Some properties of the correlation between the high-frequency financial assets

Nicolas Huth

Thèse de doctorat de l'école centrale paris

Spécialité :
Finance Quantitative

Laboratoire :
Mathématiques Appliquées aux Systèmes

Présentée par :
Nicolas HUTH

En vue d'obtenir le grade de
Docteur de l'Ecole Centrale Paris

Quelques propriétés de la corrélation entre les actifs financiers à haute fréquence

Sous la direction de Frédéric ABERGEL

Rapporteurs : Fabrizio LILLO
Mathieu ROSENBAUM

Thèse soutenue publiquement le 14 décembre 2012 devant le jury composé de

Frédéric ABERGEL	Ecole Centrale Paris	Directeur
Fabrizio LILLO	Scuola Normale Superiore di Pisa	Rapporteur
Mathieu ROSENBAUM	Université Pierre et Marie Curie	Rapporteur
Emmanuel BACRY	Ecole Polytechnique	Examinateur
Michel CROUHY	Natixis	Examinateur
Matteo MARSILI	International Centre for Theoretical Physics	Examinateur
Anirban CHAKRABORTI	Ecole Centrale Paris	Examinateur invité

Remerciements

Je tiens à remercier en premier lieu Frédéric Abergel, mon directeur de thèse, sans qui ce manuscrit n'aurait jamais vu le jour. Frédéric a toujours été présent pour me guider lors de ce périple scientifique. Ses conseils avisés m'ont permis de progresser constamment. Par ailleurs, je le remercie de m'avoir accordé sa confiance pour promouvoir mes travaux dans des conférences plus qu'idéalement localisées. Je suis heureux d'avoir croisé le chemin d'une personne aussi bien scientifiquement qu'humainement irréprochable. J'exprime ma sincère gratitude envers Michel Crouhy, directeur de la recherche et du développement à Natixis, pour m'avoir recruté dans son équipe afin d'y effectuer ma thèse. J'ai ainsi bénéficié d'un cadre de travail épanouissant et de ressources sans lesquelles je n'aurais pu obtenir certains résultats de cette thèse. Merci à Michel pour cette inlassable envie de construire un échange entre la recherche académique et le monde professionnel. Lors de ces trois années passées au sein de Natixis, j'ai eu la chance d'être encadré par Adil Reghaï, directeur de l'équipe de recherche quantitative Equity Markets. Il a cru en moi au commencement de cette thèse et m'a donné l'opportunité de faire mes preuves. Je lui suis profondément reconnaissant pour sa disponibilité sans faille et ses conseils toujours éclairés. La curiosité scientifique dont il fait preuve en a fait un interlocuteur privilégié pendant toutes ces années.

Je souhaite remercier chaleureusement Fabrizio Lillo et Mathieu Rosenbaum pour avoir accepté d'être rapporteurs de cette thèse. Leurs remarques judicieuses ont largement contribué à améliorer ce manuscrit. Je suis à la fois ravi et honoré qu'Emmanuel Bacry, Michel Crouhy et Matteo Marsili fassent partie de mon jury de thèse car leurs travaux furent une source d'inspiration indéniable pour mes recherches. Merci à Mathieu et à Emmanuel pour les nombreux séminaires FIESTA, agréables autant pour les neurones que pour les papilles. Ils m'ont permis de mieux appréhender les problématiques de la microstructure et de rencontrer les confirmés et les novices de la recherche en finance. Un grand merci également à Anirban, examinateur invité, pour l'organisation de cet inoubliable voyage à Kolkata.

Ces remerciements ne seraient évidemment pas dignes de ce nom si j'oubliais mes compagnons du "stat arb" de Natixis qui ont partagé mes errements de thésard au sein d'une banque. Ils m'ont tous beaucoup appris dans une ambiance toujours joyeuse. Merci donc à Ban Zheng pour sa gentillesse, son humour et pour m'avoir enseigné les fondamentaux de la langue chinoise. Merci à Quentin Amelot pour sa sympathie et son ouverture musicale qui nous a valu nombre de discussions "tsugistiques". Merci à Hugues-Henri Liniger pour son sixième sens du business, ses connaissances du trading et pour être le personnage haut en couleur qu'il est. Merci à Marc Souaille (a.k.a. Le Docteur) pour sa folie douce, sa bonne humeur constante qui détonne agréablement dans une salle des marchés et pour son ouverture d'esprit scientifique. Merci à Mohamed Lakhdar pour les nombreuses remarques pertinentes sur mes travaux et pour son soutien moral, même à travers le brouillard épais d'Erice. Merci à Abbas Msa pour ses anecdotes atypiques et toujours riches en détails techniques. Merci également à Julien Puvilland, Thomas Sévin et Charif Bouchemat pour l'aide qu'ils m'ont apporté durant ces trois années. Ce sera toujours un plaisir de vous retrouver autour d'un verre au Kleemend's ou ailleurs.

Je n'oublie pas que mes premières armes furent faites chez les quants dérivés. C'est pourquoi je désire remercier Majed Abdelhedi, Albert Andinaik, Ousmane Aw, Geoffrey Babiarz, Saad Bahbouhi, Adel Ben

Haj Yedder, Amine Boukhaffa, Gilles Boya, Catherine Collin, Thomas Combarel, Sylvain Corlay, Olivier Croissant, Ghada El Boury, Fatima El Khyari, Eglantine Giraud, Houari Houalef, Laurent Jacquel, Sanae Loulidi, José Luu, Johan Mabille, Marouen Messaoud, Stéphanie Mielnik, Claude Muller, Abdessamad Sahnoun et Emilie Tétard.

J'ai eu le privilège de faire partie de l'équipe FiQuant du laboratoire de Mathématiques Appliquées aux Systèmes de l'Ecole Centrale Paris lors de cette thèse. J'y ai côtoyé des chercheurs brillants et des thésards passionnés. Les groupes de travail et les séminaires organisés par cette équipe ont permis d'élargir mon spectre de connaissances et de garder une ouverture d'esprit nécessaire à la recherche. Je remercie profondément mes compagnons thésards Marouane Anane, Rémy Chicheportiche, Joao De Gama Batista, Sofiène El Aoud, Esteban Guevara, Aymen Jedidi, Mehdi Lallouache, Nicolas Millot, Fabrizio Pomoponio (que de souvenirs impérissables à Kolkata et à Sydney!), Rémi Tachet, Gayatri Tilak et Riadh Zaatour. Merci également aux chercheurs érudits que sont Anirban Chakraborti, Damien Challet, Sophie Laruelle, Ioane Muni Toke, Mauro Politi et Olaf Torné. Chaque dîner d'équipe fut un moment délicieux.

Je n'aurais jamais pu venir à bout de cette thèse sans la bouffée d'air frais que m'apportent mes amis lors de nos soirées passées ensemble à délirer sur tout et n'importe quoi, surtout n'importe quoi. Merci donc à Arnaud et Omérine (et Héloïse désormais!), et à Céline et Julio. Vous me permettez de maintenir le cap et de prendre du recul sur le milieu dans lequel j'évolue. Je suis heureux de vous compter parmi mes amis.

Mes parents et mon frère me soutiennent depuis toujours, je ne serais jamais arrivé au terme de cette thèse sans eux. Je les remercie du plus profond de mon cœur pour ce soutien indéfectible. Je ne saurais terminer ces remerciements sans une pensée pour la femme que j'aime, Sandrine, qui m'accompagne et me rend heureux depuis déjà dix ans. Rien ne serait pareil sans elle.

A mes parents,
pour leur amour sans fin et leur support sans faille

Résumé

Le but de cette thèse est d'approfondir les connaissances académiques sur les variations jointes des actifs financiers à haute fréquence en les analysant sous un point de vue novateur. Nous tirons profit d'une base de données de prix *tick-by-tick* pour mettre en lumière de nouveaux faits stylisés sur la corrélation haute fréquence, et également pour tester la validité empirique de modèles multivariés.

Dans le chapitre 1, nous discutons des raisons pour lesquelles la corrélation haute fréquence est d'une importance capitale pour le trading. Par ailleurs, nous passons en revue la littérature empirique et théorique sur la corrélation à de petites échelles de temps. Puis nous décrivons les principales caractéristiques du jeu de données que nous utilisons. Enfin, nous énonçons les résultats obtenus dans cette thèse.

Dans le chapitre 2, nous proposons une extension du modèle de subordination au cas multivarié. Elle repose sur la définition d'un temps événementiel global qui agrège l'activité financière de tous les actifs considérés. Nous testons la capacité de notre modèle à capturer les propriétés notables de la distribution multivariée empirique des rendements et observons de convaincantes similarités.

Dans le chapitre 3, nous étudions les relations *lead/lag* à haute fréquence en utilisant un estimateur de fonction de corrélation adapté aux données *tick-by-tick*. Nous illustrons sa supériorité par rapport à l'estimateur standard de corrélation pour détecter le phénomène de *lead/lag*. Nous établissons un parallèle entre le *lead/lag* et des mesures classiques de liquidité et révélons un arbitrage pour déterminer les paires optimales pour le trading de *lead/lag*. Enfin, nous évaluons la performance d'un indicateur basé sur le *lead/lag* pour prévoir l'évolution des prix à court terme.

Dans le chapitre 4, nous nous intéressons au profil saisonnier de la corrélation intra-journalière. Nous estimons ce profil sur quatre univers d'actions et observons des ressemblances frappantes. Nous tentons d'incorporer ce fait stylisé dans un modèle de prix *tick-by-tick* basé sur des processus de Hawkes. Le modèle ainsi construit capture le profil de corrélation empirique assez finement, malgré sa difficulté à atteindre le niveau de corrélation absolu.

Abstract

This thesis aims at providing insight into comovements of financial assets at high frequency from an original point of view. We take advantage of a database of tick-by-tick prices to bring to light new stylized facts on high frequency correlation as well as to check the empirical validity of multivariate modelling frameworks.

In chapter 1, we elaborate on the reasons why high frequency correlation is of the utmost importance for trading purposes. We also briefly review the empirical and theoretical literature on correlation at small time scales. Then, we describe the main features of the dataset we use. Finally, we enunciate the results obtained in this thesis.

In chapter 2, we suggest a way of extending the subordination modelling to the multivariate case. This relies on the definition of a global event time that merges the trading activity of all assets under consideration. We test the ability of our model to capture salient features of the empirical multivariate probability distribution of returns and find a convincing agreement.

In chapter 3, we study high frequency lead/lag relationships using a suitable cross-correlation estimator for tick-by-tick data. We show its superiority over the classical correlation estimator in detecting lead/lag patterns. We relate lead/lag to standard liquidity measures and exhibit a trade-off to find optimal pairs for lead/lag trading. Finally, we evaluate the performance of a lead/lag indicator in forecasting the short-term evolution of prices.

In chapter 4, we focus on the intraday correlation seasonal pattern. We estimate this pattern over four universes of stocks and observe striking similarities. We attempt to incorporate this stylized fact into a tick-by-tick price model based upon Hawkes processes. The resulting model captures the empirical profile of correlation quite well, though it doesn't match the absolute level of correlation.

Contents

Chapitre 1

Introduction

1.1 Motivation et objectifs

Cette thèse vise à mieux comprendre le comportement collectif des actifs financiers à haute fréquence. Beaucoup d'efforts ont été fournis par la recherche pour appréhender plus finement le processus de formation du prix d'un seul actif (voir par exemple [26]). *A contrario*, les interdépendances des titres financiers à l'échelle du mouvement élémentaire du prix n'ont été que très peu étudiées. Du point de vue scientifique, il semble plus raisonnable d'aborder un problème complexe par une version simplifiée. Ceci pourrait expliquer pourquoi les propriétés univariées des marchés financiers sont aujourd'hui plus apprivoisées que les aspects multivariés. Cependant, les marchés sont par nature des systèmes hautement corrélés dans l'espace. Parmi les nombreuses corrélations existant sur les marchés, nous pouvons citer ces quelques exemples représentatifs[1]

- des actions appartenant au même indice financier: Total et France Télécom (corrélés à 37%). Ces deux actions sont échangées sur Euronext Paris et appartiennent à l'indice action français CAC40;
- des actions appartenant au même secteur d'activité: BNP Paribas et Société Générale (79%), deux banques françaises majeures;
- des actions de la même entreprise échangées sur différentes places financières: Total échangée sur Euronext Paris et Total échangée sur Chi-X Europe (98%);
- des corrélations géographiques: l'indice S&P500 (Etats-Unis) et l'indice Eurostoxx50 (zone Euro) (60%)
- des corrélations entre deux taux de change, deux matières premières ou métaux précieux, ou bien deux taux d'intérêt de différente maturité: les taux de change USD/EUR et USD/JPY (26%), l'or et l'argent (76%), les bons du Trésor Américain à 2 et 10 ans (75%)
- des corrélations entre différentes classes d'actifs: l'indice S&P500 et le taux de change USD/EUR (-13%), l'indice S&P500 et le pétrole brut (17%)

Ces corrélations significatives sont bien connues des professionnels du monde financier et des chercheurs en finance, tout particulièrement à l'échelle de temps quotidienne. En effet, de telles interdépendances jouent un rôle fondamental dans de nombreuses activités du secteur de la finance quantitative. Dans la gestion d'actifs, elles sont exploitées afin de construire des portefeuilles diversifiés affichant un risque amoindri. Dans un autre registre, les traders algorithmiques fractionnent les ordres volumineux de leurs clients afin de tempérer leur impact sur le marché et ainsi améliorer le coût d'exécution final, qui dépend de la corrélation s'il s'agit d'un portefeuille d'actifs à traiter. Les traders pratiquant l'arbitrage statistique recherchent des

[1] Les corrélations présentées ci-dessous sont les corrélations des log-rendements calculés sur les prix de clôture quotidiens ajustés des opérations sur titres entre le 01/01/2000 et le 14/05/2012. Ces données sont fournies par Bloomberg.

corrélations retardées entre différents actifs, appelées aussi relations de *lead/lag*, pour identifier des opportunités de profit. Les banques offrant des produits dérivés à leurs clients sont attentives aux corrélations entre les actifs inclus dans ces produits car leur prix et leur couverture sont très sensibles à ce paramètre. L'activité dans laquelle la corrélation est d'une importance la plus cruciale reste probablement le trading de produits dérivés de crédit, compte tenu notamment du montant faramineux d'argent que ce marché représente. L'évaluation des produits financiers tels que les CDS[2] multi-noms et les CDO[3] est grandement sensible aux corrélations entre les événements de défaut de paiement des pays ou entreprises inclus dans ces contrats (*cf. The Formula That Killed Wall Street* [92] où il est expliqué comment une modélisation non réaliste de la corrélation peut conduire à une considérable sous-estimation des risques financiers).

Les travaux empiriques sur la corrélation à l'échelle de temps quotidienne sont abondants dans la littérature (voir par exemple [25] pour un panorama des différents faits stylisés sur la corrélation). Quelques tentatives de modélisation ont été développées afin de capturer ces faits stylisés sont également disponibles, citons entre autres les copules, les matrices de covariance stochastiques et les processus ponctuels multivariés. Nous décrivons succinctement chacune de ces approches ci-dessous.

Copules

Une copule caractérise le comportement d'une distribution de probabilité multivariée. En effet, le théorème de Sklar[94] prouve que la fonction de répartition d'un vecteur aléatoire $(X_1, \ldots, X_d)$ peut être formulée comme $F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d))$, où C est une copule et $F_i(x_i)$ est la fonction de répartition univariée de X_i. C est unique si chaque F_i est continue. Les copules sont théoriquement attrayantes puisqu'elles permettent de séparer explicitement les propriétés statistiques univariées et multivariées. De plus, elles décrivent entièrement la structure de dépendance d'un vecteur aléatoire, à l'opposé des mesures de corrélation classiques telles que le coefficient de corrélation de Pearson. Malgré toute l'information intéressante que renferment les copules, comme l'épaisseur et l'asymétrie de la queue de distribution multivariée, retenir telle ou telle copule est souvent un choix *ad hoc* non stimulé par une interprétation financière ou phénoménologique. Les matrices de corrélation empiriques étant hautement structurées[23], les modèles multivariés se doivent d'être0 interprétables.

Matrices de covariance stochastiques

Il est tentant de concevoir des modèles dans lesquels la matrice de covariance est gouvernée par une équation différentielle stochastique, par analogie avec les recherches menées sur un seul actif. Dans cette mouvance s'inscrivent les processus de Wishart [54], les modèles DCC (Dynamic Conditional Correlation) [48], ou bien les processus stochastiques transformés [17, 96]. Les processus de Wishart sont mathématiquement confortables puisqu'ils bénéficient d'une formule analytique pour leur fonction caractéristique. Ils peuvent être assimilés à l'extension multivariée du processus Cox-Ingersoll-Ross [40]. Les matrices de covariance résultant de ce modèle sont par essence définies positives, ce qui est une contrainte intrinsèque de toute matrice de covariance. Les modèles DCC sont quant à eux des modèles économétriques généralisant les modèles GARCH au cas multivarié. La matrice de covariance y évolue dynamiquement et est linéairement impactée par ses valeurs passées ainsi que par celles des rendements croisés. Cette approche est intéressante car elle décrit un phénomène bien connu qu'est celui des grappes de covariance [25]. Enfin, l'approche par processus stochastique transformé consiste à définir un processus latent affichant des caractéristiques statistiques similaires à celles de la corrélation empirique, puis de projeter ce processus dans l'intervalle $[-1, 1]$. Cependant, de tels modèles ne s'occupent que des corrélations par paire, laissant de côté des propriétés essentielles telles que la positivité et la structure du spectre des matrices de corrélation empiriques. Nous pensons qu'une piste intéressante à suivre dans le futur serait la modélisation directe du spectre plutôt que celle des corrélations par paire.

[2]Credit Default Swap
[3]Collateralised Debt Obligation

Processus ponctuels multivariés

Avec les processus ponctuels multivariés, l'attention est portée sur la dépendance entre les sauts des actifs financiers. C'est particulièrement pertinent lorsqu'on s'intéresse aux problématiques rencontrées dans le marché du crédit où les événements de défaut sont l'objet d'intérêt. Citons [13] pour une application au marché du crédit. Nous allons étudier cette approche en détail dans cette thèse car les processus ponctuels sont tout à fait adaptés à la dynamique des prix *tick-by-tick* (voir [84] pour un tour d'horizon des applications des processus ponctuels en finance).

L'étude de la dynamique multivariée des prix à haute fréquence en est encore à ses balbutiements, ce qui est somme toute naturelle car les données haute fréquence de qualité ne sont disponibles que depuis récemment à l'échelle de la recherche académique. Nous dressons un tableau de l'état de l'art sur la corrélation haute fréquence dans la section 1.2 afin que le lecteur puisse se faire une idée précise des avancées faites sur le sujet. Dans cette thèse, nous souhaitons aborder la problématique de la corrélation haute fréquence avec un angle d'attaque original et essentiellement empirique. Même s'il est évident que la dépendance statistique peut être mesurée de façons diverses et variées, nous choisissons de nous concentrer sur la corrélation. Des mesures plus sophistiquées et non linéaires sont pour sûr d'une grande valeur, mais nous préférons sacrifier la complexité à l'intuition et la robustesse des mesures, ainsi qu'à l'originalité de l'étude. Notre but ici est d'apporter un éclairage compréhensible et utile sur la façon dont les prix des actifs financiers réagissent entre eux à l'échelle *tick-by-tick*. Ce travail est une tentative pour atteindre cet objectif. Le reste du manuscrit s'articule autour de trois parties indépendantes qui abordent chacune un aspect différent de la corrélation à haute fréquence. Nous détaillons ci-dessous le contenu de chacun de ces trois chapitres.

Dans le chapitre 2, nous étendons au cas multivarié le cadre de la modélisation par subordination introduit en finance dans [37]. Cette extension s'appuie sur la définition adéquate d'un temps événementiel multivarié. Ce temps événementiel se trouve être une généralisation naturelle du temps événementiel univarié. Il est incrémenté à chaque fois qu'un événement se produit sur l'un des actifs considérés. Nous testons le réalisme des implications de cette modélisation des prix en utilisant des données haute fréquence. Nous trouvons que les prédictions du modèle sont en ligne avec les mesures empiriques. Ce cadre permet de trouver les origines de la queue épaisse de la distribution multivariée des rendements dans une matrice de covariance stochastique. Il nous autorise également à établir un lien entre la matrice de covariance aléatoire et l'arrivée des ordres marché et leur volume, rendant cette matrice, *a priori* inobservable, observable. Ce chapitre a donné lieu à la publication de l'article [5].

Dans le chapitre 3, nous nous intéressons à la mesure des relations de *lead/lag*. Deux actifs financiers entretiennent une relation de *lead/lag* dès lors que l'un suit l'évolution de l'autre avec un certain retard. Sur les marchés liquides, de tels phénomènes statistiques tendent à s'évaporer. Les traders en tirant profit, le temps d'ajustement de l'actif suiveur tend vers zéro. Aussi la détection des effets de *lead/lag* nécessite-t-elle des outils pour mesurer les décalages inter-événements aux plus fines échelles de temps. En pratique, le temps de décalage est de l'ordre d'une seconde sur les marchés liquides. La durée entre deux transactions sur les futures liquides étant du même ordre de grandeur, nous devons échantillonner les prix à haute fréquence. Un échantillonnage aussi fin soulève des complications statistiques pour la mesure de la corrélation. Nous résolvons une partie de ces problèmes en utilisant l'estimateur de corrélation de Hayashi-Yoshida [61]. Nous observons des relations de *lead/lag* significatives à travers des fonctions de corrélation hautement asymétriques. Ce phénomène de *lead/lag* est particulièrement flagrant pour les paires future/action. Il est crucial de comprendre que le recours à des données échantillonnées de façon régulière ne permet pas de détecter précisément ces effets de *lead/lag* haute fréquence. De plus, un échantillonnage régulier peut conduire à des conclusions fallacieuses provenant du déséquilibre intrinsèque entre les fréquences de trading des actifs financiers. Ce chapitre a conduit à la publication de l'article [3].

Dans le chapitre 4, nous étudions le profil journalier de la corrélation haute fréquence. Cette corrélation apparaît comme étant significativement fluctuante au cours de la session de trading, puisqu'elle peut varier

de une à six ou sept fois sa valeur moyenne selon l'univers d'actifs considéré. De façon assez surprenante, ce schéma est universel. Nous suggérons une extension du modèle de prix introduit dans [16] pour capturer ce fait stylisé. Ce modèle repose sur des processus de Hawkes, qui sont des processus auto- et mutuellement excitants. Nous généralisons ce modèle en y incorporant des intensités de base et des noyaux d'excitation dépendant du temps courant afin de capter le comportement non-stationnaire de la dynamique multivariée des prix. L'estimation des paramètres passe par l'algorithme EM décrit dans [74]. Le modèle résultant reproduit précisément la forme du profil journalier de corrélation mais ne parvient pas à atteindre le niveau de corrélation observé. Un niveau de corrélation réaliste peut être retrouvé en rendant le profil journalier plus grossier. Ce chapitre a abouti à la publication de l'article [4].

Avant de passer à une description plus précise des résultats de chaque chapitre, nous proposons de débuter avec une revue de la littérature sur la corrélation à haute fréquence. Dans la sous-section 1.2.1, nous énumérons les différents problèmes rencontrés lors de l'estimation de la corrélation à partir de données *tick-by-tick*. Puis, la sous-section 1.2.2 dissèque les quelques tentatives de modélisation faites pour décrire la dynamique multivariée des actifs à de petites échelles de temps, ainsi que ses propriétés asymptotiques. La rareté de tels travaux a largement motivé la présente thèse.

1.2 Etat de l'art sur la corrélation à haute fréquence

Comme nous l'avons évoqué dans la section précédente, la connaissance accumulée sur la corrélation haute fréquence est encore aujourd'hui modeste. Une partie de l'explication réside dans la disponibilité de données *tick-by-tick* fiables et le besoin des professionnels de prendre en compte la dépendance multivariée aux petites échelles de temps dans leurs modèles, qui sont tous deux somme toute récents.

Bien que l'effet Epps fut découvert en 1979 [49], la littérature statistique s'intéressant à la corrélation haute fréquence connut réellement son essor dans les années 2000. Il existe cependant deux références précurseurs, [76] en 1991 and [43] en 1997, qui s'attaquèrent les premières, à notre connaissance, à la problématique du trading asynchrone sans avoir recours à un échantillonnage régulier. Dans les années 2000, un grand nombre de chercheurs en statistique se penchèrent sur le sujet de l'impact de la microstructure des marchés sur l'estimation de la covariance, citons entre autres [19, 100]. Grâce à une facilité d'accès aux données grandissante, les académiques furent capables de tester la performance de différents estimateurs de covariance sur des données réelles. Ceci donna naissance à un large panel d'estimateurs statistiques de la corrélation réalisée que nous détaillons dans la sous-section 1.2.1.

Après la conception de mesures plus précises de la corrélation haute fréquence, l'étape suivante naturelle pour les membres de cette communauté scientifique fut la construction de modèles décrivant le comportement de la dépendance aux petites échelles de temps. A notre connaissance, ce champ de recherche est aujourd'hui encore largement inexploré. Notons qu'une contribution prometteuse fut faite dans [16]. Nous discutons de ce sujet dans la sous-section 1.2.2.

1.2.1 Problématiques liées à l'estimation: l'effet Epps

La mesure est une étape fondamentale avant de passer à la modélisation. Cette étape s'avère être spécialement complexe lorsqu'il s'agit de la corrélation à très court terme. T.W. Epps fut la premier à rapporter, en ses propres termes, que "*Correlations among price changes [...] are found to decrease with the length of the interval for which the price changes are measured*"[4]. Le tableau 1.1, extrait de [49], illustre la découverte de Epps.

	Pairs of Stocks					
Interval	AMC-Chrysler	AMC-Ford	AMC-GM	Chrysler-Ford	Chrysler-GM	Ford-GM
10 minutes	001	.009	– 009	– 014	007	055
20 minutes	009	018	011	017	026	118
40 minutes	.006	012	.014	041	040	197
One hour	–.043	.057	064	.023	065	.294
Two hours	029	.060	094	.112	129	383
Three hours	031	.158	111	361	518	519
One day	–.067	.170	078	342	442	571
Two days	– 020	223	186	336	449	572
Three days	– 098	203	.100	334	542	645

Figure 1.1: Corrélations entre les log-rendements de AMC, Chrysler, Ford et GM. Extrait de [49].

Les corrélations reportées dans l'article d'Epps sont calculées selon la méthodologie de l'échantillonnage régulier. A partir d'une série temporelle de prix, un sous-échantillon de prix est extrait après avoir échantillonné les données toutes les Δt minutes. En supposant que la session de trading commence à la date t_0 et

[4] Les corrélations entre les changements de prix [...] diminuent avec la longueur de l'intervalle au cours duquel les changements de prix sont mesurés.

s'achève en T, nous prélevons des prix aux temps $t_i = t_{i-1} + \Delta t$ pour $i = 1, \ldots, n$ où $n = \max\{i \mid t_i \leq T\}$. Le prix à la date t est par convention choisi comme le dernier connu avant t. Puis, en utilisant le sous-échantillon de prix $\{p_0, \ldots, p_n\}$, les log-rendements $r_i = \ln(p_i/p_{i-1})$ sont calculés pour $i = 1, \ldots, n$. La corrélation entre les log-rendements $\left\{r_i^1, \ldots, r_n^1\right\}$ et $\left\{r_i^2, \ldots, r_n^2\right\}$ de deux actifs à l'échelle de temps Δt est finalement donnée par

$$\rho(\Delta t) = \frac{\sum_{i=1}^{n} r_i^1 r_i^2}{\sqrt{\sum_{i=1}^{n} (r_i^1)^2 \sum_{i=1}^{n} (r_i^2)^2}}$$

A l'époque où Epps mena ses recherches, la plus haute fréquence accessible était 10 minutes. Les corrélations mesurées à cette échelle sont de l'ordre de $0-5\%$ d'après le tableau 1.1. Elles augmentent graduellement et se stabilisent autour de $30-60\%$ (selon la paire d'actifs, noter le cas des corrélations impliquant AMC qui sont considérablement plus faibles, mais toutefois fortement croissantes en fonction de Δt) à l'échelle de temps quotidienne. La dépendance de la corrélation à l'échelle de temps est par conséquent flagrante.

Depuis la découverte de l'effet Epps et sa confirmation par d'autres auteurs ([59, 85] par exemple), beaucoup se sont essayés à déterminer la source de ce phénomène. Nous avons identifié trois raisons mises en avant dans la littérature:

- le trading asynchrone [76]
- les relations de *lead/lag* [71]
- la discrétisation des prix [57]

Le trading asynchrone renvoie au caractère aléatoire des temps d'arrivée des ordres marché, limite et des annulations. Par conséquent, il n'y a aucune chance que les prix de deux actifs changent exactement en même temps. Donc la probabilité que les prix de deux actifs changent au cours d'une période de temps donnée tend vers zéro lorsque la durée de cette période devient de plus en plus courte, et ce, plus vite que la probabilité de changement de prix d'un seul actif. Cela conduit à sommer un grand nombre de zéros lorsqu'on estime la covariance de deux actifs échantillonnés à haute fréquence comme la somme des produits des rendements de ces actifs. En effet, ces zéros proviennent de la nullité d'au moins un des deux rendements multipliés. Ce phénomène de trading asynchrone, qui est intrinsèque aux marchés dirigé par les ordres, biaise la covariance estimée vers zéro sans impacter les volatilités réalisées, ce qui aboutit à une sous-estimation de la corrélation.

Il convient de dire que deux actifs entretiennent une relation de *lead/lag* si le prix de l'un d'entre eux est statistiquement en avance sur celui de l'autre. L'actif suiveur a besoin d'un peu de temps pour incorporer l'information incluse dans le prix du meneur. Si nous estimons la corrélation à une échelle de temps inférieure à celle du temps caractéristique de la relation *lead/lag*, alors une partie du mouvement joint des deux actifs n'est pas prise en compte. Evidemment, cela tire la corrélation estimée à haute fréquence à la baisse.

Les ordres d'achat et de vente ne peuvent pas être soumis à n'importe quel prix sur les marchés financiers. Les prix admissibles pour un instrument financier sont des multiples d'une quantité appelée le *tick*. Le prix d'un instrument financier vit donc sur une grille discrète dont la maille a pour taille le *tick*. Dans [57], les auteurs démontrent que la taille du *tick* joue un rôle prépondérant dans l'effet Epps. A basse fréquence, les variations de prix sont si grandes par rapport au *tick* que la discrétisation des prix devient négligeable dans la distribution de probabilité des rendements, si bien que celle-ci peut être modélisée par une distribution continue. *A contrario*, à l'échelle *tick-by-tick*, cette distribution est hautement discrète et concentrée autour de valeurs typiques. Selon [57], la discrétisation conduit à une perte d'information par rapport au cadre de modélisation continue, ce qui contribue à l'effet Epps. Notons que la discrétisation des prix impacte aussi bien l'estimation de la covariance que celle des volatilités. Il est prouvé dans [58] que c'est la surestimation des volatilités, plutôt qu'une sous-estimation de la covariance, qui conduit à une sous-estimation de la corrélation.

De nombreuses solutions furent suggérées dans la littérature pour se passer outre ces biais d'estimation que l'on regroupe généralement sous l'appellation "bruit de microstructure". Ces effet peuvent être assimilés à du bruit puisqu'ils créent un fossé entre la réalité et le cadre théorique des semi-martingales pour l'estimation de la covariance [100]. Ceci dit, les trois problématiques évoquées plus haut n'en restent pas moins des caractéristiques intrinsèques des marchés électroniques, si bien que le terme "bruit" peut paraître inapproprié. Nous classifions la plupart des estimateurs proposés dans la littérature pour vaincre le bruit de microstructure en cinq catégories qui sont les suivantes:

- les estimateurs de sous-échantillonnage
- les estimateurs de Fourier
- l'estimateur Hayashi-Yoshida
- les estimateurs de *lead/lag*
- les estimateurs d'interpolation du prix efficient

L'ensemble de ces estimateurs cherche à approcher la corrélation asymptotique, disons quotidienne, à partir de données haute fréquence. Nous décrivons brièvement ces estimateurs ci-dessous.

Estimateurs de sous-échantillonnage

Les estimateurs de sous-échantillonnage sont similaires à la méthodologie utilisée par Epps et exposée précédemment. Les prix sont échantillonnés le long d'une grille régulière avec une fréquence donnée et la corrélation classique entre les deux séries temporelles de rendements résultantes est retournée. Deux degrés de liberté s'offrent à nous: la période d'échantillonnage Δt et la procédure d'interpolation des prix.

Le choix de la période d'échantillonnage n'est pas trivial car il met en jeu un arbitrage biais/variance. En effet, à haute fréquence, nous disposons de plus de données mais également de plus de bruit de microstructure, tandis qu'à plus basse fréquence les données se font plus rares mais le bruit de microstructure s'évapore. Beaucoup de travaux ont été conduits dans cette direction pour estimer la volatilité réalisée, voir par exemple [15, 91], et ont ensuite été étendus à l'estimation de la covariance. Tout ce pan de la littérature cherche à trouver la fréquence d'échantillonnage optimale, de façon à minimiser un critère d'erreur d'estimation comme le MSE (*Mean Square Error*) asymptotique. Il suggère également de combiner des estimateurs calculés à des échelles de temps différentes pour améliorer le MSE. Remarquons que l'échantillonnage pour l'estimation de la covariance peut être réalisé en temps événementiel plutôt qu'en temps calendaire. Cela nécessite au passage une définition adéquate du temps événementiel multivarié (voir [5] pour une proposition d'un tel temps événementiel et [18, 14] pour une autre et son application à l'estimation de la covariance).

La technique d'interpolation a également un rôle à jouer ici. En effet, quel prix doit être considéré comme le prix à la date t? Prendre le dernier connu peut se révéler problématique pour l'estimation de la covariance à cause du trading asynchrone (voir [100] pour une étude rigoureuse du schéma dit *previous-tick*). Quelques auteurs prônent l'utilisation d'une interpolation linéaire entre les prix juste avant et juste après la date t [41]. Notons cependant que cela revient à faire appel à de l'information future, ce qui est rédhibitoire pour l'application au trading en temps réel.

Estimateurs de Fourier

En 2002, les auteurs de [77] introduisirent un estimateur de covariance prenant en compte l'ensemble des données disponibles, qu'elles soient régulièrement échantillonnées ou non. Cet estimateur est basé sur une reconstruction de Fourier de la matrice de covariance sous l'hypothèse que le processus générateur des données est une semi-martingale. Considérons deux actifs ayant la dynamique de log-prix suivante

$$\begin{aligned} \mathrm{d}p_1(t) &= \sigma_{11}(t)\mathrm{d}W_1(t) + \sigma_{12}(t)\mathrm{d}W_2(t) \\ \mathrm{d}p_2(t) &= \sigma_{21}(t)\mathrm{d}W_1(t) + \sigma_{22}(t)\mathrm{d}W_2(t) \end{aligned}$$

où (W_1, W_2) est un mouvement brownien standard bidimensionnel. Nous définissons la matrice de covariance instantanée $\Sigma_{ij}(t) = \sum_{k=1}^{2} \sigma_{ik}(t)\sigma_{jk}(t)$ et sa version intégrée $\Sigma_{ij} = \int_0^T \Sigma_{ij}(t)\mathrm{d}t$. Supposons maintenant que nous observons deux séries temporelles de log-prix $p_1(t_1^1), \ldots, p_1(t_{n_1}^1)$ et $p_2(t_1^2), \ldots, p_2(t_{n_2}^2)$ pouvant être irrégulièrement espacées dans le temps. Les coefficients de Fourier des rendements sont définis par

$$
\begin{aligned}
a_0^\ell &= \frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}p_\ell(t) = \frac{p_\ell(2\pi) - p_\ell(0)}{2\pi} \\
a_k^\ell &= \frac{1}{\pi}\int_0^{2\pi} \cos(kt)\mathrm{d}p_\ell(t) \quad \text{for } k > 0 \\
&= \frac{p_\ell(2\pi) - p_\ell(0)}{\pi} + \frac{1}{\pi}\sum_{i=1}^{n_\ell} p_\ell(t_{i-1}^\ell)(\cos(kt_i^\ell) - \cos(kt_{i-1}^\ell)) \\
b_k^\ell &= \frac{1}{\pi}\int_0^{2\pi} \sin(kt)\mathrm{d}p_\ell(t) = -\frac{1}{\pi}\sum_{i=1}^{n_\ell} p_\ell(t_{i-1}^\ell)(\sin(kt_i^\ell) - \sin(kt_{i-1}^\ell))
\end{aligned}
$$

pour $\ell \in \{1, 2\}$. Le temps a été redimensionné de façon à appartenir à l'intervalle $[0, 2\pi]$. Ces coefficients de Fourier sont calculés en utilisant l'ensemble des points des séries temporelles. La covariance intégrée est déduite de ces coefficients *via* la relation suivante

$$
\Sigma_{ij} = 2\pi \lim_{N\to+\infty} \frac{\pi}{N}\sum_{k=1}^{N}(a_k^i a_k^j + b_k^i b_k^j)
$$

La corrélation intégrée est définie comme suit, $\rho = \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}}$. En pratique, un entier N fini doit être déterminé pour calculer le développement de Fourier ci-dessus. Les auteurs de [80] remarquent avec raison que ce choix est critique.

Estimateur de Hayashi-Yoshida

L'estimateur de Hayashi-Yoshida fut présenté dans [61]. Il est conçu pour régler le problème du trading asynchrone. Les auteurs prouvent que cet estimateur est à la fois non biaisé pour la covariance et consistent pour la corrélation sous des hypothèses souples d'échantillonnage asynchrone de deux semi-martingales. En utilisant les mêmes notations que précédemment et en supposant que $\sigma_{ij}(t) = \sigma_{ij}$, la corrélation Hayashi-Yoshida est définie ainsi

$$
\begin{aligned}
\rho_{\mathrm{HY}} &= \frac{\sum_{i=1}^{n_1-1}\sum_{j=1}^{n_2-1} r_i^1 r_j^2 \mathbb{1}_{\{]t_i^1, t_{i+1}^1] \cap]t_j^2, t_{j+1}^2] \neq \emptyset\}}}{\sqrt{\sum_{i=1}^{n_1-1}(r_i^1)^2 \sum_{i=1}^{n_2-1}(r_i^2)^2}} \\
r_i^\ell &= p_\ell(t_{i+1}^\ell) - p_\ell(t_i^\ell)
\end{aligned}
$$

L'idée sous-jacente à cet estimateur est que, dans le cadre des semi-martingales, les incréments de prix de deux actifs sont corrélés seulement s'ils partagent une fenêtre de temps. Le rendement de l'actif 1 entre les temps t_i^1 et t_{i+1}^1 est corrélé avec le rendement de l'actif 2 entre t_j^2 et t_{j+1}^2 à travers la partie commune entre les intervalles de temps, c'est-à-dire durant l'intervalle $]t_i^1, t_{i+1}^1] \cap]t_j^2, t_{j+1}^2]$. S'il n'existe aucune intersection entre ces deux intervalles de temps, alors les rendements sont considérés comme non corrélés. Cet estimateur est libre de tout paramètre d'ajustement puisque les seuls ingrédients requis sont les séries temporelles de prix. Cependant, la corrélation ainsi calculée peut sortir de l'intervalle $[-1, 1]$ puisque la normalisation par les volatilités ne respecte pas l'inégalité de Cauchy-Schwarz. Cela peut notamment arriver lorsque les séries de prix sont singulièrement autocorrélées [20]. C'est rarement le cas en pratique puisqu'il est

notoire que les séries temporelles de rendements sont faiblement autocorrélées [1]. De plus, rien ne garantit que la matrice de corrélation estimée sera définie positive, ce qui constitue un autre inconvénient de cette méthode. Remarquons enfin que l'estimateur de Hayashi-Yoshida est conceptuellement proche de celui créé par les auteurs de [43] en 1997 et qui est, à notre connaissance, le premier à traiter le problème du trading asynchrone d'un point de vue empirique.

Estimateurs de *lead/lag*

Il est aujourd'hui acquis que certains actifs financiers mènent les autres. Ces relations de *lead/lag* contredisent le postulat de non-arbitrage. Elles ont tendance à disparaître, dans le sens où leur temps caractéristique tend vers zéro au fur et à mesure que les traders profitent d'elles. Elles ont été largement étudiées dans la littérature empirique. Au départ, les chercheurs s'intéressaient à des effets de *lead/lag* hebdomadaire [82], puis quotidien [35], et maintenant c'est le *lead/lag* intra-journalier qui est inspecté. Ce phénomène de *lead/lag* peut participer à l'effet Epps, comme il est mentionné dans [71]. En effet, le calcul de la corrélation à une fréquence plus haute que celle de la relation de *lead/lag* ne peut pas inclure les corrélations retardées dues au temps d'ajustement et qui contribuent à la corrélation à l'échelle asymptotique. Par conséquent, de nouveaux estimateurs palliant ce problème furent conçus. Tous ces estimateurs sont assimilables à des moyennes pondérées des corrélations retardées calculées avec un échantillonnage régulier, *i.e.*

$$\rho(\Delta t, H) = \sum_{h=-H}^{H} \kappa\left(\frac{|h|}{H}\right) \rho_{\Delta t,h}$$
$$\rho_{\Delta t,h} = \frac{\sum_{i=1+|h|}^{n-|h|} r_i^1 r_{i+h}^2}{\sqrt{\sum_{i=1}^{n}(r_i^1)^2 \sum_{i=1}^{n}(r_i^2)^2}}$$

où $\rho_{\Delta t,h}$ est la corrélation retardée de h unités de temps et calculée avec un échantillonnage régulier de période Δt. Cet estimateur dépend donc du choix de la fréquence Δt^{-1}, du nombre de retards H pris en compte et du noyau de moyennisation κ. Remarquons que l'échantillonnage peut être en temps calendaire [71, 56] ou bien événementiel [18].

Estimateurs d'interpolation du prix efficient

Les auteurs de [57] mettent en lumière l'impact critique de la taille du *tick* sur l'estimation de la corrélation à haute fréquence. En particulier, ils prouvent que la sous-estimation de la corrélation provient essentiellement de la surestimation des volatilités due à la discrétisation des prix. Ils suggèrent l'utilisation du modèle suivant pour se débarrasser des erreurs d'arrondi

$$p_1(t) = \overline{p}_1(t) + \varepsilon_1(t)$$
$$p_2(t) = \overline{p}_2(t) + \varepsilon_2(t)$$

où $\overline{p}_i$ est la série temporelle des prix observés de l'actif i et ε_i est l'erreur de discrétisation à valeurs dans l'intervalle $[-\delta_i/2, \delta_i/2]$, δ_i étant la taille du *tick* de l'actif i. Ainsi p_i peut être vu comme le prix efficient nettoyé de l'erreur de discrétisation. Les auteurs établissent une relation analytique entre la corrélation des rendements des prix efficients et la matrice de covariance des rendements observés et des erreurs de discrétisation. La procédure d'estimation fait appel à l'estimation au préalable de la distribution de probabilité marginale des rendements continus qui provient d'une interpolation de la distribution discrète. Dans [58], les mêmes auteurs précisent qu'une version simplifiée de leur estimateur de corrélation consiste à ne corriger que l'estimation des volatilités de l'erreur de discrétisation et conserver l'estimateur standard de covariance.

La problématique de la découverte du prix efficient à partir du prix observé est abordée de façon originale dans [90]. Le processus de prix efficient est supposé être une semi-martingale. Le prix observé vit quant à lui sur la grille définie par le *tick*. Une donnée de prix est enregistrée par le statisticien dès lors que le

prix efficient sort d'un tunnel qui est centré autour du dernier prix observé. Dans la formulation la plus simple du modèle, le prix observé saute vers le haut (resp. bas) d'un *tick* si le prix efficient brise une barrière haute (resp. basse) qui est égale au dernier prix observé auquel s'additionne (resp. se soustrait) la moitié du *tick* et un seuil de petite valeur. Si ce seuil vaut zéro, alors le prix observé saute à chaque fois que le prix efficient croise la *midquote* la plus proche. Les temps d'observation sont donc définis comme $\tau_0 = 0$ puis, récursivement pour $i \geq 0$,

$$\tau_{i+1} = \inf \left\{ t > \tau_i : |p_{\text{efficient}}(t) - p_{\text{observé}}(\tau_i)| = \delta(\frac{1}{2} + \eta) \right\}$$

où δ est la taille du *tick* et η est le seuil additionnel (en taille de *tick*) qui doit être dépassé pour déclencher l'enregistrement d'une observation. En supposant que η est estimé à partir des données, nous pouvons en déduire le prix efficient aux temps d'observation

$$p_{\text{efficient}}(\tau_i) = p_{\text{observé}}(\tau_i) - \delta(\frac{1}{2} - \eta)\text{sgn}(p_{\text{observé}}(\tau_i) - p_{\text{observé}}(\tau_{i-1}))$$

Les auteurs proposent un estimateur pour η et prouvent sa convergence uniforme en probabilité. Cela permet d'estimer la volatilité du processus de prix efficient en utilisant l'estimateur standard de volatilité réalisée calculé sur ces valeurs estimées. Dans le cas de deux actifs, il est montré qu'une extension de l'estimateur de Hayashi-Yoshida ayant recours aux valeurs estimées des prix efficients converge vers la covariance des rendements des prix efficients.

L'effet Epps dans nos données

Dans ce qui suit, nous mesurons l'effet Epps sur des séries temporelles *tick-by-tick* provenant de notre base de données (voir section 1.3 pour une description détaillée). Nous considérons deux actions du secteur bancaire français, BNPP.PA et SOGN.PA, respectivement BNP Paribas et Société Générale. Ces deux actions font partie de l'indice action français CAC40 et sont très liquides. La période de temps choisie débute le 01/03/2010 et s'achève le 31/05/2010, soit 64 jours d'activité financière. Nous étudions deux types de séries temporelles de prix: les prix de transaction et les *midquotes*. Les *midquotes* sont échantillonnées juste avant les transactions afin de rendre le nombre d'observations comparable dans les deux cas. Nous calculons plusieurs estimateurs de corrélation pour illustrer l'effet Epps:

- la corrélation à échantillonnage régulier sans intersection et calculée avec les *midquotes*
- la corrélation à échantillonnage régulier avec intersection et calculée avec les *midquotes*
- la corrélation Hayashi-Yoshida calculée avec les *midquotes*
- la corrélation Hayashi-Yoshida calculée avec les prix de transaction

L'échantillonnage régulier sans intersection consiste en un échantillonnage le long d'une grille uniformément espacée de Δt unités de temps et qui commence à partir du premier point observé dans la série temporelle. La contrepartie avec intersection prend en compte toutes les grilles sans intersection possibles commençant à partir de chaque point de la série temporelle, ce qui implique que les rendements consécutifs s'entrecroisent si Δt est plus grand que la durée moyenne entre deux observations. Par construction, la méthode avec intersection aboutit à un plus grand nombre de points. Néanmoins, ces points peuvent être hautement autocorrélés. Par conséquent, le nombre effectif de points indépendants généré par ce schéma n'est pas simple à déterminer.

Le graphique 1.2 présente ces quatre estimateurs en fonction de la période d'échantillonnage Δt. L'effet Epps est très prononcé dans nos données. La corrélation vaut $0,2$ quand $\Delta t = 1$ seconde, atteint rapidement $0,75$ à $\Delta t = 300$ secondes pour ensuite saturer à sa valeur asymptotique de $0,8$ peu de temps après. Notons

que la corrélation quotidienne calculée avec les prix de clôture ajustés sur la même période vaut $0,94$, ce qui signifie qu'une partie non négligeable de la corrélation entre ces deux actifs est réalisée durant la période de fermeture du marché. Les corrélations avec ou sans intersection se comportent de façon similaire mais la première apparaît comme plus stable et plus précise. C'est particulièrement vrai pour de très grandes valeurs de Δt, par exemple $\Delta t = 7200$ secondes (deux heures), pour laquelle l'intervalle de confiance à 95% autour de la corrélation sans intersection est déraisonnablement large. La corrélation Hayashi-Yoshida des *midquotes* vaut 0.49, ce qui est remarquable si on garde à l'esprit que la durée inter-événement est de $3,317$ secondes pour BNPP.PA et $3,351$ pour SOGN.PA. Le passage des *midquotes* aux prix de transaction réduit la corrélation à 0.38. Cela provient du phénomène d'oscillation *bid/ask*. En effet, il y a 50% de chances que deux transaction consécutives soient de signe (achat ou vente) opposés, ce qui résulte en une variation de prix de transaction égale au *spread bid/ask* si nous supposons que les prix cotés sont identiques entre les deux transactions. A l'inverse, la *midquote* juste avant ces transactions ne varie pas, d'où un rendement égal à zéro. Ce phénomène fait enfler la volatilité réalisée à haute fréquence[5] et diminue ainsi la corrélation lorsqu'on utilise les prix de transaction. Nous concluons donc que même le choix de la convention de prix retenue est critique dans l'estimation de la corrélation à haute fréquence.

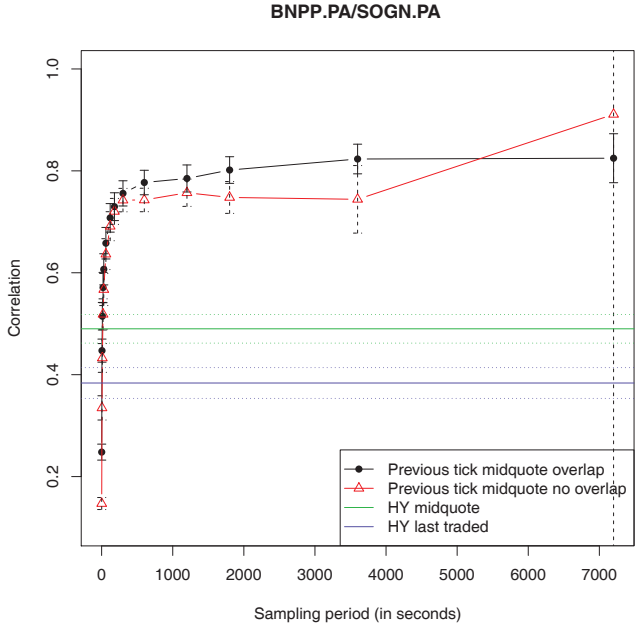

Figure 1.2: L'effet Epps sur nos données.

Comme nous l'avons mentionné précédemment, le trading asynchrone contribue à l'effet Epps. Néanmoins, il n'en est pas l'unique raison. En effet, si tel était le cas, l'estimateur Hayashi-Yoshida atteindrait la valeur asymptotique de la corrélation, ce qui n'est pas le cas d'après le graphique 1.2. De façon plus directe, nous pouvons aussi mesurer la probabilité que les prix de deux actifs varient dans une période de temps donnée. Nous estimons donc

$$P(\Delta t) = \mathbb{P}(p_1(t + \Delta t) \neq p_1(t) \cap p_2(t + \Delta t) \neq p_2(t))$$

[5]Notons par ailleurs que cela augmente également la covariance mais dans une moindre mesure que pour la volatilité, probablement car d'autres phénomènes, tels que le trading asynchrone, viennent compenser l'effet positif sur la covariance.

en fonction de Δt et nous la comparons avec la corrélation réalisée. Notons que $P(\Delta t) = 0$ n'implique pas que les prix n'ont pas bougé du tout entre t et $t + \Delta t$. Si le prix de l'un des deux actifs a varié depuis t mais est revenu à sa valeur initiale en $t + \Delta t$ alors $P(\Delta t) = 0$. En pratique, il s'avère que ce cas est extrêmement rare mais possible, ce qui explique pourquoi la valeur empirique de $P(\Delta t)$ ne converge pas exactement vers 1. Le graphique 1.3 représente $P(\Delta t)$ et la corrélation, toutes deux calculées sur les séries de *midquotes* avec intersection. Les deux courbes se comportent de façon assez similaire, signe que le trading asynchrone est un facteur clé de l'effet Epps. Cependant, $P(\Delta t)$ converge plus rapidement que la corrélation, laissant ainsi la place à d'autres explications de l'effet Epps, telles que les relations de *lead/lag* et la discrétisation des prix.

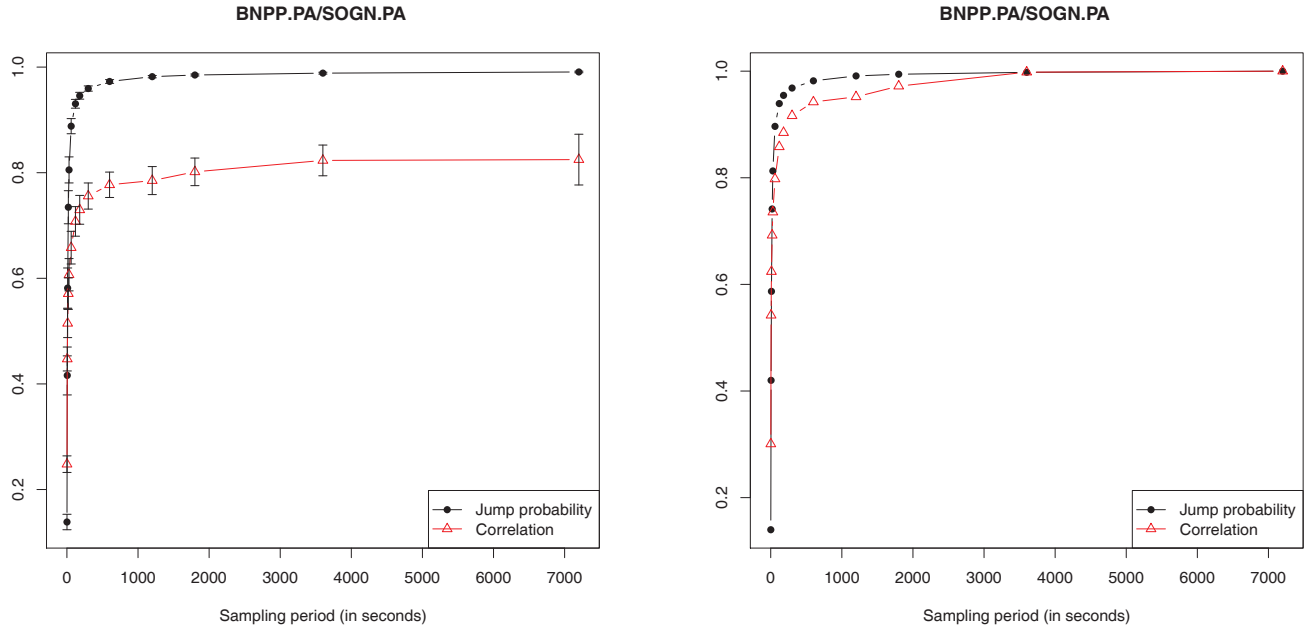

Figure 1.3: La mesure du trading asynchrone dans nos données. Gauche: $P(\Delta t)$ et la corrélation. Droite: $P(\Delta t)$ et la corrélation normalisées par leurs valeurs respectives à 7200 secondes.

Le trading asynchrone et le *lead/lag* impactent la corrélation à travers la covariance. Ils ne modifient pas la volatilité réalisée. Néanmoins, la convergence de la volatilité doit aussi être prise en compte pour comprendre pleinement l'effet Epps. La discrétisation des prix est une source majeure d'écart par rapport au cadre des semi-martingales lorsqu'on estime le coefficient de diffusion [91]. Le graphique 1.2 apporte un premier élément de preuve grâce à la différence considérable entre les corrélations calculées avec les prix de transaction ou bien les *midquotes*. Afin de jauger les contributions respectives de la covariance et de la volatilité à l'effet Epps, nous séparons la variation de la corrélation en deux parties. Notons respectivement $C(\Delta t)$ et $V_i(\Delta t)$ la covariance et la variance (de l'actif i) à l'échelle de temps Δt. La corrélation est $\rho(\Delta t) = \frac{C(\Delta t)}{\sqrt{V_1(\Delta t)V_2(\Delta t)}}$. Par différenciation, nous obtenons

$$\begin{aligned}
\frac{\mathrm{d}\rho(\Delta t)/\mathrm{d}\Delta t}{\rho(\Delta t)} &= \frac{\mathrm{d}\ln(\rho(\Delta t))}{\mathrm{d}\Delta t} \\
&= \frac{\mathrm{d}C(\Delta t)/\mathrm{d}\Delta t}{C(\Delta t)} - \frac{1}{2}\left(\frac{\mathrm{d}V_1(\Delta t)/\mathrm{d}\Delta t}{V_1(\Delta t)} + \frac{\mathrm{d}V_2(\Delta t)/\mathrm{d}\Delta t}{V_2(\Delta t)}\right) \\
&= \text{Contribution covariance} - \text{Contribution variance}
\end{aligned}$$

Puisque l'évaluation numérique de ces contributions nécessiteraient les valeurs de $C(\Delta t)$ et $V_i(\Delta t)$ le long d'une fine grille de valeurs de Δt, nous optons plutôt pour une approche paramétrique en calibrant ces courbes avec des fonctions adéquates. Nous posons $\tilde{C}(\Delta t) = \frac{C(\Delta t)}{\Delta t}$ et $\tilde{V}_i(\Delta t) = \frac{V_i(\Delta t)}{\Delta t}$. Alors, $\rho(\Delta t) = \frac{\tilde{C}(\Delta t)}{\sqrt{\tilde{V}_1(\Delta t)\tilde{V}_2(\Delta t)}}$ et les contributions en variance et covariance peuvent être définies de la même façon que précédemment. Nous calibrons $\tilde{C}(\Delta t)$ et $\tilde{V}_i(\Delta t)$ avec des fonctions du type $f(x) = Cx^a e^{bx}$ par la méthode des moindres carrés ordinaires. De simples calculs montrent que

$$\text{Contribution covariance} = \frac{a_C}{\Delta t} + b_C$$
$$\text{Contribution variance} = \frac{1}{2}\left(\frac{a_{V_1} + a_{V_2}}{\Delta t} + b_{V_1} + b_{V_2}\right)$$

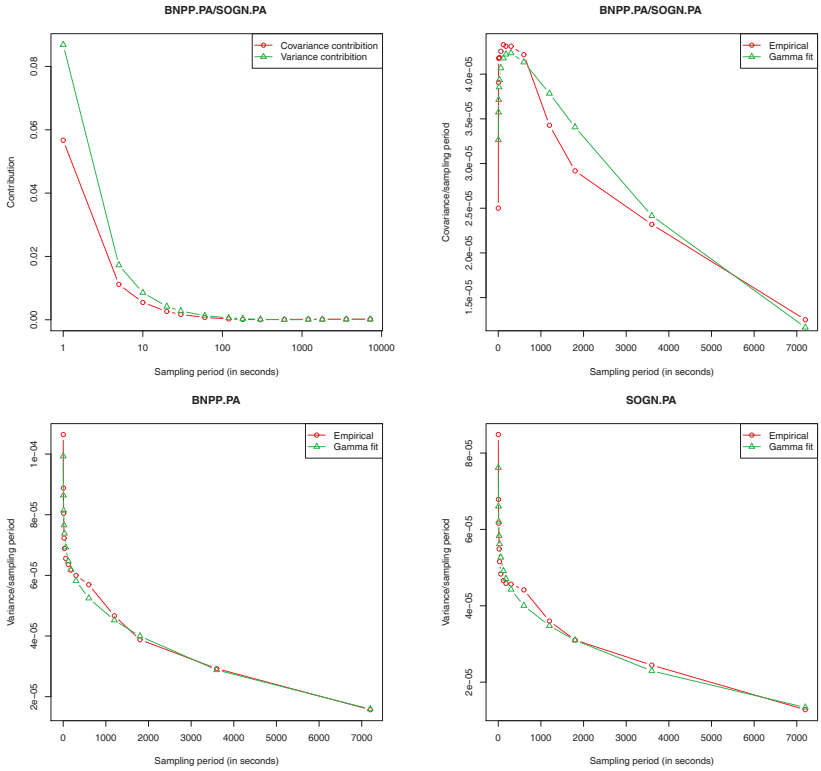

Figure 1.4: Haut gauche: contributions de la covariance et de la variance à l'effet Epps. Haut droite: $\tilde{C}(\Delta t)$ et calibration. Bas gauche: $\tilde{V}_{\text{BNPP.PA}}(\Delta t)$ et calibration. Bas droite: $\tilde{V}_{\text{SOGN.PA}}(\Delta t)$ et calibration.

Le graphique 1.4 représente les contributions en valeur absolue de la covariance et de la variance ainsi que les calibrations mentionnées au-dessus. Les calibrations sont de bonne qualité, indiquant que notre choix d'approximation est judicieux. Les deux contributions décroissent rapidement comme attendu puisque l'effet Epps sature en peu de temps. La contribution de la variance est systématiquement supérieure à celle de la covariance. A la plus haute fréquence, $\Delta t = 1$ seconde, la variance domine significativement la covariance. A $\Delta t = 5$ secondes, les deux contributions diminuent fortement et se rapprochent l'une de l'autre. L'écart entre les deux contributions se resserre à mesure que Δt tend vers sa valeur la plus élevée. Ce graphique prouve que même si le trading asynchrone et les relations de *lead/lag* jouent un rôle déterminant dans l'effet Epps, les biais de microstructure gonflant la volatilité réalisée, tels que la discrétisation des prix, pourraient en être le facteur prépondérant à très haute fréquence.

1.2.2 Modélisation: de la corrélation haute fréquence à la corrélation quotidienne

La mesure précise de la corrélation haute fréquence étant toujours un domaine de recherche actif, les modèles décrivant la dynamique jointe de plusieurs instruments financiers au niveau *tick-by-tick* sont rares. De par la nature discrète des prix en temps et en espace, nous croyons que le cadre de modélisation le plus approprié à cette échelle de temps est celui des processus ponctuels. En effet, d'un point de vue exclusivement mécanique, les prix bougent à cause des événements impactant le carnet d'ordres et qui correspondent soit à des ordres marché, soit à des ordres limite ou soit à des annulations. Ces événements surviennent à des temps aléatoires et donc irrégulièrement espacés. De plus, les prix vivent sur une grille discrète dont la maille est la taille du *tick*. Toutes ces propriétés intrinsèques des marchés font des processus ponctuels un outil privilégié pour la modélisation à l'échelle du mouvement de prix.

Critères d'un modèle de prix multivarié

Ces premières observations étant faites, il convient maintenant de savoir ce qu'on est en droit d'attendre d'un bon modèle multivarié du point de vue statistique. Nous souhaiterions que ce modèle reproduise le plus précisément possible les principaux faits stylisés observés dans les données. En ce qui concerne la dépendance à haute fréquence, l'effet Epps ainsi que les relations de *lead/lag* doivent être prises en compte. L'effet Epps est décrit dans la sous-section 1.2.1 et les relations de *lead/lag* sont au cœur du chapitre 3. A de plus larges échelles de temps, disons quotidiennes, nous connaissons beaucoup plus de choses au sujet de la corrélation. Le graphique 1.5 illustre plusieurs faits stylisés sur la corrélation quotidienne. Nous utilisons ici des prix de clôture ajustés des indices S&P500 et Eurostoxx50, à l'exception près du graphique bas droite qui est basé sur l'univers des actions composant l'indice Footsie100.

Le graphique 1.5 montre

- les séries temporelles de prix et de corrélation glissante sur une fenêtre de 120 jours
- la densité de probabilité des corrélations inconditionnelle et conditionnelles au signe des deux rendements (les deux positifs ou négatifs ou bien de signes opposés)
- la densité de probabilité du nombre de jours de trading consécutifs avec les deux rendements positifs ou bien négatifs
- la corrélation des excès[6]: $f(u) = \mathbb{C}\mathrm{orr}(r_1, r_2 \,|\, |r_1| \geq F_1^{-1}(u), |r_2| \geq F_2^{-1}(u))$
- la corrélation en boîte: $f(u, \Delta u) = \mathbb{C}\mathrm{orr}(r_1, r_2 | F_1^{-1}(u) \leq r_1 \leq F_1^{-1}(u + \Delta u), F_2^{-1}(u) \leq r_2 \leq F_2^{-1}(u + \Delta u))$
- la densité de probabilité des valeurs propres de la matrice de corrélation de l'univers d'actions composant l'indice Footsie100, comparée avec la distribution de Marchenko-Pastur [79]

Nous retrouvons un grand nombre de faits stylisés de la corrélation quotidienne à partir de ces statistiques. En particulier, la corrélation est corrélée avec le signe des rendements. Les rendements des actifs financiers deviennent plus corrélés pendant les tendances baissières. Nous n'observons pas d'effet de grappe prononcé sur la corrélation, contrairement à la variance ou à la covariance, puisque la probabilité qu'il y ait cinq jours consécutifs de corrélation positive (resp. négative) vaut 5% (resp. 1%). La corrélation des excès montre que les rendements se corrèlent de plus en plus lorsque le marché varie fortement. La corrélation en boîte illustre de façon encore plus saisissante le fait que les mouvements anodins, *i.e.* ceux appartenant au centre de la distribution de probabilité, sont à peine (et plutôt négativement) corrélés. Etant donné que la corrélation inconditionnelle est positive, cela signifie que la plupart de la corrélation est réalisée pendant les périodes turbulentes (celles correspondant aux premiers et derniers déciles de la distribution des rendements). Le spectre de la matrice de corrélation empirique englobe une information significative puisqu'il diffère de la

[6]Nous notons $F_i(u) = \mathbb{P}(r_i \leq u)$ la fonction de répartition des rendements de l'actif i et F_i^{-1} son inverse généralisé.

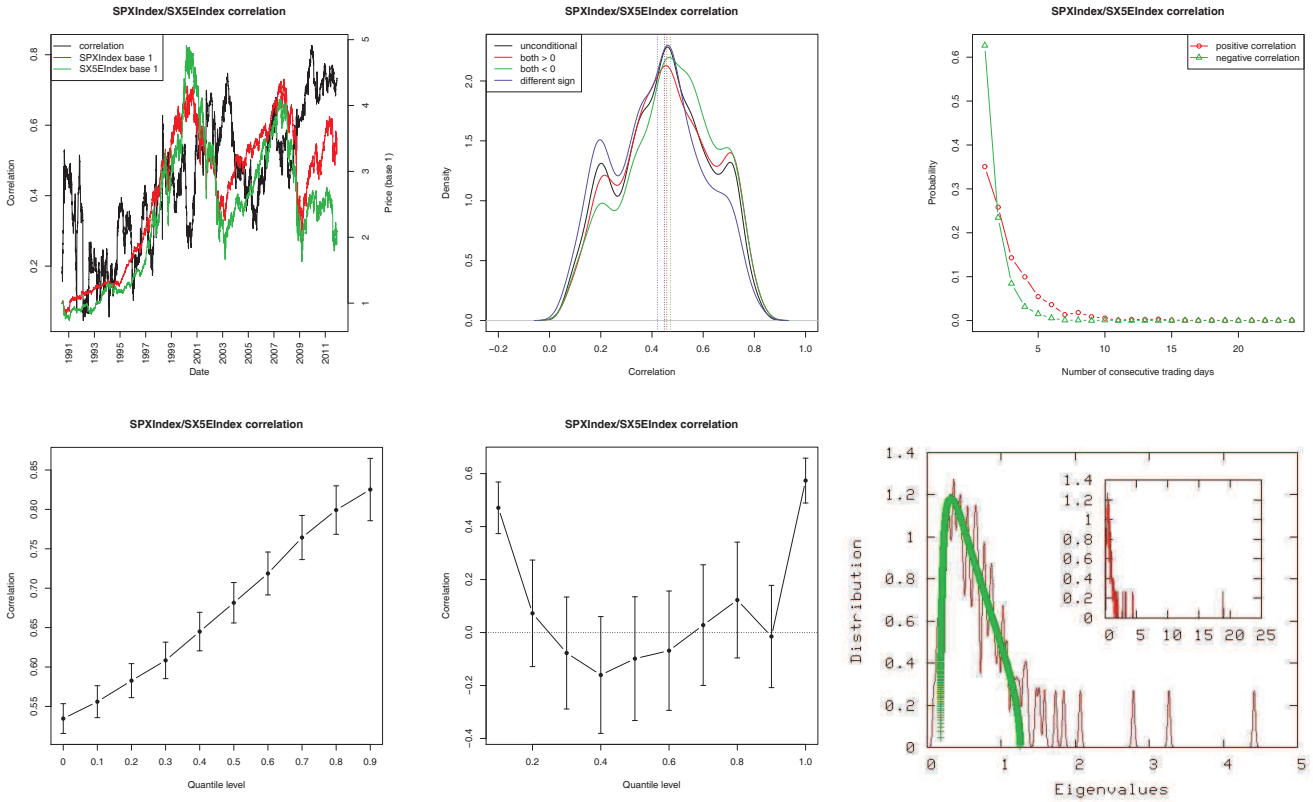

Figure 1.5: Haut gauche: évolution jointe des prix et de la corrélation. Haut centre: densité de probabilité des corrélations inconditionnelle et conditionnelles. Haut droite: densité de probabilité du nombre de jours consécutifs de corrélation positive ou négative. Bas gauche: corrélation des excès. Bas centre: corrélation en boîte avec $\Delta u = 0.1$. Bas droite: densité de probabilité des valeurs propres de la matrice de corrélation sur l'univers Footsie100 (aimablement fournie par Marc Souaille de Natixis).

prédiction faite par la théorie des matrices aléatoires. La valeur propre la plus élevée est un ordre de grandeur au-dessus des autres et correspond au mode appelé communément mode de marché, dans lequel l'ensemble des actions se déplacent dans la même direction. Les autres valeurs propres significatives sont associées à des vecteurs propres mettant en lumière des corrélations inter- et intra-secteurs d'activité.

Un premier modèle

Gardant à l'esprit ces faits stylisés à diverses échelles de temps, considérons tout d'abord un modèle simple basé sur des processus ponctuels. Nous supposons que la dynamique jointe de deux actifs est régie par l'équation suivante

$$\mathrm{d}P_i(t) = \delta_i(\mathrm{d}N_i^+(t) - \mathrm{d}N_i^-(t)) \quad \text{pour } i \in \{1,2\}$$

où δ_i est une constante positive représentant la taille d'un saut du prix de l'actif i et $N_i^\pm$ est un processus de Cox [39] qui compte le nombre de sauts à la hausse/à la baisse du prix de l'actif i. Nous notons $\lambda_i^\pm(t)$ l'intensité du processus $N_i^\pm$, définie par

$$\lambda_i^{\pm}(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{E}(N_i^{\pm}(t+\Delta t) - N_i^{\pm}(t) | \mathcal{F}_i^{\pm}(t))$$

où $\mathcal{F}_i^{\pm}$ est la filtration engendrée par $N_i^{\pm}$. Dans ce modèle, le prix de l'actif i saute vers le haut (resp. bas) entre t et $t+\Delta t$ avec une probabilité égale à $\lambda_i^+(t)\Delta t$ (resp. $\lambda_i^-(t)\Delta t$) pour un petit incrément de temps Δt. La probabilité qu'il ne saute pas vaut donc $1-(\lambda_i^+(t)+\lambda_i^-(t))\Delta t$. Si un saut survient, son amplitude vaut δ_i, qui peut être apparentée à la taille du *tick* de l'actif i. Seul le signe du saut est aléatoire, sa taille est fixée. Calculons le moment d'ordre deux des changements de prix dans ce cadre. En posant $\mathrm{d}_{\Delta t} P_i(t) = P_i(t+\Delta t) - P_i(t)$ et $\mathbb{V}\mathrm{ar}_t(X) = \mathbb{V}\mathrm{ar}(X|\mathcal{F}(t))$, $\mathcal{F}$ étant la filtration générée par $(N_1^+, N_1^-, N_2^+, N_2^-)$, nous avons

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_t(\mathrm{d}_{\Delta t} P_i(t)) &= \delta_i^2 \left(\mathbb{V}\mathrm{ar}_t(\mathrm{d}_{\Delta t} N_i^+(t) - \mathrm{d}_{\Delta t} N_i^-(t)) \right) \\
&= \delta_i^2 \left(\mathbb{V}\mathrm{ar}_t(\mathrm{d}_{\Delta t} N_i^+(t)) + \mathbb{V}\mathrm{ar}_t(\mathrm{d}_{\Delta t} N_i^-(t)) - 2\mathbb{C}\mathrm{ov}_t(\mathrm{d}_{\Delta t} N_i^+(t), \mathrm{d}_{\Delta t} N_i^-(t)) \right) \\
&= \delta_i^2 \left(\mathbb{E}_t(\Lambda_i(t,\Delta t)) + \mathbb{V}\mathrm{ar}_t(\Lambda_i^+(t,\Delta t) - \Lambda_i^-(t,\Delta t)) \right)
\end{aligned}$$

où $\Lambda_i^{\pm}(t,\Delta t) = \int_t^{t+\Delta t} \lambda_i^{\pm}(s)\mathrm{d}s$ et $\Lambda_i(t,\Delta t) = \Lambda_i^+(t,\Delta t) + \Lambda_i^-(t,\Delta t)$. Nous avons fait appel à la relation $\mathbb{C}\mathrm{ov}(X,Y) = \mathbb{E}(\mathbb{C}\mathrm{ov}(X,Y|\mathcal{F})) + \mathbb{C}\mathrm{ov}(\mathbb{E}(X|\mathcal{F}), \mathbb{E}(Y|\mathcal{F}))$ et au fait que $\mathrm{d}_{\Delta t} N_i^{\pm}(t)$ suit une distribution de Poisson de paramètre $\Lambda_i^{\pm}(t,\Delta t)$ conditionnellement à $\Lambda_i^{\pm}(t,\Delta t)$. Le même raisonnement s'applique à la covariance des changements de prix et conduit à

$$\mathbb{C}\mathrm{ov}_t(\mathrm{d}_{\Delta t} P_1(t), \mathrm{d}_{\Delta t} P_2(t)) = \delta_1 \delta_2 \mathbb{C}\mathrm{ov}(\Lambda_1^+(t,\Delta t) - \Lambda_1^-(t,\Delta t), \Lambda_2^+(t,\Delta t) - \Lambda_2^-(t,\Delta t))$$

si bien que la corrélation est donnée par

$$\rho(t,\Delta t) = \frac{\mathbb{C}\mathrm{ov}_t(\Lambda_1^+(t,\Delta t) - \Lambda_1^-(t,\Delta t), \Lambda_2^+(t,\Delta t) - \Lambda_2^-(t,\Delta t))}{\sqrt{\left(\mathbb{E}_t(\Lambda_1(t,\Delta t)) + \mathbb{V}\mathrm{ar}_t(\Lambda_1^+(t,\Delta t) - \Lambda_1^-(t,\Delta t))\right)\left(\mathbb{E}_t(\Lambda_2(t,\Delta t)) + \mathbb{V}\mathrm{ar}_t(\Lambda_2^+(t,\Delta t) - \Lambda_2^-(t,\Delta t))\right)}}$$

La corrélation des changements de prix est donc complètement caractérisée par le moment d'ordre deux du vecteur $\Lambda = (\Lambda_1^+, \Lambda_1^-, \Lambda_2^+, \Lambda_2^-)$. Si Λ est stationnaire en covariance alors $\rho(t,\Delta t)$ ne dépend pas de t. Considérons le cas dans lequel λ est aléatoire mais sa distribution de probabilité est indépendante du temps. Nous avons alors

$$\begin{aligned}
\rho(\Delta t) &= \frac{a}{\sqrt{b_0 + \frac{b_1}{\Delta t} + \frac{b_2}{\Delta t^2}}} \\
a &= \mathbb{C}\mathrm{ov}(\lambda_1^+ - \lambda_1^-, \lambda_2^+ - \lambda_2^-) \\
b_0 &= \mathbb{V}\mathrm{ar}(\lambda_1^+ - \lambda_1^-)\mathbb{V}\mathrm{ar}(\lambda_2^+ - \lambda_2^-) \\
b_1 &= \mathbb{E}(\lambda_1)\mathbb{V}\mathrm{ar}(\lambda_2^+ - \lambda_2^-) + \mathbb{E}(\lambda_2)\mathbb{V}\mathrm{ar}(\lambda_1^+ - \lambda_1^-) \\
b_2 &= \mathbb{E}(\lambda_1)\mathbb{E}(\lambda_2)
\end{aligned}$$

La corrélation satisfait

$$\begin{aligned}
\lim_{\Delta t \to 0} \rho(\Delta t) &= 0 \\
\lim_{\Delta t \to +\infty} \rho(\Delta t) &= \frac{a}{\sqrt{b_0}} = \mathbb{C}\mathrm{orr}(\lambda_1^+ - \lambda_1^-, \lambda_2^+ - \lambda_2^-)
\end{aligned}$$

Si $a > 0$ alors $\rho(\Delta t)$ est une fonction positive et croissante, ce qui réplique l'effet Epps. Cependant, il est impossible d'obtenir des relations de *lead/lag* dans ce modèle car $\mathrm{d}_{\Delta t}P_1(t_0)$ est indépendant de $\mathrm{d}_{\Delta t}P_2(t_1)$ pour $|t_1 - t_0| \geq \Delta t$. Afin d'inclure du *lead/lag*, λ doit dépendre du temps courant et donc être spécifié en tant que processus stochastique. Nous présentons un cas intéressant ci-dessous.

Processus de Hawkes

Les processus de Hawkes furent introduits dans [60] et représentent une classe particulière de processus ponctuels. Ils sont depuis utilisés en mathématiques appliquées pour décrire divers phénomènes tels que les tremblements de terre, les épidémies, l'activité neuronale ou bien les insurrections. L'intensité d'un processus de Hawkes univarié peut être séparée en deux parties. D'un côté, l'intensité de base μ assure l'arrivée de nouveaux événements comme pour un processus de Poisson standard. D'autre part, à l'instant t, les arrivées passés $t_j < t$ augmentent les chances d'une nouvelle arrivée d'une quantité $\phi(t - t_j)$ où ϕ est une fonction positive et intégrable. Les processus de Hawkes sont appelés processus auto-excitants à cause de ce mécanisme endogène de déclenchement de nouvelles arrivées. Rappelons que, comme pour tout processus ponctuel non-marqué, leur dynamique est caractérisée par un processus d'intensité. L'intensité λ d'un processus ponctuel est définie par

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}(\text{un événement survient entre } t \text{ et } t + \Delta t | \mathcal{F}(t))$$

où $\mathcal{F}$ est la filtration engendrée par le processus ponctuel, *i.e.* par la connaissance des temps d'arrivée passés. Dans le cas d'un processus de Hawkes univarié, l'intensité du processus est spécifiée comme suit

$$\lambda(t) = \mu + \sum_{t_j < t} \phi(t - t_j)$$

Le graphique 1.6 montre un échantillon d'une trajectoire d'un processus de Hawkes simulée *via* l'algorithme de dégrossissement présenté dans [87]. Une trajectoire de Poisson est également représentée pour la comparaison. Le noyau de déclenchement choisi est exponentiellement décroissant, c'est-à-dire $\phi(x) = \alpha e^{-\beta x} \mathbb{1}_{\mathbb{R}_+}(x)$. Ce graphique illustre clairement le caractère auto-excitant des processus de Hawkes. Les événements surviennent par grappes qui correspondent aux périodes de forte intensité. Cela contraste avec les arrivées poissoniennes qui sont approximativement équiréparties.

Un tel comportement auto-excitant amène la question de la stabilité du processus. Si le processus est trop auto-excité alors il peut diverger. Intuitivement, le processus n'explosera pas si le noyau d'excitation ϕ est assez petit. Afin de mieux comprendre cette problématique de stabilité, calculons le nombre moyen d'événements par unité de temps

$$\begin{aligned}
\mathbb{E}(\lambda(t)) &= \mathbb{E}(\mu + \sum_{t_j < t} \phi(t - t_j)) \\
&= \mu + \int_{-\infty}^{t} \phi(t - s) \mathbb{E}(\lambda(s)) \mathrm{d}s
\end{aligned}$$

Dans le régime stationnaire, s'il existe, nous avons $\mathbb{E}(\lambda(t)) = \overline{\lambda}\ \forall t$. Cela conduit à

$$\begin{aligned}
\overline{\lambda} &= \mu + \overline{\lambda} \int_{-\infty}^{t} \phi(t - s) \mathrm{d}s \\
&= \frac{\mu}{1 - \int_0^{+\infty} \phi(x) \mathrm{d}x}
\end{aligned}$$

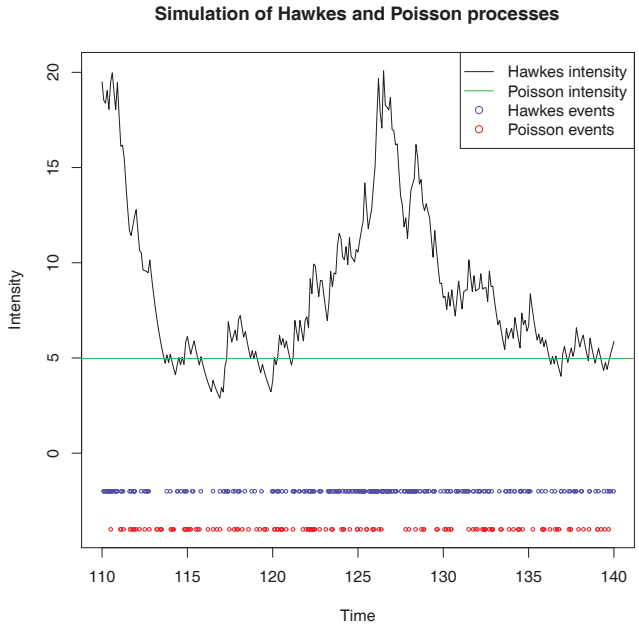

Figure 1.6: Echantillon d'une trajectoire d'un processus de Hawkes univarié ayant pour paramètres $\mu = 1$, $\alpha = 0.8$ et $\beta = 1$. L'horizon de la trajectoire est $T = 1000$ et le graphique se focalise sur l'intervalle de temps $[110, 140]$. L'intensité du processus de Poisson est $\frac{4975}{T} = 4.975$. Elle est choisie de façon à ce que les deux trajectoires contiennent le même nombre de points en moyenne.

Par conséquent, le noyau ϕ doit satisfaire $\int_0^{+\infty} \phi(x)\mathrm{d}x < 1$ pour garantir la finitude de $\overline{\lambda}$. Cette condition donne un sens quantitatif à l'intuition de départ disant que l'effet d'auto-excitation doit être suffisamment petit pour assurer la stabilité du processus.

Les processus de Hawkes multivariés peuvent être définis de la même façon. L'intensité $\lambda = (\lambda_1, \ldots, \lambda_d)^T$ d'un processus de Hawkes multivarié est

$$\lambda_i(t) = \mu_i + \sum_{k=1}^{d} \sum_{t_j^k < t} \phi_{ik}(t - t_j^k)$$

où t_j^k est le $j^{\text{ième}}$ temps d'arrivée de la composante k. Ces processus permettent de modéliser les phénomènes d'excitation croisée grâce aux noyaux $\phi_{ik}, k \neq i$.

Les processus de Hawkes sont aujourd'hui de plus en plus utilisés par les chercheurs en finance, notamment pour les problématiques de microstructure des marchés, voir par exemple [27]. Récemment, le modèle simple décrit dans la sous-section précédente a été étendu en ayant recours à des processus de Hawkes pour modéliser les intensités des changements de prix à la hausse et à la baisse [16]. Les intensités sont gouvernées par la dynamique auto- et mutuellement excitée suivante

$$\lambda_i^{\pm}(t) = \mu_i^{\pm} + \int_0^t \phi_i^{\mathrm{TF},\pm}(t-s)\mathrm{d}N_i^{\pm}(s) + \int_0^t \phi_i^{\mathrm{MR},\pm}(t-s)\mathrm{d}N_i^{\mp}(s)$$
$$+ \int_0^t \phi_i^{\mathrm{CTF},\pm}(t-s)\mathrm{d}N_j^{\pm}(s) + \int_0^t \phi_i^{\mathrm{CMR},\pm}(t-s)\mathrm{d}N_j^{\mp}(s)$$

où $\mu_i^{\pm} > 0$ est une constante appelée intensité de base et $\phi_i^{\mathrm{TF},\pm}, \phi_i^{\mathrm{MR},\pm}, \phi_i^{\mathrm{CTF},\pm}, \phi_i^{\mathrm{CMR},\pm}$ sont des fonctions positives appelées noyaux de déclenchement. TF (resp. MR, CTF, CMR) est l'abréviation de *Trend Following* (resp. *Mean Reverting*, *Cross Trend Following*, *Cross Mean Reverting*). L'intuition justifiant cette notation est qu'un saut de N_1^+ (resp. N_1^-, N_2^+, N_2^-) à l'instant s impacte $\lambda_1^+(t)$ d'une quantité positive $\phi_1^{\mathrm{TF},+}(t-s)$ (resp. $\phi_1^{\mathrm{MR},+}(t-s)$, $\phi_1^{\mathrm{CTF},+}(t-s)$, $\phi_1^{\mathrm{CMR},+}(t-s)$) rendant ainsi un changement de prix de l'actif 1 à la hausse plus probable. Les changements de prix sont donc auto- et mutuellement excités par leurs ancêtres grâce aux noyaux de déclenchement. Ces noyaux capturent les propriétés d'autocorrélation et de corrélation croisée des changements de prix.

Dans [16], les auteurs expliquent que ce modèle est particulièrement bien adapté aux séries temporelles de prix *tick-by-tick* puisqu'il reproduit le comportement de la variance réalisée en fonction de l'échelle de temps et l'effet Epps. Ils établissent des formules analytiques pour ces deux quantités en fonction des paramètres du modèle moyennant quelques restrictions de symétrie ($\mu_i^{\pm} = \mu_i$, $\phi_i^{\mathrm{MR},\pm} = \phi_i^{\mathrm{MR}}$, $\phi_i^{\mathrm{CTF},\pm} = \phi_i^{\mathrm{CTF}}$, $\phi_i^{\mathrm{TF},\pm} = \phi_i^{\mathrm{CMR},\pm} = 0$). Ces courbes analytiques sont calibrées sur leurs contreparties empiriques avec succès. Des relations de *lead/lag* peuvent également être incorporées au modèle grâce à une spécification adéquate des noyaux de déclenchement. Par exemple, rendre les événements passés de l'actif 1 plus influents sur les sauts de l'actif 2 plutôt que l'inverse *via* des noyaux asymétriques conduit à faire de l'actif 1 un meneur.

Il est prouvé que les prix ainsi modélisés convergent vers un mouvement brownien multivarié avec une matrice de covariance s'exprimant explicitement en fonction des paramètres du modèle [16], ce qui permet de relier la covariance diffusive aux paramètres de microstructure. Les prix ont donc un comportement diffusif à long terme, ce qui s'accorde parfaitement avec les observations empiriques. Cependant, plusieurs aspects sont absents comme le caractère non-gaussien de la distribution multivariée des rendements [34], la corrélation négative entre les rendements et la corrélation réalisée, la corrélation supérieure des mouvements de prix extrêmes et la structure du spectre de la matrice de corrélation empirique, comme illustré par le graphique 1.5. Concernant ce dernier point, une solution consisterait en l'ajout d'une contribution endogène du marché $\int_0^t \phi_i^{\mathrm{Market},\pm}(t-s)\mathrm{d}N^{\pm}(s)$ à la dynamique des intensités, où $N^{\pm} = \sum_{i=1}^d N_i^{\pm}$ est le mouvement global à la hausse ou à la baisse du marché.

1.3 Description de notre jeu de données

Cette thèse est constituée en majeure partie de travaux empiriques. Aussi les données sont-elles notre matière première et un soin tout particulier doit donc être apporté pour comprendre leurs caractéristiques intrinsèques. Les données *intraday* utilisées dans cette thèse sont fournies par Thomson Reuters[7] *via* leur produit TRTH (Thomson Reuters Tick History[8]). TRTH est une base de données donnant accès à une masse importante de données financières couvrant un grand nombre de produits échangés, chaque instrument financier étant identifié par son RIC (Reuters Instrument Code). Pour les thèmes de la microstructure, nous nous intéressons plus spécialement à trois sources de données:

- les fichiers de transaction: chaque transaction avec son prix, son volume et son heure à la milliseconde près
- les fichiers de cotation: chaque modification de cotation (meilleurs achat et vente), en prix ou en quantité, à la milliseconde près
- les fichiers de carnet d'ordres: chaque modification de carnet d'ordre, en prix ou en quantité, à la milliseconde près, jusqu'à une profondeur donnée, typiquement dix limites de chaque côté du carnet d'ordres.

Chacune de ces sources de données est enregistrée *via* un canal différent. Par conséquent, ces trois types de données sont asynchrones entre elles. Par exemple, si une transaction est associée à un instant t dans le fichier de transactions, la modification de cotation correspondante peut apparaître dans le fichier de cotations à un instant différent de t.

Ces trois niveaux de données englobent différents types d'événements. Seules les transactions apparaissent dans les fichiers de transaction, elles correspondent à des ordres marché ou bien des ordres limite spécifié avec un prix croisant la meilleure limite opposée disponible au moment de la réception de l'ordre par le marché. Les transactions *block* ou OTC (*Over-The-Counter*) sont également enregistrées et indiquées comme telles. D'autres types de transactions sont enregistrés en fonction des règles du marché où l'actif est négocié. Par exemple, sur le NYSE, les ordres inter-marché (*intermarket sweep orders*) sont fréquents. Ce sont des ordres exécutés dans plusieurs plateformes de trading dans le but d'obtenir la quantité désirée au meilleur prix disponible. Les fichiers de cotation regroupent toutes les modifications des meilleurs prix achat/vente et des quantités associées. De tels événements peuvent être dus aussi bien à des ordres marché qu'à des ordres limite à la meilleure limite ou dans le *spread bid/ask* ou bien des annulations à la meilleure limite. Enfin, les fichiers de carnet d'ordres listent tous les événements modifiant l'état du carnet d'ordres jusqu'à une profondeur donnée. Une fois encore, ces événements peuvent être des ordres marché, limite ou des annulations.

La précision horaire de nos données est la milliseconde. Si plusieurs événements se produisent au cours de la même milliseconde, ils sont étiquetés avec le même temps. Le cas échéant, l'ordre des événements enregistrés est supposé conforme à la réalité. Dans les fichiers de transaction, les événements avec le même horaire peuvent également correspondre à des transactions traversantes. Une transaction est dite traversante si elle est le résultat d'un ordre marché dont la quantité attachée est strictement supérieure à celle disponible au meilleur prix proposé au même moment. Un trader à l'origine d'une telle transaction reçoit la quantité disponible au meilleur prix, puis la quantité restante est graduellement exécutée contre les ordres présents plus loin dans le carnet d'ordres jusqu'à épuisement de la quantité demandée. Si la transaction traversante est un ordre d'achat (resp. vente), le prix d'exécution effectivement obtenu est plus (resp. moins) élevé que le meilleur prix de vente (resp. achat). Dès lors qu'une transaction de ce genre consomme la liquidité présente aux différents niveaux du carnet d'ordres, elle est enregistrée comme une séquence de transactions avec le même horaire mais avec des prix et des quantités correspondant aux ordres limite exécutés. Par

[7] http://thomsonreuters.com
[8] http://thomsonreuters.com/products_services/financial/financial_products/a-z/tick_history

exemple, supposons que les meilleurs vendeurs offrent en tout 100 actions à un prix de 10 euros l'unité, puis 200 au prix de $10,01$ euros. Un ordre marché d'achat arrive alors pour une quantité de 150 actions. L'exécution de cet ordre sera enregistrée comme deux lignes avec le même horaire dans le fichier de transactions, la première étant 100 actions à 10 euros et la seconde 50 à $10,01$ euros. Nous traitons en amont ce type d'enregistrement en agrégeant les lignes avec des horaires identiques dans les fichiers de transactions. Le prix agrégé est le VWAP (*Volume Weighted Average Price*) sur toute la transaction et le volume est la somme des quantités consommées à chaque ligne. Dans l'exemple précédent, le prix de transaction serait donc $(100 * 10 + 50 * 10.01)/(100 + 50) = 10.00333$ euros et le volume $100 + 50 = 150$.

1.4 Résultats de la thèse

Cette section est dédiée à la présentation des résultats obtenus au cours de la thèse. Nous récapitulons les questions soulevées dans chaque partie et exposons brièvement les réponses apportées. Celles-ci seront détaillées dans les trois prochains chapitres de ce manuscrit.

1.4.1 Subordination multivariée

Depuis la mise en lumière des failles empiriques du célèbre modèle de Black&Scholes [21] dans [78], de nombreuses tentatives d'intégration de la queue épaisse de la distribution des rendements dans les modèles furent observées. Une des approches les plus populaires pour combler ce manque du modèle log-normal est l'ajout d'aléa dans la volatilité des rendements. De tels modèles peuvent être formulés comme des mouvements browniens subordonnés. Ils furent en premier introduits dans [37]. Considérons le modèle suivant pour le prix S d'un actif financier

$$\mathrm{d}S(t) = S(t)\sqrt{V(t)}\mathrm{d}W(t)$$

où W est un mouvement brownien standard. V représente la variance stochastique des rendements. Elle est supposée être définie par un processus d'Itô. Il peut être démontré que $S(t)$ suit la même distribution de probabilité que

$$\begin{aligned} X(t) &= S(0)\exp\left(-\frac{1}{2}\int_0^t V(s)\mathrm{d}s + B(\int_0^t V(s)\mathrm{d}s)\right) \\ &= S(0)\exp\left(-\frac{T(t)}{2} + B(T(t))\right) \end{aligned}$$

où B est un mouvement brownien standard. Le processus $\tilde{B}(t) = B(T(t))$ est un mouvement brownien subordonné. T est souvent appelé temps d'activité puisque sa cadence accélère ou ralentit par rapport au temps usuel selon la variance accumulée. L'équation ci-dessus signifie que le prix suit un mouvement brownien géométrique lorsqu'il est mesuré avec l'horloge stochastique appropriée. Ce modèle peut reproduire la queue épaisse observée empiriquement sur la distribution des log-rendements en temps calendaire. Il implique également plusieurs propriétés, telles que la normalité des log-rendements échantillonnés en temps d'activité, que nous pouvons aisément confronter aux données moyennant une approximation observable du temps d'activité. Les auteurs de [36] fournissent une étude empirique approfondie du mécanisme de subordination et observent une adéquation satisfaisante avec les propriétés statistiques des rendements intra-journaliers en utilisant le nombre de transactions pour approcher le temps d'activité.

Notre objectif est d'étendre le cadre précédent à un vecteur d'actifs. Théoriquement, cela nécessiterait la subordination d'un mouvement brownien multivarié avec autant d'horloges stochastiques que d'actifs, ce qui engendre des difficultés pour tester la validité empirique du modèle. Pour illustrer pourquoi, considérons le cas de deux actifs et du nombre de transactions comme approximation de l'horloge stochastique. Les deux actifs s'échangent à des fréquences de trading différentes donc l'échantillonnage des prix après N transactions pour chacun des deux actifs conduit à des prix asynchrones et potentiellement beaucoup moins corrélés qu'ils ne le sont réellement. Une solution possible consiste à échantillonner les prix de chaque actif avec sa propre fréquence de façon à rendre les prix synchrones en moyenne, *i.e.* choisir N_1 et N_2 tels que le temps qu'il faut pour observer N_1 transactions sur l'actif 1 est en moyenne égal au temps qu'il faut pour observer N_2 transactions sur l'actif 2. Cependant, cette méthode n'est pas entièrement satisfaisante car l'activité financière est hautement variable au cours de la session de trading.

Afin de résoudre ce problème, nous suggérons de recourir à un temps événementiel global qui agrège l'activité de tous les actifs. Cela revient à considérer un unique temps événementiel $N = f(N_1, \ldots, N_d)$. Les prix échantillonnés après N événements sont ainsi synchrones. L'étape suivante est la spécification de

la fonction f. Nous choisissons $N = N_1 + \ldots + N_d$, ce qui revient à incrémenter le temps événementiel global d'une unité dès lors qu'un événement se produit sur l'un des d actifs considérés. Nous retiendrons deux types de temps événementiel dans la suite, le temps de transaction et le temps de volume. Le temps de transaction augmente d'une unité à chaque transaction, quels qu'en soient ses termes. Le temps de volume quant à lui se voit croître du nombre de contrats échangés à chaque transaction. Tous deux reflètent l'activité financière. Du point de vue mathématique, le processus stochastique associé aux log-rendements est alors un mouvement brownien multivarié avec une matrice de covariance stochastique dirigée par un seul facteur.

Notre but ici est de vérifier l'adéquation entre ce modèle multivarié et les données *tick-by-tick*. Les principaux résultats sont résumés ci-dessous:

- **Normalité multivariée des rendements en temps événementiel**: nous tirons avantage de la décomposition sphérique d'un vecteur gaussien pour évaluer la normalité multivariée des rendements selon la notion de temps utilisée (calendaire, transaction, volume). Cette procédure se compose de trois étapes: adéquation de la distribution des rayons (au carré) des rendements à une loi du chi-deux, adéquation de la distribution des angles des rendements à une loi uniforme, et test d'indépendance entre les rayons et les angles. En appliquant cette procédure, nous observons que les rendements convergent vers une loi gaussienne multivariée lorsque leur échelle de temps croît, observation justifiée théoriquement par l'application du théorème central limite. Cette convergence a cependant lieu plus précocement en temps événementiel (transaction ou volume) qu'en temps calendaire. Ceci nous amène à penser que notre définition du temps événementiel multivarié est judicieuse car elle permet d'inférer la composante aléatoire de la matrice de covariance des rendements. Nous énonçons également des conditions de consistance entre les versions univariée et multivariée du modèle de subordination, qui mettent en lumière l'importance des ratios de liquidité entre les différents actifs financiers considérés.

- **Croissance linéaire de la matrice de covariance en fonction de l'échelle de temps**: l'approche par subordination implique que les éléments de la matrice de covariance sont linéairement reliés à l'échelle de temps événementiel des rendements. Les données confirment cette propriété du modèle. D'autre part, les rendements financiers exhibant un comportement diffusif, le temps événementiel est en moyenne du même ordre que le temps calendaire.

- **Distribution de probabilité du temps événementiel**: nous montrons que le comportement de la queue de distribution des rendements en temps calendaire, qui est notoirement plus épaisse que celle d'une distribution gaussienne, est dicté par celui du temps événementiel dans le modèle de subordination. Aussi nous intéressons-nous à cette distribution, et observons une forme oscillant entre loi gamma (queue exponentielle) et loi inverse gamma (queue loi-puissance). Une fois encore, ceci est en accord avec les travaux empiriques aboutissant pour la plupart à une queue de type exponentielle ou loi-puissance pour la distribution des rendements en temps calendaire.

- **Vitesse de convergence vers la loi gaussienne**: la croissance du moment d'ordre quatre de la distribution des rendements nous donne une indication sur la vitesse de convergence vers la loi gaussienne. Or, le modèle de subordination nous permet d'établir une relation entre le moment d'ordre quatre des rendements en temps calendaire et le moment d'ordre deux du temps événementiel. Nous comparons les vitesses de convergence empirique et induite par le modèle et trouvons qu'elles sont extrêmement proches, ce qui rend encore plus plausible le mécanisme de subordination.

1.4.2 Relations de *lead/lag* à haute fréquence

Il est fréquent de voir les indicateurs économiques être séparés en trois classes: les indicateurs meneurs, suiveurs et concomitants. Les indicateurs meneurs (resp. suiveurs, concomitants) réagissent avant (resp. après, en même temps) que l'économie change. Parmi les indicateurs meneurs, nous pouvons citer le nombre de permis de construire, l'offre monétaire ou bien les rendements du marché action. Pour ce qui est des indicateurs suiveurs, il y a entre autres le taux de chômage, le nombre de prêts commerciaux et industriels

ou le taux auquel les banques prêtent. Ce phénomène de *lead/lag* est également observable sur les marchés financiers. Certains actifs tendent à mener les autres. Evidemment, ce type de situation crée des possibilités d'arbitrage statistique. Un trader sachant quantifier ce phénomène peut utiliser l'histoire passée des actifs meneurs à des fins lucratives en investissant sur les suiveurs. Pour que cette stratégie soit rentable, l'effet de *lead/lag* doit être suffisamment prononcé pour couvrir les frais de transaction associés. Du simple fait de l'existence de telles stratégies, les relations de *lead/lag* s'amenuisent, voire disparaissent. En effet, alors que les auteurs étudiaient le *lead/lag* à l'échelle hebdomadaire dans le passé [82], il fut ensuite sujet de *lead/lag* quotidien [35]. Aujourd'hui, seules les relations de *lead/lag* à haute fréquence semblent réellement exploitables sur les marchés liquides [69].

Ce travail tire profit d'un estimateur de la fonction de corrélation croisée récemment introduit dans [63] pour mener une étude empirique approfondie du phénomène de *lead/lag* à haute fréquence. Les principaux résultats sont énoncés ci-dessous:

- **Notre définition du *lead/lag***: malgré le caractère intuitif de la notion de *lead/lag*, il n'en existe aucune définition ultime. Il est d'usage dans la littérature de le quantifier grâce à un test de causalité de Granger [55]. Malheureusement, ce cadre est difficilement transposable aux données *tick-by-tick* à cause de l'effet Epps et la nécessité de recourir à des données échantillonnées régulièrement pour conduire une régression linéaire. C'est pourquoi nous avons recours à l'estimateur de fonction de corrélation croisée proposé dans [63] pour mesurer la corrélation retardée entre deux séries temporelles de prix d'actifs financiers à haute fréquence. En gardant à l'esprit l'idée latente au test de Granger, nous considérons qu'un actif X mène un autre actif Y s'il permet de prédire l'évolution future de Y mieux que Y ne le ferait pour X. Nous prouvons que, sous des hypothèses raisonnables, cette caractérisation du *lead/lag* est entièrement retranscrite dans la fonction de corrélation entre X et Y. Plus précisément, X mène Y si et seulement si $\text{LLR} := \frac{\sum_{i=1}^{p} \rho^2(\ell_i)}{\sum_{i=1}^{p} \rho^2(-\ell_i)} > 1$ où $\rho(\ell) = \mathbb{C}\text{orr}(r_t^X, r_{t+\ell}^Y)$ représente la corrélation entre les rendements de X et ceux de Y ℓ unités de temps plus tard (ou plus tôt si $\ell < 0$). Nous expliquons pourquoi cette approche est appropriée à haute fréquence et montrons sa supériorité par rapport à la méthode d'estimation classique de la corrélation.

- **Lien entre liquidité et *lead/lag***: grâce à une étude extensive des relations de *lead/lag* au sein d'un large univers d'actions, nous établissons un lien entre la liquidité d'un actif et son potentiel de meneur ou suiveur. Nous concluons que les actifs les plus liquides, en termes de durée inter-transaction, de masse monétaire échangée par transaction, de *spread bid/ask* et de volatilité, tendent à mener les autres. La taille du *tick* ainsi que la proportion de transactions traversantes ne semblent pas jouer un rôle prépondérant dans les relations de *lead/lag* selon notre méthodologie.

- **Profil journalier du *lead/lag***: nous analysons le comportement de nos indicateurs de *lead/lag* au cours de la session de trading. Ces indicateurs varient sensiblement dans la journée et réagissent notamment à l'arrivée de chiffres macroéconomiques et à l'ouverture du marché américain.

- ***Lead/lag* et mouvements extrêmes**: les relations de *lead/lag* se renforcent lors des périodes turbulentes, c'est-à-dire lorsque des changements de prix significatifs interviennent. Nous explorons ce phénomène en utilisant une version conditionnelle de nos indicateurs.

- **Evaluation d'un outil de prévision basé sur le *lead/lag***: nous construisons un signal de prévision du sens du prochain mouvement de prix de l'actif suiveur en utilisant l'information des changements de prix passés du meneur. Ce signal est une moyenne mobile des rendements de l'actif meneur pondérés par la fonction de corrélation croisée. Nous atteignons une précision de 60% dans cette tâche de prédiction, soit à la fois mieux qu'un signal aléatoire et un signal uniquement fondé sur l'historique du suiveur lui-même. Les stratégies associées et purement à base d'ordres marché ne sont pas rentables du fait du *spread bid/ask* qui est sensiblement plus large que le profit réalisé.

1.4.3 Profil intra-journalier de la corrélation à haute fréquence

Les marchés financiers sont saisonniers à bien des égards. Auparavant, les auteurs se sont intéressés aux saisonnalités mensuelles telles que les effets décembre et janvier [97, 33]. Un argument de législation fiscale a souvent été avancé pour justifier ces effets. Des éléments de preuve d'existence d'autres types de saisonnalité, comme des effets de fin de mois et de week-end [12, 52], d'envolée pré-vacances [51] et de baisse du lundi [50], ont également été rapportés dans la littérature empirique. Ces profils statistiques tendent à s'uniformiser au fur et à mesure que les investisseurs en prennent conscience et en tirent profit.

Cependant, certaines structures saisonnières persistent au niveau intra-journalier. Elles ne concernent pas directement les rendements des actifs financiers mais plutôt des quantités microstructurales. Nous pensons par exemple au volume de transaction, au *spread bid/ask* ou au flux d'ordres, voir [1] pour une revue récente de ces phénomènes. Ces saisonnalités intra-journalières sont bien connues à la fois des intervenants sur le marché et des académiques. Elles demeurent car elles ne conduisent pas à des arbitrages triviaux, contrairement celles impliquant les rendements. Une autre raison de leur subsistance réside dans leur lien étroit avec les habitudes de l'industrie financière. Par exemple, la forme en U du volume échangé peut être mise en parallèle avec certaines caractéristiques intrinsèques des marchés telles que l'ajustement des positions prises la veille, l'incorporation d'informations apprises après la clôture, l'heure du déjeuner, le planning de tombée des informations économiques et la pratique courante consistant à évaluer les contrats financiers sur les prix de clôture.

A notre connaissance, seuls deux articles se focalisent sur le profil intra-journalier de la corrélation [8, 22]. Les auteurs de [8] ont recours à des rendements à cinq minutes pour calculer le profil intra-journalier moyen du spectre de la matrice de corrélation empirique d'actions liquides cotées aux Etats-Unis. Ils observent la croissance du mode de marché au cours de la session de trading tandis que les autres modes décroissent. Etant donné qu'ils fondent leur étude sur des données échantillonnées toutes les cinq minutes, ils ne disposent que d'une observation par jour pour chaque tranche de cinq minutes. Par conséquent, ils doivent faire appel à une série temporelle étalée sur dix années, de 2000 à 2009 pour être précis. Cette démarche soulève des problèmes d'ordre statistique tels que la stationnarité de la série temporelle et la pertinence des observations remontant à plusieurs années. Dans [22], l'auteur s'intéresse à la corrélation entre un ETF (*Exchange Traded Fund*) sur le S&P500 et d'autres ETF indexés sur des secteurs d'activité économique. Le profil de corrélation est calculé sur une base horaire et présente une forme en U. Nous croyons que le choix d'un découpage horaire n'est pas adapté à la récente accélération de l'activité sur les marchés financiers, où des changements brusques de régime peuvent survenir en quelques minutes, voire secondes.

Notre méthodologie est la suivante. Nous découpons une journée de trading en tranches de cinq minutes, au sein desquelles nous disposons de données de transaction et de cotation. Nous estimons la corrélation réalisée entre deux actifs sur une tranche de cinq minutes par la méthode Hayashi-Yoshida [61]. Cette opération est répétée pour chaque jour sur la période de temps considérée et pour un grand nombre de paires d'actifs financiers. Enfin, pour une tranche horaire donnée, nous calculons la moyenne de la corrélation réalisée lors de cette tranche horaire sur l'ensemble des journées de trading, ce qui nous permet de construire un profil intra-journalier moyen de la corrélation à haute fréquence.

Ce travail a pour objectif de mesurer la saisonnalité intra-journalière des mouvements joints des actifs financiers en utilisant des données *tick-by-tick*. Nous cherchons également à introduire ce phénomène dans un modèle de prix s'appuyant sur des processus de Hawkes. Les principaux résultats sont récapitulés ci-dessous:

- **Profil intra-journalier de la corrélation haute fréquence**: la corrélation à haute fréquence varie significativement au cours d'une journée de trading, de 20% à 140% de son niveau moyen. Le profil mesuré est similaire d'un univers d'actifs à un autre. Les actifs sont très peu corrélés à l'ouverture, puis la corrélation augmente fortement dans la première heure de trading, continue à croître mais à une vitesse plus modérée jusqu'au milieu de la journée. Lorsque des chiffres macroéconomiques sont annoncés ou bien lorsque d'autres marchés ouvrent, la corrélation saute soudainement à la hausse.

Enfin, nous observons un amenuisement substantiel de la corrélation entre les actifs pendant la dernière heure de trading, même si elle ne retrouve pas le niveau particulièrement bas de l'ouverture. Ce profil de corrélation est soumis à des tests de robustesse et de significativité, tous deux passés avec succès.

- **Modèle de Hawkes non-stationnaires**: nous étendons le modèle de prix *tick-by-tick* de [16] afin qu'il puisse capturer le profil intra-journalier de la corrélation. Nous faisons appel pour cela à des processus de Hawkes non-stationnaires dont l'intensité de base et les noyaux d'excitation dépendent du temps courant. Les paramètres du modèle sont estimés par une extension de l'algorithme EM présenté dans [74]. Nous considérons quatre spécifications de notre modèle, de complexité croissante, et estimons chaque modèle sur les séries temporelles de prix de deux actions corrélées du CAC40. Le modèle stationnaire reproduit le niveau de corrélation empirique mais son profil est par définition plat. Il nous faut à la fois rendre l'intensité de base et le noyau d'excitation mutuelle dépendant du temps pour reproduire le profil de corrélation empirique. Cependant, le niveau de corrélation du modèle est deux fois plus bas que celui observé. Nous suggérons des pistes de recherche pour permettre au modèle d'atteindre des niveaux de corrélation réalistes tout en capturant son profil intra-journalier.

Chapter 2

The Times Change: Multivariate Subordination

2.1 Introduction

Since the celebrated Black&Scholes model started its career as a benchmark model in derivatives houses, many attempts have been made to check whether its assumptions are in agreement with real world data. Among those assumptions, the most heavily rejected is that returns in calendar time are normally distributed, or in other words, that volatility does not fluctuate randomly[1]. Indeed, non-gaussian tails (such as power-law or exponential tails) are well documented in empirical literature, see [25] for a comprehensive review.

However, one is prone to ask whether, as studied in the pioneering work of Clark [37], returns may indeed be normally distributed *modulo* a stochastic change of time. In the one-dimensional case considered by Clark and several other authors thereafter, see e.g. [11, 28], a stochastic clock is introduced, the so-called business time, speeding up the usual calendar time as market activity rushes and *vice versa*. Therefore, the random intensity of the market activity flow, through the arrival of trades, provides an explanation for the fluctuating aspect of volatility. In a more mathematical formulation: returns are then driven by a subordinated Brownian motion, the subordinator being the business time, leading to (semi-)heavy-tailed distributions[2]. Moreover, the degree of heaviness is easily seen to depend on the distribution of arrival times. It is noteworthy that - to the best of our knowledge - no such approach has been previously applied to the multidimensional case, and the aim of the work presented in this paper is precisely to do so. Inspired by a multidimensional version of the Central Limit Theorem, we introduce a suitable, rather natural multidimensional business time and study some of its elementary properties. We then determine the influence of this stochastic change of time on the joint distribution of returns, our main objective being to introduce a finer-grain, more accurate description of the origin of covariance between pairs of assets. Both the statistical estimates and the models that we present in this note are new, and we strongly believe that the use of multidimensional trade time for the modelling of correlated assets at the microstructure level is prone to many interesting new developments.

The paper is organized as follows: in 2.2, we provide a very simple theoretical background supporting Clark's model as well as the empirical and statistical studies presented in [36] and extend it to the multivariate case. After a brief description, in section 2.3, of the data sets we use, we confront our approach to high frequency data in section 2.4, unveiling a strikingly convincing agreement. Section 2.5 states our temporary conclusions as well as questions for future research.

[1] assuming the absence of jumps in the dynamics of prices.

[2] A probability distribution with survival function $\bar{F}$ is said to exhibit heavy tails if $\lim_{x\to+\infty} \bar{F}(x)e^{\lambda x} = +\infty$ for every $\lambda > 0$. It has semi-heavy tails if $\lim_{x\to+\infty} \bar{F}(x)e^{\lambda x} < +\infty$ and $\lim_{x\to+\infty} \bar{F}(x)e^{\lambda x^2} = +\infty$ for every $\lambda > 0$.

2.2 Multivariate event time

2.2.1 Univariate case

Let us consider an asset whose price fluctuates randomly during the trading hours. Between the $(i-1)^{th}$ and the i^{th} trade, the performance of the asset is simply $F_i := P_i/P_{i-1}$. Then, N trades away from the opening price P_0, the total variation is given by the product of these elementary ratios

$$\frac{P_N}{P_0} = \prod_{i=1}^{N} F_i$$

We are then left with the following expression for the relative price increment, *i.e.* the asset return

$$R_N := \ln(\frac{P_N}{P_0}) = \sum_{i=1}^{N} \ln(F_i)$$

The asset return is clearly the sum of random variables and we would like to apply a version of the Central Limit Theorem (CLT). The basic CLT is stated for independent and identically distributed random variables, a property which clearly fails to hold if one considers the returns of a financial asset: it has been well documented, and is easily verified experimentally, that absolute values of returns are autocorrelated [26]. However, the CLT can be extended to the more general case of weakly dependent variables $X_1, \ldots, X_n$ [98]. The main condition for it to hold is the existence of the asymptotic variance

$$\lim_{n\to+\infty} \mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\right) = \mathbb{V}\mathrm{ar}\left(X_1\right) + 2\sum_{k=1}^{+\infty} \mathbb{C}\mathrm{ov}\left(X_1, X_{1+k}\right)$$

assuming that the X_i's form a weak-sense stationary sequence.

If the sum above is finite, *i.e.* if the autocorrelation function of $\ln(F_i)$ decays fast enough[3], then the CLT yields, as $N \to +\infty$

$$\frac{R_N}{\sqrt{N}} \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

where $\mathcal{N}(\mu, s^2)$ denotes the Gaussian distribution with mean μ and variance s^2 and $\sigma^2 := \lim_{N\to+\infty} \mathbb{V}\mathrm{ar}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \ln(F$
We have assumed $\mathbb{E}\left(\ln(F_1)\right) = 0$[4]. Hence, there holds, for $N \sim \infty$

$$R_N \sim \mathcal{N}(0, N\sigma^2)$$

i.e. returns are asymptotically normally distributed with variance proportional to the number of trades when they are sampled in trade time. Recast in the context of stochastic processes, the returns can therefore be viewed as a Brownian motion in a stochastic clock (such processes are called subordinated Brownian motions), the clock being the number of trades.

Note that the same line of reasoning applies with the traded volume as the stochastic clock. Indeed, the return after a volume V has been traded is

[3] In the case of price returns $\ln(F_i)$, the autocorrelation function decays very fast and can be considered as statistically insignificant after some lag k close to one, even for small time scales [1, 2]. Therefore the above sum is finite in practice.

[4] It is a reasonable assumption since we are dealing with high frequency data.

$$R_V = \sum_{i=1}^{N_V} \ln(F_i)$$
$$V = \sum_{i=1}^{N_V} V_i$$

where V_i is the volume traded during transaction i. N_V is the number of transactions needed to reach an aggregated volume V. When scaling with the square root of the traded volume, we get

$$\begin{aligned}\frac{R_V}{\sqrt{V}} &= \sqrt{\frac{N_V}{V}} \frac{\sum_{i=1}^{N_V} \ln(F_i)}{\sqrt{N_V}} \\ &= \frac{1}{\sqrt{\frac{\sum_{i=1}^{N_V} V_i}{N_V}}} \frac{\sum_{i=1}^{N_V} \ln(F_i)}{\sqrt{N_V}}\end{aligned}$$

Since the volume of each transaction is finite, $V \to +\infty$ implies $N_V \to +\infty$. From the law of large numbers, we have $\frac{\sum_{i=1}^{N_V} V_i}{N_V} \to \mathbb{E}(V_1)$ as $N_V \to +\infty$, which is the average volume traded in a single transaction. Thus, applying Slutsky's theorem, we have, when $V \to +\infty$

$$\frac{R_V}{\sqrt{V}} \xrightarrow{d} \mathcal{N}(0, \frac{\sigma^2}{\mathbb{E}(V_1)})$$

Note that the asymptotic variance is rescaled by the average volume traded in a single transaction. This is natural when comparing several assets because the number of shares can vary widely between stocks, take for instance a high priced asset and a penny stock. Also when dealing with a single asset, it means that the magnitude of returns is smaller during periods with large volume traded on a single transaction, though the aggregated volume is identical. Periods with large volumes traded by transaction can be associated to periods with high volume available at the best quantity since trades-through are not frequent [6]. Therefore, volume time models tell us that it is not the absolute volume traded that matters to exlpain volatility. It is rather the volume traded compared to the volume traded in a single transaction, which might implicitly reflect the liquidity available in the order book.

Now, the number of trades or the traded volume over a time period is obviously random. Therefore, the Δt-return $R_{\Delta t} := \ln(\frac{P_{\Delta t}}{P_0})$ in calendar time exhibits a random variance $\sigma^2 X_{\Delta t}$ where $X_{\Delta t}$ is either the number of trades or the traded volume occurring during a time period of length Δt. The distribution of calendar time returns $R_{\Delta t}$ can be recovered from subordinated returns $R_{X_{\Delta t}}$ through the application of Bayes′ formula[5]

$$P_{R_{\Delta t}}(r) = \int_0^{+\infty} P_{\mathcal{N}(0, x\sigma^2)}(r) P_{X_{\Delta t}}(x) \mathrm{d}x$$

where $P_Z(z)$ is the probability density function of the random variable Z. It is easy to show that the resulting distribution has fatter tails than the Gaussian (see appendix 2.7).

Slightly rephrasing the equations above, one has that $R_{\Delta t} = \sigma\sqrt{X_{\Delta t}} Z$ in distribution, where $Z \sim \mathcal{N}(0,1)$ independently from the value of $X_{\Delta t}$ and σ is a scaling constant. In order to derive the final expression

[5]We assume that the distribution of the trading activity $X_{\Delta t}$ (number of trades or traded volume) can be approximated by a continuous distribution.

for $P_{R_{\Delta t}}$, $P_{X_{\Delta t}}$ must be specified. Nevertheless, some model-free properties can be established, such as the computation of moments. Defining $\lambda_k(X) := \frac{\mathbb{E}(X^k)}{\mathbb{E}(X^2)^{k/2}}$ the k^{th} dimensionless moment of X, there holds:

$$\begin{aligned}\mathbb{E}\left(R_{\Delta t}^k\right) &= \mathbb{E}\left(X_{\Delta t}^{k/2}\mathbb{E}\left(Z^k|N_{\Delta t}\right)\right)\\ &= \sigma^k\mathbb{E}\left(X_{\Delta t}^{k/2}\right)\mathbb{E}\left(\mathcal{N}(0,1)^{k/2}\right)\\ &= \sigma^k\mathbb{E}\left(X_{\Delta t}^{k/2}\right)\left(\frac{k}{2}\right)!!\mathbb{1}_{\{\exists m\in\mathbb{N}^*:k=2m\}}\end{aligned}$$

where $n!! := 1\times 3\times\ldots\times(n-1)$ for n even (see appendix 2.6). It leads to

$$\lambda_k\left(R_{\Delta t}\right) = \frac{\mathbb{E}\left(X_{\Delta t}^{k/2}\right)}{\mathbb{E}\left(X_{\Delta t}\right)^{k/2}}\left(\frac{k}{2}\right)!!\mathbb{1}_{\{\exists m\in\mathbb{N}^*:k=2m\}}$$

For $k=4$, we obtain

$$\lambda_4\left(R_{\Delta t}\right) = 3\frac{\mathbb{E}\left(X_{\Delta t}^2\right)}{\mathbb{E}\left(X_{\Delta t}\right)^2} \geq 3 = \lambda_4\left(\mathcal{N}(0,1)\right)$$

thanks to Jensen's inequality[6]. As the fourth moment provides information on the heaviness of the tails, we see that such a model predicts distribution tails fatter than those of the normal distribution for $R_{\Delta t}$. Therefore, stochastic volatility, *via* the random fluctuations of the number of trades or traded volume, provides an explanation for the fat tails of returns sampled in calendar time[7].

Moreover, the shape of the tails of $P_{R_{\Delta t}}$ is given by the asymptotic behaviour of $P_{X_{\Delta t}}$. For instance, it can be shown that if $X_{\Delta t}$ obeys an inverse gamma (resp. gamma) distribution, which has heavy (resp. semi-heavy) tails, then $P_{R_{\Delta t}}$ is a Student (resp. Variance Gamma) law [25, 29], therefore exhibiting heavy (resp. semi-heavy) tails. A sketch of the proof is given in appendix 2.7 using the saddle point approximation [45].

As mentioned in the introduction, this mechanism has been extensively studied in the finance literature [36], but only in the univariate framework. We want to generalize it to the multivariate case, which would turn into a stochastic covariance model to take into account the deviation of the empirical multivariate distribution of returns from the multivariate Gaussian.

2.2.2 Multivariate case

We now turn to the more interesting case of a basket of $d\in\mathbb{N}^*$ assets. How can we extend the previous framework to take several assets into account? Let us assume that an event time N is defined, and that we sample returns $R_N = (R_N^1,\ldots,R_N^d)^T$ according to this time. Using a multivariate CLT, we follow the same line of reasoning than in the previous section and obtain for $N\sim\infty$

$$R_N \sim \mathcal{N}_d\left(0, N\Sigma\right)$$

where $\mathcal{N}_d\left(\mu, M\right)$ denotes the d-variate Gaussian distribution with mean μ and covariance matrix M and $\Sigma = \lim_{N\to+\infty}\mathbb{V}\text{ar}\left(\frac{1}{\sqrt{N}}\sum_{i=1}^N\ln(F_i)\right)$.

[6] The equality is reached iff $X_{\Delta t}$ is almost surely constant.

[7] As pointed out in [53], there might be other factors explaining the volatility of financial markets on a microscopic time scale, such as gaps present in the order book. However, the number of trades or the traded volume tends to be a reliable proxy for variance and covariance according to [36] and our study.

For a given time interval Δt, the respective numbers of trades or traded volumes for each asset $N^1_{\Delta t}, \ldots, N^d_{\Delta t}$ are obviously different, and we need to define a global event time $N_{\Delta t} = f\left(N^1_{\Delta t}, \ldots, N^d_{\Delta t}\right)$. Indeed, sampling prices in their respective univariate event times would result in asynchronous returns. We suggest to use $N_{\Delta t} = \sum_{i=1}^d N^i_{\Delta t}$, which amounts to increment time as soon as a trade occurs on any one of the d assets. The increment is either one if we consider trading time or the number of shares traded at that transaction in the case of volume time. This choice seems to us the simplest and most intuitive generalization of the univariate case, since it amounts to considering a single asset that pools together trading activities of all assets: we aggregate the time series of prices of each asset in chronological order, and then count trades or volumes as in the univariate case[8].

Since univariate central limit theorems apply for each asset under consideration, there should be restrictions on diagonal terms of the covariance matrix. Indeed, univariate CLTs imply that

$$\frac{R^i_{N^i}}{\sqrt{N^i}} \xrightarrow{d} \mathcal{N}\left(0, \sigma_i^2\right) \quad \text{as } N^i \to +\infty \quad \forall i \in \{1, \ldots, d\}$$

where $\sigma_i^2 = \lim_{N^i \to +\infty} \frac{1}{N^i} \mathbb{V}\text{ar}(\sum_{k=1}^{N^i} \ln(F^i_k))$.

Then, as we showed earlier

$$\frac{R_N}{\sqrt{N}} \xrightarrow{d} \mathcal{N}\left(0, \Sigma\right) \quad \text{as } N \to +\infty$$

where $R_N = (R^1_N, \ldots, R^d_N)^T$ and $\Sigma = \lim_{N \to +\infty} \frac{1}{N} \mathbb{V}\text{ar}(\sum_{k=1}^{N} \ln(F_k))$ is the $d \times d$ limiting covariance matrix. In particular,

$$\begin{aligned}
\Sigma_{ii} &= \lim_{N \to +\infty} \frac{1}{N} \mathbb{V}\text{ar}(\sum_{k=1}^{N} \ln(F^i_k)) \\
&= \lim_{N \to +\infty} \frac{N^i}{N} \frac{1}{N^i} \mathbb{V}\text{ar}(\sum_{k=1}^{N^i} \ln(F^i_{g^i(k)}))
\end{aligned}$$

where

$$\begin{aligned}
g^i(1) &= \inf\left\{p > 1 \mid F^i_p \neq 1\right\} \\
g^i(k) &= \inf\left\{p > g^i(k-1) \mid F^i_p \neq 1\right\} \quad \forall k \geq 2
\end{aligned}$$

The second equality comes from the fact that only events of type i make the price of the asset i change. Let us define

$$C_i = \lim_{N \to +\infty} \frac{N^i}{N} = \lim_{N \to +\infty} \frac{N^i(N)}{N}$$

which is the limiting proportion of events of type i among all events. $C_i > 0$ requires that $N^i(N)$ goes to infinity with N but the ratio is fixed. If $C_i = 0$ then the i^{th} marginal of the limiting multivariate Gaussian distribution is trivial, which is why we impose the condition $C_i > 0$. Under this condition,

[8] Such a representation makes sense when the orders of magnitude of the liquidity of each of the d assets are similar: consider the extreme case of one heavily traded asset and another one with only one trade per day. In this case, the normality of the couple as $N_{\Delta t}$ increases may fail to appear before an unrealistically large number of trades. A more appropriate clock might thus be to wait that each asset has been updated at least N times, which amounts to choose $N_{\Delta t} = \min(N^1_{\Delta t}, \ldots, N^d_{\Delta t})$.

$$\lim_{N\to+\infty}\frac{1}{N^i}\mathbb{V}\mathrm{ar}(\sum_{k=1}^{N^i}\ln(F^i_{g^i(k)})) = \lim_{N^i\to+\infty}\frac{1}{N^i}\mathbb{V}\mathrm{ar}(\sum_{k=1}^{N^i}\ln(F^i_{g^i(k)}))$$
$$= \sigma_i^2$$

Then, we have

$$\Sigma_{ii} = C_i\sigma_i^2$$

Therefore, diagonal terms of the covariance matrix are imposed by the variances σ_i^2 stemming from univariate CLTs and the liquidity ratios C_i. This insures that $\mathbb{V}\mathrm{ar}(R^i_N) \sim N\Sigma_{ii} \sim NC_i\sigma_i^2 \sim N_i\sigma_i^2$, which is consistent with the univariate CLT for asset i.

There are also consistency conditions on nondiagonal elements of the covariance matrix. Since elements of the covariance matrix depict pairwise linear dependencies, they are imposed by pairwise bivariate CLTs. We have

$$\begin{aligned}\Sigma_{ij} &= \lim_{N\to+\infty}\frac{1}{N}\mathbb{C}\mathrm{ov}(\sum_{k=1}^{N}\ln(F^i_k),\sum_{k=1}^{N}\ln(F^j_k))\\ &= \lim_{N\to+\infty}\frac{N^{ij}}{N}\frac{1}{N^{ij}}\mathbb{C}\mathrm{ov}(\sum_{k=1}^{N^i}\ln(F^i_{g^i(k)}),\sum_{k=1}^{N^j}\ln(F^j_{g^j(k)}))\\ &= C_{ij}\tilde{\Sigma}_{ij} = (C_i+C_j)\tilde{\Sigma}_{ij}\end{aligned}$$

where $N^{ij} = N^i + N^j$ is the bivariate event time on assets i and j, $C_{ij} = \lim_{N\to+\infty}\frac{N^{ij}}{N} = C_i + C_j$ is the proportion of events of type i or j among all events and $\tilde{\Sigma}_{ij}$ is the limiting covariance coming from the bivariate CLT involving assets i and j.

When dealing with several assets, the correlation matrix is an important statistical quantity which might be used for portfolio diversification for instance. Since we are looking at price returns in either calendar or event time, it is natural to compare correlations computed on either clocks. The covariance matrix of returns in calendar time is given by

$$\begin{aligned}C_{\Delta t} &= \mathbb{E}(R^T_{\Delta t}R_{\Delta t})\\ &= \mathbb{E}(\mathbb{E}(R^T_{\Delta t}R_{\Delta t}|N_{\Delta t}))\\ &= \mathbb{E}(N_{\Delta t})\Sigma\end{aligned}$$

Therefore, the correlation matrix in calendar time is[9]

$$\begin{aligned}\rho_{\Delta t} &= \mathrm{diag}(C_{\Delta t})^{-\frac{1}{2}}C_{\Delta t}\mathrm{diag}(C_{\Delta t})^{-\frac{1}{2}}\\ &= \mathrm{diag}(\Sigma)^{-\frac{1}{2}}\Sigma\mathrm{diag}(\Sigma)^{-\frac{1}{2}}\end{aligned}$$

Note that the correlation matrix does not depend on the time scale Δt. This is in contradiction with empirical evidence for small Δt: the Epps effect states that "*Correlations among price changes* [...] *are found to decrease with the length of the interval for which the price changes are measured.*" [49]. However,

[9]For a given matrix M, we note $\mathrm{diag}(M)$ the diagonal matrix with the same diagonal elements as M.

the correlation is found to reach its asymptotic value quickly, in a few minutes on the assets we consider as we shall see in section 2.4.4. The correlation of returns sampled in event time is equal to the correlation in calendar time since

$$C_N = \mathbb{E}(R_N^T R_N) = N\Sigma$$

which leads

$$\begin{aligned}\rho_N &= \mathrm{diag}(C_N)^{-\frac{1}{2}} C_N \mathrm{diag}(C_N)^{-\frac{1}{2}} \\ &= \mathrm{diag}(\Sigma)^{-\frac{1}{2}} \Sigma \mathrm{diag}(\Sigma)^{-\frac{1}{2}}\end{aligned}$$

2.3 Data description

We have access to the Thomson Reuters Tick History database (see section 1.3) which provides tick-by-tick data on many financial assets (equities, fixed income, forex, futures, commodities, *etc*). Three levels of data are available:

- trades files: each transaction price and quantity timestamped up to the millisecond
- quotes files: each quote (best bid and ask) price or quantity modification timestamped up to the millisecond
- order book files: each limit price or quantity modification timestamped up to the millisecond, up to a given depth, typically ten limits on each side of the order book.

Throughout this study, we will use trades and quotes files. We will sample quotes on a trading time basis so that for each trade we have access to the quotes right before this trade. We do this because the returns we will consider are the returns of the midquote in order to get rid of the bid/ask bounce. Therefore, it might sound more natural to consider every best quote modification since events other than trades, such as limit orders placed in the bid/ask spread, cancellation orders for the whole quantity available at the best limit and trades-though, can affect the midquote. However, we favor trading time sampling over best quotes sampling because in our opinion trades represent more significant events than quotes changes. Indeed, only trades involve money exchanges.

When a trade walks the order book up or down by hitting consecutive limit prices, it is recorded as a sequence of trades with the same timestamp but with prices and quantities corresponding to each limit hit. For instance, assume that the best ask offers 100 shares at price 10 and 200 shares at price 10.01, and that a buy trade arrives for 150 shares. This is recorded as two lines in the trades file with the same timestamp, the first line being 100 shares at 10 and the second line 50 shares at 10.01. As a pre-processing step, we aggregate identical timestamps in trades files by replacing the price by the volume weighted average price (VWAP) over the whole transaction and the volume by the sum of all quantities consumed. In the previous example, the trade price will thus be $(100*10+50*10.01)/(100+50) = 10.00333$ and the trade quantity $100+50 = 150$.

The assets we consider are French stocks between 01/03/2010 and 31/05/2010. We drop the first and last hours of trading because they show a very diferent trading pattern than the rest of the day. By doing so, we limit seasonality effects.

2.4 Empirical results

In this section, we test our theory against high-frequency multivariate data. The main statements are:

- Do returns become jointly normal when sampled in trade (resp. volume) time as N (resp. V) grows ?
- Is the empirical covariance matrix of returns scaling linearly with N or V? Is the empirical covariance scaling down with $\mathbb{E}(V_1)$?
- What do $P_{N_{\Delta t}}$ and $P_{V_{\Delta t}}$ look like?
- Is the correlation of returns independent of the time scale and definition of time?

We focus on four pairs of assets:

- BNPP.PA/SOGN.PA: BNP Paribas/Société Générale
- RENA.PA/PEUP.PA: Renault/Peugeot
- EDF.PA/GSZ.PA: EDF/GDF Suez
- TOTF.PA/EAD.PA: Total/EADS

For sampling in trading time, we choose to sample every 2^i trades for $i = 0, 1, \ldots, 12$ ($2^{12} = 4096$). The samplings in calendar and volume time are chosen so that they match the trading time sampling on average. This means that if we sample every N trades, then the corresponding time scale Δt is the average time that it takes to observe N consecutive trades. In the same way, V is the average volume traded after N trades. Table 2.1 describes the corresponding average durations and volumes for the four aforementioned pairs of assets. Since we want to avoid the use of overnight returns[10] and to get approximately the same number of points for each sampling period, we sample prices with overlap. For instance, assume that we have a series of prices P_1, P_2, P_3, P_4, P_5, and that we sample prices every two trades. Then we get 3 returns if we sample with overlap, namely $\ln(P_3/P_1), \ln(P_4/P_2), \ln(P_5/P_3)$. On the contrary, if we sample without overlap, we get 2 returns $\ln(P_3/P_1), \ln(P_5/P_3)$. The drawback of sampling with overlap is that our points cannot be considered as independent, which is a central assumption of many statistical testing procedures.

Table 2.1: Average durations $\langle \Delta t \rangle$ and traded volumes $\langle V \rangle$ corresponding to a given number of trades N

	BNPP.PA/SOGN.PA		RENA.PA/PEUP.PA		EDF.PA/GSZ.PA		TOTF.PA/EAD.PA	
N	$\langle \Delta t \rangle$	$\langle V \rangle$	$\langle \Delta t \rangle$	$\langle V \rangle$	$\langle \Delta t \rangle$	$\langle V \rangle$	$\langle \Delta t \rangle$	$\langle V \rangle$
1	1.643	705	3.61	904	3.146	908	2.514	1168
2	3.286	1057	7.22	1355	6.292	1361	5.029	1752
4	6.573	1761	14.441	2259	12.583	2269	10.058	2920
8	13.148	3170	28.883	4066	25.169	4084	20.119	5255
16	26.3	5988	57.785	7680	50.346	7716	40.246	9927
32	52.616	11625	115.651	14908	100.724	14980	80.527	19271
64	105.305	22897	231.576	29356	201.64	29509	161.208	37958
128	210.879	45443	464.387	58233	404.257	58561	323.013	75326
256	422.928	90522	933.208	115957	812.387	116708	648.534	150035
512	850.21	180642	1888.43	231274	1638.738	233025	1306.37	299628
1024	1718.409	360728	3871.216	461378	3330.381	465452	2649.743	598957
2048	3507.494	721162	8046.594	921883	6878.099	933056	5437.687	1196058
4096	7220.702	1445566	15417.677	1842793	13919.229	1861140	10909.276	2388092

[10]The use of overnight returns would be inappropriate in our framework because there is a significant trading activity during the opening and closing auctions, which would bias the event time clocks. Moreover, we want to focus only on the intraday behavior of returns when there is continuous trading between the starting and ending timestamps of returns.

2.4.1 Multivariate normality

Testing for a multivariate distribution is a complex issue. Extensions of standard univariate procedures such as the Kolmogorov-Smirnov test [72] or the Cramer-von-Mises test [10] are problematic because there is no straightforward definition of a distance between two multivariate cumulative distribution functions. Several solutions have been suggested in the statistical literature for the assessment of multivariate normality, see [62] for a comprehensive review. We choose to tackle this issue by using the spherical decomposition of a Gaussian random vector. Let us consider a random vector $X \sim \mathcal{N}_d(\mu, \Sigma)$. Then $X \overset{d}{=} \Sigma^{\frac{1}{2}} Y + \mu$ where $Y \sim \mathcal{N}_d(0, I)$. The spherical decomposition of Y is given by

$$\begin{cases} Y \overset{d}{=} RU(\theta) \\ R^2 \sim \chi^2(d) \\ \theta \sim p(\theta_1, \dots, \theta_{d-1}) = \frac{\Gamma(\frac{d}{2})}{2\pi^{\frac{d}{2}}} \prod_{i=1}^{d-2} \sin(\theta_{d-1-i})^i \mathbb{1}_{[0,\pi]}(\theta_1, \dots, \theta_{d-2}) \mathbb{1}_{[0,2\pi]}(\theta_{d-1}) \\ \theta \text{ and } R \text{ are independent} \\ U(\theta) = \begin{pmatrix} \sin(\theta_1)\sin(\theta_2)\dots\sin(\theta_{d-2})\sin(\theta_{d-1}) \\ \sin(\theta_1)\sin(\theta_2)\dots\sin(\theta_{d-2})\cos(\theta_{d-1}) \\ \sin(\theta_1)\sin(\theta_2)\dots\sin(\theta_{d-3})\cos(\theta_{d-2}) \\ \vdots \\ \sin(\theta_1)\cos(\theta_2) \\ \cos(\theta_1) \end{pmatrix} \end{cases}$$

where $\chi^2(d)$ is the chi-square distribution with d degrees of freedom and $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} \mathrm{d}t$ is the Gamma function. The proof is given in appendix 2.8.
In the case $d = 2$, we have $U = (\cos(\theta), \sin(\theta))^T$ where θ is uniformly distributed on $[0, 2\pi]$. Intuitively, in the (Y^1, Y^2)-plane, R is the radius of Y, representing its amplitude, and θ is its angle, describing the correlation between Y^1 and Y^2. The pair (R, θ) can be identified from Y since

$$\begin{aligned} R &= \|Y\|_2 = \sqrt{(Y^1)^2 + (Y^2)^2} \\ \theta &= \begin{cases} \arctan(Y^2/Y^1) & \text{if } Y^1 > 0 \text{ and } Y^2 \geq 0 \\ \arctan(Y^2/Y^1) + 2\pi & \text{if } Y^1 > 0 \text{ and } Y^2 < 0 \\ \arctan(Y^2/Y^1) + \pi & \text{if } Y^1 < 0 \\ \frac{\pi}{2} & \text{if } Y^1 = 0 \text{ and } Y^2 > 0 \\ \frac{3\pi}{2} & \text{if } Y^1 = 0 \text{ and } Y^2 < 0 \end{cases} \\ &= f(Y) \end{aligned}$$

If $Y^1 = Y^2 = 0$ then θ can take any value in $[0, 2\pi]$ which we choose to set at zero. Let us assume that we have bidimensional i.i.d. data $X_1, \dots, X_n$. The bijection between Y and (R, θ) suggests the following procedure to assess the bivariate normality of X

1. compute the sample average $\mu = \frac{1}{n} \sum_{i=1}^n X_i$ and the sample covariance matrix $\Sigma = \frac{1}{n} \sum_{i=1}^n X_i X_i^T$
2. compute the standardized and centralized observations $Y_1, \dots, Y_n$ where $Y_i = \Sigma^{-\frac{1}{2}}(X_i - \mu)$
3. compute the squared radii (also known as Mahalanobis distances) $R_1, \dots, R_n$ where $R_i = Y_i^T Y_i$
4. compute the angles $\theta_1, \dots, \theta_n$ where $\theta_i = f(Y_i)$
5. test $R^2 \sim \chi^2(2)$ using $R_1, \dots, R_n$

6. test $\theta \sim \mathcal{U}([0, 2\pi])$ using $\theta_1, \ldots, \theta_n$

7. test the independence between R and θ using $R_1, \ldots, R_n$ and $\theta_1, \ldots, \theta_n$

The test of bivariate normality is thus composed of three univariate tests : two goodness-of-fit tests and one independence test.

Figure 2.1 displays the QQ-plot of the squared Mahalanobis distance of returns sampled in volume time against chi-square quantiles with two degrees of freedom for increasing sampling periods. We drop duplicates in the vector $(R_1^2, \ldots, R_n^2)$, which creates a new effective sample size n. We then plot the sorted R_i^2 against the chi-square quantiles $F^{-1}_{\chi^2(2)}(\frac{i-0.5}{n})$ for $i = 1, \ldots, n$. Since n is of the order of 10^5 for our sample, we plot a point every 10^3 points and we keep every point in the last 10^3. If the empirical distribution agrees with the theoretical one, then points should lie on the 45° line.

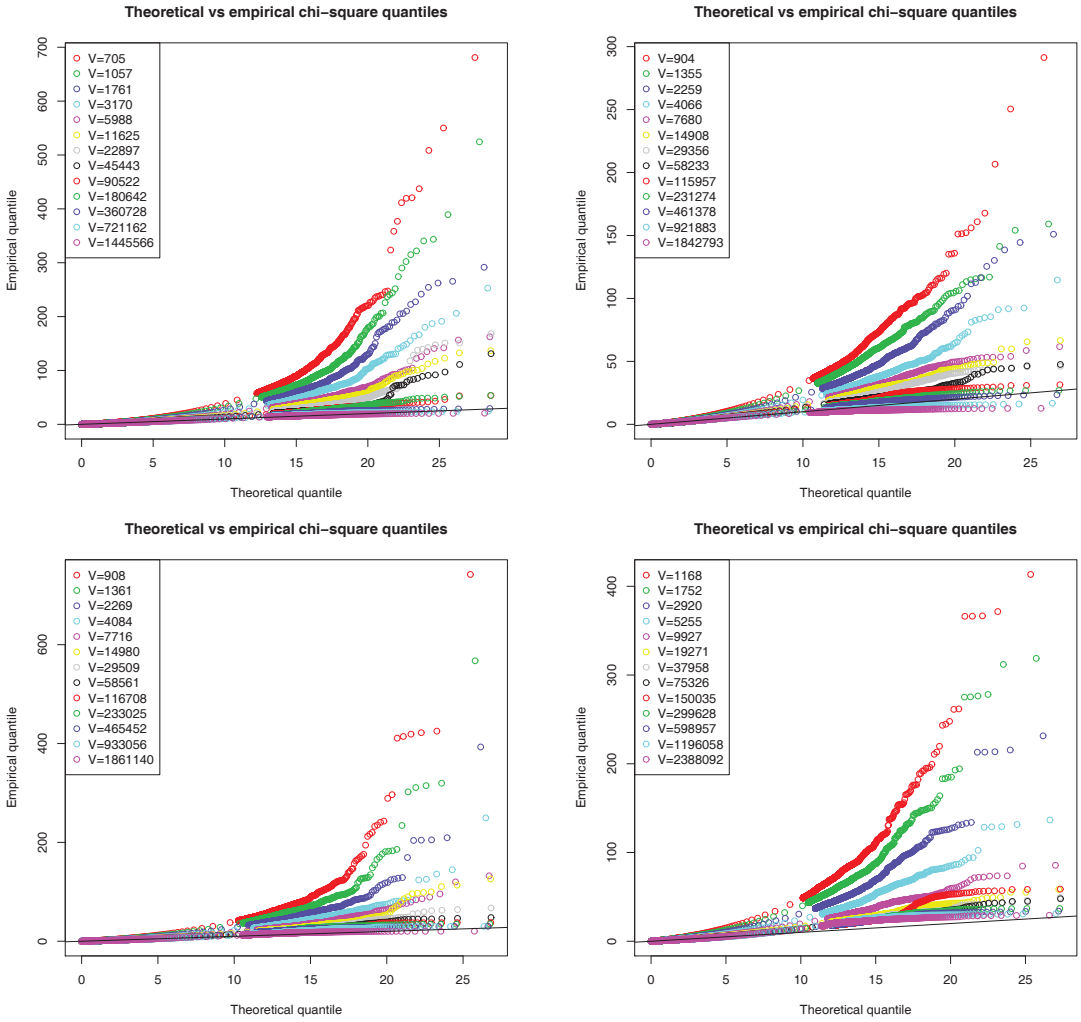

Figure 2.1: QQ-plot of the squared Mahalanobis distance of returns sampled in volume time against chi-square quantiles with two degrees of freedom. The straight black line depicts the 45° line. Top left panel: BNPP.PA/SOGN.PA. Top right panel: RENA.PA/PEUP.PA. Bottom left panel: EDF.PA/GSZ.PA. Bottom right panel: TOTF.PA/EAD.PA.

We see that the empirical Mahalanobis distances measured in volume time become closer to the chi-square distribution as the sampling period increases. This means that the probability of accepting the hypothesis of bivariate Gaussianity with success becomes larger with the event time scale. Note that for TOTF.PA/EAD.PA, returns sampled with the highest time scale under consideration ($V = 2.38 \times 10^6$) are a bit farer from the straight line than for other pairs. This might come from the difference of trading activity between these two stocks. The average duration between two trades for TOTF.PA is 3.283 seconds while it is 12.152 seconds for EAD.PA, which is 3.7 times longer. For BNPP.PA/SOGN.PA (resp. RENA.PA/PEUP.PA, EDF.PA/GSZ.PA), this ratio is 1.01 (resp. 1.7, 1.5). Thus the bivariate event time might be affected mostly by TOTF.PA, leaving EAD.PA pretty much unchanged. As we mentioned it before, the solution could be either to wait more time (measured in events) or to consider a new event time such as the time it takes to have both assets being updated at least a given number of times.

On figure 2.2 we plot the estimated probability density function of normalized angles $\frac{\theta_1}{2\pi}, \ldots, \frac{\theta_n}{2\pi}$ of returns sampled in volume time. For small volumes the density is multimodal and strongly peaked, thus exhibiting obvious deviations from uniformity. The density flattens toward a uniform distribution as the traded volume increases.

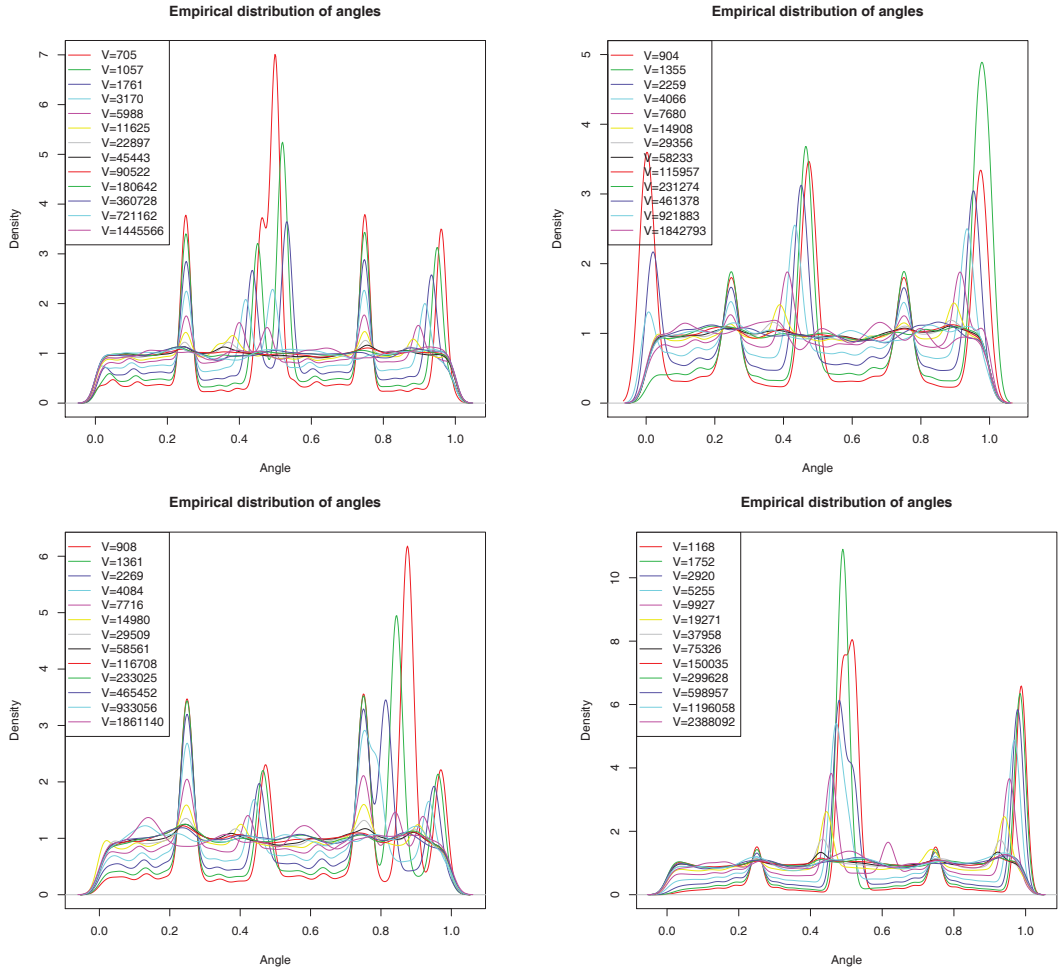

Figure 2.2: Probability density function of the normalized angles $\frac{\theta}{2\pi}$ of returns sampled in volume time. Top left panel: BNPP.PA/SOGN.PA. Top right panel: RENA.PA/PEUP.PA. Bottom left panel: EDF.PA/GSZ.PA. Bottom right panel: TOTF.PA/EAD.PA.

In comparison with volume time sampling, we plot on figure 2.3 the same QQ-plot than on figure 2.1, along with trading and calendar time sampling. We only show the results for the pair BNPP.PA/SOGN.PA and for four selected sampling periods for visual clarity. The findings are similar on the other pairs tested. From this figure it is clear that Mahalanobis distances of returns sampled in volume or trading time are closer to a chi-square distribution than calendar time returns, especially as the sampling period increases. It seems even more plausible for volume time returns than for trading time returns.

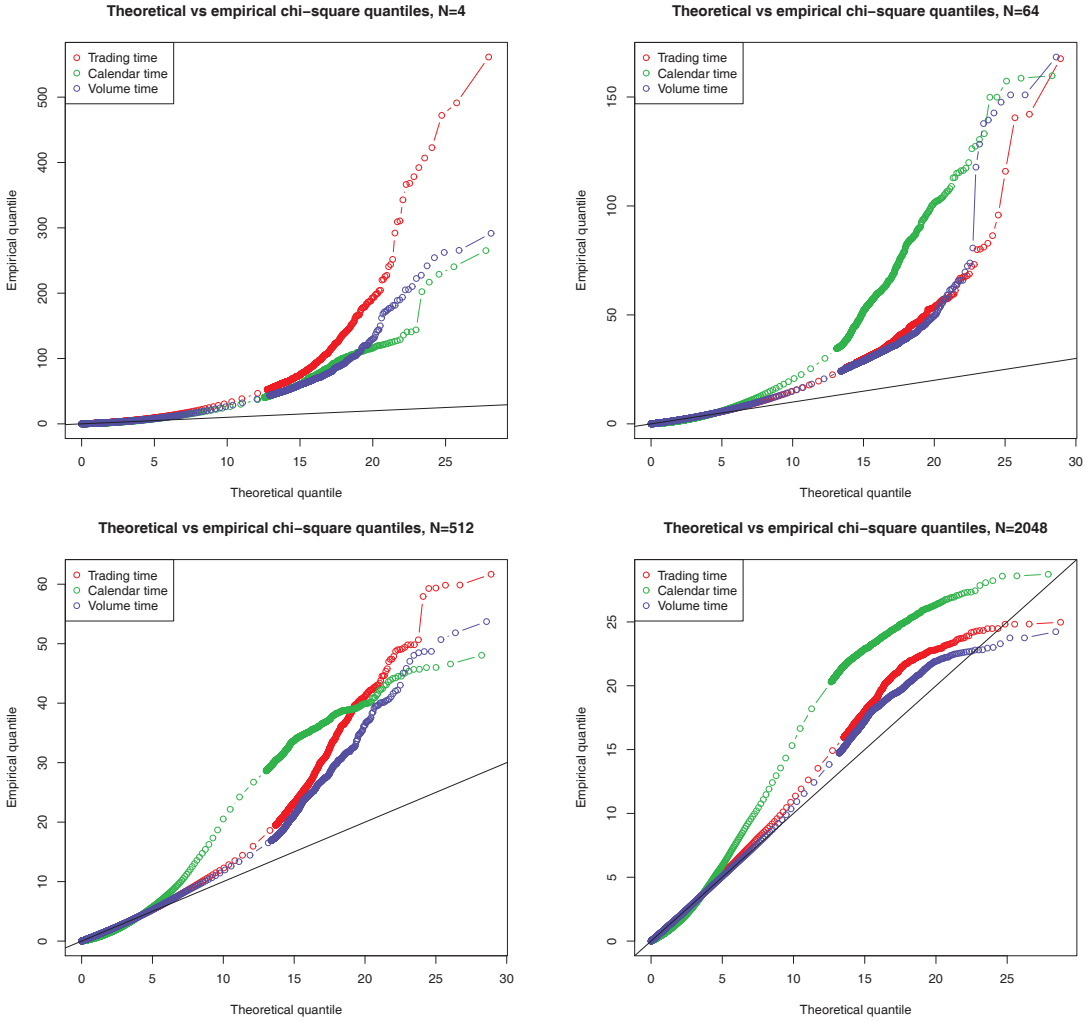

Figure 2.3: QQ-plot of the squared Mahalanobis distance of returns sampled in trading, volume and calendar time against chi-square quantiles with two degrees of freedom for BNPP.PA/SOGN.PA. Top left panel: $N = 4$. Top right panel: $N = 64$. Bottom left panel: $N = 512$. Bottom right panel: $N = 2048$.

In the same way, figure 2.4 compares the distribution of normalized angles computed according to different clocks. It also indicates that the convergence towards unformity happens faster when returns are sampled in event time.

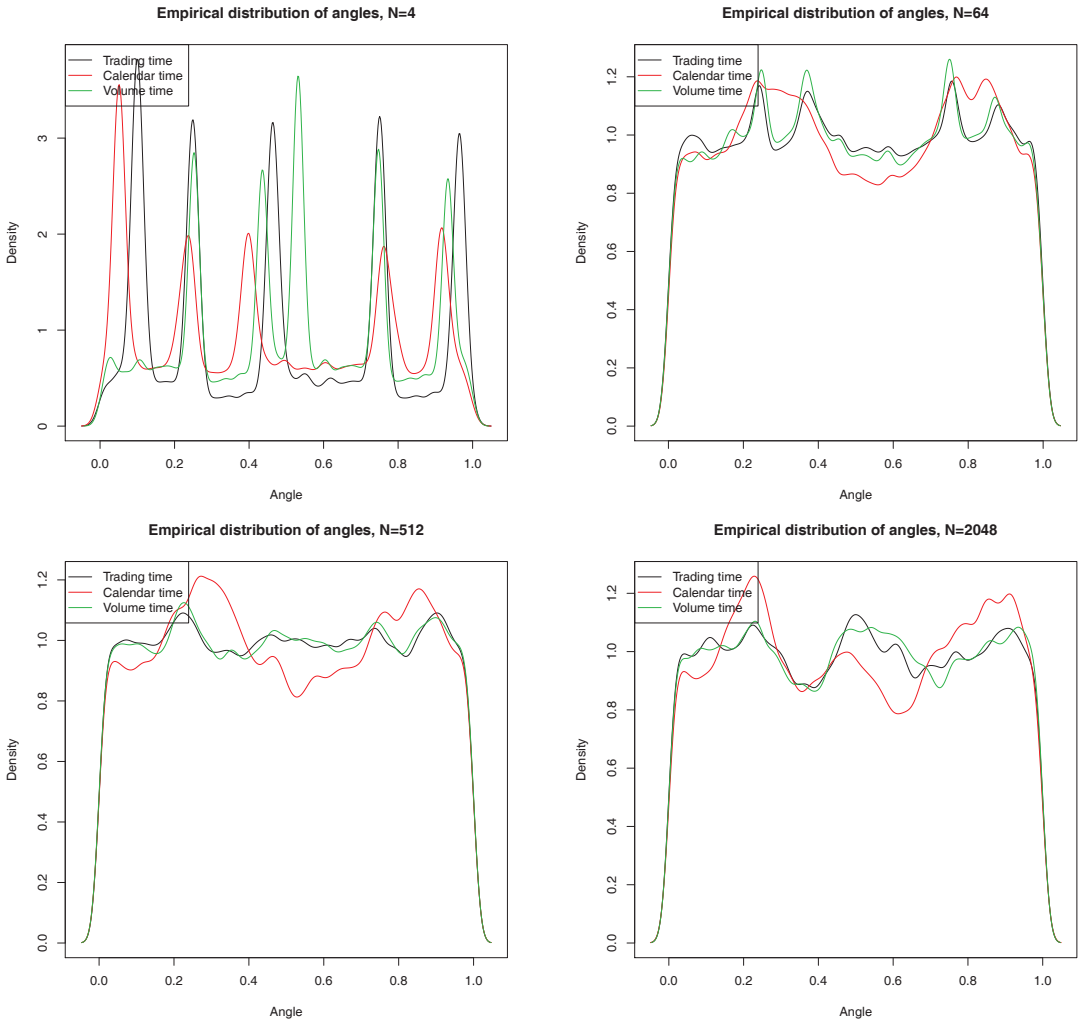

Figure 2.4: Probability density function of the normalized angles $\frac{\theta}{2\pi}$ of returns sampled in trading, volume and calendar time against chi-square quantiles with two degrees of freedom for BNPP.PA/SOGN.PA. Top left panel: $N = 4$. Top right panel: $N = 64$. Bottom left panel: $N = 512$. Bottom right panel: $N = 2048$.

Finally, we depict the behaviour of the dependency between the radii and angles as a function of the time scale of returns on figure 2.5. Multivariate normality requires the independence between radial and angular parts. Since both these quantities are quantitative variables, usual independency tests such as the chi-square test don't apply. We rather compute the correlation between radii and angles to measure their dependency. We use the standard Pearson correlation coefficient as well as Spearman correlation [95]. The Spearman correlation coefficient can detect any non-linear correlation as long as it is monotonic. We observe that these two correlations converge towards zero, which tends to confirm the bivariate normality assumption. We remark that correlations computed in volume time are significantly smaller than in trading time or calendar time. Once again, this leads to favor traded volume time as an appropriate subordinator for prices.

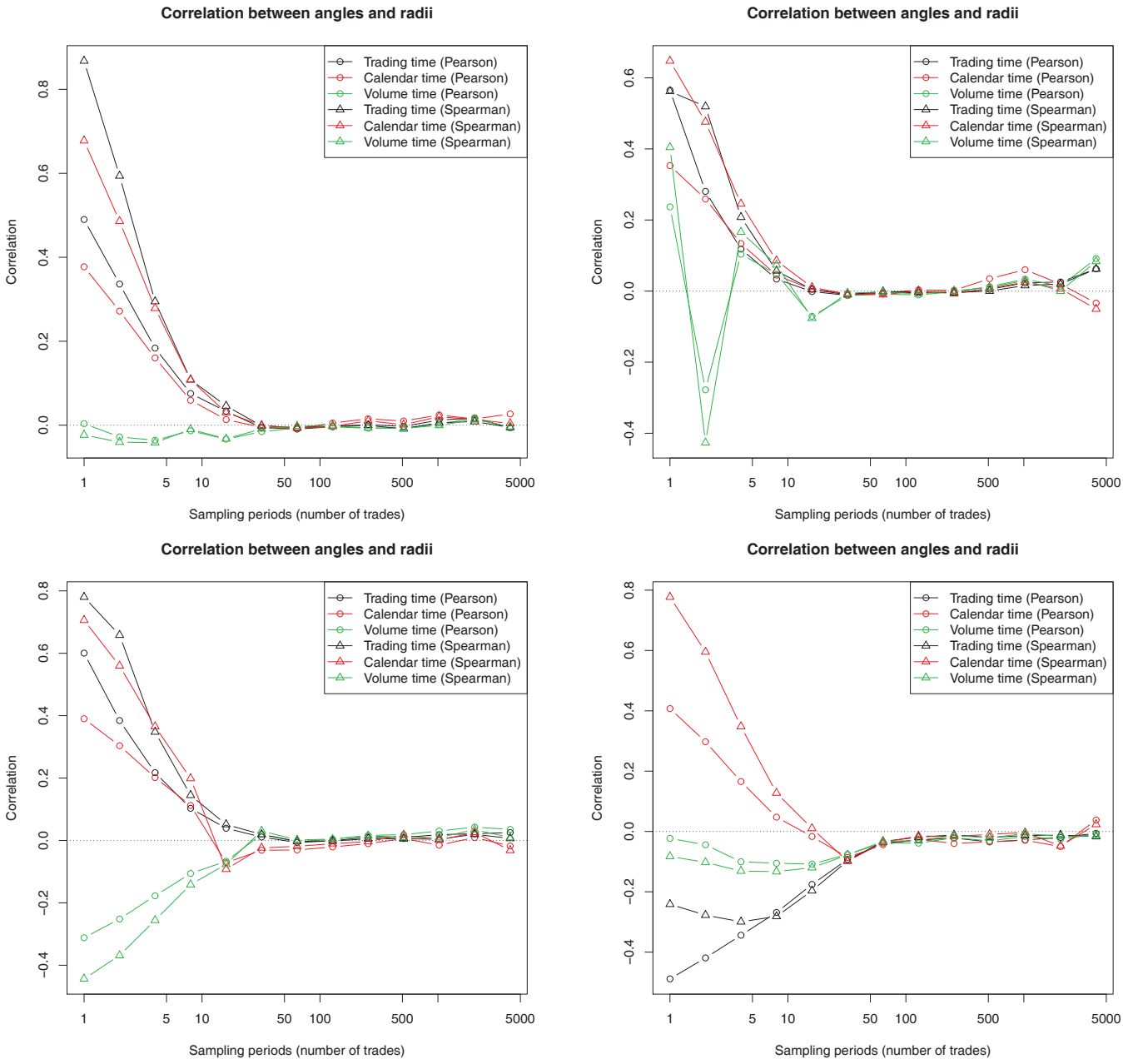

Figure 2.5: Correlation between the radii and the angles of returns sampled in trading, volume and calendar time. Top left panel: BNPP.PA/SOGN.PA. Top right panel: RENA.PA/PEUP.PA. Bottom left panel: EDF.PA/GSZ.PA. Bottom right panel: TOTF.PA/EAD.PA.

2.4.2 Scaling of the covariance matrix

We now consider the scaling of the covariance matrix of returns in event time. Figure 2.6 (resp. 2.7) plots the realized variance (resp. covariance) of returns sampled in trading and volume time as a function of the sampling period, along with a linear fit. Clearly, the linearity of the covariance matrix as a function of either the volume or trading time is in agreement with the data. However, for very large sampling periods, the linearity tends to break down but this might come from the shortage of data. We obtain a much better fit by leaving aside the last sampling period as blue lines show on figures 2.6 and 2.7.

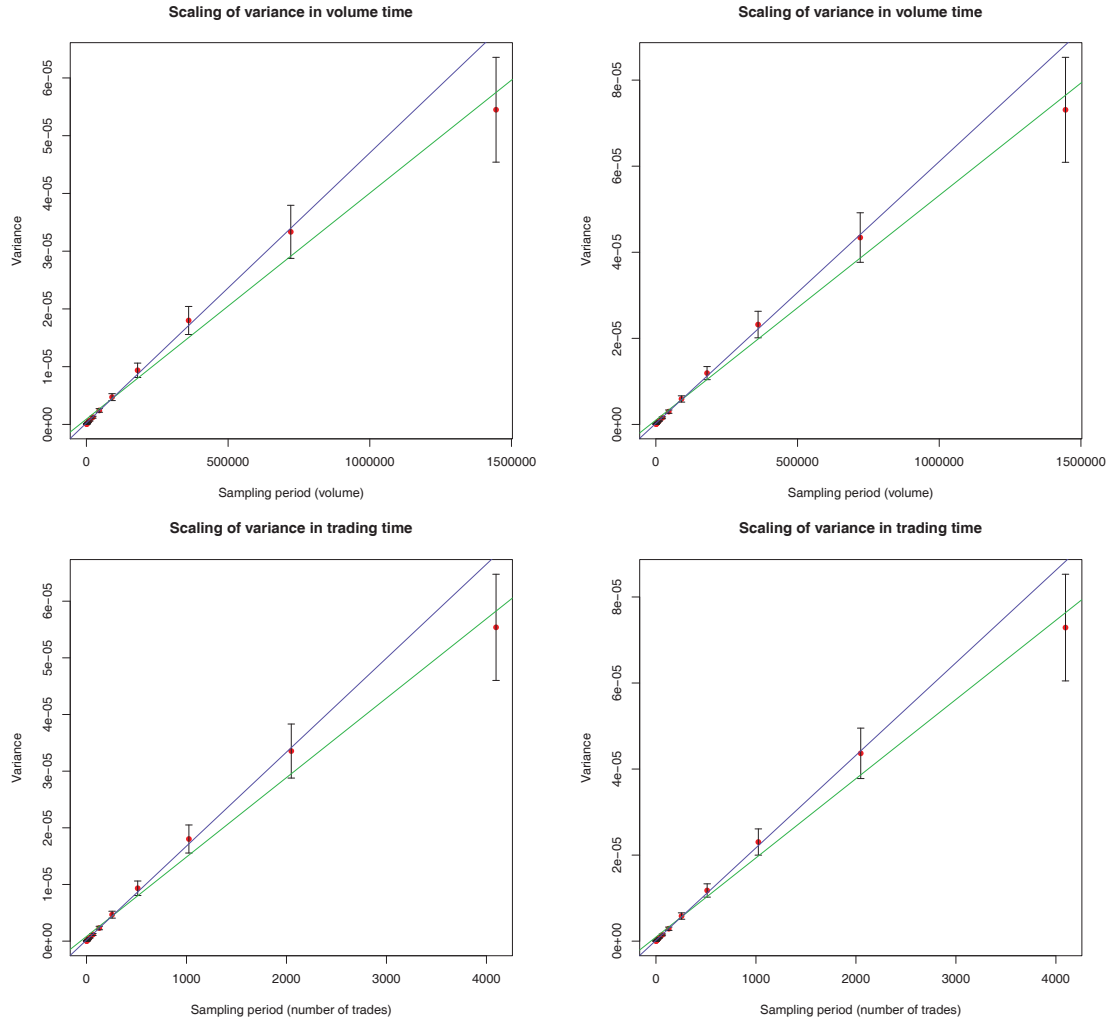

Figure 2.6: Scaling of the realized variance of returns sampled in volume (top panel) and trading (bottom panel) time as a function of the sampling period. The green (resp. blue) line shows the best linear fit with ordinary least squares on all points (resp. on all points except the last one). Left panel: BNPP.PA. Right panel: SOGN.PA.

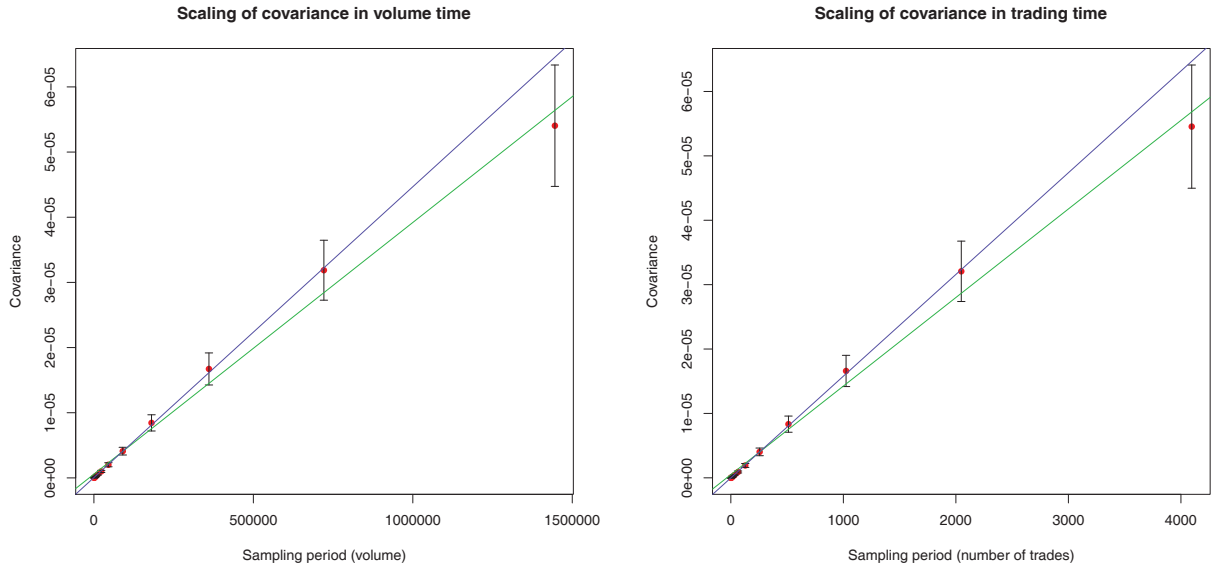

Figure 2.7: Scaling of the realized covariance of returns sampled in volume and trading time as a function of the sampling period for BNPP.PA/SOGN.PA. The green (resp. blue) line shows the best linear fit with ordinary least squares on all points (resp. on all points except the last one). Left panel: volume time. Right panel: trading time.

Another interesting feature of volume time models is the scaling of the asymptotic variance of returns with the average volume traded in a single transaction. Indeed, we have in the univariate case

$$\frac{R_V}{\sqrt{V}} \xrightarrow{d} \mathcal{N}(0, \frac{\sigma^2}{\mathbb{E}(V_1)}) \text{ as } V \to +\infty$$

For a same aggregated volume V, the variance of $\frac{R_V}{\sqrt{V}}$ decreases with $\mathbb{E}(V_1)$. In order to test this result, we sample returns in volume time and we compute the average volume traded per transaction during the time period it takes to have a volume V traded for each return. We thus obtain a sample (R_i, V_i) where R_i is the i^{th} volume time return sampled and V_i is the corresponding average volume traded in a single transaction. Then we bin average volumes by quantiles and compute the associated covariance matrix of returns. Figure 2.8 plots the covariance of returns as a function of the average volume traded in a single transaction for four pairs of assets and increasing aggregated volumes. The covariance is decreasing on average, which is in agreement with the prediction of volume time models, except for very small average traded volumes. Tiny traded volumes in a single transaction might correspond to periods with very low liquidity at the best limit, which generates trades-through and thus higher amplitude of returns, especially if there are large gaps in the order book, leading to higher covariance.

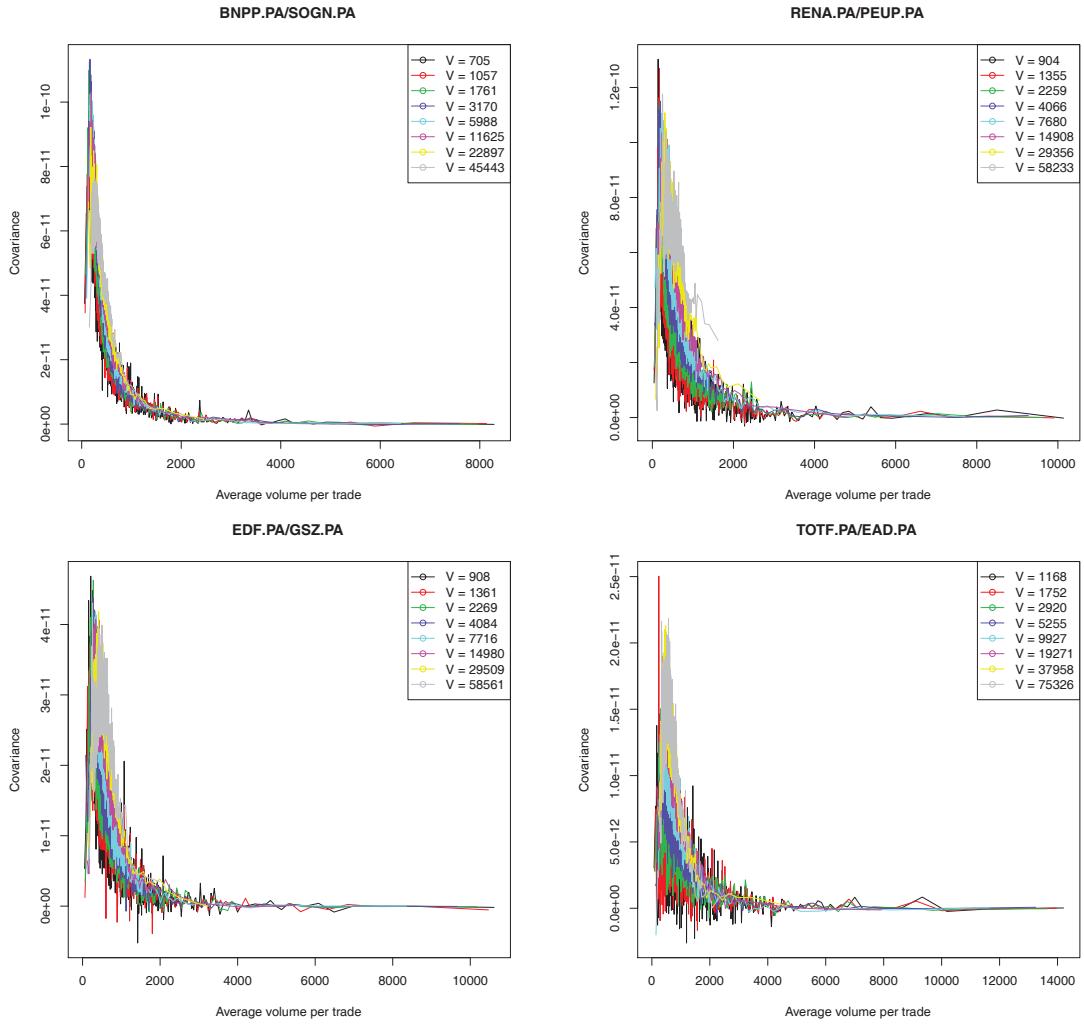

Figure 2.8: Scaling of covariance as a function of the average volume traded in a single transaction. Top left panel: BNPP.PA/SOGN.PA. Top right panel: RENA.PA/PEUP.PA. Bottom left panel: EDF.PA/GSZ.PA. Bottom right panel: TOTF.PA/EAD.PA.

2.4.3 Probability distribution and scaling properties of the event time

Figure 2.9 presents the survival function[11] of the aggregated volume $V_{\Delta t}$ on a semi-log scale for four sampling periods Δt. We consider four candidate theoretical distributions for comparison with the empirical one: Poisson, gamma, inverse-gamma and log-normal. We fit the parameters in order to match the first or first two moments of the empirical distribution. The empirical distribution lies somewhere in between the gamma distribution (exponential tail) and the inverse-gamma or log-normal distribution (heavy tails). Such a distribution for the subordinator respectively results in either an hyperbolic (exponential tails) or power-law (in the tails) distribution for price returns, which is in agreement with the deviation of the empirical distribution from the Gaussian.

[11] The survival function of a random variable X is defined as $S(x) = \mathbb{P}(X > x)$.

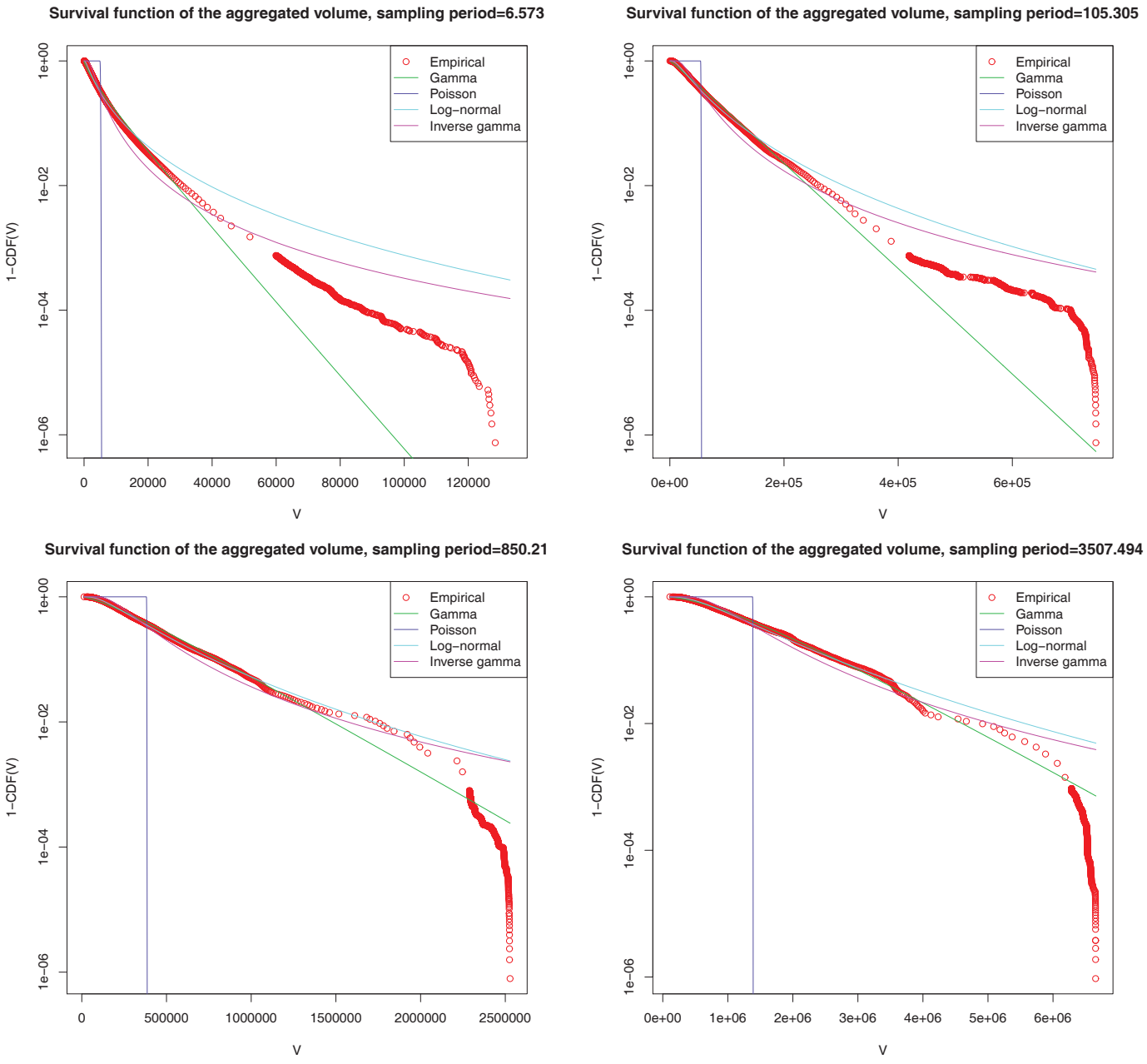

Figure 2.9: Distribution of the aggregated volume $V_{\Delta t}$ for BNPP.PA/SOGN.PA and four sampling periods Δt. Top left panel: $\Delta t = 6.573$ seconds. Top right panel: $\Delta t = 105.305$ seconds. Bottom left panel: $\Delta t = 850.21$ seconds. Bottom right panel: $\Delta t = 3507.494$ seconds.

On figure 2.10, we have plotted the second moment of $V_{\Delta t}$ and $N_{\Delta t}$, which impacts the kurtosis of returns in calendar time, against Δt in log-log coordinates[12]. More precisely, we respectively multiply these quantities by constants σ_V^4 and σ_N^4 to have an estimate for $\mathbb{E}(R_{\Delta t}^4)$ in our model, see section 2.2.1. We also plot the empirical counterpart of $\mathbb{E}(R_{\Delta t}^4)$ computed with calendar time returns for comparison[13].

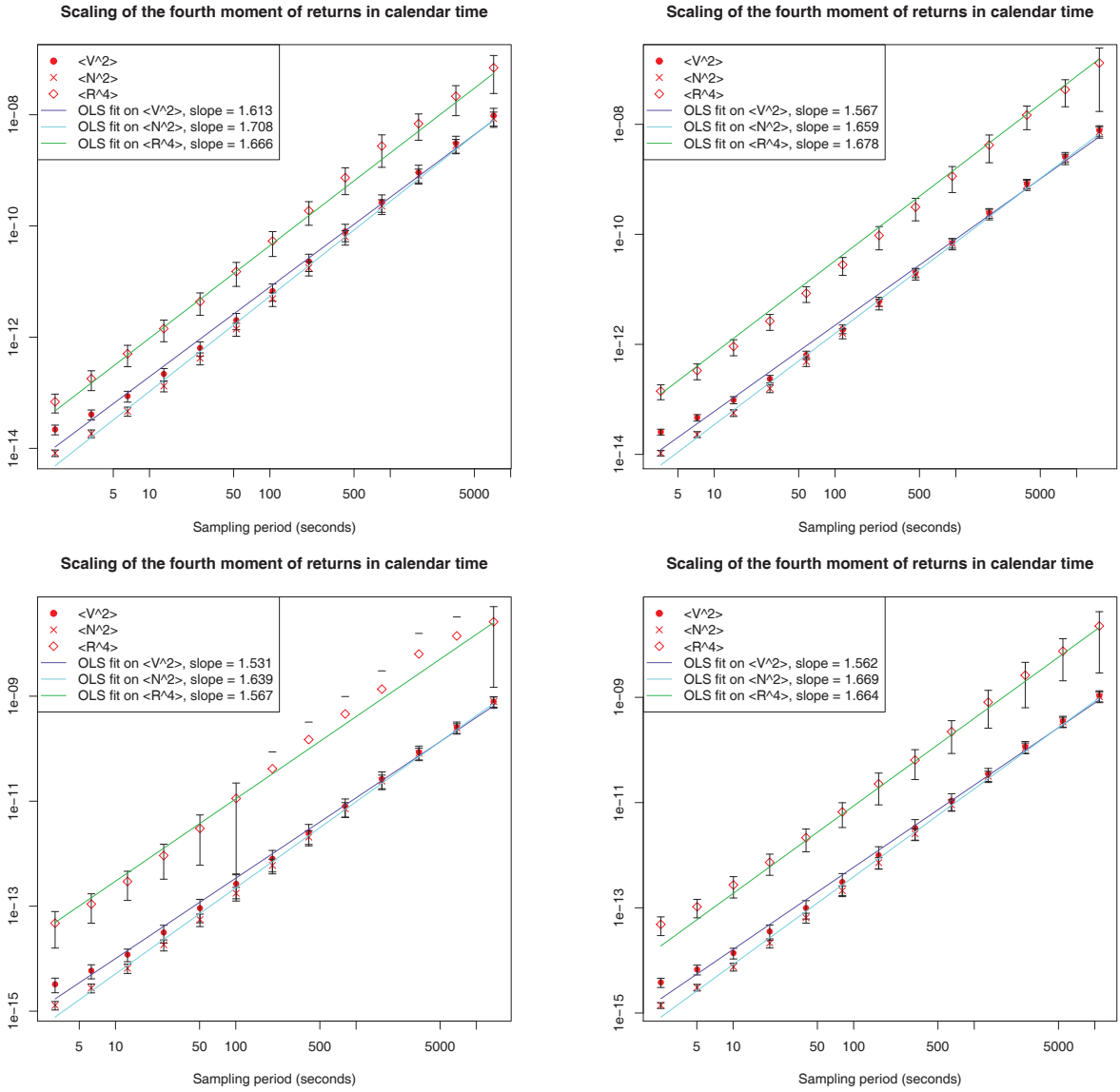

Figure 2.10: Log-log plot of $\mathbb{E}\left(R_{\Delta t}^4\right)$, $\mathbb{E}\left(V_{\Delta t}^2\right)$ and $\mathbb{E}\left(N_{\Delta t}^2\right)$ *versus* Δt. The green line shows the best linear fit with ordinary least squares. Top left panel: BNPP.PA/SOGN.PA. Top right panel: RENA.PA/PEUP.PA. Bottom left panel: EDF.PA/GSZ.PA. Bottom right panel: TOTF.PA/EAD.PA. Some lower error bars are not depicted for EDF.PA/GSZ.PA because they cross zero, so their logarithm is not defined.

[12]Note that the linear scaling of $\mathbb{V}\mathrm{ar}(R_{\Delta t}) \propto \Delta t$ and the equality $\mathbb{E}(R_{\Delta t}^2) = \sigma_X^2 \mathbb{E}(X_{\Delta t})$ imply that $\mathbb{E}(X_{\Delta t}) = \alpha \Delta t$ for some $\alpha > 0$, where X is the event time. Therefore, we know that event time scales linearly with time on average, which is why we directly study second order moments.

[13]Since $R_{\Delta t}$ is a bivariate vector, we define $\mathbb{E}(R_{\Delta t}^4)$ as the average between $\mathbb{E}((R_{\Delta t}^1)^4)$, $\mathbb{E}((R_{\Delta t}^2)^4)$ and $\mathbb{E}((R_{\Delta t}^1 R_{\Delta t}^2)^2)$.

We find that

$$\begin{aligned}
\mathbb{E}\left(V_{\Delta t}^2\right) &\propto \Delta t^{\beta_V} \quad \beta_V \approx 1.57 \\
\mathbb{E}\left(N_{\Delta t}^2\right) &\propto \Delta t^{\beta_N} \quad \beta_N \approx 1.67 \\
\mathbb{E}\left(R_{\Delta t}^4\right) &\propto \Delta t^{\beta_R} \quad \beta_R \approx 1.64
\end{aligned}$$

for the four pairs of stocks under consideration. The sub-quadratic scaling of the second moment implies a slow decay for the kurtosis of the returns in calendar time, *i.e.* a slow convergence of the returns distribution in calendar time towards a Gaussian limit. Actually, we have

$$\frac{\mathbb{E}\left(R_{\Delta t}^4\right)}{\mathbb{E}\left(R_{\Delta t}^2\right)^2} \propto \frac{\mathbb{E}\left(X_{\Delta t}^2\right)}{\mathbb{E}\left(X_{\Delta t}\right)^2} \propto \Delta t^{\beta_X - 2} \text{ for } X \in \{N, V\}$$

with $(\beta_X - 2) \approx -0.4$. It is in agreement with the existing literature [25], as well as with the fact that an i.i.d modelling of returns in calendar time is unappropriate, since it forecasts a scaling of the kurtosis in Δt^{-1}.

Note that the empirical fourth moment of returns $\mathbb{E}(R_{\Delta t}^4)$ is surprisingly close to its model counterpart $\sigma_V^4 \mathbb{E}(V_{\Delta t}^2)$ or $\sigma_N^4 \mathbb{E}(N_{\Delta t}^2)$ both in terms of scaling, *i.e.* β_V and β_N are close to β_R, and amplitude, especially for the traded volume. This comforts us in the subordination hypothesis to explain fat tails in the probability distribution of price returns and in the role of the traded volume to account for stochastic volatility. Going one step further, subordination models allow to infer the hidden stochastic volatility since traded volume is observable.

Finally, it is noteworthy to say that error bars on the estimated "model fourth moments" $\sigma_V^4 \mathbb{E}(V_{\Delta t}^2)$ and $\sigma_N^4 \mathbb{E}(N_{\Delta t}^2)$ are much sharper than for the empirical $\mathbb{E}(R_{\Delta t}^4)$. Indeed, estimates of high order moments are notably noisy [25]. Therefore, replacing fourth order moment $\mathbb{E}(R_{\Delta t}^4)$ estimates by second order such as $\mathbb{E}(V_{\Delta t}^2)$ yields more accurate estimations.

2.4.4 Correlation of returns

We are now interested in the adequacy of the empirical correlation with the implications of our model. We know from subsection 2.2.2 that the correlation is independent of the time scale and the clock in the subordination framework. Figure 2.11 shows how the emprical correlation changes with the time scale according to three different definition of time.

The decline in correlation as we increase the sampling frequency is well documented and is known as the Epps effect [49]. Correlation converges after 15 minutes roughly, though it might be a bit longer depending on the liquidity of the pair of assets, TOTF.PA/EAD.PA being such a case. Thus, a constant correlation is only legitimate in the "long run". The same applies for the equality of correlations computed on either calendar or volume or trading time.

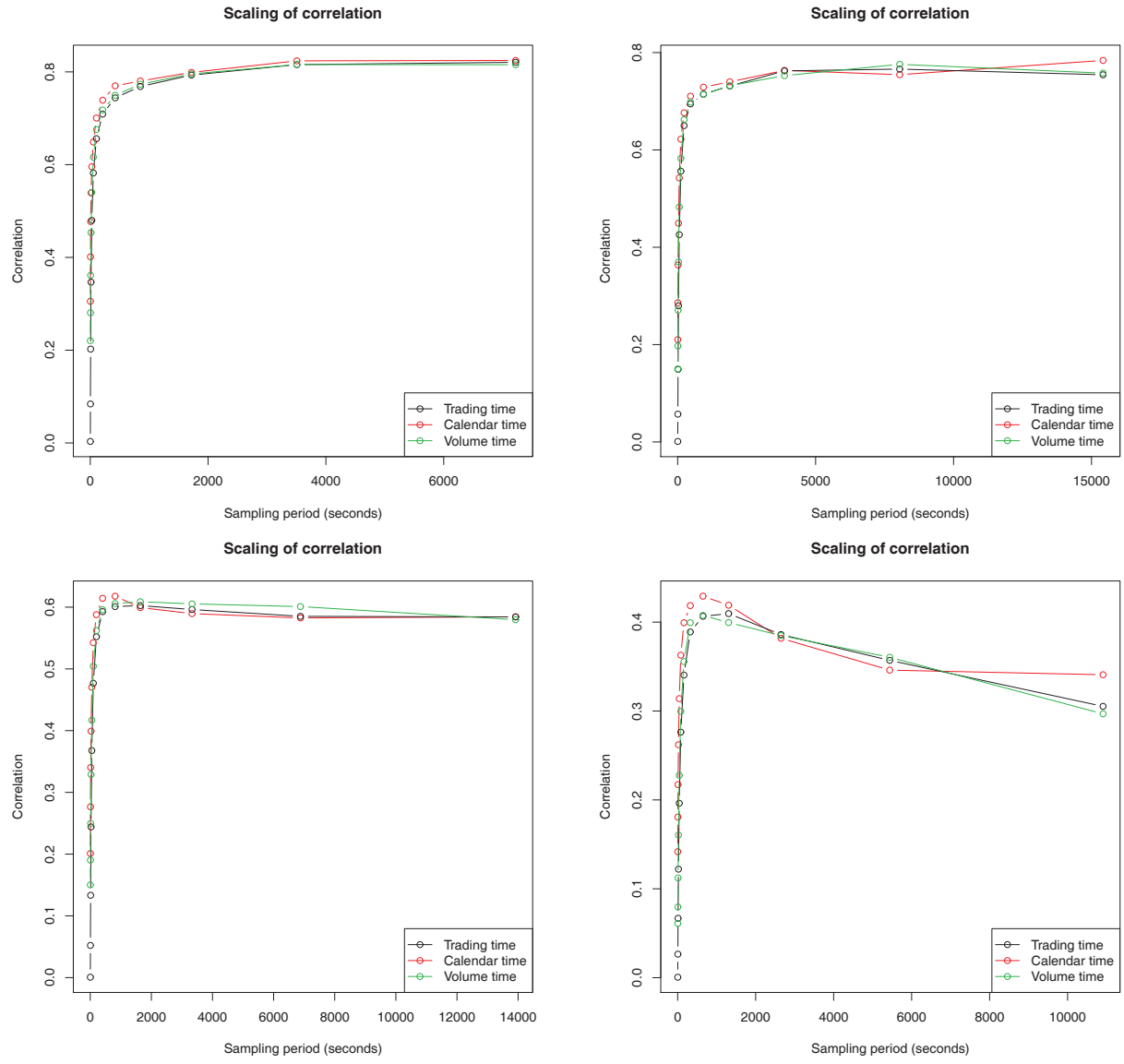

Figure 2.11: Scaling of the correlation of returns sampled in calendar, volume and trading time as a function of the sampling period. Top left panel: BNPP.PA/SOGN.PA. Top right panel: RENA.PA/PEUP.PA. Bottom left panel: EDF.PA/GSZ.PA. Bottom right panel: TOTF.PA/EAD.PA.

2.5 Conclusion and further research

In this paper, we have presented empirical evidence that demonstrates the stochastic behaviour of the covariance matrix in financial markets. A simple mechanism has been described, that accounts for this stochastic behaviour: as in the classical subordination approach of Clark [37], the randomness of the covariance matrix stems from that of the arrival times of market orders and their volume. Moving to continuous-time finance, one can think of returns being driven by a subordinated multivariate Brownian motion. The stochastic clock, *i.e.* the event time, is distributed according to a gamma law for large enough time periods. The resulting

distribution for stocks returns in calendar time is known as the multivariate Variance Gamma model [73], see also [86] for interesting extensions. The price of asset i evolves as follows

$$
\begin{aligned}
&S_t^i = S_0^i.e^{X_t^i}.e^{BS_t^i}.e^{\omega_t^i} \\
&X_t^i = \theta^i G_t + \eta^i W^i\left(G_t\right) \\
&G_t \sim \text{Gamma}\left(\frac{t}{\nu}, \frac{1}{\nu}\right) \\
&BS_t^i = \int_0^t \sigma^i(s)\mathrm{d}W_s^{i,\perp} - \frac{1}{2}\int_0^t \sigma^i(s)^2\mathrm{d}s \\
&\mathrm{d} < W^i, W^j >_t = \rho^{ij}\mathrm{d}t
\end{aligned}
$$

where $W = (W^1, \ldots, W^d)$ and $W^\perp = (W^{1,\perp}, \ldots, W^{d,\perp})$ are independent d-dimensional standard Brownian motions. The common stochastic clock G refers to the multivariate event time. As far as option pricing is concerned, this model has several advantages such as:

- an intuitive and parsimonious parameterization: θ^i provides control over the skewness of $\ln\left(\frac{S_t^i}{S_0^i}\right)$ while ν^i impacts its kurtosis;
- a straightforward marginal calibration: the Variance Gamma belongs to the class of Lévy processes, leading to an analytical formula for the characteristic function of $\ln\left(\frac{S_t^i}{S_0^i}\right)$. Therefore, vanillas can be quickly priced using FFT techniques [30];
- correlation calibration: since the joint characteristic function of returns is available in closed form in this model, the Brownian correlation matrix can be easily calibrated with the historical correlation matrix of returns [73].

But it also exhibits very serious shortcomings from the empirical point of view, which are:

- the lack of correlation between the price and the instantaneous covariance: the only correlation between spot and covariance in this model goes through the parameter θ. This might not be sufficient to capture the strong functional dependency between spot price and variance/covariance, especially during market crashes [24];
- there is no real covariance dynamics. The model cannot explain long range covariance dependency or covariance clustering [78];
- the volatilities of the assets are perfectly correlated due to the common stochastic clock;
- the variance of G_t scales as t, contrary to empirical data scaling as $t^{1.6}$;
- full independence cannot be reached in this model even if Brownian correlations are set to zero;
- the joint probability distribution of returns is elliptic, which is not in agreement with empirical data [34]. This is due to the common random factor to all covariances of price returns. It is however not clear how to relate multi-factor stochastic covariance matrices to event time.

In our opinion, one of the main questions for future work is to understand, from the microstructure point of view, the origins of the spot/covariance correlation as well as covariance clustering and include these stylized facts in the subordination framework. For instance, using an event time that speeds up when prices decrease might reproduce the negative correlation between returns and covariance.

2.6 Appendix: Moments of the Gaussian distribution

Let $X \sim \mathcal{N}(0,1)$. We state that $\mathbb{E}(X^{2m}) = (2m)!!$ and that $\mathbb{E}(X^{2m-1}) = 0$ for $m \in \mathbb{N}^*$. We prove it by forward induction. It is clear that for $m = 1$, $\mathbb{E}(X) = 0$ and $\mathbb{E}\left(X^2\right) = 1 = 2!!$. Then assume it is true for $m-1$. We get

$$\begin{aligned}
\mathbb{E}\left(X^{2m}\right) &= \int_{\mathbb{R}} x^{2m-1}\left(xe^{-x^2/2}\right)\frac{\mathrm{d}x}{\sqrt{2\pi}} \\
&= (2m-1)\int_{\mathbb{R}} x^{2m-2}e^{-x^2/2}\frac{\mathrm{d}x}{\sqrt{2\pi}} \\
&= (2m-1)\mathbb{E}\left(X^{2(m-1)}\right) = (2m-1).(2(m-1))!! = (2m)!!
\end{aligned}$$

The same computations lead to

$$\mathbb{E}\left(X^{2m-1}\right) = (2m-2)\mathbb{E}\left(X^{2(m-1)-1}\right) = 0$$

so that the initial statement is true for any $m \in \mathbb{N}^*$.

2.7 Appendix: Probability distribution of returns in calendar time

Given the probability distribution of the event time over the time period Δt, the probability distribution of returns in calendar time can be computed as

$$\begin{aligned}
P_{R_{\Delta t}}(r) &= \int_0^{+\infty} P_{\mathcal{N}(0,x.\sigma^2)}(r)P_{X_{\Delta t}}(x)\mathrm{d}x \\
&= \int_0^{+\infty} \frac{e^{-|r|(\frac{|r|}{2\sigma^2 x}-\frac{\ln(P_{X_{\Delta t}}(x))}{|r|})}}{\sqrt{2\pi\sigma^2 x}}\mathrm{d}x \\
&= \int_0^{+\infty} e^{-|r|f(x)}g(x)\mathrm{d}x
\end{aligned}$$

where $f(x) = \frac{|r|}{2\sigma^2 x} - \frac{\ln(P_{X_{\Delta t}}(x))}{|r|}$ and $g(x) = (2\pi\sigma^2 x)^{-\frac{1}{2}}$. An approximation of this integral can be provided for $|r| \sim +\infty$ by performing the integration around the minimizer $x^\star = x^\star(r)$ of f. Indeed, the saddle point method [45] states that

$$P_{R_{\Delta t}}(r) \sim e^{-|r|f(x^\star)}g(x^\star)\sqrt{\frac{2\pi}{|r|f''(x^\star)}}$$

$x^\star$ is defined through the following equation

$$-\frac{|r|}{2\left(\sigma x^\star\right)^2} - \frac{P'_{X_{\Delta t}}(x^\star)}{P_{X_{\Delta t}}(x^\star)|r|} = 0$$

Let us consider two cases for $P_{X_{\Delta t}}(x)$: exponential and power-law tails, *i.e.* for $x \sim +\infty$

$$P_{X_{\Delta t}}(x) \propto e^{-\lambda x} \tag{2.1}$$

$$P_{X_{\Delta t}}(x) \propto x^{-(\mu+1)} \tag{2.2}$$

In the first case, $x^\star = b|r|$, while in the second case, $x^\star = cr^2$. This yields, for $|r| \sim +\infty$

$$P_{R_{\Delta t}}(r) \propto e^{-d|r|} \quad \text{for case 1}$$
$$P_{R_{\Delta t}}(r) \propto \frac{1}{|r|^{2\mu+1}} \quad \text{for case 2}$$

So in both cases, the tail behavior of the event time, *i.e.* of stochastic volatility, is propagated over the tails of returns with adjusted shape coefficients, leading to non-Gaussian tails.

2.8 Appendix: Spherical decomposition of Gaussian vectors

Let $X \sim \mathcal{N}_d(0, I)$ be a standard Gaussian random vector and $g : \mathbb{R}^d \mapsto \mathbb{R}$ a measurable function. Then we have

$$\mathbb{E}(g(X)) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}^d} g(x) e^{-\frac{1}{2}x^T x} \mathrm{d}x$$

We switch to hyperspheric variables $r \in \mathbb{R}_+, \theta_1 \in [0, \pi], \theta_2 \in [0, \pi], \ldots, \theta_{d-2} \in [0, \pi], \theta_{d-1} \in [0, 2\pi]$

$$\begin{aligned}
x_1 &= r\sin(\theta_1)\sin(\theta_2)\ldots\sin(\theta_{d-2})\sin(\theta_{d-1}) \\
x_2 &= r\sin(\theta_1)\sin(\theta_2)\ldots\sin(\theta_{d-2})\cos(\theta_{d-1}) \\
x_3 &= r\sin(\theta_1)\sin(\theta_2)\ldots\sin(\theta_{d-3})\cos(\theta_{d-2}) \\
&\vdots \\
x_{d-1} &= r\sin(\theta_1)\cos(\theta_2) \\
x_d &= r\cos(\theta_1)
\end{aligned}$$

or equivalently in vector form $x = ru(\theta)$ with

$$u(\theta) = \begin{pmatrix}
\sin(\theta_1)\sin(\theta_2)\ldots\sin(\theta_{d-2})\sin(\theta_{d-1}) \\
\sin(\theta_1)\sin(\theta_2)\ldots\sin(\theta_{d-2})\cos(\theta_{d-1}) \\
\sin(\theta_1)\sin(\theta_2)\ldots\sin(\theta_{d-3})\cos(\theta_{d-2}) \\
\vdots \\
\sin(\theta_1)\cos(\theta_2) \\
\cos(\theta_1)
\end{pmatrix}$$

This leads to

$$\mathbb{E}(g(X)) = \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}_+} \int_{[0,\pi]^{d-2}\times[0,2\pi]} g(ru(\theta)) e^{-\frac{r^2}{2}} \left|\det(J_d)\right| \mathrm{d}r\mathrm{d}\theta_1\mathrm{d}\theta_2\ldots\mathrm{d}\theta_{d-1}$$

where J_d is the Jacobian matrix

$$
\begin{aligned}
J_d &= \begin{pmatrix}
\frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \theta_1} & \dots & \frac{\partial x_1}{\partial \theta_{d-2}} & \frac{\partial x_1}{\partial \theta_{d-1}} \\
\frac{\partial x_2}{\partial r} & \frac{\partial x_2}{\partial \theta_1} & \dots & \frac{\partial x_2}{\partial \theta_{d-2}} & \frac{\partial x_2}{\partial \theta_{d-1}} \\
\vdots & \vdots & \dots & \vdots & \\
\frac{\partial x_{d-1}}{\partial r} & \frac{\partial x_{d-1}}{\partial \theta_1} & \dots & \frac{\partial x_{d-1}}{\partial \theta_{d-2}} & \frac{\partial x_{d-1}}{\partial \theta_{d-1}} \\
\frac{\partial x_d}{\partial r} & \frac{\partial x_d}{\partial \theta_1} & \dots & \frac{\partial x_d}{\partial \theta_{d-2}} & \frac{\partial x_d}{\partial \theta_{d-1}}
\end{pmatrix} \\
&= \begin{pmatrix}
s_1 s_2 \dots s_{d-1} & r c_1 s_2 \dots s_{d-1} & \dots & r s_1 s_2 \dots s_{d-3} c_{d-2} s_{d-1} & r s_1 s_2 \dots s_{d-2} c_{d-1} \\
s_1 s_2 \dots c_{d-1} & r c_1 s_2 \dots c_{d-1} & \dots & r s_1 s_2 \dots s_{d-3} c_{d-2} c_{d-1} & -r s_1 s_2 \dots s_{d-2} s_{d-1} \\
\vdots & \vdots & \dots & \vdots & \vdots \\
s_1 c_2 & r c_1 c_2 & \dots & 0 & 0 \\
c_1 & -r s_1 c_2 & \dots & 0 & 0
\end{pmatrix}
\end{aligned}
$$

where $c_i = \cos(\theta_i)$ and $s_i = \sin(\theta_i)$. We expand $\det(J_d)$ over the last column, which leads

$$
\begin{aligned}
\det(J_d) &= (-1)^{d+1}(r s_1 \dots s_{d-2} c_{d-1})(c_{d-1} \det(J_{d-1})) + (-1)^{d+2}(-r s_1 \dots s_{d-2} s_{d-1})(s_{d-1} \det(J_{d-1})) \\
&= (-1)^{d+1} r s_1 \dots s_{d-2} \det(J_{d-1}) \\
&= (-1)^{\frac{(n-1)(n+4)}{2}} r^{d-1} \prod_{i=1}^{d-2} s_{d-1-i}^i
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\mathbb{E}(g(X)) &= \frac{1}{(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}_+} \int_{[0,\pi]^{d-2}\times[0,2\pi]} g(ru(\theta)) e^{-\frac{r^2}{2}} r^{d-1} \prod_{i=1}^{d-2} \sin(\theta_{d-1-i})^i \mathrm{d}r \mathrm{d}\theta_1 \mathrm{d}\theta_2 \dots \mathrm{d}\theta_{d-1} \\
&= \frac{1}{2(2\pi)^{\frac{d}{2}}} \int_{\mathbb{R}_+} \int_{[0,\pi]^{d-2}\times[0,2\pi]} g(\sqrt{x}u(\theta)) e^{-\frac{x}{2}} x^{\frac{d}{2}-1} \prod_{i=1}^{d-2} \sin(\theta_{d-1-i})^i \mathrm{d}x \mathrm{d}\theta_1 \mathrm{d}\theta_2 \dots \mathrm{d}\theta_{d-1} \\
&= \frac{\Gamma(\frac{d}{2})}{2\pi^{\frac{d}{2}}} \int_{\mathbb{R}_+} \mathrm{d}x \frac{e^{-\frac{x}{2}} x^{\frac{d}{2}-1}}{\Gamma(\frac{d}{2}) 2^{\frac{d}{2}}} \int_0^\pi \mathrm{d}\theta_1 \sin(\theta_1)^{d-2} \int_0^\pi \mathrm{d}\theta_2 \sin(\theta_2)^{d-3} \dots \int_0^\pi \mathrm{d}\theta_{d-2} \sin(\theta_{d-2}) \int_0^{2\pi} \mathrm{d}\theta_{d-1} g(\sqrt{x}u(\theta)) \\
&= \int_{\mathbb{R}_+} \mathrm{d}x \frac{e^{-\frac{x}{2}} x^{\frac{d}{2}-1}}{\Gamma(\frac{d}{2}) 2^{\frac{d}{2}}} \int_0^\pi \mathrm{d}\theta_1 \frac{\sin(\theta_1)^{d-2}}{2W_{d-2}} \int_0^\pi \mathrm{d}\theta_2 \frac{\sin(\theta_2)^{d-3}}{2W_{d-3}} \dots \int_0^\pi \mathrm{d}\theta_{d-2} \frac{\sin(\theta_{d-2})}{2W_1} \int_0^{2\pi} \mathrm{d}\theta_{d-1} \frac{1}{2\pi} g(\sqrt{x}u(\theta)) \\
&= \int_{\mathbb{R}_+} \int_{[0,\pi]^{d-2}\times[0,2\pi]} g(ru(\theta)) p_{R^2,\theta}(x,\theta) \mathrm{d}x \mathrm{d}\theta_1 \mathrm{d}\theta_2 \dots \mathrm{d}\theta_{d-1} \\
&= \int_{\mathbb{R}_+} \int_{[0,\pi]^{d-2}\times[0,2\pi]} g(ru(\theta)) p_{R^2}(x) p_{\theta_1}(\theta_1) p_{\theta_2}(\theta_2) \dots p_{\theta_{d-1}}(\theta_{d-1}) \mathrm{d}x \mathrm{d}\theta_1 \mathrm{d}\theta_2 \dots \mathrm{d}\theta_{d-1} \\
&= \mathbb{E}(g(RU(\theta)))
\end{aligned}
$$

where

- $p_{R^2}(x) = \frac{e^{-\frac{x}{2}} x^{\frac{d}{2}-1}}{\Gamma(\frac{d}{2}) 2^{\frac{d}{2}}} \mathbb{1}_{\mathbb{R}_+}(x)$ is the chi-square distribution with d degrees of freedom
- $p_{\theta_i}(\theta) = \frac{\sin(\theta)^{d-1-i}}{2W_{d-1-i}} \mathbb{1}_{[0,\pi]}(\theta)$ for $i < d-1$ and $p_{\theta_{d-1}}(\theta) = \frac{1}{2\pi} \mathbb{1}_{[0,2\pi]}(\theta)$

The second equality comes from the change of variable $x = r^2$. The fourth equality comes from the fact that $\frac{\Gamma(\frac{d}{2})}{2\pi^{\frac{d}{2}}} = \frac{1}{2\pi}\prod_{i=1}^{d-2}(2W_i)^{-1}$, where $W_i = \int_0^{\frac{\pi}{2}} \sin(\theta)^i \mathrm{d}\theta = \frac{\sqrt{\pi}\Gamma(\frac{i+1}{2})}{2\Gamma(\frac{i+2}{2})}$ is the i^{th} term of the Wallis' integrals sequence. This ends the proof.

Chapter 3

High Frequency Lead/lag Relationships

3.1 Introduction

The standard financial theory assumes that there is no arbitrage on financial markets[1] [47]. In particular, it does not allow for predictability of asset returns. As a result, lead/lag relationships (assets driving others in advance) should not exist according to this theory. Figure 3.1 plots the cross-correlation function between the daily returns of the French equity index CAC40 (.FCHI)[2] and those of the French stock Renault (RENA.PA) which is part of the CAC40, between 2003/01/02 and 2011/03/04[3]. The cross-correlation function is close to a Dirac delta function (times the average daily correlation). Thus, the absence of lead/lag relationships on a daily time scale seems to be quite reasonable.

The availability of high frequency financial data allows us to zoom on microscopic fluctuations of the order flow. In this paper, we study the existence of lead/lag relationships between assets on small time scales. Lead/lag relationships are measured with the Hayashi-Yoshida cross-correlation estimator [61, 63]. This estimator deals with the issue of asynchronous trading and makes use of all the available tick-by-tick data, so that we can theoretically measure lags down to the finest time scale. We report evidence of highly asymmetric cross-correlation functions as a witness of lead/lag relationships. These are not statistical artefacts due to differences in levels of trading activity. We provide a descriptive picture of the microstructural factors that discriminate leaders from laggers. We find an intraday profile of lead/lag that reacts to market news and openings. We also study how this lead/lag phenomenon evolves when we focus on extreme events. We backtest forecasting devices using these lead/lag relationships and find that they are statistically successful, with an average accuracy of about 60% for predicting variations of the midquote. These lead/lag relationships tend to disappear as we move to larger time scales.

The paper is organized as follows. Section 3.2 introduces the dataset and provides basic but insightful statistics on the assets under focus. Section 3.3 describes the methodology used to measure lead/lag relationships. Section 3.4 presents our empirical results. Finally, section 3.5 concludes by summarizing the main findings and giving the directions for further research.

[1] An arbitrage opportunity occurs when one can set up a zero-cost portfolio allowing for strictly positive wealth in the future with non-zero probability.

[2] We indicate the Reuters Instrument Code (RIC) of each financial asset in brackets.

[3] The data used for figure 3.1 are adjusted closing prices and can be downloaded for free at fr.finance.yahoo.com.

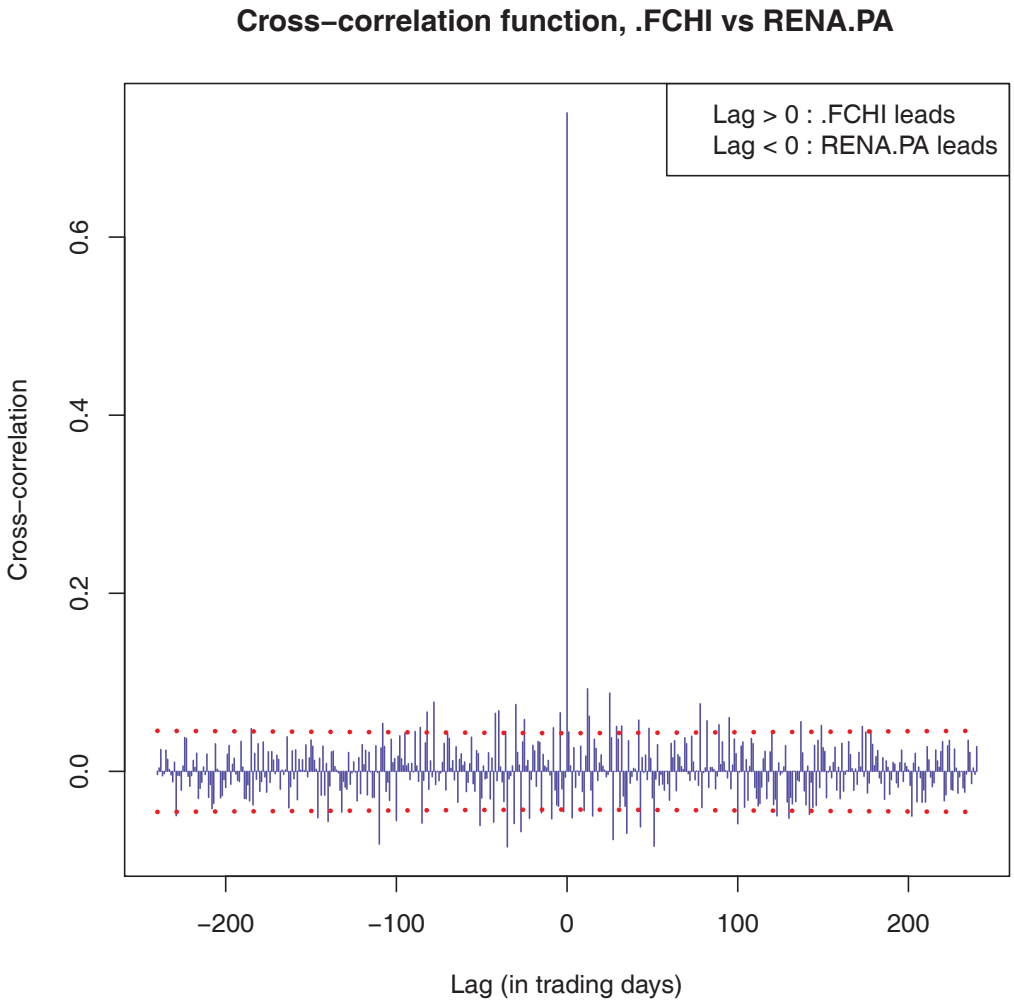

Figure 3.1: Cross-correlation function between .FCHI and RENA.PA, 2003/01/02 – 2011/03/04

3.2 Data description and summary statistics

We have access to the Thomson Reuters Tick History database (see section 1.3) which provides tick-by-tick data on many financial assets (equities, fixed income, forex, futures, commodities, *etc*). Three levels of data are available:

- trades files: each transaction price and quantity timestamped up to the millisecond
- quotes files: each quote (best bid and ask) price or quantity modification timestamped up to the millisecond
- order book files: each limit price or quantity modification timestamped up to the millisecond, up to a given depth, typically ten limits on each side of the order book.

Throughout this study, we will only use trades and quotes files. We will sample quotes on a trading time basis so that for each trade we have access to the quotes right before this trade. We do this because the returns we will consider are the returns of the midquote in order to get rid of the bid/ask bounce. Therefore, it might sound more natural to consider every best quote modification since events other than trades, such as limit orders placed in the bid/ask spread, cancellation orders for the whole quantity available at the best limit and trades-though, can affect the midquote. However, we favor trading time sampling over best quotes sampling because in our opinion trades represent more significant events than quotes changes. Indeed, only trades involve money exchanges.

When a trade walks the order book up or down by hitting consecutive limit prices, it is recorded as a sequence of trades with the same timestamp but with prices and quantities corresponding to each limit hit. For instance, assume that the best ask offers 100 shares at price 10 and 200 shares at price 10.01, and that a buy trade arrives for 150 shares. This is recorded as two lines in the trades file with the same timestamp, the first line being 100 shares at 10 and the second line 50 shares at 10.01. As a pre-processing step, we aggregate identical timestamps in trades files by replacing the price by the volume weighted average price (VWAP) over the whole transaction and the volume by the sum of all quantities consumed. In the previous example, the trade price will thus be $(100*10+50*10.01)/(100+50) = 10.00333$ and the trade quantity $100+50 = 150$.

Table 3.1 describes the scope of assets for our empirical study. We only consider equities and equity index futures. The futures are nearby-maturity futures and are rolled the day before the expiration date. The time period is 2010/03/01-2010/05/31 and the trading hours are specified in table 3.1. On each day, we drop the first and last half hours of trading. We only consider regular trades to avoid outliers such as block trades or OTC trades (see [6] for a more detailed description). When studying lead/lag relationships between assets traded on different exchanges, we only consider hours of simultaneous trading.

Table 3.2 gives some insight into the liquidity of each of these assets. It displays the following summary statistics[4]:

- the average duration between two consecutive trades $\langle \Delta t \rangle$
- the average tick size δ in percentage of the midquote $\langle \delta / m \rangle$
- the average bid/ask spread expressed in tick size $\langle s \rangle / \delta$
- the frequency of unit bid/ask spread $\left\langle \mathbb{1}_{\{s=\delta\}} \right\rangle$
- the frequency of trades hitting more than the best limit price available[5] $\left\langle \mathbb{1}_{\{\text{trade through}\}} \right\rangle$ [6]
- a proxy for the volatility expressed in tick size : $\langle |\Delta m| \rangle / \delta$, where Δm is the midquote variation between two consecutive trades
- the average turnover per trade $\langle P_{\text{trade}} V_{\text{trade}} \rangle$

Every average is computed independently on a daily basis, and then averaged over all days available : $\langle x \rangle = \frac{1}{n_{\text{days}}} \sum_{d=1}^{n_{\text{days}}} \frac{\sum_{i=1}^{n_d} x_{i,d}}{n_d}$, where n_d is the number of observations on day d and $x_{i,d}$ is the i^{th} observation on day d.

[4]For assets traded in an other currency than EUR (that is to say VOD.L, FFI, NESN.VX and FSMI), we convert the average turnover per trade in EUR by using the closing price of the corresponding exchange rate (GBP/EUR for the first two and CHF/EUR for the last two).

[5]In our data, we detect a so-called trade-through as a sequence of trades with the same timestamp and at least two different consecutive execution prices.

Table 3.1: Description of the scope of assets.

RIC	Description	Exchange	Trading hours (CET)	Currency
ACCP.PA	Accor	NYSE Euronext Paris	09:00-17:30	EUR
AIRP.PA	Air Liquide	NYSE Euronext Paris	09:00-17:30	EUR
ALSO.PA	Alstom	NYSE Euronext Paris	09:00-17:30	EUR
ALUA.PA	Alcatel Lucent	NYSE Euronext Paris	09:00-17:30	EUR
AXAF.PA	Axa	NYSE Euronext Paris	09:00-17:30	EUR
BNPP.PA	BNP Paribas	NYSE Euronext Paris	09:00-17:30	EUR
BOUY.PA	Bouygues	NYSE Euronext Paris	09:00-17:30	EUR
CAGR.PA	Crédit Agricole	NYSE Euronext Paris	09:00-17:30	EUR
CAPP.PA	Cap Gemini	NYSE Euronext Paris	09:00-17:30	EUR
CARR.PA	Carrefour	NYSE Euronext Paris	09:00-17:30	EUR
DANO.PA	Danone	NYSE Euronext Paris	09:00-17:30	EUR
DEXI.BR	Dexia	NYSE Euronext Brussels	09:00-17:30	EUR
EAD.PA	EADS	NYSE Euronext Paris	09:00-17:30	EUR
EDF.PA	EDF	NYSE Euronext Paris	09:00-17:30	EUR
ESSI.PA	Essilor	NYSE Euronext Paris	09:00-17:30	EUR
FTE.PA	France Télécom	NYSE Euronext Paris	09:00-17:30	EUR
GSZ.PA	GDF Suez	NYSE Euronext Paris	09:00-17:30	EUR
ISPA.AS	Arcelor Mittal	NYSE Euronext Amsterdam	09:00-17:30	EUR
LAFP.PA	Lafarge	NYSE Euronext Paris	09:00-17:30	EUR
LAGA.PA	Lagardère	NYSE Euronext Paris	09:00-17:30	EUR
LVMH.PA	LVMH	NYSE Euronext Paris	09:00-17:30	EUR
MICP.PA	Michelin	NYSE Euronext Paris	09:00-17:30	EUR
OREP.PA	L'Oréal	NYSE Euronext Paris	09:00-17:30	EUR
PERP.PA	Pernod Ricard	NYSE Euronext Paris	09:00-17:30	EUR
PEUP.PA	Peugeot	NYSE Euronext Paris	09:00-17:30	EUR
PRTP.PA	PPR	NYSE Euronext Paris	09:00-17:30	EUR
RENA.PA	Renault	NYSE Euronext Paris	09:00-17:30	EUR
SASY.PA	Sanofi Aventis	NYSE Euronext Paris	09:00-17:30	EUR
SCHN.PA	Schneider Electric	NYSE Euronext Paris	09:00-17:30	EUR
SEVI.PA	Suez Environnement	NYSE Euronext Paris	09:00-17:30	EUR
SGEF.PA	Vinci	NYSE Euronext Paris	09:00-17:30	EUR
SGOB.PA	Saint-Gobain	NYSE Euronext Paris	09:00-17:30	EUR
SOGN.PA	Société Générale	NYSE Euronext Paris	09:00-17:30	EUR
STM.PA	StMicroelectronics	NYSE Euronext Paris	09:00-17:30	EUR
TECF.PA	Technip	NYSE Euronext Paris	09:00-17:30	EUR
TOTF.PA	Total	NYSE Euronext Paris	09:00-17:30	EUR
UNBP.PA	Unibail-Rodamco	NYSE Euronext Paris	09:00-17:30	EUR
VIE.PA	Veolia Environnement	NYSE Euronext Paris	09:00-17:30	EUR
VIV.PA	Vivendi	NYSE Euronext Paris	09:00-17:30	EUR
VLLP.PA	Vallourec	NYSE Euronext Paris	09:00-17:30	EUR
VOD.L	Vodafone	London Stock Exchange	09:00-17:30	GBP
NESN.VX	Nestlé	SIX Swiss Exchange	09:00-17:30	CHF
DTEGn.DE	Deutsche Telekom	XETRA	09:00-17:30	EUR
FCE	CAC40 future	NYSE Liffe Paris	08:00-22:00	EUR
FFI	Footsie100 future	NYSE Liffe London	02:00-08:50, 09:00-22:00	GBP
FSMI	SMI future	Eurex	07:50-22:00	CHF
FDX	DAX future	Eurex	07:50-22:00	EUR

Table 3.2: Summary statistics on the scope of assets.

RIC	$\langle \Delta t \rangle$ (sec)	$\langle \delta/m \rangle$ (bp)	$\langle s \rangle / \delta$	$\langle \mathbb{1}_{\{s=\delta\}} \rangle$ (%)	$\langle \mathbb{1}_{\{\text{trade-through}\}} \rangle$ (%)	$\langle \lvert \Delta m \rvert \rangle / \delta$	$\langle P_{\text{trade}} V_{\text{trade}} \rangle$ (EUR$\times 10^3$)
ACCP.PA	13.350	1.22	3.98	16	5	1.18	11
AIRP.PA	7.328	1.15	2.76	16	5	0.88	13
ALSO.PA	6.620	1.11	3.48	19	6	1.01	13
ALUA.PA	8.276	4.34	1.55	58	4	0.37	12
AXAF.PA	5.580	3.30	1.37	69	3	0.38	16
BNPP.PA	3.317	1.58	2.13	47	6	0.65	18
BOUY.PA	9.585	1.36	2.77	27	5	0.99	12
CAGR.PA	5.911	3.35	2.34	59	4	0.70	13
CAPP.PA	10.390	1.35	3.23	21	5	1.01	12
CARR.PA	7.668	1.39	2.31	34	4	0.76	15
DANO.PA	6.158	1.15	2.35	35	5	0.70	14
DEXI.BR	22.106	2.44	5.3	8	8	1.35	6
EAD.PA	12.153	3.34	1.73	50	3	0.44	12
EDF.PA	8.026	1.29	2.61	30	4	0.78	12
ESSI.PA	14.452	1.08	2.83	26	4	0.76	10
FTE.PA	6.581	2.97	1.18	83	2	0.23	20
GSZ.PA	5.427	1.85	1.76	50	4	0.47	15
ISPA.AS	3.013	1.68	2.02	39	6	0.62	22
LAFP.PA	9.037	1.60	3.17	26	5	0.98	14
LAGA.PA	17.009	1.74	2.95	22	4	0.86	8
LVMH.PA	6.102	1.16	2.64	19	6	0.88	15
MICP.PA	9.106	1.84	2.58	28	4	0.76	12
OREP.PA	10.051	1.28	2.66	19	5	0.84	16
PERP.PA	12.214	1.61	2.18	35	3	0.65	12
PEUP.PA	10.112	2.36	2.81	22	5	0.73	12
PRTP.PA	13.259	2.39	3.49	33	4	0.92	17
RENA.PA	5.795	1.51	3.16	21	6	0.99	13
SASY.PA	5.270	1.71	1.96	47	4	0.49	20
SCHN.PA	6.709	1.19	2.95	15	5	0.94	15
SEVI.PA	21.398	3.10	1.82	45	3	0.48	8
SGEF.PA	5.582	1.22	2.64	28	5	0.81	13
SGOB.PA	6.432	1.41	2.85	23	5	0.92	13
SOGN.PA	3.351	1.20	2.91	28	7	0.95	15
STM.PA	13.462	1.44	3.65	10	6	1.10	10
TECF.PA	12.496	1.70	4.17	12	5	0.99	12
TOTF.PA	3.283	1.21	1.88	48	5	0.57	21
UNBP.PA	14.968	3.53	1.39	68	2	0.35	20
VIE.PA	9.905	2.10	1.98	43	3	0.52	13
VIV.PA	7.864	2.63	1.36	70	3	0.35	17
VLLP.PA	11.355	3.36	1.68	52	3	0.47	17
VOD.L	6.766	3.47	1.13	88	2	0.26	21
NESN.VX	13.056	9.42	1.01	99	1	0.05	71
DTEGn.DE	7.319	1.56	2.58	19	7	0.70	29
FCE0	1.414	1.32	1.14	87	5	0.37	150
FFI0	1.372	0.91	1.19	82	4	0.42	217
FSMI0	5.265	1.50	1.15	86	3	0.38	178
FDX0	1.215	0.83	1.28	74	11	0.38	593

3.3 Methodology

3.3.1 The Hayashi-Yoshida cross-correlation function

In [61], the authors introduce a new estimator of the linear correlation coefficient between two asynchronous diffusive processes[6]. Given two Itô processes X, Y such that

$$\begin{aligned}
\mathrm{d}X_t &= \mu_t^X \mathrm{d}t + \sigma_t^X \mathrm{d}W_t^X \\
\mathrm{d}Y_t &= \mu_t^Y \mathrm{d}t + \sigma_t^Y \mathrm{d}W_t^Y \\
\mathrm{d}\left\langle W^X, W^Y \right\rangle_t &= \rho_t \mathrm{d}t
\end{aligned}$$

and observation times $0 = t_0 \leq t_1 \leq \ldots \leq t_{n-1} \leq t_n = T$ for X and $0 = s_0 \leq s_1 \leq \ldots \leq s_{m-1} \leq s_m = T$ for Y, which must be independent from X and Y, they show that the following quantity

$$\begin{aligned}
&\sum_{i,j} r_i^X r_j^Y \mathbb{1}_{\{O_{ij} \neq \emptyset\}} \\
&O_{ij} =]t_{i-1}, t_i] \cap]s_{j-1}, s_j] \\
&r_i^X = X_{t_i} - X_{t_{i-1}} \\
&r_j^Y = Y_{s_j} - Y_{s_{j-1}}
\end{aligned}$$

is an unbiased and consistent estimator of $\int_0^T \sigma_t^X \sigma_t^Y \rho_t \mathrm{d}t$ as the largest mesh size goes to zero, as opposed to the standard previous-tick covariance estimator [56, 100]. In practice, it amounts to sum every product of increments as soon as they share an overlap of time. In the case of constant volatilities and correlation, it provides a consistent estimator for the correlation

$$\hat{\rho} = \frac{\sum_{i,j} r_i^X r_j^Y \mathbb{1}_{\{O_{ij} \neq \emptyset\}}}{\sqrt{\sum_i (r_i^X)^2 \sum_j (r_j^Y)^2}}$$

Recently, in [63], the authors generalize this estimator to the whole cross-correlation function. They use a lagged version of the original estimator

$$\begin{aligned}
\hat{\rho}(\ell) &= \frac{\sum_{i,j} r_i^X r_j^Y \mathbb{1}_{\{O_{ij}^\ell \neq \emptyset\}}}{\sqrt{\sum_i (r_i^X)^2 \sum_j (r_j^Y)^2}} \\
O_{ij}^\ell &=]t_{i-1}, t_i] \cap]s_{j-1} - \ell, s_j - \ell]
\end{aligned}$$

It can be computed by shifting all the timestamps of Y and then using the Hayashi-Yoshida estimator. They define the lead/lag time as the lag that maximizes $|\hat{\rho}(\ell)|$. In the following we will not estimate the lead/lag time but rather decide if one asset leads the other by measuring the asymmetry of the cross-correlation function between positive and negative lags. More precisely, we state that X leads Y if X forecasts Y more accurately than Y does for X. Formally speaking, X leads Y if

$$\begin{aligned}
&\frac{\left\| r_t^Y - \mathrm{Proj}(r_t^Y | \vec{r}_{t^-}^X) \right\|}{\|r^Y\|} < \frac{\left\| r_t^X - \mathrm{Proj}(r_t^X | \vec{r}_{t^-}^Y) \right\|}{\|r^X\|} \\
&\iff \frac{\left\| \varepsilon^{YX} \right\|}{\|r^Y\|} < \frac{\left\| \varepsilon^{XY} \right\|}{\|r^X\|}
\end{aligned}$$

[6] In fact, a very similar estimator was already designed in [43].

where $\text{Proj}(r_t^Y|\vec{r}_{t^-}^X)$ denotes the projection of r_t^Y on the space spanned by $\vec{r}_{t^-}^X := \left\{r_s^X, s < t\right\}$. We will only consider the ordinary least squares setting, *i.e.* $\text{Proj}(r_t^Y|\vec{r}_{t^-}^X) = \mu + \int_{]0,\bar{\ell}]} \beta_s r_{t-s}^X \mathrm{d}s$ and $\|X\|^2 = \mathbb{V}\text{ar}(X)$. In practice, we compute the cross-correlation function on a discrete grid of lags so that $\int_{]0,\bar{\ell}]} \beta_s r_{t-s}^X \mathrm{d}s = \sum_{i=1}^p \tilde{\beta}_i r_{t-\ell_i}^X$, where $\tilde{\beta}_i = \beta_i(\ell_i - \ell_{i-1})$. It is easy to show (see appendix 3.6) that

$$\begin{aligned}
\frac{\left\|\varepsilon^{YX}\right\|^2}{\left\|r^Y\right\|^2} &= 1 - (C^{YX})^T (C^{XX} C^{YY})^{-1} C^{YX} \\
C^{YX} &= (\mathbb{C}\text{ov}(r_t^Y, r_{t-\ell_1}^X), \ldots, \mathbb{C}\text{ov}(r_t^Y, r_{t-\ell_p}^X))^T \\
C^{YY} &= \mathbb{V}\text{ar}(r_t^Y) \\
C^{XX} &= (\mathbb{C}\text{ov}(r_{t-\ell_i}^X, r_{t-\ell_j}^X), i, j = 1, \ldots, p)
\end{aligned}$$

$(C^{YX})^T (C^{XX} C^{YY})^{-1} C^{YX}$ measures the squared correlation between r^Y and r^X. Indeed, r^X is a good predictor of r^Y if both are highly correlated. If we assume furthermore that the predictors r^X are uncorrelated, we can prove (see appendix 3.6) that

$$\begin{aligned}
&\frac{\left\|\varepsilon^{YX}\right\|}{\left\|r^Y\right\|} < \frac{\left\|\varepsilon^{XY}\right\|}{\left\|r^X\right\|} \\
\iff &\sum_{i=1}^p \rho^2(\ell_i) > \sum_{i=1}^p \rho^2(-\ell_i) \\
\iff &\text{LLR} := \frac{\sum_{i=1}^p \rho^2(\ell_i)}{\sum_{i=1}^p \rho^2(-\ell_i)} > 1
\end{aligned}$$

The asymmetry of the cross-correlation function, as defined by the LLR (standing for Lead/Lag Ratio) measures lead/lag relationships. This definition of lead/lag is closely related to the notion of Granger causality [55]. Given two stochastic processes X and Y, X is said to cause Y (in the Granger sense) if, in the following linear regression

$$Y_t = c + \sum_{k=1}^p a_k^{YX} . Y_{t-k} + \sum_{\ell=1}^q b_\ell^{YX} . X_{t-\ell} + \varepsilon_t^{YX}$$

some of the estimated coefficients $\hat{b}_\ell^{YX}$ are found to be statistically significant. Since these coefficients are closely linked to the cross-correlation function, there is indeed a strong similarity between this approach and ours. The Granger regression includes lags of the lagger in order to control for its autocorrelation, which we do not take into account. Note that there is *a priori* no obstacle to find that X Granger-causes Y and Y Granger-causes X. Our approach amounts to compare in some sense the signficance of all $\hat{b}_\ell^{YX}$ to the one of all $\hat{b}_\ell^{XY}$.

Our indicator tells us which asset is leading the other for a given pair, but we might also wish to consider the strength and the characteristic time of this lead/lag relationship. Therefore, the maximum level of the cross-correlation function and the lag at which it occurs must also be taken into account.

In the following empirical study, we measure the cross-correlation function between variations of midquotes of two assets, *i.e.* X and Y are midquotes. The observation times will be tick times. Tick time is defined as the clock that increments as soon as there is a non-zero variation of the midquote between two (not necessarily consecutive) trades. The resulting set of timestamps does not take into account the nil variations of the midquote, contrary to trading time. Computing the Hayashi-Yoshida correlation in trade time or in

tick time does not yield the same result. Indeed, consider the trading sequence shown on figure 3.2. It is easily seen that the trade time covariance is zero while the tick time covariance is not. We will be interested in forecasting the midquote variation of the lagging asset, so we prefer to use tick time rather than trade time. Indeed, classification in tick time is binary: either the midquote moves up or it moves down.

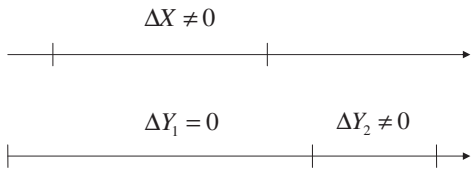

Figure 3.2: Difference between the trade time and tick time Hayashi-Yoshida correlations

3.3.2 Simulation study: artificial lead/lag due to different levels of trading activity

When looking at two assets, a natural bet is to say that the most traded asset leads the other. In the literature, empirical research focusing on lead/lag relationships [43, 44, 64, 66, 67, 69, 70, 75, 82] often concludes that the most liquid assets drive the others[7]. Intuitively, the most heavily traded assets tend to incorporate information into prices faster than others so they should lead.

A very simple simulation framework [56, 61] provides some insight into this liquidity lead/lag effect. Let us consider two correlated Brownian motions B_1, B_2 with correlation ρ and two series of random timestamps $0 = t_0 < t_1 < \ldots < t_n = T$ and $0 = s_0 < s_1 < \ldots < s_m = T$ independent of B_1 and B_2. For instance, let the timestamps be the jumping times of two independent Poisson processes with respective intensities λ_1 and λ_2. We define two time series of price as the Brownian motions sampled along the Poisson timestamps, that is

$$\begin{aligned} X(u) &= B_1(t(u)) \\ Y(u) &= B_2(s(u)) \\ t(u) &= \max\{t_i | t_i \leq u\} \\ s(u) &= \max\{s_i | s_i \leq u\} \end{aligned}$$

There shouldn't be any lead/lag relationship between X and Y since they are sampled from two synchronous Brownian motions. The cross-correlation function should thus be a Dirac delta function with level ρ at lag zero. Figure 3.3 illustrates the behaviour of the cross-correlation function computed with either the previous-tick or the Hayashi-Yoshida estimator for various levels of $\frac{\lambda_1}{\lambda_2}$. We simulate two synchronously correlated Brownian motions on $[0, T = 30600]$ with time step $\Delta t = 5$ and correlation $\rho = 0.8$. Then we sample them along two independent Poisson time grids with parameters λ_1 and λ_2. We repeat this simulation 64 times and average the cross-correlation functions computed independently over each simulation. In the Hayashi-Yoshida case, we also plot on figure 3.3 the average cross-correlation function computed using

[7] Liquidity does not necessarily mean more transactions, it can be measured with other microstructure statistics such as bid/ask spread, market impact *etc...*

the closed-form formula shown in appendix 3.7.

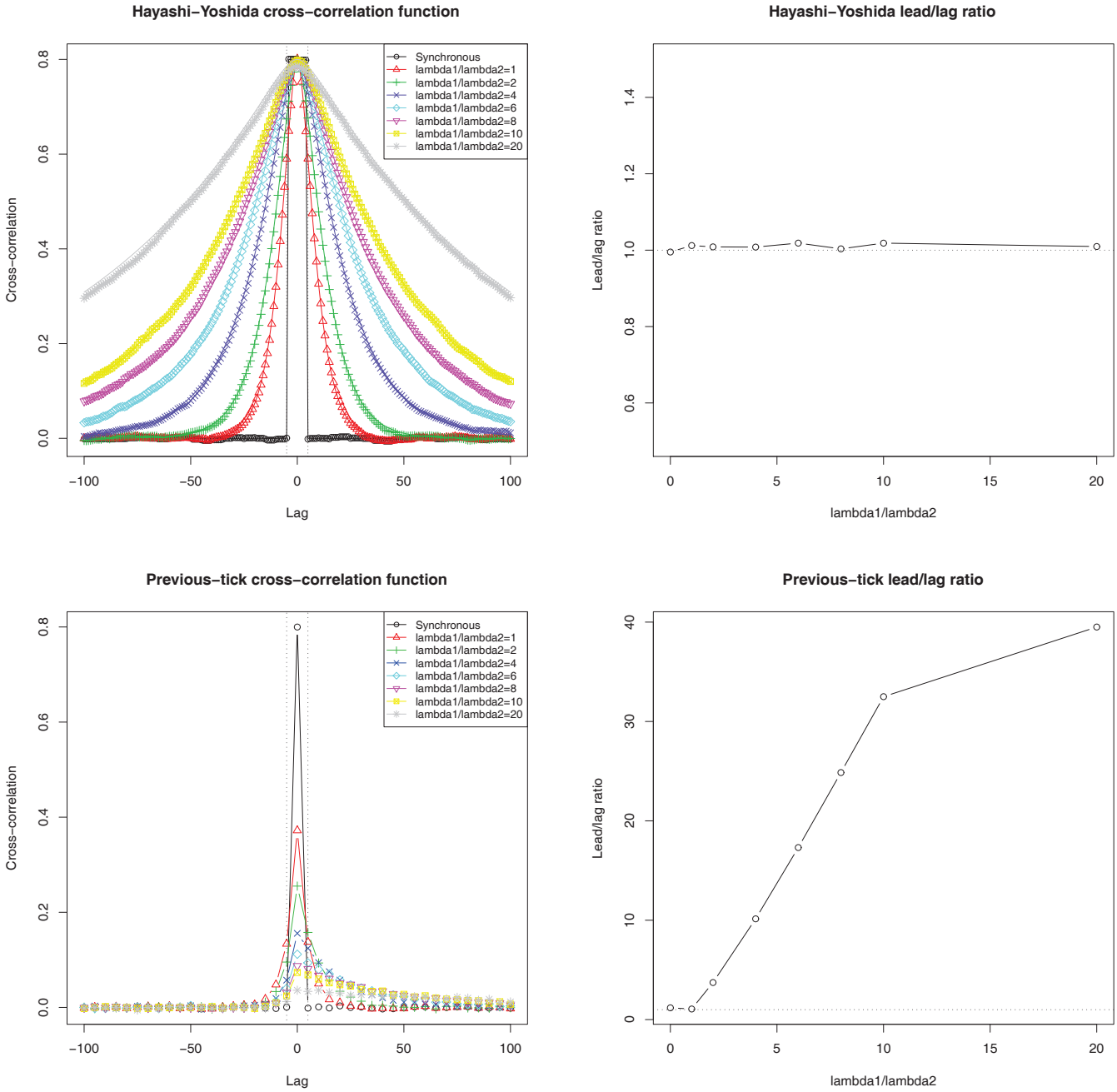

Figure 3.3: Cross-correlation of two synchronously correlated Brownian motions sampled along Poisson time grids for various levels of $\frac{\lambda_1}{\lambda_2}$, with $\lambda_1 = \frac{1}{\Delta t} = 0.2$ kept fixed. Top left panel: Hayashi-Yoshida cross-correlation function (symbols) and its closed-form expression (straight lines). Top right panel: Hayashi-Yoshida LLR. Bottom left panel: Previous-tick cross-correlation function. Bottom right panel: Previous-tick LLR.

From figure 3.3, we observe that the cross-correlation function is always peaked at zero, whatever the method of computation. The previous-tick correlation function shrinks to zero as the level of asynchrony increases, but we get rid of this problem with the Hayashi-Yoshida estimator. Moreover, the previous-tick correlation function is blurred by spurious liquidity effects: the asymmetry grows significantly with $\frac{\lambda_1}{\lambda_2}$, yielding the most active Brownian motion to lead the other systematically. In the contrary, the Hayashi-Yoshida LLR is not impacted by the level of $\frac{\lambda_1}{\lambda_2}$, the cross-correlation function remains symmetric (see appendix 3.7 for the proof). Even though the Hayashi-Yoshida cross-correlation remains symmetric, we notice that it is not exactly a Dirac mass. The irregular sampling creates correlation at non-zero lags (see appendix 3.7 for the proof).

As a result, we choose to use the Hayashi-Yoshida cross-correlation function to measure lead/lag relationships in our empirical study. This allows us to avoid being fooled by liquidity effects, yielding the most traded assets to be automatically leaders.

3.4 Empirical results

3.4.1 Empirical cross-correlation functions

We now turn to measuring lead/lag relationships on our dataset. Figure 3.4 shows the tick time Hayashi-Yoshida cross-correlation functions computed on four pairs of assets

- FCE/FSMI: future/future
- FCE/TOTF.PA: future/stock
- RENA.PA/PEUP.PA: stock/stock
- FSMI/NESN.VX: future/stock

We choose the following grid of lags (in seconds)

$$0, 0.01, 0.02, \ldots, 0.1, 0.2, \ldots, 1, 2, \ldots, 10, 15, 20, 30, \ldots, 120, 180, 240, 300$$

We consider that there is no correlation after five minutes of trading on these assets, which seems to be empirically justified on figure 3.4, except for the FSMI/NESN.VX case. Figure 3.5 is similar to figure 3.4, but it zooms on lags smaller than 10 seconds. In order to assess the robustness of our empirical results against the null hypothesis of no genuine lead/lag relationship but only artificial liquidity lead/lag, we build a surrogate dataset[8]. For two assets and for a given trading day, we generate two synchronously correlated Brownian motions with the same correlation as the two assets, *i.e.* $\rho = \hat{\rho}_{HY}(0)$, on $[0, T]$ with a mesh of 0.01 second, T being the duration of a trading day. Then we sample these Brownian motions along the true timestamps of the two assets, so that the surrogate data have the same timestamp structure as the original data. The error bars indicate the 95%-confidence interval for the average correlation over all trading days[9].

For the FCE/FSMI pair (future *vs* future), the cross-correlation vanishes very quickly, there is less than 5% of correlation at 30 seconds. We observe that there is more weight on the side where FCE leads with a LLR of 1.17 and a maximum correlation at 0.16 seconds. The FCE/TOTF.PA pair involves a future on an index and a stock being part of this index. Not surprisingly, the future leads by 0.37 seconds. This

[8] Even though we have proven that the Hayashi-Yoshida estimator is not impacted by the level of trading activity in appendix 3.7, we require events to happen according to Poisson arrivals. This might not be realistic for empirical data. However, we think that the Hayashi-Yoshida cross-correlation function remains symmetric under more general random sampling schemes. Results presented on figures 3.4 and 3.5 support this conjecture.

[9] Assuming our dataset is made of D uncorrelated trading days, the confidence interval for the average correlation $\bar{\rho}_D = \frac{1}{D}\sum_{d=1}^{D}\rho_d$ is $\left[\bar{\rho}_D \pm 1.96\frac{\sigma_D}{\sqrt{D}}\right]$ where $\sigma_D^2 = \frac{1}{D}\sum_{d=1}^{D}\rho_d^2 - \bar{\rho}_D^2$.

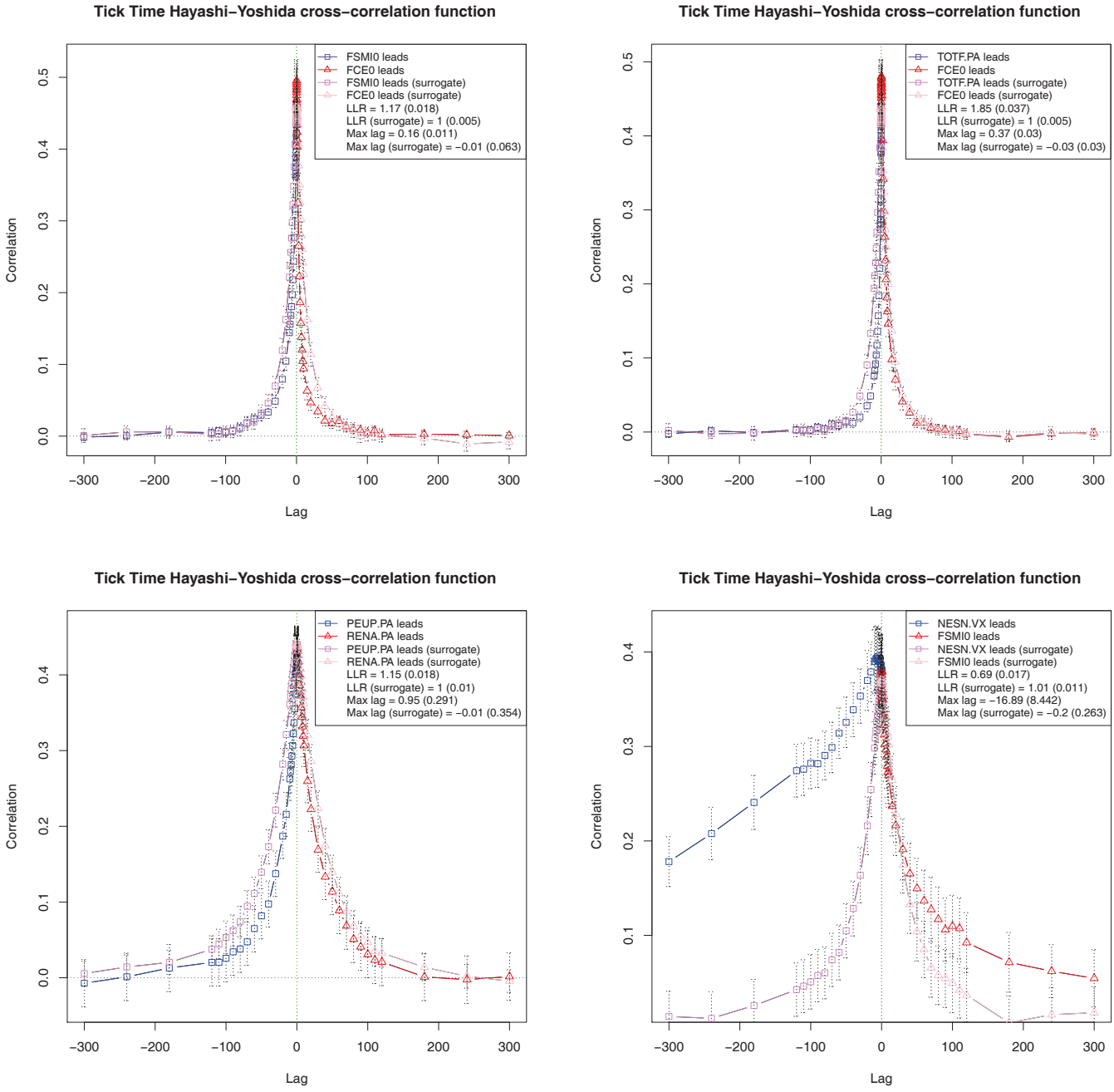

Figure 3.4: Tick time Hayashi-Yoshida cross-correlation function. Top left panel: FCE/FSMI. Top right panel: FCE/TOTF.PA. Bottom left panel: RENA.PA/PEUP.PA. Bottom right panel: FSMI/NESN.VX. Standard deviations are indicated between brackets.

pair shows the biggest amount of lead/lag as measured by the LLR (1.85). The RENA.PA/PEUP.PA case compares two stocks in the French automobile industry. The cross-correlation is the most symmetric of the four shown with a LLR of 1.15. Finally, the FSMI/NESN.VX pair is interesting because the stock leads the future on the index where it belongs. This result might be explained by the fact NESN.VX is the largest market capitalization in the SMI, about 25%. The asymmetry is quite strong (LLR = 0.69) and the maximum lag is 16.89 seconds, which is very large but we remark there is a significant amount

of noise on this average lag. We also see that there still seems to be correlation after five minutes. The difference between the maximum correlation and the correlation at lag zero is 0.045 for FCE/FSMI, 0.047 for FCE/TOTF.PA, 0.014 for RENA.PA/PEUP.PA and 0.024 for FSMI/NESN.VX, which confirms that the lead/lag relationship is less pronounced for RENA.PA/PEUP.PA. The LLR for surrogate data is equal to one and the maximum lag is statistically zero with a usual confidence level of 95% for the four pairs of assets considered. This strong contrast between real and surrogate data suggests that there are genuine lead/lag relationships between these assets that are not solely due to the difference in the levels of trading activity.

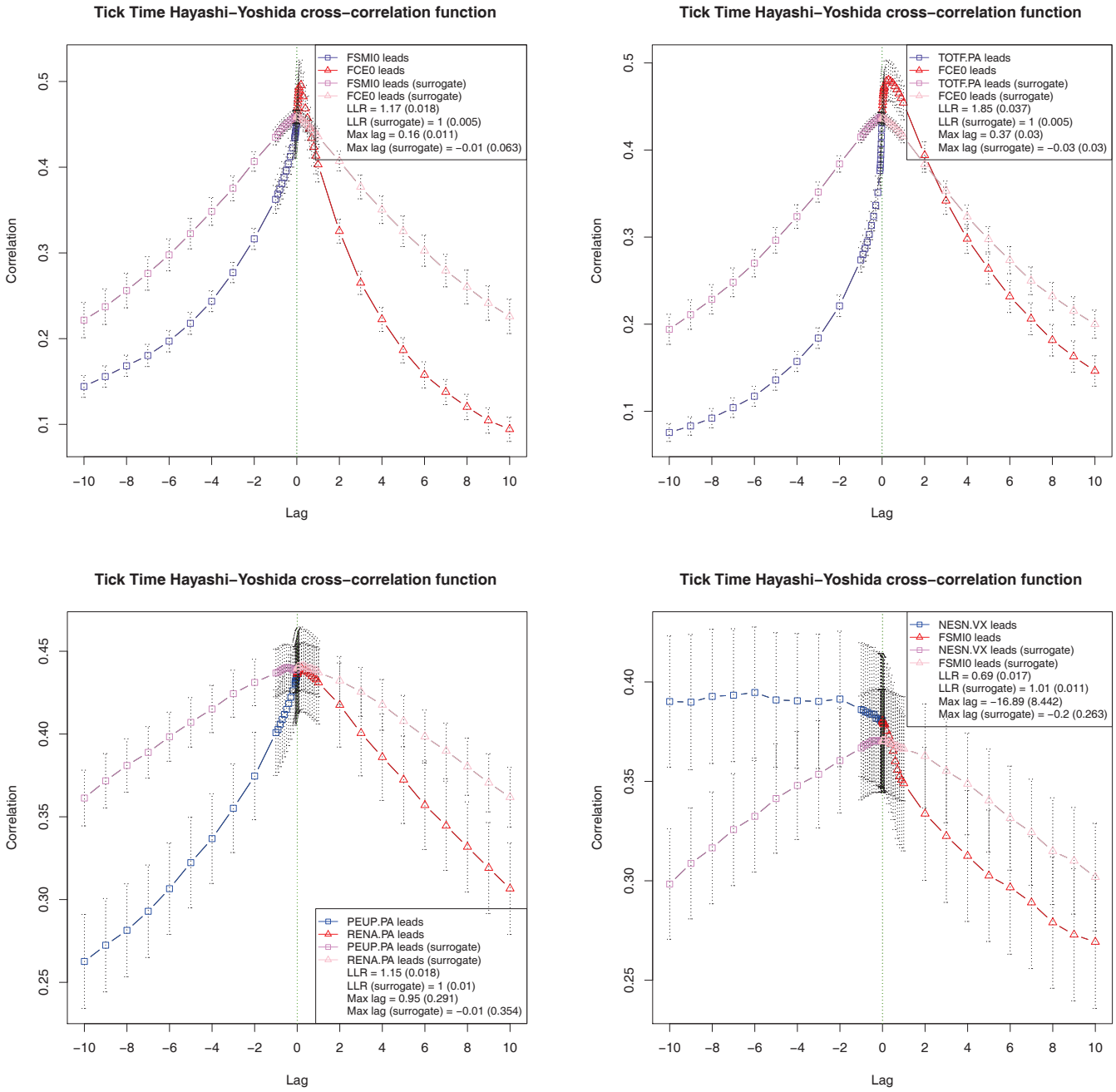

Figure 3.5: Zoom on lags smaller than 10 seconds in figure 3.4.

3.4.2 Microstructure features of leading assets

In this section, we investigate what are the common features of leading assets. It is often claimed in the literature [43, 44, 64, 66, 67, 69, 70, 75, 82] that the most liquid assets tend to be leaders, which sounds intuitive because it should take more time to illiquid assets to incorporate information into prices.

In section 3.2, we have presented several indicators measuring liquidity from the point of view of market microstructure. We now look for the dependency of our lead/lag indicator LLR on these liquidity indicators. In order to make an extensive study, we have computed the LLR and six liquidity indicators[10] for all pairs in the universe made from the CAC40 components and its future, which amounts to $41 * 40/2 = 820$ pairs. To be more precise, for all pairs (X, Y), and for each indicator I, we plot the LLR against the ratio $\mathrm{IR} = \frac{\mathrm{I_X}}{\mathrm{I_Y}}$. Remembering that the LLR is the ratio of the squared correlations when X leads over those when Y leads, it means that if $\mathrm{sign}(\mathrm{LLR} - 1) = \mathrm{sign}(\mathrm{IR} - 1)$, then the higher this indicator, the more X leads and *vice versa*. The results are shown on figure 3.6. Table 3.3 illustrates the discriminatory power of each of these indicators by counting the proportion of points falling into the four quadrants delimited by the straight lines $x = 1$ and $y = 1$.

Table 3.3: Discriminatory power of liquidity indicators.

	N^{++}	N^{--}	N^{+-}	N^{-+}	$N^{++} + N^{--}$	$N^{+-} + N^{-+}$
$\langle \Delta t \rangle$	7%	7%	44%	42%	14%	86%
$\langle P_{\text{trade}}.V_{\text{trade}} \rangle$	35%	44%	16%	5%	79%	21%
$\langle s \rangle / \delta$	17%	12%	33%	38%	29%	71%
$\langle \lvert \Delta m \rvert \rangle / \delta$	20%	13%	31%	36%	33%	67%
$\left\langle \mathbb{1}_{\{\text{trade-through}\}} \right\rangle$	29%	26%	22%	23%	55%	45%
$\langle \delta / m \rangle$	22%	26%	28%	24%	48%	52%

The most discriminatory indicators are the intertrade duration, the average turnover per trade, the average bid/ask spread and the midquote volatility. The tick size and the probability of having a trade-through do not seem to play any direct role in determining who leads or lags. The most liquid assets appear to be leaders, which is in agreement with common market knowledge. Indeed, assets which trade faster, or involve bigger exchanges of money, or have a narrower bid/ask spread, or are less volatile tend to lead on average. Even though the number of trades-through (resp. the tick size) does not emerge as a key feature at first sight, we see a decreasing (resp. increasing) trend if we focus on the future/stock pairs, *i.e.* the blue points on figure 3.6, which means that stocks having a bigger probability of trade-through (resp. a smaller tick size) are less led by the future. This is still in agreement with the intuition that the most liquid assets tend to lead.

On figure 3.7, we plot the (cross-sectional) average maximum correlation per decile of ratio of liquidity indicators. In other words, we bin pairs of assets according to the cross-sectional distribution of ratio of liquidity indicators, and we compute the average maximum correlation in each bin. Most of the weight of these distributions is concentrated around the decile where ratios of liquidity are close to one[11]. It means that highly correlated stocks tend to have a similar level of liquidity, as measured by the six indicators above. As a result, there is a trade-off in lead/lag relationships: while liquid *vs* illiquid pairs exhibit highly asymmetric cross-correlation functions, they tend to be less correlated than pairs with similar liquidity.

Figure 3.8 provides a network of stock/stock lead/lag relationships in the CAC40 universe (see [69] for another network based upon lead/lag). We use a minimum spanning tree [31] to plot the network, which

[10] We omit the probability of unit bid/ask spread because it essentially gives the same information as the average spread.

[11] The conditional distribution of correlation given the decile of the average turnover tends to be peaked on the left because of future/stock pairs. Indeed, the average turnover of the future is one order beyond the average turnover of stocks so that the decile containing the unit turnover ratio does not take into account these pairs, which are highly correlated.

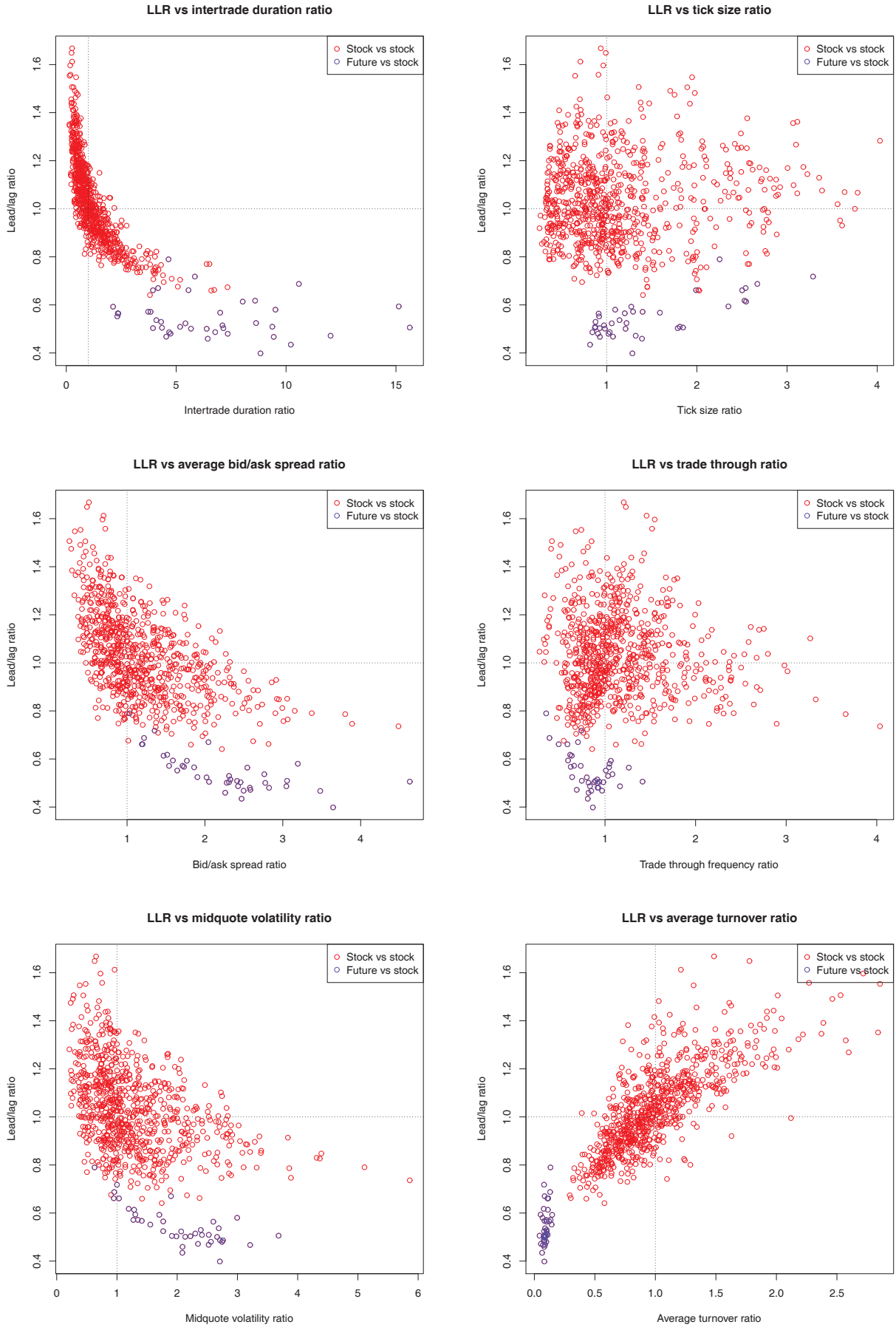

Figure 3.6: Scatterplot of LLR against pairwise ratios of various liquidity indicators for all the pairs in the CAC40 universe and its future.

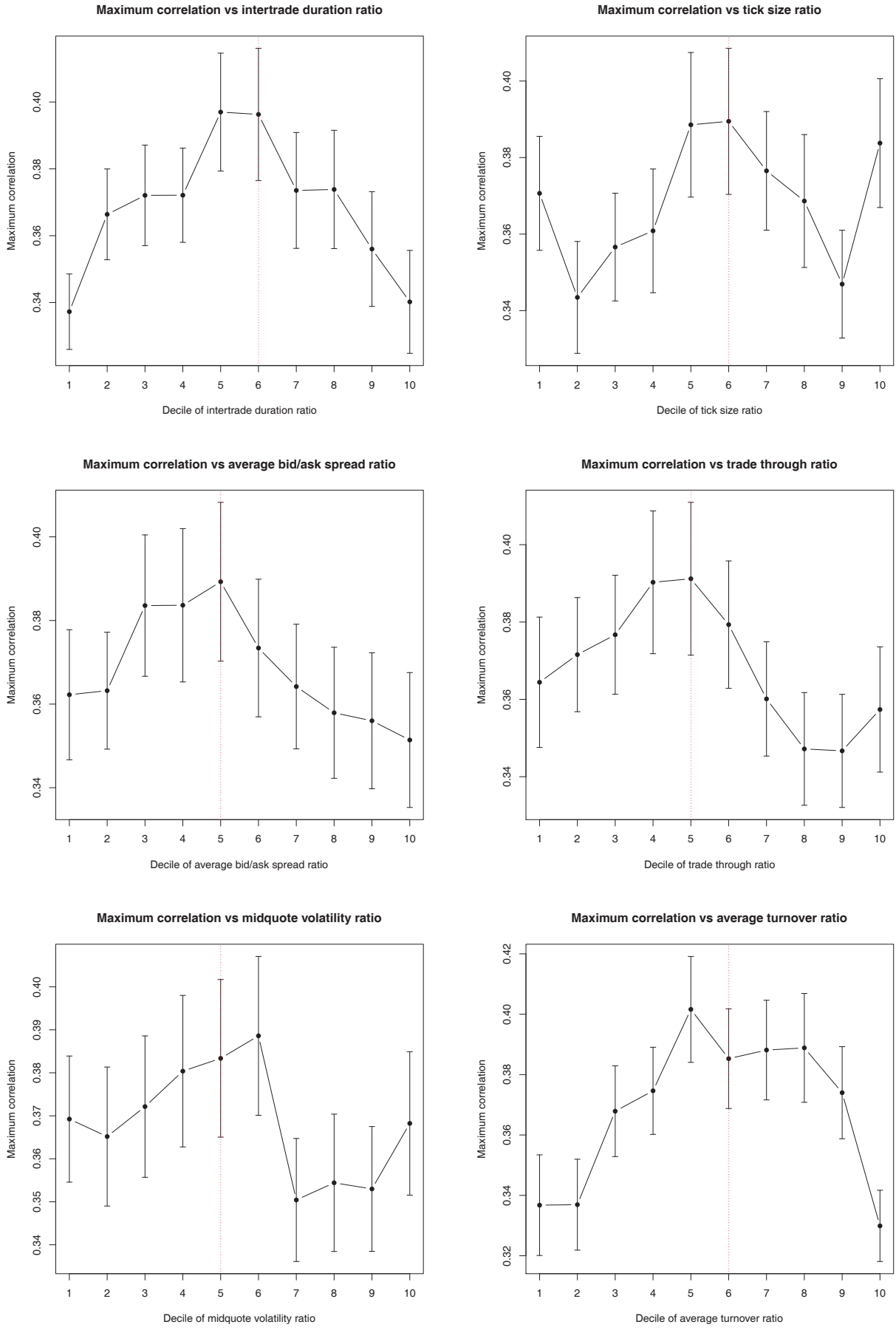

Figure 3.7: Average maximum correlation against deciles of pairwise ratios of various liquidity indicators for all the pairs in the CAC40 universe and its future. A dotted red line indicates the decile that includes one. Error bars represent Gaussian 95%-confidence intervals.

only keeps the most significant correlations by construction. We draw a directed edge from stock X to stock Y if stock X leads stock Y, *i.e.* $\text{LLR} = \frac{\sum_{\ell>0} \rho^2(\text{X leads Y by } \ell)}{\sum_{\ell>0} \rho^2(\text{Y leads X by } \ell)} > 1$. The color of the edge indicates the level of LLR of the associated pair. This network can be useful to find optimal pairs of assets for lead/lag arbitrage. Indeed, good candidates are close nodes (high correlation) with red links (high LLR).

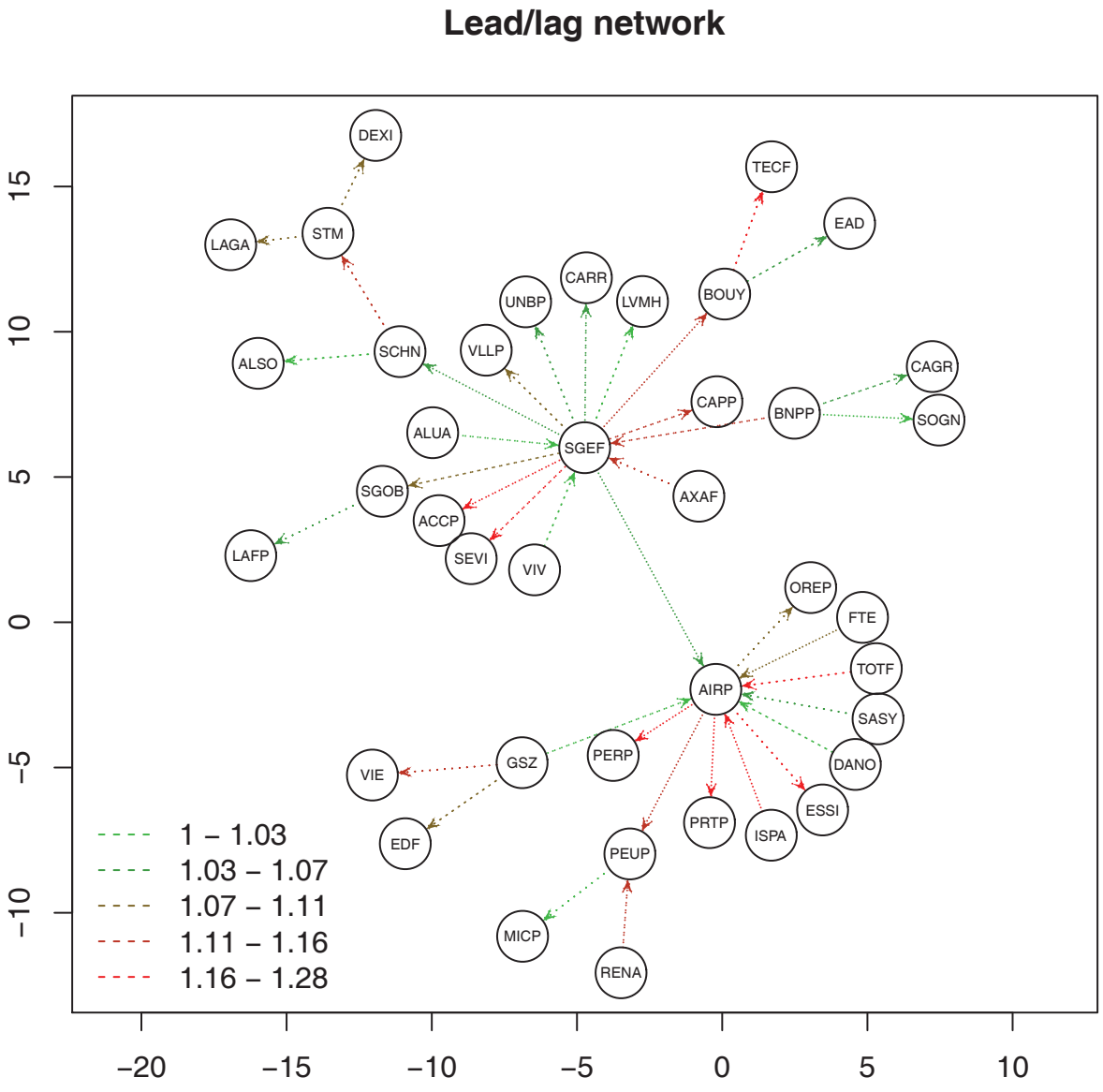

Figure 3.8: Lead/lag network on the CAC40 universe. The axes units are arbitrary.

3.4.3 Intraday profile of lead/lag

A well known stylized fact about financial markets activity is that it strongly changes over the day [4]. For instance, the intraday volatility exhibits a so-called asymmetric U-shape: massive volatility at the open, then it decreases to reach a minimum during lunch time, it peaks at macroeconomic figures announcements, and even experiences a change of regime in Europe after the opening of the US market, and finally rallies again at the close, but doesn't recover the opening level.

We study the same phenomenon for lead/lag relationships by computing our three lead/lag indicators (LLR, maximum lag measured in seconds and maximum correlation) into 5-minute slices from the open to the close and averaging over all days for each slice[12]. Figure 3.9 plots the results for future/stock pairs with whisker plots describing the cross-sectional distribution (*i.e.* the distribution among assets) for each time slice[13]. Figure 3.10 does the same for stock/stock pairs.

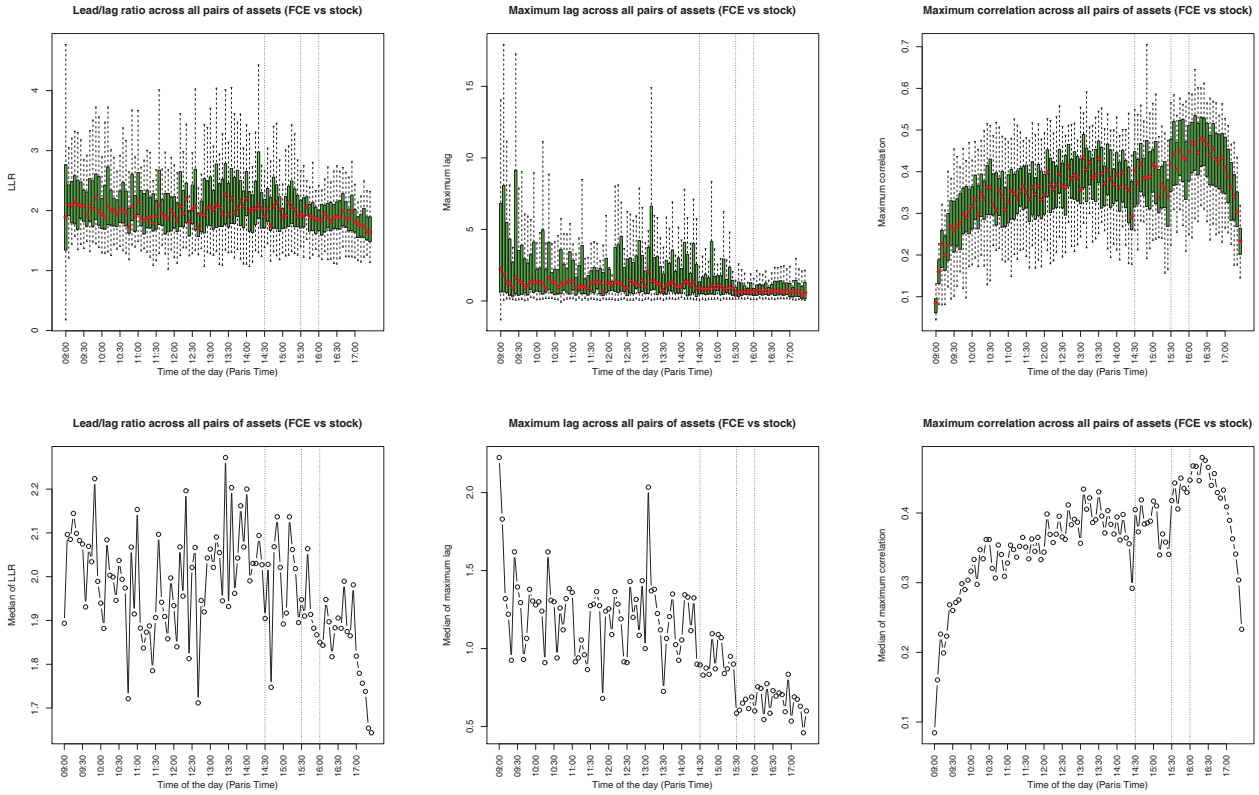

Figure 3.9: Intraday profile of lead/lag for future/stock pairs. Top panel: cross-sectional distribution of LLR, maximum lag and maximum correlation. Bottom panel: zoom on cross-sectional medians. Blue dotted lines are drawn 14:30 and 16:00 (announcement of US macroeconomic figures) and 15:30 (NYSE and NASDAQ opening).

[12] More precisely, since we don't have so much data during 5 minutes, we only consider lags no larger than a minute and we use average (over days) cross-correlation functions per slice. We also interpolate these cross-correlation functions with a spline on a regular grid of lags with mesh equal to 0.1 second (function `spline` of `R`). The maximum lag and correlation are computed with these smooth cross-correlation functions. However, the LLR is computed with correlations on the non-interpolated grid to make it comparable with values obtained in previous sections. The same approach is used in subsection 3.4.4 but we consider lags up to 300 seconds.

[13] The whisher plots we present display a box ranging from the first to the third quartile with the median in between, and whiskers extending to the most extreme point that is no more than 1.5 times the interquartile range from the box. These are the default settings of the `boxplot` function of `R`.

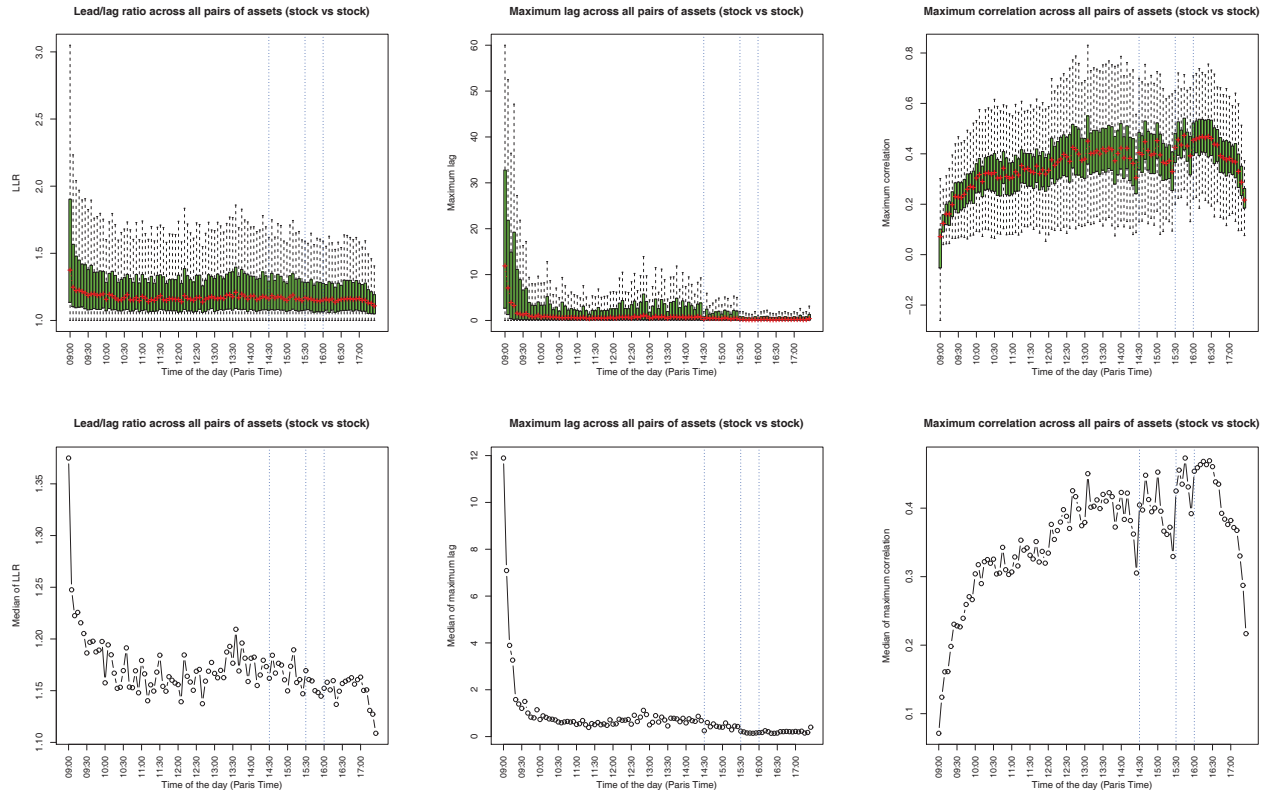

Figure 3.10: Same as figure 3.9 for stock/stock pairs.

For future/stock pairs, the LLR is always above one so that the future leads stocks all day long. The LLR jumps up five minutes after the opening, it exhibits a noisy U-shape from 09:30 to 14:00; it drops between 14:20 and 14:30, between 15:15 and 15:30, and between 15:45 and 16:00, *i.e.* before news arrivals or market openings; finally it decreases half an hour before the market closes to reach its lowest level. The maximum lag is always positive, which confirms that the future always leads stocks. We also notice that lead/lag becomes faster at 14:30 and 16:00 (announcement of US macroeconomic figures) and 15:30 (US market opening), where the maximum lag reaches local minima. There is a global upward trend in the maximum correlation as we move forward on the timeline (see [4] for more details on the intraday correlation pattern), still with significant peaks at the aforementioned specific event times, and a decorrelation as the market closing approaches.

Stock/stock pairs also show a varying intraday profile of lead/lag which is less noisy than for future/stock pairs. We first remark that lead/lag relationships are far less pronounced than in the future/stock case. Indeed, the LLR is around 1.2 on average, while it is 2.0 for future/stock pairs. The average level of correlation is similar in both cases, though there are seldom uncorrelated future/stock pairs in comparison with stock/stock pairs. Indeed, two stocks belonging to different business sectors might be little correlated, while both are strongly correlated with the future. For instance, the percentage of stock/stock pairs having a correlation less than 0.3 is 19% while it is 10% for future/stock pairs. The LLR is at its highest level at the open, which might reflect the fact that some corporate news are discovered when the market is closed, then it decreases until 10:00 and stays roughly constant until 17:00 after which it drops until the close at its lowest level. The decay of the maximum lag towards zero is similar to the one observed for LLR and shows that stock/stock cross-correlation functions tend to be symmetric around zero as time goes by. Finally, the

maximum correlation shows the same rising profile than for future/stock pairs. This comes from the fact that most of the correlation comes from the so-called "market mode" [1]. However, it is reported in [4] that idiosyncratic correlations (*i.e.* once the market mode is statistically removed) tend to decrease during the day.

3.4.4 Lead/lag conditional to extreme events

In the previous sections, we measured lead/lag relationships taking into account every non-zero price variation, whatever its magnitude. However, it sounds reasonable that large returns are more informative than small ones [6]. Therefore, we introduce a thresholded version of the cross-correlation estimator

$$\begin{aligned}
\hat{\rho}_\theta(\ell > 0) &= \frac{\sqrt{N_\theta^X N_0^Y}}{N_\theta^X(\ell)} \frac{\sum_{i,j} r_i^X r_j^Y \mathbb{1}_{\{O_{ij}^\ell \neq \emptyset\}} \mathbb{1}_{\{|r_i^X| \geq \theta\}}}{\sigma_\theta^X \sigma_0^Y} \\
\hat{\rho}_\theta(\ell < 0) &= \frac{\sqrt{N_0^X N_\theta^Y}}{N_\theta^Y(\ell)} \frac{\sum_{i,j} r_i^X r_j^Y \mathbb{1}_{\{O_{ij}^\ell \neq \emptyset\}} \mathbb{1}_{\{|r_j^Y| \geq \theta\}}}{\sigma_0^X \sigma_\theta^Y} \\
\sigma_\theta^k &= \sqrt{\sum_i (r_i^k)^2 \mathbb{1}_{\{|r_i^k| \geq \theta\}}} \\
N_\theta^k &= \sum_i \mathbb{1}_{\{|r_i^k| \geq \theta\}} \\
N_\theta^k(\ell) &= \sum_{i,j} \mathbb{1}_{\{|r_i^k| \geq \theta\}} \mathbb{1}_{\{O_{ij}^\ell \neq \emptyset\}}
\end{aligned}$$

We only take into account variations of the price of the leading asset that are greater or equal than some threshold θ. For $\ell = 0$, we compute both possibilities. Since prices of assets lie on a large scale, we use relative returns rather than price differences for this section only, *i.e.* $r_i^k = \frac{P_{t_i}^k - P_{t_{i-1}}^k}{P_{t_{i-1}}^k}$. We consider thresholds θ up to four basis points (bps), $\theta = i * 10^{-4}, i = 0, 0.5, \ldots, 4$ for stock/stock pairs. The average volatility on the universe we focus on being roughly 1.3 bps, it means that we consider events up to three standard deviations for stock/stock pairs. The future is signifcantly less volatile than stocks, its average volatiltiy being approximately 0.5 bps, so we consider thresholds not exceeding 2.5 bps, which amounts to five standard deviations of future midquote returns. For this section, we use trading time sampling rather than tick time in order to get as many events as possible. Figure 3.11 plots the LLR, maximum lag and maximum correlation as a function of θ for future/stock and stock/stock pairs.

The overall trend is that lead/lag becomes more and more pronounced as we focus on larger price variations. Indeed, both the LLR and the maximum correlation increase with the threshold θ. There is roughly three times more correlation in future/stock pairs when θ goes from 0 to 2.5 bps. The maximum lag is quite independent from θ. Figure 3.11 suggests that one should filter out insignificant moves of the leader when trying to build up a forecast of the lagger.

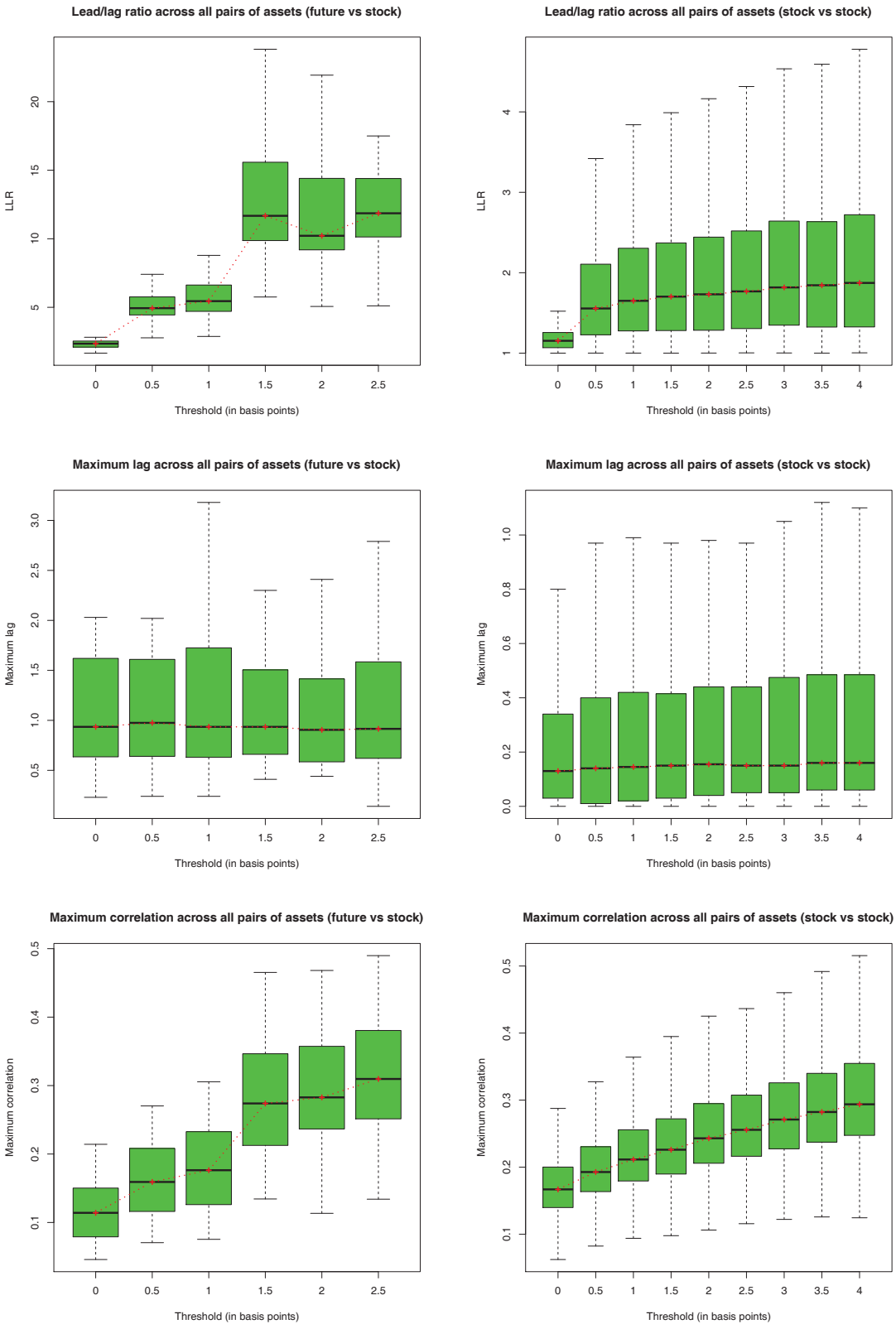

Figure 3.11: Lead/lag measures as functions of the threshold θ. Left panel: cross-sectional distribution for LLR, maximum lag and maximum correlation for future/stock pairs. Right panel: Idem for stock/stock pairs.

3.4.5 Lead/lag response functions

Cross-correlation helps us in gaining insight into lead/lag relationships. However, it does not tell us the amplitude of the variation of the lagger following a variation of the midquote of the leader. From a practical point of view, it is of great importance because an arbitrage strategy based on market orders is only profitable if it generates enough profit to bypass the bid/ask spread. As a result, we study the so-called lead/lag response functions

$$R_{v,\substack{\leq\\>}}(\ell,\theta) = \left\langle v_{t+\ell}^{\text{stock}} - v_t^{\text{stock}} | r_t^{\text{future}} \substack{\leq\\>} \theta \right\rangle$$

for v being any relevant variable in the order book, such as the bid and ask quotes or the bid/ask spread. The main issue in measuring such a function is the following: after a jump of the future, one can only record the trajectory of the stock before any other jump of the future happens if one wants to isolate the impact of that particular jump. Since futures are much more actively traded than stocks, this can lead to a substantial lack of data for large lags, which is why we show empirical measures for lags less than 5 seconds. We need to monitor the state of the best quotes of the stock continuously so we use the quotes files (see section 3.2). As in the previous section, we use trading time to compute the returns of the future.

Figure 3.12 plots the response function for the FCE/TOTF.PA pair with v being the bid/ask quotes and the bid/ask spread. The same graphs for FDX/DTEGn.DE, FFI/VOD.L and FSMI/NESN.VX are displayed in appendix 3.8.

The first row of figure 3.12 measures how much the bid/ask quotes of TOTF.PA move away from their initial level after a change in the midquote of the future FCE. Since TOTF.PA and FCE are positively correlated, the deviation is positive (resp. negative) for positive (resp. negative) thresholds θ. For $|\theta| \leq 2$, it tends to saturate after approximately 1 second, the same order of magnitude than the lag where the cross-correlation function reaches its maximum. The deviation is quite small, typically less than half a tick, and it increases with $|\theta|$, meaning that the larger the return of the future, the bigger the impact on the stock, which sounds intuitive (see [6] for a similar study on single stock response functions). The curves for $|\theta| > 2$ are a bit messy because of the lack of such events but qualitative results remain unchanged.

The middle panel plots the deviation of the bid (resp. ask) quote from the initial best opposite quote, *i.e.* the ask (resp. bid) quote, after a variation of the midquote of the future. If the bid (resp. ask) quote becomes larger (resp. smaller) than the initial ask (resp. bid) quote, then it is possible to make money with market orders on the stock by buying (resp. selling) at time zero and unwinding the position afterwards. On the middle panel, we see that curves with dots (resp. triangles) are always below (resp. above) zero, so it is not possible to make money with market orders. We would rather lose about two ticks on average, which is the average bid/ask spread of TOTF.PA over this period.

Finally, the bottom panel depicts the trajectory of the bid/ask spread of the stock after a move of the future. The variation of the spread is not so big for $|\theta| \leq 2$, typically smaller than five percent of the tick size. We observe a relaxation of the spread towards its average value at inception. For $|\theta| > 2$, the spread narrows for small lags before being wider a few seconds after. This can be due to high frequency market making algorithms of index arbitrage traders, who try to replicate the future with stocks and post quotes accordingly. These people act at very high frequency, often less than a second. The widening of the spread for larger lags might come from agents who follow the evolution of the future as a signal for arbitrage strategies and send market orders on stocks once the future has moved significantly.

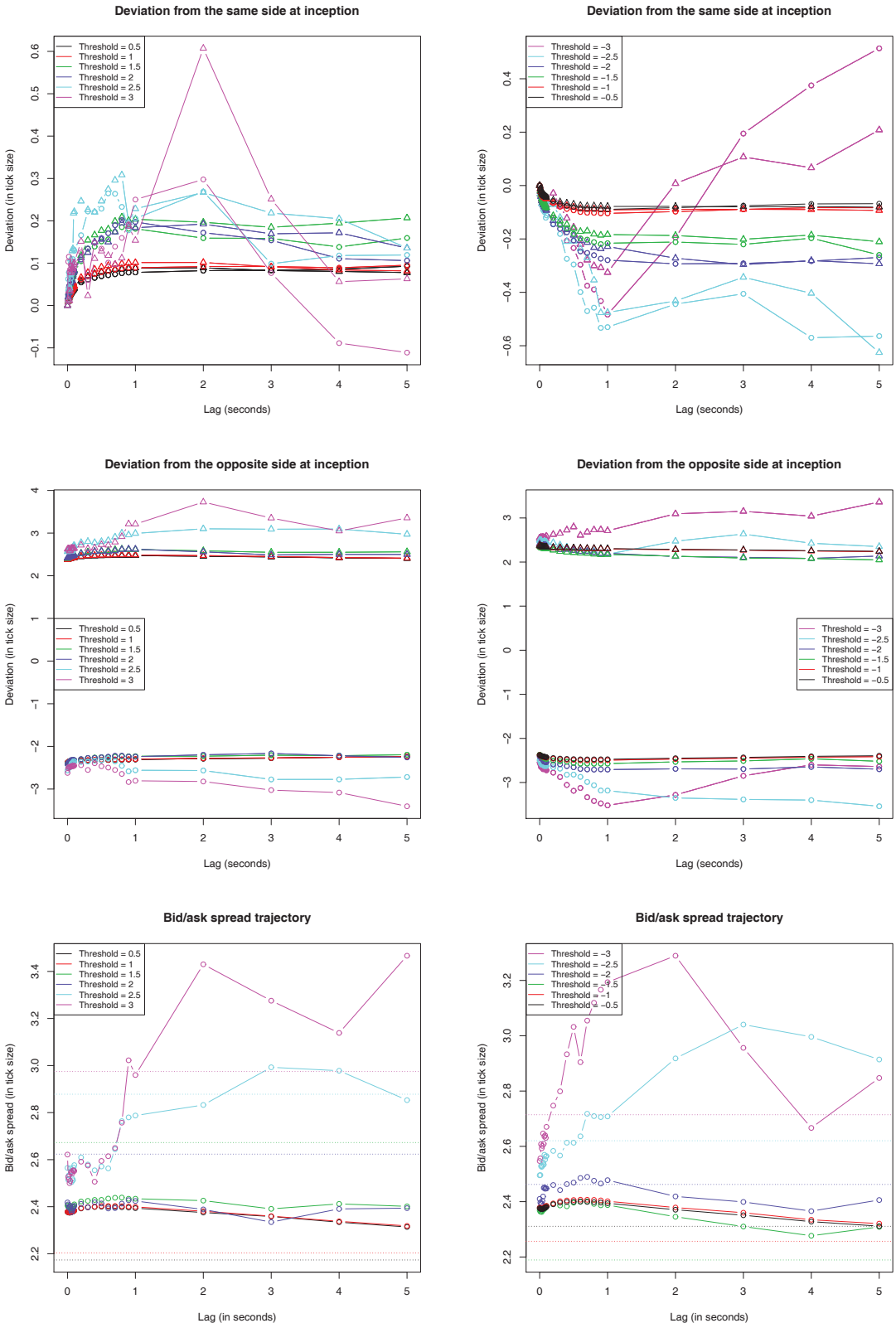

Figure 3.12: Lead/lag response functions for FCE/TOTF.PA. Results for positive (resp. negative) thresholds are shown on the left (resp. right) panel. Top panel: deviation of the bid (dots) and ask (triangles) quotes w.r.t. their initial level. Middle panel: deviation of the bid (dots) and ask (triangles) quotes w.r.t. the opposite best quote at inception. Bottom panel: Bid/ask spread. Dotted lines are the average bid/ask spreads of the stock at time zero.

3.4.6 Backtest of forecasting devices

The knowledge of lead/lag relationships can be used to forecast the short-term evolution of lagging assets and thus build statistical arbitrage strategies. More precisely, the cross-correlation functions shown in subsection 3.4.1 enable us to forecast whether the midquote will move up or down on the next price change. This forecast is built using the past evolution of the leading asset. For instance, if we assume that X is the leader, and that we are at time s_{j-1}, our estimation of the next midquote return of the lagger $r_j^Y = Y_{s_j} - Y_{s_{j-1}}$ is

$$\hat{r}_j^Y = \sum_{k=1}^{p} \beta_k \sum_i r_i^X \mathbb{1}_{\left\{O_{ij}^{\ell_k} \neq \emptyset\right\}}$$

In the following, we will only be interested in the sign of the midquote return $\text{sign}(\hat{r}_j^Y)$, so we set $\beta_k = \hat{\rho}(\ell_k)$, which is estimated on the last 20 trading days. In practice, we set the last regression lag ℓ_p to be the last statistically significant lag. Since our clock is running in tick time, our classification is binary: upward or downward move of the midquote. Note that we also need an estimate of the next tick timestamp of the lagger to compute $\mathbb{1}_{\left\{O_{ij}^{\ell_k} \neq \emptyset\right\}}$. We choose it to be the current timestamp plus the average duration between two ticks over the last 20 trading days[14].

Once the prediction of the next midquote move is computed, then if it is positive (resp. negative) we buy (resp. sell) one monetary unit of the lagger, and then we sell (resp. buy) it back after the next tick of the lagger occurs. Regarding the execution costs, we consider two scenarios: execution at the midquote and execution taking into account the bid/ask spread. Midquote execution is clearly not realistic at all but it gives an upper bound for the P&L of the strategy. Note that even with a perfect forecasting device, we need the opposite quote to move more than the initial bid/ask spread at the next tick to make money with market orders, which is highly unlikely according to subsection 3.4.5.

Figure 3.13 (resp. 3.14) plots the accuracy, defined as the percentage of good predictions of our forecasting device, a random forecast and a forecast based on the autocorrelation function of TOTF.PA (resp. ESSI.PA[15]) over the 44 test days if we take the future FCE to be the leader. It also shows the probability distribution of returns of the strategy[16]. Figure 3.15 shows the distribution of returns when taking into account the bid/ask spread[17]. The lead/lag prediction is right in 60% (resp. 63%) of cases on average for TOTF.PA (resp. ESSI.PA), which is much better than a random forecast[18]. As a result, being able to trade at the midquote yields a profitable strategy with an average return of 0.39 (resp. 0.52) bps per trade. The average daily return is 20.2% (resp. 5.7%) and the standard deviation is 15% (resp. 4.2%), thus the annualized Sharpe ratio is 21 (resp. 22), which is extremely high. The distribution of returns of the lead/lag strategy is significantly different from the random strategy as judged by the Kolmogorov-Smirnov test, with a distance $D = 0.0904$ (resp. 0.1127)[19] yielding a p-value of the order of 10^{-16}. The lead/lag strategy also performs better than the autocorrelation strategy (accuracy of 57% for TOTF.PA and 59% for ESSI.PA), itself performing significantly better than a purely random strategy. The Student t-test concludes that both average prediction rates are significantly different from each other (p-value$=2 \times 10^{-9}$ (resp. 8×10^{-6})). This shows that using the past information from the leader yields a significant improvement in the forecasting process. This is close to the notion of Granger causality [55]. However, all the profits made by the lead/lag strategy is lost when taking into account the bid/ask spread: the average return is -2.74 (resp. -3.45) bps per trade. The overwhelming returns obtained by trading at the midquote are unfortunately unreachable

[14] Clearly, this could be improved by taking into account the highly seasonal pattern of interevent duration.

[15] We choose TOTF.PA and ESSI.PA because TOTF.PA is highly liquid in contrast with ESSI.PA.

[16] The flame-like shape of the probability distribution comes from the fact that returns can be written as $\frac{m_1 - m_0}{m_0} = \frac{i\delta/2}{m_0}$ where $i \in \mathbb{N}$ is the midquote variation expressed in half-ticks and $m_0 \in \mathbb{R}^+$ is the midquote at the inception of the trade.

[17] That is buying at the ask price and selling at the bid price.

[18] More precisely, the Student t-test for equality of the average prediction rates yields a p-value of 10^{-16} for both stocks.

[19] The sample size is 226142 (resp. 48295) returns over the 44 test days.

in real life. The bid/ask spread insures that such naive lead/lag strategies are not profitable with market orders only.

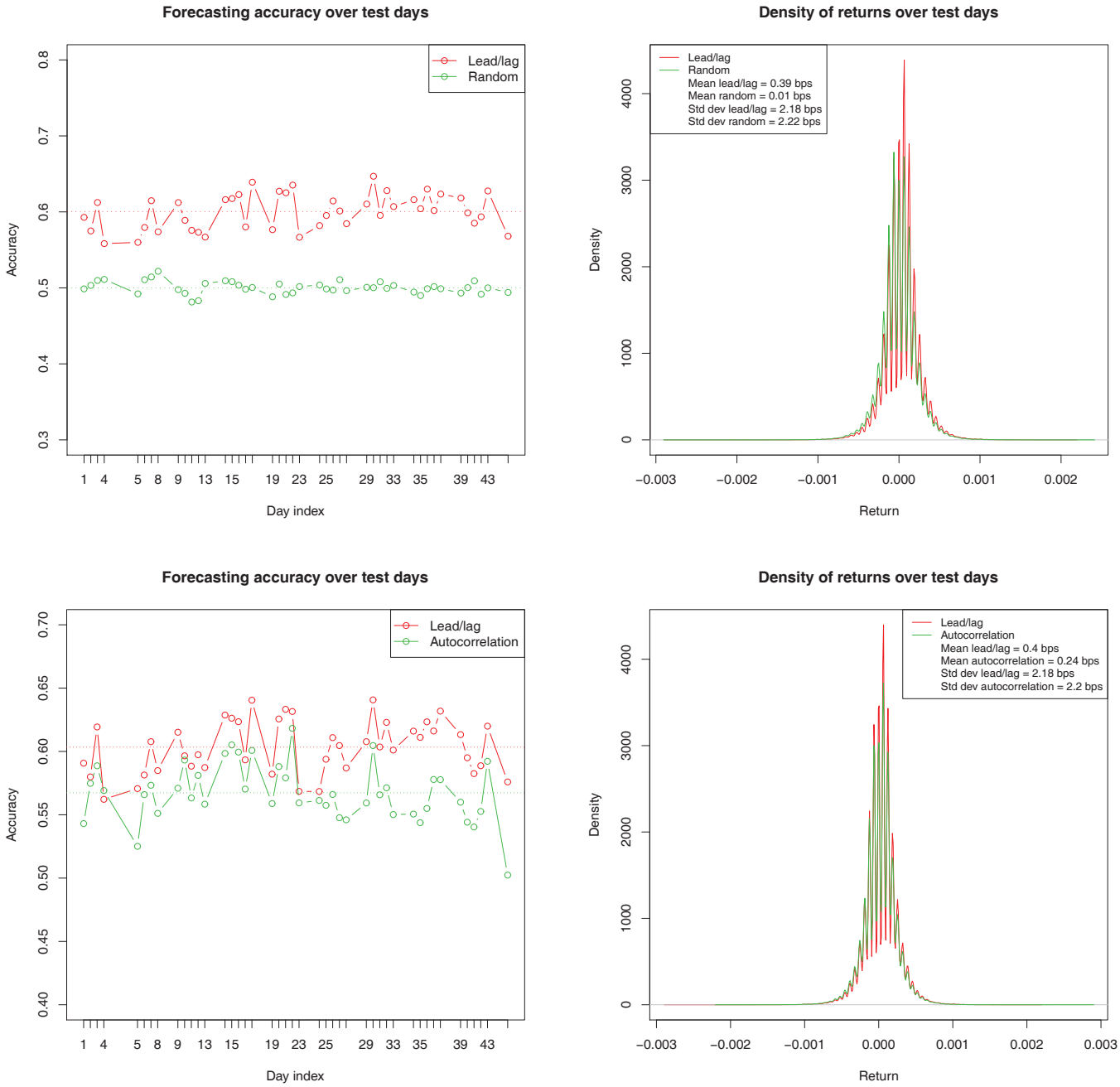

Figure 3.13: Backtest of the lead/lag strategy (*versus* a random forecast and a forecast based on autocorrelation) on the pair FCE/TOTF.PA over the 44-day test period 2010/03/29 – 2010/05/31. Top left panel: Forecasting accuracy of lead/lag *vs* random. Top right panel: Density of the returns of the strategy lead/lag *vs* random. Bottom left panel: Forecasting accuracy of lead/lag *vs* autocorrelation. Bottom right panel: Density of the returns of the strategy lead/lag *vs* autocorrelation.

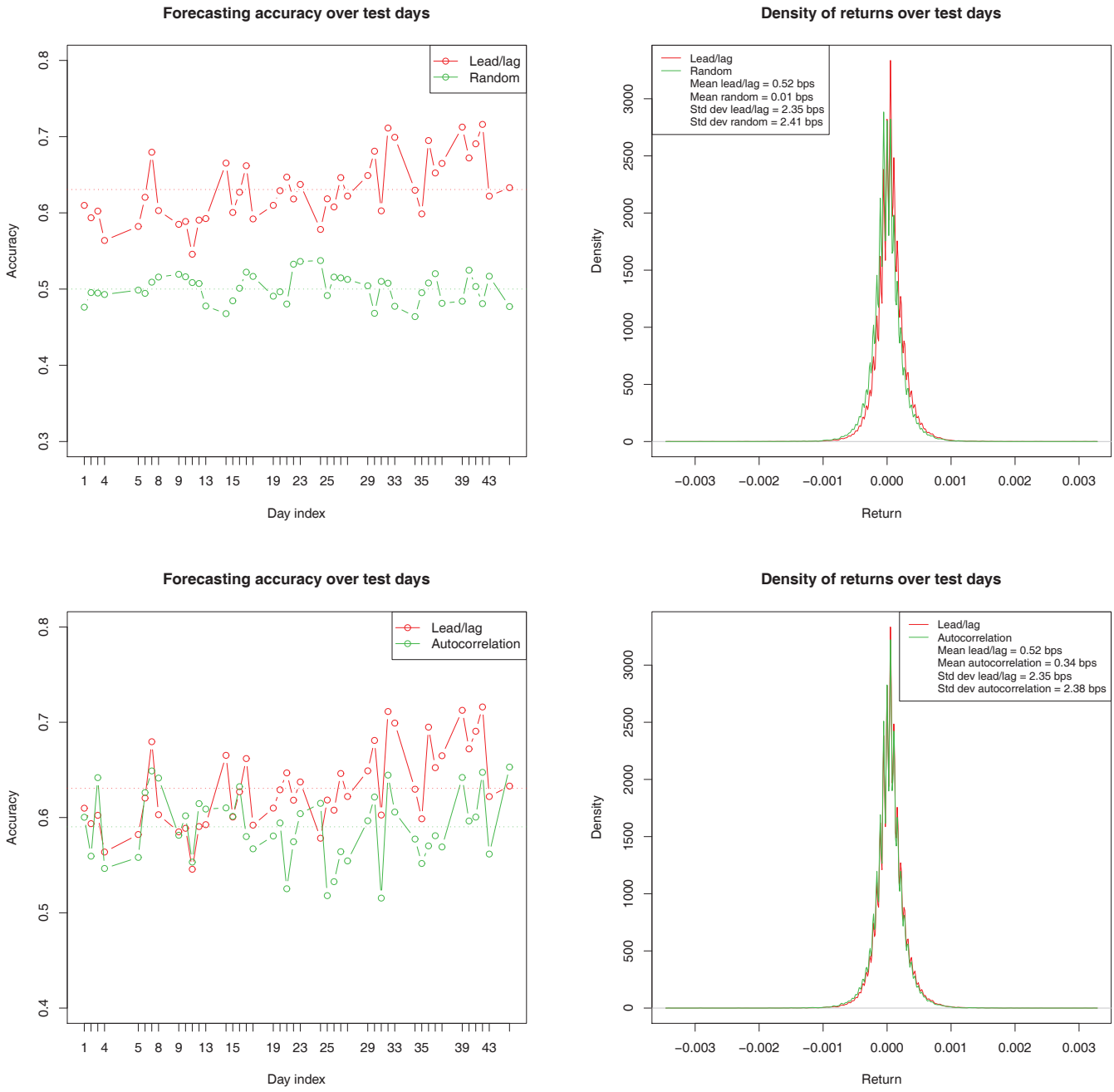

Figure 3.14: Same as figure 3.13 for FCE/ESSI.PA.

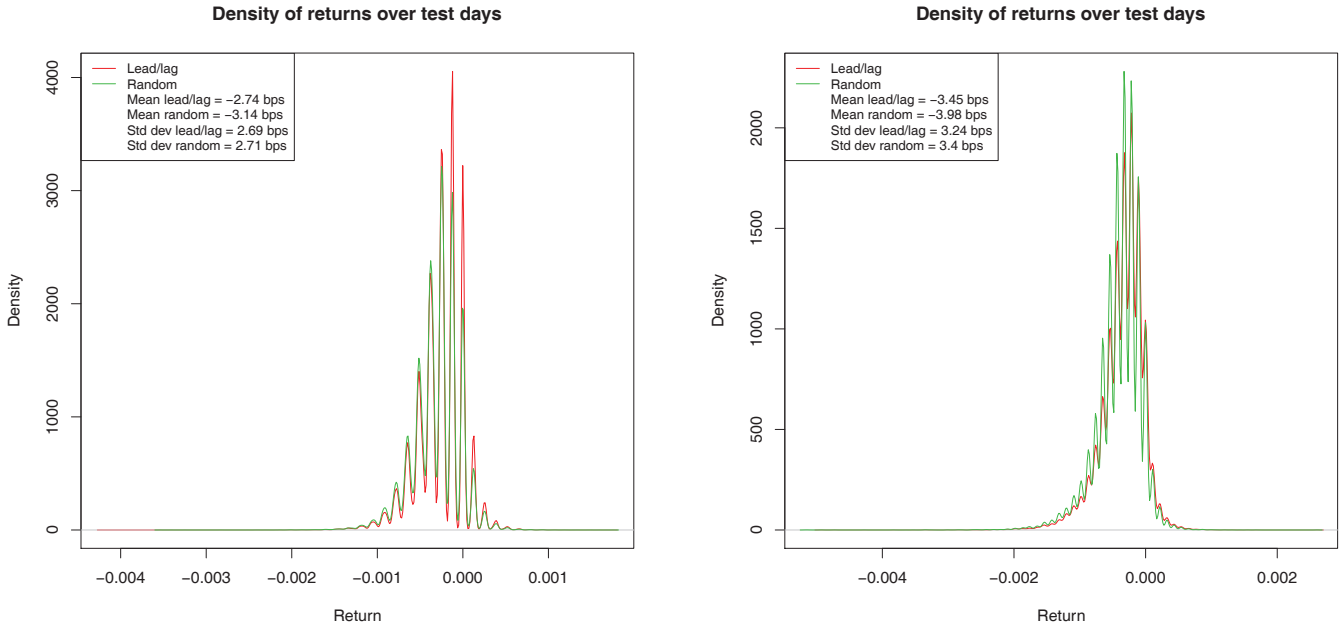

Figure 3.15: Density of the returns of the lead/lag strategy over the 44-day test period 2010/03/29 – 2010/05/31 when taking into account the bid/ask spread. Left panel: FCE/TOTF.PA. Right panel: FCE/ESSI.PA

In order to bypass massive losses due to the bid/ask spread, we try to predict the midquote of the lagger at a longer horizon than the next tick. For instance, we sample data on a bigger tick time basis, *i.e.* once the midquote has moved of θ ticks, with $\theta > 0.5$. Typically, we need to have a tick time bigger than the bid/ask spread which is 1.88 (resp. 2.83) ticks on average for TOTF.PA (resp. ESSI.PA). We use the cross-correlation computed at this time scale to forecast the midquote of the lagger. Figure 3.16 plots the accuracy and the distribution of returns for $\theta = i/2, i = 1, \ldots, 6$. Clearly, the forecasting accuracy deteriorates and the distribution of returns gets wider as the time scale increases, in agreement with the absence of arbitrage. Note that the median return remains negative for any θ.

3.5 Conclusion and further research

We study high frequency lead/lag relationships on the French equity market. We use the Hayashi-Yoshida cross-correlation function estimator because it bypasses the issues of asynchrony and artificial liquidity lead/lag. Lead/lag relationships between two stocks or between an equity index future and a stock belonging to this index show different behaviours. The later are far more pronounced than the former. From a more general point of view, we find that the most liquid assets, in terms of short intertrade duration, high trading turnover, narrow bid/ask spread and small volatility tend to lead the others. However, the highest correlations on the market appear for assets displaying similar levels of liquidity. Lead/lag relationships display a non-constant intraday profile that is different for future/stock and stock/stock pairs. Lead/lag becomes more pronounced when focusing on extreme price movements, in terms of both level of asymmetry and correlation. The examination of response functions shows that the average response time of a stock after a move of the index future is of the order of one second and that there is no chance to make money from this effect with market orders. Finally, we obtain an average prediction rate of 60% when forecasting the next midquote variation of a stock with the past evolution of the index future. As said earlier, it does not allow to make profits by sending market orders only but it could be used for other trading purposes such as market-making or best execution.

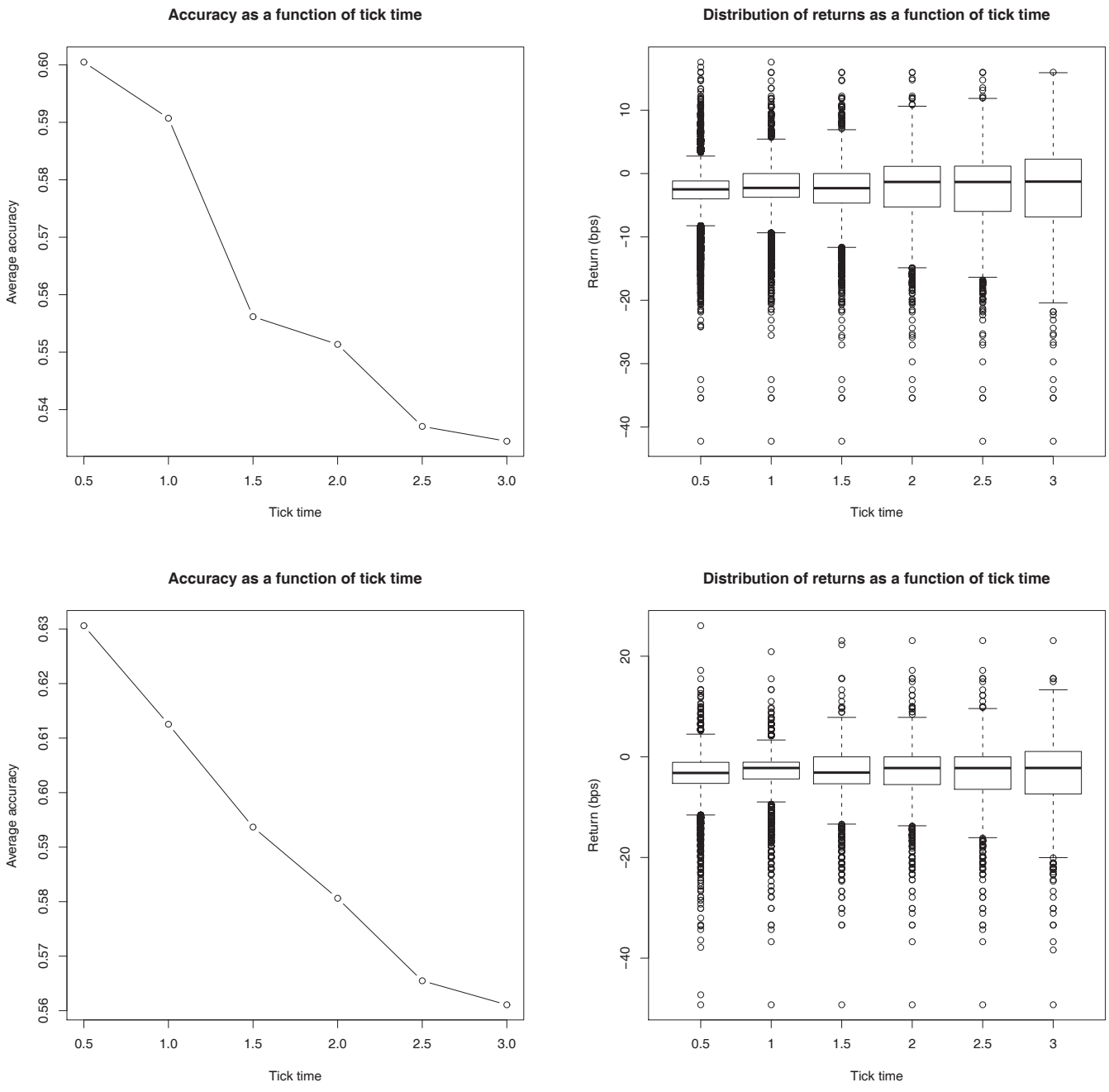

Figure 3.16: Top left panel: Forecasting accuracy as a function of tick time for FCE/TOTF.PA. Top right panel: Distribution of returns per trade as a function of tick time for FCE/TOTF.PA. Bottom panel: idem for FCE/ESSI.PA

3.6 Appendix: Explicit computation of LLR

We prove how to end up with the formulation of the LLR as a ratio of squared correlations. We use the notations introduced in subsection 3.3.1. Recall that X leads Y if X forecasts Y more accurately than Y does for X, *i.e.*

$$\frac{\left\|\varepsilon^{YX}\right\|}{\|r^Y\|} < \frac{\left\|\varepsilon^{XY}\right\|}{\|r^X\|}$$

By definition,

$$\varepsilon^{YX} = r_t^Y - r_{t^-}^X \beta = Y - r_{t^-}^X (C^{XX})^{-1} C^{YX}$$

where $r_{t^-}^X = (r_{t-\ell_1}^X, \ldots, r_{t-\ell_p}^X)^T$ so that

$$\frac{\left\|\varepsilon^{YX}\right\|^2}{\|Y\|^2} = 1 - (C^{YX})^T (C^{XX} C^{YY})^{-1} C^{YX}$$

If we expand $(C^{YX})^T (C^{XX} C^{YY})^{-1} C^{YX}$, we get

$$\frac{\left\|\varepsilon^{YX}\right\|^2}{\|r^Y\|^2} = 1 - \sum_{i,j=1}^{p} C_i^{YX} C_j^{YX} \left(C^{XX}\right)_{ij}^{-1} \left(C^{YY}\right)^{-1}$$

Assuming C^{XX} is diagonal, then we get

$$\begin{aligned}\frac{\left\|\varepsilon^{YX}\right\|^2}{\|r^Y\|^2} &= 1 - \sum_{i=1}^{p} \left(\frac{C_i^{YX}}{\sqrt{C_{ii}^{XX} C^{YY}}} \right)^2 \\ &= 1 - \sum_{i=1}^{p} \rho^2(\ell_i)\end{aligned}$$

which ends the computation.

3.7 Appendix: Explicit computation of $\mathbb{E}(\hat{C}(\ell))$

Let us consider two standard Brownian motions B_1, B_2 with contemporary correlation $\rho(0) = \rho \in [-1, 1]$. These two Brownian motions are sampled along respective time grids $0 = t_0 \leq t_1 \leq \ldots \leq t_n = T$ and $0 = s_0 \leq s_1 \leq \ldots \leq s_m = T$. The time grids are respectively the jumping times of two independent Poisson processes N_1 and N_2 and are also independent of (B_1, B_2). This results in piecewise constant processes X and Y

$$\begin{aligned}X(u) &= B_1(t(u)) \\ Y(u) &= B_2(s(u)) \\ t(u) &= \max\{t_i | t_i \leq u\} \\ s(u) &= \max\{s_i | s_i \leq u\}\end{aligned}$$

We want to compute $\mathbb{E}(\hat{\rho}(\ell))$ for any lag ℓ. In fact, we are only interested in the covariance function $\hat{C}(\ell) = \hat{\rho}(\ell)\hat{\sigma}_X\hat{\sigma}_Y$ where $\hat{\sigma}_k^2 = \frac{1}{T}\sum_i (r_i^k)^2$ since standard results show that $\hat{\sigma}_k^2$ is an unbiased and consistent estimator of the realized variance in this framework [65]. Let us assume $\ell \geq 0$. We have, using the notations from subsection 3.3.1

$$\begin{aligned} T.\mathbb{E}(\hat{C}(\ell)) &= \mathbb{E}(\sum_{i,j} r_i^X r_j^Y \mathbb{1}_{\left\{O_{ij}^\ell \neq \emptyset\right\}}) \\ &= \rho\mathbb{E}(\sum_{i,j} \mathbb{1}_{\{\ell < s_j - t_{i-1}\}}(t_i \wedge s_j - t_{i-1} \vee s_{j-1})^+) \end{aligned}$$

where $x^+ = x \vee 0$. Similarly,

$$T.\mathbb{E}(\hat{C}(-\ell)) = \rho\mathbb{E}(\sum_{i,j} \mathbb{1}_{\{\ell < t_i - s_{j-1}\}}(t_i \wedge s_j - t_{i-1} \vee s_{j-1})^+)$$

Let us remark that for $\ell = 0$ the covariance function is unbiased [61] since

$$\begin{aligned} &\mathbb{E}(\sum_{i,j} \mathbb{1}_{\{0 < s_j - t_{i-1}\}}(t_i \wedge s_j - t_{i-1} \vee s_{j-1})^+) \\ =&\mathbb{E}(\sum_{i,j} (t_i \wedge s_j - t_{i-1} \vee s_{j-1})^+) = \mathbb{E}(T) = T \end{aligned}$$

It is also clear that $\mathbb{E}(\hat{C}(T)) = \mathbb{E}(\hat{C}(-T)) = 0$.

$(t_i \wedge s_j - t_{i-1} \vee s_{j-1})^+$ is the length of the overlap between $]t_{i-1}, t_i]$ and $]s_{j-1}, s_j]$. If there is indeed an overlap, it can also be seen as the duration between two consecutive events of the Poisson process resulting from the merge of the two initial Poisson processes. A standard result on the Poisson process states that the merge of two independent Poisson processes is also a Poisson process with an intensity that is the sum of the two [42]. Therefore, we have, for $0 < \ell < T$

$$T.\mathbb{E}(\hat{C}(\ell)) = \rho\mathbb{E}(\sum_{k=1}^{N} \mathbb{1}_{\left\{\ell < \tau_{\overline{i}_2(k)} - \tau_{\underline{i}_1(k)}\right\}}(\tau_k - \tau_{k-1}))$$

where $\{\tau_k, k = 0, \ldots, N = n + m\}$ are the jumping times of the merged Poisson process and

$$\begin{aligned} \underline{i}_p(k) &= \operatorname*{argmax}_j \{\tau_j \leq \tau_{k-1} : \tau_j \text{ is of type } p\} \vee 0 \\ \overline{i}_p(k) &= \operatorname*{argmin}_j \{\tau_j \geq \tau_k : \tau_j \text{ is of type } p\} \wedge N \end{aligned}$$

for $p = 1, 2$. Then, since $N - 1$ is Poisson distributed with parameter $(\lambda_1 + \lambda_2)T$,

$$\mathbb{E}(\hat{C}(\ell)) = \rho e^{-(\lambda_1+\lambda_2)T} \sum_{n=1}^{+\infty} \frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!} \sum_{k=1}^{n} \sum_{i_1=0}^{k-1} \sum_{i_2=k}^{n} \mathbb{P}(\underline{i}_1(k) = i_1 | N = n)\mathbb{P}(\overline{i}_2(k) = i_2 | N = n) f_{k,h}(n, i_1, i_2)$$

$$\begin{aligned} f_{k,h}(n, i_1, i_2) &= \mathbb{E}((\frac{\tau_k}{T} - \frac{\tau_{k-1}}{T})\mathbb{1}_{\left\{h < \frac{\tau_{i_2}}{T} - \frac{\tau_{i_1}}{T}\right\}} | N = n, \underline{i}_1(k) = i_1, \overline{i}_2(k) = i_2) \\ &= \frac{(n-1)! \int_{[0,1]^4} (y-x)\mathbb{1}_{\{h<v-u\}}\mathbb{1}_{\{u<x<y<v\}} u^{i_1-1}(x-u)^{k-i_1-2}(v-y)^{i_2-1-k}(1-v)^{n-i_2-1} \mathrm{d}x\mathrm{d}y\mathrm{d}u\mathrm{d}v}{(i_1-1)!(k-i_1-2)!(i_2-1-k)!(n-i_2-1)!} \end{aligned}$$

where $h = \frac{\ell}{T} \in [0,1[$ and the convention $((-1)!0)^{-1} = 1$ to take into account the boundary cases $i_1 = 0$ and $i_2 = n$. We have used another standard result on the Poisson process which tells that, conditionally to $N_T = n$, the arrival times $t_1, \ldots, t_n$ follow the distribution of the order statistics of the uniform distribution on $[0,T]$ [42]. It means that $(u_0 = 0, u_1 = t_1/T, \ldots, u_{n-1} = t_{n-1}/T, u_n = 1)$ has the following probability density function

$$p(u_0, u_1, \ldots, u_{n-1}, u_n) = (n-1)!\mathbb{1}_{\{0=u_0<u_1<\ldots u_{n-1}<u_n=1\}}$$

which implies that the probability density function of $(u_{i_1}, u_{k-1}, u_k, u_{i_2})$ is

$$p(u_{i_1}, u_{k-1}, u_k, u_{i_2}) = \frac{(n-1)!\mathbb{1}_{\{0\le u_{i_1}<u_{k-1}<u_k<u_{i_2}\le 1\}} u_{i_1}^{i_1-1}(u_{k-1}-u_{i_1})^{k-i_1-2}(u_{i_2}-u_k)^{i_2-1-k}(1-u_{i_2})^{n-i_2-1}\delta(u_0)\delta(u_n-1)}{(i_1-1)!(k-i_1-2)!(i_2-1-k)!(n-i_2-1)!}$$

It is easily seen that

$$\begin{aligned}
\mathbb{P}(\bar{i}_p(k) = i | N = n) &= \mathbb{P}(\text{every jump between } \tau_k \text{ and } \tau_{i-1} \text{ is of type } q \text{ and } \tau_i \text{ is a jump of type } p) \\
&= \frac{\lambda_p}{\lambda_p+\lambda_q}\left(\frac{\lambda_q}{\lambda_p+\lambda_q}\right)^{i-k}\mathbb{1}_{\{k\le i\le n\}} + \delta(i-n)\left(\frac{\lambda_q}{\lambda_p+\lambda_q}\right)^{n-k+1} \\
\mathbb{P}(\underline{i}_p(k) = i | N = n) &= \mathbb{P}(\text{every jump between } \tau_{i+1} \text{ and } \tau_{k-1} \text{ is of type } q \text{ and } \tau_i \text{ is a jump of type } p) \\
&= \frac{\lambda_p}{\lambda_p+\lambda_q}\left(\frac{\lambda_q}{\lambda_p+\lambda_q}\right)^{k-i-1}\mathbb{1}_{\{0\le i\le k-1<n\}} + \delta(i)\left(\frac{\lambda_q}{\lambda_p+\lambda_q}\right)^{k}
\end{aligned}$$

for $p \neq q$. Thus,

$$\begin{aligned}
\mathbb{E}(\hat{C}(\ell)) = {} & \rho e^{-(\lambda_1+\lambda_2)T}\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}\sum_{n=3}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{k=2}^{n-1}\sum_{i_1=1}^{k-1}\sum_{i_2=k}^{n-1}\frac{\lambda_1^{i_2-k}\lambda_2^{k-i_1-1}}{(\lambda_1+\lambda_2)^{i_2-i_1}}f_{k,h}(n,i_1,i_2) \\
& + \rho e^{-(\lambda_1+\lambda_2)T}\frac{\lambda_2}{\lambda_1+\lambda_2}\sum_{n=2}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{k=1}^{n-1}\sum_{i_2=k}^{n-1}\frac{\lambda_1^{i_2-k}\lambda_2^{k-1}}{(\lambda_1+\lambda_2)^{i_2-1}}f_{k,h}(n,0,i_2) \\
& + \rho e^{-(\lambda_1+\lambda_2)T}\frac{\lambda_1}{\lambda_1+\lambda_2}\sum_{n=2}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{k=2}^{n}\sum_{i_1=1}^{k-1}\frac{\lambda_1^{n-k}\lambda_2^{k-i_1-1}}{(\lambda_1+\lambda_2)^{n-i_1-1}}f_{k,h}(n,i_1,n) \\
& + \rho e^{-(\lambda_1+\lambda_2)T}\sum_{n=1}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{k=1}^{n}\frac{\lambda_1^{n-k}\lambda_2^{k-1}}{(\lambda_1+\lambda_2)^{n-1}}f_{k,h}(n,0,n)
\end{aligned}$$

Similarly,

$$\mathbb{E}(\hat{C}(-\ell)) = \rho e^{-(\lambda_1+\lambda_2)T}\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}\sum_{n=3}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{k=2}^{n-1}\sum_{i_1=k}^{n-1}\sum_{i_2=1}^{k-1}\frac{\lambda_1^{k-i_2-1}\lambda_2^{i_1-k}}{(\lambda_1+\lambda_2)^{i_1-i_2}}\tilde{f}_{k,h}(n,i_1,i_2)$$
$$+\rho e^{-(\lambda_1+\lambda_2)T}\frac{\lambda_1}{\lambda_1+\lambda_2}\sum_{n=2}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{k=1}^{n-1}\sum_{i_1=k}^{n-1}\frac{\lambda_1^{k-1}\lambda_2^{i_1-k}}{(\lambda_1+\lambda_2)^{i_1-1}}\tilde{f}_{k,h}(n,i_1,0)$$
$$+\rho e^{-(\lambda_1+\lambda_2)T}\frac{\lambda_2}{\lambda_1+\lambda_2}\sum_{n=2}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{k=2}^{n}\sum_{i_2=1}^{k-1}\frac{\lambda_1^{k-i_2-1}\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^{n-i_2-1}}\tilde{f}_{k,h}(n,n,i_2)$$
$$+\rho e^{-(\lambda_1+\lambda_2)T}\sum_{n=1}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{k=1}^{n}\frac{\lambda_1^{k-1}\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^{n-1}}\tilde{f}_{k,h}(n,n,0)$$

where

$$\begin{aligned}\tilde{f}_{k,h}(n,i_1,i_2) &= \mathbb{E}((\frac{\tau_k}{T}-\frac{\tau_{k-1}}{T})\mathbb{1}_{\left\{h<\frac{\tau_{i_1}}{T}-\frac{\tau_{i_2}}{T}\right\}}|N=n,\overline{i}_1(k)=i_1,\underline{i}_2(k)=i_2)\\ &= \frac{(n-1)!\int_{[0,1]^4}(y-x)\mathbb{1}_{\{h<v-u\}}\mathbb{1}_{\{u<x<y<v\}}u^{i_2-1}(x-u)^{k-i_2-2}(v-y)^{i_1-1-k}(1-v)^{n-i_1-1}\mathrm{d}x\mathrm{d}y\mathrm{d}u\mathrm{d}v}{(i_2-1)!(k-i_2-2)!(i_1-1-k)!(n-i_1-1)!}\\ &= f_{k,h}(n,i_2,i_1)\end{aligned}$$

which leads

$$\mathbb{E}(\hat{C}(-\ell)) = \rho e^{-(\lambda_1+\lambda_2)T}\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}\sum_{n=3}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{k=2}^{n-1}\sum_{i_1=1}^{k-1}\sum_{i_2=k}^{n-1}\frac{\lambda_1^{k-i_1-1}\lambda_2^{i_2-k}}{(\lambda_1+\lambda_2)^{i_2-i_1}}f_{k,h}(n,i_1,i_2)$$
$$+\rho e^{-(\lambda_1+\lambda_2)T}\frac{\lambda_1}{\lambda_1+\lambda_2}\sum_{n=2}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{k=1}^{n-1}\sum_{i_2=k}^{n-1}\frac{\lambda_1^{k-1}\lambda_2^{i_2-k}}{(\lambda_1+\lambda_2)^{i_2-1}}f_{k,h}(n,0,i_2)$$
$$+\rho e^{-(\lambda_1+\lambda_2)T}\frac{\lambda_2}{\lambda_1+\lambda_2}\sum_{n=2}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{k=2}^{n}\sum_{i_1=1}^{k-1}\frac{\lambda_1^{k-i_1-1}\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^{n-i_1-1}}f_{k,h}(n,i_1,n)$$
$$+\rho e^{-(\lambda_1+\lambda_2)T}\sum_{n=1}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{k=1}^{n}\frac{\lambda_1^{k-1}\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^{n-1}}f_{k,h}(n,0,n)$$

We now need to compute the function $f_{k,h}(n,i_1,i_2)$. After integrating over y, we have

$$\begin{aligned}f_{k,h}(n,i_1,i_2) &= \frac{(n-1)!\int_{[0,1]^3}\mathbb{1}_{\{h<v-u\}}\mathbb{1}_{\{u<x<v\}}u^{i_1-1}(x-u)^{k-i_1-2}(v-x)^{i_2-k+1}(1-v)^{n-i_2-1}\mathrm{d}x\mathrm{d}u\mathrm{d}v}{(i_1-1)!(k-i_1-2)!(i_2+1-k)!(n-i_2-1)!}\\ &= \frac{(n-1)!\int_{[0,1]^2}(\int_u^v(x-u)^{k-i_1-2}(v-x)^{i_2-k+1}\mathrm{d}x)\mathbb{1}_{\{h<v-u\}}\mathbb{1}_{\{u<v\}}u^{i_1-1}(1-v)^{n-i_2-1}\mathrm{d}u\mathrm{d}v}{(i_1-1)!(k-i_1-2)!(i_2+1-k)!(n-i_2-1)!}\end{aligned}$$

We use the following lemma that can be proven by successive integration by parts.

Lemma 1. *Let $k\in\mathbb{N}^*$, $(i_1,i_2)\in\mathbb{N}^2$ such that $i_1+1<k\leq i_2$. Let $(u,v)\in\mathbb{R}^2$ such that $u\leq v$. Then,*

$$\int_u^v (x-u)^{k-i_1-2}(v-x)^{i_2-k+1}\mathrm{d}x = \frac{(k-i_1-2)!}{\prod_{p=0}^{k-i_1-2}(i_2-k+2+p)}(v-u)^{i_2-i_1}$$

Using this lemma, we get

$$\begin{aligned} f_{k,h}(n,i_1,i_2) &= \frac{(n-1)!\int_{[0,1]^2}\mathbb{1}_{\{h<v-u\}}\mathbb{1}_{\{u<v\}}u^{i_1-1}(v-u)^{i_2-i_1}(1-v)^{n-i_2-1}\mathrm{d}u\mathrm{d}v}{(i_1-1)!(i_2+1-k)!(n-i_2-1)!\prod_{p=0}^{k-i_1-2}(i_2-k+2+p)} \\ &= \frac{(n-1)!\int_h^1(\int_0^{v-h}u^{i_1-1}(v-u)^{i_2-i_1}\mathrm{d}u)(1-v)^{n-i_2-1}\mathrm{d}v}{(i_1-1)!(i_2-i_1)!(n-i_2-1)!} \end{aligned}$$

We now use the following lemma, that can also be proven by successive integration by parts.

Lemma 2. *Let $(i_1,i_2)\in\mathbb{N}^2$ such that $i_1<i_2$. Let $(h,v)\in\mathbb{R}^2$ such that $0\le h<v$. Then,*

$$\int_0^{v-h}u^{i_1-1}(v-u)^{i_2-i_1}\mathrm{d}u = \frac{(i_1-1)!}{\prod_{p=1}^{i_1}(i_2-i_1+p)}v^{i_2} - \sum_{p=1}^{i_1}\frac{\prod_{m=1}^{p-1}(i_1-m)}{\prod_{m=1}^{p}(i_2-i_1+m)}h^{i_2-i_1+p}(v-h)^{i_1-p}$$

We now have

$$\begin{aligned} f_{k,h}(n,i_1,i_2) = &-\frac{(n-1)!}{(i_1-1)!(i_2-i_1)!(n-i_2-1)!}\sum_{p=1}^{i_1}\frac{\prod_{m=1}^{p-1}(i_1-m)}{\prod_{m=1}^{p}(i_2-i_1+m)}h^{i_2-i_1+p}\int_h^1(1-v)^{n-i_2-1}(v-h)^{i_1-p}\mathrm{d}v \\ &+\frac{(n-1)!}{i_2!(n-i_2-1)!}\int_h^1(1-v)^{n-i_2-1}v^{i_2}\mathrm{d}v \end{aligned}$$

We need to use the two following lemmas. The first one can be proven by the change of variable $u=\frac{v-h}{1-h}$ and the second one by succcessive integration by parts.

Lemma 3. *Let $(p,i_1,i_2,n)\in\mathbb{N}^4$ such that $0<p\le i_1<i_2<n$. Let $h\in\mathbb{R}$ such that $h<1$. Then,*

$$\int_h^1(1-v)^{n-i_2-1}(v-h)^{i_1-p}\mathrm{d}v = \frac{(n-i_2-1)!(i_1-p)!}{(n-p-(i_2-i_1))!}(1-h)^{n-p-(i_2-i_1)}$$

Lemma 4. *Let $(i_2,n)\in\mathbb{N}^2$ such that $i_2<n$. Let $h\in\mathbb{R}$ such that $h<1$. Then,*

$$\int_h^1(1-v)^{n-i_2-1}v^{i_2}\mathrm{d}v = \sum_{p=0}^{i_2}\frac{\prod_{m=1}^{p}(i_2+1-m)}{\prod_{m=0}^{p}(n-i_2+m)}h^{i_2-p}(1-h)^{n-i_2+p}$$

As a result, we have for $h\in[0,1[$

$$
\begin{aligned}
f_{k,h}(n,i_1,i_2) =& -(n-1)!\sum_{p=1}^{i_1}\frac{h^{i_2-i_1+p}(1-h)^{n-p-(i_2-i_1)}}{(n-p-(i_2-i_1))!(i_2-i_1+p)!}\\
&+(n-1)!\sum_{p=0}^{i_2}\frac{h^{i_2-p}(1-h)^{n-i_2+p}}{(i_2-p)!(n-i_2+p)!}\\
=&(n-1)!\sum_{p=i_1}^{i_2}\frac{h^{i_2-p}(1-h)^{n-i_2+p}}{(i_2-p)!(n-i_2+p)!}\\
=&(n-1)!\sum_{p=0}^{i_2-i_1}\frac{h^{i_2-i_1-p}(1-h)^{n-(i_2-i_1)+p}}{(i_2-i_1-p)!(n-(i_2-i_1)+p)!}\\
=&g_h(n,i_2-i_1)
\end{aligned}
$$

where $g_h(n,i)=(n-1)!\sum_{p=0}^{i}\frac{h^{i-p}(1-h)^{n-i+p}}{(i-p)!(n-i+p)!}=(n-1)!\sum_{k=0}^{i}\frac{h^k(1-h)^{n-k}}{k!(n-k)!}$. Note that $g_h(n,n)=\frac{1}{n}\ \forall h$. Therefore, the average covariance reads

$$
\begin{aligned}
&\mathbb{E}(\hat{C}(\ell))=\rho e^{-(\lambda_1+\lambda_2)T}(S_1(\ell)+S_2(\ell)+S_3(\ell)+S_4(\ell))\\
&S_1(\ell)=\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}\sum_{n=3}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{k=2}^{n-1}\sum_{i_1=1}^{k-1}\sum_{i_2=k}^{n-1}\frac{\lambda_1^{i_2-k}\lambda_2^{k-i_1-1}}{(\lambda_1+\lambda_2)^{i_2-i_1}}g_h(n,i_2-i_1)\\
&S_2(\ell)=\frac{\lambda_2}{\lambda_1+\lambda_2}\sum_{n=2}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{k=1}^{n-1}\sum_{i_2=k}^{n-1}\frac{\lambda_1^{i_2-k}\lambda_2^{k-1}}{(\lambda_1+\lambda_2)^{i_2-1}}g_h(n,i_2)\\
&S_3(\ell)=\frac{\lambda_1}{\lambda_1+\lambda_2}\sum_{n=2}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{k=2}^{n}\sum_{i_1=1}^{k-1}\frac{\lambda_1^{n-k}\lambda_2^{k-i_1-1}}{(\lambda_1+\lambda_2)^{n-i_1-1}}g_h(n,n-i_1)\\
&S_4(\ell)=\sum_{n=1}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{n!}\sum_{k=1}^{n}\frac{\lambda_1^{n-k}\lambda_2^{k-1}}{(\lambda_1+\lambda_2)^{n-1}}
\end{aligned}
$$

Let us consider the case $\lambda_1\neq\lambda_2$ first. Then,

$$
\begin{aligned}
S_1(\ell)&=\frac{\lambda_1\lambda_2}{\lambda_1+\lambda_2}\sum_{n=3}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{i_1=1}^{n-2}\sum_{i_2=i_1+1}^{n-1}\frac{\lambda_1^{i_2}\lambda_2^{-i_1-1}}{(\lambda_1+\lambda_2)^{i_2-i_1}}g_h(n,i_2-i_1)\sum_{k=i_1+1}^{i_2}(\frac{\lambda_2}{\lambda_1})^k\\
&=\frac{\lambda_1\lambda_2}{\lambda_1^2-\lambda_2^2}\sum_{n=3}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{i=1}^{n-2}\sum_{j=1}^{n-1-i}\frac{\lambda_1^j-\lambda_2^j}{(\lambda_1+\lambda_2)^j}g_h(n,j)
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
S_1(-\ell)&=\frac{\lambda_1\lambda_2}{\lambda_2^2-\lambda_1^2}\sum_{n=3}^{+\infty}\frac{((\lambda_1+\lambda_2)T)^{n-1}}{(n-1)!}\sum_{i=1}^{n-2}\sum_{j=1}^{n-1-i}\frac{\lambda_2^j-\lambda_1^j}{(\lambda_1+\lambda_2)^j}g_h(n,j)\\
&=S_1(\ell)
\end{aligned}
$$

and more generally we have $S_i(-\ell) = S_i(\ell)$ for $i \in \{1,2,3,4\}$. Therefore the covariance function is symmetric on average, *i.e.* $\mathbb{E}(\hat{C}(-\ell)) = \mathbb{E}(\hat{C}(\ell))$. Let us carry on the computations,

$$
\begin{aligned}
S_1(\ell) &= \frac{\lambda_1\lambda_2}{\lambda_1^2-\lambda_2^2}\sum_{n=3}^{+\infty}((\lambda_1+\lambda_2)T)^{n-1}\sum_{i=1}^{n-2}\sum_{k=1}^{n-1-i}\frac{h^k(1-h)^{n-k}}{k!(n-k)!}\sum_{j=k}^{n-1-i}\frac{\lambda_1^j-\lambda_2^j}{(\lambda_1+\lambda_2)^j} + \frac{(1-h)^n}{n!}\sum_{j=1}^{n-1-i}\frac{\lambda_1^j-\lambda_2^j}{(\lambda_1+\lambda_2)^j} \\
&= \frac{\lambda_1}{\lambda_1-\lambda_2}\sum_{n=3}^{+\infty}(\lambda_1+\lambda_2)T)^{n-1}\sum_{k=1}^{n-2}\frac{h^k(1-h)^{n-k}}{k!(n-k)!}(n-1-k)(\frac{\lambda_1}{\lambda_1+\lambda_2})^k \\
&\quad - \frac{\lambda_1}{\lambda_1-\lambda_2}\sum_{n=3}^{+\infty}(\lambda_1+\lambda_2)T)^{n-1}(\frac{\lambda_1}{\lambda_1+\lambda_2})^n\sum_{k=1}^{n-2}\frac{h^k(1-h)^{n-k}}{k!(n-k)!}\sum_{i=1}^{n-1-k}(\frac{\lambda_1+\lambda_2}{\lambda_1})^i \\
&\quad + \frac{\lambda_1^2}{\lambda_1^2-\lambda_2^2}\sum_{n=3}^{+\infty}(\lambda_1+\lambda_2)T)^{n-1}\frac{(1-h)^n}{n!}(n-2) \\
&\quad - \frac{\lambda_1}{\lambda_1-\lambda_2}\sum_{n=3}^{+\infty}(\lambda_1+\lambda_2)T)^{n-1}\frac{(1-h)^n}{n!}(\frac{\lambda_1}{\lambda_1+\lambda_2})^n\sum_{i=1}^{n-2}(\frac{\lambda_1+\lambda_2}{\lambda_1})^i \\
&\quad - \frac{\lambda_2}{\lambda_1-\lambda_2}\sum_{n=3}^{+\infty}(\lambda_1+\lambda_2)T)^{n-1}\sum_{k=1}^{n-2}\frac{h^k(1-h)^{n-k}}{k!(n-k)!}(n-1-k)(\frac{\lambda_2}{\lambda_1+\lambda_2})^k \\
&\quad + \frac{\lambda_2}{\lambda_1-\lambda_2}\sum_{n=3}^{+\infty}(\lambda_1+\lambda_2)T)^{n-1}(\frac{\lambda_2}{\lambda_1+\lambda_2})^n\sum_{k=1}^{n-2}\frac{h^k(1-h)^{n-k}}{k!(n-k)!}\sum_{i=1}^{n-1-k}(\frac{\lambda_1+\lambda_2}{\lambda_2})^i \\
&\quad - \frac{\lambda_2^2}{\lambda_1^2-\lambda_2^2}\sum_{n=3}^{+\infty}(\lambda_1+\lambda_2)T)^{n-1}\frac{(1-h)^n}{n!}(n-2) \\
&\quad + \frac{\lambda_2}{\lambda_1-\lambda_2}\sum_{n=3}^{+\infty}((\lambda_1+\lambda_2)T)^{n-1}\frac{(1-h)^n}{n!}(\frac{\lambda_2}{\lambda_1+\lambda_2})^n\sum_{i=1}^{n-2}(\frac{\lambda_1+\lambda_2}{\lambda_2})^i
\end{aligned}
$$

$$
\begin{aligned}
S_1(\ell) &= \frac{\lambda_1}{\lambda_2(\lambda_1-\lambda_2)T}(e^{\lambda_1\ell}(e^{\lambda_1(T-\ell)}-1-\lambda_1(T-\ell)) - \frac{1}{2}(\lambda_1(T-\ell))^2) \\
&\quad - \frac{\lambda_2}{\lambda_1(\lambda_1-\lambda_2)T}(e^{\lambda_2\ell}(e^{\lambda_2(T-\ell)}-1-\lambda_2(T-\ell)) - \frac{1}{2}(\lambda_2(T-\ell))^2) \\
&\quad + \frac{\lambda_1(e^{\lambda_1\ell}-1)-\lambda_2(e^{\lambda_2\ell}-1)}{(\lambda_1^2-\lambda_2^2)T}(1+e^{(\lambda_1+\lambda_2)(T-\ell)}((\lambda_1+\lambda_2)(T-\ell)-1)) \\
&\quad + \frac{\frac{\lambda_2^2}{\lambda_1}(e^{\lambda_2\ell}-1)-\frac{\lambda_1^2}{\lambda_2}(e^{\lambda_1\ell}-1)}{(\lambda_1^2-\lambda_2^2)T}(e^{(\lambda_1+\lambda_2)(T-\ell)}-1-(\lambda_1+\lambda_2)(T-\ell)) \\
&\quad - \frac{\lambda_1^2+\lambda_2^2}{\lambda_1\lambda_2(\lambda_1+\lambda_2)T}(e^{(\lambda_1+\lambda_2)(T-\ell)}-1-(\lambda_1+\lambda_2)(T-\ell)-\frac{((\lambda_1+\lambda_2)(T-\ell))^2}{2}) \\
&\quad + \frac{(e^{(\lambda_1+\lambda_2)(T-\ell)}((\lambda_1+\lambda_2)(T-\ell)-2)+(\lambda_1+\lambda_2)(T-\ell)+2)}{(\lambda_1+\lambda_2)T}
\end{aligned}
$$

Similarly, one can prove that

$$
\begin{aligned}
S_2(\ell) &= \frac{(e^{(\lambda_1+\lambda_2)(T-\ell)}-1)}{(\lambda_1-\lambda_2)T}(e^{\lambda_1\ell}-1-\frac{\lambda_2}{\lambda_1}(e^{\lambda_2\ell}-1)) \\
&\quad - \frac{(e^{\lambda_1(T-\ell)}-1)(e^{\lambda_1\ell}-1)+e^{\lambda_1(T-\ell)}-1-\lambda_1(T-\ell)-\frac{\lambda_2}{\lambda_1}((e^{\lambda_2(T-\ell)}-1)(e^{\lambda_2\ell}-1)+e^{\lambda_2(T-\ell)}-1-\lambda_2(T-\ell))}{(\lambda_1-\lambda_2)T} \\
&\quad + \frac{(e^{(\lambda_1+\lambda_2)(T-\ell)}-1-(\lambda_1+\lambda_2)(T-\ell))}{\lambda_1 T} \\
S_3(\ell) &= \frac{\lambda_1}{\lambda_2}S_2(\ell) \\
S_4(\ell) &= \frac{e^{\lambda_1 T}-e^{\lambda_2 T}}{(\lambda_1-\lambda_2)T}
\end{aligned}
$$

The case $\lambda_1 = \lambda_2 = \lambda$ coincides with the limit $\lambda_2 \to \lambda_1$. In this case, we have

$$
\begin{aligned}
\mathbb{E}(\hat{C}(\ell)) &= \rho e^{-2\lambda T}(S_1(\ell)+2S_2(\ell)+S_4(\ell)) \\
S_1(\ell) &= e^{\lambda T}(\frac{2}{\lambda T}+1)-e^{\lambda\ell}(\frac{2}{\lambda T}+\frac{\ell}{T}+(1-\frac{\ell}{T})(3+\lambda\ell))-\frac{2\lambda(T-\ell)^2}{T} \\
&\quad + (1+e^{2\lambda(T-\ell)}(2\lambda(T-\ell)-1))(\frac{e^{\lambda\ell}(1+\lambda\ell)-1}{2\lambda T}) \\
&\quad + (e^{2\lambda(T-\ell)}-1-2\lambda(T-\ell))(\frac{3-e^{\lambda\ell}(3+\lambda\ell)}{2\lambda T}) \\
&\quad - \frac{(e^{2\lambda(T-\ell)}-1-2\lambda(T-\ell)-2(\lambda(T-\ell))^2)}{\lambda T} \\
&\quad + \frac{e^{2\lambda(T-\ell)}(\lambda(T-\ell)-1)+\lambda(T-\ell)+1}{\lambda T} \\
S_2(\ell) &= \frac{(e^{2\lambda(T-\ell)}-1)}{T}(e^{\lambda\ell}(\ell+\frac{1}{\lambda})-\frac{1}{\lambda})+\frac{e^{2\lambda(T-\ell)}-1-2\lambda(T-\ell)}{\lambda T} \\
&\quad - e^{\lambda T}(1+\frac{1}{\lambda T})+\frac{e^{\lambda\ell}}{T}(\frac{1}{\lambda}+\ell)+2(1-\frac{\ell}{T}) \\
S_4(\ell) &= e^{\lambda T}
\end{aligned}
$$

3.8 Appendix: Lead/lag response functions

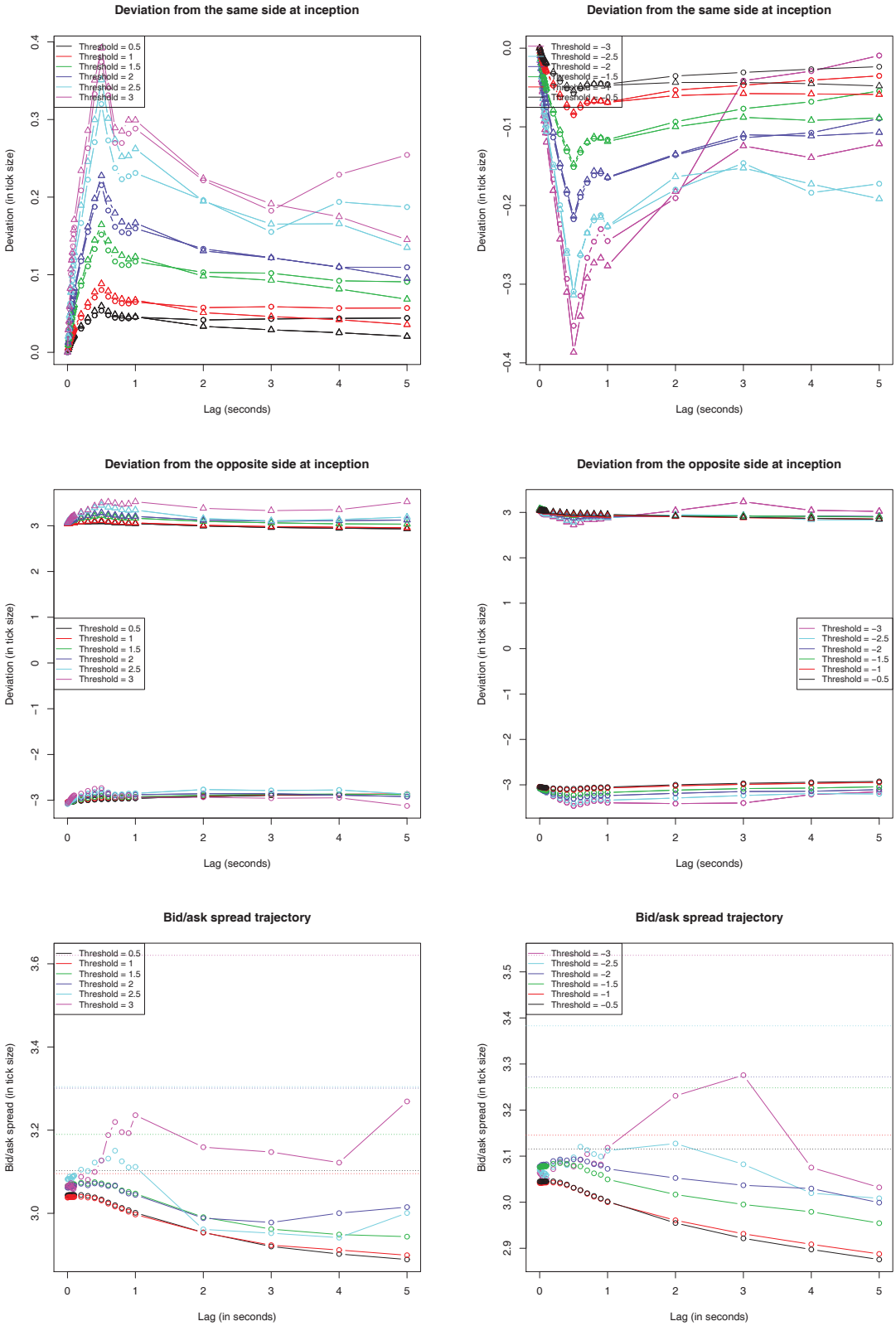

Figure 3.17: Same as figure 3.12 for FDX/DTEGn.DE.

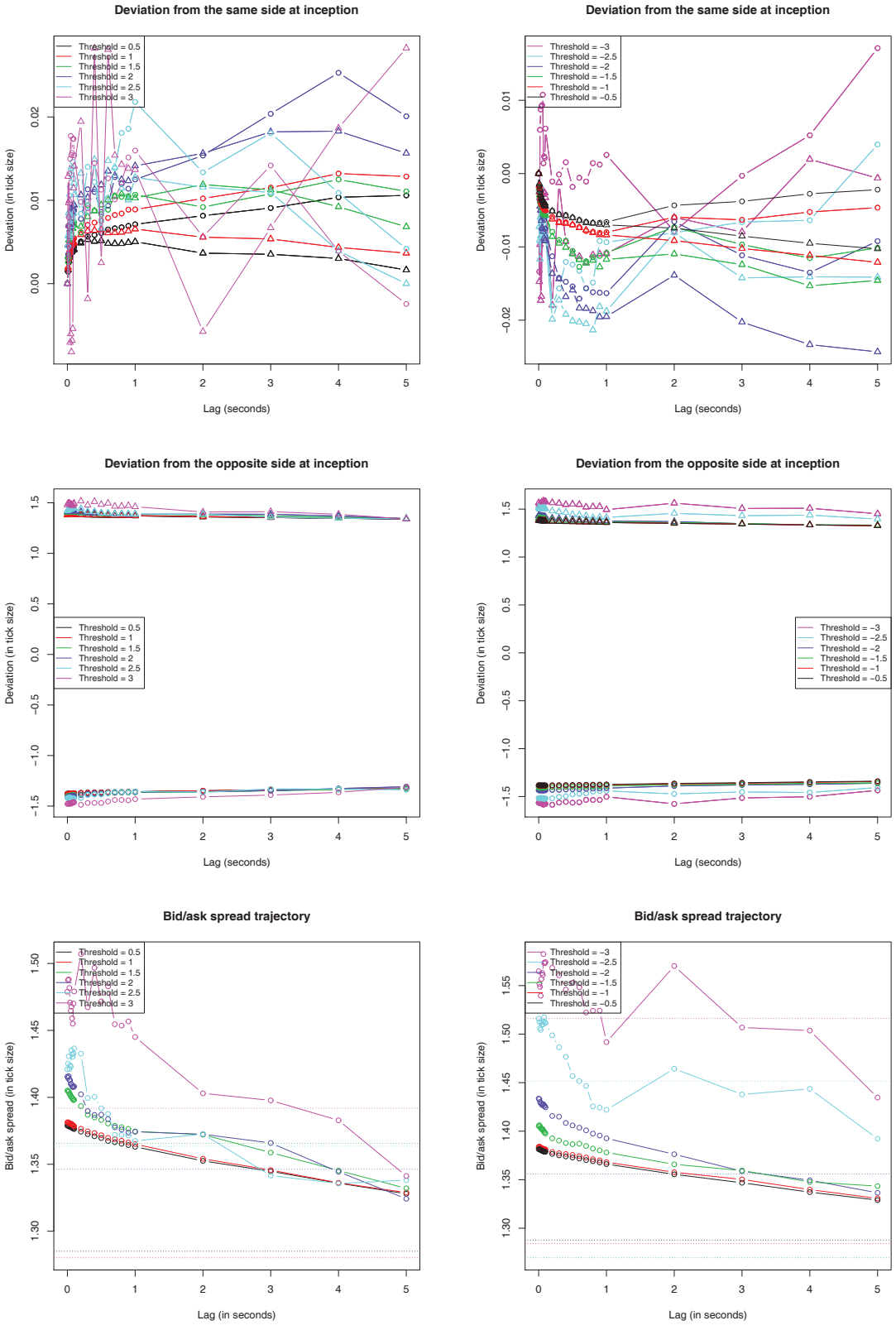

Figure 3.18: Same as figure 3.12 for FFI/VOD.L.

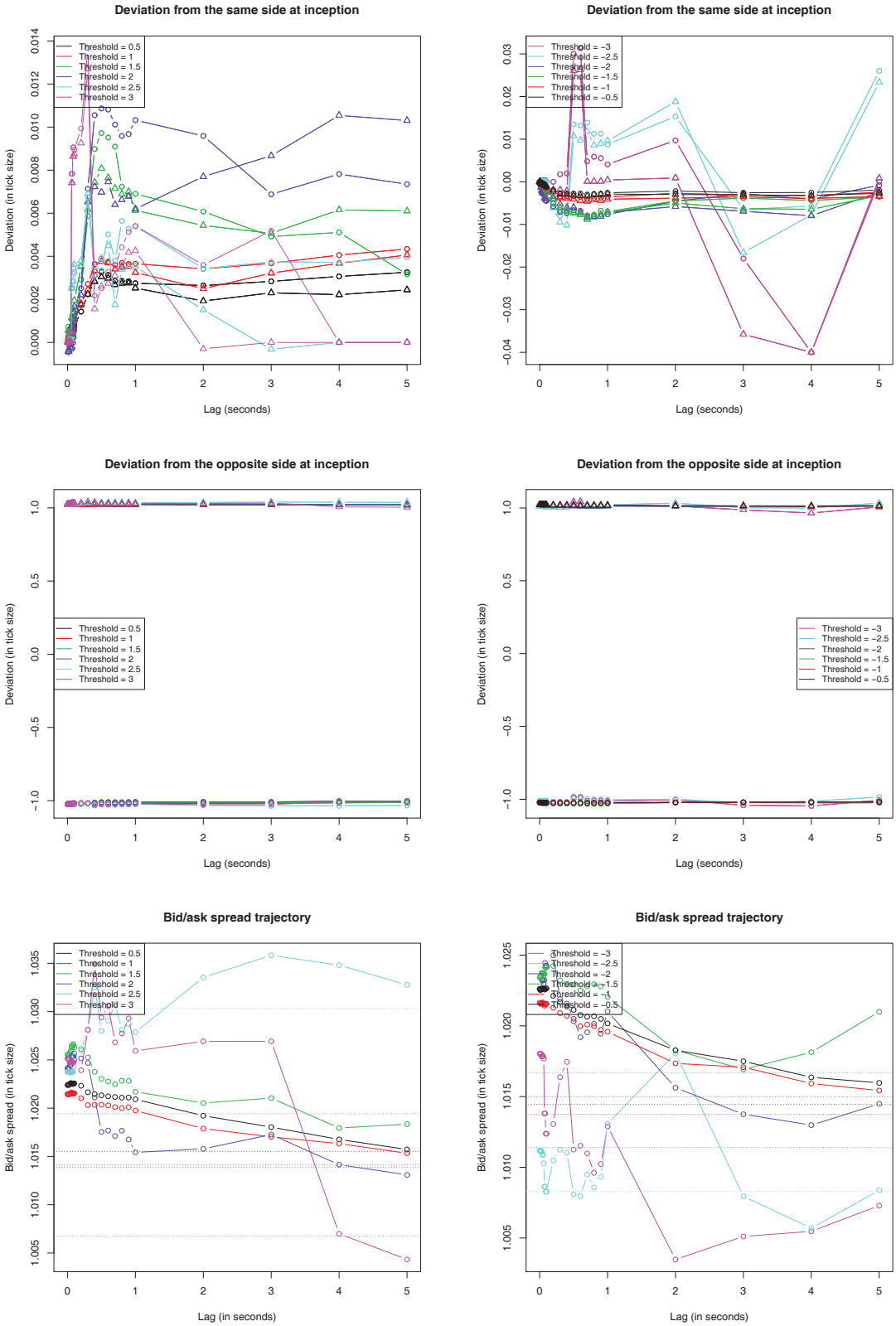

Figure 3.19: Same as figure 3.12 for FSMI/NESN.VX.

Chapter 4

Intraday Correlation Pattern

4.1 Introduction

Trading activity on financial markets is well known to display intraday seasonality [1, 7, 9, 32]. Seasonality appears on many quantities of interest such as volatility, transaction volume, bid/ask spread, market and limit orders arrivals *etc*. Figure 4.1 plots the intraday pattern of the traded volume on the CAC40 future[1] (Reuters Instrument Code: FCE). We observe an asymmetric U-shape during the trading session of the underlying market: big volumes at the opening (9:00)[2], then it decreases until lunch time where it reaches its mimimum, after 13:30 it starts to rise again, peaks at 14:30 (announcement of US macroeconomic figures), it changes of regime at 15:30 (NYSE and NASDAQ openings), peaks at 16:00 (other US macroeconomic figures) and finally rallies at the close of the market. The trading activity outside the underlying market session is fairly small, with the exception of the futures opening session at 8:00. The intraday seasonality seems to be highly connected with both human activity and arrival of significant information. We believe that huge volumes at the opening are due to both the adjustment of positions taken the day before and the discovery of corporate news and figures (earnings, dividends, *etc*) before the market opens, while the peak at the close is made by large investors such as derivatives traders and portfolio managers who benchmark on closing prices.

The intraday correlation between assets is a key quantity for agents acting at high frequency such as market-makers, statistical arbitrage traders and algorithmic brokers in order to manage multi-asset strategies. However, the study of the intraday profile of correlation has not been widely adressed, with the slight exceptions of [8, 22]. In [8], the authors compute correlation on 5-minute bins using 5-minute price returns, so they only have one observation per time slice and per day. As a result, they resort to a dataset that spans 10 years, a very large period of time which is subject to changes in regulation of markets and strong macroeconomic fluctuations. In [22], the focus is on correlation between the S&P500 ETF[3] and business sectors ETFs. Moreover, the binning is done on an hourly basis, which seems too rough for us to account for the speed-up of financial activity in recent years.

In this paper, we investigate the shape of the intraday correlation profile with tick-by-tick data. We use the Hayashi-Yoshida correlation estimator [61] in order to make use of all the available transaction data. It allows us to consider short time periods such as three months of trading and to bin results into 5-minute time slices. As a result, we clearly identity times of the day where there correlation shows a specific behavior. We also consider several stock exchanges: Paris, London, New-York and Tokyo. Besides, we compute the intraday profile of idiosyncratic correlations using a CAPM (Capital Asset Pricing Model) approach [93].

[1] The CAC40 is the largest French equity index.

[2] In the following, we will always use Paris Time.

[3] An ETF (Exchange-Traded-Fund) is an investment fund traded on stock exchanges that tracks the performance of a given underlying.

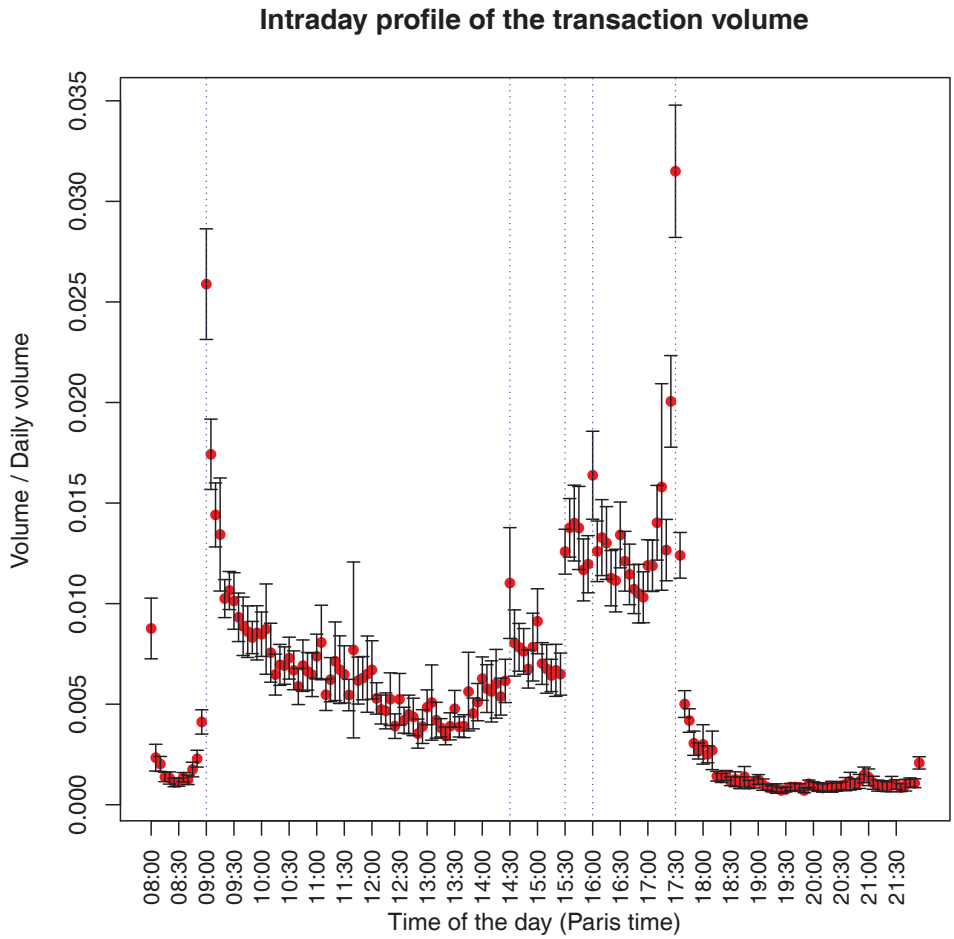

Figure 4.1: Intraday profile of transaction volume of FCE, 2010/03/01 – 2010/05/31. The y-axis represents the proportion of the daily traded volume for a given 5-minute bin. Error bars indicate 95% Gaussian confidence intervals. Blue dotted lines are drawn at 14:30 and 16:00 (US macroeconomic figures annoucements) and 15:30 (NYSE and NASDAQ openings)

Finally, we examine how Hawkes processes with time-dependent parameters can reproduce the observed intraday correlation profile.

This paper is organized as follows. Section 4.2 introduces the dataset and provides basic but insightful statistics on the assets under focus. Section 4.3 describes the methodology used to measure the intraday correlation pattern. Section 4.4 provides empirical results. Section 4.5 elaborates on Hawkes processes with time dependent parameters and their ability to capture the intraday correlation profile. Finally, section 4.6 concludes and announces further research.

4.2 Data description and summary statistics

We have access to the Thomson Reuters Tick History database (see section 1.3) which provides tick-by-tick data on many financial assets (equities, fixed income, forex, futures, commodities, *etc*). Three levels of data are available:

- trades files: each transaction price and quantity timestamped up to the millisecond
- quotes files: each quote (best bid and ask) price or quantity modification timestamped up to the millisecond
- order book files: each limit price or quantity modification timestamped up to the millisecond, up to a given depth, typically ten limits on each side of the order book

Throughout this study, we will only use trades and quotes files. We will sample quotes on a trading time basis so that for each trade we have access to the quotes right before this trade. We do this because the returns we will consider are the returns of the midquote in order to get rid of the bid/ask bounce. Therefore, it might sound more natural to consider every best quote modification since events other than trades, such as limit orders placed in the bid/ask spread, cancellation orders for the whole quantity available at the best limit and trades-though, can affect the midquote. However, we favor trading time sampling over best quotes sampling because in our opinion trades represent more significant events than quotes changes. Indeed, only trades involve money exchanges.

When a trade walks the order book up or down by hitting consecutive limit prices, it is recorded as a sequence of trades with the same timestamp but with prices and quantities corresponding to each limit hit. For instance, assume that the best ask offers 100 shares at price 10 and 200 shares at price 10.01, and that a buy trade arrives for 150 shares. This is recorded as two lines in the trades file with the same timestamp, the first line being 100 shares at 10 at 50 shares at 10.01. As a pre-processing step, we aggregate identical timestamps in trades files by replacing the price by the volume weighted average price (VWAP) over the whole transaction and the volume by the sum of all quantities consumed. In the previous example, the trade price will thus be $(100*10+50*10.01)/(100+50) = 10.00333$ and the trade quantity $100+50 = 150$.

We only consider equities and equity index futures for our empirical study. The futures are nearby-maturity futures and are rolled the day before the expiration date. We focus on four universes of stocks

- the 40 assets composing the French index CAC40 on 2010/03/01 (CAC universe)
- the 30 most liquid[4] assets from the English index Footsie100 on 2010/03/01 (FTSE universe)
- the 30 assets composing the US index DJIA on 2010/03/01, plus major financial and IT stocks, 40 US stocks altogether (NY universe)
- the 30 assets[5] composing the Japanese index TopixCore30 on 2010/03/01 (TOPIX universe)

The description of the Paris' stock universe is provided in table 4.1 and the three others are displayed in appendix 4.7. The time period is 2010/03/01-2010/05/31.

[4] We use here the average daily number of trades as a rough criterion for liquidity.
[5] In fact, we only consider 29 assets because we couldn't get data for 4974.T (Nintendo Co Ltd).

Table 4.1: Description of the scope of assets (CAC universe).

RIC	Description	Exchange	Trading hours (Paris time)
ACCP.PA	Accor	NYSE Euronext Paris	09:00-17:30
AIRP.PA	Air Liquide	NYSE Euronext Paris	09:00-17:30
ALSO.PA	Alstom	NYSE Euronext Paris	09:00-17:30
ALUA.PA	Alcatel Lucent	NYSE Euronext Paris	09:00-17:30
AXAF.PA	Axa	NYSE Euronext Paris	09:00-17:30
BNPP.PA	BNP Paribas	NYSE Euronext Paris	09:00-17:30
BOUY.PA	Bouygues	NYSE Euronext Paris	09:00-17:30
CAGR.PA	Crédit Agricole	NYSE Euronext Paris	09:00-17:30
CAPP.PA	Cap Gemini	NYSE Euronext Paris	09:00-17:30
CARR.PA	Carrefour	NYSE Euronext Paris	09:00-17:30
DANO.PA	Danone	NYSE Euronext Paris	09:00-17:30
DEXI.BR	Dexia	NYSE Euronext Brussels	09:00-17:30
EAD.PA	EADS	NYSE Euronext Paris	09:00-17:30
EDF.PA	EDF	NYSE Euronext Paris	09:00-17:30
ESSI.PA	Essilor	NYSE Euronext Paris	09:00-17:30
FTE.PA	France Télécom	NYSE Euronext Paris	09:00-17:30
GSZ.PA	GDF Suez	NYSE Euronext Paris	09:00-17:30
ISPA.AS	Arcelor Mittal	NYSE Euronext Amsterdam	09:00-17:30
LAFP.PA	Lafarge	NYSE Euronext Paris	09:00-17:30
LAGA.PA	Lagardère	NYSE Euronext Paris	09:00-17:30
LVMH.PA	LVMH	NYSE Euronext Paris	09:00-17:30
MICP.PA	Michelin	NYSE Euronext Paris	09:00-17:30
OREP.PA	L'Oréal	NYSE Euronext Paris	09:00-17:30
PERP.PA	Pernod Ricard	NYSE Euronext Paris	09:00-17:30
PEUP.PA	Peugeot	NYSE Euronext Paris	09:00-17:30
PRTP.PA	PPR	NYSE Euronext Paris	09:00-17:30
RENA.PA	Renault	NYSE Euronext Paris	09:00-17:30
SASY.PA	Sanofi Aventis	NYSE Euronext Paris	09:00-17:30
SCHN.PA	Schneider Electric	NYSE Euronext Paris	09:00-17:30
SEVI.PA	Suez Environnement	NYSE Euronext Paris	09:00-17:30
SGEF.PA	Vinci	NYSE Euronext Paris	09:00-17:30
SGOB.PA	Saint-Gobain	NYSE Euronext Paris	09:00-17:30
SOGN.PA	Société Générale	NYSE Euronext Paris	09:00-17:30
STM.PA	StMicroelectronics	NYSE Euronext Paris	09:00-17:30
TECF.PA	Technip	NYSE Euronext Paris	09:00-17:30
TOTF.PA	Total	NYSE Euronext Paris	09:00-17:30
UNBP.PA	Unibail-Rodamco	NYSE Euronext Paris	09:00-17:30
VIE.PA	Veolia Environnement	NYSE Euronext Paris	09:00-17:30
VIV.PA	Vivendi	NYSE Euronext Paris	09:00-17:30
VLLP.PA	Vallourec	NYSE Euronext Paris	09:00-17:30
FCE	CAC40 future	NYSE Liffe Paris	08:00-22:00

Table 4.2 gives some insight into the liquidity of each of these assets. It displays the following summary statistics:

- the average duration between two consecutive trades $\langle \Delta t \rangle$
- the average tick size δ in percentage of the midquote $\langle \delta / m \rangle$
- the average bid/ask spread as a multiple of the tick size $\langle s \rangle / \delta$
- the frequency of unit bid/ask spread $\left\langle \mathbb{1}_{\{s=\delta\}} \right\rangle$

- the frequency of trades hitting more than the best limit available, also known as trades-through[6] $\left\langle \mathbb{1}_{\{\text{trade through}\}} \right\rangle$ [6]
- a proxy for the daily volatility expressed in tick size : $\langle |\Delta m| \rangle / \delta$, where Δm is the midquote variation between two consecutive trades
- the average turnover per trade $\langle P_{\text{trade}} V_{\text{trade}} \rangle$

Every average is computed independently on a daily basis, and then averaged over all days available : $\langle x \rangle = \frac{1}{n_{\text{days}}} \sum_{d=1}^{n_{\text{days}}} \frac{\sum_{i=1}^{n_d} x_{i,d}}{n_d}$, where n_d is the number of observations on day d and $x_{i,d}$ is the i^{th} observation on day d.

Table 4.2: Summary statistics on the scope of assets (CAC universe).

RIC	$\langle \Delta t \rangle$ (sec)	$\langle \delta/m \rangle$ (bp)	$\langle s \rangle / \delta$	$\left\langle \mathbb{1}_{\{s=\delta\}} \right\rangle$ (%)	$\left\langle \mathbb{1}_{\{\text{trade through}\}} \right\rangle$ (%)	$\langle \lvert \Delta m \rvert \rangle / \delta$	$\langle P_{\text{trade}} V_{\text{trade}} \rangle$ (EUR$\times 10^3$)
ACCP.PA	12.563	1.22	4.09	17	5	1.19	11
AIRP.PA	6.821	1.15	2.90	16	5	0.89	13
ALSO.PA	6.161	1.11	3.65	20	7	1.02	13
ALUA.PA	7.264	4.34	1.66	56	4	0.38	12
AXAF.PA	5.109	3.30	1.43	68	3	0.38	17
BNPP.PA	3.056	1.59	2.27	46	6	0.65	19
BOUY.PA	8.902	1.36	2.89	28	5	0.99	12
CAGR.PA	5.400	3.35	2.46	58	4	0.70	13
CAPP.PA	9.659	1.35	3.38	22	5	1.03	12
CARR.PA	7.071	1.39	2.42	35	5	0.77	16
DANO.PA	5.724	1.15	2.54	36	5	0.73	15
DEXI.BR	20.070	2.44	5.39	8	8	1.36	6
EAD.PA	11.189	3.34	1.79	50	4	0.44	12
EDF.PA	7.553	1.29	2.71	31	5	0.79	12
ESSI.PA	13.282	1.08	3.06	27	5	0.79	10
FTE.PA	5.952	2.97	1.25	81	2	0.24	20
GSZ.PA	4.989	1.85	1.86	50	4	0.48	15
ISPA.AS	2.760	1.68	2.11	39	6	0.63	22
LAFP.PA	8.330	1.60	3.30	28	5	0.99	14
LAGA.PA	16.236	1.74	3.08	23	4	0.87	8
LVMH.PA	5.747	1.16	2.78	20	6	0.90	16
MICP.PA	8.503	1.84	2.71	29	5	0.77	13
OREP.PA	9.151	1.28	2.79	20	5	0.85	16
PERP.PA	11.329	1.61	2.31	36	4	0.67	13
PEUP.PA	9.248	2.36	2.90	23	5	0.74	12
PRTP.PA	12.371	2.39	3.63	33	4	0.95	17
RENA.PA	5.371	1.51	3.32	22	6	1.00	14
SASY.PA	4.817	1.71	2.07	47	4	0.51	20
SCHN.PA	6.152	1.19	3.08	16	5	0.95	15
SEVI.PA	20.129	3.10	1.91	45	4	0.49	8
SGEF.PA	5.188	1.22	2.77	29	6	0.82	13
SGOB.PA	6.011	1.41	2.95	25	6	0.93	14
SOGN.PA	3.082	1.2	3.04	29	7	0.97	15
STM.PA	11.978	1.44	3.85	10	6	1.14	10
TECF.PA	11.504	1.70	4.30	14	5	1.00	13
TOTF.PA	3.000	1.21	1.99	48	6	0.58	22
UNBP.PA	14.005	3.53	1.46	67	2	0.36	20
VIE.PA	9.147	2.10	2.08	44	4	0.53	13
VIV.PA	7.239	2.63	1.44	69	3	0.35	18
VLLP.PA	10.486	3.36	1.79	51	4	0.48	17
FCE	1.271	1.32	1.14	87	5	0.36	152

[6] In our data, we detect a trade through as a sequence of trades with the same timestamp and at least two different consecutive execution prices.

4.3 Methodology

In order to compute the intraday profile of correlation, or any other quantity of interest, we cut trading days into 5-minute bins spanning the whole day from the open to the close of the market, excluding the auction phases. For instance, on the CAC40 universe, the regular trading session opens at 9:00 and closes at 17:30, which leads to $8.5 \times 12 = 102$ bins. On each day, we compute statistics on each time slice and then we average over days for each slice. We end up with an average statistics for each 5-minute bin. More precisely, for a given bin $b = 1, \dots, B$, the resulting statistics is

$$S(b) = \frac{1}{n_{\text{days}}} \sum_{d=1}^{n_{\text{days}}} S_d(b)$$

where $S_d(b)$ is the statistics computed on day d and bin b. In order to compare results between assets universes, we normalize the intraday profile by the average value of the profile

$$\tilde{S}(b) = \frac{S(b)}{\frac{1}{B}\sum_{b=1}^{B} S(b)}$$

The average interevent duration across assets we consider ranges from 2.760 to 20.129 seconds, so that the average number of events in a bin roughly ranges from 15 to 109 and averages to 44. Note that this number fluctuates depending on the bin since intraday trading activity varies wildly. It leads to approximately 2816 observations per asset for each bin since we resort to a dataset spanning a time period of 64 trading sessions. To further increase the number of data, we display cross-sectional results, *i.e.* averaged over assets, which leads to $2816 \times 40 = 112640$ data per bin.

In the following, we will focus on second-order moments of price returns, such as covariance and variance. We use the Hayashi-Yoshida estimator [61] for the covariance C and the standard realized volatility estimator σ defined as follows

$$\begin{aligned}
C^{XY} &= \frac{1}{T} \sum_{i,j} r_i^X r_j^Y \mathbb{1}_{\{O_{ij} \neq \emptyset\}} \\
r_i^X &= P_{t_i}^X - P_{t_{i-1}}^X \\
r_j^Y &= P_{s_j}^Y - P_{s_{j-1}}^Y \\
O_{ij} &=]t_{i-1}, t_i] \cap]s_{j-1}, s_j] \\
\rho^{XY} &= \frac{C^{XY}}{\sigma^X \sigma^Y} \\
\sigma^k &= \sqrt{\frac{1}{T} \sum_i (r_i^k)^2}
\end{aligned}$$

We take P as the midquote sampled in trading time, *i.e.* $P_{t_i}^X$ (resp. $P_{s_j}^Y$) is the midquote of asset X (resp. Y) just before the trade that occured at time t_i (resp. s_j) on X (resp. Y). The Hayashi-Yoshida estimator avoids the choice of a large enough time scale to attenuate the Epps effect in correlation estimation [49]. As a result, we can use all the data available in a given bin, which should make our estimation more accurate than if we sample data on a regular time grid.

4.4 Empirical results

4.4.1 Correlation intraday profile

Figure 4.2 plots the intraday profile of correlation for the four universes and figure 4.3 zooms on cross-sectional medians of profiles normalized by their average[7].

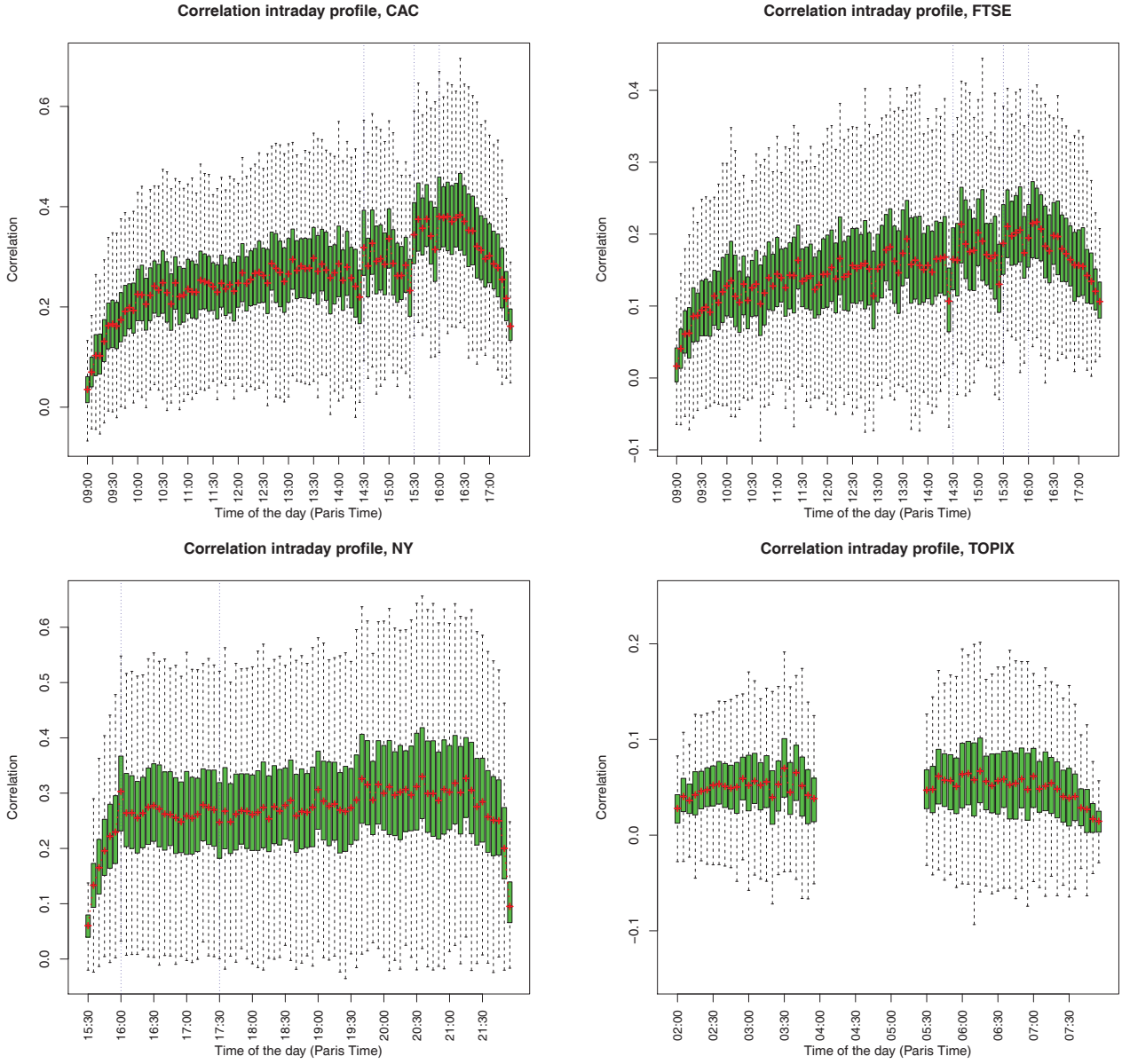

Figure 4.2: Intraday profile of correlation (cross-sectional distribution). Top left panel: CAC universe. Top right panel: FTSE universe. Bottom left panel: NY universe. Bottom right panel: TOPIX universe.

[7]The whisher plots we present display a box ranging from the first to the third quartile with the median in the middle, and whiskers extending to the most extreme point that is no more than 1.5 times the interquartile range from the box. These are the default settings of the `boxplot` function of `R`.

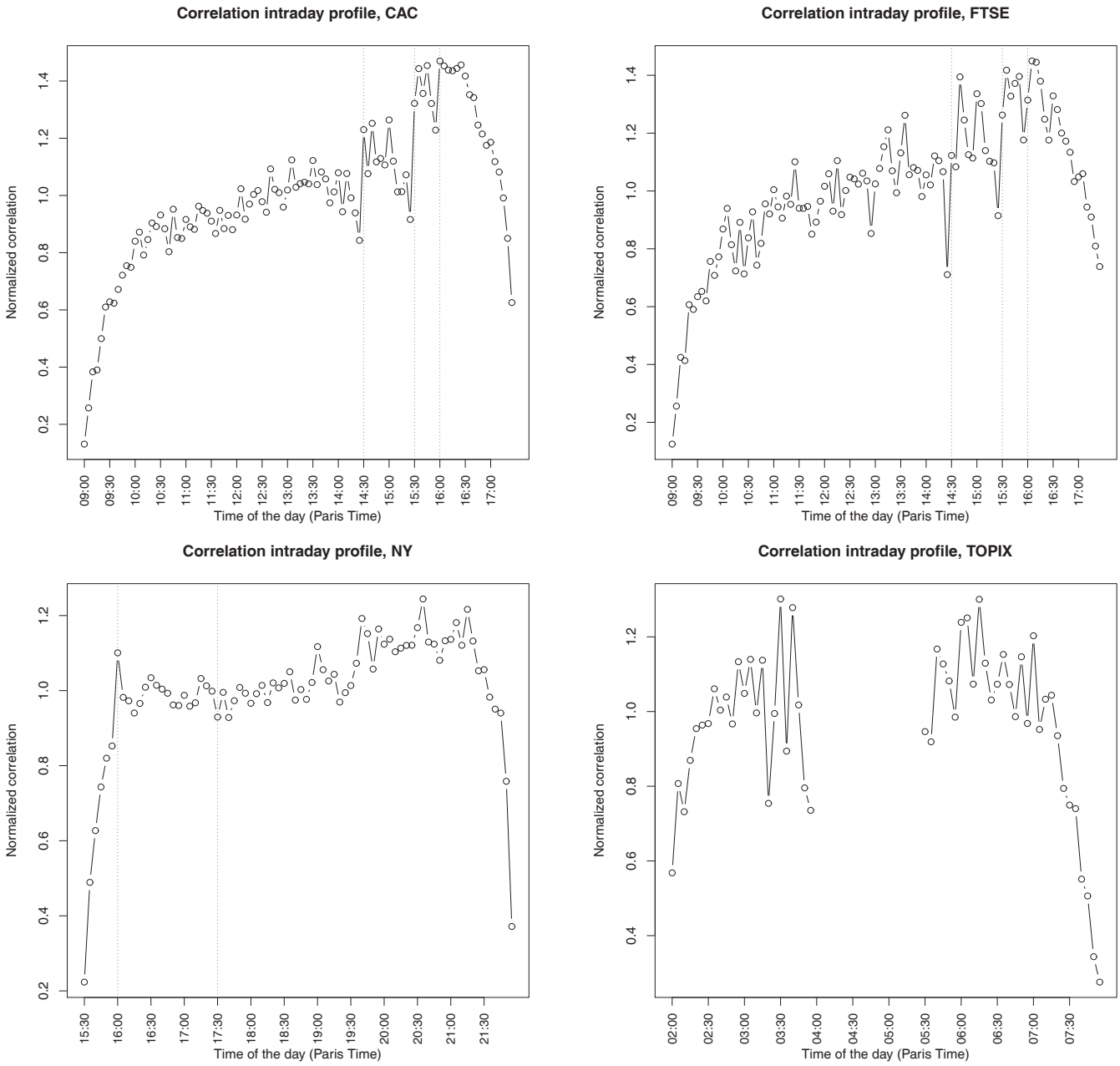

Figure 4.3: Intraday profile of normalized correlation (cross-sectional median). Top left panel: CAC universe. Top right panel: FTSE universe. Bottom left panel: NY universe. Bottom right panel: TOPIX universe.

The CAC and FTSE correlation profiles are very similar and show an upward trend. Correlation is substantially weak at the open and it steadily increases until 14:00, then there is a downward drift until 14:25 and a sudden strong upward jump at 14:30 when most of the US macroeconomic figures are announced[8].

[8]Figures released at 14:30 are the Consumer Price Index (Bureau of Labor Statistics), the Durable Goods Report (U.S. Census Bureau), the Employment Cost Index (Bureau of Labor Statistics), the Existing Home Sales (National Association of Realtors), the Factory Orders Report (U.S. Census Bureau), the Gross Domestic Product (Bureau of Economic Analysis), the

At 15:30 (NYSE and NASDAQ markets opening), we observe another sharp jump in correlation, followed by another jump at 16:00 when other US macroeconomic figures are known by market participants. After 16:00, correlation substantially decreases to roughly 60% (resp. 70%) of its daily average at the close for the CAC (resp. FTSE) universe. A remarkable pattern in this profile is that there is always a downard trend starting 15 or 20 minutes before abrupt jumps.

Regarding the NY universe, correlation is also very low at the open and increases up to 16:00, when the aforementioned jump related to US figures occurs. We do not observe a significant impact when European stock markets close (17:30), in contrary to the noteworthy impact of US markets opening on Europe. The upward trend continues until 21:15, after which the correlation drops to reach 40% of its average daily level. Most of this massive decrease in correlation is realized during the last 10 minutes of the trading session.

The intraday profile for the TOPIX universe is cut into two pieces because of the trading halt on the Tokyo Stock Exchange between 3:00 and 4:30. Both trading sessions tend to be similar with the patterns already observed on other universes: a global upward trend and a drop at the close. Note that the gap between the correlation at the first close and at the second open is quite small, roughly 10% the average daily level. Finally, we remark that the TOPIX cross-sectional distribution of correlation is the tightest and smallest, with an average correlation of about 5% and many negatively correlated stocks, while the three other universes exhibit about 25% of average correlation and not that much anti-correlated pairs.

The intraday correlation patterns of the four universes we consider present striking similarities. They all start with low correlation that increases as the trading session goes on and jumps up when significant pieces of news are released. A substantial decorrelation before the close is also observed on all four universes. This somehow universal pattern might be the consequence of underlying trading habits common to all market venues. We think that the small correlation at the open is due to the absence of market information at this time[9] and to the readjustment of positions taken the day before. Market participants get new pieces of information as time elapses, which can be either exogenous (economic figures and news) or endogenous (recent evolution of the market index). This tends to increase correlation as traders invest on all assets that they think are influenced by these news. This phenomenon is reinforced at times when it is known that significant news are released (14:30, 15:30 and 16:00 for example). Regarding the drop in correlation at the close, we believe that it is related to the intervention of big players such as asset managers and derivative houses, who often benchmark funds or derivatives on closing prices. These agents take colossal positions that are dictated by models for hedging or performance purposes and that might not be correlated with the market. Moreover, traders who don't want to bear overnight risk, such as high frequency market makers, unwind their position at the close whatsoever. To sum up, because lots of market participants have trading constraints at the end of the session, there is a natural decorrelation of assets near the close.

In order to assess the statistical significance of the pattern we find in intraday correlation, we use the Kolmogorov-Smirnov test and the Wilcoxon rank-sum test, also known as the U-test [72, 99]. We first aggregate the correlation profile into 15-minute bins. For instance, for the CAC universe, we get $(8.5 \times 60)/15 = 34$ samples of $40 \times 39/2 \times (15/5) = 2340$ correlations. For each time slice, we compare the cross-sectional distributions of correlation coefficients into this slice and into the next one. Figure 4.4 plots the p-values of both statistical tests as a function of time. The null hypothesis is that both samples are drawn from the same distribution, which is rejected at level α if the p-value is less than α.

Housing Starts (U.S. Census Bureau), the Jobless Claims Report (U.S. Department of Labor), the Personal Income and Outlays (Bureau of Economic Analysis), the Producer Price Index (Bureau of Labor Statistics), the Productivity Report (Bureau of Labor Statistics), the Retail Sales Report (Census Bureau and U.S. Department of Commerce) and the Trade Balance Report (Bureau of Economic Analysis). Those released at 16:00 are the Consumer Confidence Index (The Conference Board), the Non-Manufacturing Report (Institute for Supply Management), the Purchasing Managers Index (Institute for Supply Management) and the Wholesale Trade Report (U.S. Census Bureau). All of these figures are released on a monthly basis, except GDP which is announced quarterly.

[9] This lack of information is also reflected by large bid/ask spreads at the open, which illustrate the uncertainty of market participants on what the fair price should be.

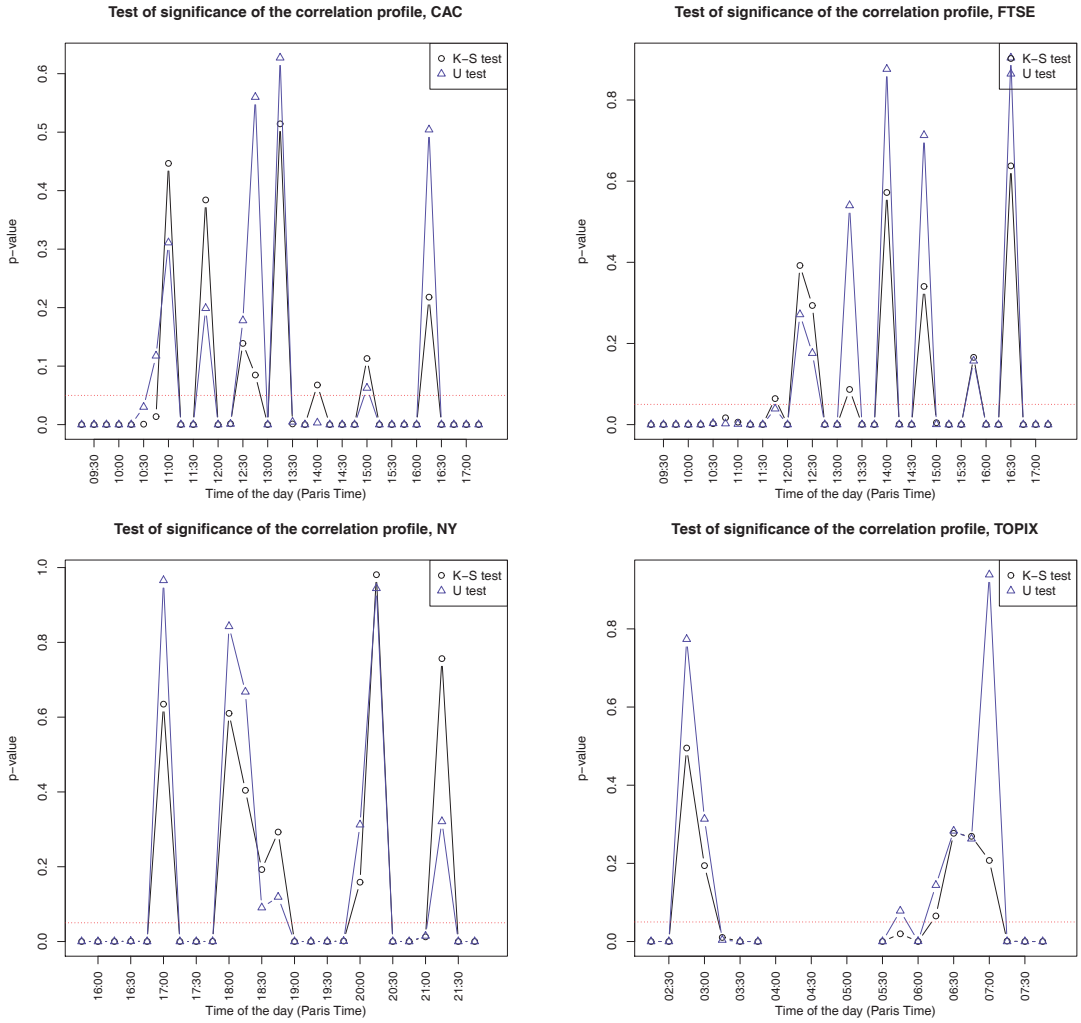

Figure 4.4: P-values of the tests of significance of the intraday correlation profile. The x-axis denotes the time between the two corresponding time slices. For instance, at 9:30 we compare the distribution of correlations from 9:15 to 9:30 with the distribution of correlations from 9:30 to 9:45. The red dotted line is the 5% level. Top left panel: CAC universe. Top right panel: FTSE universe. Bottom left panel: NY universe. Bottom right panel: TOPIX universe.

For the CAC and FTSE universes, we find that correlation does not significantly change 8 times out of 33, that is to say 24% of the times if we consider a standard type I error level $\alpha = 5\%$[10]. For the NY universe, this ratio is $8/25 = 32\%$. For the TOPIX, it is $6/17 \approx 35\%$ ($7/17 \approx 41\%$ with the U-test). It is thus confirmed that the correlation is significantly varying during the trading session. The correlation pattern in Europe (CAC and FTSE) is the wildest, followed by the US, which is itself less flat than in Japan. The half-hour time slots in which correlation can be regarded as statistically constant, taking the K-S test as a reference, are:

- CAC: 10:45-11:15, 11:30-12:00, 12:15-12:45, 12:30-13:00, 13:00-13:30, 13:45-14:15, 14:45-15:15, 16:00-16:30
- FTSE: 12:00-12:30, 12:15-12:45, 13:00-13:30, 13:45-14:15, 14:30-15:00, 15:30-16:00, 16:15-16:45
- NY: 16:45-17:15, 17:45-18:15, 18:00-18:30, 18:15-18:45, 18:30-19:00, 19:45-20:15, 20:00-20:30, 21:00-21:30
- TOPIX: 2:30-3:00, 2:45-3:15, 6:00-6:30, 6:15-6:45, 6:30-7:00, 6:45-7:15

[10]To be exact, there are 7 out of 33 p-values greater than α according to the U-test on the FTSE universe.

This statistically certifies that changes in correlation happen either next to the open and close auctions or when new market information arrives.

4.4.2 Comovement probabilities

The intraday profile of correlation provides insight into the evolution of linear dependencies during the trading session. However, it has been shown that the correlation matrix might not be enough to describe the complex dependencies between assets returns [34]. In order to get a more general picture of comovements, we compute the intraday profile of the following probability ratios

$$p^{ij}_{\pm\pm} = \frac{\mathbb{P}(\pm r^i > 0 \cap \pm r^j > 0)}{\mathbb{P}(\pm r^i > 0)\mathbb{P}(\pm r^j > 0)}$$

for two given assets i and j. This statistics measures the degree of dependency of the sign of returns of two assets in comparison with the null hypothesis of independent returns. If the two assets are independent, then $p^{ij}_{\pm\pm} = 1$. If they are positively (resp. negatively) correlated, then $p^{ij}_{++} > 1$ and $p^{ij}_{--} > 1$ (resp. < 1) and $p^{ij}_{+-} < 1$ and $p^{ij}_{-+} < 1$ (resp. > 1). It is easily seen that this measure does not depend on volatility. Indeed, assume that $r^k = \sigma^k \varepsilon^k$ where ε is a unit standard deviation noise. Then

$$p^{ij}_{\pm\pm} = \frac{\mathbb{P}(\pm \varepsilon^i > 0 \cap \pm \varepsilon^j > 0)}{\mathbb{P}(\pm \varepsilon^i > 0)\mathbb{P}(\pm \varepsilon^j > 0)}$$

To compute this quantity on our data, we switch to tick time rather than trade time, *i.e.* we increment time only when there is a non-zero midquote variation between two trades. As a result, the probabilities we consider cover all the possibilities. We compute joint probabilities with a straightforward modification of the Hayashi-Yoshida estimator.

Figure 4.5 depicts the intraday profile of the four comovement probability ratios on the four universes. The patterns are very similar to those of the correlation and thus allow to extend the conclusions of previous sections to a more general notion of bivariate comovement, not necessarily linear.

4.4.3 Idiosyncratic correlation

It is well known that the spectrum of the correlation matrix of returns has a non-trivial probability distribution [23]. In particular, most of the correlation comes from a so-called market mode that drives all assets in the same direction. The remaining significant correlation comes from sector-specific information or other less obvious (in terms of financial interpretation) sources of information. The intraday profile of the spectrum of the correlation matrix is studied in [8]. The authors find that the largest eigenvalue (market mode) increases throughout the day while other eignvalues decrease, and that this not a trivial consequence stemming from the fact that the trace of the correlation matrix is constant.

In the following, we use another approach to separate market and idiosyncratic correlations. Let us consider the standard CAPM model [93]

$$r^i_t = \alpha^i + \beta^i r^{\text{market}}_t + \varepsilon^i_t$$

where r^{market} is the return of the market portfolio and ε^i is the idiosyncratic randomness which is uncorrelated with r^{market}. Since the market potfolio is not a traded instrument, we use returns of the nearby-maturity future written on the index of the market under consideration as a proxy for r^{market}. The standard estimate for β^i is the ordinary least squares estimate

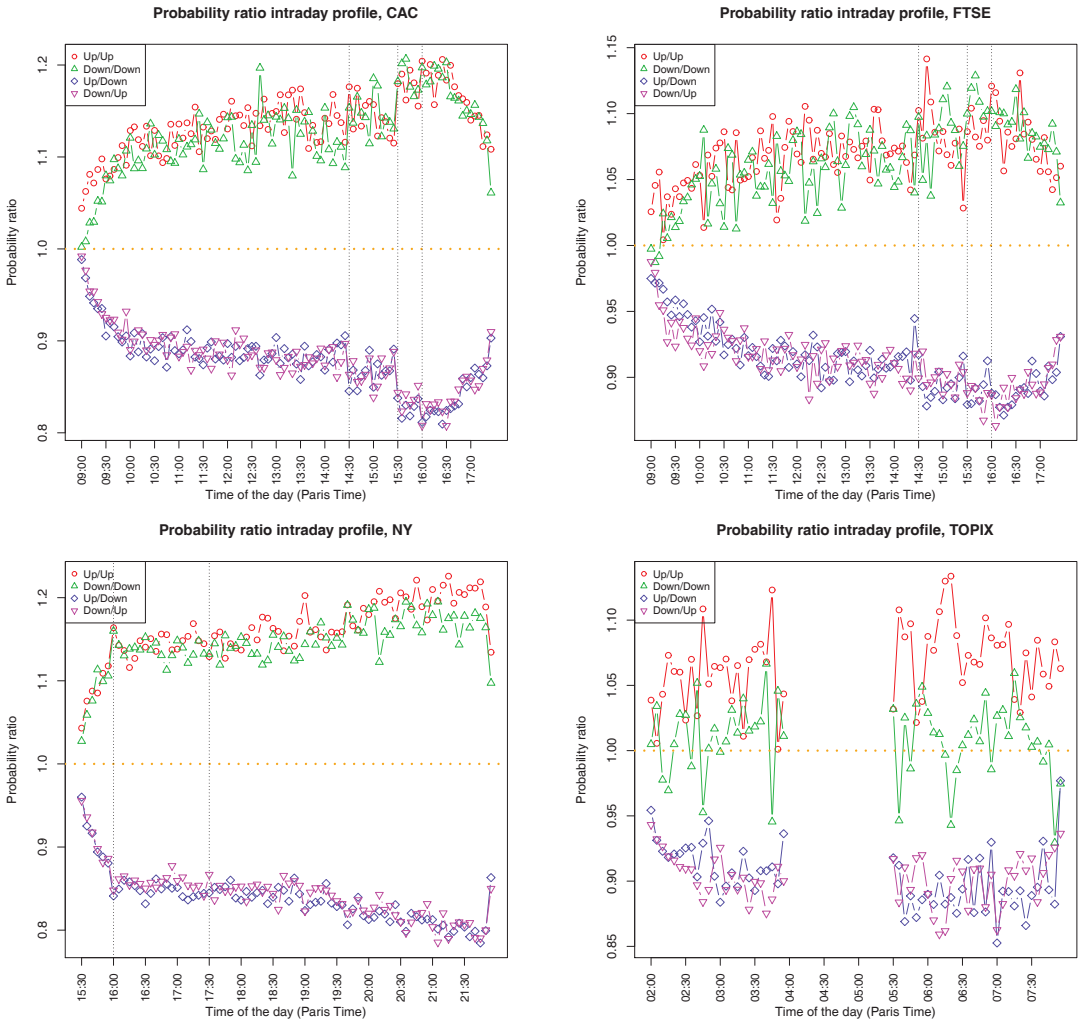

Figure 4.5: Intraday profile of comovement probability ratios (cross-sectional median) on the four universes. Top left panel: CAC. Top right panel: FTSE. Bottom left panel: NY. Bottom right panel: TOPIX.

$$\beta^i = \frac{\mathbb{C}\text{ov}(r^i, r^{\text{market}})}{\mathbb{V}\text{ar}(r^i, r^{\text{market}})}$$

Computations detailed in appendix 4.8 show that idiosyncratic correlations are given by

$$\mathbb{C}\text{orr}(\varepsilon^i, \varepsilon^j) = \frac{\rho^{i,j} - \rho^{i,\text{market}}\rho^{i,\text{market}}}{\sqrt{(1-(\rho^{i,\text{market}})^2)(1-(\rho^{j,\text{market}})^2)}}$$

where $\rho^{i,j} = \mathbb{C}\text{orr}(r^i, r^j)$.

Figure 4.6 plots the intraday profile of $\frac{\mathbb{C}\text{orr}(\varepsilon^i,\varepsilon^j)}{\rho^{i,j}}$, which is the proportion of correlation of returns that comes from the idiosyncratic correlation. This ratio is computed as a ratio of averages, not as an average of

ratios because the latter can result in outliers due to the possiblity of having a correlation close to zero while the idiosyncratic correlation is not small enough to compensate for it during a given five-minute time window. This typically happens during lunch time when data is scarcer. In Europe, the first value is significantly higher than others. In our opinion, this comes from the fact that corporate news are released when the market is closed and thus are incorporated into prices when the next trading session starts. Indeed, corporate news and figures are company-specific and are therefore associated to idiosyncratic returns, contrary to macroeconomic figures or market openings that are rather related to market information. The overall trend is downward sloping so that the market mode becomes more and more important as the trading session advances. This is intuitive since new information from the market becomes available.

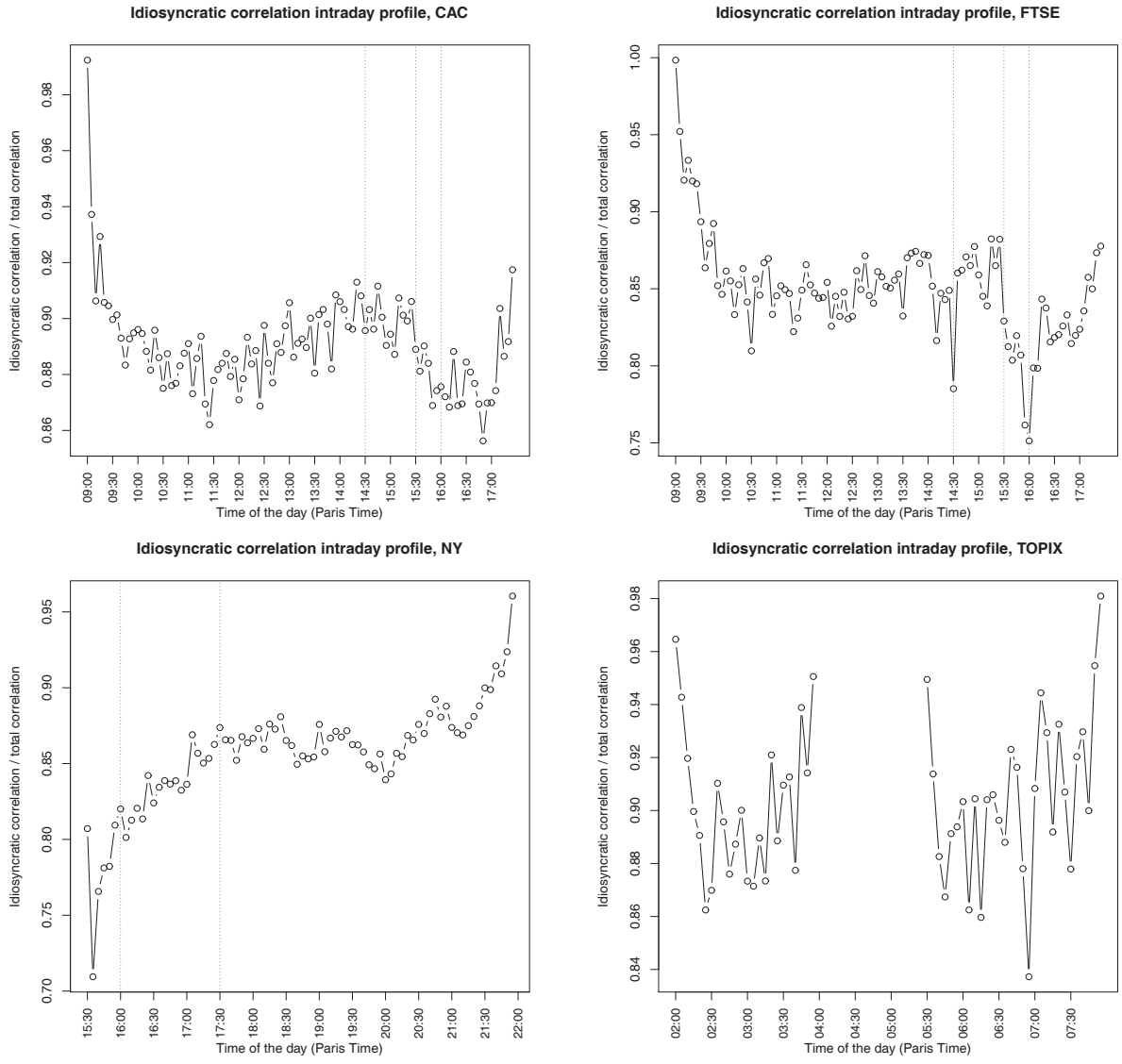

Figure 4.6: Intraday profile of idiosyncratic correlation (cross-sectional median). Top left panel: CAC universe. Top right panel: FTSE universe. Bottom left panel: NY universe. Bottom right panel: TOPIX universe.

4.5 Calibration of the intraday correlation profile with non-stationary Hawkes processes

4.5.1 Non-stationary Hawkes model

The discrete nature of tick-by-tick data, in both time and space, makes diffusion processes inappropriate for modelling prices at high frequency. The recent years have seen point processes emerge as an appealing alternative. In particular, Hawkes processes are now widely used in the empirical literature on microstructure (see [83] and references therein). In [16], the authors introduce a model for prices that can reproduce the signature plot and the Epps effect. However, they consider time-independent background intensities and triggering kernels. As a result, the correlation profile is flat in this setting. We extend their model to time-dependent background intensities and kernels in order to reproduce the varying correlation profile shown in subsection 4.4.1.

We consider two assets with the following coupled dynamics

$$\begin{aligned}
\mathrm{d}P_i(t) =& \delta_i(\mathrm{d}N_i^+(t) - \mathrm{d}N_i^-(t)) \\
\lambda_i^\pm(t) =& \mu_i^\pm(t) + \int_0^t \phi_i^{\mathrm{TF},\pm}(t, t-s)\mathrm{d}N_i^\pm(s) + \int_0^t \phi_i^{\mathrm{MR},\pm}(t, t-s)\mathrm{d}N_i^\mp(s) \\
&+ \int_0^t \phi_i^{\mathrm{CTF},\pm}(t, t-s)\mathrm{d}N_j^\pm(s) + \int_0^t \phi_i^{\mathrm{CMR},\pm}(t, t-s)\mathrm{d}N_j^\mp(s)
\end{aligned}$$

for $i \in \{1, 2\}$ and where

- $\delta_i > 0$ denotes the jump size of the price of the asset i. It is assumed to be constant.
- N_i^+ (resp.N_i^-) is the point process describing the upward (resp. downward) jumps of the price of the asset i
- $\mu_i^\pm$ is the background intensity of the point process $N_i^\pm$. It is a non-negative function.
- $\phi_i^{\mathrm{TF},\pm}$ (resp. $\phi_i^{\mathrm{MR},\pm}$, $\phi_i^{\mathrm{CTF},\pm}$, $\phi_i^{\mathrm{CMR},\pm}$) is the triggering kernel describing the influence of past jumps of $N_i^\pm$ (resp. $N_i^\mp$, $N_j^\pm$, $N_j^\mp$) on the intensity of $N_i^\pm$. TF (resp. MR, CTF, CMR) stands for Trend Following (resp. Mean Reverting, Cross Trend Following, Cross Mean Reverting). These are non-negative functions.

In this model, prices display temporal correlation through the kernels ϕ^{TF} and ϕ^{MR} as well as spatial correlation due to the kernels ϕ^{CTF} and ϕ^{CMR}. For instance, the kernel ϕ^{TF} creates a trend following behaviour because an upward jump of the price at time s will impact the upward intensity of the same asset at time t by a positive amount $\phi^{\mathrm{TF}}(t, t-s)$.

The price model described above can be rewritten in vector form

$$\begin{aligned}
\mathrm{d}P(t) =& J\mathrm{d}N(t) \\
\lambda(t) =& \mu(t) + \int_0^t \phi(t, t-s)\mathrm{d}N(s)
\end{aligned}$$

where $P = (P_1, P_2)^T$, $N = (N_1^+, N_1^-, N_2^+, N_2^-)^T$, $\lambda = (\lambda_1^+, \lambda_1^-, \lambda_2^+, \lambda_2^-)^T$, $\mu = (\mu_1^+, \mu_1^-, \mu_2^+, \mu_2^-)^T$, and

$$J = \begin{pmatrix} \delta_1 & -\delta_1 & 0 & 0 \\ 0 & 0 & \delta_2 & -\delta_2 \end{pmatrix}$$

$$\phi = \begin{pmatrix} \phi_1^{\mathrm{TF},+} & \phi_1^{\mathrm{MR},+} & \phi_1^{\mathrm{CTF},+} & \phi_1^{\mathrm{MCR},+} \\ \phi_1^{\mathrm{TF},-} & \phi_1^{\mathrm{MR},-} & \phi_1^{\mathrm{CTF},-} & \phi_1^{\mathrm{MCR},-} \\ \phi_2^{\mathrm{TF},+} & \phi_2^{\mathrm{MR},+} & \phi_2^{\mathrm{CTF},+} & \phi_2^{\mathrm{MCR},+} \\ \phi_2^{\mathrm{TF},-} & \phi_2^{\mathrm{MR},-} & \phi_2^{\mathrm{CTF},-} & \phi_2^{\mathrm{MCR},-} \end{pmatrix}$$

The increments of the resulting point process are obviously nonstationary, in the sense that the distributions of $N_{t_0+h} - N_{t_0}$ and $N_{t_1+h} - N_{t_1}$ are not identical for $t_0 \neq t_1$. The following proposition states a sufficient condition for the process to be well-defined on average.

Proposition 4.5.1. *Let us consider a multivariate point process $N = (N_1, \ldots, N_d)^T$ with intensity $\lambda = (\lambda_1, \ldots, \lambda_d)^T$ defined as*

$$\lambda(t) = \mu(t) + \int_{-\infty}^{t} \phi(t, t-s) \mathrm{d}N_s$$

where $\mu = (\mu_1, \ldots, \mu_d)^T$ is a vector function valued in $\mathbb{R}^{d,+}$ and $\phi(t,x) = (\phi_{ij}, 1 \leq i, j \leq d)$ is a matrix function with each entry ϕ_{ij} valued in $\mathbb{R}^+$. If $\sup_t \mu(t) < +\infty$ and the largest eigenvalue of the matrix $\left\|\overline{\phi}\right\|_1$ is strictly less than one, where

$$\left\|\overline{\phi}\right\|_1 = \int_0^{+\infty} \sup_t \phi(t,x) \mathrm{d}x$$

where the supremum and the integration are componentwise, then $\mathbb{E}(\lambda(t)) < +\infty$ for all t.

Proof. We set $f(t) = \mathbb{E}(\lambda(t))$. Taking the expectation on both sides of the definition of $\lambda(t)$ yields

$$\begin{aligned} f(t) &= \mu(t) + \int_{-\infty}^{t} \phi(t, t-s) f(s) \mathrm{d}s \\ &= \mu(t) + \int_0^{+\infty} \phi(t,x) f(t-x) \mathrm{d}x \end{aligned}$$

We define $\|f\|_\infty = (\|f_1\|_\infty, \ldots, \|f_d\|_\infty)^T$ where $\|f_i\|_\infty = \sup_t |f_i(t)|$ is the uniform convergence norm. We have

$$\begin{aligned} \|f\|_\infty &\leq \|\mu\|_\infty + \sup_t \int_0^{+\infty} \phi(t,x) f(t-x) \mathrm{d}x \\ &\leq \|\mu\|_\infty + \int_0^{+\infty} \sup_t \left\{\phi(t,x) f(t-x)\right\} \mathrm{d}x \\ &\leq \|\mu\|_\infty + \int_0^{+\infty} \sup_t \left\{\phi(t,x)\right\} \sup_t \left\{f(t-x)\right\} \mathrm{d}x \\ &\leq \|\mu\|_\infty + \int_0^{+\infty} \sup_t \phi(t,x) \mathrm{d}x \, \|f\|_\infty \\ &= \|\mu\|_\infty + \left\|\overline{\phi}\right\|_1 \|f\|_\infty \end{aligned}$$

The third inequality comes from the fact that $\sup_t f(t)g(t) \leq \sup_t f(t) \sup_t g(t)$. Indeed,

$$\begin{aligned}
\sup_t f(t)g(t) &= \exp(\sup_t \ln(f(t)g(t))) \\
&= \exp(\sup_t \ln(f(t)) + \ln(g(t))) \\
&\leq \exp(\sup_t \ln(f(t)) + \sup_t \ln(g(t))) \\
&= \exp(\sup_t \ln(f(t))) \exp(\sup_t \ln(g(t))) \\
&= \sup_t f(t) \sup_t g(t)
\end{aligned}$$

If $\|\mu\|_\infty < +\infty$ and if the largest eigenvalue of $\left\|\overline{\phi}\right\|_1$ is strictly less than one, which implies that the matrix $I - \left\|\overline{\phi}\right\|_1$ is positive definite, then we have

$$\|f\|_\infty \leq (I - \left\|\overline{\phi}\right\|_1)^{-1} \|\mu\|_\infty < +\infty$$

□

4.5.2 Estimation of the parameters

Constant parameters

The standard approach for estimating the parameters of a Hawkes process is the maximum likelihood estimator (MLE) [88]. An analytical formula can be derived for the likelihood of a Hawkes process but it has to be maximized with a numerical routine. Recently, an Expectation-Maximization (EM) algorithm was suggested in [74]. This is an iterative algorithm that converges towards the MLE. Contrary to the MLE approach, it is entirely analytical and thus does not require the use of a numerical optimizer. We briefly recall how it works below.

Let us consider a series of timestamps $0 < t_1 < \ldots < t_n < T$ as the outcome of a univariate Hawkes process with parameters $\theta = (\mu, \phi)$ where μ is the background intensity and ϕ is the triggering kernel. The log-likelihood of this outcome is given by (see [42] e.g.)

$$\ell(t_1, \ldots, t_n; \theta) = \sum_{t_i < T} \ln(\mu + \sum_{t_j < t_i} \phi(t_i - t_j)) - \mu T - \int_0^T \sum_{t_i < t} \phi(t - t_i) \mathrm{d}t + T$$

In the case of an exponential triggering kernel[11], *i.e.* $\phi(x) = \alpha\beta e^{-\beta x}$, it becomes[12]

$$\ell(t_1, \ldots, t_n; \mu, \alpha, \beta) = \sum_{i=1}^{n} \ln(\mu + \alpha\beta \sum_{j=1}^{i-1} e^{-\beta(t_i - t_j)}) - \mu T - \alpha \sum_{i=1}^{n} (1 - e^{-\beta(T - t_i)})$$

Unfortunately, there is no closed form solution to the MLE equation $\nabla_\theta \ell = 0$, which calls for the need of a numerical optimizer.

[11] We specifiy the exponential triggering kernel as $\phi(x) = \alpha\beta e^{-\beta x}$. In the literature, the form $\phi(x) = \alpha e^{-\beta x}$ is often used instead. We rather prefer the former specification for numerical stability of the estimation procedure when $\beta << 1$ (see appendix 4.9 for further details).

[12] We get rid of the last term equal to T because it does not depend on the parameters θ and thus has no impact on the maximization of the likelihood.

In [74], the authors take advantage of the branching structure of Hawkes processes to derive an EM algorithm. Indeed, the log-likelihood can be rewritten as follows

$$\begin{aligned}
\ell(t_1,\dots,t_n;\theta) =& \sum_{i=1}^{n} \ln(\lambda(t_i))(\chi_{ii} + \sum_{j=1}^{i-1}\chi_{ij}) - \mu T - \int_0^T \sum_{t_i<t} \phi(t-t_i)\mathrm{d}t \\
=& \ln(\mu)\sum_{i=1}^{n}\chi_{ii} + \sum_{i=2}^{n}\sum_{j=1}^{i-1}\chi_{ij}\ln(\phi(t_i-t_j)) - \mu T - \int_0^T \sum_{t_i<t}\phi(t-t_i)\mathrm{d}t \\
\chi_{ii} =& \mathbb{1}_{\{i \text{ is a background event}\}} \\
\chi_{ij} =& \mathbb{1}_{\{i \text{ is caused by } j\}}
\end{aligned}$$

since for all i, $\chi_{ii} + \sum_{j=1}^{i-1}\chi_{ij} = 1$. In the exponential case, the log-likelihood becomes

$$\begin{aligned}
\ell(t_1,\dots,t_n;\mu,\alpha,\beta) =& \ln(\mu)\sum_{i=1}^{n}\chi_{ii} + (\ln(\alpha)+\ln(\beta))\sum_{i=2}^{n}\sum_{j=1}^{i-1}\chi_{ij} - \beta\sum_{i=2}^{n}\sum_{j=1}^{i-1}\chi_{ij}(t_i-t_j) \\
& - \mu T - \alpha\sum_{i=1}^{n}(1-e^{-\beta(T-t_i)})
\end{aligned}$$

The branching structure corresponds to the terms χ_{ii} and χ_{ij}. Of course, it is not observable from the outcome $(t_1,\dots,t_n)$. However we can compute its expected value given the parameters (μ,ϕ)

$$\begin{aligned}
p_{ii} = \mathbb{E}(\chi_{ii}) &= \frac{\mu}{\mu + \sum_{r=1}^{i-1}\phi(t_i-t_r)} \\
p_{ij} = \mathbb{E}(\chi_{ij}) &= \frac{\phi(t_i-t_j)}{\mu + \sum_{r=1}^{i-1}\phi(t_i-t_r)}
\end{aligned}$$

This gives rise to the following EM algorithm:

Algorithm 1 EM estimation

Require: $\theta^0 = (\mu^0,\phi^0)$, $\epsilon \in \mathbb{R}^+$, $k_{\max} \in \mathbb{N}^*$

1: Start with initial parameters $\theta^0 = (\mu^0,\phi^0)$ and set $k=0$.
2: E-step: compute p_{ii}^k, p_{ij}^k using θ^k to get $\mathbb{E}(\ell^k(t_1,\dots,t_n;\mu,\phi))$.
3: M-step: solve $\frac{\partial\mathbb{E}(\ell^k)}{\partial\mu}=0$ and $\frac{\partial\mathbb{E}(\ell^k)}{\partial\phi}=0$ to get $\theta^{k+1}=(\mu^{k+1},\phi^{k+1})$.
4: set $k=k+1$
5: set $\Delta^k = \frac{\mathbb{E}(\ell^k(t_1,\dots,t_n;\theta^k)) - \mathbb{E}(\ell^{k-1}(t_1,\dots,t_n;\theta^{k-1}))}{|\mathbb{E}(\ell^{k-1}(t_1,\dots,t_n;\theta^{k-1}))|}$
6: **if** $\Delta^k \le \epsilon$ or $k = k_{\max}$ **then**
7: go to 11
8: **else**
9: go to 2
10: **end if**
11: **return** θ^k

In the exponential case, the M-step corresponds to the three following equations

$$\frac{\partial \mathbb{E}(\ell)}{\partial \mu} = \frac{\sum_{i=1}^{n} p_{ii}}{\mu} - T = 0$$
$$\frac{\partial \mathbb{E}(\ell)}{\partial \alpha} = \frac{\sum_{i=2}^{n}\sum_{j=1}^{i-1} p_{ij}}{\alpha} - \sum_{i=1}^{n}(1 - e^{-\beta(T-t_i)}) = 0$$
$$\frac{\partial \mathbb{E}(\ell)}{\partial \beta} = \frac{\sum_{i=2}^{n}\sum_{j=1}^{i-1} p_{ij}}{\beta} - \sum_{i=2}^{n}\sum_{j=1}^{i-1} p_{ij}(t_i - t_j) - \alpha\sum_{i=1}^{n} e^{-\beta(T-t_i)}(T - t_i) = 0$$

which translate into the following estimates at iteration $k+1$

$$\mu^{k+1} = \frac{\sum_{i=1}^{n} p_{ii}^{k}}{T}$$
$$\alpha^{k+1} = \frac{\sum_{i=2}^{n}\sum_{j=1}^{i-1} p_{ij}^{k}}{\sum_{i=1}^{n}(1 - e^{-\beta^k(T-t_i)})}$$
$$\beta^{k+1} = \frac{\sum_{i=2}^{n}\sum_{j=1}^{i-1} p_{ij}^{k}}{\sum_{i=2}^{n}\sum_{j=1}^{i-1} p_{ij}^{k}(t_i - t_j) + \alpha^k\sum_{i=1}^{n} e^{-\beta^{k+1}(T-t_i)}(T - t_i)}$$

The estimate for β^{k+1} is implicit so we have to solve the equation with a numerical root finder. However, as suggested in [74], we replace β^{k+1} by β^k in the right-hand side of the equation, which makes the estimate for β^{k+1} explicit.

We illustrate the respective performances of the MLE and EM approaches in a simulation framework in appendix 4.10.

Time-dependent parameters

In this paragraph, we extend the previous EM algorithm to estimate time-dependent parameters. We allow the background intensity and the triggering kernel to be time-dependent functions, *i.e.*

$$\lambda(t) = \mu(t) + \int_0^t \phi(t, t-s)\mathrm{d}N_s$$

In the following, we consider exponential decay kernels and we only allow the amplitude coefficient α to be time-dependent in the kernel, *i.e.* (in the univariate framework)

$$\lambda(t) = \mu(t) + \alpha(t)\beta\sum_{t_i<t} e^{-\beta(t-t_i)}$$

Since β is the most difficult parameter to estimate, it seems more reasonable to keep it constant. Moreover we shall see below that the estimation procedure would be much more complex if we allow β to vary. We specify the time dependency of μ and α as follows

$$\mu(t) = \sum_{p=0}^{P} \mu_p N_{p,d}(t), \quad \alpha(t) = \sum_{p=0}^{P} \alpha_p N_{p,d}(t)$$

where $(N_{p,d}, p = 0, \dots, P)$ is a B-spline basis of order d [46]. B-spline functions are piecewise polynomial functions of order d that are $d-1$ continuously differentiable. Their recursive definition makes them computationally appealing. The `splines` package of `R` provides a framework to compute them in an efficient way. We give some details on this family of functions in appendix 4.11. Note that the stationary case corresponds to $\mu_p = \mu$ and $\alpha_p = \alpha \ \forall p$ since $\sum_{p=0}^{P} N_{p,d}(t) = \mathbb{1}_{\{[0,T]\}}(t)$.

The log-likelihood is now

$$\begin{aligned}\ell(t_1, \dots, t_n; \mu, \alpha, \beta) = &\sum_{i=1}^{n} \chi_{ii} \ln(\sum_{p=0}^{P} \mu_p N_{p,d}(t_i)) + \sum_{i=2}^{n} (\ln(\sum_{p=0}^{P} \alpha_p N_{p,d}(t_i)) + \ln(\beta)) \sum_{j=1}^{i-1} \chi_{ij} - \beta \sum_{i=2}^{n} \sum_{j=1}^{i-1} \chi_{ij}(t_i - t_j) \\ &- \sum_{p=0}^{P} \mu_p \int_0^T N_{p,d}(t)\mathrm{d}t - \beta \sum_{p=0}^{P} \alpha_p \sum_{i=1}^{n} \int_{t_i}^{T} e^{-\beta(t-t_i)} N_{p,d}(t)\mathrm{d}t\end{aligned}$$

The E-step estimators of the branching structure become

$$\begin{aligned}p_{ii} = \mathbb{E}(\chi_{ii}) &= \frac{\mu(t_i)}{\mu(t_i) + \alpha(t_i)\beta \sum_{r=1}^{i-1} e^{-\beta(t_i - t_r)}} \\ p_{ij} = \mathbb{E}(\chi_{ij}) &= \frac{\alpha(t_i)\beta e^{-\beta(t_i - t_j)}}{\mu(t_i) + \alpha(t_i)\beta \sum_{r=1}^{i-1} e^{-\beta(t_i - t_r)}}\end{aligned}$$

The M-step corresponds to the following set of $2(P+1)+1$ M-step equations

$$\begin{aligned}\frac{\partial \mathbb{E}(\ell)}{\partial \mu_k} &= \frac{1}{\mu_k} \sum_{i=1}^{n} \frac{p_{ii} N_{k,d}(t_i)}{\sum_{p \neq k} \frac{\mu_p}{\mu_k} N_{p,d}(t_i) + N_{k,d}(t_i)} - \int_0^T N_{k,d}(t)\mathrm{d}t = 0 \\ \frac{\partial \mathbb{E}(\ell)}{\partial \alpha_k} &= \frac{1}{\alpha_k} \sum_{j<i} \frac{p_{ij} N_{k,d}(t_i)}{\sum_{p \neq k} \frac{\alpha_p}{\alpha_k} N_{p,d}(t_i) + N_{k,d}(t_i)} - \beta \sum_{i=1}^{n} \int_{t_i}^{T} e^{-\beta(t-t_i)} N_{k,d}(t)\mathrm{d}t = 0 \\ \frac{\partial \mathbb{E}(\ell)}{\partial \beta} &= \frac{\sum_{j<i} p_{ij}}{\beta} - \sum_{j<i} p_{ij}(t_i - t_j) - \sum_{p=0}^{P} \alpha_p (\sum_{i=1}^{n} \int_{t_i}^{T} e^{-\beta(t-t_i)} N_{p,d}(t)\mathrm{d}t - \beta \sum_{i=1}^{n} \int_{t_i}^{T} e^{-\beta(t-t_i)}(t-t_i) N_{p,d}(t)\mathrm{d}t) = 0\end{aligned}$$

It turns into the following estimators

$$\begin{aligned}\mu_k &= \frac{\sum_{i=1}^{n} \frac{p_{ii} N_{k,d}(t_i)}{\sum_{p=0}^{P} \mu_p^k N_{p,d}(t_i)}}{\int_0^T N_{k,d}(t)\mathrm{d}t} \\ \alpha_k &= \frac{\sum_{i=2}^{n} \frac{N_{k,d}(t_i)}{\sum_{p=0}^{P} \alpha_p^k N_{p,d}(t_i)} \sum_{j=1}^{i-1} p_{ij}}{\beta \sum_{i=1}^{n} \int_{t_i}^{T} e^{-\beta(t-t_i)} N_{k,d}(t)\mathrm{d}t} \\ \beta &= \frac{\sum_{j<i} p_{ij}}{\sum_{j<i} p_{ij}(t_i - t_j) + \sum_{p=0}^{P} \alpha_p (\sum_{i=1}^{n} \int_{t_i}^{T} e^{-\beta(t-t_i)} N_{p,d}(t)\mathrm{d}t - \beta \sum_{i=1}^{n} \int_{t_i}^{T} e^{-\beta(t-t_i)}(t-t_i) N_{p,d}(t)\mathrm{d}t)}\end{aligned}$$

where $\mu_p^k = \frac{\mu_p}{\mu_k}$ and $\alpha_p^k = \frac{\alpha_p}{\alpha_k}$. Note that all these equations are implicit. In our implementation of the EM algorithm, we choose to evaluate the right-hand side of these equations with the values of the previous

iteration. All the integrals present in these formulas can be computed analytically and rapidly (see appendix 4.11). We remark that our approximation is exact in the case of piecewise constant parameters, which corresponds to $d = 0$. Indeed, $N_{k,0}(x) = \mathbb{1}_{[u_k,u_{k+1})}(x)$ where $u_0, u_1, \ldots, u_{P+1}$ are knots (see appendix 4.11). This leads to

$$\mu_k = \frac{\sum_{i=1}^{n} p_{ii}\mathbb{1}_{[u_k,u_{k+1})}(t_i)}{u_{k+1} - u_k}$$

$$\alpha_k = \frac{\sum_{i=2}^{n}\sum_{j=1}^{i-1} p_{ij}\mathbb{1}_{[u_k,u_{k+1})}(t_i)}{\beta \sum_{i=1}^{n}\int_{\max(t_i,u_k)}^{u_{k+1}} e^{-\beta(t-t_i)}\mathrm{d}t\mathbb{1}_{\{t_i<u_{k+1}\}}}$$

$$\beta = \frac{\sum_{j<i} p_{ij}}{\sum_{j<i} p_{ij}(t_i - t_j) + \sum_{p=0}^{P}\alpha_p \sum_{i=1}^{n}\int_{\max(t_i,u_p)}^{u_{p+1}} e^{-\beta(t-t_i)}(1-\beta(t-t_i))\mathrm{d}t\mathbb{1}_{\{t_i<u_{p+1}\}}}$$

μ_k and α_k can be interpreted as local estimators using only branching probabilities of timestamps falling into the set $[u_k, u_{k+1})$. In the following, we will always set $d = 0$ for the sake of simplicity. However, it is straightforward to implement formulas with $d > 0$ using results of appendix 4.11 if smoother functions are needed.

Price model

The price model introduced in subsection 4.5.1 can be seen as a superposition of four coupled point processes. Therefore, the log-likelihood of the model can be split into four independent parts corresponding to each point process [68]

$$\ell(\mathcal{T}_1^+, \mathcal{T}_1^-, \mathcal{T}_2^+, \mathcal{T}_2^-; \theta_1^+, \theta_1^-, \theta_2^+, \theta_2^-) = \ell_1^+(\mathcal{T}_1^+; \theta_1^+) + \ell_1^-(\mathcal{T}_1^-; \theta_1^-) + \ell_2^+(\mathcal{T}_2^+; \theta_2^+) + \ell_2^-(\mathcal{T}_2^-; \theta_2^-)$$

$$\ell_i^a(\mathcal{T}_i^a; \theta_i^a) = \sum_{j=1}^{n_i^a} \ln(\lambda_i^a(t_{i,j}^a; \theta_i^a)) - \int_0^T \lambda_i^a(t; \theta_i^a)\mathrm{d}t$$

$$\mathcal{T}_i^a = \left\{t_{i,1}^a, t_{i,2}^a, \ldots, t_{i,n_i^a}^a\right\}$$

where $t_{i,j}^a$ is the j^{th} jump of type $a \in \{+,-\}$ of the price of the asset i, and $\theta_i^a = (\mu_i^a, \phi_i^{\text{TF},a}, \phi_i^{\text{MR},a}, \phi_i^{\text{CTF},a}, \phi_i^{\text{CMR},a})$[13]. This decomposition is very useful from a computational point of view because it allows to switch from one optimization problem with d parameters to four independent optimization problems with $d/4$ parameters. Each optimization problem can be solved with the EM algorithm presented in the above paragraph.

4.5.3 Empirical results

We consider the two French stocks AIRP.PA (gas company) and TOTF.PA (oil company). We estimate the parameters of the price model presented in subsection 4.5.1 using the EM algorithm described in subsection 4.5.2[14]. We report average parameters (over trading days) along with standard deviations in brackets. We make four estimations corresponding to different specifications of the model

- Model 1: $\mu_i^a(t) = \mu_i^a \ \forall t \in [0,T]$ and $\alpha_i^{\text{CTF},a}(t) = \alpha_i^{\text{CTF},a} \ \forall t \in [0,T]$
- Model 2: $\mu_i^a(t)$ as specified in subsection 4.5.2 (B-spline expansion) and $\alpha_i^{\text{CTF},a}(t) = \alpha_i^{\text{CTF},a} \ \forall t \in [0,T]$

[13] Note that ℓ_i^a also depends on timestamps other than $\mathcal{T}_i^a$ due to excitation effects. We omit this dependency in the above equations for notational simplicity.

[14] We set the relative tolerance $\epsilon = 10^{-10}$, the maximum number of iterations $k_{\max} = 4000$ and the starting parameters $\mu^0 = \frac{0.01+10}{2\langle\Delta t\rangle}$, $\alpha^0 = 0.5$, $\beta = \frac{0.001+100}{2}$.

- Model 3: $\mu_i^a(t) = \mu_i^a \ \forall t \in [0,T]$ and $\alpha_i^{\mathrm{CTF},a}(t)$ as specified in subsection 4.5.2 (B-spline expansion)
- Model 4: $\mu_i^a(t)$ and $\alpha_i^{\mathrm{CTF},a}(t)$ as specified in subsection 4.5.2 (B-spline expansion)

When we specifiy either μ or α as a B-spline expansion, we choose a sequence ranging from 9:00 to 17:30 with a mesh of five minutes for the knots of the B-splines, which amounts to $P+1 = 105$ parameters for each B-spline expansion. The timestamps $\mathcal{T}_i^+$ (resp. $\mathcal{T}_i^-$) are the upward (resp. downward) jumps of the midquote of asset i sampled in trading time (we set TOTF.PA as asset 1 and AIRP.PA as asset 2). We only set μ and α^{CTF} as time-dependent functions because we are interested in the intraday profile of comovements of these two assets that are strongly positively correlated.

Once parameters have been estimated, we simulate 1000 bivariate paths according to the corresponding model and we compute the associated average intraday correlation profile. Paths of assets are simulated using the thinning algorithm introduced in [87]. Our aim is to reproduce the intraday correlation profile observed in subsection 4.4.1. It is shown on figure 4.7 in the case of AIRP.PA/TOTF.PA.

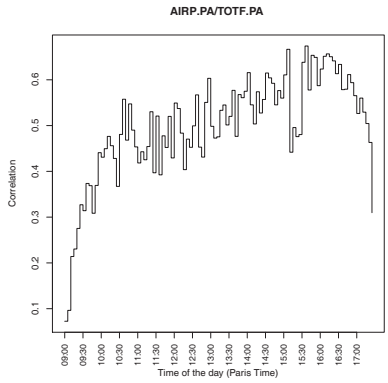

Figure 4.7: Intraday correlation profile for the pair AIRP.PA/TOTF.PA.

Model 1

Table 4.3 shows the average estimated parameters of model 1. We see that the background intensity of TOTF.PA is twice higher than the AIRP.PA one. The largest excitation kernel is the TF effect, followed by the CTF. The MR kernel is roughly three times less than the TF. The weakest effect is the CMR. This means that returns are positively correlated both in time and space dimensions at this time scale (one price move). Though positive correlation between these two assets sounds intuitive, positive autocorrelation seems less obvious. Indeed, evidence of negative autocorrelation of futures tick-by-tick returns is reported in the literature (see [38] e.g.). As a sanity check, we have run the estimation of a bivariate model for the CAC40 future (asset 1) and the stock BNPP.PA (asset 2). We have found that the mean-reverting effect strongly dominates for future returns contrarily to the stock[15]. Therefore, the trend-following behavior of stock tick-by-tick returns seems to be genuine over this time period. The values of β^{CTF} indicate that the impact of the past jumps of AIRP.PA are incorporated faster in price moves of TOTF.PA than the converse. Indeed, the half-life time of the influence of AIRP.PA on TOTF.PA is $\frac{\ln(2)}{2}((\beta_1^{\mathrm{CTF},+})^{-1} + (\beta_1^{\mathrm{CTF},-})^{-1}) \approx 0.096$ seconds, while the converse is $\frac{\ln(2)}{2}((\beta_2^{\mathrm{CTF},+})^{-1} + (\beta_2^{\mathrm{CTF},-})^{-1}) \approx 0.145$ seconds. This is as expected since TOTF.PA is more liquid than AIRP.PA. Note that upward and downward intensities are highly symmetric, which suggests

[15] $\alpha_1^{\mathrm{TF},+} = 0.132$, $\alpha_1^{\mathrm{TF},-} = 0.137$, $\alpha_1^{\mathrm{MR},+} = 0.417$, $\alpha_1^{\mathrm{MR},-} = 0.399$ and $\alpha_2^{\mathrm{TF},+} = 0.212$, $\alpha_2^{\mathrm{TF},-} = 0.221$, $\alpha_2^{\mathrm{MR},+} = 0.054$, $\alpha_2^{\mathrm{MR},-} = 0.052$

a reduction of the number of parameters by imposing $\mu_i^+ = \mu_i^-$ and $\phi_i^{e,+} = \phi_i^{e,-}$ for each asset $i \in \{1,2\}$ and effect $e \in \{\text{TF}, \text{MR}, \text{CTF}, \text{CMR}\}$. We report the results of the estimation of the symmetrized model in table 4.4. The estimated parameters are very close to the values presented in table 4.3 and display smaller standard deviations because more data (upward and downward timestamps) are used for the estimation. As a result, we will consider the symmetrized model in the following. This is even more appropriate as we increase the dimension of the model.

Table 4.3: Estimation of the parameters of model 1.

	$1,+$	$1,-$	$2,+$	$2,-$
μ	0.044 (0.003)	0.045 (0.003)	0.022 (0.001)	0.022 (0.001)
α^{TF}	0.269 (0.021)	0.237 (0.016)	0.22 (0.017)	0.227 (0.017)
α^{MR}	0.088 (0.012)	0.107 (0.015)	0.083 (0.016)	0.065 (0.007)
α^{CTF}	0.19 (0.033)	0.223 (0.031)	0.094 (0.008)	0.104 (0.007)
α^{CMR}	0.016 (0.001)	0.019 (0.001)	0.011 (0.001)	0.01 (0.001)
β^{TF}	2.588 (0.325)	2.574 (0.254)	2.624 (0.414)	3.487 (0.593)
β^{MR}	7.705 (0.527)	6.889 (0.525)	9.029 (0.692)	8.153 (0.676)
β^{CTF}	7.071 (1.023)	7.41 (0.961)	5.463 (0.844)	4.245 (0.674)
β^{CMR}	58.868 (13.937)	41.963 (4.338)	32.378 (3.144)	36.804 (3.452)

Table 4.4: Estimation of the parameters of the symmetric version of model 1.

	1	2
μ	0.043 (0.003)	0.023 (0.001)
α^{TF}	0.268 (0.02)	0.222 (0.014)
α^{MR}	0.081 (0.008)	0.057 (0.007)
α^{CTF}	0.243 (0.037)	0.107 (0.009)
α^{CMR}	0.017 (0.001)	0.011 (0)
β^{TF}	2.298 (0.271)	2.371 (0.368)
β^{MR}	7.454 (0.469)	8.25 (0.53)
β^{CTF}	5.736 (0.721)	3.584 (0.613)
β^{CMR}	37.561 (3.014)	32.385 (2.908)

Furthermore, we decide to get rid of trend-following, mean-reverting and cross-mean-reverting kernels. The two first mainly impact marginal probability distributions of assets, leaving bivariate quantities such as correlation almost unchanged. Regarding the cross-mean-reverting effect, we assume that it can be set to zero due to the strong positive correlation between assets returns. Table 4.4 supports this assumption. This allows us to decrease the number of parameters by six. Table 4.5 reports the estimation of this shrinked version of model 1. Results are qualitatively similar to those for the full symmetric model. Quantitatively speaking, μ and α^{CTF} increase while β^{CTF} decreases. The increase in μ and α^{CTF} comes from the absorption of previous excitation effects. We note that the increase in α^{CTF} is much stronger than in μ. The estimated model is stable, in the sense that $\alpha^{\text{CTF}} < 1$ (see subsection 4.5.1).

Table 4.5: Estimation of the parameters of the symmetric and shrinked version of model 1.

	1	2
μ	0.065 (0.005)	0.031 (0.002)
α^{CTF}	0.509 (0.062)	0.181 (0.011)
β^{CTF}	1.208 (0.173)	0.95 (0.143)

Figure 4.8 depicts the average intraday correlation profile in this model obtained through simulation. It is flat as expected since price returns are stationary in this specification of the model. The average correlation is 0.5, which is close to the empirical value of correlation between tick time returns of AIRP.PA and TOTF.PA, 0.46.

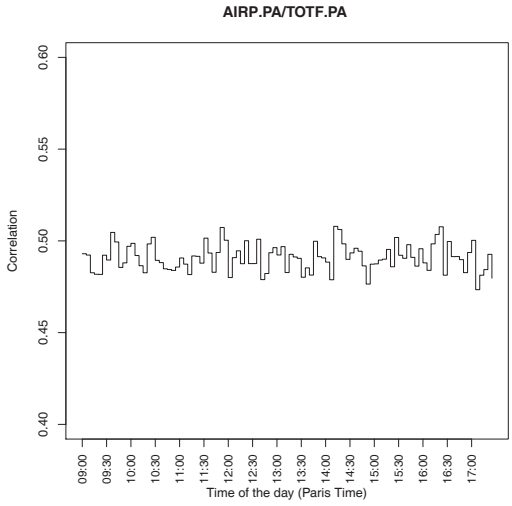

Figure 4.8: Intraday correlation profile for the symmetric and shrinked version of model 1.

Model 2

We present the results of the estimation of model 2 in table 4.6. We define the average background intensity

$$\begin{aligned}\overline{\mu}_i &= \frac{1}{T}\int_0^T \mu_i(t)\mathrm{d}t \\ &= \frac{1}{(d+1)T}\sum_{p=0}^{P}\mu_{i,p}(u_{p+d+1}-u_p)\end{aligned}$$

where $u_0, u_1, \ldots, u_{P+d+1}$ are the knots of the B-spline basis (see appendix 4.11).

Table 4.6: Estimation of the parameters of model 2.

	1	2
$\overline{\mu}$	0.083 (0.005)	0.04 (0.002)
α^{CTF}	0.145 (0.006)	0.085 (0.003)
β^{CTF}	3.75 (0.308)	3.073 (0.249)

Figure 4.9 plots the estimated background intensities as functions of time on the left panel. The shape of background intensities depicts the empirical pattern of trading activity as shown on the right panel. The fine size of the mesh of the knots grid (five minutes) allows us to capture sudden jumps due to news arrivals at 14:30 and 16:00 as well as the change of regime when the US market opens at 15:30. The conclusions about triggering kernels are qualitatively similar to those for model 1. It is however noteworthy that the excitation effect is smaller than in model 1. This is compensated by higher background intensities, as measured by the average background intensity. Estimated background intensities are far from being constant, thus making model 1 unrealistic. Nevertheless, the estimated μ can be substantially high at certain times of the day in order to capture these strong variations, which tends to bias α^{CTF} downwards.

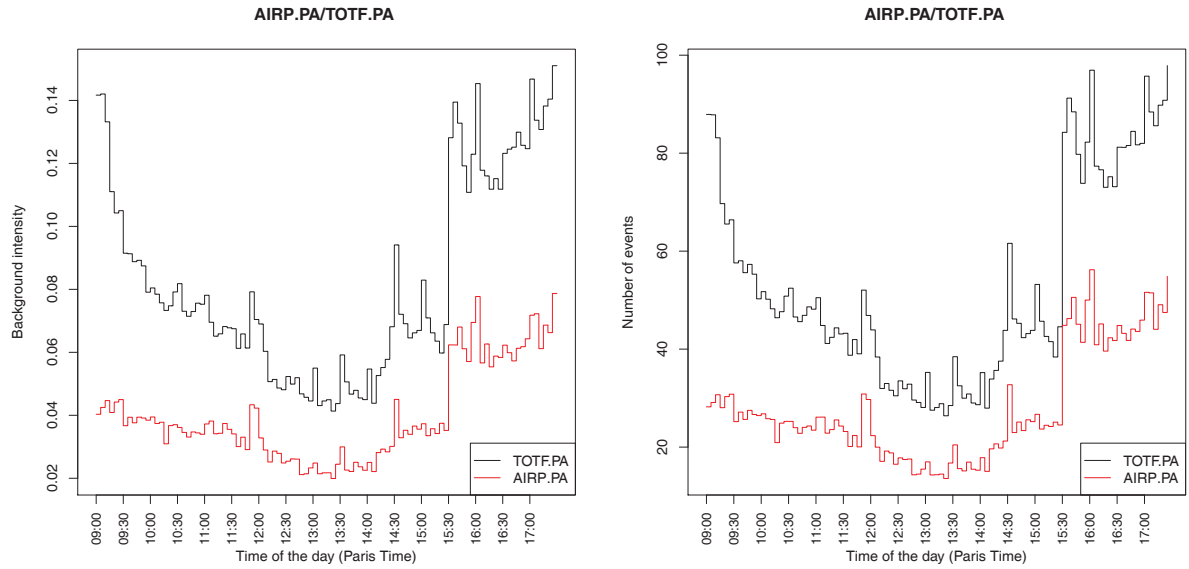

Figure 4.9: Left panel: estimated time-dependent μ_1, μ_2. Right panel: empirical intraday profile of the number of events.

The intraday correlation profile in model 2 is illustrated on figure 4.10. The correlation profile is found to be flat as in model 1. Its average value is much lower, 0.2, which comes from the significanlty smaller value of α^{CTF}. We think that the flatness of the correlation profile comes from the compensation between covariance and variance profiles. Indeed, both variance and covariance profiles follow the shape of background intensities in this model, which leaves correlation profile flat. Figure 4.10 illustrates covariance and variance profiles as well to support this reasoning.

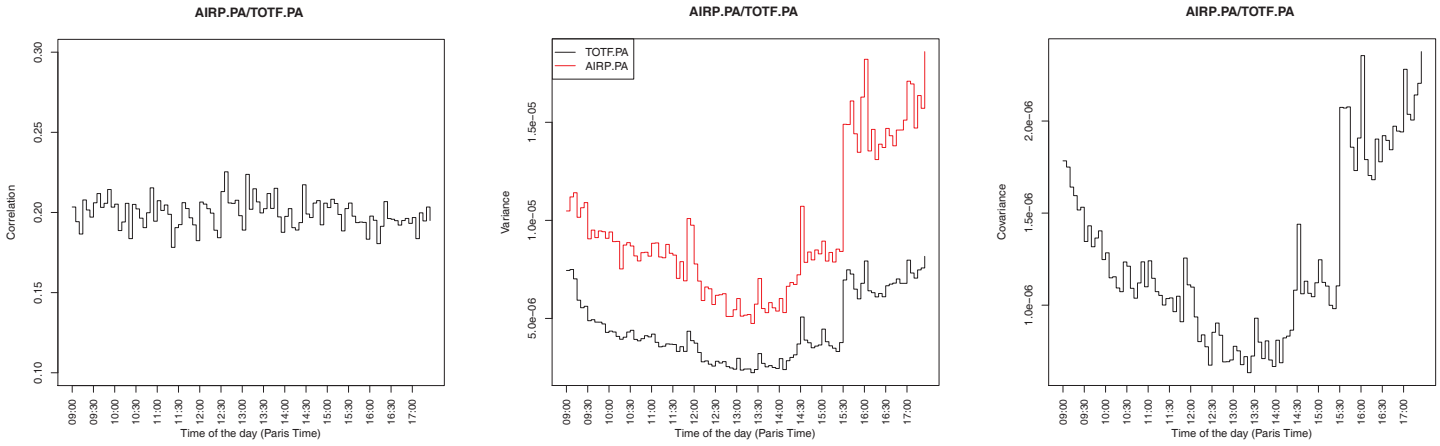

Figure 4.10: Left panel: intraday correlation profile in model 2. Middle panel: intraday variance profile in model 2. Right panel: intraday covariance profile in model 2.

Model 3

Figure 4.11 reports the results of the estimation of parameters in model 3. The values of μ_1 and μ_2 are closer to those in model 1 than in model 2, which is logical since models 1 and 3 assume constant background intensities. $\overline{\alpha}_2^{\mathrm{CTF}}$ is also close to its value in model 1. In the contrary, we observe that $\overline{\alpha}_1^{\mathrm{CTF}}$ is substantially bigger than in previous specifications of the model. It is higher than one and the function α_1^{CTF} is above five during the first 15 minutes of trading, which means that the resulting dynamics is highly unstable according to subsection 4.5.1. This overestimation of cross-excitation effects might come from the shortage of data with a mesh of five minutes. Therefore, we present the results when using meshes of 15 and 30 minutes on figures 4.12 and 4.13 respectively. The severe upward bias tends to vanish as we increase the mesh size. When it is 30 minutes we recover average results close to those for model 1.

Figure 4.14 plots the simulated variance, covariance and correlation intraday profiles as well as the empirical covariance profile. This model does not success in reproducing the empirical correlation pattern. The correlation pattern in the model is very similar to the covariance profile since variances don't vary much relatively to covariance in this model. It seems that the cross-excitation function α_1^{CTF} captures the variance profile of TOTF.PA while α_2^{CTF} mixes the patterns of empirical covariance and variance. Since α_2^{CTF} is much lower than α_1^{CTF}, the model fails in reproducing the empirical correlation intraday profile. Thus we think that we need to make background intensities and triggering kernels time-dependent to capture the empirical correlation pattern. We go in this direction in the next paragraph.

	1	2
μ	0.052 (0.004)	0.028 (0.002)
$\overline{\alpha}^{\text{CTF}}$	1.229 (0.251)	0.179 (0.006)
β^{CTF}	1.316 (0.578)	0.479 (0.086)

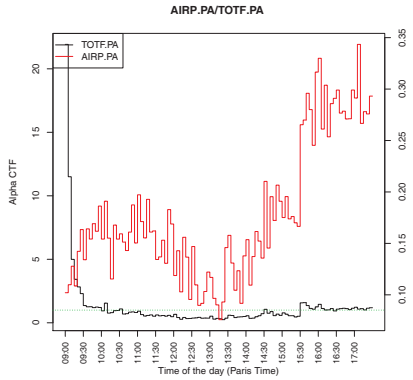

Figure 4.11: Estimation of parameters of the symmetric version of model 3 with a mesh of 5 minutes. Left panel: constant parameters. Right panel: time-dependent α_1^{CTF} (left scale), α_2^{CTF} (right scale).

	1	2
μ	0.056 (0.004)	0.029 (0.002)
$\overline{\alpha}^{\text{CTF}}$	0.629 (0.055)	0.171 (0.006)
β^{CTF}	0.623 (0.137)	0.556 (0.097)

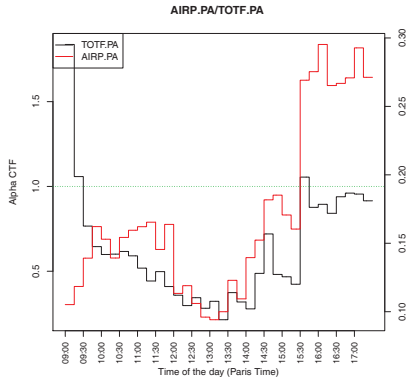

Figure 4.12: Estimation of the parameters of the symmetric version of model 3 with a mesh of 15 minutes. Left panel: constant parameters. Right panel: time-dependent α_1^{CTF} (left scale), α_2^{CTF} (right scale).

	1	2
μ	0.06 (0.004)	0.03 (0.002)
$\overline{\alpha}^{\text{CTF}}$	0.534 (0.044)	0.169 (0.007)
β^{CTF}	0.709 (0.138)	0.598 (0.104)

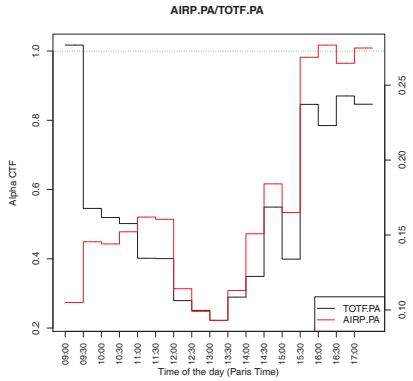

Figure 4.13: Estimation of the parameters of the symmetric version of model 3 with a mesh of 30 minutes. Left panel: constant parameters. Right panel: time-dependent α_1^{CTF} (left scale), α_2^{CTF} (right scale).

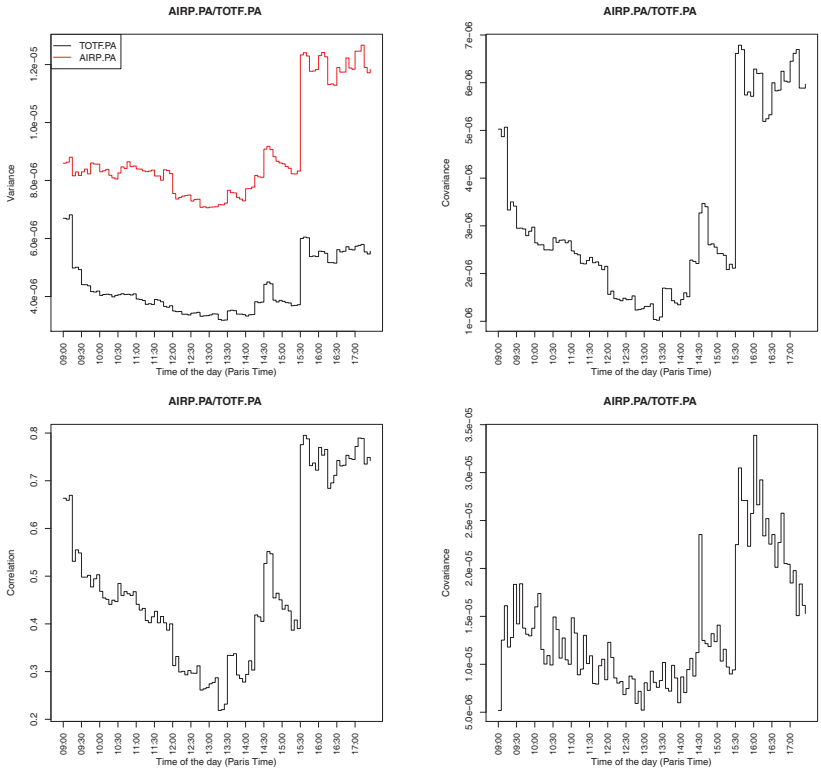

Figure 4.14: Top left panel: intraday variance profile in model 3 with a mesh of 15 minutes. Top right panel: intraday covariance profile in model 3 with a mesh of 15 minutes. Bottom left panel: intraday correlation profile in model 3 with a mesh of 15 minutes. Bottom right panel: empirical intraday covariance profile.

Model 4

So far we have seen that

- Model 1 displays flat variance, covariance and correlation profiles.
- Model 2 displays varying variance and covariance profiles but flat correlation.
- Model 3 displays a varying correlation profile similar to covariance.

In model 4, by specifying both μ and α^{CTF} as time-dependent functions, we think that a part of the variance pattern will be absorbed by background intensities, which allows triggering kernels to capture correlation. We show the estimated parameters with a mesh of 5 minutes on figure 4.15. As in model 2, making background intensities time-varying lowers cross-excitation effects. The shape of μ is identical to the one observed for model 2. Interestingly, cross-excitation functions α^{CTF} behave in a very similar way than the empirical correlation.

	1	2
$\overline{\mu}$	0.082 (0.005)	0.039 (0.002)
$\overline{\alpha}^{\text{CTF}}$	0.145 (0.006)	0.092 (0.003)
β^{CTF}	4.007 (0.562)	2.419 (0.215)

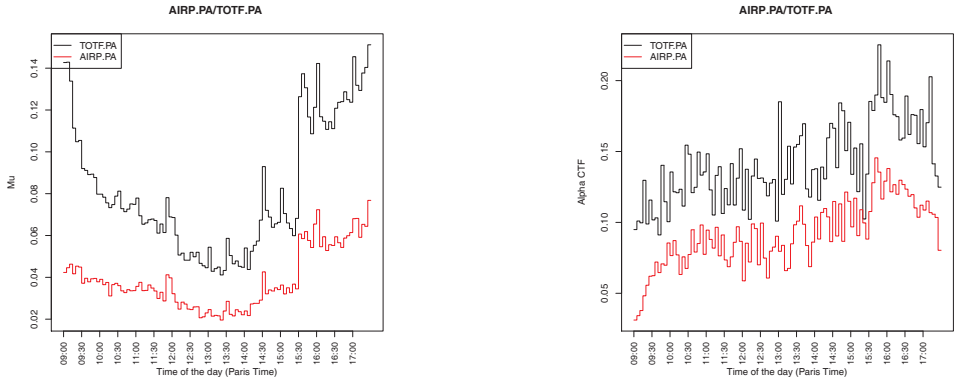

Figure 4.15: Estimation of the parameters of model 4 with a mesh of 5 minutes. Top panel: constant parameters. Bottom left panel: time-dependent μ_1, μ_2. Bottom right panel: time-dependent $\alpha_1^{\text{CTF}}, \alpha_2^{\text{CTF}}$.

On figures 4.16 and 4.17, we test the robustness of these results with respect to the mesh size by using meshes of 15 and 30 minutes respectively. Results are seen to be robust when increasing the mesh size. The decrease (resp. increase) in $\overline{\mu}$ (resp. $\overline{\alpha}^{\text{CTF}}$) is quite small, which is an issue to reach the average level of empirical correlation.

	1	2
$\overline{\mu}$	0.081 (0.005)	0.038 (0.002)
$\overline{\alpha}^{\text{CTF}}$	0.158 (0.006)	0.098 (0.004)
β^{CTF}	2.807 (0.248)	2.107 (0.198)

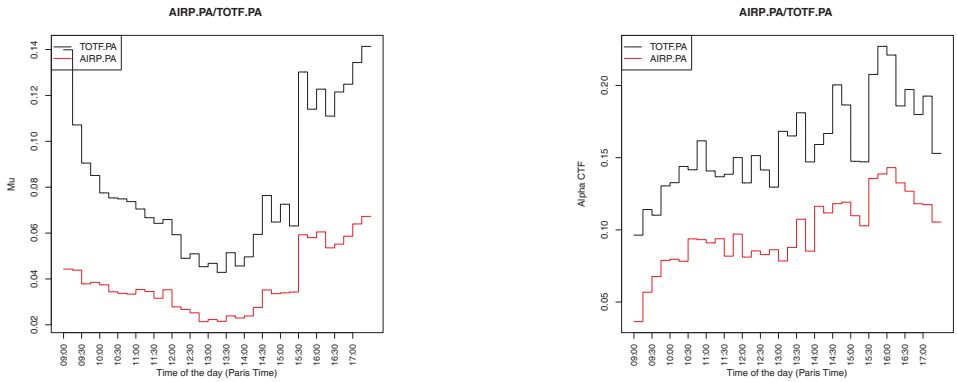

Figure 4.16: Estimation of the parameters of model 4 with a mesh of 15 minutes. Top panel: constant parameters. Bottom left panel: time-dependent μ_1, μ_2. Bottom right panel: time-dependent $\alpha_1^{\text{CTF}}, \alpha_2^{\text{CTF}}$.

	1	2
$\overline{\mu}$	0.081 (0.005)	0.038 (0.002)
$\overline{\alpha}^{\mathrm{CTF}}$	0.17 (0.007)	0.102 (0.004)
β^{CTF}	2.627 (0.252)	1.986 (0.193)

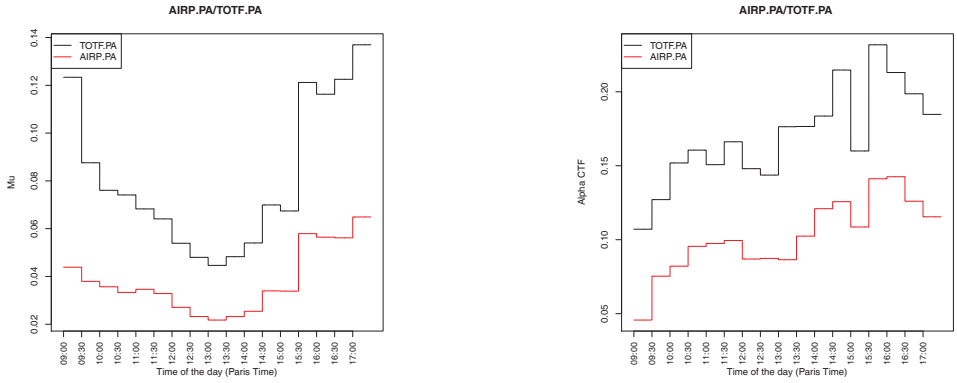

Figure 4.17: Estimation of the parameters of model 4 with a mesh of 30 minutes. Top panel: constant parameters. Bottom left panel: time-dependent μ_1, μ_2. Bottom right panel: time-dependent $\alpha_1^{\mathrm{CTF}}, \alpha_2^{\mathrm{CTF}}$.

We further test the scaling of parameters with the mesh size on figure 4.18 by increasing the mesh size up to the whole trading session. We naturally recover results of model 1 for the largest mesh. The scaling of parameters is monotonic: $\overline{\alpha}^{\mathrm{CTF}}$ increases while $\overline{\mu}$ and β^{CTF} decrease with the mesh size. There is a significant jump of parameters when moving from a mesh of 10200 (4 equally spaced knots) to 15300 (3 equally spaced knots) seconds. This might come from the sharp upward jumps in trading activity at 14:30, 15:30 and 16:00. When using three knots, all these jumps are merged within a single estimation of parameters while this is not the case with tighter meshes.

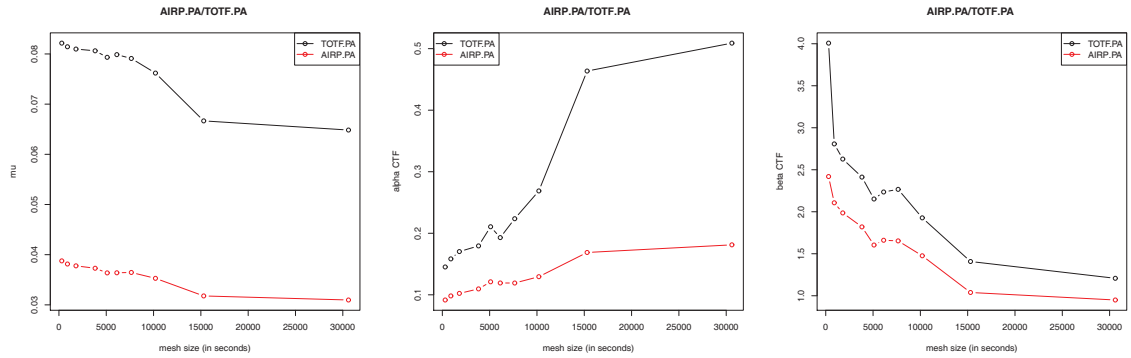

Figure 4.18: Scaling of the parameters of model 4 with the mesh size. Left panel: $\overline{\mu}_1, \overline{\mu}_2$. Middle panel: $\overline{\alpha}_1^{\mathrm{CTF}}, \overline{\alpha}_2^{\mathrm{CTF}}$. Right panel: $\beta_1^{\mathrm{CTF}}, \beta_2^{\mathrm{CTF}}$.

Figure 4.19 plots the simulated correlation profiles in model 4 with mesh sizes varying from five minutes to the whole trading session. It appears that the empirical correlation profile is quite well reproduced in this model. There is an upward trend in correlation with a jump when the US market opens. This jump is preceeded by a decrease in correlation as it is empirically seen. The model also succeeds in capturing the dive in correlation during the last hour of trading. We note that the upward trend of correlation in the

model is not as sharp as the empirical one. Furthermore, the average level of correlation in the model with a mesh of five minutes is 0.21, approximately twice less than its empirical counterpart. We have to increase the mesh size to 15300 seconds to recover realistic correlation values, but we fail to reproduce the flexibility of the correlation profile in this case. A potential solution to this issue might be adding an either endogenous or exogenous noise common to both assets in order to take into account the so-called market mode that is known to be responsible for a large part of correlation [23]. This would generate extra correlation that will add to the endogenous correlation created by cross-excitation effects.

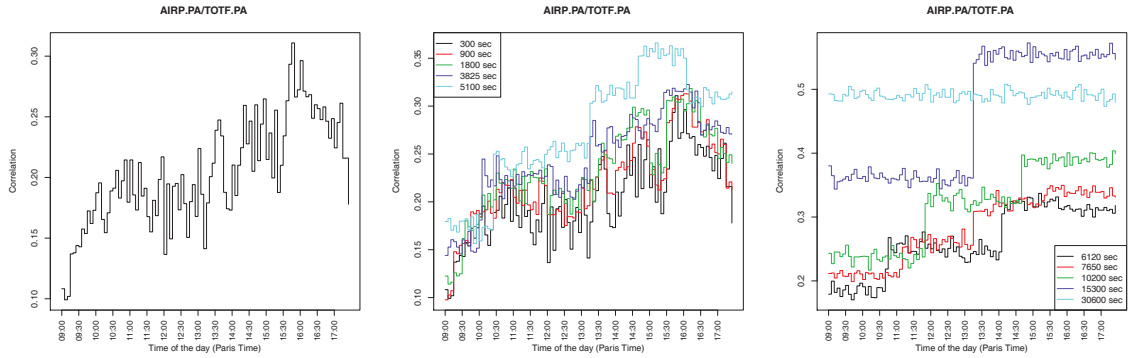

Figure 4.19: Intraday correlation profile in model 4. Left panel: Mesh size of 5 minutes. Middle panel: Mesh size from 300 to 5100 seconds. Right panel: Mesh size from 10200 to 30600 seconds.

4.6 Conclusion and further research

We study how high frequency correlation evolves during the trading session in equity markets. We use the Hayashi-Yoshida estimator to compute the intraday correlation profile. We find that it is highly varying and cannot be regarded as statistically constant. Correlation starts from very moderate values, close to zero. Then it gradually increases and jumps upwards when economic or financial news arrive. A substantial decline in correlation is observed during the last hour of trading. This pattern is found to be statistically significant and similar on four market venues. The idiosyncratic correlation, *i.e.* correlation once the market mode is removed, sharply falls during the first minutes of the trading session.
We use nonstationary Hawkes processes to build a model for prices at high frequency that captures the intraday correlation profile. We apply an EM algorithm to estimate parameters and a simulation algorithm to compute resulting correlation profiles. We find that we have to specifiy both background intensities and triggering kernels as time-dependent functions in order to reproduce the empirical correlation pattern. Though the correlation pattern is quite well fitted with our model, the level of correlation is substantially lower than the empirical one. Realistic levels of correlation are recovered as we flatten time-dependent parameters, but we lose flexibility in the correlation profile in return. We believe that the introduction of an either endogenous or exogenous noise common to both assets in our model could help to reach the level of empirical correlation while still capturing the empirical correlation profile. This is left for further research.
Another interesting direction for future research is to understand why correlation falls off during the last hour of trading. Though the increase in correlation as the day evolves can be related to the accumulation of information by traders, we have not found satisfying explanations (as well as statistical tests to support these explanations) for this decorrelation before market closes.

4.7 Appendix: Description of the stock universes

Table 4.7: Description of the scope of assets (FTSE universe).

RIC	Description	Exchange	Trading hours (Paris time)
AAL.L	Anglo American	London Stock Exchange	09:00-17:30
ANTO.L	Antofagasta	London Stock Exchange	09:00-17:30
AV.L	Aviva	London Stock Exchange	09:00-17:30
AZN.L	AstraZeneca	London Stock Exchange	09:00-17:30
BAES.L	BAE Systems	London Stock Exchange	09:00-17:30
BARC.L	Barclays	London Stock Exchange	09:00-17:30
BATS.L	British American Tobacco	London Stock Exchange	09:00-17:30
BG.L	BG Group	London Stock Exchange	09:00-17:30
BLT.L	Bhp Billiton	London Stock Exchange	09:00-17:30
BP.L	BP	London Stock Exchange	09:00-17:30
BT.L	BT Group	London Stock Exchange	09:00-17:30
EMG.L	Man Group	London Stock Exchange	09:00-17:30
GSK.L	GlaxoSmithKline	London Stock Exchange	09:00-17:30
HSBA.L	HSBC Holdings	London Stock Exchange	09:00-17:30
IMT.L	Imperial Tobacco Group	London Stock Exchange	09:00-17:30
LLOY.L	Lloyds Banking Group	London Stock Exchange	09:00-17:30
MKS.L	Marks And Spencer Group	London Stock Exchange	09:00-17:30
PRU.L	Prudential	London Stock Exchange	09:00-17:30
RB.L	Reckitt Benckiser Group	London Stock Exchange	09:00-17:30
RBS.L	Royal Bank of Scotland Group	London Stock Exchange	09:00-17:30
RDSb.L	Royal Dutch Shell	London Stock Exchange	09:00-17:30
RIO.L	Rio Tinto	London Stock Exchange	09:00-17:30
SAB.L	SABMiller	London Stock Exchange	09:00-17:30
STAN.L	Standard Chartered	London Stock Exchange	09:00-17:30
TSCO.L	TESCO	London Stock Exchange	09:00-17:30
ULVR.L	Unilever	London Stock Exchange	09:00-17:30
VED.L	Vedanta Resources	London Stock Exchange	09:00-17:30
VOD.L	Vodafone Group	London Stock Exchange	09:00-17:30
WPP.L	WPP	London Stock Exchange	09:00-17:30
XTA.L	Xstrata	London Stock Exchange	09:00-17:30
FFI	Footsie100 future	NYSE Liffe London	02:00-08:50, 09:00-22:00

Table 4.8: Summary statistics on the scope of assets (FTSE universe).

RIC	$\langle \Delta t \rangle$ (sec)	$\langle \delta / m \rangle$ (bp)	$\langle s \rangle / \delta$	$\left\langle \mathbb{1}_{\{s=\delta\}} \right\rangle$ (%)	$\left\langle \mathbb{1}_{\{\text{trade through}\}} \right\rangle$ (%)	$\langle \lvert \Delta m \rvert \rangle / \delta$	$\langle P_{\text{trade}} V_{\text{trade}} \rangle$ (GBP$\times 10^3$)
AAL.L	5.074	1.84	2.22	36	4	0.97	17
ANTO.L	15.332	6.79	1.46	65	2	0.54	10
AV.L	9.968	2.76	2.06	44	3	0.65	10
AZN.L	8.413	1.72	1.71	56	3	0.54	19
BAES.L	13.926	2.78	1.69	59	2	0.52	11
BARC.L	3.373	1.48	3.11	26	6	1.00	15
BATS.L	9.966	2.31	1.49	68	2	0.51	15
BG.L	10.081	4.15	1.53	78	2	0.53	17
BLT.L	3.804	2.39	1.45	67	3	0.55	18
BP.L	4.445	1.60	1.85	45	4	0.55	20
BT.L	15.260	8.04	1.12	90	1	0.28	12
EMG.L	14.114	4.15	1.70	60	3	0.57	7
GSK.L	10.102	4.07	1.19	87	2	0.29	22
HSBA.L	3.406	1.49	1.94	41	4	0.62	19
IMT.L	16.784	5.12	1.14	89	1	0.29	18
LLOY.L	4.225	1.62	3.62	21	5	1.05	11
MKS.L	14.336	2.80	1.80	55	3	0.59	9
PRU.L	11.767	8.70	1.28	89	1	0.34	19
RB.L	16.028	2.88	1.44	70	2	0.50	15
RBS.L	7.284	3.37	3.20	30	5	0.95	10
RDSb.L	8.609	2.72	1.26	81	2	0.44	15
RIO.L	3.432	1.41	2.63	28	6	0.97	20
SAB.L	17.601	5.16	1.19	85	1	0.33	17
STAN.L	7.900	2.90	1.57	63	2	0.60	16
TSCO.L	8.312	1.16	2.37	41	4	0.81	12
ULVR.L	16.370	5.18	1.15	88	1	0.32	18
VED.L	12.912	3.86	1.76	54	3	0.78	12
VOD.L	6.125	3.47	1.20	85	2	0.27	19
WPP.L	17.134	7.48	1.09	93	1	0.27	16
XTA.L	4.436	3.65	2.13	60	3	0.83	17
FFI	1.219	0.91	1.19	82	5	0.41	191

Table 4.9: Description of the scope of assets (NY universe).

RIC	Description	Exchange	Trading hours (Paris time)
AA.N	Alcoa Inc	New York Stock Exchange	15:30-22:00
AIG.N	American International Group Inc	New York Stock Exchange	15:30-22:00
AXP.N	American Express Co	New York Stock Exchange	15:30-22:00
BA.N	Boeing Co	New York Stock Exchange	15:30-22:00
BAC.N	Bank of America Corp	New York Stock Exchange	15:30-22:00
C.N	Citigroup Inc	New York Stock Exchange	15:30-22:00
CAT.N	Caterpillar Inc	New York Stock Exchange	15:30-22:00
CSCO.OQ	Cisco Systems Inc	NASDAQ	15:30-22:00
CVX.N	Chevron Corp	New York Stock Exchange	15:30-22:00
DD.N	E. I. Du Pont De Nemours And Co	New York Stock Exchange	15:30-22:00
DIS.N	Walt Disney Co	New York Stock Exchange	15:30-22:00
DOW.N	Dow Chemical Co	New York Stock Exchange	15:30-22:00
F.N	Ford Motor Co	New York Stock Exchange	15:30-22:00
GE.N	General Electric Company	New York Stock Exchange	15:30-22:00
HD.N	Home Depot Inc	New York Stock Exchange	15:30-22:00
HPQ.N	Hewlett Packard Co	New York Stock Exchange	15:30-22:00
IBM.N	International Business Machines Corp	New York Stock Exchange	15:30-22:00
INTC.OQ	Intel Corp	NASDAQ	15:30-22:00
JNJ.N	Johnson & Johnson	New York Stock Exchange	15:30-22:00
JPM.N	JPMorgan Chase & Co	New York Stock Exchange	15:30-22:00
KFT.N	Kraft Foods Inc	New York Stock Exchange	15:30-22:00
KMB.N	Kimberly Clark Corp	New York Stock Exchange	15:30-22:00
KO.N	Coca Cola Co	New York Stock Exchange	15:30-22:00
MCD.N	McDonalds Corp	New York Stock Exchange	15:30-22:00
MMM.N	3M Co	New York Stock Exchange	15:30-22:00
MRK.N	Merck & Co Inc	New York Stock Exchange	15:30-22:00
MS.N	Morgan Stanley	New York Stock Exchange	15:30-22:00
MSFT.OQ	Microsoft Corp	NASDAQ	15:30-22:00
NKE.N	Nike Inc	New York Stock Exchange	15:30-22:00
PEP.N	Pepsico Inc	New York Stock Exchange	15:30-22:00
PFE.N	Pfizer Inc	New York Stock Exchange	15:30-22:00
PG.N	Procter & Gamble Co	New York Stock Exchange	15:30-22:00
T.N	AT&T Inc	New York Stock Exchange	15:30-22:00
TRV.N	Travelers Companies Inc	New York Stock Exchange	15:30-22:00
TWX.N	Time Warner Inc	New York Stock Exchange	15:30-22:00
TXN.N	Texas Instruments Inc	New York Stock Exchange	15:30-22:00
UTX.N	United Technologies Corp	New York Stock Exchange	15:30-22:00
VZ.N	Verizon Communications Inc	New York Stock Exchange	15:30-22:00
WMT.N	Wal-Mart Stores Inc	New York Stock Exchange	15:30-22:00
XOM.N	Exxon Mobil Corp	New York Stock Exchange	15:30-22:00
YM	E-mini Dow future	Chicago Board Of Trade	00:00-22:15, 22:30-23:30

Table 4.10: Summary statistics on the scope of assets (NY universe).

RIC	$\langle \Delta t \rangle$ (sec)	$\langle \delta/m \rangle$ (bp)	$\langle s \rangle / \delta$	$\langle \mathbb{1}_{\{s=\delta\}} \rangle$ (%)	$\langle \mathbb{1}_{\{\text{trade through}\}} \rangle$ (%)	$\langle \lvert \Delta m \rvert \rangle / \delta$	$\langle P_{\text{trade}} V_{\text{trade}} \rangle$ (USD$\times 10^3$)
AA.N	6.031	7.51	1.02	98	0	0.16	12
AIG.N	4.413	2.78	2.49	30	4	1.06	10
AXP.N	4.834	2.37	1.31	80	1	0.52	13
BA.N	4.609	1.43	2.26	25	3	0.76	18
BAC.N	2.869	5.79	1.02	98	1	0.12	35
C.N	3.448	24.24	1.00	100	0	0.02	34
CAT.N	4.186	1.58	1.93	41	2	0.74	16
CSCO.OQ	3.341	3.87	1.02	96	1	0.16	38
CVX.N	3.365	1.30	1.63	42	2	0.60	20
DD.N	6.083	2.67	1.23	85	1	0.44	12
DIS.N	4.825	2.89	1.08	95	1	0.26	14
DOW.N	6.240	3.42	1.23	85	1	0.46	11
F.N	4.287	7.82	1.02	98	0	0.14	20
GE.N	3.513	5.63	1.02	98	1	0.12	24
HD.N	4.932	2.98	1.10	93	1	0.26	16
HPQ.N	4.491	1.95	1.20	89	1	0.38	20
IBM.N	3.576	0.78	2.59	12	3	0.84	25
INTC.OQ	3.084	4.52	1.01	95	1	0.15	38
JNJ.N	4.977	1.56	1.24	63	1	0.35	24
JPM.N	2.609	2.33	1.17	89	1	0.36	21
KMB.N	9.664	1.63	1.65	64	1	0.61	14
KO.N	5.387	1.86	1.24	85	1	0.37	20
KFT.N	7.237	3.35	1.08	96	0	0.22	15
MCD.N	5.189	1.47	1.42	49	1	0.46	20
MMM.N	5.330	1.20	2.21	25	2	0.77	18
MRK.N	4.734	2.81	1.15	91	1	0.31	16
MS.N	4.403	3.43	1.16	89	1	0.39	12
MSFT.OQ	2.902	3.42	1.01	96	1	0.16	41
NKE.N	8.052	1.36	2.39	23	2	0.94	15
PEP.N	4.828	1.53	1.48	50	2	0.48	19
PFE.N	4.522	5.96	1.01	99	0	0.08	26
PG.N	4.937	1.59	1.25	85	1	0.39	21
T.N	5.336	3.89	1.03	98	0	0.14	22
TRV.N	6.492	1.93	1.37	78	1	0.46	14
TWX.N	6.918	3.18	1.17	91	1	0.36	13
TXN.N	6.326	3.97	1.12	92	1	0.33	12
UTX.N	6.084	1.38	1.78	38	2	0.66	18
VZ.N	5.713	3.41	1.05	97	1	0.17	18
WMT.N	3.986	1.85	1.13	91	1	0.25	22
XOM.N	2.528	1.51	1.23	60	1	0.35	27
YM	0.850	0.93	1.09	93	6	0.32	198

Table 4.11: Description of the scope of assets (TOPIX universe).

RIC	Description	Exchange	Trading hours (Paris time)[16]
2914.T	Japan Tobacco Inc	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
3382.T	Seven & I Holdings Co Ltd	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
4063.T	Shin Etsu Chemical Co Ltd	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
4502.T	Takeda Pharmaceutical Co Ltd	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
4503.T	Astellas Pharma Inc	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
5401.T	Nippon Steel Corp	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
5411.T	JFE Holdings Inc	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
6301.T	Komatsu Ltd	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
6502.T	Toshiba Corp	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
6752.T	Panasonic Corp	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
6758.T	Sony Corp	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
7201.T	Nissan Motor Co Ltd	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
7203.T	Toyota Motor Corp	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
7267.T	Honda Motor Co Ltd	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
7751.T	Canon Inc	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
8031.T	Mitsui & Co Ltd	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
8058.T	Mitsubishi Corp	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
8306.T	Mitsubishi UFJ Financial Group Inc	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
8316.T	Sumitomo Mitsui Financial Group Inc	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
8411.T	Mizuho Financial Group Inc	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
8604.T	Nomura Holdings Inc	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
8766.T	Tokio Marine Hldgs Inc	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
8802.T	Mitsubishi Estate Co Ltd	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
9020.T	East Japan Railway Co	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
9432.T	Nippon Telegraph And Telephone Corp	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
9433.T	Kddi Corp	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
9437.T	NTT Docomo Inc	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
9501.T	Tokyo Electric Power Co Incorporated	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
9503.T	Kansai Electric Power Co Inc	Tokyo Stock Exchange	02:00-04:00, 05:30-08:00
2JTI	Topix future	Tokyo Stock Exchange	02:00-04:00, 05:30-08:10

[16]The trading hours are those that applied during the time period March-May 2010. They have been extended since then. See `http://www.tse.or.jp/english/` for further details. There is also an evening trading session for the TOPIX future taking place from $09:30$ to $12:00$ (it has been extended too nowadays). This corresponds to another RIC in our database, 1JTI namely.

Table 4.12: Summary statistics on the scope of assets (TOPIX universe).

RIC	$\langle \Delta t \rangle$ (sec)	$\langle \delta/m \rangle$ (bp)	$\langle s \rangle / \delta$	$\left\langle \mathbb{1}_{\{s=\delta\}} \right\rangle$ (%)	$\left\langle \mathbb{1}_{\{\text{trade through}\}} \right\rangle$ (%)	$\langle \lvert \Delta m \rvert \rangle / \delta$	$\langle P_{\text{trade}} V_{\text{trade}} \rangle$ (JPY$\times 10^3$)
2914.T	14.967	13.57	1.07	94	1	0.11	3452
3382.T	9.098	4.52	1.10	91	1	0.21	2522
4063.T	17.858	16.28	1.05	95	0	0.08	4772
4502.T	12.719	12.42	1.00	100	0	0.03	4098
4503.T	19.236	13.71	1.07	95	1	0.10	3354
5401.T	14.980	29.20	1.00	99	0	0.03	4687
5411.T	12.018	12.93	1.15	93	1	0.13	3984
6301.T	5.785	5.45	1.05	95	1	0.16	2251
6502.T	7.032	20.77	1.00	100	0	0.03	5570
6752.T	6.633	7.59	1.03	97	1	0.10	2235
6758.T	5.976	12.47	1.11	94	1	0.09	4396
7201.T	6.653	13.26	1.01	99	0	0.05	2997
7203.T	5.918	14.05	1.00	100	0	0.03	6763
7267.T	8.577	12.52	1.24	91	1	0.11	5127
7751.T	8.713	12.22	1.01	99	0	0.06	6197
8031.T	4.905	6.83	1.03	97	1	0.12	2819
8058.T	4.986	4.42	1.07	93	1	0.15	2756
8306.T	4.126	21.06	1.00	100	0	0.01	4508
8316.T	4.193	8.05	1.14	92	2	0.14	5264
8411.T	4.838	55.72	1.00	93	0	0.00	3556
8604.T	5.317	15.50	1.00	100	0	0.03	3331
8766.T	9.845	3.79	1.23	82	2	0.32	2172
8802.T	17.201	6.60	1.13	88	2	0.30	4973
9020.T	21.411	15.97	1.00	100	0	0.07	4895
9432.T	13.890	12.94	1.00	100	0	0.04	4313
9433.T	18.172	10.87	1.01	99	0	0.07	3892
9437.T	8.258	7.03	1.01	99	0	0.06	2611
9501.T	10.698	4.17	1.08	93	1	0.18	2319
9503.T	15.012	4.75	1.08	92	1	0.20	1930
2JTI	4.736	5.31	1.00	99	0	0.06	63541

4.8 Appendix: Proof of the formula for the idiosyncratic correlation

We consider the CAPM model [93]

$$r^i = \alpha^i + \beta^i r^{\text{market}} + \varepsilon^i$$

Let us compute the covariance between the idiosyncratic returns of two assets

$$\begin{aligned}
\mathbb{C}\text{ov}(\varepsilon^i, \varepsilon^j) =& \mathbb{C}\text{ov}(r^i - \alpha^i - \beta^i r^{\text{market}}, r^j - \alpha^j - \beta^j r^{\text{market}}) \\
=& \mathbb{C}\text{ov}(r^i, r^j) - \beta^j \mathbb{C}\text{ov}(r^i, r^{\text{market}}) - \beta^i \mathbb{C}\text{ov}(r^j, r^{\text{market}}) + \beta^i \beta^j \mathbb{V}\text{ar}(r^{\text{market}}) \\
=& \mathbb{C}\text{ov}(r^i, r^j) - \frac{\mathbb{C}\text{ov}(r^i, r^{\text{market}}) \mathbb{C}\text{ov}(r^j, r^{\text{market}})}{\mathbb{V}\text{ar}(r^{\text{market}})}
\end{aligned}$$

since $\beta^i = \frac{\mathbb{C}\text{ov}(r^i, r^{\text{market}})}{\mathbb{V}\text{ar}(r^{\text{market}})}$. Let us note $\sigma^i = \sqrt{\mathbb{V}\text{ar}(r^i)}$, $\rho^{i,j} = \mathbb{C}\text{orr}(r^i, r^j)$. Then,

$$\mathbb{C}\text{ov}(\varepsilon^i, \varepsilon^j) = \sigma^i \sigma^j (\rho^{i,j} - \rho^{i,\text{market}} \rho^{j,\text{market}})$$

Taking $j = i$, we get

$$\mathbb{V}\text{ar}(\varepsilon^i) = (\sigma^i)^2 (1 - (\rho^{i,\text{market}})^2)$$

As a result,

$$\begin{aligned}
\mathbb{C}\text{orr}(\varepsilon^i, \varepsilon^j) &= \frac{\mathbb{C}\text{ov}(\varepsilon^i, \varepsilon^j)}{\sqrt{\mathbb{V}\text{ar}(\varepsilon^i) \mathbb{V}\text{ar}(\varepsilon^j)}} \\
&= \frac{\rho^{i,j} - \rho^{i,\text{market}} \rho^{j,\text{market}}}{\sqrt{(1 - (\rho^{i,\text{market}})^2)(1 - (\rho^{j,\text{market}})^2)}}
\end{aligned}$$

4.9 Appendix: Comparison between two specifications of the exponential kernel

We compare two specifications of the exponential triggering kernel of a Hawkes process. The two possibilities are $\phi_{\text{Classic}}(x) = \alpha e^{-\beta x} \mathbb{1}_{\mathbb{R}^+}(x)$ and $\phi_{\text{Proportional}}(x) = \beta \phi_{\text{Classic}}(x)$. We simulate 1000 trajectories of a Hawkes process with 5000 points on average for each trajectory. This is done by finding the time horizon T such that $\mathbb{E}(N_T | \mu, \alpha, \beta) = \frac{\mu}{1-\alpha/\beta}(T + \frac{e^{-(\beta-\alpha)T}-1}{\beta-\alpha}) = 5000$. The default parameters of the simulations are $\mu = 1$, $\alpha = 0.5$ and $\beta = 1$. We use the MLE approach to estimate the parameters. The `nlminb` function of `R` with default parameters is called to maximize the log-likelihood. The starting parameters $\theta^0 = (\mu^0, \alpha^0, \beta^0)$ are the average of lower and upper bounds, respectively $(0.001/\langle \Delta t \rangle, 0.001, 0.001)$ and $(100/\langle \Delta t \rangle, 100, 100)$. We report the median estimates along with standard deviations[17] rather than the average because the probability distribution of the estimated parameters is strongly right-skewed. This is due to the fact that starting

[17] We follow the approximation of [81] to compute a confidence interval on the sample median. The approximate 95% confidence interval is $\left[m \pm 1.58 \frac{\text{IQR}}{\sqrt{n}}\right]$ where m is the sample median and IQR is the sample interquartile range. Therefore the approximate standard deviation of the sample median is $\frac{1.58}{1.96} \frac{\text{IQR}}{\sqrt{n}} \approx 0.8 \frac{\text{IQR}}{\sqrt{n}}$.

parameters are far greater than the true ones and that the numerical optimizer stays stuck close to the starting parameters on some trajectories. The results are reported in table 4.13. Clearly, for small β the proportional specification yields far better results than the classic one.

In order to understand where this striking difference comes from, let us write the log-likelihood for both specifications

$$\ell_{\text{Classic}}(t_1,\dots,t_n;\mu,\alpha,\beta) = \sum_{i=1}^{n} \ln(\mu + \alpha \sum_{j=1}^{i-1} e^{-\beta(t_i-t_j)}) - \mu T - \frac{\alpha}{\beta}\sum_{i=1}^{n}(1-e^{-\beta(T-t_i)})$$
$$\ell_{\text{Proportional}}(t_1,\dots,t_n;\mu,\alpha,\beta) = \sum_{i=1}^{n} \ln(\mu + \alpha\beta \sum_{j=1}^{i-1} e^{-\beta(t_i-t_j)}) - \mu T - \alpha\sum_{i=1}^{n}(1-e^{-\beta(T-t_i)})$$

which gives the following derivatives w.r.t. β

$$\frac{\partial \ell_{\text{Classic}}}{\partial \beta} = -\alpha\left(\sum_{i=1}^{n}\frac{\sum_{j=1}^{i-1}(t_i-t_j)e^{-\beta(t_i-t_j)}}{\mu+\alpha\sum_{j=1}^{i-1}e^{-\beta(t_i-t_j)}} + \frac{1}{\beta}\sum_{i=1}^{n}(T-t_i)e^{-\beta(T-t_i)} - \frac{1}{\beta^2}\sum_{i=1}^{n}(1-e^{-\beta(T-t_i)})\right)$$
$$\frac{\partial \ell_{\text{Proportional}}}{\partial \beta} = \alpha\left(\sum_{i=1}^{n}\frac{\sum_{j=1}^{i-1}e^{-\beta(t_i-t_j)} - \beta\sum_{j=1}^{i-1}(t_i-t_j)e^{-\beta(t_i-t_j)}}{\mu+\alpha\beta\sum_{j=1}^{i-1}e^{-\beta(t_i-t_j)}} - \sum_{i=1}^{n}(T-t_i)e^{-\beta(T-t_i)}\right)$$

We think that the poor performance of the estimation when β is small in the classic setting comes from the term in $\frac{1}{\beta^2}$, which goes to infinity when β goes to zero. Thanks to the proportional specification, this term disappears in the log-likelihood.

	$\beta = 0.01$	$\beta = 0.1$	$\beta = 1$	$\beta = 10$
Classic	100 (0.000)	99.995 (2.546)	1.009 (0.003)	10.004 (0.013)
Proportional	0.009 (0.000)	0.093 (0.001)	1.007 (0.003)	9.939 (0.016)

Table 4.13: Comparison between the two specifications of the exponential kernel. The standard deviation of the estimates (computed over simulations) is indicated in brackets.

4.10 Appendix: Comparison between the MLE and EM estimators

We test the accuracy of the MLE and EM algorithms in a simulation framework. The simulation settings are identical to those presented in appendix 4.9. The maximum number of iterations of the EM algorithm is set to $k_{\max} = 150$ and the tolerance to $\epsilon = 1.5 \times 10^{-8}$ since these are the default settings of the MLE numerical optimizer. The results of the estimation are reported in table 4.14 and the computation time required in table 4.15.

In most cases, the MLE and EM algorithms provide satisfactory results and perform similarly . This is as expected because the EM estimate converges towards the MLE. We note that the EM algorithm runs slighlty faster than the MLE in terms of overall CPU time. This comes from two competiting effects. On the one hand, one iteration of EM takes roughly 4.6 times less time than a MLE iteration. On the other hand, the EM algorithm needs 4 times more iterations to converge. There are two cases where the MLE performs better than the EM estimate: $\mu = 10$ and $\beta = 0.1$. These are two extreme cases, the first begin a very exogenously active process, while the second describes a process with a very slowly decaying self-excitation.

In both cases, the EM algorithm reaches its maximum number of iterations, which suggests that it has not converged yet. Thus the solution could be increasing the number of EM iterations. We show the results when we increase the maximum number of EM iterations to $k_{\max} = 500$ in tables 4.16 and 4.17. The results for $\mu = 10$ and $\beta = 0.1$ are then much closer to the MLE. However, the execution time becomes significantly larger than for the MLE, respectively 2.6 and 4.7 times more.

	$\mu = 0.1$	$\mu = 1$	$\mu = 10$
MLE	0.100 (0.000)	1.001 (0.002)	10.228 (0.047)
EM	0.100 (0.000)	1.006 (0.002)	15.866 (0.048)
	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.9$
MLE	0.006 (0.001)	0.096 (0.001)	0.894 (0.001)
EM	0.006 (0.000)	0.073 (0.001)	0.890 (0.001)
	$\beta = 0.1$	$\beta = 1$	$\beta = 10$
MLE	0.093 (0.001)	1.007 (0.003)	9.939 (0.016)
EM	1.526 (0.035)	1.017 (0.003)	10.008 (0.013)

Table 4.14: Comparison between the MLE and EM algorithms. The standard deviation of the estimates (computed over simulations) is indicated in brackets.

	$\mu = 0.1$	$\mu = 1$	$\mu = 10$
k MLE	44 (0.3)	34 (0.2)	33 (0.1)
k EM	32 (0.1)	148 (0.3)	150 (0.0)
CPU MLE	540 (3.7)	400 (2.6)	303 (0.9)
CPU EM	85 (0.2)	347 (1.0)	315 (0.9)
CPU/k MLE	12.4 (0.0)	11.7 (0.0)	9.1 (0.0)
CPU/k EM	2.7 (0.0)	2.4 (0.0)	2.1 (0.0)
	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.9$
k MLE	35 (0.3)	28 (0.2)	29 (0.1)
k EM	150 (0.5)	150 (0.0)	150 (0.0)
CPU MLE	418 (4.7)	329 (3.1)	304 (3.4)
CPU EM	395 (1.7)	401 (0.5)	243 (2.4)
CPU/k MLE	12 (0.0)	11.9 (0.1)	10.3 (0.1)
CPU/k EM	2.7 (0.0)	2.7 (0.0)	1.6 (0.0)
	$\beta = 0.1$	$\beta = 1$	$\beta = 10$
k MLE	26 (0.2)	34 (0.2)	22 (0.3)
k EM	150 (0.0)	148 (0.3)	29 (0.0)
CPU MLE	201 (1.3)	400 (2.6)	295 (2.6)
CPU EM	393 (0.9)	347 (1.0)	81 (0.2)
CPU/k MLE	8.3 (0.0)	11.7 (0.0)	12.7 (0.1)
CPU/k EM	2.6 (0.0)	2.4 (0.0)	2.8 (0.0)

Table 4.15: Comparison between the computation times of the MLE and EM algorithms. The CPU time is expressed in seconds. The standard deviation of the estimates (computed over simulations) is indicated in brackets.

	$\mu = 0.1$	$\mu = 1$	$\mu = 10$
EM	0.100 (0.000)	1.004 (0.002)	11.244 (0.062)
	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.9$
EM	0.007 (0.000)	0.097 (0.001)	0.893 (0.001)
	$\beta = 0.1$	$\beta = 1$	$\beta = 10$
EM	0.186 (0.003)	1.016 (0.003)	10.008 (0.013)

Table 4.16: Results of the estimation using the EM algorithm with $k_{\max} = 500$. The standard deviation of the estimates (computed over simulations) is indicated in brackets.

	$\mu = 0.1$	$\mu = 1$	$\mu = 10$
k EM	32 (0.1)	148 (0.4)	500 (0.0)
CPU EM	83 (0.2)	341 (1.2)	783 (2.0)
CPU/k EM	2.6 (0.0)	2.3 (0.0)	1.6 (0.0)
	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.9$
k EM	194 (4.7)	334 (3.6)	186 (0.7)
CPU EM	490 (12.1)	833 (8.7)	279 (3.3)
CPU/k EM	2.6 (0.0)	2.5 (0.0)	1.5 (0.0)
	$\beta = 0.1$	$\beta = 1$	$\beta = 10$
k EM	500 (0.0)	148 (0.4)	29 (0.0)
CPU EM	945 (3.7)	341 (1.2)	79 (0.2)
CPU/k EM	1.9 (0.0)	2.3 (0.0)	2.7 (0.0)

Table 4.17: Computation time of the EM algorithm with $k_{\max} = 500$. The CPU time is expressed in seconds. The standard deviation of the estimates (computed over simulations) is indicated in brackets.

4.11 Appendix: Some details on the B-spline functions

B-spline functions are polynomial functions on a finite support defined with knots $u_0 \leq u_1 \leq \ldots \leq u_n$. The B-spline functions of degree d form a family of functions $N_{k,d}, 0 \leq k \leq n-d-1$ that are recursively defined as follows

$$N_{k,d}(t) = \frac{t-u_k}{u_{k+d}-u_k} N_{k,d-1}(t) + \frac{u_{k+d+1}-t}{u_{k+d+1}-u_{k+1}} N_{k+1,d-1}(t) \qquad 1 \leq d \leq n-1$$
$$N_{k,0}(t) = \mathbb{1}_{[u_k, u_{k+1})}(t)$$

with the convention $\frac{0}{0} = 0$. Figure 4.20 plots the cubic B-spline basis $N_{k,3}, 0 \leq k \leq n-4$ with knots $0, 0, 0, 0, 0.1, 0.2, \ldots, 0.9, 1, 1, 1, 1$, which corresponds to $n = 16$.

In the estimation of the time-dependent parameters of the model introduced in subsection 4.5.1, we have to compute three integrals involving B-spline functions. These three integrals are

$$\int_0^T N_{k,d}(t)\mathrm{d}t$$
$$\sum_{i=1}^n \int_{t_i}^T e^{-\beta(t-t_i)} N_{k,d}(t)\mathrm{d}t$$
$$\sum_{i=1}^n \int_{t_i}^T e^{-\beta(t-t_i)} (t-t_i) N_{k,d}(t)\mathrm{d}t$$

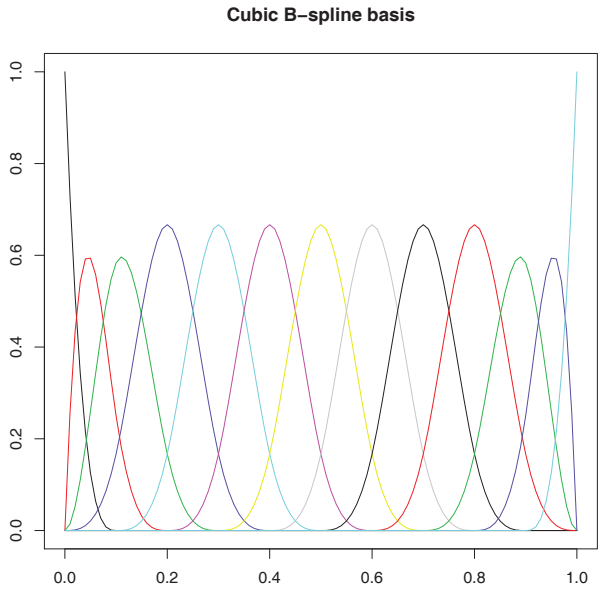

Figure 4.20: Cubic B-spline basis computed with the `splines` package of `R`.

Regarding the integral $\int_0^T N_{k,d}(t)\mathrm{d}t$, we use the formula of [46], page 9, and we get

$$\int_0^T N_{k,d}(t)\mathrm{d}t = \frac{u_{k+d+1} - u_k}{d+1}$$

We now consider the integral $I(k,d) = \sum_{i=1}^n \int_{t_i}^T e^{-\beta(t-t_i)} N_{k,d}(t)\mathrm{d}t$. Integrating by parts yields

$$\begin{aligned} I(k,d) &= \frac{1}{\beta}\left(\sum_{i=1}^n (N_{k,d}(t_i) - N_{k,d}(T)e^{-\beta(T-t_i)}) + \sum_{i=1}^n \int_{t_i}^T e^{-\beta(t-t_i)} N'_{k,d}(t)\mathrm{d}t\right) \\ &= \frac{1}{\beta}\left(\sum_{i=1}^n (N_{k,d}(t_i) - N_{k,d}(T)e^{-\beta(T-t_i)}) + d\left(\frac{I(k,d-1)}{u_{k+d} - u_k} - \frac{I(k+1,d-1)}{u_{k+d+1} - u_{k+1}}\right)\right) \end{aligned}$$

where we have used the following relationship (see [89])

$$N'_{k,d}(t) = d\left(\frac{N_{k,d-1}(t)}{u_{k+d} - u_k} - \frac{N_{k+1,d-1}(t)}{u_{k+d+1} - u_{k+1}}\right)$$

Therefore $I(k,d)$ can be computed recursively with initial condition

$$I(k,0) = \frac{1}{\beta}\sum_{i=1}^n \left(e^{-\beta(\max(u_k,t_i)-t_i)} - e^{-\beta(u_{k+1}-t_i)}\right)\mathbb{1}_{\{t_i<u_{k+1}\}}$$

In the same fashion, we define $J(k,d) = \sum_{i=1}^n \int_{t_i}^T e^{-\beta(t-t_i)}(t-t_i)N_{k,d}(t)\mathrm{d}t$. We have

$$\begin{aligned} J(k,d) =& \frac{1}{\beta^2}\sum_{i=1}^n \left(N_{k,d}(t_i) - N_{k,d}(T)e^{-\beta(T-t_i)}(1+\beta(T-t_i))\right) \\ &+ \frac{d}{\beta^2}\left(\frac{I(k,d-1)}{u_{k+d} - u_k} - \frac{I(k+1,d-1)}{u_{k+d+1} - u_{k+1}}\right) + \frac{d}{\beta}\left(\frac{J(k,d-1)}{u_{k+d} - u_k} - \frac{J(k+1,d-1)}{u_{k+d+1} - u_{k+1}}\right) \end{aligned}$$

which can also be computed recursively with initial condition

$$J(k,0) = \sum_{i=1}^n \left(\frac{e^{-\beta(\max(u_k,t_i)-t_i)} - e^{-\beta(u_{k+1}-t_i)}}{\beta^2} - \frac{(u_{k+1}-t_i)e^{-\beta(u_{k+1}-t_i)} - (\max(u_k,t_i)-t_i)e^{-\beta(\max(u_k,t_i)-t_i)}}{\beta}\right)\mathbb{1}_{\{t_i<u_{k+1}\}}$$

Bibliography

[1] F. Abergel, A. Chakraborti, I. Muni Toke, and M. Patriarca. Econophysics review: i. Empirical facts. *Quantitative Finance*, 11(7):991–1012, 2011.

[2] F. Abergel, A. Chakraborti, I. Muni Toke, and M. Patriarca. Econophysics review: ii. Agent-based models. *Quantitative Finance*, 11(7):1013–1041, 2011.

[3] F. Abergel and N. Huth. High frequency lead/lag relationships. *Forthcoming*, 2012.

[4] F. Abergel and N. Huth. Intraday correlation pattern. *Forthcoming*, 2012.

[5] F. Abergel and N. Huth. The times change: Multivariate subordination. Empirical facts. *Quantitative Finance*, 12(1):1–10, 2012.

[6] F. Abergel and F. Pomponio. Trade-throughs: Empirical facts - Application to lead-lag measures. *To Appear in Quantitative Finance*, 2010. `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1694103`.

[7] A. Admati and P. Pfleiderer. A theory of intraday patterns: Volume and price variability. *The Review of Financial Studies*, 1(1):3–40, 1988.

[8] R. Allez and J.-P. Bouchaud. Individual and collective stock dynamics: Intra-day seasonalities. *New Journal of Physics*, 13(2), 2011.

[9] T. Andersen and T. Bollerslev. Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance*, 4(2-3):115–158, 1997.

[10] T.W. Anderson and D.A. Darling. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, 23(2):193–212, 1952.

[11] T. Ane and H. Geman. Stochastic subordination. *Risk*, 9:145–149, 1996.

[12] R. Ariel. The monthly effect in stock returns. *Journal of Financial Economics*, 18(1):161–174, 1987.

[13] Y. Aït-Sahalia, J. Cacho-Diaz, and R.J.A. Laeven. Modeling financial contagion using mutually exciting jump processes. *National Bureau of Economics Research Working Paper*, 2010. `http://www.nber.org/papers/w15850.pdf?new_window=1`.

[14] Y. Aït-Sahalia, Fan J., and D. Xiu. High frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association*, 105(492):1504–1517, 2010.

[15] Y. Aït-Sahalia, Mykland P.A., and L. Zhang. A tale of two time scales: Determining integrated volatility with noisy high frequency data. *Journal of the American Statistical Association*, 100(472):1394–1411, 2005.

[16] E. Bacry, S. Delattre, M. Hoffman, and J.-F. Muzy. Modelling microstructure noise with mutually exciting point processes. *To Appear in Quantitative Finance*, 2011. `http://www.cmap.polytechnique.fr/~bacry/ftpPapers1.php?paper=BDHM1.pdf`.

[17] C.A. Ball and W.N. Torous. Stochastic correlation across international stock markets. *Journal of Empirical Finance*, 7(3-4):373–388, 2000.

[18] O.E. Barndorff-Nielsen, Hansen P.R., A. Lunde, and N. Shephard. Multivariate realised kernels: consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading. 2008. `http://economics.ouls.ox.ac.uk/15055/1/2008OMI05.pdf`.

[19] O.E. Barndorff-Nielsen and N. Shephard. Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica*, 72(3):885–925, 2004.

[20] L. Bergomi. Correlations in asynchronous markets. *Risk*, pages 76–82, 2010.

[21] F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3):637–654, 1973.

[22] M.J. Bommarito II. Intraday correlation patterns between the S&P 500 and sector indices. 2010. `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1677915`.

[23] J.-P. Bouchaud, P. Cizeau, L. Laloux, and M. Potters. Noise dressing of financial correlation matrices. *Physical Review Letters*, 83(7):1467–1470, 1998.

[24] J.-P. Bouchaud, A. Matacz, and M. Potters. Leverage effect in financial markets: The retarded volatility model. *Physical Review Letters*, 87(22):228701, 2001.

[25] J. P. Bouchaud and M. Potters. *Theory of Financial Risk and Derivative Pricing, From Statistical Physics to Risk Management*. Cambridge University Press, second edition, 2004.

[26] J.-P. Bouchaud, M. Potters, Y. Gefen, and M. Wyart. Fluctuations and response in financial markets: The subtle nature of 'random' price changes. *Quantitative Finance*, 4(2):176–190, 2004.

[27] C.G. Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912, 2007.

[28] P. Carr, H. Geman, D. Madan, and M. Yor. The fine structure of asset returns: An empirical investigation. *Journal of Business*, 75(2):305–332, 2002.

[29] P. Carr, D. Madan, and E. C. Chang. The Variance Gamma process and option pricing. *European Finance Review*, 2(1):79–105, 1998.

[30] P. Carr, D.B. Madan, and R.H. Smith. Option valuation using the fast Fourier transform. *Journal of Computational Finance*, 2(4):61–73, 1999.

[31] A. Chakraborti, K. Kaski, J. Kertész, and J.-P. Onnela. Dynamic asset trees and portfolio analysis. *European Physical Journal B*, 30:285–288, 2002.

[32] K. Chan, Y. P. Chung, and H. Johnson. The intraday behavior of bid-ask spreads for NYSE stocks and CBOE options. *Journal of Financial and Quantitative Analysis*, 30(3):329–346, 1995.

[33] H. Chen and V. Singal. A December effect with tax-gain selling? *Financial Analysts Journal*, 59(4):78–90, 2003.

[34] R. Chicheportiche and J.-P. Bouchaud. The joint distribution of stock returns is not elliptical. 2011. `http://arxiv.org/abs/1009.1100`.

[35] T. Chordia and B. Swaminathan. Trading volume and cross-autocorrelations in stock returns. *Journal of Finance*, 55(2):913–935, 2000.

[36] A. Christian Silva and V.M. Yakovenko. Stochastic volatility of financial markets as the fluctuating rate of trading: An empirical study. *Physica A: Statistical Mechanics and its Applications*, 382(1):278–285, 2007.

[37] P. K. Clark. A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, 41(1):135–155, 1973.

[38] R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

[39] D.R. Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 17(2):129–164, 1955.

[40] J.C. Cox, J.E. Ingersoll, and S.A. Ross. A theory of the term structure of interest rates. *Econometrica*, 53(2):385–407, 1985.

[41] M. Dacorogna, R. Gencay, U.A. Müler, R.B. Olsen, and O.V. Pictet. *An Introduction to High Frequency Finance*. Academic Press, 2001.

[42] D.J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes. Volume 1: Elementary Theory and Methods*. Springer, second edition, 2003.

[43] F. De Jong and T. Nijman. High frequency analysis of lead-lag relationships between financial markets. *Journal of Empirical Finance*, 4(2-3):259–277, 1997.

[44] F. De Jong and T. Nijman. Intraday lead-lag relationships between the futures-, options and stock market. *European Finance Review*, 1:337–359, 1998.

[45] P. Debye. Näherungsformeln für die zylinderfunktionen für groSSe werte des arguments und unbeschränkt veränderliche werte des index. *Mathematische Annalen*, 67(4):535–558, 1909.

[46] P. Dierckx. *Curve and Surface Fitting with Splines*. Clarendon Press, 1995.

[47] N. El Karoui. Couverture des risques dans les marchés financiers. *Lectures Notes of Probability and Finance Master from Université Paris 6*, 2003-2004.

[48] R. Engle. Dynamic conditional correlation - a simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3):339–350, 2002.

[49] T. W. Epps. Comovements in stock prices in the very short-run. *Journal of the American Statistical Association*, 74(366):291–298, 1979.

[50] J. Erickson, Y. Li, and K. Wang. A new look at the Monday effect. *Journal of Finance*, 52(5):2171–2186, 1997.

[51] M.J. Fields. Security prices and stock exchange holidays in relation to short selling. *Journal of Business of the University of Chicago*, 7(4):328–338, 1934.

[52] K. French. Stock returns and the weekend effect. *Journal of Financial Economics*, 8(1):55–69, 1980.

[53] L. Gillemot, J. D. Farmer, and F. Lillo. There's more to volatility than volume. *Quantitative Finance*, 6(5):371–384, 2006.

[54] C. Gouriéroux. Continuous time Wishart process for stochastic risk. *Econometric Reviews*, 25(2-3):177–217, 2006.

[55] C.W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[56] J.E. Griffin and R.C.A. Oomen. Covariance measurement in the presence of non-synchronous trading and market microstructure noise. *Journal of Econometrics*, 160(1):58–68, 2011.

[57] T. Guhr, M.C. Munnix, and R. Schäfer. Impact of the tick-size on financial returns and correlations. *Physica A: Statistical Mechanics and its Applications*, 389(21):4828–4843, 2010.

[58] T. Guhr, M.C. Munnix, and R. Schäfer. Statistical causes for the Epps effect in microstructure noise. *Accepted for publication in International Journal of Theoretical and Applied Finance*, 2012. `http://arxiv.org/abs/1009.6157`.

[59] D. M. Guillaume, M.M. Dacorogna, R. Dave, U. A. Müller, R. B. Olsen, and O. V. Pictet. From the bird's eye to the microscope: A survey of new stylized facts of the intradaily foreign exchange markets. *Finance and Stochastics*, 1(2):95–129, 1997.

[60] A.G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[61] T. Hayashi and N. Yoshida. On covariance estimation of non-synchronously observed diffusion processes. *Bernoulli*, 11(2):359–379, 2005.

[62] N. Henze. Invariant tests for multivariate normality: a critical review. *Statistical Papers*, 43(4):467–506, 2002.

[63] M. Hoffmann, M. Rosenbaum, and N. Yoshida. Estimation of the lead-lag parameter from non-synchronous data. *To Appear in Bernoulli*, 2010. `http://www.crest.fr/ckfinder/userfiles/files/Pageperso/rosenbaum/HRY_Rev_03_12_2010.pdf`.

[64] K. Hou, N. Barberis, and X. Chen. Information diffusion and asymmetric cross-autocorrelations in stock returns. 2007. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.8.6147&rep=rep1&type=pdf`.

[65] J. Jacod and A. Shirayev. *Limit Theorems for Stochastic Processes*. Springer, 1987.

[66] E. Jarnecic. Trading volume lead/lag relations between the ASX and ASX option market: Implications of market microstructure. *Australian Journal of Management*, 24(1):77–94, 1999.

[67] G.B. Kadlec and D.M. Patterson. A transaction data analysis of nonsynchronous trading. *The Review of Financial Studies*, 12(3):609–630, 1999.

[68] A. F. Karr. *Point Processes and Their Statistical Inference*. Dekker, New York, 1991.

[69] K. Kaski, J. Kertész, and L. Kullman. Time-dependent cross-correlations between different stock returns: A directed network of influence. *Physical Review E*, 66(2):026125.1–026125.6, 2002.

[70] J. Kertész and B. Toth. Increasing market efficiency: Evolution of cross-correlations of stock returns. *Physica A*, 360(2):505–515, 2006.

[71] J. Kertész and B. Toth. The Epps effect revisited. *Quantitative Finance*, 9(7):793–802, 2009.

[72] A. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giorn. dell' Inst. Ital. Attuari*, 4:83–91, 1933.

[73] P. Leoni and W. Schoutens. Multivariate smiling. *Wilmott Magazine*, 8(March):82–91, 2008.

[74] E. Lewis and G. Mohler. A nonparametric EM algorithm for multiscale Hawkes processes. 2011. `http://math.scu.edu/~gmohler/EM_paper.pdf`.

[75] A.W. Lo and A.C. MacKinlay. When are contrarian profits due to stock market overreaction? *The Review of Financial Studies*, 3(2):175–205, 1990.

[76] A.W. Lo and A.C. MacKinlay. An econometric analysis of nonsynchronous trading. *National Bureau of Economics Research Working Paper*, 1991. http://www.nber.org/papers/w2960.pdf.

[77] P. Malliavin and M.E. Mancino. Fourier series method for measurement of multivariate volatilities. *Finance and Stochastics*, 6(1):49–61, 2002.

[78] B. Mandelbrot. The variation of certain speculative prices. *Journal of Business*, 36(4):394–419, 1963.

[79] V.A. Marchenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

[80] V. Mattiussi and Iori G. A nonparametric approach to estimate volatility and correlation dynamics. https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=SNDE2008&paper_id=36.

[81] R. McGill, J. W. Tukey, and W. A. Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.

[82] T.S. Mech. Portfolio return autocorrelation. *Journal of Financial Economics*, 34(3):307–344, 1993.

[83] I. Muni Toke. "Market making" behaviour in an electronic order book and its impact on the bid-ask spread. 2010. http://arxiv.org/abs/1003.3796.

[84] I. Muni Toke. An introduction to Hawkes processes with applications to finance. *Lectures Notes from Ecole Centrale Paris, BNP Paribas Chair of Quantitative Finance*, 2011. http://fiquant.mas.ecp.fr/ioane_files/HawkesCourseSlides.pdf.

[85] J. Muthuswamy, S. Sarkar, A. Low, and E. Terry. Time variation in the correlation structure of exchange rates: High frequency analysis. *Journal of Futures Markets*, 21(2):127–144, 2001.

[86] E. Nicolato. Multivariate modelling via matrix subordination. *Quantitative Methods in Finance Conference, Sydney*, 2009 December 16-19.

[87] Y. Ogata. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981.

[88] T. Ozaki. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Annals of the Institute of Statistical Mathematics*, 31(Part B):145–155, 1979.

[89] J. Prochazkova. Derivative of B-spline function. 2005. http://mat.fsv.cvut.cz/gcg/sbornik/prochazkova.pdf.

[90] C.Y. Robert and M. Rosenbaum. Volatility and covariation estimation when microstructure noise and trading times are endogenous. *Mathematical Finance*, 22(1):133–164, 2012.

[91] M. Rosenbaum. Integrated volatility and round-off error. *Bernoulli*, 15(3):687–720, 2009.

[92] F. Salmon. Recipe for disaster: The formula that killed Wall Street, 2009. http://www.wired.com/techbiz/it/magazine/17-03/wp_quant?currentPage=all.

[93] W. F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3):425–442, 1964.

[94] A. Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de L'Institut de Statistiques de l'Université de Paris*, 8:229–231, 1959.

[95] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101, 1904.

[96] C. van Emmerich. Modelling correlation as a stochastic process. 2006. http://www-num.math.uni-wuppertal.de/fileadmin/mathe/www-num/preprints/amna_06_03.pdf.

[97] S.B. Wachtel. Certain observations on seasonal movements in stock prices. *Journal of Business of the University of Chicago*, 15(2):184–193, 1942.

[98] W. Whitt. *Stochastic-Process Limits*. Springer, first edition, 2002.

[99] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[100] L. Zhang. Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics*, 160(1):33–47, 2011.