariance differences are obtained for the random forests algorithm, closely followed by the K-NN and IPCA, which obtain slightly higher (but still low) deviations. As before, the Amelia, MICE, and MIPCA algorithms are slightly less efficient here.

Conversely, the Brownian bridge leads to covariance gaps that are too large and can be six times larger than those of the other methods. This method, which did not obtain good results until now, must be avoided if the objective is to reproduce the covariance matrix.

Figure 3.5-14 shows the averaged results for each proportion of missing data. The scale of the graph below has been modified to make the results visible. The figure with the original scale is available in Appendix I.5.

Fig. 3.5-14: Averaged covariance matrix differences, according to the Frobenius norm, based on original historical returns and the imputed returns from historical data based on a heuristic approach containing MCAR data (only in the first column) according to the missingness probability



The Brownian bridge method obtains much higher average deviations than those obtained by the other methods; nevertheless, it remains preferable to the process of making the pairwise matrix positive semidefinite, at least if less than half of the sample is missing.

As in the other sections, interpolation is the most effective for a low proportion of missing data (below 15%) but greatly distorts the covariance matrix beyond that.

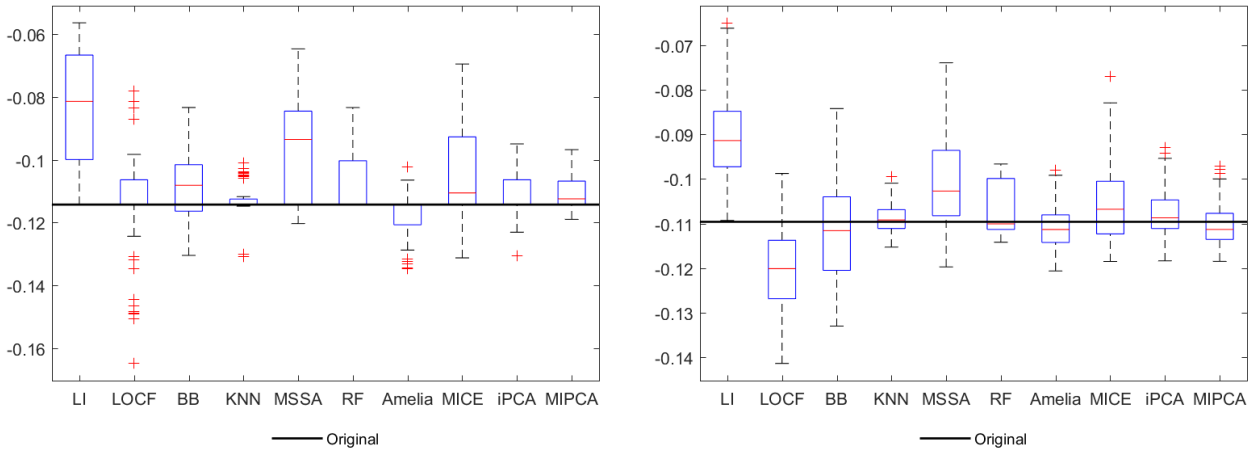
The Amelia algorithm obtains results comparable to the other methods, but beyond 50% missing data, this method tends to significantly distort the covariance matrix. This phenomenon was already observed in the case of simulated data (see Section 3.2.1) but for a higher proportion. Here, with historical data, the covariance matrix deforms more quickly and with a lower proportion of missing data.

The other methods obtain relatively close results, but the algorithm that best preserves the covariance matrix is random forests. In the previous sections, this method was effective at minimizing covariance matrix deviations, and this is also the case for historical data.

Value-at-risk and Expected Shortfall

Finally, the completion methods can also be compared in terms of risk measures. Figure 3.5-15 represents the 1-day VaR and ES for a 99% and 97.5% confidence level, respectively, for the 100 scenarios containing 30% of data imputed by the different algorithms.

Fig. 3.5-15: Distribution of the 1-day risk measures computed from the 100 scenarios of historical sample based on a heuristic approach containing 30% of MCAR data (only in the first column)



(a) 1-day VaR at 99%

(b) 1-day ES at 97.5%

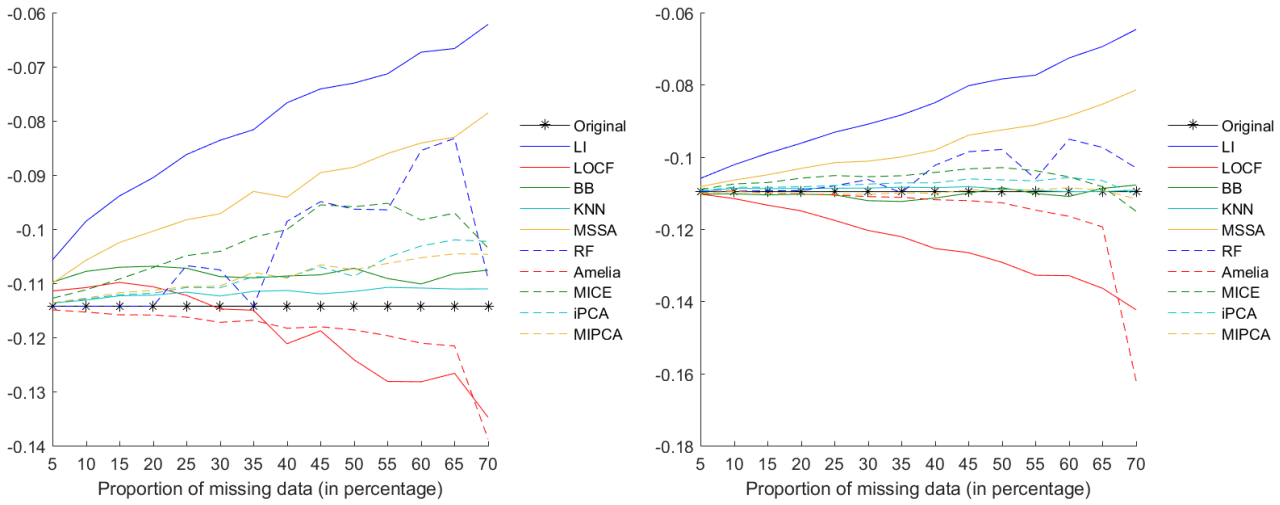
The K-NN method obtains the most stable risk measures from one scenario to another, which is not the case for the other methods. It is the only one with such a small distribution. Linear interpolation, LOCF, Brownian bridge, MSSA, and even MICE obtain the most spread-out distributions, which means the associated risk measures vary more strongly than the others methods from one missingness scenario to another.

Even if the distributions are different from one method to another, all of them succeed in being relatively close to the true risk measures for at least one scenario. These spread-out distributions indicate that the result depends on the scenario. The methods that tend to estimate the risk measures most accurately are the K-NN, followed by the Amelia (where 75% of the scenarios lead to a conservative VaR), IPCA, and MIPCA algorithms.

Here the random forests do not appear to be particularly efficient, as was observed in the case of simulated data (see Section 3.2.1), but they do tend to overestimate the risk measures.

Figure 3.5-16 represents the average risk measures (among the 100 scenarios) for the different proportions of missing data.

Fig. 3.5-16: Average 1-day risk measures computed from the historical sample based on a heuristic approach containing MCAR data (only in the first column) according to the missingness probability



(a) 1-day VaR at 99% *(b)* 1-day ES at 97.5%

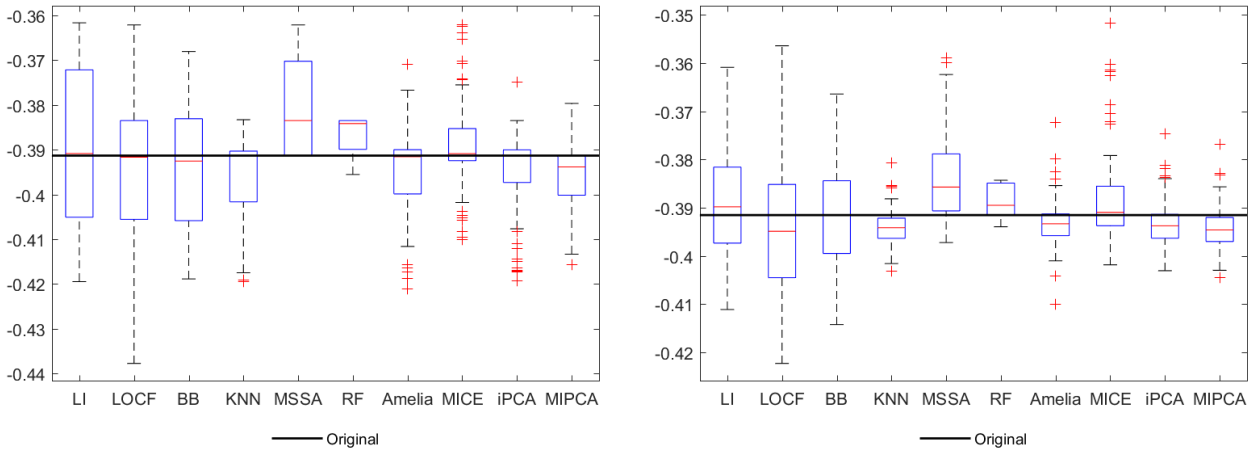
Like the results previously obtained for a proportion of missing data of 30%, K-NN appears to be closest to the true risk measures, as it correctly reproduces the level of VaR and ES at 1 day (even if it tends to slightly overestimate these levels).

The same is true for the Amelia algorithm but for a proportion of missing data lower than 45%, which was also the case with proximity measures. However, this is the only method that leads to conservative risk measures (excluding LOCF) and is, therefore, acceptable to the regulator.

The Brownian bridge is also particularly efficient at reproducing the ES level but overestimates the VaR level even more than the K-NN.

The performance of the algorithms in terms of risk measures at 10 days and for a proportion of 30% of missingness is presented in Figure 3.5-17.

Fig. 3.5-17: Distribution of the 10-day risk measures, computed from the 100 scenarios of historical sample based on a heuristic approach containing 30% of MCAR data (only in the first column)



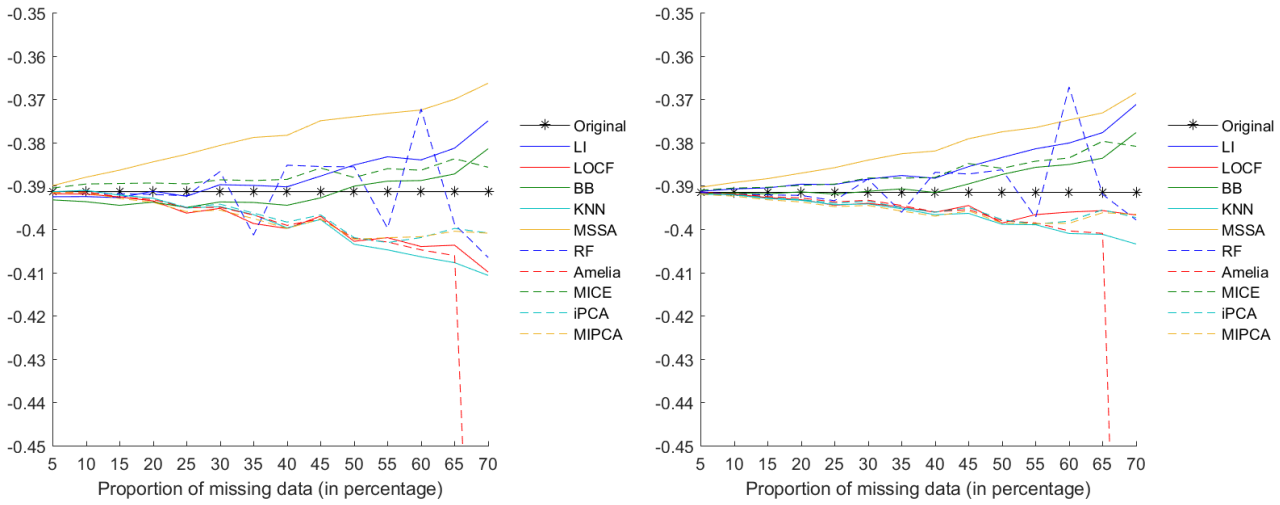
(a) 10-day VaR at 99%

(b) 10-day ES at 97.5%

The general trend for all algorithms reveals a greater instability: It appears that the risk measures at 10 days are much more variable from one scenario to another, but this is due to the 10-day returns.

If the results are now averaged for each of the proportions of missing data studied, the results obtained are those in Figure 3.5-18. The scale of these figures has been modified to correctly analyze the results. The results with the initial scale are observed in Appendix I.7.

Fig. 3.5-18: Average 10-day risk measures, computed from the historical sample based on a heuristic approach containing MCAR data (only in the first column) according to the missingness probability



(a) 10-day VaR at 99%

(b) 10-day ES at 97.5%

Here, no method appears to be able to reproduce either the VaR or the ES in a stable way for any proportion of missing data.

While the K-NN method was the best-performing for a 1-day horizon, here it tends to underestimate the risk measures, just like Amelia, MIPCA, IPCA, and even LOCF.

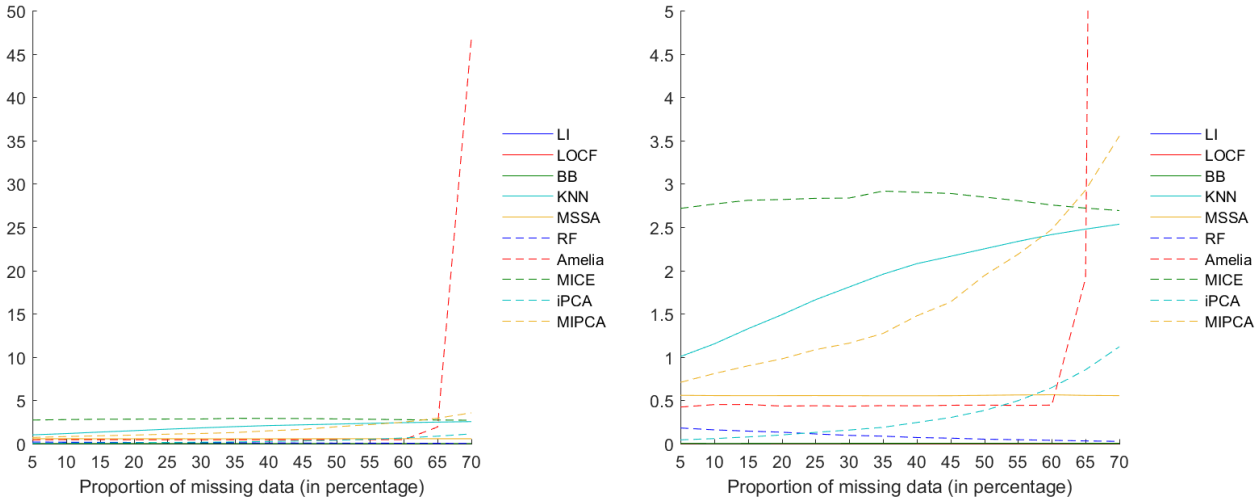
By contrast, the random forests oscillate between underestimating and overestimating the risk measures from one proportion of missing data to another. This is due to the specific sample but not very convincing for the regulator.

Thus, correctly reproducing 10-day returns is much more complex for all methods.

Computation time

Finally, the average computation time (in seconds) required for each algorithm to impute missing data can be observed in Figure 3.5-19.

Fig. 3.5-19: Average computation time of the imputation of MCAR data on only the first series on historical sample based on a heuristic approach (with two different scales) according to the missingness probability



As with simulated data, the Amelia algorithm sees its computation time exploding when the proportion of missing data reaches 70%. It is twice as large as with simulated data (see 3.2-14).

Otherwise, the computation times of the other algorithms are similar to those observed in the case of simulated data and sometimes slightly higher for MICE and MIPCA.

The sample used here was built from a heuristic approach (stocks from the same sector and highly correlated) and makes it possible to see the reaction of the algorithms on real market data. The first observation concerns the orders of magnitude of the comparative tools. The imputation quality seems to be strongly degraded compared with the results obtained on imputed samples. Thus, the orders of magnitude previously obtained reflect the case where the sample is highly correlated with constant volatility. However, this type of sample is difficult to obtain in practice, which has an impact on the quality of the imputation. By contrast, while the application of the completion methods on this historical sample confirms the performance of certain methods, it also calls into question the performance of other methods.

As in the case of simulated data, the Amelia and random forests algorithms remain among the best-performing, although Amelia seems to lose some efficiency with this sample. The MIPCA and K-NN methods also remain among the best performers.

Moreover, the MICE algorithm, which was very often among the worst-performing (if not the worst-performing), obtains results here that are close to those of the Amelia

algorithm. Since the analyses were made from a single sample, sampling errors are possible, and the results of this method improved when the sample was slightly modified (Section 3.2.3 and Section 3.2.4). Thus, when MICE is applied to historical data, these performances improve considerably (to be comparable to Amelia) compared with the results obtained with the simulated sample. Finally, the results of MICE applied to the simulated sample do not seem to be representative because of the sample itself or the lack of heteroskedasticity, and the results of the second historical sample should confirm this. This results emphasizes the need for further analysis of this analysis by using multiple simulated samples, which will be done in future research.

3.5.3 Impact on a sample based on the graphical Lasso

The first sample was built from a heuristic approach, which is not necessarily the most relevant in terms of completion methods. This type of approach aims to make economic sense and relies on the experience of the expert, but it is not necessarily optimal when dealing with multi-step and complex completion methods. As mentioned in Section 2.3, the choice of the sample is one of the preliminary steps that are key to a method's success. Choosing a sample where the columns are not related to each other and where no cross-sectional information would be useful for completion would not make sense.

Presentation of the sample based on the graphical Lasso

A second sample was constructed, this time based on a more sophisticated approach and still using Euro Stoxx 300 data. Here, the procedure to define this sample refers to Kahneman's System 2 [62] because the approach will be analytical and rational but more costly cognitively and in time. According to Kahneman [62], this type of approach would be more reliable than a heuristic approach because it is not impacted by external stimuli.

The selection of stocks (columns) for this second sample uses the graphical Lasso methodology as presented by Friedman, Hastie, and Tibshirani [87] in 2007. This approach was introduced in Section 2.3 and is based on the covariance selection method developed by Dempster [67] in 1972. The idea of this method is to set to 0 the elements of the inverse covariance matrix (also called the precision matrix) that correspond to the weakest dependencies. Thus, the graphical Lasso involves combining this method with a Lasso (l_1 norm) penalty to estimate the strong dependencies of the covariance matrix.

Thus, the *glasso* (from the package of the same name) function implemented on R software by Friedman, Hastie, and Tibshirani [86] was used here to constitute the second historical sample of this PhD thesis. This function was implemented based on the work by Friedman, Hastie, and Tibshirani [87] and Witten, Friedman, and Simon [211].

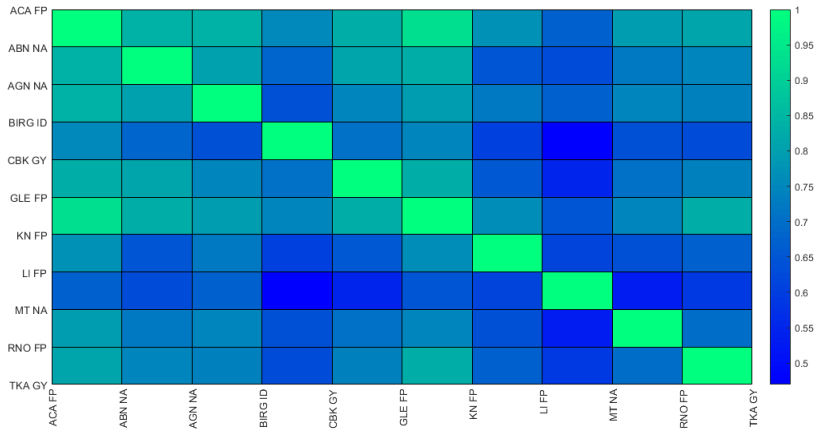
This function has been applied to the pairwise covariance matrix of the 293 stocks (data from 1 January 2020 to 1 February 2021) composing the Euro Stoxx 300. Moreover, this method requires a penalty parameter to be set, which directly impact the dependency level to be considered or not. Therefore, this penalty parameter has been fixed to build a sample containing 11 stocks, including Crédit Agricole Bank, in line with the sample of the previous section. For a penalty parameter equal to 0.00089, the precision matrix finds 10 stocks dependent on the one of Crédit Agricole Bank. This new historical sample now consists of the following stocks (the columns of the data matrix are in this specific order):

- Crédit Agricole Bank(ACA FP)
- ABN Amro Bank(ABN NA)
- Aegon NV (AGN NA)
- Bank of Ireland Group (BIRG ID)
- Commerzbank (CBK GY)
- Société Générale S.A. (GLE FP)
- Natixis S.A. (KN FP)
- Klepierre S.A.(LI FP)
- Arcelor Mittal S.A. (MT NA)
- Renault S.A. (RNO FP)
- Thyssenkrupp (TKA GY)

While the previous sample comprised 100% "financial" stocks (according to level 1 of the BICS classification), this is not the case for this new sample. It comprises eight "financial" stocks (ABN Amro Bank, Aegon NV, Bank of Ireland Group, Commerzbank, Société Générale S.A., Natixis S.A., Klepierre S.A. and Crédit Agricole Bank), with one of them (Klepierre S.A.) in the "real estate" sector according to level 1 of the GICS categorization, two in the "basic materials" sector (Arcelor Mittal S.A. and Thyssenkrupp), and one in the "consumer-cyclical" sector (Renault S.A.).

Thus, this new sample contains only four stocks in common with the one based on a heuristic approach (i.e., ABN Amro Bank, Aegon NV, Société Générale S.A. and Crédit Agricole Bank). Its pairwise correlations are, therefore, necessarily different: They are now between 47% and 93% for this sample (see Figure 3.5-20).

Fig. 3.5-20: Correlation of stock returns of the historical sample based on graphical Lasso

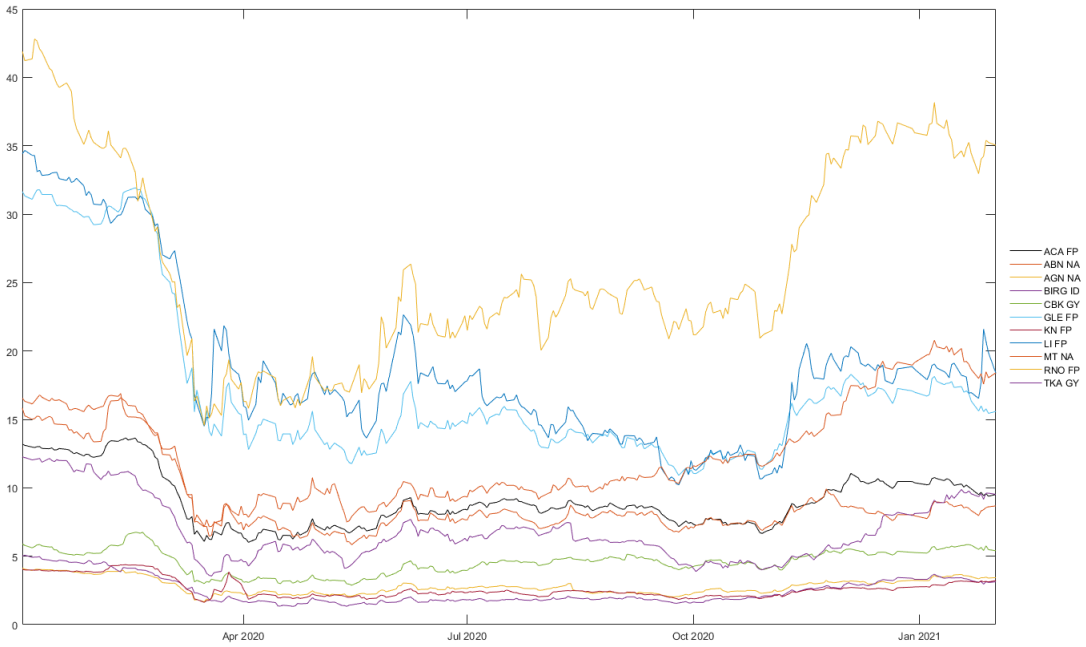


As for the previous historical sample, Little's [142] and Jamshidian and Jalal's [123] tests were applied to these true missing data. As before, Little's test [142] accepts the null hypothesis and, therefore, concludes that the data are MCAR. After deleting the observations where all the variables are missing, Jamshidian and Jalal test [123] does not have enough missing data patterns that can be calculated.

As before, a listwise deletion was applied to this sample to use the complete data later. This led to a complete matrix of 11 columns and 274 rows. Thus, the sample based on a heuristic approach is the same size as the one based on a graphical Lasso approach; however, the dates are not exactly the same: 28 December 2020 is part of the first sample but not of the second, and vice versa for 26 October 2020; hence, there was a date shift over 43 days. This shift (identified ex-post) implies a slight bias when comparing these two historical samples.

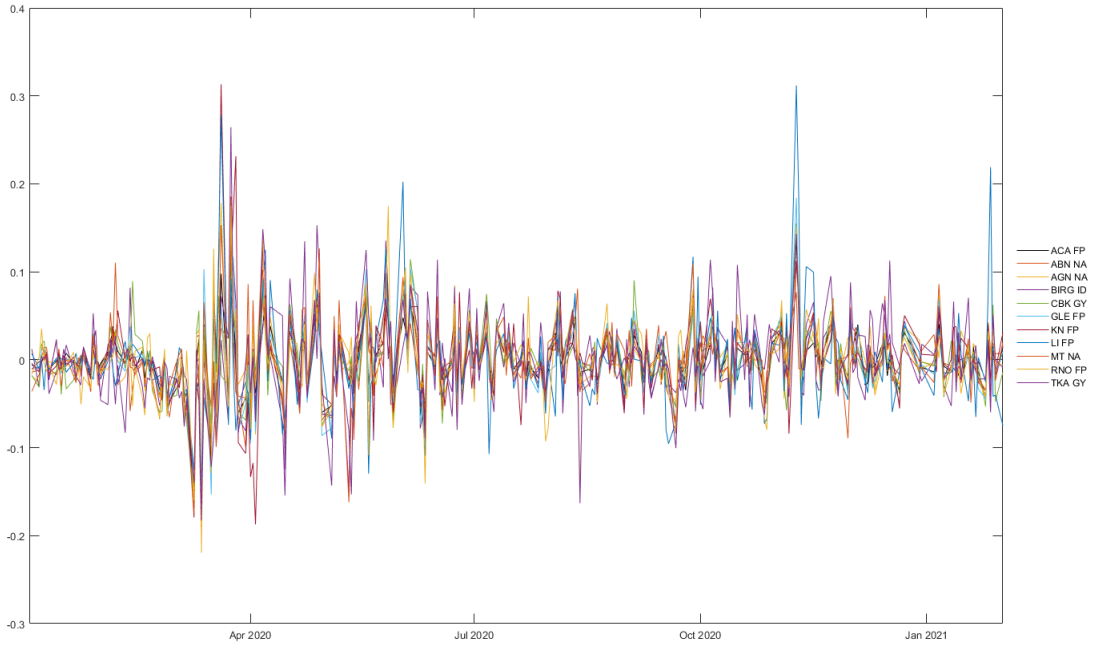
The price evolution between 1 January 2020 and 1 February 2021 of the stocks in this second historical sample is represented in Figure 3.5-22. As before, periods of greater or lesser volatility are clearly visible.

Fig. 3.5-21: Final sample based on graphical Lasso: 11 financial stocks from 1 January 2020 to 1 February 2021



These periods are particularly visible on the returns presented below (Figure 3.5-22).

Fig. 3.5-22: Price returns of the historical sample based on graphical Lasso



Thus, once the sample was built, the data were voluntarily deleted in the first column of this matrix by using an MCAR mechanism and following the same process as in Section 3.2.1. For each proportion of missing data (from 5% to 70%, in increments of 5%), this leads to 100 missingness scenarios different from each other but identical to those used in the first historical sample (Section 3.5.2) and the simulated sample (Section 3.2.1).

Since this historical sample is exactly the same size as the previous one, the missing data will be in the same rows (because the missingness process uses the same seed). Thus, the proportion of missing returns for this historical sample is the same as that observed in the previous section (see Table 3.5-3).

Tab. 3.5-3: Average proportion (number) of missing returns (among the 100 missingness scenarios) associated with the proportion of MCAR raw data injected into the first column of the historical sample based on a graphical Lasso approach of length 274 (273 for return sample)

	Proportion (and number) of missing returns associated with missing data													
Data	**5%** (13)	**10%** (27)	**15%** (41)	**20%** (54)	**25%** (68)	**30%** (82)	**35%** (95)	**40%** (109)	**45%** (123)	**50%** (136)	**55%** (150)	**60%** (164)	**65%** (178)	**70%** (191)
Return	10% (25)	20% (52)	29% (76)	37% (97)	46% (119)	53% (138)	60% (156)	67% (174)	73% (189)	78% (204)	83% (217)	88% (229)	92% (239)	95% (248)

Finally, each comparison tool is computed as defined in Figure 3.1-2. The templates of all the graphs presented in this section have already been presented and detailed (i.e., what they represent and how they were obtained) in Section 3.1.4.

MCAR tests

As before, the analysis begins with the application of MCAR tests. Thus, Little's test [142] and Jamshidian and Jalal's test [123] have been applied to the returns of this second historical sample. Here, both tests are perfectly calculable for all missingness scenarios for each proportion of missing data (see Appendix J.1). The results in Table 3.5-4 represent the proportion of tests that accepted the null hypothesis that the data are MCAR.

Tab. 3.5-4: Confidence level (probability of not rejecting H_0 when H_0 is true) for both MCAR tests applied to price return matrices containing MCAR on the first column of the historical matrix based on the graphical Lasso, for a 5% significance level

	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Little's test	91%	94%	91%	96%	93%	96%	96%	95%	93%	93%	89%	96%	94%	87%
J&J's test	86%	92%	95%	91%	90%	85%	86%	87%	95%	91%	93%	91%	92%	95%

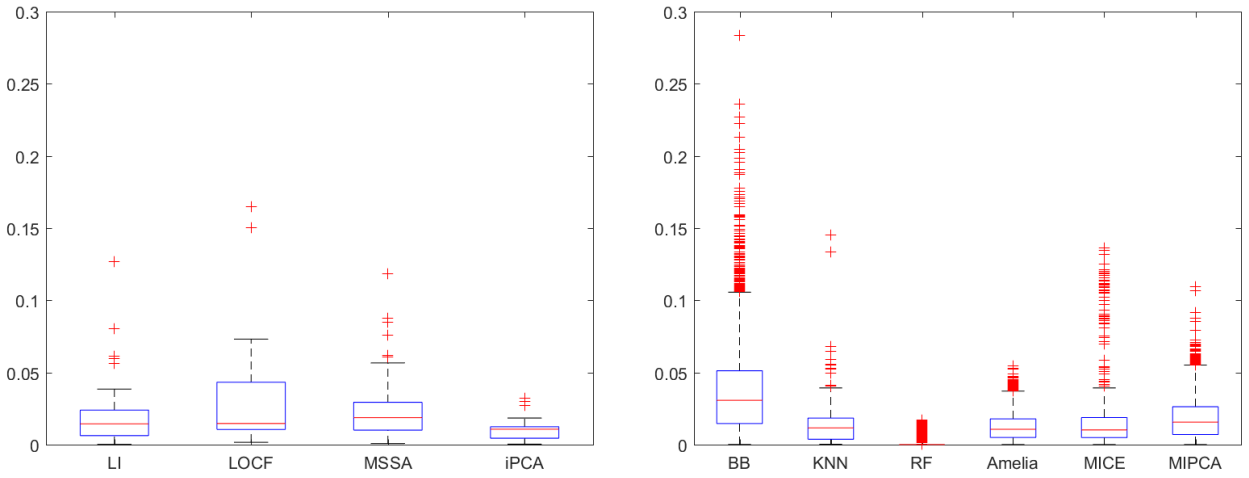
The results above are close to those obtained for the previous historical sample: Little's test [142] tends not to reject the null hypothesis that the data are MCAR, whatever the missingness proportion (confidence levels between 87% and 96%); Jamshidian and Jalal's test [123] also obtains high confidence levels, but they are slightly lower (between 85% and 95%) than those of Little's test [142].

Thus, the conclusions are the same as before: Little's test [142] and Jamshidian and Jalal's test [123] make it possible to detect MCAR data. Jamshidian and Jalal's test [123] is efficient at detecting MCAR data but more sensitive to heteroskedasticity. Further studies should still be done to see whether it is also effective at rejecting the null hypothesis when they are not MCAR.

Preliminary results

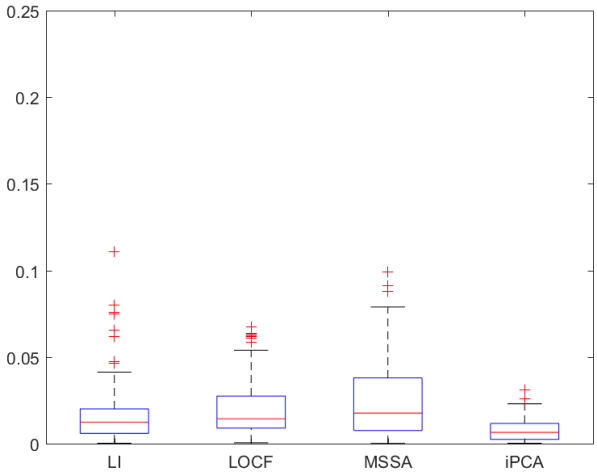
Figure 3.5-23 represents the absolute differences between the original series and the imputed series (of the first column of the matrix) for the first scenario containing 10% missing data (at the top) and 30% (at the bottom).

Fig. 3.5-23: Distribution of absolute return differences between the imputed historical series and original historical series (based on the graphical Lasso) for a single scenario containing 10% (at the top) and 30% (at the bottom) MCAR data (only in the first column)

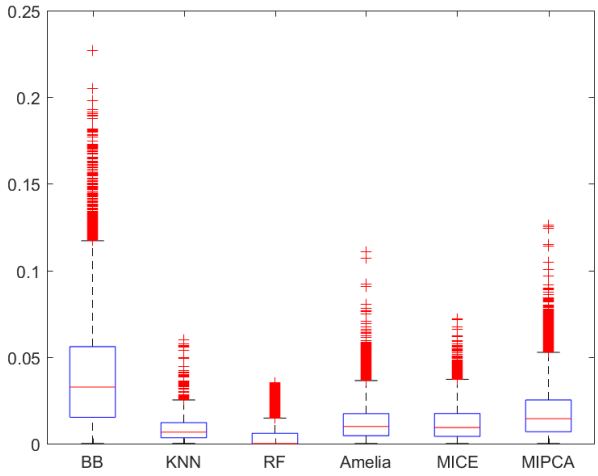


(a) Simple imputation methods with 10% missingness

(b) Methods with a random component and with 10% missingness



(c) Simple imputation methods with 30% missingness



(d) Methods with a random component and with 30% missingness

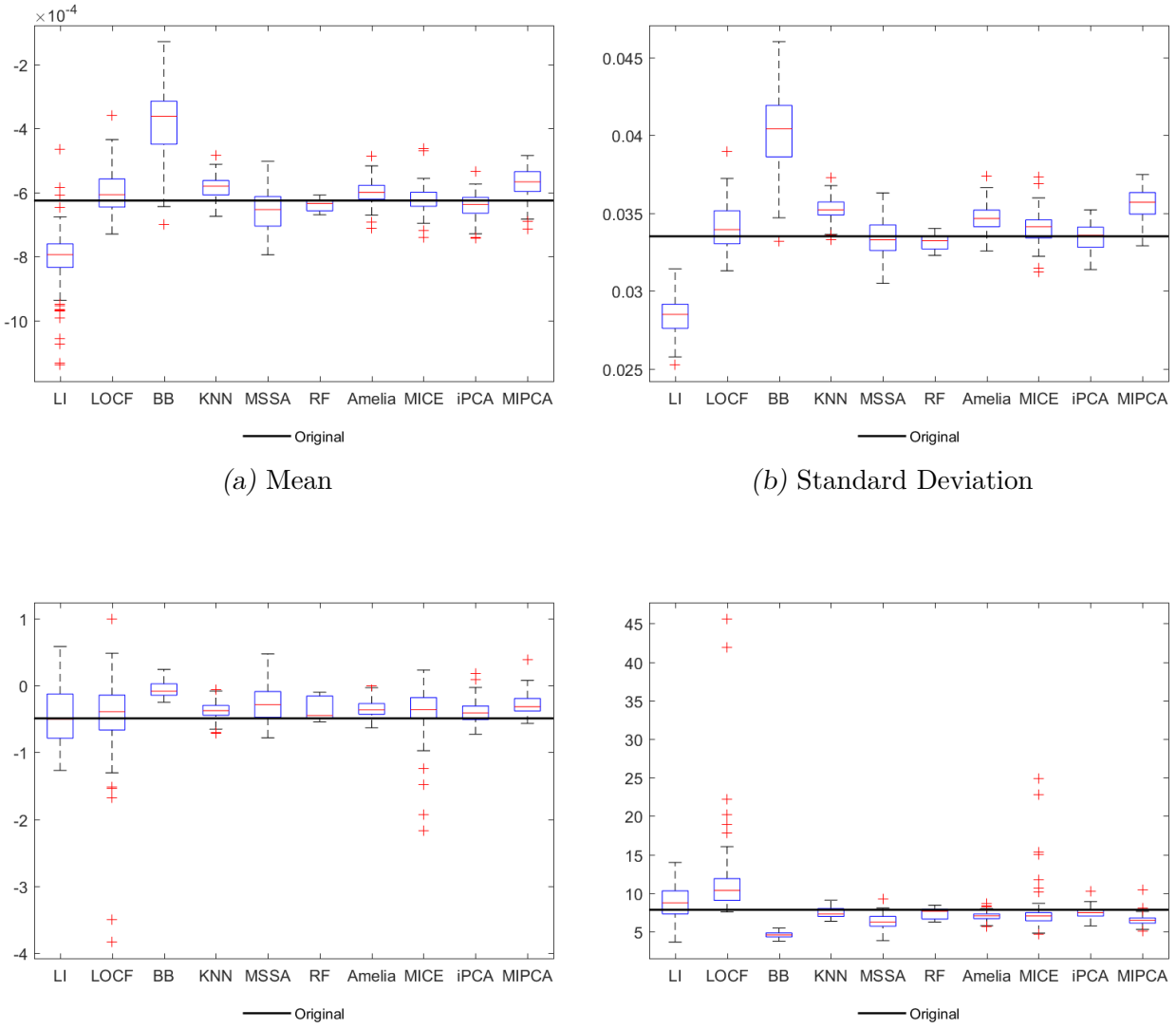
These results are very similar to those observed in the previous sample. The random forests algorithm gives the closest absolute deviations to zero (at least, for this missing data scenario), followed by IPCA, K-NN, Amelia, MICE, and MIPCA. The MICE algorithm obtains results comparable to those of Amelia, which calls into question the results of this method that were previously obtained with the simulated sample. The Brownian bridge and the MSSA do even worse than the usual methods.

Given these results, it seems the sample based on a heuristic approach and the one based on a graphical Lasso give the same results.

Statistical moments

Figure 3.5-24 represents the first four statistical moments obtained for the 100 scenarios containing 30% missing data.

Fig. 3.5-24: Distribution of the first four statistical moments obtained for the 100 scenarios based on historical sample based on the graphical Lasso, with 30% of MCAR data (only in the first column)



(a) Mean

(b) Standard Deviation

(c) Skewness

(d) Kurtosis

The results below are very similar to those obtained with the first historical sample (from Section 3.5.2), but one method obtains slightly different results: the K-NN method.

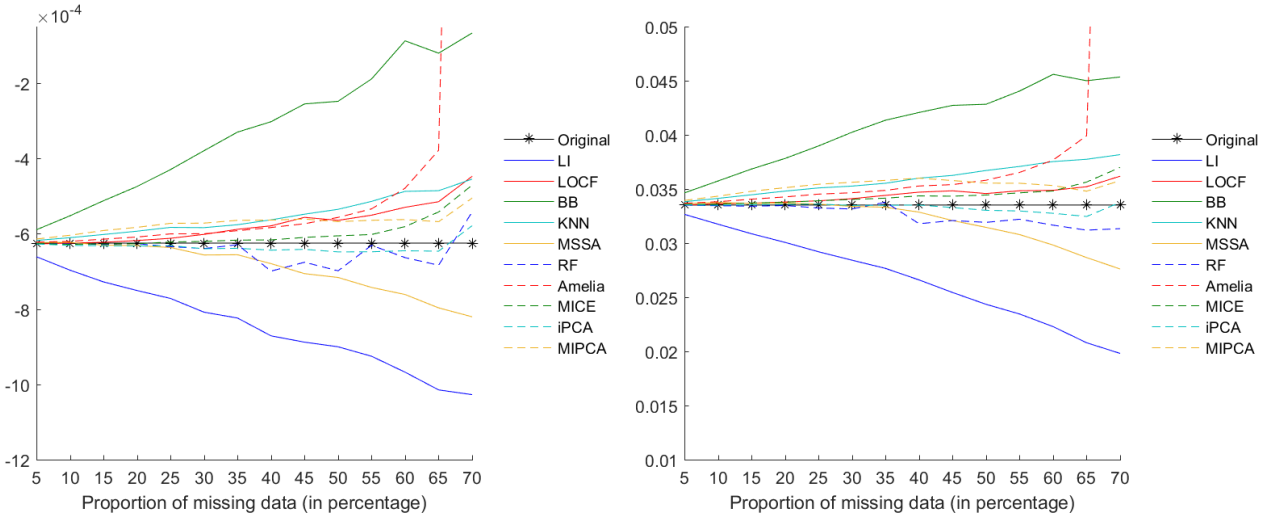
While the K-NN method obtains a mean here that tends to slightly deviate from that of the original series, it mainly tends to increase the volatility of the original series. In the previous sample, the median of the standard deviations obtained for imputations made by K-NN was almost similar to that of the original series, whereas, for this sample,

the K-NN method systematically leads to imputations that increase the volatility of the series.

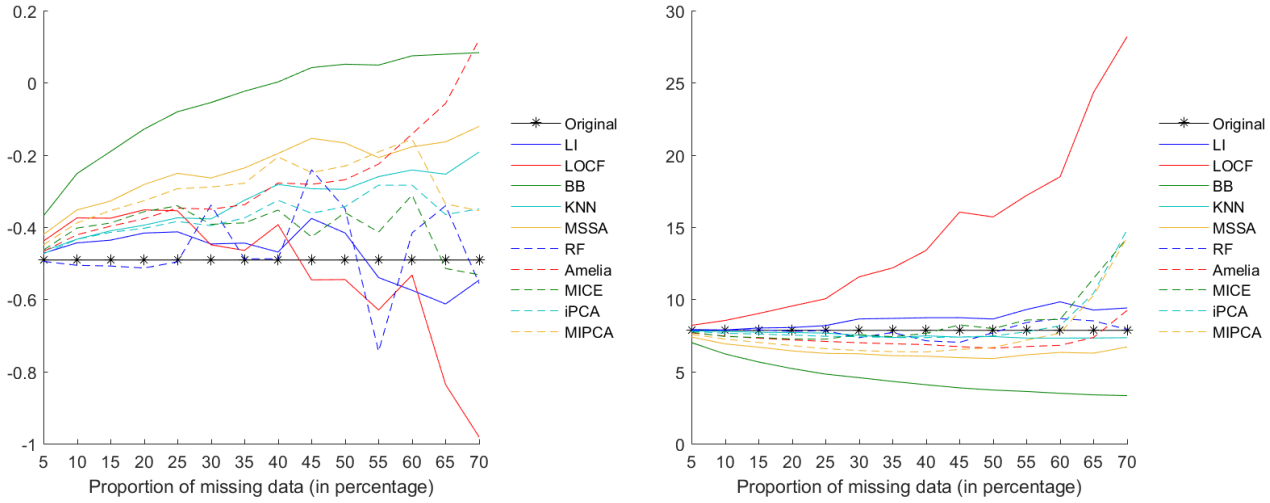
Since the cross-validation step is not very constraining, K-NN almost systematically uses all (or almost all) of the available columns to impute the missing data of the first column. Thus, it is clear that using the heuristic approach (used in the previous section) is preferable to the graphical Lasso when imputing missing data from the Crédit Agricole Bank series with K-NN. The graphical Lasso approach introduces stocks into the sample with specific returns that are different than the Crédit Agricole Bank stock, which results in higher volatility.

The previous results focused on a specific proportion of missing data, hence the interest in observing the results as averages for several proportions of missing data. Figure 3.5-25 represents the average evolution of the first four statistical moments for an increasing proportion of missing data. The standard deviation for each of the statistical moments (among the 100 missingness scenarios) is available in Appendix J.3 for all proportions of missing data.

Fig. 3.5-25: Average of the first four statistical moments of the returns of the historical imputed data matrix based on the graphical Lasso containing MCAR data (only in the first column) according to the missingness probability



(a) Mean *(b)* Standard Deviation



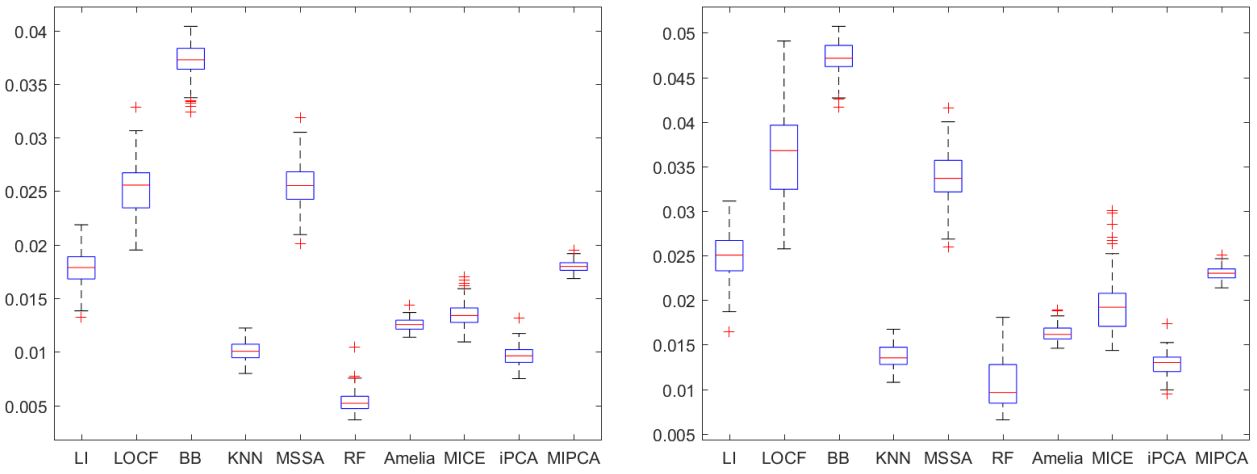
(c) Skewness *(d)* Kurtosis

As before, the results above are relatively close to those obtained in the previous section (Figure 3.5-10 of Section 3.5.2). The only notable differences concern the K-NN method, which tends to increase the standard deviation of the series more and more with the appearance of missing data. Apart from that, the behaviors of the other methods coincide perfectly with the previous sample (based on a heuristic approach). The Amelia algorithm always tends to explode when the proportion of missing data is too high; as before, this will be the case for all other criteria to come.

Proximity metrics

Concerning the proximity measures, a first analysis is made below from the 100 scenarios containing 30% missing data and presented in Figure 3.5-26.

Fig. 3.5-26: Distribution of the MAE and RMSE computed from the 100 scenarios containing 30% of MCAR data (only in the first column) on the historical sample based on the graphical Lasso



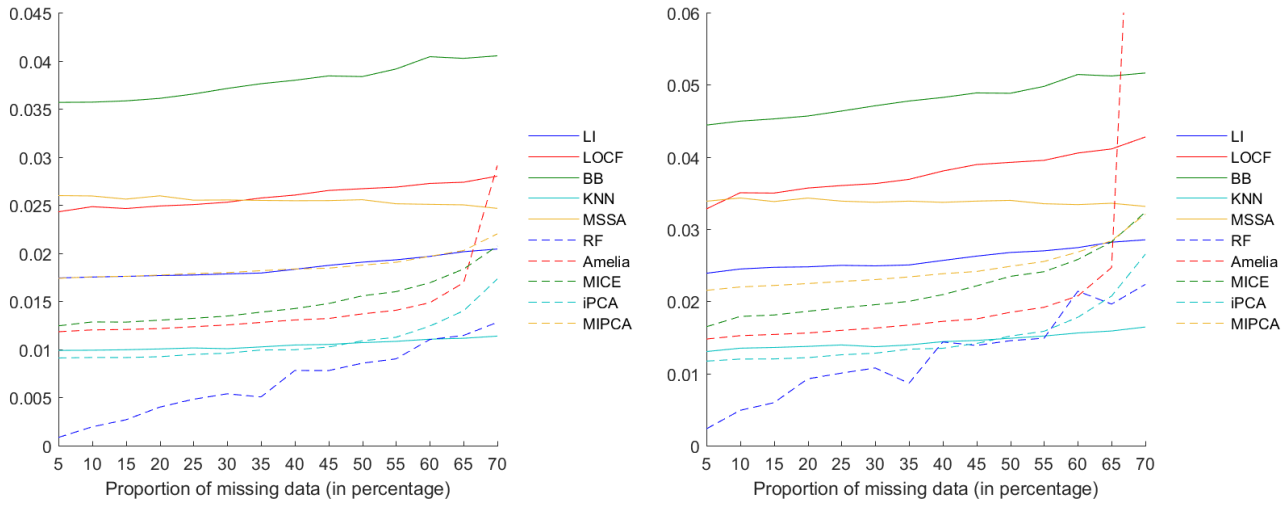
(a) MAE *(b)* RMSE

As for the sample based on a heuristic approach (see Section 3.5.2), the results obtained here are much higher than those obtained with a simulated sample (see Section 3.2.1). This shows that the results are conditional on the sample and that the orders of magnitude observed on the simulated sample are unrealistic but make it possible to illustrate a theoretical framework.

The change in methodology regarding sample formation has little effect on these proximity measures. Only a few differences are observable for K-NN, with MAE and RMSE slightly higher here than in the previous section.

As for the other proportions of missing data, the average MAE and RMSE (among the 100 missingness scenarios) is shown in Figure 3.5-27 below. The standard deviation of each proximity measure (among the 100 missingness scenarios) is available in Appendix J.6 for all proportions of missing data.

Fig. 3.5-27: Average MAE and RMSE between the return of the imputed data from the historical sample containing MCAR data (only in the first column) and the original historical sample based on the graphical Lasso, according to the missingness probability



(a) Averaged MAE *(b)* Averaged RMSE

As in the previous section, the proximity measures from historical data are much higher than those from simulated data (from Section 3.2.1). Once again, only the random forests can fill in the missing data with a deviation much lower than 1% for any missing data proportion. Thus, imputation methods are less efficient on historical data, even with highly correlated stock return series.

The results obtained here are similar to those observed for the sample constituted through a heuristic approach, except for K-NN and MIPCA.

Here, the MAE and RMSE from K-NN are higher than those obtained by IPCA for proportions of missing data below 40%, whereas previously, these two methods gave the same proximity measures. Nevertheless, the differences are almost insignificant here.

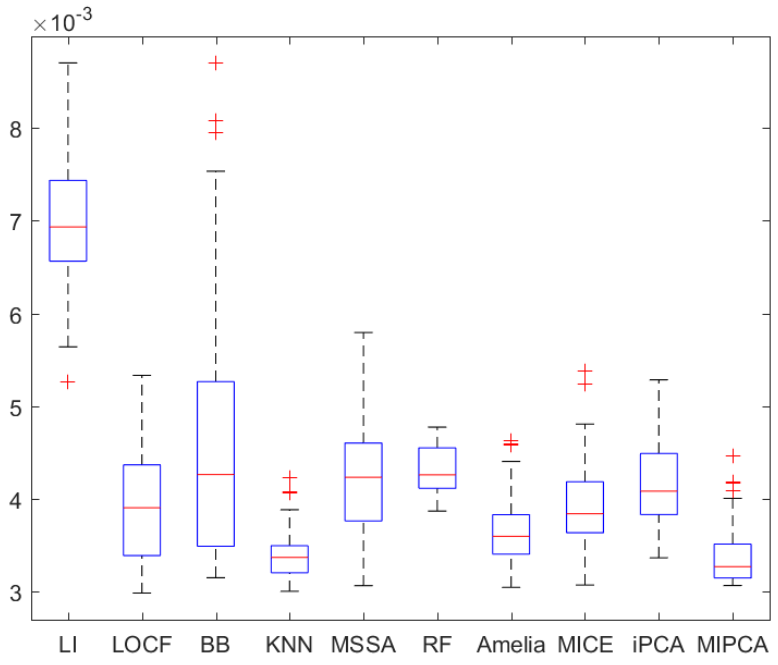
The MIPCA method obtains higher MAE and RMSE. While in the previous section the proximity measures obtained for this method were close to those of MICE, here they appear to be closer to the linear interpolation; hence, there is a deterioration of the quality of the imputation with this graphical Lasso approach.

This is also true for the MSSA and IPCA but to a lesser extent. The number of principal components used by the IPCA and MIPCA algorithms has increased compared with those used with the sample based on a heuristic approach. Here, IPCA uses on average three principal components against two previously, and MIPCA uses two against only one previously (see Appendix J.4).

Covariance matrices comparison

Figure 3.5-28 represents the differences in covariance matrices, according to a Frobenius norm, for the 100 scenarios containing 30% missing data.

Fig. 3.5-28: Covariance matrix differences, according to the Frobenius norm, based on original historical returns and the imputed returns from historical data (based on the graphical Lasso) containing 30% of MCAR data (only in the first column)



The covariance matrices between the heuristic sample and the graphical Lasso sample are necessarily different, which is why the results above are hardly comparable to those obtained in the previous section (from Figure 3.5-13 in Section 3.5.2). However, it is still possible to compare the methods that are the most likely to minimize these differences.

In the previous sample (heuristic), the random forests method was the method that imputed with the least distortion of the covariance matrix, followed closely by K-NN, IPCA, and Amelia. Conversely, the Brownian bridge was the method that distorted the covariance matrix the most.

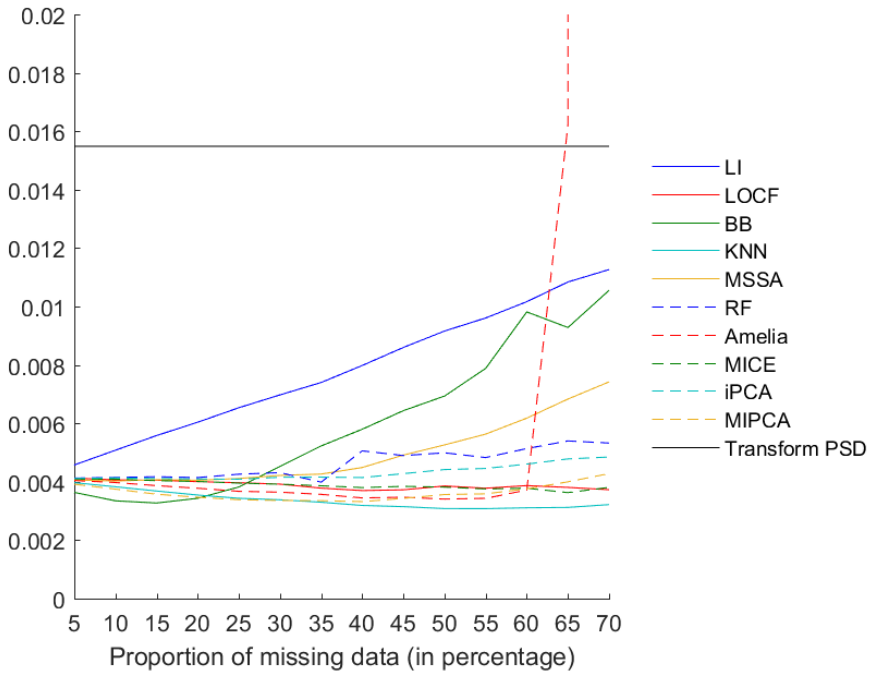
Here, with this sample constituted by the graphical Lasso approach, the ranking is quite different. The random forests method is far from being the method that distorts the least the covariance matrix - quite the contrary. This method obtains a median

comparable to that of the Brownian bridge, which remains among the least efficient here too. The same observation can be made for the MSSA and the IPCA, which are now among the least efficient methods.

The K-NN, MIPCA, and Amelia methods are much more efficient at preserving the original covariance matrix here. Moreover, the Amelia method obtains covariance differences almost similar to those obtained in the previous section, which is not surprising, given that the method aims to estimate the law of the matrix and, thus, the covariance matrix.

To analyze the results for different proportions of missing data, Figure 3.5-29 presents the average evolution of the covariance matrix deviations, based on a Frobenius norm. The standard deviation of these covariance differences (among the 100 missingness scenarios) is available in Appendix J.8 for all proportions of missing data.

Fig. 3.5-29: Average covariance matrix differences, according to the Frobenius norm, based on original historical returns and the imputed returns from historical data based on the graphical Lasso containing MCAR data (only in the first column) according to the missingness probability



The Brownian bridge gives less significant covariance differences on this matrix but is still among the least efficient and comparable to linear interpolation. Moreover, the

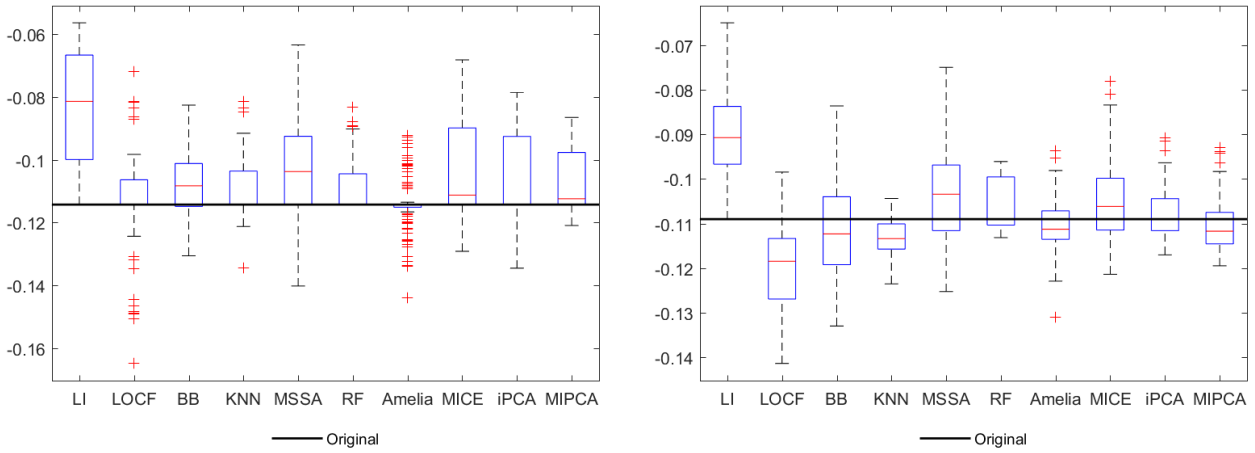
observations made previously are confirmed here: Random forests, MSSA, and IPCA are among the least efficient methods of the group.

By contrast, the results of the other methods remain relatively comparable to each other.

Value-at-risk and expected shortfall

Finally, Figure 3.5-30 represents the 1-day risk measures obtained for each of the 100 scenarios containing 30% missing data.

Fig. 3.5-30: Distribution of the 1-day risk measures computed from the 100 scenarios of historical sample based on the graphical Lasso containing 30% of MCAR data (only in the first column)



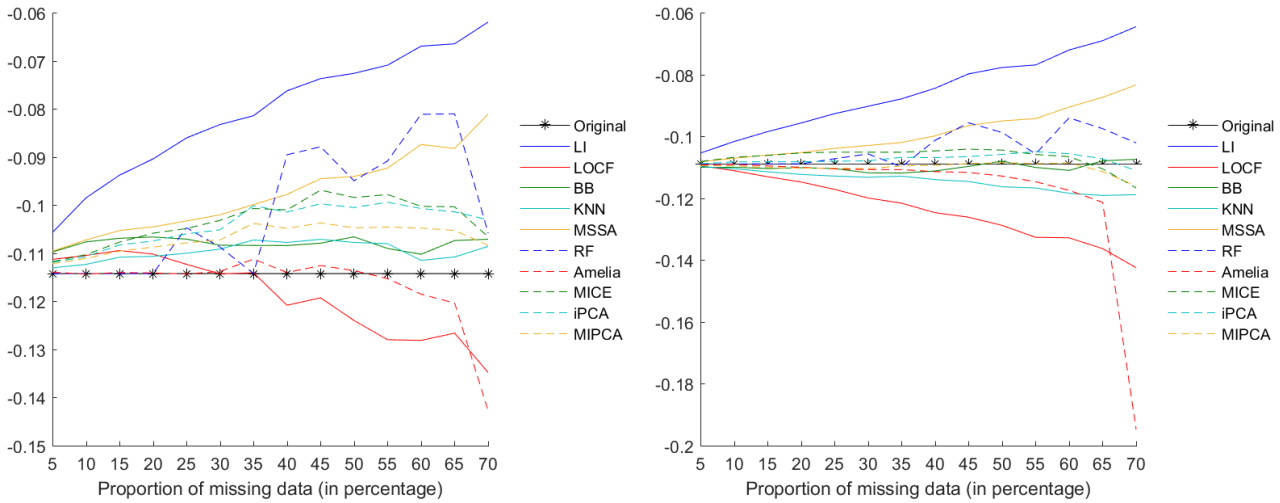
(a) 1-day VaR at 99%

(b) 1-day ES at 97.5%

Here, the Amelia algorithm manages to reproduce the VaR of the original series very faithfully when the sample contains 30% missing data; otherwise, the other methods tend to overestimate the level of VaR, sometimes very strongly. Amelia obtains a conservative median for the ES.

Figure 3.5-31 represents the average of the risk measures obtained among the 100 missingness scenarios for all the missing data proportions tested. The standard deviation of each 1-day risk measure (among the 100 missingness scenarios) is available in the first part of Appendix J.10 for all proportions of missing data.

Fig. 3.5-31: Average 1-day risk measures computed from historical sample based on the graphical Lasso containing MCAR data (only in the first column) according to the missingness probability



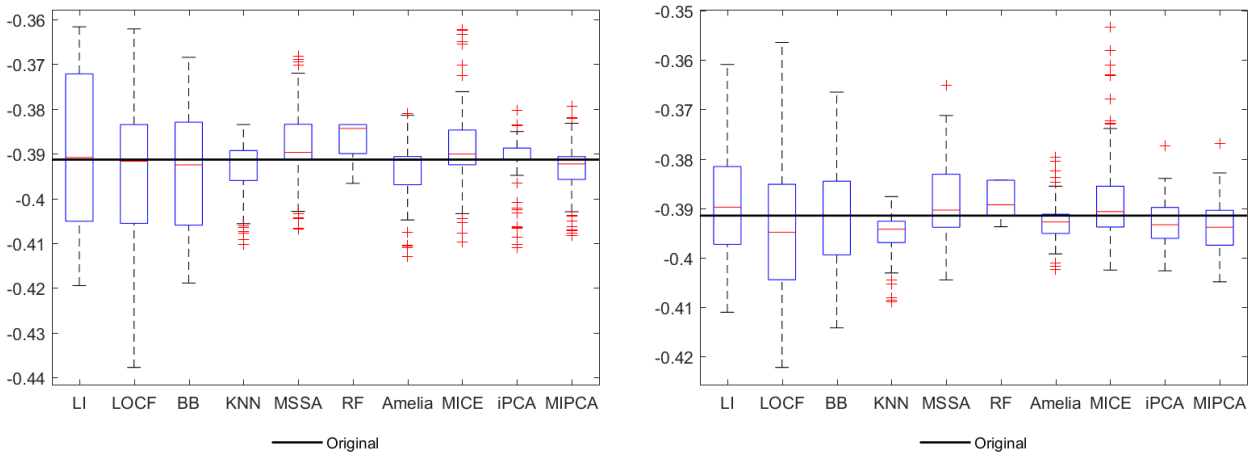
(a) 1-day VaR at 99% *(b)* 1-day ES at 97.5%

As in the previous section, Amelia is the only method that on average obtains conservative risk measures for all proportions of missing data. It appears here as the most relevant method to submit to the regulator, as long as the proportion of missing data is not excessive.

The other completion methods tend to underestimate the risk, with risk measures above the real level.

Figure 3.5-32 represents the 10-day risk measures for the 100 missingness scenarios containing 30% missing data.

Fig. 3.5-32: Distribution of the 10-day risk measures, computed from the 100 scenarios of historical sample based on the graphical Lasso containing 30% of MCAR data (only in the first column)



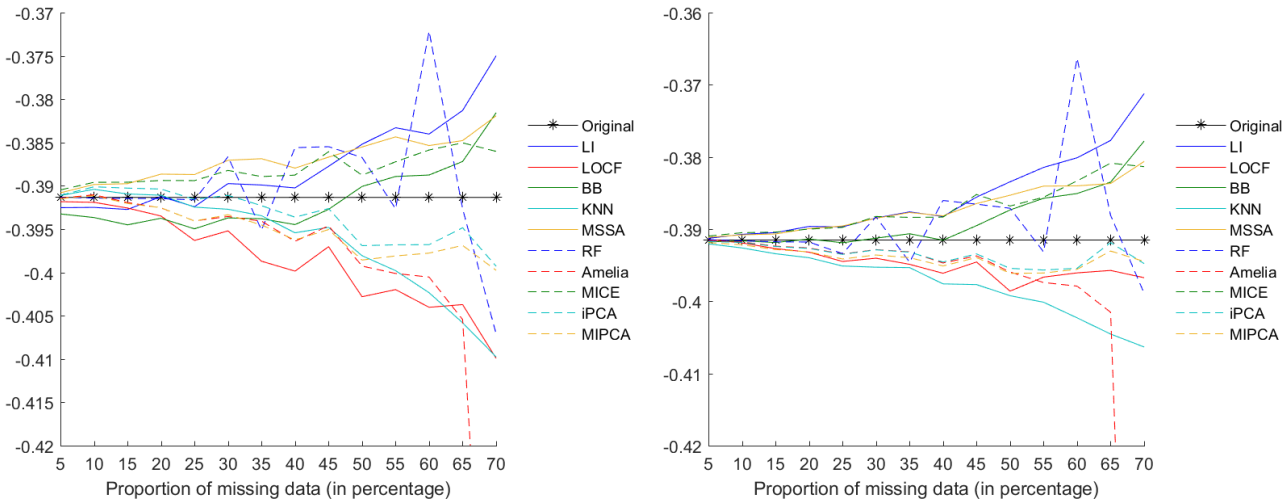
(a) 10-day VaR at 99%

(b) 10-day ES at 97.5%

As is often the case, the 10-day risk measures are more variable than the 1-day risk measures. Moreover, Amelia here also appears to be a good imputation method and conservative in terms of risk measures since its median is below the level of the risk measures of the original series. This phenomenon can also be observed for MIPCA.

Figure 3.5-33 represents the average risk measures (obtained from the 100 missingness scenarios) for each proportion of missing data. The standard deviation of each 10-day risk measure (among the 100 missingness scenarios) is available in the second part of Appendix J.10 for all proportions of missing data.

Fig. 3.5-33: Average 10-day risk measures, computed from the historical sample based on the graphical Lasso containing MCAR data (only in the first column) according to the missingness probability



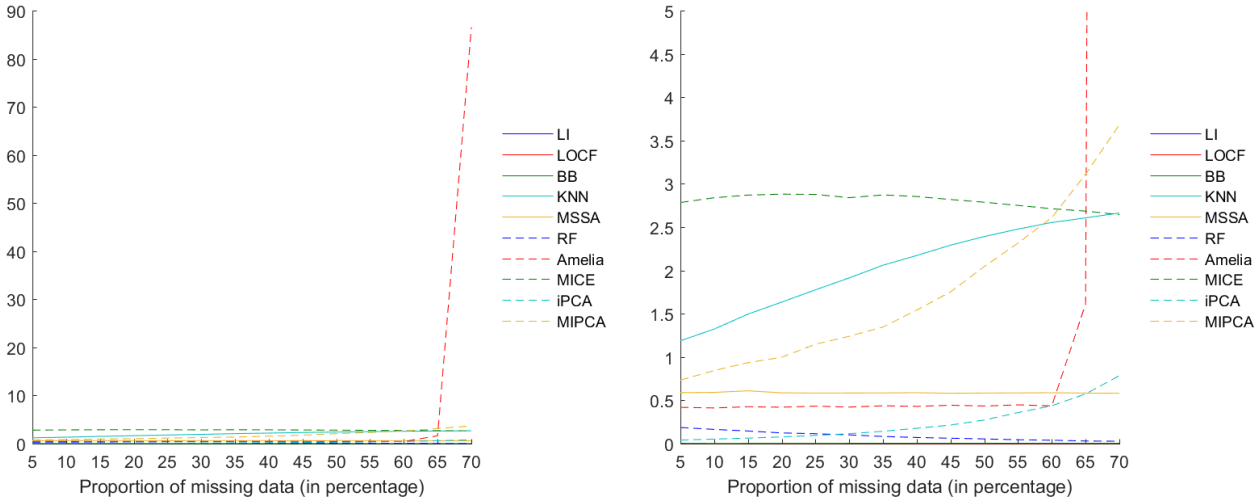
(a) 10-day VaR at 99% *(b)* 10-day ES at 97.5%

As in the previous section, the risk measures associated with the Crédit Agricole Bank stock are conservative for all proportions of missing data if the methodology used is Amelia, MIPCA, IPCA, and K-NN (or even LOCF). The results obtained here are, in fact, comparable to those of the previous section.

Computation time

The computation time (in seconds) required to impute a sample for each method is presented in Figure 3.5-34.

Fig. 3.5-34: Average computation time of the imputation of MCAR data on only the first series on historical sample based on the graphical Lasso (with two different scales) according to the missingness probability



Not surprisingly, the computation time observed here is exactly the same as in the previous section.

Thus, this last empirical study is based on a sample of historical data constituted quite differently than those observed in the previous section. Previously, the idea was to follow a heuristic approach, which an expert could have used through his experience. Here, the idea is to use a very precise statistical model to constitute the columns of the sample. This method is based on Dempster's model [67], which uses the precision matrix to deduce the strong dependencies between variables. The method used here is that of Friedman, Hastie, and Tibshirani [87] to obtain the dependencies from a Lasso (l_1 norm) penalty performed on the precision matrix.

This method was applied to the Euro Stoxx 300 data to constitute a sample comparable (in dimension) to that of the previous section and to analyze the differences in terms of imputation. This new sample is not correlated as much and, therefore, necessarily leads some methods to impute differently. However, as for the sample based on a heuristic sample, the orders of magnitude obtained are much larger than those observed on the simulated sample (see Section 3.2.1), suggesting that, in practice, the orders of magnitude are larger than those observed in the case of the simulated sample.

Logically, the one-dimensional methods (linear interpolation, LOCF, and Brownian bridge) obtain the same results given that, for both samples, the methods only use the Crédit Agricole Bank data. By contrast, the other methods can be impacted by the

choice of the columns that make up the data matrix. This is the case for the K-NN method, which directly uses the values of the neighbors to impute the missing data. Thus, if the values of the other columns are very different, this will have an impact on the imputations. This is why the proximity measures obtained by K-NN are higher for this sample than for the previous one because it is not constructed in such a way that the series are similar. Similarly, the MIPCA, IPCA, and MSSA methods, which use the whole sample, see their proximity measures deteriorate slightly for this second historical sample.

Moreover, although the random forests algorithm obtains proximity measures that are as satisfactory as with the first historical sample, it tends to seriously distort the covariance matrix of the second sample.

The Amelia algorithm is not affected by this change of sample since it looks for the conditional distribution of the missing values to the observed values and not the missing values directly. Thus, all the results obtained for this method are very close for both samples, which leads us to believe in the robustness of this method, although future work will be necessary to confirm this.

3.6 Imputation of data: MAR on historical data

The historical data used in this PhD thesis was presented in the previous section (Section 3.5). However, Figure 3.5-1 and Figure 3.5-2 reveal that the true missing data in this sample follow a pattern of data being successively missing at the end of the series and not an MCAR mechanism. The previous section aimed to study the impact of MCAR data on historical samples, but a mechanism where data are successively missing at the end of the series seems more realistic.

That is why this section uses the two historical samples created in the previous sections to apply the missing data mechanism that involves successively deleting the data at the end of the series. Thus, the mechanism, which was presented in Section 3.3.3), involves successively deleting the data at the end of the first column of each of the two historical samples: the first one based on a heuristic approach (see Section 3.5.2) and the second one based on a graphical Lasso approach (see Section 3.5.3).

In this way, data were removed from the first column of price data with proportions ranging from 5% to 70%, in increments of 5%. This missingness mechanism makes it possible to have a unique missingness scenario for each proportion of missing data, as no random component is part of it.

As the missing data are at the end of the series, the number of missing prices is equal to the number of missing returns (see Table 3.6-1).

Tab. 3.6-1: Proportion (number) of missing returns associated with the proportion of MAR raw data injected into the first column of the historical sample of length 274 (273 for return sample)

	Proportion (and number) of missing returns associated with missing data													
Data	**5%** (14)	**10%** (27)	**15%** (41)	**20%** (55)	**25%** (69)	**30%** (82)	**35%** (96)	**40%** (110)	**45%** (123)	**50%** (137)	**55%** (151)	**60%** (164)	**65%** (178)	**70%** (192)
Return	**5%** (14)	**10%** (27)	**15%** (41)	**20%** (55)	**25%** (69)	**30%** (82)	**35%** (96)	**40%** (110)	**45%** (123)	**50%** (137)	**55%** (151)	**60%** (164)	**65%** (178)	**70%** (192)

Finally, each comparison tool is computed as defined in Figure 3.1-2. The templates of all the graphs presented in this section have already been presented and detailed (i.e., what they represent and how they were obtained) in Section 3.1.4.

3.6.1 Impact on a sample based on a heuristic approach

The interest of this section is to remove data successively at the end of the first column of the sample based on a heuristic approach. The sample used here is the one presented in Section 3.5.2, which consists of highly correlated financial stocks. The missingness mechanism is the one presented in Section 3.3.3.

MCAR tests

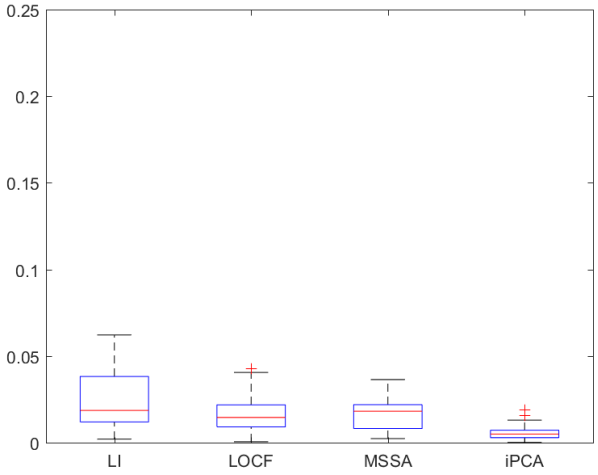
As before, the analysis starts with the application of the MCAR tests. As explained in Section 3.3.3, this missingness mechanism is categorized as MCAR according to Little and Rubin [145]. Little's test [142] and Jamshidian and Jalal's test [123] are based on Little and Rubin's categorization, which means they are efficient if they do not reject the null hypothesis that the data are MCAR.

As in Section 3.3.3, the tests are always calculable as the missing data are only in the first column. However, the results are not the same. Little's test [142] never rejects the null hypothesis for all missingness proportions. By contrast, Jamshidian and Jalal's test [123] does not reject it only for two missingness proportions (25% and 30%). When these tests were applied to simulated data, Jamshidian and Jalal's test [123] accepted the null hypothesis for all missingness proportions while Little's test [142] generally rejected it. These results contradict those from Section 3.3.3, thereby confirming that this missingness mechanism should be applied to other samples before any conclusions are drawn.

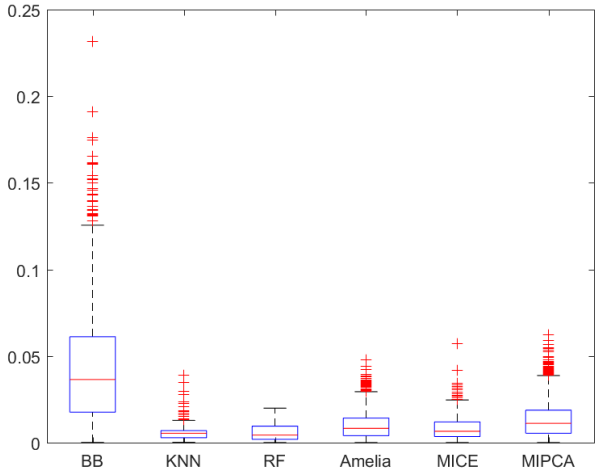
Preliminary results

As previously, these preliminary results aim to present the distribution of absolute return differences between original data and imputed data (see Figure 3.6-1). The results from methods with a random component are based on 100 imputations of reach missing data (no repetition is needed with methods without a random component as the imputations are always the same).

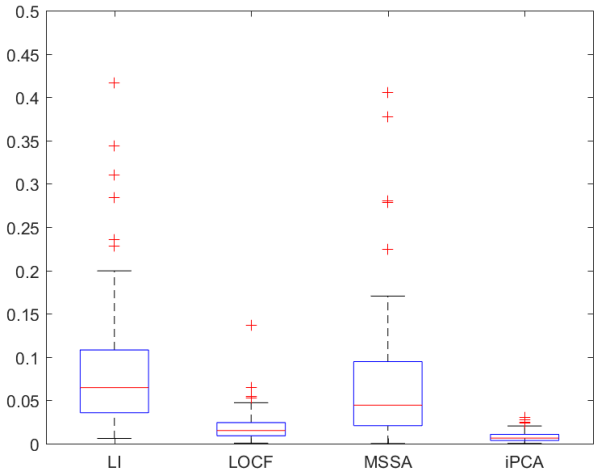
Fig. 3.6-1: Distribution of absolute return differences between the imputed series and original series for a sample containing 10% (at the top) and 30% (at the bottom) MAR data (successive missing data at the end of the first series of the sample based on a heuristic approach)



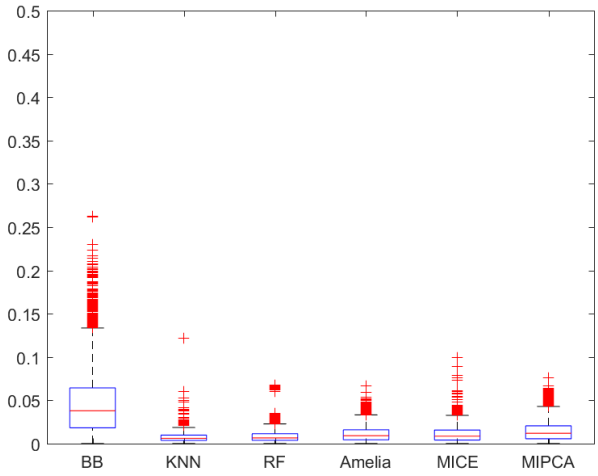
(a) Methods without a random component for 10% missingness



(b) Methods with a random component for 10% missingness



(c) Methods without a random component for 30% missingness



(d) Methods with a random component for 30% missingness

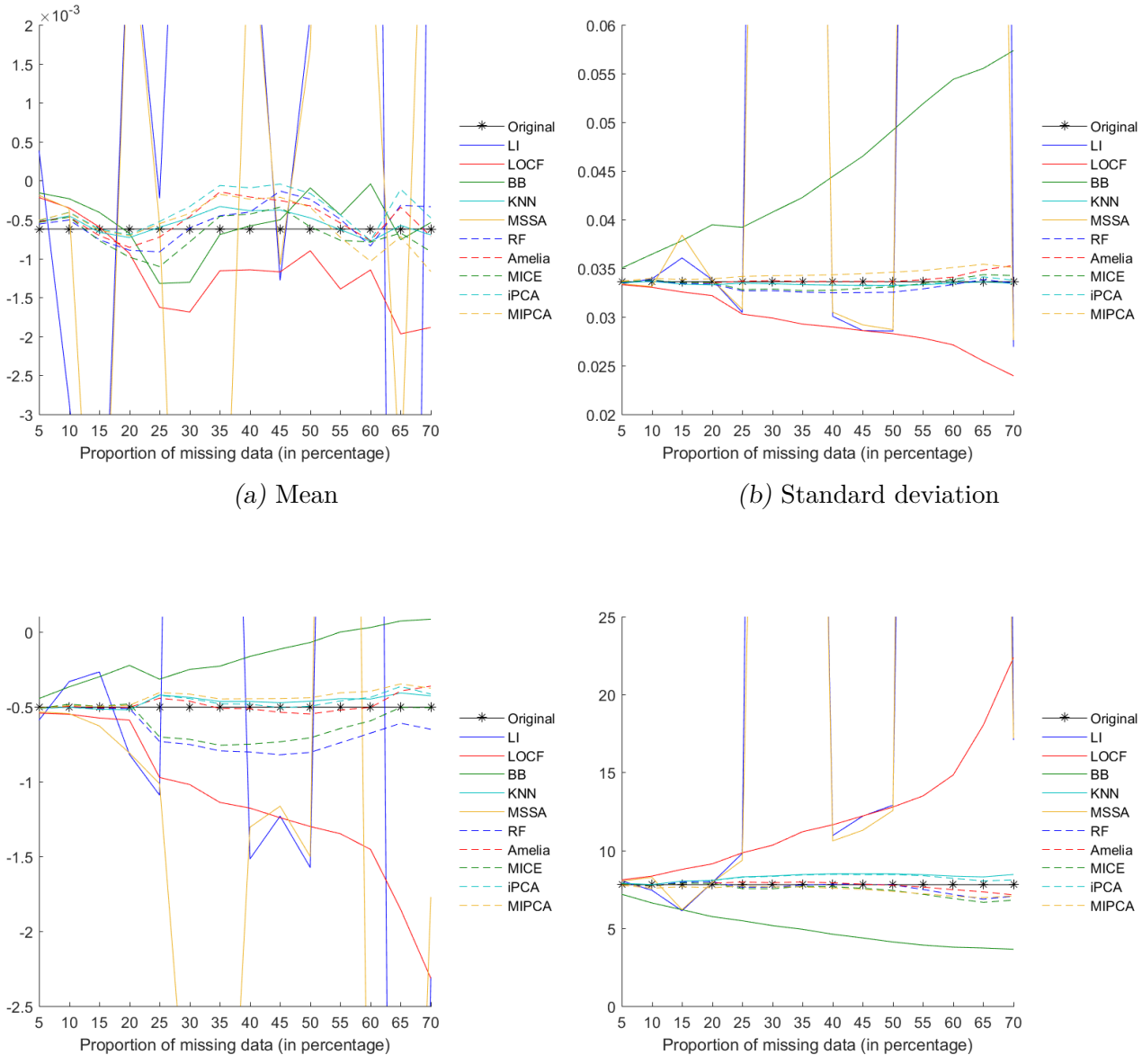
These results are different from those obtained in Section 3.3.3, where the same missing data mechanism was applied to simulated data. The absolute differences are much smaller when the completion methods are applied to simulated data. However, these differences are comparable to those obtained when imputation is applied to MCAR data on the same historical sample (see Section 3.5.2).

The methods obtaining the lowest absolute deviations are K-NN, random forests, and IPCA for both 10% and 30% missing data. Conversely, the worst-performing methods are linear interpolation, MSSA, and Brownian bridge. Moreover, the linear interpolation and the MSSA see their performance degrading strongly from 10% to 30% missing data. These two methods had already been removed from the graphical analysis in Section 3.3.3 due to their catastrophic results. The MSSA tended to behave like a linear interpolation, which extrapolated the data for this type of missingness pattern. Finally, the MICE algorithm appears to perform much better on the historical sample than on the simulated sample.

Statistical moments

Figure 3.6-2 represents the evolution of the performances of the completion methods in terms of statistical moments, according to the proportion of missing data.

Fig. 3.6-2: The first four statistical moments of the returns of the imputed data based on a matrix containing MAR data (successive missing data at the end of the first series of the sample based on a heuristic approach) according to the missingness proportion



(a) Mean

(b) Standard deviation

(c) Skewness

(d) Kurtosis

The first observation concerns the results of linear interpolation and MSSA, which tend to impute data in ways that deviate greatly from the original series. No matter the missing data proportion, these completion methods seem to completely distort the

series. The original graphs are available in Appendix K.1. This is not surprising given the results obtained with simulated data (see Section 3.3.3), where these two methods were removed from the graphical analysis because they obtained large deviations from the original series. The linear interpolation extrapolates data when the data are available at the end of the series. These results show that extrapolation can lead to disastrous imputations.

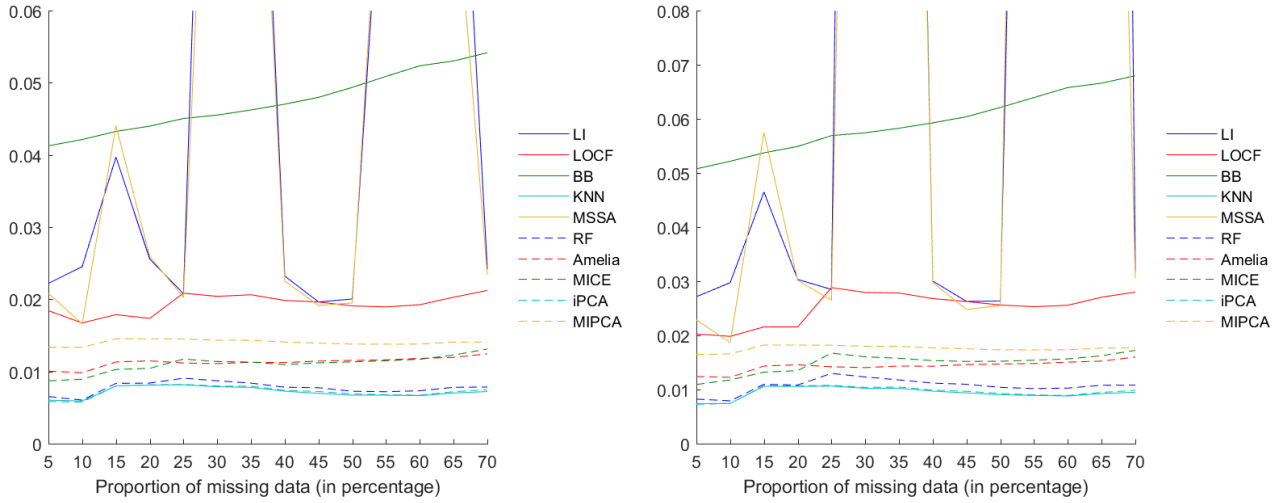
The Brownian bridge tends to make volatility explode. This is because the method is calibrated to the available values. The sample spans the COVID-19 crisis (March and April 2020). If this historical sample is divided into two parts (especially the first column), the annualized volatility of the first part (January to July 2020) is equal to 64%, and that of the second part (July 2020 to January 2021) is 41%. The highest variations are found at the beginning of the series, which means the volatility of observed values is even higher when the missingness proportion is high (as data are removed from the end). As the Brownian bridge aims to reproduce the law of observed values, it imputes data with high volatility when the missingness proportion is high.

Apart from these three methods (i.e., linear interpolation, MSSA, and Brownian bridge), the results are comparable to those on simulated data (see Section 3.3.3). Thus, random forests, Amelia, and MIPCA are still among the best-performing methods. MICE also provides good results. This method, which was one of the worst-performing ones on simulated data, is among the best when it comes to historical data (as in Section 3.5).

Proximity metrics

The completion methods are now compared in terms of closeness to the original series and, in particular, MAE and RMSE. The results associated with the imputation of missing data at the end of the series on the historical sample based on a heuristic approach are presented in Figure 3.6-3.

Fig. 3.6-3: MAE and RMSE between the return of the imputed data from a matrix containing MAR data (successive missing data at the end of the first series of the sample based on a heuristic approach) and the original data matrix according to the missingness proportion



(a) MAE *(b)* RMSE

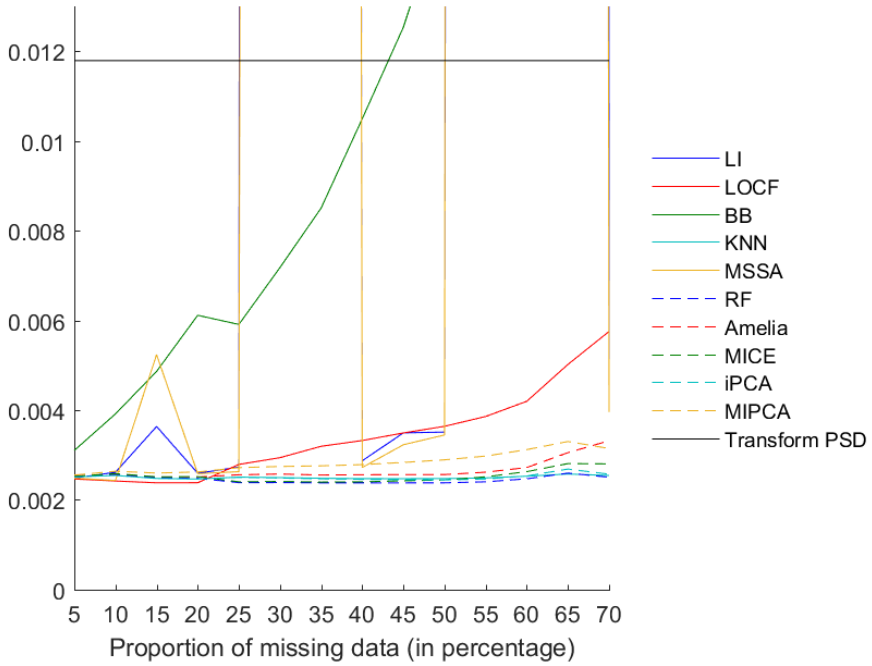
As for the MCAR mechanism, the orders of magnitude obtained here for this MAR mechanism are much larger than those obtained with the simulated sample (see Section 3.3.3). As before, this is because the simulated sample is highly correlated and has constant volatility, which is not the case for this historical sample.

As for the statistical moments, linear interpolation and MSSA obtain results that can be very high depending on the proportion of missing data in the series (original graphs in Appendix K.3). Finally, the K-NN, IPCA, and random forests are the methods that minimize both proximity measures. These measures are followed by MICE and Amelia, which obtain comparable results, and MIPCA. The IPCA obtains smaller proximity measures than MIPCA, even if both methods use the same number of principal components (see Appendix K.2). Brownian bridge and linear interpolation provide some of the worst results, as with simulated data (see Section 3.3.3).

Covariance matrices comparison

Figure 3.6-4 presents the differences between the covariance matrix of the imputed data and that of the original data, using a Frobenius norm.

Fig. 3.6-4: Covariance matrix differences, according to the Frobenius norm, based on original returns and the imputed returns from a matrix containing MAR data (successive missing data at the end of the first series of the sample based on a heuristic approach) according to the missingness proportion



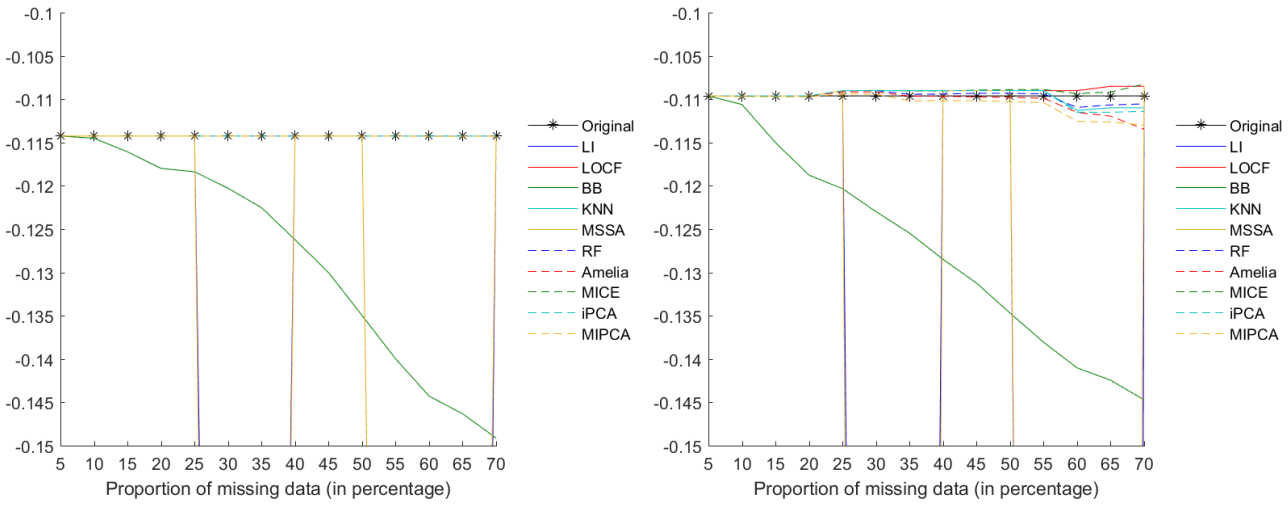
As expected, linear interpolation and MSSA distort the covariance matrix significantly (original graph in Appendix K.4). The Brownian bridge obtains high differences as was the case for the simulated sample (see Section 3.3.3).

For the other methods, the results are in keeping with those obtained with the simulated sample (see Section 3.3.3). Amelia, IPCA, MIPCA, K-NN, and MICE allow for the good preservation of the covariance matrix.

Value-at-risk and expected shortfall

The 1-day risk measures of the imputed data based on each completion method are shown in Figure 3.6-5.

Fig. 3.6-5: The 1-day risk measures computed from a matrix containing MAR data (successive missing data at the end of the first series of the sample based on a heuristic approach) according to the missingness proportion



(a) 1-day VaR at 99%

(b) 1-day ES at 99%

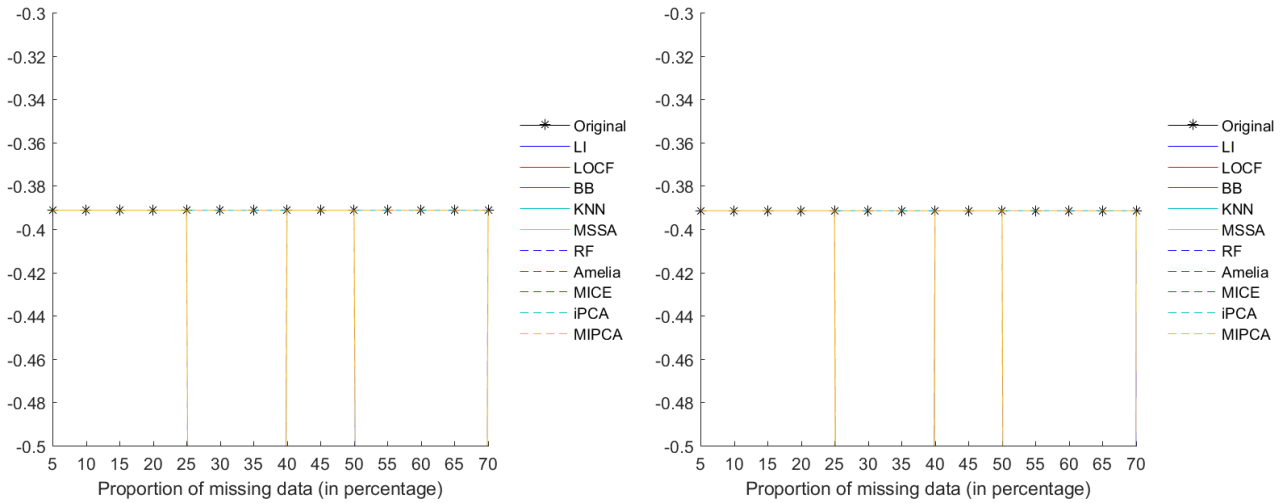
Since the beginning of the section, linear interpolation and MSSA (original graphs in Appendix K.5) have performed poorly on each of the criteria presented. This is also the case in terms of risk measures. Linear interpolation and MSSA tend to overestimate the level of VaR and ES.

The Brownian bridge tends to overestimate the level of risk measures as missing data appear in the sample. This is because the non-missing data used to calibrate the Brownian bridge are highly volatile (i.e., they correspond to the COVID-19 crisis).

Finally, the other completion methods manage to reproduce the correct level of VaR for all proportions of missing data. The methods also manage to reproduce ES fairly well but have a little more difficulty when the proportion of missing data is very high.

The 10-day risk measures computed from the imputation of missing data at the end of the series are presented in Figure 3.6-6

Fig. 3.6-6: The 10-day risk measures, computed from a matrix containing MAR data (successive missing data at the end of the first series of the sample based on a heuristic approach) according to the missingness proportion



(a) 10-day VaR at 99%

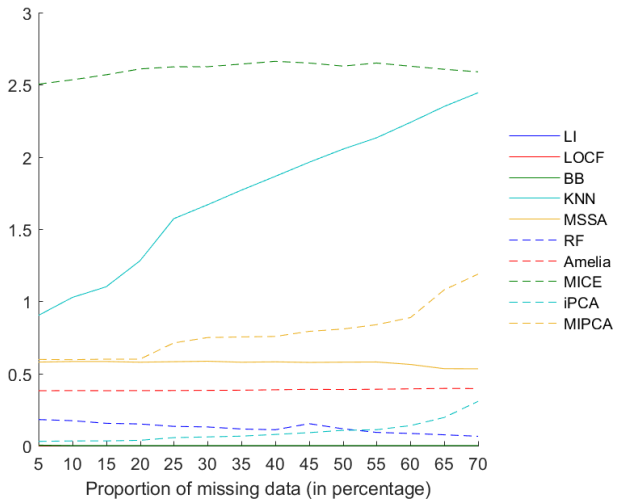
(b) 10-day ES at 97.5%

The 10-day risk measures are correctly reproduced by all the imputation methods, except for linear interpolation and MSSA. As for the 1-day horizon, but also the other analysis criteria, these two methods obtain results that are far from those of the original series.

Computation time

Finally, the completion methods are compared in terms of computation time in Figure 3.6-7.

Fig. 3.6-7: Computation time of the imputation of MAR data (successive missing data at the end of the first series of the sample based on a heuristic approach) according to the missingness proportion



The computation times are comparable to those obtained for imputing the simulated sample, except for the IPCA and MIPCA methods. These two methods impute this missing data mechanism more quickly on this historical sample than on the simulated sample (see Section 3.3.3).

Thus, these results confirm that linear interpolation and MSSA should be completely banned in the case of successive missing data at the end of the series. They also confirm the effectiveness of random forests, Amelia, MIPCA, and MICE on these historical data.

To compare the results on a less correlated sample, the same exercise must also be done with the historical sample based on a graphical Lasso approach.

3.6.2 Impact on a sample based on the graphical Lasso

The goal of this section is to use the sample constructed by a graphical Lasso approach, which was presented at the beginning of Section 3.5.3, to remove successive missing data at the end of the first column of the data matrix. This missing data mechanism was presented in Section 3.3.3 and involves removing a certain proportion of successive missing data from the end of the series.

As this same exercise was performed on the sample based on a heuristic approach, it can help in comparing the results obtained in the previous section (see Section 3.6.2) for this same missing data mechanism and putting them in perspective.

Finally, each comparison tool is computed as defined in Figure 3.1-2. The templates of all the graphs presented in this section have already been presented and detailed (i.e., what they represent and how they were obtained) in Section 3.1.4.

MCAR tests

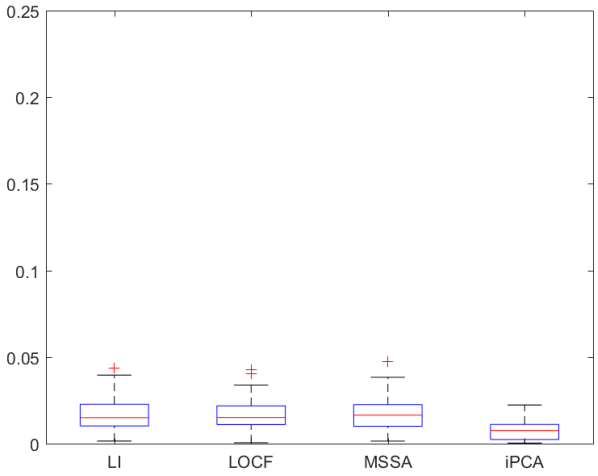
First of all, MCAR tests are applied to this second historical sample to see whether the results are comparable to those from the heuristic sample. As a reminder, this missingness mechanism is categorized as MCAR according to Little and Rubin [145], which means the tests are efficient if they do not reject the null hypothesis.

As previously, both tests were calculable. Jamshidian and Jalal's test [123] does not reject the null hypothesis, regardless of the missingness proportion, nor does Little's test [142], except for 30% missing data. These results are different from those based on the heuristic sample (see Section 3.6.1), which means further analysis on other samples is required before drawing general conclusions.

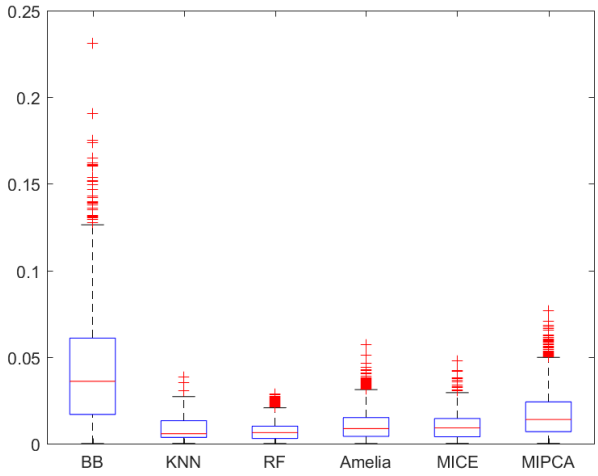
Preliminary results

First of all, the distribution of absolute return differences between original data and imputed data is presented in Figure 3.6-8. The results from methods with a random component are based on 100 imputations of reach missing data (no repetition is needed with methods without a random component).

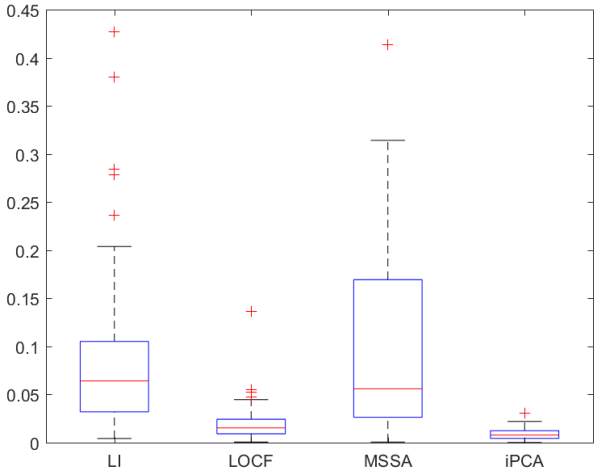
Fig. 3.6-8: Distribution of absolute return differences between the imputed series and original series for a sample containing 10% (at the top) and 30% (at the bottom) MAR data (successive missing data at the end of the first series of the sample based on the graphical Lasso)



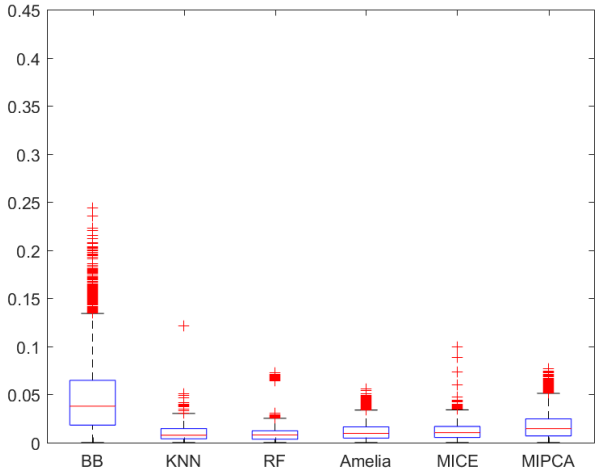
(a) Methods without a random component for 10% missingness



(b) Methods with a random component for 10% missingness



(c) Methods without a random component for 30% missingness



(d) Methods with a random component for 30% missingness

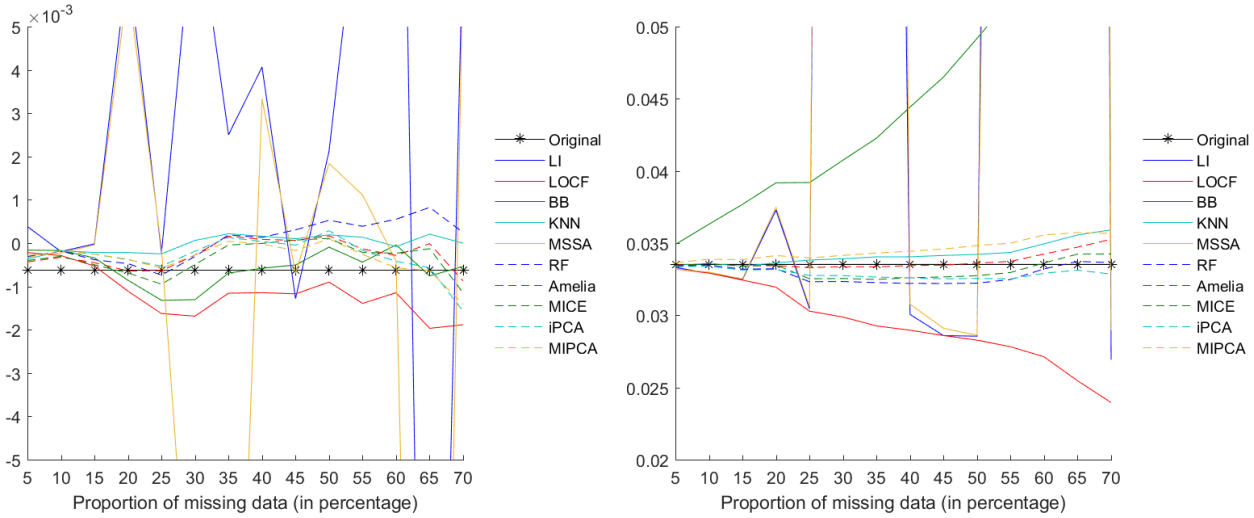
The results obtained for this historical sample are comparable to those obtained for the sample based on a heuristic approach (see Section 3.6.1). First of all, the

absolute differences are comparable in terms of scale. Moreover, the K-NN, random forests, and IPCA methods obtain the lowest absolute differences, while the linear interpolation, MSSA, and Brownian bridge methods obtain the highest differences. As before, the MICE algorithm obtains results comparable to those of Amelia and MIPCA on historical data, which was not the case on simulated data (see Section 3.3.3).

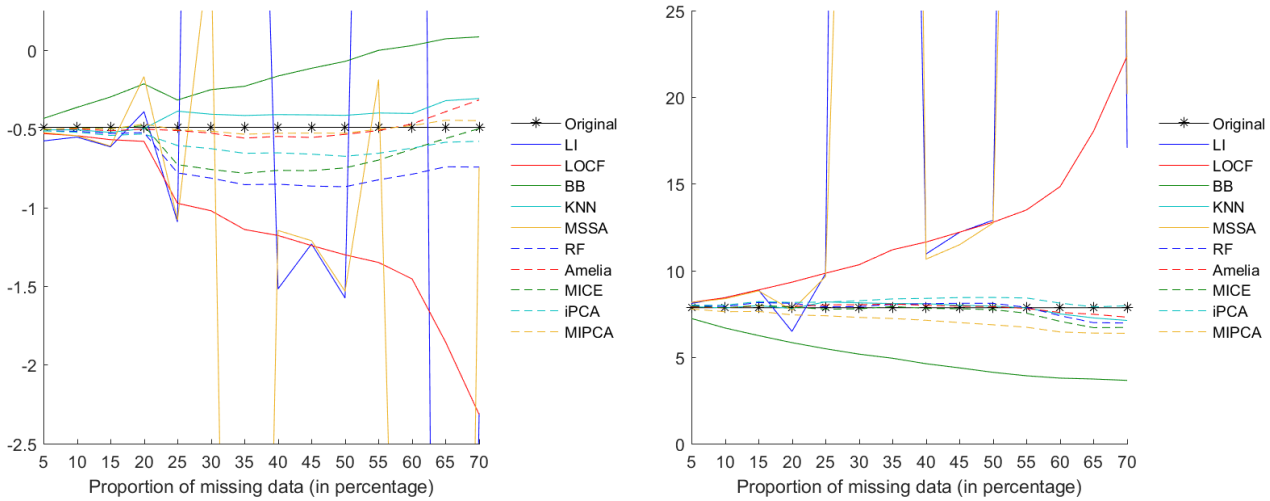
Statistical moments

Figure 3.6-9 represents the evolution of the performances of the completion methods in terms of statistical moments according to the proportion of missing data.

Fig. 3.6-9: The first four statistical moments of the returns of the imputed data based on a matrix containing MAR data (successive missing data at the end of the first series of the sample based on the graphical Lasso) according to the missingness proportion



(a) Mean *(b)* Standard deviation



(c) Skewness *(d)* Kurtosis

As in Section 3.6.1, linear interpolation and MSSA obtain statistical moments very far from those of the original series (original graphs in Appendix L.1).

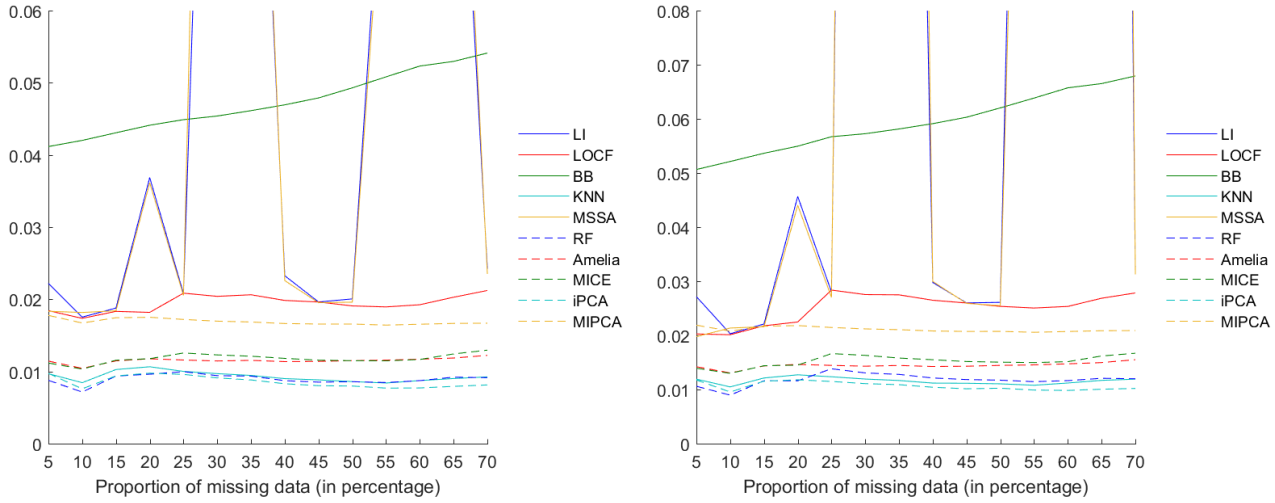
As in the previous section, the Brownian bridge is calibrated according to the non-missing values. Because the first part of the sample includes the COVID-19 crisis (i.e., high variations), the volatility of observed values increases as data are removed at the end of the series. This is why the volatility of data imputed by the Brownian bridge also increases.

Moreover, the random forests, Amelia, MIPCA, IPCA, K-NN,and MICE algorithms manage to stay close to the statistical moments of the original series, no matter how much data is missing.

Proximity metrics

The completion methods are now compared in terms of MAE and RMSE. The results from imputation methods applied to the end of the first column of the historical sample based on a graphical Lasso approach are presented in Figure 3.6-10.

Fig. 3.6-10: MAE and RMSE between the return of the imputed data from a matrix containing MAR data (successive missing data at the end of the first series of the sample based on the graphical Lasso) and the original data matrix, according to the missingness proportion



(a) MAE *(b)* RMSE

First of all, the orders of magnitude observed here are much larger than those obtained with the simulated sample (see Section 3.2.1). In a practical case on historical data, an expert should expect orders of magnitude comparable to those obtained above.

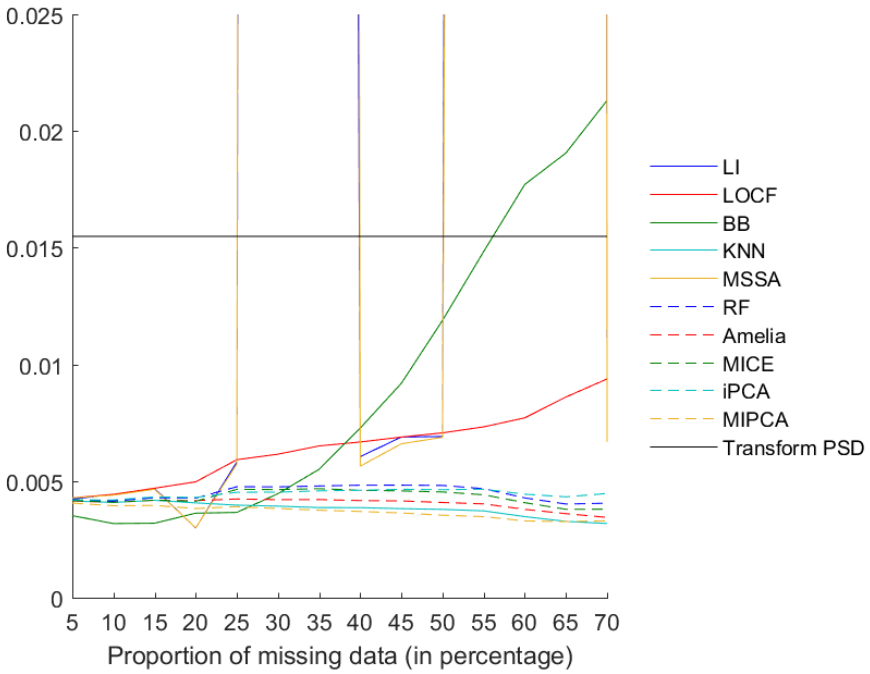
As before, the linear interpolation and MSSA methods obtain very unsatisfactory results. Their MAE and RMSE tend to explode for many missingness proportions, and they produce some of the worst results (original graphs in Appendix L.3).

The random forests, IPCA, Amelia, K-NN, and MICE algorithms obtain relatively close proximity measures. These methods are the most satisfactory here. IPCA and K-NN are as efficient as the random forests. However, IPCA obtains among the highest proximity measures when this mechanism is applied to simulated data (see Section 3.3.3). The performance of this method still has to be confirmed by studying a larger number of samples. The performance of IPCA is more satisfactory than that of MIPCA. However, these two methods use the same number of principal components (see Appendix L.2). Their performance differences are, therefore, due to the multiple imputations and regularized versions included in MIPCA.

Covariance matrices comparison

Figure 3.6-4 presents the differences between the covariance matrix of the imputed data and that of the original data by using the Frobenius norm.

Fig. 3.6-11: Covariance matrix differences, according to the Frobenius norm, based on original returns and the imputed returns from a matrix containing MAR data (successive missing data at the end of the first series of the sample based on the graphical Lasso) according to the missingness proportion

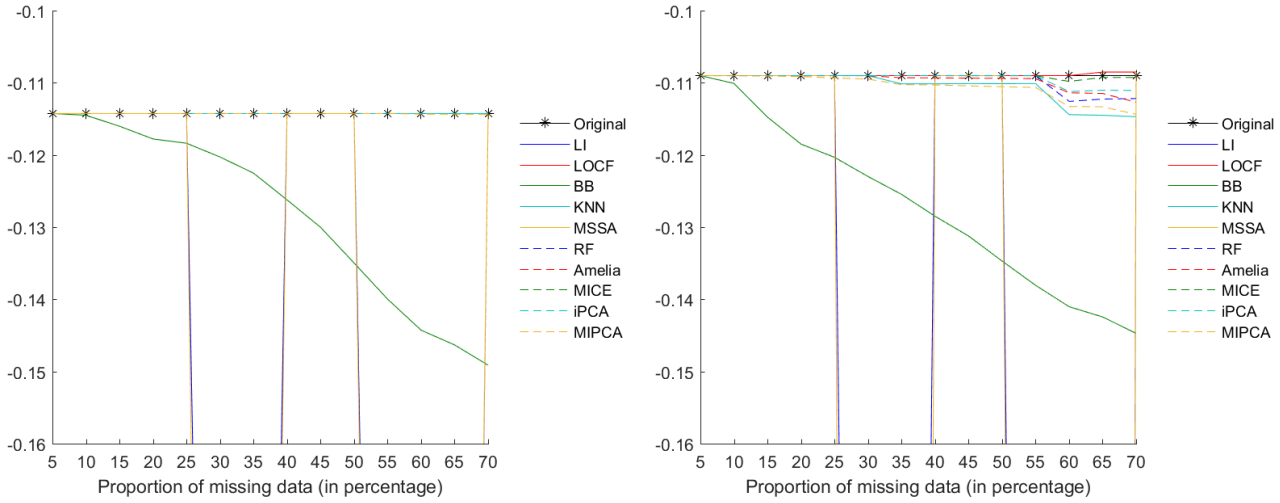


Unsurprisingly, both linear interpolation and MSSA strongly distort the covariance matrix (original graph in Appendix L.4). LOCF and the Brownian bridge tend to distort the covariance matrix with at least 25% missingness. The other completion methods obtain results comparable to those presented in the case of simulated data (see Section 3.3.3), but 10 times higher.

Value-at-risk and expected shortfall

The 1-day VaR and ES of the imputed data based on each completion method are shown in Figure 3.6-5.

Fig. 3.6-12: The 1-day risk measures computed from a matrix containing MAR data (successive missing data at the end of the first series of the sample based on the graphical Lasso) according to the missingness proportion



(a) 1-day VaR at 99%

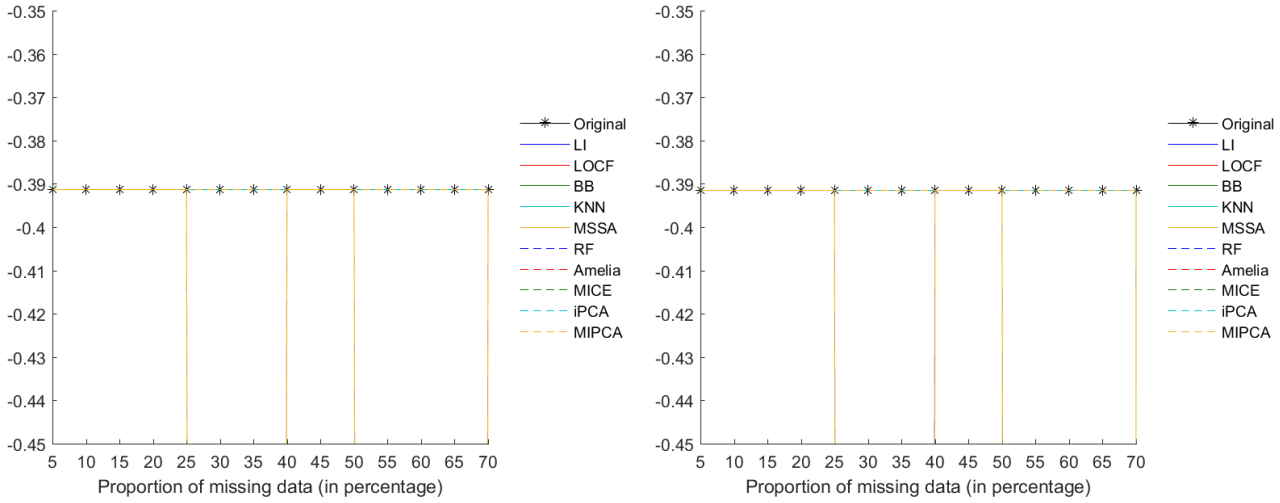
(b) 1-day ES at 99%

For this missingness mechanism and this simulated sample, all methods can reproduce the true VaR, except for linear interpolation, MSSA, and Brownian bridge. Linear interpolation and Brownian bridge extrapolate to impute missing data, and since the beginning, the MSSA obtain results comparable to those of linear interpolation.

In terms of ES, the completion methods have more difficulty reproducing the original ES, when the proportion of missing data becomes very important. Amelia, random forests, K-NN, IPCA, and MIPCA tend to move away slightly from the original ES. Only LOCF and MICE manage to reproduce the ES, even for the highest missingness proportions.

The 10-day risk measures are presented in Figure 3.6-13

Fig. 3.6-13: The 10-day risk measures computed from a matrix containing MAR data (successive missing data at the end of the first series of the sample based on the graphical Lasso) according to the missingness proportion



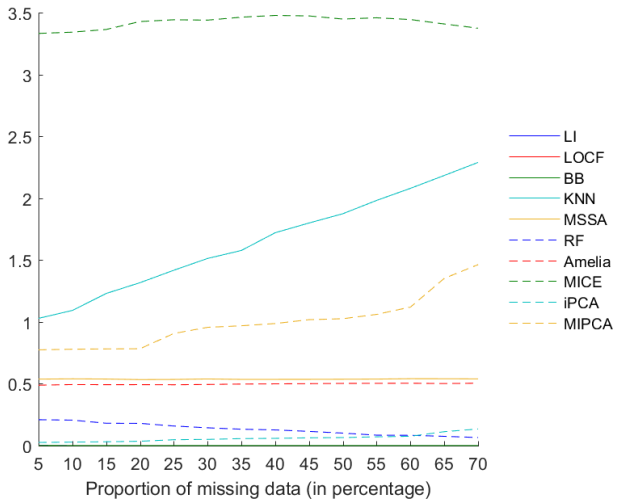
(a) 10-day VaR at 99% *(b)* 10-day ES at 97.5%

Here, the results are very close to those obtained in the historical sample based on a heuristic approach (see Section 3.6-13). All the methods, except linear interpolation and MSSA, can reproduce the 10-day risk measures efficiently (original graphs in Appendix L.5).

Computation time

The computation times needed by each algorithm are presented in Figure 3.6-14.

Fig. 3.6-14: Computation time of the imputation of MAR data (successive missing data at the end of the first series of the sample based on the graphical Lasso) according to the missingness proportion



The computation times presented here are comparable to those of the previous sample, except for the MICE algorithm, which requires 1 second for the imputation of the sample based on a heuristic approach.

Thus, these results remain comparable to those obtained for the sample based on a heuristic approach. Linear interpolation and MSSA obtain catastrophic results, regardless of the analysis criterion when the data are successively missing at the end of the series. Conversely, methods such as random forests, Amelia, MIPCA, and even MICE manage to handle this missing data mechanism efficiently, although the results obtained are often much higher than those obtained on simulated data.

Although the analysis must be repeated on a large number of samples to generalize the results, this section gives an idea of the good and bad algorithms that can be adopted in this case of missingness.

3.7 Discussion

Now that all the analyses have been conducted and all the results have been presented, this last section discusses the fundamental arguments in this chapter and the conclusions that emerged from it. The purpose of the exposition that follows is to revisit some observations which were made on the preceding pages and to suggest avenues for future research.

3.7.1 Results are conditioned by samples

First of all, the results in in Section 3.2, Section 3.3 and Section 3.4 naturally depend on the simulation of the samples, that is, their random generators, their log-normal processes, their particularly high correlation matrices, their Poisson jump processes and such like. This sensitivity was especially pronounced in the results of the MICE algorithm, which were among the worst for the simulated sample (without jumps and with constant volatility) but whose performance seemed to improve as the sample was distorted (with jumps and heteroskedasticity). Finally, its performance was better when it was applied to historical samples. The results were presented for illustrative purposes and highlight how certain algorithms are affected by specific samples as well as their functioning and specificities. Nevertheless, no general conclusions should be drawn from the results — they are based on only a few samples.

It has been noted on numerous occasions that future work ought to aim at reproducing these analyses across many samples to enable generalization. The most efficient completion methods will be identified as a result, and it will become possible to establish a confidence interval around imputed values or even around criteria (statistical moments, risk measures and such like). In the case of an illiquid financial product, it would be possible to use an imputation method to deduce a VaR and to identify confidence intervals around it. A quantile (among the tested scenarios) of this VaR may be deemed conservative by regulators. This type of process would be acceptable to them because it is based on statistically well-founded results.

Furthermore, the observed orders of magnitude for the historical samples are much larger than those observed for the simulated sample. The latter was, in fact, constructed so that it was very highly correlated and with constant volatility. In practice, such a sample is complicated to construct from historical data. Thus, even if the results are comparable, the orders of magnitude obtained on the historical samples appear to be more realistic.

3.7.2 Non-replicability of results

Although non-replicability was not discussed in this chapter, the results that were presented in it depend on execution. As explained in the previous chapter, algorithms that contain a random component, be it because of the bootstrap or because of the method itself, yield non-replicable results. These include the Brownian bridge, K-NN, random forests, Amelia, MICE and MIPCA, which were implemented in this PhD thesis. If the calculations are repeated, the results will be different. The same would also be true of VaR and ES. A problem arises, in that regulators typically want to reproduce results to confirm their validity. A method that provides non-replicable results may tempt banks to select executions that favor them, biasing risk analyses. Consequently, regulators may refuse to accept such imputation methods. One solution

would be to freeze results by always using the same seed value. Since the data are not perfectly random but pseudo-random, that is, dependent on the seed, it would be possible to always obtain the same results. The non-replicability problem can be solved in this fashion.

Moreover, this PhD thesis deals with the non-replicability of results on several levels, an issue that is present in the literature under the terms of repeatability and reproducibility. In 1972, Mandel [148] is the first to propose a definition. He defines repeatability as "the variability between replicate results obtained on the same material within a single laboratory" and the reproductibility by "the variability between results obtained on the same material in different laboratory". Thus, in the context of this PhD thesis, repeatability would consist in measuring the variability of the results obtained from the same completion algorithm and the same data sample, whereas reproducibility consists in the variability of the results obtained from the same completion algorithm but from different data samples.

The subject is widely exploited in the medical field (Gillard, Antoun, Burnet and Pickard [92], Bratlett and Frost [13], Traverso, Wee, Dekker and Gillies [200], etc.), with the aim of making their tools more reliable. On the other hand, it is a theme rarely discussed in finance. However, the use of many algorithms (completion or others) leading to non-repeatable or non-reproducible results is not rare. Moreover, the use of such methods would be more easily accepted by the regulator if studies concerning their repeatability and reproducibility were known.

3.7.3 Imputation method depends on criteria

As discussed in Section 2.2.1, it is important to identify the nature of the missing data. This can provide clues to the expert about how to complete missing data. But, one of the main learnings of this PhD thesis refers to the choice of the completion method which must be made according to the use. The expert must keep in mind what he wants to do with the data before imputing it. Imputation methods can be efficient for some purposes but not necessarily for others. Moreover, it is possible to distinguish two main objectives: replacing the missing data by a precise value, and replacing the missing data in order to preserve the correct distribution.

In the first case, the objective is to find the best possible prediction of the missing value, conditional on the missing data pattern and the observable data. The idea is to find a value in order to minimize the distance with the original value. This objective aims at being efficient in terms of proximity measures (MAE, RMSE, etc.). However, this objective is contradictory with the second one, which aims at imputing in order to preserve the good distribution of data. The goal is therefore to replace the missing data by a value that will be in the true conditional distribution, so as not to distort the distribution. This objective is thus more efficient to reproduce the true risk measures (VaR and ES). These two objectives are contradictory because an imputation that aims

to preserve the correct law does not necessarily imply an imputation that minimizes the deviation from the original data. Thus, the expert must be sure of his objective before imputing the missing data. As Carpenter and Kenward write in a book edited by Molenberghs, Fitzmaurice, Kenward, Tsiatis and Verbeke [157], in such scenarios, "the appropriate distribution should be carried forward," not the observations.

When the objective is to calculate VaR or ES, the expert is indeed interested in the distribution tails of a series and thus in its law. Therefore, methods such as Amelia or Brownian bridge, which attempt to reproduce a distribution, seem to be more relevant. However, if the goal is to reproduce the values of the original series as accurately as possible, with accuracy measured through proximity measures, then methods such as K-NN, that aim to find the right level of the missing value, might be more adapted.

Thus, some completion methods obtain, by definition, good or bad results according to certain criteria. This is particularly clear in the case of RMSE computed from data imputed by linear interpolation and that of data imputed by Brownian bridge. The RMSE can be written as

$$\begin{aligned} RMSE &= \sqrt{\frac{1}{n^{miss}} \sum_{i=1}^{n^{miss}} (y_i - y_i^*)^2} \\ &= \sqrt{\frac{1}{n^{miss}} \sum_{i=1}^{n^{miss}} (y_i^2) + \frac{1}{n^{miss}} \sum_{i=1}^{n^{miss}} \left(y_i^{*2}\right) - 2\frac{1}{n^{miss}} \sum_{i=1}^{n^{miss}} (y_i y_i^*)}, \end{aligned} \tag{3.7-1}$$

where y_i and y_i^* corresponds to the original values, and the imputed values, respectively, for all $i = 1, \ldots, n^{miss}$, the set of observations that are missing before imputation.

By supposing that the mean of returns (for y_i and y_i^*) is zero (as it is frequently the case with financial data), this expression is equivalent to writing the RMSE as

$$RMSE = \sqrt{\sigma^2 + \sigma^{*2} - 2\rho\sigma^2\sigma^{*2}}, \tag{3.7-2}$$

where σ^2 and σ^{*2} are the variance of y_i and y_i^*, respectively, for all $i = 1, \ldots, n^{miss}$, and ρ the correlation between y_i and y_i^*.

Moreover, a method such as linear interpolation, which aims to predict the closest value to the original price (i.e. to minimize the proximity measures), imputes data by without adding noise in the imputed return series. In this case, the variance of imputed data is null ($\sigma^{*2} = 0$), which means that its RMSE is equal to $\sqrt{\sigma^2}$. On the other hand, a method such as the Brownian bridge, which attempts to predict missing returns by reproduce a law, has a variance of imputed data equal to that of the original data (in theory). Therefore, its RMSE is equal to $\sqrt{2(1-\rho)\sigma^2}$.

Therefore, if the correlation between original data and imputed data lower than 0.5 then, the RMSE of linear interpolation is always lower than that of Brownian bridge. On the other hand, the Brownian bridge performs better only when the y_i are highly correlated with y_i^*, for all $i = 1, \ldots, n^{miss}$. Since the Brownian bridge imputes by

randomly drawing from the distribution of non-missing data, it is not expected that the correlation between the original data and the imputed data (by this method) is close to 1. Thus, a method such as linear interpolation is expected to be more efficient in terms of MAE and RMSE than the Brownian bridge. This is true for all methods aiming at estimating the law of the observed data, in order to impute the missing data by drawing randomly in this law, as it is the case for Amelia for example. This example shows that algorithm selection must accord with the criteria of the analysis.

As a general matter, a method can be particularly close to the original series according to one or more criteria, but not on others. This highlight the fact that the choice of completion method is sensitive to its intended application. The most appropriate method for estimating regulatory risk measures is unlikely to coincide with the most appropriate method for reproducing a covariance matrix to solve portfolio management issues or to forecast data. In the last case, proximity measures are probably more suitable.

Table 3.7-1 summarizes the imputation methods to be used according to the intended application and to the missingness proportion, based on the results observed in this chapter.

Tab. 3.7-1: Completion methods to be used depending on the application and the proportion of missing data

	Analysis criteria	**5% to 25% of missing data**	**30% to 50% of missing data**	**55% to 70% of missing data**
Risk Management and forecast	-Statistical moments -Risk measures	Random Forests Amelia MIPCA	Random Forests Amelia MIPCA	Random Forests
Portfolio Management	-Covariance differences	Linear interpolation Random Forests Amelia MIPCA	Random Forests Amelia MIPCA	Random Forests MIPCA
Data replication and forecast	-Proximity measures	Random Forests	Random Forests Amelia MIPCA K-NN	Random Forests MIPCA K-NN

If the expert wants to use a completion method for risk management purposes, then the analysis criteria will be statistical moments and risk measures (VaR and ES). Random forests are considered to be efficient, even when the proportion of missing data

is important, which is not the case for Amelia and MIPCA, which are efficient but tend to be affected by a high proportion of missing data. These choices reflect the banks' point of view, but the regulator may also use completion methods. From the regulator's point of view, the LOCF method is the most conservative in terms of risk measures. If the impact is non-massive on risk measures, then this method would be preferred by the regulator, out of mistrust of the banks. Indeed, LOCF imputations generally lead to an overestimation of the VaR and the ES, which implies more capital charges.

On the other hand, if the objective is related to portfolio management and in particular to the preservation of the covariance matrix, the analysis criterion will be based on the differences in terms of the covariance matrix. The random forests and MIPCA methods have often shown good results concerning this criterion, for any missingness proportion. This is also the case for Amelia, except when the proportion of missing data becomes too important. Finally, the linear interpolation has often been satisfactory for low proportion of missing data.

Finally, if the expert's objective is to replace missing data or to forecast, then the analysis will rely on proximity measures, such as the MAE and the RMSE. In this case, random forests remain among the most efficient methods. When the missingness proportion is low, this method tends to obtain proximity measures far below those of the other methods, but when the proportion is higher (beyond 30%), the Amelia, MIPCA and K-NN methods obtain results comparable to those of random forests. On the other hand, as for the other criteria, Amelia is not a good solution when more than half of the sample is missing.

The MICE algorithm needs to be analyzed on more samples to ensure its performance. This algorithm is indeed as efficient as Amelia and MIPCA on historical samples, but it is also one of the least efficient on simulated samples. Thus, reproducing the analysis on a large number of samples remains essential to confirm its performance.

3.7.4 A method is not an algorithm

One of the main insights from this PhD thesis concerns the distinction between methods and algorithms. A method makes little sense when taken in isolation, without steps such as data preparation, preprocessing, result formatting and so on. These steps convert a method into an algorithm, and they are essential to effective imputation.

It emerged in the course of the exposition that MSSA can lead to very different results depending on its implementation and its algorithm. In the previous chapter, two possible uses of MSSA were presented, namely forecasting and imputation. The two involve two different algorithms. In this way, the same method was applied to two different processes with different objectives. Unsurprisingly, it obtained different results.

The same argument applies to random forests. Random forests has been implemented by both Breiman [43] (the *rfimpute* function) and by Stekhoven and Bühlmann

[194] (the *missForest* function). They use the random forests method in two very different imputation algorithms (as discussed in Section 2.4.5). The function used in this PhD thesis is the one of Stekhoven and Bühlmann [194]. However, when the research process began, the analysis was premised on a (homemade) implementation that was aligned with Breiman's [43], which was used for comparative purposes. The results of this first version were very far from the performance obtained here (in fact, they were very poor). The Stekhoven and Bühlmann [194] algorithm is much more efficient for imputing missing data. This development highlights the importance of documentation. Without accurate documentation, results can vary between the highly satisfactory and the utterly catastrophic.

3.7.5 Operational risk: non-calculability and documentation

The analyses in this chapter also highlighted an important operational risk that attends to the use of completion methods. Despite the implementation of significant precautions, Section 3.2.2 revealed that K-NN, Amelia and MIPCA could not run beyond a certain proportion of missing data. The K-NN was non-calculable on a sample that contained more than 10% missing data, the Amelia algorithm was not suitable for a missingness proportion of more than 30%, and MIPCA failed when missingness exceeded 70%. These methods had been among the best performers in the course of the analysis. However, in most sections of the chapter, the analyses were conducted with missing data restricted to a single column, which is obviously not very realistic for historical data. However, the application of these methods on a sample that contains missing data in all (or almost all) of its columns revealed a fundamental limitation that can be generalized to all models: a method may be inapplicable.

In an automated system, where these methods may be integrated into the calculation of daily risk measures or even just daily P&L in an automated system, where these methods may be integrated into the calculation of daily risk measures or even just daily K-NN, Amelia and MIPCA. The MCAR tests were also not always calculable, foreclosing the possibility of analysis on certain occasions. In addition, there are still many cases in which imputation methods are incalculable, a serious operational risk for banks. For banks, it is important to have access to data at all times. The use of an imputation method therefore implies a necessary operational risk for the bank. That risk should not be underestimated.

The operational risk could also be observed in the implementation and documentation of the methods. Documentation is even more important than implementation. The results obtained by the MSSA method are relatively disappointing, but this may be due to the implementation rather than the method. No package allows imputation of missing data by MSSA. The implementation that was presented here was developed specifically for the purposes of the PhD thesis. Implementing a sophisticated algorithm

is not always simple, and it is possible that some elements were missing. The documentation of Dash and Zhang [65], the Bloomberg researchers in charge of the model, was used. That documentation may have been left incomplete voluntarily for commercial reasons. Complete implementation would allow any bank to implement the method, depressing sales and reducing profits.

Commercial considerations aside, the foregoing offers a perfect illustration of the operational risk that can result from incomplete or poor documentation. It is important to be able to explain why an algorithm works as well as why it fails. However, these explanations are not always obvious, especially with algorithms growing in sophistication and including ever-increasing numbers of steps. Precise documentation can help a bank avoid operational risk and attain regulatory approval. Whatever it was that prompted Dash and Zhang to leave their documentation incomplete, questions remain about their implementation, their choice of parameters and their input (as discussed in Section 2.4.4).

3.7.6 Amelia's sensitivity to a high proportion of missing data

The Amelia algorithm was among the best-performing imputation methods on many criteria and for many types of missing data. This said, its performance suffers when the proportion of missing data becomes too significant. It appears that this fall in imputation quality is always accompanied by a sharp increase in computation time. When the algorithm is scrutinized carefully, it becomes apparent that for high proportions of missing data, some bootstrapped samples lead the EM algorithm to a larger number of iterations. This tendency can be attributed to the initialization of the parameters of the EM algorithm. The staring values are based on the mean and the covariance matrix of the complete data from the bootstrapped sample, that is, the parameters obtained after a listwise deletion of the bootstrapped sample is performed, as recommended by Honaker, Joseph, King, Scheve and Sigh [112] and as indicated by Little and Rubin [145].

However, when the proportion of missing data is high, the proportion of complete data in the bootstrapped samples can be low. The Amelia algorithm is applied to price returns and the missing data are injected into the prices. Consequently, the algorithm only uses cases in which two successive prices are observed, equivalent to $30\%^2 = 9\%$ fully observed returns when 70% of prices in the first column are missing. This coincides with the proportion in the historical sample (see Table 3.7-2). It follows that some bootstrapped samples may contain insufficient observed data, which impacts the initialization of the parameters. The initial parameters that are based on few observations can lead the algorithm to a local optimum that is far from the global one.

Tab. 3.7-2: Average proportion (number) of missing returns (among the 100 missingness scenarios) associated with the proportion of MCAR raw data injected into the first column of the simulated sample of length 261 (260 for return sample)

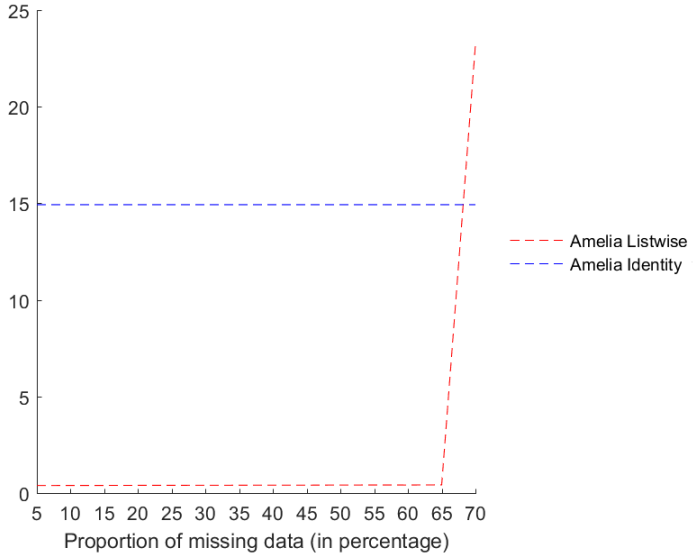
	Proportion (and number) of missing returns associated with missing data													
Data	**5%** (13)	**10%** (26)	**15%** (39)	**20%** (52)	**25%** (65)	**30%** (78)	**35%** (91)	**40%** (104)	**45%** (117)	**50%** (130)	**55%** (143)	**60%** (156)	**65%** (169)	**70%** (182)
Return	10% (25)	19% (49)	28% (72)	36% (93)	44% (113)	51% (132)	57% (149)	64% (165)	69% (180)	75% (194)	79% (207)	84% (218)	87% (227)	91% (236)

Some bootstrapped samples lead the EM algorithm to estimate the wrong parameters and thus the wrong law. The resultant imputations are far from the original series. The algorithm loses much of its efficiency when the proportion of missing data is too significant. Very large standard deviations from one missingness scenario to another are also observed, which is also related to high proportions of missing data. Imputation quality is highly sensitive to missingness.

Honaker, Joseph, King, Scheve and Sigh [112] have implemented two possibilities for parameter initialization into the Amelia function. One is based on listwise deletion; the other uses an identity matrix. If the covariance matrix of Amelia is initialized by an identity matrix, then the rapid degradation of imputation quality when high proportions of data are missing should no longer be observed. Moreover, beginning with an identity matrix means that the EM algorithm starts without any information about the correlations between the columns. Therefore, the Amelia algorithm with identity matrix initialization is applied to simulated data that only contains MCAR data in the first column. This approach enables the results to be compared to those obtained previously.

Firstly, computation time (presented in Figure 3.7-1) is constant when the algorithm is always initiated by the same matrix. It soars when parameters that are obtained after listwise deletion are used.

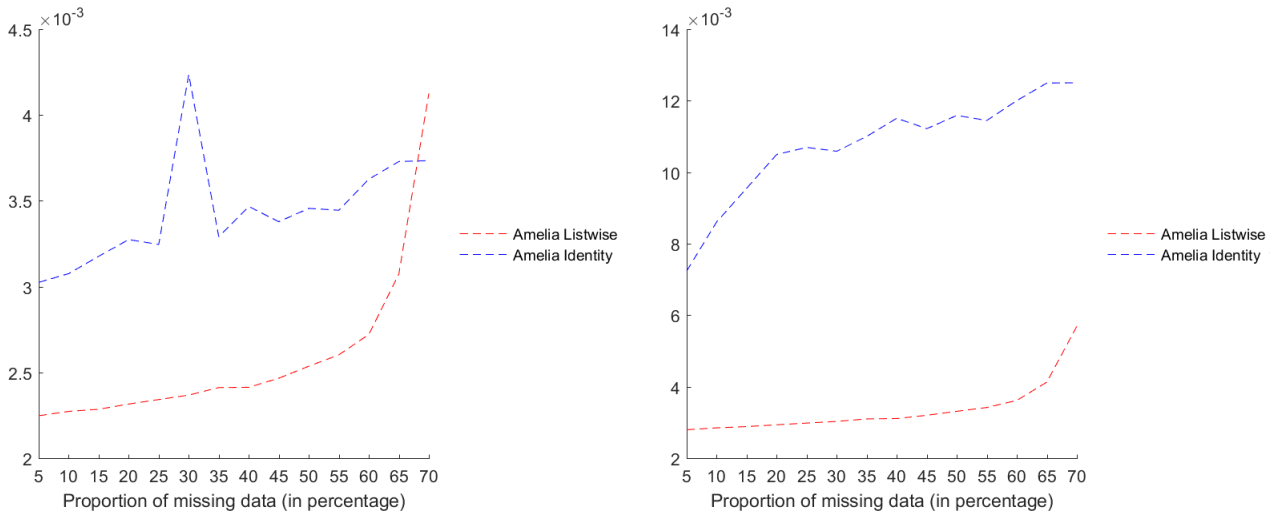
Fig. 3.7-1: Computation time (in seconds) of Amelia algorithm initialized by list-wise deletion* and by identity matrix



* The results for Amelia Listwise are those from Section 3.2.1

Furthermore, both initialization methods affect the fluctuations of the proximity measures (see Figure 3.7-2).

Fig. 3.7-2: Proximity measures of Amelia algorithm initialized by listwise deletion* and by identity matrix



(a) MAE *(b)* RMSE

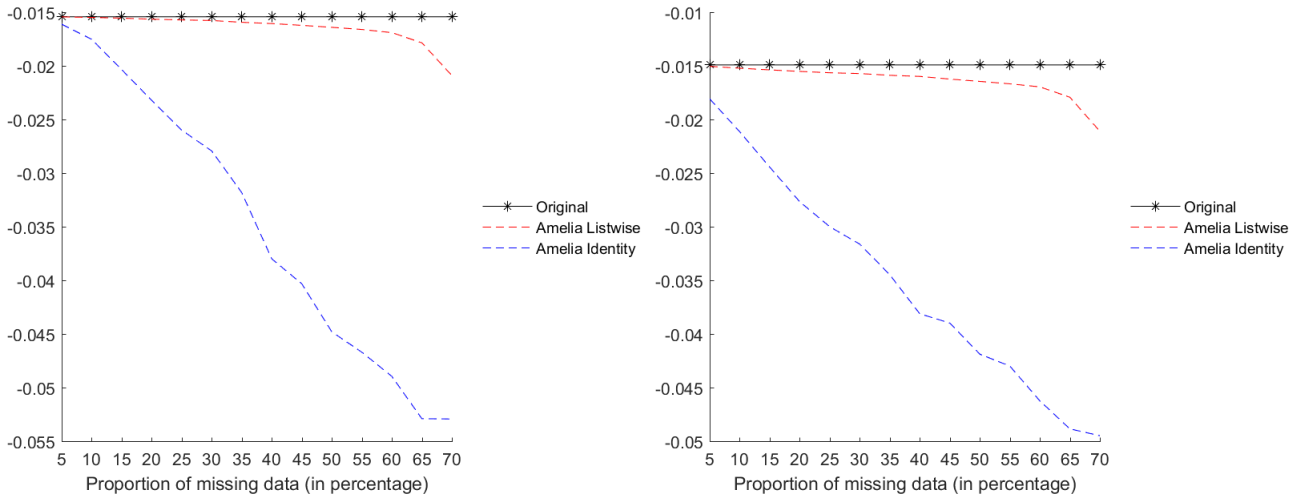
* The results for Amelia Listwise are those from Section 3.2.1

When the EM algorithm is initialized by listwise deletion, the proximity measures increase slightly at first. When the proportion of missing data becomes too significant, the increase intensifies. When the model is initialized by identity matrix, the degradation is progressive. The jump in MAE for the method initialized by identity matrix seems to be attributable to sampling error. The fluctuation of the proximity measures when initialization is by identity matrix seems to follow a stable upward trend; with listwise deletion, proximity measures increase slightly before exploding.

The difference between the proximity measures obtained from the two initializations is important, even for 5% missing data. In that case, MAE is about 30% higher when initialization is by identity matrix. RMSE is more than doubled. These results may serve as a reminder of the sensitivity of the method to parameter initialization and of the propensity of the EM algorithm to converge to a local maximum that does not coincide with the global one.

The initialization of the parameters also impacts risk measures. The 1-day VaR and ES are represented in Figure 3.7-3.

Fig. 3.7-3: 1-day risk measures of Amelia algorithm initialized by listwise deletion* and by identity matrix



(a) 1-day VaR at 99% *(b)* 1-day ES at 97.5%

* The results for Amelia Listwise are those from Section 3.2.1

It is possible to observe that the method that uses the identity matrix tends to underestimate the amount of VaR and ES very strongly. The worst loss, which occurs at a threshold of 99%, and the average loss, which occurs at a threshold of 97.5%, is much higher when the EM algorithm is initialized by identity matrix than when it is initiated by listwise deletion. Moreover, the fluctuation of the listwise deletion-initialized method is slightly conservative over most of the missingness proportion, and it tends to be more conservative when that proportion is particularly high. Identity matrix initialization causes VaR to become approximately 3.5 times more negative (and ES to become 2.7 times more negative) than with listwise deletion for missingness proportions that range between 5% and 70%. In addition, identity matrix initialization leads to increasingly conservative risk measures in an almost linear fashion; listwise deletion initialization tends to remain close to the original risk measures. The steep decline in imputation quality when the missingness proportion is particularly high is therefore due to the use of the listwise deletion-initialized method. Nevertheless, that method is still preferable to using an identity matrix, which degrades the performance of the algorithm considerably.

Parameter initialization is a well-known problem of the EM algorithm, and it remains relevant. In 2016, Shiremann, Steinley and Brusco [188] showed the impact of different parameter initialization methods on the performance of the EM algorithm. In 2018, Li and Chen [140] combined K-NN and k-means to determine its initial parameters. It is possible to use a method of initialization that differs from the one

recommended by Honaker, Joseph, King, Scheve and Sigh [112] to stabilize the results of Amelia.

Another step in the Amelia algorithm that receives little coverage in Honaker, Joseph, King, Scheve and Sigh's documentation [112] concerns data normalization. One of the preliminary steps of the EM algorithm consists in normalizing the bootstrapped sample data by using the means and standard deviations of the non-missing data of that very sample. Each bootstrapped sample is normalized by using the parameters of the observed data, which implies the same issues as those of parameter initialization. If the parameters are based on too few observations, which is the case for the highest proportions of missing data, then the normalization distorts the data rather than removing noise. Moreover, once the imputation is complete, denormalization is performed from the same starting parameters (means and standard deviations from non-missing data), which yields the same values for the non-missing data before and after the application of the Amelia algorithm. However, if the parameters obtained by the EM algorithm are too different from those of the non-missing data, the distribution of the imputed data is distorted.

It would be possible to denormalize the data from the parameters obtained after the convergence of the EM algorithm, but the imputed sample, including its non-missing values, would be entirely different from the initial one. That an imputation algorithm returns different non-missing data after completion is problematic, and it would impact proximity measures considerably. However, if the purpose of an imputation is to compute risk measures, then the different non-missing data are irrelevant because the criterion is based on the distribution. Once more, it emerges that the algorithm must be defined in line with the needs of the analysis, especially the intended use of the imputed data.

Unfortunately, the Amelia algorithm, in the form in which it is implemented by Honaker, Joseph, King, Scheve and Sigh [112], does not return any variables with the details of the bootstrapped samples, which makes it impossible to compare the parameters of the bootstrapped sample with the parameters of the EM algorithm. Reimplementation is necessary to observe the magnitude of the effect of data normalization on imputation quality. It follows that, like the initialization of the parameters, the normalization of the data affects the quality of an imputation. Its effect can be sizable when the proportion of missing data is high.

3.7.7 Paradoxical results

As a reminder, missing data from Section 3.2.1 were injected only in the first column for given proportions, whereas in Section 3.2.2, the same missingness proportions where injected in each column of the data matrix (except the last one). For example, in Section

3.2.1, 10% MCAR data were injected only in the first column, whereas in Section 3.2.2, 10% MCAR data were injected in the first column, then 10% in the second column, and so forth. Considering the whole sample, the missingness proportion used in Section 3.2.1 are much lower than those used in Section 3.2.2.

This procedure were made in order to observe the decline in the performance of the completion methods Section Section 3.2.1 and Section Section 3.2.2. It transpired that the presence of missing data across the whole sample leads to better results for some methods. In other words, the less information the methods have, the better their imputation quality. This observation applies to IPCA, MIPCA, MICE and Amelia in particular. The proximity measures were especially revealing. Table 3.7-3 (already presented in Section 3.2.2) shows the results for some proportions of missing data. For example, the MAE of returns imputed from IPCA is equal to 0.188% when 10% of missing data are in the first column whereas it is equal to 0.052% when 10% of missing data are spread across all columns.

Tab. 3.7-3: Average MAE and RMSE for 10%, 30%, 50% and 70% of missing data in the first columns versus missing data across the whole matrix of the simulated sample

		Missingness proportion in the first column*				**Missingness proportion in all columns****			
		10%	**30%**	**50%**	**70%**	**10%**	**30%**	**50%**	**70%**
MAE (10^{-3})	**IPCA**	1.88	2.42	3.87	4.16	0.52	1.53	2.89	4.65
	MICE	6.38	6.79	7.52	8.26	1.32	3.56	5.70	7.38
	MIPCA	2.20	2.24	2.52	3.40	0.50	1.67	3.37	-
	Amelia	2.27	2.37	2.54	4.13	0.50	1.64	-	-
RMSE (10^{-3})	**IPCA**	2.39	3.10	5.05	5.69	1.34	2.85	4.39	6.45
	MICE	7.85	8.58	9.81	11.43	3.32	6.17	8.33	10.41
	MIPCA	2.77	2.88	3.29	4.70	1.28	2.93	5.06	-
	Amelia	2.85	3.03	3.31	5.69	1.29	2.92	-	-

* Results from Section 3.2.1
** Results from Section 3.2.2

MICE always obtains lower measures of closeness when the data have been imputed from the whole sample (right part of the table) than when they have been imputed from the first column only (left part of the table). The same is true of Amelia in the calculable scenarios as well as of IPCA and MIPCA, except for missingness proportions that are too high. This tendency is especially pronounced when the proportion of missing data is low. As that proportion increases, the two patterns of the missing data tend to have the same of proximity measures.

It is not obvious to find a logical explanation for these paradoxical results, since they imply that information must be removed from complete columns in order to improve the imputation quality.

There is no obvious explanation for these paradoxical results. They imply that information must be removed from complete columns in order to improve imputation quality. The results are, as is usual with MCAR data, based on the imputation of 100 missingness scenarios. It is therefore important to analyze variability between missingness scenarios and to compare it across missing data patterns. Table 3.7-4 presents the standard deviations of the MAE and RMSE obtained from the 100 missing data scenarios that were tested for 10%, 30%, 50% and 70% missing data. Results for other proportions are available in Appendix A.4 and Appendix B.6.

Tab. 3.7-4: Standard deviation of the proximity measures among all missingness scenarios, for 10%, 30%, 50%, and 70% missing data in the first column versus missing data across the whole matrix of the simulated sample

		Missingness proportion in the first column*				**Missingness proportion in all columns****			
		10%	**30%**	**50%**	**70%**	**10%**	**30%**	**50%**	**70%**
σ_{MAE} (10^{-3})	**IPCA**	0.28	0.34	0.46	0.83	0.32	0.36	0.28	0.73
	MICE	0.61	0.59	1.08	1.27	0.71	0.49	0.84	1.16
	MIPCA	0.17	0.11	0.22	0.43	0.28	0.23	0.22	-
	Amelia	0.16	0.08	0.12	1.19	0.29	0.23	-	-
σ_{RMSE} (10^{-3})	**IPCA**	0.33	0.44	0.61	1.19	0.66	0.65	0.45	1.08
	MICE	0.78	0.77	1.53	2.15	1.00	0.60	1.32	1.90
	MIPCA	0.23	0.17	0.28	0.65	0.46	0.28	0.31	-
	Amelia	0.20	0.12	0.17	1.61	0.47	0.26	-	-

* Results from Section 3.2.1
** Results from Section 3.2.2

While MAE and RMSE tend to be much lower when missing data are distributed throughout the sample, this is usually (but not always) accompanied by an increase in the variability of these measures from one missing data scenario to another. In other words, the proximity measures may be higher when the missing data are in the first column, but they are also more representative of the trend observed among the 100 missingness scenarios. Missingness scenarios tend to produce risk measures that are further away from their average when the missing data are distributed across the whole sample. Moreover, the lower the proportion of missing data, the greater the variability from one missingness scenario to another. The largest standard deviation differences are in fact observed for low proportions of missing data.

These algorithms therefore lead to more stable measures of proximity from one

missingness scenario to another when the missing data are in a single column. They tend to be much more variable when the data to be imputed are spread across the whole sample.

This tendency is less pronounced when the analysis focuses on risk measures. Table 3.7-5 represents the 1-day VaR and ES of the first column when the missing data is concentrated (results from Section 3.2.1 1 on the left) and when it is spread across the matrix (results from Section 3.2.2 on the right). For example, the original VaR means that the maximum loss for the first column of the original matrix is -1.54%. In the case of IPCA imputation, the VaR is equal to -1.55% when 10% missing data are in the first column only, whereas it is -1.54% when 10% missing data are in each columns of the data matrix.

Tab. 3.7-5: Average 1-day risk measures for 10%, 30%, 50% and 70% of MCAR data in the first columns versus missing data across the whole simulated sample

		Missingness proportion in the first column*				**Missingness proportion in all columns****			
		10%	**30%**	**50%**	**70%**	**10%**	**30%**	**50%**	**70%**
$VaR_{99\%}^{1-day}$ (10^{-2})	**Original**	-1.54				-1.54			
	IPCA	-1.55	-1.54	-1.61	-1.8	-1.54	-1.53	-1.65	-2.06
	MICE	-1.67	-2.01	-2.23	-2.54	-1.55	-1.76	-2.08	-2.4
	MIPCA	-1.55	-1.55	-1.58	-1.82	-1.54	-1.54	-1.77	-
	Amelia	-1.55	-1.58	-1.64	-2.09	-1.54	-1.56	-	-
$ES_{97.5\%}^{1-day}$ (10^{-2})	**Original**	-1.49				-1.49			
	IPCA	-1.51	-1.52	-1.6	-1.82	-1.49	-1.52	-1.71	-2.03
	MICE	-1.69	-2.02	-2.26	-2.55	-1.53	-1.77	-2.12	-2.39
	MIPCA	-1.52	-1.54	-1.6	-1.85	-1.5	-1.53	-1.82	-
	Amelia	-1.52	-1.57	-1.64	-2.11	-1.5	-1.55	-	-

* Results from Section 3.2.1
** Results from Section 3.2.2

Evidently, the risk measures are less affected than the proximity measures. Overall, the risk measures that are obtained when the sample contains missing data across the whole matrix are about equivalent or more distant than those obtained when only the first column contains missing data. The results are more intuitive here, except for the MICE algorithm. The results presented in this chapter also revealed that MICE is highly variable from sample to sample.

Thus, the proximity measures reveal that MICE, Amelia, IPCA and MIPCA can be further away from the original series when the missing data are concentrated in the first column, compared to when they are spread over the whole sample. On the other hand,

this deterioration in the proximity measures is not associated with a deterioration in the risk measures. The risk measures obtained when the missing data are in a single column are, in fact, closer to those of the original series than those obtained when the missing data are spread over the whole sample.

A possible explanation would be that the proximity measure were reduced by a degradation of the estimated distribution of imputed data. Previously, the RMSE An underestimation of the variance can reduce artificially proximity measures. It is, indeed, possible to write the RMSE was written as a function of the variance of the original data and the variance of the imputed data (see Equation 3.7-2). Then, if completion methods underestimate the variance, then it may reduce the proximity measures and especially the RMSE. In order to observe if the distribution of imputed data are close to that of original data, a Wasserstein distance is used. Its purpose is to evaluate the distance between two distributions in order to see if degradation is observable in practice. The Wasserstein distance l_2 between two distributions $\mathcal{P}_1$ and $\mathcal{P}_2$ is written as follows:

$$\mathcal{W}_2(\mathcal{P}_1, \mathcal{P}_2) = \left(\int_0^1 \left(F_1^{-1}(t) - F_2^{-1}(t) \right)^2 dt \right)^{1/2}, \tag{3.7-3}$$

where F_1^{-1} and F_2^{-1} are the inverse cumulative distribution function $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively.

The Wasserstein distances of the algorithms for the simulated series are presented in Table 3.7-6.

Tab. 3.7-6: Wasserstein distance (10^{-6}) between the first column of the original simulated sample and that of the imputed data for 10%, 30%, 50% and 70% of missing data in the first columns versus missing data across the whole matrix

	Missingness proportion in the first column*				**Missingness proportion in all columns****			
	10%	**30%**	**50%**	**70%**	**10%**	**30%**	**50%**	**70%**
IPCA	1.25	3.77	20.77	70.81	0.30	2.63	8.92	175.62
MICE	6.60	42	138.66	312.91	0.78	14.78	81.35	261.71
MIPCA	1.31	3.53	8.79	53.61	0.28	2.70	14.26	-
Amelia	1.39	4.10	7.93	114.04	0.28	2.80	-	-

* Results from Section 3.2.1
** Results from Section 3.2.2

It appears that the algorithms estimate the distributions better when the missing data are spread over the whole data matrix. The trend that was observed for the proximity measures is also observed here: the algorithms perform better with fewer observations.

The exercise was repeated with the first historical sample that was studied on the basis of a heuristic approach (Section 3.5.2). Missing data were removed in a completely random way from all columns of the historical sample except the last. In other words, the missingness mechanism from Section 3.2.2 was followed. Then, the IPCA, MIPCA, MICE and Amelia algorithms were applied to these data to ensure that the tendency described in the preceding paragraphs is not attributable to the sample. The average proximity measures for the methods are presented in Table 3.7-7. The left-hand side of the table presents the proximity measures when the missing data are in a single column; the right-hand side presents the proximity measures when the missing data are spread across the whole matrix.

Tab. 3.7-7: Average MAE and RMSE for 10%, 30%, 50% and 70% missing data in the first columns versus missing data across the whole matrix of the first historical data sample (based on a heuristic approach)

		Missingness proportion in the first column*				**Missingness proportion in all columns**			
		10%	**30%**	**50%**	**70%**	**10%**	**30%**	**50%**	**70%**
MAE (10^{-2})	**IPCA**	0.86	0.87	0.98	1.61	0.20	0.48	0.88	1.41
	MICE	1.22	1.25	1.45	1.99	0.23	0.75	1.45	2.33
	MIPCA	1.51	1.54	1.62	1.90	0.33	0.87	1.44	1.91
	Amelia	1.18	1.21	1.28	2.18	0.24	0.71	-	-
RMSE (10^{-2})	**IPCA**	1.21	1.24	1.43	2.50	0.51	0.94	1.41	2.23
	MICE	1.79	1.87	2.24	3.15	0.60	1.48	2.37	3.61
	MIPCA	1.93	2.00	2.15	2.76	0.83	1.53	2.17	2.78
	Amelia	1.55	1.62	1.78	4.29	0.62	1.28	-	-

* Results from Section 3.5.2

These results are in agreement with those observed in the case of simulated data (presented in Table 3.7-3). Therefore, this tendency is not specific to the simulated sample: when the methods in question are used, the imputation of the first column of the historical matrix is typically closer to the original when the other columns contain missing data than when they are complete. The implication is that it is preferable to impute a series from other incomplete series when these methods are used, a paradox. An analyst might be tempted to delete data from the other columns of the matrix (following the pattern of the column that they wish to complete) in order to improve imputation quality.

The proximity measures of the historical sample (based on a heuristic approach) are also affected by this paradox. It has no impact on the risk measures presented below (see Table 3.7-8).

Tab. 3.7-8: Average 1-day risk measures for 10%, 30%, 50% and 70% of missing data in the first columns versus missing data across the whole matrix of the first historical data sample (based on a heuristic approach)

		Missingness proportion in the first column*				**Missingness proportion in all columns****			
		10%	**30%**	**50%**	**70%**	**10%**	**30%**	**50%**	**70%**
$VaR_{99\%}^{1-day}$ (10^{-2})	**Original**	-0.11				-0.11			
	IPCA	-0.11	-0.11	-0.11	-0.10	-0.11	-0.11	-0.11	-0.11
	MICE	-0.11	-0.10	-0.10	-0.10	-0.11	-0.11	-0.10	-0.13
	MIPCA	-0.11	-0.11	-0.11	-0.10	-0.11	-0.11	-0.12	-0.13
	Amelia	-0.12	-0.12	-0.12	-0.14	-0.11	-0.11	-	-
$ES_{97.5\%}^{1-day}$ (10^{-2})	**Original**	-0.11				-0.11			
	IPCA	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.12
	MICE	-0.11	-0.11	-0.10	-0.12	-0.11	-0.11	-0.10	-0.12
	MIPCA	-0.11	-0.11	-0.11	-0.11	-0.11	-0.11	-0.12	-0.13
	Amelia	-0.11	-0.11	-0.11	-0.16	-0.11	-0.11	-	-

* Results from Section 3.5.2

The risk measures for both missing data schemes are comparable, as was the case for the simulated sample. Once more, it is not possible to explain the results by a degradation of the law because the Wasserstein distances (presented in Table 3.7-9) are lower when the data are missing in all columns.

Tab. 3.7-9: Wasserstein distance (10^{-6}) between the first column of the original historical sample (based on a heuristic approach) and that of imputed data for 10%, 30%, 50% and 70% missing data in the first columns versus missing data across the whole matrix

	Missingness proportion in the first column*				**Missingness proportion in all columns**			
	10%	**30%**	**50%**	**70%**	**10%**	**30%**	**50%**	**70%**
IPCA	4.93	13.64	29.24	304.94	1.22	7.09	26.63	181.21
MICE	7.53	24.44	57.13	198.41	1.42	12.14	60.54	245.44
MIPCA	10.88	39.09	59.56	169.64	1.97	19.04	92.37	247.59
Amelia	7.42	23.23	46.98	510.69	1.46	11.36	-	-

* Results from Section 3.5.2

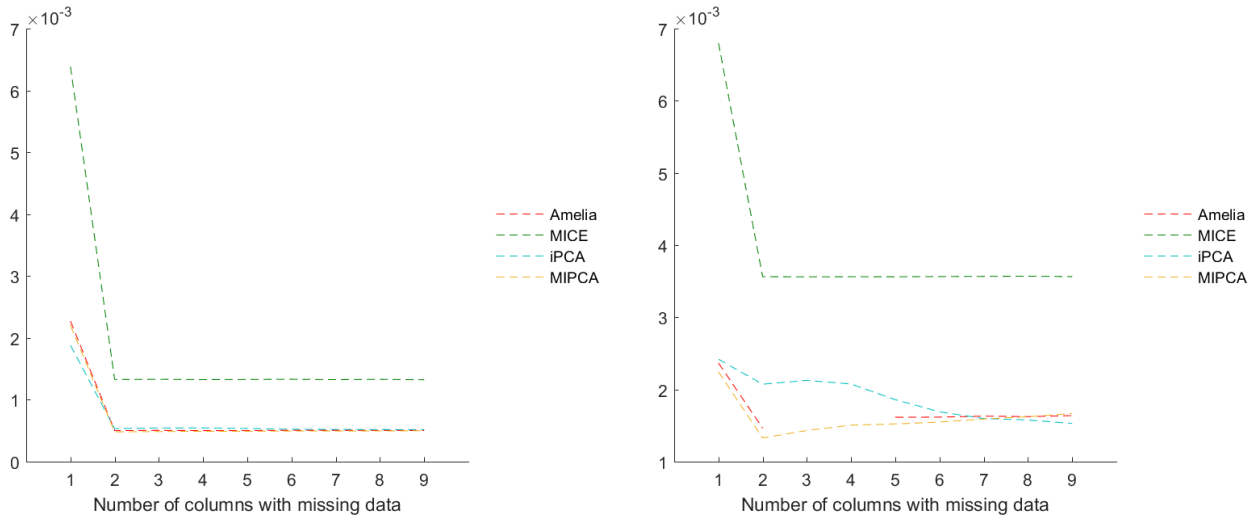
The comparison is made between the results from a matrix containing missing data on the first column and those from a matrix containing missing data on almost all its columns (9 columns out of 10). These two scenarios represent two extremes. However,

it is interesting to analyze the evolution of the criteria (obtained from the first column) according to an increasing number of columns containing missing data. If the number of columns with missing data is really the cause of an improvement in imputation quality, then a increasing monotone function is expected.

For this purpose, data were deleted and then imputed for a number of columns ranging from 1 to 9. Figure 3.7-4 represents the evolution of the average MAE (y-axis) as a function of the number of columns (x-axis) with 10%, 30%, 50% and 70% of missing data, respectively. The average MAEs are based on 100 missingness scenarios with MCAR data. The results associated with 1 and 9 columns are the same as those discussed earlier, computed for Section 3.2.1 and Section 3.2.2, respectively.

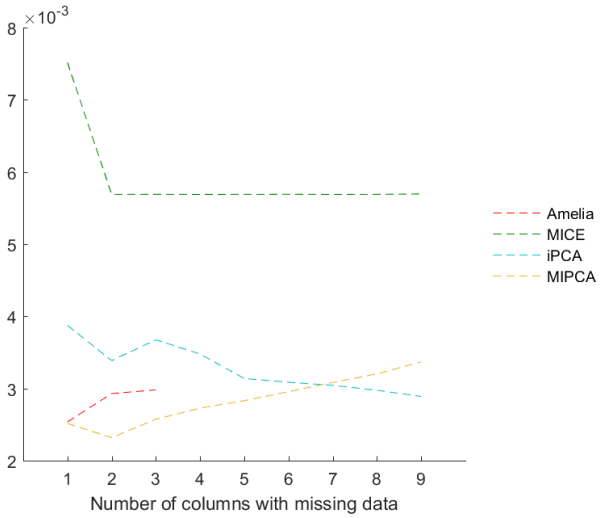
As observed in Section 3.2.2, the Amelia algorithm is not computable when almost all columns contain 70% missing data, but this is also the case as soon as 2 columns (see Figure 3.7-4d). When this proportion is 50%, the algorithm manages 3 columns at most (see Figure 3.7-4c). Moreover, some scenarios containing 30% missing data in 3 and 4 columns were not computable by Amelia, hence the absence of results (see Figure 3.7-4b). Moreover, the MIPCA algorithm is not always computable, notably when more than 7 columns of the matrix contain 70% missing data.

Fig. 3.7-4: Average MAE (of the first column) obtained according to the number of columns containing 10%, 30%, 50% and 70% of missing data on simulated sample

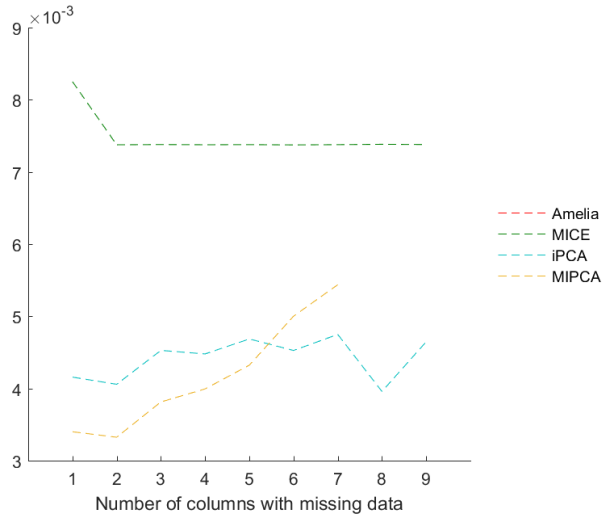


(a) 10% missing data

(b) 30% missing data



(c) 50% missing data



(d) 70% missing data

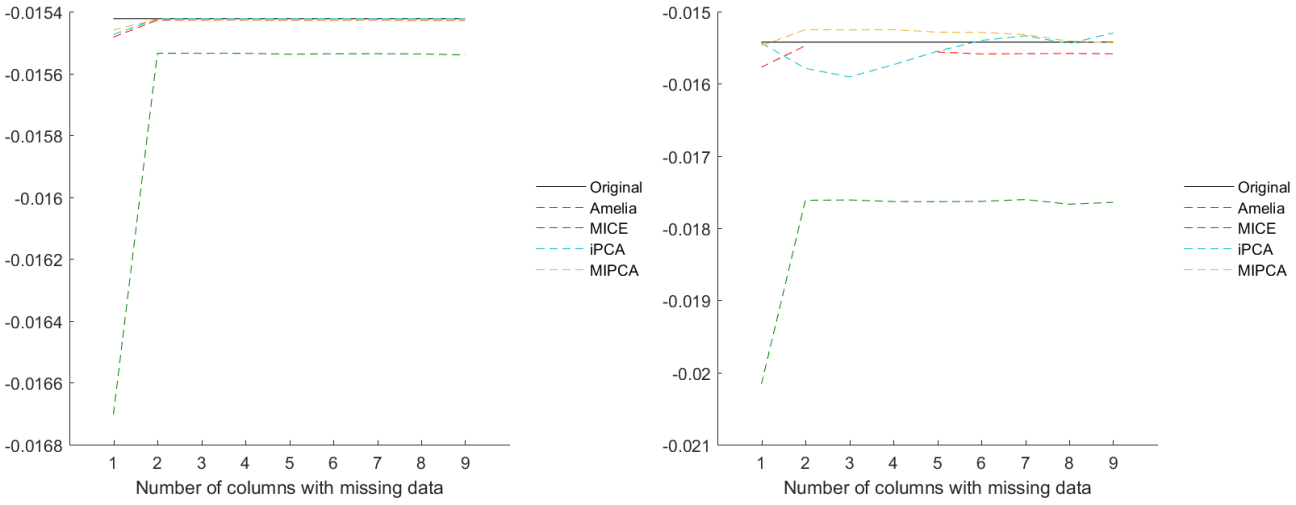
When the proportion of missing data is low, a common trend shared by all algorithms is observable (see Figure 3.7-4a). The average MAEs obtained when the missing data are only in 1 column are significantly higher than those obtained when they are in 2 or more columns. On the other hand, this increase in imputation quality (decrease of average MAE) is only observable between 1 and 2 columns. The results obtained for 2 columns are comparable to those obtained for more than 2 columns.

This dynamic persists for all proportions of missing data for the MICE algorithm, even though it is less important. However, this is not the case of other algorithms. If the MIPCA algorithm seems to see its imputation quality improved between 1 to 2 columns, it also seems to degrade with more than 2 columns. The evolution of the average MAE is first decreasing then increasing, and this phenomenon is accentuated with the proportion of missing data. The evolution of the IPCA's average MAE is not monotonous either. The algorithm sees its imputation quality successively improving and deteriorating with an additional column containing missing data, when the proportion of missing data is 70%. Finally, since the Amelia algorithm is not always computable, its analysis is not obvious. Nevertheless, it is possible to notice that the trend observed, between 1 and 2 columns, for 10% missing data (an increase of the imputation quality) is opposite to the one observed for 50% missing data (a decrease of the imputation quality).

The same results have been computed from the RMSE (see Appendix M.1), and the same observations can be made: an improvement of the imputation quality is observable when the number of columns containing missing data increases from 1 to 2, but not beyond; this improvement loses in amplitude (and becomes insignificant for some algorithms) as the proportion of missingness increases.

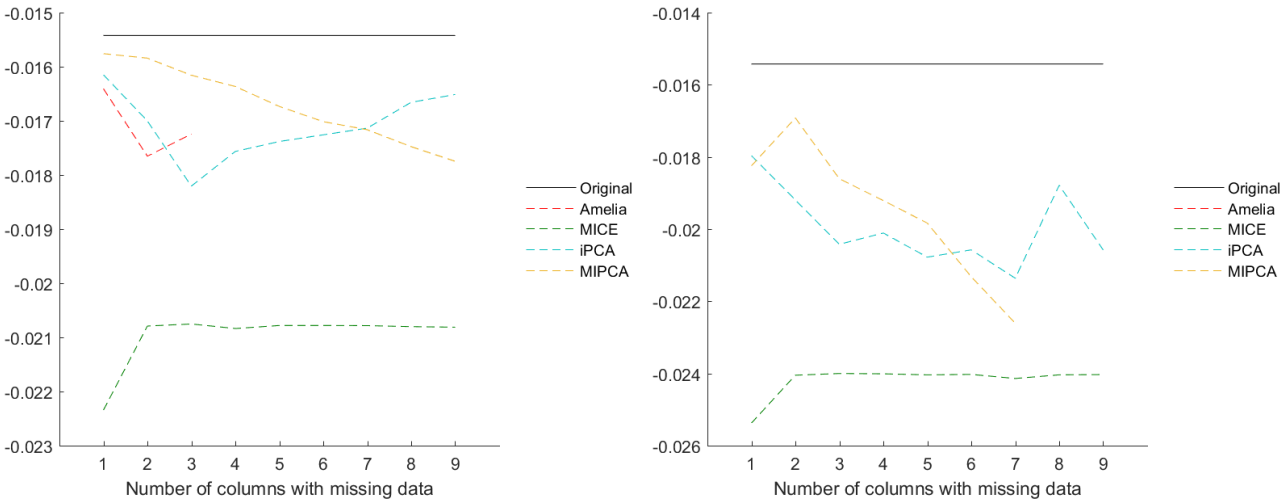
For the risk measures, in particular the 1-day VaR (see Figure 3.7-5), this improvement in imputation quality is particularly observable for the MICE algorithm. This method obtains a VaR closer to the original one with 2 (or more) columns than 1, and this is true for all proportions of missing data. On the other hand, it is not possible to draw any conclusion for the other methods. These results are comparable to those obtained for the 1-day ES (available in Appendix M.2).

Fig. 3.7-5: Average 1-day VaR (of the first column) obtained according to the number of columns containing 10%, 30%, 50% and 70% of missing data on simulated sample



(a) 10% missing data

(b) 30% missing data



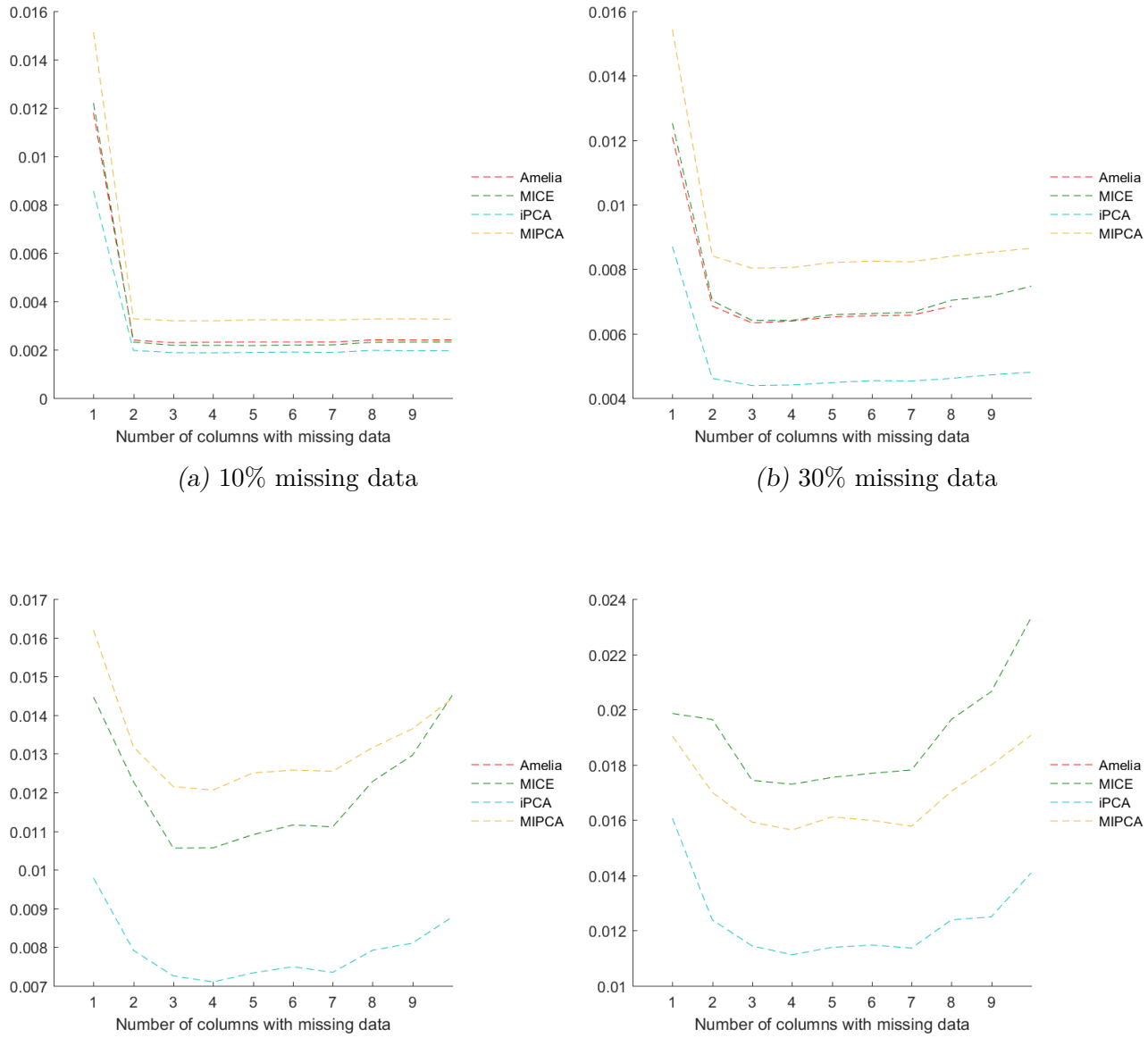
(c) 50% missing data

(d) 70% missing data

These results from proximity and risk measures show that the improvement of imputation quality is observed when going from 1 incomplete column to 2 (or more). On the other hand, this is not observed when going from 2 to 3 columns, and so on. Moreover, this phenomenon is particularly visible for the MICE algorithm. However, this chapter revealed that the results of this imputation method were not representative on this specific sample. This algorithm performed very poorly on this simulated data compared to the others. That is why no conclusion should be drawn based on this algorithm. If only the three other methods are considered (Amelia, IPCA and MIPCA), the improvement in imputation quality is visible when the proportion of missing data is low, but does not persist when this proportion increases. The proximity and risk measures increase and decrease depending on the number of columns with missing data, making it impossible to draw any conclusions.

The same exercise was done on the historical sample based on a heuristic approach. Missing data were removed in 1 column, then 2 columns, and so on up to 10 columns (the historical sample consists of 11 columns). The results obtained in terms of MAE are presented in Figure 3.7-6, and those of RMSE are presented in Appendix M.3.

Fig. 3.7-6: Average MAE (of the first column) obtained according to the number of columns containing 10%, 30%, 50% and 70% of missing data on the heuristic historical sample (based on a heuristic approach)



(a) 10% missing data

(b) 30% missing data

(c) 50% missing data

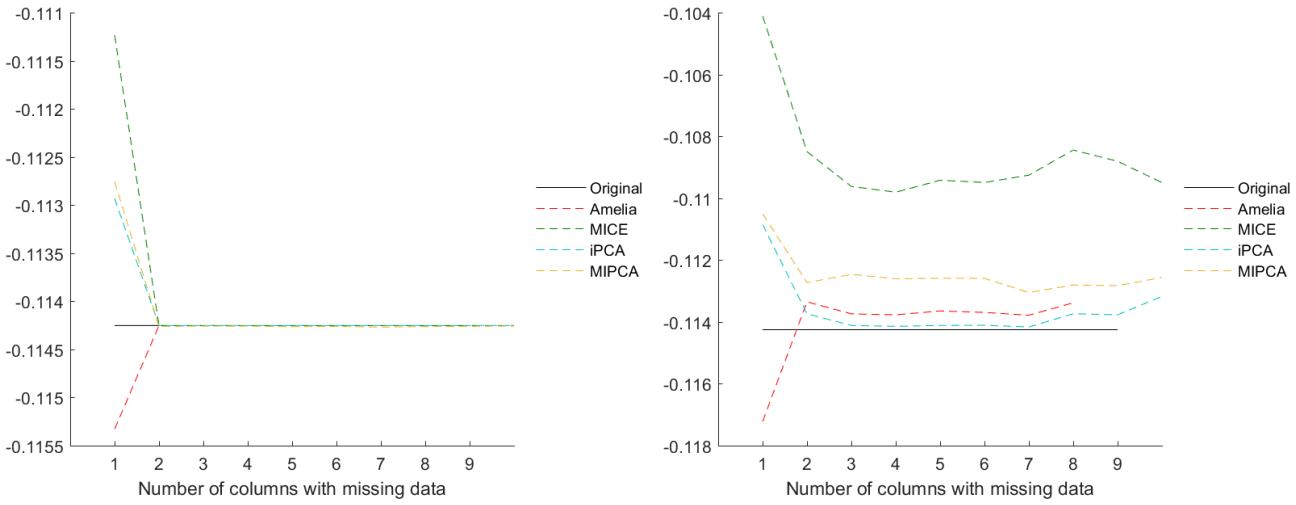
(d) 70% missing data

For 10% missing data, the results obtained on the historical sample have the same trend as those obtained for the simulated sample. On the other hand, for a higher proportion of missing data, the evolution of the proximity measures are not the same. The proximity measures tend to decrease, and therefore the imputation quality improves,

when the missing data are from 1 to 4 columns. On the other hand, beyond 4 columns, the proximity measures stagnate or increase. As was observable on the Table 3.7-7, Amelia is not computable for 50% or more missing data.

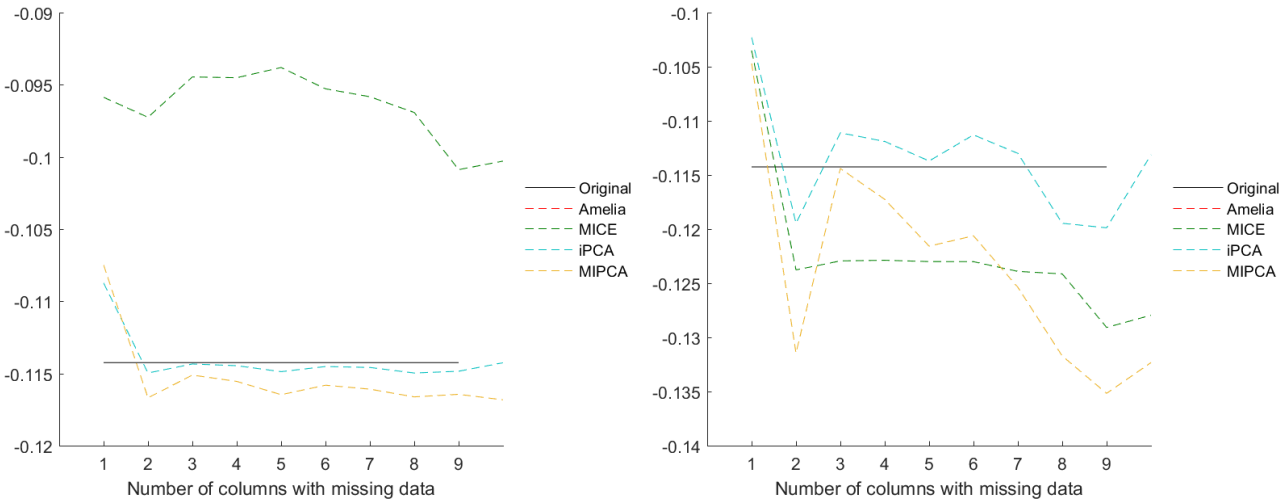
In terms of risk measures, the VaR results obtained in terms of VaR are presented in Figure 3.7-7 and those in terms of ES are presented in Appendix M.4.

Fig. 3.7-7: Average 1-day VaR (of the first column) obtained according to the number of columns containing 10%, 30%, 50% and 70% of missing data on the heuristic historical sample



(a) 10% missing data

(b) 30% missing data



(c) 50% missing data

(d) 70% missing data

When 2 or more columns contain 10% missing data, then all methods manage to reproduce the VaR of the original series. But this is not true when the missing data are only in one column. Moreover, there is a real difference between 1 and 2 (or more) columns. On the other hand, when the proportion of missing data is higher, the VaR levels vary in a non-monotonic way. Nevertheless, there is a variation (more or less important) between 1 and 2 columns containing missing data.

Overall, the results observed on the simulated sample and the historical sample based on a heuristic approach show that the imputation quality improves when 2 columns contain (each) 10% of missing data compared to a single column. However, this is not necessarily true for larger proportions of missing data. Further studies on a large number of samples are therefore necessary to generalized this phenomenon.

Moreover, Amelia, MICE, IPCA and MIPCA are iterative methods, and it is clear that the number of iterations that is needed to achieve convergence is more important when the missing data are in several columns. Therefore, the improvement in imputation quality may be due to the number of iterations needed for the convergence of imputations. Since the number of iterations is different, the path to convergence is not the same. For reasons that remain unclear, this leads to more satisfactory results when all columns are biased by the presence of missing data.

These results could also be explained by steps that are specific to each algorithm. As seen previously, the initialization of the parameters for the EM algorithm of Amelia impacts the imputation quality. The initialization of parameters based on several incomplete columns could lead to a global maximum, while the initialization based on a single incomplete column could lead to a local maximum. In the case of IPCA and

MIPCA, the number of principal components used could be different, but after verification for 10% missing data, these methods use the same number of principal components irrespective of whether the missing data are in one column or in all columns. Furthermore, these algorithms were executed step by step in order to observe in detail each of their operations. However, no additional steps were observed when the missing data are in several columns compared to a single column. The improvement of imputation quality seems to be due to the different path of convergence based on a different number of iteration.

This paradox is observed in only two samples (the simulated sample and the one based on heuristic approach). As noted on numerous occasions, it is necessary to repeat the experiment on more samples in order to draw more general conclusions.

It emerges that the results are difficult to explain, emphasizing the importance of documentation for completion methods. One may hold well-grounded expectations about the behavior of an algorithm in a certain situation. Those expectations must be tested to ensure transparency and to anticipate all possible scenarios. It is also important for banks to implement or reimplement completion methods internally in order to avoid becoming dependent on algorithms and to guarantee that they are able to react when corrections become necessary.

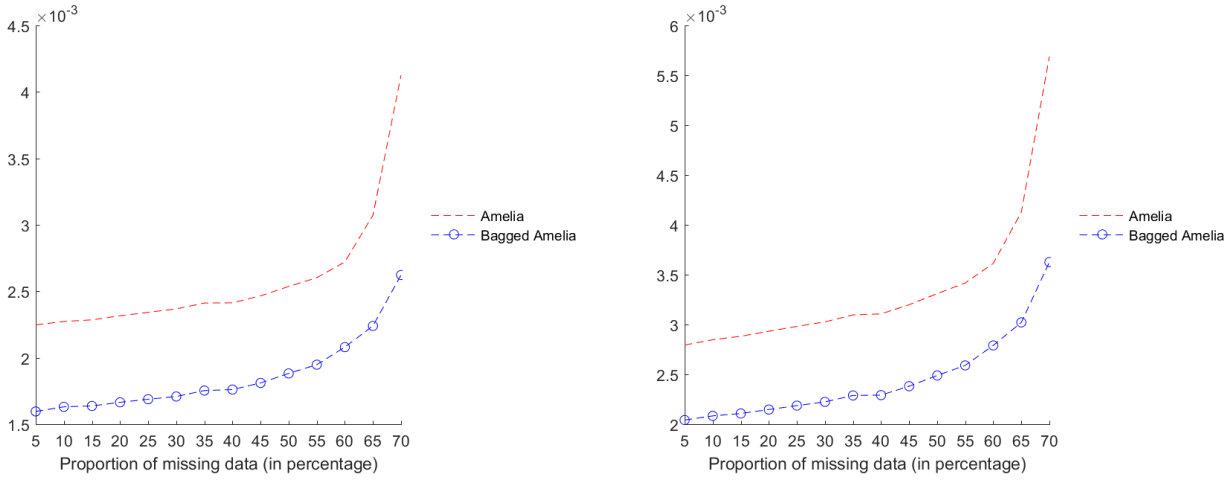
3.7.8 Amelia versus random forests

In this chapter, one of the best-performing algorithms is random forests, especially for low missingness proportions. It outperforms Amelia on Gaussian data. On first impression, this is surprising. What causes the random forests algorithm to do better than Amelia? How can Amelia be improved? The two methods operate very differently: one seeks to estimate the law of the missing data, while the other seeks to ascertain the value of the missing data directly. This difference in procedure undoubtedly affects imputation quality, but its impact remains difficult to quantify. Other avenues of exploration ought to be pursued.

The first approach focuses on the bagging procedure. Bagging consists of averaging the results from multiple imputation before calculating an analysis criterion. Breiman [41] uses it and recommends it for decision trees in order to stabilize and improve results. It could be that bagging also improves the imputation quality of other multiple imputation methods, such as Amelia. The benefits of bagging should be observable if the average MAE of each imputed sample is compared to the MAE of the average of the imputed samples. A similar method may be followed for the RMSE. Drawing on the results that were presented, especially those from the simulated data in Section 3.2.1, Figure 3.7-8 represents the benefits of bagging for Amelia. For a given proportion of missing data, the results that are designated as “Amelia” correspond to an average of

the MAEs (and RMSEs) calculated from each of the imputed samples for each of the 100 missingness scenarios (Section 3.2.1). The results designated as "bagged Amelia" correspond to the average of the MAEs (and RMSEs) calculated from the average of the 100 imputed samples for each of the 100 missingness scenarios. In the first case, 100 MAEs (and RMSEs) are calculated for each missingness scenario. In the second case, only one MAE (and RMSE) is necessary.

Fig. 3.7-8: Comparison between multiple imputation results and bagged results, for Amelia applied to a simulated sample containing MCAR data in the first column*

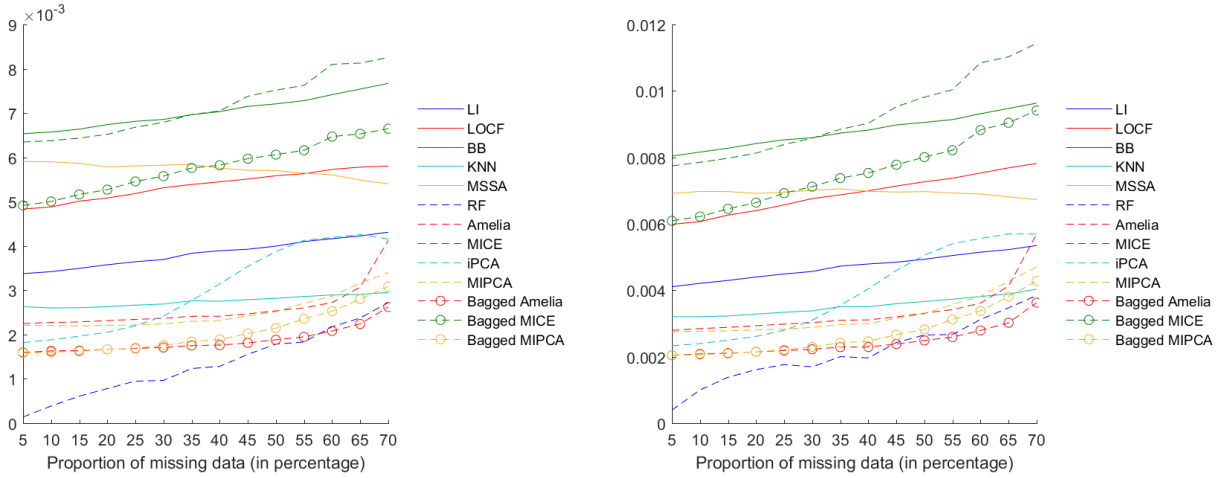


(a) MAE *(b)* RMSE

* Results for Amelia from Section 3.2.1

Bagging tends to reduce proximity measures considerably. Moreover, it allows results to be stabilized when the proportion of missing data is too significant. This may explain, at least partially, the exceptional performances of random forests. It is also possible that bagging improves other multiple imputation methods, such as MICE and MIPCA. Therefore, in order to understand whether introducing bagging to Amelia would be sufficient to cause it to perform as well as random forests, these results (as well as those for the "bagged MICE" and "bagged MIPCA", following the same procedure as for "bagged Amelia") are reported alongside those of the other methods presented in Section 3.2.1. Thus, Figure 3.7-9 includes the results presented in Section 3.2.1 and adds those of Amelia, MICE and MIPCA with bagging.

Fig. 3.7-9: Average MAE and RMSE from simulated sample containing MCAR data in the first column* according to missingness probability for all methods, including Amelia, MICE and MIPCA with bagging



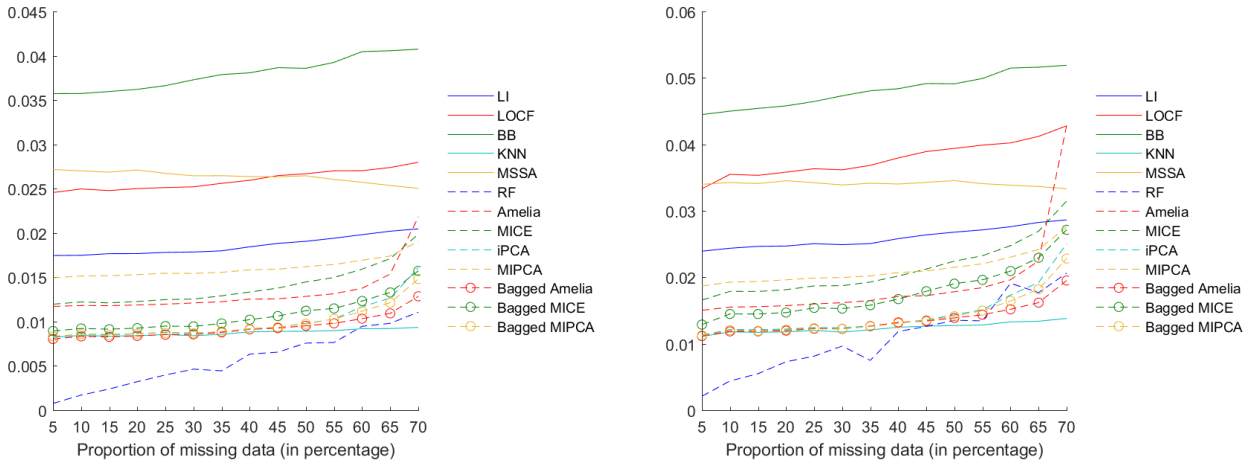
(a) MAE *(b)* RMSE

* Results from Section 3.2.1, except those for bagged Amelia, bagged MICE and bagged MIPCA

Overall, the additional bagged methods improve the results of the original methods. Like Amelia, the "bagged MICE" and "bagged MIPCA" enable a reduction in the level of the proximity measures. However, integrating bagging into Amelia is not sufficient for it to match the performance of random forests, especially when the missingness proportions are low.

The same exercise was repeated with the first historical sample (based on a heuristic approach). Table 3.7-10 presents the results of the application of all the methods, including those with bagging, to the data from the sample used in Section 3.5.2.

Fig. 3.7-10: Average MAE and RMSE from historical sample (based on heuristic approach) containing MCAR data in the first column* according to missingness probability, for all methods, including Amelia, MICE and MIPCA with bagging

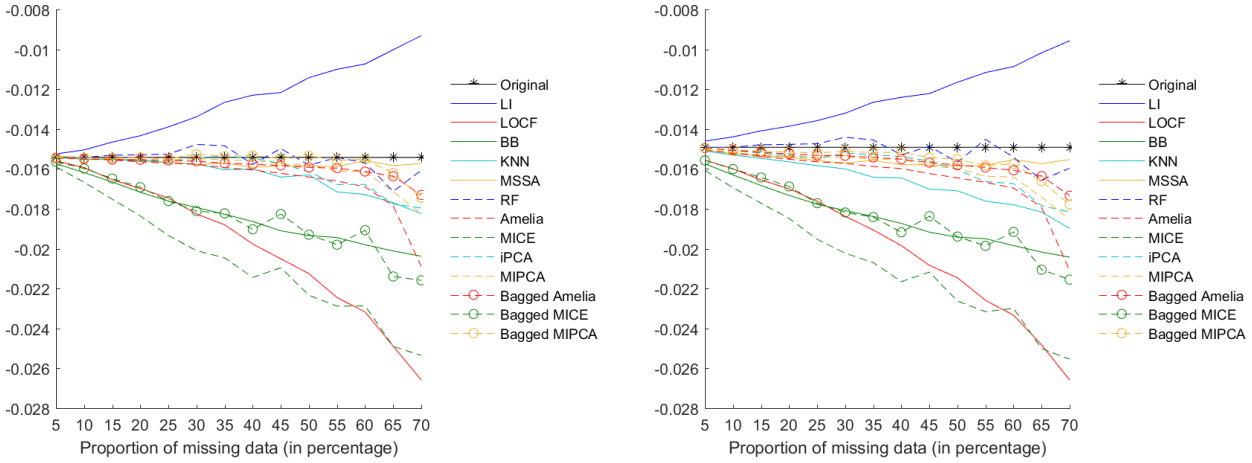


(a) MAE *(b)* RMSE

* Results from Section 3.5.2, except those for bagged Amelia, bagged MICE and bagged MIPCA

The conclusion is the same: bagging improves the results of multiple imputation methods, but it does not cause them to perform as well as random forests. Bagging also impacts risk measures. Figure 3.7-11 highlights the behavior of bagged methods with 1-day risk measures.

Fig. 3.7-11: Average 1-day risk measures from simulated sample containing MCAR data in the first column* according to missingness probability for all methods, including Amelia, MICE and MIPCA with bagging



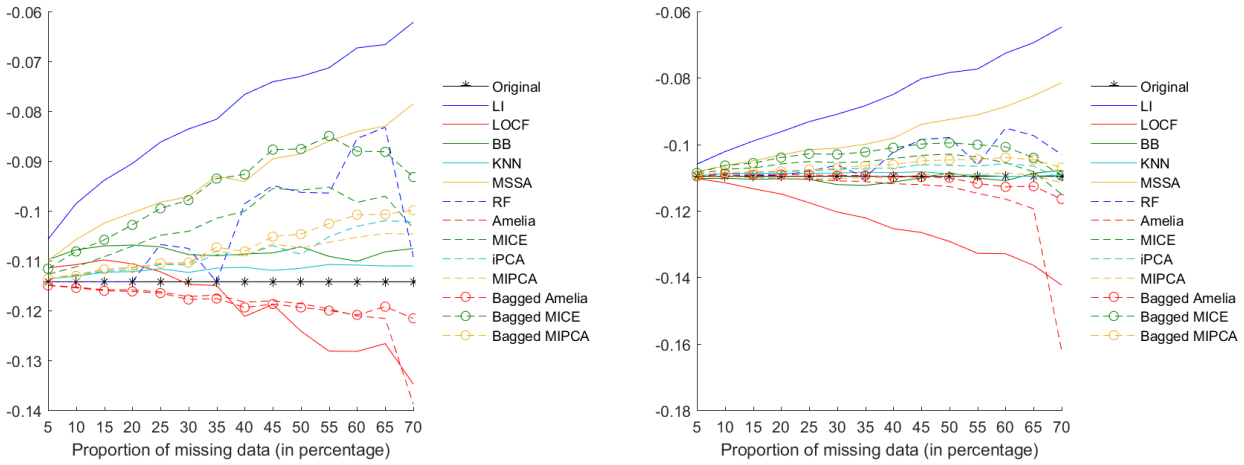
(a) 1-day VaR at 99%

(b) 1-day ES at 97.5%

* Results from Section 3.2.1, except those for bagged Amelia, bagged MICE and bagged MIPCA

It is clear that bagging can reduce the overestimation of risk considerably, especially for the most significant missing data proportions. The same observation can be made about the risk measures obtained from the historical sample on the basis of a heuristic approach (see Figure 3.7-12), but only for the Amelia algorithm.

Fig. 3.7-12: Average 1-day risk measures from historical sample (based on a heuristic approach) that contains MCAR data in the first column* according to missingness probability for all methods, including Amelia, MICE and MIPCA with bagging



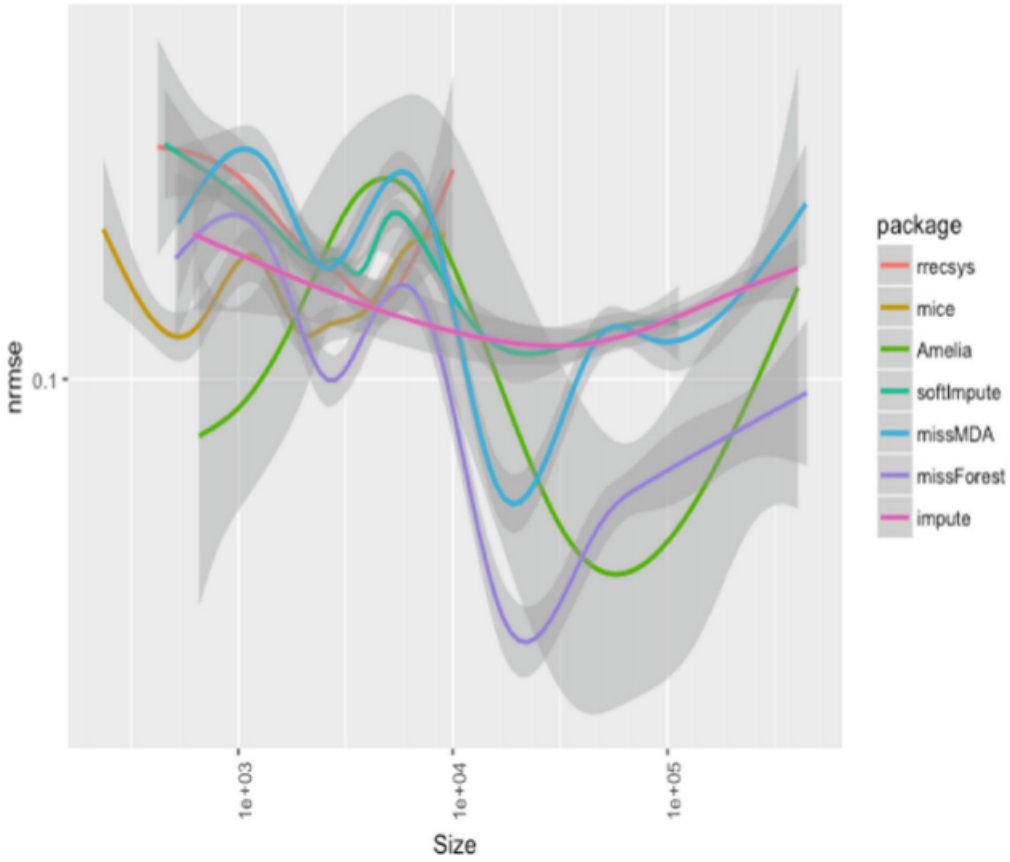
(a) 1-day VaR at 99% *(b)* 1-day ES at 97.5%

* Results from Section 3.5.2, except those of bagged Amelia, bagged MICE and bagged MIPCA

The risk measures obtained by MICE and MIPCA tend to underestimate risk: they produce more optimistic risk measures than the original series. Bagging aggravates this problem. For the historical sample, the results with bagging are closer to the original series in terms of MAE and RMSE. However, bagging can lead to risk measures that are increasingly distant from the original, as is the case for MICE and MIPCA. It follows that bagging explains the good performance of the random forests only partially. It would be logical to expect Amelia to be particularly efficient when applied to simulated data, given its Gaussian framework and Amelia's use of the EM algorithm. In actuality, random forests outperforms Amelia. The reasons are unclear.

Yet another approach to the problem circles on sample size. In 2019, Modarresi and Diner [156] showed that the performances of imputation methods varies with sample size. They worked on 100 bootstrapped samples from several research domains. They selected a random number of rows, a random number of columns and a random number of missing values. Then, they imputed these missing data by different methods, including random forests (to be precise, they used *missForest*, which was also used in this chapter) and Amelia and compared their performance in terms of normalized RMSE (NRMSE). Their results are presented in Figure 3.7-13.

Fig. 3.7-13: Impact of sample size on imputation quality (Source: Modarresi and Diner,2019 [156]



Amelia minimizes the normalized RMSE for a sample size at 1,000, while it is random forests for a sample size at 10,000.

Evidently, random forests (*missForest*) and Amelia are the two methods that allow the NRMSE to be minimized whatever the size of the sample. This said, performance does depend on sample size. These results may explain the differences in performance, especially since the number of samples used in this chapter is not large. The results also cohere with the argument about sample-dependent results that was developed on the preceding pages. The random forests algorithm may be particularly effective when applied to these samples or to a sample that covers daily data over a year, but it may run into difficulties with other samples or sample sizes. Future work should be directed at the impact of sample size on completion methods as a potential explanation of the superior performances of random forests.

The last approach involves analyzing preprocessing methods. Amelia and random forests entail different steps. Their preprocessing procedures are different, too. As

a general matter, the term "preprocessing" covers all the microsteps that are taken to prepare the data before the main method of an algorithm is applied. For Amelia, that main method is the EM algorithm. For random forests, it is the application of decision trees. Preprocessing aims to optimize performance. In the Amelia algorithm, the data are normalized. In random forests, the data is preimputed naively through mean imputation before the trees are grown.

These preprocessing steps undoubtedly affect imputation quality. A priori, the normalization of the data should facilitate convergence to the correct law in the EM algorithm. However, the preimputation in random forests may not be optimal. The the implementations of Breiman [43] (*rfimpute* function) and Stekhoven and Bühlmann [194] (*missForest* function) differ on this point. Breiman's [43] version uses the median of the observed data, while Stekhoven and Bühlmann [194] use the mean. It was noted early that in the preparatory stage of this research, a homemade version of Breiman's [43] algorithm yielded worse results than Stekhoven and Bühlmann's version [194]. Of course, the processes of the two algorithms differ substantially. Nevertheless, the preprocessing step may affect imputation quality. Unfortunately, Stekhoven and Bühlmann [194] did not enable users to modify the preprocessing steps, and preimputation in particular, in their R function. Therefore, their algorithm should be reimplemented to gauge the impact of preprocessing on imputation quality.

In the final section of this chapter, recurrent problems in data completion were revisited to highlight the potential benefits and the risks that attach to the use of completion methods. Many avenues for further research remain to be explored. They will be treated in future work.

This chapter presented a practical approach to the theory presented in the previous ones. Several completion methods were applied to different historical and simulated samples. Different missing data mechanisms that can occur in financial data were examined. As noted, MCAR, MAR and MNAR data are liable to appear in financial time series. Historical samples were used to verify the results from the simulated data. Amelia and random forests were usually the most efficient completion methods. They were often followed by MIPCA, IPCA and K-NN. The MICE algorithm obtained satisfactory results with the historical samples but not with the simulated ones. Further investigation is necessary to draw additional conclusions. The Brownian bridge and MSSA were the worst performers. They were frequently less relevant than the usual methods.

Some methods were efficient when measured against some criteria but inefficient when measured against others. The implication is that completion methods must be chosen to reflect the intended use of the imputed data. The results also highlighted the operational risk that inheres in the use of completion methods, which may be incalculable or intransparent. The latter point reflects the proposition that the results of using a method mean nothing if the procedure is not recorded. Every step of an algorithm, including preprocessing, affects imputation quality.

Conclusion

This PhD thesis explored the impact of missing data in financial regulation and the influence that completion methods exercise on risk assessment. Missing data affect all research fields that draw on historical records. The subject has been studied in fields as varied as medicine, oceanology, image analysis and economics. The richness of the literature made it possible to adumbrate the impact of missing data on statistical analyses. Thereafter, the exposition turned to financial series. Although missing data have always affected finance, their importance has increased in recent years, particularly in the banking sector. Since the 2007 crisis, many new regulations have emerged across the world. Their requirements and their principles routinely address data quality and missing data. Some regulations mention data quality in passing; others are dedicated exclusively to that problem. Regulators are alive to the importance of data as well as to the considerable difficulties that banks may encounter in managing it. The implementation of regulations that target data quality, be it directly or indirectly, is regularly postponed due to the difficulty of ensuring compliance. Surveys, too, make it clear that missing data issues are pervasive in banking.

After the overview of the background of the research, Little and Rubin's [145] missing data framework was presented and connected to the financial data environment. Numerous cases were discussed in an attempt to explain the presence of missing data in financial series. Confronted with missing data and ever-intensifying regulation, banks have become interested in imputation and thus in completion methods. It was important to mention the data encoding stage, in which a sample is chosen, before presenting and comparing methods. Empirical studies show that data encoding is essential to algorithm performance. A description of the relevant completion methods and algorithms followed. The empirical analyses and their results were presented in the last chapter.

Banks must select the most suitable completion methods. Many empirical analyses have been performed. The methods that this PhD thesis considered were first applied to simulated data so that the behavior and the performance of the methods could be analyzed by reference to assorted missing data mechanisms (MCAR, MAR and MNAR) and the characteristics of series, in particular heteroskedasticity and the jumps that typify financial data. The exercise was then repeated with two samples of European historical data using an MCAR data mechanism.

The results show that conventional methods are far from being the most satisfactory. Although banks use these methods frequently, they can produce very poor estimates of risk. It is clear that other methods can meet the requirements of banks and regulators better. The LOCF method is the most conservative in risk-measure terms, and it could

replace the standard approach. Its use ensures that risk is overestimated, which is likely to be appreciated by regulators.

They also revealed that the choice of completion method should be driven by the intended use of the completed data. That a method performs well according to one criterion does not ensure that it will perform well when measured against others. Completion methods, which can be complex, act on many variables simultaneously. Consequently, the interpretation of their results can pose difficulties.

It also emerged that the use of completion methods implies operational risks at several levels. Despite precautions, some methods were incalculable. That some completion methods still yield missing data is a matter of serious concern. The results for some methods are disappointing. The results of others are hard to decipher, highlighting the importance of documentation. It must be possible to grasp the steps of an algorithm, the choice and the calibration of its parameters and to ensure that it can be reimplemented by regulators and by those who wish to improve it. The accompanying documentation must be clear and transparent. Its preparation is demanding. It is even more difficult to anticipate all possible scenarios, especially ones that have never occurred.

The comparative analysis demonstrated that some methods perform particularly well. The most efficient algorithms in this study are Amelia, implemented by Honaker, King and Scheve [112] and random forests, implemented by Stekhoven and Bühlmann [194]. They share few commonalities. Amelia estimates the distribution of missing data; random forests estimates missing values. Their performances, sometimes far apart, were not always clear because the comparison involved not only two methods but also two algorithms and the steps that they use, including preprocessing. Random forests benefits from bagging. However, bagging does not explain the performance of random forests completely. The performance of the two algorithms was triturated at length in order to arrive at answers.

Many avenues for improvement and further research were identified throughout the substantive exposition. The results that were presented are based on a single sample of simulated data. The results for another sample may be different. In the case of the MICE algorithm, the difference between simulated and historical samples is very pronounced. The procedure presented here allows the effects of using different completion methods on a specific sample (which could be considered historical) to be highlighted. The analysis of the historical samples confirms some of the results. This said, it would be appropriate to repeat the analysis with numerous simulated samples in order to eliminate sampling effects.

It is not easy to interpret the performance of some algorithms because they comprise multiple successive steps that can affect imputation quality. These black-box algorithms can pose an operational risk for banks. Regulators may refuse to validate

them. Understanding and explaining their behavior is a common problem in artificial intelligence. It has led to the emergence of reverse engineering as a candidate solution. In 2012, Wiesinger, Sornette and Satinover [209] published a study that models financial markets using a reverse engineering method that is based on genetic algorithms. In 2019, Herzog and Osamah [108] employed artificial intelligence-based reverse engineering to study option pricing. Reverse engineering can also be applied to unclear imputation methods. Its use may allow methods to become completely transparent and to be validated by regulators.

Finally, if the completion algorithms are used here directly on the series in order to complete the databases to then calculate their P&L or their risk measures, it is also possible to consider other applications for these completion methods. The most obvious one, already mentioned in this PhD thesis, consists in using a powerful completion method for data forecasting. If tomorrow's data is given to the algorithms as missing data, then some of them will be able to give a reliable forecast.

Finally, in this PhD thesis, completion methods were applied directly to series in order to complete databases and to calculate P&L and risk measures. The methods also have other applications. The most obvious one is data forecasting. Some algorithms may be able to treat future data as missing data and produce reliable forecasts. Completion methods can also be used in risk management, particularly in stress test scenarios. The generation of stress test scenarios is not always easy, especially if the cross-sectional dimension is large. When several thousand risk factors are involved, setting shocks is fraught with difficulty. To use completion methods in this domain, it would be necessary to define a representative group of risk factors first. Shocks would be applied to that group only, and the completion methods would be used to extend their application to the remaining risk factors. The missing data would then correspond to the dates of the shocks. It would thus be possible to obtain a stress scenario that is defined by reference to the main risk factors. The procedure can also be used to understand how one risk factor varies in relation to another, providing important information about hedging. Evidently, completion methods can solve data quality issues while featuring in active risk management processes or even dynamic hedging, which makes them even more appealing to banks.

Bibliography

[1] A. Abraham and D. L. Ikenberry. "The individual investor and the weekend effect". In: *Journal of Financial and Quantitative Analysis* (1994), pages 263–277.

[2] J. Adams, D. Hayunga, S. Mansi, D. Reeb, and V. Verardi. "Identifying and treating outliers in finance". In: *Financial Management* 48.2 (2019), pages 345–384.

[3] C. Agostinelli, A. Leung, V. J. Yohai, and R. H. Zamar. "Robust estimation of multivariate location and scatter in the presence of cellwise and casewise contamination". In: *Test* 24.3 (2015), pages 441–461.

[4] H. Akaike. "Maximum likelihood identification of Gaussian autoregressive moving average models". In: *Biometrika* 60.2 (1973), pages 255–265.

[5] P. D. Allison. *Missing data*. Sage publications, 2001.

[6] L. Bachelier. "Théorie de la spéculation". In: *Annales scientifiques de l'École normale supérieure*. Volume 17. 1900, pages 21–86.

[7] Bank for International Settlements. *Statistical release : OTC derivatives statistics at end-December 2015 - Monetary and Economic Department*. May 2016. URL: https://www.bis.org/publ/otc_hy1605.pdf.

[8] Bank of International Settlements. *Consultative Document - Fundamental review of the trading book*. May 2012. URL: https://www.bis.org/publ/bcbs219.pdf.

[9] Bank of International Settlements. *Consultative Document - Fundamental review of the trading book. A revised market risk framework.* Oct. 2013. URL: https://www.bis.org/publ/bcbs265.pdf.

[10] Bank of International Settlements. *Consultative Document - Fundamental review of the trading book: outstanding issues*. Dec. 2014. URL: https://www.bis.org/publ/d305.pdf.

[11] Bank of International Settlements. *Minimum capital requirements for market risk*. Jan. 2016. URL: https://www.bis.org/bcbs/publ/d352.pdf.

[12] Bank of International Settlements. *Minimum capital requirements for market risk*. Jan. 2019. URL: https://www.bis.org/bcbs/publ/d457.pdf.

[13] J. Bartlett and C. Frost. "Reliability, repeatability and reproducibility: analysis of measurement errors in continuous variables". In: *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology* 31.4 (2008), pages 466–475.

[14] Basel Committee on Banking Supervision. *Amendment to the capital accord to incorporate market risks*. Bank for International Settlements, 1996. URL: https://www.bis.org/publ/bcbs24.pdf.

[15] Basel Committee on Banking Supervision. "Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework—Comprehensive Version". In: (2006). URL: https://www.bis.org/publ/bcbs128.pdf.

[16] Basel Committee on Banking Supervision. *Revisions to the Basel II Market Risk Framework*. Bank for International Settlements, 2009. URL: https://www.bis.org/publ/bcbs158.pdf.

[17] Basel Committee on Banking Supervision. *Basel III: A global regulatory framework for more resilient banks and banking systems*. Bank for International Settlements, 2010. URL: https://www.bis.org/publ/bcbs189_dec2010.pdf.

[18] Basel Committee on Banking Supervision. *Basel III: International framework for liquidity risk measurement, standards and monitoring*. Bank for International Settlements, 2010. URL: https://www.bis.org/publ/bcbs188.pdf.

[19] Basel Committee on Banking Supervision. *Progress in adopting the principles for effective risk data aggregation and risk reporting*. Dec. 2013. URL: https://www.bis.org/publ/bcbs268.pdf.

[20] Basel Committee on Banking Supervision. *Progress in adopting the principles for effective risk data aggregation and risk reporting*. Jan. 2015. URL: https://www.bis.org/bcbs/publ/d308.pdf.

[21] Basel Committee on Banking Supervision. *Progress in adopting the principles for effective risk data aggregation and risk reporting*. Dec. 2015. URL: https://www.bis.org/publ/d348.pdf.

[22] Basel Committee on Banking Supervision. *Frequently asked questions on market risk capital requirements*. Mar. 2018. URL: https://www.bis.org/bcbs/publ/d437.pdf.

[23] Basel Committee on Banking Supervision. *Progress in adopting the principles for effective risk data aggregation and risk reporting*. June 2018. URL: https://www.bis.org/publ/d443.pdf.

[24] Basel Committee on Banking Supervision. *Supervisory review process - Risk data aggregation and risk reporting*. Dec. 2019. URL: https://www.bis.org/publ/bcbs239.pdf.

[25] Basel Committee on Banking Supervision. *Progress in adopting the principles for effective risk data aggregation and risk reporting*. Apr. 2020. URL: https://www.bis.org/publ/d443.pdf.

[26] D. S. Bates. "The crash of '87: was it expected? The evidence from options markets". In: *The journal of finance* 46.3 (1991), pages 1009–1044.

[27] J. Bauer, O. Angelini, and A. Denev. "Imputation of multivariate time series data-performance benchmarks for multiple imputation and spectral techniques". In: *Available at SSRN 2996611* (2017).

[28] A. E. Beaton. "The use of special matrix operators in statistical calculus". In: *ETS Research Bulletin Series* 1964.2 (1964), pages i–222.

[29] J.-M. Beckers and M. Rixen. "EOF calculations and data filling from incomplete oceanographic datasets". In: *Journal of Atmospheric and oceanic technology* 20.12 (2003), pages 1839–1856.

[30] C. Bénard, G. Biau, S. Veiga, and E. Scornet. "Interpretable random forests via rule extraction". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pages 937–945.

[31] G. Biau, F. Cérou, and A. Guyader. "On the Rate of Convergence of the Bagged Nearest Neighbor Estimate." In: *Journal of Machine Learning Research* 11.2 (2010).

[32] G. Biau and L. Devroye. "On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification". In: *Journal of Multivariate Analysis* 101.10 (2010), pages 2499–2518.

[33] G. Biau, L. Devroye, and G. Lugosi. "Consistency of random forests and other averaging classifiers." In: *Journal of Machine Learning Research* 9.9 (2008).

[34] M. Billio, M. Costola, F. Mazzari, and L. Pelizzon. "The European Repo Market, ECB Intervention and the COVID-19 Crisis". In: *A New World Post COVID-19* 58 (2020).

[35] F. Black and M. Scholes. "The valuation of options and corporate liabilities". In: *Journal of Political Economy* 81.3 (1973), pages 637–654.

[36] T. E. Bodner. "What improves with increased missing data imputations?" In: *Structural Equation Modeling: A Multidisciplinary Journal* 15.4 (2008), pages 651–675.

[37] T. Bollerslev. "Generalized autoregressive conditional heteroskedasticity". In: *Journal of econometrics* 31.3 (1986), pages 307–327.

[38] C. Borio, C. Furfine, P. Lowe, et al. "Procyclicality of the financial system and financial stability: issues and policy options". In: *BIS papers* 1.3 (2001), pages 1–57.

[39] A. N. Borodin and P. Salminen. *Handbook of Brownian motion-facts and formulae*. Birkhäuser, 2012.

[40] H. J. Breaux. *On stepwise multiple linear regression*. Technical report. Army Ballistic Research Lab Aberdeen Proving Ground MD, 1967.

[41] L. Breiman. "Bagging predictors". In: *Machine learning* 24.2 (1996), pages 123–140.

[42] L. Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pages 5–32.

[43] L. Breiman. "Manual on setting up, using, and understanding random forests v3. 1. 2002". In: 1 (2002). URL: http://oz.%20berkeley.%20edu/users/breiman/Using%5C_random%5C_forests%5C_V3.

[44] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and regression trees*. CRC press, 1984.

[45] H. Briceño, C. M. Rocco, and E. Zio. "Singular spectrum analysis for forecasting of electric load demand". In: *Chemical Engineering* 33 (2013).

[46] D. S. Broomhead and G. P. King. "Extracting qualitative dynamics from experimental data". In: *Physica D: Nonlinear Phenomena* 20.2-3 (1986), pages 217–236.

[47] D. Broomhead and G. P. King. "On the qualitative analysis of experimental dynamical systems". In: *Nonlinear phenomena and chaos* 113 (1986), page 114.

[48] P. R. Brown. "Independent auditor judgment in the evaluation of internal audit functions". In: *Journal of accounting research* (1983), pages 444–455.

[49] H. Buehler, B. Horvath, T. Lyons, I. Perez Arribas, and B. Wood. "A data-driven market simulator for small data environments". In: *Available at SSRN 3632431* (2020).

[50] H. Buehler, B. Horvath, T. Lyons, I. Perez Arribas, and B. Wood. "Generating financial markets with signatures". In: *Available at SSRN* (2020).

[51] P. Bühlmann. "Bagging, boosting and ensemble methods". In: *Handbook of computational statistics*. Springer, 2012, pages 985–1022.

[52] P. L. Bühlmann and B. Yu. "Explaining bagging". In: *Research report/Seminar für Statistik, Eidgenössische Technische Hochschule Zürich*. Volume 92. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH). 2000.

[53] A. Buja and W. Stuetzle. "Observations on bagging". In: *Statistica Sinica* (2006), pages 323–351.

[54] S. v. Buuren and K. Groothuis-Oudshoorn. "mice: Multivariate imputation by chained equations in R". In: *Journal of statistical software* (2010), pages 1–68.

[55] R. D. Camino, C. A. Hammerschmidt, and R. State. "Improving missing data imputation with deep generative models". In: *arXiv preprint arXiv:1902.10666* (2019).

[56] P. Cerda, G. Varoquaux, and B. Kégl. "Similarity encoding for learning with dirty categorical variables". In: *Machine Learning* 107.8-10 (2018), pages 1477–1494.

[57] Committee on Payement and Settlement Systems - Board of the International Organization of Securities Commissions. *Authorities' access to trade repository data*. Aug. 2013. URL: https://www.bis.org/cpmi/publ/d110.pdf.

[58] Committee on Payment and Settlement Systems and Technical Committee of the International Organization of Securities Commissions. *Principles for financial market infrastructures*. Apr. 2012. URL: https://www.bis.org/cpmi/publ/d101a.pdf.

[59] J. Cortiñas Abrahantes, C. Sotto, G. Molenberghs, G. Vromman, and B. Bierinckx. "A comparison of various software tools for dealing with missing data via imputation". In: *Journal of Statistical Computation and Simulation* 81.11 (2011), pages 1653–1675.

[60] A. d'Aspremont, O. Banerjee, and L. El Ghaoui. "First-order methods for sparse covariance selection". In: *SIAM Journal on Matrix Analysis and Applications* 30.1 (2008), pages 56–66.

[61] G. E. Dallal. "PC-SIZE: A program for sample-size determinations". In: *The American Statistician* 40.1 (1986), pages 52–52.

[62] K. Daniel. *Thinking, fast and slow*. 2011.

[63] J. Dash, X. Yang, H. J. Stein, and M. Bondioli. "Stable Reduced-Noise'Macro'SSA-Based Correlations for Long-Term Counterparty Risk Management". In: *Available at SSRN 2808015* (2016).

[64] J. Dash and Y. Zhang. "Cleaning Financial Data Using SSA and MSSA". In: *Available at SSRN 2808156* (2016).

[65] J. Dash and Y. Zhang. "MSSA vs. Multivariate Regularized Expectation Maximization for Data Cleaning". In: *Multivariate Regularized Expectation Maximization for Data Cleaning (July 19, 2016)* (2016).

[66] S. C. Dass and V. N. Nair. "Edge detection, spatial smoothing, and image reconstruction with partially observed multivariate data". In: *Journal of the American Statistical Association* 98.461 (2011), pages 77–89.

[67] A. P. Dempster. "Covariance selection". In: *Biometrics* (1972), pages 157–175.

[68] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society. Series B (methodological)* (1977), pages 1–38.

[69] W. J. Dixon. "Analysis of extreme values". In: *The Annals of Mathematical Statistics* 21.4 (1950), pages 488–506.

[70] Y. Dodge and D. Commenges. *The Oxford dictionary of statistical terms*. Oxford University Press on Demand, 2006.

[71] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[72] F. Eckerli. "Generative Adversarial Networks in finance: an overview". In: *Available at SSRN 3864965* (2021).

[73] B. Efron. "Missing data, imputation, and the bootstrap". In: *Journal of the American Statistical Association* 89.426 (1994), pages 463–475.

[74] J. B. Elsner and A. A. Tsonis. *Singular spectrum analysis: a new tool in time series analysis*. Springer Science & Business Media, 2013.

[75] E. J. Elton, M. r. J. Gruber, and C. R. Blake. "A first look at the accuracy of the CRSP mutual fund database and a comparison of the CRSP and Morningstar mutual fund databases". In: *Investments And Portfolio Performance*. World Scientific, 2011, pages 99–114.

[76] E. J. Elton, M. J. Gruber, and C. R. Blake. "The persistence of risk-adjusted mutual fund performance". In: *Journal of business* (1996), pages 133–157.

[77] R. F. Engle. "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". In: *Econometrica: Journal of the econometric society* (1982), pages 987–1007.

[78] Eurex Reseach. *Euro Stoxx 500 Index implied repo trading at Eurex*. Jan. 2017. URL: https://www.eurexchange.com/resource/blob/34718/622649dd378e4338d2f56cae1adf data/eurostoxx-repo-trf.pdf.

[79] European Banking Authority. *EBA report on Credit Valuation Adjustment (CVA) under Article 456(2) of Regulation (EU) No 575/2013 (Capital Requirements Regulation — CRR) and EBA review on the application of CVA charges to non-financial counterparties established in a third country under Article 382(5) of Regulation (EU) No 575/2013 (Capital Requirements Regulation — CRR)*. Jan. 2015. URL: https://eba.europa.eu/sites/default/documents/files/documents/10180/950548/1ab4df48-03c6-431b-a754-1b5e7efdfebd/EBA%20Report%20on%20CVA.pdf.

[80] European Banking Authority. *Final Draft RTS on the calculation of the stress scenario risk measure under Article 325bk(3) of Regulation (EU) No 575/2013 (Capital Requirements Regulation 2 – CRR2)*. Dec. 2020. URL: https://www.eba.europa.eu/sites/default/documents/files/document_library/Publications/Draft%20Technical%20Standards/2020/RTS/961600/Final%20draft%20RTS%20on%20the%20calculation%20of%20stress%20scenario%20risk%20measure.pdf.

[81] European Central Bank - Banking Supervision. *ECB guide to internal models*. Oct. 2019. URL: https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.guidetointernalmodels_consolidated_201910~97fd49fb08.en.pdf.

[82] Financial Stability Board. *Strengthening Oversight and Regulation of Shadow Banking, Policy Framework for Strengthening Oversight and Regulation of Shadow Banking Entities*. Aug. 2013. URL: https://www.fsb.org/wp-content/uploads/r_130829c.pdf.

[83] E. Fix and J. L. Hodges Jr. *Discriminatory analysis-nonparametric discrimination: consistency properties*. Technical report. California Univ Berkeley, 1951.

[84] A. Folch-Fortuny, F. Arteaga, and A. Ferrer. "Missing data imputation toolbox for MATLAB". In: *Chemometrics and Intelligent Laboratory Systems* 154 (2016), pages 93–100.

[85] K. R. French. "Stock returns and the weekend effect". In: *Journal of financial economics* 8.1 (1980), pages 55–69.

[86] J. Friedman, T. Hastie, R. Tibshirani, and M. R. Tibshirani. *Package 'glasso'*. 2015.

[87] J. Friedman, T. Hastie, and R. Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3 (2008), pages 432–441.

[88] J. H. Friedman and P. Hall. "On bagging and nonlinear estimation". In: *Journal of statistical planning and inference* 137.3 (2007), pages 669–683.

[89] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal. "Pattern classification with missing data: a review". In: *Neural Computing and Applications* 19.2 (2010), pages 263–282.

[90] A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[91] M. R. Gibbons and P. Hess. "Day of the week effects and asset returns". In: *Journal of business* (1981), pages 579–596.

[92] J. Gillard, N. Antoun, N. Burnet, and J. Pickard. "Reproducibility of quantitative CT perfusion imaging". In: *The British journal of radiology* 74.882 (2001), pages 552–555.

[93] P. Giot and S. Laurent. "Modelling daily value-at-risk using realized volatility and ARCH type models". In: *Journal of empirical finance* 11.3 (2004), pages 379–398.

[94] N. Golyandina, V. Nekrutkin, and A. A. Zhigljavsky. *Analysis of time series structure: SSA and related techniques*. CRC press, 2001.

[95] N. Golyandina and A. Zhigljavsky. *Singular Spectrum Analysis for time series*. Volume 120. Springer, 2013.

[96] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[97] M. M. Grabka and C. Westermeier. *Editing and multiple imputation of item non-response in the wealth module of the German Socio-Economic Panel*. Technical report. SOEP Survey Papers, 2015.

[98] J. W. Graham. *Missing data: Analysis and design*. Springer Science & Business Media, 2012.

[99] Y. Grandvalet. "Bagging equalizes influence". In: *Machine Learning* 55.3 (2004), pages 251–270.

[100] F. E. Grubbs et al. "Sample criteria for testing outlying observations". In: *The Annals of Mathematical Statistics* 21.1 (1950), pages 27–58.

[101] P. S. Hagan, D. Kumar, A. S. Lesniewski, and D. E. Woodward. "Managing smile risk". In: *The Best of Wilmott* 1 (2002), pages 249–296.

[102] P. Hall and R. J. Samworth. "Properties of bagged nearest neighbour classifiers". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.3 (2005), pages 363–379.

[103] H. Hassani. "Singular spectrum analysis: methodology and comparison". In: (2007).

[104] H. Hassani and R. Mahmoudvand. "Multivariate singular spectrum analysis: A general view and new vector forecasting approach". In: *International Journal of Energy and Statistics* 1.01 (2013), pages 55–83.

[105] H. Hassani, R. Mahmoudvand, and M. Zokaei. "Separability and window length in singular spectrum analysis". In: *Comptes rendus mathematique* 349.17-18 (2011), pages 987–990.

[106] D. M. Hawkins. "A new test for multivariate normality and homoscedasticity". In: *Technometrics* 23.1 (1981), pages 105–110.

[107] P. Henry-Labordere. "Generative models for financial data". In: *Available at SSRN 3408007* (2019).

[108] B. Herzog and S. Osamah. "Reverse engineering of option pricing: an AI application". In: *International Journal of Financial Studies* 7.4 (2019), page 68.

[109] S. L. Heston. "A closed-form solution for options with stochastic volatility with applications to bond and currency options". In: *The review of financial studies* 6.2 (1993), pages 327–343.

[110] N. J. Higham. "Computing the nearest correlation matrix—a problem from finance". In: *IMA journal of Numerical Analysis* 22.3 (2002), pages 329–343.

[111] P. J. Hoffman, P. Slovic, and L. G. Rorer. "An analysis-of-variance model for the assessment of configural cue utilization in clinical judgment". In: *Psychological bulletin* 69.5 (1968), page 338.

[112] J. Honaker, A. Joseph, G. King, K. Scheve, and N. Singh. *Amelia: a program for missing data. 2001*. 2002.

[113] J. Honaker, G. King, M. Blackwell, et al. "Amelia II: A program for missing data". In: *Journal of statistical software* 45.7 (2011), pages 1–47.

[114] J. Huang, J. W. Keung, F. Sarro, Y.-F. Li, Y.-T. Yu, W. Chan, and H. Sun. "Cross-validation based K nearest neighbor imputation for software quality datasets: An empirical study". In: *Journal of Systems and Software* 132 (2017), pages 226–252.

[115] M. Hubert, P. J. Rousseeuw, and W. Van den Bossche. "MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers". In: *Technometrics* 61.4 (2019), pages 459–473.

[116] M. Hubert, P. J. Rousseeuw, and K. Vanden Branden. "ROBPCA: a new approach to robust principal component analysis". In: *Technometrics* 47.1 (2005), pages 64–79.

[117] S. Iizuka, E. Simo-Serra, and H. Ishikawa. "Globally and locally consistent image completion". In: *ACM Transactions on Graphics (ToG)* 36.4 (2017), pages 1–14.

[118] N. Ikeda and S. Watanabe. *Stochastic differential equations and diffusion processes*. Volume 24. Elsevier, 2014.

[119] International Swaps and Derivatives Association. *CDS Market Summary: Market Risk Transaction Activity*. Oct. 2013. URL: https://www.isda.org/a/HPDDE/cds-research-note-final-2013-10-01.pdf.

[120] R. Ioan. "Portfolio Selection During Crises Using Principal Component Analysis". In: *Timisoara Journal of Economics and Business* 13.2 (2020), pages 129–144.

[121] Irving Fisher Committee on Central Bank Statistics. *Central banks and trade repositories derivatives data*. Oct. 2018. URL: https://www.bis.org/ifc/publ/ifc_report_cb_trade_rep_deriv_data.pdf.

[122] V. Jain and R. Gupta. "Identification of linear systems through a Grammian technique". In: *International Journal of Control* 12.3 (1970), pages 421–431.

[123] M. Jamshidian and S. Jalal. "Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data". In: *Psychometrika* 75.4 (2010), pages 649–674.

[124] J. M. Jerez, I. Molina, P. J. Garcièa-Laencina, E. Alba, N. Ribelles, M. Martièn, and L. Franco. "Missing data imputation using statistical and machine learning methods in a real breast cancer problem". In: *Artificial intelligence in medicine* 50.2 (2010), pages 105–115.

[125] L. Josifovski, M. Cooke, P. Green, and A. Vizinho. "State based imputation of missing data for robust speech recognition and speech enhancement". In: *Sixth European Conference on Speech Communication and Technology*. 1999.

[126] J. Josse and F. Husson. "Handling missing values in exploratory multivariate data analysis methods". In: *Journal de la Société Française de Statistique* 153.2 (2012), pages 79–99.

[127] J. Josse, J. Pagès, and F. Husson. "Multiple imputation in principal component analysis". In: *Advances in data analysis and classification* 5.3 (2011), pages 231–246.

[128] G. Keppel. *Design and analysis: A researcher's handbook*. Prentice-Hall, Inc, 1991.

[129] H. A. Kiers. "Weighted least squares fitting using ordinary least squares algorithms". In: *Psychometrika* 62.2 (1997), pages 251–266.

[130] J.-O. Kim and J. Curry. "The treatment of missing data in multivariate analysis". In: *Sociological Methods & Research* 6.2 (1977), pages 215–240.

[131] K. H. Kim and P. M. Bentler. "Tests of homogeneity of means and covariance matrices for multivariate incomplete data". In: *Psychometrika* 67.4 (2002), pages 609–623.

[132] G. A. Klein. *Sources of power: How people make decisions*. MIT press, 2017.

[133] P. Kofman and I. G. Sharpe. "Using multiple imputation in the analysis of incomplete observations in finance". In: *Journal of Financial Econometrics* 1.2 (2003), pages 216–249.

[134] A. Kondratyev and C. Schwarz. "The market generator". In: *Available at SSRN 3384948* (2019).

[135] J. Kruger and D. Dunning. "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments." In: *Journal of personality and social psychology* 77.6 (1999), page 1121.

[136] H. R. Kunsch. "The jackknife and the bootstrap for general stationary observations". In: *The annals of Statistics* (1989), pages 1217–1241.

[137] F. Lagona and M. Picone. "A latent-class model for clustering incomplete linear and circular data in marine studies". In: *Journal of Data Science* 9.4 (2011), pages 585–605.

[138] K. J. Lau, Y. K. Goh, and A. C. Lai. "An empirical study on asymmetric jump diffusion for option and annuity pricing". In: *PloS one* 14.5 (2019), e0216529.

[139] O. Ledoit and M. Wolf. "Honey, I shrunk the sample covariance matrix". In: *The Journal of Portfolio Management* 30.4 (2004), pages 110–119.

[140] Y. Li and Y. Chen. "Research on Initialization on EM Algorithm Based on Gaussian Mixture Model". In: *Journal of Applied Mathematics and Physics* 6.1 (2018), pages 11–17.

[141] M. Lichman et al. *UCI machine learning repository, 2013*. 2013. URL: https://archive.ics.uci.edu/.

[142] R. J. Little. "A test of missing completely at random for multivariate data with missing values". In: *Journal of the American statistical Association* 83.404 (1988), pages 1198–1202.

[143] R. J. Little. "Missing-data adjustments in large surveys". In: *Journal of Business & Economic Statistics* 6.3 (1988), pages 287–296.

[144] R. J. Little and D. B. Rubin. "The analysis of social science data with missing values". In: *Sociological Methods & Research* 18.2-3 (1989), pages 292–326.

[145] R. J. Little and D. B. Rubin. *Statistical analysis with missing data*. Volume 793. John Wiley & Sons, 2019.

[146] J. Liu, S. Kumar, and D. P. Palomar. "Parameter estimation of heavy-tailed AR model with missing data via stochastic EM". In: *IEEE Transactions on Signal Processing* 67.8 (2019), pages 2159–2172.

[147] Q. Liu and Y. An. "Risk contributions of trading and non-trading hours: Evidence from Chinese commodity futures markets". In: *Pacific-Basin Finance Journal* 30 (2014), pages 17–29.

[148] J. Mandel. "Repeatability and reproducibility". In: *Journal of Quality Technology* 4.2 (1972), pages 74–85.

[149] A. Marshall, D. G. Altman, and R. L. Holder. "Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study". In: *BMC medical research methodology* 10.1 (2010), page 112.

[150] A. Marshall, D. G. Altman, P. Royston, and R. L. Holder. "Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study". In: *BMC medical research methodology* 10.1 (2010), page 7.

[151] M. Martin et al. "Epidemiological study of the GEICAM group about breast cancer in Spain (1990-1993): El Alamo project". In: *Medicina clinica* 122.1 (2004), pages 12–17.

[152] F. Mealli and D. B. Rubin. "Clarifying missing at random and related definitions, and implications when coupled with exchangeability". In: *Biometrika* 102.4 (2015), pages 995–1000.

[153] P. E. Meehl. "Clinical versus statistical prediction: A theoretical analysis and a review of the evidence". In: (1954).

[154] R. C. Merton. "Option pricing when underlying stock returns are discontinuous". In: *Journal of financial economics* 3.1-2 (1976), pages 125–144.

[155] E. M. Miller. "Why a weekend effect?" In: *Journal of Portfolio Management* 14.4 (1988), page 43.

[156] K. Modarresi and J. Diner. "An evaluation metric for content providing models, recommendation systems, and online campaigns". In: *International Conference on Computational Science*. Springer. 2019, pages 550–563.

[157] G. Molenberghs, G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke. *Handbook of missing data methodology*. CRC Press, 2014.

[158] J. N. Morgan and J. A. Sonquist. "Problems in the analysis of survey data, and a proposal". In: *Journal of the American statistical association* 58.302 (1963), pages 415–434.

[159] Morgan Stanley Capital International. *The Global Industry Classification Standard*. URL: https://www.msci.com/gics.

[160] C. I. Mosier et al. "Symposium: The need and means of cross-validation". In: *Educational and Psychological Measurement* 11.1 (1951), pages 5–11.

[161] C. R. Nelson and A. F. Siegel. "Parsimonious modeling of yield curves". In: *Journal of business* (1987), pages 473–489.

[162] P. R. Nelson, P. A. Taylor, and J. F. MacGregor. "Missing data methods in PCA and PLS: Score calculations with incomplete observations". In: *Chemometrics and intelligent laboratory systems* 35.1 (1996), pages 45–65.

[163] L. Pan, J. Li, et al. "K-nearest neighbor based missing data estimation algorithm in wireless sensor networks". In: *Wireless Sensor Network* 2.02 (2010), page 115.

[164] K. Pearson. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pages 559–572.

[165] B. Peirce. "On Peirce's criterion". In: *Proceedings of the American Academy of Arts and Sciences*. Volume 13. JSTOR. 1877, pages 348–351.

[166] Q. A. Raaijmakers. "Effectiveness of different missing data treatments in surveys with Likert-type data: Introducing the relative mean substitution approach". In: *Educational and Psychological Measurement* 59.5 (1999), pages 725–748.

[167] Regulation, EC. *No 575/2013 of the European Parliament and of the Council of 26 June 2013 on prudential requirements for credit institutions and investment firms and amending Regulation (EU) No 648/2012 Text with EEA relevance.* 2013.

[168] Regulation, EU. *No 648/2012 of the European Parliament and of the Council on OTC derivatives, central counterparties and trade repositories, Pub. L. No. 648/2012*. July 2012. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32012R0648&from=EN.

[169] Regulation, EU. *No 2015/2365 of the European Parliament and of the Council on transparency of securitues financing transactions and of reuse and amending Regulation, Pub. L. No. 648/2012*. Nov. 2015. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32015R2365&from=FR.

[170] P. C. Rodrigues and R. Mahmoudvand. "The benefits of multivariate singular spectrum analysis over the univariate version". In: *Journal of the Franklin Institute* 355.1 (2018), pages 544–564.

[171] B. Rosner. "Percentage points for a generalized ESD many-outlier procedure". In: *Technometrics* 25.2 (1983), pages 165–172.

[172] P. L. Roth, F. S. Switzer III, and D. M. Switzer. "Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques". In: *Organizational research methods* 2.3 (1999), pages 211–232.

[173] P. J. Rousseeuw and G. Molenberghs. "Transformation of non positive semidefinite correlation matrices". In: *Communications in Statistics–Theory and Methods* 22.4 (1993), pages 965–984.

[174] D. B. Rubin. "Inference and missing data". In: *Biometrika* 63.3 (1976), pages 581–592.

[175] D. B. Rubin. "Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse". In: *Proceedings of the survey research methods section of the American Statistical Association*. Volume 1. American Statistical Association. 1978, pages 20–34.

[176] D. B. Rubin. "Statistical matching using file concatenation with adjusted weights and multiple imputations". In: *Journal of Business & Economic Statistics* 4.1 (1986), pages 87–94.

[177] D. B. Rubin. *Multiple imputation for nonresponse in surveys*. Volume 81. John Wiley & Sons, 1987.

[178] L. H. Rubin, K. Witkiewitz, J. S. Andre, and S. Reilly. "Methods for handling missing data in the behavioral neurosciences: Don't throw the baby rat out with the bath water". In: *Journal of Undergraduate Neuroscience Education* 5.2 (2007), A71.

[179] Z. Sahri, R. Yusof, and J. Watada. "FINNIM: Iterative imputation of missing values in dissolved gas analysis dataset". In: *IEEE Transactions on Industrial Informatics* 10.4 (2014), pages 2093–2102.

[180] J. Salma. "Construction du taux de rachat structurel en épargne : approximation non-linéaire et agrégation de modèles". Mémoire IA. ISFA, 2016.

[181] J. L. Schafer. *Analysis of incomplete multivariate data*. Chapman and Hall/CRC, 1997.

[182] J. L. Schafer and J. W. Graham. "Missing data: our view of the state of the art." In: *Psychological methods* 7.2 (2002), page 147.

[183] N. Schenker and J. M. Taylor. "Partially parametric techniques for multiple imputation". In: *Computational statistics & data analysis* 22.4 (1996), pages 425–446.

[184] F. W. Scholz and M. A. Stephens. "K-sample Anderson–Darling tests". In: *Journal of the American Statistical Association* 82.399 (1987), pages 918–924.

[185] G. Schwarz et al. "Estimating the dimension of a model". In: *The annals of statistics* 6.2 (1978), pages 461–464.

[186] E. Scornet, G. Biau, J.-P. Vert, et al. "Consistency of random forests". In: *The Annals of Statistics* 43.4 (2015), pages 1716–1741.

[187] A. D. Shah, J. W. Bartlett, J. Carpenter, O. Nicholas, and H. Hemingway. "Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study". In: *American journal of epidemiology* 179.6 (2014), pages 764–774.

[188] E. Shireman, D. Steinley, and M. J. Brusco. "Examining the effect of initialization strategies on the performance of Gaussian mixture modeling". In: *Behavior research methods* 49.1 (2017), pages 282–293.

[189] Siemens. *Spin-off of Siemens Energy approved by large majority of Siemens shareholders*. July 2020. URL: https://press.siemens.com/global/en/pressrelease/spin-siemens-energy-approved-large-majority-siemens-shareholders.

[190] Société de Calcul Mathémtatique SA. *Reconstruction de données manquantes dans des séries chronologiques*. July 2006. URL: http://www.scmsa.com/RMM/SCMSA_reconstruction_Vendee_2006.pdf.

[191] M. S. Srivastava and M. Dolatabadi. "Multiple imputation and other resampling schemes for imputing missing observations". In: *Journal of Multivariate Analysis* 100.9 (2009), pages 1919–1937.

[192] N. Städler and P. Bühlmann. "Pattern alternating maximization algorithm for high-dimensional missing data". In: *Arxiv preprint arXiv* 1005 (2010).

[193] D. J. Stekhoven. *missForest: Nonparametric Missing Value Imputation using Random Forest*. R package version 1.4. 2013. URL: http://www.r-project.org/.

[194] D. J. Stekhoven and P. Bühlmann. "MissForest—non-parametric missing value imputation for mixed-type data". In: *Bioinformatics* 28.1 (2011), pages 112–118.

[195] E. Stewart. "How Mike Bloomberg made his billions: a computer system you've probably never seen. The Bloomberg Terminal, explained." In: *Vox* (Dec. 2019). URL: https://www.vox.com/2020-presidential-election/2019/12/11/21005008/michael-bloomberg-terminal-net-worth-2020#:~:text=According%20to%20data%20from%20Burton,now%20Refinitiv%2C%20had%20%246.7%20billion..

[196] L. E. Svensson. *Estimating and interpreting forward interest rates: Sweden 1992-1994*. Technical report. National bureau of economic research, 1994.

[197] W. R. Thompson. "On a criterion for the rejection of observations and the distribution of the ratio of deviation to sample standard deviation". In: *The Annals of Mathematical Statistics* 6.4 (1935), pages 214–219.

[198] R. Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pages 267–288.

[199] G. L. Tietjen and R. H. Moore. "Some Grubbs-type statistics for the detection of several outliers". In: *Technometrics* 14.3 (1972), pages 583–597.

[200] A. Traverso, L. Wee, A. Dekker, and R. Gillies. "Repeatability and reproducibility of radiomic features: a systematic review". In: *International Journal of Radiation Oncology* Biology* Physics* 102.4 (2018), pages 1143–1158.

[201] N. Tsikriktsis. "A review of techniques for treating missing data in OM survey research". In: *Journal of operations management* 24.1 (2005), pages 53–62.

[202] J. W. Tukey. *Exploratory data analysis*. Volume 2. Reading, Mass., 1977.

[203] United States. Congress. *Dodd-Frank Wall Street Reform and Consumer Protection Act: Conference Report (to Accompany HR 4173)*. Volume 111. 517. US Government Printing Office, 2010.

[204] X. Urli. "Algorithme d'espérance maximisation (EM) pour l'imputation de données Utilisation sur des séries financières". In: (2007).

[205] S. Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.

[206] S. Van Buuren and K. Oudshoorn. *Flexible multivariate imputation by MICE*. Leiden: TNO, 1999.

[207] R. Verma and J. C. Goodale. "Statistical power in operations management research". In: *Journal of Operations Management* 13.2 (1995), pages 139–152.

[208] A. Warga. "Bond returns, liquidity, and missing data". In: *Journal of Financial and Quantitative Analysis* 27.4 (1992), pages 605–617.

[209] J. Wiesinger, D. Sornette, and J. Satinover. "Reverse engineering financial markets with majority and minority games using genetic algorithms". In: *Computational Economics* 41.4 (2013), pages 475–492.

[210] D. R. Wilson and T. R. Martinez. "Improved heterogeneous distance functions". In: *Journal of artificial intelligence research* 6 (1997), pages 1–34.

[211] D. M. Witten, J. H. Friedman, and N. Simon. "New insights and faster computations for the graphical lasso". In: *Journal of Computational and Graphical Statistics* 20.4 (2011), pages 892–900.

[212] C. J. Wu. "On the convergence properties of the EM algorithm". In: *The Annals of statistics* (1983), pages 95–103.

[213] J. Yoon, J. Jordon, and M. Schaar. "Gain: Missing data imputation using generative adversarial nets". In: *International Conference on Machine Learning*. PMLR. 2018, pages 5689–5698.

[214] J. Young. "Imputation for Random Forests". In: (2017).

[215] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. "Generative image inpainting with contextual attention". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pages 5505–5514.

[216] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. "Free-form image inpainting with gated convolution". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pages 4471–4480.

[217] S. Zhang, X. Wu, and M. Zhu. "Efficient missing data imputation for supervised learning". In: *9th IEEE International Conference on Cognitive Informatics (ICCI'10)*. IEEE. 2010, pages 672–679.

[218] R. Zhou, J. Liu, S. Kumar, and D. P. Palomar. "Student's t VAR Modeling With Missing Data Via Stochastic EM and Gibbs Sampling". In: *IEEE Transactions on Signal Processing* 68 (2020), pages 6198–6211.

APPENDIX

Appendix A: MCAR data on the first column

A.1 Number of MCAR tests that are calculable, when applied to return matrices containing MCAR data on the first column, for a 5% significance level

	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Little's test	100	100	100	100	100	100	100	100	100	100	100	100	100	100
J&J's test	100	100	100	100	100	100	100	100	100	100	100	100	100	100

A.2 Standard deviation of each statistical moment among all missingness scenarios, with MCAR data in the first column

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Mean (10^{-5})	**LI**	0.84	0.84	1.54	1.94	2.32	3.59	4.28	4.22	5.51	5.66	5.92	7.10	7.49	6.69
	LOCF	1.21	1.26	1.73	2.15	2.34	2.99	2.52	2.78	3.22	3.11	3.22	3.25	3.17	3.18
	MSSA	0.73	0.83	1.13	1.47	1.73	2.52	2.21	2.64	3.25	3.32	3.55	3.69	4.02	4.09
	BB	1.25	1.34	1.79	2.28	2.48	3.06	2.63	2.93	3.34	3.25	3.40	3.50	3.34	3.40
	IPCA	0.47	0.50	0.56	0.70	0.77	1.05	1.10	1.47	1.81	1.88	2.53	2.32	3.97	3.62
	KNN	0.20	0.21	0.30	0.38	0.41	0.67	0.67	0.61	0.64	0.79	0.93	0.92	1.02	1.15
	RF	0.00	0.01	0.02	0.03	0.03	1.00	0.55	0.05	0.70	0.83	1.59	0.82	1.35	1.99
	MICE	1.25	1.28	1.80	2.21	2.46	3.20	3.07	3.19	4.17	4.37	4.78	5.19	5.25	7.07
	MIPCA	0.32	0.36	0.43	0.53	0.59	0.77	0.77	0.78	1.02	1.08	1.31	1.58	1.75	2.25
	Amelia	0.28	0.35	0.42	0.50	0.54	0.70	0.68	0.74	0.76	0.81	0.98	1.10	2.28	3.55
Std (10^{-3})	**LI**	0.05	0.07	0.09	0.10	0.11	0.11	0.12	0.16	0.12	0.13	0.18	0.15	0.18	0.16
	LOCF	0.08	0.11	0.14	0.16	0.17	0.21	0.24	0.27	0.23	0.29	0.32	0.32	0.38	0.44
	MSSA	0.11	0.15	0.17	0.17	0.17	0.19	0.23	0.24	0.22	0.24	0.24	0.26	0.30	0.28
	BB	0.06	0.09	0.12	0.14	0.18	0.20	0.26	0.32	0.37	0.43	0.55	0.63	0.80	1.05
	IPCA	0.04	0.06	0.07	0.09	0.11	0.11	0.14	0.17	0.18	0.20	0.34	0.43	0.60	0.83
	KNN	0.06	0.07	0.09	0.11	0.11	0.10	0.13	0.14	0.16	0.15	0.18	0.16	0.18	0.21
	RF	0.00	0.02	0.03	0.04	0.06	0.06	0.07	0.08	0.12	0.10	0.18	0.20	0.27	0.34
	MICE	0.09	0.13	0.18	0.23	0.32	0.41	0.63	0.73	1.04	1.26	1.43	1.85	2.08	2.17
	MIPCA	0.03	0.05	0.06	0.08	0.09	0.09	0.11	0.12	0.16	0.17	0.25	0.25	0.29	0.41
	Amelia	0.03	0.05	0.06	0.08	0.09	0.09	0.11	0.13	0.15	0.16	0.23	0.27	0.49	1.35

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Skewness	**LI**	0.04	0.06	0.08	0.10	0.11	0.13	0.16	0.17	0.20	0.24	0.27	0.27	0.32	0.39
	LOCF	0.09	0.14	0.17	0.21	0.25	0.32	0.35	0.40	0.51	0.55	0.54	0.58	0.64	0.75
	MSSA	0.07	0.11	0.11	0.12	0.13	0.11	0.12	0.13	0.16	0.13	0.14	0.17	0.18	0.23
	BB	0.05	0.06	0.06	0.07	0.06	0.06	0.07	0.06	0.06	0.06	0.05	0.05	0.04	0.06
	IPCA	0.02	0.03	0.04	0.05	0.06	0.08	0.13	0.25	0.31	0.37	0.50	0.70	0.81	1.17
	KNN	0.04	0.06	0.07	0.08	0.10	0.10	0.12	0.14	0.17	0.18	0.20	0.19	0.19	0.22
	RF	0.00	0.01	0.03	0.03	0.04	0.05	0.05	0.05	0.06	0.10	0.08	0.24	0.28	0.32
	MICE	0.11	0.19	0.26	0.37	0.48	0.61	0.78	0.85	0.97	1.18	1.26	1.42	1.62	1.86
	MIPCA	0.02	0.04	0.04	0.05	0.06	0.06	0.08	0.10	0.14	0.16	0.23	0.34	0.44	0.76
	Amelia	0.02	0.03	0.04	0.05	0.06	0.05	0.06	0.07	0.08	0.09	0.11	0.16	0.18	0.35
Kurtosis	**LI**	0.07	0.10	0.14	0.18	0.20	0.24	0.33	0.31	0.42	0.48	0.57	0.72	0.81	1.18
	LOCF	0.28	0.43	0.50	0.64	0.84	1.07	1.20	1.34	1.80	2.84	2.45	2.09	2.72	3.61
	MSSA	0.14	0.18	0.20	0.23	0.25	0.24	0.30	0.36	0.40	0.50	0.50	0.72	1.01	1.12
	BB	0.12	0.10	0.11	0.10	0.08	0.09	0.08	0.07	0.06	0.06	0.06	0.04	0.04	0.15
	IPCA	0.06	0.07	0.10	0.10	0.12	0.15	0.31	0.79	0.93	1.10	1.19	1.98	2.04	2.90
	KNN	0.07	0.16	0.18	0.21	0.31	0.27	0.45	0.40	0.68	0.60	0.67	0.63	0.66	1.02
	RF	0.00	0.02	0.05	0.06	0.07	0.08	0.08	0.11	0.13	0.30	0.22	0.56	0.60	0.79
	MICE	0.35	0.51	0.53	0.73	0.80	1.16	1.34	1.45	2.03	3.29	3.35	3.33	3.72	3.57
	MIPCA	0.05	0.08	0.10	0.11	0.15	0.14	0.26	0.26	0.39	0.47	0.63	0.87	1.43	2.34
	Amelia	0.05	0.08	0.10	0.11	0.13	0.12	0.15	0.16	0.21	0.20	0.24	0.43	0.44	0.94

A.3 Average number of principal components used by IPCA and MIPCA, among the 100 missingness scenarios, for imputation of MCAR data in the first column

	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
IPCA	4	4	4	4	4	4	4	4	4	3.99	4.02	4.04	4.01	4.01
MIPCA	4	4	4	4	4	4	4	3.98	3.89	3.73	3.4	3.28	3.06	3.01

A.4 Standard deviation of the proximity measures among all missingness scenarios, with MCAR data in the first column

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
MAE (10^{-3})	**LI**	0.74	0.49	0.41	0.32	0.28	0.23	0.25	0.23	0.19	0.16	0.17	0.15	0.15	0.12
	LOCF	1.04	0.77	0.65	0.54	0.45	0.43	0.34	0.33	0.29	0.30	0.31	0.28	0.27	0.22
	MSSA	0.97	0.72	0.59	0.49	0.42	0.38	0.34	0.32	0.27	0.26	0.26	0.21	0.18	0.22
	BB	0.38	0.21	0.16	0.16	0.17	0.20	0.23	0.23	0.26	0.34	0.39	0.44	0.56	0.75
	IPCA	0.34	0.28	0.22	0.23	0.27	0.34	0.50	0.58	0.57	0.46	0.49	0.48	0.68	0.83
	KNN	0.49	0.34	0.28	0.23	0.20	0.18	0.16	0.17	0.16	0.16	0.16	0.12	0.14	0.15
	RF	0.19	0.24	0.23	0.22	0.21	0.17	0.15	0.13	0.14	0.14	0.13	0.16	0.22	0.31
	MICE	0.78	0.61	0.58	0.55	0.57	0.59	0.76	0.73	0.95	1.08	1.11	1.32	1.31	1.27
	MIPCA	0.26	0.17	0.15	0.14	0.12	0.11	0.17	0.18	0.21	0.22	0.22	0.24	0.32	0.43
	Amelia	0.23	0.16	0.12	0.11	0.09	0.08	0.07	0.09	0.10	0.12	0.14	0.21	0.42	1.19
RMSE (10^{-3})	**LI**	0.77	0.53	0.43	0.35	0.30	0.26	0.28	0.26	0.20	0.17	0.19	0.17	0.16	0.13
	LOCF	1.22	0.90	0.74	0.62	0.51	0.55	0.46	0.44	0.43	0.45	0.44	0.40	0.41	0.42
	MSSA	1.09	0.84	0.69	0.57	0.47	0.45	0.40	0.38	0.33	0.32	0.33	0.29	0.24	0.28
	BB	0.46	0.25	0.19	0.20	0.22	0.26	0.30	0.30	0.33	0.43	0.49	0.57	0.71	0.95
	IPCA	0.44	0.33	0.27	0.28	0.32	0.44	0.64	0.76	0.78	0.61	0.67	0.69	0.97	1.19
	KNN	0.53	0.39	0.33	0.29	0.27	0.25	0.24	0.24	0.28	0.24	0.25	0.22	0.26	0.30
	RF	0.42	0.47	0.38	0.31	0.27	0.21	0.19	0.14	0.22	0.20	0.22	0.23	0.35	0.49
	MICE	0.98	0.78	0.72	0.68	0.69	0.77	0.95	1.01	1.33	1.53	1.57	1.99	2.08	2.15
	MIPCA	0.34	0.23	0.20	0.20	0.18	0.17	0.24	0.24	0.29	0.28	0.31	0.34	0.47	0.65
	Amelia	0.30	0.20	0.16	0.15	0.13	0.12	0.11	0.12	0.14	0.17	0.20	0.29	0.57	1.61

A.5 Standard deviation of the covariance matrix differences among all missingness scenarios, with MCAR data in the first column

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Cov (10^{-4})	**LI**	0.03	0.01	0.03	0.06	0.08	0.09	0.11	0.14	0.11	0.11	0.14	0.12	0.13	0.11
	LOCF	0.07	0.09	0.12	0.14	0.16	0.20	0.21	0.24	0.24	0.31	0.34	0.33	0.40	0.50
	MSSA	0.11	0.16	0.19	0.19	0.20	0.23	0.25	0.25	0.24	0.24	0.20	0.16	0.13	0.12
	BB	0.06	0.10	0.16	0.19	0.26	0.29	0.38	0.48	0.57	0.67	0.88	1.02	1.32	1.70
	IPCA	0.03	0.05	0.06	0.08	0.10	0.10	0.13	0.13	0.16	0.17	0.31	0.42	0.63	0.91
	KNN	0.05	0.06	0.08	0.10	0.11	0.10	0.14	0.15	0.18	0.16	0.20	0.19	0.21	0.26
	RF	0.00	0.02	0.03	0.04	0.04	0.04	0.05	0.06	0.09	0.09	0.09	0.17	0.24	0.24
	MICE	0.09	0.15	0.22	0.30	0.44	0.60	1.04	1.22	1.85	2.39	2.80	3.93	4.77	5.02
	MIPCA	0.03	0.04	0.05	0.07	0.08	0.08	0.09	0.10	0.13	0.14	0.22	0.21	0.23	0.45
	Amelia	0.03	0.04	0.05	0.07	0.09	0.09	0.11	0.14	0.17	0.19	0.26	0.32	0.73	6.08

A.6 Standard deviation of each risk measures among all missingness scenarios, with MCAR data in the first column

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
$VaR_{99\%}^{1day}$ (10^{-3})	**LI**	0.33	0.68	1.03	1.25	1.40	1.50	1.40	1.47	1.57	1.34	1.44	1.24	1.00	1.13
	LOCF	0.46	0.84	1.30	1.38	1.58	1.86	2.17	2.57	2.38	2.62	2.63	3.00	3.41	3.93
	MSSA	0.39	0.67	0.79	0.84	0.87	0.91	0.95	1.18	1.20	1.06	1.09	1.34	1.32	1.59
	BB	0.24	0.32	0.41	0.45	0.44	0.49	0.70	0.78	0.90	1.10	1.32	1.53	1.92	2.46
	IPCA	0.19	0.23	0.41	0.43	0.60	0.51	0.91	1.21	1.58	1.89	2.37	3.03	3.94	5.26
	KNN	0.24	0.42	0.51	0.52	0.59	0.65	1.11	1.06	1.14	1.33	1.99	2.01	2.22	2.76
	RF	0.00	0.00	0.22	0.23	0.25	0.73	0.59	0.59	0.78	0.74	0.02	1.81	2.55	2.02
	MICE	0.52	0.90	1.47	1.99	2.62	3.09	4.01	4.14	4.41	5.21	5.77	7.24	8.31	9.66
	MIPCA	0.14	0.23	0.32	0.34	0.47	0.49	0.65	0.63	0.88	0.81	1.18	1.58	2.11	3.51
	Amelia	0.13	0.21	0.30	0.36	0.45	0.45	0.62	0.57	0.67	0.65	0.78	0.98	1.54	4.26
$ES_{99\%}^{1day}$ (10^{-3})	**LI**	0.37	0.49	0.64	0.72	0.71	0.88	0.94	0.92	1.13	1.01	1.06	1.10	1.00	1.09
	LOCF	0.67	0.84	1.03	1.19	1.30	1.53	1.75	2.05	2.12	2.22	2.26	2.27	2.67	3.21
	MSSA	0.57	0.79	0.83	0.88	0.86	0.84	0.81	1.04	1.06	0.94	1.00	1.14	1.04	1.32
	BB	0.38	0.38	0.42	0.46	0.44	0.50	0.69	0.78	0.89	1.08	1.30	1.54	1.92	2.45
	IPCA	0.26	0.33	0.44	0.47	0.59	0.53	0.77	1.05	1.35	1.62	2.20	2.89	3.76	5.24
	KNN	0.42	0.54	0.64	0.67	0.78	0.77	1.07	1.12	1.45	1.26	1.57	1.52	1.57	2.00
	RF	0.00	0.00	0.37	0.39	0.43	0.60	0.44	0.43	0.66	0.70	0.34	1.39	1.76	1.86
	MICE	0.76	1.07	1.53	1.97	2.57	3.09	4.08	4.28	4.60	5.44	6.00	7.32	8.17	9.68
	MIPCA	0.22	0.31	0.41	0.46	0.59	0.58	0.69	0.73	1.00	1.02	1.35	1.70	2.15	3.40
	Amelia	0.23	0.31	0.41	0.46	0.55	0.53	0.63	0.61	0.76	0.74	0.83	1.06	1.56	4.29

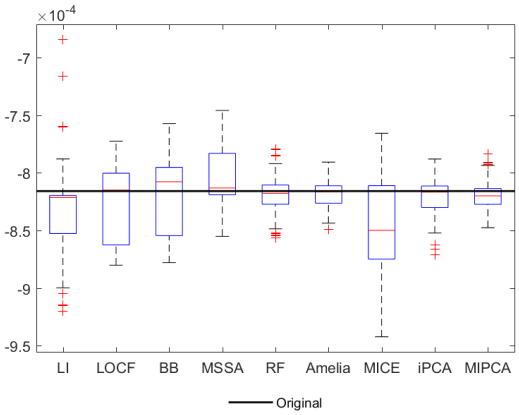
		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
$VaR^{10days}_{97.5\%}$ (10^{-3})	**LI**	0.89	1.14	1.35	1.73	2.11	2.32	2.29	2.47	2.74	2.85	3.07	3.23	3.07	4.02
	LOCF	1.01	1.26	1.56	1.94	2.44	2.96	3.20	3.84	4.49	5.02	4.90	5.42	4.77	6.22
	MSSA	1.66	2.07	2.12	2.43	2.53	2.53	2.42	2.56	2.47	2.62	2.52	2.38	2.53	2.53
	BB	0.91	1.00	1.09	1.27	1.43	1.44	1.54	1.64	1.80	1.67	2.09	2.30	2.25	2.82
	IPCA	0.45	0.64	0.83	0.97	1.24	1.24	1.64	2.04	2.70	2.92	3.51	3.59	4.76	5.96
	KNN	1.16	1.43	1.57	1.70	1.74	1.89	2.14	2.16	2.09	2.13	2.24	2.24	2.52	2.29
	RF	0.00	0.00	0.48	0.56	1.14	1.37	0.33	0.42	0.76	0.57	1.02	1.73	0.98	2.24
	MICE	1.01	1.42	1.75	1.94	2.40	2.78	2.92	3.57	3.91	4.84	5.72	7.01	7.28	8.72
	MIPCA	0.37	0.50	0.55	0.80	0.90	1.11	1.39	1.32	1.84	1.66	1.80	2.04	2.61	3.40
	Amelia	0.29	0.35	0.43	0.55	0.59	0.74	0.94	0.89	1.29	1.01	1.36	1.67	2.31	5.03
$ES^{10days}_{97.5\%}$ (10^{-3})	**LI**	0.47	0.68	0.86	1.22	1.53	1.62	1.63	1.91	2.12	2.23	2.38	2.76	2.49	3.38
	LOCF	0.74	1.14	1.48	1.74	2.15	2.57	2.57	3.15	3.42	3.96	4.43	4.49	4.14	5.39
	MSSA	0.86	1.16	1.26	1.64	1.77	1.79	1.67	1.83	2.03	2.06	2.06	1.97	2.18	2.35
	BB	0.59	0.67	0.87	1.08	1.25	1.29	1.38	1.45	1.67	1.54	1.94	2.18	2.12	2.73
	IPCA	0.24	0.32	0.40	0.50	0.76	0.74	1.09	1.32	2.00	2.22	2.91	2.89	4.17	5.34
	KNN	0.49	0.71	0.83	0.96	1.09	1.42	1.45	1.54	1.76	1.71	1.71	1.97	2.00	2.09
	RF	0.00	0.13	0.46	0.45	0.54	0.66	0.47	0.44	0.71	0.67	0.65	1.67	0.89	1.28
	MICE	0.75	1.14	1.58	1.73	2.33	2.52	2.74	3.38	3.65	4.50	5.48	6.66	6.90	8.37
	MIPCA	0.18	0.24	0.28	0.41	0.50	0.73	0.93	0.98	1.48	1.31	1.48	1.76	2.23	3.14
	Amelia	0.19	0.23	0.30	0.37	0.44	0.60	0.77	0.76	1.15	0.87	1.20	1.47	2.07	5.04

Appendix B: MCAR data in the whole sample

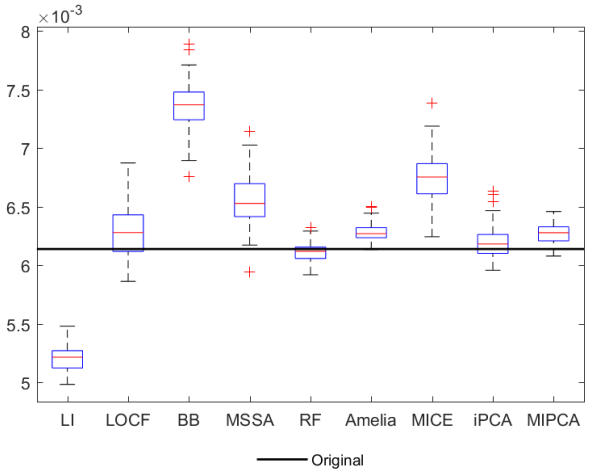
B.1 Number of MCAR tests that are calculable, when applied to return matrices containing MCAR data on almost the whole data matrix, for a 5% significance level

	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Little's test	100	100	100	100	100	100	100	100	100	100	100	100	100	100
J&J's test	100	100	97	6	0	0	0	4	80	100	100	100	100	100

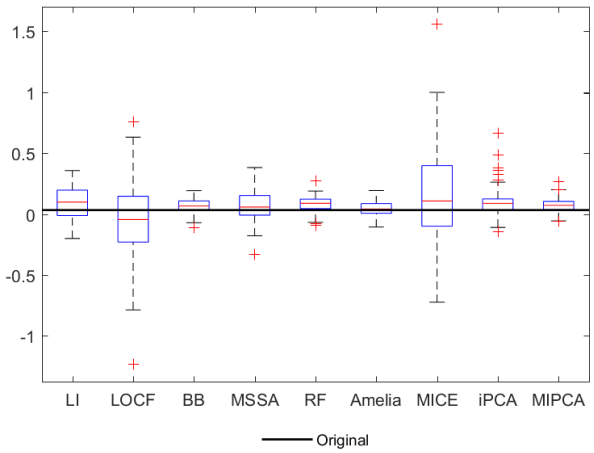
B.2 Distribution of the first four statistical moments obtained for the 100 scenarios with 30% of MCAR data in the whole sample



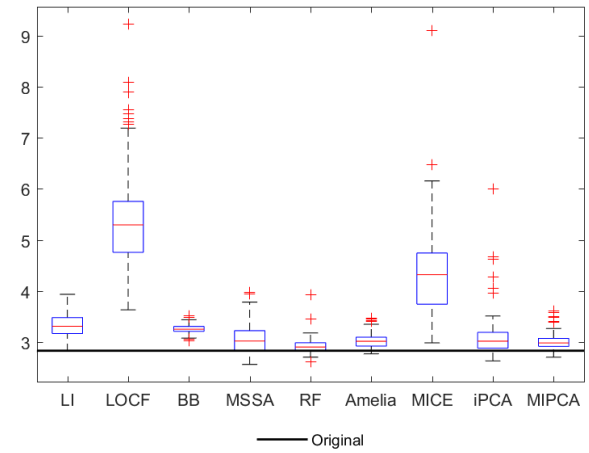
(a) Mean



(b) Standard Deviation



(c) Skewness



(d) Kurtosis

B.3 Standard deviation of each statistical moment among all missingness scenarios, with MCAR data in the whole matrix

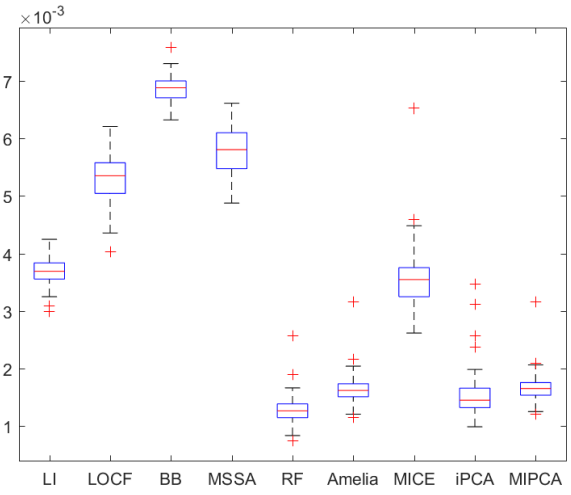
		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Mean (10^{-5})	**LI**	0.84	0.84	1.54	1.94	2.32	3.59	4.28	4.22	5.51	5.66	5.92	7.10	7.49	6.69
	LOCF	1.21	1.26	1.73	2.15	2.34	2.99	2.52	2.78	3.22	3.11	3.22	3.25	3.17	3.18
	MSSA	0.75	0.83	1.14	1.54	1.80	2.54	2.42	2.87	3.41	3.70	4.11	4.13	4.71	4.87
	BB	1.29	1.34	1.81	2.26	2.51	3.12	2.53	2.90	3.40	3.26	3.36	3.44	3.33	3.40
	IPCA	0.01	0.02	0.03	0.05	0.05	1.64	0.91	0.10	1.71	1.93	1.99	0.89	1.76	3.31
	KNN	0.20	0.22	-	-	-	-	-	-	-	-	-	-	-	-
	RF	0.00	0.01	0.02	0.03	0.04	1.63	0.64	0.08	1.10	1.53	2.63	1.46	2.33	3.31
	MICE	0.02	0.03	0.05	0.08	0.12	3.81	2.70	0.31	2.65	3.03	4.74	2.32	3.59	5.01
	MIPCA	0.00	0.01	0.02	0.03	0.04	1.31	0.61	0.09	1.09	1.45	1.94	0.81	1.60	-
	Amelia	0.00	0.01	0.02	0.03	0.04	1.29	-	-	-	-	-	-	-	-
Std (10^{-3})	**LI**	0.05	0.07	0.09	0.10	0.11	0.11	0.12	0.16	0.12	0.13	0.18	0.15	0.18	0.16
	LOCF	0.08	0.11	0.14	0.16	0.17	0.21	0.24	0.27	0.23	0.29	0.32	0.32	0.38	0.44
	MSSA	0.11	0.15	0.18	0.17	0.18	0.21	0.23	0.24	0.23	0.27	0.25	0.28	0.32	0.30
	BB	0.06	0.09	0.13	0.14	0.19	0.20	0.26	0.32	0.37	0.43	0.56	0.63	0.80	1.04
	IPCA	0.01	0.03	0.05	0.08	0.09	0.14	0.11	0.16	0.22	0.28	0.20	0.30	0.42	0.55
	KNN	0.06	0.07	-	-	-	-	-	-	-	-	-	-	-	-
	RF	0.01	0.02	0.04	0.06	0.07	0.08	0.11	0.13	0.19	0.20	0.31	0.36	0.51	0.64
	MICE	0.03	0.06	0.09	0.12	0.18	0.20	0.38	0.41	0.79	1.03	1.24	1.59	1.72	1.89
	MIPCA	0.01	0.02	0.04	0.05	0.06	0.08	0.12	0.13	0.22	0.22	0.30	0.38	0.57	-
	Amelia	0.01	0.02	0.03	0.05	0.06	0.07	-	-	-	-	-	-	-	-

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Skewness	**LI**	0.04	0.06	0.08	0.10	0.11	0.13	0.16	0.17	0.20	0.24	0.27	0.27	0.32	0.39
	LOCF	0.09	0.14	0.17	0.21	0.25	0.32	0.35	0.40	0.51	0.55	0.54	0.58	0.64	0.75
	MSSA	0.07	0.11	0.12	0.13	0.13	0.12	0.13	0.14	0.18	0.16	0.16	0.19	0.22	0.27
	BB	0.05	0.06	0.06	0.07	0.06	0.06	0.07	0.06	0.06	0.05	0.05	0.05	0.04	0.06
	IPCA	0.01	0.01	0.04	0.05	0.06	0.11	0.11	0.13	0.16	0.34	0.19	0.52	0.56	1.04
	KNN	0.04	0.06	-	-	-	-	-	-	-	-	-	-	-	-
	RF	0.00	0.01	0.03	0.04	0.04	0.06	0.06	0.09	0.13	0.23	0.19	0.45	0.47	0.51
	MICE	0.02	0.04	0.08	0.18	0.22	0.38	0.45	0.50	0.68	0.97	1.00	1.27	1.36	1.44
	MIPCA	0.00	0.01	0.02	0.02	0.03	0.06	0.06	0.08	0.09	0.21	0.14	0.34	0.39	-
	Amelia	0.00	0.01	0.02	0.03	0.03	0.06	-	-	-	-	-	-	-	-
Kurtosis	**LI**	0.07	0.10	0.14	0.18	0.20	0.24	0.33	0.31	0.42	0.48	0.57	0.72	0.81	1.18
	LOCF	0.28	0.43	0.50	0.64	0.84	1.07	1.20	1.34	1.80	2.84	2.45	2.09	2.72	3.61
	MSSA	0.15	0.19	0.22	0.26	0.30	0.30	0.36	0.49	0.50	0.66	0.72	0.98	1.26	1.49
	BB	0.12	0.10	0.10	0.10	0.09	0.09	0.08	0.06	0.06	0.06	0.05	0.05	0.05	0.15
	IPCA	0.01	0.04	0.11	0.15	0.18	0.47	0.36	0.46	0.52	1.36	0.53	1.00	1.00	2.09
	KNN	0.08	0.16	-	-	-	-	-	-	-	-	-	-	-	-
	RF	0.00	0.03	0.06	0.06	0.10	0.15	0.12	0.21	0.28	0.64	0.53	1.26	1.16	1.02
	MICE	0.02	0.06	0.17	0.35	0.39	0.84	0.81	0.77	1.23	1.85	1.91	1.82	1.91	1.71
	MIPCA	0.01	0.02	0.05	0.05	0.07	0.16	0.16	0.22	0.26	0.54	0.30	0.82	0.63	-
	Amelia	0.01	0.02	0.04	0.05	0.08	0.14	-	-	-	-	-	-	-	-

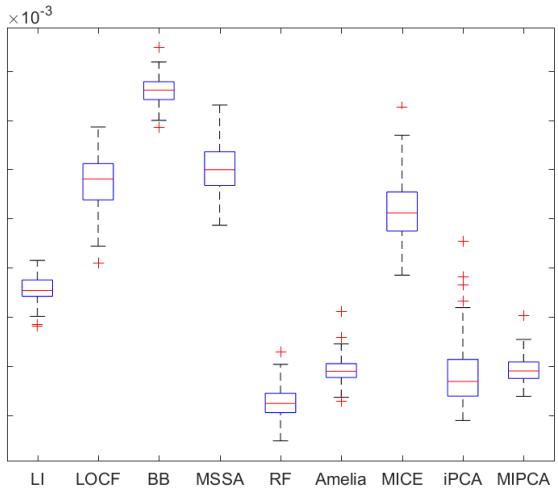
B.4 Average number of principal components used by IPCA and MIPCA, among the 100 missingness scenarios, for imputation of MCAR data in the whole sample

	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
IPCA	4	4	3.94	3.97	2.97	2.65	2.93	1.99	1.86	1.55	1.03	1	1	3.26
MIPCA	4	3.99	3.49	3.08	2.94	2.46	2.03	2	1.97	1.75	1.03	1	1	-

B.5 Distribution of proximity measures obtained for the 100 scenarios with 30% of MCAR data in the whole sample



(a) MAE

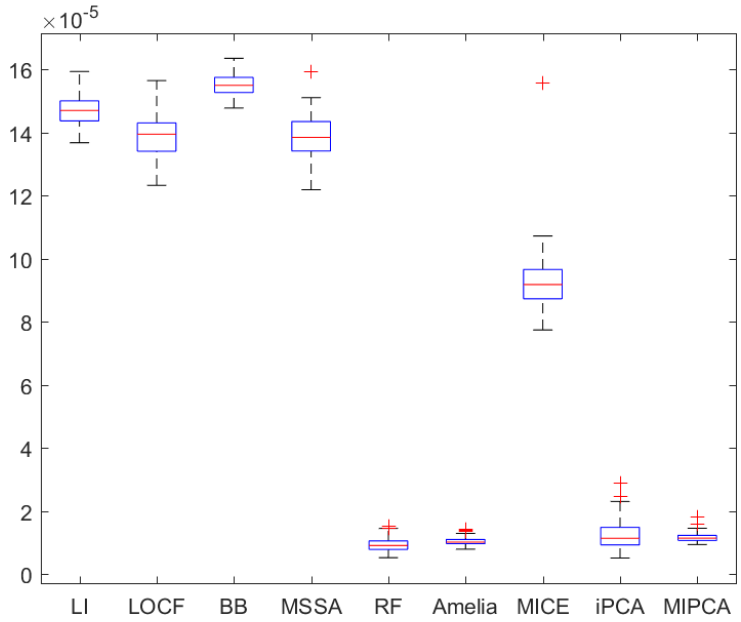


(b) RMSE

B.6 Standard deviation of the proximity measures among all missingness scenarios, with MCAR data in the whole matrix

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
MAE (10^{-3})	**LI**	0.74	0.49	0.41	0.32	0.28	0.23	0.25	0.23	0.19	0.16	0.17	0.15	0.15	0.12
	LOCF	1.04	0.77	0.65	0.54	0.45	0.43	0.34	0.33	0.29	0.30	0.31	0.28	0.27	0.22
	MSSA	0.97	0.72	0.60	0.49	0.42	0.38	0.34	0.33	0.27	0.27	0.26	0.22	0.19	0.20
	BB	0.36	0.21	0.19	0.16	0.17	0.21	0.23	0.24	0.27	0.35	0.38	0.44	0.55	0.75
	IPCA	0.31	0.32	0.38	0.33	0.32	0.36	0.27	0.23	0.36	0.28	0.22	0.20	0.23	0.73
	KNN	0.49	0.35	-	-	-	-	-	-	-	-	-	-	-	-
	RF	0.20	0.25	0.27	0.24	0.26	0.22	0.23	0.22	0.26	0.22	0.26	0.25	0.30	0.43
	MICE	0.75	0.71	0.69	0.64	0.56	0.49	0.56	0.48	0.77	0.84	0.95	1.08	1.13	1.16
	MIPCA	0.26	0.28	0.28	0.25	0.24	0.23	0.25	0.21	0.28	0.22	0.21	0.24	0.34	-
	Amelia	0.27	0.29	0.28	0.26	0.25	0.23	-	-	-	-	-	-	-	-
RMSE (10^{-3})	**LI**	0.77	0.53	0.43	0.35	0.30	0.26	0.28	0.26	0.20	0.17	0.19	0.17	0.16	0.13
	LOCF	1.22	0.90	0.74	0.62	0.51	0.55	0.46	0.44	0.43	0.45	0.44	0.40	0.41	0.42
	MSSA	1.09	0.84	0.70	0.57	0.48	0.46	0.41	0.39	0.34	0.34	0.33	0.33	0.27	0.29
	BB	0.43	0.25	0.23	0.20	0.21	0.27	0.29	0.31	0.34	0.44	0.49	0.56	0.71	0.96
	IPCA	0.58	0.66	0.70	0.60	0.58	0.65	0.42	0.36	0.57	0.45	0.31	0.27	0.36	1.08
	KNN	0.53	0.41	-	-	-	-	-	-	-	-	-	-	-	-
	RF	0.43	0.49	0.42	0.35	0.33	0.30	0.31	0.29	0.37	0.33	0.41	0.38	0.46	0.65
	MICE	1.44	1.00	0.86	0.73	0.64	0.60	0.74	0.67	1.13	1.32	1.37	1.75	1.78	1.90
	MIPCA	0.48	0.46	0.38	0.32	0.28	0.28	0.29	0.25	0.35	0.31	0.29	0.32	0.48	-
	Amelia	0.51	0.47	0.38	0.32	0.29	0.26	-	-	-	-	-	-	-	-

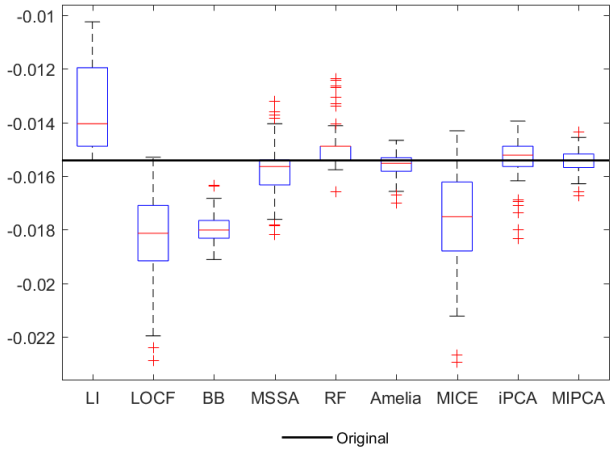
B.7 Distribution of covariance differences,according to the Froebinus norm, obtained for the 100 scenarios with 30% of MCAR data in the whole sample



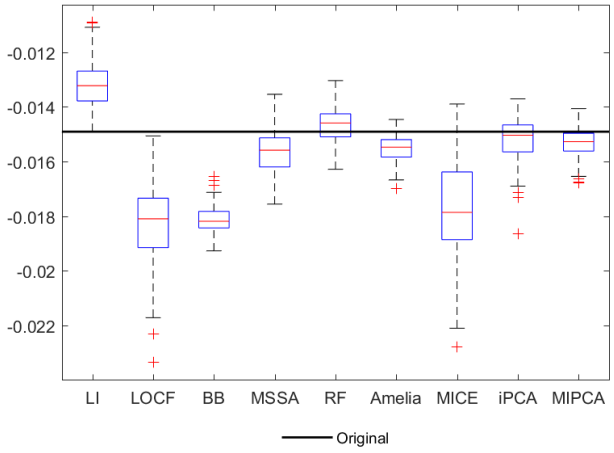
B.8 Standard deviation of the covariance matrix differences among all missingness scenarios, with MCAR data in the whole matrix

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Cov (10^{-4})	**LI**	0.38	0.47	0.47	0.49	0.48	0.47	0.43	0.43	0.44	0.41	0.39	0.38	0.30	0.30
	LOCF	0.46	0.54	0.54	0.66	0.63	0.67	0.77	0.73	0.66	0.62	0.64	0.68	0.57	0.62
	MSSA	0.51	0.57	0.65	0.70	0.64	0.67	0.66	0.63	0.64	0.64	0.54	0.52	0.43	0.43
	BB	0.33	0.39	0.39	0.38	0.38	0.34	0.31	0.30	0.30	0.30	0.35	0.43	0.59	0.82
	IPCA	0.05	0.08	0.17	0.23	0.33	0.44	0.31	0.47	0.72	1.04	0.69	0.67	1.07	7.26
	KNN	0.11	0.14	-	-	-	-	-	-	-	-	-	-	-	-
	RF	0.06	0.07	0.08	0.09	0.13	0.21	0.18	0.28	0.34	0.39	0.49	0.55	0.95	1.41
	MICE	0.30	0.58	0.70	0.81	0.85	0.90	0.70	0.83	0.92	0.98	1.27	1.77	3.38	4.13
	MIPCA	0.04	0.05	0.06	0.07	0.10	0.14	0.15	0.19	0.28	0.46	0.49	0.63	0.95	-
	Amelia	0.04	0.05	0.06	0.08	0.10	0.11	-	-	-	-	-	-	-	-

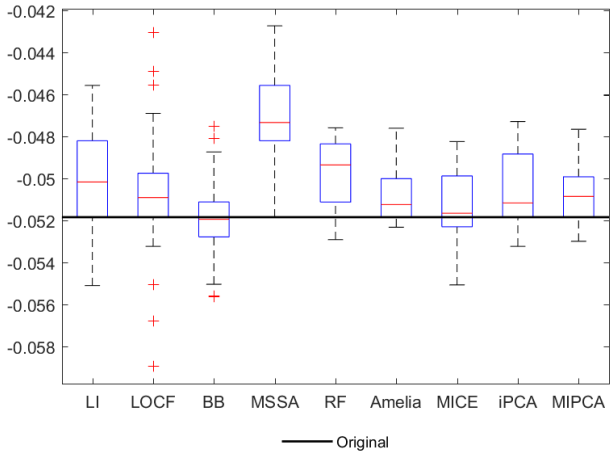
B.9 Distribution of the 1-day risk measures, and of 10-day risk measures computed from the 100 scenarios containing 30% of MCAR data in the whole sample



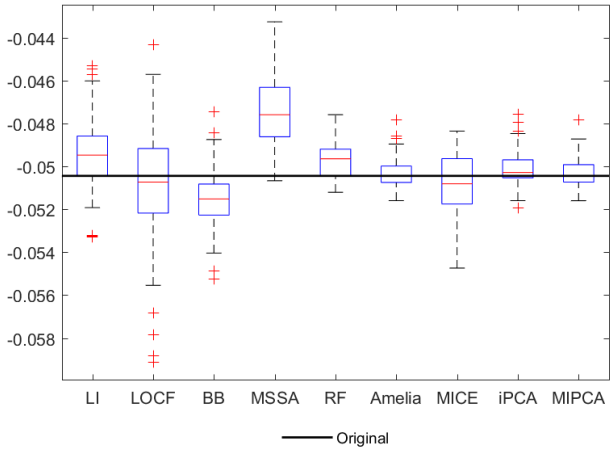
(a) 1-day VaR at 99%



(b) 1-day ES at 97.5%



(c) 10-day VaR at 99%



(d) 10-day ES at 97.5%

B.10 Standard deviation of each risk measures among all missingness scenarios, with MCAR data in the whole matrix

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
$VaR_{99\%}^{1day}$ (10^{-3})	**LI**	0.33	0.68	1.03	1.25	1.40	1.50	1.40	1.47	1.57	1.34	1.44	1.24	1.00	1.13
	LOCF	0.46	0.84	1.30	1.38	1.58	1.86	2.17	2.57	2.38	2.62	2.63	3.00	3.41	3.93
	MSSA	0.39	0.64	0.79	0.85	0.87	0.94	0.98	1.19	1.18	1.14	1.15	1.32	1.26	1.52
	BB	0.24	0.33	0.40	0.46	0.48	0.51	0.70	0.81	0.91	1.07	1.27	1.54	1.93	2.48
	IPCA	0.00	0.00	0.19	0.30	0.30	0.74	0.59	0.98	1.22	1.34	0.30	3.52	3.32	4.50
	KNN	0.23	0.45	-	-	-	-	-	-	-	-	-	-	-	-
	RF	0.00	0.00	0.22	0.23	0.35	0.71	0.50	0.82	1.24	1.25	0.29	2.90	3.00	3.53
	MICE	0.04	0.16	0.27	0.62	0.90	1.72	1.76	2.38	3.07	3.78	4.01	6.22	6.82	7.24
	MIPCA	0.00	0.01	0.17	0.20	0.24	0.42	0.33	0.62	1.08	1.14	0.69	1.88	2.30	-
	Amelia	0.00	0.01	0.16	0.17	0.22	0.44	-	-	-	-	-	-	-	-
$ES_{99\%}^{1day}$ (10^{-3})	**LI**	0.37	0.49	0.64	0.72	0.71	0.88	0.94	0.92	1.13	1.01	1.06	1.10	1.00	1.09
	LOCF	0.67	0.84	1.03	1.19	1.30	1.53	1.75	2.05	2.12	2.22	2.26	2.27	2.67	3.21
	MSSA	0.57	0.78	0.83	0.88	0.85	0.86	0.81	1.04	1.04	1.04	1.09	1.18	1.05	1.38
	BB	0.38	0.38	0.41	0.46	0.48	0.50	0.69	0.81	0.92	1.06	1.27	1.54	1.94	2.47
	IPCA	0.02	0.08	0.29	0.38	0.36	0.83	0.55	0.69	1.25	1.33	0.63	2.63	2.61	4.47
	KNN	0.42	0.54	-	-	-	-	-	-	-	-	-	-	-	-
	RF	0.01	0.02	0.40	0.42	0.54	0.67	0.50	0.61	1.08	1.28	0.70	2.31	2.71	3.56
	MICE	0.15	0.32	0.52	0.83	1.06	1.75	1.96	2.37	3.12	3.77	4.43	6.64	7.04	7.42
	MIPCA	0.04	0.06	0.24	0.24	0.30	0.53	0.44	0.59	1.11	1.12	0.82	1.71	2.25	-
	Amelia	0.04	0.06	0.22	0.26	0.35	0.50	-	-	-	-	-	-	-	-

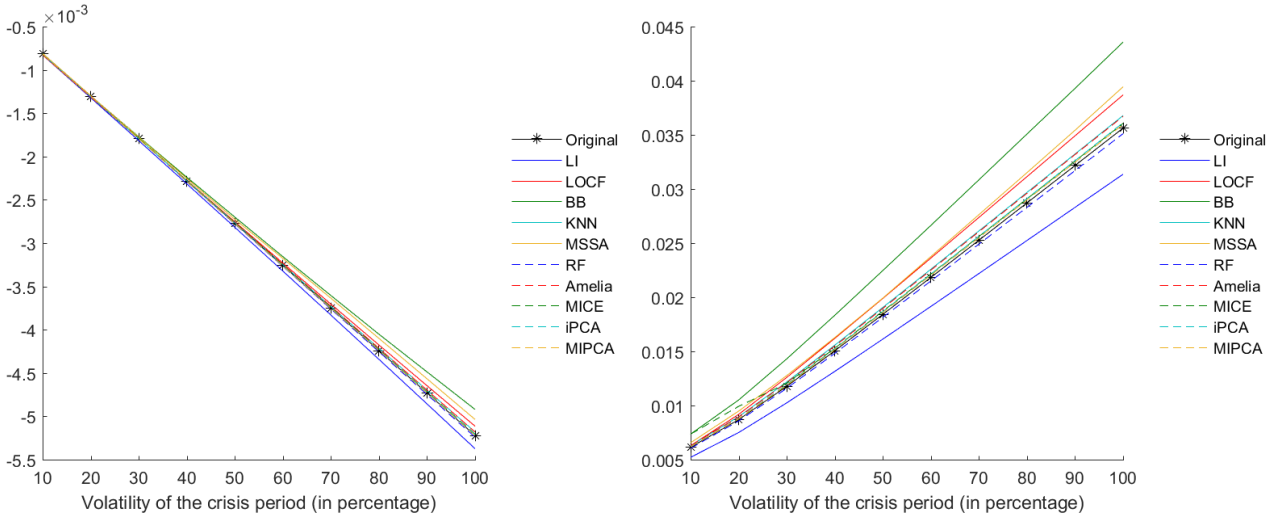
		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
$VaR^{10days}_{97.5\%}$ (10^{-3})	**LI**	0.89	1.14	1.35	1.73	2.11	2.32	2.29	2.47	2.74	2.85	3.07	3.23	3.07	4.02
	LOCF	1.01	1.26	1.56	1.94	2.44	2.96	3.20	3.84	4.49	5.02	4.90	5.42	4.77	6.22
	MSSA	1.66	2.07	2.12	2.42	2.54	2.53	2.45	2.57	2.38	2.54	2.32	2.29	2.44	2.50
	BB	0.86	0.98	1.10	1.31	1.48	1.42	1.55	1.63	1.86	1.72	2.09	2.28	2.20	2.74
	IPCA	0.00	0.00	0.91	0.70	1.61	1.62	2.46	1.06	1.55	1.75	1.75	2.88	1.40	2.31
	KNN	1.19	1.42	-	-	-	-	-	-	-	-	-	-	-	-
	RF	0.00	0.00	0.74	0.81	1.65	1.47	1.52	1.22	1.51	1.53	1.90	3.73	2.00	1.92
	MICE	0.00	0.00	1.57	1.64	1.72	1.63	2.85	1.57	1.91	2.72	2.23	9.64	6.21	3.96
	MIPCA	0.00	0.00	0.44	0.34	1.36	1.11	1.88	1.00	1.33	1.13	1.02	2.25	1.77	-
	Amelia	0.00	0.00	0.40	0.39	1.28	1.13	-	-	-	-	-	-	-	-
$ES^{10days}_{97.5\%}$ (10^{-3})	**LI**	0.47	0.68	0.86	1.22	1.53	1.62	1.63	1.91	2.12	2.23	2.38	2.76	2.49	3.38
	LOCF	0.74	1.14	1.48	1.74	2.15	2.57	2.57	3.15	3.42	3.96	4.43	4.49	4.14	5.39
	MSSA	0.86	1.16	1.27	1.64	1.77	1.77	1.68	1.84	2.01	2.03	2.00	1.97	2.15	2.34
	BB	0.55	0.65	0.87	1.08	1.30	1.29	1.41	1.44	1.70	1.58	1.92	2.15	2.09	2.67
	IPCA	0.00	0.08	0.55	0.52	0.63	0.76	1.76	1.24	1.12	1.17	0.77	2.58	1.59	2.14
	KNN	0.50	0.74	-	-	-	-	-	-	-	-	-	-	-	-
	RF	0.00	0.09	0.49	0.52	0.79	0.81	1.18	1.02	1.16	1.20	1.32	3.21	2.05	2.03
	MICE	0.00	0.16	0.87	0.90	1.09	1.48	2.84	1.43	1.84	2.50	2.43	7.79	5.98	3.98
	MIPCA	0.00	0.08	0.31	0.29	0.57	0.63	1.46	1.15	1.16	0.99	0.97	1.97	1.78	-
	Amelia	0.00	0.09	0.30	0.32	0.59	0.67	-	-	-	-	-	-	-	-

Appendix C: Impact of heteroskedasticity

C.1 Number of MCAR tests that are calculable, when applied to heteroskedastic return matrices containing MCAR data on the first column, for a 5% significance level

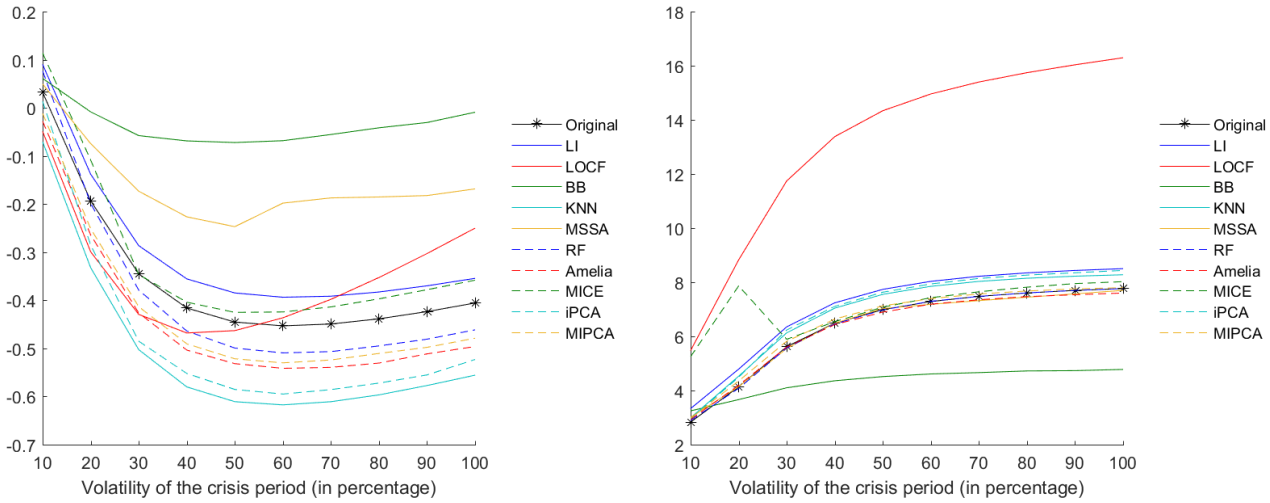
	Volatility of the crisis period									
	10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Little's test	100	100	100	100	100	100	100	100	100	100
J&J's test	100	100	100	100	100	100	100	100	100	100

C.2 Average of the first four statistical moments of the returns of the imputed data based on a matrix containing MCAR data according to the volatility of the crisis period



(a) Averaged mean

(b) Averaged standard deviation



(c) Averaged skewness

(d) Averaged kurtosis

C.3 Standard deviation of each statistical moment among all missingness scenarios, with MCAR data in the first column of an heteroskedastic sample

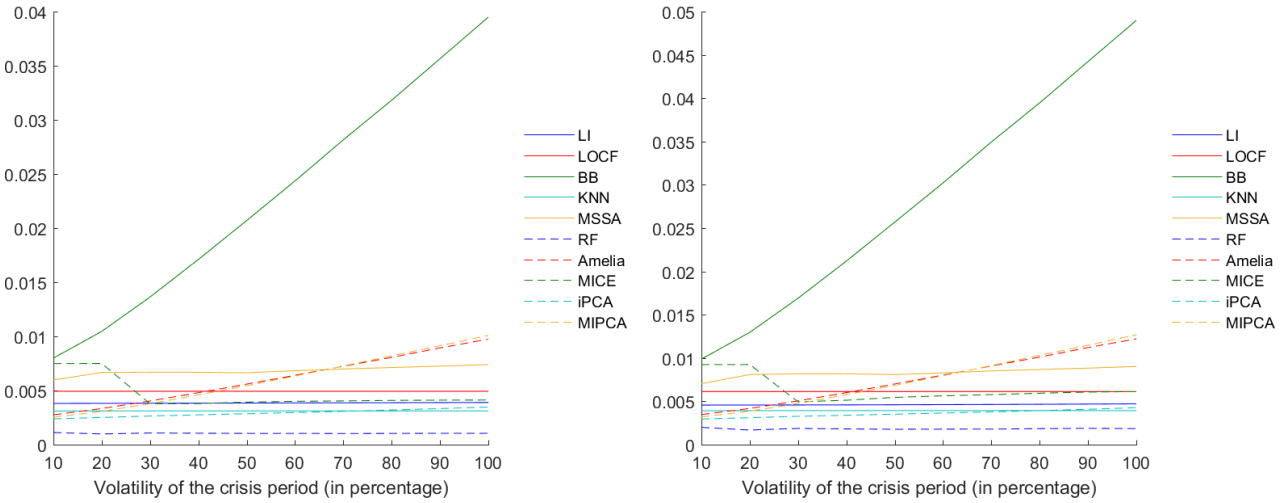
		Volatility of the crisis period									
		10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Mean (10^{-5})	**LI**	3.59	3.60	3.62	3.66	3.74	3.86	4.05	4.30	4.63	5.05
	LOCF	2.99	3.06	3.20	3.46	3.87	4.48	5.28	6.27	7.46	8.84
	MSSA	2.52	4.32	4.63	4.87	5.10	6.29	7.48	8.27	8.64	8.66
	BB	3.11	3.10	3.19	3.57	3.93	4.86	6.09	7.14	9.12	10.76
	IPCA	1.05	1.26	1.45	1.58	1.80	2.10	2.54	3.06	3.68	4.49
	KNN	0.67	0.68	0.73	0.87	1.11	1.43	1.85	2.34	2.85	3.51
	RF	1.00	1.11	1.03	1.14	1.20	1.26	1.46	1.79	2.24	2.56
	MICE	3.21	3.28	1.08	1.01	1.24	1.54	1.88	2.32	2.68	3.21
	MIPCA	0.77	0.89	1.06	1.24	1.44	1.71	2.03	2.49	3.01	3.49
	Amelia	0.70	0.75	0.85	0.98	1.30	1.57	1.85	2.29	2.99	3.52
Std (10^{-3})	**LI**	0.11	0.17	0.28	0.39	0.50	0.61	0.72	0.82	0.93	1.03
	LOCF	0.21	0.36	0.56	0.77	0.98	1.20	1.42	1.66	1.89	2.14
	MSSA	0.19	0.32	0.48	0.66	0.84	1.01	1.19	1.38	1.58	1.79
	BB	0.20	0.39	0.66	0.92	1.19	1.43	1.73	1.92	2.19	2.46
	IPCA	0.11	0.18	0.32	0.40	0.51	0.63	0.75	0.86	0.97	1.11
	KNN	0.10	0.17	0.26	0.36	0.46	0.55	0.64	0.72	0.79	0.89
	RF	0.06	0.11	0.17	0.24	0.31	0.38	0.45	0.51	0.58	0.63
	MICE	0.41	0.46	0.47	0.32	0.41	0.49	0.58	0.67	0.76	0.84
	MIPCA	0.09	0.15	0.22	0.30	0.38	0.47	0.54	0.65	0.73	0.80
	Amelia	0.09	0.14	0.22	0.32	0.40	0.51	0.58	0.68	0.78	0.85

		Missingness proportion									
		10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Skewness	**LI**	0.13	0.19	0.25	0.28	0.29	0.30	0.31	0.31	0.32	0.32
	LOCF	0.32	0.57	0.74	0.83	0.89	0.93	0.97	1.00	1.03	1.06
	MSSA	0.11	0.16	0.22	0.26	0.28	0.29	0.29	0.29	0.30	0.30
	BB	0.06	0.07	0.08	0.08	0.08	0.09	0.09	0.09	0.09	0.09
	IPCA	0.08	0.15	0.23	0.24	0.26	0.26	0.27	0.28	0.28	0.30
	KNN	0.10	0.12	0.15	0.17	0.17	0.17	0.17	0.17	0.17	0.17
	RF	0.05	0.07	0.09	0.10	0.11	0.11	0.11	0.11	0.11	0.11
	MICE	0.61	0.68	0.19	0.18	0.19	0.20	0.21	0.22	0.22	0.22
	MIPCA	0.06	0.09	0.11	0.13	0.14	0.15	0.15	0.15	0.15	0.15
	Amelia	0.06	0.08	0.11	0.12	0.13	0.13	0.14	0.14	0.14	0.14
Kurtosis	**LI**	0.24	0.52	0.74	0.86	0.93	0.97	1.00	1.01	1.03	1.04
	LOCF	1.07	2.78	3.91	4.55	5.02	5.41	5.78	6.13	6.49	6.86
	MSSA	0.24	0.44	0.68	0.81	0.89	0.96	0.98	0.99	1.01	1.04
	BB	0.08	0.13	0.20	0.26	0.29	0.30	0.32	0.33	0.34	0.35
	IPCA	0.15	0.48	1.00	0.83	0.89	0.93	0.96	0.99	1.01	1.06
	KNN	0.27	0.49	0.67	0.73	0.75	0.74	0.73	0.71	0.69	0.67
	RF	0.08	0.18	0.25	0.30	0.33	0.34	0.35	0.35	0.34	0.34
	MICE	1.17	2.63	1.50	0.76	0.87	0.93	0.97	1.04	1.08	1.12
	MIPCA	0.14	0.34	0.44	0.48	0.53	0.55	0.55	0.56	0.56	0.57
	Amelia	0.12	0.24	0.36	0.41	0.44	0.47	0.47	0.47	0.47	0.46

C.4 Average number of principal components used by IPCA and MIPCA, among the 100 missingness scenarios, for imputation of MCAR data in the first column of a sample with heteroskedasticity

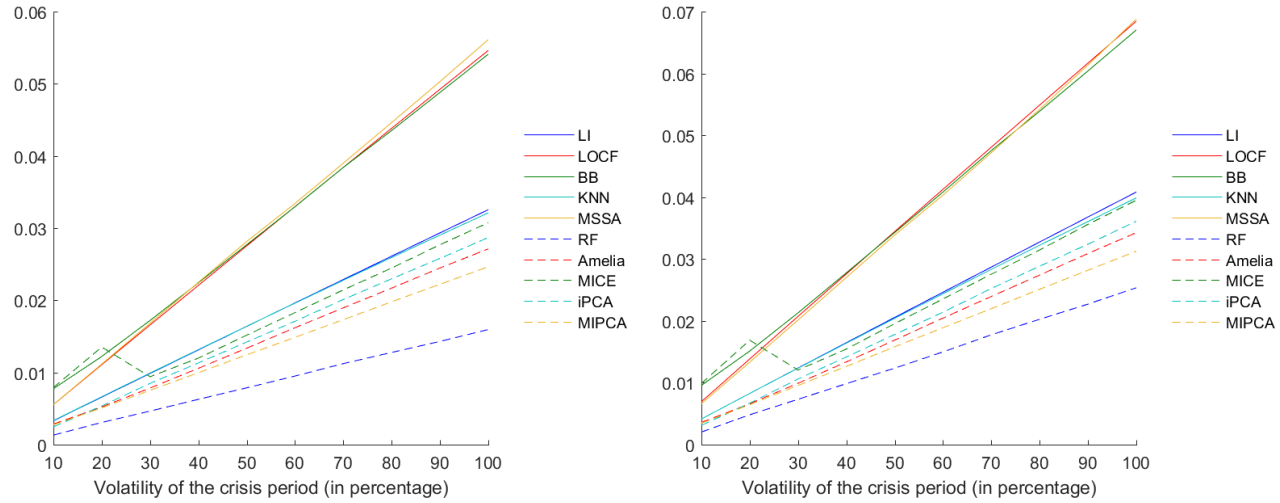
	Volatility of the crisis period									
	10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
IPCA	4	3.99	3.98	3.98	3.98	3.98	4	4	4	4
MIPCA	4	3.99	3.97	3.97	3.96	3.96	3.96	3.95	3.95	3.95

C.5 Average MAE and RMSE between the original returns and the returns of the imputed data based on a matrix containing MCAR data, for each periods, according to the volatility of the crisis period



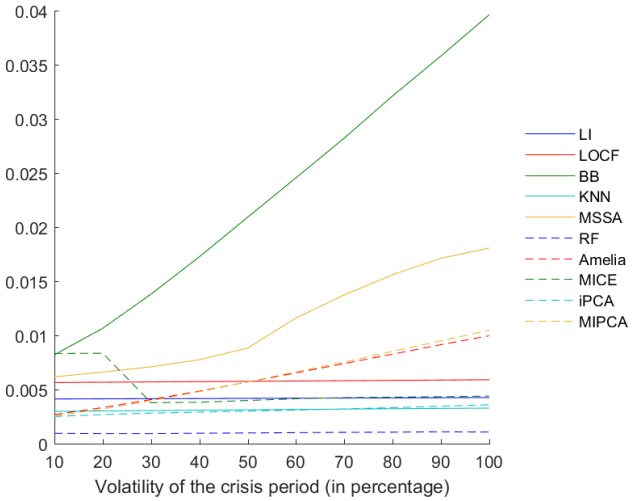
(a) Averaged MAE of the first period

(b) Averaged RMSE of the first period

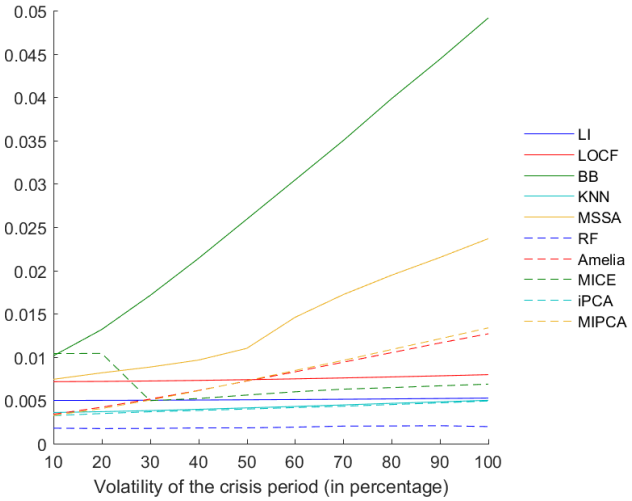


(c) Averaged MAE of the second period

(d) Averaged RMSE of the second period



(e) Averaged MAE of the third period



(f) Averaged RMSE of the third period

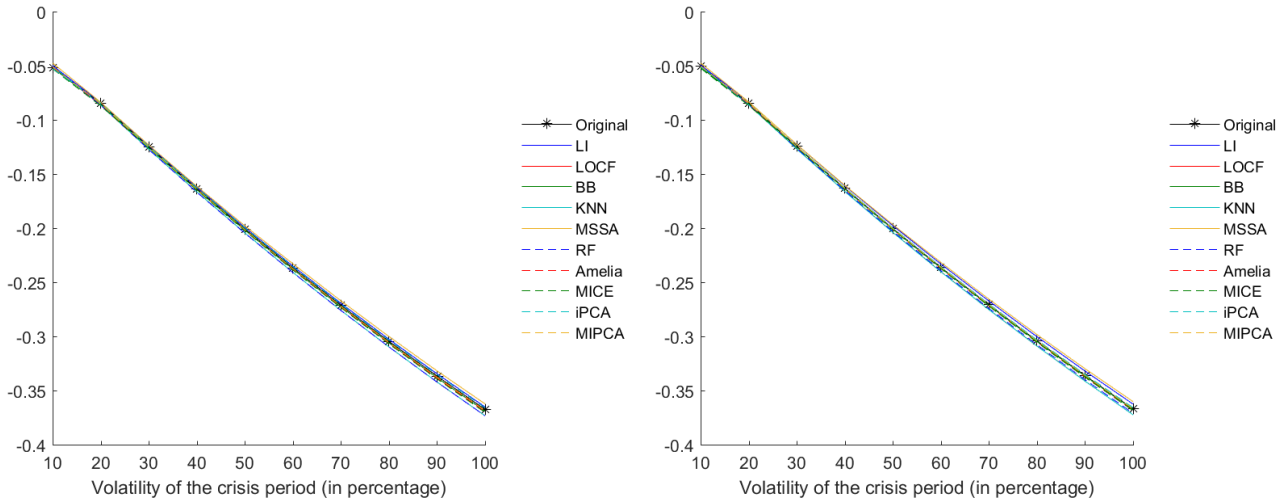
C.6 Standard deviation of the proximity measures among all missingness scenarios, with MCAR data in the first column of an heteroskedastic sample

		Volatility of the crisis period									
		10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
MAE (10^{-3})	**LI**	0.23	0.33	0.48	0.64	0.81	0.98	1.15	1.33	1.50	1.67
	LOCF	0.43	0.67	0.97	1.29	1.62	1.95	2.28	2.61	2.94	3.28
	MSSA	0.38	0.60	0.83	1.08	1.33	1.60	1.88	2.18	2.50	2.86
	BB	0.21	0.39	0.67	0.90	1.18	1.43	1.73	1.91	2.15	2.44
	IPCA	0.34	0.65	1.11	1.29	1.56	1.84	2.21	2.50	2.79	3.12
	KNN	0.18	0.28	0.40	0.54	0.67	0.81	0.95	1.08	1.20	1.33
	RF	0.17	0.28	0.43	0.56	0.70	0.84	0.99	1.12	1.26	1.38
	MICE	0.58	0.75	1.51	0.54	0.69	0.85	1.02	1.19	1.36	1.44
	MIPCA	0.11	0.24	0.31	0.36	0.49	0.56	0.67	0.80	0.90	1.01
	Amelia	0.08	0.12	0.18	0.27	0.35	0.46	0.53	0.63	0.77	0.82
RMSE (10^{-3})	**LI**	0.26	0.42	0.70	1.00	1.29	1.57	1.85	2.13	2.40	2.67
	LOCF	0.55	1.04	1.66	2.30	2.93	3.56	4.19	4.81	5.45	6.08
	MSSA	0.45	0.73	1.13	1.57	2.00	2.38	2.80	3.25	3.73	4.24
	BB	0.27	0.50	0.83	1.10	1.46	1.74	2.09	2.32	2.59	2.92
	IPCA	0.44	0.92	1.78	2.05	2.55	3.07	3.76	4.31	4.86	5.49
	KNN	0.25	0.46	0.73	1.01	1.27	1.53	1.77	2.01	2.23	2.46
	RF	0.21	0.48	0.78	1.07	1.37	1.67	1.94	2.25	2.53	2.80
	MICE	0.76	1.11	2.10	0.95	1.22	1.50	1.84	2.15	2.46	2.67
	MIPCA	0.17	0.42	0.63	0.78	1.03	1.23	1.46	1.69	1.90	2.11
	Amelia	0.12	0.21	0.35	0.52	0.68	0.89	1.06	1.24	1.50	1.61

C.7 Standard deviation of the covariance matrix differences among all missingness scenarios, with MCAR data in the first column of an heteroskedastic sample

		Volatility of the crisis period									
		10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Cov (10^{-4})	**LI**	0.09	0.20	0.44	0.77	1.19	1.71	2.32	3.02	3.81	4.68
	LOCF	0.20	0.56	1.24	2.22	3.50	5.11	7.04	9.34	12.01	15.1
	MSSA	0.23	0.55	1.10	1.91	2.97	4.31	5.96	7.89	10.18	12.97
	BB	0.29	0.80	1.83	3.24	5.10	7.29	10.2	12.86	16.47	20.47
	IPCA	0.10	0.23	0.56	0.86	1.35	1.96	2.71	3.53	4.45	5.69
	KNN	0.10	0.24	0.51	0.90	1.37	1.94	2.60	3.33	4.06	5.01
	RF	0.04	0.09	0.17	0.29	0.46	0.66	0.91	1.18	1.48	1.76
	MICE	0.59	0.89	1.03	0.68	1.09	1.50	2.02	2.60	3.33	4.02
	MIPCA	0.08	0.17	0.33	0.57	0.88	1.28	1.69	2.38	2.97	3.65
	Amelia	0.09	0.20	0.41	0.74	1.15	1.73	2.28	3.01	3.90	4.61

C.8 Average 10-day risk measures, computed from a data matrix containing MCAR data in the whole sample, according to the volatility of the crisis period



(a) 10-day VaR at 99%

(b) 10-day ES at 97.5%

C.9 Standard deviation of each risk measures among all missingness scenarios, with MCAR data in the first column of an heteroskedastic sample

		Volatility of the crisis period									
		10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
$VaR_{99\%}^{1day}$ (10^{-3})	**LI**	1.50	3.39	5.04	6.64	8.21	9.73	11.2	12.65	14.05	15.4
	LOCF	1.86	2.71	3.94	5.19	6.42	7.62	8.80	9.94	11.05	12.13
	MSSA	0.91	2.07	3.21	4.27	5.08	5.96	6.60	7.29	8.26	8.81
	BB	0.53	1.38	2.49	3.41	4.22	5.09	6.04	6.58	7.58	8.42
	IPCA	0.51	1.95	2.86	3.89	4.86	5.79	6.84	7.76	8.68	9.48
	KNN	0.65	1.50	2.22	2.94	3.62	4.26	4.94	5.57	6.17	6.75
	RF	0.73	2.09	3.00	3.90	4.86	5.76	6.74	7.61	8.51	9.29
	MICE	3.10	3.24	3.96	3.88	4.76	5.69	6.44	7.27	7.87	8.42
	MIPCA	0.49	1.53	2.26	3.03	3.72	4.46	5.04	5.70	6.41	6.94
	Amelia	0.46	1.57	2.38	3.11	3.79	4.49	5.17	5.78	6.44	7.06
$ES_{99\%}^{1day}$ (10^{-3})	**LI**	0.88	1.90	2.84	3.74	4.61	5.46	6.29	7.09	7.88	8.63
	LOCF	1.53	2.57	3.86	5.09	6.26	7.40	8.51	9.57	10.6	11.6
	MSSA	0.84	1.61	2.54	3.44	4.32	4.97	5.74	6.53	7.39	8.32
	BB	0.54	1.23	2.07	2.83	3.50	4.19	4.94	5.40	6.15	6.82
	IPCA	0.53	1.16	2.13	2.41	3.07	3.71	4.34	4.93	5.45	5.92
	KNN	0.77	1.58	2.38	3.12	3.83	4.52	5.18	5.82	6.39	7.03
	RF	0.60	1.10	1.58	2.12	2.68	3.21	3.72	4.23	4.72	5.15
	MICE	3.11	3.39	3.76	2.76	3.35	3.95	4.59	5.11	5.61	6.08
	MIPCA	0.58	1.01	1.48	1.95	2.47	2.97	3.35	3.90	4.36	4.77
	Amelia	0.52	1.01	1.57	2.09	2.62	3.14	3.53	4.02	4.57	4.96

		Volatility of the crisis period									
		10%	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
$VaR^{10days}_{97.5\%}$ (10^{-3})	**LI**	2.32	1.79	2.56	3.27	3.91	4.49	5.02	5.49	5.91	6.29
	LOCF	2.96	4.27	6.14	7.85	9.41	10.82	12.09	13.23	14.24	15.14
	MSSA	2.53	2.41	3.28	4.08	4.85	5.48	6.07	6.58	7.08	7.60
	BB	1.42	1.51	2.20	2.85	3.32	3.75	4.30	4.75	5.05	5.21
	IPCA	1.24	2.21	3.51	3.72	4.42	5.06	5.71	6.24	6.65	7.08
	KNN	1.89	2.61	3.74	4.78	5.72	6.57	7.33	8.01	8.58	9.09
	RF	1.37	1.17	1.89	2.43	2.90	3.27	3.65	4.00	4.34	4.63
	MICE	2.79	4.38	2.22	2.61	3.08	3.58	3.92	4.16	4.64	4.86
	MIPCA	1.12	1.25	1.85	2.10	2.60	2.99	3.37	3.65	3.94	4.10
	Amelia	0.72	1.61	2.34	3.02	3.62	4.19	4.67	5.06	5.38	5.80
$ES^{10days}_{97.5\%}$ (10^{-3})	**LI**	1.62	1.73	2.49	3.17	3.79	4.35	4.85	5.29	5.68	6.03
	LOCF	2.57	3.76	5.40	6.89	8.20	9.41	10.48	11.44	12.28	13.02
	MSSA	1.79	2.10	3.00	3.80	4.49	5.00	5.61	6.19	6.73	7.23
	BB	1.25	1.44	2.11	2.76	3.23	3.71	4.16	4.65	4.93	5.07
	IPCA	0.74	1.62	2.65	2.95	3.54	4.07	4.76	5.19	5.57	5.91
	KNN	1.42	1.98	2.84	3.61	4.30	4.92	5.48	5.94	6.33	6.65
	RF	0.67	0.69	1.09	1.42	1.66	1.97	2.24	2.46	2.63	2.79
	MICE	2.54	3.76	2.01	2.22	2.67	3.11	3.39	3.65	4.01	4.22
	MIPCA	0.74	0.99	1.49	1.82	2.26	2.57	2.87	3.16	3.37	3.59
	Amelia	0.60	1.33	1.94	2.53	3.05	3.57	4.06	4.42	4.72	4.98

Appendix D: Impact of jumps

D.1 Number of MCAR tests that are calculable, when applied to return series containing jumps and MCAR data on the first column, for a 5% significance level

	Number of jumps in the series					
	0	**2**	**5**	**8**	**9**	**11**
Little's test	100	100	100	100	100	100
J&J's test	100	100	100	100	100	100

D.2 Standard deviation of each statistical moment among all missingness scenarios, with MCAR data in the first column, from data matrix containing jumps

		Number of jumps in the series					
		0	**2**	**5**	**8**	**9**	**11**
Mean (10^{-5})	**LI**	3.59	3.59	3.62	3.62	3.65	3.66
	LOCF	2.99	2.98	2.98	3.00	3.03	3.04
	MSSA	2.52	2.45	2.02	2.45	6.35	8.65
	BB	3.11	3.06	3.52	3.63	3.71	3.88
	IPCA	1.05	1.22	2.21	2.56	3.00	2.97
	KNN	0.67	0.99	2.40	2.49	2.74	2.86
	RF	1.00	0.98	0.98	1.00	1.00	0.98
	MICE	3.19	3.23	2.90	2.74	2.40	2.15
	MIPCA	0.79	0.95	1.26	1.75	2.02	2.06
	Amelia	0.70	0.74	0.83	0.86	0.81	0.82
Std (10^{-3})	**LI**	0.11	0.24	0.62	0.63	0.66	0.66
	LOCF	0.21	0.20	0.24	0.26	0.48	0.46
	MSSA	0.19	0.28	0.62	0.64	0.81	0.82
	BB	0.20	0.50	1.48	1.53	1.62	1.62
	IPCA	0.11	0.18	0.24	0.24	0.43	0.44
	KNN	0.10	0.87	1.80	1.71	1.77	1.80
	RF	0.06	0.05	0.11	0.17	0.14	0.14
	MICE	0.42	0.40	0.47	0.51	0.66	0.64
	MIPCA	0.09	0.14	0.21	0.21	0.23	0.28
	Amelia	0.09	0.19	0.28	0.22	0.20	0.18

		Number of jumps in the series					
		0	**2**	**5**	**8**	**9**	**11**
Skewness	**LI**	0.13	0.59	1.33	1.09	0.84	0.83
	LOCF	0.32	0.31	0.33	0.42	0.66	0.64
	MSSA	0.11	0.29	0.78	0.65	0.31	0.31
	BB	0.06	0.18	0.29	0.27	0.20	0.21
	IPCA	0.08	0.28	0.46	0.46	0.72	0.72
	KNN	0.10	0.91	0.86	0.83	0.72	0.80
	RF	0.05	0.04	0.16	0.18	0.10	0.10
	MICE	0.61	0.58	0.54	0.58	0.53	0.53
	MIPCA	0.06	0.18	0.24	0.32	0.32	0.39
	Amelia	0.05	0.23	0.21	0.13	0.09	0.09
Kurtosis	**LI**	0.24	3.60	14.81	9.53	6.62	5.94
	LOCF	1.07	0.99	3.14	2.10	5.87	5.46
	MSSA	0.24	1.26	7.03	5.00	1.72	1.55
	BB	0.08	0.75	2.70	1.75	1.24	1.16
	IPCA	0.15	1.54	4.34	2.33	5.82	5.55
	KNN	0.27	8.79	14.59	10.34	8.10	7.24
	RF	0.08	0.21	1.13	0.80	0.44	0.41
	MICE	1.17	0.95	3.95	2.91	4.25	3.88
	MIPCA	0.15	1.26	2.51	1.81	2.18	2.81
	Amelia	0.12	1.94	2.44	1.19	0.80	0.64

D.3 Average number of principal components used by IPCA and MIPCA, among the 100 missingness scenarios, for imputation of MCAR data in the first column of a sample with jumps

	Number of jumps in the series					
	0	**2**	**5**	**8**	**9**	**11**
IPCA	4	4	4.01	4.01	4	3.97
MIPCA	4	4	3.97	3.98	3.9	3.93

D.4 Standard deviation of the proximity measures among all missingness scenarios, with MCAR data in the first column, from data matrix containing jumps

		Number of jumps in the series					
		0	**2**	**5**	**8**	**9**	**11**
MAE (10^{-3})	**LI**	0.23	0.29	0.55	0.64	0.70	0.72
	LOCF	0.43	0.54	0.88	1.00	1.05	1.11
	MSSA	0.38	0.45	0.65	0.64	0.81	0.86
	BB	0.21	0.33	1.12	1.14	1.30	1.28
	IPCA	0.34	0.72	1.26	1.37	1.42	1.37
	KNN	0.18	0.45	0.74	0.80	0.84	0.90
	RF	0.16	0.16	0.35	0.46	0.37	0.37
	MICE	0.60	0.65	2.49	2.84	2.87	2.87
	MIPCA	0.11	0.30	0.86	1.54	1.95	2.21
	Amelia	0.08	0.09	0.12	0.11	0.12	0.11
RMSE (10^{-3})	**LI**	0.26	0.53	1.39	1.45	1.60	1.58
	LOCF	0.55	1.00	2.74	2.83	2.98	3.01
	MSSA	0.45	0.59	1.02	0.95	1.15	1.19
	BB	0.27	0.36	1.04	1.12	1.33	1.32
	IPCA	0.44	1.29	3.11	3.25	3.50	3.44
	KNN	0.25	1.41	2.96	2.79	3.01	3.04
	RF	0.21	0.21	1.88	2.05	1.65	1.65
	MICE	0.78	0.99	4.21	4.61	4.77	4.76
	MIPCA	0.17	0.40	1.50	2.85	3.88	4.42
	Amelia	0.12	0.17	0.28	0.22	0.24	0.22

D.5 Standard deviation of the covariance matrix differences among all missingness scenarios, with MCAR data in the first column, from data matrix containing jumps

		Number of jumps in the series					
		0	**2**	**5**	**8**	**9**	**11**
Cov (10^{-4})	**LI**	0.09	0.25	0.85	0.98	1.12	1.15
	LOCF	0.20	0.22	0.38	0.47	0.95	0.95
	MSSA	0.23	0.28	1.00	1.15	1.72	1.81
	BB	0.29	0.23	2.39	2.85	3.33	3.46
	IPCA	0.10	0.20	0.38	0.43	0.85	0.89
	KNN	0.10	0.47	2.58	3.01	3.54	3.77
	RF	0.04	0.06	0.17	0.29	0.27	0.27
	MICE	0.60	0.26	0.77	0.94	1.34	1.33
	MIPCA	0.08	0.15	0.34	0.37	0.44	0.56
	Amelia	0.09	0.19	0.47	0.41	0.40	0.37

D.6 Standard deviation of each risk measures among all missingness scenarios, with MCAR data in the first column, from data matrix containing jumps

		Number of jumps in the series					
		0	**2**	**5**	**8**	**9**	**11**
$VaR_{99\%}^{1day}$ (10^{-3})	**LI**	1.50	1.57	8.81	11.12	11.98	11.98
	LOCF	1.86	2.53	3.34	2.07	3.54	3.54
	MSSA	0.91	1.73	6.37	8.98	10.41	9.95
	BB	0.53	1.52	4.55	8.05	8.78	8.61
	IPCA	0.51	1.13	6.00	5.74	4.64	4.10
	KNN	0.65	1.83	13.39	9.37	9.37	9.37
	RF	0.73	0.57	0.42	6.79	3.55	3.56
	MICE	3.07	3.74	4.73	4.04	3.02	3.00
	MIPCA	0.49	0.59	3.34	3.86	3.43	3.53
	Amelia	0.47	0.54	3.56	2.34	1.95	1.79
$ES_{99\%}^{1day}$ (10^{-3})	**LI**	0.88	2.04	3.99	5.19	5.99	5.99
	LOCF	1.53	1.54	1.62	1.53	2.12	2.12
	MSSA	0.84	1.78	4.11	4.93	5.18	5.29
	BB	0.54	2.44	5.84	6.28	6.41	6.39
	IPCA	0.53	1.22	1.71	1.75	2.03	2.12
	KNN	0.77	3.55	6.83	7.37	7.75	7.92
	RF	0.60	0.52	0.41	1.18	1.13	1.13
	MICE	3.08	3.12	3.08	2.83	2.98	2.88
	MIPCA	0.58	0.82	1.24	1.46	1.50	1.61
	Amelia	0.53	0.95	1.48	1.29	1.22	1.07

		Number of jumps in the series					
		0	**2**	**5**	**8**	**9**	**11**
$VaR_{97.5\%}^{10days}$ (10^{-3})	**LI**	2.32	2.10	1.81	1.80	1.57	1.57
	LOCF	2.96	2.85	2.56	2.56	1.80	1.80
	MSSA	2.53	2.77	2.89	3.13	5.68	5.97
	BB	1.42	1.22	2.01	2.13	1.99	2.07
	IPCA	1.24	1.34	1.55	1.95	1.90	1.91
	KNN	1.89	4.08	9.37	8.96	12.79	12.84
	RF	1.37	0.67	0.16	0.17	0.37	0.36
	MICE	2.84	2.54	2.22	1.88	1.71	1.66
	MIPCA	1.12	0.82	1.01	1.27	1.23	1.28
	Amelia	0.76	0.83	0.88	0.81	1.24	1.34
$ES_{97.5\%}^{10days}$ (10^{-3})	**LI**	1.62	1.61	1.63	1.63	1.49	1.49
	LOCF	2.57	2.58	1.92	1.92	2.03	2.03
	MSSA	1.79	2.29	2.63	2.69	4.62	4.55
	BB	1.25	1.26	2.05	2.14	2.18	2.22
	IPCA	0.74	0.88	1.21	1.33	1.92	2.04
	KNN	1.42	4.58	6.62	6.32	9.64	9.62
	RF	0.66	0.64	0.42	0.43	0.06	0.06
	MICE	2.56	2.41	1.94	1.61	1.83	1.80
	MIPCA	0.75	0.90	0.73	0.90	1.28	1.45
	Amelia	0.61	1.06	0.65	0.58	1.26	1.19

Appendix E: Impact of missing values depending on extremes values of another series

E.1 Number of principal components used by IPCA and MIPCA for imputation of MAR data in the first column

	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
IPCA	3	3	3	3	3	3	3	3	3	3	3	3	3	3
MIPCA	2	2	2	2	2	2	2	2	2	2	2	2	2	2

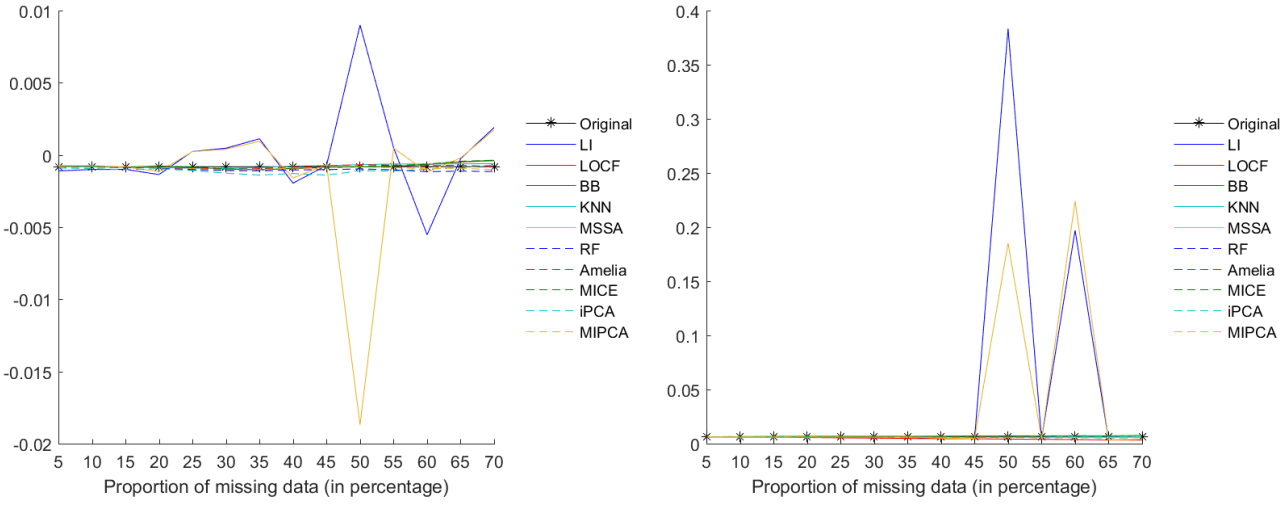
Appendix F: Impact of successive missing data in the middle of the series

F.1 Number of principal components used by IPCA and MIPCA for imputation of MAR data corresponding to successive missing data in the middle of the first series

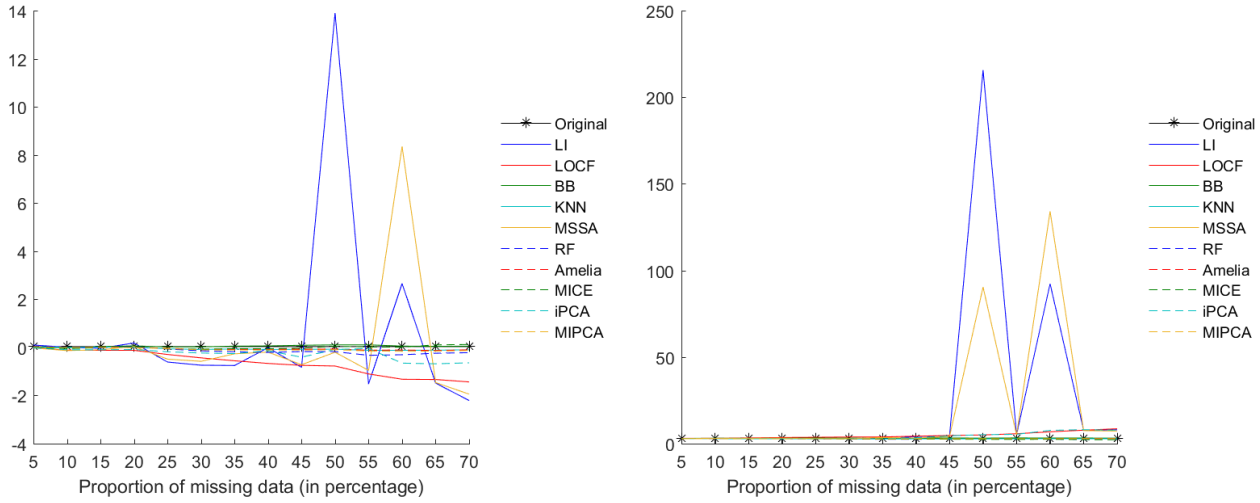
	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
IPCA	4	4	4	4	4	4	4	4	4	4	4	4	4	4
MIPCA	4	4	4	4	4	4	4	4	4	4	4	4	4	4

Appendix G: Impact of successive missing data at the end of the series

G.1 The first four statistical moments of the returns of the imputed data based on a matrix containing MAR data (successive missing data at the end of the first series) according to the missingness proportion



(a) Mean *(b)* Standard deviation

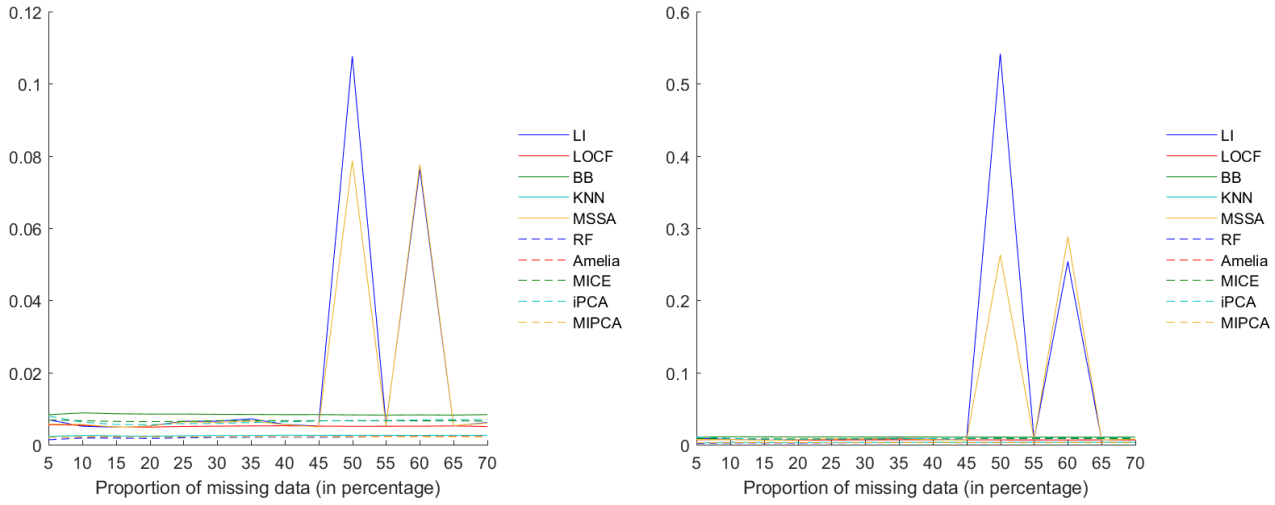


(c) Skewness *(d)* Kurtosis

G.2 Number of principal components used by IPCA and MIPCA for imputation of MAR data corresponding to successive missing data at the end of the first series

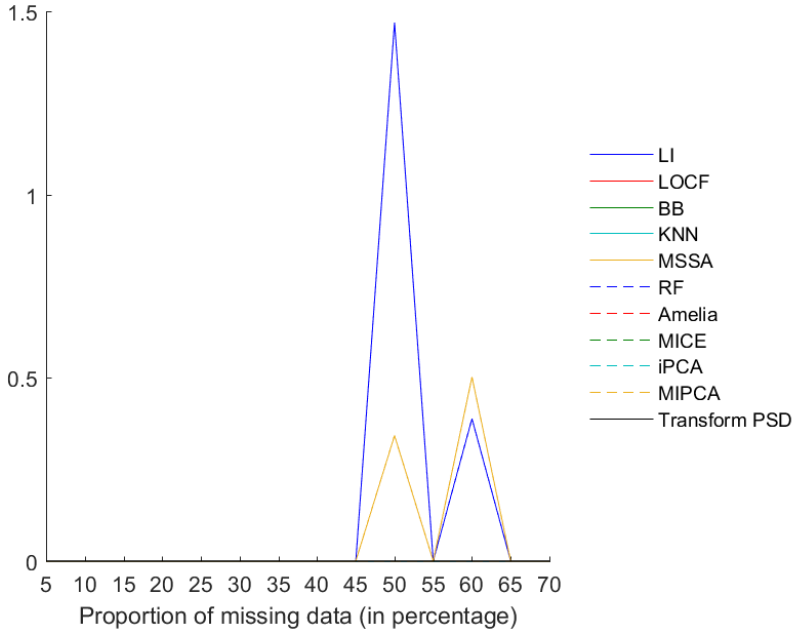
	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
IPCA	4	4	4	4	4	4	4	4	4	4	4	4	4	4
MIPCA	4	4	4	4	4	4	4	4	4	4	4	4	4	4

G.3 MAE and RMSE between the return of the imputed data from a matrix containing MAR data (successive missing data at the end of the first series) and the original data matrix, according to the missingness proportion

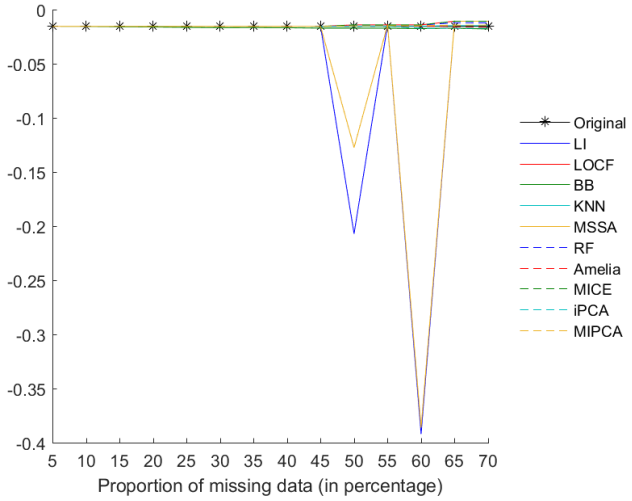


(a) MAE *(b)* RMSE

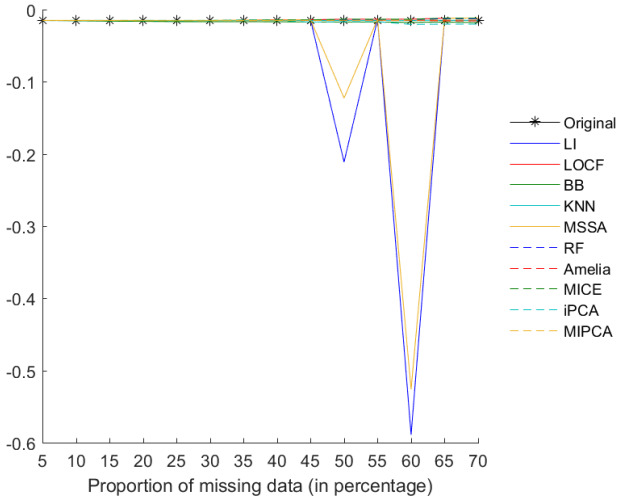
G.4 Covariance matrices differences, according to Frobenius norm, based on original returns and the imputed returns from a matrix containing MAR data (successive missing data at the end of the first series) according to the missingness proportion



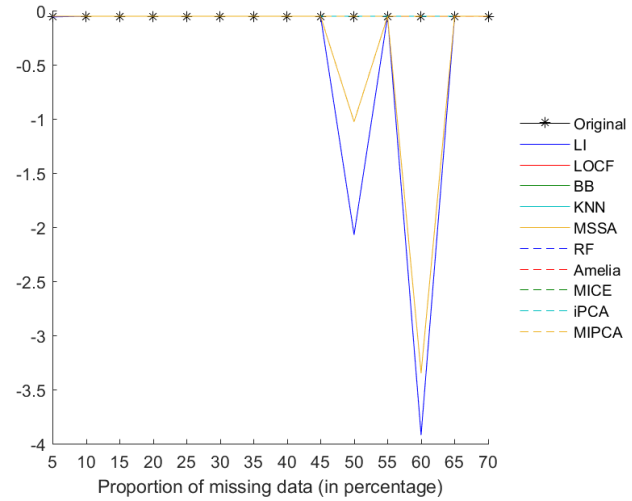
G.5 The 1-day risk measures and 10-day risk measures, computed from a matrix containing MAR data (successive missing data at the end of the first series) according to the missingness proportion



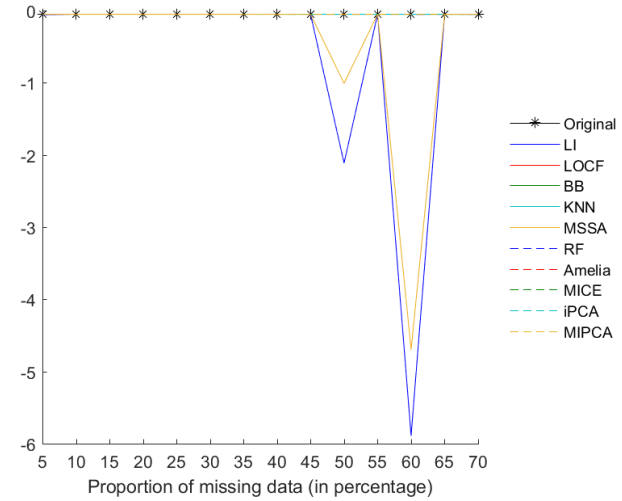
(a) The 1-day VaR at 99%



(b) The 1-day ES at 97.5%



(c) The 10-day VaR at 99%



(d) The 10-day ES at 97.5%

Appendix H: Impact of extreme values missing

H.1 Number of principal components used by IPCA and MIPCA for imputation of MNAR data corresponding to extreme values of the first series

	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
IPCA	3	3	3	3	3	3	3	3	3	3	3	3	3	3
MIPCA	2	2	2	2	2	2	2	2	2	2	2	2	2	2

Appendix I: Impact on a sample of historical data based on a heuristic approach

I.1 Number of MCAR tests that are calculable, when applied to historical return matrices based on a heuristic approach containing MCAR data on the first column, for a 5% significance level

	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Little's test	100	100	100	100	100	100	100	100	100	100	100	100	100	100
J&J's test	100	100	100	100	100	100	100	100	100	100	100	100	100	100

I.2 Standard deviation of each statistical moment among all missingness scenarios, with MCAR data in the first column of historical data based on a heuristic approach

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Mean (10^{-4})	**LI**	0.33	0.43	0.52	0.51	0.80	1.01	0.88	1.28	1.19	1.14	1.43	1.91	1.48	2.05
	LOCF	0.20	0.27	0.36	0.41	0.46	0.68	0.63	0.66	1.00	1.04	0.83	1.25	1.33	1.66
	MSSA	0.38	0.55	0.60	0.79	0.87	0.95	1.03	1.14	1.16	1.23	1.31	1.28	1.36	1.32
	BB	0.25	0.40	0.51	0.65	0.79	1.15	1.22	1.54	1.85	2.35	2.43	3.49	3.59	4.67
	IPCA	0.11	0.18	0.22	0.26	0.30	0.35	0.34	0.37	0.54	0.56	0.62	0.83	0.89	1.75
	KNN	0.10	0.16	0.19	0.21	0.23	0.25	0.25	0.26	0.36	0.29	0.36	0.33	0.34	0.40
	RF	0.02	0.04	0.05	0.05	0.08	0.19	0.14	0.25	0.28	0.31	0.39	0.37	0.39	0.91
	MICE	0.12	0.20	0.24	0.29	0.31	0.40	0.42	0.41	0.58	0.65	0.56	0.75	0.82	1.23
	MIPCA	0.11	0.19	0.23	0.27	0.32	0.37	0.37	0.40	0.56	0.59	0.61	0.80	0.87	1.26
	Amelia	0.10	0.18	0.21	0.25	0.31	0.34	0.37	0.41	0.52	0.54	0.71	0.84	1.17	37.38
Std (10^{-3})	**LI**	0.46	0.66	0.78	0.89	0.92	1.12	1.15	1.24	1.32	1.29	1.19	1.30	1.26	1.32
	LOCF	0.51	0.67	0.93	1.02	1.18	1.36	1.30	1.57	1.96	2.08	1.86	2.21	2.51	2.78
	MSSA	0.56	0.84	0.90	1.03	1.10	1.16	1.26	1.31	1.47	1.45	1.23	1.53	1.64	1.60
	BB	0.53	0.90	1.16	1.51	1.77	2.56	2.86	3.40	4.03	5.18	5.42	7.30	7.47	9.27
	IPCA	0.27	0.45	0.56	0.66	0.76	0.86	0.89	1.01	1.36	1.42	1.69	2.02	2.19	4.14
	KNN	0.25	0.40	0.47	0.50	0.57	0.56	0.61	0.65	0.86	0.64	0.82	0.75	0.86	0.85
	RF	0.06	0.13	0.15	0.16	0.24	0.51	0.24	0.61	0.81	0.83	0.77	0.91	1.09	0.99
	MICE	0.29	0.47	0.55	0.67	0.73	0.96	1.05	1.05	1.41	1.49	1.34	1.50	1.70	2.47
	MIPCA	0.27	0.47	0.58	0.69	0.81	0.89	0.95	1.04	1.39	1.51	1.66	1.95	2.24	2.87
	Amelia	0.26	0.46	0.54	0.62	0.80	0.80	0.93	1.01	1.26	1.21	1.75	1.81	2.53	68.48

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Skewness	**LI**	0.18	0.26	0.31	0.35	0.37	0.43	0.46	0.50	0.59	0.59	0.67	0.69	0.78	0.80
	LOCF	0.18	0.25	0.30	0.36	0.39	0.70	0.67	0.72	1.03	0.88	0.92	0.88	1.28	1.53
	MSSA	0.16	0.21	0.22	0.25	0.26	0.25	0.27	0.25	0.28	0.27	0.29	0.28	0.32	0.33
	BB	0.12	0.15	0.15	0.14	0.13	0.12	0.11	0.09	0.09	0.08	0.08	0.06	0.06	0.08
	IPCA	0.07	0.12	0.13	0.15	0.17	0.17	0.22	0.19	0.24	0.25	0.34	0.44	0.65	1.52
	KNN	0.07	0.11	0.12	0.14	0.15	0.14	0.18	0.16	0.16	0.15	0.17	0.18	0.19	0.27
	RF	0.02	0.03	0.03	0.03	0.04	0.17	0.03	0.18	0.17	0.18	0.18	0.29	0.28	0.36
	MICE	0.09	0.14	0.18	0.25	0.28	0.33	0.35	0.43	0.58	0.50	0.52	0.51	0.72	1.10
	MIPCA	0.07	0.11	0.13	0.14	0.14	0.14	0.19	0.16	0.22	0.23	0.28	0.37	0.54	1.15
	Amelia	0.07	0.10	0.12	0.15	0.18	0.16	0.19	0.20	0.21	0.20	0.24	0.27	0.31	0.43
Kurtosis	**LI**	0.66	1.06	1.35	1.62	1.82	2.10	2.05	2.67	3.05	3.43	3.68	4.75	4.69	5.68
	LOCF	0.78	1.05	1.36	1.88	2.45	5.54	5.27	5.72	8.81	6.57	6.41	6.77	10.04	14.04
	MSSA	0.55	0.77	0.85	0.96	1.00	1.02	0.96	1.07	1.06	1.14	1.16	1.29	1.22	1.51
	BB	0.38	0.56	0.52	0.50	0.44	0.39	0.33	0.33	0.29	0.24	0.19	0.20	0.15	0.15
	IPCA	0.25	0.50	0.51	0.61	0.64	0.77	1.08	0.87	1.39	1.38	1.93	2.07	4.45	12.12
	KNN	0.20	0.45	0.47	0.51	0.54	0.58	0.75	0.66	0.86	0.71	0.75	0.71	0.78	1.71
	RF	0.02	0.06	0.07	0.12	0.17	0.67	0.16	0.77	0.66	0.68	0.89	2.56	1.34	1.54
	MICE	0.43	0.62	0.90	1.47	1.67	2.57	2.50	3.33	4.79	3.61	3.40	3.09	6.12	8.66
	MIPCA	0.22	0.46	0.45	0.52	0.54	0.64	1.00	0.81	1.27	1.21	1.45	1.88	3.16	8.80
	Amelia	0.23	0.43	0.44	0.53	0.64	0.67	0.83	0.72	1.06	0.89	1.04	1.10	1.49	3.17

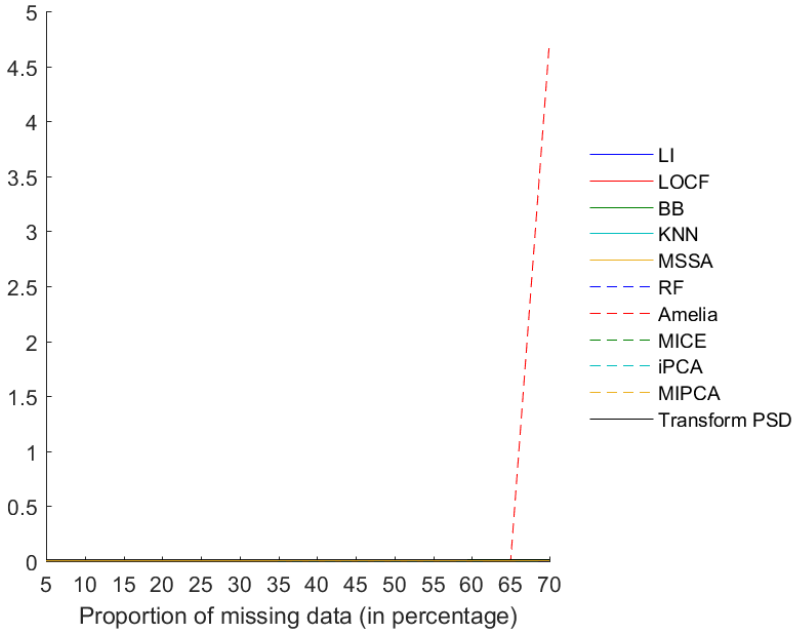
I.3 Average number of principal components used by IPCA and MIPCA for imputation of MCAR data in the first column of historical data based on a heuristic approach

	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
IPCA	2	2	2	2	2	2	2	2	2	1.98	1.94	1.95	1.9	1.89
MIPCA	1	1	1	1	1	1	1	1	1	1	1	1	1	1

I.4 Standard deviation of the proximity measures among all missingness scenarios, with MCAR data in the first column of historical data based on a heuristic approach

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
MAE (10^{-3})	**LI**	4.13	2.89	2.32	2.03	1.67	1.57	1.43	1.28	1.24	1.11	0.88	0.79	0.70	0.70
	LOCF	7.58	4.90	3.56	3.32	2.88	2.63	2.29	2.24	1.77	1.91	1.72	1.76	1.63	1.34
	MSSA	5.50	3.70	3.09	2.52	2.00	1.91	1.68	1.63	1.40	1.36	1.12	1.14	1.02	1.02
	BB	1.96	0.99	0.91	1.08	1.30	1.76	2.05	2.50	2.84	3.96	4.08	5.47	5.61	6.81
	IPCA	2.37	1.57	1.14	1.10	0.98	0.96	0.74	0.76	0.82	1.18	1.41	2.23	2.83	4.57
	KNN	2.29	1.51	1.09	1.01	0.88	0.77	0.65	0.62	0.61	0.56	0.51	0.50	0.48	0.42
	RF	1.08	1.06	0.87	0.84	0.82	0.97	0.56	0.96	0.85	0.79	0.78	1.17	0.87	1.47
	MICE	3.23	2.05	1.55	1.52	1.32	1.12	0.94	1.06	1.08	1.26	1.28	1.31	1.50	2.42
	MIPCA	1.55	1.04	0.80	0.78	0.70	0.63	0.52	0.51	0.50	0.46	0.63	0.86	0.92	1.87
	Amelia	1.81	1.10	0.78	0.70	0.62	0.55	0.52	0.61	0.63	0.75	0.88	1.04	1.81	8.58
RMSE (10^{-3})	**LI**	6.65	4.74	3.69	3.21	2.61	2.68	2.36	2.07	1.91	1.69	1.30	1.30	1.10	1.04
	LOCF	11.1	8.10	6.32	6.23	5.32	5.12	4.38	4.43	3.75	4.29	3.72	4.07	3.94	3.32
	MSSA	6.73	4.43	3.75	3.24	2.58	2.59	2.23	2.37	1.88	1.94	1.65	1.68	1.59	1.52
	BB	2.89	1.46	1.17	1.24	1.51	1.99	2.43	2.96	3.43	4.77	4.94	6.68	6.86	8.40
	IPCA	3.75	2.56	1.93	1.79	1.59	1.68	1.33	1.36	1.52	2.14	2.49	3.65	4.73	7.41
	KNN	3.49	2.38	1.80	1.63	1.45	1.29	1.08	1.09	1.08	1.00	0.94	0.94	0.90	0.88
	RF	2.44	1.87	1.43	1.84	1.51	3.08	0.76	2.99	2.87	2.49	1.81	5.12	2.59	3.37
	MICE	6.08	4.77	3.85	3.77	3.46	3.38	2.75	3.00	3.02	3.19	2.98	3.16	3.34	3.73
	MIPCA	2.23	1.60	1.24	1.15	1.05	0.97	0.79	0.82	0.78	0.92	1.13	1.64	1.80	3.63
	Amelia	2.84	1.86	1.33	1.19	1.18	1.00	0.91	1.11	1.06	1.16	1.37	1.65	2.89	72.0

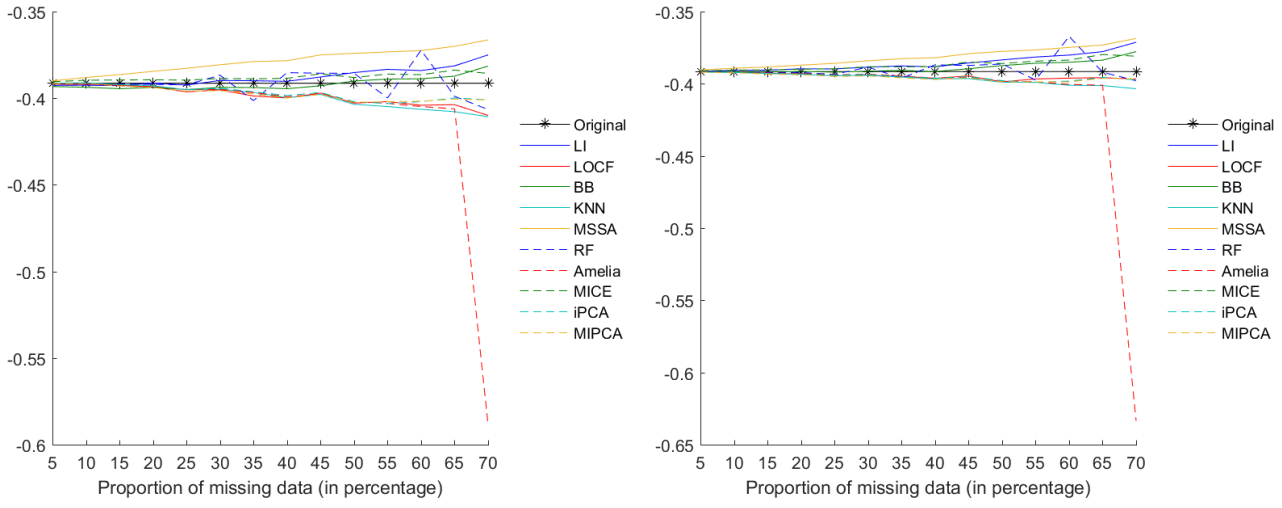
I.5 Average covariance matrices differences, according to Frobenius norm, based on original historical returns and the imputed returns from historical data based on a heuristic approach containing MCAR data (only in the first column) according to the missingness probability



I.6 Standard deviation of the covariance matrix differences among all missingness scenarios, with MCAR data in the first column of historical data based on a heuristic approach

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Cov (10^{-3})	**LI**	0.03	0.10	0.21	0.31	0.38	0.51	0.54	0.61	0.64	0.62	0.56	0.59	0.54	0.55
	LOCF	0.14	0.19	0.28	0.32	0.40	0.53	0.55	0.71	0.97	1.00	0.94	1.06	1.29	1.72
	MSSA	0.16	0.25	0.27	0.32	0.32	0.35	0.37	0.25	0.25	0.23	0.27	0.52	0.65	0.73
	BB	0.22	0.49	0.74	1.07	1.36	2.07	2.46	2.93	3.62	4.80	5.01	7.05	7.42	9.41
	IPCA	0.07	0.11	0.15	0.17	0.20	0.23	0.20	0.27	0.32	0.30	0.48	0.49	0.46	2.20
	KNN	0.06	0.10	0.12	0.14	0.16	0.15	0.16	0.20	0.25	0.21	0.28	0.29	0.29	0.32
	RF	0.02	0.03	0.03	0.04	0.05	0.10	0.07	0.06	0.08	0.09	0.08	0.14	0.13	0.27
	MICE	0.07	0.12	0.15	0.18	0.23	0.34	0.36	0.35	0.52	0.45	0.50	0.53	0.77	1.56
	MIPCA	0.08	0.17	0.23	0.30	0.38	0.43	0.45	0.52	0.67	0.67	0.78	0.89	0.80	1.50
	Amelia	0.07	0.14	0.18	0.23	0.33	0.36	0.41	0.53	0.63	0.65	1.05	1.24	2.58	41882

I.7 Average 10-day risk measures, computed from historical sample based on a heuristic approach containing MCAR data (only in the first column) according to the missingness probability



(a) Average of the 10-day VaR at 99%

(b) Average of the 10-day ES at 97.5%

I.8 Standard deviation of each risk measures among all missingness scenarios, with MCAR data in the first column of historical data based on a heuristic approach

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
$VaR_{99\%}^{1day}$ (10^{-2})	**LI**	1.46	1.71	1.78	1.71	1.62	1.78	1.69	1.31	1.39	1.29	1.29	1.16	1.31	1.17
	LOCF	0.89	0.95	1.14	1.25	1.42	1.76	1.87	2.04	2.03	2.23	2.08	2.01	2.28	2.53
	MSSA	1.09	1.35	1.49	1.50	1.56	1.52	1.42	1.40	1.47	1.29	1.16	1.10	1.27	1.15
	BB	0.93	0.93	0.91	0.89	0.88	1.08	1.08	1.11	1.25	1.42	1.44	1.78	1.72	2.08
	IPCA	0.24	0.34	0.40	0.42	0.52	0.60	0.70	0.67	0.68	1.00	1.05	1.24	1.30	2.04
	KNN	0.23	0.31	0.41	0.41	0.45	0.47	0.53	0.58	0.62	0.53	0.57	0.55	0.64	0.69
	RF	0.00	0.00	0.00	0.00	0.75	1.00	0.00	1.45	1.18	1.57	1.31	1.56	1.70	2.28
	MICE	0.71	0.88	1.10	1.27	1.36	1.54	1.46	1.52	1.53	1.65	1.46	1.75	1.76	2.12
	MIPCA	0.22	0.32	0.44	0.46	0.48	0.53	0.56	0.53	0.67	0.87	0.78	0.87	1.06	1.62
	Amelia	0.22	0.31	0.38	0.40	0.45	0.62	0.51	0.78	0.69	0.89	0.93	1.03	1.17	3.87
$ES_{99\%}^{1day}$ (10^{-2})	**LI**	0.59	0.79	0.87	0.91	0.91	1.06	1.04	1.02	1.06	1.02	0.97	0.93	1.06	0.99
	LOCF	0.42	0.55	0.65	0.74	0.78	0.89	0.88	1.11	1.09	1.34	1.09	1.12	1.20	1.14
	MSSA	0.53	0.72	0.79	0.90	0.95	1.03	0.96	0.96	1.09	1.04	0.92	1.03	1.05	1.05
	BB	0.49	0.72	0.83	0.91	0.95	1.10	1.07	1.15	1.27	1.44	1.42	1.77	1.69	2.01
	IPCA	0.16	0.31	0.35	0.41	0.45	0.48	0.52	0.46	0.65	0.71	0.66	0.94	1.01	1.77
	KNN	0.14	0.27	0.32	0.34	0.37	0.32	0.43	0.35	0.45	0.39	0.41	0.43	0.42	0.48
	RF	0.00	0.03	0.03	0.03	0.14	0.59	0.17	0.67	0.76	0.64	0.46	0.80	0.61	1.03
	MICE	0.31	0.48	0.55	0.71	0.79	0.82	0.80	0.90	0.94	1.15	1.00	1.29	1.09	1.51
	MIPCA	0.16	0.32	0.36	0.41	0.45	0.47	0.52	0.48	0.69	0.73	0.68	0.89	0.96	1.36
	Amelia	0.14	0.29	0.31	0.37	0.44	0.47	0.50	0.52	0.62	0.62	0.66	0.79	0.81	18.8

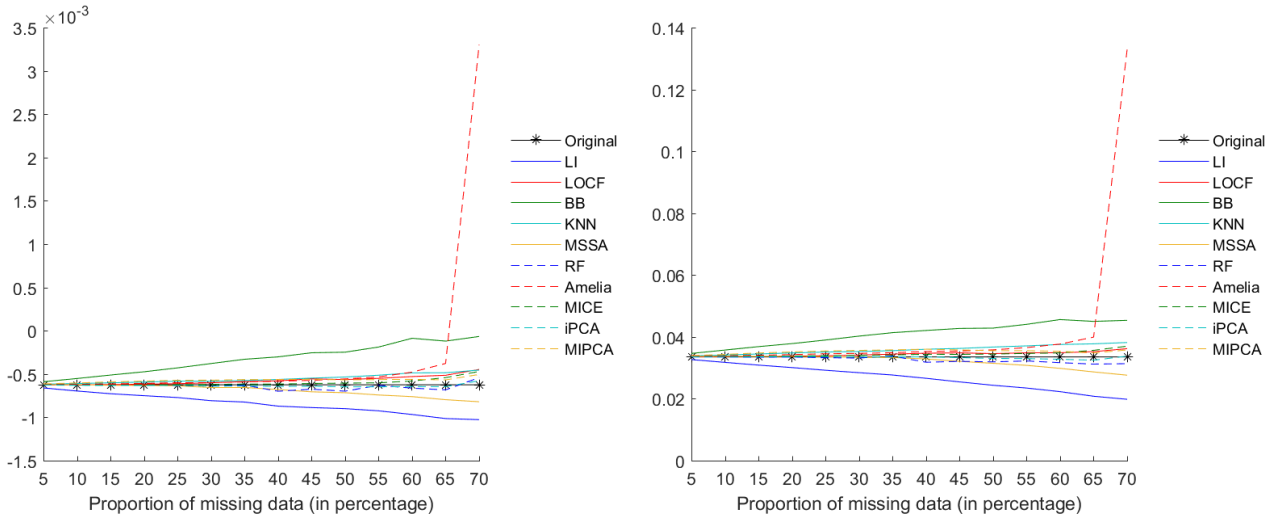
		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
$VaR_{97.5\%}^{10days}$ (10^{-2})	**LI**	0.69	1.03	1.23	1.44	1.56	1.70	1.70	1.84	1.89	1.91	1.81	2.36	2.74	3.01
	LOCF	0.61	0.82	1.05	1.35	1.56	1.66	1.95	2.31	2.47	3.04	2.99	3.62	3.59	4.14
	MSSA	0.33	0.53	0.68	0.79	0.87	0.98	0.98	1.10	1.19	1.07	0.96	1.17	1.50	1.48
	BB	0.65	0.87	1.00	1.16	1.23	1.35	1.29	1.51	1.53	1.69	1.65	1.99	2.35	2.71
	IPCA	0.21	0.42	0.56	0.68	0.87	0.90	1.17	1.20	1.31	1.37	1.33	1.87	1.96	3.37
	KNN	0.23	0.44	0.57	0.69	0.91	0.95	1.18	1.28	1.26	1.35	1.26	1.42	1.41	1.36
	RF	0.00	0.10	0.11	0.12	0.16	0.38	0.78	0.38	0.35	0.30	1.24	0.93	1.71	1.59
	MICE	0.29	0.44	0.50	0.59	0.76	0.94	0.91	1.10	1.41	1.37	1.48	2.20	2.49	3.14
	MIPCA	0.22	0.44	0.53	0.61	0.77	0.76	0.98	1.00	1.08	1.22	1.20	1.73	1.99	2.94
	Amelia	0.27	0.50	0.61	0.70	0.86	0.90	1.07	1.16	1.20	1.20	1.11	1.53	1.61	148
$ES_{97.5\%}^{10days}$ (10^{-2})	**LI**	0.44	0.58	0.73	0.89	1.01	1.14	1.13	1.41	1.49	1.52	1.54	2.00	2.50	2.79
	LOCF	0.44	0.64	0.85	1.04	1.20	1.46	1.70	1.99	2.19	2.63	2.61	3.16	3.38	4.18
	MSSA	0.37	0.48	0.60	0.71	0.79	0.89	0.84	1.09	1.14	1.13	1.07	1.31	1.64	1.58
	BB	0.40	0.56	0.69	0.84	0.94	1.06	1.06	1.33	1.40	1.49	1.49	1.87	2.29	2.61
	IPCA	0.12	0.25	0.30	0.34	0.42	0.48	0.62	0.69	0.89	0.94	1.02	1.52	1.78	3.17
	KNN	0.11	0.23	0.28	0.31	0.38	0.36	0.57	0.61	0.67	0.74	0.66	0.92	0.99	1.09
	RF	0.02	0.09	0.10	0.10	0.14	0.35	0.30	0.40	0.54	0.35	0.91	1.29	1.50	1.39
	MICE	0.27	0.37	0.45	0.53	0.66	0.92	0.88	1.09	1.46	1.32	1.50	2.21	2.57	3.05
	MIPCA	0.12	0.25	0.29	0.34	0.42	0.46	0.67	0.72	0.88	1.01	1.02	1.54	1.90	2.82
	Amelia	0.13	0.22	0.27	0.31	0.42	0.47	0.62	0.65	0.84	0.87	0.84	1.20	1.39	190

Appendix J: Impact on a sample of historical data based on the graphical Lasso

J.1 Number of MCAR tests that are calculable, when applied to historical return matrices based on the graphical Lasso containing MCAR data on the first column, for a 5% significance level

	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Little's test	100	100	100	100	100	100	100	100	100	100	100	100	100	100
J&J's test	100	100	100	100	100	100	100	100	100	100	100	100	100	100

J.2 Average mean and standard deviation of the returns of the historical imputed data matrix based on the graphical Lasso containing MCAR data (only in the first column) according to the missingness probability



(a) Mean *(b)* Standard Deviation

J.3 Standard deviation of each statistical moment among all missingness scenarios, with MCAR data in the first column of historical data based on the graphical Lasso

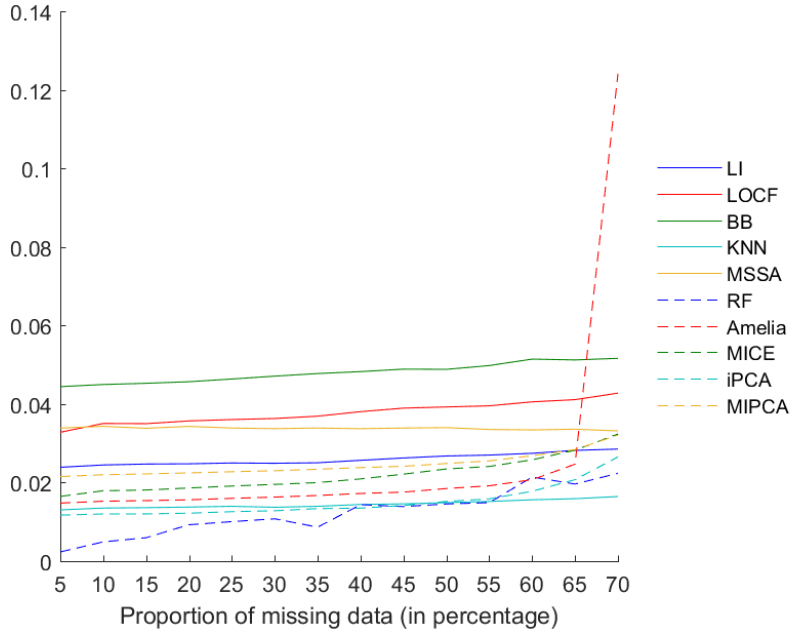
		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Mean (10^{-4})	**LI**	0.34	0.44	0.53	0.51	0.81	1.01	0.87	1.27	1.19	1.15	1.44	1.93	1.50	2.04
	LOCF	0.20	0.26	0.33	0.38	0.44	0.71	0.64	0.65	0.97	1.03	0.87	1.30	1.33	1.66
	MSSA	0.31	0.43	0.48	0.58	0.59	0.66	0.73	0.79	0.79	1.01	1.00	1.14	1.22	1.16
	BB	0.26	0.41	0.51	0.64	0.81	1.13	1.21	1.50	1.76	2.23	2.50	3.51	3.52	4.84
	IPCA	0.14	0.20	0.24	0.27	0.33	0.38	0.42	0.42	0.56	0.60	0.66	1.03	1.05	1.88
	KNN	0.18	0.25	0.28	0.30	0.35	0.37	0.39	0.39	0.54	0.42	0.61	0.65	0.63	0.87
	RF	0.02	0.04	0.04	0.06	0.08	0.18	0.17	0.28	0.27	0.36	0.52	0.43	0.43	1.21
	MICE	0.13	0.21	0.26	0.30	0.33	0.41	0.41	0.51	0.69	0.65	0.64	0.83	1.00	1.49
	MIPCA	0.16	0.22	0.28	0.33	0.39	0.46	0.48	0.48	0.64	0.69	0.73	1.02	1.12	1.84
	Amelia	0.12	0.19	0.22	0.25	0.31	0.36	0.39	0.37	0.56	0.56	0.69	0.93	2.26	379
Std (10^{-3})	**LI**	0.49	0.68	0.75	0.89	0.95	1.10	1.17	1.16	1.30	1.34	1.26	1.29	1.36	1.29
	LOCF	0.49	0.62	0.79	0.91	1.10	1.43	1.41	1.56	1.90	2.06	2.06	2.42	2.58	2.83
	MSSA	0.69	0.89	0.93	1.02	1.12	1.16	1.25	1.24	1.55	1.54	1.41	1.56	1.76	1.63
	BB	0.55	0.93	1.14	1.51	1.85	2.46	2.82	3.28	3.85	5.06	5.51	7.33	7.31	9.60
	IPCA	0.31	0.48	0.55	0.62	0.69	0.81	1.02	0.96	1.24	1.50	1.49	1.99	2.34	4.10
	KNN	0.38	0.53	0.60	0.63	0.63	0.72	0.72	0.80	0.85	0.78	0.87	0.91	0.86	0.98
	RF	0.05	0.12	0.13	0.17	0.25	0.46	0.24	0.72	0.80	0.90	0.79	0.91	0.97	1.19
	MICE	0.31	0.47	0.55	0.67	0.74	0.98	0.92	1.12	1.32	1.45	1.38	1.62	1.70	2.37
	MIPCA	0.31	0.48	0.56	0.70	0.78	0.94	1.09	1.07	1.38	1.72	1.72	2.09	2.31	3.35
	Amelia	0.26	0.42	0.49	0.55	0.64	0.82	0.88	0.87	1.17	1.21	1.43	1.84	4.24	621

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Skewness	**LI**	0.18	0.28	0.34	0.37	0.39	0.44	0.45	0.51	0.61	0.59	0.65	0.73	0.74	0.86
	LOCF	0.17	0.23	0.29	0.33	0.40	0.64	0.66	0.76	1.08	0.93	0.94	0.93	1.35	1.50
	MSSA	0.17	0.22	0.26	0.28	0.27	0.29	0.27	0.29	0.32	0.28	0.29	0.33	0.32	0.38
	BB	0.12	0.16	0.16	0.15	0.13	0.12	0.11	0.10	0.09	0.07	0.08	0.06	0.07	0.08
	IPCA	0.06	0.12	0.14	0.15	0.16	0.16	0.20	0.20	0.23	0.30	0.38	0.45	0.83	1.55
	KNN	0.09	0.13	0.15	0.15	0.16	0.14	0.15	0.17	0.14	0.16	0.17	0.15	0.17	0.18
	RF	0.02	0.03	0.03	0.05	0.06	0.16	0.03	0.22	0.17	0.16	0.16	0.37	0.29	0.40
	MICE	0.09	0.15	0.18	0.24	0.28	0.36	0.32	0.43	0.54	0.50	0.49	0.51	0.73	1.11
	MIPCA	0.06	0.11	0.12	0.13	0.14	0.15	0.18	0.19	0.25	0.32	0.39	0.47	0.80	1.36
	Amelia	0.05	0.09	0.10	0.11	0.13	0.13	0.15	0.16	0.15	0.18	0.19	0.25	0.31	0.42
Kurtosis	**LI**	0.67	1.08	1.32	1.58	1.82	2.12	2.13	2.64	2.93	3.45	3.74	4.72	4.53	5.47
	LOCF	0.81	1.03	1.35	1.84	2.38	5.35	5.29	5.79	9.23	7.06	6.23	6.95	10.37	13.98
	MSSA	0.57	0.80	0.87	0.96	1.00	1.11	1.03	1.16	1.14	1.18	1.28	1.42	1.44	1.63
	BB	0.39	0.56	0.51	0.50	0.45	0.40	0.33	0.33	0.29	0.25	0.19	0.18	0.14	0.15
	IPCA	0.24	0.51	0.57	0.58	0.62	0.74	1.01	0.89	1.23	1.37	1.62	2.73	4.65	12.08
	KNN	0.39	0.52	0.63	0.68	0.70	0.66	0.70	0.75	0.75	0.68	0.71	0.66	0.62	0.60
	RF	0.01	0.06	0.07	0.19	0.19	0.68	0.16	0.88	0.71	0.67	0.94	3.18	1.39	1.45
	MICE	0.47	0.69	0.84	1.27	1.55	2.82	2.33	3.11	4.48	3.34	2.89	3.13	5.60	8.16
	MIPCA	0.22	0.51	0.54	0.55	0.57	0.70	0.91	0.89	1.44	1.42	1.78	2.66	4.11	9.59
	Amelia	0.19	0.36	0.36	0.41	0.48	0.54	0.63	0.56	0.68	0.73	0.73	0.86	1.40	2.81

J.4 Average number of principal components used by IPCA and MIPCA for imputation of MCAR data in the first column of historical data based on the graphical Lasso

	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
IPCA	3	3	3	3	3	3	3	3	3	3	3	3	3	3
MIPCA	2	2	2	2	2	2	2	2	2	2	2	2	2	2

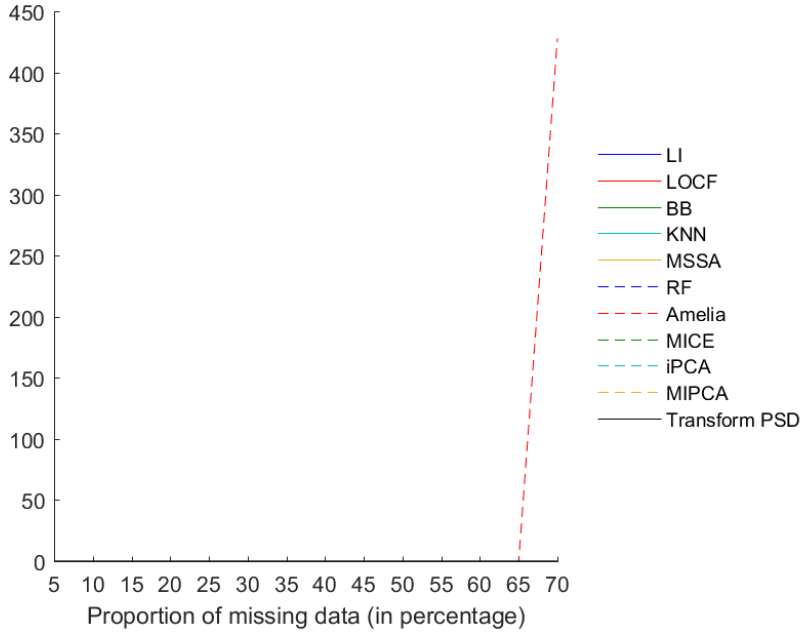
J.5 Average MAE and RMSE between the return of the imputed data from historical sample containing MCAR data (only in the first column) and the original historical sample based on the graphical Lasso, according to the missingness probability



J.6 Standard deviation of the proximity measures among all missingness scenarios, with MCAR data in the first column of historical data based on the graphical Lasso

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
MAE (10^{-3})	**LI**	4.48	3.05	2.25	2.06	1.73	1.64	1.42	1.18	1.25	1.12	0.89	0.77	0.77	0.72
	LOCF	7.11	4.90	3.60	3.19	2.88	2.42	2.25	2.22	1.77	1.78	1.74	1.82	1.58	1.36
	MSSA	5.80	3.77	2.80	2.56	2.05	2.10	1.64	1.59	1.41	1.43	1.08	1.15	1.11	0.85
	BB	2.15	1.07	0.85	1.01	1.28	1.74	1.98	2.49	2.69	3.81	4.19	5.56	5.38	7.04
	IPCA	2.40	1.64	1.24	1.10	0.93	0.95	0.72	0.75	0.84	1.05	1.38	2.25	2.88	4.43
	KNN	2.83	1.75	1.30	1.04	0.88	0.92	0.70	0.78	0.70	0.61	0.63	0.55	0.52	0.50
	RF	1.15	1.15	0.97	1.17	1.19	1.02	0.57	1.14	0.89	0.78	0.83	1.22	1.14	1.47
	MICE	2.88	1.92	1.53	1.38	1.19	1.19	1.09	1.12	1.04	1.16	1.25	1.26	1.52	1.97
	MIPCA	1.27	0.92	0.69	0.62	0.55	0.53	0.48	0.48	0.49	0.48	0.64	1.06	1.05	2.01
	Amelia	1.44	0.93	0.71	0.61	0.51	0.54	0.47	0.52	0.55	0.73	0.82	1.09	1.98	43.3
RMSE (10^{-3})	**LI**	7.01	4.86	3.57	3.24	2.70	2.64	2.41	1.97	1.92	1.77	1.39	1.30	1.19	1.02
	LOCF	10.79	8.11	6.46	5.86	5.23	4.95	4.29	4.34	3.57	3.85	3.69	4.19	3.85	3.50
	MSSA	7.50	4.85	3.49	3.25	2.59	2.88	2.15	2.11	1.87	1.86	1.53	1.57	1.56	1.24
	BB	3.13	1.56	1.09	1.17	1.46	2.00	2.33	2.92	3.23	4.56	5.08	6.79	6.56	8.68
	IPCA	2.92	2.05	1.57	1.45	1.23	1.34	1.06	1.06	1.38	1.75	2.40	3.77	4.73	7.29
	KNN	4.02	2.58	1.95	1.63	1.36	1.43	1.14	1.27	1.19	0.98	1.13	0.95	1.02	1.04
	RF	2.21	1.89	1.30	2.76	2.47	2.84	0.76	2.80	2.60	2.30	1.50	4.65	2.84	3.11
	MICE	5.51	4.47	3.84	3.49	3.27	3.39	2.81	3.07	2.85	2.90	2.78	3.07	3.43	3.53
	MIPCA	1.66	1.19	0.91	0.86	0.78	0.77	0.68	0.71	0.76	1.01	1.26	2.10	2.19	4.01
	Amelia	1.89	1.26	0.97	0.88	0.80	0.86	0.72	0.82	0.91	1.25	1.30	1.90	4.59	653

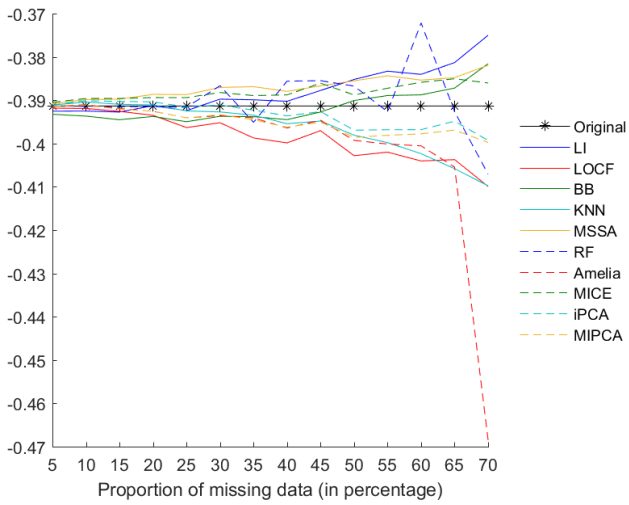
J.7 Average covariance matrices differences, according to Frobenius norm, based on original historical returns and the imputed returns from historical data based on the graphical Lasso containing MCAR data (only in the first column) according to the missingness probability



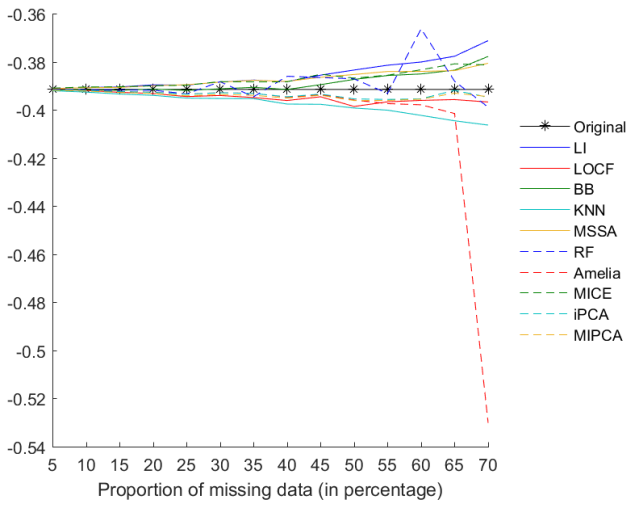
J.8 Standard deviation of the covariance matrix differences among all missingness scenarios, with MCAR data in the first column of historical data based on the graphical Lasso

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
Cov (10^{-3})	**LI**	0.27	0.38	0.43	0.51	0.54	0.62	0.65	0.63	0.68	0.68	0.61	0.60	0.60	0.55
	LOCF	0.25	0.31	0.39	0.44	0.51	0.60	0.59	0.58	0.67	0.74	0.66	0.76	0.88	0.87
	MSSA	0.34	0.42	0.44	0.49	0.55	0.58	0.62	0.66	0.86	0.85	0.80	0.88	0.99	0.90
	BB	0.23	0.26	0.16	0.31	0.67	1.25	1.80	2.26	2.72	3.59	4.45	6.16	6.26	8.74
	IPCA	0.16	0.24	0.28	0.31	0.35	0.41	0.51	0.48	0.61	0.78	0.75	0.99	1.18	1.64
	KNN	0.18	0.24	0.25	0.24	0.23	0.24	0.22	0.17	0.16	0.10	0.11	0.15	0.17	0.29
	RF	0.03	0.06	0.07	0.09	0.13	0.24	0.12	0.41	0.44	0.50	0.43	0.51	0.55	0.65
	MICE	0.16	0.24	0.28	0.33	0.35	0.43	0.41	0.47	0.51	0.61	0.50	0.56	0.45	0.74
	MIPCA	0.15	0.21	0.22	0.24	0.25	0.28	0.28	0.25	0.36	0.52	0.49	0.61	0.69	1.25
	Amelia	0.13	0.20	0.23	0.25	0.27	0.32	0.32	0.28	0.34	0.31	0.29	0.57	92.07	4143023.79

J.9 Average 10-day risk measures, computed from historical sample based on the graphical Lasso containing MCAR data (only in the first column) according to the missingness probability



(a) Average of the 10-day VaR at 99%



(b) Average of the 10-day ES at 97.5%

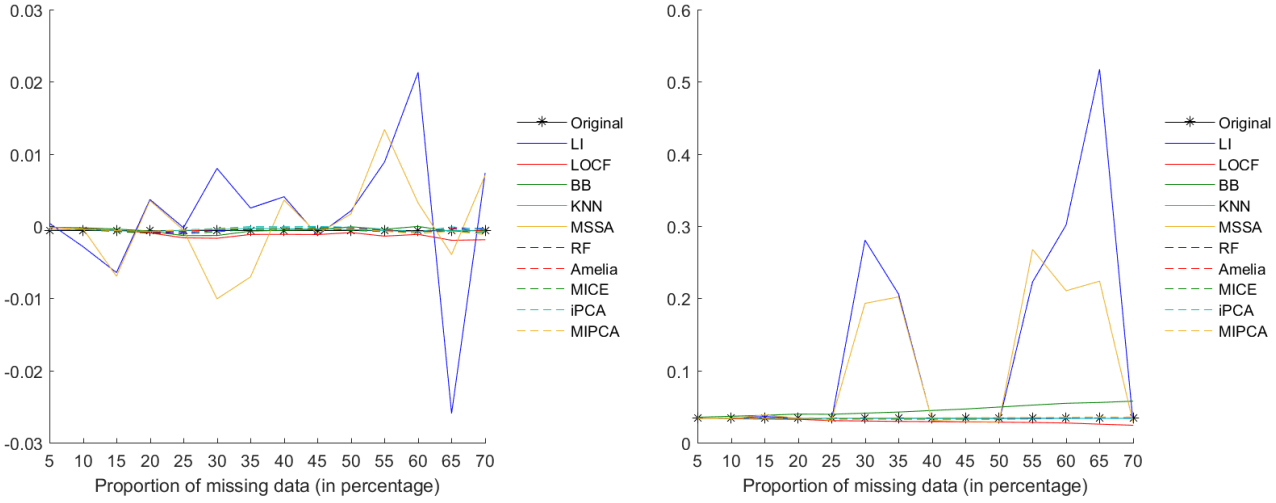
J.10 Standard deviation of each risk measures among all missingness scenarios, with MCAR data in the first column of historical data based on the graphical Lasso

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
$VaR_{99\%}^{1day}$ (10^{-2})	**LI**	1.47	1.71	1.80	1.72	1.64	1.81	1.71	1.35	1.42	1.33	1.33	1.19	1.31	1.18
	LOCF	0.91	1.00	1.18	1.28	1.40	1.81	1.98	2.09	2.02	2.25	2.10	2.02	2.26	2.52
	MSSA	1.06	1.20	1.32	1.32	1.31	1.48	1.18	1.25	1.31	1.20	1.10	1.10	1.25	1.24
	BB	0.93	0.94	0.91	0.89	0.90	1.07	1.08	1.10	1.19	1.39	1.48	1.78	1.68	2.12
	IPCA	0.72	0.83	1.04	1.13	1.19	1.30	1.37	1.42	1.37	1.51	1.33	1.43	1.47	2.31
	KNN	0.38	0.45	0.74	0.76	0.84	0.79	0.87	1.14	0.96	1.11	1.18	1.53	1.52	1.45
	RF	0.00	0.00	0.00	0.00	0.95	0.91	0.00	1.52	1.11	1.71	1.40	1.31	1.59	2.51
	MICE	0.84	1.05	1.27	1.38	1.46	1.63	1.55	1.56	1.45	1.62	1.46	1.72	1.74	2.15
	MIPCA	0.64	0.69	0.81	0.87	0.87	0.95	0.96	0.90	0.93	1.15	0.88	1.13	1.30	2.12
	Amelia	0.38	0.41	0.57	0.69	0.75	0.87	0.94	1.08	0.98	1.12	1.04	1.23	1.28	3.29
$ES_{99\%}^{1day}$ (10^{-2})	**LI**	0.58	0.78	0.86	0.90	0.91	1.06	1.03	1.02	1.05	1.02	0.98	0.95	1.05	0.98
	LOCF	0.38	0.54	0.64	0.73	0.78	0.90	0.91	1.10	1.07	1.33	1.12	1.13	1.14	1.14
	MSSA	0.62	0.82	0.88	0.94	0.93	1.07	0.97	1.08	1.09	0.96	0.85	0.98	1.00	1.05
	BB	0.49	0.73	0.83	0.90	0.95	1.09	1.09	1.14	1.22	1.41	1.45	1.77	1.66	2.07
	IPCA	0.24	0.34	0.41	0.46	0.49	0.55	0.62	0.54	0.67	0.79	0.63	0.90	1.13	1.88
	KNN	0.22	0.31	0.38	0.41	0.42	0.40	0.41	0.45	0.45	0.50	0.55	0.58	0.57	0.60
	RF	0.00	0.02	0.02	0.02	0.16	0.56	0.16	0.64	0.67	0.68	0.46	0.86	0.62	1.08
	MICE	0.33	0.50	0.59	0.73	0.82	0.87	0.80	0.93	0.94	1.20	1.03	1.32	1.19	1.67
	MIPCA	0.23	0.34	0.40	0.44	0.48	0.56	0.60	0.54	0.73	0.89	0.75	0.97	1.15	1.80
	Amelia	0.16	0.28	0.33	0.38	0.43	0.52	0.48	0.52	0.59	0.63	0.56	0.79	1.02	31.4

		Missingness proportion													
		5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
$VaR^{10days}_{97.5\%}$ (10^{-2})	**LI**	0.69	1.03	1.23	1.44	1.56	1.70	1.70	1.84	1.89	1.91	1.81	2.36	2.74	3.01
	LOCF	0.61	0.82	1.05	1.35	1.56	1.66	1.95	2.31	2.47	3.04	2.99	3.62	3.59	4.14
	MSSA	0.38	0.54	0.77	0.92	0.96	0.97	1.05	1.21	1.25	1.22	1.16	1.38	1.64	1.58
	BB	0.65	0.88	1.00	1.16	1.23	1.35	1.29	1.51	1.53	1.69	1.65	2.00	2.35	2.68
	IPCA	0.12	0.28	0.30	0.34	0.52	0.59	0.91	0.97	1.04	1.18	1.25	1.78	2.20	3.15
	KNN	0.15	0.25	0.34	0.37	0.54	0.63	0.85	0.94	0.93	1.09	1.04	1.27	1.42	1.49
	RF	0.00	0.03	0.04	0.05	0.07	0.38	0.46	0.46	0.35	0.41	0.58	0.94	1.47	1.69
	MICE	0.28	0.42	0.51	0.58	0.73	0.90	0.91	1.20	1.32	1.40	1.52	2.10	2.54	2.87
	MIPCA	0.12	0.32	0.36	0.42	0.53	0.60	0.83	0.94	1.05	1.35	1.43	1.89	2.31	3.10
	Amelia	0.16	0.36	0.44	0.53	0.71	0.67	0.82	0.86	0.92	1.03	0.97	1.19	2.27	4300
$ES^{10days}_{97.5\%}$ (10^{-2})	**LI**	0.44	0.58	0.73	0.89	1.01	1.14	1.13	1.41	1.49	1.52	1.54	2.00	2.50	2.79
	LOCF	0.44	0.64	0.85	1.04	1.20	1.46	1.70	1.99	2.19	2.63	2.61	3.16	3.38	4.18
	MSSA	0.34	0.47	0.61	0.72	0.78	0.79	0.77	1.03	1.05	1.14	1.01	1.25	1.52	1.52
	BB	0.40	0.56	0.69	0.84	0.94	1.05	1.06	1.33	1.39	1.49	1.49	1.87	2.29	2.57
	IPCA	0.13	0.23	0.30	0.35	0.41	0.45	0.59	0.69	0.84	0.94	1.02	1.58	1.97	2.96
	KNN	0.14	0.19	0.27	0.28	0.37	0.41	0.47	0.60	0.60	0.62	0.70	0.80	0.98	0.98
	RF	0.02	0.07	0.07	0.08	0.15	0.35	0.28	0.39	0.51	0.39	0.63	1.21	1.22	1.43
	MICE	0.28	0.37	0.47	0.55	0.66	0.93	0.89	1.17	1.44	1.32	1.50	2.18	2.55	2.82
	MIPCA	0.13	0.24	0.30	0.37	0.43	0.51	0.65	0.79	1.01	1.16	1.23	1.78	2.27	2.98
	Amelia	0.12	0.20	0.26	0.31	0.37	0.41	0.52	0.55	0.66	0.73	0.76	0.96	1.98	59.7

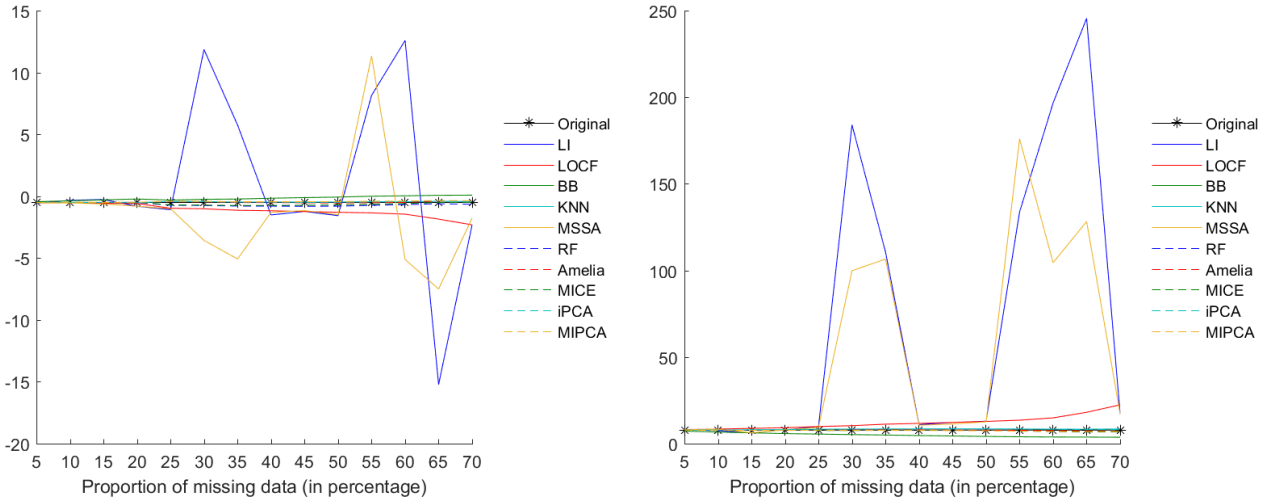
Appendix K: Impact on a sample based on a heuristic approach

K.1 The first four statistical moments of the returns of the imputed data based on a matrix containing MAR data (successive missing data at the end of the first series of historical sample based on heuristic approach) according to the missingness proportion



(a) Mean

(b) Standard deviation

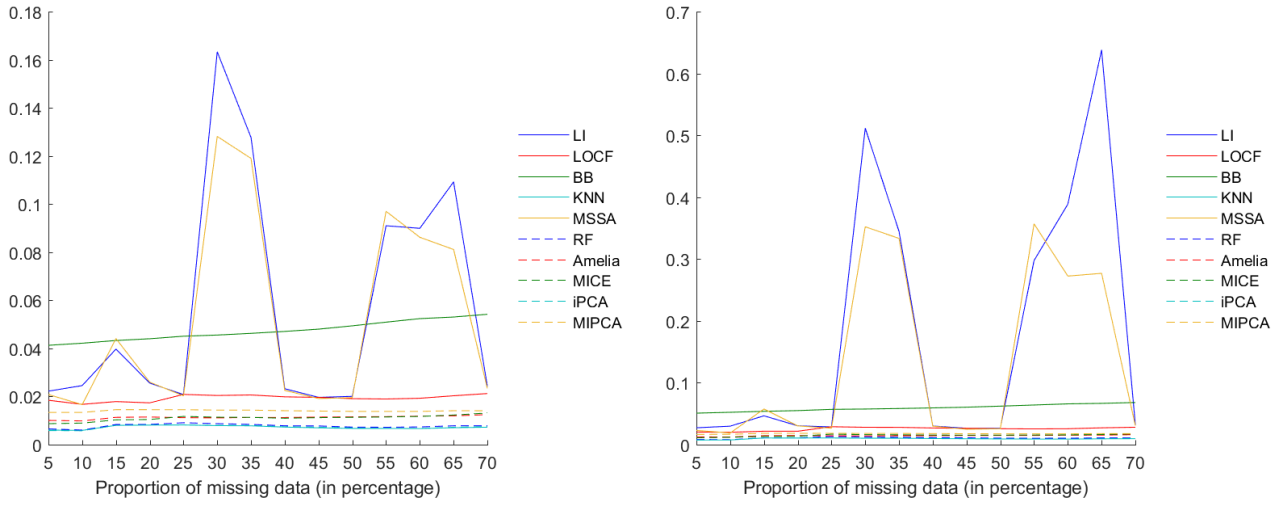


(c) Skewness *(d)* Kurtosis

K.2 Number of principal components used by IPCA and MIPCA for imputation of MAR data corresponding to successive missing data at the end of the first series

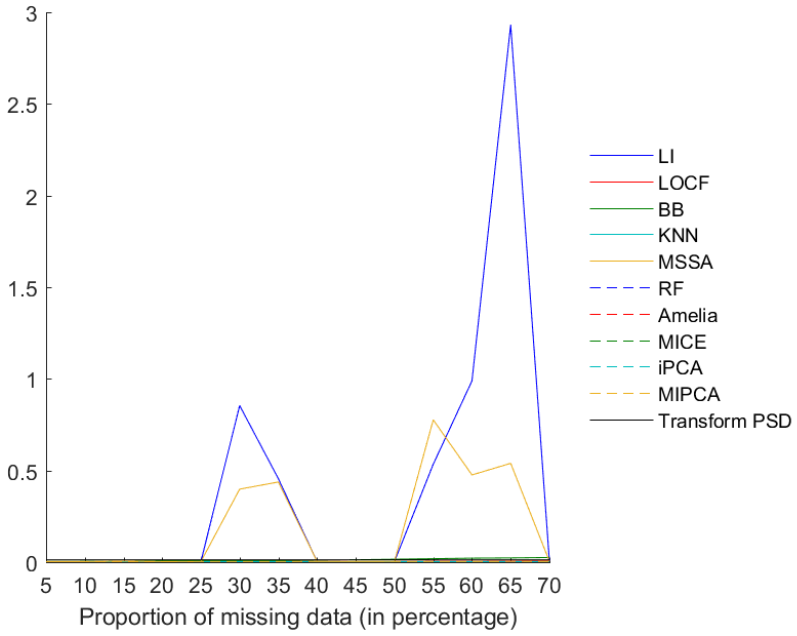
	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
IPCA	1	1	1	1	1	1	1	1	1	1	1	1	1	1
MIPCA	1	1	1	1	1	1	1	1	1	1	1	1	1	1

K.3 MAE and RMSE between the return of the imputed data from a matrix containing MAR data (successive missing data at the end of the first series of historical sample based on heuristic approach) and the original data matrix, according to the missingness proportion

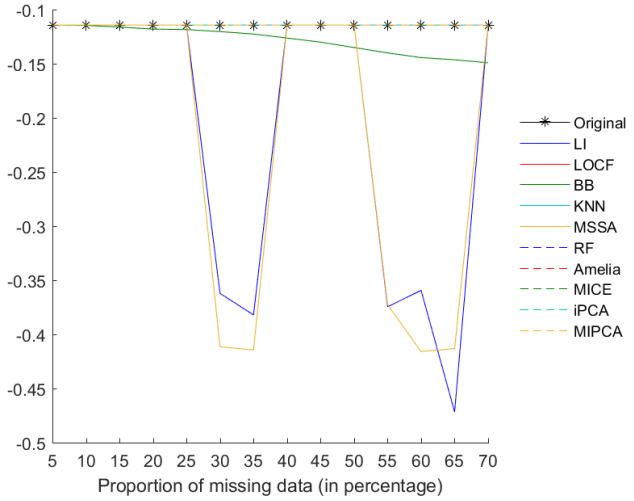


(a) MAE *(b)* RMSE

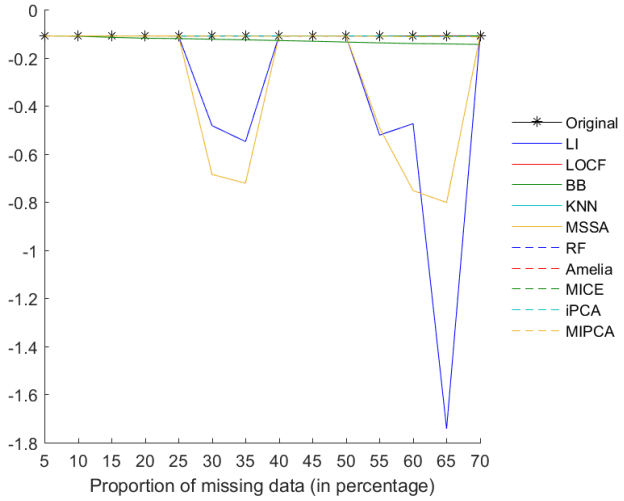
K.4 Covariance matrices differences, according to Frobenius norm, based on original returns and the imputed returns from a matrix containing MAR data (successive missing data at the end of the first series of historical sample based on heuristic approach) according to the missingness proportion



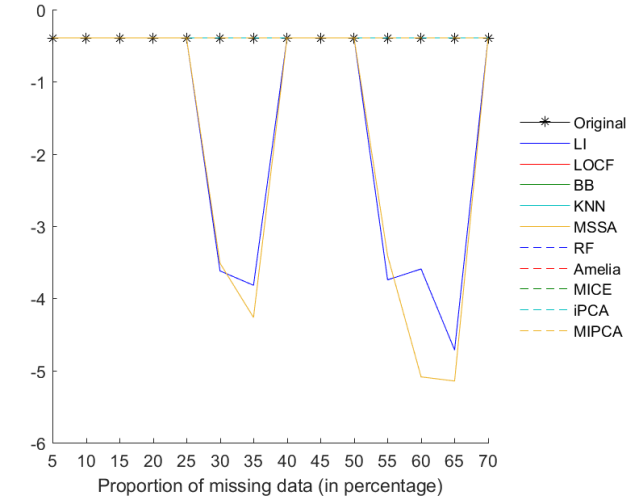
K.5 The 1-day risk measures and 10-day risk measures, computed from a matrix containing MAR data (successive missing data at the end of the first series of historical sample based on heuristic approach) according to the missingness proportion



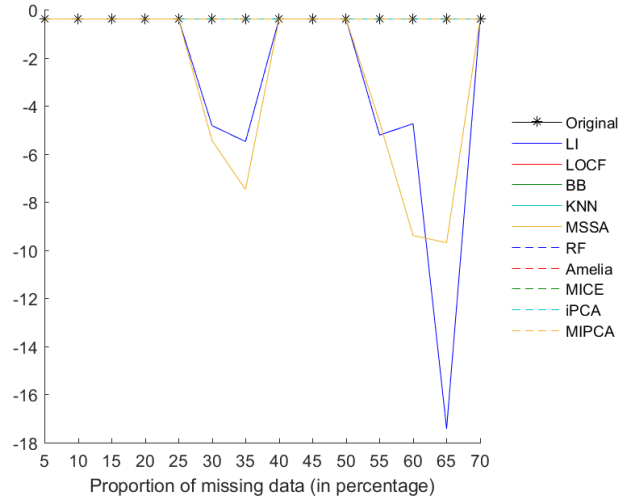
(a) The 1-day VaR at 99%



(b) The 1-day ES at 97.5%



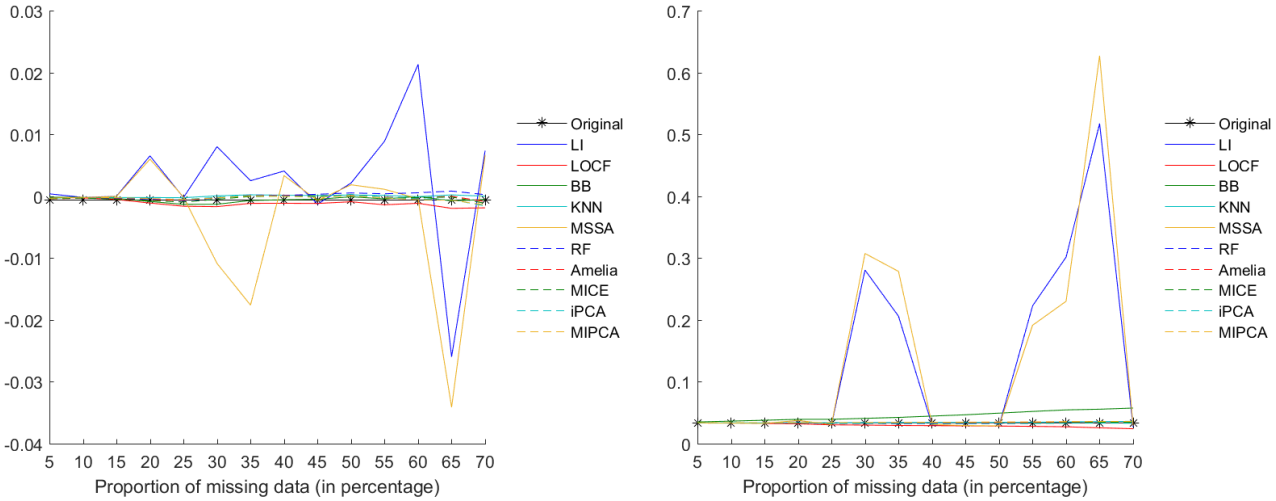
(c) The 10-day VaR at 99%



(d) The 10-day ES at 97.5%

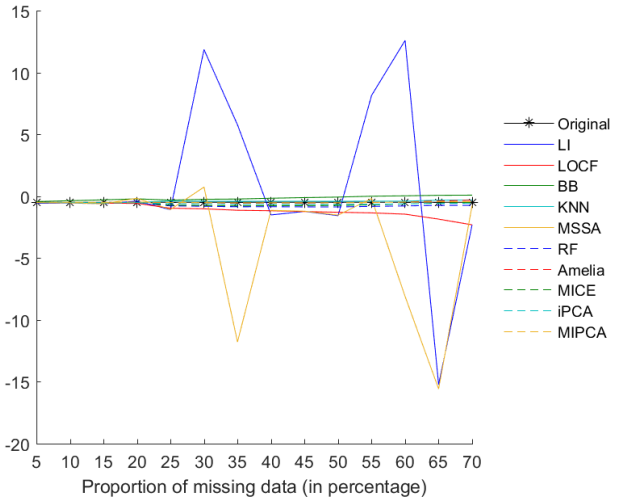
Appendix L: Impact on a sample based on the graphical Lasso

L.1 The first four statistical moments of the returns of the imputed data based on a matrix containing MAR data (successive missing data at the end of the first series of historical sample based on graphical Lasso approach) according to the missingness proportion

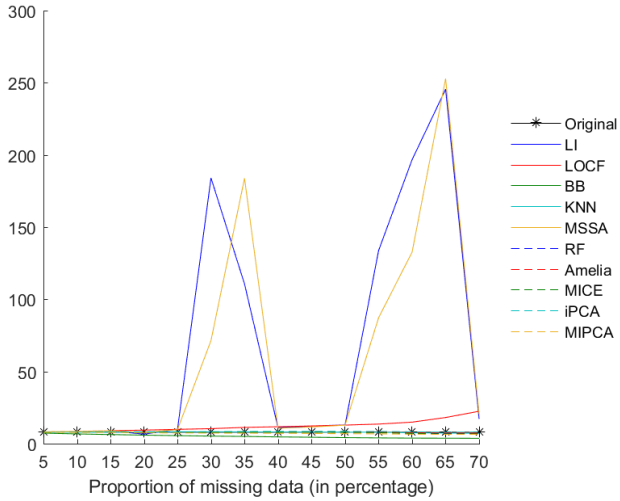


(a) Mean

(b) Standard deviation



(c) Skewness

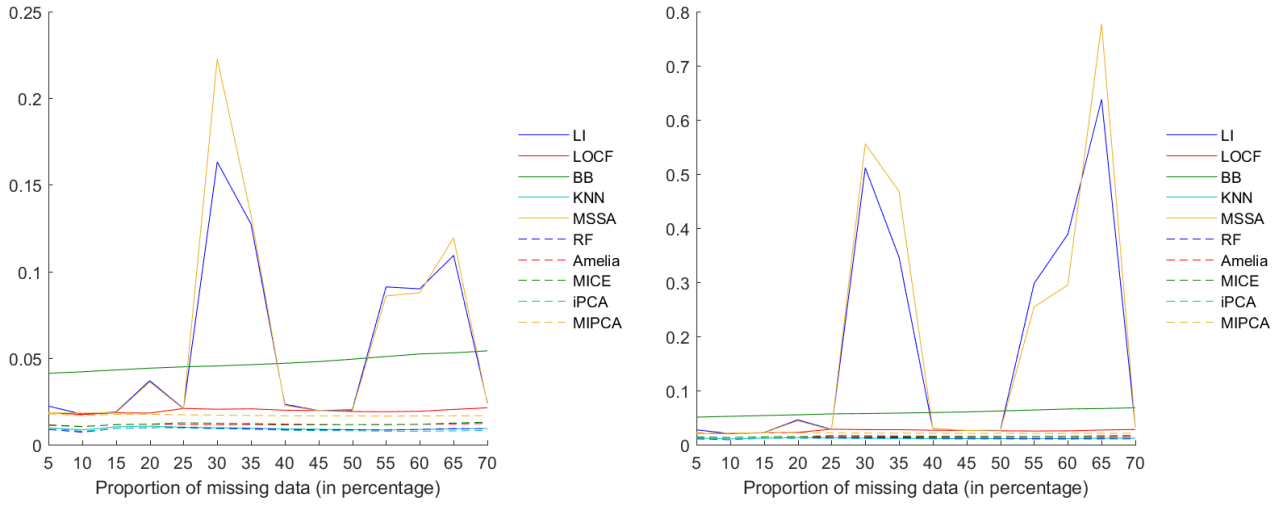


(d) Kurtosis

L.2 Number of principal components used by IPCA and MIPCA for imputation of MAR data corresponding to successive missing data at the end of the first series of historical sample based on graphical Lasso approach

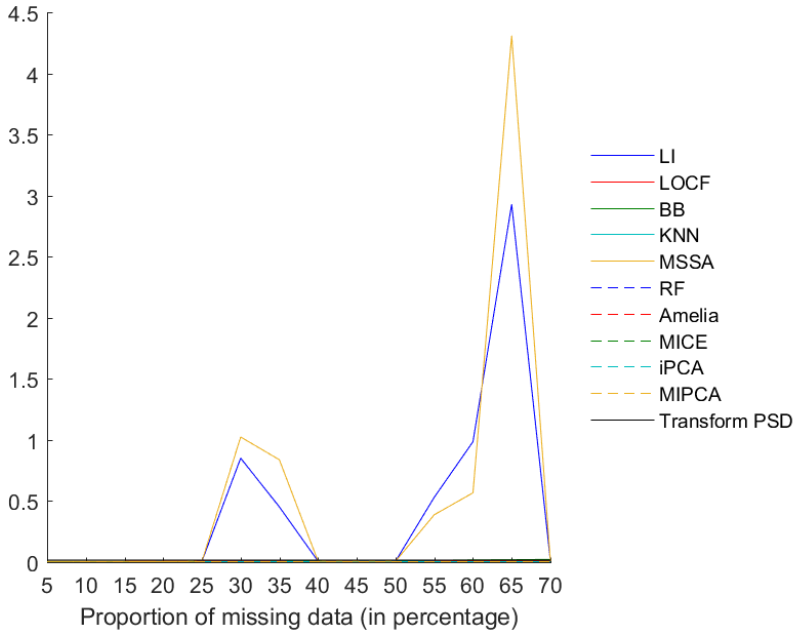
	Missingness proportion													
	5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**	**55%**	**60%**	**65%**	**70%**
IPCA	2	2	2	2	2	2	2	2	2	2	2	2	2	2
MIPCA	2	2	2	2	2	2	2	2	2	2	2	2	2	2

L.3 MAE and RMSE between the return of the imputed data from a matrix containing MAR data (successive missing data at the end of the first series of historical sample based on graphical Lasso approach) and the original data matrix, according to the missingness proportion

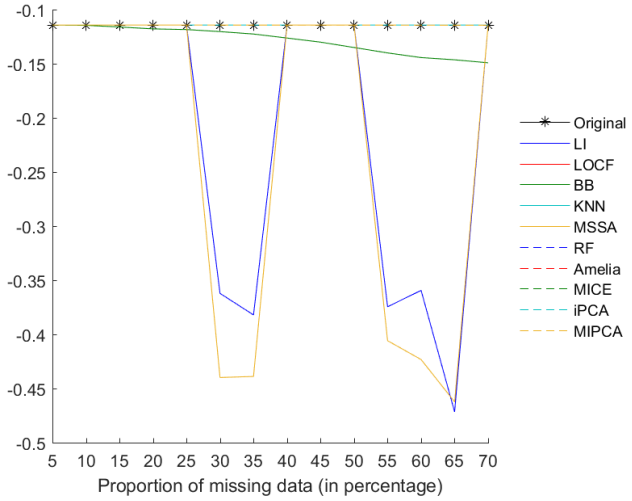


(a) MAE *(b)* RMSE

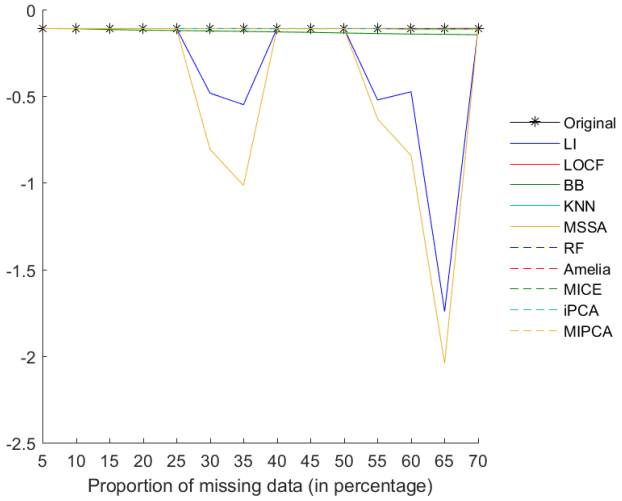
L.4 Covariance matrices differences, according to Frobenius norm, based on original returns and the imputed returns from a matrix containing MAR data (successive missing data at the end of the first series of historical sample based on graphical Lasso approach) according to the missingness proportion



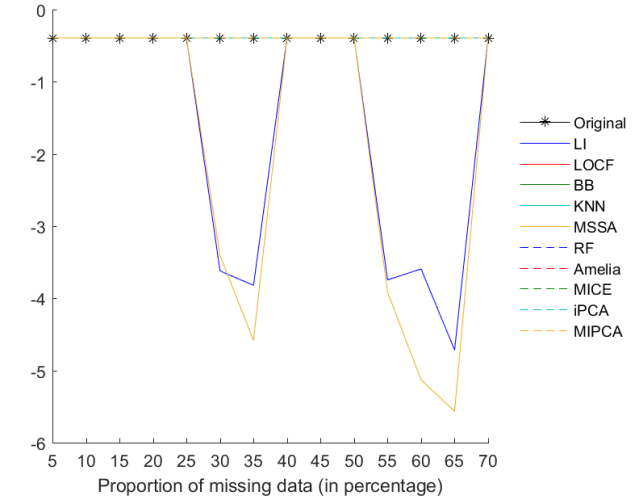
L.5 The 1-day risk measures and 10-day risk measures, computed from a matrix containing MAR data (successive missing data at the end of the first series of historical sample based on graphical Lasso approach) according to the missingness proportion



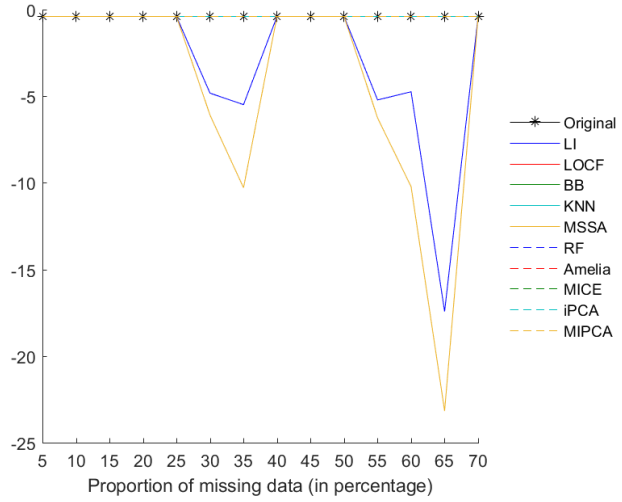
(a) The 1-day VaR at 99%



(b) The 1-day ES at 97.5%



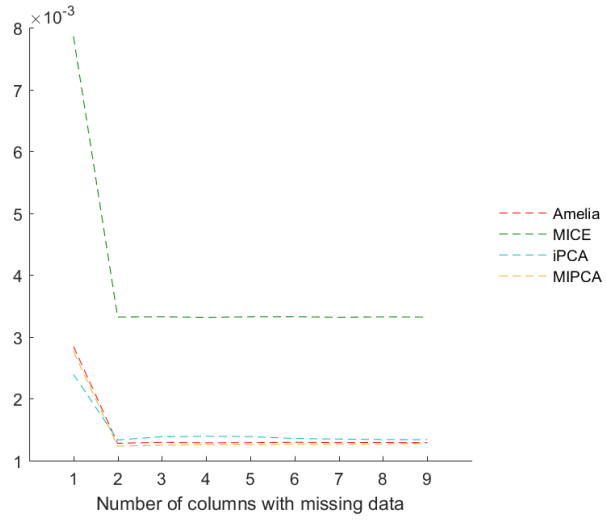
(c) The 10-day VaR at 99%



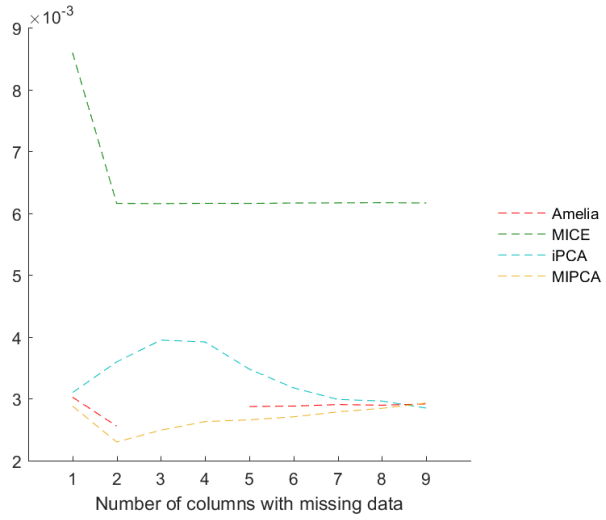
(d) The 10-day ES at 97.5%

Appendix M: Discussion

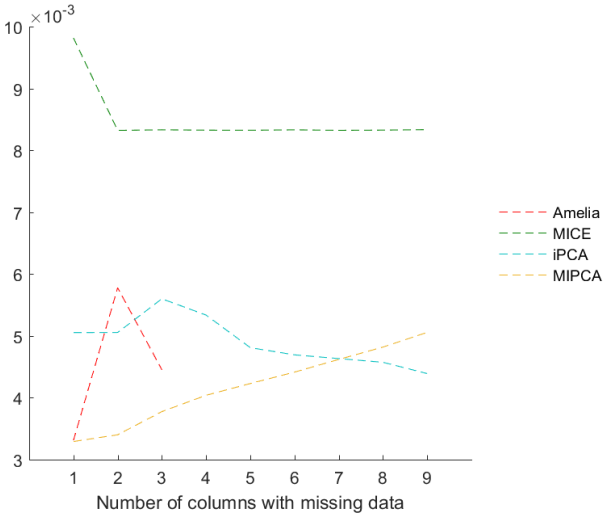
M.1 Average RMSE (of the first column) obtained according to the number of columns containing 10%, 30%, 50% and 70% of missing data on the simulated sample



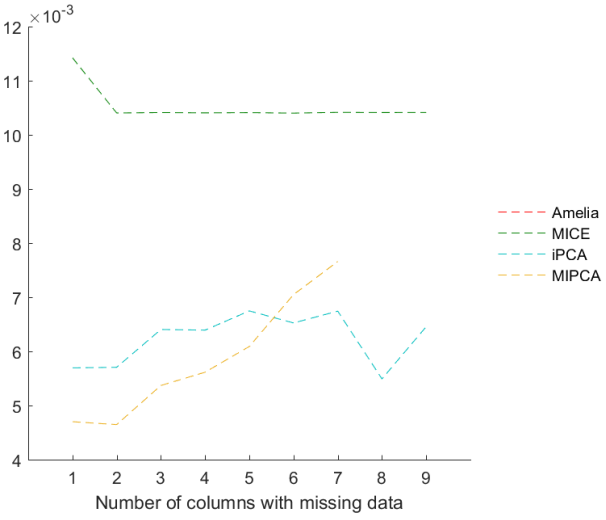
(a) 10% missing data



(b) 30% missing data

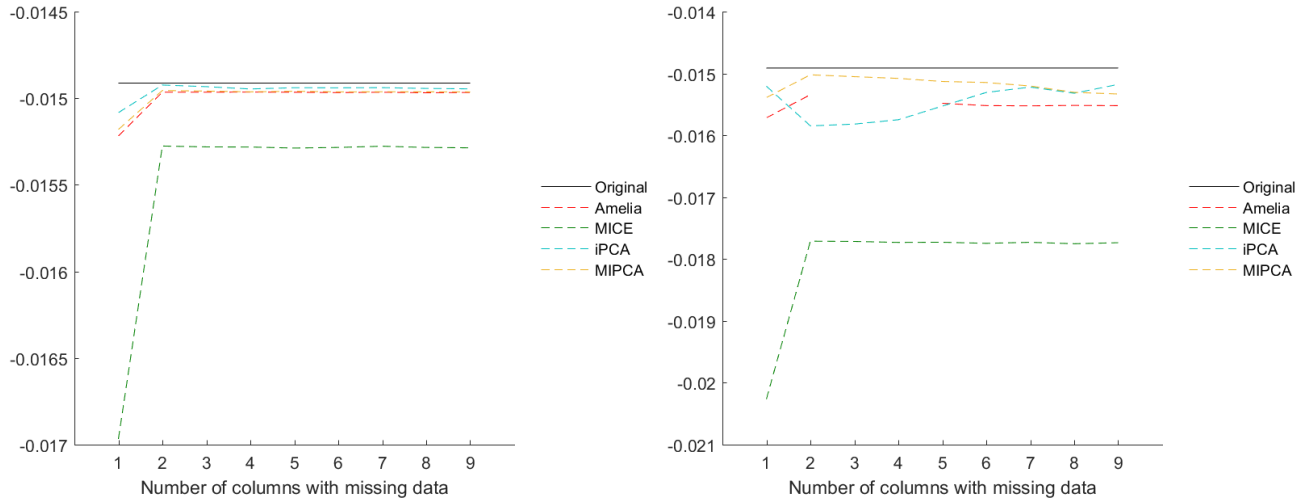


(c) 50% missing data



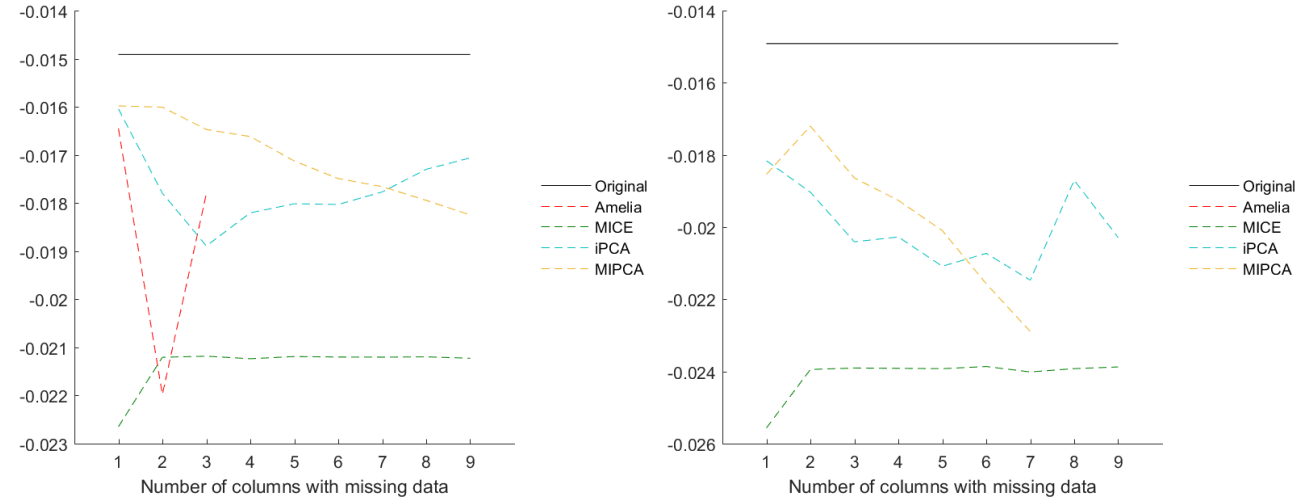
(d) 70% missing data

M.2 Average 1-day ES (of the first column) obtained according to the number of columns containing 10%, 30%, 50% and 70% of missing data on the simulated sample



(a) 10% missing data

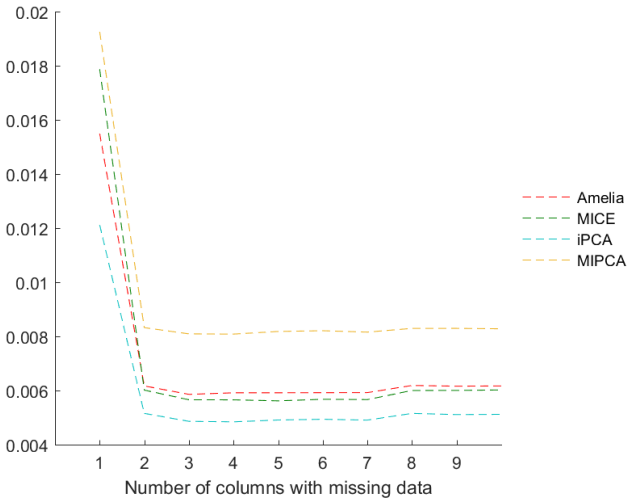
(b) 30% missing data



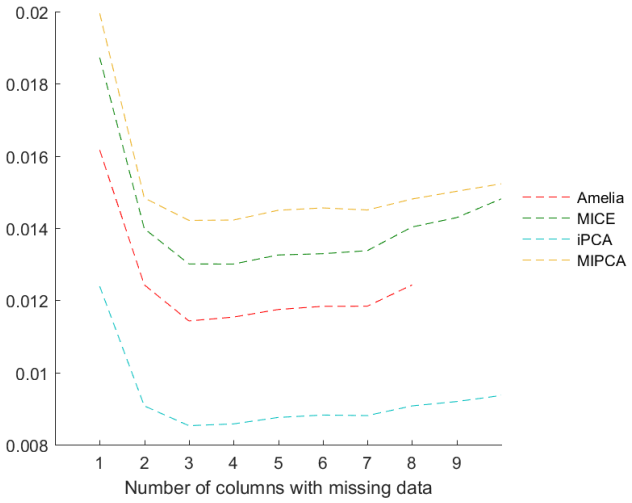
(c) 50% missing data

(d) 70% missing data

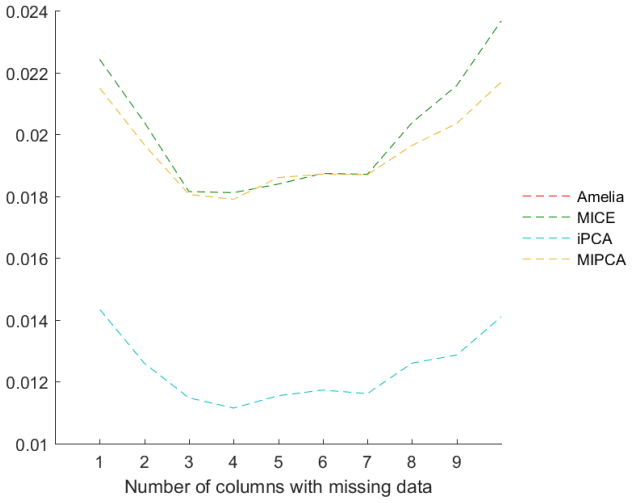
M.3 Average RMSE (of the first column) obtained according to the number of columns containing 10%, 30%, 50% and 70% of missing data on the heuristic historical sample



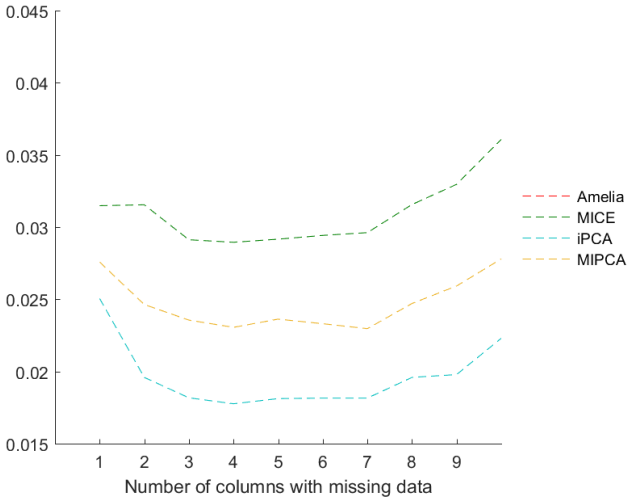
(a) 10% missing data



(b) 30% missing data

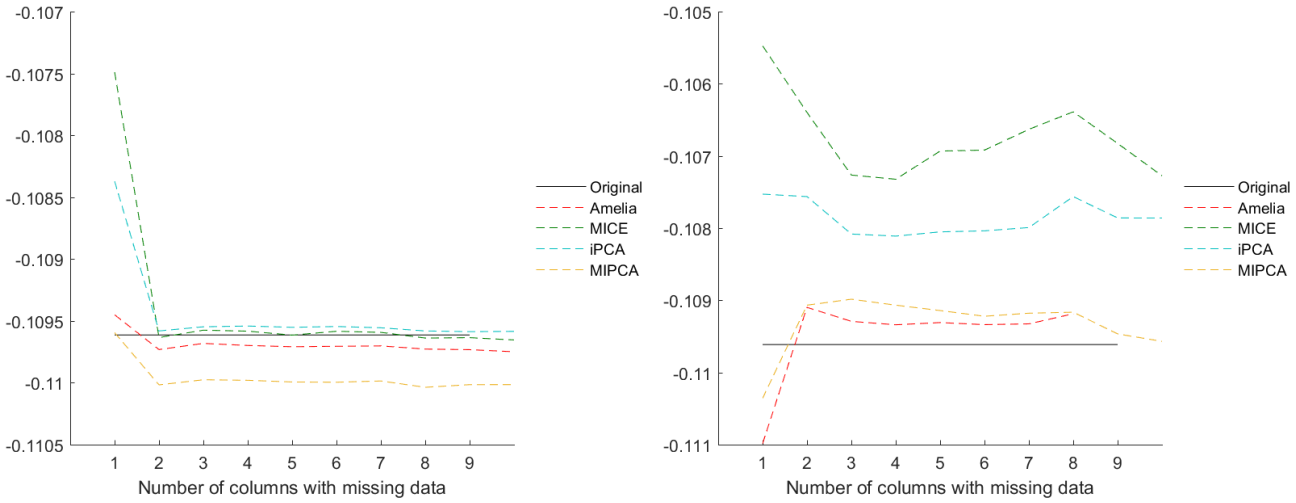


(c) 50% missing data



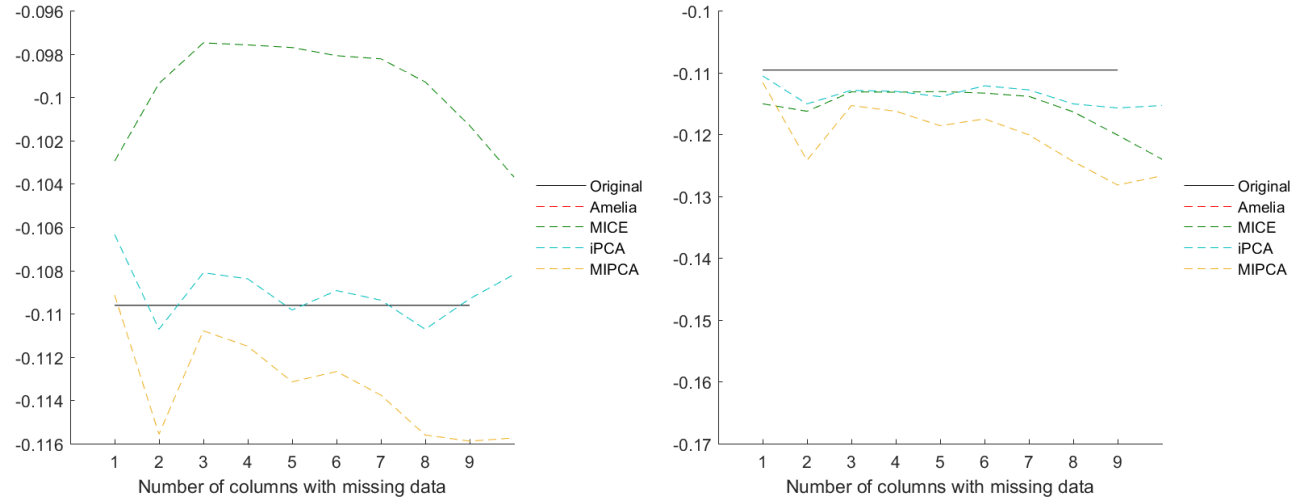
(d) 70% missing data

M.4 Average 1-day ES (of the first column) obtained according to the number of columns containing 10%, 30%, 50% and 70% of missing data on the heuristic historical sample



(a) 10% missing data

(b) 30% missing data



(c) 50% missing data

(d) 70% missing data

Résumé

Face à un contexte réglementaire toujours plus contraignant, les banques ont dû s'adapter en mettant en œuvre de nombreux moyens afin de répondre aux exigences liées à la qualité de la donnée. En effet, le régulateur a pris conscience que l'évaluation des risques ne pouvait se faire sans des données de bonne qualité, c'est pourquoi les réglementations récentes traitent de plus en plus, directement ou indirectement, de la gestion des données et notamment de la gestion des données manquantes. C'est pourquoi les banques s'intéressent aux méthodes d'imputation, à leur efficacité et surtout à leur impact sur la gestion du risque. Ainsi, cette thèse présente les enjeux généraux liés aux données manquantes, avant de se focaliser sur le cas financier et, en particulier, sur les implications réglementaires, pour ensuite mener une analyse comparative basée sur plusieurs critères (moments statistiques, mesures de proximité, matrice de covariance, mesures de risque et temps de calcul). L'analyse comparative est effectuée à partir de nombreuses méthodes d'imputation, telles que l'interpolation linéaire ou le LOCF, qui seront utilisées comme modèles de référence, ainsi que du pont brownien, des K-NN, MSSA, des forêts aléatoires, Amelia, MICE, IPCA et MIPCA. Ainsi, ces méthodes ont été appliquées à des échantillons simulés où les données ont été supprimées de l'échantillon selon un mécanisme MCAR, MAR ou MNAR, mais également à des échantillons historiques avec des données MCAR. Si cette analyse comparative révèle des résultats particulièrement satisfaisants pour l'algorithme d'Amelia et les forêts aléatoires, elle révèle également de nombreux points critiques pour les banques mais aussi pour le régulateur.

Mots clefs : données manquantes, imputation de données, gestion du risque, analyse comparative, données financières

Summary

In view of the increasingly stringent regulatory context, banks have had to adapt by implementing many methods to meet data quality requirements. Regulators understand that risk assessment is impossible without good-quality data, which is why recent regulations increasingly deal, be it explicitly or implicitly, with data management and with the management of missing data in particular. For this reason, banks are interested in imputation methods, their efficiency and in particular, their impact on risk management. Accordingly, the PhD thesis presents general issues that are related to missing data before focusing on the financial framework and regulation. Then, a comparative analysis that is based on several criteria (statistical moments, proximity measures, covariance matrix, risk measures and computation time) is conducted. The analysis compares numerous imputation methods, such as linear interpolation and LOCF, which are used as benchmarks, and the Brownian bridge, K-NN, MSSA, random forests, Amelia, MICE, IPCA and MIPCA. These methods have been applied to simulated samples whereby data are removed in line with an MCAR, MAR or MNAR mechanism and to historical samples with MCAR data. The comparative analysis produces particularly satisfactory results for the Amelia algorithm and for random forests. It also reveals many critical issues for banks and regulators.

Keywords: missing data, data imputation, risk management, comparative analysis, financial data