Handling heterogeneous and MNAR missing data in statistical learning frameworks : imputation based on low-rank models, online linear regression with SGD, and model-based clustering

Aude Sportisse

École Doctorale de Sciences Mathématiques de Paris-Centre (ED386)

THÈSE DE DOCTORAT

pour obtenir le grade de docteur délivré par

Sorbonne Université

Discipline doctorale: Mathématiques

Spécialité doctorale: Statistique

présentée et soutenue publiquement par

Aude Sportisse

le 29 juin 2021

Handling heterogeneous and MNAR missing data in statistical learning frameworks: imputation based on low-rank models, online linear regression with SGD, and model-based clustering

Sous la direction de **Claire Boyer et Julie Josse**

Jury

M. Gérard Biau,	Professeur, Sorbonne Université	Président
M. Charles Bouveyron,	Professeur, Université Nice Côte d'Azur	Examinateur
Mme. Claire Boyer,	Maitre de conférence, Sorbonne Université	Co-directrice de thèse
M. Julien Chiquet,	Professeur, Université Paris-Saclay, AgroParisTech	Rapporteur
M. Arnaud Guyader,	Professeur, Sorbonne Université	Co-directeur de thèse
M. Jes Frellsen,	Assistant professeur, Université technique du Danemark	Rapporteur
Mme. Julie Josse,	Professeur, INRIA Sophia Antipolis	Co-directrice de thèse
Mme. Olga Klopp,	Professeur, ESSEC Business School	Examinatrice
Mme. Madeleine Udell,	Assistant professeur, Université de Cornell	Invitée

Remerciements

Claire et Julie, je souhaiterais vous remercier très sincèrement pour tout ce que vous m'avez appris et toutes les portes que vous m'avez ouvertes. Merci Julie de m'avoir fait découvrir les données manquantes ; tu as toujours su me donner les bonnes cartes pour avancer et je n'aurais pas pu rêver d'une meilleure entrée dans ton domaine. Merci Claire pour ton aide minutieuse tout au long de ces trois ans et merci de t'être impliquée corps et âme dans les moments les plus difficiles ; c'est grâce à toi que la balle tombe du bon côté du filet. Merci à vous deux pour votre optimisme et votre joie qui ont su me rassurer. Vous nous avez emmenées vers des sujets fascinants, aux perspectives tout aussi passionnantes. J'espère que nous continuerons à discuter du MNAR, même autour d'un grog et de quelques raviolis aux queues de boeuf.

Je tiens à adresser ma plus sincère gratitude à mes rapporteurs de thèse, Julien Chiquet et Jes Frellsen. Julien, je suis très honorée que tu aies accepté d'être mon rapporteur et te remercie pour tes encouragements à mi-parcours, qui ont compté. I am very honoured, Jes, that you have agreed to be my reporter and I sincerely thank you. Je souhaiterais également remercier Gérard Biau, Charles Bouveyron, Arnaud Guyader, Olga Klopp et Madeleine Udell d'avoir accepté de faire partie de mon jury. Gérard, merci de m'avoir toujours soutenue, de l'obtention de ma bourse de thèse à nos discussions. Charles, merci à toi et à Pierre-Alexandre de m'avoir accordé votre confiance en m'offrant une fin de thèse pleine de motivation. Arnaud, je tiens à te remercier de m'avoir guidée dans mes choix après le M2 et d'avoir cru en moi : je n'aurais pas commencé cette thèse sans ton soutien. Merci également de m'avoir remonté le moral avec ton humour dont seul toi détiens la clé. Olga, je suis particulièrement honorée que tu fasses partie de mon jury. I would like to thank you, Madeleine, for our discussions during my PhD thesis; it is my honour to have you in my jury.

Je souhaiterais aussi remercier ceux avec qui j'ai eu la chance de travailler : Pascaline Descloux, Sylvain Sardy, Aymeric Dieuleveut, Christophe Biernacki, Matthieu Marbac-Lourdelle et Imke Mayer. Merci Aymeric, c'était un réel plaisir de travailler avec toi sur le SGD. Tu m'a énormément appris et aidée : le quatrième chapitre de ma thèse te doit beaucoup. Merci Christophe de m'avoir permis de faire partie de votre projet et de m'avoir accompagnée et beaucoup appris sur le SEM. Merci Matthieu pour ton aide indispensable sur l'identifiabilité. Un grand merci, Imke, de m'avoir permis de te rejoindre sur R-miss-tastic. Merci à François Husson, Boris Muzellec et Katarzyna Wòznica pour vos contributions à ce projet. Merci aux personnes qui m'ont aidé administrativement, en particulier Corentin Lacombe. Merci également à l'équipe de l'X-Exed.

Je voudrais aussi remercier toute l'équipe du LPSM pour sa convivialité ainsi que les doctorants du CMAP. Merci Geneviève de m'avoir motivée par ton exemple et d'avoir rempli toutes nos pauses de ta joie, qui m'a manqué. Merci Imke pour ta gentillesse et ta sérénité ; merci de m'avoir tant aidée sur la Traumabase. Merci Frédéric pour ton aide qui m'a été précieuse. Merci Rémi de ne pas trop t'être moqué de mes boxplots. Merci Constantin pour ta compagnie au CIRM et pour la confiance que tu m'as accordée. Merci Bénédicte, Marine

et Wei pour votre soutien. Thanks to the Causal Inference and Missing Data group. Thanks Zoltan. Merci Gloria, Qiming et Thibault. Merci Adeline pour nos discussions et pour ton soutien de co-chargée de TD. Merci Eva pour tous tes conseils et ta bonne humeur. Maud et Antoine, un grand merci, vous m'avez vraiment donné un billet première classe pour mes premiers pas dans l'enseignement ; cela a beaucoup compté.

Enfin, j'aimerais remercier l'Université, Jussieu, car tu m'as toujours laissé ma chance - rien n'est figé ! -, tu m'as tellement appris et tu m'as donné ma liberté. Un merci particulier à Gérard Biau, Claire Boyer, Arnaud Guyader, Vincent Lemaire, Tabéa Rebafka, Etienne Roquain, Maxime Sangnier et Maud Thomas : c'est réellement grâce à vos cours que j'ai pu me lancer dans une thèse.

Merci Rika et Lacene pour nos discussions chaleureuses et votre disponibilité. Merci Christophe et merci Marie de m'avoir fait participer à de superbes aventures.

Merci à mes amis Emilien, Fleur, François, Gaston, Guillaume et Saint-Clair. Merci LOL (vos identités protégées) : toutes ces enquêtes seront bien résolues un jour. Merci Anissa et Domitille pour votre amitié si précieuse.

Merci Isabelle, Olivier et François pour votre soutien aux doux airs bretons.

Merci mémé pour tes encouragements réconfortants. Merci Edith, Claude, Dany et Lucien pour tout ce que vous me transmettez.

Merci Bucky, sans rancune. Merci Drs Myriam et Bruno de toujours avoir cru en moi et de m'avoir donné le goût pour les sciences : un immense merci, maman, de tant nous donner de ce qui compte, et, papa, d'être le meilleur et le plus zen des taxis, sur bien des routes. Merci Marine pir tillimi di chises : merci d'être un modèle, de créer des moments plein de joie, et de savoir créer bien plus. Merci Thibaut d'être le seul, dans les moments les plus improbables, à savoir me faire sourire, avec tes farces ingénieuses : merci de savoir prendre tant de virages d'avance.

Nicolas, merci pour nos discussions mathématiques sur le canapé, elles me sont salvatrices et merci pour ta relecture attentive. Je ne peux te donner qu'un tout-petit merci pour ton infini soutien quotidien. Finalement, cette thèse, c'est un bout de l'une de nos belles aventures, palaisienne, pentue, pleine de boues, de rires et de vents.

Contents

Chapter 1

Introduction

Contents

1.1 Preamble

The increasing availability of data sets and the multiplication of sources offer hopes for understanding, interpreting and predicting many phenomena. However one of the ironies of the so-called "big data" era is that missing data are unavoidable: the more data there are, the more missing data there are. Indeed, missing data can occur for many reasons: unanswered questions in a survey, lost data, sensing machines that fail, aggregation of multiple sources, etc. Classical statistical methods can not be directly applied on the datasets which contain missing values. A naive solution is then to delete the missing values: either the incomplete variables or the missing individuals, i.e. either some columns or some rows of the dataset. However, deleting data is not a solution in most cases for two main reasons: (i) this is only possible if there is little missing data, otherwise the loss of information is too great and (ii) the kept observations can constitute a sub-population which is not necessarily representative of the overall population leading to bias in subsequent analyses. In general, this strategy is not suitable: (i) it is rare that only a few variables or individuals contain missing data and (ii) the case where a sub-population is representative of the general population is a toy case, as most of the time the process that causes the data to be missing depends on the data values themselves. For example, rich people are often less inclined to reveal their income. In this dissertation, we are interested in developing methods for handling large scale data with heterogeneous types of missing values, different natures of variables and different percentages of missing values in each variable. More particularly, our work is motivated by a public health application with a clinical register presented below.

1.1.1 Motivation: the Traumabase® dataset

Major trauma, i.e. injuries that endanger a person's life or functional integrity (such as road accidents, interpersonal violence and falls) have been qualified as a worldwide public health challenge and a principal source of mortality in the world by the World Health Organization (Hay et al., 2017). This is particularly striking in the group of people aged between 16 and 45 years for whom major trauma is the leading cause of death. A patient who has just suffered a trauma is first taken care of at the scene of the accident, then transferred to the hospital in an ambulance and finally treated in the emergency services of a medical center. This highly stressful environment involving many carers can lead to delays or errors in the decision making, with high risks for the patient. Hamada et al. (2014) and Hamada et al. (2015) showed that the patient management often exceeds acceptable time frames and that diagnoses can differ in the ambulance and at the arrival at hospital. As efficient and timely trauma management is crucial to improve patient care, 19 French trauma centers have been working together since 2012 to collect high-quality clinical measurements (250 variables) on 20,000 traumatized patients from the scene of the accident to the hospital admission.

This tabular dataset contains heterogeneous clinical measurements, with both categorical variables (sex, type of illness,...) and quantitative variables (blood pressure, hemoglobin level...). There is a high percentage of missing values in most of the variables (in the whole dataset, 80% of the individuals have missing values). Missing values can be due to the

aggregation of datasets from multiple hospitals, which typically gathers different observations on patients, or to failures of the measuring devices, or to the fact that doctors may not have time to accordingly measure health variables in emergency situations, or the missing values can be informative in the sense if the state of the patient is such that it was not possible to make the measurement. Both percentage and nature of missing data demonstrate the importance of taking appropriate account of missing data. In any case, being able to handle missing values can also avoid the unnecessary effort of collecting new complete observations, which is unfeasible in view of the financial and time costs this would entail.

Our aim is to assist doctors for their decision making in emergency situations. For example, given one patient's pre-hospital features, could we predict the risk of an hemorrhagic shock? This falls within the scope of supervised learning, as the goal is to carry out predictive models as regression or classification ones in presence of missing data. Another example of statistical analysis that doctors would like to conduct on such data, is to identify relevant groups of patients sharing similarities. Formally, this is a task of unsupervised learning that should be processed in presence of missing data. Another burning issue is that of imputation. Indeed, it would also be useful to judiciously replace each missing entry in the Traumabase by a plausible value. By doing so, we would obtain a complete dataset, on which usual statistical methods could be applied, although this may be too simplistic in some cases.

1.1.2 Outline of the introduction

Rubin (1976) laid the foundations of the missing-data formalism that is still used nowadays. Since then, one could point out the review works by Schafer (1997); Kim and Shao (2013); Molenberghs et al. (2014); Van Buuren (2018); Little and Rubin (2019) that provide a complete introduction to the main concepts and methods related to missing values. The choice of a method to deal with missing values depend on both (i) the missing-data pattern, which indicates *where the missing data are* and (ii) the missing-data mechanism which answers the difficult question *why the data are missing*. Section 1.2 presents these two key ingredients of the missing-data analysis: Section 1.2.1 defines the missing-data pattern and Section 1.2.2 the missing-data mechanism. Section 1.2.3 gives the main result of the missing-data analysis, which explains why the cause of the lack of data is so important to consider in some cases, when the missing-data mechanism is said *nonignorable*.

In addition to depending on the missing-data type, the method to choose also depends on the purpose of the statistical analysis. The learning procedure in presence of missing values can be of different natures such as imputing missing data, estimating parameters of an underlying model or predicting a target variable with missing covariates, etc. This dissertation mainly focuses on the inferential framework, when the goal is to perform parametric model estimation from incomplete data. In Section 1.3, a summary of the main techniques to deal with missing values in the inferential framework is given, assuming that the missing-data mechanism is ignorable. In Section 1.3.1, we discuss the complete-case analysis, when the missing values are deleted, which, despite an appealing simplicity, suffers from strong flaws. Section 1.3.2 presents the Expectation Maximization (EM) algorithm (Dempster et al., 1977), which allows to modify the estimation strategy to apply it to an incomplete dataset. Section

1.3.3 focuses on the methods which consist of imputing missing values to get a complete dataset (on which classical algorithms could then be applied for example to estimate some parameters). Section 1.3.4 is devoted to specific methods which impute naively the missing values and then adapt classical algorithms to account for the imputation error. In Section 1.3.5, the previously introduced methods are discussed and put into perspective.

In Section 1.4, three statistical frameworks are discussed, which call for specific methods when missing data occur. Section 1.4.1 aims to introduce the linear regression with missing covariates, i.e. when there exists an outcome variable linearly related to the covariates. The literature considered focuses on estimating the linear regression parameters or selecting relevant variables in the high-dimensional setting. Section 1.4.2 deals with supervised learning, when the data are *labeled*, i.e. there exists an outcome variable, and the goal is to know how to predict outcomes for new observations. Note that linear regression can be considered a special case of supervised learning, but the objectives differ (parameter estimation in the first case, prediction in the other). Finally, Section 1.4.3 focuses on unsupervised learning, when the data are *unlabeled* and the aim is to partition the data into different groups that make sense.

The classical methods to deal with missing data presented in Sections 1.3 and 1.4 are only valid under the assumption that the missing-data mechanism is ignorable and lead to bias if it does not hold. Nevertheless, these missing-data scenarios are often unrealistic and too restrictive. The MNAR mechanism, which encompasses the nonignorable mechanisms, allows to model a large variety of situations, because the missingness may depend on both observed and missing variables, and is often much more appropriate for real datasets. Indeed, the cause of the missing values for a variable is often related to the values of other missing variables or its values itself. For example, in the Traumabase dataset, the doctors can have no time to make the clinical measurements in emergency situations. In this case, the fact that variables related to the patient's condition may be missing is explained by the values of the variables themselves. If the patient is in a very bad condition, the heart rate may be high. Thus, the higher values of the heart rate have a high probability of being missing, and the missing mechanism will be MNAR. It is then essential to consider realistic missing-data scenarios and to propose methods compatible with the MNAR assumption: it is the focus of Section 1.5. As the MNAR mechanism is nonignorable, the missing-data mechanism has to be taken into account. It leads to the key issues under the MNAR assumption: (i) the specification of the missing-data mechanism addressed in Section 1.5.3, (ii) the identifiability of the parameters of the missing-data mechanism, see Section 1.5.2, (iii) the need for specific methods presented in Section 1.5.3, and (iv) the impossibility of testing the MNAR assumption discussed in Section 1.5.4.

In this dissertation, we propose new methods, addressing real-world problems, that are both theoretically sound and computationally efficient. Our main contributions are summarized in Section 1.6.

1.2 Key tools for missing-data analysis

1.2.1 Missing-data pattern

In the following, the data sample is denoted as X of size $n \times d$, where n is the number of observations and d the dimension (the number of variables). More precisely, one can write $X = (X_{1.}|\dots|X_{n.})^T$, in which each observation $X_{i.} = (X_{i1}, \dots, X_{id})^T$ belongs to the d-dimensional features space $\mathcal{X}$ which depends on the data type at hand (categorical, continuous or a mix of both). We distinguish the random variables from their realisations, by denoting them with capital and lower-case letters respectively. For instance, x_{ij} is a realisation of variable $X_{.j}$ for individual i. We assume that X contains missing values, i.e. some of its entries are denoted as `NA` for Not Available. The missing-data pattern, denoted by M is defined as follows:

Definition 1 (Missing-data pattern)**.** *The missing-data pattern $M \in \{0,1\}^{n\times d}$ is a (random) binary matrix, such that its realised values are*

$$\forall i \in \{1,\dots,n\},\ \forall j \in \{1,\dots,d\}, \quad m_{ij} = \begin{cases} 1 & \text{if } x_{ij} \text{ is missing,} \\ 0 & \text{otherwise.} \end{cases} \tag{1.1}$$

Until Rubin (1976) introduced the notion of missing-data mechanism and treat M as a random variable, the missing-data pattern was only viewed as a realisation and was largely ignored in the statistical analysis.

With an incomplete dataset at hand, saying where the missing data occur is not the most difficult part, as the realisation of the missing-data pattern is always observed. However, specifying the missing-data mechanism, which amounts to modelling the distribution of the missing-data pattern according to the data, is more complicated.

1.2.2 Missing-data mechanism

Rubin (1976) classifies the cause of the lack of data into three missing-data mechanisms, which describe the relationship between the missing-data pattern and the data values. The historical notations have been slightly modified to avoid overloading notations and simplify interpretation. What we called the classical definitions are widely used in all papers and textbooks on missing values. However, as discussed below, these definitions can be subject to debate (Seaman et al., 2013) and alternative definitions have been suggested. In all definitions, the missing-data mechanism is always characterized by the conditional distribution of M given X, parameterized by an unknown parameter $\phi \in \Omega_\phi$.

(a) Classical definition of missing mechanisms Some authors (including the precedent editions 1987, 2002 of Little and Rubin (2019) and Schafer (1997)) denote the observed components and the missing components of X as $X^{\text{obs}} \in \mathcal{X}^{\text{obs}}$ and $X^{\text{mis}} \in \mathcal{X}^{\text{mis}}$, where $\mathcal{X}^{\text{obs}}$ and $\mathcal{X}^{\text{mis}}$ are subsets of the space $\mathcal{X}$. In the following, the conditional distribution of M given X is written as $f_{M|X}(.|.;\phi)$, parametrized by ϕ. The three different missing-data mechanisms can be defined as follows.

Notation	**Description**
$X \in \mathcal{X}^n$ (vector of size nd)	Vectorized data containing missing values
$M \in \{0,1\}^{nd}$	Vectorized missing-data pattern
$X^{\text{obs}} \in \mathcal{X}^{\text{obs}}$	Observed component of X
$X^{\text{mis}} \in \mathcal{X}^{\text{mis}}$	Missing component of X
$X \in \mathcal{X}^n$ (matrix of size $n \times d$)	Data matrix containing missing values
$M \in \{0,1\}^{n \times d}$	Missing-data pattern
$X_{i.}^{(0)} \in \mathcal{X}_{i.}^{(0)}$	Values of the observed variables for ind. i
$X_{i.}^{(1)} \in \mathcal{X}_{i.}^{(1)}$	Values of the missing variables for ind. i

Table 1.1: Notations for the missing-data mechanisms for Definition 2 (at the top) and for the Definition 3 (at the bottom).

Definition 2 (Missing-data mechanism (classical))**.** *The missing-data mechanism is said*

- *missing completely at random (MCAR) if*

$$f_{M|X}(M|X;\phi) = f_M(M;\phi), \qquad \forall X \in \mathcal{X}^n, \forall \phi \in \Omega_\phi$$

- *missing at random (MAR) if*

$$f_{M|X}(M|X;\phi) = f_{M|X^{\text{obs}}}(M|X^{\text{obs}};\phi), \qquad \forall X^{\text{mis}} \in \mathcal{X}^{\text{mis}}, \forall \phi \in \Omega_\phi$$

- *missing not at random (MNAR) when the MAR assumption does not hold.*

This first definition is sufficient enough to get an intuition on the different missing-data mechanisms: roughly speaking, the missing-data mechanism is said (i) MCAR when the occurrence of the missing data is totally independent of the data, (ii) MAR when the unavailability of the data depends on the values of observed variables and (iii) MNAR when the process that causes the missing data depends on the values of missing variable, and possibly observed ones too. As we will see later in detail, MCAR and MAR can be handled more easily than the challenging MNAR case. We illustrate these missing scenarios on the following example.

An example of missing-data mechanisms Consider the simple situation of a survey with two variables, Income and Age, with missing values only on the Income variable. The MCAR setting implies that the missing values are independent of any value (e.g. respondents have forgotten to fill the form). The MAR situation settles that missing values on Income depend on the values of Age (e.g. younger respondents would be less incline to reveal their income). The MNAR scenario allows the occurrence of the missing values on Income to depend on the values of the income itself (e.g. poor and rich respondents would be less incline to reveal their income): even though Age and Income are related, the process that causes the missing data is not fully explained by Age. Consequently, knowing the value of Age is not enough to retrieve the value of Income. See Figure 1.3 to visualize this example.

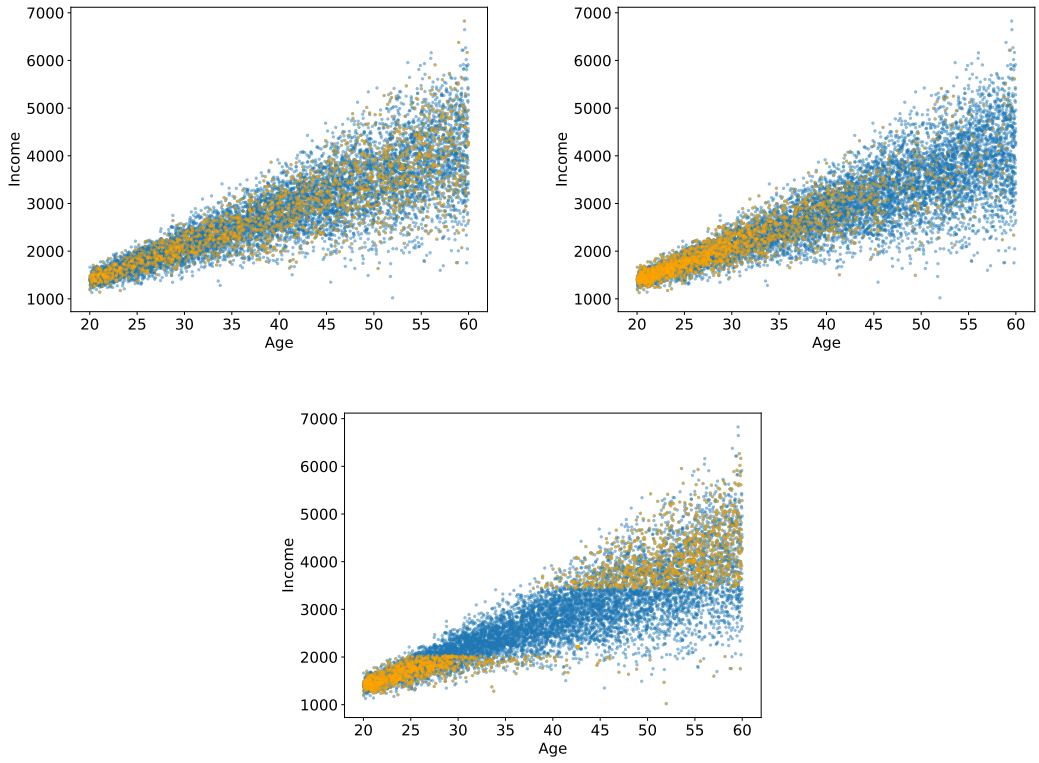


Figure 1.1: Illustration of MCAR (top left), MAR (top right) and MNAR (bottom) mechanisms for the survey example, when the Income is missing and the Age is fully observed. Individuals containing missing values (resp. fully observed) are represented in orange (resp. blue). To obtain these plots, the ground truth is assumed to be known, i.e. we know the underlying values of the missing elements in Income. The missing values for the MCAR mechanism in Income occur for any age of the respondent. For the MAR mechanism, it is clear that the younger the respondent, the more likely it is that their income is missing. For the MNAR mechanism, the low and high values of Income are missing. It corresponds to the self-masked mechanism, because the unavailability of a value depends on the value itself.

(b) Removing the ambiguity in the definition of the mechanisms Definition 2 is often subject to debates. Indeed, the notations for X^{obs} and X^{mis} are ambiguous, because both vectors depend on the missing-data pattern M. One could note that X^{obs} (resp. X^{mis}) is the matrix formed by the components x_{ij} if $m_{ij} = 0$ (resp. $m_{ij} = 1$). Thus, X^{obs} and X^{mis} are functions of M so that writting $M|X^{\text{obs}}$ is not appropriate.

Little and Rubin (2019) suggest a new definition to clarify the shadows of Definition 2. More precisely, the values of the observed (resp. missing) variables for individual i are denoted as $X_{i.}^{(0)}$ (resp. $X_{i.}^{(1)}$). The space of the observed (resp. missing) variables for individual i is $\mathcal{X}_{i.}^{(0)} = \{\tilde{X}_{i.} = (\tilde{X}_{i1}, \ldots, \tilde{X}_{id}) \in \mathcal{X} : \tilde{X}_{i.}^{(1)} = X_{i.}^{(1)}\}$ (resp. $\mathcal{X}_{i.}^{(1)} = \{\tilde{X}_{i.} = (\tilde{X}_{i1}, \ldots, \tilde{X}_{id}) \in \mathcal{X} : \tilde{X}_{i.}^{(0)} = X_{i.}^{(0)}\}$).

Definition 3 (Missing-data mechanism (Little and Rubin, 2019))**.** *Under the assumption that the pairs $(X_{i.}, M_{i.})_i$ are i.i.d., the missing-data mechanism is said*

- *missing completely at random (MCAR) if*

$$f_{M|X}(m_{i.}|x_{i.}; \phi) = f_{M|X}(m_{i.}|x_{i.}^{\star}; \phi), \qquad \forall x_{i.} \neq x_{i.}^{\star} \in \mathcal{X}, \quad \forall \phi \in \Omega_{\phi},$$

  where $x_{i.}^{\star} \in \mathcal{X}$ is a realisation of $X_{i.}$, distinct from $x_{i.} \in \mathcal{X}$,

- *missing at random (MAR) if*

$$f_{M|X}(m_{i.}|x_{i.}^{(0)}, x_{i.}^{(1)}; \phi) = f_{M|X}(m_{i.}|x_{i.}^{(0)}, x_{i.}^{\star(1)}; \phi), \quad \forall x_{i.}^{(1)} \neq x_{i.}^{\star(1)} \in \mathcal{X}^{(1)}, \ \forall \phi \in \Omega_{\phi};$$

- *missing not at random (MNAR) if the MAR assumption does not hold for some $(x_{i.}^{(1)}, x_{i.}^{\star(1)})$.*

Note that we have slightly modified the definition of Little and Rubin (2019) as we added the parameter ϕ in such a way that the statements hold for any value of ϕ. Definitions 2 and 3 are equivalent[1], only the notations differ (summed up in Table 1.1).

As pointed by Seaman et al. (2013), Definition 3 still remains restrictive: for the MAR mechanism, it requires fully observed variables. Indeed, the law of M given X should be the same for each individual i and should depend on observed variables. Therefore, some variables (at least one) must be always observed. To tackle this issue, some authors (Seaman et al., 2013; Murray et al., 2018) propose to consider a more general mechanism, called the *realised* MAR mechanism, which in fact corresponds to the historical version given in Rubin (1976). This definition does not consider the i.i.d. assumption as expected and relies on statements holding only for *realised* values of $(X_{i.}, M_{i.})$, and not for any values of $(X_{i.}, M_{i.})$. Even if this version is of particular interest, it is more canonical to use Definition 3. Indeed, in statistics, statements about the realised values are rarely used, it is often preferred to use a dedicated formalism that involves random variables directly.

In the literature, a mechanism derived from the general MNAR given in Definition 2 is often considered (Mohan, 2018), when the unavailability of a missing variable $X_{.j}$ only

[1]Note that the assumption that the rows of (X, M) are i.i.d. is implied in Definition 2. Without this assumption, a function f should be defined for each couple (X_i, M_i).

depends on the values of $X_{.j}$ themselves. It is the so-called *self-masked* MNAR mechanism, given in the definition below.

Definition 4 (Self-masked MNAR mechanism)**.** *Under the assumption that the pairs* $(X_{i.}, M_{i.})_i$ *are i.i.d., the missing-data mechanism is said self-masked MNAR if*

$$\forall j \in \{1, \dots, d\}, f_{M|X}(m_{ij}|x_{i.}^{(0)}, x_{i.}^{(1)}; \phi) = f_{M|X}(m_{ij}|x_{ij}^{(1)}). \tag{1.2}$$

Note that if a specific distribution for $f_{M|X}$ is assumed, it is often a logistic distribution or a probit one (Ibrahim et al., 1999; Morikawa et al., 2017; Tang and Ishwaran, 2017). For the logistic distribution, Equation (1.2) leads to

$$f_{M|X}(m_{ij}|x_{i.}^{(0)}, x_{i.}^{(1)}; \phi) = \left(1 + e^{-(\phi_0 + \phi_1 x_{ij}^{(1)})}\right)^{-1},$$

where $\phi = (\phi_1, \phi_2)$.

(c) Testing the missing-data mechanism In practice, it is extremely difficult to know if the missing values in a dataset are either MCAR, MAR or MNAR and it is exacerbated if there are different types of missing values within the same dataset. In Figure 1.2, we illustrate this issue by representing the observed (available) values for the simple situation of a survey with two variables, Income and Age, with missing values (MCAR, MAR or MNAR) only on the Income variable.

Most of the time, the knowledge of the missing-data mechanism relies on domain expertise: experts know why the data are missing (they may not have had time to fill in the form, the measuring device may not indicate values above a certain threshold, etc.).

Nevertheless, there are few cases where it is possible to infer the mechanism from the data (Schafer and Graham, 2002; Graham et al., 1994). Let us consider the following example with the questions "How old are you?" and "do you go and vote?". In such a case, the missing data in the answer to the second question were never intended to be collected for people below 18 years old and the mechanism can be identified as MAR as the probability to be missing on the voting variable depends on the age values.

As an alternative, there are some procedures to check the validity of the assumption but only to test whether the mechanism is MCAR. Little and Rubin (2019, Chapter 3) propose a simple procedure to verify if the MCAR assumption makes sense. For a fully observed variable $X_{.j}$ (i.e. $\forall i \in \{1, \dots, n\}, m_{ij} = 0$), the purpose is to compare the distribution of $X_{.j}$ for the "complete" individuals such that all the variables are observed (i.e. the individuals i such that $\forall j \in \{1, \dots, d\}, m_{ij} = 0$) and the distribution of $X_{.j}$ for the individuals which have missing values (i.e. the individuals i such that $\exists j \in \{1, \dots, d\}, m_{ij} = 1$). If the distributions are significantly different, the MCAR assumption is invalid.

For the M(N)AR mechanisms, there are some interesting works which assess the results sensitivity to alternate hypotheses about the missing-data mechanism (see Section 1.5.4 for more details).

✍ In this dissertation, a close attention is paid to considering realistic missing-data mechanisms. Even if in Chapter 4 only a more general MCAR missing-data mechanism is studied, Chapter 2, 3 and 5 focus on MNAR ones.

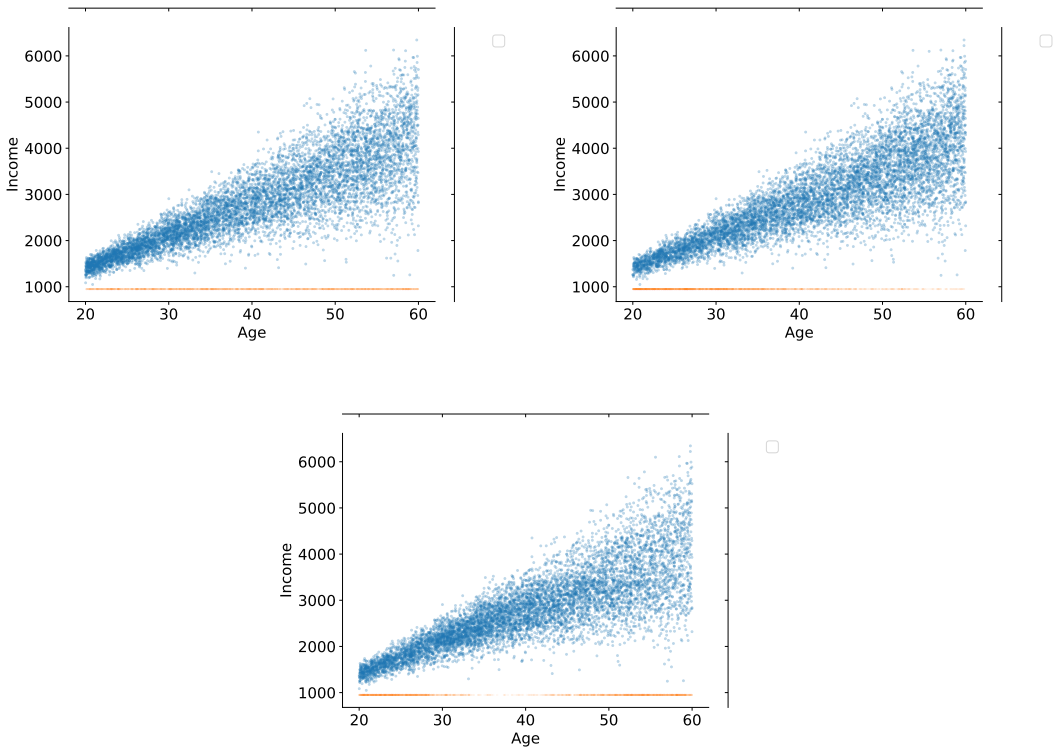


Figure 1.2: Illustration of MCAR (top left), MAR (top right) and MNAR (bottom) mechanisms for the survey example, when the Income is missing and the Age is fully observed. To obtain these plots, the ground truth is not assumed to be known (contrary to Figure 1.1). The observed individuals for Income and Age are represented in blue and individuals which are observed only for Age are represented in orange with a default value for Income. For the MCAR mechanism, the missing values are uniformly distributed according to Age. For the MAR and MNAR mechanisms, it is difficult to come to any conclusions. For the MAR mechanism, we can observe that there are fewer missing values for higher values of Age. For the MNAR mechanism, it seems that the age group between 35 and 45 years is less prone to lack in Income. In either cases, this does not tell us whether the mechanism is MAR or MNAR.

1.2.3 Notion of ignorability

The main difference between the MCAR and MAR mechanisms on the one hand, and the MNAR mechanism on the other hand, is that the former do not require to account for the missing-data mechanism, while the latter does. To see this, the simplest way is to consider the inferential framework, when the aim is to estimate the parameters of an underlying model on the data.

Rubin (1976) treats the missing-data pattern as a random variable (see Section 1.2.1). At first glance, the statistical inference should be conducted on the joint distribution of the data X and the missing-data pattern M, even if the main goal is to estimate the parameter θ of the data distribution, denoted as $f_X(.;\theta)$.

(a) Likelihood-based inference without missing values To estimate θ without missing values, a common estimation strategy in a parametric framework relies on maximizing the likelihood associated to the data. Assume that an i.i.d. sample $X = (X_{1.}, \ldots, X_{n.})$ is distributed according to $f_X(.;\theta)$, with an unknown parameter θ. The likelihood L is formed from the joint probability distribution evaluated on the observed sample, viewed as a function of the parameters only, namely

$$L(\theta; X) = \prod_{i=1}^{n} f_X(x_{i.};\theta).$$

The likelihood is thus the probability of drawing the sample obtained. Then, a standard way of estimating θ is to maximize the likelihood with respect to the parameters. More formally, this amounts to choosing the maximum likelihood estimator $\hat{\theta}$ (more details are given by Cox and Hinkley (1979)) as

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta; X).$$

(b) Likelihood-based inference with missing values With missing values, the statistical inference is conducted on the joint distribution of (X, M) denoted as $f_{X,M}(.;\theta,\phi)$. By abuse of notation, X and M are viewed as vector of size nd and no longer as matrices of size $n \times d$. As the rows of (X, M) are i.i.d., one has $f_{X,M}(x, m;\theta,\phi) = \prod_{i=1}^{n} f_{X,M}(x_{i.}, m_{i.};\theta,\phi)$. where $\theta \in \Omega_\theta$ and $\phi \in \Omega_\phi$ are the data distribution parameter and the missing-data mechanism distribution parameter. $\Omega_{\theta\times\phi}$ denotes the joint parameter space.

The full likelihood L_{full} can thus be defined as

$$L_{\text{full}}(\theta,\phi; X, M) = f_{X,M}(x, m;\theta,\phi) \tag{1.3}$$

However, in presence of missing values, this likelihood is intractable (since it involves $X^{(1)}$) and the full observed likelihood is considered instead, denoted as $L_{\text{full,obs}}$, by integrating (1.3) over the missing values as follows

$$L_{\text{full,obs}}(\theta,\phi; X^{(0)}, M) = \int_{\mathcal{X}^{(1)}} f_{X,M}(x, m;\theta,\phi)dx^{(1)}, \tag{1.4}$$

The form of the full observed likelihood is rarely closed, so that a direct computation is often impossible. To overcome this issue, the EM algorithm (Dempster et al., 1977) can be used (see Section 1.3.2).

(c) Ignorability Let the observed likelihood be denoted as L_{ign} such that

$$L_{\text{ign}}(\theta; X^{(0)}) = \int_{\mathcal{X}^{(1)}} f_X(x;\theta)dx^{(1)}. \tag{1.5}$$

The missing-data mechanism is said ignorable if the statistical inference for θ can be conducted by maximizing the observed likelihood, instead of the full observed likelihood given in (1.4). This is formalized in the definition below and more precise conditions are given in Theorem 6.

Definition 5 (Ignorable missing-data mechanism (Little and Rubin, 2019))**.** *The missing-data mechanism is said ignorable, if*

$$\forall \phi \in \Omega_\phi, \operatorname*{argmax}_{\theta\in\Omega_\theta} L_{\text{full,obs}}(\theta, \phi; X^{(0)}, M) = \operatorname*{argmax}_{\theta\in\Omega_\theta} L_{\text{ign}}(\theta; X^{(0)})$$

Theorem 6 (Ignorability of a missing-data mechanism (Little and Rubin, 2019))**.** *The missing-data mechanism is ignorable if the two following condition hold*

(i) the parameters θ and ϕ are distinct, in the sense that $\Omega_{\theta,\phi} = \Omega_\theta \times \Omega_\phi$.

(ii) the full observed likelihood can be factorized as follows

$$L_{\text{full,obs}}(\theta, \phi; X^{(0)}, M) = f_{M|X}(m|x^{(0)};\phi) L_{\text{ign}}(\theta; X^{(0)}).$$

Condition (i) means that each value $\theta \in \Omega_\theta$ is compatible with each value $\phi \in \Omega_\phi$ (it is a technical condition required to split the likelihood as in (ii)).

Theorem 6 provides one of the key results in missing-data analysis. Under condition (i) and the M(C)AR assumption, the inference about θ can be achieved by maximizing the likelihood given by (1.5), which is computationally easier than (1.4) and especially avoids any modelling of the missing-data mechanism (in particular, its specific form is not required to be modeled). Indeed, standard computation give

$$\begin{aligned}
L_{\text{full,obs}}(\theta, \phi; X^{(0)}, M) &= \int_{\mathcal{X}^{(1)}} f_{X,M}(x, m; \theta, \phi)dx^{(1)} \\
&\overset{(\star)}{=} \int_{\mathcal{X}^{(1)}} f_X(x;\theta) f_{M|X}(m|x;\phi)dx^{(1)} \\
&\overset{(\star\star)}{=} \int_{\mathcal{X}^{(1)}} f_X(x;\theta) f_{M|X}(m|x^{(0)};\phi)dx^{(1)} \qquad \text{(using Definition 3)} \\
&\overset{(\star\star\star)}{=} f_{M|X}(m|x^{(0)};\phi) \int_{\mathcal{X}^{(1)}} f_X(x;\theta)dx^{(1)}.
\end{aligned}$$

In Step ($\star$), the factorization of the joint distribution is chosen so as to show the distribution of the missing-data mechanism $f_{M|X}$ explicitly. In Step ($\star\star$), the definition of the MAR mechanism is used, implying that the missing-data mechanism does not depend on the missing values $x^{(1)}$. In Step ($\star\star\star$), this term is taken out of the integral.

Note that the likelihood-based inference theory has been presented in the frequentist framework, but the Bayesian framework could be also considered (Tanner and Wong, 1987; Little and Rubin, 2019, Chapters 6 and 10), where the parameters (θ, ϕ) are considered as random variables rather than fixed quantities.

For the sake of clarity, Table 1.2 summarises the different likelihood functions introduced so far.

Notation	**Name**	**Quantities involved**	**Comment**
L_{full} (1.3)	Full likelihood	$X^{(0)}, X^{(1)}, M$	not tractable
$L_{\text{full,obs}}$ (1.4)	Full observed likelihood	$X^{(0)}, M$	needed for MNAR data
L_{ign} (1.5)	Observed likelihood	$X^{(0)}$	sufficient for M(C)AR data

Table 1.2: Summary of introduced likelihoods in Section 1.2.3

1.3 Dealing with missing data

1.3.1 Complete-case analysis

The complete-case analysis consists of removing all the individuals containing missing values; an illustration is given in Figure 1.3. Due to its simplicity, this method is widely used in data science.

First of all, it should be noted that this method can be considered only under the MCAR assumption (when the missing pattern and the data are independent $M \perp\!\!\!\perp X$). Indeed, in this setting, the observed individuals are representative of the whole population. For M(N)AR mechanisms, this method leads to large bias in the estimates and can result in disastrous statistical analyses. Moreover, it should be noted that in most cases (even for MCAR data) removing individuals creates a huge loss of information. Graham (2009) advises against using this method when individuals with missing values represent more than 5% of the population. Zhu et al. (2019) takes the following example: they consider a data matrix $X \in \mathbb{R}^{n \times d}$ where each entry has a probability 1% to be missing independently (MCAR). If $d = 5$, around 95% of the individuals are complete; however, when dimension is larger such as p = 300, the complete-case analysis amounts to keep only 5% complete rows. Consequently, despite its simplicity, for many applications (including the Traumabase dataset), this method is not relevant and other methods should be considered.

1.3.2 EM algorithm and variants

The algorithm In the inferential framework, a general method to maximize the full likelihood (1.3) is the Expectation Maximization (EM) algorithm, introduced as is by

$$X = \begin{array}{c} \begin{array}{ccc} X_{\cdot 1} & X_{\cdot 2} & X_{\cdot 3} \end{array} \\ \begin{pmatrix} 12 & 28 & \texttt{NA} \\ 23 & \texttt{NA} & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \texttt{NA} & 3 & 7 \end{pmatrix} \end{array}, \; X^{CC} = \begin{array}{c} \begin{array}{ccc} X_{\cdot 1} & X_{\cdot 2} & X_{\cdot 3} \end{array} \\ \begin{pmatrix} \cancel{12} & \cancel{28} & \cancel{\texttt{NA}} \\ \cancel{23} & \cancel{\texttt{NA}} & \cancel{89} \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \cancel{\texttt{NA}} & \cancel{3} & \cancel{7} \end{pmatrix} \end{array}$$

Figure 1.3: Illustration of the complete-case analysis: if X is the data matrix, the statistical inference will be conducted on X^{CC}, by removing individuals (rows) which contain missing values.

Dempster et al. (1977). After initializing the algorithm with $\theta^{(0)}$, the two following steps are iteratively proceeded until convergence,

- the E-step (Expectation) (at step r): it consists of computing the expected complete likelihood knowing the observed data $X^{(0)}$ and the current value of the parameter θ^r, denoted as Q,
$$Q(\theta; \theta^r) = \mathbb{E}[L_{\text{full,ign}}(\theta; X)|X^{(0)}; \theta^r],$$
where $L_{\text{full,ign}}$ is the complete likelihood
$$L_{\text{full,ign}}(\theta; X) = f_X(x; \theta) \tag{1.6}$$
- the M-step (Maximization) (at step r): Q is maximized over θ,
$$\theta^{r+1} = \underset{\theta}{\text{argmax}}\, Q(\theta; \theta^r).$$

From a theoretical point of view, Dempster et al. (1977) have proven that the EM algorithm produces a monotonically increasing sequence for the expected likelihood $(Q(\theta; \theta^r))_{r \geqslant 0}$. However, the EM algorithm can be trapped in local maxima (for example, McLachlan and Krishnan (2007) gives more details about the EM algorithm, both on the theoretical and practical aspects).

Note that the two steps of the EM algorithm do not involve the imputation of missing values as such, although the E-step can be reduced to this in some cases, in particular if $L_{\text{full,ign}}(\theta; X)$ is linear in $X^{(1)}$ (Sportisse et al., 2020). Therefore, this algorithm is not an imputation method, it allows to modify the estimation process to handle incomplete datasets.

Variants to tackle the computational burden In the specific case of (multivariate) Gaussian data, explicit formulae can be derived (the E-step requires the computation of sufficient statistics of $X^{(1)}$ so that both steps of the EM algorithm can be written in closed-form, but it is not generally the case, as discussed by (Meng and Rubin, 1993). To fix the ideas, consider the continuous case. The E-step consists of computing

$$Q(\theta; \theta^r) = \int_{\mathcal{X}^{(1)}} f_X(x; \theta) f_{X^{(1)}|X^{(0)}}(x^{(1)}|x^{(0)}; \theta) dx^{(1)}, \tag{1.7}$$

where $f_{X^{(1)}|X^{(0)}}$ denotes the conditional distribution of the missing components given the observed ones and the missing-data pattern. As this integral can be not explicit, one can resort to sampling methods, such as Monte Carlo sampling (Ibrahim, 1990) when the conditional distribution is known, or adaptive rejection sampling (Gilks and Wild, 1992; Ibrahim et al., 1999) otherwise. However, these methods are computationally costly, as they involve many drawings from the conditional distribution at each step of the EM algorithm.

To overcome this issue, Celeux and Diebolt (1985) have proposed the stochastic EM algorithm (SEM), for which the expectation step is replaced by a "drawing" step as follows,

- the SE-step (at step r): draw the missing values $(x^{(1)})^{r+1} \sim f_{X^{(1)}|X^{(0)}}(.|x^{(0)}; \theta^r)$
- the M-step (at step r): maximize the full likelihood $L_{\text{full,ign}}$ in (1.6) and compute

$$\theta^{r+1} = \underset{\theta}{\operatorname{argmax}}\, f_X(x^{(0)}, (x^{(1)})^{r+1}; \theta).$$

Contrary to the Monte Carlo or adaptive rejection samplings, the SE-step requires the drawing of only one sample for $(x^{(1)})^{r+1}$. The SEM algorithm also has another possible advantage over the EM algorithm: it is not trapped by the first local maximum encountered of the likelihood function and it converges to the neighbourhood of the maximum likelihood (Celeux and Diebolt, 1985). Delyon et al. (1999) propose another stochastic approximation of the EM algorithm, called the SAEM algorithm, which has been proven to converge to the maximum likelihood under specific assumptions, but is more difficult to implement.

Variants to obtain confidence intervals The EM algorithm applied with its initial form does not provide any variance for the estimates, preventing from obtaining associated confidence intervals. Note that the variance of the estimates can be estimated using extensions of the EM algorithm, such as the supplemental EM algorithm (Meng and Rubin, 1991).

Further references on the EM algorithm in the specific regression framework will be given in Section 1.4.1. In addition, the EM algorithm will be derived and discussed for the MNAR case in Section 1.5.3.

1.3.3 Imputation

A popular approach consists of imputing missing values. It allows to obtain a complete dataset, for which any classical analysis method can be applied. Note that even though we impute the data, the aim is not to impute as well as possible, but to estimate parameters of an underlying model.

Single imputation A first strategy is to propose a predicted values for each missing entry, which is referred to as *single imputation methods*.

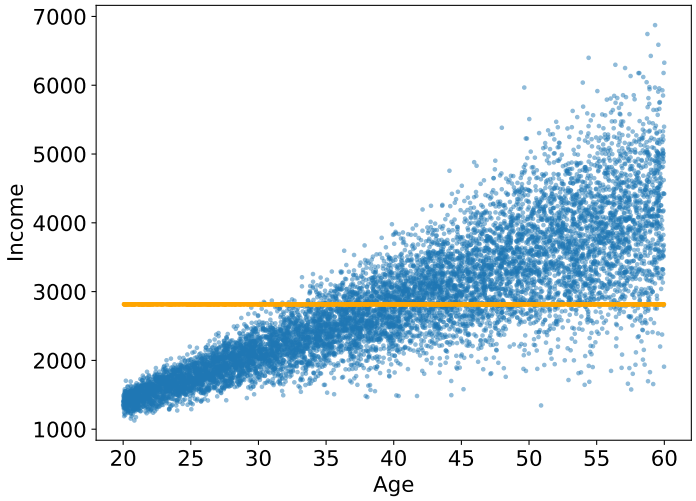


Figure 1.4: Imputation by the mean of the missing values in Income (see Figure 1.1).

Mean imputation The most popular method is the mean imputation. The principle is very simple: if the variable j of X contains missing entries, each one is replaced by the mean of the observed values of the variable $X_{\cdot j}$. Despite its simplicity of implementation, this method distorts the distribution of the data, which induces bias in estimators (Schafer and Graham, 2002), as illustrated in Figure 1.4.

Model-based methods The imputation can be performed assuming a joint model for the data $(X^{(0)}, X^{(1)})$ (see for example (Honaker et al., 2011) in the Gaussian case) or considering a fully conditional model (Van Buuren, 2018).

In nonparametric settings, some methods impute missing values using the similarities of the individuals, they include k-nearest neighbors algorithm, (Troyanskaya et al., 2001; Zhang, 2012) or a near concept called hot-decks procedures (see (Andridge and Little, 2010) for a complete review). Powerful methods relying on nonparametric assumptions also include random forest imputations (Stekhoven and Bühlmann, 2012). Besides, imputation methods based on deep learning techniques have also been proposed using generative adversarial networks (Yoon et al., 2018), denoising autoencoders (Gondara and Wang, 2018) and variational autoencoders (Mattei and Frellsen, 2019). In addition, a recent work (Muzellec et al., 2020) proposes an imputation strategy using optimal transport.

Low-rank methods The low-rank model has become very popular in a large variety of applications (genomics (Price et al., 2006), denoising (Gavish and Donoho, 2017), recommender system (Bell and Koren, 2007)). It can approximate many datasets (Udell and Townsend, 2019), as soon as individual profiles can be summarized into a limited number of

general profiles, or dependencies between variables can be established. Recently, low-rank methods have proven to be a very powerful solution for dealing with missing values (Josse et al., 2016a; Kallus et al., 2018; Robin, 2019).

A matrix $\Theta \in \mathbb{R}^{n\times d}$ has a *low rank*, if its rank, refereed to as the dimension of the vector space generated by its columns, is small compared to the dimensions n and d. More precisely, denoting the rank of Θ as $r \geqslant 1$, the matrix Θ has a *low rank* if $r \ll \min\{n, d\}$, where $\ll$ can be interpreted as $\exists r_{\max} \geqslant 1, r < r_{\max} < \min\{n, d\}$. Low rank models often assume that the dataset X is a noisy realisation of Θ, so that

$$X = \Theta + \epsilon, \tag{1.8}$$

where ϵ is a noise matrix. In this case, to estimate Θ without missing values, the most classical method for dimensionality reduction is the Principal Component Analysis (PCA) (Jolliffe, 1986), which nearly amounts to solve the optimization problem

$$\hat{\Theta} \in \text{argmin}_{\Theta} \|(X - \Theta)\|_F^2 \text{ s.t. } \text{rank}(\Theta) \leqslant r,$$

with $\|.\|_F$ the Frobenius norm. To estimate Θ, when the columns of X have been initially centered, a step of the PCA consists of computing the truncated singular value decomposition (SVD) as follows,

$$\text{SVD}_r(X) = U_{.(r)} D_{(r)(r)} V_{.(r)}^T,$$

where $U \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{d\times d}$ are orthonormal matrices containing the left and right singular vectors of X and $D \in \mathbb{R}^{n\times d}$ where the diagonal coefficients are the singular values or X and the others are zero. $U_{.(r)} = (U_{i,j})_{i\in\{1,\dots,n\},j\in\{1,\dots,r\}}$ (resp. $V_{.(r)} = (V_{i,j})_{i\in\{1,\dots,d\},j\in\{1,\dots,r\}}$) denotes the submatrix of U (resp. V) defined by its r first columns and $D_{(r)(r)} = (D_{i,j})_{i\in\{1,\dots,r\},j\in\{1,\dots,r\}}$ is the submatrix extracted from D keeping only the r first rows and and the r first columns.

Classical methods to handle missing values are based on convex relaxations of the rank such as the nuclear norm $\|.\|_\star$ and consists of solving the following penalized weighted least-squares problem

$$\hat{\Theta} \in \text{argmin}_{\Theta} \|(\mathbf{1}_{n\times d} - M) \odot (X - \Theta)\|_F^2 + \lambda \|\Theta\|_\star, \tag{1.9}$$

with $\lambda > 0$ a regularization term, $\odot$ the Hadamard product (by convention $0 \times \texttt{NA} = 0$) and $\mathbf{1}_{n\times d} \in \mathbb{R}^{n\times d}$ with each of its entry equal to 1. The estimator $\hat{\Theta}$ of Θ is then the matrix which fits the data best in the mean squared sense (first term in (1.9)) and which is likely to be of low rank (second term in (1.9)). To solve the optimization problem, Hastie et al. (2015) propose to use a proximal gradient method, leading to iterative soft thresholding algorithm (ISTA) of the SVD. More particularly, this iterative algorithm consists of two steps, given a matrix Θ^0,

- *Estimation* step (at step t): perform the threshold SVD of the complete matrix

$$X^t = (\mathbf{1}_{n\times d} - M) \odot X + M \odot \Theta^t,$$

which leads to

$$\text{SVD}_\lambda(X^t) = U^t D_\lambda^t V^t, \tag{1.10}$$

where $U^t \in \mathbb{R}^{n\times r}$, $V^t \in \mathbb{R}^{r\times d}$ are orthonormal matrices containing the singular vectors of X^t and $D^t_\lambda \in \mathbb{R}^{r\times r}$ is a diagonal matrix such that its diagonal terms are $(D^t_\lambda)_{ii} = \max((\sigma_i - \lambda), 0), i \in \{1, \dots, r\}$, with σ_i the singular values of X^t.

- *Imputation* step (at step t): the entries of Θ^t corresponding to missing values in X are replaced by the values of $\text{SVD}_\lambda(X^t)$ in (1.10),

$$\Theta^{t+1} \odot M = \text{SVD}_\lambda(X^t) \odot M.$$

Other works have suggested related algorithms (Josse et al., 2016a) or extended these methods to handle both continuous and count data (Udell et al., 2016; Robin et al., 2020).

In Equation (1.8), note that the low-rank matrix is a fixed parameter. One could consider the probabilistic principal component analysis (PPCA) (Tipping and Bishop, 1999) instead, for which the probabilistic model can be advantageously exploited. The data matrix X is a noisy realisation of the factorization of the loading coefficients $B \in \mathbb{R}^{r\times d}$ and r latent variables groupes in the matrix $W \in \mathbb{R}^{n\times r}$,

$$X = \mathbf{1}\alpha + WB + \epsilon, \tag{1.11}$$

with $\alpha \in \mathbb{R}^d$ which allows X to have non-zero means, $\mathbf{1} = (1 \dots 1)^T \in \mathbb{R}^n$ and $W \in \mathbb{R}^{n\times r}$ such that $W_{i.} \sim \mathcal{N}(0_r, \text{I}_{r\times r}), \forall i \in \{1, \dots, r\}$. As soon as $r < d$, this model is motivated by the fact that a few latent variables explain the dependencies between the variables, which can be seen as a low-rank model with random effects. Ilin and Raiko (2010) discuss several approaches to deal with missing values in the PPCA model. Note that in the methods based on a model either with fixed effects or with random effects (PPCA), dealing with missing values allows simultaneously to make the parameters estimation and to perform single imputation of the data.

✍ In this dissertation, both Chapters 2 and 3 consider low-rank methods with fixed and random effects to deal with MNAR data, whereas the methods presented in this section are only valid for M(C)AR data.

Multiple imputation The single imputation does not reflect the variability of imputation. To overcome this potential issue, multiple imputation (Rubin, 2004) can be used. The method consists of generating M plausible values for each missing value, leading to M complete datasets, $\hat{X}^1, \dots, \hat{X}^M$. The analysis is then performed on each imputed data sets and results are combined so that the final variance accounts for the variability induced by the imputation. The most popular multiple imputation is the one developed by Buuren and Groothuis-Oudshoorn (2010), which use multiple imputations by chained equations, i.e. iterative conditional distributions assuming a Bayesian framework. Indeed, multiple imputation is intrinsically linked to the Bayesian approach (see (Erler, 2019; Little and Rubin, 2019, Chapter 10) for more details).

Murray and Reiter (2016) consider a nonparametric Bayesian strategy and Audigier et al. (2016b) use a Bayesian principal component analysis. More recently, Erler et al. (2019)

propose a fully Bayesian unified framework, extension of (Buuren and Groothuis-Oudshoorn, 2010).

Simple imputation methods have also been adapted to offer multiple imputations, without necessarily giving rules for combining the different results (Honaker et al., 2011; Josse et al., 2016a; Mattei and Frellsen, 2019).

1.3.4 Naive imputation coupled with adapted algorithms

When the main goal is to estimate some parameters of an underlying model (and not to perform matrix completion), another strategy is to naively impute the missing values and then to account for the imputation error by adapting the subsequent algorithm. More precisely, if the goal is to apply an algorithm A (available in the case without missing values), the two steps are the following ones.

(i) First step (the easiest one): naively impute the missing values, say by zero, to get a complete dataset $\tilde{X}$. It leads to

$$\tilde{X} = X \odot (\mathbf{1}_{n\times d} - M).$$

(ii) Second step (the difficult part): adapt algorithm A to account for the error induced by the imputation of the missing values in (i) and apply this *debiased* version to the complete dataset $\tilde{X}$.

This strategy has been mostly studied in the linear regression setting, when the covariates are missing, i.e. $\mathbb{E}[Y|X = x] = f(x)$, with f a linear function. In a sparse regression context, Rosenbaum et al. (2010) and Loh and Wainwright (2011) adapt the Dantzig selector and LASSO by debiasing the resulting covariance matrix. Besides, in a ridge regression framework, Ma and Needell (2018) consider debiased gradients to apply the stochastic gradient descent (SGD) algorithm. More recently, in a nonlinear setting, Yi et al. (2019) propose a heuristic to debiase zero-imputation in neural networks.

✍ In this dissertation, Chapter 4 and Appendix A use this strategy to handle missing values with the averaged stochastic gradient algorithm and with the Robust Lasso-Zero algorithm.

Inverse Probability Weighting (IPW) method The approach to naively impute missing values and adapt a classical algorithm has actually *false* similarities with the IPW methods (Seaman and White, 2013). Indeed, the latter consists of keeping only complete observations and reducing the induced bias by reweighting the loss with respect to the complete observations with their probabilities of being observed. Thus, this method does not use the whole matrix, as it is the case for the strategy mentioned before. These approaches often consider simple reweighting, assuming that the covariates are fully observed and only the outcome variable may be missing.

1.3.5 Comparison of the methods

We now compare the methods that have been introduced in this section. First, the EM algorithm (Section 1.3.2) is perfectly well fitted to the aim of estimating parameters. However, it has to be established for each statistical model, meaning that if one wants to do logistic regression with missing values, one has to derive an EM algorithm and if on the same data, one wants to do unsupervised clustering, one has to develop another algorithm. In addition, it is not often an easy task to design EM algorithms. For instance, it has been seen that it can involve non-explicit integrals (see (1.7)). Moreover, as stated before, this algorithm does not provide confidence intervals of the estimates, without being coupled with other algorithms. This remains yet a powerful algorithm to handle missing data, and amenable to the MNAR case, as shown in Section 1.5.3. Note also that even though no imputation is performed, this step can be added easily.

The single imputation (Section 1.3.3) allows to obtain a complete dataset and is easier to implement. As for the multiple imputation (Section 1.3.3), the difficulty is only to propose the imputation, because the estimation part is performed with usual algorithms applied to the imputed datasets. Note that mean imputation lead to biases in the estimates (Schafer and Graham, 2002). In particular, Jones (1996) has studied the induced bias in the regression framework. The main drawback of any single imputation method is that it does not take into account the uncertainty of the imputation (Schafer and Graham, 2002). On the contrary, multiple imputation accounts for the uncertainty, but then requires specific rules for combining the results, which are not defined for each algorithm (in particular no results are given, either for regression in high-dimension or for unsupervised learning or even for variable selection).

Finally, naive imputation coupled with debiasing of classical algorithms is an easy strategy to implement, as soon as the algorithm has been debiased. The goal is not to impute missing values, but to adapt powerful algorithms (Lasso, SGD,...) to the missing values case. For the sake of clarity, Table 1.3 gives an overview of the methods comparison.

In conclusion, there is no general recommendation as to which method to choose. Keep in mind that the choice depends mostly on the goal and on both the data and missing-data types.

1.4 Specific learning frameworks with missing data

1.4.1 Linear regression with missing data

The literature on how to deal missing values is vast (see Section 1.3), but there are still some challenges even for linear regression models and M(C)AR data. Consider the classical model

$$Y = X\beta + \epsilon \tag{1.12}$$

where $Y \in \mathbb{R}^n$ is the outcome variable, $X \in \mathbb{R}^{n\times d}$ the covariates, $\beta \in \mathbb{R}^d$ the regression parameter and $\epsilon \in \mathbb{R}^d$ the noise term, traditionally assumed to be Gaussian $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{1}_n)$, with σ^2 its variance. We assume that the covariates contain missing values.

Method	**Simple to implement**	**Imputation**	**Confidence intervals**	**Main drawbacks**
Complete case	✓	naive	✗	information loss, bias estimates
EM	✗	not directly, but can be obtained	can be obtained	specific EM algorithm for each statistical model
Single imputation	✓	single	✗	possibly biased estimates (when too simple imp.)
Multiple imputation	✓	multiple	✓	specific rules for combining results
Naive imp. + debiasing	✓	not the goal	✗	debiasing each algorithm

Table 1.3: Comparison of the methods introduced in Section 1.3

Likelihood-based approaches When the aim is to estimate the parameter β, Novo and Schafer (2013) provide an implementation using the EM algorithm (Section 1.3.2). Instead of considering the likelihood given in (1.3), they consider the following one, which involves the response variable y,

$$L_{\text{full}}(\beta, \theta, \phi; Y, X, M) = f_{Y,X,M}(y, x, m; \beta, \theta, \phi).$$

Besides, Murray et al. (2018) derive the multiple imputation (Section 1.3.3) in the linear case. In both methods, strong parametric assumptions are made on the distribution of the covariates, often assumed to be Gaussian.

Stochastic gradient descent algorithm Without missing values, a powerful algorithm for linear regression models, which requires only few parametric assumptions, is the stochastic gradient descent algorithm (SGD) (Robbins and Monro, 1951). In general, two different settings are studied: (i) the streaming setting, i.e. when the data comes in as they go along, or (ii) the finite-sample setting, i.e. when the data size is fixed and form a finite design matrix X. In both cases, the observations $(X_{i.}, Y_i)$ are assumed to be i.i.d.. In order to estimate β in (1.12), the aim is to solve the least-squares optimization problem,

$$\hat{\beta} \in \operatorname{argmin}_{\beta} \mathbb{E}_{X,Y}[(Y_i - X_{i.}\beta)^2] := R(\beta), \tag{1.13}$$

where R is the theoretical risk and $\mathbb{E}_{Y,X}$ denotes the expectation over the distribution of $(X_{i.}, Y_i)$ (independent of i since the observations are i.i.d.). The SGD algorithm is an iterative algorithm which computes the current iterate by moving it in the opposite direction of a unbiased gradient as follows, for step k,

$$\beta_k = \beta_{k-1} - \alpha g_k(\beta_{k-1}),$$

where α is the step-size and $\mathbb{E}\left[g_k(\beta_{k-1})|\mathcal{F}_{k-1}\right] = \nabla R(\beta_{k-1})$, $\mathcal{F}_{k-1} = \sigma(X_{1.}, Y_1, \ldots, X_{k-1.}, Y_{k-1})$ the σ-algebra. To solve (1.13), the ordinary least-squares estimator could be considered,

$\hat{\beta} = (X^T X)^{-1} X^T Y$, but it requires the inversion of the matrix $(X^T X)^{-1}$ of size d times d which can be computationally costly. Due to its cheap computational cost and memory per iteration, the stochastic gradient is a key ingredient in machine learning.

In presence of missing values, the goal is to find the unbiased gradients g_k depending on known quantities, i.e. $\tilde{X}_{k.} = X_{k.} \odot (\mathbf{1}_{n\times d} - M)$ and Y_k. Ma and Needell (2018) propose to naively impute missing values and adapt the SGD algorithm in the linear regression model (method presented in Section 1.3.4).

However, the references given above are not suitable for the high-dimensional setting, when the dimension of the observations d is larger than the number of the observations n. This case is encountered in many applications, such as genomics, because the gene expression is naturally represented by many variables.

High-dimensional setting In order to tackle the curse of dimensionality, classical methods assume that β is s-sparse, i.e. only s out of its d entries are different from zero. To estimate β, the classical strategy is to penalize the least-squares problem $\beta \in \text{argmin}_\beta \|Y - X\beta\|^2$, with the Ridge regularization (Hastie, 2020),

$$\hat{\beta} \in \text{argmin}_\beta \|Y - X\beta\|^2 + \lambda \|\beta\|_2^2, \tag{1.14}$$

where $\|\beta\|_2 = \sum_{j=1}^d \beta_j^2$ and λ is the regularization parameter to tune. This regularization is proposed in Chapter 4 for the averaged SGD algorithm. However, this regularization does not allow to select relevant variables, as the LASSO (Tibshirani, 1996), by allowing some coefficients to be zero.

$$\hat{\beta} \in \text{argmin}_\beta \|Y - X\beta\|^2 + \lambda \|\beta\|_1^2, \tag{1.15}$$

where $\|\beta\|_1 = \sum_{j=1}^d |\beta_j|$.

Loh and Wainwright (2011); Datta et al. (2017) propose to naively impute the missing values and adapt the LASSO by debiasing the covariance matrix (same spirit as the methods proposed in Section 1.3.4). Bogdan et al. (2015) use another penalization without missing values to penalize the highest coefficients more strongly and Jiang et al. (2019) derive an algorithm to the missing values case.

Chen et al. (2013) point that if the data are naively imputed, the linear regression (1.12) can be be rewritten in the form of the sparse corruption model,

$$Y = X\beta + \sqrt{n}\omega + \epsilon,$$

where ω is a k-sparse vector. Consider the case of naive imputation by zero, a complete matrix $\tilde{X} = X \odot (\mathbf{1}_{n\times d} - M)$ is obtained and ω can represent the corruption due to imputations $\omega = \frac{1}{\sqrt{n}}(X - \tilde{X})\beta$.

✍ In this dissertation, we adapt the averaged SGD algorithm to the missing values case in Chapter 4 and makes it suitable for the high-dimensional setting. Appendix A presents a thresholded robust algorithm for model selection in the high-dimensional setting. Both methods consider naive imputation and debiased algorithms.

1.4.2 Supervised learning with missing data

In supervised learning, a key question, which comes up very often in applications, is how to deal with missing data if the goal is to predict an outcome variable Y when the covariates X contain missing values. As a reminder, the algorithms are learned from training data and the results of new observations are then predicted by applying this learning. Two scenarios can be considered, which imply different strategies:

i) the new observations do not contain missing values (efforts to better collect data have been made),

ii) the new observations contain missing values.

In the first case, the distribution of interest is that of complete data, whereas in the second case, this is the distribution of data containing missing values which should be estimated.

For i), a strategy consists of imputing the train set with imputation methods reviewed in Section 1.3.3 and applying a classical learner to the complete training dataset (depending on the cases: linear regression, random forest, gradient boosting (Hastie et al., 2009)). Another solution is to use learning algorithms adapted to the case of missing data, such as stochastic gradient algorithm (see Chapter 4).

The literature dealing with the case ii) is sparse. Josse et al. (2019) show that the mean imputation is consistent for a powerful learner (including mostly random forests). In a linear case, Le Morvan et al. (2020b) propose to specify the distribution of data containing missing values with ReLU activation functions, i.e. they want to find a linear function f such that $Y = f(X \odot (\mathbf{1}_{n\times d} - M), M)$. Le Morvan et al. (2020a) propose a general algorithm to tackle this issue for different missing-data mechanisms (including MAR and self-masked MNAR). Recently, You et al. (2020) also address the prediction task for graph representation learning (only under the MCAR assumption).

1.4.3 Model-based clustering with missing data

Unsupervised learning concerns the analysis of datasets without outcome variables (unlabeled) for which the aim is to group individuals. In particular, the model-based paradigm (McLachlan and Basford, 1988; Zhong and Ghosh, 2003; Bouveyron et al., 2019), relying on parametric assumptions for the data distribution, allows to perform clustering, by providing interpretable models, valuable to understand the connections between the constructed clusters and the features in play, by using the estimation of the parameters.

In model-based clustering, the goal is then to estimate an (unknown) partition of n individuals $X_{1.}, \ldots, X_{n.}$ into K groups. This partition can be encoded using the matrix $Z = (Z_{1.}|\ldots|Z_{n.})^T \in \{0,1\}^{n\times K}$ whose i-th row $Z_{i.} = (Z_{i1}, \ldots, Z_{iK})^T \in \{0,1\}^K$ is a group indicator vector for the i-th individual, with $z_{ik} = 1$ if $x_{i.}$ belongs to the class k , and $z_{ik} = 0$ otherwise. The model-based clustering relies on the assumption that the individuals $X_{1.}, \ldots, X_{n.}$ are an i.i.d. sample from the mixture distribution

$$f(x_{i.}; \pi, \theta) = \sum_{k=1}^{K} \pi_k f_k(x_{i.}; \theta_k), \tag{1.16}$$

where $\pi_k = \mathbb{P}(z_{ik} = 1)$ is the mixing proportion of the k-th component ($\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0$ for all $k \in \{1, \ldots, K\}$), $f_k(\cdot\,; \theta_k)$ is the distribution of the data in the k-th group parameterized by θ_k, $\pi = (\pi_1, \ldots, \pi_K)$ and $\theta = (\theta_1, \ldots, \theta_K)$ denotes the whole mixture parameter.

In order to deal with missing data, classical imputation methods are not designed for the final clustering task. Most existing methods propose to maximize the full observed likelihood given in (1.6) for M(C)AR data, written here for the specific model-based framework:

$$L(\pi, \theta; X^{(0)}) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \int_{\mathcal{X}_{i.}^{(1)}} \pi_k f_k(x_{i.}; \theta) dx_{i.}^{(1)} \right)$$

Hunt and Jorgensen (2003) implement the standard EM algorithm, Serafini et al. (2020) also propose an EM algorithm to estimate Gaussian mixture models in the presence of missing values by performing multiple imputations (with Monte Carlo methods).

✍ In this dissertation, Chapter 5 addresses the model-based clustering with MNAR data (with the selection models specification) and for mixed data (both continuous and categorical).

1.5 Dealing with MNAR data

1.5.1 MNAR specifications

The specification of the missing-data mechanism is a crucial but controversial part of MNAR data processing. In particular, the choice of the MNAR specification has a direct impact on the identifiability, the method to use and the sensitivity analysis.

As introduced in Section 1.2.3, statistical inference is conducted on the joint distribution (X, M) of the data and the missing-data pattern. However, this joint distribution is intractable. To illustrate this idea, consider the following factorization

$$f_{X,M}(x, m) = f_{X^{(1)}|X^{(0)},M}(x^{(1)}|x^{(0)}, m) f_{X^{(0)},M}(x^{(0)}, m). \tag{1.17}$$

The distribution $f_{X^{(0)},M}$ can be estimated from the data (only observed quantities) but some assumptions on $f_{X^{(1)}|X^{(0)},M}$ should be added as the final goal is to get the joint distribution $f_{X,M}$. In the literature, the two main approaches to model the joint distribution are

(I) the selection models (Heckman, 1976),

$$f_{X,M}(x, m; \theta, \phi) = f_X(x; \theta) f_{M|X}(m|x; \phi), \tag{1.18}$$

(II) the pattern-mixture models (Little, 1993),

$$f_{X,M}(x, m; \xi, \varphi) = f_{X|M}(x|m; \xi) f_M(m; \varphi), \tag{1.19}$$

where ξ and φ are the parameters of the conditional distribution of the data given the missing-data pattern $f_{X|M}$ and the missing-data pattern f_M respectively.

The selection models consider a factorization of the joint distribution involving the distribution of the data f_X and the incidence of the missingness as a function of X, with $f_{M|X}$. Regarding the definition of the missing-data mechanisms (Definition 3), this formulation could be the most natural and under parametric assumptions, modeling the data distribution also seems natural. For instance, in the continuous case, typical assumptions set the data to be Gaussian. For the missing-data mechanism, widely used distributions include the logistic or the probit one.

The pattern-mixture models factorize the joint distribution by specifying the missing-data distribution f_M and the conditional distribution of the data given the missing-data pattern $f_{X|M}$. The main advantage of the pattern-mixture models is that it is clear what information is available from the observed data and what quantities need to be extrapolated, i.e. for what quantities prior information that cannot be tested with the observed data is needed. In particular, for the univariate setting ($d = 1$),

$$f_{X,M}(x, m; \xi, \varphi) = f_{X|M=0}(x|0; \xi) f_M(0; \varphi) + f_{X|M=1}(x|1; \xi) f_M(1; \varphi), \tag{1.20}$$

the quantity which needs to be "extrapolated" is $f_{X|M=1}$ which is the distribution of the data conditionally to be missing. This factorization can be easily used to derive identifiability results. In addition, some authors point that pattern-mixture can be preferred to conduct sensitivity analysis (Glynn et al., 1986; Miao et al., 2015; Little and Rubin, 2019, Chapter 15), as discussed in Section 1.5.4.

Council et al. (2010) propose an example to understand the different uses of the specifications. In a clinical survey, *for each decrease of 0.1 in quality of life, the chance of being missing doubles*, so that the natural specification is the selection models (I), as how the occurrence of missing values is related to the data is known. If *participants with missing data have a 0.1 lower quality of life than those observed*, the pattern-mixture models (II) fits perfectly well, because in (1.19), the first term of the factorization $f_{X|M}$ represents the data distribution in the strata defined by different missing-data patterns (here, the participants do not have the same distributions for those missing and observed).

To conclude, the MNAR specification choice mainly depends on the assumptions that are the easiest to extrapolate from the data, i.e. assumptions on the missing-data mechanism or assumptions on the distribution of the observed data and the missing data separetely. However, keep in mind that the identifiability of MNAR models is not guaranteed (see Section 1.5.2), and the choice of the MNAR specification, for example between the selection models and the pattern mixture models, may facilitate proofs of identifiability.

As an alternative to selection models and pattern mixture models, some authors consider for example the shared-parameter models (Beunckens et al., 2008; Creemers et al., 2010; Kuha et al., 2018), in which a variable suject to the missingness, say $X_{\cdot j}$, and its missing-data pattern $M_{\cdot j}$ are linked through a latent (unobserved) variable.

1.5.2 Identifiability

Without loss of generality, the univariate setting ($d = 1$) is considered to present the notion of identifiability. The parameters of the joint distribution $f_{X,M}$ are said identifiable if the

joint distribution $f_{X,M}$ can be uniquely determined from the observed distribution $f_{X,M=1}$.

Definition 7 (Identifiability of the parameters)**.** *The family of parameters (θ, ϕ) is identifiable if for all (X, M) and (X', M') of distributions parametrized by (θ, ϕ) and (θ', ϕ'),*

$$f_{X,M=1}(x, m = 1; \theta, \phi) = f_{X',M'=1}(x, m = 1; \theta', \phi') \Rightarrow (\theta, \phi) = (\theta', \phi')$$

Note first that M(C)AR data preserve parameter identifiability, i.e. if the model parameters are identifiable without missing values, they remain identifiable with the missing data. To illustrate this idea, the univariate case leads to

$$f_{X,M=1}(x, m = 1; \theta, \phi) = f_X(x; \theta) f_{M=1|X}(m = 1|x; \phi) = f_X(x; \theta) f_{M=1}(m = 1; \phi),$$

with the last term proportional to $f_X(x; \theta)$, as the statistical analysis is only conducted on the parameter of interest θ and it is enough to have the identifiability of the parameters of the distribution f_X.

A key issue As pointed by Baker and Laird (1988) and more recently by Miao et al. (2016), the identifiability is not guaranteed for MNAR mechanisms and many models lead to non identifiable parameters, even if parametric assumptions are made. In particular, the identifiability of the parameters of the data distribution θ is conditional on the identifiability of the parameters of the missing-data mechanism ϕ, as illustrated in the following example.

Example 1 (Need of identifiability of the missing-data mechanism parameters[2])**.** *Let us consider a binary matrix, $X \sim \mathcal{B}(p)$, containing MNAR values $X = (1, \texttt{NA}, 0, 1, \texttt{NA}, 0)$. We cannot retrieve the parameter p of the binomial law of X without identifying the parameters of the missing-data mechanism, i.e. the conditional law of M given X. Indeed, if X is missing only if X is equal to 1, thus $X = (1, 1, 0, 1, 1, 0)$ and $p = 2/3$. If X is missing only if X is equal to 0, $X = (1, 0, 0, 1, 0, 0)$ and $p = 1/3$. Thus, the parameter p is not identifiable, because two equal observed distributions can lead to different parameters. One should consider the conditional law of M given X.*

Leveraging additional information To get identifiability guarantees, the idea is simple: prior or additional knowledge about the missing-data mechanism should be added (Molenberghs et al., 2008). The consequences of no prior information are for example discussed empirically by Ipsen et al. (2020) or Tang et al. (2014) where consistencies of the estimators are obtained only with the use of auxiliary information, even though neither work adresses the identifiability issue.

In the parametric setting, Miao et al. (2016) prove the identifiability of parameters in Gaussian data and mixture model. Their results require specific known forms of the missing-data mechanism $f_{M|X}$, such as a logistic or probit model. The example below illustrates their work for a self-masked mechanism and Gaussian data.

[2]This example is largely inspired by one of the lectures of Ilya Shipster.

Example 2 (Identifiability of the parameters if the data are Gaussian and the mechanism is probit (Miao et al., 2016))**.** *Let us consider* $X \sim \mathcal{N}(\mu, \Sigma)$ *and let us assume that* X *may contain self-masked MNAR values, i.e.*

$$f_{M|X}(m = 1|x) = F(\phi_0 + \phi_X x),$$

where $\phi = (\phi_0, \phi_X)$ *is the parameter of the missing-data mechanism. Miao et al. (2016) states the identifiability of the parameters* (μ, Σ, ϕ) *by assuming the following*

1. F *is a known and strictly monotone distribution function.*
2. *The left tail decay rate of* F *is not exponential, i.e.*

$$\forall \delta > 0, \lim_{z \to -\infty} \frac{F(z)}{e^{-\delta z}} = 0 \text{ or } +\infty.$$

The first condition holds if the missing-data mechanism distribution is logistic or probit, which are the most encountered specifications. However, the second condition is true only for the probit distribution. Note that prior information on the form of the missing-data mechanism is required to make the parameters identifiable. Besides, this result excludes a wide variety of models, which shows that the study of the identifiability for MNAR data is crucial.

More specifically, to add prior information in linear regressions models, a method to make the parameters identifiable consists of using an instrument variable, called *shadow variable*, which is independent from the missing-data pattern given the data. In particular, assume that the outcome variable Y is missing and the covariates X are fully observed. A shadow variable Z is associated with the missing variable Y, conditional on the observed data X, but independent of the missing-data pattern M given both the missing variable Y and the observed ones X. It can be formalized as follows,

$$\exists Z, \qquad Z \not\perp\!\!\!\perp Y|X \quad \text{and} \quad Z \perp\!\!\!\perp M|(Y, X).$$

A direct acyclic graph (DAG) summing up this definition is drawn in Figure 1.5. Y is caused by Z and X, i.e. $Y = f(Z, X)$. M is caused by X, Y but not Z, i.e. the missing-data mechanism $f_{M|X,Y,Z}$ depends on X and Y but not on the shadow variable Z.

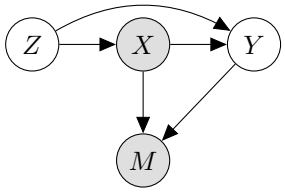


Figure 1.5: DAG for a shadow variable Z. The nodes in grey represent fully observed variables and the edges from Z to Y means that Z causes Y.

Semi-parametric models have been shown particularly adapted to address the identifiability issue by using this method (Wang et al., 2014; Zhao and Shao, 2015; Miao and Tchetgen, 2018; Zhao and Ma, 2021).

Another similar technique to identify the parameters in a regression setting is to use an *instrumental variable* Z independent of the missing variable Y conditional to X but related to M conditional to X (see for example the work of Morikawa et al. (2017)). Roughly speaking, this technique switch the role of Z and Y compared to the shadow variable strategy.

$$\exists Z, \qquad Z \not\perp\!\!\!\perp M|X \quad \text{and} \quad Z \perp\!\!\!\perp Y|X.$$

The terminology can be sometimes confusing. For example, latent variable models which consider a shared-parameter model (Beunckens et al., 2008; Creemers et al., 2010; Kuha et al., 2018) lead to identifiable parameters because they use an instrumental variable and not a shadow variable.

Graphical-based methods Another part of the identifiability literature is interested to exploit causal inference techniques by handling missing data using graphical models. The latters encode assumptions on the missing-data mechanism well (Mohan et al., 2013; Ilya et al., 2015) and allow to consider nonparametric settings. For discrete variables, Mohan et al. (2013); Mohan and Pearl (2014); Ilya et al. (2015) develop algorithms to identify the parameters. Recent works (Bhattacharya et al., 2020; Nabi et al., 2020) aim to unify results on identification for graphical models, by giving a graphical condition under which the data distribution is identified by the observed data distribution. In particular, Nabi et al. (2020) prove that if $X^{(1)}$ corresponds to the missing variables, $X^{(0)}$ to the observed ones, and M to the missing-data pattern, the full law $(M, X^{(1)}, X^{(0)})$, and thus the target data law $(X^{(1)}, X^{(0)})$ is identified if the following conditions hold:

1. The mechanism is not self-masked, i.e. in a graphical point of view, the edge between $X^{(1)}_{.j}$ and $M_{.j}$ is not allowed.

2. There is no *colluder*, i.e. the following relation is not allowed: $X^{(1)}_{.j} \to M_{.i} \leftarrow M_{.j}$.

This result gives a sufficient condition to retrieve the target law $(X^{(0)}, X^{(1)})$ but not a necessary condition. In addition, they prove that such models are sub-models of the itemwise conditionally independent nonresponse model (ICIN), which is a general manner to encode not self-masking mechanisms containing no colluders, introduced by (Shpitser, 2016; Sadinle and Reiter, 2017), when

$$\forall j \in \{1, \dots, d\}, X_{.j} \perp\!\!\!\perp M_{.j}|(X_{.k})_{k\neq j}, (M_{.k})_{k\neq j}.$$

For the sake of clarity, the DAG associated to the ICIN model has been drawn in Figure 1.6 in the case of $d = 3$.

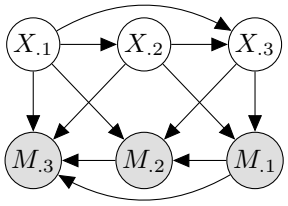


Figure 1.6: DAG for the ICIN model for $d = 3$. The nodes in grey represent fully observed variables and the edges from $X_{\cdot 1}$ to $X_{\cdot 2}$ means that $X_{\cdot 1}$ causes $X_{\cdot 2}$.

Mohan (2018) obtains identifiability guarantees for the self-masked mechanism by using an auxiliary variable, which is actually the same method as the shadow variable. Similarly, for continuous variables in a linear case, Mohan et al. (2018) show identifiability of the parameters. Besides, Sadinle and Reiter (2019) consider sequential additive nonignorable mechanism for which they also assume auxiliary information on the variables (use of auxiliary information on marginal distributions such as the moments of the variables).

Even though the point of view of graphical models is presented as dissociated from the methods of identifiability in semi-parametric models, the methods for making parameters identifiable have similarities and would benefit from being unified.

✍ In this dissertation, Chapters 3 and 5 address the identifiability issue in a PPCA model and in a model-based clustering.

1.5.3 Existing methods

Modeling the joint likelihood When the mechanism is MNAR, the maximization of the full log-likelihood given in (1.3) should be considered (and the ignored version in (1.6) discarded), the mechanism being *not ignorable* (see Section 1.2.3). In particular, as in Section 1.3.2, the EM algorithm can be derived for MNAR data and the inference is now conducted on both the data distribution parameters θ and the missing-data pattern distribution parameters ϕ (taking the notation of the selection models, but it does not exclude other specifications as (II)). Given $(\theta^{(0)}, \phi^{(0)})$, the algorithm steps have the following form

- the E-step (Expectation) (at step r):

$$Q(\theta, \phi; \theta^r, \phi^r) = \mathbb{E}[L_{\text{full}}(\theta, \phi; X, M)|X^{(0)}, M; \theta^r, \phi^r].$$

- the M-step (Maximization) (at step r):

$$\theta^{r+1}, \phi^{r+1} \in \underset{\theta,\phi}{\operatorname{argmax}}\, Q(\theta, \phi; \theta^r, \phi^r).$$

If the continuous case and the selection models are considered, the E-step is written

$$Q(\theta, \phi; \theta^r, \phi^r) = \int_{\mathcal{X}^{(1)}} f_X(x; \theta) f_{M|X}(m|x; \phi) f_{X^{(1)}|X^{(0)},M}(x^{(1)}|x^{(0)}, m; \theta, \phi) dx^{(1)},$$

where $f_{X^{(1)}|X^{(0)},M}$ denotes the conditional distribution of the missing component given the observed ones and the missing-data pattern. This integral has an explicit form only in rare cases. If sampling methods are used, the difficulty is to draw from the conditional law $f_{X^{(1)}|X^{(0)},M}$ which is hard to model. To estimate the parameters of generalized linear models when the covariates may have MNAR values, Ibrahim et al. (1999) propose an EM algorithm using an adaptive rejection sampling in the E-step to draw from the conditional law $f_{X^{(1)}|X^{(0)},M}$.

Likelihood-based approaches, especially the EM algorithm, can be computationally costly and require to model the missing-data mechanism. These parametric assumptions are often untestable and the results can be very sensitive to departure from these assumptions (see Section 1.5.4).

Besides, Tabouy et al. (2020) and Frisch et al. (2020) use a variational EM algorithm for Stochastic Block Models and for Latent Block Model in presence of MNAR values. Other likelihood-based approaches include an imputation and estimation procedure using a kernel regression method for the exponential tilting model (Tang et al., 2014) and an imputation method in a PPCA model using variational autoencoders (Ipsen et al., 2020). In addition, some authors (Marlin and Zemel, 2009; Hernández-Lobato et al., 2014; Wang et al., 2019) consider a joint modeling of (X, M) and debiase existing methods for MCAR data, for instance with inverse probability weighting approaches. Recently, De Chaumaray and Marbac (2020) propose to perform clustering via a mixture model using the pattern-mixture models to formulate the joint distribution.

Semi-parametric models The semi-parametric models (see the recent review of Tang and Ju (2018)) consist of assuming parametric assumptions on a part of the joint distribution (for example in (1.18), assuming that the data distribution is Gaussian) and of letting the other part nonparametric. Most of the works consider the regression setting, where Y is the outcome variable, which may contain MNAR values, and X is the fully observed covariates. They focus on how to prove the identifiability (in most cases, using a shadow variable), how to estimate the mean of the outcome and how to test the parametric assumptions.

For example, Miao et al. (2015); Miao and Tchetgen Tchetgen (2016) use a shadow variable Z to get the identifiability and propose doubly robust estimators. In particular, they consider the pattern-mixture model (1.19) and they introduce the odds-ratio function $\mathrm{OR}(X, Y, Z)$, which measure the deviation between the distributions of the observed data $f_{Y|M=0,X,Z}$ and of the missing data $f_{Y|M=1,X,Z}$. Their estimation is said doubly robust in the sense that they require correct specification of the odds-ratios $\mathrm{OR}(X, Y, Z)$ and either of the observed data distribution $f_{Y,Z|X,M=1}$ or of the missing-data mechanism $f_{M=1|Y,X}$, but not necessarily both.

Several authors have considered the case where the missing-data mechanism is specified (the logistic distribution is often considered) but the data distribution is nonparametric. The estimators then rely on pseudo-likelihood approaches (Zhao and Shao, 2015), empirical likelihood inference (Liu et al., 2019) or inverse propensity weighting methods (Shao and Wang, 2016; Morikawa et al., 2017). Recently, Zhao and Ma (2021) suggest a kernel estimation method where they also consider nonparametric data distribution. They do not directly

$$X = \begin{array}{c} \begin{matrix} X_{\cdot 1} & X_{\cdot 2} & X_{\cdot 3} \end{matrix} \\ \begin{pmatrix} 12 & 28 & \texttt{NA} \\ 23 & \texttt{NA} & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \texttt{NA} & 3 & 7 \end{pmatrix} \end{array}, \; X^{AC} = \begin{array}{c} \begin{matrix} X_{\cdot 1} & X_{\cdot 2} & X_{\cdot 3} \end{matrix} \\ \begin{pmatrix} 12 & 28 & \cancel{\texttt{NA}} \\ 23 & \cancel{\texttt{NA}} & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \cancel{\texttt{NA}} & 3 & 7 \end{pmatrix} \end{array}$$

Figure 1.7: Illustration of the available-case analysis: if X is the data matrix, the statistical inference will be conducted on X^{AC}, by considering only observed cells in X.

require to model the missing-data mechanism but only a working model (not necessarily containing the true mechanism).

Note that all these models exclude the case of missing covariates X, which makes them very case-specific. Besides, when some observations have outcome variables and others do not, the machine learning community usually identifies this setting as a semi-parametric one. For MNAR data, it refers to Class Distribution Mismatch introduced by Oliver et al. (2018). A parallel with this literature would be worth making.

Very few works consider missing covariates for semi-parametric models, but Miao and Tchetgen (2018) propose an inverse probability weighted estimator and address the identifiability of such models using a shadow variable.

Available-case analysis without modeling the missing-data mechanism Although semi-parametric methods may avoid the use of strong parametric assumptions on the missing data, they can still have a heavy parametric estimation. Recent works propose to estimate parameters in the available-case analysis. The latter refers to the method consisting of using all observed case in X, as illustrated in Figure 1.7. In the linear regression case, (Tang et al., 2003; Mohan et al., 2018) propose an estimation method without specifying the distribution of missing data and calculated using only the observed information. Both works consider the self-masked mechanism (1.2). This approach does not remove the rows containing missing values, as the complete-case analysis, but only the cases encoded as `NA`.

✍ In this dissertation, an EM algorithm is derived to deal with MNAR data for low-rank models in Chapter 2 and for model-based clustering in Chapter 5. In Chapter 3, we propose an estimation method for the probabilistic principal component analysis model using the available-case analysis and without modelling the missing-data mechanism.

1.5.4 Sensitivity analysis

The MNAR assumption is nearly impossible to check, because the mechanism depends on the unobserved data (d'Haultfoeuille, 2010). In particular, when the selection models (1.18) are used, the conditional distribution of the missing data-pattern given the data $f_{M|X}$ (including missing variables) is untestable. However, the results can be sensitive to deviations from

these assumptions (Kenward, 1998). What is called sensitivity analysis is assessing how sensitive the method is to a deviation to its assumptions.

For the pattern-mixture models given in (1.20), several methods have been proposed to test the assumptions. A classical strategy consists of testing if the distribution of the missing individuals $f_{X|M=1}$ differs from the one of the observed individuals $f_{X|M=0}$ (Council et al., 2010; Leurent et al., 2018; Little and Rubin, 2019, Chapter 15). Let us take the example given in Council et al. (2010) where the participants with missing data have a 0.1 lower quality of life than those observed. Thus, it can be assumed that the missing individuals are related to the observed ones by a scaled parameter c, $X^{(1)} = cX^{(0)}$. The missing variables are first imputed by using a method assuming a MAR mechanism (in this case, $c = 1$). The results are then analysed for several plausible values of c. It allows to assess how a deviation from MAR could output to different results.

To conclude this section, keep in mind that the MNAR mechanism allows to model many situations. Nevertheless, this flexibility comes at the cost of theoretical and practical challenges, in particular the identifiability of the parameters and the sensitivity of the model assumption.

1.6 Summary of the contributions

Although many methods are already available to deal with missing data, there are still great challenges, depending on the type of missing data and on the statistical task. In this dissertation, a particular attention is paid to considering realistic missing-data mechanisms such as the MNAR one. We aim at proposing innovative methods, relying on both strong theoretical and practical aspects, and meeting concrete needs in applications, especially those posed by the Traumabase dataset. Note that the corresponding code for each piece of work is available on my github account for reproducibility purposes.

In the first part of this dissertation, we consider low-rank models when MNAR values on several variables can occur. The aim is two-fold: (i) the estimation of the model parameters and (ii) the imputation of missing values. In Chapter 2, for a low-rank model with fixed effects, an (accelerated) EM algorithm is considered to maximize the joint distribution of the data and the missing-data pattern. Although this method is theoretically sound, the missing-data mechanism has to be specified and the algorithm derived can be computationally costly. To overcome this, an alternative strategy is proposed,which is free of specification for the missing-data mechanism but does not rely on theoretical guarantees. In Chapter 3, for a low-rank model with random effects, a.k.a. a probabilistic principal component analysis setting, we propose an estimation and imputation method that is free of any missing mechanism modeling and that is theoretically sound. The model parameters are proven to be identifiable and the estimators derived have the great advantage of being computed using only the observed cases (the available-case analysis).

The second part of this dissertation addresses two specific scenarios of supervised and unsupervised learning, widely encountered in the applications: the linear regression on the one hand, and the clustering on the other hand. In Chapter 4, the aim is to study online linear

Main topics	**Chapter 2**	**Chapter 3**	**Chapter 4**	**Chapter 5**	**App. A**
Mechanisms Section 1.2.2	MNAR (self-masked)	MNAR (general)	MCAR (heterogeneous)	MNAR (general)	MNAR
EM algorithm Section 1.3.2, 1.5.3	✓			✓	
Low rank models Section 1.3.3	✓	✓			
Debiasing algorithm Section 1.3.4			✓		✓
Identifiability Section 1.5.2		✓		✓	
Linear regression Section 1.4.1			✓		✓
Clustering Section 1.4.3				✓	

Table 1.4: Indications for a parsimonious reading of the introduction (Chapter 1).

regression in presence of heterogeneous MCAR values in the covariates (i.e. each variable has not the same probability of being missing). In order to estimate the model parameters, the strategy consists of naively imputing the missing values and adapting the averaged stochastic gradient algorithm to account for the imputation error. The proposed algorithm comes with strong convergence guarantees. Note that in Appendix A, we also study a standard sparse regression framework, in which the impact of missing values in the covariates is modelled as a sparse corruption problem, whatever the type of missing data encountered. To solve the latter, we derive a robust version of the Lasso-Zero strategy introduced by Descloux and Sardy (2020). Finally, Chapter 5 is dedicated to clustering individuals of a dataset containing MNAR values using a model-based approach. The identifiability question is thoroughly studied, and the estimation of the mixture model parameters (and by doing so the clustering) is performed using stochastic EM algorithms.

The third part presents our platform on missing values that bundles classical and recent references on the subject, that gives an overview on the large variety of related R packages and also gives some tutorials both on theoretical and practical questions (in R and Python).

For the sake of clarity, Table 1.4 gives indications on which parts of this introduction can be read as a prelude to the corresponding chapter.

1.6.1 A low-rank model with fixed effects for MNAR data

Using a prior of a low-rank model with fixed effect, Chapter 2 focuses on estimation and imputation with MNAR data. More precisely, the data matrix $X \in \mathbb{R}^{n\times d}$ is considered as a low-rank matrix $\Theta \in \mathbb{R}^{n\times d}$ corrupted by an additive Gaussian noise:

$$X = \Theta + \epsilon, \text{where} \begin{cases} \Theta \text{ with rank } r < \min\{n, p\}, \\ \epsilon_{i.} \overset{\perp\!\!\!\perp}{\sim} \mathcal{N}(0_n, \sigma^2 \mathrm{I}_{n\times n}), \forall i \in \{1, \dots, n\}, \end{cases}$$

and to contain MNAR values. The aim is to estimate Θ and impute missing values in X. To the best of our knowledge most of the existing methods reviewed in Section 1.3.3 do not consider the case of MNAR data.

Our contribution is two-fold. We first propose to maximize the joint distribution of the data and the missing-data pattern using an EM algorithm. The missing-data pattern is modeled with the selection models specification and a self-masked mechanism is assumed. As the E-step has no closed form, a Monte Carlo approximation is performed and coupled with the Sampling Importance Resampling (SIR) algorithm (Gordon et al., 1993). The M-step is penalized by the nuclear norm, i.e.

$$\Theta^{r+1}, \phi^{r+1} \in \underset{\Theta,\phi}{\operatorname{argmax}}\, Q(\Theta, \phi; \theta^r, \phi^r) + \lambda \|\Theta\|_\star,$$

and solved by using an accelerated proximal gradient algorithm, called Fast Iterative Soft-Thresholding Algorithm (Beck and Teboulle, 2009), which converges faster than ISTA presented in Section 1.3.3. However, the whole method can be computationally costly and relies on the specification of the missing-data mechanism. The second contribution is to suggest an efficient surrogate estimation, without specifying the missing-data mechanism, by concatenating the data matrix and the missing-data mask as $X^{\text{aug}} = [X, (1 - M)]$. A low-rank structure on this new matrix is assumed in order to take into account the relationship between the variables and the mechanism. The optimization can thus be performed as if the data were M(C)AR, because we assume that the information of the missing-data mechanism is already encoded in $(\mathbf{1}_{n\times d} - M)$. For this, we use the algorithm in (Robin et al., 2020) which deals with mixed data, as X is assumed to contain continuous variables, and as $(\mathbf{1}_{n\times d} - M)$ is a binary matrix.

Through a study on synthetic data, the model-based method proves to be extremely relevant when few variables are missing and the implicit method, which models the mask using a binomial distribution, is much less costly in terms of computation time and allows a better imputation. The performances of our methods are assessed on the Traumabase dataset, when the aim is to complete the data before using it to predict if the doctors should administrate tranexomic acid to patients with traumatic brain injury, that would limit excessive bleeding.

1.6.2 A low-rank model with random effects (PPCA) for MNAR data

In this chapter, we consider that the data matrix X is generated under a *fully-connected* PPCA model, in the sense that the loading coefficients B are of full rank. In particular,

$$X = \mathbf{1}\alpha + WB + \epsilon, \text{with} \begin{cases} W = (W_{1.}|\dots|W_{n.})^T, \text{ with } W_{i.} \overset{\perp\!\!\!\perp}{\sim} \mathcal{N}(0_r, \mathrm{Id}_{r\times r}) \in \mathbb{R}^r, \\ B \text{ of rank } r < \min\{n, d\}, \\ \alpha \in \mathbb{R}^d \text{ and } \mathbf{1} = (1\dots 1)^T \in \mathbb{R}^n, \\ \epsilon = (\epsilon_{1.}|\dots|\epsilon_{n.})^T, \text{ with } \epsilon_{i.} \overset{\perp\!\!\!\perp}{\sim} \mathcal{N}(0_d, \sigma^2 \mathrm{I}_{d\times d}) \in \mathbb{R}^d, \end{cases}$$

where σ^2 and r are known. This model implies that the rows of X are independent and Gaussian with mean α and covariance matrix $\Sigma = B^T B + \sigma^2 \mathrm{I}_{d\times d}$. From a theoretical point

of view, we first discuss and prove the identifiability of the parameters of the PPCA and of the missing-data mechanism, by assuming a self-masked MNAR mechanism.

Then, in presence of (general) MNAR values, we propose a strategy to estimate the coefficients matrix B based on estimations of the mean and the covariance matrix. We show that they can be consistently estimated in the available-case analysis when only the observed cases are used (see Section 1.5.3). In order to derive such estimators, we leverage linear connections that can be established between variables under the fully-connected PPCA assumption. Two strategies to derive mean and covariance estimators are suggested: by using algebraic arguments or graphical models. The latter is inspired by (Mohan et al., 2018), which considers linear models with a self-masked mechanism.

This method has the great advantage of being specification-free for the missing-data mechanism and of dealing with MNAR data, possibly coupled with M(C)AR data, resulting in a realistic missing scenario. To assess the proposed methodology, experiments are conducted on synthetic data and on two real datasets including the Traumabase dataset and a recommendation system dataset.

1.6.3 Debiased averaged SGD algorithm with heterogeneous MCAR data

Chapter 4 proposes a debiased averaged SGD algorithm to deal with heterogeneous MCAR data in the linear regression case. In particular, $(Y_i, X_{i.})$ are assumed to be i.i.d. observations such that

$$Y_i = X_{i.}^T \beta + \epsilon_i,$$

where $Y_i \in \mathbb{R}$, $X_{i.} \in \mathbb{R}^d$, and $\epsilon_i \in \mathbb{R}$ are respectively the outcome variable, the covariates and the noise term for the individual i, and $\beta \in \mathbb{R}^d$ is the regression parameter. The aim is to estimate the regression parameter, and then to solve the least-square optimization problem, recalled here

$$\hat{\beta} \in \operatorname{argmin}_\beta \mathbb{E}_{X,Y}[(Y_i - X_{i.}\beta)^2] := R(\beta).$$

We assume that there are incomplete variables in the covariates, and that each variable may have different missing probability but independent of the values of the data (this is the heterogeneous MCAR setting, more realistic than the MCAR one). To deal with missing values, we propose to naively impute the missing values by zero in order to get complete covariates $\tilde{X}_{i.} = X_{i.} \odot (\mathbf{1}_{n\times d} - M_{i.})$ and to account for the imputation error by debiasing the gradients of the averaged SGD algorithm. The latter has been shown to stabilise the behaviour of the algorithm and reduce the impact of noise, resulting in better convergence rates (Bach and Moulines, 2013). Instead of considering the iterates β_k, the averaged SGD uses the Polyak-Ruppert averaged iterates $\bar{\beta}_k$ (Polyak and Juditsky, 1992), which allow to account for all the iterates and not to forget previous ones,

$$\beta_k = \beta_{k-1} - \alpha \tilde{g}_k(\beta_{k-1})$$
$$\bar{\beta}_k = \frac{1}{k+1} \sum_{i=0}^{k} \beta_i.$$

with $\tilde{g}_k$ the unbiased gradients, such that $\mathbb{E}\left[\tilde{g}_k(\beta_{k-1}) \,|\, \mathcal{F}_{k-1}\right] = \nabla R(\beta_{k-1})$, $\mathcal{F}_{k-1} = (X_{1.}, Y_1, M_{1.} \ldots, X_{k-1.}, Y_{k-1}, M_{k-1.})$.

The literature considering such strategy of naively imputing the missing values and adapting existing algorithms has been reviewed in Section 1.3.4. In particular, the SGD algorithm is studied by Ma and Needell (2018) using the same strategy but this work assumes MCAR data, is restricted to the finite-sample setting and is not suitable for the high-dimensional setting. A detailed comparison is given in Chapter 4. The main contribution of our work consists of adapting a powerful supervised learning algorithm to deal with missing values, adapted both to the streaming setting, when the data arrive progressively, and to the high dimension setting, without adding strong parametric assumptions. These are the main advantages of our work compared to classical methods such as multiple imputation or the EM algorithm.

From a theoretical point of view, under weak assumptions on the observations, we derive a convergence rate of $\mathcal{O}(k^{-1})$ for our algorithm in the streaming setting. More particularly, for a constant step-size $\alpha = \frac{1}{2L}$, our algorithm ensures that, for any $k \geqslant 0$, the excess of theoretical risk is

$$\mathbb{E}\left[R\left(\bar{\beta}_k\right) - R(\beta)\right] \leqslant \frac{2}{k}\left(\sqrt{c(\beta)d} + \frac{\|\beta_0 - \beta\|}{\sqrt{\alpha}}\right)^2.$$

The expected excess risk is upper bounded by a variance term, which increases with the rate of missing values, and a bias term, which takes into account the initial distance between the starting point β^0 and the optimal point β. This convergence rate is remarkable, because it is optimal for the least-square regression and similar to the rate without missing values (Bach and Moulines, 2013).

In order to assess the convergence behavior and the relevance of our algorithm, we conduct experiments on synthetic data and on real datasets including the Traumabase dataset.

1.6.4 Model-based clustering with MNAR data

Chapter 5 addresses unsupervised learning when MNAR values occur. We consider the model-based clustering, because our aim is two-fold: (i) to cluster the individuals and (ii) to estimate the parameters of the distributions for each cluster (which also allows to impute missing values). To this end, we model the MNAR mechanism by using selection models. To our knowledge, the only work that considers MNAR data in model-based clustering (De Chaumaray and Marbac, 2020) models the mechanism with pattern-mixture models, which makes it unsuitable to estimate the density parameters or to impute missing values.

Our inference is conducted on the following full observed likelihood,

$$L(\pi, \theta, \phi; X^{(0)}, M) = \prod_{i=1}^{n}\left(\int_{\mathcal{X}_{i.}^{(1)}} f(x_{i.}; \pi, \theta) f_{M|X,Z_{.k}=1}(m_{i.} \mid x_{i.}, z_{ik} = 1; \phi) dx_{i.}^{(1)}\right),$$

where $f(x_{i.}; \pi, \theta) = \sum_{k=1}^{K} \pi_k f_k(y_i; \theta_k)$ is the mixture model, with $\pi_k = \mathbb{P}(z_{ik} = 1)$ and K the number of clusters (unknown quantity). Our first contribution is to specify a large variety of

mechanism distributions $f_{M|X,Z_{.k}=1}$, derived from the following general model,

$$f_{M|X,Z_{.k}=1}(m_{i.} \mid x_{i.}, z_{ik} = 1; \phi) = \rho(\xi^z_{kj} + \xi^x_{kj} x_{ij}), \tag{1.21}$$

where ρ is the cumulative distribution of any continuous distribution function. In this case, the parameter of the mechanism is $\phi = (\alpha, \beta)$. For all $k \in \{1, \ldots, K\}$, for all $j \in \{1, \ldots, d\}$, the parameter ξ^z_{kj} represents the average effect of the link between the presence of missing value for the variable j and the membership to the class k (i.e. this effect may not be the same for all variables). The parameters ξ^x_{kj} represent the direct effect of missingness on the variable j which depends on the class k. The model (1.21) makes it possible to consider realistic processes that cause the lack of the data, but requires to estimate $2Kd$ parameters, which can be challenging. Consequently, we propose more parsimonious versions of the MNAR model (1.21). For each of the sub-models, we discuss identifiability depending on the data type (continuous, categorical or mixed) and we propose an estimation strategy, using the EM algorithm or the SEM algorithm. The parameters for the sub-models which consider that the effect of missingness depends on x are not identifiable in both categorical and mixed cases. In addition, their estimation procedures turn out to be difficult, as they involve the use of a SEM algorithm with the introduction of a latent variable.

An interesting sub-model is the following one

$$f_{M|X,Z_{.k}=1}(m_{i.} \mid x_{i.}, z_{ik} = 1; \phi) = \rho(\xi^z_k), \tag{1.22}$$

when the only effect of missingness is on the class membership k which is the same for all the variables. The parameters are identifiable in the continuous, categorical and mixed cases.

We illustrate the methods on synthetic data. In particular, we show the flexibility of the MNAR model (1.22), for which the estimation method is based on a simple EM algorithm which does not use costly sampling methods.

1.6.5 A resource website on missing values

Chapter 6 presents R-miss-tastic, our platform which aims to provide an overview of standard missing values problems and methods, by providing relevant implementations of methodologies. In particular, several pipelines in R and Python allow for a hands-on illustration on how to handle missing values in various statistical tasks such as estimation and prediction, while ensuring reproducibility of the analyses.

Part I

Dealing with MNAR data in low-rank models

Chapter 2

Imputation and low-rank estimation with MNAR data

This chapter corresponds to the paper Imputation and low-rank estimation with Missing Not At Random data, published in Statistics and Computing, 2020, written with Claire Boyer and Julie Josse.

Abstract

Missing values challenge data analysis because many supervised and unsupervised learning methods cannot be applied directly to incomplete data. Matrix completion based on low-rank assumptions are very powerful solution for dealing with missing values. However, existing methods do not consider the case of informative missing values which are widely encountered in practice. This paper proposes matrix completion methods to recover Missing Not At Random (MNAR) data. Our first contribution is to suggest a model-based estimation strategy by modelling the missing mechanism distribution. An EM algorithm is then implemented, involving a Fast Iterative Soft-Thresholding Algorithm (FISTA). Our second contribution is to suggest a computationally efficient surrogate estimation by implicitly taking into account the joint distribution of the data and the missing mechanism: the data matrix is concatenated with the mask coding for the missing values; a low-rank structure for exponential family is assumed on this new matrix, in order to encode links between variables and missing mechanisms. The methodology that has the great advantage of handling different missing value mechanisms is robust to model specification errors.

The performances of our methods are assessed on the real data collected from a trauma registry (TraumaBase®) containing clinical information about over twenty thousand severely traumatized patients in France. The aim is then to predict if the doctors should administrate tranexomic acid to patients with traumatic brain injury, that would limit excessive bleeding.

Contents

2.1 Introduction

The problem of missing data is ubiquitous in the practice of data analysis. Main approaches for handling missing data include imputation methods and the use of Expectation-Maximization (EM) algorithm (Dempster et al., 1977) which allows to get the maximum likelihood estimators in various incomplete-data problems (Little and Rubin, 2019). The theoretical guarantees of these methods ensuring the correct prediction of missing values or the correct estimation of some parameters of interest are only valid if some assumptions are made on how the data came to be missing. Rubin (1976) introduced three types of missing-data mechanisms: (i) the restrictive assumptions of missing completely at random (MCAR) data, (ii) the missing at random (MAR) data, where the missing data may only depend on the observable variables, and (iii) the more general assumption of missing not at random (MNAR) data, *i.e.* when the unavailability of the data depends on the values of other variables and its own value. A classic example of MNAR data, which is the focus of the paper, is surveys where rich people would be less willing to disclose their income or where people would be less incline to answer sensitive questions on their addictive use. Another example would be the diagnosis of Alzheimer's disease, which can be made using a score obtained by the patient on a specific test. However, when a patient has the disease, he or she has difficulty answering questions and is more likely to abandon the test before it ends.

Missing not at random data When data are MCAR or MAR, valid inferences can be obtained by ignoring the missing-data mechanism (Little and Rubin, 2019). The MNAR data lead to selection bias, as the observed data are not representative of the population. In this setting, the missing-data mechanism must be taken into account, by considering the joint distribution of complete data matrix and the missing-data pattern. There are mainly two approaches to model the joint distribution using different factorizations:

1. selection models (Heckman, 1979), which seem preferred as it models the distribution of the data, say Y, and the incidence of missing data as a function of Y which is rather intuitive;

2. pattern-mixture models (Little, 1993), which key issue is that it requires to specify the distribution of each missing-data pattern separately.

Most of the time, in these parametric approaches, the EM algorithm is performed to estimate the parameters of interest, such as the parameters of generalized linear models (Ibrahim et al., 1999) and the missing-data mechanism distribution is usually specified by logistic regression models (Ibrahim et al., 1999; Tang and Ishwaran, 2017; Morikawa et al., 2017), in the case of selection models. In addition, the MNAR mechanism often is chosen self-masked *i.e.* the lack of a variable depends only on the variable itself and only simple models have been considered with cases where just the output variable or one or two variables are subject to missingness (Miao and Tchetgen, 2018; Ibrahim et al., 1999). Note that recent works based on graph-based approaches (Mohan and Pearl, 2021; Mohan et al., 2018) show that in some specific setting of MNAR values, it is possible to estimate parameters for simple models, such as the mean and variance in linear models, without specifying the missing value mechanism.

Low-rank models with missing values In this paper, we focus on estimation and imputation in low-rank models with MNAR data. The low-rank model has become very popular in recent years (Kishore Kumar and Schneider, 2017) and it plays a key role in many scientific and engineering tasks, including denoising (Gavish and Donoho, 2017), collaborative filtering (Yang et al., 2018), genome-wide studies (Leek and Storey, 2007; Price et al., 2006), and functional magnetic resonance imaging (Candès et al., 2013). It is also a very powerful solution for dealing with missing values (Josse et al., 2016b; Kallus et al., 2018). Indeed, the low-rank assumption can be considered as an accurate approximation for many matrices as detailed by Udell and Townsend (2017). For instance, the low-rank approximation makes sense when either, one can consider that a limited number of individual profiles exist or, dependencies between variables can be established.

Let us consider a data matrix $Y \in \mathbb{R}^{n \times p}$ which is a noisy realisation of a low-rank matrix $\Theta \in \mathbb{R}^{n \times p}$ with rank $r < \min\{n, p\}$:

$$Y = \Theta + \epsilon, \text{where} \begin{cases} \Theta \text{ has a low rank } r, \\ \epsilon \sim \mathcal{N}(0, \sigma^2 I). \end{cases} \tag{2.1}$$

In the following, σ is assumed to be known. Suppose that only partial observations are accessible. We note the mask $\Omega \in \{0,1\}^{n\times p}$ with

$$\Omega_{ij} = \begin{cases} 0 \text{ if } y_{ij} \text{ is missing,} \\ 1 \text{ otherwise.} \end{cases}$$

where y is a realisation of Y. The main objective is then to estimate the parameter matrix Θ from the incomplete data, which can be seen on the one hand as a denoising task by estimating the parameters from the observed incomplete noisy data, and on the other hand as a prediction task by imputing missing values with values given by the estimated parameter matrix. A classical approach to estimate Θ with MAR or MCAR missing values are based on convex relaxations of the rank, *i.e.* the nuclear norm and consists in solving the following penalized weighted least-squares problem:

$$\hat{\Theta} \in \text{argmin}_\Theta \|(Y-\Theta)\odot\Omega\|_F^2 + \lambda\|\Theta\|_\star, \tag{2.2}$$

where $\|.\|_F$ and $\|.\|_\star$ respectively denote the Frobenius norm and the nuclear norm and $\odot$ is the Hadamard product. The main algorithm available to solve (2.2) consists in a proximal gradient method, leading to iterative soft-thresholding algorithm (ISTA) of the singular value decomposition (SVD) (Mazumder et al., 2010; Cai et al., 2010) in the case of a regularization via the nuclear norm (note that this strategy is equivalent to perform an EM algorithm with a nuclear norm penalization in the M-step, see Appendix B.2.2). Given any initialization (for instance the missing values can be initialized to the mean of the non-missing entries), a soft-thresholding SVD is computed on the completed matrix and the predicted values of the missing entries are updated using the values given by the new estimation. The two steps of estimation and imputation are iterated until empirical stabilization of the prediction. There has been a lot of work on denoising and matrix completion with low-rank models, whether algorithmic, methodological or theoretical contributions (Candès and Recht, 2009; Candes and Plan, 2010). However, to the best of our knowledge most of the existing methods do not consider the case of MNAR data.

Contributions In order to perform low-rank estimation with MNAR data, our first contribution, detailed in Section 2.3.1, is to suggest a model-based estimation strategy by maximizing the joint distribution of the data and the missing values mechanism using an EM algorithm. More specifically, a Monte Carlo approximation is performed coupled with the Sampling Importance Resampling (SIR) algorithm. Note yet that introducing such a model for MNAR data does not prevent from handling Missing Completely At Random (MCAR) or Missing At Random (MAR) data as well. Indeed, our model can only impact variables of type MNAR, while the low-rank assumption will be enough to deal with other types of missing variables. This approach, although theoretically sound and well defined, has two drawbacks: its computational time and the need to specify an explicit model for the mechanism, so to have a strong prior knowledge about the shape of the missing-data distribution.
Our second contribution (Section 2.3.2) is to suggest an efficient surrogate estimation by

implicitly modelling the joint distribution. To do so, we suggest to concatenate the data matrix and the missing-data mask, *i.e.* the indicator matrix coding for the missing values, and to assume a low-rank structure on this new matrix in order to take into account the relationship between the variables and the mechanism. This strategy has the great advantage that it can be performed using classical methods used in the MCAR and MAR settings and that it does not require to specify a model for the mechanism. This approach can be seen as connected to the following works. Harel and Schafer (2009) present a method to handle missing data in a latent-class model where the missing covariates X are linked to the missing-data pattern M by a latent variable η. In an example, they suggests treating M as additional items alongside X, in order to make statistical inferences. Moreover, in the context of decision trees used for classification, Twala et al. (2008) suggest an approach known as missing values attribute where at each split, all the missing values can go on the right or on the left. This can be seen as cutting according to the missing value pattern so it is equivalent as implicitly adding M with the covariates X. Finally, from the optimization point of view, we also suggest (Section 2.3.3) to use an accelerated proximal gradient algorithm, also called Fast Iterative Soft-Thresholding Algorithm (FISTA) (Beck and Teboulle, 2009) which is an accelerated version of the classical iterative SVD algorithm in the case of a penalization with the nuclear norm.

The rest of the article is organized as follows. First, although the missing-data mechanism framework is widely used, there are points of ambiguity in the classical definitions, especially considering whether the statements hold for any value (from any sample) or for the realised value (from a specific sample) (Seaman et al., 2013; Murray et al., 2018). Therefore, Section 2.2 is dedicated to specify a general and clear framework of the missing-data mechanisms in order to remove ambiguities and introduce the MNAR mechanism being considered. In Section 2.3, we present both proposals to address the MNAR data issue: by explicitly modelling the missing mechanism or by implicitly taking it into account. Section 2.4 is devoted to a simulation study on synthetic data. In Section 2.5, we apply the model-based method to the TraumaBase® dataset in order to to assist doctors in making decisions about the administration of an active substance, called the tranexomic acid, to patients with traumatic brain injury. Finally, a discussion on the results and perspectives is proposed on Section 2.6.

2.2 The missing-data mechanism: notations and definitions

In the sequel, we write the complete data matrix $Y \in \mathbb{R}^{n\times p}$ of quantitative variables, whose distribution is parameterized by Θ. The missing-data pattern is denoted by $M \in \{0,1\}^{n\times p}$ and ϕ is the parameter of the conditional distribution of M given Y. We assume the distinctness of the parameters, *i.e.* the joint parameter space of (Θ, ϕ) is the product of the parameter space of Θ and the one of ϕ. We start by writing the most popular definitions of Little and Rubin (2019) for the missing-data mechanism. By writing, $Y = (Y_{\text{obs}}, Y_{\text{mis}})$, where Y_{obs} and Y_{mis} denote the observed components and the missing ones of Y respectively,

they define:

$$p(M|Y;\phi) = p(M;\phi), \quad \forall Y, \phi \tag{MCAR}$$
$$p(M|Y;\phi) = p(M|Y_{\text{obs}};\phi), \quad \forall Y_{\text{mis}}, \phi \tag{MAR}$$
$$p(M|Y;\phi) = p(M|Y_{\text{obs}}, Y_{\text{mis}};\phi), \quad \forall \phi \tag{MNAR}$$

Note that all matrices may be regarded as vectors of size $n \times p$ (see Example 3). There are mainly two ambiguities: (i) it is unclear whether the equations hold for any realisation (y, m) of (Y, M), although it is widely understood as such and (ii) Y_{obs} and Y_{mis} are actually functions of M, which is extremely confusing and explain why other attempts for definitions and notations are necessary. Seaman et al. (2013) propose two definitions of the MAR mechanism, for which they differentiate if (i) the statements hold for any values (from any sample), the everywhere case (EC) (ii) or for the realised values (from a specific sample), the realised case (RC). They also introduce a specific notation for the observed values of Y, clearly written as a function o of Y and M: $o(Y, M)$. By writing $\tilde{y}$ and $\tilde{m}$ the realised values of Y and M for a specific sample, it leads to:

$$\forall y, y^*, m \text{ such that } o(y, m) = o(y^*, m)$$
$$p(M = m|Y = y;\phi) = p(M = m|Y = y^*;\phi), \text{ (EC)}$$

$$\forall y, y^* \text{ such that } o(y, \tilde{m}) = o(y^*, \tilde{m}) = o(\tilde{y}, \tilde{m})$$
$$p(M = \tilde{m}|Y = y;\phi) = p(M = \tilde{m}|Y = y^*;\phi), \text{ (RC)}$$

We can illustrate these concepts with the following example:

Example 3. *Let* $y = \begin{pmatrix} 1 & 3 \\ 4 & 10 \end{pmatrix}$*, that can be regarded as a vector* $\text{vec}(y) = \begin{pmatrix} 1 & 3 & 4 & 10 \end{pmatrix}$*. If* $\text{vec}(y) = \begin{pmatrix} 1 & 3 & 4 & NA \end{pmatrix}$ *is observed, then* $\tilde{m} = \begin{pmatrix} 1 & 1 & 1 & 0 \end{pmatrix}$ *and* $o(\tilde{y}, \tilde{m}) = \begin{pmatrix} 1 & 3 & 4 \end{pmatrix}$*. The data are realised MAR if*

$$p(M = (1,1,1,0)|Y = y;\phi) = p(M = (1,1,1,0)|Y = y^*;\phi),$$
$$\forall y, y^*,\ o(y, \tilde{m}) = o(y^*, \tilde{m}) = (1, 3, 4)$$

$$\Updownarrow$$

$$p(M = (1,1,1,0)|Y = (1,3,4,a);\phi) = p(M = (1,1,1,0)|Y = (1,3,4,b);\phi), \forall a, b$$

By extending the framework of Seaman et al. (2013), the MNAR mechanism can be defined in the everywhere case and with the two following assumptions:

- the missing-data indicators are independent given the data,
- the MNAR mechanism is said to be self-masked, which assures that the distribution of a missing-data indicator M_{ij} given the data Y is a function of Y_{ij} only.

In the specific case of low-rank models, these both assumptions allow to have the independence by unit and to make the computations easier.

Definition 8. *The missing data are generated by the self-masked everywhere MNAR mechanism if:*

$$p(M = \Omega | Y = y; \phi) = \prod_{i=1}^{n} \prod_{j=1}^{p} p(\Omega_{ij} | y_{ij}; \phi), \quad \forall Y, \phi$$

2.3 Proposition

Our propositions for low-rank estimation with MNAR data require the following comments on the classical algorithms to solve (2.2). First, as in regression analysis there is an equivalence between minimizing least-squares and maximizing the likelihood under Gaussian noise assumption. Here as specified in Equation 2.1, the entries $(Y_{ij})_{ij}$'s are assumed to be independent and normally distributed, for all $i \in [1, n], j \in [1, p]$:

$$p(y_{ij}; \Theta_{ij}) = (2\pi\sigma^2)^{-1/2} e^{\left(-\frac{1}{2}\left(\frac{y_{ij} - \Theta_{ij}}{\sigma}\right)^2\right)}. \tag{2.3}$$

It implies that we can show (in Appendix B.2.2) that the classical proximal gradient methods to solve the penalized weighted least-squares criterion (2.2), such as iterative thresholding SVD, can be seen as a genuine EM algorithm, maximizing the observed penalized likelihood. Secondly, as detailed in Section 2.3.3, Equation 2.2 can be solved using a fast iterative soft-thresholding algorithm (FISTA) (Beck and Teboulle, 2009).

2.3.1 Modelling the mechanism

Considering the framework of selection models (Heckman, 1979), the first proposition consists in handling MNAR values in the low-rank model (2.1), by specifying a distribution for the missing-data pattern M. Here, the missing data models M_{ij} given the data Y_{ij} are assumed to be independent and distributed by a logistic model, $\forall i \in [1, n], \forall j \in [1, p]$:

$$p(\Omega_{ij} | y_{ij}; \phi) = [(1 + e^{-\phi_{1j}(y_{ij} - \phi_{2j})})^{-1}]^{(1-\Omega_{ij})} [1 - (1 + e^{-\phi_{1j}(y_{ij} - \phi_{2j})})^{-1}]^{\Omega_{ij}}, \tag{2.4}$$

where $\phi_j = (\phi_{1j}, \phi_{2j})$ denotes the parameter vector for conditional distribution of M_{ij} given Y_{ij} for all i.

Then, the joint distribution of the data and mechanism can be specified. Due to independence (see Definition (8)):

$$\begin{aligned} p(y, \Omega; \Theta, \phi) &= p(y; \Theta) p(\Omega | y; \phi) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{p} p(y_{ij}; \Theta_{ij}) p(\Omega_{ij} | y_{ij}; \phi_j). \end{aligned}$$

This leads to the joint negative log-likelihood:

$$\ell(\Theta, \phi; y, \Omega) = -\sum_{i=1}^{n}\sum_{j=1}^{p} \ell((\Theta_{ij}, \phi_j); y_{ij}, \Omega_{ij}),$$

with $\ell((\Theta_{ij}, \phi); y_{ij}, \Omega_{ij}) = \log(p((y_{ij}, \Omega_{ij}); \Theta_{ij}, \phi_j))$, $\forall i, j$. In practice, the parameters vector ϕ is unknown but viewed as a nuisance parameter, since our main interest is the estimation of Θ. To find an estimator $\hat{\Theta}$, we aim at maximizing the following penalized joint negative log-likelihood:

$$(\hat{\Theta}, \hat{\phi}) \in \text{argmin}_{\Theta,\phi} \ell(\Theta, \phi; y, \Omega) + \lambda\|\Theta\|_\star. \tag{2.5}$$

It can be achieved using a Monte-Carlo Expectation Maximization (MCEM) algorithm, whose two steps, iteratively proceeded, are given below:

- **E-step:** the expectation (taking the distribution of the missing data given the observed data and the missing-data pattern) of the complete data likelihood is computed:

$$Q(\Theta, \phi|\hat{\Theta}^{(t)}, \hat{\phi}^{(t)}) = \mathbb{E}_{Y_{\text{mis}}}\left[\ell(\Theta, \phi; y, \Omega)|Y_{\text{obs}}, M; \Theta = \hat{\Theta}^{(t)}, \phi = \hat{\phi}^{(t)}\right] \tag{2.6}$$

- **M-step:** the parameters $\hat{\Theta}^{(t+1)}$ and $\hat{\phi}^{(t+1)}$ are determined as follows:

$$\hat{\Theta}^{(t+1)}, \hat{\phi}^{(t+1)} \in \text{argmin}_{\Theta,\phi}\, Q(\Theta, \phi|\hat{\Theta}^{(t)}, \hat{\phi}^{(t)}) + \lambda\|\Theta\|_\star. \tag{2.7}$$

The E-step may be rewritten as follows:

$$Q(\Theta, \phi|\hat{\Theta}^{(t)}, \hat{\phi}^{(t)}) = -\sum_{i=1}^{n}\sum_{j=1}^{p} C_1^{\Omega_{ij}} + C_2^{1-\Omega_{ij}}$$

where

$$\begin{aligned} C_1 &= \log(p(y_{ij}, \Omega_{ij}; \Theta_{ij}, \phi_j)) \\ C_2 &= \int \log(p(y_{ij}, \Omega_{ij}; \Theta_{ij}, \phi_j)) p(y_{ij}|\Omega_{ij}; \hat{\Theta}_{ij}^{(t)}, \hat{\phi}_j^{(t)}) \mathrm{d}y_{ij} \end{aligned}$$

Note that the E-step is written as a sum of the E-steps for each (i, j)-th elements. If the (i, j)-th element is observed, we do not integrate and it leads to the first term; the second term corresponds to the missing elements. By the lack of a closed form for Q, it is approximated by using a Monte Carlo approximation, denoted as $\hat{Q}$, $\forall i \in [1, n]$, $\forall j \in [1, p]$:

$$\hat{Q}_{ij}(\Theta, \phi|\hat{\Theta}^{(t)}, \hat{\phi_j}^{(t)}) = -\frac{1}{N_s}\sum_{k=1}^{N_s} \log(p(v_{ij}^k; \Theta_{ij})) + \log(p(\Omega_{ij}|v_{ij}^k; \phi_j)),$$

where $v_{ij}^k = \begin{cases} y_{ij} \text{ if } \Omega_{ij} = 1, \\ z_{ij}^k \text{ otherwise,} \end{cases}$ with z_{ij}^k the realisation of $Z \sim p(y_{ij}|\Omega_{ij}; \hat{\Theta}_{ij}^{(t)}, \hat{\phi_j}^{(t)})$.

Note that $\hat{Q}$ is separable in the variables Θ and ϕ, so that the maximization for the M-step may be independently performed for Θ and ϕ:

$$\hat{\Theta}^{(t+1)} \in \underset{\Theta}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{1}{N_s}\sum_{k=1}^{N_s} -\log(p(v_{ij}^k;\Theta_{ij})) + \lambda\|\Theta\|_\star \tag{2.8}$$

$$\hat{\phi}^{(t+1)} \in \underset{\phi}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{1}{N_s}\sum_{k=1}^{N_s} -\log(p(\Omega_{ij}|v_{ij}^k;\phi_j)). \tag{2.9}$$

Classical algorithms can be used: (accelerated) proximal gradient method to solve (2.8) and the Newton-Raphson algorithm to solve (2.9).

Moreover, for all $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,p\}$ such that y_{ij} is missing, we suggest the use of the sampling importance resampling (SIR) algorithm (Gordon et al., 1993) to simulate the variable z_{ij}^k. The detail is given in Appendix B.3.1 and we take as a proposal distribution a Gaussian distribution.

2.3.2 Adding the mask

We now propose to directly include the information of the mask while considering the criterion (2.2), without explicitly modelling the mechanism, so that the new optimisation problem is written as follows:

$$\hat{\Theta} \in \operatorname{argmin}_\Theta \frac{1}{2}\|[\Omega\odot Y|\Omega] - [\Omega|\mathbf{1}]\odot\Theta\|_F^2 + \lambda\|\Theta\|_\star, \tag{2.10}$$

where $\mathbf{1} \in \mathbb{R}^{n\times p}$ denotes the matrix such that all its elements are equal to 1, and $[X_1|X_2]$ denotes the column-concatenation of matrices X_1 and X_2. To solve (2.10), we could use again classical algorithms such as the (accelerated) iterative (SVD) soft-thresholding algorithm (Section 2.3.3). However, this approach does not take into account that the mask is made of binary variables and suggests that the concatenated matrix $[Y\odot\Omega,\Omega]$ is Gaussian. Consequently, a better approach is to take into account the mask binary type by using the low-rank model but extended to the exponential family. There is a vast literature on how to deal with mixed matrices (containing categorical, real and discrete variables) in the low-rank model (Udell et al., 2016; Liu et al., 2018; Cai and Zhou, 2013). Robin et al. (2020) suggested such a method, by using a data-fitting term based on heterogeneous exponential family quasi-likelihood with a nuclear norm penalization:

$$\hat{\Theta} \in \operatorname{argmin}_\Theta \sum_{i=1}^{n}\sum_{j=1}^{p}\Omega_{ij}\left(Y_{ij}\Theta_{ij} + g_j(\Theta_{ij})\right) + \lambda\|\Theta\|_\star, \tag{2.11}$$

where g_j is a link function chosen according to the type of the variable j. In our case, it allows to model the joint distribution of the concatenated matrix $[Y\odot\Omega,\Omega]$ of size $n\times 2p$ as follows : (i) the data are assumed to be Gaussian, i.e. for all $j\in[1,p]$, $g_j(x) = \frac{x^2\sigma^2}{2}$ (ii) the missing-data pattern can be modelled by the Bernoulli distribution with success probability $1/(1+\exp(-\Theta_{ij}))$, i.e. for all $j\in[p+1,2p]$, $g_j(x) = \log(1+\exp(x))$. To solve (2.11), a

Penalized Iteratively Reweighted Least Squares algorithm called `mimi` (see (Robin et al., 2020, page 12)) is used. The advantage of such a strategy is to better incorporate the mask as binary features but this comes at a price of a more involved algorithm in comparison to (2.10).

2.3.3 FISTA algorithm

To solve (2.2), (2.8) and (2.10) we suggest to use the FISTA algorithm, introduced by Beck and Teboulle (2009), detailed in Appendix B.1, which corresponds to an accelerated version of the proximal gradient method. The acceleration is performed via momentum. The key advantage is that it converges to a minimizer at the rate of $\mathcal{O}(1/K^2)$ (K is the number of iterations) in the case of L-smooth functions.

This algorithm is of interest compared to the the non-accelerated proximal gradient method, that is shown in Appendix B.2.1 to be implemented in `softImpute-SVD` in the R package `softImpute` (see Hastie and Mazumder (2015)): it is known to converge only to the rate $O(1/K)$ (Beck and Teboulle, 2009, Theorem 3.1). To be more precise, another algorithm has been suggested that uses alternating least-squares (Hastie et al., 2015) and departs from the previous one by solving a non-convex problem: it relies on the maximum margin matrix factorization approach (combined with a final SVD thresholding). Therefore, although appealing numerically, the algorithm known as `softImpute-ALS` is proven to converge only to a stationary point.

2.4 Simulations

The parameter Θ is generated as a low-rank matrix of size $n \times p$ with a fixed rank $r < \min(n, p)$. The results are presented for N simulations, for each of them: (i) a noisy version Y of Θ is considered,

$$Y = \Theta + \epsilon,$$

where ϵ is a Gaussian noise matrix with i.i.d. centered entries of variance σ^2, (ii) MNAR missing values are introduced using a logistic regression, resulting in a mask Ω and (iii) only knowing $Y \odot \Omega$, we apply different methods to denoise and impute Y:

(a) Explicit method (Model): in order to take into account the missing mechanism modelling, we apply the MCEM algorithm to solve (2.5), as detailed in Section 2.3.1; note that either FISTA or `softImpute` are performed in the M-step.

(b) Implicit method (Mask): the missing mechanism is implicitly integrated by concatenating the mask to the data, as detailed in Section 2.3.2. When the binary type of the mask is neglected, FISTA or `softImpute` are used to solve (2.10). When taking into account the binary type of the mask, solving (2.11) is done by `mimi`.

(c) MAR methods: they consist in classical methods for low-rank matrix completion, proved to be efficient under the MCAR or MAR assumption, and that aim at minimizing

(2.2). The missing values mechanism is then ignored. They encompass FISTA and `softImpute`.

We also include in (b) and (c) the regularised iterative PCA algorithm (Verbanck et al., 2015; Josse et al., 2016b) which uses another penalty than the nuclear norm one. We also compare all the methods to the naive imputation by the mean (the estimation of Θ is obtained by replacing all values by the mean of the column). We performed an extended simulation study and other more heuristic methods have been tested, such as the FAMD and MFA algorithms dedicated to mixed data or blocks of variables (Audigier et al., 2016a) but they are not included in the article to make the plots more readable as the results were never convincing. The results presented are representative of all the results obtained.

The results are presented for different matrix dimensions and ranks, mechanisms of missing values (MAR and MNAR), and percentages of missing data. The code to reproduce all the simulations is available on github https://github.com/AudeSportisse/stat.

Measuring the performance To measure the methods performance, two types of normalized mean square errors (MSE) are considered:

$$\mathbb{E}\left[\left\|(\hat{\Theta} - Y) \odot (1 - \Omega)\right\|_F^2\right] \Big/ \mathbb{E}\left[\|Y \odot (1 - \Omega)\|_F^2\right] \tag{2.12}$$

$$\mathbb{E}\left[\left\|\hat{\Theta} - \Theta\right\|_F^2\right] \Big/ \mathbb{E}\left[\|\Theta\|_F^2\right], \tag{2.13}$$

that are respectively the prediction error, corresponding to the error committed when we impute values, and the total error, encompassing the prediction and the estimation error.

Some practical details on the algorithms are provided in the following paragraphs.

EM algorithm The stopping criterion used in the EM algorithm is the following:

$$\frac{\|\hat{\Theta}^{(t)} - \hat{\Theta}^{(t-1)}\|_F}{\|\hat{\Theta}^{(t-1)}\|_F + \delta} \leqslant \tau,$$

where $\delta = 10^{-3}$ and $\tau = 10^{-2}$[1]. In addition, the E-step is performed with $N_s = 1000$ Monte Carlo iterations. The key issue of this method is the run-time complexity largely due to this Monte Carlo approximation.

Tuning the algorithms hyperparameters When considering (2.2), (2.10) and (2.7), the regularisation parameter λ is chosen among some fixed grid $\mathcal{G} = \{\lambda_1, \ldots, \lambda_M\}$ to minimize either the prediction or the total errors. In the regularised iterative PCA algorithm, the hyper-parameter is the number of components to perform PCA, which can be found using

[1]Once the stopping criterion is met, $T = 10$ extra iterations are performed to assure the convergence stability.

cross-validation criteria. In the simulations, the noise level is assumed to be known. To overcome this hypothesis, one can use standard estimators of the noise level such as the ones of Gavish and Donoho (2017) and Josse et al. (2016b).

2.4.1 Univariate missing data

Let us consider a simple case with $n = 100$ and $p = 4$, the rank of the parameter matrix is $r = 1$ and $\sigma^2 = 0.8$. Assume that only one variable has missing entries. The missing values are introduced by using the self-masked MNAR mechanism. The missingness probabilities are then given as follows:

$$\forall i \in [1:n], p(\Omega_{i1} = 0 | y_{i1}; \phi) = \frac{1}{1 + e^{-\phi_1(y_{i1} - \phi_2)}} \tag{2.14}$$

The parameters of the logistic regression are chosen to mimic a cutoff effect, see Figure 2.1. Indeed, extrapolating imputed values can be challenging and classical methods are expected to introduce a large prediction bias. Given the previous parameters choice, the percentage of missing values is 50% in expectation for the missing variable, corresponding to 12.5% missing values in the whole matrix. In Figure 2.2, the three methods (a), (b) and (c) are compared in such a setting, using boxplots on MSE errors for $N = 50$ simulations. In this MNAR setting, the proposed model-based method (a), in red in Figure 2.2, aiming at minimizing (2.5) -specially designed for such a setting- gives better results globally for the total error with a significant improvement on the prediction of missing values (either when FISTA or `softImpute` is used in the M-step of the MCEM algorithm).

In addition, the implicit methods (b), in green in Figure 2.2, working on the concatenation of the mask and the data, either based on a binomial modeling of the mechanism (`mimi`, solving (2.11)), or neglecting the binary feature of the mask (FISTA and `softImpute`, solving (2.10)), do not lead to improved performance compared to the MAR method (c) (FISTA and `softImpute`) in terms of prediction or estimation errors. On the contrary, the implicit method (b) working on the concatenation of the mask and the data, based now on the regularized iterative PCA improves both estimation and prediction errors compared to the regular PCA algorithm used in the MAR method (c). However the obtained prediction error does not compete with performance of regular MAR completion algorithms (FISTA and `softImpute`).

Note also that the results of both SVD algorithms, `softImpute` and FISTA, are similar in terms of estimation and prediction error, but FISTA has the advantage to improve the numerical convergence to a minimizer.

In conclusion on the univariate case, (i) modelling the missing mechanism outperforms any other method, particularly in terms of prediction error; (ii) implicit methods (b) have limited interest, except to improve the regular PCA algorithm.

2.4.2 Bivariate missing data

We consider now a higher dimensional case: $n = 100$ and $p = 50$ and the rank of the parameter matrix is $r = 4$. The noise level is $\sigma^2 = 0.8$, as in Section (2.4.1). The missing

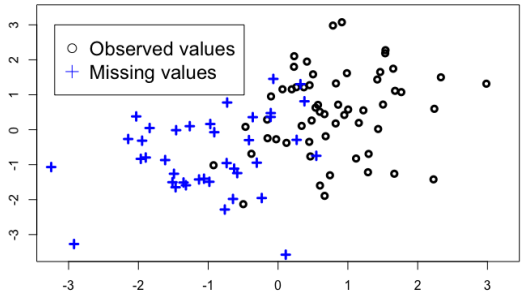


Figure 2.1: Introduction of MNAR missing values using a logistic regression (2.14), with $\phi_1 = 3$ and $\phi_2 = 0$. One can see that the the highest values of y_{i1} are missing, mimicking a cutoff effect.

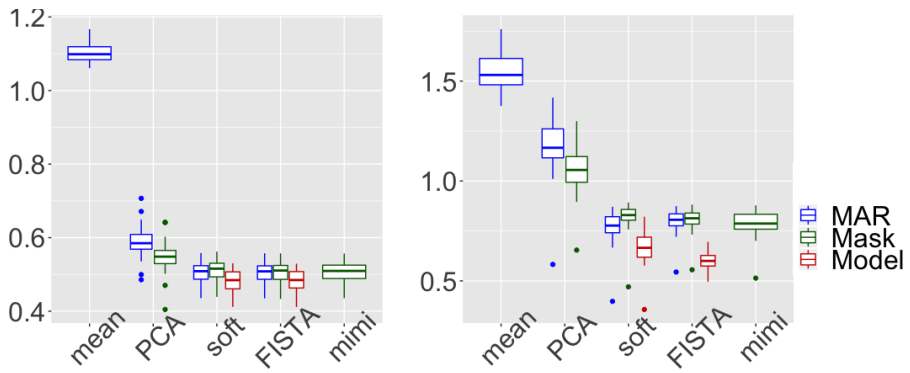


Figure 2.2: Univariate missing data: total error (left) and prediction error (right) for the methods (a) in red, (b) in green and (c) in blue.

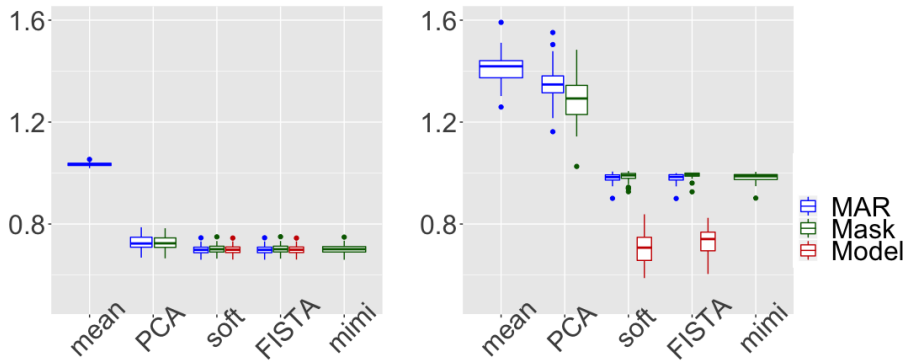


Figure 2.3: Bivariate missing data: total error (left) and prediction error (right) for the methods (a) in red, (b) in green and (c) in blue.

values are introduced on two variables by using the following MNAR mechanism, for all $i \in [1, n]$ and $j \in [1, 2]$,

$$p(\Omega_{ij} = 0 | y_{ij}; \phi) = \frac{1}{1 + e^{-\phi_{1j}(y_{ij} - \phi_{2j})}}$$

$$\text{where} \begin{cases} \phi_{1j} = 3, \phi_{2j} = 0 \text{ if } j = 1, \\ \phi_{1j} = 2, \phi_{2j} = 1 \text{ if } j = 2. \end{cases}$$

This parameters choice leads to 50% missing values in $Y_{\cdot 1}$ and 20% in $Y_{\cdot 2}$ mimicking a cutoff effect again. In Figure 2.3, the methods (a), (b) and (c) are compared in such a setting, using boxplots on MSE errors for $N = 50$ simulations.

The model-based method (a), designed for the MNAR setting, give significant better results than any other method in terms of prediction error. The mask-adding methods (b) lead to no significant improvement compared to classical MAR methods, either by solving (2.10) using FISTA, `softImpute`, or solving (2.11) via `mimi`. One can note that the PCA algorithm still benefits from the concatenation with the mask in terms of prediction error, but to a lesser extent than in the univariate case.

Overall, the poor performance of the mask-adding methods (b) can be explained by the dimensionality issue and the small weight of the added mask variables. Indeed, in this higher dimensional case with bivariate missing variables, only two informative binary variables corresponding to the mask are really concatenated to a 50-column matrix.

Note that in terms of total error, the advantages of model-based methods (a) are no longer visible, which can be explained by the very low percentage of missing data (1.5%) (see Section 2.4.3 in which more missing values are considered).

2.4.3 Multivariate missing data

We consider now a multivariate missing data case for the following dimensional setting: $n = 100$, $p = 20$ and $r = 4$. The missing values are introduced on ten variables by using the following MNAR mechanism, for all $i, j \in [1, n] \times [1, 10]$,

$$p(\Omega_{ij} = 0|y_{ij}; \phi) = \frac{1}{1 + e^{-\phi_1(y_{ij}-\phi_2)}}.$$

Note that the parameters of the missingness mechanism are the same for each element, this can be easily extended to a more general case. The parameters choice leads to 25% missing values in the whole matrix. The results are presented in Figure 2.4 for $N = 50$ simulations and different noise levels, $\sigma^2 = 0.2, 0.5$ or 0.8.

First, one can note that the model-based method (a) provides the best result both in estimation and prediction error regardless the noise level (and whatever FISTA or `softImpute` used in the MCEM). Of course, this performance improvement comes at the price of a computational cost due to the Monte Carlo approximations needed in the MCEM algorithm.

Regarding the implicit methods (b), the mask-adding techniques handling the concatenation of the data and the mask matrix as Gaussian (FISTA and `softImpute`) miss to improve both estimation and prediction errors compared to their MAR version. However, the variant `mimi` modelling the mask with a binomial distribution always largely outperforms MAR methods (c) in terms of prediction (while the improvement in terms of estimation error is only visible at a low noise level). Therefore, the mask-adding approach can implicitly capture the MNAR missing mechanism, when the mask is really considered as a matrix of binary variables. This comes at the price of a more involved algorithm `mimi` able to take into account mixed variables, but that remains far less computationally expensive than the model-based approach. Indeed, for an estimation/prediction of one parameter matrix Θ, the process time for a computer with a processor Intel Core i5 of 2,3 GHz is 0.0549 seconds for the MAR method with `softImpute`, 3.215 seconds for the implicit method with `mimi` and 13.069 minutes for the model-based method with `softImpute` when 50% of the variables are missing.

As a side comment, in this high-dimensional setting, one can note that the PCA algorithm still benefits from adding the mask, which is a variant of method (b), compared to the regular PCA method, both in estimation and prediction error. However the mask-adding PCA algorithm only compete the mask-adding methods based on iterative SVD thresholding (FISTA, `softImpute`) at a low noise level.

2.4.4 Sensitivity to model misspecifications

Deviation in the missing-data mechanism setting Here, the missing values are introduced by using the MAR mechanism. It allows to test the stability of model-based methods, designed for the MNAR setting, to a deviation in the missing mechanism. The

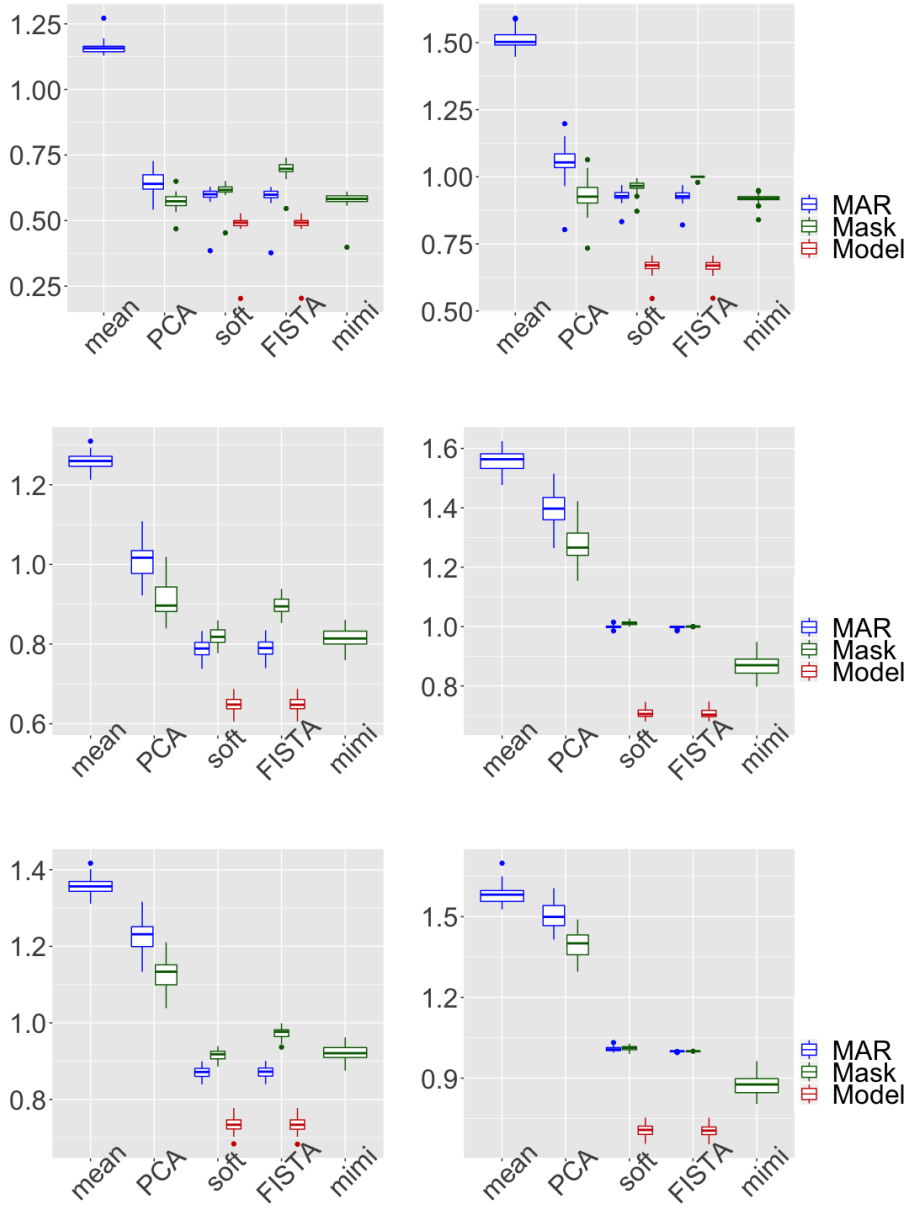


Figure 2.4: Multivariate MNAR missing data: total error (left) and prediction error (right) for the methods (a) in red, (b) in green and (c). Three noise settings are considered: on top strong signal ($\sigma^2 = 0.2$), middle noisy data ($\sigma^2 = 0.5$), bottom very noisy data ($\sigma^2 = 0.8$).

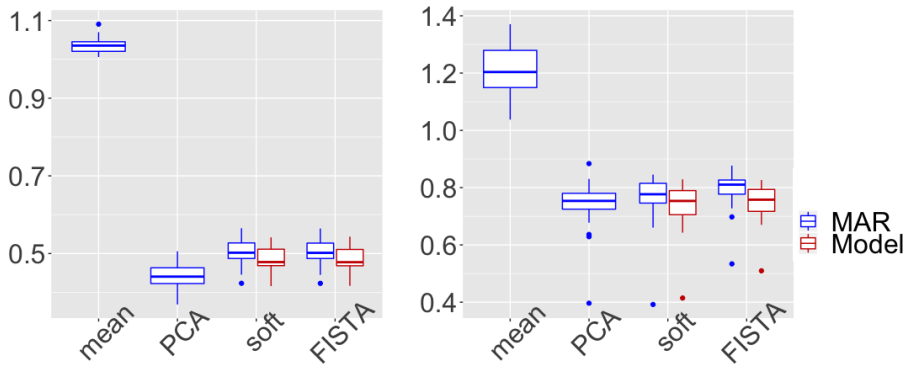


Figure 2.5: Comparison of methods performance when the missing data are of type MAR (for $N = 50$ simulations) with a rank one: total error (left) and prediction error (right) for different methods and algorithms.

missingness probabilities are given as follows in such a setting:

$$\forall i \in [1, n], p(\Omega_{i1} = 0 | y_{i2}; \phi) = \frac{1}{1 + e^{-\phi_1(y_{i2} - \phi_2)}}, \tag{2.15}$$

meaning that the probability to have a missing value in Y_1 depends on the value of Y_2.

First, let us consider the setting of Section 2.4.1, i.e. $n = 100$, $p = 4$, $r = 1$.

In Figure 2.5, we observe that the model-based method (a) improves both the estimation and the prediction, which is not expected in a MAR setting. However, this can be explained because of the rank is one which implies that there are only small differences between MNAR and MAR (the second variable's value is directly linked to the missing one's value). Consequently, modelling a MNAR mechanism is enough to retrieve information on such a MAR missing mechanism.

To avoid this case, we consider the setting of Section 2.4.2, i.e. $n = 100$, $p = 50$, $r = 4$, with a MAR missing mechanism as described by (2.15), however, the second variable involved is chosen to be decorrelated from the missing one (which is possible given the rank is 4). In such a case, there is no equivalence between the missing values that are simulated to be MAR and the mechanism we model as MNAR. Figure 2.6 shows that the model-based approach does not lead to any improvement compared to regular methods used for MAR methods; but more importantly, it does not degrade the results either which highlights the robustness of the approach with respect to deviations from the model.

Deviation in the logistic regression setting We now want to test the robustness of our model-based method (a) to a misspecification of the logistic model, given by (2.4). To do so, missing values are introduced by a MNAR missing-data mechanism based on the

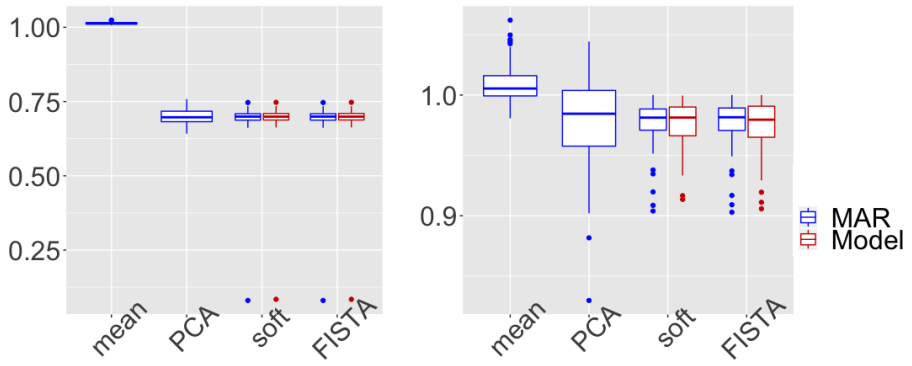


Figure 2.6: Comparison of methods performance when the missing data are of type MAR (for $N = 50$ simulations) with a rank four (the MAR mechanism depends on a decorrelated variable to the missing one): total error (left) and prediction error (right) for different methods and algorithms.

following probit model, the missingness probabilities are then:

$$\forall i \in [1, n], \qquad p(\Omega_{i1} = 0 | y_{i1}; \phi) = F(y_{i1}),$$

where F is the quantile function the standard Gaussian cumulative distribution function. Consider the setting of Section 2.4.1, i.e. $n = 100$, $p = 4$, $r = 1$. In Figure 2.7, we observe that the model-based methods (a) globally improves the results for both errors (2.13) and (2.12). Very similar results to the ones of Section 2.4.1 are obtained, meaning that the model-based method (a) behaves well to a deviation of the logistic regression modelling.

2.5 Application to clinical data

2.5.1 Motivation

Our work is motivated by a public health application with APHP TraumaBase®[2] Group (Assistance Publique - Hopitaux de Paris) on the management of traumatized patients. Major trauma, i.e. injuries that endanger a person's life or functional integrity, have been qualified as a worldwide public health challenge and a major source of mortality (first cause in the age group 16-45) in the world by the WHO (Hay et al., 2017). Hemorrhagic shock and traumatic brain injury have been identified as the lead causes of death. Effective and timely management of trauma is crucial to improve outcomes, as delays or errors entail high risks for the patient.

[2]http://www.traumabase.eu/

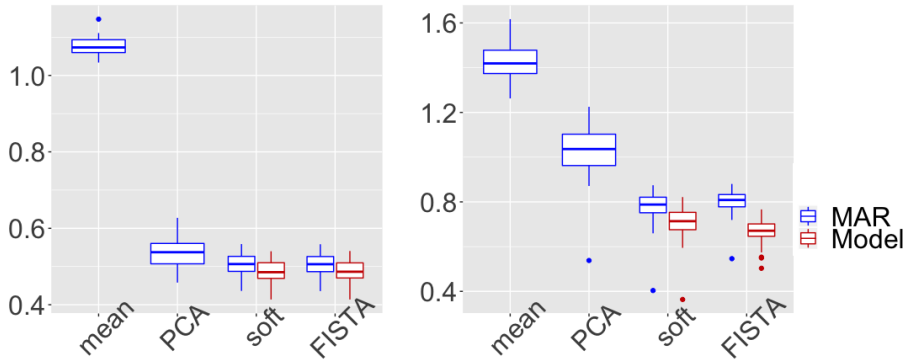


Figure 2.7: Univariate MNAR missing data parametrized with a probit model for $N = 50$ simulations: total error (left) and prediction error (right) for different methods and algorithms. Note that the methods modeling the missing mechanisms use the logistic model.

2.5.2 Data description

A subset of the trauma registry containing the clinical measurements of 3168 patients with brain trauma injury is first selected.

Our aim is to predict from pre-hospital measurements whether or not the tranexomic acid[3] should be administrated on arrival at the hospital. In the dataset, the variable *Tranexomic.acid* is the decision made by the doctors, which is considered as ground truth. This variable is equal to 1 if the doctors have decided to administrate tranexomic acid, 0 otherwise.

Nine quantitative variables containing missing values are selected by doctors. In Figure 2.8, one can see the percentage of missing values in each variable, varying from 1.5 to 30%, leading to 11% is the whole dataset. After discussion with doctors, almost all variables can be considered to have informative missingness. For example, when the patient's condition is too critical and therefore his heart rate (variable *HR.ph*) is either high or low, the heart rate may not be measured, as doctors prefer to provide emergency care. The heart rate itself can then be qualified of self-masked MNAR, and the other variables, either of MNAR or MAR. Both percentage and nature of missing data demonstrate the importance of taking appropriate account of missing data. More information on the data can be found in Appendix B.4.

In the following, two questions are addressed. Firstly, we compare the validity of the imputation methods in terms of prediction of the tranexomic acid administration based on the different imputed data. Secondly, we test the methods in terms of their imputation performance.

[3]the tranexomic acid is an antifibrinolyic agent which reduces blood loss.

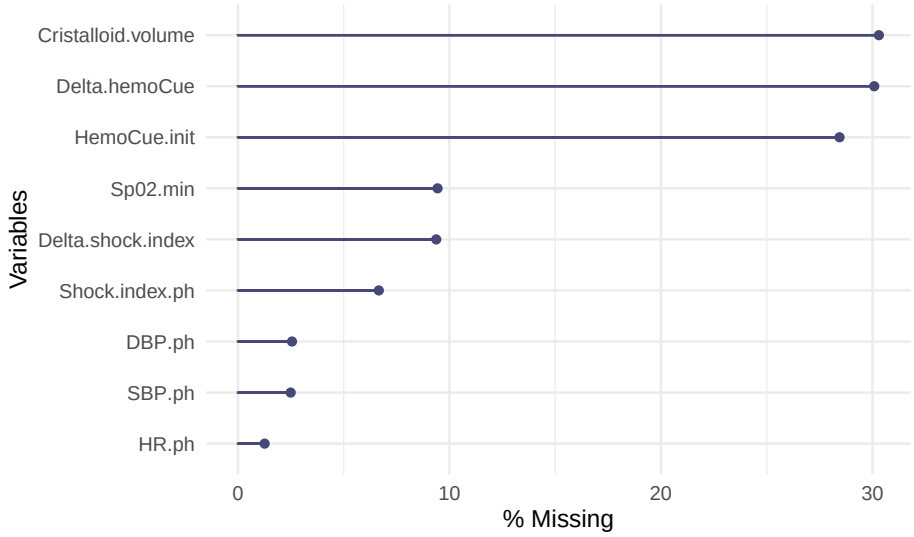


Figure 2.8: Percentage of missing values in each variable.

2.5.3 Prediction of tranexomic acid administration

We consider a two-step procedure:

- Step 1: imputation of the explanatory variables. As a preprocessing step, we impute missing data in the explanatory variables, beforehand proceeding to the classification training. Imputation is performed using the model-based method (a), the implicit methods (b) or the MAR methods (c). All these methods are compared to the naive imputation by the mean.

- Step 2: classification task which consists in predicting the administration or not of the tranexomic acid. Therefore, we are looking for the prediction function f such that

$$Z \simeq f(Y^{\text{imp}}),$$

where $Z \in \{0,1\}^n$ is equal to 1 (resp. 0) if the tranexomic acid is (resp. not) administered, and $Y^{\text{imp}} \in \mathbb{R}^{n\times p}$ represents the nine imputed explanatory variables discussed above. Based on these new-filled design matrices formed in Step 1, the classification is always done using either random forests or logistic regression.

Since not administering tranexomic acid by mistake can be vital, for the training and testing errors, we use a dissymetrized loss function where the cost of false negatives is much more than of false positives as follows

$$l(\hat{z}, z) = \frac{1}{n}\sum_{i=1}^{n} w_0 1_{\{z_i=1,\hat{z}_i=0\}} + w_1 1_{\{z_i=0,\hat{z}_i=1\}}, \tag{2.16}$$

	Model	Mask		MAR		
	soft	mimi	soft	soft	PCA	mean
error	**12.5**	16.0	15.8	14.8	13.6	**13.0**
sd	3.3	2.8	4.9	5.0	3.2	2.1
AUC	**85.4**	83.9	84.6	84.6	**85.5**	**85.2**
sd	1.6	1.7	1.8	2.0	1.4	2.2
acc	**79.5**	77.8	77.6	78.6	**79.9**	**80.7**
sd	5.0	3.2	5.0	5.2	3.4	3.1
pre	**47.5**	45.0	45.1	46.5	45.2	**48.7**
sd	6.7	4.2	8.2	8.3	5.9	5.0
sen	76.5	**78.1**	**78.2**	**77.4**	72.4	76.0
sd	6.1	3.4	5.7	5.4	3.2	4.5
spe	80.2	77.7	77.4	78.9	**80.8**	**81.7**
sd	7.2	4.4	7.2	7.3	4.6	4.6

Table 2.1: **By using random forest for the classification.** Comparison of the mean of different prediction criteria over ten simulations (values are multiplied by 100). Error corresponds to the validation error with the loss described in (2.16). AUC is the area under ROC; the accuracy (acc) is the number of true positive plus true negative divided by the total number of observations; the sensitivity (sen) is defined as the true positive rate; specificity (spe) as the true negative rate; the precision (pre) is the number of true positive over all positive predictions. The lines sd correspond to standard deviations. The three best results are in bold.

where w_0 and w_1 are the weights for the cost of false negative and false positive respectively, s.t. $w_0 + w_1 = 1$ and $\omega_0 = 5\omega_1$.

The dataset is divided into training and test sets (random selection of $80 - 20\%$) and the prediction quality on the test set is compared according to different indicators such as the accuracy, the sensitivity, etc.

Table 2.1 compares results when random forests are used as a prediction method. In this setting, mean imputation gives among the best results on all the metrics which is in agreement with recent results on its consistency when used with a powerful learner, see Josse et al. (2019). Nevertheless, the model-based method (a) is very competitive. The proposed implicit methods result in the best performances in terms of the sensitivity which is particularly relevant for the application.

Table 2.2 compares results when the prediction is performed with logistic regression. For almost all criteria, and especially on sensitivity the model-based method (a) leads to the best performances. The standard deviations are also smaller with the model based approach in comparison with the implicit methods.

Therefore, the model-based method performs well regardless of the prediction method used.

	Model	Mask		MAR		
	soft	mimi	soft	soft	PCA	mean
error	**13.5**	**13.3**	15.5	15.5	13.8	13.7
sd	2.4	4.5	3.9	3.9	3.3	2.1
AUC	**82.6**	78.7	81.9	81.9	82.1	82.0
sd	2.4	2.3	2.4	2.4	2.5	2.4
acc	**80.1**	79.3	77.6	77.6	79.6	79.8
sd	3.7	6.9	6.1	6.1	5.1	3.3
pre	**47.7**	46.2	47.0	46.0	45.1	46.9
sd	4.1	7.9	6.4	5	5.2	3.2
sen	**74.8**	67.0	73.7	73.8	73.7	73.9
sd	5.1	4.4	7.6	7.7	6.5	5.5
spe	81.3	**82.0**	78.4	81.1	81.0	78.4
sd	3.7	3.6	6.1	6.2	5.1	3.3

Table 2.2: **By using logistic regression for the classification.** Comparison of the mean of different prediction criteria over ten simulations (values are multiplied by 100). Error corresponds to the validation error with the loss described in (2.16). AUC is the area under ROC; the accuracy (acc) is the number of true positive plus true negative divided by the total number of observations; the sensitivity (sen) is defined as the true positive rate; specificity (spe) as the true negative rate; the precision (pre) is the number of true positive over all positive predictions. The lines sd correspond to standard deviations. The two best results are in bold.

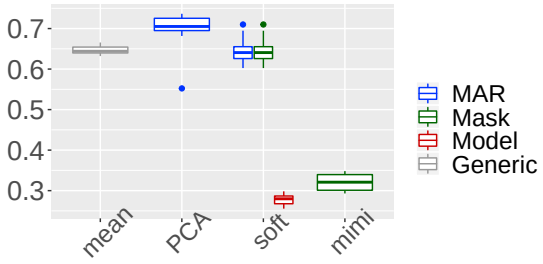


Figure 2.9: Comparison of the imputation error (for ten simulations).

2.5.4 Imputation performances

As the methods are initially designed for imputation, we perform simulations on the real dataset. In order to be able to measure the quality of the imputation, some additional MNAR values are introduced in the variable *Shock.index.ph*, which is a variable with MNAR missing values (according to doctors) that contains initially 7% of missing values. The missing values are introduced by using the self-masked mechanism described in (2.14). The choice of parameters in the logistic regression leads to 35% missing values. In the model-based method (a), the variables are scaled before each EM iteration to give the same weight to each variable. Besides, the noise level σ^2 is estimated using the residual sum of squares divided by the number of observations minus the number of estimated parameters as suggested by Josse et al. (2016b),

$$\hat{\sigma^2} = \frac{\|Y - \sum_{l=1}^{r} u_l d_l v_l\|_2^2}{np - nr - rp + r^2},$$

where u_l, v_l and d_l are the singular vectors and the singular values from the singular value decomposition of Y. We let r denote the rank of Y, estimated here using cross-validation (Josse and Husson, 2012). In Figure 2.9, the three methods (a), (b) and (c) are compared using boxplots of the prediction error over ten simulations. The proposed method (a), designed for the MNAR setting, gives significantly smaller prediction error than other methods. Besides, the other proposed methods (b), taking the mask into account, also improve prediction errors compared to the classical MAR methods (c).

2.6 Discussion

In this article two methods have been suggested for handling self-masked MNAR data in the low-rank context: explicit modeling of the mechanism or implicit consideration by adding the mask. The first method is clearly the most successful in terms of prediction or estimation errors. Moreover, it is robust to model misspecifications. However, one should note that, on the one hand it can be computationally expensive, and then hardly scalable in the high-dimensional multivariate missing setting and on the other hand, it is a parametric

approach. Therefore, the implicit method handling both the data and the mask matrices, when taking into account the binary distribution of the latter, may be regarded as the right alternative. Both methods can handle MNAR and MAR data simultaneously.

As a take-home message, one should keep in mind that (i) if there are a few missing variables, the model-based method is extremely relevant; and (ii) when many variables can be missing, the implicit method, that models the mask using a binomial distribution, has empirically proven to provide better imputation.

Note that the logistic regression assumption may seem restrictive but the proposed approach could be easily adapted to other distributions such as the probit one.

We pointed out that when the rank is one, there are few differences between MAR and MNAR, which implies that MNAR missing values could be handled without specifiying a model. This is in line with the work of Mohan et al. (2018) in regression using graphical models and it would be interesting to extend their work to low-rank models.

As directions of future research, one could also extend this work to data matrices containing mixed variables (quantitative and categorical variables) with MNAR data, so that the logistic regression model should include the case of categorical explanatory and output variables.

In addition, in this paper, we focus on single imputation techniques where a unique value is predicted for each missing value. Consequently, it can not reflect the variance of prediction. It would be very interesting to derive confidence intervals for the predicted value, for instance by considering multiple imputation methods (Rubin, 2004).

Acknowledgments

The authors are thankful for fruitful discussion with François Husson, Wei Jiang, Imke Mayer and Geneviève Robin.

Chapter 3

Estimation and imputation in PPCA with MNAR data

This chapter corresponds to the paper Estimation and imputation in Probabilistic Principal Component Analysis with Missing Not At Random data, accepted at NeurIPS, 2020, written with Claire Boyer and Julie Josse.

Abstract

Missing Not At Random (MNAR) values where the probability of having missing data may depend on the missing value itself, are notoriously difficult to account for in analyses, although very frequent in the data. One solution to handle MNAR data is to specify a model for the missing data mechanism, which makes inference or imputation tasks more complex. Furthermore, this implies a strong *a priori* on the parametric form of the distribution. However, some works have obtained guarantees on the estimation of parameters in the presence of MNAR data, without specifying the distribution of missing data (Mohan et al., 2018; Tang et al., 2003). This is very useful in practice, but is limited to simple cases such as few self-masked MNAR variables in data generated according to linear regression models. We continue this line of research, but extend it to a more general MNAR mechanism, in a more general model of the probabilistic principal component analysis (PPCA), *i.e.*, a low-rank model with random effects. We prove identifiability of the PPCA parameters. We then propose an estimation of the loading coefficients, and a data imputation method. Both are based on estimators of means, variances and covariances of missing variables, for which consistency is discussed. These estimators have the great advantage of being calculated using only the observed information, leveraging the underlying low-rank structure of the data. We illustrate the relevance of the method with numerical experiments on synthetic data and also on two datasets, one collected from a medical register and the other one from a recommendation system.

Contents

3.1 Introduction

The problem of missing data is ubiquitous in the practice of data analysis. Theoretical guarantees of estimation strategies or imputation methods rely on assumptions regarding the missing-data mechanism, *i.e.* the cause of the lack of data. Rubin (1976) introduced three missing-data mechanisms. The data are said (i) Missing Completely At Random (MCAR) if the probability of being missing does not depend on any values observed or missing, (ii) Missing At Random (MAR) if the probability of being missing only depends on observed values, (iii) Missing Not At Random (MNAR) if the unavailability of the data may depend on both observed and unobserved data such as its value itself. We focus on this later case, which is frequent in practice, and theoretically challenging. A classic example of MNAR data is surveys about salary for which rich people would be less willing to disclose their income.

When the data is MCAR or MAR, statistical inference is carried out by ignoring the missing data mechanism (Little and Rubin, 2019). In the MNAR case, the observed data are no longer representative of the population, which leads to selection bias in the sample, and therefore to bias in the parameters estimation when using for instance complete case analysis. One solution to handle MNAR data, known as *selection model* (Little and Rubin, 2019), is to model missing data distribution; most of the time, by logistic regression models (Ibrahim et al., 1999; Morikawa et al., 2017; Sportisse et al., 2020). This comes at the price of an important computational burden to perform inference and is often restricted to a limited number of MNAR variables. In the recommender system community, some authors (Marlin and Zemel, 2009; Hernández-Lobato et al., 2014; Ma and Chen, 2019; Wang et al., 2019) suggest that not MCAR values can be handled using a joint modelling of the data and mechanism distributions by matrix factorization; then they debiase existing methods for MCAR data, for instance with inverse probability weighting approaches.

In addition, a key issue of MNAR data is to establish identifiability, which is not always guaranteed (Miao et al., 2016). The literature on this topic is abundant, both in the non-parametric (Mohan et al., 2013; Mohan and Pearl, 2014; Ilya et al., 2015; Shpitser, 2016; Nabi et al., 2020), and semi-parametric settings (Wang et al., 2014; Miao and Tchetgen, 2018). For parametric models, in the case of multivariate regression, Tang et al. (2003) and Miao et al. (2016) guarantee the identifiability of the coefficients of the conditional distribution of $Y|X$, when Y is missing. Tang et al. (2003) estimate them by calculating the coefficient of the distributions of X and $X|Y$ using only observations with no missing values. Besides, in a linear model with self-masked missing mechanism, *i.e.*, the lack depends only on the missing variable itself, Mohan et al. (2018) consider a related approach based on graphical models, adopting a causal point of view. Despite the great advantage of not modeling the distribution of missing values, the assumption of a self-masked MNAR mechanism and the restriction to a linear model are yet strong.

Contributions. We consider a framework where the data are generated according to a probabilistic principal components analysis (PPCA) (Tipping and Bishop, 1999) model. Contrary to available works that handle only MAR data in PPCA (Ilin and Raiko, 2010), we consider that the missing values mechanism can be MNAR (on several variables) and we also consider the possibility of having different mechanisms in the same data (MNAR and M(C)AR).

- We prove the identifiability of the PPCA model parameters in a self-masked MNAR values setting encompassing a large set of self-masked mechanism distributions.
- For more general MNAR mechanism, we give a strategy to estimate the PPCA loading parameters without any modeling of the missing-data mechanism and use it to impute missing values.
- The proposed method is based on estimators for the mean, the variance and the covariance of the variables with MNAR values. We show that they can be consistently estimated. Two strategies lead to the proposed estimators: (i) the first one uses algebraic arguments based on partial linear models derived from the PPCA model; (ii) the second one is inspired by (Mohan et al., 2018) and uses graphical models and in particular the so-called missingness graph.
- We derive an algorithm implementing our proposal. We show that it outperforms the state-of-the-art methods on synthetic data and on two real datasets, collected from a medical registry (Traumabase®) and from a joke recommender system (the Jester Online Joke Recommender System Hahsler (2015)). The code to reproduce all the simulations and the numerical experiments is available at `https://github.com/AudeSportisse/PPCA_MNAR`.

3.2 PPCA model with informative missing values: identifiability issues

Setting. The data matrix $Y \in \mathbb{R}^{n\times p}$ is assumed to be generated under a fully-connected PPCA model (Tipping and Bishop, 1999) (a.k.a. a low-rank model with random effects), *i.e.* by the factorization of the loading matrix $B \in \mathbb{R}^{r\times p}$ and r latent variables grouped in the matrix $W \in \mathbb{R}^{n\times r}$,

$$Y = \mathbf{1}\alpha + WB + \epsilon, \text{with} \begin{cases} W = (W_{1.}|\dots|W_{n.})^T, \text{ with } W_{i.} \sim \mathcal{N}(0_r, \mathrm{Id}_{r\times r}) \in \mathbb{R}^r, \\ B \text{ of rank } r < \min\{n,p\}, \\ \alpha \in \mathbb{R}^p \text{ and } \mathbf{1} = (1\dots 1)^T \in \mathbb{R}^n, \\ \epsilon = (\epsilon_{1.}|\dots|\epsilon_{n.})^T, \text{ with } \epsilon_{i.} \sim \mathcal{N}(0_p, \sigma^2\mathrm{Id}_{p\times p}) \in \mathbb{R}^p, \end{cases} \tag{3.1}$$

for σ^2 and r known. In the sequel, $Y_{.j}$ and $Y_{i.}$ respectively denote the column j and the row i of Y. The rows of Y are identically distributed, $\forall i \in \{1,\dots,n\}$, $Y_{i.} \sim \mathcal{N}(\alpha, \Sigma)$, with $\Sigma = B^TB + \sigma^2\mathrm{Id}_{p\times p}$. We denote $\Omega \in \{0,1\}^{n\times p}$ the missing-data pattern (or mask) defined as follows:

$$\forall i \in \{1,\dots,n\},\ \forall j \in \{1,\dots,p\}, \quad \Omega_{ij} = \begin{cases} 0 & \text{if } Y_{ij} \text{ is missing,} \\ 1 & \text{otherwise.} \end{cases} \tag{3.2}$$

Some variables $Y_{.m_1},\dots,Y_{.m_d}$, indexed by $\mathcal{M} := \{m_1,\dots,m_d\} \subset \{1,\dots,p\}$ (with $d < p$), contain MNAR values. The other variables are considered to be observed (or M(C)AR see Appendix C.2.5). We define a general MNAR mechanism where the probability to have missing values may depend on the d MNAR variables but also on $p-d-r$ other variables that can be observed or M(C)AR[1]. The remaining r variables are called *pivot variables* and can be observed or MCAR. More precisely, we denote the complementary of a set $\mathcal{A}$ as $\overline{\mathcal{A}} := \{1,\dots,p\}\backslash\mathcal{A}$. The general MNAR mechanism is defined as follows, with $\mathcal{J} \subset \overline{\mathcal{M}}$ the set of indices of the r pivot variables ($|\mathcal{J}| = r$),

$$\forall m \in \mathcal{M}, \forall i \in \{1,\dots,n\}, \quad \mathbb{P}(\Omega_{im} = 1|Y_{i.}) = \mathbb{P}(\Omega_{im} = 1|(Y_{ik})_{k\in\overline{\mathcal{J}}}). \tag{3.3}$$

We also define a specific MNAR mechanism, called the self-masked MNAR mechanism as follows. We assume that d variables are self-masked MNAR indexed by $\mathcal{M}$ and the $p-d$ other variables are MCAR (or observed), indexed by $\overline{\mathcal{M}}$, i.e, $\forall i \in \{1,\dots,n\}$,

$$\forall m \in \mathcal{M}, \quad \mathbb{P}(\Omega_{im} = 1|Y_{i.}) = \mathbb{P}(\Omega_{im} = 1|Y_{im}). \tag{3.4}$$

Model identifiability. We prove the identifiability of the PPCA model (see Appendix C.1 for the complete proof), *i.e.* the joint distribution of Y can be uniquely determined from the available information, in the self-masked missing values case. More particularly, assume the following

[1]Note that it implies that $d < p - r$.

A01. d variables are self-masked MNAR as in (3.4) and the $p-d$ other variables are MCAR (or observed). The missing-data distributions $(F_m)_{m\in\mathcal{M}}$ and $(F_j)_{j\in\overline{\mathcal{M}}}$ are known strictly monotone functions with a finite support, defined as follows, $\forall i \in \{1,\dots,n\}$,

$$\begin{aligned}\forall m \in \mathcal{M}, \quad \mathbb{P}(\Omega_{im} = 1|Y_{i.}) &= F_m(\phi_m^0 + \phi_m^1 Y_{im}),\\ \forall j \in \overline{\mathcal{M}}, \quad \mathbb{P}(\Omega_{ij} = 1|Y_{i.}) &= \mathbb{P}(\Omega_{ij} = 1) = F_j(\phi_j),\end{aligned}$$

with $\phi_j \in \mathbb{R}$ and $\phi_m = (\phi_m^0, \phi_m^1) \in \mathbb{R}^2$ the mechanism parameters.

A02. $\forall (k,\ell) \in \{1,\dots,p\}^2, \quad k \neq \ell, \qquad \Omega_{.k} \perp\!\!\!\perp \Omega_{.\ell}|Y$

Note that under Assumption **A01.**, any function $F_m, m \in \mathcal{M}$ can be considered, as a logistic function while (Miao et al., 2016) presented many counterexamples when identification fails considering the logistic distribution. **A02.** requires that the missing-data patterns are independent conditionally to the data.

Proposition 9. *Under Assumptions **A01.** and **A02.**, the parameters (α, Σ) of the PPCA model (3.1) and the mechanism parameters $\phi = (\phi_\ell)_{\ell\in\{1,\dots p\}}$ are identifiable. Assuming that the noise level σ^2 is known, the parameter B is identifiable up to a row permutation.*

3.3 Estimators with theoretical guarantees

In this section, we provide estimators of the means, variances and covariances for the MNAR variables, when data are generated under the PPCA model described in (3.1). These estimators are used to derive an estimator of the loading matrix B in (3.1). This makes it possible to derive a new imputation method with MNAR data as detailed in Algorithm 1.

We denote $\mathcal{J}_{-j} := \mathcal{J}\backslash\{j\}$ and assume

A1. $\forall m \in \mathcal{M}, \forall j \in \mathcal{J}, \quad \begin{pmatrix} B_{.m} & (B_{.j'})_{j'\in\mathcal{J}_{-j}}\end{pmatrix}$ is invertible,

A2. $\forall m \in \mathcal{M}, \forall j \in \mathcal{J}, \quad Y_{.j} \perp\!\!\!\perp \Omega_{.m}|(Y_{.k})_{k\in\overline{\{j\}}}$.

Note that Assumption **A1.** implies that B has a full rank r and that any variable in Y is generated by all the latent variables[2] (named a "fully-connected" PPCA). Assumption **A2.** is implied by the general MNAR mechanism in (3.3).

We start by illustrating the methodology and the assumptions using an example in small dimension, before turning to the general case.

3.3.1 Estimation of the mean of a MNAR variable

Consider a toy dataset where $p = 3, r = 2$, in which only one variable is missing, $\mathcal{M} = \{1\}$ and there are two pivots variables $\mathcal{J} = \{2, 3\}$. Note that the MNAR mechanism is self-masked in such a context, because Equation (3.3) leads to $\mathbb{P}(\Omega_{.1} = 1|Y_{.1}, Y_{.2}, Y_{.3}) = \mathbb{P}(\Omega_{.1} = 1|Y_{.1})$, but the method can be extended to more general cases. Our aim is to estimate the mean of $Y_{.1}$, without specifying the distribution of the missing-data mechanism.

[2] It does not require that the linear combination coefficients are non-zero.

Using algebraic arguments. We proceed in three steps: (i) **A1.** allows to obtain linear link between the pivot variables $(Y_{\cdot 2}, Y_{\cdot 3})$ and the MNAR variable $Y_{\cdot 1}$. For instance,

$$Y_{\cdot 2} = \mathcal{B}_{2\to 1,3[0]} + \mathcal{B}_{2\to 1,3[1]} Y_{\cdot 1} + \mathcal{B}_{2\to 1,3[3]} Y_{\cdot 3} + \zeta, \tag{3.5}$$

with ζ a noise term, $\mathcal{B}_{2\to 1,3[0]}$, $\mathcal{B}_{2\to 1,3[1]}$ and $\mathcal{B}_{2\to 1,3[3]}$ the intercept and the coefficients in the model (the arrow $2 \to 1,3$ indicates the regression model of $Y_{\cdot 2}$ on $Y_{\cdot 1}$ and $Y_{\cdot 3}$, while the squared bracket represents the coefficient, for instance 3 for the coefficient of $Y_{\cdot 3}$) ; (ii) Assumption **A2.**, *i.e.* $Y_{\cdot 2} \perp\!\!\!\perp \Omega_{\cdot 1} | Y_{\cdot 1}, Y_{\cdot 3}$, is required to obtain identifiable and consistent parameters of the distribution of $Y_{\cdot 2}$ given $Y_{\cdot 1}, Y_{\cdot 3}$ in the complete-case when $\Omega_{\cdot 1} = 1$, denoted as $\mathcal{B}^c_{2\to 1,3[0]}$, $\mathcal{B}^c_{2\to 1,3[1]}$ and $\mathcal{B}^c_{2\to 1,3[3]}$,

$$(Y_{\cdot 2})_{|\Omega_{\cdot 1}=1} = \mathcal{B}^c_{2\to 1,3[0]} + \mathcal{B}^c_{2\to 1,3[1]} Y_{\cdot 1} + \mathcal{B}^c_{2\to 1,3[3]} Y_{\cdot 3} + \zeta^c, \tag{3.6}$$

(note that the regression of $Y_{\cdot 1}$ on $(Y_{\cdot 2}, Y_{\cdot 3})$ is prohibited, as **A2.** does not hold); (iii) using again **A2.**,

$$\mathbb{E}\left[Y_{\cdot 2} | Y_{\cdot 1}, Y_{\cdot 3}, \Omega_{\cdot 1} = 1\right] = \mathbb{E}\left[\mathcal{B}^c_{2\to 1,3[0]} + \mathcal{B}^c_{2\to 1,3[1]} Y_{\cdot 1} + \mathcal{B}^c_{2\to 1,3[3]} Y_{\cdot 3} | Y_{\cdot 1}, Y_{\cdot 3}\right],$$

and taking the expectation leads to

$$\mathbb{E}\left[Y_{\cdot 2}\right] = \mathcal{B}^c_{2\to 1,3[0]} + \mathcal{B}^c_{2\to 1,3[1]} \mathbb{E}\left[Y_{\cdot 1}\right] + \mathcal{B}^c_{2\to 1,3[3]} \mathbb{E}\left[Y_{\cdot 3}\right].$$

The latter expression can be reshuffled so that the expectation of $Y_{\cdot 1}$ can be estimated: the means of $Y_{\cdot 2}$ and $Y_{\cdot 3}$ are estimated by standard empirical estimators (it will be Assumption **A4.** in the sequel).

Using graphical arguments. The PPCA model can be represented with structural causal graphs (Pearl, 2003), as illustrated in Figure 3.1. The top left graph in which each variable is generated by a combination of all latent variables, see Assumption **A1.**, can be represented as the top right one, as $Y_{\cdot 1} \leftarrow W_{\cdot 1} \rightarrow Y_{\cdot 2}$ is equivalent to $Y_{\cdot 1} \leftrightarrow Y_{\cdot 2}$ (see (Pearl, 2003, page 52)). Then, six reduced graphical models can be derived from the top right graph (two instances are represented in the bottom). Indeed, a bidirected edge $Y_{\cdot 1} \leftrightarrow Y_{\cdot 2}$ can be interchanged (see (Pearl, 2003, rule 1, page 147)) with an oriented edge $Y_{\cdot 1} \rightarrow Y_{\cdot 2}$, if each neighbor of $Y_{\cdot 2}$ (*i.e.* $Y_{\cdot 1}$ or $Y_{\cdot 3}$) is inseparable of $Y_{\cdot 1}$ (see (Pearl, 2003, page 17)). The bottom left graph can also be represented by Equation (3.5), which gives a connection between the algebraic and graphical approaches.

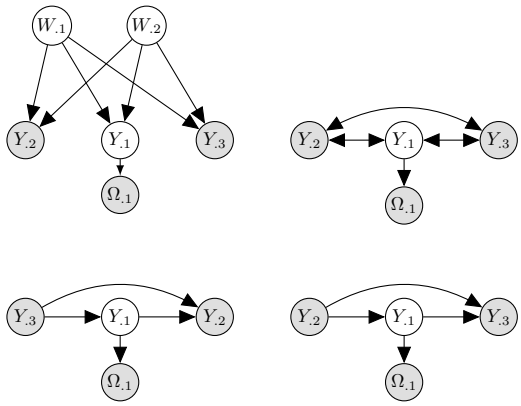


Figure 3.1: Graphical models for the toy example with one missing variable $Y_{\cdot 1}$, $p = 3$ and $r = 2$.

3.3.2 Estimation of the mean, variance and covariances of the MNAR variables

In a general case, estimators of the mean, variance and covariances of the variables with MNAR values can be computed one by one. We detail the results only for one variable $Y_{\cdot m}, m \in \mathcal{M}$, but the results hold for several variables with MNAR values. In addition, the other variables are considered to be observed for simplicity but they could contain MCAR and MAR values as well, as explained in Appendix C.2.5. We adopt the algebraic strategy here to derive estimators (see Appendix C.2 for proofs) but graphical arguments can also be used to obtain similar results (see Appendix C.6). The starting point is to exploit the linear links between variables, as described in the next lemma.

Lemma 1. *Under the PPCA model (3.1) and Assumption **A1.**, choose $j \in \mathcal{J}$. One has*

$$Y_{\cdot j} = \mathcal{B}_{j\to m,\mathcal{J}_{-j}}[0] + \sum_{j' \in \mathcal{J}_{-j}} \mathcal{B}_{j\to m,\mathcal{J}_{-j}}[j'] Y_{\cdot j'} + \mathcal{B}_{j\to m,\mathcal{J}_{-j}}[m] Y_{\cdot m} + \zeta, \tag{3.7}$$

where $\zeta = -\sum_{j' \in \mathcal{J}_{-j}} \mathcal{B}_{j\to m,\mathcal{J}_{-j}}[j'] \epsilon_{\cdot j'} - \mathcal{B}_{j\to m,\mathcal{J}_{-j}}[m] \epsilon_{\cdot m} + \epsilon_{\cdot j}$. is a noise term.

$\mathcal{B}_{j\to m,\mathcal{J}_{-j}}[0]$, $\mathcal{B}_{j\to m,\mathcal{J}_{-j}}[j']$ and $\mathcal{B}_{j\to m,\mathcal{J}_{-j}}[m]$ are given in Appendix C.2.1 and depend on the coefficients of B given in (3.1).

Then we define the regression coefficients of $Y_{\cdot j}$ on $Y_{\cdot m}$ and $Y_{\cdot k}$, for $k \in \mathcal{J}_{-j}$ in the complete case, that will be used to express the mean of a variable with MNAR values.

Definition 10 (Coefficients in the complete case)**.** *For $j \in \mathcal{J}$ and $k \in \mathcal{J}_{-j}$, let $\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[0]$, $\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[m]$ and $\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[j']$ be respectively the intercept and the coefficients standing for the effects of $Y_{\cdot j}$ on $(Y_{\cdot m}, (Y_{\cdot j'})_{j' \in \mathcal{J}_{-j}})$ in the complete case, i.e. when $\Omega_{\cdot m} = 1$:*

$$(Y_{\cdot j})_{|\Omega_{\cdot m}=1} := \mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[0] + \sum_{j' \in \mathcal{J}_{-j}} \mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[j'] Y_{\cdot j'} + \mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[m] Y_{\cdot m} + \zeta^c, \tag{3.8}$$

with $\zeta^c = -\sum_{j' \in \mathcal{J}_{-j}} \mathcal{B}^c_{j \to m, \mathcal{J}_{-j}[j']} \epsilon_{\cdot j'} - \mathcal{B}^c_{j \to m, \mathcal{J}_{-j}[m]} \epsilon_{\cdot m} + \epsilon_{\cdot j}$.

Then, we make the two following assumptions:

A3. For all $j \in \mathcal{J}$, for all $m \in \mathcal{M}$, the complete-case coefficients $\mathcal{B}^c_{j \to m, \mathcal{J}_{-j}[0]}$, $\mathcal{B}^c_{j \to m, \mathcal{J}_{-j}[m]}$ and $\mathcal{B}^c_{j \to m, \mathcal{J}_{-j}[k]}, k \in \mathcal{J}_{-j}$ can be consistently estimated.

A4. The means $(\alpha_j)_{j \in \mathcal{J}}$, variances $(\mathrm{Var}(Y_{\cdot j}))_{j \in \mathcal{J}}$ and covariances $(\mathrm{Cov}(Y_{\cdot j}, Y_{\cdot j'}))_{j \in \mathcal{J}, j' \in \mathcal{J}_{-j}}$ of the r pivot variables can be consistently estimated.

Note that Assumption **A4.** is met whether the r pivot variables are fully observed.

Proposition 11 (Mean estimator)**.** *Consider the PPCA model (3.1). Under Assumptions **A1.** and **A2.**, an estimator of the mean of a MNAR variable $Y_{\cdot m}$, for $m \in \mathcal{M}$, can be constructed as follows: choose $j \in \mathcal{J}$, and compute*

$$\hat{\alpha}_m := \frac{\hat{\alpha}_j - \hat{\mathcal{B}}^c_{j \to m, \mathcal{J}_{-j}[0]} - \sum_{j' \in \mathcal{J}_{-j}} \hat{\mathcal{B}}^c_{j \to m, \mathcal{J}_{-j}[j']} \hat{\alpha}_{j'}}{\hat{\mathcal{B}}^c_{j \to m, \mathcal{J}_{-j}[m]}}, \tag{3.9}$$

with $(\hat{\mathcal{B}}^c_{j \to m, \mathcal{J}_{-j}[k]})_{k \in \{0, m\} \cup \mathcal{J}_{-j}}$ estimators of the coefficients obtained from Definition 10.

*Under the additional Assumptions **A3.** and **A4.**, this estimator is consistent.*

The proof is given in Appendix C.2.2. Proposition 11 provides an estimator easily computable from all observed cells. Furthermore, different choices of $Y_{\cdot j}$, $j \in \mathcal{J}$ can be done in Equation (3.9) and all the resulting estimators may be aggregated to stabilize the estimation of α_m.

Proposition 12 (Variance and covariances estimators)**.** *Consider the PPCA model (3.1). Under Assumptions **A1.** and **A2.**, an estimator of the variance of a MNAR variable $Y_{\cdot m}$, for $m \in \mathcal{M}$, and its covariances with the pivot variables, can be constructed as follows: choose a pivot variable $Y_{\cdot j}$ for $j \in \mathcal{J}$ and compute*

$$\left(\widehat{\mathrm{Var}}(Y_{\cdot m}) \quad \widehat{\mathrm{Cov}}(Y_{\cdot m}, (Y_{\cdot j'})_{j' \in \mathcal{J}}) \right)^T := (\widehat{M}_j)^{-1} \hat{P}_j, \tag{3.10}$$

assuming that σ^2 tends to zero, with $\widehat{M}_j^{-1} \in \mathbb{R}^{(r+1) \times (r+1)}$, $\hat{P}_j \in \mathbb{R}^{r+1}$ detailed in Appendix C.2.3. These quantities depend on $(\hat{\alpha}_{j'})_{j' \in \mathcal{J}}$, $\hat{\alpha}_m$ given in Proposition 11, on $(\widehat{\mathrm{Var}}(Y_{\cdot j'}))_{j' \in \mathcal{J}}$ and on complete-case coefficients such as $(\hat{\mathcal{B}}^c_{j' \to m, \mathcal{J}_{-j'}[k]})_{k \in \{m\} \cup \mathcal{J}_{-j'}}$ for $j' \in \mathcal{J}$.

*Under the additional Assumptions **A3.** and **A4.**, the estimators of the variance of $Y_{\cdot m}$ and its covariances with the pivot variables given in (3.10) are consistent.*

The proof is given in Appendix C.2.3. Note that to estimate the variance of a MNAR variable, only r pivot variables are required to solve (3.10) and r tasks have to be performed for estimating the coefficients of the effects of $Y_{\cdot k}$ on $(Y_{\cdot \ell})_{\ell \in \{m\} \cup \mathcal{J}_{-k}}$ for all $k \in \mathcal{J}$.

All the ingredients can be combined to form an estimator $\hat{\Sigma}$ for the covariance matrix Σ. Define

$$\hat{\Sigma} := \left(\widehat{\mathrm{Cov}}(Y_{\cdot k}, Y_{\cdot \ell}) \right)_{k, \ell \in \{1, \ldots, p\}}, \tag{3.11}$$

- if $Y_{.k}$ and $Y_{.\ell}$ have both consistent mean/variance estimators, then $\widehat{\mathrm{Cov}}(Y_{.k}, Y_{.\ell})$ can be trivially evaluated by standard empirical covariance estimators.
- if $Y_{.k}$ is a MNAR variable and $Y_{.\ell}$ is a pivot variable, then $\widehat{\mathrm{Cov}}(Y_{.k}, Y_{.\ell})$ is given by (3.10),
- if $Y_{.k}$ is a MNAR variable and $Y_{.\ell}$ is not a pivot variable, *i.e.* $\ell \in \overline{\mathcal{J}} \backslash \{k\}$, a similar strategy as the one above can be devised. Then $\widehat{\mathrm{Cov}}(Y_{.k}, Y_{.\ell})$ is given by (C.35) detailed in Appendix C.2.4 and for which some additional assumptions similar as the ones above are required. This estimator relies on the choice of $r-1$ pivot variables indexed by j and $\mathcal{H} \subset \mathcal{J}$, and only necessitates to evaluate the effects of $Y_{.j}$ on $(Y_{.j'})_{j' \in \{k,\ell\} \cup \mathcal{H}}$ in the complete case.

3.3.3 Performing PPCA with MNAR variables

With the estimator $\hat{\Sigma}$ in (3.11) at hand, one can perform the estimation of the loading matrix B in (3.1).

Definition 13 (Estimation of the loading matrix)**.** *Given the estimator $\hat{\Sigma}$ of the covariance matrix in (3.11), let the orthogonal matrix $\hat{U} = (\hat{u}_1 | \dots | \hat{u}_p) \in \mathbb{R}^{p \times p}$ and the diagonal matrix $\hat{D} = \mathrm{diag}(\hat{d}_1, \hat{d}_2, \dots, \hat{d}_p) \in \mathbb{R}^{p \times p}$ with $\hat{d}_1 \geqslant \hat{d}_2 \geqslant \dots \geqslant \hat{d}_p \geqslant 0$ form the singular value decomposition of the following matrix $\hat{\Sigma} - \sigma^2 \mathrm{Id}_{p \times p} =: \hat{U}\hat{D}\hat{U}^T$. An estimator $\hat{B}$ of B can be defined using the r first singular values and vectors, as follows*

$$\hat{B} = \hat{D}_{|r}^{1/2} \hat{U}_{|r}^T = \mathrm{diag}(\hat{d}_1, \dots, \hat{d}_r)^{1/2} (\hat{u}_1^T | \dots | \hat{u}_r^T)^T \tag{3.12}$$

The estimation of the loading matrix is used to impute the variables with missing values. More precisely, a classical strategy to impute missing values is to estimate their conditional expectation given the observed values. One can note that with $\Sigma = B^T B + \sigma^2 \mathrm{Id}_{p \times p}$, the conditional expectation of $Y_{.m}$ for $m \in \mathcal{M}$ given $(Y_{.k})_{k \in \overline{\mathcal{M}}}$ reads as follows

$$\mathbb{E}[Y_{.m} | (Y_{.k})_{k \in \overline{\mathcal{M}}}] = \alpha_m + \Sigma_{m,\overline{\mathcal{M}}} \Sigma_{\overline{\mathcal{M}},\overline{\mathcal{M}}}^{-1} \left(Y_{.\overline{\mathcal{M}}}^T - \alpha_{\overline{\mathcal{M}}} \right),$$

with $\Sigma_{m,\overline{\mathcal{M}}} := (\Sigma_{m,k})_{k \in \overline{\mathcal{M}}}^T$, $\Sigma_{\overline{\mathcal{M}},\overline{\mathcal{M}}} := (\Sigma_{k,k'})_{k,k' \in \overline{\mathcal{M}}}$, $Y_{.\overline{\mathcal{M}}} := (Y_{.k})_{k \in \overline{\mathcal{M}}}$, and $\alpha_{\overline{\mathcal{M}}} := (\alpha_k)_{k \in \overline{\mathcal{M}}}$.

Definition 14 (Imputation of a MNAR variable)**.** *Set $\hat{\Gamma} := \hat{B}^T \hat{B} + \sigma^2 \mathrm{Id}_{p \times p}$ for $\hat{B}$ given in Definition 13. The MNAR variable $Y_{.m}$ with $m \in \mathcal{M}$ can be imputed as follows: for i such that $\Omega_{i,m} = 0$,*

$$\hat{Y}_{im} = \hat{\alpha}_m + \hat{\Gamma}_{m,\overline{\mathcal{M}}} \hat{\Gamma}_{\overline{\mathcal{M}},\overline{\mathcal{M}}}^{-1} \left(Y_{i,\overline{\mathcal{M}}}^T - \hat{\alpha}_{\overline{\mathcal{M}}} \right) \tag{3.13}$$

with $\hat{\Gamma}_{m,\overline{\mathcal{M}}} := (\hat{\Gamma}_{m,k})_{k \in \overline{\mathcal{M}}}^T$, $\hat{\Gamma}_{\overline{\mathcal{M}},\overline{\mathcal{M}}} := (\hat{\Gamma}_{k,k'})_{k,k' \in \overline{\mathcal{M}}}$, $Y_{.\overline{\mathcal{M}}} := (Y_{.k})_{k \in \overline{\mathcal{M}}}$ and $\hat{\alpha}_{\overline{\mathcal{M}}} := (\hat{\alpha}_k)_{k \in \overline{\mathcal{M}}}$.

3.3.4 Algorithm

The proposed imputation method described in Algorithm 1 can handle the different MNAR mechanisms, the self-masked MNAR case and the general MNAR cases where the probability to have missing values on variables depends on both the underlying values and values of other variables (observed or missing).

Algorithm 1 PPCA with MNAR variables.

Require: r (number of latent variables), σ^2 (noise level), $\mathcal{J}$ (pivot variables indices), Ω (mask).

1: **for** each MNAR variable $(Y_{\cdot m})_{m\in\mathcal{M}}$ **do**
2: Evaluate $\hat{\alpha}_m$ the estimator of its mean given in (3.9) using the r pivot variables indexed by $\mathcal{J}$.
3: Evaluate $\widehat{\mathrm{Var}}(Y_{\cdot m})$, and $\widehat{\mathrm{Cov}}(Y_{\cdot m}, Y_{\cdot \ell})$ for $\ell \in \mathcal{J}$, using (3.10).
4: Evaluate $\widehat{\mathrm{Cov}}(Y_{\cdot m}, Y_{\cdot \ell})$ for $\ell \in \overline{\mathcal{J}}\backslash\{m\}$ using Proposition 28.
5: **end for**
6: Form $\hat{\Sigma}$, covariance matrix estimator in (3.11).
7: Compute the loading matrix estimator $\hat{B}$ given in (3.12).
8: Compute $\hat{\Gamma} = \hat{B}^T\hat{B} + \sigma^2 \mathrm{Id}_{p\times p}$.
9: **for** each missing variable $(Y_{\cdot j})$ **do**
10: **for** i such that $\Omega_{ij} = 0$ **do**
11: $\hat{Y}_{ij} \leftarrow$ Impute Y_{ij} as in (3.13).
12: **end for**
13: **end for**

Algorithm 1 requires the set $\mathcal{J}$, *i.e.* the selection of r pivot variables on which the regressions in Propositions 11, 12 and 28 will be performed. If there are more than r variables that can be pivot, we suggest selecting a bigger set ($> r$) and computing the final estimator with the median of the estimators over all possible combinations. The efficiency of this strategy is illustrated in Appendix C.3.

The estimators associated to any missing variable in the steps 1 to 5 are computed in the complete case, i.e. with the rows for which the missing variable is observed. When the pivot variables are also missing, the complete case corresponds to discarding all rows where the pivot variables or the MNAR one are missing and not all rows containing missing values. This could be problematic in the high-dimensional setting, but here the low-rank assumption ($r < \min\{n, p\}$) ensures that the number of pivot variables is small enough, so that the complete case analysis will not result in discarding many rows of the dataset.

In order to estimate the coefficients in Definition 10, we use ordinary least squares despite that the exogeneity assumption, *i.e.* the noise term is independent of the covariates, does not hold. It still leads to accurate estimation in numerical experiments as shown in Section 3.4. Actually, the consistency required by Assumption **A3.** holds as the variance of the noise tends to 0.

3.4 Numerical experiments

3.4.1 Synthetic data

We empirically compare Algorithm 1 (**MNAR**) to the state-of-the-art methods, including

(i) **MAR**: our method which has been adapted to handle MAR data (inspired by (Mohan et al., 2018, Theorems 1, 2, 3) in linear models), see Appendix C.7 for details;

(ii) **EMMAR**: which consists in an EM algorithm to perform PPCA with MAR values (Ilin and Raiko, 2010);

(iii) **SoftMAR**: a matrix completion method using an iterative soft-thresholding singular value decomposition algorithm (Mazumder et al., 2010) relevant only for M(C)AR values;

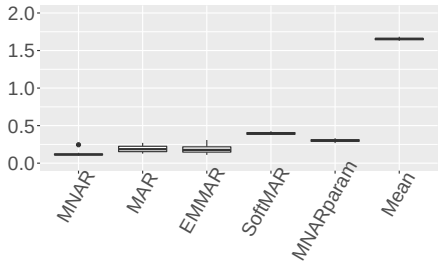
(iv) **MNARparam**: a matrix completion technique modeling the MNAR mechanism with a parametric logistic model (Sportisse et al., 2020).

Note that Method (ii) is specially designed to estimate the PPCA loading matrix and not to perform imputation, but this is possible combining Method (ii) with steps 8 and 9 in Algorithm 1. This is the other way around for completion Methods (iii) and (iv), but the loading matrix can be computed as in (3.12). Note also that Methods (iii) and (iv) are developed in a context of low-rank models with fixed effects. They require tuning a regularization parameter λ: we consider an oracle value minimizing the true imputation error. We also use oracle values for the noise level and the rank in Algorithm 1. These methods are compared with the imputation by the mean (**Mean**), which serves as a benchmark, and the naive listwise deletion method (**Del**) which consists in estimating the parameters empirically with the fully-observed data only. A comparison of the methods in terms of computational times is given in Appendix C.4.

Measuring the performance. For the loading matrix, the RV coefficient (Josse et al., 2008), which is a measure of relationship between two random vectors, is computed between the estimate $\hat{B}$ and the true B. An RV coefficient close to one means high correlation between the image spaces of $\hat{B}$ and B. Denoting the Frobenius norm as $\|.\|_F$, the quality of imputation is measured with the normalized imputation error given by $\|(\hat{Y} - Y) \odot (1 - \Omega)\|_F^2 / \|Y \odot (1 - \Omega)\|_F^2$.

Setting. We generate a data matrix of size $n = 1000$ and $p = 10$ from a PPCA model (3.1) with two latent variables ($r = 2$) and with a noise level $\sigma = 0.1$. Missing values are introduced on seven variables $(Y_{\cdot k})_{k \in [1:7]}$ according to a logistic self-masked MNAR mechanism, leading to 35% of missing values in total. Results are presented[3] for one missing variable $Y_{\cdot 1}$ (same results hold for other missing variables). All the observed variables $(Y_{\cdot k})_{k \in [8:10]}$ are considered to be pivot. Figure 3.3 shows that Algorithms 1 is the only one

[3]For a given set of PPCA parameters, the stochasticity comes from the process of drawing 20 times the latent variables, the additive noise and the missing-data pattern.



	RV
MNAR	0.997
MAR	0.993
EMMAR	0.988
SoftMAR	0.986
Mean	0.593

Figure 3.2: Imputation error (left) and median of the RV coefficients for the loading matrix (right).

which always gives unbiased estimators of the mean, variance and associated covariances of $Y_{\cdot 1}$. As expected, the listwise deletion method provides biased estimates inasmuch as the observed sample is not representative of the population with MNAR data. Method (ii), specifically designed for PPCA models but assuming MAR missing values, provides biased estimators. Method (iv) improves on the benchmark mean imputation and on Method (iii) as well as it explicitly takes into account the MNAR mechanism, but it still leads to biased estimates probably because of the fixed effects model assumption. Figure 3.2 shows that Algorithm 1 gives the best estimate of the loading matrix and the smallest imputation error. Method (i), based on the same arguments as Algorithm 1 but considering MAR data, may be considered as a second choice for this low-dimensional example as the biais is quite small (yet not in higher dimension, see Appendix C.3).

Misspecification to the PPCA model. The data matrix $Y \in \mathbb{R}^{n \times p}$ of size $n = 200$ and $p = 10$ is now generated under the fixed effects model such that $Y = \Theta + \epsilon$, with $\Theta \in \mathbb{R}^{n \times p}$ a low-rank matrix with $r = 2$ and $\epsilon \in \mathbb{R}^{n \times p}$ a Gaussian noise matrix with $\sigma = 0.1$. Figure 3.4 shows that mean and variance estimators given by Algorithm 1 have a larger variance than those given by Method (iv) precisely dedicated to this specific setting. But surprisingly, Algorithm 1 provides less biased estimates than Method (iv).

In Appendix C.3, we report further simulation results, where we vary the features dimension ($p = 50$), the rank ($r = 5$), the missing values mechanism using probit self-masking and also multivariate MNAR (when the probability to be missing for a variable depends on its underlying values and on values of other variables that can be missing) and the percentage of missing values (10%, 50%). We obtain similar results as before, and as expected, all the methods deteriorate with an increasing percentage of missing values but our method remains stable.

In addition to the model misspecification experiment (assuming a fixed effect model), we assess the robustness of the methods in terms of noise level and we evaluate the impact of under- or overestimating the number r of latent variables. When the level of noise increases, our method is very robust in terms of mean and variance estimations, and despite a bias for some covariances estimations for large noise it outperforms competitors regarding the

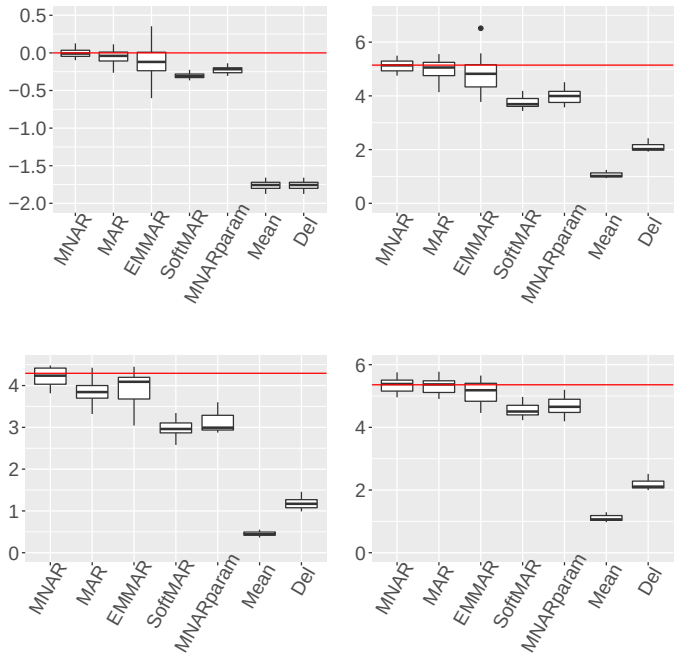


Figure 3.3: Mean (top left) and variance (top right) estimations of the missing variable and covariances (bottom) estimations of $\mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot 2})$ (*i.e.* covariance between two missing variables) and of $\mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot 8})$ (*i.e.* between one missing variable and one pivot variable). True values are indicated by red lines.

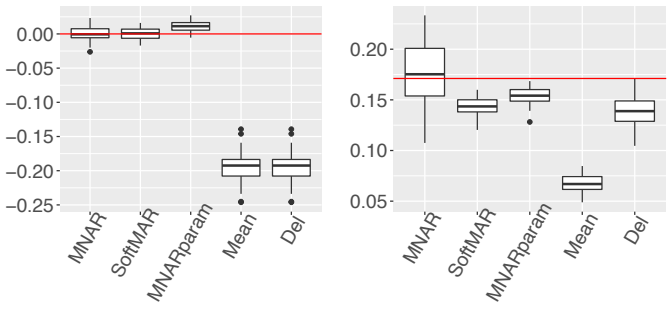


Figure 3.4: Mean (left) and variance (right) estimations of $Y_{\cdot 1}$ when data are generated under the fixed effects model.

imputation error. It also turns out that the procedure remains stable at a wrong specification of the number r of latent variables.

3.4.2 Application to recommendation system data

To show the extent and feasibility of our methodology on real data, we detail the methodology on the Jester dataset Hahsler (2015) of 5000 users who rated 100 jokes, with 27% of missing values.

Discussion on the assumptions. First, considering MNAR and self-masking values is plausible because users only rate jokes they like or dislike strongly or might be ashamed to assume their taste for sexual jokes for instance. Then, Assumption **A1.**, which can be viewed as a low-rank assumption for the loading matrix, makes sense in the rating context: any variable (i.e. user preferences) can be expressed as a linear combination of r latent variables. In particular, the first latent variable opposes individuals who like jokes about physics but dislike jokes about sexuality, and conversely. Finally, Assumption **A2.** means that a user's non-response for a sexual joke given all jokes may depend on the scores of the sexual and physical jokes but not on the musical and computer jokes.

Selecting the number r of latent variables and estimating the noise variance. In practice, to select r, one could use complete observations only but this is not possible when the number of features is large. As an alternative, we use a cross-validation strategy assuming M(C)AR mechanism as detailed in Josse and Husson (2012). Algorithm 1 is robust to a misspecification of the rank (see Appendix C.3) and thus a reasonable heuristic may already be enough. With r at hand, the noise variance is obtained directly using weighted residual sum of squares as in (Josse et al., 2016b). Without further information on the missing mechanisms, we select the r pivot variables with the lowest missing rate.

Imputation performances. To assess the quality of our method, we introduce additional MNAR values using a logistic self-masked mechanism in a chosen variable with an initial rate of 33% and a final one of 65%. The other variables are considered M(C)AR. The process is repeated 10 times. We compare our method to the EMMAR, SoftMAR and add an imputation method based on deep generative models `Deep` Gondara and Wang (2018)[4]. The parametric method `MNARparam` is not performed as it does not scale on such large data. Figure 3.5 shows that Algorithm 1 outperforms the competitors (mean imputation corresponds to an error of 1).

3.4.3 Application to clinical data

We illustrate our method on the TraumaBase® dataset containing the clinical measurements of 3159 patients with brain trauma injury (see Appendix C.5 for more information). Nine quantitative variables, selected by doctors, contain missing values (11% in the whole dataset).

[4]Note that this method requires to be trained on a complete dataset.

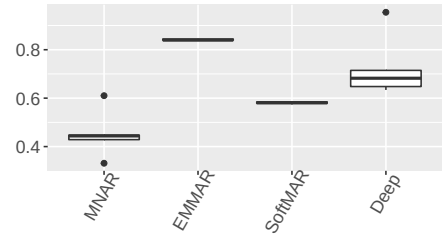


Figure 3.5: Imputation error for the Jester dataset.

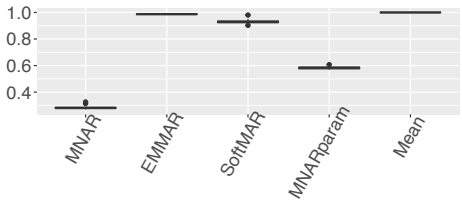


Figure 3.6: Imputation error for the TraumaBase dataset.

After discussion with doctors, some variables can be considered to have MNAR values, such as the variable *HR.ph*, which denotes the heart rate. Indeed, when the patient's condition is too critical and therefore his heart rate is either high or low, the heart rate may not be measured, as doctors prefer to provide emergency care.

As for the Jester dataset, we introduce additional MNAR values in the variable *HR.ph* (which has an initial missing rate of 1%) using a logistic self-masked mechanism leading to 50% missing values. Both the rank and the noise level are estimated using the complete-case analysis (1862 observations). The selection of the pivot variables was discussed with experts (doctors) who identified M(C)AR variables. In Figure 3.6, Algorithm 1 gives significantly smaller imputation error than other methods. In addition, a supervised learning task is also performed in Appendix C.5 for which Algorithm 1 also gives the smallest prediction error.

3.5 Discussion

In this work, we propose a new estimation and imputation method to perform PPCA with MNAR data (possibly coupled with M(C)AR data), without any need of modeling the missing mechanism. This comes with strong theoretical guarantees as identifiability and consistency, but also with an efficient algorithm. Estimating the rank in the PPCA setting with MNAR data remains non trivial. Once the number of latent variables is estimated, the noise variance can be estimated. A cross-validation strategy by additionally adding some MNAR values is a first solution, but this definitely requires further research. Another ambitious prospect would be to extend work to the exponential family to process count data, for example, which is prevalent in many application fields such as genomics.

Part II

Supervised and unsupervised framework for missing values

Chapter 4

Debiased averaged SGD algorithm with heterogeneous MCAR data

This chapter corresponds to the paper Debiasing Stochastic Gradient Descent to handle missing values, accepted at NeurIPS, 2020, written with Claire Boyer, Aymeric Dieuleveut and Julie Josse.

Abstract

Stochastic gradient algorithm is a key ingredient of many machine learning methods, particularly appropriate for large-scale learning. However, a major caveat of large data is their incompleteness. We propose an averaged stochastic gradient algorithm handling missing values in linear models. This approach has the merit to be free from the need of any data distribution modeling and to account for heterogeneous missing proportion. In both streaming and finite-sample settings, we prove that this algorithm achieves convergence rate of $\mathcal{O}(\frac{1}{n})$ at the iteration n, the same as without missing values. We show the convergence behavior and the relevance of the algorithm not only on synthetic data but also on real data sets, including those collected from medical register.

Contents

4.1 Introduction

Stochastic gradient algorithms (SGD) (Robbins and Monro, 1951) play a central role in machine learning problems, due to their cheap computational cost and memory per iteration. There is a vast literature on its variants, for example using averaging of the iterates (Polyak and Juditsky, 1992), some robust versions of SGD (Nemirovski et al., 2009; Juditsky et al., 2011) or adaptive gradient algorithms like Adagrad (Duchi et al., 2011); and on theoretical guarantees of those methods (Moulines and Bach, 2011; Bach and Moulines, 2013; Dieuleveut et al., 2017; Shamir and Zhang, 2013; Hazan and Kale, 2011; Needell et al., 2014). More globally, averaging strategies have been used to stabilize the algorithm behaviour and reduce the impact of the noise, giving better convergence rates without requiring strong convexity.

The problem of missing values is ubiquitous in large scale data analysis. One of the key challenges in the presence of missing data is to deal with the half-discrete nature of the data which can be seen as a mixed of continuous data (observed values) and categorical data (the missing values). In particular for gradient-based methods, the risk minimization with incomplete data becomes intractable and the usual results cannot be directly applied.

Context. In this paper, we consider a linear regression model, for $i \geqslant 1$,

$$y_i = X_{i:}^T \beta^\star + \epsilon_i, \tag{4.1}$$

parametrized by $\beta^\star \in \mathbb{R}^d$, where $y_i \in \mathbb{R}$, $\epsilon_i \in \mathbb{R}$ is a real-valued centered noise and $X_{i:} \in \mathbb{R}^d$ stands for the real covariates of the i-th observation. The $(X_{i:})$'s are assumed to be only partially known, since some covariates may be missing: our objective is to derive stochastic algorithms for estimating the parameters of the linear model, which handle missing data, and come with strong theoretical guarantees on excess risk.

Related works. There is a rich literature on handling missing values (Little and Rubin, 2019) and yet there are still some challenges even for linear regression models. This is all the more true as we consider such models for large sample size or in high dimension. There are very few regularized versions of regression that can deal with missing values. A classical approach to estimating parameters with missing values consists in maximizing the observed likelihood, using for instance an Expectation Maximization algorithm (Dempster et al., 1977). Even if this approach can be implemented to scale for large datasets see for instance (Cappé and Moulines, 2009), one of its main drawbacks is to rely on strong parametric assumptions for the covariates distributions. Another popular strategy to fix the missing values issue consists in predicting the missing values to get a completed data and then in applying the desired method. However matrix completion is a different problem from estimating parameters and can lead to uncontrolled bias and undervalued variance of the estimate (Little and Rubin, 2019). In the regression framework, Jones (1996) studied the bias induced by naive imputation.

In the settings of the Dantzig selector (Rosenbaum et al., 2010) and LASSO (Loh and Wainwright, 2011), another solution consists in naively imputing by 0 the incomplete matrix and modifying the algorithm used in the complete case to account for the imputation error. Such a strategy has also been studied by Ma and Needell (2018) for SGD in the context of linear regression with missing values and with finite samples: the authors used debiased gradients, in the same spirit as the covariance matrix debiasing considered by Loh and Wainwright (2011) in a context of sparse linear regression, or by Koltchinskii et al. (2011) for matrix completion. This modified version of the SGD algorithm (Ma and Needell, 2018) is conjectured to converge in expectation to the ordinary least squares estimator, achieving the rate of $\mathcal{O}(\frac{\log n}{\mu n})$ at iteration n for the excess empirical risk, assumed to be μ-strongly convex in that work. However, their algorithm requires a step choice relying on the knowledge of the strong-convexity constant μ which is often intractable for large-scale settings. In a non-linear setting, Yi et al. (2019) also propose a heuristic to debiase zero-imputation in neural networks but their proposed algorithm comes with no guarantee of convergence.

Besides, the inverse probability weighting method (IPW) consists in keeping *only* complete observations and on reducing the induced bias by reweighting the loss w.r.t. the complete observations with their probabilities of completeness (Little and Rubin, 2019; Seaman and White, 2013). However, in the IPW literature, weighting is often used to rebalance samples with missing outcome but not in cases where there may be missing values in all covariates, which would imply more complex debiasing expression than simply weighting the data.

Contributions.

- We develop a debiased averaged SGD to perform (regularized) linear regression either streaming or with finite samples, when covariates are missing. The approach consists in imputing the covariates with a simple imputation and using debiased gradients accordingly.
- Furthermore, the design is allowed to be contaminated by heterogeneous missing values: each covariate may have a different probability to be missing. This encompasses

the classical homogeneous Missing Completely At Random (MCAR) case, where the missingness is independent of any covariate value.

- This algorithm comes with theoretical guarantees: we establish convergence in terms of generalization risk at the rate $1/n$ at iteration n. This rate is remarkable as it is (i) optimal w.r.t. n, (ii) free from any bad condition number (no strong convexity constant is required), and (iii) similar to the rate of averaged SGD without any missing value. The same convergence rate is also obtained when the probabilities that variables are missing are not known but estimated.

- In terms of performance with respect to the missing entries proportion in large dimension, our strategy results in an error provably several orders of magnitude smaller than the best possible algorithm that would only rely on complete observations.

- We show the relevance of the proposed approach and its convergence behavior on numerical applications and its efficiency on real data; including the TraumaBase® dataset to assist doctors in making real-time decisions in the management of severely traumatized patients. The code to reproduce all the simulations and numerical experiments is available on https://github.com/AudeSportisse/SGD-NA.

4.2 Problem setting

In this paper, we consider either the streaming setting, i.e. when the data comes in as it goes along, or the finite-sample setting, i.e. when the data size is fixed and form a finite design matrix $X = (X_{1:}|\dots|X_{n:})^T \in \mathbb{R}^{n\times d}$ $(n > d)$. We define $\mathcal{D}_n := \sigma\left((X_{i:}, y_i), i = 1, \dots, n\right)$ the $\sigma-$field generated by n observations. We also denote $\preccurlyeq$ the partial order between self-adjoint operators, such that $A \preccurlyeq B$ if $B - A$ is positive semi-definite.

Given observations as in (4.1) and defining $f_i(\beta) := \left(\langle X_{i:}, \beta\rangle - y_i\right)^2/2$, the (unknown) linear model parameter satisfies:

$$\beta^\star = \arg\min_{\beta\in\mathbb{R}^d} \left\{R(\beta) := \mathbb{E}_{(X_{i:},y_i)}\left[f_i(\beta)\right]\right\}, \tag{4.2}$$

where $\mathbb{E}_{(X_{i:},y_i)}$ denotes the expectation over the distribution of $(X_{i:}, y_i)$ (which is independent of i as the observations are assumed to be i.i.d.).

In this work, the covariates are assumed to contain missing values, so one in fact observes $X_{i:}^{\mathrm{NA}} \in (\mathbb{R} \cup \{\mathtt{NA}\})^d$ instead of $X_{i:}$, as $X_{i:}^{\mathrm{NA}} := X_{i:} \odot D_{i:} + \mathtt{NA}(\mathbf{1}_d - D_{i:})$, where $\odot$ denotes the element-wise product, $\mathbf{1}_d \in \mathbb{R}^d$ is the vector filled with ones and $D_{i:} \in \{0,1\}^d$ is a binary vector mask coding for the presence of missing entries in $X_{i:}$, i.e. $D_{ij} = 0$ if the (i,j)-entry is missing in $X_{i:}$, and $D_{ij} = 1$ otherwise. We adopt the convention $\mathtt{NA} \times 0 = 0$ and $\mathtt{NA} \times 1 = \mathtt{NA}$. We consider a *heterogeneous* MCAR setting, i.e. D is modeled with a Bernoulli distribution

$$D = (\delta_{ij})_{1\leqslant i\leqslant n, 1\leqslant j\leqslant d} \quad \text{with} \quad \delta_{ij} \sim \mathcal{B}(p_j), \tag{4.3}$$

with $1 - p_j$ the probability that the j-th covariate is missing.

The considered approach consists in imputing the incomplete covariates by zero in $X_{i:}^{\mathrm{NA}}$, as $\tilde{X}_{i:} = X_{i:}^{\mathrm{NA}} \odot D_{i:} = X_{i:} \odot D_{i:}$, and in accounting for the imputation error in the subsequent algorithm.

4.3 Averaged SGD with missing values

The proposed method is detailed in Algorithm 2. The impact of the naive imputation by 0 directly translates into a bias in the gradient. Consequently, at each iteration we use a debiased estimate $\tilde{g}_k$. In order to stabilize the stochastic algorithm, we consider the Polyak-Ruppert Polyak and Juditsky (1992) averaged iterates $\bar{\beta}_k = \frac{1}{k+1}\sum_{i=0}^{k} \beta_i$.

Lemma 2. *Let $(\mathcal{F}_k)_{k\geqslant 0}$ be the following σ-algebra, $\mathcal{F}_k = \sigma(X_{1:}, y_1, D_{1:} \ldots, X_{k:}, y_k, D_{k:})$. The modified gradient $\tilde{g}_k(\beta_{k-1})$ in Equation (4.4) is $\mathcal{F}_k$-measurable and a.s.,*

$$\mathbb{E}\left[\tilde{g}_k(\beta_{k-1}) \,|\, \mathcal{F}_{k-1}\right] = \nabla R(\beta_{k-1}).$$

Algorithm 2 Averaged SGD for Heterogeneous Missing Data

Input: data $\tilde{X}, y, \alpha$ (step size)
Initialize $\beta_0 = 0_d$.
Set $P = \mathrm{diag}\left((p_j)_{j\in\{1,\ldots,d\}}\right) \in \mathbb{R}^{d\times d}$.
for $k = 1$ **to** n **do**

$$\tilde{g}_k(\beta_k) = P^{-1}\tilde{X}_{k:}\left(\tilde{X}_{k:}^T P^{-1}\beta_k - y_k\right) - (\mathrm{I} - P)P^{-2}\mathrm{diag}\left(\tilde{X}_{k:}\tilde{X}_{k:}^T\right)\beta_k \tag{4.4}$$

$\beta_k = \beta_{k-1} - \alpha\tilde{g}_k(\beta_{k-1})$
$\bar{\beta}_k = \frac{1}{k+1}\sum_{i=0}^{k}\beta_i = \frac{k}{k+1}\bar{\beta}_{k-1} + \frac{1}{k+1}\beta_k$
end for

Lemma 2 is proved in Section D.2.1. Note that in the case of homogeneous MCAR data, i.e. $p_1 = \ldots = p_d = p \in (0,1)$, the chosen direction at iteration k in Equation (4.4) boils down to $\frac{1}{p}\tilde{X}_{k:}\left(\frac{1}{p}\tilde{X}_{k:}^T\beta_k - y_k\right) - \frac{1-p}{p^2}\mathrm{diag}\left(\tilde{X}_{k:}\tilde{X}_{k:}^T\right)\beta_k$, where $\mathrm{diag}(A) \in \mathbb{R}^{d\times d}$ denotes the diagonal matrix containing either the diagonal of A if $A \in \mathbb{R}^{d\times d}$ or the vector A if $A \in \mathbb{R}^d$. This meets the classical debiasing terms of covariance matrices Loh and Wainwright (2011); Ma and Needell (2018); Koltchinskii et al. (2011) . Note also that in the presence of complete observations, meaning that $p = 1$, Algorithm 2 matches the standard least squares stochastic algorithm.

Remark 15 (Ridge regularization)**.** *Instead of minimizing the theoretical risk as in (4.2), we can consider a Ridge regularized formulation:* $\min_{\beta\in\mathbb{R}^d} R(\beta) + \lambda\|\beta\|^2$, *with* $\lambda > 0$. *Algorithm 2 is trivially extended to this framework: the debiasing term is not modified since the penalization term does not involve the incomplete data $\tilde{X}_{i:}$. This is useful in practice as no implementation is available for incomplete ridge regression.*

Remark 16 (Towards a more general MCAR setting)**.** *Note that we consider a specific MCAR setting in Equation (4.3) in which the missing-data patterns were independent ($D_{\cdot j} \perp\!\!\!\perp D_{\cdot j'}, j \neq j'$). However, an extended MCAR setting could allow coordinates of the missing mask to be dependently missing. In such a case, we propose a new way of constructing debiased versions of gradients, as* $\tilde{g}_k(\beta) := (W \odot (\tilde{X}_{k:}\tilde{X}_{k:}^T))\beta - y_k P^{-1}\tilde{X}_{k:}$ *with* $W \in \mathbb{R}^{d\times d}$*, and* $W_{ij} := 1/\mathbb{E}[\delta_{ki}\delta_{kj}]$ *for* $1 \leqslant i, j \leqslant d$*. Regarding practical implementation, the matrix W can be estimated, in particular using low-rank strategies on the missing pattern matrix.*

4.4 Theoretical results

In this section, we prove convergence guarantees for Algorithm 2 in terms of theoretical excess risk, in both the streaming and the finite-sample settings. For the rest of this section, assume the following.

- The observations $(X_{k:}, y_k) \in \mathbb{R}^d \times \mathbb{R}$ are independent and identically distributed.
- $\mathbb{E}[\|X_{k:}\|^2]$ and $\mathbb{E}[\|y_k\|^2]$ are finite.
- Let H be an invertible matrix, defined by $H := \mathbb{E}_{(X_{k:},y_k)}[X_{k:}X_{k:}^T]$.

The main technical challenge to overcome is proving that the noise in play due to missing values is *strutured* and still allows to derive convergence results for a debiased version of averaged SGD. This work builds upon the analysis made by Bach and Moulines (2013) for standard SGD strategies.

4.4.1 Technical results

Bach and Moulines (2013) proved that for least-squares regression, averaged SGD converges at rate n^{-1} after n iterations. In order to derive similar results, we prove in addition to Lemma 2, Lemmas 3 and 4:

- Lemma 3 shows that the noise induced by the imputation by zeros and the subsequent transformation results is a *structured noise*. This is the most challenging part technically: having a structured noise is fundamental to obtain convergence rates scaling as n^{-1} – in the unstructured case the convergence speed is only $n^{-1/2}$ (Dieuleveut et al., 2017).
- Lemma 4 shows that the adjusted random gradients $\tilde{g}_k(\beta)$ are almost surely *co-coercive* (Zhu and Marcotte, 1996) i.e., for any k, there exists a random "primitive" function $\tilde{f}_k$ which is a.s. convex and smooth, and such that $\tilde{g}_k = \nabla\tilde{f}_k$. Proving that $\tilde{f}_k$ is a.s. convex is an important step which was missing in the analysis of Ma and Needell (2018).

Lemma 3. *The additive noise process $(\tilde{g}_k(\beta^\star))_k$ with $\beta^\star$ defined in (4.2) is $\mathcal{F}_k-$measurable and,*

1. $\forall k \geqslant 0$, $\mathbb{E}[\tilde{g}_k(\beta^\star) \,|\, \mathcal{F}_{k-1}] = 0$ *a.s.*.

2. $\forall k \geqslant 0$, $\mathbb{E}[\|\tilde{g}_k(\beta^\star)\|^2 \mid \mathcal{F}_{k-1}]$ *is a.s. finite.*

3. $\forall k \geqslant 0$, $\mathbb{E}[\tilde{g}_k(\beta^\star)\tilde{g}_k(\beta^\star)^T] \preccurlyeq C(\beta^\star) = c(\beta^\star)H$, *with*

$$c(\beta^\star) = \frac{\mathrm{Var}(\epsilon_k)}{p_m^2} + \left(\frac{(2+5p_m)(1-p_m)}{p_m^3}\right)\gamma^2\|\beta^\star\|^2. \tag{4.5}$$

Sketch of proof (Lemma 3). Property 1 easily followed from Lemma 2 and the definition of $\beta^\star$. Property 2 can be obtained with similar computations as in (Ma and Needell, 2018, Lemma 4). Property 3 cannot be directly derived from Property 2, since $\tilde{g}_k(\beta^\star)\tilde{g}_k(\beta^\star)^T \preccurlyeq \|\tilde{g}_k(\beta^\star)\|^2 I$ leads to an insufficient upper bound. Proof relies on decomposing the external product $\tilde{g}_k(\beta^\star)\tilde{g}_k(\beta^\star)^T$ in several terms and obtaining the control of each, involving technical computations. □

Lemma 4. *For all $k \geqslant 0$, given the binary mask D, the adjusted gradient $\tilde{g}_k(\beta)$ is a.s. $L_{k,D}$-Lipschitz continuous, i.e. for all $u, v \in \mathbb{R}^d$, $\|\tilde{g}_k(u) - \tilde{g}_k(v)\| \leqslant L_{k,D}\|u - v\|$ a.s.. Set*

$$L := \sup_{k,D} L_{k,D} \leqslant \frac{1}{p_m^2} \max_k \|X_{k:}\|^2 \text{ a.s.}. \tag{4.6}$$

In addition, for all $k \geqslant 0$, $\tilde{g}_k(\beta)$ is almost surely co-coercive.

Lemmas 3 and 4 are respectively proved in Sections D.2.2 and D.2.3, and can be combined with Theorem 1 in Bach and Moulines (2013) in order to prove the following theoretical guarantees for Algorithm 2.

4.4.2 Convergence results

The following theorem quantifies the convergence rate of Algorithm 2 in terms of excess risk.

Theorem 17 (Streaming setting)**.** *Assume that for any i, $\|X_{i:}\| \leqslant \gamma$ almost surely for some $\gamma > 0$. For any constant step-size $\alpha \leqslant \frac{1}{2L}$, Algorithm 2 ensures that, for any $k \geqslant 0$:*

$$\mathbb{E}\left[R\left(\bar{\beta}_k\right) - R(\beta^\star)\right] \leqslant \frac{1}{2k}\left(\frac{\sqrt{c(\beta^\star)d}}{1-\sqrt{\alpha L}} + \frac{\|\beta_0 - \beta^\star\|}{\sqrt{\alpha}}\right)^2,$$

with L given in Equation (4.6), $p_m = \min_{j=1,\dots d} p_j$ and $c(\beta^\star)$ given in Equation (4.5).

Note that in Theorem 17, the expectation is taken over the randomness of the observations $(X_{i:}, y_i, D_{i:})_{1 \leqslant i \leqslant k}$. The bounded features assumption in Theorem 17 is mostly convenient for the readability, but it can be relaxed at the price of milder but more technical assumptions and proofs (typically bounds on quadratic mean instead of a.s. bounds, see e.g. Section 6.1. in Dieuleveut et al. (2020)).

Remark 18 (Finite-sample setting)**.** *Similar results as Theorem 17 can be derived in the case of finite-sample setting. For the sake of clarity, they are made explicit hereafter: for any constant step-size* $\alpha \leqslant \frac{1}{2L}$*, Algorithm 2 ensures that for any* $k \leqslant n$*:* $\mathbb{E}\left[R(\bar{\beta}_k) - R(\beta^\star)]|\mathcal{D}_n\right] \leqslant \frac{1}{2k}\left(\frac{\sqrt{c(\beta^\star)d}}{1-\sqrt{\alpha L}} + \frac{\|\beta_0-\beta^\star\|}{\sqrt{\alpha}}\right)^2$ *with* L *given in Equation (4.6) and* $c(\beta^\star) = \frac{\mathrm{Var}(\epsilon_k)}{p_m^2} + \left(\frac{(2+5p_m)(1-p_m)}{p_m^3}\right)\max_{1\leqslant i\leqslant n}\|X_{i:}\|^2\|\beta^\star\|^2$.

Remark 19 (Estimating missing probabilities $(\hat{p}_j)_j$)**.** *Algorithm 2 and the associated convergence rate established in Theorem 18 require the knowledge of the missing probabilities* $(p_j)_j$*. In practice, one could construct an estimator* $\bar{\hat{\beta}}_k$ *using our algorithm with estimated probabilities* $(\hat{p}_j)_j$*. In such a case, we can show that we preserve the convergence rate at* $1/k$*. More precisely, in the finite-sample setting, we can use the first half of the data to evaluate the* $(\hat{p}_j)_j$*'s and the second half of the data to build* $\bar{\hat{\beta}}_k$*. Under the additional assumptions of bounded iterates and strong convexity of the risk, the resulting supplementary risk w.r.t. the iterate* $\bar{\beta}_k$ *built with the true* $(p_j)_j$ *is* $\mathbb{E}[R(\bar{\hat{\beta}}_k) - R(\bar{\beta}_k)] = \mathcal{O}(1/kp_{\min}^6)$*. This is formalized in Theorem 1 of Appendix D.3, followed by its proof.*

Convergence rates for the iterates. Note that if a Ridge regularization is considered, the regularized function to minimize $R(\beta) + \lambda\|\beta\|^2$ is 2λ-strongly convex. Theorem 17 and Remark 18 then directly provide the following bound on the iterates: $\mathbb{E}\left[\left\|\overline{\beta}_k - \beta^\star\right\|^2\right] \leqslant \frac{1}{2\lambda k}\left(\frac{\sqrt{c(\beta^\star)d}}{1-\sqrt{\alpha L}} + \frac{\|\beta_0-\beta^\star\|}{\sqrt{\alpha}}\right)^2$.

Additional comments. We highlight the following points:

- In Theorem 17, the expected excess risk is upper bounded by (a) *a variance* term, that grows with the noise variance and is increased by the missing values, and (b) *a bias* term, that accounts for the importance of the initial distance between the starting point β_0 and the optimal one $\beta^\star$.
- The optimal convergence rate is achieved for a *constant* learning rate α. One could for example choose $\alpha = \frac{1}{2L}$, that does *not decrease* with the number of iterations. In such a situation, both the *bias* and *variance* terms scale as k^{-1}. Remark that convergence of the averaged SGD with constant step-size only happens for least squares regression, because the un-averaged iterates converge to a limit distribution whose mean is exactly β^* Bach and Moulines (2013); Dieuleveut et al. (2020).
- The expected risk scales as n^{-1} after n iterations, without strong convexity constant involved.
- For the generalization risk R, this rate of n^{-1} is known to be statistically optimal for least-squares regression: under reasonable assumptions, no algorithm, even more complex than averaged SGD or without missing observations, can have a better dependence in n Tsybakov (2003).

- In the complete case, i.e. when $p_1 = \ldots = p_d = 1$, Theorem 17 and Remark 18 meet the results from Bach and Moulines (2013, Theorem 1). Indeed, in such a case, $c(\beta^\star) = \mathrm{Var}(\epsilon_k)$.
- The noise variance coefficient $c(\beta^\star)$ includes (i) a first term as a classical noise one, proportional to the model variance, and increased by the missing values occurrence to $\frac{\mathrm{Var}(\epsilon_k)}{p_m^2}$; (ii) the second term is upper-bounded by $\frac{7(1-p_m)}{p_m^3} \cdot \gamma^2 \|\beta^\star\|^2$ corresponds to the multiplicative noise induced by the imputation by 0 and gradient debiasing. It naturally increases as the radius γ^2 of the observations increases (so does the imputation error), and vanishes if there are no missing values ($p_m = 1$).

Remark 20 (Only one epoch)**.** *It is important to notice that in a finite-sample setting, as covered by Remark 18, given a maximum number of n observations, our convergence rates are only valid for $k \leqslant n$: the theoretical bound holds only for one pass on the input/output pairs. Indeed, afterwards, we cannot build unbiased gradients of the risk.*

4.4.3 What about empirical risk minimization (ERM)?

Theoretical locks. Note that the translation of the results in Remark 18 in terms of empirical risk convergence is still an open issue. The heart of the problem is that it seems really difficult to obtain a sequence of unbiased gradients of the empirical risk.

- Indeed, to obtain unbiased gradients, the data should be processed *only once* in Algorithm 2: if we consider the gradient of the loss with respect to an observation k, we obviously need the binary mask D_k and the current point β_{k-1} to be independent for the correction relative to the missing entries to make sense. As a consequence, no sample can be used twice - in fact, running multiple passes over a finite sample could result in over-fitting the missing entries.
- Therefore, with a finite sample at hand, the sample used at each iteration should be chosen *without replacement* as the algorithm runs. But even in the complete data case, sampling without replacement induces a bias on the chosen direction Gürbüzbalaban et al. (2015); Jain et al. (2019). Consequently, Lemma 2 does not hold for the empirical risk instead of the theoretical one. This issue is not addressed in Ma and Needell (2018), unfortunately making the proof of their result invalid/wrong.

Comparison to Ma and Needell (2018). Leaving aside the last observation, we can still comment on the bounds in Ma and Needell (2018) for the empirical risk without averaging. As they do not use averaging but only the last iterate, their convergence rate (see Lemma 1 in their paper) is only studied for $\mu-$strongly convex problems and is expected to be larger (i) by a factor μ^{-1}, due to the choice of their decaying learning rate, and (ii) by a $\log n$ factor due to using the last iterate and not the averaged one (Shamir and Zhang, 2013). Moreover, the strategy of the present paper does not require to access the strong convexity constant, which is generally out of reach, if no explicit regularization is used. More marginally, we provide the proof of the co-coercivity of the adjusted gradients (Lemma 4), which is required

to derive the convergence results, and which was also missing in Ma and Needell (2018). A more detailed discussion on the differences between the two papers is given in Section D.1.

ERM hindered by NA. It is also interesting to point out that with missing features, *neither* the generalization risk R, *nor* the empirical risk R_n are observed (i.e., only approximations of their values or gradients can be computed). As a consequence, one cannot expect to minimize those functions with unlimited accuracy. This stands in contrast to the *complete observations setting*, in which the empirical risk R_n is known exactly. As a consequence, with missing data, empirical risk loses its main asset - being an observable function that one can minimize with high precision. Overall it is both more natural and easier to focus on the generalization risk.

4.4.4 On the impact of missing values

Marginal values of incomplete data. An important question in practice is to understand how much information has been lost because of the incompleteness of the observations. In other words, it is better to access 200 input/output pairs with a probability 50% of observing each feature on the inputs, or to observe 100 input/output pairs with complete observations?

Without missing observations, the variance bound in the expected excess risk is given by Theorem 17 with $p_m = 1$: it scales as $\mathcal{O}\left(\frac{\mathrm{Var}(\epsilon_k)d}{k}\right)$, while with missing observations it increases to $\mathcal{O}\left(\frac{\mathrm{Var}(\epsilon_k)d}{kp_m^2} + \frac{C(X,\beta^\star)}{kp_m^3}\right)$. As a consequence, the variance upper bound is larger by a factor p_m^{-1} for the estimator derived from k incomplete observations than for $k \times p_m$ complete observations. This suggests that there is a higher gain to collecting fewer complete observations (e.g., 100) than more incomplete ones (e.g., 200 with $p = 0.5$). However, one should keep in mind that this observation is made by comparing upper bounds thus does not necessarily reflect what would happen in practice.

Keeping only complete observations? Another approach to solve the missing data problem is to discard all observations that have at least one missing feature. The probability that one input is complete, under our missing data model is $\prod_{j=1}^d p_j$. In the homogeneous case, the number of complete observations k_{co} out of a $k-$sample thus follows a binomial law $k_{co} \sim \mathcal{B}(k, p^d)$. With only those few observations, the statistical lower bound is $\mathrm{Var}(\epsilon_k)d/k_{co}$. In expectation, by Jensen inequality, we get that the lower bound on the risk is larger than $\mathrm{Var}(\epsilon_k)d/kp^d$.

Our strategy thus leads to an *upper-bound* which is typically p^{d-3} times smaller than the *lower bound* on the error of any algorithm relying only on complete observations. For a large dimension or a high percentage of missing values, our strategy is thus provably several orders of magnitude smaller than the best possible algorithm that would only rely on complete observations - e.g., if $p = 0.9$ and $d = 40$, the error of our method is at least 50 times smaller.

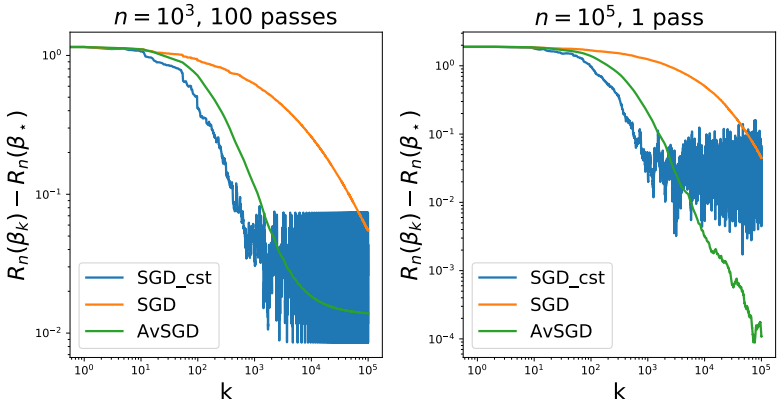


Figure 4.1: Empirical excess risk $(R_n(\beta_k) - R_n(\beta^\star))$. Left: $n = 10^3$ and 100 passes. Right: $n = 10^5$ and 1 pass. $d = 10$, 30% MCAR data. L is assumed to be known in both graphics.

Also note that in Theorem 1 and Lemma 1 in Ma and Needell (2018), the convergence rate with missing observations suffers from a similar multiplicative factor $\mathcal{O}(p^{-2} + \kappa p^{-3})$.

4.5 Experiments

4.5.1 Synthetic data

Consider the following simulation setting: the covariates are normally distributed, $X_{i:} \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$, where Σ is constructed using uniform random orthogonal eigenvectors and decreasing eigenvalues $1/k$, $k = 1, \ldots, d$. For a fixed parameter vector β, the outputs y_i are generated according to the linear model (4.1), with $\epsilon_i \sim \mathcal{N}(0, 1)$. Setting $d = 10$, we introduce 30% of missing values either with a uniform probability p of missingness for any feature, or with probability p_j for covariate j, with $j = 1, \ldots, d$. Firstly, the three following algorithms are implemented:

(1) **AvSGD** described in Algorithm 2 with a constant step size $\alpha = \frac{1}{2L}$, and L given in (4.6).

(2) **SGD** from (Ma and Needell, 2018) with iterates $\beta_{k+1} = \beta_k - \alpha_k \tilde{g}_{i_k}(\beta_k)$, and decreasing step size $\alpha_k = \frac{1}{\sqrt{k+1}}$.

(3) **SGD_cst** from (Ma and Needell, 2018) with a constant step size $\alpha = \frac{1}{2L}$, where L is given by (4.6).

Debiased averaged vs. standard SGD. Figure 4.1 compares the convergence of Algorithms (1), (2) and (3), with either multiple passes or one pass, in terms of excess empirical risk $R_n(\beta) - R_n(\beta^\star)$, with $R_n(\beta) := \frac{1}{n}\sum_{i=1}^n f_i(\beta)$. As expected (see Remark 20

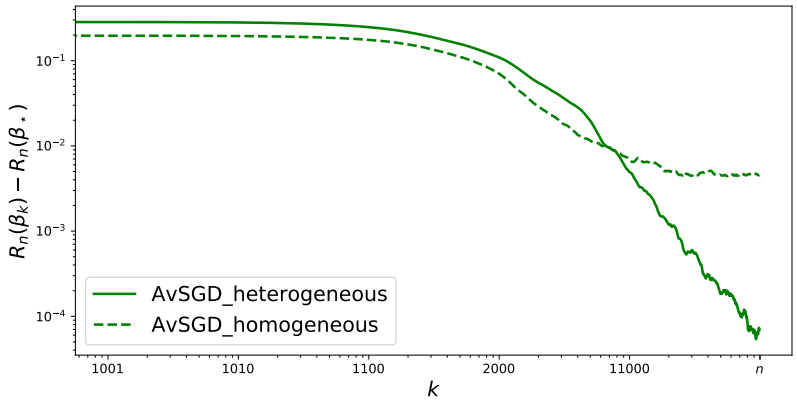


Figure 4.2: Empirical excess risk $R_n(\beta_k) - R_n(\beta^\star)$ for synthetic data where $n = 10^5$, $d = 10$ and with heterogeneous missing values either taking into account the heterogeneity (plain line) in the algorithm or not (dashed line).

and section 4.4.3), multiple passes can lead to saturation: after one pass on the observations, AvSGD does not improve anymore (Figure 4.1, left), while it keeps decreasing in the streaming setting (Figure 4.1, right). Looking at Figure 4.1 (right), one may notice that without averaging and with decaying step-size, Algorithm (2) achieves the convergence rate $\mathcal{O}\left(\sqrt{\frac{1}{n}}\right)$, whereas with constant step-size, Algorithm (3) saturates at an excess risk proportional to α after $n = 10^3$ iterations. As theoretically expected, both methods are improved with averaging. Indeed, Algorithm 2 converges pointwise with a rate of $\mathcal{O}(\frac{1}{n})$.

About the algorithm hyperparameter. Note that the Lipschitz constant L given in (4.6) can be either computed from the complete covariates, or estimated from the incomplete data, see discussion and numerical experiments in Section D.4.

Heterogeneous vs. homogeneous missingness. In Figure 4.2, the missing values are introduced with different missingness probabilities, i.e. with distinct $(p_j)_{1 \leqslant j \leqslant d}$ per feature, as described in Equation (4.3). When taking into account this heterogeneousness, Algorithm 2 achieves the same convergence rates as in Figure 4.1. However, ignoring the heterogeneous probabilities in the gradient debiasing leads to stagnation far from the optimum in terms of empirical excess risk.

Increasing missing data proportions. Figure 4.3 shows the results of Algorithm 2 with different percentage of missing values (25%, 50% and 75%). The more missing data there are, the more the convergence rate deteriorates. This was expected, as the established theoretical upper bound for the convergence in Theorem 17 increases as the probability of

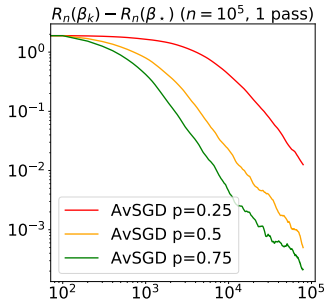


Figure 4.3: Empirical excess risk $R_n(\beta_k) - R_n(\beta^\star)$ for synthetic data where $n = 10^5$, $d = 10$ with 25% (green), 50% (orange) and 75% (red) missing values.

being observed gets smaller.

Polynomial features. Algorithm 2 can be adapted to handle missing polynomial features, see Section D.5 for a detailed discussion and numerical experiments on synthetic data.

4.5.2 Real dataset 1: Traumabase® dataset

We illustrate our approach on a public health application with the APHP TraumaBase® Group (Assistance Publique - Hopitaux de Paris) on the management of traumatized patients. Our aim is to model the level of platelet upon arrival at the hospital from the clinical data of 15785 patients. The platelet is a cellular agent responsible for clot formation and it is essential to control its levels to prevent blood loss and to decide on the most suitable treatment. A better understanding of the impact of the different features is key to trauma management. Explanatory variables for the level of platelet consist in seven quantitative (missing) variables, which have been selected by doctors. In Figure 4.4, one can see the percentage of missing values in each variable, varying from 0 to 16%, see Section D.6 for more information on the data.

Model estimation. The model parameter estimation is performed either using the AvSGD Algorithm 2 or an Expectation Maximization (EM) algorithm Dempster et al. (1977). Both methods are compared with the ordinary least squares linear regression in the complete case, i.e. keeping the fully-observed rows only (i.e. 9448 rows). The signs of the coefficients for Algorithm 2 are shown in Figure 4.4.

According to the doctors, a negative effect of shock index (*SI*), vascular filling (*VE*), blood transfusion (*RBC*) and lactate (*Lactacte*) was expected, as they all result in low platelet levels and therefore a higher risk of severe bleeding. However, the effects of delta Hemocue (*Delta.Hemocue*) and the heart rate (*HR*) on platelets are not entirely in agreement with their opinion. Note that using the linear regression in the complete case and the EM algorithm lead to the same sign for the variables effects as presented in Figure 4.4.

Variable	Effect	NA %
Lactate	−	16%
Δ.Hemo	+	16%
VE	−	9%
RBC	−	8%
SI	−	2%
HR	+	1%
Age	−	0%

Figure 4.4: Percentage of missing features, and effect of the variables on the platelet for the TraumaBase data when the AvSGD algorithm is used. "+" indicates positive effect while "−" negative.

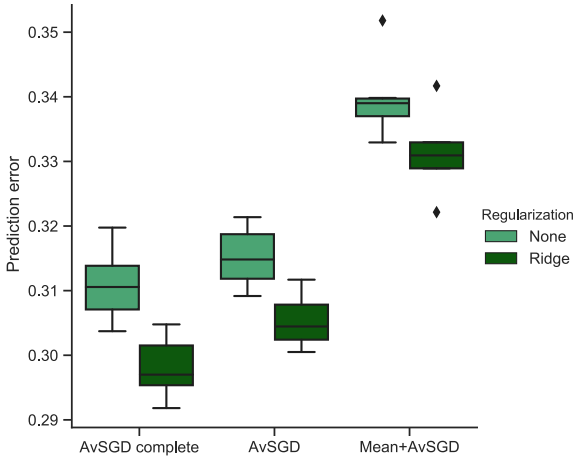


Figure 4.5: Prediction error boxplots (over 10 replications) for the Superconductivity data. AvSGD complete corresponds to applying the AvSGD on the complete data, AvSGD and Mean+AvSGD use the predictions obtained with the estimated parameters $\hat{\beta}_n^{\text{AvSGD}}$ and $\bar{\beta}_n^{\text{AvSGD}}$ respectively.

4.5.3 Real dataset 2: Superconductivity dataset

We now consider the Superconductivity dataset (available here), which contains 81 quantitative features from 21263 superconductors. The goal here is to predict the critical temperature of each superconductor. Since the dataset is initially complete, we introduce 30% of missing values with probabilities $(p_j)_{1\leqslant j\leqslant 81}$ for the covariate j, with p_j varying between 0.7 and 1. The results are shown in Figure 4.5 where a Ridge regularization has been added or not. The regularization parameter λ (see Remark 15) is chosen by cross validation.

Prediction performance. The dataset is divided into training and test sets (random selection of $70-30\%$). The test set does not contain missing values. In order to predict the critical temperature of each superconductor, we compute $\hat{y}_{n+1} = X_{n+1}^T\hat{\beta}$ with $\hat{\beta} = \beta_n^{\text{AvSGD}}$ or β_n^{EM}. We also impute the missing data naively by the mean in the training set, and apply the averaged stochastic gradient without missing data on this imputed dataset, giving a coefficient model $\bar{\beta}_n^{\text{AvSGD}}$. It corresponds to the case where the bias of the imputation has not been corrected. The prediction quality on the test set is compared according to the relative ℓ_2 prediction error, $\|\hat{y}-y\|^2/\|y\|^2$. The data is scaled, so that the naive prediction by the mean of the outcome variable leads to a prediction error equal to 1. In Figure 4.5, we observe that the SGD strategies give quite good prediction performances. The EM algorithm is not represented since it is completely out of range (the mean of its prediction error is 0.7), which indicates that it struggles with a large number of covariates. Note also that the EM algorithm requires a distributional assumption on the covariates, which is not the case of our method. As for the AvSGD Algorithm, it performs well in this setting. Indeed, with or without regularization, the prediction error with missing values is very close to the one obtained from the complete dataset. Algorithm 2 is shown to handle missing polynomial features well even in higher dimensions, see Section D.5 for a detailed discussion and large-scale experiments on the superconductivity dataset.

Comparison to other methods. For completeness, we ran the proposed algorithm on the superconductivity dataset and compare it to two-step heuristics in which first, the covariates are imputed (by the mean or by the ICE[1]iterative imputer that estimates each feature from all the others) and then linear regression (LR) is performed on the completed data. The coefficient of determination R^2 is plotted on Figure 4.6 (thus higher is better) for the Superconductivity dataset with 60% of missing values. Our method greatly outperforms all other methods, and follows closely the linear regression performed on the initial complete data. One should note that the two-step heuristics considered here come with no theoretical guarantee at all.

[1] `sklearn.impute.IterativeImputer`

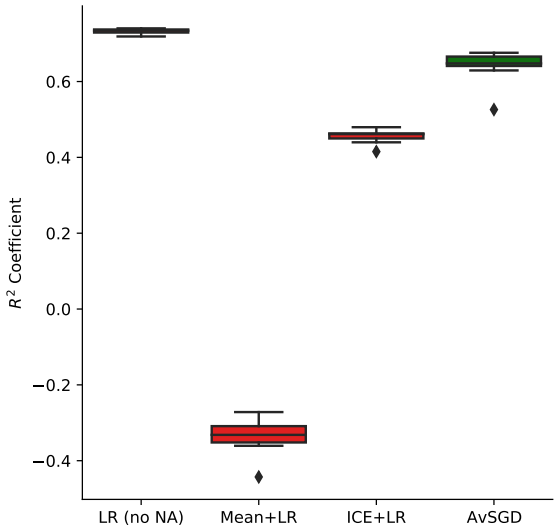


Figure 4.6: R^2 coefficients for the Superconductivity data with 60% MCAR values.

4.6 Discussion

In this work, we thoroughly study the impact of missing values for Stochastic Gradient Descent algorithm for Least Squares Regression. We leverage both the power of averaging and a simple and powerful debiasing approach to derive tight and rigorous convergence guarantees for the generalization risk of the algorithm. The theoretical study directly translates into practical recommendations for the users and a byproduct is the availability of a python implementation of regularized regression with missing values for large scale data, which was not available. Even though we have knocked down some barriers, there are still exciting perspectives to be explored as the robustness of the approach to rarely-occurring covariates, or dealing with more general loss functions as well - for which it is challenging to build a debiased gradient estimator from observations with missing values, or also considering more complex missing-data patterns such as missing-not-at-random mechanisms.

Funding disclosure

The work was partly supported by the chaire SCAI (ANR-19-CHIA-0002-0).

Chapter 5

Model-based clustering with MNAR data

This chapter is an ongoing work, realised in collaboration with Christophe Biernacki, Claire Boyer, Gilles Celeux, Julie Josse, Fabien Laporte and Matthieu Marbac-Lourdelle.

Contents

5.1 Introduction

Clustering remains a pivotal tool for readable analysis of large datasets, offering a consistent summary of datasets by grouping individuals. In particular, the model-based paradigm (McLachlan and Basford, 1988; Zhong and Ghosh, 2003; Bouveyron et al., 2019) allows to perform clustering, by providing interpretable models, valuable to understand the connections between the constructed clusters and the features in play. This parametric framework provides a certain plasticity by handling high dimensionality problems (Bouveyron et al., 2007; Bouveyron and Brunet-Saumard, 2014), mixed datasets (Marbac et al., 2017), or even time series and dependent data (Ramoni et al., 2002; Xiong and Yeung, 2004). The counterpart of performing this multifaceted model-based clustering is the involved modelling work for designing mixture models appropriate to the data structure.

In large scale data analysis, the problem of missing data is ubiquitous, since the more data we have, the more missing values we have. Classical approaches for dealing with missing data consist of working on a complete dataset (Little and Rubin, 2019), either by using only complete individuals, or by imputing missing values. Both methods can raise huge problems in the analysis. On the one hand, if we delete the missing values, the remaining observations can form a too small subset or a biased subset of the population, which increases the variance of the estimates. On the other hand, the imputation often leads to the overestimation of the correlation between the variables and the model variance is underestimated. Moreover, neither of both strategies are designed for the final clustering task. Thus, it is desirable to develop some clustering methods able to deal with missing data in an efficient way.

Notations and typology of the missing values mechanisms To define the missing values mechanisms correctly, some notations must be introduced. The full dataset consists of n individuals $Y = (y_1|\dots|y_n)^T$, where each observation $y_i = (y_{i1}, \dots, y_{id})^T$ belongs to a space $\mathcal{Y}$, depending on the kind of data, defined by d features. The pattern of missing data for the full dataset is denoted by $C = (c_1|\dots|c_n)^T \in \{0,1\}^{n\times d}$, $c_i = (c_{i1}, \dots, c_{id})^T \in \{0,1\}^d$ being the indicator pattern of missing data for the individual $i \in \{1, \dots, n\}$: $c_{ij} = 1$ indicates that the value y_{ij} is missing and $c_{ij} = 0$ otherwise. The observed variables values for individual i will be denoted by y_i^{obs}. Similarly the missing variables values for individual i is denoted by y_i^{mis}. In addition, in a clustering context, the target is to estimate an unknown partition of the whole dataset Y into K groups. This partition is denoted by $Z = (z_1|\dots|z_n)^T \in \{0,1\}^{n\times K}$ with $z_i = (z_{i1}, \dots, z_{iK})^T \in \{0,1\}^K$ and where $z_{ik} = 1$ if y_i belongs to cluster k, $z_{ik} = 0$ otherwise. Consequently, in a clustering context, the missing data are not only the values y_i^{mis} but also the partition labels z_i.

Rubin (1976) distinguish three missing values mechanisms, namely Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing not at Random (MNAR). The missing data are MCAR when the missingness is independent of all the values, missing or not, and thus can be formalized by $\mathbb{P}(c_i \mid y_i, z_i; \psi) = \mathbb{P}(c_i; \psi)$, for all (missing or observed) values (y_i, z_i), ψ generically designating a parameter of the multinomial pdf on c_i. The missing data are MAR when the missingness is independent of the missing values, even if possibly depending on some (or all) observed values, meaning that $\mathbb{P}(c_i \mid y_i, z_i; \psi) = \mathbb{P}(c_i \mid y_i^{\text{obs}}; \psi)$

for all missing values (y_i^{mis}, z_i). The M(C)AR mechanisms are said ignorable, because the inference does not require the modelisation of $\mathbb{P}(c_i \mid y_i, z_i; \psi)$. Finally, MNAR corresponds to a missing-data mechanism which is not MCAR or MAR. For such missing data, the observed variables are not representative of the population. It is well known that the MNAR mechanism is not-ignorable when the goal is to estimate the parameters of the mixture model Little and Rubin (2019). The MNAR mechanism is actually also not-ignorable when the aim is to recover the partition of the data. Therefore, as the MNAR mechanism is neither ignorable for the density, nor for the clustering, dealing with such data does require the specific modeling effort of $\mathbb{P}(c_i \mid y_i, z_i; \psi)$.

MNAR data In this paper, the data are supposed to be MNAR which is very frequent in practice (Ibrahim et al., 2001; Mohan et al., 2018). Examples may include surveys where rich people would be less willing to disclose their income or clinical data collected in emergency situations, where doctors may choose to treat patients before measuring heart rate. In both cases, the missingness of income or heart rate depends on the missing values themself.

The missing-data mechanism must be generally taken into account (Little and Rubin, 2019) by considering the joint distribution of the data and the missing-data pattern. There are mainly two approaches to formulate the joint distribution of the data and the missing-data pattern: (i) the selection model (Heckman, 1979) which factorizes it into the product of the marginal data density and the conditional density of the missing-data pattern given the data i.e. $\mathbb{P}(y_i, c_i|z_i) = \mathbb{P}(y_i|z_i)\mathbb{P}(c_i|y_i, z_i)$ (ii) the pattern-mixture model (Little, 1993) which uses the product of the marginal density of the missing-data pattern and the conditional density of the data given the missing-data pattern i.e. $\mathbb{P}(y_i, c_i|z_i) = \mathbb{P}(c_i|z_i)\mathbb{P}(y_i|c_i, z_i)$. In this paper, we adopt the selection model strategy, as it is more intuitive to model the distribution of the data (as usually done in parametric clustering approaches) and the cause of the lack according to the data. Although this point of view requires to model the missing-data mechanism, it allows to estimate the parameters of the model-based clustering and the data density and possibly to impute missing values, which are out of reach in pattern-mixture models.

Related works In order to handle missing values in a model-based clustering framework, Hunt and Jorgensen (2003) have implemented the standard EM algorithm (Dempster et al., 1977) based on the observed likelihood. More recently, Serafini et al. (2020) also propose an EM algorithm to estimate Gaussian mixture models in the presence of missing values by performing multiple imputations (with Monte Carlo methods) in the E-step. However, both works only consider M(C)AR data.

In a partition-based framework, Chi et al. (2016) propose an extension of k-means clustering for missing data, called k-Pod, without requiring the missing-data pattern to be modelled, making it suitable for MNAR data. However, like k-means clustering, the k-Pod algorithm cannot identify difficult cluster structures, since it relies on strong assumptions as equal proportions between the clusters. De Chaumaray and Marbac (2020) have proposed to perform clustering via a mixture model using the pattern-mixture model to formulate the joint distribution, which makes the method not suitable to estimate the density parameters

or to impute missing values. For longitudinal data, some authors (Beunckens et al., 2008; Kuha et al., 2018) jointly model the measurements and the dropout process by using an extension of the shared-parameter model, which is an other MNAR model assuming that both the data and the dropout process depend on shared latent variables. They introduce for this a latent-class mixture model allowing classification of the subjects into latent groups. However, the MNAR model is restricted to the case where the missingness may depend on the latent variables but not on the missing variables themselves.

For MNAR data, and specifically in selection models, the main challenge to overcome consists of proving the identifiability of the parameters of both the data and the missing-data pattern distributions. In particular, Molenberghs et al. (2008) prove that the identifiability does not hold when the models are not fixed, i.e. when there is no prior information on the type of the distribution for the missing-data pattern. For fixed models, Miao et al. (2016) provide identifiability results of Gaussian mixture and t-mixture models with MNAR data. However, their identifiability results are restricted to specific missing scenarios in a univariate case (one variable) and no estimation strategy is proposed. In this paper, their idenfiability results are extended to more complex missing scenario and to the multivariate case.

Contributions. We present and illustrate a relevant inventory of distributions for the MNAR missingness process in the context of unsupervised classification based on mixture models. We then conduct an exhaustive study of the identifiability of the mixture model parameters (π, θ) and the missingness process parameters ψ, under certain conditions (including the data type and the link functions governing the missingness mechanism distribution). This is a real issue in the context of MNAR data, as models often lead to unidentifiable parameters. In the continuous case, all models lead to identifiable parameters. In the categorical case, only the models for which the missingness depends on the class membership only have identifiable parameters. For each model or sub-model, an EM or SEM algorithm is proposed, implemented, and made available for reproducibility. We also prove that concerning MNAR models for which the missingness depends on the class membership, the statistical inference can be conducted on the augmented matrix $[Y, C]$ considering the MAR mechanism instead; which is a real advantage, especially because the missing-data mechanism does not have to be modelled in this case. Preliminary numerical experiments assess the performances of the proposed algorithms for performing clustering with MNAR data.

The rest of the chapter is organized as follows. Section 5.2 introduces the model-based clustering in presence of missing-data. In Section 5.3, we propose an exhaustive zoology of the possible MNAR specifications in the model-based clustering framework and we discuss the different models. For each model, we address the identifiability issue in Section 5.4 and we propose an estimation strategy in Section 5.5. Section 5.6 is devoted to numerical experiments on synthetic data in order to assess the performances of our methods.

5.2 Missing data in model-based clustering

5.2.1 Mixture model as foundation

Model-based clustering relies on the assumption that $y_1, \ldots, y_n$ form an i.i.d. sample from some mixture distribution (see for instance (McLachlan and Basford, 1988))

$$f(y_i; \pi, \theta) = \sum_{k=1}^{K} \pi_k f_k(y_i; \theta_k), \tag{5.1}$$

where $\pi_k = \mathbb{P}(z_{ik} = 1)$ is the mixing proportion of the k-th component ($\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0$ for all $k \in \{1, \ldots, K\}$), $f_k(\cdot\,; \theta_k)$ is the pdf of the k-th component parameterized by θ_k. The mixture distribution is then fully parameterized by $\pi = (\pi_1, \ldots, \pi_K)$ and $\theta = (\theta_1, \ldots, \theta_K)$. Different kinds of distributions can be considered, depending on the types of features at hand.

- For continuous data, the space of each observation $(y_i)_{i=1,\ldots,n}$ is $\mathcal{Y} = \mathbb{R}^d$ and a current family for $f_k(\cdot; \theta_k)$ is the d-variate Gaussian pdf, often noted $\phi(\cdot; \theta_k)$, where $\theta_k = (\mu_k, \Sigma_k)$, μ_k being the mean vector and Σ_k being the covariance matrix (for Gaussian mixture, see for example (McLachlan and Basford, 1988; Banfield and Raftery, 1993)).

- For categorical data, one defines the space of each observation $(y_i)_{i=1,\ldots,n}$ as $\mathcal{Y} = \{0,1\}^{\ell_1} \times \ldots \times \{0,1\}^{\ell_d}$ where ℓ_j is the number of levels for the feature $j \in \{1, \ldots, d\}$. More precisely, if the j-th feature is categorical then it is one-hot encoded as follows $(y_{ij}^1, \ldots, y_{ij}^{\ell_j})$, where $y_{ij}^{\ell} = 1$ if the j-th feature of the i-th individual takes the level ℓ, 0 otherwise ($\ell \in \{1, \ldots, \ell_j\}$). In addition, one has $f_k(\cdot; \theta_k) = \prod_{j=1}^{d} f_{kj}(\cdot\,; \theta_{kj})$ where $\theta_k = (\theta_{k1}, \ldots, \theta_{kd})$ and where $f_{kj}(\cdot; \theta_{kj})$ is the multinomial distribution parameterized by the vector $\theta_{kj} = (\theta_{kj}^1, \ldots, \theta_{kj}^{\ell_j})$, with $\theta_{kj}^{\ell} = \mathbb{P}(y_{ij}^{\ell} = 1 \mid z_{ik} = 1)$ for $\ell \in \{1, \ldots, \ell_j\}$. Thus, we have $f_{kj}(y_{ij}; \theta_{kj}) = \prod_{\ell=1}^{\ell_j} (\theta_{kj}^{\ell})^{y_{ij}^{\ell}}$. The product on $j = 1, \ldots, d$ in the definition of f_k indicates that the features are independently drawn conditionally to the group membership, what is often referred as the latent class model (see (Geweke et al., 1994)).

- For a combination of continuous and categorical data (the so-called mixed case, see for example (Jorgensen and Hunt, 1996)), y_{ij} denotes either a continuous feature or a categorical one and adopts the corresponding notation related to its own type. In this case, it is often simply assumed that all the variables are conditionally independent knowing the group membership (McParland and Gormley, 2016). It means that group distributions are the product of univariate Gaussian and multinomial distributions. Consequently, the covariance matrix Σ_k associated the set of continuous features is restricted to a diagonal one. This apparently stringent assumption is made to ensure that the continuous and categorical variables are treated on a fair playing field.

5.2.2 Mixture parameter estimation with missing data

The mixture parameters (π, θ) and the parameter ψ of the missing-data mechanism have to be estimated from the observed data which consist of the observed vectors $(y_i^{\text{obs}})_{1 \leqslant i \leqslant n}$ and the patterns $(c_i)_{1 \leqslant i \leqslant n}$. The full observed model likelihood of the parameters (π, θ, ψ) for the datasets (Y, C) can be written as follows

$$L(\pi, \theta, \psi; Y^{\text{obs}}, C) = \prod_{i=1}^{n} \sum_{k=1}^{K} \int_{\mathcal{Y}_i^{\text{mis}}} \mathbb{P}(c_i \mid y_i, z_{ik} = 1; \psi) \mathbb{P}(y_i, z_{ik} = 1; \pi, \theta) dy_i \quad (5.2)$$

where $\mathcal{Y}_i^{\text{mis}} = \{\tilde{y}_i = (\tilde{y}_{i1}, \ldots, \tilde{y}_{id}) \in \mathcal{Y}_i : \tilde{y}_i^{\text{obs}} = y_i^{\text{obs}}\}$. Note that in the case of the mixture model (1.16), $\mathbb{P}(y_i, z_{ik} = 1; \pi, \theta) = \pi_k f_k(y_i; \theta_k)$.

In both the MCAR and MAR paradigms, this observed likelihood can be decomposed into the following two likelihoods:

$$\begin{aligned} L(\pi, \theta, \psi; Y^{\text{obs}}, C) &= \prod_{i=1}^{n} \mathbb{P}(c_i \mid y_i^{\text{obs}}; \psi) \times \sum_{k=1}^{K} \int_{\mathcal{Y}_i^{\text{mis}}} \mathbb{P}(y_i, z_{ik} = 1; \pi, \theta) dy_i \\ &= \prod_{i=1}^{n} L(\psi; c_i \mid y_i^{\text{obs}}) \times \prod_{i=1}^{n} L(\pi, \theta; y_i^{\text{obs}}). \end{aligned} \quad (5.3)$$

In such a situation, and also provided that parameters π, θ and ψ are functionally independent, the missing mechanism is said to be *ignorable*, meaning that estimating the mixture parameters π and θ are independent of any modeling of the missing-data pattern distribution $\mathbb{P}(c_i \mid y_i^{\text{obs}}; \psi)$. Consequently, estimating π and θ can be performed just by maximizing the (usual) observed partial likelihood $L(\pi, \theta; y_i^{\text{obs}})$ (Little and Rubin, 2019). Then, maximizing this likelihood can be performed with (usual) algorithms such that the EM or the SEM ones (Celeux et al., 1996; Nielsen et al., 2000) (see Section 5.5 for details).

General ignorability vs. ignorability for clustering A necessary and sufficient condition to have an ignorable missing process for clustering is that the distributions of c_i are equal among the mixture components. Thus, we said that the missingness process is *ignorable for clustering* if

$$\forall y_i, \quad r_k(y_i^{\text{obs}}) = t_k(y_i, c_i)$$

where

$$r_k(y_i^{\text{obs}}) = \frac{\pi_k \int_{\mathcal{Y}_i^{\text{mis}}} f_k(y_i; \theta_k) dy_i}{\int_{\mathcal{Y}_i^{\text{mis}}} \sum_{\ell=1}^{K} \pi_\ell f_\ell(y_i; \theta_\ell) dy_i}$$

and

$$t_k(y_i, c_i) = \frac{\pi_k \int_{\mathcal{Y}_i^{\text{mis}}} f_k(y_i; \theta_k) \mathbb{P}(c_i \mid y_i, z_{ik} = 1; \psi) dy_i}{\int_{\mathcal{Y}_i^{\text{mis}}} \sum_{\ell=1}^{K} \pi_\ell f_\ell(y_i; \theta_\ell) \mathbb{P}(c_i \mid y_i, z_{ik} = 1; \psi) dy_i}$$

However, under the MNAR assumption the missing mechanism is no longer ignorable, even for clustering, and a specific estimation process for the vector parameter (π, θ, ψ) is needed. Obviously, it depends on the MNAR model, namely the assumptions made on the missing-pattern distribution $\mathbb{P}(c_i \mid y_i, z_i; \psi)$.

5.3 Zoology of MNAR models in clustering

First, in a parsimonious perspective, we assume that the c_j's are independent conditionally on the complete dataset

$$\mathbb{P}(c_i \mid y_i, z_{ik} = 1; \psi) = \prod_{j=1}^{d} \mathbb{P}(c_{ij} \mid y_i, z_{ik} = 1; \psi). \tag{5.4}$$

A general MNAR mechanism for c_{ij} can be written as follows, by giving the probability of missingness for the variable j given the data y_i and the class membership $z_{ik} = 1$,

$$\mathbb{P}(c_{ij} = 1 \mid y_i, z_{ik} = 1; \psi) = \rho\left(\alpha_{kj} + \beta_{kj} y_{ij} + \sum_{j' \in \{1,\dots,d\}\backslash\{j\}} \gamma_{jj'} y_{ij'}\right), \tag{5.5}$$

where ρ is the cumulative distribution function of any continuous distribution function and $\psi = (\alpha, \gamma, \beta)$ is the vector parameter of this MNAR model where

$$\begin{aligned}
\alpha &= (\alpha_{11}, \dots, \alpha_{1d}, \dots, \alpha_{K1}, \dots, \alpha_{Kd})^T \in \mathbb{R}^{Kd} \\
\gamma &= (\gamma_{12}, \dots, \gamma_{1d}, \dots, \gamma_{d1}, \dots, \gamma_{dd-1})^T \in \mathbb{R}^{d(d-1)} \\
\beta &= (\beta_{11}, \dots, \beta_{1d}, \dots, \beta_{K1}, \dots, \beta_{Kd})^T \in \mathbb{R}^{Kd},
\end{aligned}$$

if the feature j is continuous. If the feature j is categorical, then $\beta_{kj} = (\beta_{kj}^1, \dots, \beta_{kj}^{\ell_j})^T$ and $\gamma_{jj'} = (\gamma_{jj'}^1, \dots, \gamma_{jj'}^{\ell_{j'}})^T$, when fixing $\beta_{kj}^{\ell_j} = \gamma_{jj'}^{\ell_{j'}} = 0$ for identifiability reasons. By abuse of notation, in the categorical case $y_{ij} \in \mathbb{R}^{\ell_j}$, $\beta_{kj} y_{kj}$ denotes the scalar product $\langle \beta_{kj}, y_{kj} \rangle$.

This general MNAR mechanism seems to be over-parameterized. For instance, for a binary dataset $\mathcal{Y} = \{0,1\}^2 \times \cdots \times \{0,1\}^2$, the number of parameters is equal to $2Kd + d(d-1)$ while, for instance, the most parsimonious mixture model on y, namely the latent class model, has $dK + K - 1$ parameters. Thus, because $(d + K)(d - 1) > -1$ is always true, the missingness model (5.5) has more parameters than the associated mixture model. Since we are expecting that the individual data y convey more information on the partition z that the pattern c of missing data, it seems to be hazardous to allow the missing data modeling to be more complex than the mixture model itself. Consequently, dramatically sparser versions of the general MNAR model (5.5) have to be proposed.

It is firstly reasonable to assume that $\gamma_{jj'} = 0$ (for all $j' \in \{1, \dots, d\}\backslash\{j\}$), meaning that a given value is primary missing due to its own value far before the other variable values. Therefore, the most complex model that we propose is the so-called MNAR$y^k z^j$ model

$$\text{MNAR}y^k z^j: \quad \mathbb{P}(c_{ij} = 1 \mid y_i, z_{ik} = 1; \psi) = \rho(\alpha_{kj} + \beta_{kj} y_{ij}), \tag{5.6}$$

with $\psi = (\alpha, \beta)$.

The parameters α_{kj} represent the effect of missingness on the k-th class membership which depends on the variable j (i.e. the effect is not the same for all variables). The parameters β_{kj} represent the direct effect of missingness on the variable j which depends on the class k.

5.3.1 Sparser models

Effect of the missingness on both the variable and the class membership Missing model (5.6) can be broken down into the following particular cases:

$$\text{MNAR}yz^j: \quad \mathbb{P}(c_{ij} = 1 \mid y_i, z_{ik} = 1; \psi) = \rho(\alpha_{kj} + \beta_j y_{ij}), \tag{5.7}$$

where $\psi = (\alpha, (\beta_1, \ldots, \beta_d)^T)$.

$$\text{MNAR}y^kz: \quad \mathbb{P}(c_{ij} = 1 \mid y_i, z_{ik} = 1; \psi) = \rho(\alpha_k + \beta_{kj} y_{ij}) \tag{5.8}$$

where $\psi = ((\alpha_1, \ldots, \alpha_K)^T, \beta)$.

For the MNARyz^j model, the missingness has a different effect on class membership depending on the variable and it has the same effect on a particular variable regardless of the class. In the contrary, for the MNARy^kz model, we consider that the missingness has the same effect on class membership for all the variables but it has different effect on a particular variable depending on the class. Allowing the parameters β_{kj} and α_{kj} to be dependent on the classes or the variables respectively can be thought of as redundant. Thus, we can consider that the effects on a particular variable and on the class membership are respectively the same for all the classes and for all the variables. It is the purpose of the following MNARyz model.

$$\text{MNAR}yz: \quad \mathbb{P}(c_{ij} = 1 \mid y_i, z_k = 1; \psi) = \rho(\alpha_k + \beta_j y_{ij}) \tag{5.9}$$

where $\psi = ((\alpha_1, \ldots, \alpha_K)^T, (\beta_1, \ldots, \beta_d)^T)$.

Effect of the missingness only on the variable A special case of missing not at random mechanisms, that is widely used in practice (Mohan, 2018), is the self-masked case, called MNARy here, where the only effect of missingness is on the variable j and is the same regardless of the class membership,

$$\text{MNAR}y: \quad \mathbb{P}(c_{ij} = 1 \mid y_i, z_{ik} = 1; \psi) = \rho(\alpha_0 + \beta_j y_{ij}) \tag{5.10}$$

where $\psi = (\alpha_0, \beta_1, \ldots, \beta_d)^T$, with α_0 the intercept (considering $y_{ij} = 1$).

A slightly more general case can be considered by allowing the effect of missingness on the variable j to depend on the class, as in the following MNARy^k model,

$$\text{MNAR}y^k: \quad \mathbb{P}(c_{ij} = 1 \mid y_i, z_{ik} = 1; \psi) = \rho(\alpha_0 + \beta_{kj} y_{ij}) \tag{5.11}$$

where $\psi = (\alpha_0, \beta)$, with α_0 the intercept (considering $y_{ij} = 1$).

Effect of the missingness only on the class membership In the MNARz model, we consider that the only effect of missingness is on the class membership k which is the same for all variables,

$$\text{MNAR}z: \quad \mathbb{P}(c_{ij} = 1 \mid y_i, z_{ik} = 1; \psi) = \rho(\alpha_k) \tag{5.12}$$

	Effect on the variable j		Effect on the class membership k		Nb parameters	
	Depends on j	Depends on k	Depends on j	Depends on k	Continuous	Categorical
MNAR$y^k z^j$ (5.6)	✓	✓	✓	✓	$2Kd$	$K(d + \sum_{j=1}^{d}(\ell_j - 1))$
MNARyz^j (5.7)	✓	✗	✓	✓	$(K+1)d$	$Kd + \sum_{j=1}^{d}(\ell_j - 1)$
MNAR$y^k z$ (5.8)	✓	✓	✗	✓	$K(d+1)$	$K(1 + \sum_{j=1}^{d}(\ell_j - 1))$
MNARyz (5.9)	✓	✗	✗	✓	$(K+d)$	$K + \sum_{j=1}^{d}(\ell_j - 1)$
MNARy (5.10)	✓	✗	✗	✗	$d+1$	$\sum_{j=1}^{d}(\ell_j - 1) + 1$
MNARy^k (5.11)	✓	✓	✗	✗	$Kd+1$	$K\sum_{j=1}^{d}(\ell_j - 1) + 1$
MNARz (5.12)	✗	✗	✗	✓	K	K
MNARz^j (5.13)	✗	✗	✓	✓	Kd	Kd
MCAR (5.14)	✗	✗	✗	✗	1	1

Table 5.1: Effect of missingness and their dependencies for the models that we consider. The last column indicates the corresponding number of parameters for each model.

where $\psi = (\alpha_1, \ldots, \alpha_K)^T$.

Finally, the MNARz^j model is a slightly more general case than the MNARz model, because the effect of missingness on the class membership k is not the same for all the variables,

$$\text{MNAR}z^j: \quad \mathbb{P}(c_{ij} = 1 \mid y_i, z_{ik} = 1; \psi) = \rho(\alpha_{kj}) \tag{5.13}$$

where $\psi = \alpha$.

MCAR model The last model that we consider is the naive one, which assumes MCAR values, i.e. each value has the same probability to be missing.

$$\text{MCAR}: \quad \mathbb{P}(c_{ij} = 1 \mid y_i, z_{ik} = 1; \psi) = \rho(\alpha_0), \tag{5.14}$$

where $\psi = \alpha_0$. This model is included in all the others.

For more clarity, Figure 5.1 gives the embedding between the different models. In addition, Table 5.1 shows the effects of missingness and their dependencies for each model and their corresponding number of parameters.

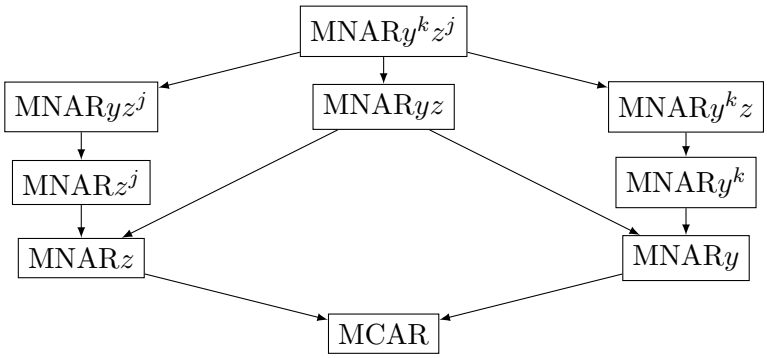


Figure 5.1: From the most general model to the sparsest one. Model A → Model B means that Model A can yield Model B, as Model B is a particular case of Model A. For example, the MNARz model is included in the MNARz^j model which is itself included in the MNARyz^j model involved by the MNAR$y^k z^j$ model.

5.3.2 Interpretation of the MNARz and MNARz^j models

The MNARz model given in (5.12) is the simplest of the MNAR models we propose. Roughly speaking, this model assumes that the proportion of missing values can vary among the clusters. However, behind this apparent simplicity, it benefits from interesting properties we underline below.

Dependency of the MNARz on y_i Although MNARz does not directly involve y_i in its ground definition (5.12), it does not mean that the pattern c_i does not depend on y_i since z_i depends itself on y_i. This can be theoretically observed through the expression

$$\mathbb{P}(c_i \mid y_i; \pi, \theta, \psi) = \sum_{k=1}^{K} \mathbb{P}(c_i \mid z_{ik} = 1; \psi)\mathbb{P}(z_{ik} = 1 \mid y_i; \pi, \theta) \neq \mathbb{P}(c_i; \theta, \psi).$$

This indirect dependency of MNARz on y_i is also numerically illustrated on Figure 5.2 by drawing $\mathbb{P}(c_i \mid y_i; \pi, \theta, \psi)$ for a a three component univariate Gaussian model with mixing proportions $\pi_1 = \pi_2 = 0.3$ and $\pi_3 = 0.4$, with centers $\mu_1 = \mu_3 = -5$ and $\mu_2 = 0$, and with variances $\sigma_k^2 = k$ $(k \in \{1, 2, 3\})$. The MNARz parameters are fixed to $\alpha_1 = 2$, $\alpha_2 = 0$ and $\alpha_3 = 1$.

Reinterpretation of the MNARz and MNARz^j models as a MAR strategy Finally it is important to mention that MNARz and MNARz^j can be linked to a MAR-like strategy commonly used in the machine learning community (Josse et al., 2019). It consists of using a MAR mixture model on the concatenated dataset $\tilde{Y}^{\text{obs}} = (Y^{\text{obs}}, C)$ as a way for

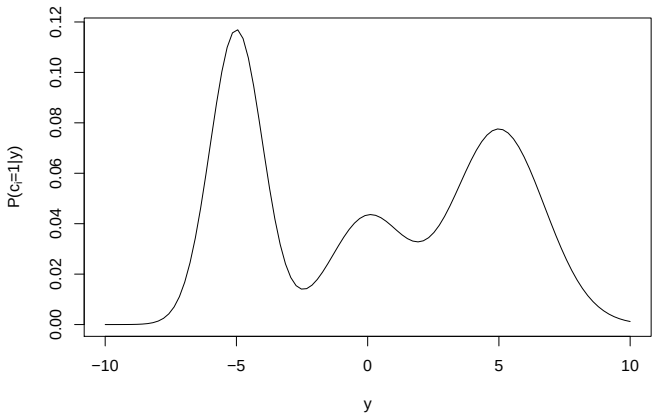


Figure 5.2: Numerical illustration of dependency between c and y in a MNARz model.

easily dealing with missing data. For instance, if Y^{obs} and C are defined as

$$Y^{\text{obs}} = \begin{pmatrix} ? & 2.6 & 5 \\ \text{blue} & 1.9 & 4 \\ \text{red} & 2.3 & ? \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

then $\tilde{Y}^{\text{obs}}$ is expressed as

$$\tilde{Y}^{\text{obs}} = \begin{pmatrix} ? & 2.6 & 5 & 1 & 0 & 0 \\ \text{blue} & 1.9 & 4 & 0 & 0 & 0 \\ \text{red} & 2.3 & ? & 0 & 0 & 1 \end{pmatrix}.$$

The MAR mixture model which is used for this new dataset $\tilde{Y}^{\text{obs}}$ assuming a MAR missing mechanism is equivalent to the mixture model for Y^{obs} given in (1.16) assuming a MNARz or MNARz^j model for C. This property is done more precise in Proposition 21 (see in particular (5.16)) for the explicit expression of the mixture associated to the dataset $\tilde{Y}^{\text{obs}}$. The proof of this proposition is given in Appendix E.1. For simplicity this proposition is particularized to maximum likelihood estimate, but it could be easily generalized to a large family of other relevant estimation strategies.

Proposition 21. *Let us consider the dataset* $(\tilde{y}_1^{\text{obs}}, \ldots, \tilde{y}_n^{\text{obs}})$ *such that* $\tilde{y}_i{}^{\text{obs}} = (y_i^{\text{obs}}, c_i)$ *for* $i \in \{1, \ldots, n\}$. *Assume that all* $\tilde{y}_i^{\text{obs}}$ *arise i.i.d. from the mixture model*

$$\tilde{f}(\tilde{y}_i^{\text{obs}}; \pi, \theta, \psi) = \sum_{k=1}^{K} \pi_k \tilde{f}_k(\tilde{y}_i^{\text{obs}}; \theta_k, \alpha_{kj}) \tag{5.15}$$

$$= \sum_{k=1}^{K} \pi_k f_k(y_i; \theta_k) \prod_{j=1}^{d} \rho(\alpha_{kj})^{c_{ij}} \rho(\alpha_{kj})^{1-c_{ij}}. \tag{5.16}$$

Then the maximum likelihood estimate of (θ, ψ) associated to the dataset $\tilde{y}_i^{\text{obs}}$ with the previous mixture model $\tilde{f}$ under the MAR assumption is the same as the maximum likelihood estimate of (π, θ, ψ) associated to the dataset y_i^{obs} with the mixture model (1.16) under the MNARz assumption (5.12) and MNARz^j assumption (5.13).

5.4 Identifiability results

5.4.1 Continuous and count data

Proving the identifiability of the parameters of a mixture model containing missing values amounts to prove that the joint distribution of (y_i, z_i, c_i) can be uniquely determined from available information. Therefore, we prove the identifiability of the parameters of the observed distribution

$$f(y_i^{\text{obs}}, c_i; \pi, \theta, \psi) = \sum_{k=1}^{K} \int_{y_i^{\text{mis}}} \pi_k f_k(y_i; \theta_k) \mathbb{P}(c_i \mid y_i, z_{ik} = 1; \psi) dy_i. \tag{5.17}$$

This section starts with Proposition 22 which gives sufficient conditions for the identifiability of the parameters for continuous or count data. We denote by f_{kj} the marginal density of the variable j for the class k and we assume

A1. The parameters (π, θ) of the marginal mixture defined by the density $\sum_{k=1}^{K} \pi_k f_k(y_i; \theta_k)$ are identifiable;

A2. There exists a total ordering $\preceq$ of $\mathcal{F}_j \times \mathcal{R}$, for $j \in \{1, \ldots, d\}$ fixed, where $\mathcal{F}_j$ is the family of the data densities $\{f_{1j}, \ldots, f_{Kj}\}$ and $\mathcal{R}$ is the family of the mechanism densities $\{\rho_1, \ldots, \rho_K\} = \{\rho(.; \psi_1), \ldots, \rho(.; \psi_K)\}$. The total ordering is such that $\forall k < \ell$, $F_k \preceq F_\ell$ (denoting $F_k = \rho_k f_{kj}$ and $F_\ell = \rho_\ell f_{\ell j}$) implies $\lim_{u \to +\infty} \frac{\rho_\ell(u) f_{\ell j}(u)}{\rho_k(u) f_{kj}(u)} = 0$;

A3. The missing-data distribution ρ is assumed to be strictly monotone.

Assumption **A1.** means that the identifiability of the parameters (π, θ, ψ) of the model (5.17) requires the identifiability of the parameters (π, θ) of the marginal mixture of (Y, Z) (i.e. considering the case without missing values). Some authors have already studied the identifiability of the mixture models, when no missing values in Y occur, especially Teicher (1963) for Gaussian mixtures and Yakowitz and Spragins (1968) for Poisson mixtures. Assumption **A2.** is the core ingredient to prove the identifiability of the parameters and we illustrate it by considering concrete examples in the following. Note that under Assumption **A3.** the probit and the logistic function may be considered, which are the most widely used for MNAR specifications.

Proposition 22. *Under Assumptions **A1.**, **A2.** and **A3.**, the parameters (π, θ, ψ) of the model given by (5.17) considering the MNARy^{k}z^j model given in (5.6) (and therefore all others MNAR models) are identifiable up to label swapping.*

The proof of this proposition is detailed in Appendix E.2 and follows the reasoning used by Teicher (1963, Theorem 2) which proves the identifiability of univariate finite mixture using a total ordering of the mixture densities. In the following, we denote by f_{kj} the marginal density of the variable j for the class k.

On the identifiability of the Gaussian mixture Proposition 22 states the identifiability of the Gaussian mixture with a probit missing-data distribution (details are given in Example 4 presented in Appendix E.2). Indeed, finite Gaussian mixtures are identifiable and, for any variable j, there is a total ordering defined by $\sigma^2_{kj} > \sigma^2_{(k+1)j}$ and $\mu_{kj} > \mu_{(k+1)j}$ if $\sigma^2_{kj} = \sigma^2_{(k+1)j}$, where μ_{kj} and σ^2_{kj} are respectively the mean and the variance of variable j under component k.

This result has been already stated, in the case of univariate distributions, by Miao et al. (2016). In particular, the identifiability conditions in Miao et al. (2016) (conditions 1 and 2) imply the existence of the total ordering defined in Proposition 22. However, these conditions excludes the case of Gaussian mixture with a logistic missing-data distribution, which is very used in practice.

Note that for such a model, a total ordering cannot be defined. Indeed, for variable j, such an ordering cannot be defined if the two univariate variances are equal (*i.e.,* $\sigma^2_{kj} = \sigma^2_{(k+1)j}$) and $\mu_{kj} - \beta_{kj} - \mu_{(k+1)j} + \beta_{(k+1)j} = 0$. Note that for the specific case of Gaussian mixture where all the univariate variances are different between the components, then conditions of Proposition 22 hold true with a logistic missing-data distribution and and so does its identifiability. In addition, for sparser MNAR models for which the effect on the variable j does not depend on the class membership k (i.e. $\beta_{kj} = \beta_{(k+1)j}$), the conditions of Proposition 22 hold true with a logistic missing-data distribution. Moreover, as stated by Corollary 1 (proved in Appendix E.2), the condition on the covariance matrices (including the case of homoscedastic Gaussian mixture) can be relaxed to obtain the generic identifiability of the model (*i.e.,* all not-identifiable parameter choices lie within a proper submanifold, and thus form a set of Lebesgue zero measure; Allman et al. (2009)).

Corollary 1. *Assume that* $\sum_{k=1}^{K} \pi_k f_k(y_i;\theta_k)$ *is a multivariate Gaussian mixture,* ρ *is the logistic function and that the missingness scenario is defined by* (5.6), (5.8) *or* (5.11), *then, the parameters* (π, θ, ψ) *of the model given by* (5.17) *are generically identifiable up to label swapping, i.e. all not-identifiable parameter choices lie within a proper submanifold, and thus form a set of Lebesgue zero measure.*

For the other MNAR models given in (5.7), (5.9), (5.10), (5.12) *and* (5.13), *the parameters* (π, θ, ψ) *of the model given by* (5.17) *are identifiable up to label swapping.*

Proposition 1 can also be applied for variables with integer value (i.e. count data), as shown in Examples 5 and 6 in Appendix E.2 for the Poisson mixture with probit or logistic missing-data distributions. The identifiability for the different missing scenarios are summarized in Table 5.2.

	Gaussian		Poisson	
	Probit	Logistic	Probit	Logistic
MNAR$y^k z^j$ (5.6) MNAR$y^k z$ (5.8) MNARy^k (5.11)	✓	generic identifiability	✓	generic identifiability
MNAR$y z^j$ (5.7) MNARyz (5.9) MNARy (5.10) MNARz (5.12) MNARz^j (5.13)	✓	✓	✓	✓

Table 5.2: Identifiability for different missing scenarios when the mixture is Gaussian or Poisson.

5.4.2 Categorical data

The case of categorical variables is not covered by Proposition 22. Consider that the vector y_i is composed of categorical variables, such that the variable y_{ij} can take ℓ_j values (i.e. $y_{ij} = (y_{ij}^1, \dots, y_{ij}^{\ell_j})$), and follows a mixture of K products of d multinomial distributions with parameters $(\theta_{k1}, \dots, \theta_{kd})_{k=1,\dots,K}$ such that $\theta_{kj} \in \mathbb{R}^{\ell_j}$. We assume the following:

A4. The feature are independently drawn conditionally to the group membership, i.e.

$$f_k(\cdot; \theta_k) = \prod_{j=1}^{d} f_{kj}(\cdot\,; \theta_{kj}); \tag{5.18}$$

A5. The dimension d of the observations is related to the number K of clusters so that

$$d \geqslant 2\lceil \log_2 K \rceil + 1.$$

Assumptions **A4.** and **A5.** are classical in the categorical case, even without missing values (Allman et al., 2009). Proposition 23 states that generic identifiability holds only for the MNARz and the MNARz^j missing scenarios and that the other missing scenarios lead to non-identifiable models. Its proof is detailed in Appendix E.2 and uses Corollary 5 of Allman et al. (2009) which gives the identifiability of finite mixtures of Bernoulli products.

Proposition 23. *Under Assumptions **A3.**, **A4.** and **A5.**, the parameters of the model given in (5.17) considering the MNARz or MNARz^j models given in (5.12) and (5.13) are generically identifiable, up to label swapping.*

For the other MNAR models, i.e. when the effect of the missingness may depend on the values of the variables, given in (5.6), (5.8), (5.7), (5.9), (5.10) and (5.11), the parameters of the model (5.17) are not identifiable.

5.4.3 Mixed data

Mixed data are a combination of continuous and categorical data. More precisely, let us denote y_i^{co} the set of continuous variables of cardinal d_{co} and y_i^{ca} the set of categorical variables of cardinal $d_{\text{ca}} = d - d_{\text{co}}$. Without loss of generality, we can consider $y_i^{\text{co}} = (y_{i1}, \dots, y_{id_{\text{co}}})$ and $y_i^{\text{ca}} = (y_{i(d_{\text{co}}+1)}, \dots, y_{id})$. Thus, $y_i = (y_i^{\text{co}}, y_i^{\text{ca}})$. By assuming the conditional independence of the features given the group membership, the idenfiability of mixed data directly follows from Proposition 22 for the continuous variables and Proposition 23 for the categorical variables.

Corollary 2.

- *For the continuous variables, assume* ***A1.*** *and* ***A2.****, i.e. the parameters* (π, θ) *of the marginal mixtures for* $(y_1, \dots, y_{d_{\text{co}}})$ *are identifiable and there exists a total ordering of* $\mathcal{F}_j \times \mathcal{R}$ *for* $j \in \{1, \dots, d_{\text{co}}\}$*. Consider the MNAR*$y^k z^j$ *model given in* (5.6) *(thus all the others are allowed).*
- *For the categorical variables, assume* ***A5.*** *i.e.* $d_{ca} \geqslant 2\lceil \log_2 K \rceil + 1$*. Consider the MNARz or MNAR*z^j *model given in* (5.12) *or* (5.13)*.*

Under Assumption ***A3.*** *and* ***A4.****, the parameters of the model in* (5.17) *are generically identifiable, up to label swapping.*

5.5 Estimation of the proposed MNAR models

As seen in Section 5.2, MNAR models are not ignorable, thus they require a specific inference procedure for estimating the parameters π, θ and ψ. This section gathers the description of the EM and SEM algorithms for Gaussian, multinomial and mixed data with MNAR models for maximum likelihood estimation. Details of the algorithms are given in Appendix E.3.

Following the expression of the observed likelihood given in (5.2), the observed log-likelihood is

$$\ell(\pi, \theta, \psi; Y^{\text{obs}}, C) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \int_{\mathcal{Y}_i^{\text{mis}}} \pi_k f_k(y_i; \theta) \mathbb{P}(c_i \mid y_i, z_{ik} = 1; \psi) dy_i \right), \tag{5.19}$$

The complete log-likelihood is then

$$\ell_{\text{comp}}(\pi, \theta, \psi; Y, Z, C) = \sum_{i=1}^{n} \sum_{k=1}^{K} \log(\pi_k f_k(y_i; \theta) \mathbb{P}(c_i \mid y_i, z_{ik} = 1; \psi)). \tag{5.20}$$

If the complete log-likelihood was known, simply maximizing it would be sufficient to estimate the parameters (π, θ, ψ). However, this quantity is unknown (since the class memberships are unknown) and maximizing it requires the use of EM or SEM algorithms, of particular interest for finding the maximum likelihood parameters in presence of latent or missing values. In the following, the iterates index of any algorithm will be r.

5.5.1 The EM algorithm

We first detail the EM algorithm for the different MNAR models at hand with Gaussian, multinomial and mixed mixture models. In its general form, the EM algorithm (Dempster et al., 1977) consists of iterating the following two steps, starting from an initial parameter value $(\pi^0, \theta^0, \psi^0)$ and until a stopping criterion is met (e.g. a given maximum iteration value $r \leqslant r_{\max}$):

- **E-step:** Computation of $Q(\pi, \theta, \psi; \pi^r, \theta^r, \psi^r)$ which is the expected complete log-likelihood ℓ_{comp} knowing the observed data and a current value of the parameters. This quantity can be decomposed into two parts (see Appendix E.3 for the full computation) as follows

$$\begin{aligned} Q(\pi, \theta, \psi; \pi^r, \theta^r, \psi^r) &= \mathbb{E}[\ell_{\text{comp}}(\pi, \theta, \psi; y, z, c) | y_i^{\text{obs}}, c_i; \pi^r, \theta^r, \psi^r] \\ &= Q_y(\pi, \theta; \pi^r, \theta^r) + Q_c(\psi; \psi^r) \end{aligned} \tag{5.21}$$

  with

$$Q_y(\pi, \theta; \pi^r, \theta^r) = \sum_{i=1}^{n} \sum_{k=1}^{K} (\tau_{ik})^r \log(\pi_k) + \sum_{i=1}^{n} \sum_{k=1}^{K} (\tau_{ik})^r E_{iy}^r(\theta). \tag{5.22}$$

$$Q_c(\psi; \psi^r) = \sum_{i=1}^{n} \sum_{k=1}^{K} (\tau_{ik})^r E_{ic}^r(\psi). \tag{5.23}$$

  where for $i = 1, \ldots, n$ and $k = 1, \ldots, K$,

$$E_{iy}^r(\theta) = \mathbb{E}\left[\log(f_k(y_i; \theta_k)) \mid y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r\right], \tag{5.24}$$

$$E_{ic}^r(\psi) = \mathbb{E}\left[\log(\mathbb{P}(c_i \mid y_i, z_{ik} = 1; \psi)) \mid y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r\right], \tag{5.25}$$

$$(\tau_{ik})^r = \mathbb{P}(z_{ik} = 1 \mid y_i^{\text{obs}}, c_i; \theta^r, \psi^r, \pi^r) \propto \pi_k^r f_k(y_i^{\text{obs}}; \theta_k^r) \mathbb{P}(c_i \mid y_i^{\text{obs}}, z_{ik} = 1; \psi^r) \tag{5.26}$$

- **M-step:** Maximization over π, θ and ψ of $Q(\pi, \theta, \psi; \pi^r, \theta^r, \psi^r)$, by respectively maximizing $Q_y(\pi, \theta; \pi^r, \theta^r)$ w.r.t. (π, θ) and $Q_c(\psi; \psi^r)$ w.r.t. ψ. This step leads to the parameters π^{r+1}, θ^{r+1} and ψ^{r+1}.

Let us note that in any case, the maximization of $Q_y(\pi, \theta; \pi^r, \theta^r)$ over π is easy, once the $(\tau_{ik})^r$'s are given. However, the computation of $E_{iy}^r(\theta)$, $E_{ic}^r(\psi)$ and $(\tau_{ik})^r$, then the maximization of $Q_y(\pi, \theta; \pi^r, \theta^r)$ over θ and $Q_c(\psi; \psi^r)$ over ψ, both depend additionally on the MNAR model at hand and thus need to be specifically detailed hereafter. It is straightforward with the MNARz and MNARz^j models given in (5.12) and (5.13) but more difficult with all the other MNAR models, called in the sequel MNAR$y*$ (modelling the effect of the missingness depending on y). Recall that MNARz and MNARz^j are the only ones which guarantee identifiability of the parameters for categorical data (see Section 5.4). For the MNAR$y*$ models, only algorithms for continuous data have to be described. For the sake of brevity, the estimation procedure in the continuous case is restricted to the case where the variables are Gaussian.

5.5.1.1 MNARz and MNARz^j models

Consider the MNARz^j model which includes the MNARz one, i.e. $\mathbb{P}(c_{ij} \mid y_i, z_{ik} = 1; \psi) = \rho(\alpha_{kj})$. Computing (5.24) requires to integrate over the distribution $\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r)$. For the MNAR$z^j$ model, by dependence of y, one has

$$\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r) = \mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik} = 1; \theta^r, \psi^r).$$

For Gaussian mixture, with the following notations

$$\begin{aligned}(y_i \mid z_{ik} = 1; \theta^r) &= \left(\begin{pmatrix} y_i^{\text{obs}} \\ y_i^{\text{mis}} \end{pmatrix} \mid z_{ik} = 1; \theta^r \right) \\ &\sim \mathcal{N}\left(\begin{pmatrix} (\mu_{ik}^{\text{obs}})^r \\ (\mu_{ik}^{\text{mis}})^r \end{pmatrix}, \begin{pmatrix} (\Sigma_{ik}^{\text{obs,obs}})^r & (\Sigma_{ik}^{\text{obs,mis}})^r \\ (\Sigma_{ik}^{\text{mis,obs}})^r & (\Sigma_{ik}^{\text{mis,mis}})^r \end{pmatrix} \right),\end{aligned}$$

one obtains

$$\left(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik} = 1; \theta^r\right) \sim \mathcal{N}\left((\tilde{\mu}_{ik}^{\text{mis}})^r, (\tilde{\Sigma}_{ik}^{\text{mis}})^r\right). \tag{5.27}$$

It makes the expectation in (5.24) classical. In addition, by independence of y, (5.25) and (5.26) have closed forms. It leads to a straightforward maximization step. The EM algorithm for the MNARz^j model is described in Algorithm 3 for Gaussian mixture. All the details are given in E.3.1.1 and E.3.1.2 for both Gaussian and categorical data. The initialization and the stopping criterion are discussed in Section 5.6.

5.5.1.2 MNAR$y*$ models

The MNAR$y*$ models consider the effect of the missingness depending on y and lead then to unfeasible computations. The distribution $\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r)$ is explicit (a truncated Gaussian as shown in Appendix E.3.2.1) if the missing-data distribution ρ is probit but it is not classical if ρ is logistic. However, to our knowledge, for both forms of missing-data distributions, Equations (5.25) and (5.26) have no closed forms. In addition, the maximization over ψ of (5.25) is a delicate issue because the function involved is not concave.

5.5.2 The SEM algorithm

While the computation for the MNARz and MNARz^j models are feasible, it is not the case for the models MNAR$y*$ with Gaussian mixtures. The SEM algorithm (Celeux and Diebolt, 1985) could avoid this difficulty, by imputing missing values using a Gibbs sampling instead of integrating over them. In addition, it has another possible advantage over the EM algorithm since it is not trapped by the first local maximum encountered of the likelihood function (Celeux and Diebolt, 1985).

The SEM algorithm consists of the following two steps for $r_{\max}$ iterations:

Algorithm 3 EM algorithm for Gaussian mixture and MNARz^j model

Input: $Y^{\text{NA}} \in \mathbb{R}^{n\times d}$, $K \geqslant 1$, $r_{\max}$.
Initialize π_k^0, μ_k^0, Σ_k^0 and ψ_k^0.
for $r = 0$ **to** $r_{\max}$ **do**
E-step:
for $i = 1$ **to** n, $k = 1$ **to** K **do**
$(\tilde{\mu}_{ik}^{\text{mis}})^r = (\mu_{ik}^{\text{mis}})^r + (\Sigma_{ik}^{\text{mis,obs}})^r \left((\Sigma_{ik}^{\text{obs,obs}})^r\right)^{-1} \left(y_i^{\text{obs}} - (\mu_{ik}^{\text{obs}})^r\right)$.
$(\tilde{\Sigma}_{ik}^{\text{mis}})^r = (\Sigma_{ik}^{\text{mis,mis}})^r - (\Sigma_{ik}^{\text{mis,obs}})^r \left((\Sigma_{ik}^{\text{obs,obs}})^r\right)^{-1} (\Sigma_{ik}^{\text{obs,mis}})^r$.
$(\tilde{y}_{i,k})^r = (y_i^{\text{obs}}, (\tilde{\mu}_{ik}^{\text{mis}})^r)$.
$\tilde{\Sigma}_{ik}^r = \begin{pmatrix} 0_i^{\text{obs,obs}} & 0_i^{\text{obs,mis}} \\ 0_i^{\text{mis,obs}} & (\tilde{\Sigma}_{ik}^{\text{mis}})^r \end{pmatrix}$.
$(\tau_{ik})^r \propto \pi_k^r \phi(y_i^{\text{obs}}; (\mu_{ik}^{\text{obs}})^r, (\Sigma_{ik}^{\text{obs,obs}})^r) \prod_{j=1}^d \rho(\alpha_{kj}^r)^{c_{ij}} (1 - \rho(\alpha_{kj}^r))^{1-c_{ij}}$
end for
M-step:
for $k = 1$ **to** K **do**
$\pi_k^{r+1} = \frac{1}{n}\sum_{i=1}^n (\tau_{ik})^r \qquad \mu_k^{r+1} = \frac{\sum_{i=1}^n (\tau_{ik})^r (\tilde{y}_{k,i})^r}{\sum_{i=1}^n (\tau_{ik})^r}$
$\Sigma_k^{r+1} = \frac{\sum_{i=1}^n \left[(\tau_{ik})^r \left((\tilde{y}_{i,k})^r - \mu_k^{r+1}\right)\left((\tilde{y}_{i,k})^r - \mu_k^{r+1}\right)^T + \tilde{\Sigma}_{ik}^r\right)\right]}{\sum_{i=1}^n (\tau_{ik})^r}$
Let ψ^{r+1} be the resulted coefficients of a GLM with a binomial link function.
end for
end for

- **SE-step**: Draw the missing data $(y_i^{\text{mis}})^{r+1}$ and z_i^{r+1} according to their current conditional distribution $\mathbb{P}(y_i^{\text{mis}}, z_i \mid y_i^{\text{obs}}, c_i; \pi^r, \theta^r, \psi^r)$. Since it is not convenient to simulate this conditional distribution, we simulate instead the following two easier conditional probabilities:

$$z_i^{r+1} \sim \mathbb{P}(\cdot \mid y_i^r, c_i; \pi^r, \theta^r, \psi^r) \quad \text{and} \quad (y_i^{\text{mis}})^{r+1} \sim \mathbb{P}(\cdot \mid y_i^{\text{obs}}, z_i^{r+1}, c_i; \theta^r, \psi^r), \tag{5.28}$$

  where $y_i^r = (y_i^{\text{obs}}, (y_i^{\text{mis}})^r)$. For the latter distribution, we can draw the membership k of z_i^{r+1} from the multinomial distribution with probabilities $(\mathbb{P}(z_{ik} = 1 \mid y_i^r, c_i; \pi^r, \theta^r, \psi^r))_{k=1,\dots,K}$.

  Note that

$$\begin{aligned}
&\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1} = 1, c_i; \theta^r, \psi^r) \\
&= \frac{\prod_{j,c_{ij}=1} \mathbb{P}(c_{ij} = 1 \mid y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}^{r+1} = 1; \psi^r) \mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1} = 1; \theta^r)}{\int_{\mathcal{Y}_i^{\text{mis}}} \prod_{j,c_{ij}=1} \mathbb{P}(c_{ij} = 1 \mid y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}^{r+1} = 1; \psi^r) \mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1} = 1; \theta^r) dy_i^{\text{mis}}},
\end{aligned} \tag{5.29}$$

  so the conditional distribution of $\left((y_i^{\text{mis}})^{r+1} \mid y_i^{\text{obs}}, z_{ik}^{r+1} = 1, c_i\right)$ may not be classical in general, this distribution will be yet made explicit in particular cases (see Section 5.5.2.1).

- **M-step**: Maximization of the completed log-likelihood $\ell_{\text{comp}}(\theta, \psi, \pi; y^r, z^r, c)$ over π, θ and ψ, which provides θ^{r+1}, ψ^{r+1} and π^{r+1}.

5.5.2.1 MNAR$y*$ models

For the MNAR$y*$ models, the conditional distribution $\left((y^{\text{mis}})^{r+1} \mid y_i^{\text{obs}}, z_{ik}^r = 1, c_i\right)$ given in (5.29) is not explicit if the missing-data distribution ρ is logistic, because the product of logistic and Gaussian distributions is not a standard law. Therefore, the SEM algorithm cannot be applied.

However, if ρ is the probit function, we can make the distribution of interest explicit. More particularly, we introduce an instrumental variable L_i such that $L_i = \alpha_k^r + \beta_k^r y_i + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0_d, I_{d\times d})$. By abuse of notation, $\beta_k^r y_i$ denotes the Hadamard product between β_k^r and y_i. Then, c_i can be viewed as an indicator for whether this latent variable is positive, i.e. for all $j = 1, \dots, d$,

$$c_{ij} = \begin{cases} 1 & \text{if } L_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5.30}$$

In the SE-step, instead of drawing $z_i^{r+1}, (y_i^{\text{mis}})^{r+1}$ as in (5.28), we draw $L_i^{r+1}, z_i^{r+1}, (y_i^{\text{mis}})^{r+1}$ as follows

- L_i^{r+1} is drawn according to $\mathbb{P}(.|y_i^r, z_{ik}^r = 1, c_i; \psi^r)$ which is the multivariate truncated Gaussian distribution

$$\mathcal{N}_t(\alpha_k^r + \beta_k^r y_i, I_{d\times d}; a, b), \tag{5.31}$$

where a and b are lower and upper bounds depending on the indicator c_i (detailed in Appendix E.3.2.1).

- The probability parameters of the multinomial distribution used for drawing the membership $k \in \{1, \dots, K\}$ of z_i^{r+1} are

$$\begin{aligned}\mathbb{P}(z_{ik} = 1 \mid L_i^{r+1}, y_i^r, c_i; \pi^r, \theta^r, \psi^r) \\ \propto \prod_{j=1}^{d} \Phi(\alpha_{kj}^r + \beta_{kj}^r (y_{ij}^{\text{mis}})^r)^{c_{ij}} \left(1 - \Phi(\alpha_{kj}^r + \beta_{kj}^r y_{ij}^{\text{obs}})\right)^{1-c_{ij}} \\ \phi_t(L_i^{r+1}; \alpha_k^r + \beta_k^r y_i^r, I_{d\times d}, a, b)\phi(y_i^r; \mu_k^r, \Sigma_k^r)\pi_k^r, \quad (5.32)\end{aligned}$$

  where Φ is the cdf of the standard Gaussian distribution, $\phi(.; \mu_k^r, \Sigma_k^r)$ is the multivariate Gaussian density with mean μ_k^r and covariance matrix Σ_k^r and $\phi_t(.; \alpha_k^r + \beta_k^r y_i, I_{d\times d}, a, b)$ is the multivariate truncated Gaussian density with mean $\alpha_k^r + \beta_k^r y_i$, identity covariance matrix and a and b as lower and upper bounds.

- $(y_i^{\text{mis}})^r$ is drawn according to $\mathbb{P}(.|L_i^{r+1}, z_i^{r+1}, y_i^{\text{obs}}, c_i; \theta^r, \psi^r)$ which is the multivariate Gaussian distribution

$$\mathcal{N}\left(\mu_{ik}^{\text{SEM}}, \Sigma_{ik}^{\text{SEM}}\right), \quad (5.33)$$

  where μ_{ik}^{SEM} and Σ_{ik}^{SEM} depend on the parameters θ^r and ψ^r (see Appendix E.3.2.1 for more details).

Eventually, when ρ is the probit function, the SEM algorithm can be derived, see Algorithm 4. The initialization and the stopping criterion are discussed in Section 5.6.

5.5.2.2 MNARz and MNARz^j models

For MNARz and MNARz^j models, the conditional distribution involved in the SE-step has already been given in (5.27). All the computations are feasible and derived in E.3.2.1 and E.3.2.2 for both Gaussian and categorical data.

5.6 Numerical experiments on synthetic data

In this section, we compare the EM algorithm (for the MNARz (5.12) model) and the SEM algorithm (for the MNARz (5.12), MNARy (5.10), MNARyz (5.9) models) on synthetic data with the SEM algorithm considering MCAR data (5.14) and several two-step heuristics detailed below. These algorithms are detailed for Gaussian variables in Algorithms 3 and 4. The two-step heuristics consist of first imputing the missing values to get a complete dataset and then applying classical model-based clustering which has been implemented for the case without missing values. Regarding the imputation methods in the two-step strategies, we consider the following ones:

Algorithm 4 SEM algorithm for Gaussian mixture, MNAR$y*$ models, ρ is probit

Input: $Y^{\text{NA}} \in \mathbb{R}^{n \times d}$, $K \geqslant 1$, $r_{\max}$.
Initialize Z^0, π_k^0, μ_k^0, Σ_k^0 and ψ_k^0.
for $r = 0$ **to** $r_{\max}$ **do**
  SE-step:
  for $i = 1$ **to** n **do**
    Draw $(L_i)^{r+1}$ from the multivariate truncated Gaussian distribution given in (5.31).
    Draw z_i^{r+1} from the multinomial distribution with probabilities detailed in (5.32).
    Draw $(y_i^{\text{mis}})^{r+1}$ from the multivariate Gaussian distribution given in (5.33).
  end for
  Let $Y^{r+1} = (y_1^{r+1} | \dots | y_n^{r+1})$ be the imputed matrix.
  Let $Z^{r+1} = (z_1^{r+1} | \dots | z_n^{r+1})$ be the partition.
  M-step:
  for $k = 1$ **to** K **do**
    Let π_k^{r+1} be the proportion of rows of Y^{r+1} belonging class k.
    Let $\mu_k^{r+1}, \Sigma_k^{r+1}$ be the mean and covariance matrix of rows of Y^{r+1} belonging to class k.
    Let ψ^{r+1} be the resulted coefficients of a GLM with a binomial link function.
  end for
end for

(a) multiple imputations by chained equations (Buuren and Groothuis-Oudshoorn, 2010): it consists of generating M plausible values for each missing value by computing expectation of the missing variables given the observed ones (`Mice`),

(b) single imputation by chained equations (Buuren and Groothuis-Oudshoorn, 2010), i.e. the same method as in (a) with $M = 1$ (`Ice`).

For the first method, M imputed datasets are computed, the model-based clustering is then performed on each complete dataset, for which the performance is measured. The final performance of this method is computed with the mean.

Measuring the performance It is possible to choose one of the proposed methods by using an information criterion such as the Bayesian Information Criterion (BIC) (Schwarz, 1978) or the Integrated Complete-data Likelihood (ICL) (Biernacki et al., 2000). As the ICL involves an integral which is generally not explicit, we can use an approximate version (Baudry et al., 2015) which we detail when there is missing data. For a model $\mathcal{M}_{\text{df}}$ with df parameters, one has

$$\text{ICL}(\text{df}) = \text{BIC}(\text{df}) + \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik}^{\text{MAP}}(\pi^{\text{MLE}}, \theta^{\text{MLE}}, \psi^{\text{MLE}}) \log(\tau_{ik}(\pi^{\text{MLE}}, \theta^{\text{MLE}}, \psi^{\text{MLE}})),$$

$$\text{BIC}(\text{df}) = \ell(\pi^{\text{MLE}}, \theta^{\text{MLE}}, \psi^{\text{MLE}}; Y^{\text{obs}}, C) - \frac{\text{df}}{2} \log(n)$$

with $\pi^{\text{MLE}}, \theta^{\text{MLE}}, \psi^{\text{MLE}}$ denoting the maximum likelihood estimators computed for the model $\mathcal{M}_{\text{df}}$ and where $\ell(\pi, \theta, \psi; Y^{\text{obs}}, C)$ is the observed log-likelihood given in (5.19) and

$$\begin{aligned} \tau_{ik}(\pi, \theta, \psi) &= \mathbb{P}(z_{ik} = 1 | y_i^{\text{obs}}, c_i; \pi, \theta, \psi) = \frac{\pi_k f_k(y_i^{\text{obs}}; \theta_k) \mathbb{P}(c_i \mid y_i^{\text{obs}}, z_{ik} = 1; \theta, \psi)}{\sum_{h=1}^{K} \pi_h f_h(y_i^{\text{obs}}; \theta_h) \mathbb{P}(c_i \mid y_i^{\text{obs}}, z_{ih} = 1; \theta, \psi)} \\ \hat{z}_{ik}^{\text{MAP}}(\pi, \theta, \psi) &= \underset{k \in \{1, \dots, K\}}{\text{argmax}}\, \tau_{ik}(\pi, \theta, \psi) \end{aligned} \tag{5.34}$$

The BIC criterion is expected to select a relevant mixture model in a density estimation perspective, while ICL is expected to select a relevant mixture model for a clustering purpose. Indeed, this latter includes an entropy term which involves the estimator of the partition, given by $\hat{Z}^{\text{MAP}} = \{\hat{z}_{ik}^{\text{MAP}}\}_{i,k} \in \mathbb{R}^{n \times K}$. In addition, the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) can be computed between the true partition given by Z and the estimator of the partition. Obviously other strategies are possible to select a sensible and useful mixture model (see Celeux et al. (2019)).

Initialization and stopping criteria for the (S)EM algorithms

Initialization For the SEM algorithm considering MCAR data, the partition matrix Z^0 is computed with an arbitrary stochastic matrix, with each row summing to 1. For Gaussian data, the parameters $\pi_k^0, \mu_k^0, \Sigma_k^0$ for each class $k \in \{1, \dots, K\}$ are initialized with the proportion, the mean and the covariance matrix of observed rows belonging to class k respectively. The mechanism parameters ψ^0 is initialized with arbitrary values. The initialization of the parameters is performed on a sub-sample of the data (by default, 30% of the observations). This random initialization is performed several times, and we keep the result which maximizes the ICL criterion. We initialize the algorithms for MNAR models with the result of the SEM algorithm considering MCAR data.

Stopping criteria The EM algorithm stops when a certain number of iteration $r_{\max}$ have been performed and the difference between the log-likelihood of the two last iterates is inferior to a certain threshold. The SEM algorithm stops when a certain number of iteration $r_{\max}$ has been performed. Note that for the SEM algorithm the set of parameters returned is the one which maximizes the observed log-likelihood.

Leveraging from MNAR data in clustering In addition to dealing with informative missing data, the expected interest of MNAR modeling in the clustering context is to improve the partition estimation. To illustrate this, let us consider a bivariate two-component Gaussian mixture with equal mixing proportions and identity covariance matrices, i.e. the observations $Y \in \mathbb{R}^2$ follows the distribution $Y \sim 0.5\mathcal{N}(\mu_1, I_{2\times 2}) + 0.5\mathcal{N}(\mu_2, I_{2\times 2})$. The difference between the centers of both mixture components is taken as $\Delta_\mu = \mu_{21} - \mu_{11} \in \{0.5, 1, \dots, 3\}$, where for any cluster $k \in \{1, 2\}$, the center $\mu_k = (\mu_{k1}, \mu_{k2})$ is chosen with equal components ($\mu_{k2} = \mu_{k1}$). This cluster overlap controls the mixture separation, which can vary from a low separation ($\Delta_\mu = 0.5$) to a high one ($\Delta_\mu = 3$). We consider the MNARz model given in

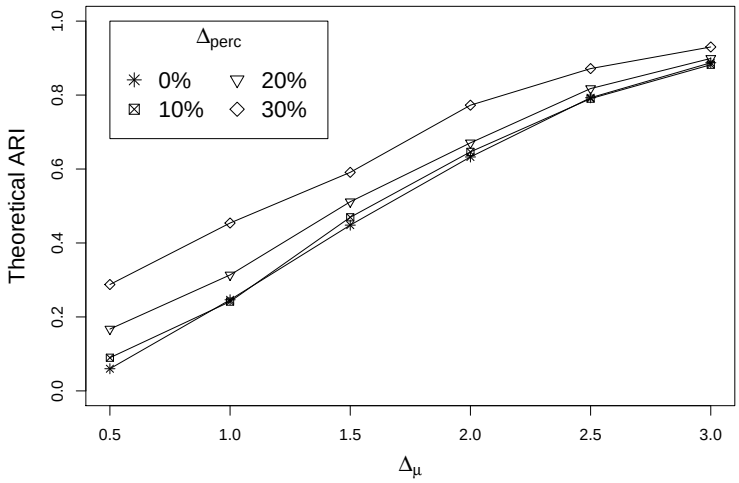


Figure 5.3: Relative effect of both the mixture component separation strength Δ_μ and the MNAR evidence Δ_{perc} on theoretical ARI. For example, if $\Delta_{\text{perc}} = 10\%$, it means that the second class has 10% more missing values than the first class.

(5.12) with ρ the cumulative distribution function of the standard Gaussian. One can play on the discrepancy between inter-cluster missing proportions $\Delta_{\text{perc}} = |\text{perc}_2 - \text{perc}_1|$, by making it vary in $\{0, 0.1, 0.2, 0.3\}$. The value Δ_{perc} means that if the percentage of missing values in the first cluster is perc_1, the percentage of missing values in the second cluster is $\text{perc}_2 = (\text{perc}_1 + \Delta_{\text{perc}})$. To have $\text{p}_1\%$ missing values in the first class, the parameter chosen is $\alpha_1 = \phi^{-1}(p_1\%)$, with ϕ^{-1} the inverse of the cumulative distribution function of the standard Gaussian. Therefore, increasing values of Δ_{perc} corresponds to increase the MNAR evidence: indeed, $\Delta_{\text{perc}} = 0$ corresponds to a MCAR model whereas a high value of Δ_{perc} corresponds to a high difference of missing pattern proportions between clusters. Finally, 15% of missing values is introduced whatever the MNAR evidence Δ_{perc} and the mixture separation are Δ_μ. Figure 5.3 gives the theoretical ARI (i.e. we compute the ARI with the theoretical parameters) as a function of the cluster overlap Δ_μ and the MNAR evidence Δ_{perc}. Even though the good classification rate is mostly influenced by the center separation Δ_μ, it also increases with the MNAR evidence Δ_{perc}. This toy example illustrates how clustering can leverage from MNAR missing values, generally considered as a true hindrance for any statistical analysis.

Toy example: MNARz in the Gaussian case We first consider a simple case with two classes and when the data are bivariate Gaussian ($n = 2000$, $d = 2$)

$$Y \sim \pi_1 \mathcal{N}(\mu_1, \Sigma_1) + \pi_2 \mathcal{N}(\mu_2, \Sigma_2),$$

with $\mu_1 = (0, 0)$, $\Delta_\mu = \mu_2 - \mu_1$, $\pi_1 = 0.3, \pi_2 = 0.7$ and $\Sigma_1 = I_{2\times 2}$, $\Sigma_2 = I_{2\times 2}$. We make the cluster overlap vary as follows: $\Delta_\mu = (2, 2)$ for a low separation between the two clusters and $\Delta_\mu = (4, 4)$ for a high separation between the two clusters (see Figure 5.4 for the visualization). We introduce missing values with a MNARz model (see (5.12)). In particular,

$$\forall j \in \{1, 2\},\ \mathbb{P}(c_{ij} = 1 | z_{i1} = 1, y_i) = \Phi(\alpha_1), \mathbb{P}(c_{ij} = 1 | z_{i2} = 1, y_i) = \Phi(\alpha_2)$$

with $\alpha = (\alpha_1, \alpha_2) = (-2, -0.2)$ and Φ the cumulative distribution function of the standard Gaussian distribution. It leads to 25% of missing values in the whole dataset: 3% of missing values in the first class and 35% in the second class. Thus, the mechanism provides some information on the clustering. In Figure 5.5, it has been illustrated numerically. Indeed, we compute the ARI with the true parameters (π, θ, ψ): (i) by considering the triplet (Y, Z, C) i.e. by computing the estimator of the partition as in (5.34) or (ii) by ignoring the missing-data pattern C and only considering (Y, Z), i.e. computing the estimator of the partition with

$$\tilde{z}_{ik}^{\text{MAP}}(\pi, \theta) = \underset{k\in\{1,\dots,K\}}{\text{argmax}} \{\tau_{ik}(\pi, \theta) = \mathbb{P}(z_{ik} = 1 | y_i^{\text{obs}}; \pi, \theta)\}.$$

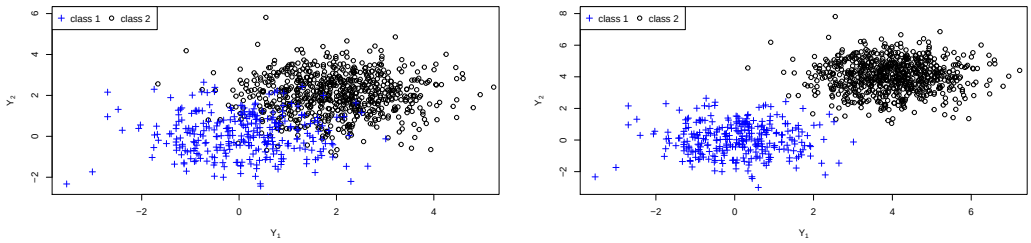


Figure 5.4: Scatterplot of the data points for a low separation between the classes when $\Delta_\mu = (2, 2)$ (left graphic) and a high separation between the classes when $\Delta_\mu = (4, 4)$ (right graphic).

Cluster separation influence In Figure 5.5, the algorithms are compared in terms of ARI in the case where the classes are well separated ($\Delta_\mu = (4, 4)$) or not ($\Delta_\mu = (2, 2)$). As expected, all the methods give better performances when the classes are well separated. In both cases, the two steps-heuristics seem not appropriate for the classification task. Note also that they rely on no theoretical guarantees at all. In the case of clear separation between the clusters ($\Delta_\mu = (4, 4)$), the MNAR models give similar performances and all outperform the MCAR model. Note that the relatively low improvement was expected because the theoretical ARI, taking into account the mechanism, does not increase so much compared to the theoretical ARI without account for it. In addition, note that the MCAR model that we consider is specific since it needs to model c in order to be able compare it with alternative MNAR models. It is not a general view of a MCAR situation, that would not include any modeling of c. In the case where the classes are not well separated ($\Delta_\mu = (2, 2)$),

the MCAR model clearly gives poorer performances than the MNAR models. Note that the MNARz and MNARyz models, which model the effect of missingness depending on the class membership, have performances close to the ARI obtained with the true parameters (by considering the triplet (Y, Z, C)). The MNARy model has poorer performances but it still outperforms the MCAR model and the two-step heuristics. This can be explained by the fact that it only takes into account the effect of the absence of data as a function of the variables.

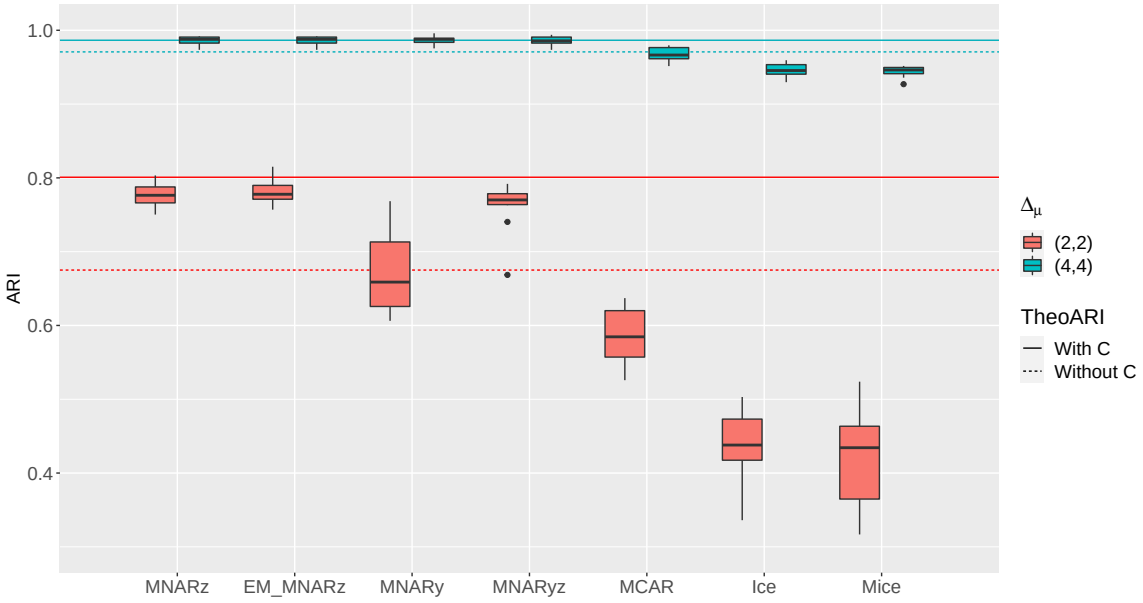


Figure 5.5: Comparison of the different algorithms in terms of ARI for bivariate Gaussian data ($n = 2000$, $d = 2$) and two classes when the true missing-data mechanism is MNARz. The red boxplot corresponds to $\Delta_\mu = (2, 2)$ when the classes are not well separated. The blue boxplot corresponds to $\Delta_\mu = (4, 4)$ when the classes are well separated. The stochasticity comes from the process of drawing 10 times the triplet (Y, Z, C). The plain and dashed lines correspond to the mean of the ARI computed with the true parameters by considering (Y, Z, C) or (Y, Z) respectively.

Automatic choice of the number of cluster In our algorithms, the number of clusters is considered known, but it can be automatically chosen by using the ICL criterion. The algorithms are performed with several values of the number of clusters $K = 1, 2, 3, 4$. The clusters number for the model with the highest ICL is then chosen. To our knowledge, no method propose an automatic choice of the number of clusters in unsupervised classification for the two-step heuristics, which is also a major drawback. Therefore, only the MNAR and the MCAR models are compared. In Figure 5.6, in the case where the classes are not well separated, only the MNARz models selects the true number of clusters $K = 2$. The other models select $K = 1$. In Figure 5.7, when the classes are well separated, all the models select

the good number of clusters $K = 2$.

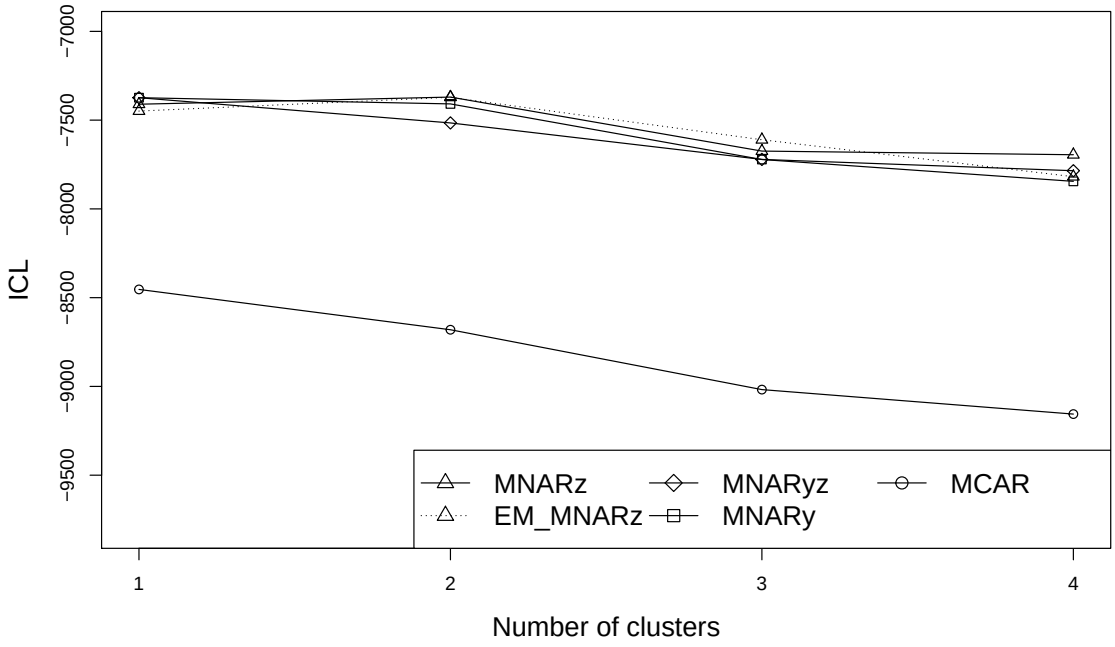


Figure 5.6: Comparison of the different algorithms in terms of ICL for bivariate Gaussian data ($n = 2000$, $d = 2$) and two classes when the true missing-data mechanism is MNARz and $\Delta_\mu = (2, 2)$. The algorithms have been performed for $K = 1, 2, 3, 4$. The dots are the means for the process of drawing 10 times the triplet (Y, Z, C).

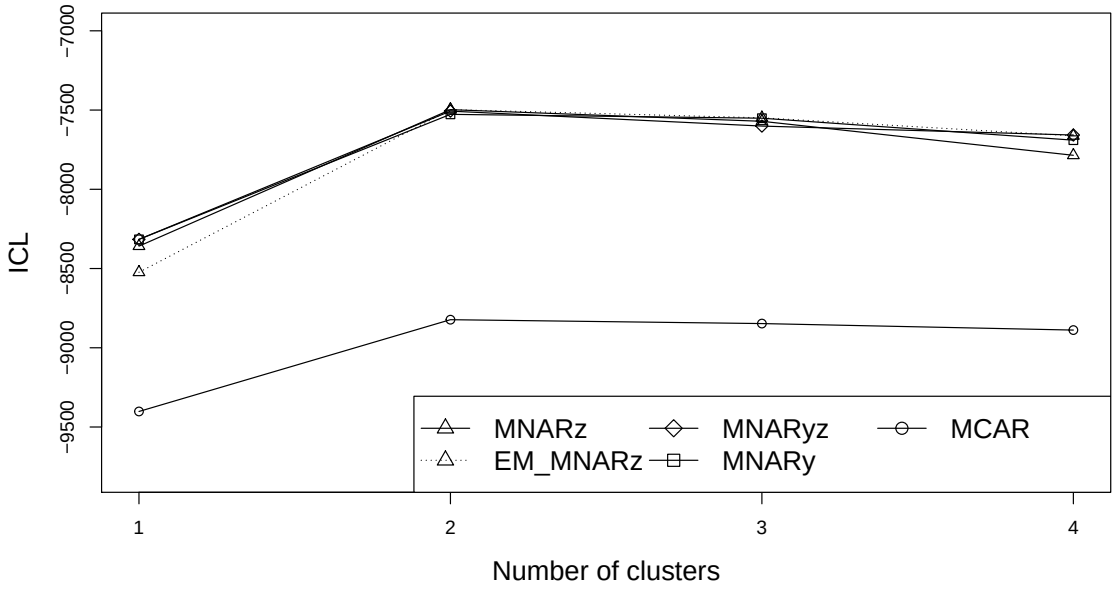


Figure 5.7: Comparison of the different algorithms in terms of ICL for bivariate Gaussian data ($n = 2000$, $d = 2$) and two classes when the true missing-data mechanism is MNARz and $\Delta_\mu = (4, 4)$. The algorithms have been performed for $K = 1, 2, 3, 4$. The dots are the means for the process of drawing 10 times the triplet (Y, Z, C).

Part III

Platform on missing values

Chapter 6

R-misstastic

This chapter is an ongoing work, led by Imke Mayer and carried out in collaboration with Julie Josse, Nicholas Tierney and Nathalie Vialaneix. I have contributed to the whole platform but mostly worked on the workflows.

Abstract

Missing values are unavoidable when working with data. Their occurrence is exacerbated as more data from different sources become available. However, most statistical models and visualization methods require complete data, and improper handling of missing data results in information loss, or biased analyses. Since the seminal work of Rubin (1976), there has been a burgeoning literature on missing values with heterogeneous aims and motivations. This has resulted in the development of various methods, formalizations, and tools (including a large number of R packages and Python modules). However, for practitioners, it remains challenging to decide which method is most suited for their problem, partially because handling missing data is still not a topic systematically covered in statistics or data science curricula.

To help address this challenge, we have launched a unified platform: "R-miss-tastic", which aims to provide an overview of standard missing values problems, methods, how to handle them in analyses, and relevant implementations of methodologies. Several pipelines in R and Python allow for an hands-on illustration of how to handle missing values in various statistical tasks such as estimation and prediction, while ensuring reproducibility of the analyses. The objective of this work is not only to collect, but also comprehensively organize materials, to create standard analysis workflows, and to unify the community. These overviews are suited for beginners, students, more advanced analysts and researchers.

Contents

6.1 Context and motivation

Missing data are unavoidable as soon as collecting or acquiring data is involved. They occur for many reasons including: individuals choose not to answer survey questions, measurement devices fail, or data have simply not been recorded. Their presence becomes even more important as data are now obtained at increasing velocity and volume, and from heterogeneous sources not originally designed to be analyzed together. As pointed out by Zhu et al. (2019), "one of the ironies of working with Big Data is that missing data play an ever more significant role, and often present serious difficulties for analysis". Despite this, the approach most commonly implemented by default in software is to toss out cases with missing values. At best, this is inefficient because it wastes information from the partially observed cases. At worst, it results in biased estimates, particularly when the distributions of the missing values are systematically different from those of the observed values (e.g., Enders, 2010, Chap. 2).

However, handling missing data in a more efficient and relevant way (than limiting the analysis on solely the complete cases) has attracted a lot of attention in the literature in the last two decades. In particular, a number of reference books have been published (Schafer and Graham, 2002; Little and Rubin, 2019; Van Buuren, 2018; Carpenter and Kenward, 2012) and the topic is an active field of research (Josse and Reiter, 2018). The diversity

of the problems of missing data means there is great variety in the methods proposed and studied. They include model-based approaches, integrating likelihoods or posterior distributions over missing values, filling in missing values in a realistic way with single, or multiple imputations, or weighting approaches, appealing to ideas from the design-based literature in survey sampling. The multiplicity of the available solutions makes sense because there is no single solution or tool to manage missing data: the appropriate methodology to handle them depends on many features, such as the objective of the analysis, type of data, the type of missing data and their pattern. Some of these methods are available in various and heterogeneous software solutions. As R (R Core Team, 2020) is one of the main softwares for statisticians and data scientists and as its development has started almost three decades ago (Ihaka, 1998), R is one of the language that offers the largest number of implemented approaches. This is also due to its ease to incorporate new methods and its modular packaging system. Currently, there are over 270 R packages on CRAN that mention missing data or imputation in their DESCRIPTION files. These packages serve many different applications, data types or types of analysis. More precisely, exploratory and visualization tools for missing data are available in packages like **naniar**, **VIM**, and **MissingDataGUI** (Tierney et al.; Tierney and Cook, 2018; Kowarik and Templ, 2016; Cheng et al., 2015). Imputation methods are included in packages like **mice**, **Amelia**, and **mi** (Buuren and Groothuis-Oudshoorn, 2010; Honaker et al., 2011; Gelman and Hill, 2011). Other packages focus on dealing with complex, heterogeneous (categorical, quantitative, ordinal variables) data or with large dimension multi-level data, such as **missMDA**, and **MixedDataImpute** (Josse et al., 2016a; Murray and Reiter, 2016). To our knowledge, R is the programming language offering the largest variety of implemented methods. However, other languages such as Python (Van Rossum and Drake, 2009), which currently only have few publicly available implementations of methods that handle missing values in statistical tasks, offer more and more solutions. Two prominent examples are: 1) the **scikit-learn** library (Pedregosa et al., 2011) which has recently added a module for missing values imputation; and 2) the **DataWig** library (Biessmann et al., 2018) which provides a framework to learn to impute incomplete data tables.

The rich variety of methods and tools for working with missing data means there are many solutions for a variety of applications. Despite the number of options, missing data are often not handled appropriately by practitioners. This may be for a few reasons. First, the plethora of options can be a double-edged sword - the sheer number of options making it challenging to navigate and find the best one. Second, the topic of missing data is often itself missing from many statistics and data science syllabuses, despite its relative omnipresence in data. So then, faced with missing data, practitioners are left powerless: quite possibly never having been taught about missing data, they do not have an idea of how to approach the problem, what are the dangers of mismanagement, navigate the methods, software, or choose the appropriate method or workflow for their problem.

To help promote better management and understanding of missing data, we have released R-miss-tastic, an open platform for missing values. The platform takes the form of a reference website `https://rmisstastic.netlify.com/`, which collects, organizes and produces material on missing data. It has been conceived by an infrastructure steering

committee (ISC; its members are authors of this article) working group, which first provided a CRAN Task View[1] on missing data [2] that lists and organizes existing R packages on the topic. The "R-miss-tastic" platform extends and builds on the CRAN Task View by collecting and organizing articles, tutorials, documentation, and workflows for analyses with missing data. The platform provides new tutorials, examples and pipelines of analyses we have developed with missing data that span the entirety of an analysis. The latter have been developed by us for this platform, implementing standard methods for generating missing values and analyzing them under different perspectives. Starting from data preparation, these pipelines also cover exploratory data analyses, establishing statistical models, analysis diagnostics, applying machine learning methods, through to communication of the results obtained from incomplete data. This platform also references publicly available datasets that are commonly used as benchmark for new missing values methodologies. It is easily extendable and well documented, so it can seamlessly incorporate future research in missing values. The intent of the platform is to foster a welcoming community, within and beyond the R community . "R-miss-tastic" has been designed to be accessible for a wide audience with different levels of prior knowledge, needs, and questions. This includes, for instance, students looking for course material to complement their studies, teachers and professors who can use a reference website for their own classes or refer to students, statisticians or researchers in a different fields using statistics searching for solutions or existing work to help with analysis, researchers wishing to understand or contribute information for specific research questions, or find collaborators.

The remainder of the article is organized as follows: Section 6.2 describes the different components of the platform, the structure that has been chosen, and the targeted audience. The section is organized as the platform itself, starting by describing materials for less advanced users then materials for researchers and finally resources for practical implementation. Section 6.3 details the implementation and use-cases of the provided workflows. Finally, in Section 6.4 we conclude with an overview of the planed future development for the platform.

6.2 Structure and content of the platform

The R-miss-tastic platform is released at `https://rmisstastic.netlify.com/`. It has been developed using the R package **blogdown** Xie et al. (2017) which wraps hugo[3]. Live examples have been included using the tool `https://rdrr.io/snippets/` provided by the website "R Package Documentation". The source code and materials of the platform have been made publicly available on GitHub[4], which provides a transparent record of the platform's development, and facilitates community contributions.

We now discuss the structure of the R-miss-tastic platform, the aim and content of each subsection, and highlight key features of the platform.

[1] `https://CRAN.R-project.org/package=ctv`
[2] `https://cran.r-project.org/web/views/MissingData.html`
[3] `https://gohugo.io/`
[4] repository `R-miss-tastic/website`

6.2.1 Workflows

An important contribution and novelty of this work is the proposal of several workflows that allow for a hands-on illustration of classical analyses with missing values, both on simulated data and on publicly available real-world data. These workflows are provided both in R and in Python code and cover the following topics:

- *How to generate missing values?* Generate missing values under different mechanisms, on complete or incomplete datasets. This is useful when performing simulations to compare methods.
- *How to do statistical inference with missing values?* We focus in particular in the different solutions (maximum likelihood or multiple imputation) that are available to estimate linear and logistic regression parameters with missing values in the covariates.
- *How to impute missing values?* We compare different single imputation/matrix completion methods (for instance using conditional models, low-rank models, etc.)
- *How to predict with missing values?* We consider establishing predictive models (for instance using random forests (Breiman, 2001)) on data with incomplete predictors. The workflows present different strategies to deal with the missing values in the covariates both in the train set and in the test set.

The aim of these workflows is threefold: 1) they provide a practical implementation of concepts and methods discussed in the lectures and bibliography sections (see Sections 6.2.2 and 6.2.3); 2) they are coded in a generic way, allowing for simple re-use on other datasets, for integration of other estimation or imputation methods; 3) the distinction between inference, imputation, and prediction lets the user keep in mind that the solutions are not the same in these cases.

Furthermore, they allow for a transparent and open discussion about the proposed implementations which can be followed on the project GitHub repository[5].

We provide a more detailed view on the proposed workflows in Section 6.3, giving code examples and their corresponding tabular or graphical outputs.

6.2.2 Lectures

Before starting to use any of the existing implementations for handling missing values in a statistical analysis, it is essential to understand why missing values are problematic. There are many lenses to view missing data through: the source of the missing values, their potential meaning, the relevance of the features they occur in – and the implications for different types of analyses. For someone unfamiliar with missing data, it is a challenge to know where to begin the journey of understanding them, and the methods to address them. This challenge is being addressed with "R-miss-tastic", which makes the material to get started easily accessible.

[5] https://github.com/R-miss-tastic/website

Teaching and workshop material takes many forms – from slides, course notes, lab workshops, video tutorials and in-depth seminars. The material is of high quality, and has been generously contributed by numerous renowned researchers who investigate the problems of missing values, many of whom are professors having designed introductory and advanced classes for statistical analysis with missing data. This makes the material on the "R-miss-tastic" platform well suited for both beginners and more experienced users.

These teaching and workshop materials are described as "lectures", and are organized into five sections:

1. General lectures: introduction to statistical analyses with missing values, theory and concepts are covered, such as missing values mechanism, likelihood methods, imputation.

2. Multiple imputation: introduction to popular methods of multiple imputation (joint modeling and fully conditional), how to correctly perform multiple imputation and limits of imputation methods.

3. Principal component methods: introduction to methods exploiting low-rank type structures in the data for visualization, imputation and estimation

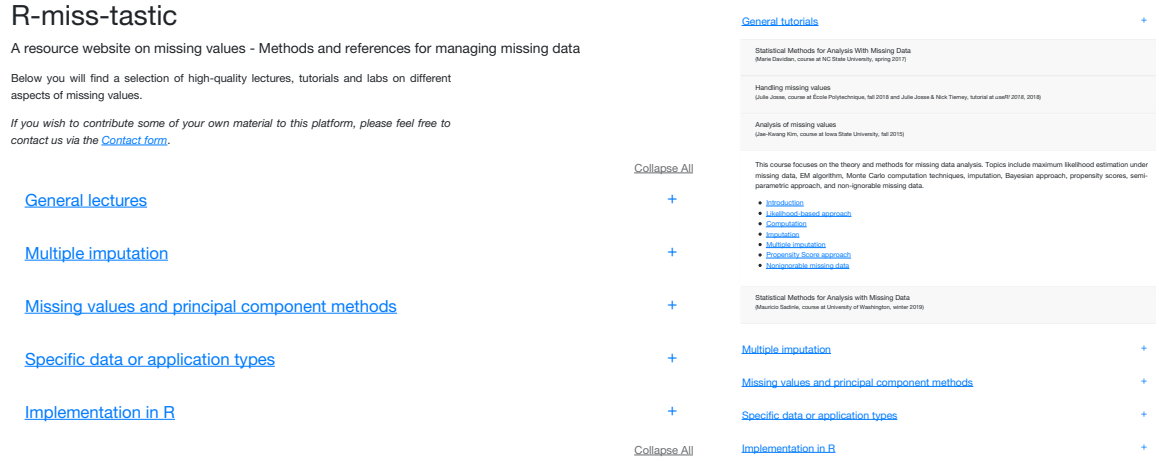
4. Specific data or applications types: lectures covering in detail various sub-problems such as missing values in time series, in surveys, or in treatment effect estimation. Indeed, certain data types require adaptations of standard missing values methods (for instance handling the time dependence in time series (Moritz and Bartz-Beielstein, 2017)) or additional assumptions about the impact of missing values (such as the impact on confounded treatment effects in the causal inference context (Mayer et al., 2020)). But also more in-depth material, for instance video recordings from a virtual workshop on *Missing Data Challenges in Computation, Statistics and Applications*[6] held in 2020.

5. Implementations: a non-exhaustive list of detailed vignettes describing functionalities of R packages and of Python modules that implement some of the statistical analysis methods covered in the other lectures. For instance the functionalities and possible applications of the **missMDA** R package are presented in a succinct summary, allowing the reader to compare the main differences between this package and the **mice** package which is also summarized using the same summary format.

Figure 6.1 is a screenshot of two views of the lectures page: Figure 6.1A shows a collapsed view presenting the different topics, Figure 6.1B shows an example of the expanded view of one topic (General tutorials), with a detailed description of one of the lecture (obtained by clicking on its title), "Analysis of missing values" by Jae-Kwang Kim. Each lecture can contain several documents (as is the case for this one) and is briefly described by a header presenting its purpose.

Lectures that we found very complete and thus recommend are:

[6] https://www.ias.edu/math/mdccsa



(a) Collapsed view (b) Extract

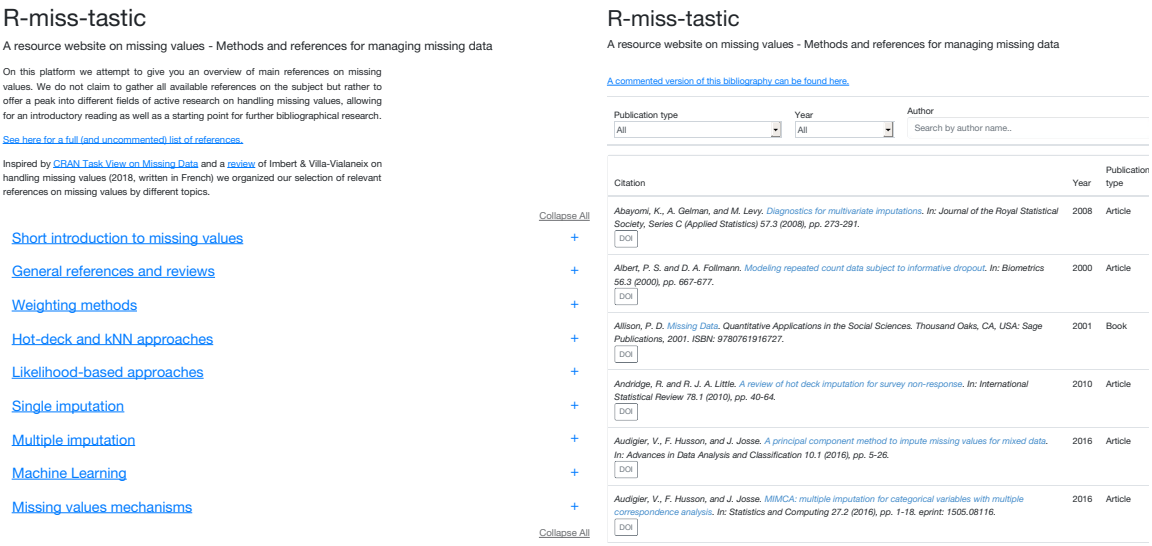
Figure 6.1: Lectures overview.

- *Statistical Methods for Analysis with Missing Data* by Mauricio Sadinle (in "General tutorials")
- *Missing Values in Clinical Research – Multiple Imputation* by Nicole Erler (in "Multiple imputation")
- *Handling missing values in PCA and MCA* by François Husson. (in "Missing values and principal component methods")

6.2.3 Bibliography

Complementary to the lectures section, this part of the platform serves as a broad overview on the scientific literature discussing missing values taxonomies and mechanisms and statistical, machine learning methods to handle them. This overview covers both classical references with books, articles, etc. such as Schafer and Graham (2002); Little and Rubin (2019); Van Buuren (2018); Carpenter and Kenward (2012) and more recent developments such as Josse et al. (2019); Gondara and Wang (2018) who study the consistency of supervised learnning with missing values. The entire (non-exhaustive) bibliography can be browsed in two ways: 1) a complete list, filtering by publication type and year, with a search option for the authors or 2), as a contextualized version. For 2), we classified the references into several domains of research or application, briefly discussing important aspects of each domain. This double representation is shown in Figure 6.2 and allows an extensive search in the existing literature, while providing some guidance for those focused on a specific topic. All references are also collected in a unique BibTex file made available on the GitHub repository[7]. This

[7] in `resources/rmisstastic_biblio.bib`



(a) Contextualized version (b) Unordered version

Figure 6.2: Bibliography overview.

centralized file allows external users to easily propose additions to the bibliography which are then reviewed by the platform maintenance team, composed of researchers with different focuses on the handling of missing values.

6.2.4 Implementations

R packages As mentioned in the introduction, prior to the platform development, the project started with the release of the *MissingData* CRAN Task View, which currently lists approximately 150 R packages. The CRAN Task View is continuously updated, adding new R packages, and removing obsolete ones. Packages are organized by topic: *exploration of missing data*, *likelihood based approaches*, *single imputation*, *multiple imputation*, *weighting methods*, *specific types of data*, *specific application fields*. We selected only those that were sufficiently mature and stable, and that were already published on CRAN or Bioconductor. This choice was made to ensure all listed packages can easily be installed and used by practitioners.

Despite the Task View classifying packages into different sub-domains, it may still be a challenge for practitioners and researchers inexperienced with missing values to choose the right package for the right application. To address this challenge, we provide a partial, less condensed overview on existing R packages on the platform, choosing the most popular and versatile. This overview is a blend of the Task View, and the individual package description pages and vignettes. For each selected package (7 at the date of writing of this article, namely **imputeTS**, **mice**, **missForest**, **missMDA**, **naniar**, **simputation** and **VIM**), we provided a category (in the style of the categorization in the Task View), a short description of

use-cases, its description (as on CRAN), the usual CRAN statistics (number of monthly downloads, last update), the handled data formats (e.g., `data.frame`, `matrix`, `vector`), a list of implemented algorithms (e.g., k-means, PCA, decision tree), and the list of available datasets, some references (such as articles and books) and a small chunk of code, ready for a direct execution on the platform via the *R package Documentation*[8]. Figure 6.3 shows the condensed view of the package page and the expanded description sheet of a given package (here **naniar**).

We believe shortlisting R packages is especially useful for practitioners new to the field, as it demonstrates data analysis that handles missing values in the data. We are aware that this selection is subjective, and welcome external suggestions for other packages to add to this shortlist.

Python modules To the best of our knowledge, there only exist very few implemented methods for handling missing data in Python. However, one of the major libraries for machine learning and data analysis, **scikit-learn** (Pedregosa et al., 2011) has recently proposed a module for simple and multiple imputations, **sklearn.impute**. And, as an alternative, the **statsmodels**[9] library now also has an implementation module for multiple imputation in Python. We regularly survey new Python implementations for handling missing values and, if pertinent from a theoretical and practical point of view, reference them on our platform. We expect this to promote their use but also additional assessment by practitioners and researchers from the missing values (statistics/machine learning) community.

6.2.5 Datasets

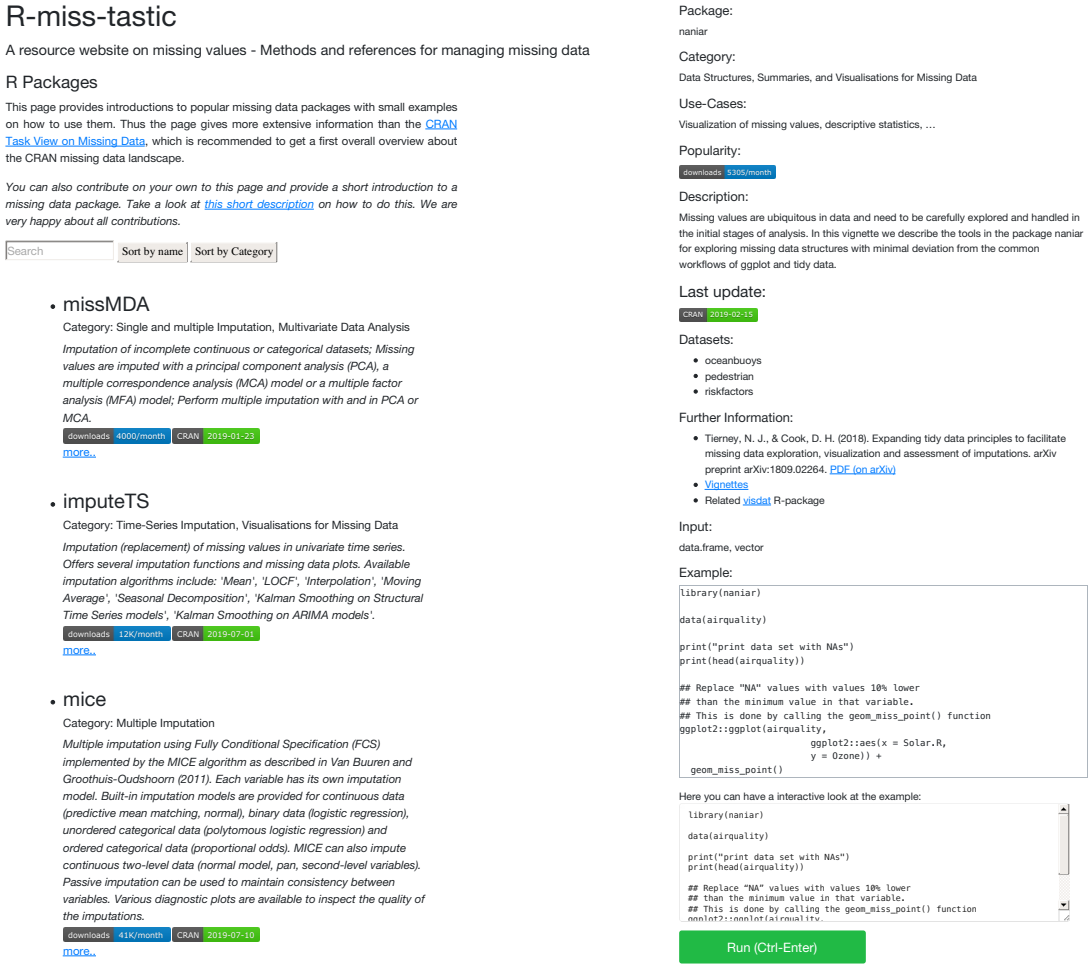
Especially in methodology research, an important aspect is the comparison of different methods to assess the respective strengths and weaknesses. There are several datasets that are recurrent in the missing values literature but, they have not been listed together yet. We gathered publicly available datasets that have been used recurrently for comparison or illustration purposes in publications, R packages and tutorials. Most of these datasets are already included in R packages but some are available on other data collections. Figure 6.4 shows how the datasets are presented, with a detailed description shown for one of the dataset ("Ozone", obtained by clicking on its name). The description follows the UCI Machine Learning Repository presentation (Dua and Graff, 2019), including a short description of the dataset, how to obtain it, external references describing the dataset in more details, and links to tutorials/lectures on our websites or to vignettes in R packages that use the dataset.

In addition, the Datasets section also references existing methods for generating missing data, given assumptions on their generation mechanisms (as in the R package **mice**). These datasets also serve as benchmark in the proposed workflows and allow for a fair and transparent comparison between the different methods.

Note however, that the list of datasets we gather here is comparatively short when compared to benchmark datasets for full data methods such as the UCI Machine Learning

[8] https://rdrr.io/snippets/

[9] https://www.statsmodels.org/stable/about.html



(a) Extract of global view (b) Description sheet

Figure 6.3: R packages overview.

Repository. Therefore, our proposed list also serves as an invitation to tackle this lack of a wider variety of common benchmark datasets in the missing data community.

6.2.6 Additional content

This unified platform collects and edits the contributions of numerous individuals who have investigated the missing values problems, and developed methods to handle them for many years. To provide an overview of some of the main actors in this field, the list of all contributors who agreed to appear on this platform is given with links to their personal or to their research lab website.

We also provide links to other interesting websites or working groups, not necessarily working with R and Python (Van Rossum and Drake, 2009) but with other programming languages such as SAS/STAT® and STATA (StataCorp., 2019).

The platform also provides two other features to engage the community:

1. a regularly updated list of events such as conferences or summer schools with special focus on missing values problems, and

2. a list of recurring questions together with short answers and links for more details for every question.

6.3 Workflows

After this general introduction to the R-miss-tastic project and platform and the overview of its structure, we now turn to a more detailed presentation of the various workflows we have developed and proposed on this platform.

To allow for both hands-on tutorials illustrating current practices and state-of-the-art and ready-to-use pipelines, we propose the workflows under different formats such as HTML, PDF, R Markdown (for R code) and IPYNB (for Python code). We encourage practitioners and researchers to use and adapt these workflows, in order to increase reproducibility and comparability of their work. Of course, we are aware that these workflows do not cover the entire spectrum of existing methods and data problems. The goal of the proposed workflows is rather to initiate a joint effort to create a larger spectrum of open-source workflows, and to encourage the use of standardized procedures to handle missing values.

With an incomplete dataset at hand, prior to embarking on an in-depth statistical analysis, a specific aim has to be defined in order to choose a specific method to use. An example of a method whose success crucially depends on the analyst's goal is *mean imputation*: this approach is strongly contra-indicated if the aim is to estimate parameters, but it can be consistent if the aim is to predict as well as possible (Josse et al., 2019). Following this observation, our workflows are divided into different parts, defined by the objective of the statistical analysis. We have tried to present and compare the main implementations available both in R and Python for each objective. Currently there are seven workflows available on the platform and we present their scope and use below.

Incomplete data

The data sets listed below are either widely used in general in the missing data community or used for illustration of different methods handling missing values in the tutorials from the Tutorials and R packages sections. This presentation scheme is inspired by the UCI Machine Learning Repository.

Click on a table entry to obtain further information about the data set.

Name	Data Types	Attribute Types	# Instances	# Attributes	% Missing entries	Complete data available	Year
Airquality	Multivariate, Time Series	Real	154	6	7	No	1973
chorizonDL	Multivariate	Integer, Real	606	110	15	Yes	1998
Health Nutrition And Population Statistics	Multivariate, Time Series	Integer, Real	15,022	397	54	No	2017
NHANES	Multivariate	Categorical, Integer, Real	10,000	75	37	No	2012
oceanbuoys	Multivariate, Time Series	Real	736	8	3	No	1997
Ozone	Multivariate	Categorical, Integer, Real	366	13	6	No	1976

Los Angeles Ozone Pollution Data, 1976. This data set contains daily measurements of ozone concentration and meteorological quantities. It can be found in R in the `mlbench` package and is loaded by calling `data(Ozone)`.

More information on the dataset.

Tutorials illustrating methods on this data:

- Julie Josse's course on missing values imputation using PC methods.
- Julie Josse's and Nick Tierney's tutorial on handling missing values. Download the data set from this tutorial: ozoneNA.csv
- Nick Tierney's `naniar` vignette for missing data visualization.

pedestrian	Multivariate, Time series	Categorical, Integer	37,700	9	2	No	2016

Figure 6.4: Datasets overview.

6.3.1 How to generate missing values?

Rubin (1976) classifies the cause for a lack of data into three missing data mechanisms. The missing data mechanism is said to be: (i) missing completely at random (MCAR) if the lack of data is totally independent of the data values, (ii) missing at random (MAR) if the process that causes the missing data may only depend on the observed values and (iii) missing not at random (MNAR) if the unavailability of the data depends on the missing variables, for instance on their values themselves and possibly on the values of observed variables. More formally, let us define the indicator matrix $R \in \mathbb{R}^{n \times p}$, which indicates the indices of the observed values in $X \in \mathbb{R}^{n \times p}$, i.e., $R_{ij} = 1$ if X_{ij} is observed and $R_{ij} = 0$ otherwise. The missing data mechanism is then the distribution of the indicator matrix R given the data X.

The goal of this workflow is to propose functions to generate missing values under these different mechanisms. The way in which the missing values are generated is crucial for comparing methods (and studies) in a fair manner and is often subject of debate (Seaman et al., 2013). This code aims to unify classical solutions to do this. Indeed, a usual strategy to compare imputation or estimation strategies is to introduce (additional) missing values in the dataset, and use the ground truth for these missing values to evaluate the strategies.

In R In the R workflow, the main function `produce_NA` allows to generate missing values under the three missing-data mechanisms outlined above. This function internally calls the `ampute` function of the **mice** package (Buuren and Groothuis-Oudshoorn, 2010) but we chose to simplify its use while adding some additional options to specify the missing values generation. In addition, the original `ampute` function generates missing values only for complete dataset. In our workflow, the user can easily introduce (additional) missing values in a complete or incomplete dataset composed of quantitative, categorical or mixed variables. The three main arguments are the initial dataset (`data`) in which the missing values are introduced using a given missing data mechanism (`mechanism`) and a given percentage of missing values (`perc.missing`). For example, to introduce 20% of MCAR values in the dataset X, the code is detailed below.

```
X.miss.mcar <- produce_NA(data = X,
                          mechanism = "MCAR",
                          perc.missing = 0.2)
X.mcar <- X.miss.mcar$data.incomp
#incomplete matrix containing (additional) missing values
R.mcar <- X.miss.mcar$idx_newNA # indicator matrix
```

Listing 6.1: Generating MCAR missing values in R.

The function returns the data matrix containing the new missing values (and the previously present missing values of the input data) and the indicator matrix R.

To introduce missing values under the MAR and MNAR mechanisms, there are several options detailed and illustrated in the workflow. We consider the definitions of the missing data mechanisms of Rubin (1976). For instance, if X contains three variables denoted as X_1, X_2, X_3, two options are available to generate MAR values:

1. the first option consists of generating missing values in X_1 by using a logistic model depending on (X_2, X_3), which are observed, i.e.

$$\mathbb{P}(R_1 = 0|X; \phi) = 1/(1 + \exp(-(\phi_2 X_2 + \phi_3 X_3)),$$

where $\phi = (\phi_2, \phi_3)$ is the parameter of the missing-data mechanism. In our function, ϕ is chosen so that the given percentage of missing values is reached. This allows to obtain missing values in the first variable X_1^{NA}. Then, the same strategy is performed to introduce missing values in X_2 and X_3, by using a logistic model depending on (X_1, X_3) (which are observed) and (X_1, X_2) (which are observed) respectively. To get the final matrix containing missing values, we concatenate X_1^{NA}, X_2^{NA} and X_3^{NA} by handling the rows containing only missing values. To use this option, the code is detailed below.

```
X.miss.mar <- produce_NA(data = X, mechanism = "MAR",
                         perc.missing = 0.2, by.patterns = FALSE)
```

Listing 6.2: Generating MAR values in R.

2. the second option consists in generating the missing values *by pattern*, i.e., by rows. In this case, the combinations of which variables are observed and missing are specified in a pattern matrix. For the MAR mechanism, at least one variable must be observed. An example (the choice by default) of such pattern matrix is

$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix},$$

where 0 indicates that the variable should have missing values whereas 1 means that it should be observed. For example, the first pattern means that the observation has a missing values in the first variable X_1 which depends on the values of X_2 and X_3 which are observed. The code below allows to choose this option.

```
X.miss.marpat <- produce_NA(data = X, mechanism = "MAR",
                            perc.missing = 0.2, by.patterns = TRUE)
```

Listing 6.3: Generating MAR values by patterns in R.

We propose several ways to generate missing values, under the MNAR mechanism, including the most general one when the missingness depends on both the missing variables and the observed variables. A particular case of a MNAR mechanism that we consider is the self-masked one, if the unavailability of the data only depends on their values themselves. The following code allows to introduce such missing values using a quantile censorship for which the form is precised by the argument `self.mask`. The argument `idx.incomplete` gives the indexes of the variables for which self-masked MNAR values should be introduced.

```
X.miss.mnar <- produce_NA(X.complete, mechanism = "MNAR",
                          perc.missing = 0.2, self.mask = "lower",
                          idx.incomplete = c(1,1,1,1))
```

Listing 6.4: Generating self-masked MNAR values in R.

The choice `self.mask = "lower"` in the above example specifies that the values are amputed based on a quantile from the lower tail of the empirical distribution chosen such that the target proportion of missing values is achieved.

In Python To our knowledge, there does not exist a specific module in Python to generate missing values. The workflow we present now has therefore been developed in collaboration with the principal author Boris Muzellec of the paper Muzellec et al. (2020) which suggests a single imputation method based on optimal transport. To stay close to the R workflow, we propose a function `produce_NA` which allows to easily generate missing values for a specific missing-data mechanisms (`mecha`) and a given percentage of missing values (`p_miss`) in a complete dataset (`X`), currently only allowed to contain quantitative variables. If the aim is to introduce 20% of MCAR values, the following code can be used.

```
X_miss_mcar = produce_NA(X = X, p_miss = 0.2, mecha = "MCAR")
X_mcar = X_miss_mcar['X_incomp']
#incomplete matrix containing missing values
R_mcar = X_miss_mcar['mask'] #indicator matrix
```

Listing 6.5: Generating MCAR values in Python.

The outputs of this function are the incomplete matrix with 20% MCAR values and the indicator matrix R.

For the MAR mechanism, by contrast with the R workflow, the Python code relies on the definition of Little and Rubin (2019). For instance, if we have $X = (X_1, X_2, X_3)$, then at least one variable should be fully-observed (say X_3) and missing values in X_1 and X_2 are introduced with the following logistic model,

$$\mathbb{P}(R_1 = 0|X; \phi) = \mathbb{P}(R_2 = 0|X; \phi) = 1/(1 + \exp(-\phi_3 X_3),$$

with $\phi = \phi_3$ the parameter of the missing-data mechanism chosen to reach the given percentage of missing values. To introduce 20% missing values in each missing variable (i.e., X_1 and X_2), the following code can be used, with `p_obs` the proportion of fully observed variables.

```
X_miss_mar = produce_NA(X=X, p_miss=0.2,
                        mecha="MAR", p_obs=0.3)
```

Listing 6.6: Generating MAR values in Python.

For the MNAR mechanism, three main options are available, using a logistic model, a quantile censorship or a logistic model for a self-masked mechanism (for their exact definition, we refer to the workflow). For example, to introduce 20% of self-masked missing values (in all the variables), we can use the code below.

```
X_miss_selfmasked = produce_NA(X=X, p_miss=0.4,
                               mecha="MNAR", opt="selfmasked")
```

Listing 6.7: Generating self-masked MNAR values in Python.

6.3.2 How to impute missing values?

There exists a vast literature on how to impute missing values. The aim of these workflows (in R and Python) is to compare the most classical imputation methods and to propose a reference pipeline for comparison on simulated and real datasets, which can be easily extended with other imputation methods. Different types of methods are included:

1. imputation by the mean, which serves as a naive benchmark.
2. conditional models, if, roughly speaking, the imputation relies on the distributions of each variable given the others.
    - in R:
        - **mice** (Buuren and Groothuis-Oudshoorn, 2010): it allows to compute multiple imputations by chained equations and thus returns several imputed datasets. We use the predictive mean matching method (default method) and aggregate the complete datasets using the mean of the imputations to get a simple imputation.
        - **missForest** (Stekhoven and Bühlmann, 2012): it imputes missing values iteratively by training random forests.
    - in Python:
        - **IterativeImputer** of scikit-learn library (Pedregosa et al., 2011): this function is inspired by mice, but it uses (iterative) regularized imputation using conditional expectation and provides a simple imputation. We also use the **ExtraTreesRegressor** estimator of **IterativeImputer**, which trains iterative random forests.
3. low-rank based models, if the data matrix to impute is assumed to be low rank and the similarities between the variables (or the observations) may inform the imputation,
    - in R:
        - **softImpute** (Hastie et al., 2015): it minimizes the reweighted least squares error penalized by the nuclear norm.
        - **missMDA** (Josse et al., 2016a): it minimizes the reweighted least squares error penalized by a mix between the ℓ_2-norm and ℓ_0-norm.
    - in Python: **softImpute** (coded in Python by ourselves).
4. recent methods (for the Python workflow only) using optimal transport or variational autoencoders, variables (or the observations) may inform the imputation,
    - in Python:
        - **MIWAE** (Mattei and Frellsen, 2019): it imputes missing values with a deep latent variable model based on importance weighted variational inference.

– **Sinkhorn** (Muzellec et al., 2020): it extracts randomly several batches and consists in minimizing optimal transport distances between batches to impute missing values.

Other methods such as GAIN (Yoon et al., 2018) which uses generative adversarial nets, have not yet been compared, as they are close to those already being compared, but will be added in the future.

The metric we choose to compare the methods is the mean squared error (MSE), which can be calculated if the ground truth of the missing values is known. More precisely, the procedure is the following one: (i) we have access to a complete dataset X, (ii) missing values are introduced in X and we get an incomplete dataset X^{NA}, (iii) this incomplete dataset is imputed and we obtain an imputed dataset X^{imp}. The MSE for X^{imp} is computed as follows

$$MSE(X^{\text{imp}}, X) = \frac{1}{n_{\text{NA}}} \sum_i \sum_j 1_{\{X_{ij}^{\text{NA}} = \text{NA}\}} (X_{ij}^{\text{imp}} - X_{ij})^2$$

where $n_{\text{NA}} = \sum_i \sum_j 1_{\{X_{ij}^{\text{NA}} = \text{NA}\}}$ is the number of missing entries in X^{NA}. Note that this procedure can also be performed on an incomplete dataset by introducing additional missing values. However, for now, both R and Python notebooks only consider complete datasets.

In R This workflow first presents the main imputation methods available in R, including **mice**, **missForest**, **softImpute** and **missMDA**.

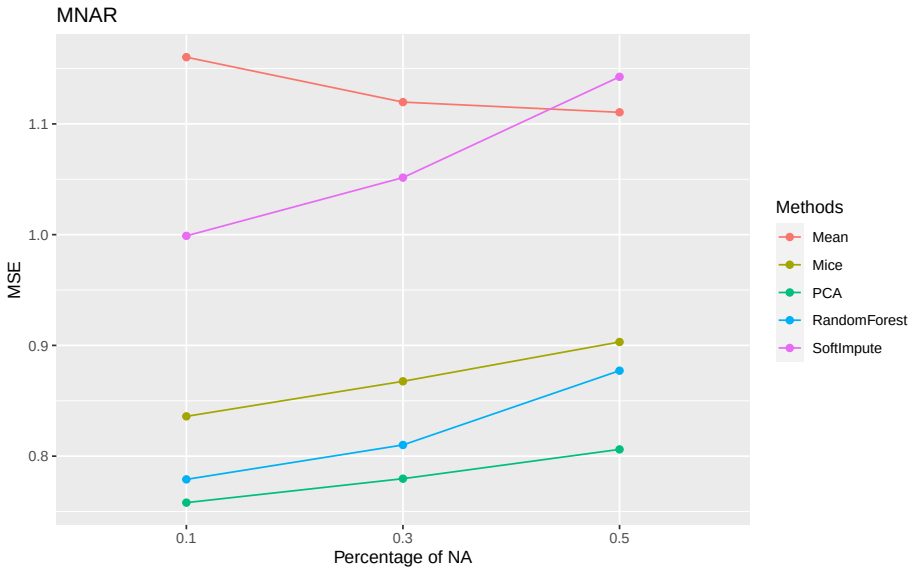
We compare the methods on a simulated dataset $X \in \mathbb{R}^{n \times d}$ under the multivariate Gaussian law $X \sim \mathcal{N}(\mu, \Sigma)$, with μ the mean vector and Σ the covariance matrix. The function `HowtoImpute` compares the imputation methods by introducing missing values in a complete dataset (`X`) using different percentages of missing values (`perc.list`) and missing data mechanisms (`mecha.list`). It returns the mean of the methods' MSEs for the different missing values settings by taking the average over `nbsim` repetitions. The code to use this function is given below. For the sake of clarity, in the workflow, all the code is detailed and commented. The output of this function and its associated plot are shown in Figure 6.5, when $n = 1000, d = 10$, $\mu_i = 1$, $\forall i \in \{1, \dots, d\}$ and $\Sigma_{ij} = 0.5$ if $i \neq j \in \{1, \dots, d\}$ and $\Sigma_{ij} = 1$ if $i = j$. For the MCAR mechanism, the methods perform well, while for the MNAR mechanism, the results are close to those of the naive imputation by the mean. As expected, most methods give worse results for high percentages of missing values.

```
perc.list = c(0.1, 0.3, 0.5)
#list of the percentages of missing values
mecha.list = c("MCAR", "MAR", "MNAR") #list of the missing-data mechanisms
res <- HowToImpute(X = X, perc.list = perc.list,
                   mecha.list = mecha.list, nbsim = 10)
```

Listing 6.8: Code to compare imputation methods for different missing-values settings in R.

	0.1 MCAR	0.3 MCAR	0.5 MCAR	0.1 MAR	0.3 MAR	0.5 MAR	0.1 MNAR	0.3 MNAR	0.5 MNAR
X.pca	0.74	0.76	0.78	0.75	0.78	0.81	0.76	0.78	0.81
X.forest	0.77	0.8	0.86	0.78	0.81	0.87	0.78	0.81	0.88
X.mice	0.82	0.83	0.86	0.83	0.86	0.9	0.84	0.87	0.9
X.soft	0.93	0.86	0.87	0.97	1	1.1	1	1.1	1.1
X.mean	1	0.99	1	1.1	1.1	1.1	1.2	1.1	1.1

(a) Output of the function `HowtoImpute` in R. The results are truncated to two decimals.



(b) Example of plot for the MNAR mechanism (one plot per mechanism).

Figure 6.5: Tabular and graphical output of the R function `HowtoImpute`. The methods **mice**, **missForest**, **softImpute** and **missMDA** are compared with the naive imputation by the mean for several percentage of missing values (10%, 30%, 50%). The mean of the MSEs computed for several introductions of missing values are given. In the tabular, the results are showed for several mechanisms (MCAR, MAR, MNAR) and the plot corresponds to the MNAR mechanism.

We also propose a function `HowToImpute_real` which gives the comparison of the imputation methods for a list of datasets (`datasets_list`) and for a given missing data mechanism (`mech`) and a given percentage of missing values (`perc`). This can be particularly useful for practitioners who would like to have an indication which method might be most suited for a or several specific datasets. This function returns a table containing the mean of the MSEs for the simulations performed and a table for the summary plot showed in Figure 6.6. An example of how to use this function in practice is detailed below. Here, the real datasets are taken from the UCI repository.

```
datasets_list <- list(
                  wine_white = wine_white,
                  wine_red = wine_red,
```

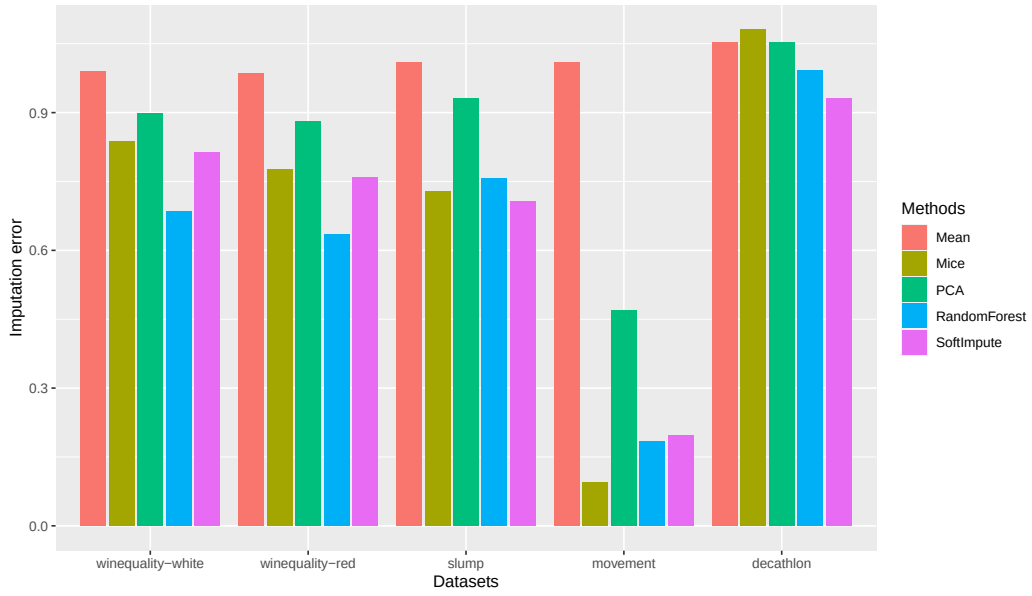


Figure 6.6: Graphical output of the R function `HowToImpute_real`. The methods **mice**, **missForest**, **softImpute** and **missMDA** for several real datasets in which 10% MCAR missing values have been introduced.

```
                    slump = slump,
                    movement = movement,
                    decathlon = decathlon
                   ) # list of different datasets
names_dataset <- c("winequality-white", "winequality-red", "slump",
                    "movement", "decathlon")
                    # names of the different datasets
perc <- 0.2 # percentage of missing values to introduce
mecha <- "MCAR" # missing data mechanism to use
howimp_real <-  HowToImpute_real(
                    datasets_list = datasets_list ,
                    perc = perc,
                    mech = mecha,
                    nbsim = 10,
                    names_dataset = names_dataset
                  )
plotdf_fin <- howimp_real$plot
res <- howimp_real$res
```

Listing 6.9: Code to compare imputation methods for different datasets in R.

In Python The Python workflow is very similar to its R counterpart. The classical imputation methods that we consider are **softImpute, IterativeImputer** and we compare them to very recent approaches using optimal transport with the **Sinkhorn** module and

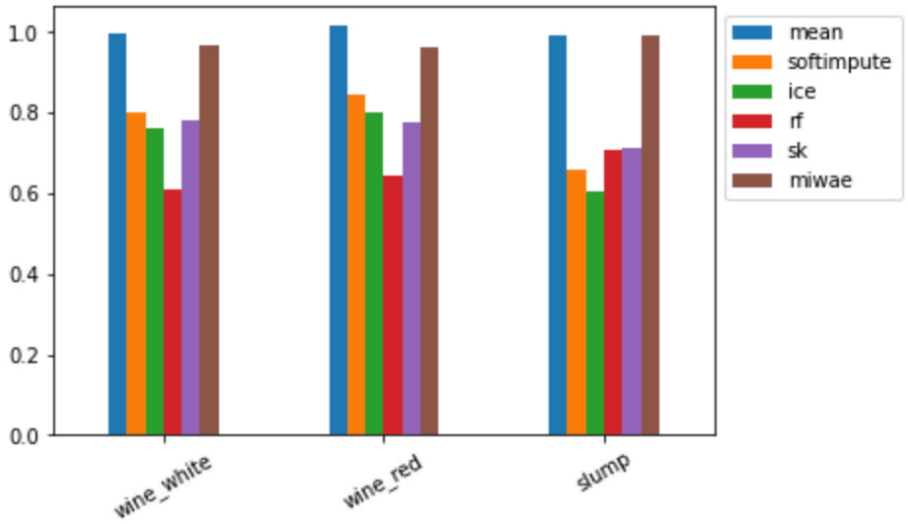


Figure 6.7: Graphical output of the Python function `HowToImpute_real`. The methods **softImpute**, **IterativeImputer**, **Sinkhorn**, **MIWAE** and the imputation by the mean are compared for several real datasets in which 10% MCAR missing values have been introduced.

autoencoders with the **MIWAE** module. The code for the function Python `HowToImpute` is provided below.

```
perc_list = [0.1, 0.3, 0.5] #list of the percentages of missing values
mecha_list = ["MCAR", "MAR", "MNAR"] #list of the missing-data mechanisms
results_how_to_impute = HowToImpute(x_comp=x_comp ,
                                    perc_list=perc_list,
                                    mecha_list=mecha_list , nbsim=10)
```

Listing 6.10: Code to compare imputation methods for different missing-values settings in Python.

Similarly, the following code for the function `HowToImpute_real` in Python can also be used. The graphical output of this code is given in Figure 6.7.

```
datasets_list = dict(wine_white=wine_white, wine_red=wine_red, slump=slump)
    # dictionary of different datasets
names_dataset = ['wine_white','wine_red','slump']
# names of the different datasets
perc = [0.1] # percentage of missing values to introduce
mecha = ["MCAR"] # missing-data mechanism to use
results_how_to_impute_real = HowToImpute_real(
                             datasets_list=datasets_list,
                             perc=perc, mecha=mecha, nbsim=10,
                             names_dataset=names_dataset)
```

Listing 6.11: Code to compare imputation methods for different datasets in Python.

An additional workflow has been written by an external contributor, François Husson (Professor in Statistics, France) and reviewed by us. It specifically compares imputation

methods using variational and denoising autoencoders (Gondara and Wang, 2018; Mattei and Frellsen, 2019; Abiri et al., 2019) with classical methods such as the low-rank based method (Josse et al., 2016a). These deep learning methods often require parameter settings. In this workflow, an automatic tuning is suggested. In addition, the methods are compared for several simulation scenarios, when the variables of the dataset are linearly linked or not. In both cases, deep learning methods do not outperform the low-rank method, although they are known to be able to handle non-linear relationships. This workflow is available on our website.[10]

6.3.3 How to estimate parameters with missing values in R?

This workflow is dedicated to a specific inferential framework when the aim is to estimate linear and logistic regression parameters for multivariate normal data. It is currently only available in R, as there are no analogous implementations available in Python to our knowledge.

In this workflow, two classical methods are compared, using available R implementations: the EM algorithm for logistic and linear regression with the package **misaem** (Jiang et al., 2020) which uses the SAEM algorithm, Stochastic Approximation of EM algorithm (Delyon et al., 1999) and multiple imputation with the package **mice**. Both strategies are valid under the MAR missing data mechanism.

The EM algorithm (Dempster et al., 1977) allows to handle MAR missing values in maximum likelihood estimation by integrating over the missing values distribution, conditionally on the observed values. A drawback of this approach is that it requires a separate derivation of the expectation and maximization steps for each model and data type, such as linear regression and logistic regression on multivariate normal covariates. More particularly, multiple imputation allows any method to be applied once the imputation is done, whereas the EM algorithm requires a new variant of the algorithm for each statistical method. Besides, note that **mice** does not rely on parametric assumptions about the data distribution, whereas **misaem** assumes Gaussian covariates.

If we assume that we have a binary response variable `y` and incomplete covariate matrix `X_NA` composed of five covariates and whose full data counterpart follows a multivariate normal distribution and where the missing values are MAR, we can fit a logistic regression with missing values using the following lines of code.

```
df_NA <- data.frame(y, X_NA)
miss_list <- miss.glm(y~., data=df_NA)
```

Listing 6.12: Code to fit a logistic regression model with incomplete covariates using the EM algorithm in R.

This functions `miss.glm` resembles the standard `glm` function both in terms of its signature and output. Below we provide an example of output when applying the function `summary` to the output of the above call to `miss.glm`.

[10] https://rmisstastic.netlify.app/how-to/external/comparison_imputation_deep_classical

```
# Summary
print(summary(miss_list))
##
## Call:
## miss.glm(formula = y ~ ., data = df_NA)
##
## Coefficients:
##              Estimate  Std. Error
## (Intercept)   0.05128   0.31942
## X1            1.05798   0.35989
## X2           -0.99347   0.19620
## X3            1.07606   0.13937
## X4           -0.02258   0.06604
## X5           -1.01527   0.13353
## Log-likelihood: -132.14
```

Listing 6.13: Summary of a fitted logistic regression model with incomplete covariates in R.

The rationale behind the popular multiple imputation approach is to create $M > 1$ complete datasets by imputing the missing values with "plausible" values, and then to estimate a parameter of interest θ on each of the imputed datasets. The multiple estimation of θ and their variability allow to reflect uncertainty due to the unknown missing values. The parameter estimation is performed by applying the analytic method we would have used had the data been complete. We assume that this provides, for each imputed dataset, an estimate of the parameter θ and an estimate of the corresponding variance. These quantities are finally "pooled" by using specific rules named "Rubin's rules" (Rubin, 2004), leading to a final point estimate with a corresponding estimation of its variance that takes into account the uncertainty due to the missing values. In the following, we will compare this method to EM and illustrate the bias and variance of estimation by an example of simulated dataset.

Using the same example as for the EM algorithm, we can fit a logistic regression model using multiple imputation and inspect its summary as follows:

```
mi <- mice(data.frame(y, X_NA), m=20)  # imputation of 20 complete datasets
fit <- with(data = mi, exp = glm(y ~ X1+X2+X3+X4+X5, family = binomial)) #
    fit
beta.mi <- mice::pool(fit) # pool the results using Rubin's rules
summary(beta.mi)

##          term    estimate  std.error  statistic        df      p.value
## 1 (Intercept)  0.04006508 0.32034287  0.1250694 325.01965 9.005460e-01
## 2          X1  0.85919319 0.35413092  2.4262021 178.92213 1.625036e-02
## 3          X2 -0.85098985 0.19626265 -4.3359745 123.30329 2.983626e-05
## 4          X3  0.99568077 0.14825886  6.7158263  85.76393 1.934425e-09
## 5          X4 -0.04100766 0.06938153 -0.5910457 126.65685 5.555431e-01
## 6          X5 -0.92834313 0.14636424 -6.3426908  65.36987 2.423562e-08
```

Listing 6.14: Code to fit a logistic regression model with incomplete covariates using multiple imputation in R.

For this simulated data set, which follows the multivariate normal distribution, **misaem** gives less biased results than **mice**. This was expected as **misaem** fits here perfectly with the

parametric assumptions.

The workflow allows to directly apply and compare these two approaches, using either a simulated dataset, or a custom dataset that the user believes to satisfy the above stated assumptions about the missing data mechanism and distribution of the covariates.

6.3.4 How to predict in the presence of missing values?

A key task in supervised learning is prediction. Knowing how to predict in the presence of missing values is thus crucial for many practitioners. More precisely, we assume that the missing values occur in the covariates X. In this context, the goal is to predict an outcome variable y such that $y = f(X) + \epsilon$, where ϵ is a noise term. As a reminder, in supervised learning, the algorithms learn on a training set where the outcome variable is assumed to be known and the results of new (incomplete) observations in the test set are then predicted by applying this learning. Both R and Python workflows present different strategies to deal with the missing values in X (in the train set and in the test set). This task has been studied in detail by Josse et al. (2019). The recommended method is to impute the train set and the test set with the same constant, as the mean, and then apply a universally consistent learner, i.e. very powerful and able to learn any function f (linear or not, etc), such as the gradient boosting. This method has been shown asymptotically consistent. Besides, when random forests are used to impute the missing values, the authors recommend to use the Missing Incorporated in Attributes method (Twala et al., 2008), which allows imputation and prediction to be performed in a single step.

In R This R workflow has been written by an external contributor of the website, Katarzyna Woźnica (PhD student, Poland). It assesses a popular strategy (two-step strategy) which consists of imputing the train set and the test set independently with the same imputation method and of using usual learning algorithms to predict a target variable. Several imputation methods are compared, such as **mice, missForest** and **softImpute**. This work is also available on our website.[11] Note that until recently, using the popular **mice** package for learning predictive models for incomplete data in R was hindered by the fact that it did not allow to use the same imputation model for the train ad the test set. This has however been addressed and the detail of this recent extension can be found on GitHub.[12]

In Python The Python workflow proposes to compare two strategies when the aim is to predict a target variable and the covariates may contain missing values:

1. The *two-step* strategy consists in imputing the missing values both in the train and the test set like mean imputation and `IterativeImputer` of the **scikit-learn** library, and to apply usual learning algorithms (such as random forests, gradient boosting, linear regression) on the complete dataset. This learning algorithm can be applied on X but also by adding the response pattern R to the covariates: $[X, R]$.

[11] https://rmisstastic.netlify.app/how-to/external/how_to_predict_in_r
[12] https://github.com/amices/mice/issues/32

2. The *one-step* strategy aims at predicting with learning methods adapted to the missing data without necessarily imputing them, such as the Missing Incorparated in Attributes (*MIA*) method (Twala et al., 2008).

We propose a function, `score_pred`, which compares these strategies in terms of prediction performances by introducing missing values in the covariates (`x_comp`) under a specific missing data mechanism (`mecha`) and a given percentage of missing values (`p`). This introduction of missing values is repeated several times (`nbsim`) which leads to a stochasticity in the results. The dataset is then split into the train set and the test set (75% in the train set, 25% in the test set) and the methods presented below are applied by considering a specific learning algorithm (`learner`), e.g. the random forests, gradient boosting, linear regression. It returns the prediction error on the test set, by comparing the ground truth (`y`) and the predicted outcome values on the test set for each simulation (introduction of missing values) in a tabular (see Figure 6.8) The code of this function is given below, when the learning algorithm is the gradient boosting and 20% of MCAR values are introduced. The covariates $X \in \mathbb{R}^{1000\times 3}$ are generated under the multivariate Gaussian distribution, the parameter of the regression $\beta \in \mathbb{R}^3$ is a random uniform distribution. y is generated considering a linear regression such that $y = X\beta + \epsilon$, with ϵ a Gaussian noise.

```
learner = HistGradientBoostingRegressor() #learning algorithm to use
p = 0.2
res = score_pred(x_comp=X, y = y, learner=learner , p=p,
                 nbsim=10, mecha="MCAR")
```

Listing 6.15: Code to compare different strategies to predict an output variable in Python.

Figure 6.9 shows the graphical output of this function performed for different learning algorithms and for different missing-data mechanisms. When the learner is the linear regression, the two-steps methods with added mask, both for the MCAR mechanism and the MNAR mechanism, performs well. Note that the simulated dataset is generated considering a linear regression, which explains why the linear regression gives better results than other learners. In addition, for the MNAR mechanism, the one-step strategy *MIA* (especially when the gradient boosting is performed) seems to be a good choice. Note that MIA or mean imputation are recommended asymptotically but when having limited data in the prediction setting, other methods such as multiple imputation can outperform these asymptotically consistent methods (Josse et al., 2019).

Mean	Iterative	Mean + Mask	Iterative + Mask	MIA
0.892	0.895	0.892	0.895	0.892
0.885	0.896	0.885	0.896	0.891
0.895	0.89	0.895	0.89	0.903
0.866	0.836	0.866	0.836	0.873
0.881	0.863	0.881	0.863	0.887
0.859	0.862	0.859	0.862	0.865
0.9	0.91	0.9	0.91	0.905
0.86	0.852	0.86	0.852	0.847
0.897	0.899	0.897	0.899	0.911
0.879	0.872	0.879	0.872	0.883

Figure 6.8: Output of the function `score_pred` to compare different strategies when the aim is to predict in Python. 20% of missing values are introduced in a simulated dataset using the MCAR mechanism. The two-steps strategies (**IterativeImputer** and the mean imputation) with or without adding a mask and the one-step strategy *MIA* are compared in terms of prediction error, and then a gradient boosting is performed. The closer the result is to 1, the more accurate the prediction is (1 corresponds to the perfect prediction, 0 to the worst prediction). The results in the tabular corresponds to the prediction error for several simulations (introduction of missing values).

Another function `plot_score_realdatasets` is specifically designed to handle datasets which already contain missing values. The main arguments are the dataset (`X`), the outcome variable (`y`) and the learning algorithm to use (`learner`) In this case, the stochasticity comes from the way we split the dataset into a train set and a test set. This splitting and subsequent learning is repeated several times.

```
learner = HistGradientBoostingRegressor() #learning algorithm to
p = plot_score_realdatasets(X = X, y = y, learner = learner)
```

Listing 6.16: Code to compare different strategies for a real dataset to predict an output variable in Python.

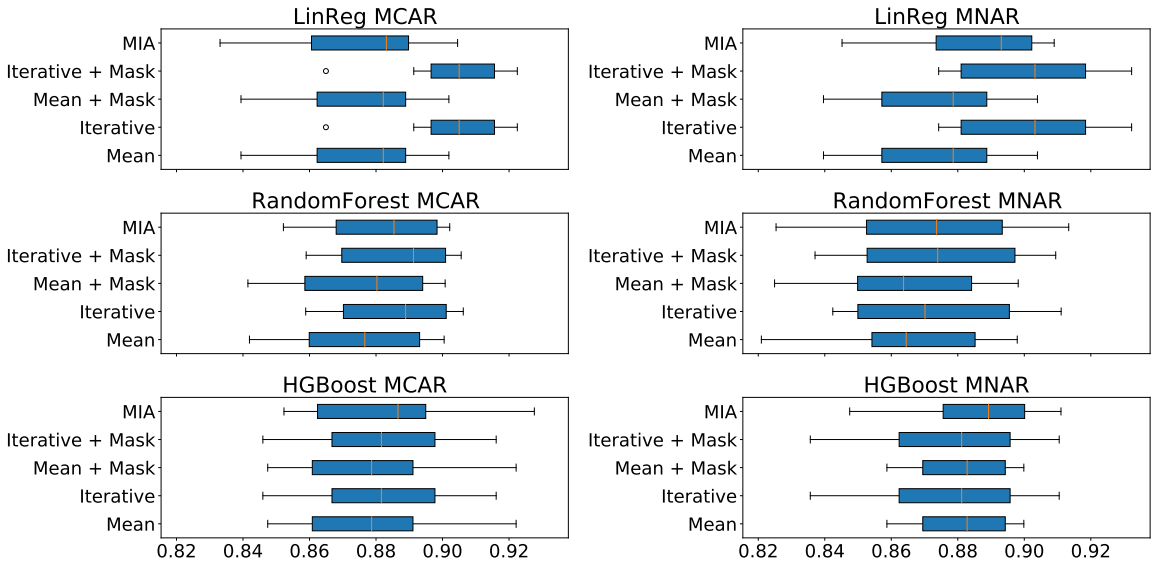


Figure 6.9: Plot of the function `score_pred` to compare different strategies when the aim is to predict in Python. 20% of missing values are introduced in a simulated dataset using the MCAR mechanism or the MNAR mechanism. The two-steps strategies (**IterativeImputer** and the mean imputation) with or without adding a mask and the one-step strategy *MIA* are compared in terms of prediction error, and several learner are performed (linear regression, random forests, gradient boosting). The closer the result is to 1, the more accurate the prediction is (1 corresponds to the perfect prediction, 0 to the worst prediction).

This concludes the overview over the workflows which have been developed in this project and which we consider as an invitation to other practitioners and researchers to use them for better comparability between methodologies when suggesting new methodologies in research articles and to extend them by providing feedback and ways of improvement which can benefit all potential users.

6.4 Perspectives and future extensions

By providing a platform and community to discuss missing data, software, approaches and workflows, we are providing a base from which we can grow.

6.4.1 Towards uniformization and reproducibility

One way to promote and encourage practitioners and researchers in their work with missing values is to provide community benchmarks and workflows centered around missing data. We will continue working on our workflows and the corresponding source code. In doing so we hope to encourage users to continue benchmarking new methods and to present the results in a clear, fair, and reproducible way.

In addition, we plan to propose two types of data challenges - 1) imputation and estimation, and 2) analysis workflows. For the first challenge, the objective is to find the best imputation or estimation strategy. The community would be given a dataset with missing values, for which there is actually a hidden copy of the real values. The community is then tasked with creating imputed values, which are assessed against the original dataset with complete values, to determine which imputation is best. This is similar in spirit to the Netflix prize (Bennett et al., 2007) and the M4 challenge in the time series domain (Makridakis et al., 2018). This benchmarking could be extended to other areas, such as parameter estimation, and predictive modeling with missing data.

Analysis workflow could form another community challenge, assessed in a similar way to existing "datathon" events where entries are assessed by an expert panel. Here the challenge could be to develop workflows and data visualizations from complex data. The data could have challenging features, and be combined from various data with complex structure, such as those data with several types of missingness, images, text, data, longitudinal data, and time series.

As has been demonstrated with data competition, involving the community brings forth many creative solutions and discussions that advance the field, and challenge existing strategies.

6.4.2 Pedagogical and practical guidance

In addition to the benchmarks and data challenges, we also plan to provide further guidance for both learning or teaching, and for applying the existing methods. For instance, an FAQ section is available on the platform, which answers prominent questions recurrent in lectures and talks, and for which the answers cannot always be found directly in the bibliography and lecture materials. Questions we receive on our platform, via the contact form or at other occasions, will also be posted there together with a concise answer.

6.4.3 Outreach

Despite the initial anchoring in the R community, we work in close collaboration with other communities, in particular with several developer teams of the **scikit-learn** library Pedregosa et al. (2011). Such collaborations have allowed us to integrate new workflows, to share respective experiences with missing values and to reach an even larger audience.

6.4.4 Participation and interaction

This platform is aimed to be for the community, in the sense that we welcome every comment and question, encourage submissions of new work, theoretical or practical, either through the provided contact form or directly via the GitHub project repository[13]. We have already received much useful feedback and contributions from the community, organized several remote calls and working sessions at statistics conferences. We are planning on regularly

[13] https://github.com/R-miss-tastic/website

relaunching calls for new material for the platform, for instance through the R consortium blog[14], R-bloggers[15] and social media platforms. We also intend to use these channels to communicate more generally about the platform and the topic of missing values.

In order for the platform to be a reference to the community, it needs to provide regularly updated user friendly content. Crucial to this is proposing sustainable and accessible solutions for the maintenance of the R-miss-tastic platform. We hope that the well documented code source of the platform invites contributions and community feedback on this project.

The aim of this platform is to go further than only community participation, namely to seed meaningful community interactions, and make it a hub of communication among groups that rarely exchange, both within, and between academia and industry communities.

6.4.5 Future extensions

Potential extensions that could be added in future releases of the platform and for which we welcome suggestions and contributions are the following: a workflow with a focus on MNAR data and different solutions that can handle such data (as diversity of existing solutions is large, such a unified workflow will be a consequential contribution); for more applied users, a comparison of computation times of different methods, benchmarked on various types of data; a more and more often encountered problem of missing values in data integration: questions such as *what do I do when I have clinical data from multiple centers with different mechanisms of missing values or with systematically missing values in certain data?* or *what do I do when I have time series and missing values in one of the groups of variables?*

More generally, these are examples to explain how we intend to update the platform with relevant results and recommendations from current research around missing values.

Acknowledgements This work has partially been funded by the R Consortium, Inc. We would like to thank Steffen Moritz for his active support and feedback, all contributors who have generously made their course and tutorial materials available, as well as the contributors to the workflows in R and Python code.

[14] https://www.r-consortium.org/news/blog
[15] https://www.r-bloggers.com/

Conclusion

Summary

The key objective of this dissertation was to propose theoretically sound methods with associated efficient implementations for dealing with missing data in realistic scenarios from different statistical frameworks. In particular, I studied low-rank models, PPCA, averaged SGDs, robust lasso and clustering methods with missing values. I considered missing data mechanisms which go beyond the classical MCAR or MNAR on one variable, by considering heterogeneous MCAR, MNAR on several variables and MNAR coupled with M(C)AR variables. In addition, I also considered heterogeneous data (categorical, continuous, mixed). A constant motivation was to make the methods applicable to real-world problems, such as those raised by the Traumabase dataset.

In the first part, several methods have been proposed for dealing with several MNAR or MAR variables in low-rank models by leveraging their ability to summarise a dataset by a few important variables (and individuals profiles). In Chapter 2, as a starting point, two approaches have been suggested to take into account self-masked MNAR data in a low-rank model (with fixed effects), in order to recover the underlying low-rank structure and impute the data accordingly. These approaches consist in modeling, either explicitly or implicitly, the joint distribution of the data and the missing-data mechanism. On the one hand, an accelerated EM algorithm has been derived, coupled with the Sampling Importance Resampling algorithm. It provides a suitable framework for theoretical analysis, at the price of an expensive computational cost. On the other hand, a heuristics that does not model the missing-data mechanism has been suggested. It consists of concatenating the data matrix and the missing-data pattern and of assuming a low-rank structure on this augmented matrix, in order to take into account the relationships between the variables and the mechanism. Classical strategies assuming M(C)AR data can be then performed. Although this heuristics is computationally efficient, it has no theoretical grounding. Our experimental study has shown that even though our accelerated EM algorithm provides better results in terms of imputation and estimation errors, our heuristics is a relevant alternative computationally efficient, especially useful when many variables are missing. In Chapter 3, this work has been

extended to the low rank model with random effects (PPCA) in order to handle (general) MNAR data, addressing both the theoretical challenges and the computational burden. First, the identifiability of the parameters is studied, which is a key issue for the MNAR mechanism. Then, consistent estimators of the parameters and an imputation method have been proposed, which do not require the knowledge of the missing data mechanism and use only all available observed cells. Although this method leads to an efficient algorithm relying on strong theoretical guarantees, there are still points of improvement to be worked on, as it requires the knowledge of both the rank of the loading coefficients and the noise variance.

In the second part, the crucial issue of handling missing data in learning tasks has been addressed. In Chapter 4, one of the most popular learning algorithms, the averaged SGD, has been adapted to handle missing values to perform linear regression, when the covariates may contain heterogeneous MCAR values. The powerful properties of the averaged SGD without missing data have been exploited to propose an easy-to-implement algorithm, suitable to the high-dimensional and online setting: it consists of naively imputing the missing values by zeros, and of using debiased gradients to account for the imputation error. This algorithm remains computationally cheap per iteration and relies on weak assumptions on the data distribution. I then established the convergence of this algorithm in terms of excess risk at the rate $1/k$ at iteration k. This rate is remarkable as it is optimal and similar to the rate of the averaged SGD without any missing value. By considering a simple linear case with heterogeneous MCAR data, this work also aimed to take a first step towards solving two open questions, particularly relevant in real data analysis: the treatment of missing values in large-scale datasets, and in an online setting (when the data come as they go along). In Appendix A, in a high-dimensional setting, a problematic reformulation for dealing with MNAR data in the case of sparse linear regression has also been proposed. In Chapter 5, the focus is made on the model-based clustering framework. New algorithms are derived to cluster individuals and to estimate the parameters of the mixture model in presence of several MNAR variables of different types (continuous, categorical). This work also includes an exhaustive catalog of possible MNAR specifications in the model-based clustering, accompanied with a detailed study of each model both in terms of identifiability and resulting estimation strategy.

During this PhD, wishing to push the reproducibility cursor further, it was important to me to promote the study of missing data and to facilitate its management and understanding. To this end, I took part in the development of an open source platform, Rmisstastic, which is an ongoing collaborative project. In particular, I have created several workflows, both in R and Python, which address the main issues raised by missing values (e.g. imputation or prediction). These are targeted to students or researchers who are familiar with missing data but also at data scientists and practitioners for whom the choice of a particular method for handling missing data is often a crucial and delicate issue.

Perspectives

While this work answers some questions regarding the processing of missing data, it also opens up exciting new perspectives.

(i) A first extension of the work presented on low-rank models in Chapters 2 and 3 to the exponential family would allow to deal with (missing) count data, which could be a great improvement. Indeed, this could be especially useful to make these methods even more suitable for real datasets. Furthermore, estimations of the rank and of the noise variance in low-rank models with MNAR data remains non-trivial. Note that a preliminary noise variance estimation allows a rank estimation, so that a cross-validation strategy to estimate the noise variance by adding MNAR values is a line of work deserving further research.

(ii) Focusing on SGDs for linear models with heterogeneous MCAR data, the work presented in Chapter 4 definitely paves the way for promising future research, as stochastic algorithms are at the heart of modern machine learning techniques. This work differs from the rest of this dissertation, as it does not deal with MNAR data, and a first ambitious extension will be to adapt the averaged SGD to more complex missing-data types. In addition, deep learning models, for which the training relies on the SGD algorithm, are undoubtedly flexible for dealing with more complex data types in different learning tasks (Goodfellow et al., 2016). Another extension of our work could be therefore to deal with more general loss functions, such as the logistic one, widely used to train neural networks in a classification setting.

These two extensions are challenging, because the bias introduced is not the same as the one considered in the case of linear regression with heterogeneous MCAR data. Preliminary numerical experiments show that it is difficult to adapt the SGD algorithm to more complex cases, and even more to theoretically study the new algorithm. Achieving these extensions would make two great contributions.

(iii) As the clustering with MNAR data is an undergoing work (Chapter 5), there is still numerical work to do. For instance, the application of our method to the Traumabase dataset will be extremely interesting, as it can be genuinely useful to form groups of similarly-behaving patients and by doing so to improve their care.

(iv) Building on the strengths of different scientific communities, another exciting and useful perspective for MNAR data could be to provide a unified reviewing work on identifiability methods, by bringing together the literature considering graphical models and the one on semi-parametric models.

(v) Going further than the themes studied in this manuscript, my bibliographic work on MNAR data has led me to believe that there are also interesting bridges to exploit between (a) the literature based on purely statistical methods, such as semi-parametric models in the framework of linear regression where only the output variable may contain MNAR values, and (b) the literature of semi-supervised learning when the unlabeled

data correspond to classes not present in the labeled data (Oliver et al., 2018). Note that in the statistical literature, the inferential framework is often considered, when the aim is to estimate the model parameters (such as regression coefficients). Meanwhile, in the semi-supervised learning, the goal is the prediction of the observations labels (for example with image classification tasks). Even though there are some differences, we can leverage both literatures: (a) by benefiting from the theoretical background of the statistical literature on the one hand, and (b) exploiting the flexibility of recent semi-supervised learning algorithms to efficiently handle complex data types on the other hand.

List of publications

- Imputation and low-rank estimation with Missing Not At Random data, A. Sportisse, C. Boyer, J. Josse, *Statistics & Computing, Springer*, 2020
- Estimation and Imputation in Probabilistic Principal Component Analysis with Missing Not At Random Data, A. Sportisse, C. Boyer, J. Josse, *Advances in Neural Information Processing Systems*, 2020
- Debiasing Stochastic Gradient Descent to handle missing values, A. Sportisse, C. Boyer, A. Dieuleveut, J. Josse, *Advances in Neural Information Processing Systems*, 2020

Submitted papers

- Robust Lasso-Zero for sparse corruption and model selection with missing covariates, led by Pascaline Descloux, and in collaboration with Claire Boyer, Julie Josse and Sylvain Sardy (submitted in 2020, in review)
- R-miss-tastic: a unified platform for missing values methods and workflows, led by Imke Mayer, and in collaboration with Julie Josse, Nicholas Tierney, Nathalie Vialaneix

Ongoing work

- Model-based Clustering with Missing Not At Random Data, initiated by Christophe Biernacki, Gilles Celeux, Julie Josse, Fabien Laporte, and reworked with Christophe Biernacki, Claire Boyer, Julie Josse and Matthieu Marbac

Appendix A

Robust Lasso-Zero

This chapter is an ongoing work, led by Pascaline Descloux, in which I collaborated with Claire Boyer, Julie Josse and Sylvain Sardy.

Abstract

We propose Robust Lasso-Zero, an extension of the Lasso-Zero methodology (Descloux and Sardy, 2020), initially introduced for sparse linear models, to the sparse corruptions problem. We give theoretical guarantees on the sign recovery of the parameters for a slightly simplified version of the estimator, called Thresholded Justice Pursuit. The use of Robust Lasso-Zero is showcased for variable selection with missing values in the covariates. In addition to not requiring the specification of a model for the covariates, nor estimating their covariance matrix or the noise variance, the method has the great advantage of handling missing not-at random values without specifying a parametric model. Numerical experiments and a medical application underline the relevance of Robust Lasso-Zero in such a context with few available competitors. The method is easy to use and implemented in the R library `lass0`.

A.1 Introduction

Let us consider the widely used framework of sparse linear models for high dimension,

$$y = X\beta^0 + \epsilon, \tag{A.1}$$

where $\epsilon \in \mathbb{R}^n$ is a (dense) Gaussian noise vector with variance σ^2, X has a number of columns p larger than the number of rows n, and the parameters of interest $\beta^0 \in \mathbb{R}^p$ is s-sparse (only s out of its p entries are different from zero). To take into account additional occasional

corruptions, the sparse corruption problem is

$$y = X\beta^0 + \sqrt{n}\omega^0 + \epsilon, \tag{A.2}$$

where $\omega^0 \in \mathbb{R}^n$ is a k-sparse corruption vector; see for instance Chen et al. (2013). Noting that (A.2) can be rewritten as

$$y = \begin{bmatrix} X & \sqrt{n}I_n \end{bmatrix} \begin{bmatrix} \beta^0 \\ \omega^0 \end{bmatrix} + \epsilon,$$

the sparse corruption model can be seen as a sparse linear model with an augmented design matrix and an augmented sparse vector. We are interested in theoretical guarantees of support recovery for β^0 in (A.2), with interesting consequences for variable selection with missing covariates.

Related litterature. To recover β^0 when $\epsilon = 0$, several authors proposed *Justice Pursuit* (JP), name coined by Laska et al. (2009), by solving

$$\min_{\beta \in \mathbb{R}^p, \omega \in \mathbb{R}^n} \quad \|\beta\|_1 + \|\omega\|_1 \tag{A.3}$$

$$\text{s.t.} \quad y = X\beta + \omega, \tag{A.4}$$

which is nothing else than the *Basis Pursuit* (BP) problem, with the augmented matrix $\begin{bmatrix} X & I_n \end{bmatrix}$ (modulo the renormalization by $\sqrt{n}$ in (A.3)) (Wright et al., 2009). Wright and Ma (2010) analyzed JP for Gaussian measurements, providing support recovery results when $n \simeq p$ using cross-polytope arguments. Besides, Laska et al. (2009) and Li et al. (2010) proved that if the entries of X are i.i.d. standard Gaussian as well, then the matrix $\begin{bmatrix} X & I_n \end{bmatrix}$ satisfies some restricted isometry property with high probability, implying exact recovery of both β^0 and ω^0, provided that $n \gtrsim (s + k)\log(p)$. However, in these works, the sparsity level k of ω^0 cannot be fixed to a proportion of the sample size n. Therefore, Li (2013) and Nguyen and Tran (2013b) introduced a tuning parameter $\lambda > 0$ and solve

$$\min_{\beta \in \mathbb{R}^p, \omega \in \mathbb{R}^n} \|\beta\|_1 + \lambda\|\omega\|_1 \quad \text{s.t.} \quad y = X\beta + \omega. \tag{A.5}$$

In a sub-orthogonal or Gaussian design, they both proved exact recovery, even for a large proportion of corruption.

In the case of sparse ($\omega^0 \neq 0$) and dense noise ($\epsilon \neq 0$), Nguyen and Tran (2013a) proposed to jointly estimate β^0 and ω^0 by solving

$$\min_{\beta \in \mathbb{R}^p, \omega \in \mathbb{R}^n} \frac{1}{2}\|y - X\beta - \omega\|_2^2 + \lambda_\beta\|\beta\|_1 + \lambda_\omega\|\omega\|_1. \tag{A.6}$$

In the special case where $\lambda_\beta = \lambda_\omega$, problem (A.6) boils down to the Lasso (Tibshirani, 1996) applied to the response y and the design matrix $\begin{bmatrix} X & I_n \end{bmatrix}$. Assuming a standard Gaussian

design and the invertibility and incoherence properties for the covariance matrix, they obtained sign recovery guarantee for an arbitrarily large fraction of corruption, provided that $n \geqslant Ck\log(p)\log(n)$. In addition, the required number of samples is proven to be optimal. More recently in the case of a Gaussian design with an invertible covariance matrix, Dalalyan and Thompson (2019) obtained an optimal rate of estimation of β^0 when considering an ℓ_1-penalized Huber's M-estimator, which is actually equivalent to (A.6) (Sardy et al., 2001).

Contributions. To estimate the support of the parameter vector β^0 in the sparse corruption problem, we study an extension of the Lasso-Zero methodology (Descloux and Sardy, 2020), initially introduced for standard sparse linear models, to the sparse corruptions problem. We provide theoretical guarantees on the sign recovery of β^0 for a slightly simplified version of Robust Lasso-Zero, that we call Thresholded Justice Pursuit (TJP). These garantees are extensions of recent results on Thresholded Basis Pursuit. The first one extends a result of Tardivel and Bogdan (2019), providing a necessary and sufficient condition for consistent recovery in a setting where the design matrix is fixed but the nonzero absolute coefficients tend to infinity. The second one extends a result of Descloux and Sardy (2020), proving sign consistency for correlated Gaussian designs when p, s and k grow with n, allowing a positive fraction of corruptions.

Showing that missing values in the covariates can be reformulated into a sparse corruption problem, we recommand Robust Lasso-Zero for dealing with missing data. For support recovery, this approach requires neither to specify a model for the covariates or the missing data mechanism, nor an estimation of the covariates covariance matrix or of the noise variance, and hence provides a simple method for the user. Numerical experiments and a medical application also underline the effectiveness of Robust Lasso-Zero with respect to few competitors.

Organization. After defining Robust Lasso-Zero in Section A.2, we analyse the sign recovery properties of Thresholded Justice Pursuit in Section A.2.3. Section A.3.1 is dedicated to variable selection with missing values and the selection of tuning parameters is discussed in Section A.3.2. Numerical experiments are presented in Section A.4 and an application in Section A.5.

Notation. Define $[p] := \{1, \dots, p\}$, and the complement of a subset $S \subset [p]$ is denoted $\bar{S}$. For a matrix A of size $u \times v$ and a set $T \subset [v]$, we use A_T to denote the submatrix of size $n \times |T|$ with columns indexed by T. We define the missing value indicator matrix $M \in \mathbb{R}^{n\times p}$ by $M_{ij} = \mathbb{1}_{\{X_{ij}^{\text{NA}} = \texttt{NA}\}}$, and the set of incomplete rows by $\mathcal{M} := \{i \in [n] \mid M_{ij} = 1 \text{ for some } j \in [p]\}$.

A.2 Robust Lasso-Zero

A.2.1 Lasso-Zero in a nutshell

Under linear model (A.1), Thresholded Basis Pursuit (TBP) estimates β^0 by setting the small coefficients of the BP solution to zero. Since BP fits the observations y exactly, noise is generally overfitted. Lasso-Zero (Descloux and Sardy, 2020) alleviates this issue by solving repeated BP problems, respectively fed with the augmented matrices $[X|G^{(k)}]$, where $G^{(k)} \in \mathbb{R}^{n\times n}$, $k = 1, \dots, M$, are different i.i.d. Gaussian noise dictionaries. Hence, some columns of $G^{(k)}$ can be used to fit the noise term. The obtained estimates $\hat{\beta}^{(1)}, \dots, \hat{\beta}^{(M)}$ are then aggregated by taking the component-wise medians, further thresholded at level $\tau > 0$. Descloux and Sardy (2020) show that Lasso-Zero tuned by Quantile Universal Thresholding (Giacobino et al., 2017) achieves a very good trade-off between high power and low false discovery rate compared to competitors.

A.2.2 Definition of Robust Lasso-Zero

Robust Lasso-Zero arises by applying Lasso-Zero to Justice Pursuit, instead of Basis Pursuit. Consider the sparse corruption model (A.2), for which S^0 and T^0 denote the respective supports of β^0 and ω^0, and $s := |S^0|$ and $k := |T^0|$ denote their respective sparsity degrees. To fix notation, we then consider the following parametrization of Justice Pursuit (JP):

$$(\hat{\beta}^{\text{JP}}_\lambda, \hat{\omega}^{\text{JP}}_\lambda) \in \underset{\beta\in\mathbb{R}^p,\ \omega\in\mathbb{R}^n}{\arg\min} \|\beta\|_1 + \lambda\|\omega\|_1 \quad \text{s.t.} \quad y = X\beta + \sqrt{n}\omega. \tag{A.7}$$

Renormalization by $\sqrt{n}$ balances the augmented design matrix $\begin{bmatrix} X & \sqrt{n}I_n \end{bmatrix}$: in practice the columns of X are often standardized so that $\|X_j\|_2^2 = n$ for every $j \in [p]$, and this way, all columns of $\begin{bmatrix} X & \sqrt{n}I_n \end{bmatrix}$ have same norm.

Robust Lasso-Zero applied to (A.7) is fully described in Algorithm 5. Attention has been drawn to the estimation of the support of β^0. However the estimation of the corruption support is also possible by computing the corresponding vectors $\hat{\omega}^{\text{med}}_\lambda$ and $\hat{\omega}^{\text{Rlass0}}_{(\lambda,\tau)}$, at stages 2)) and 3)).

Since the minimization problem (A.8) in Algorithm 5 can be recast as a linear program, any relevant solver can be used (e.g., proximal methods). Algorithm 5 includes two hyper-parameters: the regularization parameter λ of (A.7), and the thresholding parameter τ of the Robust Lasso-Zero methodology. Their choice in practice is discussed in Section A.3.2.

A.2.3 Theoretical guarantees on Thresholded Justice Pursuit

Discarding the noise dictionaries in Algorithm 5 amounts to thresholding the solution $(\hat{\beta}^{\text{JP}}_\lambda, \hat{\omega}^{\text{JP}}_\lambda)$ to the Justice Pursuit problem (A.7). Robust Lasso-Zero can therefore be regarded as an extension of this simpler estimator, which we call *Thresholded Justice Pursuit* (TJP):

$$\hat{\beta}^{\text{TJP}}_{(\lambda,\tau)} = \eta_\tau(\hat{\beta}^{\text{JP}}_\lambda) \qquad \text{and} \qquad \hat{\omega}^{\text{TJP}}_{(\lambda,\tau)} = \eta_\tau(\hat{\omega}^{\text{JP}}_\lambda). \tag{A.9}$$

We present two results about sign consistency of TJP.

Algorithm 5 Robust Lasso-Zero

Given data (y, X), for fixed hyper-parameters $\lambda > 0, \tau \geqslant 0$ and $M \in \mathbb{N}^*$:

1) For $k = 1, \dots, M$:

   i) generate a matrix $G^{(k)}$ of size $n \times n$ with i.i.d. $\mathcal{N}(0,1)$ entries

   ii) compute the solution $(\hat{\beta}^{(k)}_\lambda, \hat{\omega}^{(k)}_\lambda, \hat{\gamma}^{(k)}_\lambda)$ to the augmented JP problem

$$(\hat{\beta}^{(k)}_\lambda, \hat{\omega}^{(k)}_\lambda, \hat{\gamma}^{(k)}_\lambda) \in \begin{array}{ll} \underset{\beta\in\mathbb{R}^p,\ \omega\in\mathbb{R}^n,\ \gamma\in\mathbb{R}^n}{\arg\min} & \|\beta\|_1 + \lambda\|\omega\|_1 + \|\gamma\|_1 \\ \text{s.t.} & y = X\beta + \sqrt{n}\omega + G^{(k)}\gamma. \end{array} \tag{A.8}$$

2) Define the vector $\hat{\beta}^{\text{med}}_\lambda$ by

$$\hat{\beta}^{\text{med}}_{\lambda,j} := \text{median}\{\hat{\beta}^{(k)}_{\lambda,j}, k = 1, \dots, M\} \quad \text{for every } j \in [p].$$

3) Calculate the estimate $\hat{\beta}^{\text{Rlass0}}_{(\lambda,\tau)} := \eta_\tau(\hat{\beta}^{\text{med}}_\lambda)$, where $\eta_\tau(x) = x\mathbb{1}_{(\tau,+\infty)}(|x|)$ hard-thresholds component-wise.

A.2.3.1 Identifiability as a necessary and sufficient condition for consistent sign recovery

First introduced in Tardivel and Bogdan (2019) for the TBP, we propose the following extension of the identifiability notion for the TJP.

Definition 24. *The pair $(\beta^0, \omega^0) \in \mathbb{R}^p \times \mathbb{R}^n$ is said to be identifiable with respect to $X \in \mathbb{R}^{n\times p}$ and the parameter $\lambda > 0$ if it is the unique solution to JP (A.7) when $y = X\beta^0 + \sqrt{n}\omega^0$.*

It is worth noting that identifiability of (β^0, ω^0) can be shown to depend only on $\text{sign}(\beta^0)$ and $\text{sign}(\omega^0)$, as highlighted in the following result.

Lemma 5. *The pair $(\beta^0, \omega^0) \in \mathbb{R}^p \times \mathbb{R}^n$ is identifiable with respect to $X \in \mathbb{R}^{n\times p}$ and the parameter $\lambda > 0$ if and only if for every pair $(\beta, \omega) \neq (0, 0)$ such that $X\beta + \sqrt{n}\lambda^{-1}\omega = 0$,*

$$|\,\text{sign}(\beta^0)^T\beta + \text{sign}(\omega^0)^T\omega| < \|\beta_{\overline{S^0}}\|_1 + \|\omega_{\overline{T^0}}\|_1.$$

Proof. See Appendix A.6. □

In order to show that identifiability is necessary and sufficient for TJP to consistently recover $\text{sign}(\beta^0)$ and $\text{sign}(\omega^0)$, assume that for a fixed matrix $X \in \mathbb{R}^{n\times p}$ and a sequence $\{(\beta^{(r)}, \omega^{(r)})\}_{r\in\mathbb{N}*}$, the following holds:

(i) there exist sign vectors $\theta \in \{1, -1, 0\}^p$ and $\tilde{\theta} \in \{1, -1, 0\}^n$ such that $\text{sign}(\beta^{(r)}) = \theta$ and $\text{sign}(\omega^{(r)}) = \tilde{\theta}$ for every $r \in \mathbb{N}^*$,

(ii) $\lim_{r\to+\infty} \min\{\beta^{(r)}_{\min}, \omega^{(r)}_{\min}\} = +\infty$, where $\beta_{\min} := \min_{j\in\mathrm{supp}(\beta)} |\beta_j|$,

(iii) there exists $q > 0$ such that $\frac{\min\{\beta^{(r)}_{\min}, \omega^{(r)}_{\min}\}}{\max\{\|\beta^{(r)}\|_\infty, \|\omega^{(r)}\|_\infty\}} \geqslant q$.

These assumptions are similar to the ones of Tardivel and Bogdan (2019). We use the notation $S^0 := \mathrm{supp}(\theta) = \mathrm{supp}(\beta^{(r)})$ and $T^0 := \mathrm{supp}(\tilde{\theta}) = \mathrm{supp}(\omega^{(r)})$. We denote by $(\hat{\beta}^{\mathrm{JP}(r)}_\lambda, \hat{\omega}^{\mathrm{JP}(r)}_\lambda)$ the JP solution when $y = y^{(r)} := X\beta^{(r)} + \sqrt{n}\omega^{(r)} + \epsilon$, and $(\hat{\beta}^{\mathrm{TJP}(r)}_{(\lambda,\tau)}, \hat{\omega}^{\mathrm{TJP}(r)}_{(\lambda,\tau)})$ the corresponding TJP estimates.

Theorem 25. *Let $\lambda > 0$ and let X be a matrix of size $n \times p$ such that for any $y \in \mathbb{R}^n$, the solution to JP (A.7) is unique. Let $\{(\beta^{(r)}, \omega^{(r)})\}_{r\in\mathbb{N}^*}$ be a sequence satisfying assumptions ((i))-((iii)) above. If the pair of sign vectors $(\theta, \tilde{\theta})$ is identifiable with respect to X and λ, then for every $\epsilon \in \mathbb{R}^n$, there exists $R = R(\epsilon) > 0$ such that for every $r \geqslant R$ there is a threshold $\tau = \tau(r) > 0$ for which*

$$\mathrm{sign}(\hat{\beta}^{\mathrm{TJP}(r)}_{(\lambda,\tau)}) = \theta \quad \textit{and} \quad \mathrm{sign}(\hat{\omega}^{\mathrm{TJP}(r)}_{(\lambda,\tau)}) = \tilde{\theta}. \tag{A.10}$$

Conversely, if for some $\epsilon \in \mathbb{R}^n$ and $r \in \mathbb{N}^*$ there is a threshold $\tau > 0$ such that (A.10) holds, then $(\theta, \tilde{\theta})$ is identifiable with respect to X and λ.

Proof. See Appendix A.6. □

Remark 26. *One might be interested in recovering the signs of the sparse corruption. If $\omega^{(r)}$ is considered as noise, then only the recovery of $\mathrm{sign}(\beta^{(r)})$ matters. In this case one could weaken assumptions ((ii)) and ((iii)) above by replacing $\min\{\beta^{(r)}_{\min}, \omega^{(r)}_{\min}\}$ by $\beta^{(r)}_{\min}$, and identifiability of $(\theta, \tilde{\theta})$ would be sufficient for recovering $\mathrm{sign}(\beta^0)$. However, recovery of both $\mathrm{sign}(\beta^{(r)})$ and $\mathrm{sign}(\omega^{(r)})$ is needed for proving necessity of identifiability.*

Identifiability of sign vectors is necessary and sufficient for sign recovery when the nonzero coefficients are large. However, Theorem 25 does not provide a lower bound indicating how large these coefficients should scale to be correctly detected. In the next section, we make this explicit in particular for (correlated) Gaussian designs and prove that sign consistency holds, allowing p, s and k to grow with the sample size n.

A.2.3.2 Sign consistency of TJP for correlated Gaussian designs

We make the following assumptions:

(iv) the rows of $X \in \mathbb{R}^{n\times p}$ (with $n < p$) are random and i.i.d. $\mathcal{N}(0, \Sigma)$;

(v) The smallest eigenvalue of the covariance matrix Σ is assumed to be positive: $\lambda_{\min}(\Sigma) > 0$,

(vi) the variance of the covariates is equal to one: $\Sigma_{ii} = 1$ for every $i \in [p]$;

(vii) the noise is assumed to be Gaussian: $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

Assumptions ((iv)) and ((v)) imply that almost surely $\operatorname{rank} X = n$.

Theorem 27. *Under Assumptions ((iv))-((vii)), choosing* $\lambda = \frac{1}{\sqrt{\log p}}$ *ensures with probability greater than* $1 - ce^{-c'n} - 1.14^{-n} - 2e^{-\frac{1}{8}(\sqrt{p}-\sqrt{n})^2}$*, that there exists a value of* $\tau > 0$ *such that*

$$\operatorname{sign}(\hat{\beta}^{\mathrm{TJP}}_{(\lambda,\tau)}) = \operatorname{sign}(\beta^0),$$

provided that

$$n \geqslant C \frac{\kappa(\Sigma)}{\lambda_{\min}(\Sigma)} s \log p, \tag{A.11}$$

$$\frac{n}{k} \geqslant \max\left\{ \frac{1}{C'}, \frac{\kappa(\Sigma)}{C''} \right\}, \tag{A.12}$$

$$\beta^0_{\min} > \frac{10\sqrt{2}\max\{1,\lambda\}\sigma\sqrt{p+n}}{\left(\frac{\lambda_{\min}(\Sigma)}{4}(\sqrt{p/n}-1)^2+1\right)^{1/2}}, \tag{A.13}$$

where $\kappa(\Sigma) := \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$ *is the conditioning number of* Σ*, and* C, C', C'' *are some numerical constants with* $C \geqslant 144^2$.

Proof. See Appendix A.7. □

Theorem 27 ensures that, for correlated Gaussian designs and signal-to-noise ratios high enough, TJP successfully recovers $\operatorname{sign}(\beta^0)$ with high probability, even with a positive fraction of corruptions. As a consequence, if Σ is well-conditioned, (i.e. the eigenvalues of Σ are bounded: $0 < \gamma_1 \leqslant \lambda_{\min}(\Sigma) \leqslant \lambda_{\max}(\Sigma) \leqslant \gamma_2$) and $p/n \to \delta > 1$, TJP achieves sign consistency provided that $n = \Omega(s \log p)$, $k = \mathcal{O}(n)$ and $\beta^0_{\min} = \Omega(\sqrt{n})$. The lower-bound required on $\beta^0_{\min}$ in Theorem 27 is of the same order as the one required for TBP in Descloux and Sardy (2020). One can remark that the analysis of TJP in the sparse corruption setting makes the condition number of Σ come into play in the lower-bounds required on n and k. This quantity seems natural to arise in the sparse corruption problem helping discriminating design instability from corruptions.

A.3 Model selection with missing covariates

In practice the matrix of covariates X is often partially known due to manual errors, poor calibration, insufficient resolution, etc., and one only observes an incomplete matrix, denoted X^{NA}.

Theoretical guarantees of estimation strategies or imputation methods rely on assumptions regarding the missing-data mechanism, i.e. the cause of the lack of data. Three missing-data mechanisms have been introduced by Rubin (1976): the restrictive assumptions of data (a) missing completely at random (MCAR), and (b) missing at random (MAR), where the missing data may only depend on the observed variables, and (c) the more general assumption of data missing not at random (MNAR), when data missingness depends on the

values of other variables, but also on its own value. Complete case analysis, which discards all incomplete rows, is the most common method for facing missing values in applications. Additionally to the induced estimation bias (especially under the MNAR missing mechanism (c)), with high-dimensional data this procedure has the big disadvantage that missingness of a single entry causes the loss of an entire row, which contains a lot of information when p is large.

High dimensional variable selection with missing values turns out to be a challenging problem and very few solutions are available, not to mention implementations. Available solutions either require strong assumptions on the missing value mechanism, a lot of parameters tuning or strong assumption on the covariates distribution which is hard in high dimensions. They include the Expectation-Maximization algorithm (Dempster et al., 1977) for sparse linear regression (Garcia et al., 2010) and regression imputation methods (Van Buuren, 2018). A method combining penalized regression techniques with multiple imputation and stability selection has been developed (Liu et al., 2016). Yet, aggregating different models for the resulting multiple imputed data sets becomes increasingly complex as the number of data grows. Rosenbaum et al. (2013) modified the Dantzig selector by using a consistent estimation of the design covariance matrix. Following the same idea, Loh and Wainwright (2012) and Datta et al. (2017) reformulated the Lasso also using an estimate of the design covariance matrix, possibly resulting in a non-convex problem. Chen and Caramanis (2013) presented a variant of orthogonal matching pursuit which recovers the support and achieves the minimax optimal rate. Jiang et al. (2019) proposed Adaptive Bayesian SLOPE, combining SLOPE and Spike-and-Slab Lasso. While some of these methods have interesting theoretical guarantees, they all require an estimation of the design covariance matrix, which is often obtained under the restrictive MCAR assumption.

A.3.1 Relation to the sparse corruption model

To tackle the problem of estimating the support when the design matrix is incomplete, we suggest an easy-to-implement solution for the user, which consists in imputing the missing entries in X^{NA} with the imputation of his choice to get a completed matrix $\tilde{X}$, and to take into account the impact of the possibly occasional poor imputation as follows. Given the matrix $\tilde{X}$, the linear model (A.1) can be rewritten in the form of the sparse corruption model (A.2), where $\omega^0 := \frac{1}{\sqrt{n}}(X - \tilde{X})\beta^0$ is the (unknown) corruption due to imputations. In classical (i.e. non-sparse) regression, one could not say much about ω^0 without any prior knowledge of the distribution of the covariates or the missing data mechanism. Since the key point here is that when β^0 is sparse, then so is ω^0, even if all rows of the design matrix contain missing entries. Indeed, for every $i \in [n]$,

$$\omega_i^0 = \frac{1}{\sqrt{n}} \sum_{j=1}^{p} (X_{ij} - \tilde{X}_{ij})\beta_j^0 = \frac{1}{\sqrt{n}} \sum_{j \in S^0} (X_{ij} - \tilde{X}_{ij})\beta_j^0, \tag{A.14}$$

so ω_i^0 is nonzero only if the i^{th} row of X^{NA} contains missing value(s) on the support S^0, since $(X_{ij} - \tilde{X}_{ij}) = 0$ if X_{ij} is observed. So the problem of missing covariates can be rephrased

as a sparse corruption problem, as already pointed out in Chen et al. (2013). We propose to use Robust Lasso-Zero presented in Section A.2.2, which comes with strong theoretical guarantees, to tackle this sparse corruption reformulation, see Algorithm6.

Note that if the i^{th} row of X is fully observed, then $\omega_i^0 = 0$ by (A.14). Thus the dimension of ω^0 can be reduced by restricting it to the incomplete rows of X^{NA}. The corruption vector ω^0 is now of size $|\mathcal{M}|$ and (A.2) becomes

$$y = X\beta^0 + \sqrt{n} I_{\mathcal{M}} \omega^0 + \epsilon. \tag{A.15}$$

Algorithm 6 Robust Lasso-Zero for missing data

Given data (y, X^{NA}), for fixed hyper-parameters $\lambda > 0, \tau \geqslant 0$ and $M \in \mathbb{N}^*$:

1) Impute X^{NA} and rescale the imputed matrix X such that all columns have Euclidean norm equal to $\sqrt{n}$.

2) Run Algorithm 5 with the design matrix X.

A.3.2 Selection of tuning parameters

Algorithm 6 required selection of two hyper-parameters. Under the null model, no sparse corruption exists: indeed if $\beta^0 = 0$, so is ω^0 since $\omega^0 = \frac{1}{\sqrt{n}}(X - \tilde{X})\beta^0 = 0$. This property allows us to opt for the Quantile Universal Threshold (QUT) methodology (Giacobino et al., 2017), generally driven by model selection rather than prediction.

QUT selects the tuning parameter so that under the null model ($\beta^0 = 0$), the null vector $\hat{\beta} = 0$ is recovered with probability $1 - \alpha$. Under the null model, $y = \epsilon$ whatever the missing data pattern is. Then given a fixed value of λ and a fixed imputed matrix $\tilde{X}$, the corresponding QUT value of τ is the upper α-quantile of $\|\hat{\beta}^{\text{med}}_\lambda(\epsilon)\|_\infty$, where $\hat{\beta}^{\text{med}}_\lambda(\epsilon)$ is the vectors of medians obtained at stage 2)) of Algorithm 5 applied to $\tilde{X}$ and $y = \epsilon$. To free ourselves from preliminary estimation of the noise level σ, we exploit the noise coefficients $\hat{\gamma}^{(k)}$ of Robust Lasso-Zero to pivotize the statistic $\|\hat{\beta}^{\text{med}}_\lambda(\epsilon)\|_\infty$, as explained in Descloux and Sardy (2020).

For every $\lambda > 0$, there is a pair of QUT parameters $(\lambda, \tau^{\text{QUT}}_\alpha(y; \lambda))$ at level α. The remaining question is how to choose λ. For a fair isotropic penalty on β, ω and γ, we fix $\lambda = 1$.

A.4 Numerical experiments

We evaluate the performance of Robust-Lasso Zero when missing data affect the design matrix. The code reproducing these experiments is available at `https://github.com/pascalinedescloux/robust-lasso-zero-NA`.

A.4.1 Simulation settings

Simulation scenarios. We generate data according to model (A.1) with the covariates matrix obtained by drawing $n = 200$ observations from a Gaussian distribution $\mathcal{N}(0, \Sigma)$, where$\Sigma \in \mathbb{R}^{200\times 200}$ is a Toeplitz matrix, such that $\Sigma_{ij} = \rho^{|i-j|}$; the variance of the noise $\sigma = 0.5$ and the coefficient β^0 are drawn uniformly from $\{\pm 1\}$. We vary the following parameters:

- Correlation structures indexed by ρ with $\rho = 0$ (uncorrelated) and $\rho = 0.75$ (correlated);
- Sparsity degrees indexed by s with $s \in \{3, 10\}$.

Before generating the response vector y, all columns of X are mean-centered and standardized; Missing data are then introduced in X according to two different mechanisms, MCAR or MNAR, and in two different proportions. Any entry of X is missing according to the following logistic model

$$\mathbb{P}(X_{ij}^{\mathrm{NA}} = \texttt{NA} \mid X_{ij} = x) = \frac{1}{1 + e^{-a|x|-b}},$$

where $a \geqslant 0$ and $b \in \mathbb{R}$. Choosing $a = 0$ yields MCAR data, whereas $a = 5$ leads to MNAR setting in which high absolute entries are more likely to be missing. For a fixed a, the value of b is chosen so that the overall average proportion of missing values is π, with $\pi = 5\%$ and $\pi = 20\%$.

Two sets of simulations are run. The first one is "s-oracle", meaning that the tuning parameters of the different methods are chosen so that the estimated support has correct size s. In the second set, no knowledge of s, β^0 or σ is provided.

Estimators considered. We compare the following estimators:

- **Rlass0:** the Robust Lasso-Zero described in Algorithm 6 using M equal to 30. The tuning parameters are obtained using $\lambda = 1$ and selecting τ by quantile universal threshold (QUT) at level $\alpha = 0.05$.
- **lass0:** the Lasso-Zero proposed in Descloux and Sardy (2020). The automatic tuning is performed by QUT, at level $\alpha = 0.05$.
- **lasso:** the Lasso (Tibshirani, 1996) performed on the mean-imputed matrix whre the regularization parameter is tuned by cross-validation.
- **NClasso:** the nonconvex ℓ_1 estimator of Loh and Wainwright (2012). It is only included under the s-oracle setting, as selection of the tuning parameter in practice is not discussed in their work.
- **ABSLOPE:** Adaptive Bayesian SLOPE of Jiang et al. (2019).

Performance evaluation. The performance of each estimator is assessed in terms of the following criteria, averaged over 100 replications:

- the Probability of Sign Recovery (PSR), PSR = $\mathbb{P}(\text{sign}(\hat{\beta}) = \text{sign}(\beta^0))$,
- the signed True Positive Rate (sTPR), where s-TPR = $\mathbb{E}$(s-TPP) with

$$\text{s-TPP} := \frac{|\{j \mid \beta_j^0 > 0, \hat{\beta}_j > 0\}| + |\{j \mid \beta_j^0 < 0, \hat{\beta}_j < 0\}|}{|S^0|}, \tag{A.16}$$

  which is the proportion of nonzero coefficients whose sign is correctly identified;

- the signed False Discovery Rate (sFDR): s-FDR = $\mathbb{E}$(s-FDP) with

$$\text{s-FDP} := \frac{|\hat{S}| - |\{j \mid \beta_j^0 > 0, \hat{\beta}_j > 0\}| - |\{j \mid \beta_j^0 < 0, \hat{\beta}_j < 0\}|}{\max\{1, |\hat{S}|\}}, \tag{A.17}$$

  which is the proportion of incorrect signs among all discoveries.

A.4.2 Results

A.4.2.1 With *s*-oracle hyperparameter tuning

Under the s-oracle tuning, an s-TPP (A.16) of one means that the signs of β^0 are exactly recovered, and the s-TPP is related to the s-FDP (A.17) through s-FDP = 1 − s-TPP. That is why, in Figure A.1, only the average s-TPP and the estimated probability of sign recovery are reported.

Small missingness – High sparsity (5% of NA and $s = 3$). In the non-correlated case, in Figure A.1 (a) and (c), MCAR and MNAR results are similar across methods. With correlation, in FigureA.1 (b) and (d), Rlass0 improves PSR and sTPR, specially with MNAR data.

Increasing missingness – High sparsity (20% of NA and $s = 3$). The benefit of Rlass0 is noticeable when increasing the percentage of missing data to 20%, for both performance indicators. Indeed, with no correlation (Figure A.1 (a)(c)(bottom left)), the improvement is clear when dealing with MNAR. With correlation (Figure A.1 (b)(d)(bottom left)), Rlass0 outperforms the other methods: while the improvement can be marginal when compared to lass0 for MCAR, it becomes significant for MNAR.

Lower sparsity ($s = 10$). The performance of all estimators tends to deteriorate. One can identify two groups of estimators: Rlass0 and lass0 generally outperforms lasso and NClasso, except with a high proportion (20%) of MNAR missing data for which they all behave the same. While comparable when $s = 10$, Rlass0 proves to be better than lass0 in the case of a small proportion of MNAR missing data (5%).

A.4.2.2 With automatic hyperparameter tuning

Figures A.2 and A.3 point to the poor performance of lasso in terms of PSR for all experimental settings. The automatic tuning, being done by cross-validation, is known to lead to support overestimation. Indeed, its very good performance in sTPR is made at the cost of a very high sFDR.

Small missingness – High sparsity (5% of NA and $s = 3$). In Figures A.2(a)(top left) and A.3(a)(c)(top left), for the non-correlated case, Rlass0, lass0 and ABSlope performs very well, providing a PSR and s-TPR of one, and a s-FDR of zero, either when dealing with MCAR or MNAR data (the lasso being already out of the game). In Figures A.2(b)(top left) and A.3(b)(d)(top left), adding correlation in the design matrix seems beneficial for ABSlope, at the price of high FDR, however.

Increasing missingness – High sparsity (20% of NA and $s = 3$). With no correlation, one sees in Figure A.2(a)(bottom left) that Rlass0 provides the best PSR, whatever the type of missing data is. One could also note that the performances in terms of PSR of either lass0 or ABSLOPE are extremely variable depending on the type of missing data (MCAR or MNAR) considered: the PSR of lass0 is comparable to the one of Rlass0 when facing MCAR data and is much lower than the one of Rlass0 when facing MNAR data; the converse is true for ABSLOPE.

Regarding the s-TPR and s-FDR results in Figure A.3 (a-d)(bottom left), the following observations hold in both correlated or non-correlated cases:

(i) With MCAR data, all the methods behave similarly in terms of s-TPR, identifying correctly signs and coefficient locations in the support of β^0, see Figure A.3(a)(b)(bottom left);

(ii) With MNAR data, lasso and ABSLOPE remain stable in terms of s-TPR, providing an s-TPR of one, whereas the s-TPR of Rlass0 deteriorates (to 0.6 and 0.5 respectively for the non-correlated and correlated cases), and even worse for lass0, see Figure A.3(a)(b)(bottom left);

(iii) Lasso and ABSLOPE lead to high s-FDR, while lass0 and Rlass0 always give the best s-FDR, see Figure A.3(c)(d)(bottom left).

Lower sparsity ($s = 10$). For low missingness (5%), see Figure A.2 (a)(b) (top right), ABSLOPE gives high PSR. In terms of s-TPR, lasso and ABSLOPE have high TPR. Moreover Rlass0 improves s-TPR compared to lass0 specially for a small proportion of MNAR missing data. In terms of s-FDR, lass0 and Rlass0 bring very low s-FDR, proving their FDR stability with respect to MCAR/MNAR data, and correlation.

Variable	Rlass0	lass0	lasso	ABSLOPE
Age	−	0	−	−
SI	0	0	0	−
Delta.hemo	0	0	0	+
Lactates	0	0	0	+
Temperature	0	0	0	+
VE	−	0	−	0
RBC	−	0	0	−
DBP.min	0	0	−	+
HR.max	0	0	−	0
SI.amb	0	0	0	+

Table A.1: Sign of estimated effects on the platelet for Rlass0, lass0, lasso or ABSLOPE. Variables not shown here are not selected by any method.

A.4.2.3 Summary and discussion

The results of experiments with s-oracle tuning (Section A.4.2.1) show that Robust Lasso-Zero performs better than competitors for sign recovery, and is more robust to MNAR data compared to its nonrobust counterpart when the sparsity index and/or proportion of missing entries is low. In particular, Robust Lasso-Zero performs better than NClasso, one of the rare existing ℓ_1-estimator designed to handle missing values.

While not designed to handle MNAR data, ABSLOPE appears to be a valid competitor in terms of s-TPR or PSR when the model complexity increases, and when dealing with MNAR data. Its poor performance in FDR in such settings reveals its tendency to overestimate the support of β^0, under higher sparsity degrees, and with informative MNAR missing data.

With automatic tuning (Section A.4.2.2), Robust Lasso-Zero is the best method overall. Moreover, our results show that the choice of Robust Lasso-Zero tuned by QUT, with its low s-FDR, is particularly appropriate in cases where one wants to maintain a low proportion of false discoveries.

A.5 Application to the Traumabase® dataset

We illustrate our approach on the public health APHP (Assistance Publique Hopitaux de Paris) TraumaBase® Group for traumatized patients. Effective and timely management of trauma is crucial to improve outcomes, as delays or errors entail high risks for the patient.

In our analysis, we focuse on one specific challenge: selecting a sparse model from data containing missing covariates in order to explain the level of platelet. This model can aid creating an innovative response to the public health challenge of major trauma. Explanatory variables for the level of platelet consist in fifteen quantitative variables containing missing values, which have been selected by doctors. They give clinical measurements on 490 patients. In Figure A.4, one sees the percentage of missing values in each variable, varying from 0 to 45% and leading to 20% is the whole dataset. Based on discussions with doctors, some

variables may have informative missingness (M(N)AR variables). Both percentage and nature of missing data demonstrate the importance of taking appropriate account of missing data. More information can be found in Appendix A.8.

We compare Robust Lasso-Zero to Lasso-Zero, Lasso and ABSLOPE. The signs of the coefficients are shown in Table A.1. Lass0 does not select any variable, whereras its robust counterpart selects three. According to doctors, Robust Lasso-Zero is the most coherent. Indeed, a negative effect of age (*Age*), vascular filling (*VE*) and blood transfusion (*RBC*) was expected, as they all result in low platelet levels and therefore a higher risk of severe bleeding. Lasso similarly selects *Age* and *VE*, but also minimum value of diastolic blood pressure *DBP.min* and the maximum heart rate *HR.max*. The effect of *DBP.min* is not what doctors expected. For ABSLOPE, the effects on platelets of delta Hemocue (*Delta.Hemocue*), the lactates (*Lactates*), the temperature (*Temperature*) and the shock index measured on ambulance (*SI.amb*), at odds with the effect of the shock index at hospital (*SI*), are not in agreement with the doctors opinion either.

A.6 Proof of Theorem 25

Lemma 5 implies that under the sign invariance assumption ((i)), identifiability of $(\beta^{(r)}, \omega^{(r)})$ is equivalent to identifiability of $(\theta, \tilde{\theta})$.

Proof of Lemma 5. Note that $(\hat{\beta}^{\mathrm{JP}}_\lambda, \hat{\omega}^{\mathrm{JP}}_\lambda)$ is a solution to JP (A.7) if and only if $(\hat{\beta}^{\mathrm{JP}}_\lambda, \hat{\omega}^{\mathrm{JP}}_\lambda) = (\tilde{\beta}, \lambda^{-1}\tilde{\omega})$, where $(\tilde{\beta}, \tilde{\omega})$ is a solution to

$$\min_{(\beta,\ \omega)\in\mathbb{R}^p\times\mathbb{R}^n} \|\beta\|_1 + \|\omega\|_1 \quad \text{s.t.} \quad y = X\beta + \sqrt{n}\lambda^{-1}\omega. \tag{A.18}$$

So (β^0, ω^0) is identifiable with respect to X and $\lambda > 0$ if and only if the pair $(\beta^0, \lambda\omega^0)$ is the unique solution of (A.18) when $y = X\beta^0 + \sqrt{n}\omega^0$. But (A.18) is just Basis Pursuit with response vector $y \in \mathbb{R}^n$ and augmented matrix $\begin{bmatrix} X & \sqrt{n}\lambda^{-1} I_n \end{bmatrix}$, so by a result of Daubechies et al. (2010) this is the case if and only if for every $(\beta, \omega) \neq (0, 0)$ such that $X\beta + \sqrt{n}\lambda^{-1}\omega = 0$, we have $|\operatorname{sign}(\beta^0)^T\beta + \operatorname{sign}(\omega^0)^T\omega| < \|\beta_{\overline{S^0}}\|_1 + \|\omega_{\overline{T^0}}\|_1$, which proves our statement. □

We will need the following auxiliary lemma.

Lemma 6. *Under assumptions ((i)) and ((ii)), if the pair $(\theta, \tilde{\theta})$ is identifiable with respect to X and λ, then for any $\epsilon \in \mathbb{R}^n$,*

$$\lim_{r\to+\infty} \frac{1}{u_r} \begin{bmatrix} \hat{\beta}^{\mathrm{JP}(r)}_\lambda - \beta^{(r)} \\ \hat{\omega}^{\mathrm{JP}(r)}_\lambda - \omega^{(r)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

where $u_r := \|\beta^{(r)}\|_1 + \lambda\|\omega^{(r)}\|_1$.

Proof. First note that by assumption ((ii)), $\lim_{r\to+\infty} u_r = +\infty$. Now let $\epsilon \in \mathbb{R}^n$ and denote by $(\hat{\beta}^{\mathrm{JP}}_\lambda(\epsilon), \hat{\omega}^{\mathrm{JP}}_\lambda(\epsilon))$ the JP solution when $y = \epsilon$. In particular, one has $\epsilon = X\hat{\beta}^{\mathrm{JP}}_\lambda(\epsilon) + \sqrt{n}\hat{\omega}^{\mathrm{JP}}_\lambda(\epsilon)$, so for every $r \in \mathbb{N}^*$,

$$y^{(r)} = X(\beta^{(r)} + \hat{\beta}^{\mathrm{JP}}_\lambda(\epsilon)) + \sqrt{n}(\omega^{(r)} + \hat{\omega}^{\mathrm{JP}}_\lambda(\epsilon)).$$

Hence $(\beta^{(r)} + \hat{\beta}_\lambda^{\mathrm{JP}}(\epsilon), \omega^{(r)} + \hat{\omega}_\lambda^{\mathrm{JP}}(\epsilon))$ is feasible for JP when $y = y^{(r)}$, so

$$\begin{aligned}
&\frac{\|\hat{\beta}_\lambda^{\mathrm{JP}(r)}\|_1 + \lambda\|\hat{\omega}_\lambda^{\mathrm{JP}(r)}\|_1}{u_r} \\
&\qquad \leqslant \frac{\|\beta^{(r)} + \hat{\beta}_\lambda^{\mathrm{JP}}(\epsilon)\|_1 + \lambda\|\omega^{(r)} + \hat{\omega}_\lambda^{\mathrm{JP}}(\epsilon)\|_1}{u_r} \\
&\qquad \leqslant \frac{(\|\beta^{(r)}\|_1 + \lambda\|\omega^{(r)}\|_1) + (\|\hat{\beta}_\lambda^{\mathrm{JP}}(\epsilon)\|_1 + \lambda\|\hat{\omega}_\lambda^{\mathrm{JP}}(\epsilon)\|_1)}{u_r} \\
&\qquad = 1 + \frac{\|\hat{\beta}_\lambda^{\mathrm{JP}}(\epsilon)\|_1 + \lambda\|\hat{\omega}_\lambda^{\mathrm{JP}}(\epsilon)\|_1}{u_r}.
\end{aligned} \tag{A.19}$$

Therefore

$$\begin{aligned}
&\frac{1}{u_r}(\|\hat{\beta}_\lambda^{\mathrm{JP}(r)} - \beta^{(r)}\|_1 + \lambda\|\hat{\omega}_\lambda^{\mathrm{JP}(r)} - \omega^{(r)}\|_1) \\
&\qquad \leqslant \frac{1}{u_r}((\|\beta^{(r)}\|_1 + \lambda\|\omega^{(r)}\|_1) + (\|\hat{\beta}_\lambda^{\mathrm{JP}(r)}\|_1 + \lambda\|\hat{\omega}_\lambda^{\mathrm{JP}(r)}\|_1)) \\
&\qquad = 1 + \frac{\|\hat{\beta}_\lambda^{\mathrm{JP}(r)}\|_1 + \lambda\|\hat{\omega}_\lambda^{\mathrm{JP}(r)}\|_1}{u_r} \\
&\qquad \leqslant 2 + \frac{\|\hat{\beta}_\lambda^{\mathrm{JP}}(\epsilon)\|_1 + \lambda\|\hat{\omega}_\lambda^{\mathrm{JP}}(\epsilon)\|_1}{u_r},
\end{aligned} \tag{A.20}$$

using (A.19) for last inequality. Since $\lim_{r\to+\infty} \frac{\|\hat{\beta}_\lambda^{\mathrm{JP}}(\epsilon)\|_1 + \lambda\|\hat{\omega}_\lambda^{\mathrm{JP}}(\epsilon)\|_1}{u_r} = 0$, and since $\begin{bmatrix}\beta\\ \omega\end{bmatrix} \mapsto \|\beta\|_1 + \lambda\|\omega\|_1$ defines a norm on $\mathbb{R}^{p+n}$, one deduces that the sequence $\frac{1}{u_r}\begin{bmatrix}\hat{\beta}_\lambda^{\mathrm{JP}(r)} - \beta^{(r)}\\ \hat{\omega}_\lambda^{\mathrm{JP}(r)} - \omega^{(r)}\end{bmatrix}$ is bounded. Therefore we need to check that every convergent subsequence converges to zero. Let

$$\frac{1}{u_{\phi(r)}}\begin{bmatrix}\hat{\beta}_\lambda^{\mathrm{JP}(\phi(r))} - \beta^{(\phi(r))}\\ \hat{\omega}_\lambda^{\mathrm{JP}(\phi(r))} - \omega^{(\phi(r))}\end{bmatrix}$$

(with $\phi : \mathbb{N}^* \to \mathbb{N}^*$ strictly increasing) be an arbitrary convergent subsequence. Since

$$\frac{\|\beta^{(r)}\|_1 + \lambda\|\omega^{(r)}\|_1}{u_r} = 1 \tag{A.21}$$

for every r, and by (A.19), the sequences $\frac{1}{u_r}\begin{bmatrix}\beta^{(r)}\\ \omega^{(r)}\end{bmatrix}$ and $\frac{1}{u_r}\begin{bmatrix}\hat{\beta}_\lambda^{\mathrm{JP}(r)}\\ \hat{\omega}_\lambda^{\mathrm{JP}(r)}\end{bmatrix}$ are bounded as well. Hence without loss of generality (otherwise, reduce the subsequence),

$$\lim_{r\to+\infty} \frac{1}{u_{\phi(r)}}\begin{bmatrix}\beta^{(\phi(r))}\\ \omega^{(\phi(r))}\end{bmatrix} = \begin{bmatrix}\nu_1\\ \nu_2\end{bmatrix}, \tag{A.22}$$

and

$$\lim_{r\to+\infty} \frac{1}{u_{\phi(r)}} \begin{bmatrix} \hat{\beta}_\lambda^{\mathrm{JP}(\phi(r))} \\ \hat{\omega}_\lambda^{\mathrm{JP}(\phi(r))} \end{bmatrix} = \begin{bmatrix} \nu_1' \\ \nu_2' \end{bmatrix} \tag{A.23}$$

for some $\begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}, \begin{bmatrix} \nu_1' \\ \nu_2' \end{bmatrix} \in \mathbb{R}^{p+n}$. By (A.21), one necessarily has

$$\|\nu_1\|_1 + \lambda\|\nu_2\|_1 = 1, \tag{A.24}$$

and (A.19) implies that

$$\|\nu_1'\|_1 + \lambda\|\nu_2'\|_1 \leqslant 1. \tag{A.25}$$

Now

$$\begin{aligned} \lim_{r\to+\infty} \frac{X(\hat{\beta}_\lambda^{\mathrm{JP}(r)} - \beta^{(r)}) + \sqrt{n}(\hat{\omega}_\lambda^{\mathrm{JP}(r)} - \omega^{(r)})}{u_r} &= \lim_{r\to+\infty} \frac{y^{(r)} - (X\beta^{(r)} + \sqrt{n}\omega^{(r)})}{u_r} \\ &= \lim_{r\to+\infty} \frac{\epsilon}{u_r} = 0, \end{aligned}$$

so one deduces that

$$\lim_{r\to+\infty} \begin{bmatrix} X & \sqrt{n}I_n \end{bmatrix} \begin{bmatrix} \hat{\beta}_\lambda^{\mathrm{JP}(\phi(r))}/u_{\phi(r)} \\ \hat{\omega}_\lambda^{\mathrm{JP}(\phi(r))}/u_{\phi(r)} \end{bmatrix} = \lim_{r\to+\infty} \begin{bmatrix} X & \sqrt{n}I_n \end{bmatrix} \begin{bmatrix} \beta^{(\phi(r))}/u_{\phi(r)} \\ \omega^{(\phi(r))}/u_{\phi(r)} \end{bmatrix},$$

so by (A.22) and (A.23),

$$\begin{bmatrix} X & \sqrt{n}I_n \end{bmatrix} \begin{bmatrix} \nu_1' \\ \nu_2' \end{bmatrix} = \begin{bmatrix} X & \sqrt{n}I_n \end{bmatrix} \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}. \tag{A.26}$$

Assuming for now that (ν_1, ν_2) is identifiable with respect to X and λ, equality (A.26) together with (A.24) and (A.25) imply that $\begin{bmatrix} \nu_1' \\ \nu_2' \end{bmatrix} = \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}$, hence

$$\lim_{r\to+\infty} \frac{1}{u_{\phi(r)}} \begin{bmatrix} \hat{\beta}_\lambda^{\mathrm{JP}(\phi(r))} - \beta^{(\phi(r))} \\ \hat{\omega}_\lambda^{\mathrm{JP}(\phi(r))} - \omega^{(\phi(r))} \end{bmatrix} = \begin{bmatrix} \nu_1' \\ \nu_2' \end{bmatrix} - \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

It remains to check that (ν_1, ν_2) is identifiable with respect to X and λ, which we will do using Lemma 5. Note that (A.22) and assumption ((i)) imply

$$\operatorname{sign}(\nu_1) = \theta - \theta', \tag{A.27}$$

$$\operatorname{sign}(\nu_2) = \tilde{\theta} - \tilde{\theta}', \tag{A.28}$$

where $\theta_j' := \theta_j \mathbb{1}_{\{\nu_{1,j}=0, \theta_j \neq 0\}}$, and $\tilde{\theta}_j' = \tilde{\theta}_j \mathbb{1}_{\{\nu_{2,j}=0, \tilde{\theta}_j \neq 0\}}$, and hence

$$\overline{\operatorname{supp}(\nu_1)} = \overline{\operatorname{supp}(\theta)} \sqcup \operatorname{supp}(\theta') = \overline{S^0} \sqcup \operatorname{supp}(\theta'), \tag{A.29}$$

$$\overline{\operatorname{supp}(\nu_2)} = \overline{\operatorname{supp}(\tilde{\theta})} \sqcup \operatorname{supp} \tilde{\theta}' = \overline{T^0} \sqcup \operatorname{supp} \tilde{\theta}'. \tag{A.30}$$

Consider a pair $(\beta, \omega) \neq (0,0)$ such that $X\beta + \sqrt{n}\lambda^{-1}\omega = 0$. By (A.27) and (A.28),

$$\begin{aligned} |\operatorname{sign}(\nu_1)^T\beta + \operatorname{sign}(\nu_2)^T\omega| &= |(\theta-\theta')^T\beta + (\tilde{\theta}-\tilde{\theta}')^T\omega| \\ &\leqslant |\theta^T\beta + \tilde{\theta}^T\omega| + |(\theta')^T\beta| + |(\tilde{\theta}')^T\omega|. \end{aligned} \tag{A.31}$$

But since $(\theta, \tilde{\theta})$ is identifiable with respect to X and λ, Lemma 5 implies $|\theta^T\beta + \tilde{\theta}^T\omega| < \|\beta_{\overline{S^0}}\|_1 + \|\omega_{\overline{T^0}}\|_1$. Plugging this into (A.31) gives

$$\begin{aligned} |\operatorname{sign}(\nu_1)^T\beta + \operatorname{sign}(\nu_2)^T\omega| &< \|\beta_{\overline{S^0}}\|_1 + \|\omega_{\overline{T^0}}\|_1 + |(\theta')^T\beta| + |(\tilde{\theta}')^T\omega| \\ &\leqslant \|\beta_{\overline{S^0}}\|_1 + \|\beta_{\operatorname{supp}(\theta')}\|_1 + \|\omega_{\overline{T^0}}\|_1 + \|\omega_{\operatorname{supp}(\tilde{\theta}')}\|_1 \\ &= \|\beta_{\overline{\operatorname{supp}(\nu_1)}}\|_1 + \|\omega_{\overline{\operatorname{supp}(\nu_2)}}\|_1, \end{aligned}$$

where the equality comes from (A.29) and (A.30). By Lemma 5, one concludes that (ν_1, ν_2) is identifiable with respect to X and λ. □

Proof of Theorem 25. Let us assume that $(\theta, \tilde{\theta})$ is identifiable with respect to X and λ, and let $\epsilon \in \mathbb{R}^n$. By Lemma 6,

$$\lim_{r\to+\infty} \frac{1}{u_r} \begin{bmatrix} \hat{\beta}_\lambda^{\mathrm{JP}(r)} - \beta^{(r)} \\ \hat{\omega}_\lambda^{\mathrm{JP}(r)} - \omega^{(r)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{A.32}$$

Since

$$\min\{1,\lambda\} \max\{\|\beta^{(r)}\|_\infty, \|\omega^{(r)}\|_\infty\} \leqslant u_r \leqslant (|S^0| + \lambda|T^0|) \max\{\|\beta^{(r)}\|_\infty, \|\omega^{(r)}\|_\infty\},$$

(A.32) is equivalent to $\lim_{r\to+\infty} \frac{1}{\max\{\|\beta^{(r)}\|_\infty, \|\omega^{(r)}\|_\infty\}} \begin{bmatrix} \hat{\beta}_\lambda^{\mathrm{JP}(r)} - \beta^{(r)} \\ \hat{\omega}_\lambda^{\mathrm{JP}(r)} - \omega^{(r)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Therefore there exists $R > 0$ such that for every $r \geqslant R$,

$$\|\hat{\beta}_\lambda^{\mathrm{JP}(r)} - \beta^{(r)}\|_\infty < \frac{q}{2} \max\{\|\beta^{(r)}\|_\infty, \|\omega^{(r)}\|_\infty\} \tag{A.33}$$

and

$$\|\hat{\omega}_\lambda^{\mathrm{JP}(r)} - \omega^{(r)}\|_\infty < \frac{q}{2} \max\{\|\beta^{(r)}\|_\infty, \|\omega^{(r)}\|_\infty\}. \tag{A.34}$$

Setting $\tau := \frac{q}{2} \max\{\|\beta^{(r)}\|_\infty, \|\omega^{(r)}\|_\infty\}$, (A.33) implies that $|\hat{\beta}_{\lambda,j}^{\mathrm{JP}(r)}| < \tau$ for every $j \notin S^0$, hence $\hat{\beta}_{(\lambda,\tau),j}^{\mathrm{TJP}(r)} = 0$. If $j \in S^0$, assumption ((iii)) implies

$$|\beta_j^{(r)}| \geqslant \beta_{\min}^{(r)} \geqslant 2\tau, \tag{A.35}$$

and by (A.33), we have

$$|\hat{\beta}_{\lambda,j}^{\mathrm{JP}(r)} - \beta_j^{(r)}| < \tau, \tag{A.36}$$

so (A.35) and (A.36) together imply $|\hat{\beta}_{\lambda,j}^{\mathrm{JP}(r)}| > \tau$ and $\operatorname{sign}(\hat{\beta}_{\lambda,j}^{\mathrm{JP}(r)}) = \operatorname{sign}(\beta_j^{(r)})$. So we conclude that $\operatorname{sign}(\hat{\beta}_{(\lambda,\tau)}^{\mathrm{TJP}(r)}) = \operatorname{sign}(\beta^{(r)})$. Analogously, (A.34) implies $\operatorname{sign}(\hat{\omega}_{(\lambda,\tau)}^{\mathrm{TJP}(r)}) = \operatorname{sign}(\omega^{(r)})$.

Conversely, let us assume that for some $\epsilon \in \mathbb{R}^n$, $r \in \mathbb{N}^*$ and $\tau > 0$,

$$\operatorname{sign}(\hat{\beta}^{\mathrm{TJP}(r)}_{(\lambda,\tau)}) = \theta, \quad \operatorname{sign}(\hat{\omega}^{\mathrm{TJP}(r)}_{(\lambda,\tau)}) = \tilde{\theta}. \tag{A.37}$$

Note that the JP solution $(\hat{\beta}^{\mathrm{JP}(r)}_\lambda, \hat{\omega}^{\mathrm{JP}(r)}_\lambda)$ is unique by assumption, hence $(\hat{\beta}^{\mathrm{JP}(r)}_\lambda, \hat{\omega}^{\mathrm{JP}(r)}_\lambda)$ is identifiable with respect to X and λ. Now by (A.37), all nonzero components of θ and $\tilde{\theta}$ must have the same sign as the corresponding entries of $\hat{\beta}^{\mathrm{JP}(r)}_\lambda$ and $\hat{\omega}^{\mathrm{JP}(r)}_\lambda$ respectively. Hence

$$\begin{aligned} \theta &= \operatorname{sign}(\theta) = \operatorname{sign}(\hat{\beta}^{\mathrm{JP}(r)}_\lambda) - \delta, \\ \tilde{\theta} &= \operatorname{sign}(\tilde{\theta}) = \operatorname{sign}(\hat{\omega}^{\mathrm{JP}(r)}_\lambda) - \tilde{\delta}, \end{aligned} \tag{A.38}$$

where $\delta_j = \operatorname{sign}(\hat{\beta}^{\mathrm{JP}(r)}_{\lambda,j}) \mathbb{1}_{\{\hat{\beta}^{\mathrm{JP}(r)}_{\lambda,j} \neq 0, \theta_j = 0\}}$ and $\tilde{\delta}_i = \operatorname{sign}(\hat{\omega}^{\mathrm{JP}(r)}_{\lambda,i}) \mathbb{1}_{\{\hat{\omega}^{\mathrm{JP}(r)}_{\lambda,i} \neq 0, \tilde{\theta}_i = 0\}}$, and

$$\begin{aligned} \overline{S^0} &= \overline{\operatorname{supp}(\theta)} = \overline{\operatorname{supp}(\hat{\beta}^{\mathrm{JP}(r)}_\lambda)} \sqcup \operatorname{supp}(\delta) \\ \overline{T^0} &= \overline{\operatorname{supp}(\tilde{\theta})} = \overline{\operatorname{supp}(\hat{\omega}^{\mathrm{JP}(r)}_\lambda)} \sqcup \operatorname{supp}(\tilde{\delta}). \end{aligned} \tag{A.39}$$

In order to apply Lemma 5, let us consider a pair $(\beta, \omega) \neq (0,0)$ such that $X\beta + \sqrt{n}\lambda^{-1}\omega = 0$. By (A.38), one has

$$\begin{aligned} |\theta^T\beta + \tilde{\theta}^T\omega| &= |\operatorname{sign}(\hat{\beta}^{\mathrm{JP}(r)}_\lambda)^T\beta - \delta^T\beta + \operatorname{sign}(\hat{\omega}^{\mathrm{JP}(r)}_\lambda)^T\omega - \tilde{\delta}^T\omega| \\ &\leqslant |\operatorname{sign}(\hat{\beta}^{\mathrm{JP}(r)}_\lambda)^T\beta + \operatorname{sign}(\hat{\omega}^{\mathrm{JP}(r)}_\lambda)^T\omega| + |\delta^T\beta| + |\tilde{\delta}^T\omega| \\ &\leqslant \|\beta_{\overline{\operatorname{supp}(\hat{\beta}^{\mathrm{JP}(r)}_\lambda)}}\|_1 + \|\omega_{\overline{\operatorname{supp}(\hat{\omega}^{\mathrm{JP}(r)}_\lambda)}}\|_1 + \|\beta_{\operatorname{supp}(\delta)}\|_1 + \|\omega_{\operatorname{supp}(\tilde{\delta})}\|_1 \\ &= \|\beta_{\overline{S^0}}\|_1 + \|\omega_{\overline{T^0}}\|_1, \end{aligned}$$

where we have used Lemma 5 and the fact that $(\hat{\beta}^{\mathrm{JP}(r)}_\lambda, \hat{\omega}^{\mathrm{JP}(r)}_\lambda)$ is identifiable with respect to X and λ in the last inequality, and (A.39) for the last equality. Lemma 5 concludes our proof. □

A.7 Proof of Theorem 27

Proof of Theorem 27. We define $\tilde{X} := \begin{bmatrix} X & \sqrt{n} I_n \end{bmatrix}$, and $\tilde{\nu} = \begin{bmatrix} \tilde{\beta} \\ \tilde{\omega} \end{bmatrix} := \tilde{X}^T(\tilde{X}\tilde{X}^T)^{-1}\epsilon$. We will assume for now that the following properties hold.

i) Every pair (β, ω) such that $X\beta + \sqrt{n}\omega = 0$ satisfies

$$\|\beta_{S^0}\|_1 + \lambda\|\omega_{T^0}\|_1 \leqslant \frac{1}{3}(\|\beta_{\overline{S^0}}\|_1 + \lambda\|\omega_{\overline{T^0}}\|_1),$$

ii) $\|\tilde{\nu}\|_2 \leqslant \frac{\sqrt{2}\sigma}{\left(\frac{\lambda_{\min}(\Sigma)}{4}(\sqrt{p/n}-1)^2+1\right)^{1/2}}$.

Since $\tilde{X}\tilde{\nu} = X\tilde{\beta} + \sqrt{n}\tilde{\omega} = \epsilon$, one can rewrite model (A.2) as

$$y = X(\beta^0 + \tilde{\beta}) + \sqrt{n}(\omega^0 + \tilde{\omega}).$$

By property i)) and Lemma 7 below, one has

$$\|\hat{\beta}^{\mathrm{JP}}_\lambda - (\beta^0 + \tilde{\beta})\|_1 + \lambda\|\hat{\omega}^{\mathrm{JP}}_\lambda - (\omega^0 + \tilde{\omega})\|_1 \leqslant 4(\|\tilde{\beta}_{\overline{S^0}}\|_1 + \lambda\|\tilde{\omega}_{\overline{T^0}}\|_1), \tag{A.40}$$

and therefore $\|\hat{\beta}^{\mathrm{JP}}_\lambda - (\beta^0 + \tilde{\beta})\|_1 \leqslant 4(\|\tilde{\beta}\|_1 + \lambda\|\tilde{\omega}\|_1)$. Consequently, for any $j \in [p]$ one has

$$\begin{aligned}
|\hat{\beta}^{\mathrm{JP}}_{\lambda,j} - \beta^0_j| &\leqslant |\hat{\beta}^{\mathrm{JP}}_{\lambda,j} - (\beta^0_j + \tilde{\beta}_j)| + |\tilde{\beta}_j| \leqslant \|\hat{\beta}^{\mathrm{JP}}_\lambda - (\beta^0 + \tilde{\beta})\|_1 + \|\tilde{\beta}\|_1 \\
&\leqslant 4(\|\tilde{\beta}\|_1 + \lambda\|\tilde{\omega}\|_1) + \|\tilde{\beta}\|_1 \leqslant 5(\|\tilde{\beta}\|_1 + \lambda\|\tilde{\omega}\|_1) \\
&\leqslant 5\max\{1,\lambda\}(\|\tilde{\beta}\|_1 + \|\tilde{\omega}\|_1) = 5\max\{1,\lambda\}\|\tilde{\nu}\|_1 \\
&\leqslant 5\max\{1,\lambda\}\sqrt{p+n}\|\tilde{\nu}\|_2 \leqslant \frac{5\sqrt{2}\max\{1,\lambda\}\sigma\sqrt{p+n}}{(\frac{\lambda_{\min}(\Sigma)}{4}(\sqrt{p/n}-1)^2+1)^{1/2}}
\end{aligned}$$

where we have used property ii)) in the last inequality. Now setting

$$\tau := \frac{5\sqrt{2}\max\{1,\lambda\}\sigma\sqrt{p+n}}{(\frac{\lambda_{\min}(\Sigma)}{4}(\sqrt{p/n}-1)^2+1)^{1/2}},$$

one gets

$$|\hat{\beta}^{\mathrm{JP}}_{\lambda,j} - \beta^0_j| \leqslant \tau \tag{A.41}$$

for every $j \in [p]$. If $j \in \overline{S^0}$, we have $|\hat{\beta}^{\mathrm{JP}}_{\lambda,j}| \leqslant \tau$, hence $\hat{\beta}^{\mathrm{TJP}}_{(\lambda,\tau),j} = 0$. If $j \in S^0$, assumption (A.13) implies $|\beta^0_j| > 2\tau$, which together with (A.41) gives $\mathrm{sign}(\hat{\beta}^{\mathrm{TJP}}_{(\lambda,\tau),j}) = \mathrm{sign}(\beta^0_j)$.

It remains to prove that properties i)) and ii)) hold with high probability. First, Lemma 1 in Nguyen and Tran (2013a), implies that with probability greater than $1 - ce^{-c'n}$ the matrix X satisfies the extended restricted eigenvalue property

$$\begin{aligned}
&\|\beta_{\overline{S^0}}\|_1 + \lambda\|\omega_{\overline{T^0}}\|_1 \leqslant 3(\|\beta_{S^0}\|_1 + \lambda\|\omega_{T^0}\|_1) \\
&\qquad\qquad \Downarrow \\
&\frac{1}{n}\|X\beta + \sqrt{n}\omega\|_2^2 \geqslant \gamma^2(\|\beta\|_2^2 + \|\omega\|_2^2),
\end{aligned} \tag{A.42}$$

with $\gamma^2 = \frac{\min\{\lambda_{\min}(\Sigma),1\}}{16^2}$. Property (A.42) clearly implies i)). Finally, Lemma 8 below proves that ii)) holds with probability at least $1 - 1.14^{-n} - 2e^{-\frac{1}{8}(\sqrt{p}-\sqrt{n})^2}$, which concludes our proof. □

Lemma 7. *Assume that for some sets $S^0 \subset [p]$ and $T^0 \subset [n]$, and some constant $\rho \in (0,1)$, the matrix $X \in \mathbb{R}^{n\times p}$ satisfies*

$$\|\beta_{S^0}\|_1 + \lambda\|\omega_{T^0}\|_1 \leqslant \rho(\|\beta_{\overline{S^0}}\|_1 + \lambda\|\omega_{\overline{T^0}}\|_1), \tag{A.43}$$

for every pair $(\beta, \omega) \in \mathbb{R}^p \times \mathbb{R}^n$ *such that* $X\beta + \sqrt{n}\omega = 0$. *Then for every pair* $(\tilde{\beta}, \tilde{\omega}) \in \mathbb{R}^p \times \mathbb{R}^n$, *the solution* $(\hat{\beta}^{\mathrm{JP}}_\lambda, \hat{\omega}^{\mathrm{JP}}_\lambda)$ *to JP (A.7) with* $y = X\tilde{\beta} + \sqrt{n}\tilde{\omega}$ *satisfies*

$$\|\hat{\beta}^{\mathrm{JP}}_\lambda - \tilde{\beta}\|_1 + \lambda\|\hat{\omega}^{\mathrm{JP}}_\lambda - \tilde{\omega}\|_1 \leqslant \frac{2(1+\rho)}{1-\rho}(\|\tilde{\beta}_{\overline{S^0}}\|_1 + \lambda\|\tilde{\omega}_{\overline{T^0}}\|_1).$$

Proof. This proof is a simple extension of the one of Theorem 4.14 in Foucart and Rauhut (2013). Let us consider $y = X\tilde{\beta} + \sqrt{n}\tilde{\omega}$ for an arbitrary pair $(\tilde{\beta}, \tilde{\omega})$, and define $\beta' := \hat{\beta}^{\mathrm{JP}}_\lambda - \tilde{\beta}$ and $\omega' := \hat{\omega}^{\mathrm{JP}}_\lambda - \tilde{\omega}$. Clearly $X\beta' + \sqrt{n}\omega' = 0$, so by (A.43),

$$\|\beta'_{S^0}\|_1 + \lambda\|\omega'_{T^0}\|_1 \leqslant \rho(\|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1). \tag{A.44}$$

We also have

$$\begin{aligned}
\|\tilde{\beta}\|_1 + \lambda\|\tilde{\omega}\|_1 &= \|\tilde{\beta}_{S^0}\|_1 + \|\tilde{\beta}_{\overline{S^0}}\|_1 + \lambda(\|\tilde{\omega}_{T^0}\|_1 + \|\tilde{\omega}_{\overline{T^0}}\|_1) \\
&= \|\hat{\beta}^{\mathrm{JP}}_{\lambda,S^0} - \beta'_{S^0}\|_1 + \|\tilde{\beta}_{\overline{S^0}}\|_1 + \lambda(\|\hat{\omega}^{\mathrm{JP}}_{\lambda,T^0} - \omega'_{T^0}\|_1 + \|\tilde{\omega}_{\overline{T^0}}\|_1) \\
&\leqslant \|\hat{\beta}^{\mathrm{JP}}_{\lambda,S^0}\|_1 + \|\beta'_{S^0}\|_1 + \|\tilde{\beta}_{\overline{S^0}}\|_1 + \lambda(\|\hat{\omega}^{\mathrm{JP}}_{\lambda,T^0}\|_1 + \|\omega'_{T^0}\|_1 + \|\tilde{\omega}_{\overline{T^0}}\|_1),
\end{aligned}$$

and

$$\|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1 \leqslant (\|\hat{\beta}^{\mathrm{JP}}_{\lambda,\overline{S^0}}\|_1 + \|\tilde{\beta}_{\overline{S^0}}\|_1) + \lambda(\|\hat{\omega}^{\mathrm{JP}}_{\lambda,\overline{T^0}}\|_1 + \|\tilde{\omega}_{\overline{T^0}}\|_1).$$

Adding the last two inequalities yields

$$\begin{aligned}
\|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1 + \|\tilde{\beta}\|_1 + \lambda\|\tilde{\omega}\|_1 \leqslant \|\hat{\beta}^{\mathrm{JP}}_\lambda\|_1 + \|\beta'_{S^0}\|_1 + 2\|\tilde{\beta}_{\overline{S^0}}\|_1 \\
+ \lambda(\|\hat{\omega}^{\mathrm{JP}}_\lambda\|_1 + \|\omega'_{T^0}\|_1 + 2\|\tilde{\omega}_{\overline{T^0}}\|_1),
\end{aligned}$$

and rearranging terms gives

$$\begin{aligned}
\|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1 \leqslant (\|\hat{\beta}^{\mathrm{JP}}_\lambda\|_1 + \lambda\|\hat{\omega}^{\mathrm{JP}}_\lambda\|_1) - (\|\tilde{\beta}\|_1 + \lambda\|\tilde{\omega}\|_1) \\
+ (\|\beta'_{S^0}\|_1 + \lambda\|\omega'_{T^0}\|_1) + 2(\|\tilde{\beta}_{\overline{S^0}}\|_1 + \lambda\|\tilde{\omega}_{\overline{T^0}}\|_1).
\end{aligned}$$

Using (A.44) and the fact that $\|\hat{\beta}^{\mathrm{JP}}_\lambda\|_1 + \lambda\|\hat{\omega}^{\mathrm{JP}}_\lambda\|_1 \leqslant \|\tilde{\beta}\|_1 + \lambda\|\tilde{\omega}\|_1$ by minimality of the JP solution, we get

$$\|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1 \leqslant \rho(\|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1) + 2(\|\tilde{\beta}_{\overline{S^0}}\|_1 + \lambda\|\tilde{\omega}_{\overline{T^0}}\|_1),$$

hence

$$\|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1 \leqslant \frac{2}{1-\rho}(\|\tilde{\beta}_{\overline{S^0}}\|_1 + \lambda\|\tilde{\omega}_{\overline{T^0}}\|_1). \tag{A.45}$$

Now inequality (A.44) also implies

$$\begin{aligned}
\|\beta'\|_1 + \lambda\|\omega'\|_1 &= \|\beta'_{S^0}\|_1 + \lambda\|\omega'_{T^0}\|_1 + \|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1 \\
&\leqslant (1+\rho)(\|\beta'_{\overline{S^0}}\|_1 + \lambda\|\omega'_{\overline{T^0}}\|_1)
\end{aligned} \tag{A.46}$$

and continuing (A.46) with (A.45) gives the desired inequality. □

Lemma 8. *Let* $\tilde{X} := \begin{bmatrix} X & \sqrt{n} I_n \end{bmatrix}$. *Under assumptions (iv)), (v)), (vi)) and (vii)),*

$$\|\tilde{X}^T (\tilde{X}\tilde{X}^T)^{-1} \epsilon\|_2 \leqslant \frac{\sqrt{2}\sigma}{(\frac{\lambda_{\min}(\Sigma)}{4}(\sqrt{p/n} - 1)^2 + 1)^{1/2}},$$

with probability at least $1 - 1.14^{-n} - 2e^{-\frac{1}{8}(\sqrt{p}-\sqrt{n})^2}$.

Proof. We have

$$\|\tilde{X}^T (\tilde{X}\tilde{X}^T)^{-1} \epsilon\|_2^2 = \epsilon^T (\tilde{X}\tilde{X}^T)^{-1} \epsilon \leqslant \frac{\|\epsilon\|_2^2}{\lambda_{\min}(\tilde{X}\tilde{X}^T)} = \frac{\sigma \|\frac{1}{\sigma}\epsilon\|_2^2}{\lambda_{\min}(\tilde{X}\tilde{X}^T)}.$$

Since $\|\frac{1}{\sigma}\epsilon\|_2^2 \sim \chi^2_n$, it is upper bounded by $2n$ with probability larger than $1 - 1.14^{-n}$ (a corollary of Lemma 1 in Laurent and Massart (2000)). So

$$\mathbb{P}\left(\|\tilde{\nu}\|_2 \leqslant \frac{\sqrt{2n}\sigma}{\sigma_{\min}(\tilde{X})} \right) \geqslant 1 - 1.14^{-n}. \tag{A.47}$$

Let us now bound $\sigma_{\min}(\tilde{X})$. One has

$$\sigma^2_{\min}(\tilde{X}) = \lambda_{\min}(\tilde{X}\tilde{X}^T) = \lambda_{\min}(XX^T + nI_n) = \sigma^2_{\min}(X) + n. \tag{A.48}$$

One can write $X = G\Sigma^{1/2}$ where $G \in \mathbb{R}^{n \times p}$ with $G_{ij} \overset{\text{i.i.d.}}{\sim} N(0,1)$, thus

$$\sigma_{\min}(X) \geqslant \sigma_{\min}(G)\sigma_{\min}(\Sigma^{1/2}) = \sigma_{\min}(G)\sqrt{\lambda_{\min}(\Sigma)}. \tag{A.49}$$

Now it is known (see Rudelson and Vershynin (2010), eq. (2.3)) that

$$\sigma_{\min}(G) \geqslant \frac{1}{2}(\sqrt{p} - \sqrt{n}) = \frac{\sqrt{n}}{2}(\sqrt{p/n} - 1)$$

with probability at least $1 - 2e^{-\frac{1}{8}(\sqrt{p}-\sqrt{n})^2}$. Together with (A.48) and (A.49) this gives

$$\mathbb{P}\left(\sigma_{\min}(\tilde{X}) \geqslant \left(\frac{n\lambda_{\min}(\Sigma)}{4}(\sqrt{p/n} - 1)^2 + n \right)^{1/2} \right) \geqslant 1 - 2e^{-\frac{1}{8}(\sqrt{p}-\sqrt{n})^2}.$$

With (A.47), this implies

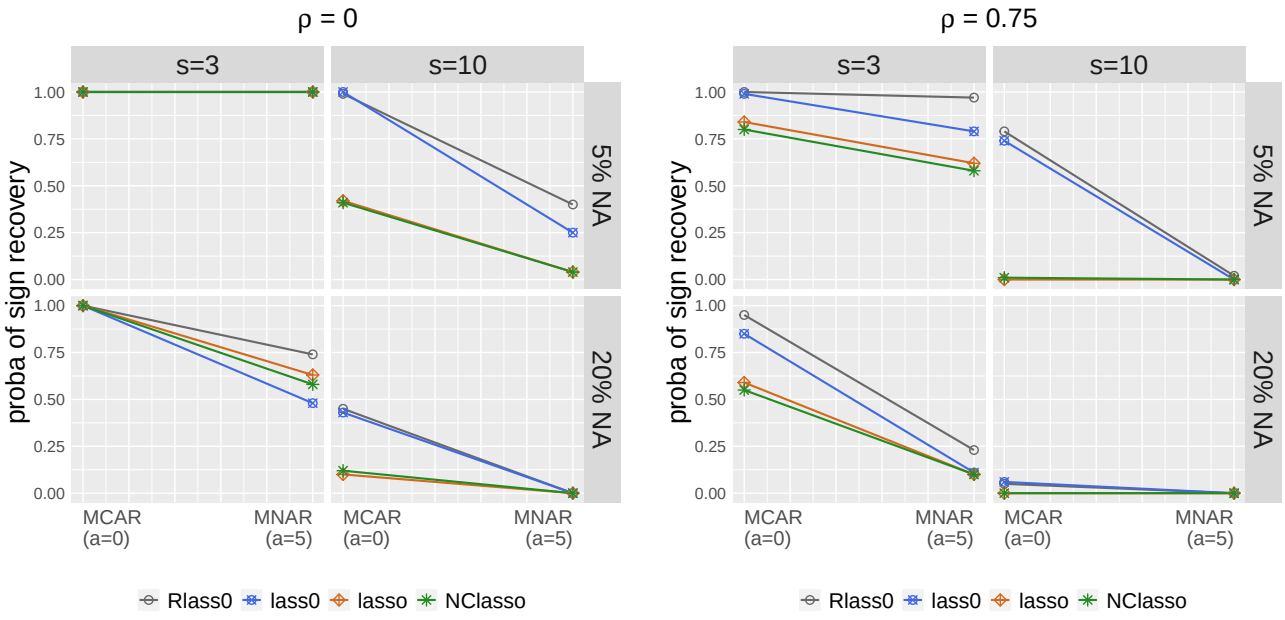
$$\mathbb{P}\left(\|\tilde{\nu}\|_2 \leqslant \frac{\sqrt{2}\sigma}{(\frac{\lambda_{\min}(\Sigma)}{4}(\sqrt{p/n} - 1)^2 + 1)^{1/2}} \right) \geqslant 1 - 1.14^{-n} - 2e^{-\frac{1}{8}(\sqrt{p}-\sqrt{n})^2}.$$

□

A.8 Variables in the Traumabase® dataset

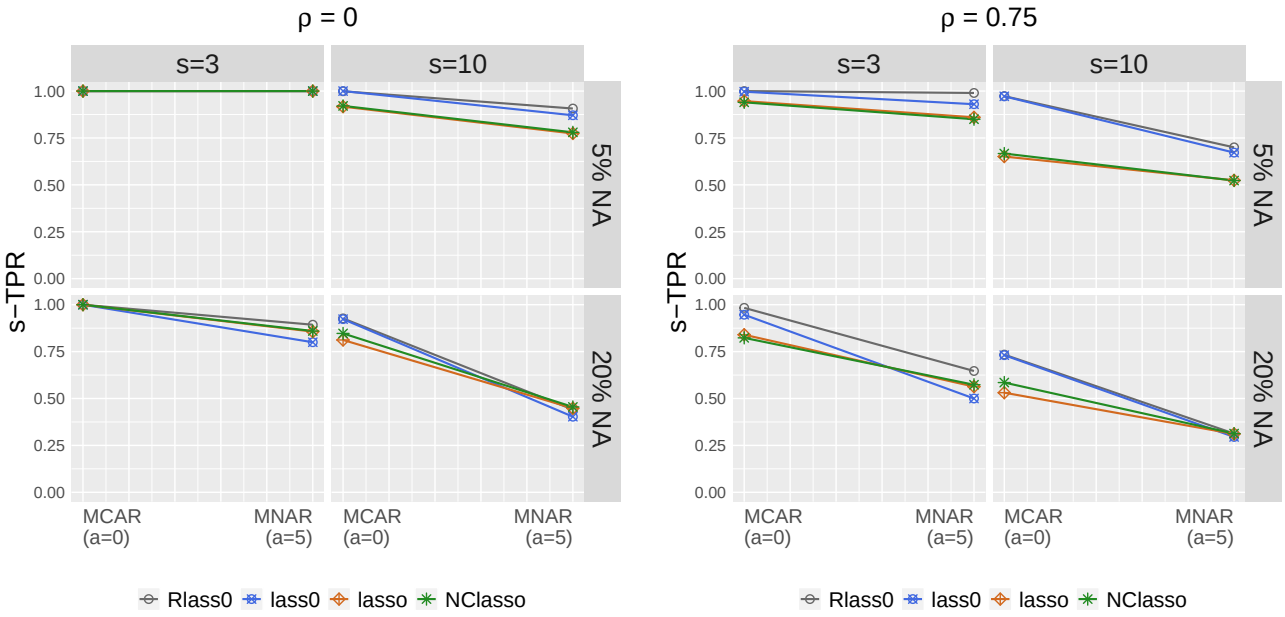
The variables of the Traumabase dataset are:

- *Time.amb*: Time spent in the ambulance, *i.e.*, transportation time from accident site to hospital, in minutes.
- *Lactate*: The conjugate base of lactic acid.
- *Delta.Hemo*: The difference between the homoglobin on arrival at hospital and that in the ambulance.
- *RBC*: A binary index which indicates whether the transfusion of Red Blood Cells Concentrates is performed.
- *SI.amb*: Shock index measured on ambulance.
- *DBP.min*: Minimum value of measured diastolic blood pressure in the ambulance.
- *SBP.min*: Minimum value of measured systolic blood pressure in the ambulance.
- *HR.max*: Maximum value of measured heart rate in the ambulance.
- *VE*: A volume expander is a type of intravenous therapy that has the function of providing volume for the circulatory system.
- *MBP.amb*: Mean arterial pressure measured in the ambulance.
- *Temp*: Patient's body temperature.
- *SI*: Shock index $SI = HR/SBP$ indicates level of occult shock based on heart rate and systolic blood pressure on arrival at hospital.
- *MBP*: Mean arterial pressure $MBP = (2DBP + SPB)/3$ is an average blood pressure in an individual during a single cardiac cycle.
- *HR*: Heart rate measured on arrival of hospital.
- *Age*: Age.



(a) PSR in the non-correlated case

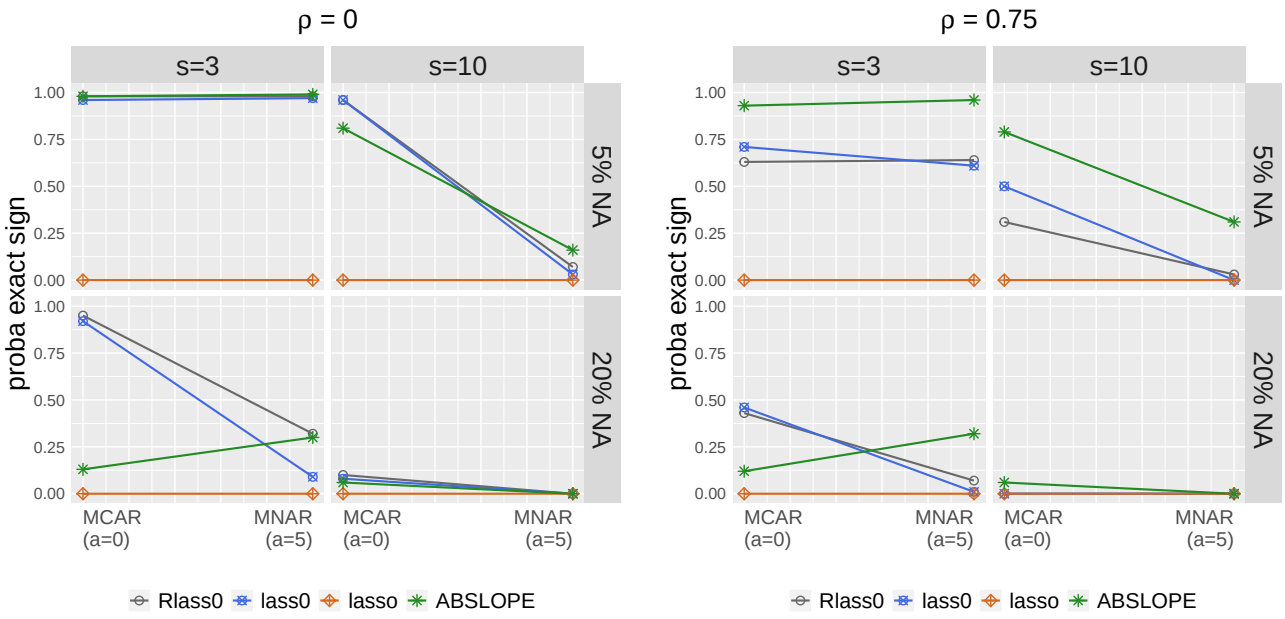
(b) PSR in the correlated case



(c) s-TPR in the non-correlated case

(d) s-TPR in the correlated case

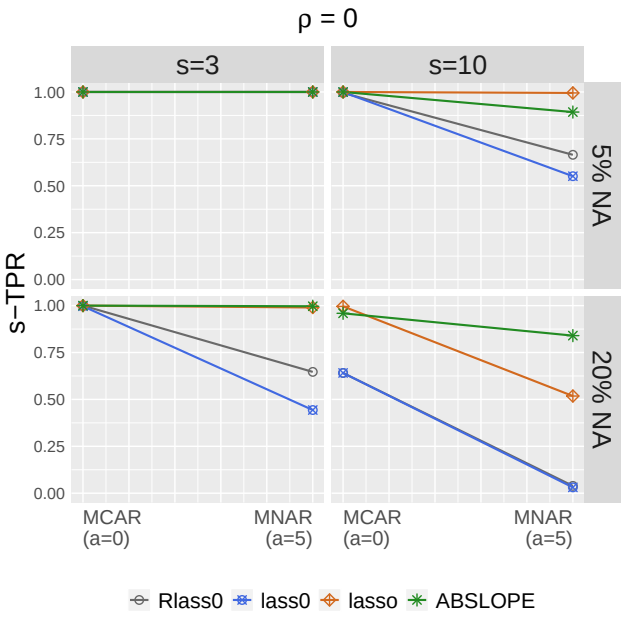
Figure A.1: PSR and s-TPR with an s-oracle tuning, for sparsity levels , $s = 3$ and $s = 10$ (subplots columns), proportions of missing values 5% or 20% (subplots rows), and two missing data mechanisms (MCAR vs MNAR).



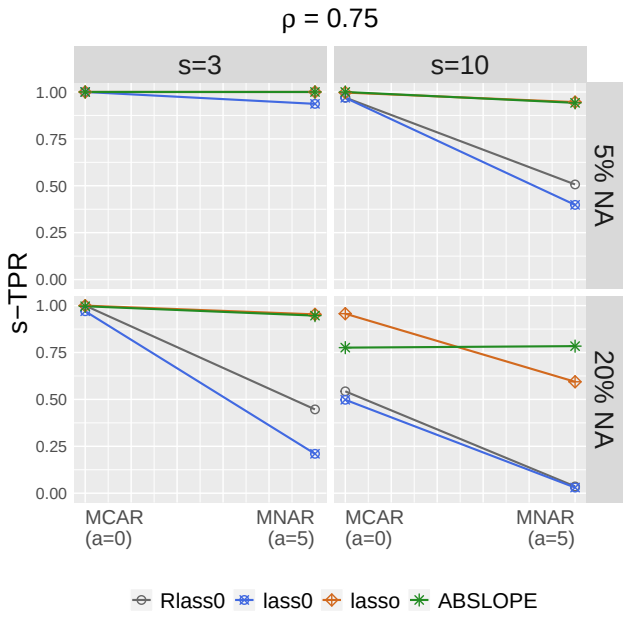
(a) PSR in the non-correlated case

(b) PSR in the correlated case

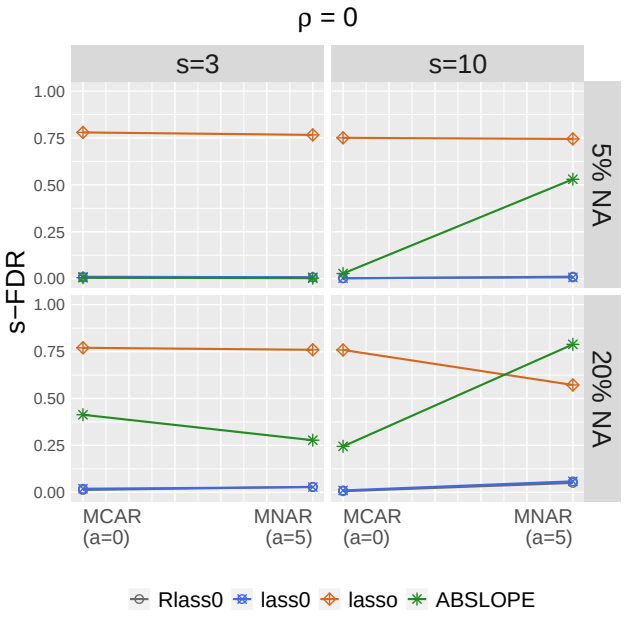
Figure A.2: PSR with automatic tuning, for sparsity levels $s = 3$ and $s = 10$ (subplots columns), proportions of missing values 5% or 20% (subplots rows), and two missing data mechanisms (MCAR vs MNAR).



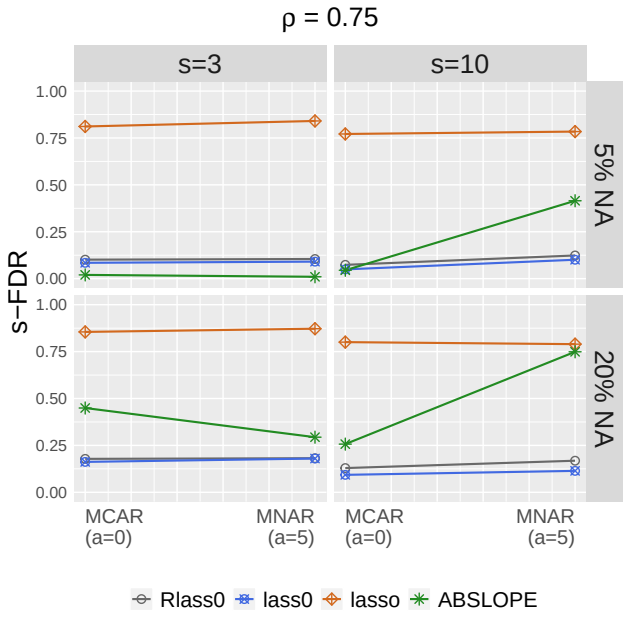
(a) s-TPR in the non-correlated case



(b) s-TPR in the correlated case



(c) s-FDR in the non-correlated case



(d) s-FDR in the correlated case

Figure A.3: s-FDR and s-TPR with automatic tuning, for sparsity levels $s = 3$ and $s = 10$ (subplots columns), proportions of missing values 5% or 20% (subplots rows), and two missing data mechanisms (MCAR vs MNAR).

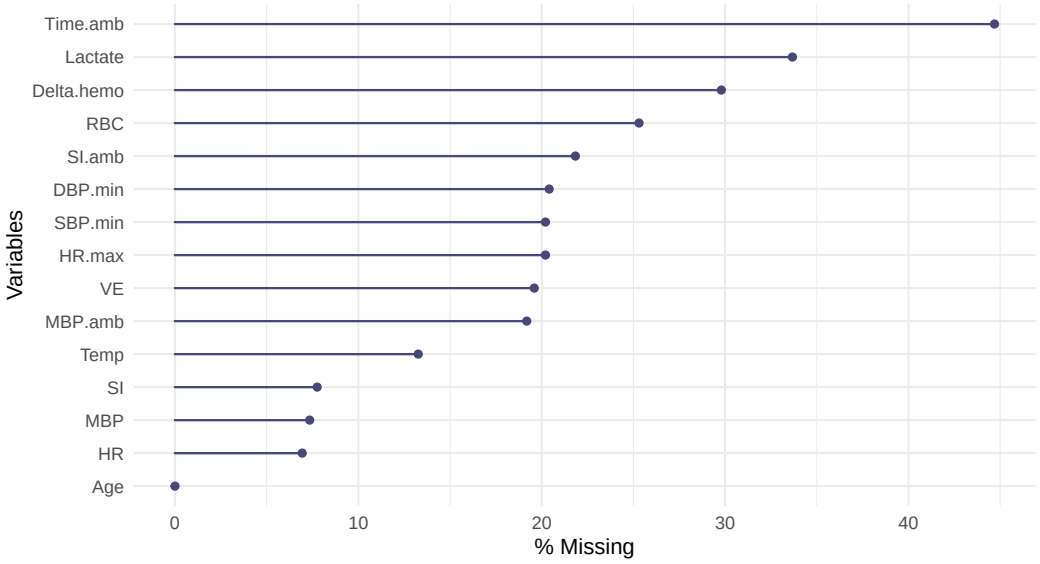


Figure A.4: Percentage of missing values in the Traumbase dataset.

Appendix B

Appendix of Chapter 2

B.1 The FISTA algorithm

We first present the proximal gradient method. The following optimisation problem is considered:

$$\hat{\Theta} \in \operatorname{argmin}_{\Theta} h_1(\Theta) + h_2(\Theta),$$

where h_1 is a convex function, h_2 a differentiable and convex function and L the gradient Lipschitz of h_2.

Algorithm 7 Proximal gradient method

Step 0: $\hat{\Theta}^{(0)}$ the null matrices.
Step $t+1$**:** $\hat{\Theta}^{(t+1)} = \operatorname{prox}_{\lambda(1/L)h_1}(\hat{\Theta}^{(t)} - (1/L)\nabla h_2(\hat{\Theta}^{(t)}))$

The main trick of the FISTA algorithm is to add a momentum term to the proximal gradient method, in order to yield smoother trajectory towards the convergence point. In addition, the proximal operator is performed on a specific linear combination of the previous two iterates, rather than on the previous iterate only.

Algorithm 8 FISTA (accelerated proximal gradient method)

Step 0: $\kappa_0 = 0.1$, $\hat{\Theta}^{(0)}$ and Ξ_0 the null matrices.
Step $t+1$**:**

$$\hat{\Theta}^{(t+1)} = \operatorname{prox}_{\lambda(1/L)h_1}(\Xi_t - (1/L)\nabla h_2(\Xi_t))$$
$$\kappa_{k+1} = \frac{1+\sqrt{1+4\kappa_k^2}}{2}$$
$$\Xi_{t+1} = \hat{\Theta}^{(t+1)} + \frac{\kappa_k - 1}{\kappa_{k+1}}(\hat{\Theta}^{(t+1)} - \hat{\Theta}^{(t)})$$

In our specific model, to solve (2.2), $h_1(\Theta) = \|\Theta\|_\star$ and $h_2(\Theta) = \|\Omega \odot (\Theta - Y)\|_F^2$. Let us precise that:

$$\frac{\partial h_2(\Theta)}{\partial \Theta_{ij}} = \nabla_{\Theta_{ij}} h_2(\Theta) = \Omega_{ij}\left(\Theta_{ij} - Y_{ij}\right).$$

Therefore,

$$\nabla h_2(\Theta) = \Omega \odot (\Theta - Y)$$

and L is equal to 1.

B.2 softImpute

We start by describing `softImpute`.

Algorithm 9 `softImpute`

Step 0**:** $\hat{\Theta}^{(0)}$ the null matrix.
Step $t+1$**:** $\hat{\Theta}^{(t+1)} = \text{prox}_{\lambda\|.\|_\star}(\Omega \odot Y + (1-\Omega) \odot \hat{\Theta}^{(t)})$

The proximal operator of the nuclear norm of a matrix X consists in a soft-thresholding of its singular values: we perform the SVD of X and we obtain the matrices U, V and D. Then

$$\text{prox}_{\lambda\|.\|_\star}(X) = U D_\lambda V.$$

D_λ is the diagonal matrix such that for all i,

$$D_{\lambda,ii} = \max((\sigma_i - \lambda), 0)$$

, where the (σ_{ii})'s are the singular values of X.

B.2.1 Equivalence between softImpute and the proximal gradient method

By using the same functions h_1 and h_2 as above, one has:

$$\begin{aligned}\hat{\Theta}^{(t+1)} &= \text{prox}_{\lambda(1/L)h_1}(\hat{\Theta}^{(t)} - (1/L)\nabla h_2(\hat{\Theta}^{(t)})) \\ &= \text{prox}_{\lambda\|.\|_\star}(\hat{\Theta}^{(t)} - \Omega \odot (\hat{\Theta}^{(t)} - Y)) \\ &= \text{prox}_{\lambda\|.\|_\star}(\Omega \odot Y + (1-\Omega) \odot \hat{\Theta}^{(t)}),\end{aligned}$$

so that `softImpute` and the proximal gradient method are similar.

B.2.2 Equivalence between the EM algorithm and iterative SVD in the MAR case

We prove here that in the MAR setting, `softImpute` is similar to the EM algorithm. Let us recall that in the MAR setting the model of the joint distribution is not needed but only the one of the data distribution, so that the E-step is written as follows:

$$\begin{aligned}Q(\Theta|\hat{\Theta}^{(t)}) &= \mathbb{E}_{Y_{\text{mis}}}\left[\log(p(\Theta; y))|Y_{\text{obs}}; \Theta = \hat{\Theta}^{(t)}\right] \\ &\propto -\sum_{i=1}^{n}\sum_{j=1}^{p}\mathbb{E}\left[\left(\frac{y_{ij} - \Theta_{ij}}{\sigma}\right)^2 |\hat{\Theta_{ij}}^{(t)}\right],\end{aligned}$$

by using (2.3) and the independance of Y_{ij}, $\forall i,j$). Then, by splitting into the observed and the missing elements,

$$Q(\Theta|\hat{\Theta}^{(t)}) \propto -\sum_{i=1}^{n}\sum_{j,\,\Omega_{ij}=0}\mathbb{E}\left[\left(\frac{y_{ij}-\Theta_{ij}}{\sigma}\right)^2|\hat{\Theta_{ij}}^{(t)}\right] - \sum_{i=1}^{n}\sum_{j,\,\Omega_{ij}=1}\left(\frac{y_{ij}-\Theta_{ij}}{\sigma}\right)^2$$

Therefore,

$$Q(\Theta|\hat{\Theta}^{(t)}) \propto -\sum_{i=1}^{n}\sum_{j,\,\Omega_{ij}=0}\mathbb{E}[y_{ij}^2 - 2\Theta_{ij}y_{ij} + \Theta_{ij}^2|\hat{\Theta_{ij}}^{(t)}]^2 - \sum_{i=1}^{n}\sum_{j,\,\Omega_{ij}=1}\left(\frac{y_{ij}-\Theta_{ij}}{\sigma}\right)^2$$

$$Q(\Theta|\hat{\Theta}^{(t)}) \propto -\sum_{i=1}^{n}\sum_{j,\,\Omega_{ij}=0}\left(\sigma^2 + \hat{\Theta_{ij}}^{(t)2} - 2\hat{\Theta_{ij}}^{(t)}\Theta_{ij} + \Theta_{ij}^2\right) - \sum_{i=1}^{n}\sum_{j,\,\Omega_{ij}=1}\left(\frac{y_{ij}-\Theta_{ij}}{\sigma}\right)^2$$

which implies $Q(\Theta|\hat{\Theta}^{(t)}) \propto \|\Omega \odot Y + (1-\Omega)\odot\hat{\Theta}^{(t)} - \Theta\|_F^2$

The M-step is then written as follows:

$$\hat{\Theta}^{(t+1)} \in \text{argmin}_\Theta \|\Omega \odot Y + (1-\Omega)\odot\hat{\Theta}^{(t)} - \Theta\|_F^2 + \lambda\|\Theta\|_\star$$

The proximal gradient method is applied with

$$h_1(\Theta) = \lambda\|\Theta\|_\star \text{ and } h_2(\Theta) = \|\Omega \odot Y + (1-\Omega)\odot\hat{\Theta}^{(t)} - \Theta\|_F^2.$$

Therefore, the EM algorithm in the MAR case is the same one as `softImpute`.

B.3 The EM algorithm in the MNAR case

For the sake of clarity, we present below the EM algorithm in the MNAR and low dimension case.

Algorithm 10 The EM algorithm in the MNAR case

Step 0**:** $\hat{\Theta}^{(0)}$ and $\hat{\phi}^{(0)}$.
Step $t+1$**:**
for $(i,j) \in \Omega_{\text{mis}} = \{(l,k), l \in [1,n], k \in [1,p], \Omega_{lk} = 0\}$ **do**
  draw $z_{ij}^1, \ldots, z_{ij}^{N_s} \overset{\text{i.i.d.}}{\sim} \left[y_{ij}|\Omega_{ij}; \hat{\phi}^{(t)}, \hat{\Theta}_{ij}^{(t)}\right]$ with the SIR algorithm.
end for
Compute $\hat{\Theta}^{(t+1)}$ by using `softImpute` or the FISTA algorithm with the imputed matrix V (given by (B.1)).
Compute $\hat{\phi}^{(t+1)}$ by using the function glm with a binomial link, which perform a logistic regression of $J_{\cdot 1}^{(j)}$ on $J_{\cdot 2}^{(j)}$, with the matrix $J^{(j)}$ given above ((B.2)), for all j such that $\exists i \in \{1,\ldots n\}$, $\Omega_{ij} = 0$.

We already have given details for the stopping criterium.
We clarify the maximization step given by (2.8) and (2.9).

$$\begin{aligned}
\hat{\Theta} &\in \underset{\Theta}{\operatorname{argmin}} \sum_{i,j} \left(\frac{1}{N_s} \sum_{k=1}^{N_s} \frac{1}{2\sigma^2} (v_{ij}^{(k)} - \Theta_{ij})^2 \right) + \lambda \|\Theta\|_\star \\
&\in \underset{\Theta}{\operatorname{argmin}} \sum_{i,j} \left(\frac{1}{N_s \sigma^2} \sum_{k=1}^{N_s} -v_{ij}^{(k)} \Theta_{ij} + \frac{1}{2} \Theta_{ij}^2 \right) + \lambda \|\Theta\|_\star \\
&\in \underset{\Theta}{\operatorname{argmin}} \|V - \Theta\|_F^2 + \lambda \|\Theta\|_\star, \text{ where:}
\end{aligned}$$

$$V = \begin{pmatrix} \frac{1}{N_s} \sum_{k=1}^{Ns} v_{11}^{(k)} & \cdots & \frac{1}{N_s} \sum_{k=1}^{Ns} v_{1p}^{(k)} \\ \vdots & \ddots & \vdots \\ \frac{1}{N_s} \sum_{k=1}^{Ns} v_{n1}^{(k)} & \cdots & \frac{1}{N_s} \sum_{k=1}^{Ns} v_{np}^{(k)} \end{pmatrix} \tag{B.1}$$

$$\begin{aligned}
\hat{\phi}^{(t+1)} &\in \underset{\phi}{\operatorname{argmin}} \sum_{i,j} \frac{1}{N_s} \sum_{k=1}^{N_s} (1 - \Omega_{ij}) C_3 - \Omega_{ij} C_4 \\
&\in \underset{\phi}{\operatorname{argmin}} \sum_{i,j} \frac{1}{N_s} \sum_{k=1}^{N_s} C_3 + \Omega_{ij} \phi_{1j} (v_{ij}^k - \phi_{2j})
\end{aligned}$$

with:

$$\begin{aligned}
C_3 &= \log(1 + e^{-\phi_{1j}(v_{ij}^k - \phi_{2j})}) \\
C_4 &= \log(1 - (1 + e^{-\phi_{1j}(v_{ij}^k - \phi_{2j})})^{-1})
\end{aligned}$$

For all $j \in \{1, \ldots, p\}$ such that $\exists i \in \{1, \ldots, n\}$, $\Omega_{ij} = 0$, estimating the coefficients ϕ_{1j} and ϕ_{2j} remains to fit a generalized linear model with the binomial link function for the matrix $J^{(j)}$:

$$J^{(j)} = \begin{pmatrix} \Omega_{1j} & v_{1j}^{(1)} \\ \vdots & \vdots \\ \Omega_{nj} & v_{nj}^{(1)} \\ \vdots & \vdots \\ \Omega_{1j} & v_{1j}^{(N_s)} \\ \vdots & \vdots \\ \Omega_{nj} & v_{nj}^{(N_s)} \end{pmatrix} \tag{B.2}$$

B.3.1 SIR

In the Monte Carlo approximation, the distribution of interest is $\left[y_{ij}|\Omega_{ij};\hat{\phi}_j^{(t)},\hat{\Theta}_{ij}^{(t)}\right]$. By using the Bayes rules:

$$\begin{aligned} &p\left(y_{ij}|\Omega_{ij};\hat{\phi}_j^{(t)},\hat{\Theta}_{ij}^{(t)}\right) =: f(y_{ij}) \\ &\propto p\left(y_{ij};\hat{\Theta}_{ij}^{(t)}\right) p\left(\Omega_{ij}|y_{ij};\hat{\phi}_j^{(t)}\right) =: g(y_{ij}) \end{aligned}$$

Denoting the Gaussian density function of mean $\Theta_{ij}^{(t)}$ and variance σ^2 by $\varphi_{\Theta_{ij}^{(t)},\sigma^2}$, if $\sigma > (2\pi)^{-1/2}$, the following condition holds:

$$f(y_{ij}) = cg(y_{ij}) \leqslant \varphi_{\Theta_{ij}^{(t)},\sigma^2}(x).$$

For M large, the SIR algorithm to simulate

$$z \sim \left[y_{ij}|\Omega_{ij};\hat{\phi}_j^{(t)},\hat{\Theta}_{ij}^{(t)}\right]$$

is described as follows.

Algorithm 11 `SIR`

Draw: a sample $x_1,\ldots,x_M \sim \mathcal{N}(\Theta_{ij}^{(t)},\sigma^2)$.
Compute the weights:

$$\omega(x_m) = \frac{g(x_m)}{\varphi_{\Theta_{ij}^{(t)},\sigma^2}(x_m)},$$

for $m = 1,\ldots,M$.
Draw z from the original sample $x_1,\ldots,x_M$ with probability proportional to $\omega(x_1),\ldots,\omega(x_M)$.

B.4 Details on the variables in Traumabase®

A description of the variables which are used in Section 2.5 is given. The indications given in parentheses ph (pre-hospital) and h (hospital) mean that the measures have been taken before the arrival at the hospital and at the hospital.

- *SBP.ph*, *DBP.ph*, *HR.ph*: systolic and diastolic arterial pressure and heart rate during pre-hospital phase. (ph)
- *HemoCue.init*: prehospital capillary hemoglobin concentration. (ph)
- *SpO2.min*: peripheral oxygen saturation, measured by pulse oxymetry, to estimate oxygen content in the blood. (ph)

- *Cristalloid.volume*: total amount of prehospital administered cristalloid fluid resuscitation (volume expansion). (ph)
- *Shock.index.ph*: ratio of heart rate and systolic arterial pressure during pre-hospital phase. (ph)
- *Delta.shock.index*: Difference of shock index between arrival at the hospital and arrival on the scene. (h)
- *Delta.hemoCue*: Difference of hemoglobin level between arrival at the hospital and arrival on the scene. (h)

Appendix C

Appendix of Chapter 3

C.1 Proof of Proposition 9

Proposition 9. *Under Assumptions **A01.** and **A02.**, the parameters (α, Σ) of the PPCA model (3.1) and the mechanism parameters $\phi = (\phi_\ell)_{\ell \in \{1,\dots p\}}$ are identifiable. Assuming that the noise level σ^2 is known, the parameter B is identifiable up to a row permutation.*

For the sake of readability, we first present the proof of Proposition 9 in the case of the toy example presented in Section 3.3.1 with $p = 3$ and $r = 2$. The proof in the general setting follows.

C.1.1 Proof of Proposition 9 in the case of the toy example presented in Section 3.3.1

Consider the setting of the toy example presented in Section 3.3.1 with $p = 3$ and $r = 2$. The PPCA model in (3.1) reads

$$\begin{cases} Y &= \begin{pmatrix} Y_1 & Y_2 & Y_3 \end{pmatrix} = \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 \end{pmatrix} + \begin{pmatrix} W_1 & W_2 \end{pmatrix} B + \epsilon, \\ Y &\sim \mathcal{N}(\alpha, \Sigma),\ \Sigma = B^T B + \sigma^2 I. \end{cases}$$

Y_2 and Y_3 are assumed to be observed and Y_1 is self-masked MNAR, *i.e.*

$$\mathbb{P}(\Omega_1 = 1 | Y_1, Y_2, Y_3; \phi_1) = \mathbb{P}(\Omega_1 = 1 | Y_1; \phi_1) = F_1(\phi_1^0 + \phi_1^1 y_1), \tag{C.1}$$

where F_1 is strictly monotone with a positive finite support and where $\phi_1 = (\phi_1^0, \phi_1^1)$.

Proof. Assume that (Y, Ω) and (Y', Ω') have distributions respectively parameterized by (α, Σ, ϕ_1) and $(\alpha', \Sigma', \phi_1')$. Assume that Y and Y' have the same observed distribution, *i.e.*

$$\mathcal{L}(Y_1, \Omega_1 = 1; \alpha_1, \Sigma_{11}, \phi_1) = \mathcal{L}(Y_1', \Omega_1' = 1; \alpha_1', \Sigma_{11}', \phi_1') \tag{C.2}$$

$$\mathcal{L}(Y_1, Y_j, \Omega_1 = 1; \alpha_1, \alpha_j, \Sigma_{(1j)}, \phi_1) = \mathcal{L}(Y_1', Y_j', \Omega_1' = 1; \alpha_1', \alpha_j', \Sigma_{(1j)}', \phi_1') \qquad j \in \{2, 3\}, \tag{C.3}$$

where $\Sigma_{(1j)}$ is the covariance matrix $\begin{pmatrix}\Sigma_{11} & \Sigma_{1j}\\ \Sigma_{1j} & \Sigma_{jj}\end{pmatrix}$. In order to show that parameters identifiability holds, we need to show that (C.2) and (C.3) imply that $\alpha = \alpha'$, $\Sigma = \Sigma'$ and $\phi_1 = \phi_1'$. Then, under a known noise level σ^2, we prove that B and B' are equal up to a row permutation.

As (Y_2, Y_3) and (Y_2', Y_3') are fully observed, the parameters of the distributions $\mathcal{L}(Y_2)$, $\mathcal{L}(Y_2')$, $\mathcal{L}(Y_3)$, $\mathcal{L}(Y_3')$, $\mathcal{L}(Y_2, Y_3)$ and $\mathcal{L}(Y_2', Y_3')$ are identifiable. It trivially implies that $\alpha_2 = \alpha_2'$, $\Sigma_{22} = \Sigma_{22}'$, $\alpha_3 = \alpha_3'$, $\Sigma_{33} = \Sigma_{33}'$ and $\Sigma_{23} = \Sigma_{23}'$.

Identifiability of the MNAR variable variance. Equation (C.2) can be rewritten in terms of density function as follows

$$f_{Y_1,\Omega_1=1}(y_1;\alpha_1,\Sigma_{11},\phi_1) = f_{Y_1',\Omega_1'=1}(y_1;\alpha_1',\Sigma_{11}',\phi_1') \qquad \forall y_1 \in \mathbb{R}.$$

Given the missing mechanism in (C.1) and that $Y_{\cdot 1} \sim \mathcal{N}(\alpha_1, \Sigma_{11})$, (Miao et al., 2016, Theorem 1 a)) ensures that $\Sigma_{11} = \Sigma_{11}'$.

Identifiability of the Mean and the MNAR mechanism parameter. Using (C.2) and (C.3), the previous computations entail that

$$\mathcal{L}(Y_2|Y_1,\Omega_1=1;\alpha_1,\alpha_2,\Sigma_{(12)},\phi_1) = \mathcal{L}(Y_2'|Y_1',\Omega_1'=1;\alpha_1',\alpha_2',\Sigma_{(12)}',\phi_1'),$$

noting that

$$f_{Y_2|Y_1=y_1,\Omega_1=1}(y_2;\alpha_1,\alpha_2,\Sigma_{(12)},\phi_1) = \frac{f_{Y_1,Y_2,\Omega_1=1}(y_1,y_2;\alpha_1,\alpha_2,\Sigma_{(12)},\phi_1)}{f_{Y_1,\Omega_1=1}(y_1;\alpha_1,\Sigma_{11},\phi_1)} \qquad \forall (y_1,y_2) \in \mathbb{R}^2$$

One obtains

$$\begin{aligned}&\frac{\mathbb{P}(\Omega_1=1|Y_1=y_1,Y_2=y_2;\phi_1) f_{Y_2|Y_1=y_1}(y_2;\alpha_1,\alpha_2,\Sigma_{(12)})}{\mathbb{P}(\Omega_1=1|Y_1=y_1;\phi_1)}\\ &\quad= \frac{\mathbb{P}(\Omega_{\cdot 1}'=1|Y_1'=y_1,Y_2'=y_2;\phi_1') f_{Y_2'|Y_1'=y_1}(y_2;\alpha_1',\alpha_2',\Sigma_{(12)}')}{\mathbb{P}(\Omega_1'=1|Y_1=y_1;\phi_1')} \qquad \forall (y_1,y_2) \in \mathbb{R}^2\end{aligned}$$

Yet,

$$\begin{aligned}\mathbb{P}(\Omega_1=1|Y_1=y_1,Y_2=y_2;\phi_1) &= \mathbb{E}[\mathbb{E}[1_{\Omega_1=1}|Y_1=y_1,Y_2=y_2,Y_3=y_3;\phi_1]|Y_1=y_1,Y_2=y_2]\\ &= \mathbb{E}[\mathbb{P}(\Omega_1=1|Y=y;\phi_1)|Y_1=y_1,Y_2=y_2]\\ &= \mathbb{E}[\mathbb{P}(\Omega_1=1|Y_1=y_1;\phi_1)|Y_1=y_1,Y_2=y_2]\\ &= \mathbb{P}(\Omega_1=1|Y=y_1;\phi_1) \qquad \text{(C.4)}\end{aligned}$$

by measurability. It implies for all $y_1 \in \mathbb{R}$ and $y_2 \in \mathbb{R}$

$$f_{Y_2|Y_1=y_1}(y_2;\alpha_1,\alpha_2,\Sigma_{(12)}) = f_{Y_2'|Y_1'=y_1}(y_2;\alpha_1',\alpha_2',\Sigma_{(12)}')$$

which leads to the equality of the conditional expectations and variances associated to the above densities:

$$\alpha_2 + \Sigma_{12}\Sigma_{11}^{-1}(\alpha_1 - y_1) = \alpha_2 + \Sigma'_{12}\Sigma_{11}^{-1}(\alpha'_1 - y_1) \qquad \forall y_1 \in \mathbb{R}$$
$$\Sigma_{22} - \Sigma_{12}^2\Sigma_{11}^{-1} = \Sigma_{22} - (\Sigma'_{12})^2\Sigma_{11}^{-1}.$$

It implies that

$$\Sigma_{12}^2 = (\Sigma'_{12})^2 \Longrightarrow |\Sigma_{12}| = |\Sigma'_{12}| \tag{C.5}$$
$$\frac{\Sigma_{21}}{\Sigma'_{21}} = \frac{(\alpha'_1 - y_1)}{(\alpha_1 - y_1)} \Longrightarrow |\alpha_1 - y_1| = |\alpha'_1 - y_1| \qquad \forall y_1 \in \mathbb{R} \tag{C.6}$$

Equation (C.6) implies that $\alpha_1 = \alpha'_1$, since for $y_1 = \alpha'_1$, one has $\alpha_1 - \alpha'_1 = 0$.
Using (C.3), one has

$$\begin{aligned}&\mathbb{P}(\Omega_1 = 1|Y_1 = y_1, Y_2 = y_2; \phi_1) f_{(Y_1,Y_2)}(y_1, y_2; \alpha_1, \alpha_2, \Sigma_{(12)}) \\ &\quad = \mathbb{P}(\Omega'_1 = 1|Y'_1 = y_1, Y'_2 = y_2; \phi'_1) f_{(Y'_1,Y'_2)}(y_1, y_2; \alpha'_1, \alpha'_2, \Sigma'_{(12)}) \qquad \forall (y_1, y_2) \in \mathbb{R}^2\end{aligned} \tag{C.7}$$

Using (C.4),

$$\frac{\exp\left(-\frac{1}{2}\begin{pmatrix} y_1 - \alpha_1 & y_2 - \alpha_2 \end{pmatrix} \Sigma_{(12)}^{-1} \begin{pmatrix} y_1 - \alpha_1 \\ y_2 - \alpha_2 \end{pmatrix}\right)}{\exp\left(-\frac{1}{2}\begin{pmatrix} y_1 - \alpha_1 & y_2 - \alpha_2 \end{pmatrix} (\Sigma'_{(12)})^{-1} \begin{pmatrix} y_1 - \alpha_1 \\ y_2 - \alpha_2 \end{pmatrix}\right)} \frac{\mathbb{P}(\Omega_1 = 1|Y_1 = y_1; \phi_1)}{\mathbb{P}(\Omega'_1 = 1|Y'_1 = y_1; \phi'_1)} = \frac{\sqrt{\det(\Sigma_{(12)})}}{\sqrt{\det(\Sigma'_{(12)})}},$$

where $\det(\Sigma_{(12)})$ denotes the determinant of the matrix $\Sigma_{(12)}$.

With (C.5), one has $\Sigma_{11}\Sigma_{22} - \Sigma_{12}^2 = \Sigma_{11}\Sigma_{22} - (\Sigma'_{12})^2$ and $\frac{\sqrt{\det(\Sigma_{(12)})}}{\sqrt{\det(\Sigma'_{(12)})}} = 1$.

It leads to $\forall (y_1, y_2) \in \mathbb{R}^2$,

$$K \cdot \frac{\mathbb{P}(\Omega_1 = 1|Y_1 = y_1; \phi_1)}{\mathbb{P}(\Omega'_1 = 1|Y'_1 = y_1; \phi'_1)} = 1,$$

with

$$K := \frac{\exp\left(-\frac{1}{2\det(\Sigma_{(12)})}\left((y_1 - \alpha_1)^2\Sigma_{11} + (y_2 - \alpha_2)^2\Sigma_{22} - 2(y_1 - \alpha_1)(y_2 - \alpha_2)\Sigma_{12}\right)\right)}{\exp\left(-\frac{1}{2\det(\Sigma_{(12)})}\left((y_1 - \alpha_1)^2\Sigma_{11} + (y_2 - \alpha_2)^2\Sigma_{22} - 2(y_1 - \alpha'_1)(y_2 - \alpha_2)\Sigma'_{12}\right)\right)}.$$

The quantity K is equal to one, because

$$(y_2 - \alpha_2)\big((y_1 - \alpha_1)\Sigma_{12} - (y_1 - \alpha'_1)\Sigma'_{12}\big) = 0$$

using (C.6). Thus,

$$\frac{\mathbb{P}(\Omega_1 = 1|Y_1 = y_1; \phi_1)}{\mathbb{P}(\Omega'_1 = 1|Y'_1 = y_1; \phi'_1)} = 1 \iff F_1(\phi_1^0 + \phi_1^1 y_1) = F_1((\phi')_1^0 + (\phi')_1^1 y_1) \qquad \forall y_1 \in \mathbb{R}$$

As F_1 is strictly monotone, it is an injective function. Thus,

$$\phi_1^0 + \phi_1^1 y_1 = (\phi')_1^0 + (\phi')_1^1 y_1 \qquad \forall y_1 \in \mathbb{R} \qquad \iff (\phi_1^0 - (\phi')_1^0) + ((\phi')_1^1 - \phi_1^1) y_1 = 0 \qquad \forall y_1 \in \mathbb{R}$$

It implies $\phi_1 = \phi'_1$.

Identifiability of the Covariances of the MNAR variable. Equation (C.7) thus leads to

$$f_{(Y_1,Y_2)}(y_1, y_2; \alpha_1, \alpha_2, \Sigma_{(12)}) = f_{(Y'_1,Y'_2)}(y_1, y_2; \alpha'_1, \alpha'_2, \Sigma'_{(12)}) \qquad \forall (y_1, y_2) \in \mathbb{R}^2$$

One can conclude that $\Sigma_{12} = \Sigma'_{12}$. The same reasoning may be done for the covariance between Y_1 and Y_3.

Identifiability of the loading matrix. One wants to prove $B = B'$ up to row permutation. One has

$$\begin{aligned} \Sigma = \Sigma' &\Leftrightarrow \Sigma - \sigma^2 I_{p\times p} = \Sigma' - \sigma^2 I_{p\times p} \\ &\Leftrightarrow B^T B = (B')^T B' \end{aligned} \tag{C.8}$$

As $B^T B$ is a positive symetric matrix of rank 2, one has the following singular value decomposition,

$$B^T B = (B')^T B' = U D U^T,$$

where $U = (u_1|u_2|u_3) \in \mathbb{R}^{3\times 3}$ the orthogonal matrix of singular vector and

$$D = \begin{pmatrix} \sqrt{d_1} & 0 & 0 \\ 0 & \sqrt{d_2} & 0 \\ 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^{3\times 3}$$

with $d_1 \geqslant d_2 \geqslant 0$. One can choose

$$B = \begin{pmatrix} \sqrt{d_1} u_1^T \\ \hline \sqrt{d_2} u_2^T \end{pmatrix}$$

noting that a row permutation of B would not change the product $B^T B$. Therefore, $B = B'$ up to a row permutation.

□

C.1.2 Proof of Proposition 9 in the general case

We present the proof of Proposition 9 in the general case where d variables are self-masked MNAR and $p-d$ variables are MCAR.

Proof. Assume that (Y,Ω) and (Y',Ω') have distributions respectively parameterized by (α,Σ,ϕ) and (α',Σ',ϕ'). Assume that Y and Y' have the same following observed distributions

$$\mathcal{L}(Y_j,\Omega_j=1;\alpha_j,\Sigma_{jj},\phi_j)=\mathcal{L}(Y'_j,\Omega'_j=1;\alpha'_j,\Sigma'_{jj},\phi'_j)\qquad\forall j\in\{1,\dots,p\},\tag{C.9}$$

$$\begin{aligned}&\mathcal{L}(Y_j,Y_k,\Omega_j=1,\Omega_k=1;\alpha_j,\alpha_k,\Sigma_{(jk)},\phi_j,\phi_k)\\&\quad=\mathcal{L}(Y'_j,Y'_k,\Omega'_j=1,\Omega'_k=1;\alpha'_j,\alpha'_k,\Sigma'_{(jk)},\phi'_j,\phi'_k)\qquad\forall j\neq k\in\{1,\dots,p\},\end{aligned}\tag{C.10}$$

where $\Sigma_{(jk)}$ denotes the covariance matrix $\begin{pmatrix}\Sigma_{jj}&\Sigma_{jk}\\\Sigma_{jk}&\Sigma_{kk}\end{pmatrix}$.

In order to show that parameters identifiability holds, we need to show that (C.9) and (C.10) implies that $\alpha=\alpha'$, $\Sigma=\Sigma'$ and $\phi=\phi'$. Then, under a known noise level σ^2, we will prove that B and B' are equal up to row permutations.

In what follows, $f_{Y_{\cdot j}}$ or $f_{(Y_{\cdot j},Y_{\cdot k})}$ respectively denote the density function of $Y_{\cdot j}$, and of $(Y_{\cdot j},Y_{\cdot k})$.

In the following, we will use the following tip, for any $l\in\{1,\dots,p\}$ and $\mathcal{K}\subset\{1,\dots,p\}\backslash\{l\}$ such that $0\leqslant|\mathcal{K}|\leqslant p-1$,

$$\begin{aligned}\mathbb{P}(\Omega_l=1|Y_l=y_l,Y_\mathcal{K}=y_\mathcal{K};\phi_l)&=\mathbb{E}[\mathbb{E}[\mathbf{1}_{\Omega_l=1}|Y;\phi_l]|Y_l=y_l,Y_\mathcal{K}=y_\mathcal{K}]\\&=\mathbb{E}[\mathbb{P}(\Omega_l=1|Y=y;\phi_l)|Y_l=y_l,Y_\mathcal{K}=y_\mathcal{K}]\end{aligned}$$

Thus, using the mechanisms in **A01.**,

$$\begin{aligned}&\mathbb{P}(\Omega_l=1|Y_l=y_l,Y_\mathcal{K}=y_\mathcal{K};\phi_l)\\&\quad=\begin{cases}\mathbb{E}[\mathbb{P}(\Omega_l=1|Y_l=y_l;\phi_l)|Y_l=y_l,Y_\mathcal{K}=y_\mathcal{K}] & \text{if } Y_l \text{ is self-masked MNAR}\\ \mathbb{E}[\mathbb{P}(\Omega_l=1;\phi_l)|Y_l=y_l,Y_\mathcal{K}=y_\mathcal{K}] & \text{if } Y_l \text{ is MCAR}\end{cases}\end{aligned}$$

Thus,

$$\mathbb{P}(\Omega_l=1|Y_l=y_l,Y_\mathcal{K}=y_\mathcal{K};\phi_l)=\begin{cases}\mathbb{P}(\Omega_l=1|Y_l=y_l;\phi_l) & \text{if } Y_l \text{ is self-masked MNAR}\\ \mathbb{P}(\Omega_l=1;\phi_l) & \text{if } Y_l \text{ is MCAR}\end{cases}\tag{C.11}$$

(C.12)

by measurability if Y_l is self-masked MNAR and by independence if Y_l is MCAR.

Identifiability of the parameters for the not-MNAR variables $(Y_j)_{j\in\overline{\mathcal{M}}}$.

Mechanism parameter, Mean and Variance of $Y_j, j \in \overline{\mathcal{M}}$. Equation (C.9) leads to

$$\mathbb{P}(\Omega_j = 1|Y_j = y_j; \phi_j) f_{Y_j}(y_j; \alpha_j, \Sigma_{jj}) = \mathbb{P}(\Omega'_j = 1|Y'_j = y_j; \phi'_j) f_{Y'_j}(y_j; \alpha'_j, \Sigma'_{jj}) \qquad \forall y_j \in \mathbb{R}.$$

Using (C.12), $P(\Omega_j = 1) = \mathbb{P}(\Omega_j = 1|Y_j = y_j; \phi_j) = F_j(\phi_j)$. This distribution is identifiable since it pertains to a conditional distribution of the observed data. As F_j is strictly monotone, it implies that

$$F_j(\phi_j) = F_j(\phi'_j) \iff \phi_j = \phi'_j.$$

As $\phi_j = \phi'_j$, one obtains

$$f_{Y_j}(y_j; \alpha_j, \Sigma_{jj}) = f_{Y'_j}(y_j; \alpha'_j, \Sigma'_{jj}) \qquad \forall y_j \in \mathbb{R}$$

which directly implies that $\alpha_j = \alpha'_j$ and $\Sigma_{jj} = \Sigma'_{jj}$, since Y_j and Y'_j are Gaussian variables.

Covariance between two not MNAR variables Y_j and Y_k, $j \neq k \in \overline{\mathcal{M}}$. Equation (C.10) gives that for all $(y_j, y_k) \in \mathbb{R}^2$

$$\begin{aligned}\mathbb{P}(\Omega_j = 1, \Omega_k = 1|Y_j = y_j, Y_k = y_k; \phi_j, \phi_k) f_{(Y_j,Y_k)}(y_j, y_k; \alpha_j, \alpha_k, \Sigma_{(j,k)}) \\ = \mathbb{P}(\Omega'_j = 1, \Omega'_k = 1|Y'_j = y_j, Y'_k = y_k; \phi'_j, \phi'_k) f_{(Y'_j,Y'_k)}(y_j, y_k; \alpha'_j, \alpha'_k, \Sigma'_{(j,k)}), \quad \text{(C.13)}\end{aligned}$$

and one has as well that

$$\mathbb{P}(\Omega_j = 1, \Omega_k = 1|Y_j = y_j, Y_k = y_k; \phi_j, \phi_k) = \mathbb{P}(\Omega_j = 1|Y_j = y_j; \phi_j)\mathbb{P}(\Omega_k = 1|Y_k = y_k; \phi_k),$$

using **A02.**. Likewise,

$$\mathbb{P}(\Omega'_j = 1, \Omega'_k = 1|Y'_j = y_j, Y'_k = y_k; \phi'_j, \phi'_k) = \mathbb{P}(\Omega'_j = 1|Y'_j = y_j; \phi'_j)\mathbb{P}(\Omega'_k = 1|Y'_k = y_k; \phi'_k).$$

Given that $\phi_j = \phi'_j$ and $\phi_k = \phi'_k$, one obtains

$$\mathbb{P}(\Omega_j = 1, \Omega_k = 1|Y_j = y_j, Y_k = y_k; \phi_j, \phi_k) = \mathbb{P}(\Omega'_j = 1, \Omega'_k = 1|Y'_j = y_j, Y'_k = y_k; \phi_j, \phi_k).$$

Thus, Equation (C.13) leads to, for all $(y_j, y_k) \in \mathbb{R}^2$,

$$f_{(Y_j,Y_k)}(y_j, y_k; \alpha_j, \alpha_k, \Sigma_{(j,k)}) = f_{(Y'_j,Y'_k)}(y_j, y_k; \alpha'_j, \alpha'_k, \Sigma'_{(j,k)}),$$

and $\Sigma_{jk} = \Sigma'_{jk}$.

Identifiability of the parameters for the MNAR variables.

Variance of $Y_m, m \in \mathcal{M}$. Equation (C.9) gives that

$$f_{(Y_m,\Omega_m=1)}(y_m; \alpha_m, \Sigma_{mm}, \phi_m) = f_{(Y'_m,\Omega'_m=1)}(y_m; \alpha'_m, \Sigma'_{mm}, \phi'_m) \qquad \forall y_m \in \mathbb{R}.$$

Given the self-masked missing mechanism in **A01.** and that $Y_{.m} \sim \mathcal{N}(\alpha_m, \Sigma_{mm})$, (Miao et al., 2016, Theorem 1 a)) ensures that $\Sigma_{mm} = \Sigma'_{mm}$.

Mean and mechanism parameter of $Y_m, m \in \mathcal{M}$. Let $j \in \overline{\mathcal{M}}$ (a not MNAR variable). One has

$$\mathcal{L}(Y_j, \Omega_j = 1 | Y_m, \Omega_m = 1; \alpha_j, \alpha_m, \Sigma_{(jm)}, \phi_j, \phi_m) = \mathcal{L}(Y'_j, \Omega'_j = 1 | Y'_m, \Omega'_m = 1; \alpha'_j, \alpha'_m, \Sigma'_{(jm)}, \phi'_j, \phi'_m) \quad \text{(C.14)}$$

using (C.9) and (C.10) and noting that

$$f_{(Y_j,\Omega_j=1)|Y_m=y_m,\Omega_m=1}(y_j; \alpha_j, \alpha_m, \Sigma_{(jm)}, \phi_j, \phi_m) = \frac{f_{(Y_j,\Omega_j=1,Y_m,\Omega_m=1)}(y_j, y_m; \alpha_j, \alpha_m, \Sigma_{(jm)}, \phi_j, \phi_m)}{f_{(Y_m,\Omega_m=1)}(y_m; \alpha_m, \Sigma_{mm}, \phi_m)} \qquad \forall (y_j, y_m) \in \mathbb{R}^2.$$

Equation (C.14) implies that $\forall (y_j, y_m) \in \mathbb{R}^2$,

$$\mathbb{P}(\Omega_j = 1 | Y_j = y_j, Y_m = y_m, \Omega_m = 1; \phi_j) \frac{\mathbb{P}(\Omega_m = 1 | Y_j = y_j, Y_m = y_m; \phi_m) f_{Y_j|Y_m=y_m}(y_j; \alpha_j, \alpha_m, \Sigma_{(jm)})}{\mathbb{P}(\Omega_m = 1 | Y_m = y_m; \phi_m)}$$
$$= \mathbb{P}(\Omega'_j = 1 | Y'_j = y_j, Y'_m = y_m, \Omega'_m = 1; \phi'_j) \frac{\mathbb{P}(\Omega'_m = 1 | Y'_j = y_j, Y'_m = y_m; \phi'_m) f_{Y'_j|Y'_m=y_m}(y_j; \alpha'_j, \alpha'_m, \Sigma'_{(jm)})}{\mathbb{P}(\Omega'_m = 1 | Y'_m = y_m; \phi'_m)} \quad \text{(C.15)}$$

One can note that

$$\mathbb{P}(\Omega_j = 1 | Y_j = y_j, Y_m = y_m, \Omega_m = 1; \phi_j) = \mathbb{P}(\Omega_j = 1 | Y_j = y_j; \phi_j).$$

Indeed,

$$\begin{aligned} \mathbb{P}(\Omega_j = 1 | Y_j = y_j, Y_m = y_m, \Omega_m = 1; \phi_j) &= \frac{\mathbb{P}(\Omega_j = 1 \cap \Omega_m = 1 | Y_j = y_j, Y_m = y_m; \phi_j, \phi_m)}{\mathbb{P}(\Omega_m = 1 | Y_j = y_j, Y_m = y_m; \phi_m)} \\ &= \frac{\mathbb{P}(\Omega_j = 1 | Y_j = y_j; \phi_j) \mathbb{P}(\Omega_m = 1 | Y_m = y_m; \phi_m)}{\mathbb{P}(\Omega_m = 1 | Y_j = y_j, Y_m = y_m; \phi_m)} \\ &= \mathbb{P}(\Omega_j = 1 | Y_j = y_j; \phi_j), \end{aligned}$$

using **A02.** in the second step. Likewise,

$$\mathbb{P}(\Omega'_j = 1 | Y'_j = y_j, Y'_m = y_m, \Omega'_m = 1; \phi'_j) = \mathbb{P}(\Omega'_j = 1 | Y'_j = y_j; \phi'_j).$$

Given that $\phi_j = \phi'_j$,

$$\mathbb{P}(\Omega_j = 1 | Y_j = y_j, Y_m = y_m, \Omega_m = 1; \phi_j) = \mathbb{P}(\Omega'_j = 1 | Y'_j = y_j, Y'_m = y_m, \Omega'_m = 1; \phi'_j)$$

Thus, Equation (C.15) leads to

$$\frac{\mathbb{P}(\Omega_m = 1 | Y_j = y_j, Y_m = y_m; \phi_m) f_{Y_j|Y_m=y_m}(y_j; \alpha_j, \alpha_m, \Sigma_{(jm)})}{\mathbb{P}(\Omega_m = 1 | Y_m = y_m; \phi_m)} = \frac{\mathbb{P}(\Omega'_m = 1 | Y'_j = y_j, Y'_m = y_m; \phi'_m) f_{Y'_j|Y'_m=y_m}(y_j; \alpha'_j, \alpha'_m, \Sigma'_{(jm)})}{\mathbb{P}(\Omega'_m = 1 | Y'_m = y_m; \phi'_m)} \qquad \forall (y_j, y_m) \in \mathbb{R}^2.$$

As $\mathbb{P}(\Omega_m = 1|Y_j = y_j, Y_m = y_m; \phi_m) = \mathbb{P}(\Omega_m = 1|Y_m = y_m; \phi_m)$ by using (C.11), one obtains

$$f_{Y_j|Y_m=y_m}(y_j; \alpha_j, \alpha_m, \Sigma_{(jm)}) = f_{Y'_j|Y'_m=y_m}(y_j; \alpha'_j, \alpha'_m, \Sigma'_{(jm)}) \qquad \forall (y_j, y_m) \in \mathbb{R}^2,$$

which leads to the equality of the conditional expectation and variance, as follows:

$$\alpha_j + \Sigma_{mj}\Sigma_{mm}^{-1}(\alpha_m - y_m) = \alpha'_j + \Sigma'_{mj}(\Sigma'_{mm})^{-1}(\alpha'_m - y_m) \qquad \forall (y_j, y_m) \in \mathbb{R}^2$$
$$\Sigma_{jj} - \Sigma_{mj}^2\Sigma_{mm}^{-1} = \Sigma'_{jj} - (\Sigma'_{mj})^2(\Sigma'_{mm})^{-1}$$

As $\alpha_j = \alpha'_j$ and $\Sigma_{mm} = \Sigma'_{mm}$,

$$\Sigma_{mj}^2 = (\Sigma'_{mj})^2 \Longrightarrow |\Sigma_{mj}| = |\Sigma'_{mj}| \tag{C.16}$$
$$\frac{\Sigma_{mj}}{\Sigma'_{mj}} = \frac{(\alpha'_m - y_m)}{(\alpha_m - y_m)} \Longrightarrow |\alpha_m - y_m| = |\alpha'_m - y_m| \qquad \forall y_m \in \mathbb{R} \tag{C.17}$$

Equation (C.17) implies that $\alpha_m = \alpha'_m$, since for $y_m = \alpha'_m$, one has $\alpha_m - \alpha'_m = 0$.
In addition, using (C.10), one has for all $(y_j, y_m) \in \mathbb{R}^2$,

$$\begin{aligned}&\mathbb{P}(\Omega_j = 1, \Omega_m = 1|Y_j = y_j, Y_m = y_m; \phi_j, \phi_m) f_{(Y_j,Y_m)}(y_j, y_m; \alpha_j, \alpha_m, \Sigma_{(jm)}) \\ &\quad = \mathbb{P}(\Omega'_j = 1, \Omega'_m = 1|Y'_j = y_j, Y'_m = y_m; \phi'_j, \phi'_m) f_{(Y'_j,Y'_m)}(y_j, y_m; \alpha'_j, \alpha'_m, \Sigma'_{(jm)})\end{aligned} \tag{C.18}$$

One can note that

$$\begin{aligned}&\mathbb{P}(\Omega_j = 1, \Omega_m = 1|Y_j = y_j, Y_m = y_m; \phi_j, \phi_m) \\ &\qquad = \mathbb{P}(\Omega_j = 1; \phi_j)\mathbb{P}(\Omega_m = 1|Y_m = y_m; \phi_m),\end{aligned}$$

using **A02.** and the tips given in (C.11) and (C.12). The same equation holds for $(Y'_j, Y'_m, \Omega'_j, \Omega'_m)$ with the parameters (ϕ'_j, ϕ'_m). Using $\phi_j = \phi'_j$, Equation (C.18) leads to

$$\begin{aligned}&\mathbb{P}(\Omega_m = 1|Y_m = y_m; \phi_m) f_{(Y_j,Y_m)}(y_j, y_m; \alpha_j, \alpha_m, \Sigma_{(jm)}) = \\ &\quad \mathbb{P}(\Omega'_m = 1|Y'_m = y_m; \phi'_m) f_{(Y'_j,Y'_m)}(y_j, y_m; \alpha'_j, \alpha'_m, \Sigma'_{(jm)}) \qquad \forall (y_j, y_m) \in \mathbb{R}^2.\end{aligned} \tag{C.19}$$

It implies that, $\forall (y_j, y_m) \in \mathbb{R}^2$,

$$\frac{\exp\left(-\frac{1}{2}\begin{pmatrix} y_j - \alpha_j & y_m - \alpha_m \end{pmatrix}\Sigma_{(jm)}^{-1}\begin{pmatrix} y_j - \alpha_j \\ y_m - \alpha_m \end{pmatrix}\right)}{\exp\left(-\frac{1}{2}\begin{pmatrix} y_j - \alpha'_j & y_m - \alpha'_m \end{pmatrix}(\Sigma'_{(jm)})^{-1}\begin{pmatrix} y_j - \alpha'_j \\ y_m - \alpha'_m \end{pmatrix}\right)} \frac{\mathbb{P}(\Omega_m = 1|Y_m = y_m; \phi_m)}{\mathbb{P}(\Omega'_m = 1|Y'_m = y_m; \phi'_m)} = \frac{\sqrt{\det(\Sigma_{(jm)})}}{\sqrt{\det(\Sigma'_{(jm)})}},$$

where $\det(\Sigma_{(jm)})$ denotes the determinant of the covariance matrix $\Sigma_{(jm)}$.

With $\Sigma_{jj} = \Sigma'_{jj}$, $\Sigma_{mm} = \Sigma'_{mm}$ and Equation (C.16), one has

$$\Sigma_{jj}\Sigma_{mm} - \Sigma_{mj}^2 = \Sigma_{jj}\Sigma_{mm} - (\Sigma'_{mj})^2 \qquad \Longrightarrow \qquad \frac{\sqrt{\det(\Sigma_{(jm)})}}{\sqrt{\det(\Sigma'_{(jm)})}} = 1.$$

Besides, using $\alpha_j = \alpha'_j$, $\Sigma_{jj} = \Sigma'_{jj}$ and $\Sigma_{mm} = \Sigma'_{mm}$, one obtains that for all $(y_j, y_m) \in \mathbb{R}^2$,

$$K \cdot \frac{\mathbb{P}(\Omega_m = 1 | Y_m = y_m; \phi_m)}{\mathbb{P}(\Omega'_m = 1 | Y'_m = y_m; \phi'_m)} = 1,$$

with

$$K := \frac{\exp\left(-\frac{1}{2\det(\Sigma_{(jm)})}\left((y_j-\alpha_j)^2\Sigma_{jj} + (y_m-\alpha_m)^2\Sigma_{mm} - 2(y_j-\alpha_j)(y_m-\alpha_m)\Sigma_{mj}\right)\right)}{\exp\left(-\frac{1}{2\det(\Sigma_{(jm)})}\left((y_j-\alpha_j)^2\Sigma_{jj} + (y_m-\alpha_m)^2\Sigma_{mm} - 2(y_j-\alpha_j)(y_m-\alpha'_m)\Sigma'_{mj}\right)\right)}.$$

The quantity K is equal to one, because

$$(y_j - \alpha_j)((y_m - \alpha_m)\Sigma_{mj} - (y_m - \alpha'_m)\Sigma'_{mj}) = 0$$

using (C.17). Thus, for all $y_m \in \mathbb{R}$,

$$\frac{\mathbb{P}(\Omega_m = 1 | Y_m = y_m; \phi_m)}{\mathbb{P}(\Omega'_m = 1 | Y'_m = y_m; \phi'_m)} = 1 \quad \Longleftrightarrow \quad F_m(\phi_m^0 + \phi_m^1 y_m) = F_m((\phi')_m^0 + (\phi')_m^1 y_m).$$

As F is strictly monotone, it is an injective function. Thus,

$$\phi_m^0 + \phi_m^1 y_m = (\phi')_m^0 + (\phi')_m^1 y_m \Leftrightarrow ((\phi')_m^0 - \phi_m^0) + ((\phi')_m^1 - \phi_m^1) y_m = 0 \qquad \forall y_1 \in \mathbb{R}$$

It implies that $\phi_m = \phi'_m$.

Covariance between Y_j and Y_m with $j \in \overline{\mathcal{M}}, m \in \mathcal{M}$. Using (C.19) and $\phi_m = \phi'_m$, one has

$$f_{(Y_j,Y_m)}(y_j, y_m; \alpha_j, \alpha_m, \Sigma_{(jm)}) = f_{(Y'_j,Y'_m)}(y_j, y_m; \alpha'_j, \alpha'_m, \Sigma'_{(jm)}) \qquad \forall (y_j, y_m) \in \mathbb{R}^2$$

One can conclude that $\Sigma_{mj} = \Sigma'_{mj}$.

Covariance between Y_ℓ and Y_m with $\ell \neq m \in \mathcal{M}$. Using (C.10), one has for all $(y_\ell, y_m) \in \mathbb{R}^2$,

$$\begin{aligned}\mathbb{P}(\Omega_\ell = 1, \Omega_m = 1 | Y_j = y_j, Y_m = y_m; \phi_\ell, \phi_m) f_{(Y_\ell,Y_m)}(y_\ell, y_m; \alpha_\ell, \alpha_m, \Sigma_{(\ell m)}) \\ = \mathbb{P}(\Omega'_\ell = 1, \Omega'_m = 1 | Y'_\ell = y_\ell, Y'_m = y_m; \phi'_\ell, \phi'_m) f_{(Y'_\ell,Y'_m)}(y_\ell, y_m; \alpha'_\ell, \alpha'_m, \Sigma'_{(\ell m)}) \quad \text{(C.20)}\end{aligned}$$

One can note that

$$\begin{aligned}\mathbb{P}(\Omega_\ell = 1, \Omega_m = 1 | Y_\ell = y_\ell, Y_m = y_m; \phi_\ell, \phi_m) \\ = \mathbb{P}(\Omega_\ell = 1 | Y_\ell = y_\ell; \phi_\ell) \mathbb{P}(\Omega_m = 1 | Y_m = y_m; \phi_m),\end{aligned}$$

using **A02.** and the tip given in (C.11). The same equation holds for $(Y'_\ell, Y'_m, \Omega'_\ell, \Omega'_m)$ with the parameters (ϕ'_ℓ, ϕ'_m). Yet $\phi_\ell = \phi'_\ell$ and $\phi_m = \phi'_m$, which gives, for all $(y_j, y_m) \in \mathbb{R}^2$,

$$\mathbb{P}(\Omega_\ell = 1, \Omega_m = 1 | Y_\ell = y_\ell, Y_m = y_m; \phi_\ell, \phi_m) = \mathbb{P}(\Omega'_\ell = 1, \Omega'_m = 1 | Y'_\ell = y_\ell, Y'_m = y_m; \phi_\ell, \phi'_m).$$

Equation (C.20) leads to

$$f_{(Y_\ell,Y_m)}(y_\ell, y_m; \alpha_\ell, \alpha_m, \Sigma_{(\ell m)}) = f_{(Y'_\ell,Y'_m)}(y_\ell, y_m; \alpha'_\ell, \alpha'_m, \Sigma'_{(\ell m)}) \qquad \forall (y_\ell, y_m) \in \mathbb{R}^2,$$

which implies that $\Sigma_{\ell m} = \Sigma'_{\ell m}$.

Identifiability of the loading matrix. One wants to prove that $B = B'$ up to a row permutation. One has

$$\begin{aligned}\Sigma = \Sigma' &\iff \Sigma - \sigma^2 I_{p\times p} = \Sigma' - \sigma^2 I_{p\times p} \\ &\iff B^T B = (B')^T B' \end{aligned} \tag{C.21}$$

As $B^T B$ is a positive symetric matrix of rank r, its singular value decomposition reads

$$B^T B = (B')^T B' = U D U^T,$$

where $U = (u_1|\dots|u_p) \in \mathbb{R}^{p\times p}$ is an orthogonal matrix containing the singular vectors and

$$D = \begin{pmatrix} \sqrt{d_1} & & & & \\ & \ddots & & 0 & \\ & & \sqrt{d_r} & & \\ & 0 & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix} \in \mathbb{R}^{p\times p}$$

with $d_1 \geqslant \dots \geqslant d_r \geqslant 0$. One can choose

$$B = \begin{pmatrix} \sqrt{d_1} u_1^T \\ \hline \vdots \\ \hline \sqrt{d_r} u_r^T \end{pmatrix}$$

A row permutation of B does not change the product $B^T B$. Therefore, $B = B'$ up to a row permutation.

□

C.2 Proof for Section 3.3

C.2.1 Proof of Lemma 1

Lemma 1. *Under the PPCA model (3.1) and Assumption **A1.**, choose $j \in \mathcal{J}$. Denote $B^{-1} \in \mathbb{R}^{r\times r}$ the inverse of $\begin{pmatrix} B_{\cdot m} & (B_{\cdot j'})_{j'\in\mathcal{J}_{-j}} \end{pmatrix}$. One has*

$$Y_{\cdot j} = \mathcal{B}_{j\to m,\mathcal{J}_{-j}[0]} + \sum_{j'\in\mathcal{J}_{-j}} \mathcal{B}_{j\to m,\mathcal{J}_{-j}[j']} Y_{\cdot j'} + \mathcal{B}_{j\to m,\mathcal{J}_{-j}[m]} Y_{\cdot m} + \zeta$$

with:

$$\begin{aligned}
\mathcal{B}_{j\to m,\mathcal{J}_{-j}[j']} &:= \sum_{k\in\{m\}\cup\mathcal{J}_{-j}} B^{-1}_{kj'}B_{jk}, \forall j' \in \mathcal{J}_{-j} \\
\mathcal{B}_{j\to m,\mathcal{J}_{-j}[m]} &:= \sum_{k\in\{m\}\cup\mathcal{J}_{-j}} B^{-1}_{km}B_{jk}, \\
\mathcal{B}_{j\to m,\mathcal{J}_{-j}[0]} &:= \mathbf{1}\alpha_j - \sum_{j'\in\mathcal{J}_{-j}} \mathcal{B}_{j\to m,\mathcal{J}_{-j}[j']}\mathbf{1}\alpha_{j'} - \mathcal{B}_{j\to m,\mathcal{J}_{-j}[m]}\mathbf{1}\alpha_m \\
\zeta &= -\sum_{j'\in\mathcal{J}_{-j}} \mathcal{B}_{j\to m,\mathcal{J}_{-j}[j']}\epsilon_{\cdot j'} - \mathcal{B}_{j\to m,\mathcal{J}_{-j}[m]}\epsilon_{\cdot m} + \epsilon_{\cdot j}
\end{aligned}$$

Proof. Starting from the PPCA model written in (3.1) and recalled here

$$Y = \mathbf{1}\alpha + WB + \epsilon$$

and the matrix $B \in \mathbb{R}^{r\times p}$ being of full rank r, solving this linear system is the same as solving the following reduced system

$$\begin{pmatrix} Y_{\cdot m} & (Y_{\cdot j'})_{j'\in\mathcal{J}_{-j}} \end{pmatrix} = \mathbf{1}\alpha_{|r} + \begin{pmatrix} W_{\cdot 1} & \dots & W_{\cdot r} \end{pmatrix} B_{|r} + \epsilon_{|r},$$

where $B_{|r} \in \mathbb{R}^{r\times r}$ denotes the reduced matrix $\begin{pmatrix} B_{\cdot m} & (B_{\cdot j'})_{j'\in\mathcal{J}_{-j}} \end{pmatrix}$ of B. Similarly, $\alpha_{|r} \in \mathbb{R}^r$ and $\epsilon_{|r} \in \mathbb{R}^{n\times r}$ denote the reduced matrices of α and ϵ. With a slight abuse of notation, B^{-1} denotes the inverse of the reduced matrix $\begin{pmatrix} B_{\cdot m} & (B_{\cdot j'})_{j'\in\mathcal{J}_{-j}} \end{pmatrix}$ which exists using **A1.**.

Then, one can derive that

$$\begin{pmatrix} W_{\cdot 1} & \dots & W_{\cdot r} \end{pmatrix} = \left(\begin{pmatrix} Y_{\cdot m} & (Y_{\cdot j'})_{j'\in\mathcal{J}_{-j}} \end{pmatrix} - \mathbf{1}\alpha_{|r} - \epsilon_{|r}\right) B^{-1}.$$

The expression of $Y_{\cdot j}$ as a function of the latent variables is

$$\begin{aligned}
Y_{\cdot j} &= \mathbf{1}\alpha_j + \begin{pmatrix} W_{\cdot 1} & \dots & W_{\cdot r} \end{pmatrix} B_{j\cdot} + \epsilon_{\cdot j} \\
&= \mathbf{1}\alpha_j + \left(\begin{pmatrix} Y_{\cdot m} & (Y_{\cdot j'})_{j'\in\mathcal{J}_{-j}} \end{pmatrix} - \mathbf{1}\alpha_{|r} - \epsilon_{|r}\right) B^{-1}B_{j\cdot} + \epsilon_{\cdot j},
\end{aligned}$$

so that

$$\begin{aligned}
Y_{\cdot j} = &\sum_{\ell\in\{m\}\cup\mathcal{J}_{-j}} \left(\sum_{k\in\{m\}\cup\mathcal{J}_{-j}} B^{-1}_{k\ell}B_{jk}\right) Y_{\cdot\ell} \\
&- \sum_{\ell\in\{m\}\cup\mathcal{J}_{-j}} \left(\sum_{k\in\{m\}\cup\mathcal{J}_{-j}} B^{-1}_{k\ell}B_{jk}\right)(\mathbf{1}\alpha_\ell + \epsilon_{\cdot\ell}) + \epsilon_{\cdot j} + \mathbf{1}\alpha_j.
\end{aligned}$$

which leads to the desired solution.

□

C.2.2 Proof of Proposition 11

Proposition 11 (Mean estimator)**.** *Consider the PPCA model (3.1). Under Assumptions **A1.** and **A2.**, an estimator of the mean of a MNAR variable $Y_{\cdot m}$, for $m \in \mathcal{M}$, can be constructed as follows: choose $j \in \mathcal{J}$, and compute*

$$\hat{\alpha}_m := \frac{\hat{\alpha}_j - \hat{\mathcal{B}}^c_{j\to m,\mathcal{J}_{-j}}[0] - \sum_{j'\in\mathcal{J}_{-j}} \hat{\mathcal{B}}^c_{j\to m,\mathcal{J}_{-j}}[j']\hat{\alpha}_{j'}}{\hat{\mathcal{B}}^c_{j\to m,\mathcal{J}_{-j}}[m]},$$

with the $(\hat{\mathcal{B}}_{j\to m,\mathcal{J}_{-j}}[k])$'s estimators of the coefficients given in Definition 10 and assuming that the coefficient $\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[m]$ estimated by $\hat{\mathcal{B}}^c_{j\to m,\mathcal{J}_{-j}}[m]$ is non zero.

*Under the additional Assumptions **A3.** and **A4.**, this estimator is consistent.*

Proof. The main goal is to obtain a formula for $\alpha_{\cdot m}$, *i.e.*

$$\alpha_m = \frac{\alpha_j - \mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[0] - \sum_{j'\in\mathcal{J}_{-j}} \mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[j']\alpha_{j'}}{\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[m]}, \tag{C.22}$$

from which an estimator can be deduced. The idea is to express α_j from α_m and $(\alpha_{j'})_{j'\in\mathcal{J}_{-j}}$. Note that $\mathbb{E}[Y_{\cdot j}] = \mathbb{E}[\mathbb{E}[Y_{\cdot j}|(Y_{\cdot k})_{k\in\overline{\{j\}}}]]$. Assumption **A2.** leads to

$$\mathbb{E}[Y_{\cdot j}|(Y_{\cdot k})_{k\in\overline{\{j\}}}] = \mathbb{E}[Y_{\cdot j}|(Y_{\cdot k})_{k\in\overline{\{j\}}}, \Omega_{\cdot m} = 1].$$

Then, by Definition 10 which gives $(Y_{\cdot j})_{|\Omega_{\cdot m}=1}$,

$$\begin{aligned}
&\mathbb{E}[Y_{\cdot j}|(Y_{\cdot k})_{k\in\overline{\{j\}}}, \Omega_{\cdot m} = 1] \\
&= \mathbb{E}\left[\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[0] + \sum_{k\in\{m\}\cup\mathcal{J}_{-j}} \mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[k]Y_{\cdot k} + \zeta^c \middle| (Y_{\cdot k})_{k\in\overline{\{j\}}}\right] \\
&= \mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[0] + \sum_{k\in\{m\}\cup\mathcal{J}_{-j}} \mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[k]Y_{\cdot k} + \mathbb{E}\left[\zeta^c \middle| (Y_{\cdot k})_{k\in\overline{\{j\}}}\right]
\end{aligned}$$

Thus, by taking the mean and given that $\mathbb{E}[\epsilon_{\cdot k}] = 0, \forall k \in \{m\}\cup\mathcal{J}_{-j}$, one has

$$\alpha_j = \mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[0] + \sum_{j'\in\mathcal{J}_{-j}} \mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[j']\alpha_{j'} + \mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[m]\alpha_m,$$

implying Equation (C.22), provided that $\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[m] \neq 0$.

From this formula for the mean α_m, one define its estimator $\hat{\alpha}_m$ as in (3.9). It is trivially consistent as the linear combination of consistent quantities under **A3.** and **A4.** □

C.2.3 Proof of Proposition 12

Proposition 12 (Variance and covariances estimators)**.** *Consider the PPCA model (3.1). Under Assumptions **A1.** and **A2.**, an estimator of the variance of a MNAR variable $Y_{\cdot m}$ for*

$m \in \mathcal{M}$ and its covariances with the pivot variables, can be constructed as follows: choose $j \in \mathcal{J}$ and compute

$$\left(\widehat{\mathrm{Var}}(Y_{.m}) \quad \widehat{\mathrm{Cov}}(Y_{.m}, (Y_{.k})_{k\in\mathcal{J}})\right)^T := (\widehat{M}_j)^{-1}\hat{P}_j,$$

assuming that σ^2 tends to zero and the inverse of the matrix M_j estimated by $(\widehat{M}_j)^{-1}$ exists, with

$$\widehat{M}_j = \begin{matrix} \\ \in \mathbb{R} \\ \in \mathbb{R}^r \end{matrix} \left\{ \begin{bmatrix} \overbrace{(\hat{\mathcal{B}}^c_{j\to m,\mathcal{J}_{-j}[m]})^2}^{\in\mathbb{R}} & \overbrace{0 \quad 2\hat{\mathcal{B}}^c_{j\to m,\mathcal{J}_{-j}[m]}\left(\hat{\mathcal{B}}^c_{j\to m,\mathcal{J}_{-j}[\mathcal{J}_{-j}]}\right)^T}^{\in\mathbb{R}^r} \\ -(\hat{\mathcal{B}}^c_{k\to m,\mathcal{J}_{-k}[m]})_{k\in\mathcal{J}} & (\widehat{M}^k)_{k\in\mathcal{J}} \end{bmatrix} \right.$$

Let us precise that $\widehat{M}_j \in \mathbb{R}^{(r+1)\times(r+1)}$. One has $(\hat{\mathcal{B}}^c_{k\to m,\mathcal{J}_{-k}[m]})_{k\in\mathcal{J}} = \begin{pmatrix} \hat{\mathcal{B}}^c_{j_1\to m,\mathcal{J}_{-j_1}[m]} \\ \vdots \\ \hat{\mathcal{B}}^c_{j_r\to m,\mathcal{J}_{-j_r}[m]} \end{pmatrix}$.

One details $\widehat{M}^k$ for $k = j_1$ and the same definition is valid for all $k \in \mathcal{J}$.

$$\widehat{M}^{j_1} = \begin{pmatrix} 1 & -\hat{\mathcal{B}}^c_{j_1\to m,\mathcal{J}_{-j_1}[j_2]} & \cdots & -\hat{\mathcal{B}}^c_{j_1\to m,\mathcal{J}_{-j_1}[j_r]} \end{pmatrix} \in \mathbb{R}^r$$

$$\hat{P}_j = \begin{bmatrix} \overbrace{(\widehat{\mathrm{Var}}(Y_{.j}) - Q^c - (\hat{\mathcal{B}}^c_{j\to m,\mathcal{J}_{-j}[\mathcal{J}_{-j}]})^T\widehat{\mathrm{Var}}(Y_{\mathcal{J}_{-j}})\hat{\mathcal{B}}^c_{j\to m,\mathcal{J}_{-j}[\mathcal{J}_{-j}]}}^{\in\mathbb{R}} \\ \left(((\hat{\mathcal{B}}^c_{k\to m,\mathcal{J}_{-k}})^T \begin{pmatrix} 1 & \hat{\alpha}_m & (\hat{\alpha}_\ell)_{\ell\in\mathcal{J}_{-k}} \end{pmatrix}^T - \hat{\alpha}_k)\hat{\alpha}_m\right)_{k\in\mathcal{J}} \end{bmatrix} \begin{matrix} \\ \} \in \mathbb{R} \\ \} \in \mathbb{R}^r \end{matrix}$$

$$\hat{Q}^c = \left(\widehat{\mathrm{Var}}(Y_{.j})\middle|\Omega_{.m} = 1\right) - \left(\widehat{\mathrm{Cov}}((Y_{.k})_{k\in\overline{\{j\}}}, Y_{.j})\widehat{\mathrm{Var}}((Y_{.k})_{k\in\overline{\{j\}}})^{-1}\widehat{\mathrm{Cov}}((Y_{.k})_{k\in\overline{\{j\}}}, Y_{.j})^T\middle|\Omega_{.m} = 1\right).$$

*Under the additional Assumptions **A3.** and **A4.**, the estimators for the variance of $Y_{.m}$ and its covariances with the pivot variables given in (3.11) are consistent.*

Proof. As for the mean, to derive some estimator of the variance and the covariances, we want to obtain a formula as

$$M_j \left(\mathrm{Var}(Y_{.m}) \quad \mathrm{Cov}(Y_{.m}, (Y_{.k})_{k\in\mathcal{J}})\right)^T = \left(P_j - \mathcal{O}(\sigma^2)\right), \tag{C.23}$$

with

$$M_j = \begin{matrix} \\ \in \mathbb{R} \\ \in \mathbb{R}^r \end{matrix} \left\{ \begin{bmatrix} \overbrace{(\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[m]})^2}^{\in\mathbb{R}} & \overbrace{0 \quad 2\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[m]}\left(\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[\mathcal{J}_{-j}]}\right)^T}^{\in\mathbb{R}^r} \\ -(\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[m]})_{k\in\mathcal{J}} & (M^k)_{k\in\mathcal{J}} \end{bmatrix} \right.$$

Let us precise that $M_j \in \mathbb{R}^{(r+1)\times(r+1)}$. One has $(\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[m]})_{k\in\mathcal{J}} = \begin{pmatrix} \mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[m]} \\ \vdots \\ \mathcal{B}^c_{j_r\to m,\mathcal{J}_{-j_r}[m]} \end{pmatrix}$.

One details M^k for $k = j_1$ and the same definition is valid for all $k \in \mathcal{J}$.

$$M^{j_1} = \begin{pmatrix} 1 & -\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[j_2]} & \dots & -\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[j_r]} \end{pmatrix} \in \mathbb{R}^r$$

$$P_j = \begin{bmatrix} \overbrace{(\mathrm{Var}(Y_{\cdot j}) - Q^c - (\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[\mathcal{J}_{-j}]})^T \mathrm{Var}(Y_{\mathcal{J}_{-j}})\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[\mathcal{J}_{-j}]}}^{\in\mathbb{R}} \\ \left(((\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}})^T \begin{pmatrix} 1 & \mathbb{E}[Y_{\cdot m}] & (\mathbb{E}[Y_{\cdot \ell}])_{\ell\in\mathcal{J}_{-k}} \end{pmatrix}^T - \mathbb{E}[Y_{\cdot k}])\mathbb{E}[Y_{\cdot m}]\right)_{k\in\mathcal{J}} \end{bmatrix} \begin{matrix} \} \in \mathbb{R} \\ \} \in \mathbb{R}^r \end{matrix}$$

$$\mathcal{O}(\sigma^2) = \begin{bmatrix} \overbrace{o_{\mathrm{var}}(\sigma^2)}^{\in\mathbb{R}} \\ -\left(o_{\mathrm{cov},k}(\sigma^2)\right)_{k\in\mathcal{J}} \end{bmatrix} \begin{matrix} \} \in \mathbb{R} \\ \} \in \mathbb{R}^r \end{matrix},$$

with $o_{\mathrm{var}}(\sigma^2)$ and $o_{\mathrm{cov},k}(\sigma^2)$ detailed in (C.29) and (C.32) respectively.

$$\begin{aligned} Q^c = &\left(\mathrm{Var}(Y_{\cdot j})|\Omega_{\cdot m} = 1\right) \\ &- \left(\mathrm{Cov}((Y_{\cdot k})_{k\in\overline{\{j\}}}, Y_{\cdot j})\mathrm{Var}((Y_{\cdot k})_{k\in\overline{\{j\}}})^{-1}\mathrm{Cov}((Y_{\cdot k})_{k\in\overline{\{j\}}}, Y_{\cdot j})^T|\Omega_{\cdot m} = 1\right). \end{aligned} \quad \text{(C.24)}$$

The strategy is to prove each equality of the linear system in (C.23).

Deriving an equation for the variance. The idea is first to express $\mathrm{Var}(Y_{\cdot j})$ from $\mathrm{Var}(Y_{\cdot m})$, $(\mathrm{Var}(Y_{\cdot j'}))_{j'\in\mathcal{J}_{-j}}$ and $(\mathrm{Cov}(Y_{\cdot k}, Y_{\cdot \ell}))_{k\neq\ell\in\{m\}\cup\mathcal{J}_{-j}}$. The law of total variance reads as

$$\mathrm{Var}(Y_{\cdot j}) = \mathbb{E}[\mathrm{Var}(Y_{\cdot j}|Z)] + \mathrm{Var}(\mathbb{E}[Y_{\cdot j}|Z]), \quad \text{(C.25)}$$

with $Z = (Y_{\cdot k})_{k\in\overline{\{j\}}}$.

For the first term in (C.25), using Assumption **A2.**, one has

$$Y_{\cdot j} \perp\!\!\!\perp (\Omega_{\cdot m} = 1)|Z$$

which leads to

$$\mathrm{Var}(Y_{\cdot j}|Z) = \mathrm{Var}(Y_{\cdot j}|Z, \Omega_{\cdot m} = 1).$$

The conditional variance for a Gaussian vector gives

$$\mathrm{Var}(Y_{\cdot j}|Z) = \mathrm{Var}(Y_{\cdot j}) - \mathrm{Cov}(Z, Y_{\cdot j})\mathrm{Var}(Z)^{-1}\mathrm{Cov}(Z, Y_{\cdot j})^T,$$

implying that

$$\mathrm{Var}(Y_{.j}|Z,\Omega_{.m}=1)=\left(\mathrm{Var}(Y_{.j})-\mathrm{Cov}(Z,Y_{.j})\mathrm{Var}(Z)^{-1}\mathrm{Cov}(Z,Y_{.j})^T|\Omega_{.m}=1\right)$$

and then, as deterministic quantity,

$$\mathbb{E}[\mathrm{Var}(Y_{.j}|Z)]=\left(\mathrm{Var}(Y_{.j})-\mathrm{Cov}(Z,Y_{.j})\mathrm{Var}(Z)^{-1}\mathrm{Cov}(Z,Y_{.j})^T|\Omega_{.m}=1\right).$$

One has

$$\mathrm{Cov}(Z,Y_{.j})\mathrm{Var}(Z)^{-1}\mathrm{Cov}(Z,Y_{.j})^T = \mathrm{Cov}((Y_{.k})_{k\in\overline{\{j\}}},Y_{.j})\mathrm{Var}((Y_{.k})_{k\in\overline{\{j\}}})^{-1}\mathrm{Cov}((Y_{.k})_{k\in\overline{\{j\}}},Y_{.j})^T$$

leading to

$$\mathbb{E}[\mathrm{Var}(Y_{.j}|Z)]=Q^c, \tag{C.26}$$

where Q^c is defined in (C.24).

For the second term of (C.25), remark that **A2.** implies that

$$\mathrm{Var}(\mathbb{E}[Y_{.j}|Z])=\mathrm{Var}(\mathbb{E}[Y_{.j}|Z,\Omega_{.m}=1]),$$

and

$$\mathrm{Var}(\mathbb{E}[Y_{.j}|Z,\Omega_{.m}=1])=\mathrm{Var}\left(\mathbb{E}\left[\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[0]+\sum_{k\in\{m\}\cup\mathcal{J}_{-j}}\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[k]Y_{.k}+\zeta^c\middle|Z\right]\right),$$

i.e.

$$\mathrm{Var}(\mathbb{E}[Y_{.j}|Z,\Omega_{.m}=1]) = \mathrm{Var}\left(\sum_{k\in\{m\}\cup\mathcal{J}_{-j}}\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[k]Y_{.k}-\sum_{k\in\{m\}\cup\mathcal{J}_{-j}}\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[k]\mathbb{E}[\epsilon_{.k}|Z]+\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[0]+\mathbb{E}[\epsilon_{.j}]\right)$$

In the variance, the first term is obtained using that the variables $(Y_{.k})_{k\in\{m\}\cup\mathcal{J}_{-j}}$ are $Z-$measurable. The two last terms use that $\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}}[0]$ is a constant and $\epsilon_{.j}$ is independent of Z. To calculate the second term, involving $\mathbb{E}[\epsilon_{.k}|Z]$, one first shows that the vector $\left((Y_{.k})_{k\in\{m\}\cup\mathcal{J}_{-j}}\quad(\epsilon_{.k})_{k\in\{m\}\cup\mathcal{J}_{-j}}\right)^T$ is Gaussian. Indeed,

- $(Y_{.k})_{k\in\{m\}\cup\mathcal{J}_{-j}}$ is a Gaussian vector, using the model (3.1).
- $(\epsilon_{.k})_{k\in\{m\}\cup\mathcal{J}_{-j}}$ is a Gaussian vector, because its components are independent Gaussian variables.
- for $k\neq\ell\in\{m\}\cup\mathcal{J}_{-j}$, $\left(WB_{k.}\quad\epsilon_{.\ell}\right)^T$ is a Gaussian vector, because $Y_{.k}\perp\!\!\!\perp\epsilon_{.\ell}$.
- for $k\in\{m\}\cup\mathcal{J}_{-j}$, $\left(Y_{.k}\quad\epsilon_{.k}\right)^T$ is a Gaussian vector, given that $Y_{.k}$ is a linear combination of $\left(WB_{k.}\quad\epsilon_{.k}\right)^T$ which is Gaussian, as $WB_{k.}$ and $\epsilon_{.k}$ are independent Gaussian variables.

Thus,

$$\begin{aligned}\mathbb{E}[\epsilon_{.k}|Z] &= \mathbb{E}[\epsilon_{.k}] + \mathrm{Cov}(\epsilon_{.k}, Z)\mathrm{Var}(Z)^{-1}(Z - \mathbb{E}[Z]) \\ &= \mathrm{Cov}(\epsilon_{.k}, Y_{.k})(\mathrm{Var}(Z)^{-1})_{k.}(Z - \mathbb{E}[Z]),\end{aligned}$$

using $\mathrm{Cov}(\epsilon_{.k}, Y_{.l}) = 0$, for $k \neq l$. $\Gamma_Z = \mathrm{Var}(Z)^{-1}$ denotes the inverse of the covariance matrix of Z and $(\Gamma_Z)_{k.}$ is its k-th row. It leads to

$$\mathbb{E}[\epsilon_{.k}|Z] = \sigma^2(\Gamma_Z)_{k.}(Z - \mathbb{E}[Z]). \tag{C.27}$$

given that $\mathrm{Cov}(\epsilon_{.k}, Y_{.k}) = \mathrm{Cov}(\epsilon_{.k}, WB_{k.} + \epsilon_{.k}) = \mathrm{Var}(\epsilon_{.k})$.

Therefore,

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[Y_{.j}|Z, \Omega_{.m} = 1]) = &\sum_{k\in\{m\}\cup\mathcal{J}_{-j}} (\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[k]})^2\mathrm{Var}(Y_{.k}) \\ &+ \sum_{(k<\ell)\in\{m\}\cup\mathcal{J}_{-j}} 2\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[k]}\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[\ell]}\mathrm{Cov}(Y_{.k}, Y_{.\ell}) + o_{\mathrm{var}}(\sigma^2),\end{aligned} \tag{C.28}$$

where

$$\begin{aligned}o_{\mathrm{var}}(\sigma^2) = &-2\sigma^2 \sum_{(k,\ell)\in\{m\}\cup\mathcal{J}_{-j}} \mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[k]}\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[\ell]} \sum_{\ell'\in\{m\}\cup\mathcal{J}_{-j}} (\Gamma_Z)_{\ell\ell'}\mathrm{Cov}(Y_{.k}, Y_{.\ell'}) \\ &+ \sigma^4 \sum_{k\in\{m\}\cup\mathcal{J}_{-j}} (\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[k]})^2 \left(\sum_{(\ell<\ell')\in\{m\}\cup\mathcal{J}_{-j}} (\Gamma_Z)^2_{k\ell}\mathrm{Var}(Y_{.\ell}) - 2(\Gamma_Z)_{k\ell}(\Gamma_Z)_{k\ell'}\mathrm{Cov}(Y_{.\ell}, Y_{.\ell'}) \right) \\ &- 2\sigma^4 \sum_{(k<\ell)\in\{m\}\cup\mathcal{J}_{-j}} \mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[k]}\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[\ell]} \sum_{(k',\ell')\in\{m\}\cup\mathcal{J}_{-j}} (\Gamma_Z)_{kk'}(\Gamma_Z)_{\ell\ell'}\mathrm{Cov}(Y_{.k'}, Y_{.\ell'})\end{aligned} \tag{C.29}$$

Combining (C.26) with (C.28), one get the following expression for the first line of the linear system

$$\begin{aligned}(\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[m]})^2\mathrm{Var}(Y_{.m}) + &\sum_{j'\in\mathcal{J}_{-j}} 2\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[j']}\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[m]}\mathrm{Cov}(Y_{.j'}, Y_{.m}) \\ &= \mathrm{Var}(Y_{.j}) - Q^c - (\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[\mathcal{J}_{-j}]})^T\mathrm{Var}(Y_{\mathcal{J}_{-j}})\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[\mathcal{J}_{-j}]} - o_{\mathrm{var}}(\sigma^2)\end{aligned} \tag{C.30}$$

Deriving equations for the covariances. Let k be an element of $\mathcal{J}$, our objective is to express $\mathrm{Cov}(Y_{.m}, Y_{.k})$ from $\mathrm{Var}(Y_{.m})$, α_m, $(\alpha_k)_{k\in\mathcal{J}}$ and $(\mathrm{Cov}(Y_{.m}, Y_{.k}))_{k\in\{m\}\cup\mathcal{J}}$.

$$\begin{aligned}\mathrm{Cov}(Y_{.m}, Y_{.k}) &= \mathbb{E}[Y_{.m}Y_{.k}] - \mathbb{E}[Y_{.m}]\mathbb{E}[Y_k] \\ &= \mathbb{E}[\mathbb{E}[Y_{.m}Y_{.k}|Z]] - \mathbb{E}[Y_{.m}]\mathbb{E}[Y_{.k}] \\ &= \mathbb{E}[Y_{.m}\mathbb{E}[Y_{.k}|Z]] - \mathbb{E}[Y_{.m}]\mathbb{E}[Y_{.k}],\end{aligned} \tag{C.31}$$

with $Z = (Y_{.\ell})_{\ell\in\overline{\{k\}}}$.

For the first term in (C.31), one has

$$\begin{aligned}\mathbb{E}[Y_{\cdot m}\mathbb{E}[Y_{\cdot k}|Z]] &\overset{(i)}{=}\mathbb{E}[Y_{\cdot m}\mathbb{E}[Y_{\cdot k}|Z,\Omega_{\cdot m}=1]]\\ &\overset{(ii)}{=}\mathbb{E}\left[Y_{\cdot m}\left(\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[0]}+\sum_{\ell\in\{m\}\cup\mathcal{J}_{-k}}\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[\ell]}Y_{\cdot\ell}+\mathbb{E}[\zeta^c_k|Z]\right)\right]\\ &\overset{(iii)}{=}\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[0]}\mathbb{E}[Y_{\cdot m}]+\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[m]}\mathbb{E}[Y^2_{\cdot m}]\\ &\quad+\sum_{\ell\in\mathcal{J}_{-k}}\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[\ell]}\mathbb{E}[Y_{\cdot m}Y_{\cdot\ell}]+o_{\text{cov},k}(\sigma^2)\end{aligned}$$

with $\zeta^c_k=-\sum_{\ell\in\mathcal{J}_{-k}}\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[\ell]}\epsilon_{\cdot\ell}-\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[m]}\epsilon_{\cdot m}+\epsilon_{\cdot k}$.

Assumption **A2.** and Definition 10 are used for (i) and (ii) respectively. For (iii), using (C.27), one has

$$\mathbb{E}[Y_{\cdot m}\mathbb{E}[\zeta^c_k|Z]]=\mathbb{E}\left[Y_{\cdot m}\left(-\sum_{\ell\in\{m\}\cup\mathcal{J}_{-k}}\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[\ell]}\sigma^2(\Gamma_Z)_{\ell\cdot}(Z-\mathbb{E}[Z])\right)\right],$$

given that $\mathbb{E}[\epsilon_{\cdot k}|Z]=\mathbb{E}[\epsilon_{\cdot k}]=0$ by independence.

$$\begin{aligned}&\mathbb{E}[Y_{\cdot m}\mathbb{E}[\zeta^c_k|Z]]\\ &\qquad=-\sigma^2\mathbb{E}\left[\sum_{\ell\in\{m\}\cup\mathcal{J}_{-k}}\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[\ell]}Y_{\cdot m}\sum_{\ell'\{m\}\cup\in\mathcal{J}_{-k}}(\Gamma_Z)_{\ell\ell'}(Y_{\cdot\ell'}-\mathbb{E}[Y_{\cdot\ell'}])\right].\end{aligned}$$

In addition,

$$\begin{aligned}&\mathbb{E}\left[\sum_{\ell\in\{m\}\cup\mathcal{J}_{-k}}\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[\ell]}Y_{\cdot m}\sum_{\ell'\in\{m\}\cup\mathcal{J}_{-k}}(\Gamma_Z)_{\ell\ell'}(Y_{\cdot\ell'}-\mathbb{E}[Y_{\cdot\ell'}])\right]\\ &=\sum_{\ell\in\{m\}\cup\mathcal{J}_{-k}}\sum_{\ell'\in\{m\}\cup\mathcal{J}_{-k}}(\Gamma_Z)_{\ell\ell'}\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[\ell]}\left(\text{Cov}\left(Y_{\cdot m},Y_{\cdot\ell'}\right)+\mathbb{E}[Y_{\cdot m}]\mathbb{E}[(Y_{\cdot\ell'}-\mathbb{E}[Y_{\cdot\ell'}])]\right)\\ &=\sum_{\ell\in\{m\}\cup\mathcal{J}_{-k}}\sum_{\ell'\in\{m\}\cup\mathcal{J}_{-k}}(\Gamma_Z)_{\ell\ell'}\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[\ell]}\text{Cov}\left(Y_{\cdot m},Y_{\cdot\ell'}\right)\end{aligned}$$

It implies that, in (iii),

$$o_{\text{cov},k}(\sigma^2)=-\sigma^2\sum_{\ell\in\{m\}\cup\mathcal{J}_{-k}}\sum_{\ell'\in\{m\}\cup\mathcal{J}_{-k}}(\Gamma_Z)_{\ell\ell'}\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[\ell]}\text{Cov}\left(Y_{\cdot m},Y_{\cdot\ell'}\right)\tag{C.32}$$

Equation (C.31) leads thus to

$$\begin{aligned}\text{Cov}(Y_{\cdot m},Y_{\cdot k})&=\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[0]}\mathbb{E}[Y_{\cdot m}]+\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[m]}(\text{Var}(Y_{\cdot m})+\mathbb{E}[Y_{\cdot m}]^2)\\ &+\sum_{\ell\in\mathcal{J}_{-k}}\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[\ell]}(\text{Cov}(Y_{\cdot m},Y_{\cdot\ell})+\mathbb{E}[Y_{\cdot m}]\mathbb{E}[Y_{\cdot\ell}])-\mathbb{E}[Y_{\cdot m}]\mathbb{E}[Y_{\cdot k}]+o_{\text{cov},k}(\sigma^2),\end{aligned}\tag{C.33}$$

which can be rewritten as

$$\begin{aligned}\mathrm{Cov}(Y_{.m},Y_{.k}) - \mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[m]}\mathrm{Var}(Y_{.m}) - \sum_{\ell\in\mathcal{J}_{-k}} \mathcal{B}^c_{k\to m,\mathcal{J}_{-k}[\ell]}\mathrm{Cov}(Y_{.m},Y_{.\ell})\\ = ((\mathcal{B}^c_{k\to m,\mathcal{J}_{-k}})^T \begin{pmatrix}1 & \mathbb{E}[Y_{.m}] & (\mathbb{E}[Y_{.\ell}])_{\ell\in\mathcal{J}_{-k}}\end{pmatrix}^T - \mathbb{E}[Y_{.k}])\mathbb{E}[Y_{.m}] + o_{\mathrm{cov},k}(\sigma^2), \quad \text{(C.34)}\end{aligned}$$

Combining Equations (C.30) and (C.34) forms the desired matrix system (C.23).

From these formulae for $\begin{pmatrix}\mathrm{Var}(Y_{.m}) & \mathrm{Cov}(Y_{.m},(Y_{.k})_{k\in\mathcal{J}})\end{pmatrix}^T$, assuming that M_j is invertible and that σ^2 tends to zero, one get their estimators $\begin{pmatrix}\widehat{\mathrm{Var}}(Y_{.m}) & \widehat{\mathrm{Cov}}(Y_{.m},(Y_{.k})_{k\in\mathcal{J}})\end{pmatrix}^T$ defined in (3.10).

As for the consistency, $\hat{\alpha}_m$ is a consistent estimator for α_m by using Proposition 11. The estimators in (3.10) are consistent, under Assumption **A3.** and **A4.**. □

C.2.4 Proof of Proposition 28

For deriving the covariance between a MNAR variable and a MNAR or not pivot variable, we assume the following

A5. $\forall m\in\mathcal{M}$, $\forall \ell\in\overline{\mathcal{J}}_{-m}$, for all set $\mathcal{H}\subset\mathcal{J}_{-j}$ such that $|\mathcal{H}|=r-2$, $\begin{pmatrix}B_{.m} & B_{.\ell} & (B_{.j'})_{j'\in\mathcal{H}}\end{pmatrix}$ is invertible,

A6. $\forall k\in\overline{\mathcal{J}}\backslash\mathcal{M}$, $\forall j\in\mathcal{J}$, $Y_{.j}\perp\!\!\!\perp \Omega_{.k}|(Y_{.\ell})_{\ell\in\overline{\{j\}}}$.

A7. $\forall k,\ell\in\overline{\mathcal{J}}$, $k\neq l$, $\Omega_{.k}\perp\!\!\!\perp\Omega_{.\ell}|Y$

A8. $\forall j\in\mathcal{J},\forall m\in\mathcal{M},\forall\ell\in\overline{\mathcal{J}}_{-m}$, for all set $\mathcal{H}\subset\mathcal{J}_{-j}$ such that $|\mathcal{H}|=r-2$, the complete-case coefficients $\mathcal{B}^c_{j\to m,\ell,\mathcal{H}[0]}$ and $\mathcal{B}^c_{j\to m,\ell,\mathcal{H}[k]}, k\neq j, k\in\{m,\ell\}\cup\mathcal{H}$ can be consistently estimated. (Here, note that the complete case is when $\Omega_{.m}=1$ and $\Omega_{.\ell}=1$.)

A9. For the variables neither MNAR nor pivot, their means $(\alpha_k)_{k\in\overline{\mathcal{J}}\backslash\mathcal{M}}$, variances $(\mathrm{Var}(Y_{.k}))_{k\in\overline{\mathcal{J}}\backslash\mathcal{M}}$ and covariances $(\mathrm{Cov}(Y_{.k},Y_{.k'}))_{k\neq k'\in\overline{\mathcal{J}}\backslash\mathcal{M}}$ can be consistently estimated. The covariances between these variables and the pivot variables $(\mathrm{Cov}(Y_{.j},Y_{.k}))_{j\in\mathcal{J},k\in\overline{\mathcal{J}}\backslash\mathcal{M}}$ are also consistent.

Proposition 28 (Covariance between a MNAR variable and a MNAR or not pivot variable)**.** *Consider the PPCA model (3.1). Under Assumptions **A2.**, **A5.**, **A6.** and **A7.**, an estimator of the covariance between a MNAR variable $Y_{.m}$, for $m\in\mathcal{M}$, and a variable $Y_{.\ell}$, for $\ell\in\overline{\mathcal{J}}\backslash\{m\}$, can be constructed as follows: choose $j\in\mathcal{J}$ and $r-2$ variable indexes in $\mathcal{J}_{-j}$ and compute:*

$$\begin{aligned}\widehat{\mathrm{Cov}}(Y_{.m},Y_{.\ell}) = \frac{1}{\hat{K}}\widehat{\mathrm{Var}}(Y_{.j}) - \hat{q}^c - \sum_{k\in\{m,\ell\}\cup\mathcal{H}} (\hat{\mathcal{B}}^c_{j\to m,\ell,\mathcal{H}[k]})^2\widehat{\mathrm{Var}}(Y_{.k})\\ - \sum_{k<k',k\in\{m,\ell\}\cup\mathcal{H},k'\in\mathcal{H}} 2\hat{\mathcal{B}}^c_{j\to m,\ell,\mathcal{H}[k]}\hat{\mathcal{B}}^c_{j\to m,\ell,\mathcal{H}[k']}\widehat{\mathrm{Cov}}(Y_{.k},Y_{.k'}), \quad \text{(C.35)}\end{aligned}$$

assuming that σ^2 *tends to zero and with* $\hat{K} = 2\hat{\mathcal{B}}^c_{j\to m,\ell,\mathcal{H}[m]}\hat{\mathcal{B}}^c_{j\to m,\ell,\mathcal{H}[\ell]}$ *and*

$$\hat{q}^c = \left(\widehat{\mathrm{Var}}(Y_{.j})\middle|\Omega_{.m}=1,\Omega_{.\ell}=1\right) - \left(\widehat{\mathrm{Cov}}((Y_{.k})_{k\in\overline{\{j\}}},Y_{.j})\widehat{\mathrm{Var}}((Y_{.k})_{k\in\overline{\{j\}}})^{-1}\widehat{\mathrm{Cov}}((Y_{.k})_{k\in\overline{\{j\}}},Y_{.j})^T\middle|\Omega_{.m}=1,\Omega_{.\ell}=1\right),$$

given that K *estimated by* $\hat{K}$ *is non zero.*

Under the additional Assumptions ***A3.***, ***A8.*** *and* ***A9.****. this estimator given in* (C.35) *is consistent.*

Proof. Let $\mathcal{H}$ be the set of the $r-2$ variable indexes. One has $\mathcal{H}\subset\mathcal{J}_{-j}$. We use the same strategy as the proof for Proposition 12 (paragraph for deriving an equation for the variance).

To derive a formula for $\mathrm{Cov}(Y_{.m},Y_{.\ell})$ with $m\in\mathcal{M}$ and $\ell\in\overline{\mathcal{J}}_{-m}$, the idea is to express $\mathrm{Var}(Y_{.j})$ from $(\mathrm{Var}(Y_{.k}))_{k\in\{m,\ell\}\cup\mathcal{H}}$ and $(\mathrm{Cov}(Y_{.k},Y_{.k'}))_{k\neq k'\in\{m,\ell\}\cup\mathcal{H}}$.

The law of total variance reads as

$$\mathrm{Var}(Y_{.j}) = \mathbb{E}[\mathrm{Var}(Y_{.j}|Z)] + \mathrm{Var}(\mathbb{E}[Y_{.j}|Z]), \tag{C.36}$$

with $Z=(Y_{.k})_{k\in\overline{\{j\}}}$.

For the first term in (C.36), one uses

$$Y_{.j}\perp\!\!\!\perp\Omega_{.m},\Omega_{.\ell}|Z.$$

If $Y_{.m}$ and $Y_{.\ell}$ are both MNAR variables, this conditional independance is obtained using Assumption **A2.** and **A7.**. Otherwise, if $Y_{.\ell}$ is not a MNAR variable, Assumption **A6.** and **A7.** lead to the desired result. It implies

$$\mathrm{Var}(Y_{.j}|Z) = \mathrm{Var}(Y_{.j}|Z,\Omega_{.m}=1,\Omega_{.\ell}=1).$$

The conditional variance for a Gaussian vector gives

$$\mathrm{Var}(Y_{.j}|Z) = \mathrm{Var}(Y_{.j}) - \mathrm{Cov}(Z,Y_{.j})\mathrm{Var}(Z)^{-1}\mathrm{Cov}(Z,Y_{.j})^T,$$

implying that

$$\mathrm{Var}(Y_{.j}|Z,\Omega_{.m}=1,\Omega_{.\ell}=1) = \left(\mathrm{Var}(Y_{.j}) - \mathrm{Cov}(Z,Y_{.j})\mathrm{Var}(Z)^{-1}\mathrm{Cov}(Z,Y_{.j})^T\middle|\Omega_{.m}=1,\Omega_{.\ell}=1\right)$$

and then, as deterministic quantity,

$$\mathbb{E}[\mathrm{Var}(Y_{.j}|Z)] = q^c \tag{C.37}$$

with

$$q^c = \left(\mathrm{Var}(Y_{.j})\middle|\Omega_{.m}=1,\Omega_{.\ell}=1\right) - \left(\mathrm{Cov}((Y_{.k})_{k\in\overline{\{j\}}},Y_{.j})\mathrm{Var}((Y_{.k})_{k\in\overline{\{j\}}})^{-1}\mathrm{Cov}((Y_{.k})_{k\in\overline{\{j\}}},Y_{.j})^T\middle|\Omega_{.m}=1,\Omega_{.\ell}=1\right).$$

For the second term of (C.25), remark that **A2.**, **A6.** and **A7.** implies that

$$\mathrm{Var}(\mathbb{E}[Y_{.j}|Z]) = \mathrm{Var}(\mathbb{E}[Y_{.j}|Z, \Omega_{.m} = 1, \Omega_{.\ell} = 1]),$$

and

$$\mathrm{Var}(\mathbb{E}[Y_{.j}|Z, \Omega_{.m} = 1, \Omega_{.\ell} = 1]) = \mathrm{Var}\left(\mathbb{E}\left[\mathcal{B}^c_{j\to m,\ell,\mathcal{H}[0]} + \sum_{k\in\{m,\ell\}\cup\mathcal{H}} \mathcal{B}^c_{j\to m,\ell,\mathcal{H}[k]} Y_{.k} + \zeta^c_j \middle| Z\right]\right),$$

i.e.

$$\mathrm{Var}(\mathbb{E}[Y_{.j}|Z, \Omega_{.m} = 1, \Omega_{.\ell} = 1]) = \mathrm{Var}\left(\sum_{k\in\{m,\ell\}\cup\mathcal{H}} \mathcal{B}^c_{j\to m,\ell,\mathcal{H}[k]} Y_{.k} - \sum_{k\in\{m,\ell\}\cup\mathcal{H}} \mathcal{B}^c_{j\to m,\ell,\mathcal{H}[k]} \mathbb{E}[\epsilon_{.k}|Z] + \mathcal{B}^c_{j\to m,\ell,\mathcal{H}[0]} + \mathbb{E}[\epsilon_{.j}]\right)$$

One uses the same reasoning as in the proof of Proposition 12 (paragraph for deriving an equation for the variance) to get

$$\mathrm{Var}(\mathbb{E}[Y_{.j}|Z, \Omega_{.m} = 1, \Omega_{.\ell} = 1]) = \sum_{k\in\{m,\ell\}\cup\mathcal{H}} (\mathcal{B}^c_{j\to m,\ell,\mathcal{H}[k]})^2 \mathrm{Var}(Y_{.k}) + \sum_{k<k'\in\{m,\ell\}\cup\mathcal{H}} 2\mathcal{B}^c_{j\to m,\ell,\mathcal{H}[k]} \mathcal{B}^c_{j\to m,\ell,\mathcal{H}[k']} \mathrm{Cov}(Y_{.k}, Y_{.k'}) + o_{\text{covmiss}}(\sigma^2), \quad \text{(C.38)}$$

where

$$\begin{aligned} o_{\text{covmiss}}(\sigma^2) &= -2\sigma^2 \sum_{(k,k')\in\{m,\ell\}\cup\mathcal{H}} \mathcal{B}^c_{j\to m,\ell,\mathcal{H}[k]} \mathcal{B}^c_{j\to m,\ell,\mathcal{H}[k']} \sum_{\ell'\in\{m,\ell\}\cup\mathcal{H}} (\Gamma_Z)_{k'\ell'} \mathrm{Cov}(Y_{.k}, Y_{.\ell'}) \\ &+ \sigma^4 \sum_{k\in\{m,\ell\}\cup\mathcal{H}} (\mathcal{B}^c_{j\to m,\ell,\mathcal{H}[k]})^2 \left(\sum_{(k'<\ell')\in\{m,\ell\}\cup\mathcal{H}} (\Gamma_Z)^2_{kk'} \mathrm{Var}(Y_{.k'}) - 2(\Gamma_Z)_{kk'} (\Gamma_Z)_{k\ell'} \mathrm{Cov}(Y_{.k'}, Y_{.\ell'}) \right) \\ &- 2\sigma^4 \sum_{(k<k')\in\{m,\ell\}\cup\mathcal{H}} \mathcal{B}^c_{j\to m,\ell,\mathcal{H}[k]} \mathcal{B}^c_{j\to m,\ell,\mathcal{H}[k']} \sum_{(k'',\ell')\in\{m,\ell\}\cup\mathcal{H}} (\Gamma_Z)_{kk''} (\Gamma_Z)_{k'\ell'} \mathrm{Cov}(Y_{.k''}, Y_{.\ell'}) \quad \text{(C.39)} \end{aligned}$$

Combining (C.36), (C.37) and (C.38), one get the following formula for $\mathrm{Cov}(Y_{.m}, Y_{.\ell})$,

$$2\mathcal{B}^c_{j\to m,\ell,\mathcal{H}[m]} \mathcal{B}^c_{j\to m,\mathcal{H}[\ell]} \mathrm{Cov}(Y_{.m}, Y_{.\ell}) = \mathrm{Var}(Y_{.j}) - q^c - \sum_{k\in\{m,\ell\}\cup\mathcal{H}} (\mathcal{B}^c_{j\to m,\ell,\mathcal{H}[k]})^2 \mathrm{Var}(Y_{.k}) - \sum_{k<k', k\in\{m,\ell\}\cup\mathcal{H}, k'\in\mathcal{H}} 2\mathcal{B}^c_{j\to m,\ell,\mathcal{H}[k]} \mathcal{B}^c_{j\to m,\ell,\mathcal{H}[k']} \mathrm{Cov}(Y_{.k}, Y_{.k'}) - o_{\text{covmiss}}(\sigma^2)$$

An estimator of $\mathrm{Cov}(Y_{.m}, Y_{.\ell})$ is then derived as in (C.35), given that σ^2 tends to zero and $K = \mathcal{B}^c_{j\to m,\ell,\mathcal{H}[m]} \mathcal{B}^c_{j\to m,\ell,\mathcal{H}[\ell]}$ is non zero.

We use the consistent estimators defined in Proposition 12 for estimating $\mathrm{Var}(Y_{.m})$ and $\mathrm{Cov}(Y_{.m}, Y_{.k})_{k\in\mathcal{H}}$. If $Y_{.\ell}$ is also a MNAR variable, Proposition 12 is applied for estimating $\mathrm{Var}(Y_{.\ell})$ and $\mathrm{Cov}(Y_{.\ell}, Y_{.k})_{k\in\mathcal{H}}$. Otherwise, if $Y_{.\ell}$ is not a MNAR variable, we use **A9.**.

Eventually, **A3.** and **A8.** lead to the consistency of $\widehat{\mathrm{Cov}}(Y_{.m}, Y_{.\ell})$. □

C.2.5 Extension to more general mechanisms for the not MNAR variables

The results of Proposition 11, 12 and 28 can be extended to a more general setting than the one presented in Section 3.2. The pivot variables may be assumed to be MCAR (or observed). The variables which are neither MNAR nor pivot may be observed or satisfying

$$\forall \ell \in \overline{\mathcal{J}} \backslash \mathcal{M}, \forall i \in \{1, \dots, n\}, \quad \mathbb{P}(\Omega_{i\ell} = 1 | Y_{i.}) = \mathbb{P}(\Omega_{i\ell} = 1 | (Y_{ik})_{k \in \overline{\mathcal{J}} \backslash \{\ell\} \cup \mathcal{M}}), \tag{C.40}$$

i.e. they are MCAR or MAR but their missing-data mechanisms may not depend on the pivot variables.

The proofs are similar and not presented here for the sake of brevity.

Note that the main difference is that the complete case has to be extended. For instance, for $j \in \mathcal{J}$ and $k \in \mathcal{J}_{-j}$, the coefficients standing respectively for the intercept and the effects of $Y_{.j}$ on $(Y_{.m}, (Y_{.j'})_{j' \in \mathcal{J}_{-j}})$ in the complete case, *i.e.* when $\Omega_{.m} = 1, (\Omega_j = 1)_{j \in \mathcal{J}}$ are in this general setting defined as follows

$$(Y_{.j})_{|\Omega_{.m}=1, (\Omega_j=1)_{j\in\mathcal{J}}} := \mathcal{B}^c_{j \to m, \mathcal{J}_{-j}}[0] + \sum_{j' \in \mathcal{J}_{-j}} \mathcal{B}^c_{j \to m, \mathcal{J}_{-j}}[j'] Y_{.j'} + \mathcal{B}^c_{j \to m, \mathcal{J}_{-j}}[m] Y_{.m} + \zeta^c,$$

with $\zeta^c = -\sum_{j' \in \mathcal{J}_{-j}} \mathcal{B}^c_{j \to m, \mathcal{J}_{-j}}[j'] \epsilon_{.j'} - \mathcal{B}^c_{j \to m, \mathcal{J}_{-j}}[m] \epsilon_{.m} + \epsilon_{.j}$.

C.3 Other numerical experiments

Robustness to noise. Considering the same setting as in Section 3.4.1 ($n = 1000$, $p = 10$, $r = 2$ and seven self-masked MNAR variables), the methods are tried for different noise levels $\sigma^2 \in \{0.1, 0.3, 0.5, 0.7, 1\}$. The results are presented for one missing variable and for all the other ones, the results are similar. In Figure C.1, Algorithm 1 is the only method that does not give a biased estimate of the mean and the variance regardless of the noise level. In Figure C.2, despite a larger bias in the estimation of the covariance between a missing variable and a pivot one as the noise level increases, Algorithm 1 outperforms all the other methods, regarding the estimation of the covariance between two missing variables. Note that the formula for the estimate of the covariance between two missing variables relies on the one for the estimate of the variance, but both differ from the one used for the covarance estimation between a missing variable and a pivot one. As expected, in Figure C.3, estimation deteriorates as the data gets noisier and then the loading matrix estimation and the imputation error get closer to the results of mean imputation. In term of imputation error, the proposed method yet remains competitive in regards of the approaches (ii) and (iii). Overall, when the noise level increases, the exogeneity will be worse and that ignoring it in practice can be made to the detriment of performance.

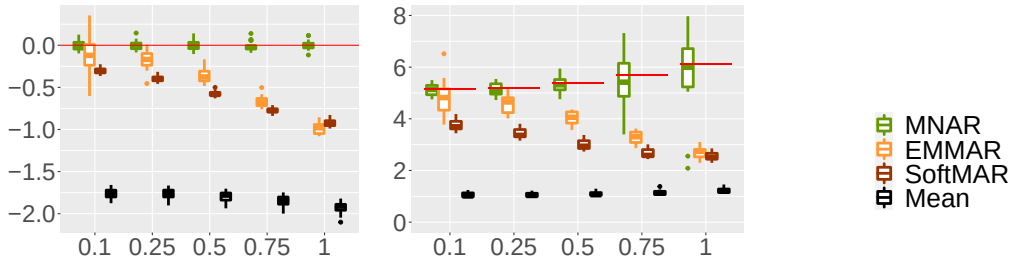


Figure C.1: Mean estimation (left graphic) and variance estimation (right graphic) of one missing variable for different values of the level of noise when $r = 2$, $n = 1000$, $p = 10$ and seven variables are MNAR. True values to be estimated are indicated by red lines.

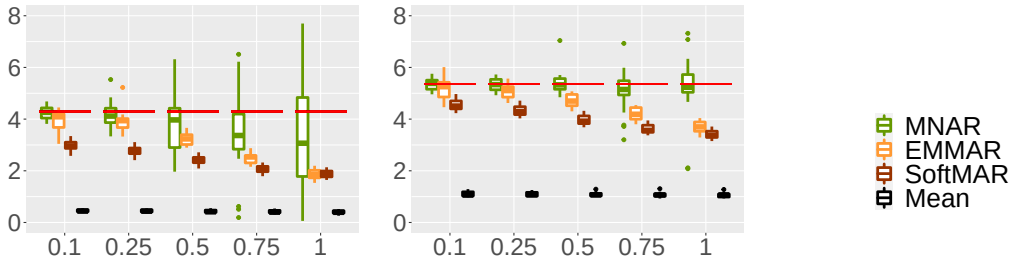


Figure C.2: Covariance estimation beetween a missing variable and a pivot one (left graphic) and two missing variables (right graphic) for different values of the level of noise when $r = 2$, $n = 1000$, $p = 10$ and seven variables are MNAR. True values to be estimated are indicated by red lines.

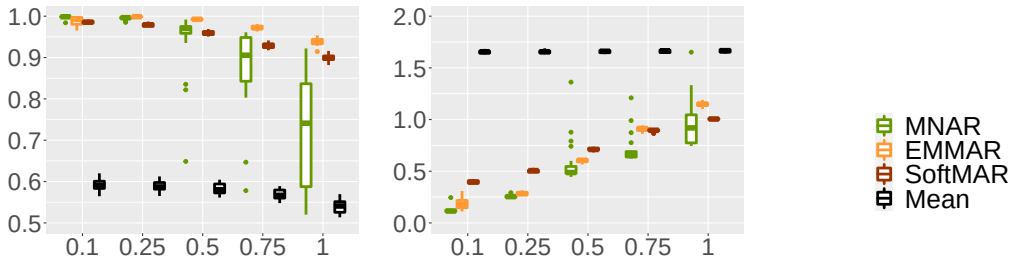


Figure C.3: RV coefficients for the loading matrix (left graphic) and imputation error (right graphic) for different values of the level of noise when $r = 2$, $n = 1000$, $p = 10$ and seven variables are MNAR.

Varying the percentage of missing values. Considering the same setting as in Section 3.4.1 ($n = 1000$, $p = 10$, $r = 2$, $\sigma = 0.1$ and seven self-masked MNAR variables), the

methods are tried for different percentages of missing values (10%, 30%, 50%). The results are presented in Figure C.4. As expected, all the methods deteriorate with an increasing percentage of missing values but our method is stable.

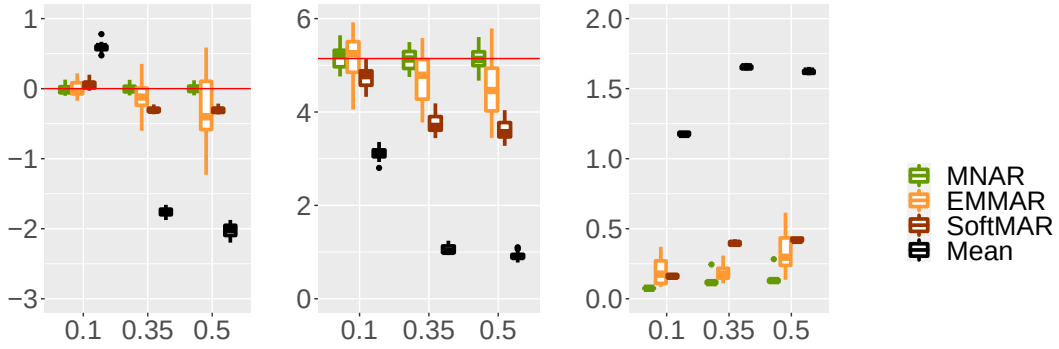


Figure C.4: Mean estimation (left graphic), variance estimation (middle graphic) and imputation error (right graphic) for different percentages of missing values when $r = 2$, $n = 1000$, $p = 10$ and seven variables are MNAR.

Misspecification to the rank. The misspecification to the parameter r has been evaluated: under a model generated with $r = 3$ latent variables ($n = 1000$, $p = 20$, $\sigma = 0.8$ and ten MNAR self-masked variables), the rank is either underestimated, well estimated or overestimated by giving to Algorithm 1 the information that $r = 2$, $r = 3$ or $r = 4$. Both estimation of the loading matrix and imputation error are shown in Figure C.5. The results for an underestimated ($r = 2$) or overestimated ($r = 4$) rank are comparable to the case where the accurate rank is considered instead ($r = 3$), showing a stability of Algorithm 1 to rank misspecification.

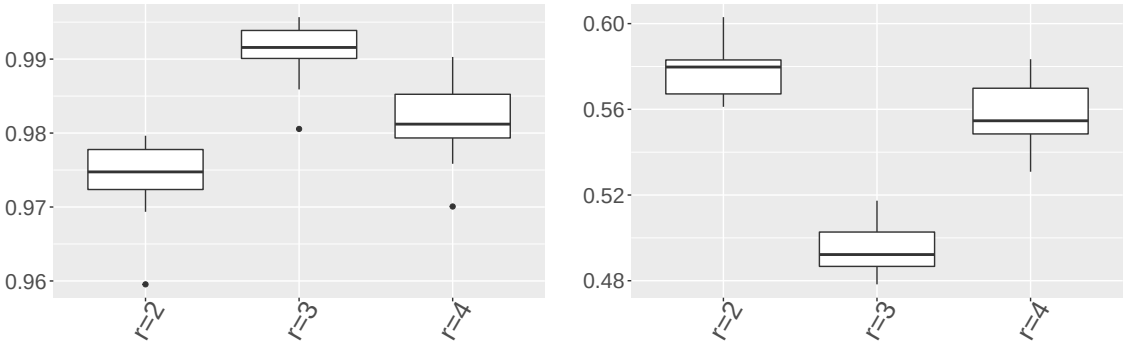


Figure C.5: RV coefficients for the loading matrix (left) and imputation error (right) when $r = 3$, $n = 1000$, $p = 20$ and ten variables are MNAR for different cases where the rank is either underestimated, well estimated or overestimated.

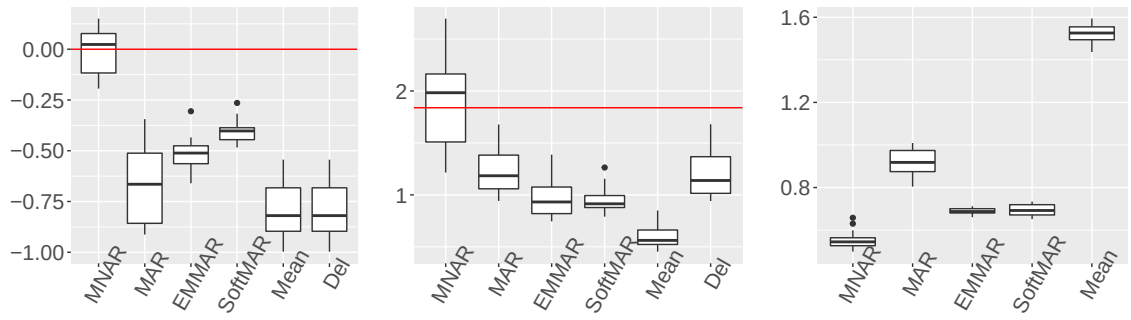


Figure C.6: Mean estimation (left), variance estimation (middle) of one missing variable and imputation error (right) when $r = 3$, $n = 1000$, $p = 20$ and ten variables are MNAR as in (C.41). True values are indicated in red lines.

General MNAR mechanism. We consider the setting $n = 1000$, $p = 20$, $r = 3$ and $\sigma = 0.8$. Here, missing values are introduced on ten variables $(Y_{.k})_{k\in[1:10]}$ using a more general MNAR mechanism (see (3.3)) than the self-masked one. In particular, the MNAR mechanism we consider is defined as follows,

$$\forall m \in [1:10], \forall i \in \{1,\dots,n\},\ \mathbb{P}(\Omega_{im} = 1|Y_{i.}) = \mathbb{P}(\Omega_{im} = 1|Y_{im}, Y_{ik}, Y_{i\ell}), \tag{C.41}$$

where k and ℓ are indexes of MNAR variables randomly chosen such that $k \neq \ell \in [1:10]\backslash\{m\}$. In Figure C.6, Algorithm 1 provides the best estimators of the mean and the variance (in term of bias) and the smallest imputation error.

Higher dimension and variation of the rank. The performance of the different methods for higher dimension is assessed. A data matrix of size $n = 1000$ and $p = 50$ is generated from two latent variables ($r = 2$) and with a noise level $\sigma = 1$. Missing values are introduced on twenty variables according to a self-masked MNAR mechanism, leading to 20% of missing values in total. Without loss of generality, the results are presented for one missing variable. Method (iv) has been discarded, as its computational time is too high for this setting.

In Figure C.7, as for the estimated mean and variance, Methods (i), (ii) and (iii) suffer from a large bias, while Algorithm 1 gives unbiased estimators. The same comment can be done for the estimation of the covariance between two missing values in Figure C.8. As for the covariance estimation between a missing variable and a pivot one Figure C.8, Algorithm 1 suffers from a variability, which can be due to the fact that in this higher dimension setting, not all the possible combinations of pivot variables are considered. Indeed, instead of taking the set of pivot variables of all the not MNAR variables *i.e.* $\mathcal{J} = \overline{\mathcal{M}}$, we choose $\mathcal{J} \subset \overline{\mathcal{M}}$ such that $|\mathcal{J}| = 10$. For the mean, 270 combinations of the pivot variables are aggregated over 870 possible combinations if $\mathcal{J} = \overline{\mathcal{M}}$.

Despite this dispersed estimator of the covariance between a MNAR variable and a pivot one, Algorithm 1 gives in Figure C.9 a high RV coefficient, by improving Methods (i), (iii)

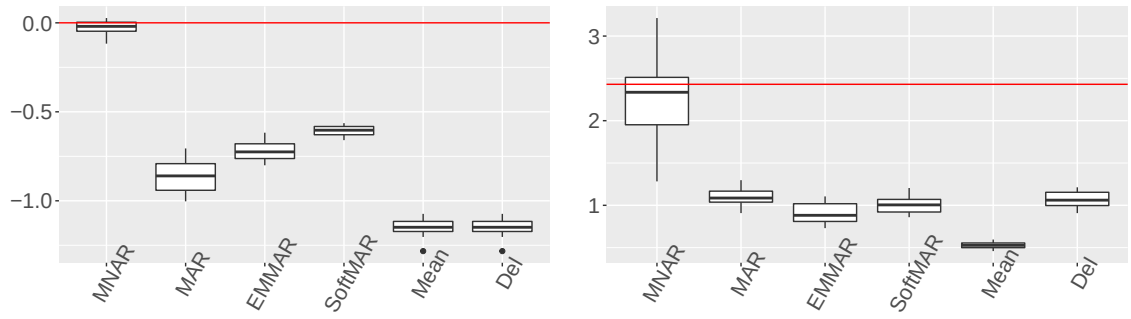


Figure C.7: Mean estimation (left) and variance estimation (right) of one missing variable when $r = 2$, $n = 1000$, $p = 50$ and twenty variables are MNAR. True values to be estimated are indicated by red lines.

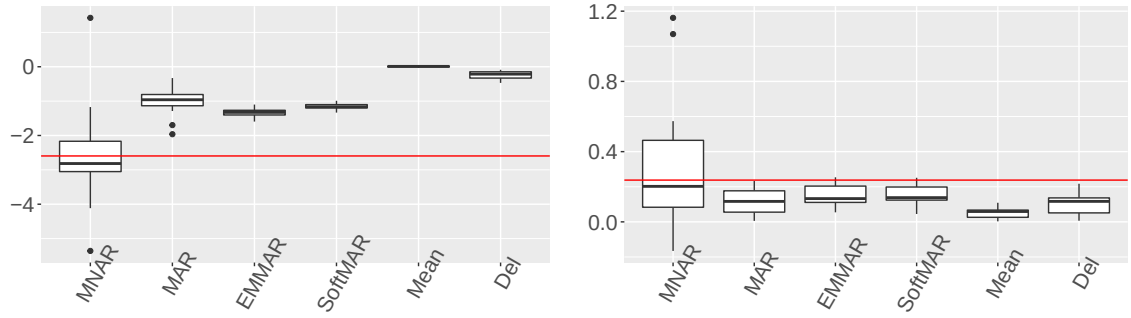


Figure C.8: Covariance estimation beetween two missing variable (left) and a missing variable and a pivot one (right) when $r = 2$, $n = 200$, $p = 10$ and seven variables are MNAR. True values to be estimated are indicated by red lines.

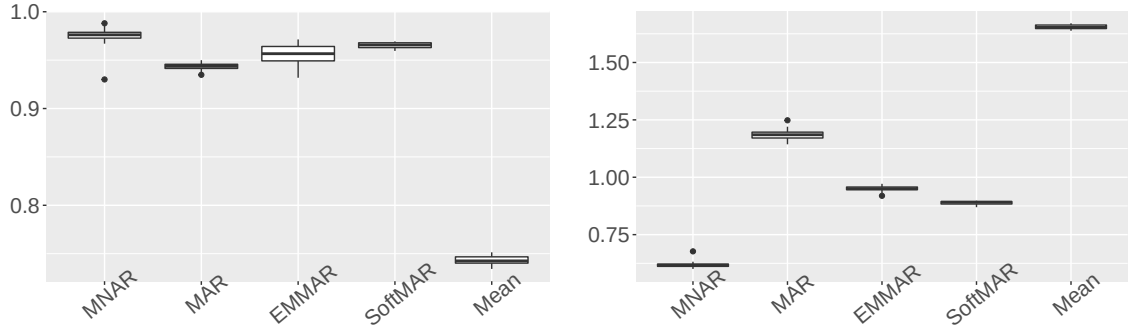


Figure C.9: RV coefficients for the loading matrix (left) and imputation error (right) when $r = 2$, $n = 1000$, $p = 50$ and twenty variables are MNAR.

and (ii). Concerning the imputation performance, Algorithm 1 strongly improves Methods (ii) and (iii).

For the same dimension setting ($n = 1000$, $p = 50$) and the same noise level ($\sigma = 1$), we vary the rank to $r = 5$. Similarly as before, missing values are introduced on twenty variables according to a self-masked MNAR mechanism, leading to 20% of missing values in total. In Figure C.10, for the mean and the variable estimations, Algorithm 1 gives unbiased estimators. In Figure C.11, the covariance between a missing variable and a pivot one estimated by Algorithm 1 is biased but still less than the other methods. In addition, the covariance between two missing variables is unbiased but suffers from a high variability. Note that once again we have chosen $\mathcal{J} \subset \mathcal{M}$ such that $|\mathcal{J}| = 10$. For the mean, 1260 combinations of the pivot variables are aggregated over 712530 possible combinations if $\mathcal{J} = \overline{\mathcal{M}}$. In Figure C.12, despite such results for the covariance estimators, Algorithm 1 gives a similar RV coefficient than Methods (ii) and (iii) but strongly improves all the methods in term of imputation error.

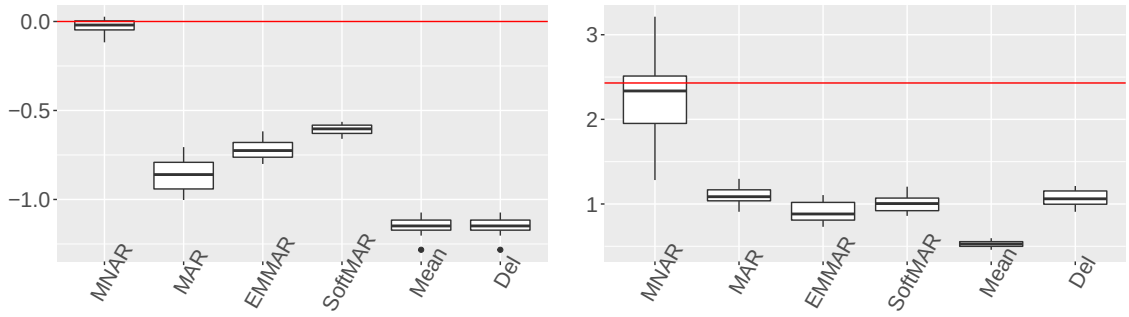


Figure C.10: Mean estimation (left) and variance estimation (right) of one missing variable when $r = 5$, $n = 1000$, $p = 50$ and twenty variables are MNAR. True values to be estimated are indicated by red lines.

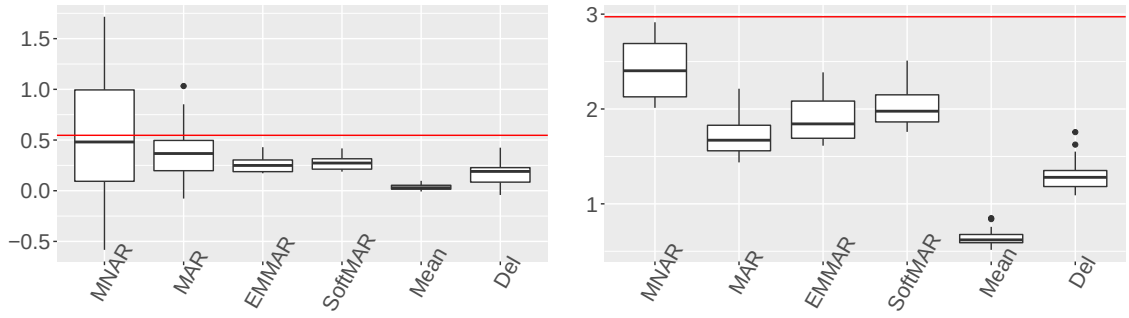


Figure C.11: Covariance estimation beetween two missing variable (left) and a missing variable and a pivot one (right) when $r = 5$, $n = 1000$, $p = 50$ and twenty variables are MNAR. True values to be estimated are indicated by red lines.

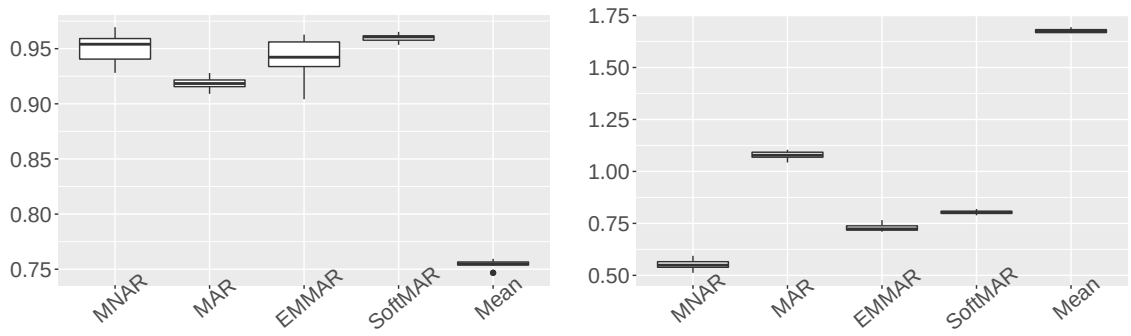


Figure C.12: RV coefficients for the loading matrix (left) and imputation error (right) when $r = 5$, $n = 1000$, $p = 50$ and twenty variables are MNAR.

Efficiency of the *aggregation* approach in the selection of the pivot variables. As described in Section 3.3.4, Algorithm 1 requires the selection of r pivot variables (considered M(C)AR) on which the regressions will be performed. To reduce the error committed by the selection pivot variables, we propose to select a bigger set of pivot variables (with a cardinal superior to r) and the final estimator will be computed with the median of the estimators over all possible combinations of r pivot variables (this is called the *aggregation* approach). In Figure C.13, we consider the same setting as in Section 3.4.1 ($n = 1000$, $p = 10$, $r = 2$ and seven self-masked MNAR variables) and we perform Algorithm 1 by using the *aggregation* (`MNARagg`) method or not (`MNARnoagg`). By discarding outliers, this *aggregation* approach is more robust than selecting only r pivot variables.

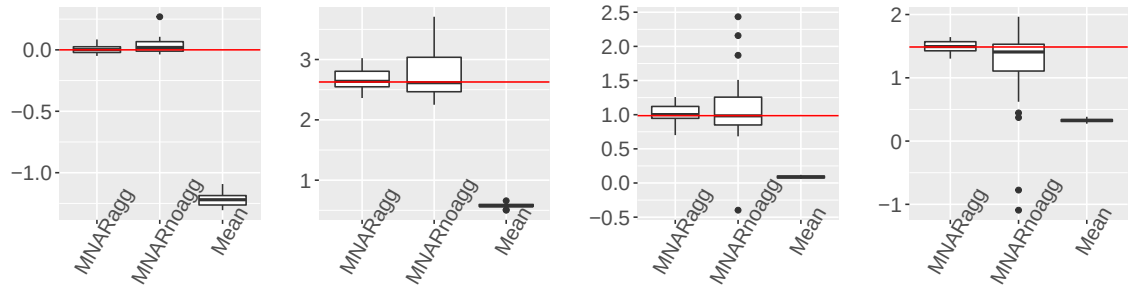


Figure C.13: Mean (left) and variance (middle left) estimations of $Y_{.1}$ and covariances estimations of $\mathrm{Cov}(Y_{.1}, Y_{.2})$ (between two missing variables) (middle right) and of $\mathrm{Cov}(Y_{.1}, Y_{.8})$ (between one missing variable and one pivot variable) (right). True values are indicated in red lines.

C.4 Computation time

Table C.1 gathers computation times of the different methods, for both settings considered in Sections 3.4 and C.3.

Method	$r = 2, p = 10, n = 1000$ 35% MNAR values in 7 variables	$r = 5, p = 50, n = 1000$ 20% MNAR values in 20 variables
MNAR algebraic	0,1 s	11 min 48 s (1260 aggregations)
SoftMAR	5,5 s	28 s
EMMAR	50,8 s	2 min 9 s
Param	5 h 15 min	not evaluated

Table C.1: Computation time for simulations in Sections 3.4 and Appendix C.3. The process time is obtained for a computer with a processor Intel Core i5 of 2,3 GHz.

C.5 Additional information on the Traumabase® dataset

C.5.1 Description of the variables

A description of the variables which are used in Section 3.4.2 is given. The indications given in parentheses ph (pre-hospital) and h (hospital) mean that the measures have been taken before the arrival at the hospital and at the hospital.

- *SBP.ph*, *DBP.ph*, *HR.ph*: systolic and diastolic arterial pressure and heart rate during pre-hospital phase. (ph)
- *HemoCue.init*: prehospital capillary hemoglobin concentration. (ph)
- *SpO2.min*: peripheral oxygen saturation, measured by pulse oxymetry, to estimate oxygen content in the blood. (ph)
- *Cristalloid.volume*: total amount of prehospital administered cristalloid fluid resuscitation (volume expansion). (ph)
- *Shock.index.ph*: ratio of heart rate and systolic arterial pressure during pre-hospital phase. (ph)
- *Delta.shock.index*: Difference of shock index between arrival at the hospital and arrival on the scene. (h)
- *Delta.hemoCue*: Difference of hemoglobin level between arrival at the hospital and arrival on the scene. (h)

The percentage of missing values in each variable is given in Figure C.14.

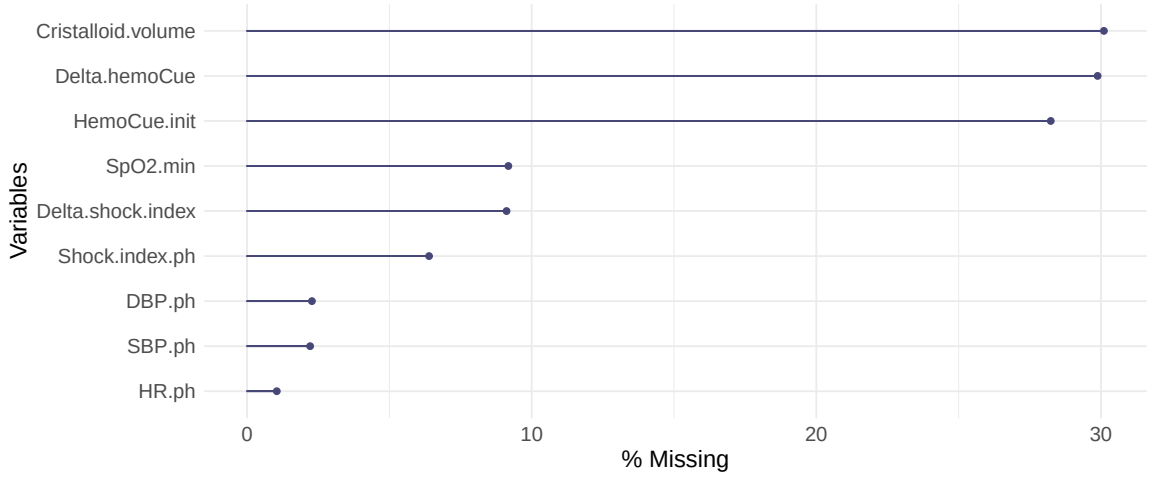


Figure C.14: Percentage of missing values in each variable for the TraumaBase data.

C.5.2 Supervised learning task

To predict the administration or not of the tranexomic acid (binary variable), we impute explanatory variables before proceeding to the classification task. In Table C.2, Algorithm 1 gives the smallest prediction error.

MNAR	5.06%
EMMAR	5.82%
SoftMAR	5.45%
MNARparam	5.39%
Mean	5.27%

Table C.2: Mean of prediction error over 10 repetitions.

C.6 Graphical approach

C.6.1 Preliminaries

Lemmas of Mohan et al. (2018) are used to construct some estimators of the mean, variance and covariances for a MNAR variable based on a graphical approach.

Lemma 9 (Lemma 2 (Mohan et al., 2018))**.** *Let us consider the m-graph G. The coefficient of the linear regression of $Y_{\cdot j}$ on $Y_{\cdot k}, k \neq j$, denoted as $\beta_{j\to k,k\neq j}$ is recoverable (i.e. they are consistent in the complete-case analysis) if $Y_{\cdot j} \perp\!\!\!\perp \Omega | Y_{\cdot k}, k \neq j$ and one has*

$$\beta_{j\to k,k\neq j} = \beta^c_{j\to k,k\neq j}.$$

Lemma 10 (Lemma 1). *(Mohan et al., 2018)](Graphical approach for computing the covariance) Let G be a m-graph with k unblocked paths $p_1, \dots, p_k$ between two variables $Y_{.\tau}$ and $Y_{.\delta}$. Let A_{p_i} be the ancestor of all nodes on path p_i. Let the number of nodes on p_i be n_{p_i}. One can derive that*

$$\mathrm{Cov}(Y_{.\tau}, Y_{.\delta}) = \sum_{i=1}^{k} \mathrm{Var}(A_{p_i}) \prod_{j=1}^{n_{p_i}-1} \alpha_j^{p_i},$$

where $\prod_{j=1}^{n_{p_i}-1} \alpha_j^{p_i}$ is the product of all causal parameters on path p_i.

In addition, let us recall the basic formula,

$$\beta_{Y \to X} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}, \tag{C.42}$$

where Y and X are two variables of a linear model.

C.6.2 Estimation of the mean, variance and covariances of the MNAR variables

The graphical approach to construct an estimator of α_1 is based on the transformation illustrated in Figure 3.1 of the graphical model of PPCA as structural causal graphs, whose context is introduced in (Pearl, 2003). This latter framework allows to directly apply the results of Mohan et al. (2018) who consider the associated (linear) structural causal equations under the exogeneity assumption with MNAR missing values for one variable.

For the sake of brevity, the results are presented for the toy example in Section 3.3.1 where $p = 3$, $r = 2$, $Y_{.1}$ is self-masked MNAR and the other variables are observed.

Then, one can associate to Figure 3.1 (bottom right graph) the structural equation model detailled in the following lemma.

Lemma 11. *Assuming $\mathbb{E}[\epsilon_{.2}|Y_{.1}, Y_{.3}] = 0$, the structural equation model associated with the bottom right graph in Figure 3.1 is*

$$Y_{.2} = \beta_{2\to 1,3[0]} + \beta_{2\to 1,3[1]} Y_{.1} + \beta_{2\to 1,3[3]} Y_{.3} + \epsilon_{.2}, \tag{C.43}$$

where $\beta_{2\to 1,3[0]}$, $\beta_{2\to 1,3[1]}$ and $\beta_{2\to 1,3[3]}$ are the intercept and the coefficients of the linear regression of $Y_{.2}$ on $Y_{.1}$ and $Y_{.3}$.

Using Equation (C.43) and Lemma 9, we apply the results of Mohan et al. (2018) to get an estimator for the mean of the MNAR variable.

Proposition 29 (Mean estimator for the graphical approach). *Under Equation (C.43), assuming **A1.** and $\beta^c_{2\to 1,3[1]} \neq 0$, one can construct an estimator of the mean α_1 of the MNAR variable $Y_{.1}$ as follows*

$$\hat{\alpha}_1 := \frac{\hat{\alpha}_2 - \hat{\beta}^c_{2\to 1,3[0]} - \hat{\beta}^c_{2\to 1,3[3]} \hat{\alpha}_3}{\hat{\beta}^c_{2\to 1,3[1]}}, \tag{C.44}$$

where $\hat{\beta}^c_{2\to 1,3[0]}$, $\hat{\beta}^c_{2\to 1,3[1]}$ *and* $\hat{\beta}^c_{2\to 1,3[3]}$ *denote some estimators of* $\beta^c_{2\to 1,3[0]}$, $\beta^c_{2\to 1,3[1]}$ *and* $\beta^c_{2\to 1,3[3]}$ *given in Lemma 11. This estimator is consistent under additional Assumption* ***A4.****.*

Proof. To derive some estimator of the mean, we want to obtain the following formula

$$\alpha_1 = \frac{\alpha_2 - \beta^c_{2\to 1,3[0]} - \beta^c_{2\to 1,3[3]}\alpha_3}{\beta^c_{2\to 1,3[1]}}. \tag{C.45}$$

Indeed, one has:

$$\begin{aligned}
\mathbb{E}[Y_{\cdot 2}] &= \mathbb{E}[\mathbb{E}[Y_{\cdot 2}|Y_{\cdot 1}, Y_{\cdot 3}] \\
&= \mathbb{E}[\mathbb{E}[Y_{\cdot 2}|Y_{\cdot 1}, Y_{\cdot 3}, \Omega_{\cdot 1} = 1]] && \text{(by using } \mathbf{A1.}\text{)} \\
&= \mathbb{E}[\mathbb{E}[\beta^c_{2\to 1,3[0]} + \beta^c_{2\to 1,3[1]}Y_{\cdot 1} + \beta^c_{2\to 3,1[3]}Y_{\cdot 3} + \epsilon_{\cdot 2}|Y_{\cdot 1}, Y_{\cdot 3}]] \\
&= \beta^c_{2\to 1,3[0]} + \beta^c_{2\to 1,3[1]}\mathbb{E}[Y_{\cdot 1}] + \beta^c_{2\to 3,1[3]}\mathbb{E}[Y_{\cdot 3}],
\end{aligned}$$

which leads to the desired Equation (C.45), provided that $\beta^c_{2\to 1,3[1]} \neq 0$. A natural estimator fo α_1 is then given by (C.44). It is consistent given that all the quantities involved are consistent, by using **A4.** (for the consistency of $\hat{\alpha}_2$ and $\hat{\alpha}_3$) and Lemma 9 (for the consistency of the coefficients $\hat{\beta}^c_{2\to 1,3[0]}$, $\hat{\beta}^c_{2\to 1,3[1]}$ and $\hat{\beta}^c_{2\to 1,3[3]}$). □

Remark 30 (Mean estimation: algebraic vs. graphical approach)**.** *In both approaches, the PPCA model is translated into a linear model. However, both estimators in Equations (3.9) and (C.44) theoretically differ. The exogeneity assumption and approximation is not made at the same step. In the algebraic approach, the results are first derived without using any approximation. It gives linear models that do not comply with the standard exogeneity assumption. Consequently, an approximation is done at the estimation step since the parameters* $\hat{\mathcal{B}}^c_{2\to 1,3[0]}$, $\hat{\mathcal{B}}^c_{2\to 1,3[1]}$ *and* $\hat{\mathcal{B}}^c_{2\to 1,3[3]}$ *are estimated with the standard linear regression coefficients. In the graphical approach, an approximation is made at the first step when a structural equation model is associated with the graphical model by assuming the exogeneity, i.e.* $\mathbb{E}[\epsilon_{\cdot 2}|Y_{\cdot 1}, Y_{\cdot 3}] = 0$*. In practice, for both approaches, the same coefficients are naturally computed, i.e.* $\hat{\beta}^c_{j\to k,\ell} = \hat{\mathcal{B}}^c_{j\to k,\ell}$*, which leads to the same computed estimators for the mean of* $Y_{\cdot 1}$*.*

While only one simplified graphical model between $Y_{\cdot 1}$, $Y_{\cdot 2}$ and $Y_{\cdot 3}$, displayed in the bottom right graph of Figure 3.1, was required to construct an estimator of the mean of $Y_{\cdot 1}$, the variance and covariance estimations rely on Equation (C.43) and the following one (associating to the bottom left graph of Figure 3.1),

$$Y_{\cdot 3} = \beta_{3\to 1,2[0]} + \beta_{3\to 1,2[1]}Y_{\cdot 1} + \beta_{3\to 1,2[2]}Y_{\cdot 2} + \epsilon_{\cdot 3}, \tag{C.46}$$

assuming $\mathbb{E}[\epsilon_{\cdot 3}|Y_{\cdot 1}, Y_{\cdot 2}] = 0$ and where $\beta_{3\to 1,2[0]}$, $\beta_{3\to 1,2[1]}$ and $\beta_{3\to 1,2[2]}$ are the intercept and the coefficients of the linear regression of $Y_{\cdot 3}$ on $Y_{\cdot 1}$ and $Y_{\cdot 2}$.

Using Equations (C.43) and (C.46) and Lemmas 9, 10, one can derive some estimators for the variance and the covariances of Y_1.

Proposition 31 (Variance and covariances formulae resulting from the graphical approach when $p = 3$ and $r = 2$)**.** *Under the two equations* (C.43) *and* (C.46)*, assuming* ***A1.*** *and also* $\beta^c_{3\to 1} \neq 0$, $\beta^c_{2\to 1,3[1]} \neq 0$ *and* $\mathrm{Var}(Y_{.3}) \neq 0$*, one can construct an estimator of the variance of the MNAR variable* $Y_{.1}$ *and its covariances as follows*

$$\widehat{\mathrm{Var}}(Y_{.1}) := \frac{\widehat{\mathrm{Var}}(Y_{.3})}{\hat{\beta}^c_{3\to 1}} \frac{1}{\hat{\beta}^c_{2\to 1,3[1]}} \left(\frac{\widehat{\mathrm{Cov}}(Y_{.2}, Y_{.3})}{\widehat{\mathrm{Var}}(Y_{.3})} - \hat{\beta}^c_{2\to 1,3[3]} \right), \tag{C.47}$$

$$\widehat{\mathrm{Cov}}(Y_{.1}, Y_{.2}) := \frac{1}{\hat{\beta}^c_{3\to 1,2[1]}} \left(\frac{\widehat{\mathrm{Cov}}(Y_{.2}, Y_{.3})}{\widehat{\mathrm{Var}}(Y_{.2})} - \hat{\beta}^c_{3\to 1,2[2]} \right) \widehat{\mathrm{Var}}(Y_{.2}), \tag{C.48}$$

$$\widehat{\mathrm{Cov}}(Y_{.1}, Y_{.3}) := \frac{1}{\hat{\beta}^c_{2\to 1,3[1]}} \left(\frac{\widehat{\mathrm{Cov}}(Y_{.2}, Y_{.3})}{\widehat{\mathrm{Var}}(Y_{.3})} - \hat{\beta}^c_{2\to 1,3[3]} \right) \widehat{\mathrm{Var}}(Y_{.3}), \tag{C.49}$$

where $\hat{\beta}^c_{3\to 1,2[1]}$, $\hat{\beta}^c_{3\to 1,2[2]}$ *and* $\hat{\beta}^c_{3\to 1}$ *are some estimators of* $\beta^c_{3\to 1,2[1]}$, $\beta^c_{3\to 1,2[2]}$ *and* $\beta^c_{3\to 1}$ *given in* (C.46)*.*

These estimators are consistent under additional Assumption ***A4.****.*

Proof. To derive some estimators of the variance and covariances of the MNAR variable $Y_{.1}$, one want to obtain the following formulae:

$$\mathrm{Var}(Y_{.1}) = \frac{\mathrm{Var}(Y_{.3})}{\beta^c_{3\to 1}} \frac{1}{\beta^c_{2\to 1,3[1]}} \left(\frac{\mathrm{Cov}(Y_{.2}, Y_{.3})}{\mathrm{Var}(Y_{.3})} - \beta^c_{2\to 1,3[3]} \right), \tag{C.50}$$

$$\mathrm{Cov}(Y_{.1}, Y_{.2}) = \frac{1}{\beta^c_{3\to 1,2[1]}} \left(\frac{\mathrm{Cov}(Y_{.2}, Y_{.3})}{\mathrm{Var}(Y_{.2})} - \beta^c_{3\to 1,2[2]} \right) \mathrm{Var}(Y_{.2}), \tag{C.51}$$

$$\mathrm{Cov}(Y_{.1}, Y_{.3}) = \frac{1}{\beta^c_{2\to 1,3[1]}} \left(\frac{\mathrm{Cov}(Y_{.2}, Y_{.3})}{\mathrm{Var}(Y_{.3})} - \beta^c_{2\to 1,3[3]} \right) \mathrm{Var}(Y_{.3}). \tag{C.52}$$

Using Equation (C.42), one has

$$\mathrm{Cov}(Y_{.1}, Y_{.3}) = \mathrm{Var}(Y_{.1})\beta_{3\to 1},$$

$$\mathrm{Cov}(Y_{.3}, Y_{.1}) = \mathrm{Var}(Y_{.3})\beta_{1\to 3},$$

so

$$\mathrm{Var}(Y_{.1}) = \frac{\mathrm{Var}(Y_{.3})\beta_{1\to 3}}{\beta_{3\to 1}}.$$

Considering the graphical model in the bottom left graph of Figure 3.1,

$$\begin{aligned} \mathrm{Cov}(Y_{.2}, Y_{.3}) &= \beta_{2\to 1,3[1]}\beta_{1\to 3}\mathrm{Var}(Y_{.3}) + \beta_{2\to 1,3[3]}\mathrm{Var}(Y_{.3}) \qquad \text{(by Lemma 10)} \\ \Rightarrow \beta_{1\to 3} &= \frac{1}{\beta_{2\to 1,3[1]}} \left(\frac{\mathrm{Cov}(Y_{.2}, Y_{.3})}{\mathrm{Var}(Y_{.3})} - \beta_{2\to 1,3[3]} \right) \\ \Rightarrow \beta_{1\to 3} &= \frac{1}{\beta^c_{2\to 1,3[1]}} \left(\frac{\mathrm{Cov}(Y_{.2}, Y_{.3})}{\mathrm{Var}(Y_{.3})} - \beta^c_{2\to 1,3[3]} \right) \end{aligned} \tag{C.53}$$

where the last implication is given by Lemma 9 and Assumption **A1.**, giving also

$$\beta_{3\to 1} = \beta^c_{3\to 1},$$

which leads to Equation (C.50).

By (C.42), the covariances can be expressed in two different ways,

$$\mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot 2}) = \beta_{2\to 1}\mathrm{Var}(Y_{\cdot 1}) \quad \text{and} \quad \mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot 3}) = \beta_{3\to 1}\mathrm{Var}(Y_{\cdot 1}), \tag{C.54}$$

$$\mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot 2}) = \beta_{1\to 2}\mathrm{Var}(Y_{\cdot 2}) \quad \text{and} \quad \mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot 3}) = \beta_{1\to 3}\mathrm{Var}(Y_{\cdot 3}). \tag{C.55}$$

In (C.54), the coefficients $\beta_{2\to 1}$ and $\beta_{3\to 1}$ can be estimated on the complete case using Lemma 9, but the variance of $Y_{\cdot 1}$ has still to be taken care of. Instead of potentially propagate error from (C.50), we propose to favor the expressions given in (C.55) to evaluate the covariances.

Focusing on (C.55), the coefficient $\beta_{1\to 3}$ is given in (C.53) and $\beta_{1\to 2}$ can be obtained using the same method, based on the reduced graphical model in the bottom right graph of Figure 3.1 (by Assumption **A1.**), so that

$$\beta_{1\to 2} = \frac{1}{\beta^c_{3\to 1,2[1]}}\left(\frac{\mathrm{Cov}(Y_{\cdot 2}, Y_{\cdot 3})}{\mathrm{Var}(Y_{\cdot 2})} - \beta^c_{3\to 1,2[2]}\right).$$

Therefore, by plugging it in (C.55), Equations (C.51) and (C.52) are obtained.

The natural estimators for $\mathrm{Var}(Y_{\cdot 1})$, $\mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot 2})$ and $\mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot 3})$ are then given by (C.47), (C.48) and (C.49). They are consistent given that all the quantites involved are consistent, by using **A4.** (for the consistency of $\widehat{\mathrm{Var}}(Y_{\cdot 2})$, $\widehat{\mathrm{Var}}(Y_{\cdot 3})$ and $\widehat{\mathrm{Cov}}(Y_{\cdot 2}, Y_{\cdot 3})$) and Lemma 9 (for the consistency of $\hat{\beta}^c_{j\to k,\ell}$). □

Remark 32 (Var-covariance estimation: algebraic vs. graphical approach)**.** *As for the mean, the exogeneity assumption is required in the last step of the algebraic approach to estimate coefficients and in the first step of the graphical approach to obtain structural equation models. However, contrary to the estimator suggested for the mean, the estimators in both graphical and algebraic approaches here differ (compare* (3.10) *with* (C.47)*,* (C.48) *and* (C.49)*). Indeed, the algebraic approach is based on the use of conditionality, while the graphical one relies on graphical results standing for the linear models when exogeneity holds.*

C.7 PPCA with MAR data

The following proposition is an adaptation of our method to handle MAR data, called **MAR** in Section 3.4.1, inspired by (Mohan et al., 2018, Theorems 1, 2, 3). In this case, the missing variables are assumed to be MAR indexed by $\mathcal{M}$. We assume the following:

A1$_{\mathbf{MAR}}$. $(B_{\cdot j'})_{j'\in\mathcal{J}}$ is invertible.

A2$_{\mathbf{MAR}}$. $\forall m \in \mathcal{M}$, $Y_{\cdot m} \perp\!\!\!\perp \Omega_{\cdot m} | (Y_k)_{k\in\overline{\{m\}}}$

A3$_{\mathbf{MAR}}$. $\forall m \in \mathcal{M}$, the complete-case coefficients $\mathcal{B}^c_{m\to\mathcal{J}[0]}$ and $\mathcal{B}^c_{m\to\mathcal{J}[k]}, k \in \mathcal{J}$ can be consistently estimated.

A5$_{\text{MAR}}$. $\forall \ell \in \overline{\mathcal{J}}$, for all set $\mathcal{H} \subset \mathcal{J}_{-j}$ such that $|\mathcal{H}| = r-1$, $\begin{pmatrix} B_{.\ell} & (B_{.j'})_{j' \in \mathcal{H}} \end{pmatrix}$ is invertible,

A6$_{\text{MAR}}$. $\forall m \in \mathcal{M}, \forall \ell \in \overline{\mathcal{J}} \backslash \mathcal{M}$, $\forall j \in \mathcal{J}$, $Y_{.m} \perp\!\!\!\perp \Omega_{.\ell} | (Y_{.k})_{k \in \overline{\{m\}}}$.

A8$_{\text{MAR}}$. $\forall m \in \mathcal{M}, \forall \ell \in \overline{\{m\}} \backslash \mathcal{J}$, for all set $\mathcal{H} \subset \mathcal{J}$ such that $|\mathcal{H}| = r-1$, the complete-case coefficients $\mathcal{B}^c_{m \to \ell, \mathcal{H}[0]}$ and $\mathcal{B}^c_{m \to \ell, \mathcal{H}[k]}, k \in \{\ell\} \cup \mathcal{H}$ can be consistently estimated.

Proposition 33 (Expectation, variance and covariances formulae for a MAR variable when $p = 3$ and $r = 2$)**.** *Consider the PPCA model (3.1). Under Assumptions **A1$_{\text{MAR}}$.** and **A2$_{\text{MAR}}$.**, one can construct the estimators of the mean, the variance and the covariances with a pivot variable for any MAR variable $Y_{.m}, m \in \mathcal{M}$, as follows*

– *the mean of the missing variable*

$$\hat{\alpha}_m = \hat{\mathcal{B}}^c_{m \to \mathcal{J}[0]} + \sum_{j \in \mathcal{J}} \hat{\mathcal{B}}^c_{m \to \mathcal{J}[j]} \hat{\alpha}_j,$$

with $\mathcal{J}$ the pivot variables set,

– *the variance of the missing variable*

$$\widehat{\mathrm{Var}}(Y_{.m}) = \hat{Q}^c_{\mathrm{MAR}} + \sum_{j \in \mathcal{J}} (\hat{\mathcal{B}}^c_{m \to \mathcal{J}[j]})^2 \widehat{\mathrm{Var}}(Y_{.j}) + 2 \sum_{(j<k) \in \mathcal{J}} \hat{\mathcal{B}}^c_{m \to \mathcal{J}[j]} \hat{\mathcal{B}}^c_{m \to \mathcal{J}[k]} \widehat{\mathrm{Cov}}(Y_{.j}, Y_{.k}),$$

with

$$\hat{Q}^c_{\mathrm{MAR}} = \left(\widehat{\mathrm{Var}}(Y_{.m}) \middle| \Omega_{.m} = 1 \right) - \left(\widehat{\mathrm{Cov}}((Y_{.j})_{j \in \overline{\{m\}}}, Y_{.m}) \widehat{\mathrm{Var}}((Y_{.j})_{j \in \overline{\{m\}}})^{-1} \widehat{\mathrm{Cov}}((Y_{.j})_{j \in \overline{\{m\}}}, Y_{.m})^T \middle| \Omega_{.m} = 1 \right).$$

– *the covariances between the missing variable and a pivot variable, for all $\ell \in \mathcal{J}$,*

$$\widehat{\mathrm{Cov}}(Y_{.m}, Y_{.\ell}) = \hat{\mathcal{B}}^c_{m \to \mathcal{J}[0]} \hat{\alpha}_\ell + \hat{\mathcal{B}}^c_{m \to \mathcal{J}[\ell]} (\widehat{\mathrm{Var}}(Y_{.\ell}) + \hat{\alpha}^2_\ell) + \sum_{k \in \mathcal{J}_{-\ell}} \hat{\mathcal{B}}^c_{m \to \mathcal{J}[k]} (\widehat{\mathrm{Cov}}(Y_{.\ell}, Y_{.k}) + \hat{\alpha}_\ell \hat{\alpha}_k) - \hat{\alpha}_m \hat{\alpha}_\ell$$

*Under Assumption **A3$_{\text{MAR}}$.** and **A4.**, these estimators are consistent.*

*In addition, under Assumption **A5$_{\text{MAR}}$.**, **A6$_{\text{MAR}}$.** and **A7.**, one can construct the estimator of the covariance between a MAR variable $Y_{.m}$ for $m \in \mathcal{M}$ and any not pivot variable as follows*

– *the covariances between the missing variable and any not pivot variable, for all* $\ell \in \overline{\{m\}}\backslash\mathcal{J}$*, choose* $r-1$ *variable indexes in* $\mathcal{J}$ *to form the set* $\mathcal{H} \cup \mathcal{J}$ *such that* $|\mathcal{H}| = r-1$

$$\widehat{\text{Cov}}(Y_{\cdot m}, Y_{\cdot \ell}) = \mathcal{B}^c_{m\to\ell,\mathcal{H}[0]}\hat{\alpha}_\ell + \hat{\mathcal{B}}^c_{m\to\ell,\mathcal{H}[\ell]}(\widehat{\text{Var}}(Y_{\cdot\ell}) + \hat{\alpha}^2_\ell) + \sum_{k\in\mathcal{H}} \hat{\mathcal{B}}^c_{m\to\ell,\mathcal{H}[k]}(\widehat{\text{Cov}}(Y_{\cdot\ell}, Y_{\cdot k}) + \hat{\alpha}_\ell\hat{\alpha}_k) - \hat{\alpha}_m\hat{\alpha}_\ell$$

Under the additional Assumptions $\boldsymbol{A8_{MAR}}$*.* *and* $\boldsymbol{A9}$*.* *this estimator is consistent.*

Proof. The proof follows exactly the same direction than in Proposition 11, 12 and 28. The only difference is that the regressions used are not the same.

For the sake of clarity, consider the same toy example as in Section 3.3.1 where $p = 3, r = 2$, in which only one variable can be missing (at random), and fix $\mathcal{M} = \{1\}$ and $\mathcal{J} = \{2, 3\}$. Note that here the MAR mechanism leads to $\mathbb{P}(\Omega_{\cdot 1} = 0|Y_{\cdot 1}, Y_{\cdot 2}, Y_{\cdot 3}) = \mathbb{P}(\Omega_{\cdot 1} = 0|Y_{\cdot 2}, Y_{\cdot 3})$.. The goal is to estimate the mean of $Y_{\cdot 1}$, without specifying the distribution of the missing-data mechanism and using only the observed data.

Assumption $\mathbf{A1_{MAR}}$**.** allows to obtain linear link between the MAR variable $Y_{\cdot 1}$ and the pivot variables $(Y_{\cdot 2}, Y_{\cdot 3})$. In particular, one has

$$Y_{\cdot 1} = \beta_{1\to 2,3[0]} + \beta_{1\to 2,3[2]}Y_{\cdot 2} + \beta_{1\to 2,3[3]}Y_{\cdot 3} + \zeta,$$

with $\beta_{1\to 2,3[0]}$, $\beta_{1\to 2,3[2]}$ and $\beta_{1\to 2,3[3]}$ the intercept and coefficients standing for the effects of $Y_{\cdot 1}$ on $Y_{\cdot 2}$ and $Y_{\cdot 3}$, and with

$$\zeta = -\mathcal{B}_{1\to 2,3[2]}\epsilon_{\cdot 2} - \mathcal{B}_{1\to 2,3[3]}\epsilon_{\cdot 3} + \epsilon_{\cdot 1}$$

Assumption $\mathbf{A2_{MAR}}$**.**, *i.e.* $Y_{\cdot 1} \perp\!\!\!\perp \Omega_{\cdot 1}|Y_{\cdot 2}, Y_{\cdot 3}$, is required to obtain identifiable and consistent parameters of the distribution of $Y_{\cdot 1}$ given $Y_{\cdot 2}, Y_{\cdot 3}$ in the complete-case when $\Omega_{\cdot 1} = 1$, denoted as $\beta^c_{1\to 2,3[0]}$, $\beta^c_{1\to 2,3[2]}$ and $\beta^c_{1\to 2,3[3]}$,

$$(Y_{\cdot 1})_{|\Omega_{\cdot 1}=1} = \beta^c_{1\to 2,3[0]} + \beta^c_{1\to 2,3[2]}Y_{\cdot 2} + \beta^c_{1\to 2,3[3]}Y_{\cdot 3} + \zeta^c,$$

with

$$\zeta^c = -\mathcal{B}^c_{1\to 2,3[2]}\epsilon_{\cdot 2} - -\mathcal{B}^c_{1\to 2,3[3]}\epsilon_{\cdot 3} + \epsilon_{\cdot 1}$$

(In the MNAR case, the regression of $Y_{\cdot 1}$ on $(Y_{\cdot 2}, Y_{\cdot 3})$ is prohibited, as $\mathbf{A2_{MAR}}$**.** does not hold. That is why we used the regression of $Y_{\cdot 2}$ on $Y_{\cdot 1}$ and $Y_{\cdot 3}$.);

Using again $\mathbf{A2_{MAR}}$**.**, one has

$$\mathbb{E}\left[Y_{\cdot 1}|Y_{\cdot 2}, Y_{\cdot 3}, \Omega_{\cdot 1} = 1\right] = \mathbb{E}\left[\beta^c_{1\to 2,3[0]} + \beta^c_{1\to 2,3[2]}Y_{\cdot 2} + \beta^c_{1\to 2,3[3]}Y_{\cdot 3}|Y_{\cdot 2}, Y_{\cdot 3}\right] + \mathbb{E}[\zeta^c|Y_{\cdot 2}, Y_{\cdot 3}],$$

and taking the expectation leads to

$$\mathbb{E}\left[Y_{\cdot 1}\right] = \beta^c_{1\to 2,3[0]} + \beta^c_{1\to 2,3[2]}\mathbb{E}\left[Y_{\cdot 2}\right] + \beta^c_{1\to 2,3[3]}\mathbb{E}\left[Y_{\cdot 3}\right],$$

given that $\mathbb{E}[\epsilon_{\cdot k}] = 0,\ \forall k \in \{1, 2, 3\}$.

One obtains

$$\alpha_1 = \beta^c_{1\to 2,3[0]} + \beta^c_{1\to 2,3[2]}\alpha_2 + \beta^c_{1\to 2,3[3]}\alpha_3$$

A natural estimator for α_1 is

$$\hat{\alpha}_1 = \hat{\beta}^c_{1\to 2,3[0]} + \hat{\beta}^c_{1\to 2,3[2]}\hat{\alpha}_2 + \hat{\beta}^c_{1\to 2,3[3]}\hat{\alpha}_3,$$

which is consistent using Assumption $\mathbf{A3_{MAR}}$. and **A4.**. □

Appendix D

Appendix of Chapter 4

D.1 Discussion on the paper of Ma and Needell (2018)

In this section, we make the theoretical issues unlocked in Ma and Needell (2018) explicit. For clarity, we directly refer to the lemmas and theorems as numbered in the published version (http://www.global-sci.org/uploads/online_news/NMTMA/201809051633-2442.pdf), the numbering being slightly different than the arXiv version. For readability, we translate their method and results with the notation used in the present paper. In their paper, they consider the finite-sample setting, with at hand $(D_{i:}, \tilde{X}_{i:})_{1 \leqslant i \leqslant n}$, in view of minimizing the empirical risk.

As a preamble, let us remind that the contributions of the present paper go far beyond correcting the approach in (Ma and Needell, 2018): we propose a different algorithm using averaging, that converges faster and in a non-strongly convex regime, with a different proof technique, requiring a more technical proof on the second order moment of the noise, and we allow for heterogeneity in the missing data mechanism.

D.1.1 Hurdles to get unbiased gradients of the empirical risk

The stochastic gradients in Ma and Needell (2018) are not unbiased gradients of the empirical risk (which makes their main result wrong). Indeed, their algorithm uses the debiased direction (4.4) by sampling uniformly with replacement the $(\tilde{X}_{k:})_k$'s.

For clarity, we highlight both why the result is not technically correct in their paper, and why it is not intuitively possible to achieve the result they give.

Technically. The proof of the main Theorem 2.2 (Theorem 2.1 being a direct corollary), corresponds to the classical proof in which one upper bounds the expectation of the mean-squared distance from the iterate at iteration $k+1$ to the optimal point **conditionally to the iterate at iteration** k, or more precisely, conditionally to a σ-algebra making this iterate measurable. This is typically written

$$\mathbb{E}\left[||\beta_{k+1} - \beta_*^n||^2 | \mathcal{F}_k\right],$$

where β_*^n is the minimizer of the empirical risk R_n and β_k is $\mathcal{F}_k$-measurable.

The crux of the proof is then to use **unbiased gradients conditionally to $\mathcal{F}_k$**: the property needed is that

$$\mathbb{E}\left[g_{i_{k+1}}(\beta_k)|\mathcal{F}_k\right] = \nabla R_n(\beta_k).$$

In classical ERM (without missing value) it is done by sampling uniformly at iteration $k+1$ one observation indexed by $i_{k+1} \sim \mathcal{U}[\![1; n]\!]$, **independently from** β_k.

In regression with missing data, one has to deal with another source of randomness, the randomness of the mask D. In Ma and Needell (2018), Lemma A.1 states that for a random $i \sim \mathcal{U}[\![1; n]\!]$ and a matrix row A_i, for a random mask D associated to this row,

$$\mathbb{E}_D[\mathbb{E}_i g_i(\beta)] = \nabla R_n(\beta).$$

This lemma is valid. Unfortunately, its usage in the proof of Theorem 2.2 (page 18, line (ii)), is not, as one *does not have*:

$$\mathbb{E}[g_{i_{k+1}}(\beta_k)|\mathcal{F}_k] = \nabla R_n(\beta_k),$$

indeed,

- either the sample i_{k+1} is chosen uniformly at random in $[\![1; n]\!]$ and $D_{i_{k+1}}$ **is not** independent from β_k.
- or the sample i is not chosen uniformly in $[\![1; n]\!]$ (for example without replacement, as we do) and then the gradient is not an unbiased gradient of R_n as the sampling is not uniform anymore.

In other words, the proof would only be valid if **the mask for the missing entries was re-sampled each time the point is used**, which is of course not realistic for a missing data approach (that would mean that the data has in fact been collected without missing entries).

Intuition on why it is hard. A way to understand the impossibility of having a bound for multiple pass on ERM in the context of missing data is to underline that the empirical risk, in the presence of missing data, is an **unknown function**: its value cannot be computed exactly (see Section 4.4.3).

As a consequence we can hardly expect that one could minimize it to unlimited accuracy. This is very similar to the situation for the *generalization risk* in a situation *without missing data*: as the function is not observed, it is impossible to minimize it exactly. Given only n observations, no algorithm can achieve 0-generalization error (and statistical lower bounds Tsybakov (2003) prove so).

Conclusion. This highlights how difficult it is to be rigorous when dealing with multiple sources of randomness. Unfortunately, none of these limits are discussed in the current version of (Ma and Needell, 2018).This makes the approach and the main theorem of (Ma

and Needell, 2018) mathematically invalid. In the present paper, the *generalization risk* is decaying *during the first pass*, and as a consequence, the empirical risk also probably does, but this has not been proved yet.

In the following paragraph, we give details on the missing technical Lemma.

D.1.2 Missing key Lemma in the proof.

Proving that $(\tilde{f}_k)$ is a.s. convex is an important step for convergence, which was missing in the analysis of Ma and Needell (2018). More precisely, in Lemma A.4. in Ma and Needell (2018), a condition is missing on $G(x)$: G needs to be smooth *and convex* for its gradient to satisfy the co-coercivity inequality. Note that this condition was also missing in the paper they refer to Needell et al. (2014) (Indeed, at the third line of the proof of Lemma A.1. in Needell et al. (2014), one needs f to be convex for G to be convex). Co-coercivity of the gradient is indeed a characterization of the fact that the function is smooth and convex, see for example Zhu and Marcotte (1996).

D.2 Proofs of technical lemmas

Recall that we aim at minimizing the theoretical risk in both streaming and finite-sample settings.

$$\beta^\star = \arg\min_{\beta\in\mathbb{R}^d} R(\beta) = \arg\min_{\beta\in\mathbb{R}^d} \mathbb{E}_{(X_{i:},y_i)}\left[f_i(\beta)\right]. \tag{D.2}$$

In the sequel, one consider the following modified gradient direction

$$\tilde{g}_k(\beta_k) = P^{-1}\tilde{X}_{k:}\left(\tilde{X}_{k:}^T P^{-1}\beta_k - y_k\right) - (\mathrm{I}-P)P^{-2}\mathrm{diag}\left(\tilde{X}_{k:}\tilde{X}_{k:}^T\right)\beta_k. \tag{D.4}$$

Note that for all k, $D_{k:}$ is independent from $(X_{k:}, y_k)$. In what follows, the proofs are derived considering

$$\mathbb{E} = \mathbb{E}_{(X_{k:},y_k),D_{k:}} = \mathbb{E}_{(X_{k:},y_k)}\mathbb{E}_{D_{k:}}$$

where $\mathbb{E}_{(X_{k:},y_k)}$ and $\mathbb{E}_{D_{k:}}$ denotes the expectation with respect to the distribution of $(X_{k:}, y_k)$ and $D_{k:}$ respectively.

D.2.1 Proof of Lemma 2

Lemma 12. *Let $(\mathcal{F}_k)_{k\geqslant 0}$ be the following σ-algebra,*

$$\mathcal{F}_k = \sigma(X_{1:}, y_1, D_{1:} \ldots, X_{k:}, y_k, D_{k:}).$$

The modified gradient $\tilde{g}_k(\beta_{k-1})$ in Equation (D.4) is $\mathcal{F}_k$-measurable and

$$\mathbb{E}\left[\tilde{g}_k(\beta_{k-1}) \,|\, \mathcal{F}_{k-1}\right] = \nabla R(\beta_{k-1}) \quad a.s.$$

Proof.

$$
\begin{aligned}
&\mathbb{E}_{(X_{k:},y_k),D_{k:}}\left[\tilde{g}_k(\beta_{k-1})|\mathcal{F}_{k-1}\right] \\
&\overset{(i)}{=} \mathbb{E}_{(X_{k:},y_k),D_{k:}}\left[P^{-1}\tilde{X}_{k:}\tilde{X}_{k:}^T P^{-1}\right]\beta_{k-1} - \mathbb{E}_{(X_{k:},y_k),D_{k:}}\left[P^{-1}\tilde{X}_{k:}y_k\right] \\
&\quad - \mathbb{E}_{(X_{k:},y_k),D_{k:}}\left[(\mathrm{I}-P)P^{-2}\mathrm{diag}\left(\tilde{X}_{k:}\tilde{X}_{k:}^T\right)\right]\beta_{k-1} \\
&\overset{(ii)}{=} \mathbb{E}_{(X_{k:},y_k)}\left[P^{-1}PX_{k:}X_{k:}^T PP^{-1}\beta_{k-1} + P^{-2}(P-P^2)\mathrm{diag}(X_{k:}X_{k:}^T)\beta_{k-1} - P^{-1}PX_{k:}y_k\right] \\
&\quad - \mathbb{E}_{(X_{k:},y_k)}\left[(\mathrm{I}-P)P^{-2}P\mathrm{diag}\left(X_{k:}X_{k:}^T\right)\beta_{k-1}\right] \\
&= \nabla R(\beta_{k-1}),
\end{aligned}
$$

In step (i), we use that β_{k-1} is $\mathcal{F}_{k-1}$-measurable and (X_k, y_k, D_k) is independent from $\mathcal{F}_{k-1}$. Step (ii) follows from

$$
\begin{cases}
\mathbb{E}_{D_{k:}}\left[\tilde{X}_{k:}\tilde{X}_{k:}^T\right] &= PX_{k:}X_{k:}^T P + (P-P^2)\mathrm{diag}(X_{k:}X_{k:}^T), \\
\mathbb{E}_{D_{k:}}\left[\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)\right] &= P\mathrm{diag}(X_{k:}X_{k:}^T), \\
\mathbb{E}_{D_{k:}}\left[\tilde{X}_{k:}\right] &= PX_{k:}.
\end{cases}
$$

□

D.2.2 Proof of Lemma 3

Lemma 13. *The additive noise process $(\tilde{g}_k(\beta^\star))_k$ with $\beta^\star$ defined in Equation (D.2) is $\mathcal{F}_k-$measurable and has the following properties:*

1. $\forall k \geqslant 0$, $\mathbb{E}[\tilde{g}_k(\beta^\star) \mid \mathcal{F}_{k-1}] = 0$ *a.s.,*
2. $\forall k \geqslant 0$, $\mathbb{E}[\|\tilde{g}_k(\beta^\star)\|^2 \mid \mathcal{F}_{k-1}]$ *is a.s. finite,*
3. $\forall k \geqslant 0$, $\mathbb{E}[\tilde{g}_k(\beta^\star)\tilde{g}_k(\beta^\star)^T] \preccurlyeq C(\beta^\star) = c(\beta^\star)H$, *where $\preccurlyeq$ denotes the order between self-adjoint operators ($A \preccurlyeq B$ if $B - A$ is positive semi-definite).*

Proof. 1 The first point is easily verified using Lemma 2 combined with $\nabla R(\beta^\star) = 0$ by (D.2).

2 Let us first remark that by independence $\mathbb{E}[\|\tilde{g}_k(\beta^\star)\|^2 \mid \mathcal{F}_{k-1}] = \mathbb{E}[\|\tilde{g}_k(\beta^\star)\|^2]$. Then,

$$
\mathbb{E}[\|\tilde{g}_k(\beta^\star)\|^2] \leqslant \frac{1}{p_m^2}\mathbb{E}\left[\|X_{k:}\|^2\left(\tilde{X}_{k:}^T P^{-1}\beta^\star - y_k\right)^2\right] + \frac{(1-p_m)^2}{p_m^2}\mathbb{E}\left[\|P^{-1}\mathrm{diag}\left(\tilde{X}_{k:}\tilde{X}_{k:}^T\right)\beta^\star\|^2\right].
$$

We decompose the computation with respect to $\mathbb{E}_{D_{k:}}$ first,

$$\begin{aligned}
\mathbb{E}_{D_{k:}}\left[\left(\tilde{X}_{k:}^T P^{-1}\beta^\star - y_k\right)^2\right] &= \mathbb{E}_{D_{k:}}\left[(\tilde{X}_{k:}^T P^{-1}\beta^\star)^2\right] - 2y_k\mathbb{E}_{D_{k:}}\left[\tilde{X}_{k:}^T P^{-1}\beta^\star\right] + y_k^2 \\
&= \mathbb{E}_{D_{k:}}\left[\left(\sum_{j=1}^d \tilde{X}_{kj}p_j^{-1}\beta_j^\star\right)^2\right] - 2y_k\mathbb{E}_{D_{k:}}\left[\sum_{j=1}^d \tilde{X}_{kj}p_j^{-1}\beta_j^\star\right] + y_k^2 \\
&= \sum_{j=1}^d \mathbb{E}_{D_{k:}}\left[\tilde{X}_{kj}^2 p_j^{-2}\beta_j^{\star 2}\right] + 2\sum_{l<j}\mathbb{E}_{D_{k:}}\left[\tilde{X}_{kj}\tilde{X}_{kl}p_j^{-1}p_l^{-1}\beta_j^\star\beta_l^\star\right] \\
&\qquad - 2y_k\sum_{j=1}^d X_{kj}\beta_j^\star + y_k^2 \\
&= \sum_{j=1}^d p_j^{-1}X_{kj}^2\beta_j^{\star 2} + 2\sum_{l<j}X_{kj}X_{kl}\beta_j^\star\beta_l^\star - 2y_k\sum_{j=1}^d X_{kj}\beta_j^\star + y_k^2 \\
&= (X_{k:}^T\beta^\star - y_k)^2 + \sum_{j=1}^d (p_j^{-1}-1)X_{kj}^2\beta_j^{\star 2},
\end{aligned}$$

which gives

$$\mathbb{E}_{D_{k:}}\left[\left(\tilde{X}_{k:}^T P^{-1}\beta^\star - y_k\right)^2\right] \leqslant (X_{k:}^T\beta^\star - y_k)^2 + \frac{1-p_m}{p_m}\beta^{\star T}\mathrm{diag}(X_{k:}X_{k:}^T)\beta^\star. \tag{D.8}$$

As for the second term,

$$\begin{aligned}
\mathbb{E}_{D_{k:}}\left[\|P^{-1}\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)\beta^\star\|^2\right] &= \mathbb{E}_{D_{k:}}\left[\sum_{j=1}^d \tilde{X}_{kj}^4 p_j^{-2}\beta_j^{\star 2}\right] \\
&= \sum_{j=1}^d X_{kj}^4 p_j^{-1}\beta_j^{\star 2} \\
&\leqslant \frac{1}{p_m}\sum_{j=1}^d X_{kj}^4\beta_j^{\star 2} \\
&\leqslant \frac{1}{p_m}\left(\sum_{j=1}^d X_{kj}^2\right)\left(\sum_{j=1}^d X_{kj}^2\beta_j^{\star 2}\right) \\
&= \frac{1}{p_m}\|X_{k:}\|^2\beta^{\star T}\mathrm{diag}(X_{k:}X_{k:}^T)\beta^\star
\end{aligned}$$

Finally, one obtains

$$\begin{aligned}
\mathbb{E}[\|\widetilde{g}_k(\beta^\star)\|^2 \mid \mathcal{F}_{k-1}] \leqslant{}& \frac{1}{p_m^2}\mathbb{E}_{(X_{k:},y_k)}\left[(\epsilon_k)^2\|X_{k:}\|^2\right] \\
&+ \frac{(1-p_m)+(1-p_m)^2}{p_m^3}\mathbb{E}_{(X_{k:},y_k)}\left[\|X_{k:}\|^2\beta^{\star T}\mathrm{diag}(X_{k:}X_{k:}^T)\beta^\star\right].
\end{aligned}$$

3 We aim at proving there exists H such that

$$\mathbb{E}[\tilde{g}_k(\beta^\star)\tilde{g}_k(\beta^\star)^T] \preccurlyeq C = cH.$$

Simple computations lead to:

$$\mathbb{E}[\tilde{g}_k(\beta^\star)\tilde{g}_k(\beta^\star)^T] = \mathbb{E}[T_1 + T_2 + T_2^T + T_3],$$

with:

$$\begin{aligned}
T_1 &= (\tilde{X}_{k:}^T P^{-1}\beta^\star - y_k)^2 P^{-1}\tilde{X}_{k:}\tilde{X}_{k:}^T P^{-1},\\
T_2 &= -(\tilde{X}_{k:}^T P^{-1}\beta^\star - y_k)P^{-1}\tilde{X}_{k:}\beta^{\star T}\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)P^{-2}(I-P),\\
T_3 &= (I-P)P^{-2}\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)\beta^\star\beta^{\star T}\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)P^{-2}(I-P).
\end{aligned}$$

Bound on T_1. For the first term, we use

$$P^{-1}\tilde{X}_{k:}\tilde{X}_{k:}^T P^{-1} \preccurlyeq \frac{1}{p_m^2}\tilde{X}_{k:}\tilde{X}_{k:}^T, \tag{D.9}$$

since for all vector $v \neq 0$, $v^T\left(\frac{1}{p_m^2}\tilde{X}_{k:}\tilde{X}_{k:}^T - P^{-1}\tilde{X}_{k:}\tilde{X}_{k:}^T P^{-1}\right)v \geqslant 0$,

$$\begin{aligned}
&\sum_{j=1}^d \left(\frac{1}{p_m^2} - \frac{1}{p_j^2}\right)\tilde{X}_{kj}^2 v_j^2 + 2\sum_{1\leqslant j<l\leqslant d}\left(\frac{1}{p_m^2} - \frac{1}{p_j p_l}\right)\tilde{X}_{kj}\tilde{X}_{kl}v_j v_l\\
&\overset{(iii)}{\geqslant} \sum_{j=1}^d \left(\frac{1}{p_m^2} - \frac{1}{p_j^2}\right)\tilde{X}_{kj}^2 v_j^2 + 2\sum_{1\leqslant j<l\leqslant d}\sqrt{\left(\frac{1}{p_m^2} - \frac{1}{p_j^2}\right)\left(\frac{1}{p_m^2} - \frac{1}{p_l^2}\right)}\tilde{X}_{kj}\tilde{X}_{kl}v_j v_l\\
&= \left(\sum_{j=1}^d \sqrt{\left(\frac{1}{p_m^2} - \frac{1}{p_j^2}\right)}\tilde{X}_{kj}v_j\right)^2 \geqslant 0.
\end{aligned}$$

Step (iii) uses $\left(\frac{1}{p_m^2} - \frac{1}{p_j p_l}\right) \geqslant \sqrt{\left(\frac{1}{p_m^2} - \frac{1}{p_j^2}\right)\left(\frac{1}{p_m^2} - \frac{1}{p_l^2}\right)}$. Indeed,

$$\begin{aligned}
\left(\frac{1}{p_m^2} - \frac{1}{p_j p_l}\right)^2 &\geqslant \left(\frac{1}{p_m^2} - \frac{1}{p_j^2}\right)\left(\frac{1}{p_m^2} - \frac{1}{p_l^2}\right)\\
&\Leftrightarrow \left(\frac{1}{p_m^4} - 2\frac{1}{p_j p_l}\frac{1}{p_m^2} + \frac{1}{p_j^2 p_l^2}\right) - \frac{1}{p_m^4} + \frac{1}{p_m^2 p_l^2} + \frac{1}{p_m^2 p_j^2} - \frac{1}{p_j^2 p_l^2} \geqslant 0\\
&\Leftrightarrow \left(\frac{1}{p_m p_j} - \frac{1}{p_m p_l}\right)^2 \geqslant 0.
\end{aligned}$$

Let us now prove that

$$\frac{1}{p_m^2}\tilde{X}_{k:}\tilde{X}_{k:}^T \preccurlyeq \frac{1}{p_m^2}X_{k:}X_{k:}^T$$

i.e.

$$\tilde{X}_{k:}\tilde{X}_{k:}^T \preccurlyeq X_{k:}X_{k:}^T. \tag{D.10}$$

Indeed, for all vector $v \neq 0$, $v^T(X_{k:}X_{k:}^T - \tilde{X}_{k:}\tilde{X}_{k:}^T)v \geqslant 0$:

$$\begin{aligned} v^T(X_{k:}X_{k:}^T - \tilde{X}_{k:}\tilde{X}_{k:}^T)v &= \sum_{j=1}^{d}(1-\delta_{kj}^2)X_{kj}^2v_j^2 + 2\sum_{1\leqslant j<l\leqslant d}(1-\delta_{kj}\delta_{kl})X_{kj}X_{kl}v_jv_l \\ &\overset{(iv)}{\geqslant} \sum_{j=1}^{d}(1-\delta_{kj}^2)X_{kj}^2v_j^2 + 2\sum_{1\leqslant j<l\leqslant d}\sqrt{(1-\delta_{kj}^2)(1-\delta_{kl}^2)}X_{kj}X_{kl}v_jv_l \\ &= \left(\sum_{j=1}^{d}\sqrt{(1-\delta_{kj}^2}X_{kj}v_j\right)^2 \geqslant 0 \end{aligned}$$

Step (iv) is obtained using $(1-\delta_{kj}\delta_{kl}) \geqslant \sqrt{(1-\delta_{kj}^2)(1-\delta_{kl}^2)}$. Indeed,

$$\begin{aligned} (1-\delta_{kj}\delta_{kl})^2 \geqslant (1-\delta_{kj}^2)(1-\delta_{kl}^2) &\Leftrightarrow (1-2\delta_{kl}\delta_{kj}+\delta_{kj}^2\delta_{kl}^2) - 1 + \delta_{kj}^2 - \delta_{kj}^2\delta_{kl}^2 + \delta_{kl}^2 \geqslant 0 \\ &\Leftrightarrow (\delta_{kj}-\delta_{kl})^2 \geqslant 0. \end{aligned}$$

Then, by (D.8) and $(X_{k:}^T\beta^\star - y_k)^2 = \epsilon_k^2$,

$$\mathbb{E}_{(X_{k:},y_k)}[T_1] = \mathbb{E}_{(X_{k:},y_k)}\left[\frac{1}{p_m^2}\epsilon_k^2X_{k:}X_{k:}^T\right] + \mathbb{E}_{(X_{k:},y_k)}\left[\frac{1-p_m}{p_m^3}\left(\beta^{\star T}\mathrm{diag}(X_{k:}X_{k:}^T)\beta^\star\right)X_{k:}X_{k:}^T\right].$$

Noting that

$$\|\mathrm{diag}(X_{k:})\beta^\star\|^2 \leqslant \|X_{k:}\|^2\|\beta^\star\|^2, \tag{D.11}$$

$$\mathbb{E}[T_1] \preccurlyeq \frac{1}{p_m^2}\mathrm{Var}(\epsilon_k)H + \frac{1-p_m}{p_m^3}\|X_{k:}\|^2\|\beta^\star\|^2H \tag{D.12}$$

Bound on T_3. Using the resulting matrix structure of

$$(I-P)P^{-2}\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)\beta^\star\beta^{\star T}\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)P^{-2}(I-P),$$

detailed as follows

$$\begin{pmatrix} (\beta_1^\star)^2\delta_{k1}^4X_{k1}^4 & \beta_1^\star\beta_2^\star\delta_{k1}^2\delta_{k2}^2X_{k1}^2X_{k2}^2 & \\ & \ddots & \\ & & (\beta_d^\star)^2\delta_{kd}^4X_{kd}^4 \end{pmatrix},$$

one obtains

$$\begin{aligned} \mathbb{E}_{D_{k:}}[T_3] = &\underbrace{(I-P)P^{-2}P\mathrm{diag}(X_{k:}X_{k:}{}^T)\beta^\star\beta^{\star T}\mathrm{diag}(X_{k:}X_{k:}{}^T)PP^{-2}(I-P)}_{=:T_{3a}} \\ &+ \underbrace{(I-P)P^{-2}(P-P^2)\mathrm{diag}(X_{k:}X_{k:}{}^T)\mathrm{diag}(\beta^\star\beta^{\star T})\mathrm{diag}(X_{k:}X_{k:}{}^T)P^{-2}(I-P)}_{=:T_{3b}}. \end{aligned} \tag{D.13}$$

Using similar arguments as in (D.9), both terms in (D.13) are bounded as follows

$$T_{3a} \leqslant \frac{(1-p_m)^2}{p_m^2}\mathrm{diag}(X_{k:}{X_{k:}}^T)\beta^\star\beta^{\star T}\mathrm{diag}(X_{k:}{X_{k:}}^T)$$
$$T_{3b} \leqslant \frac{(1-p_m)^3}{p_m^3}\mathrm{diag}(X_{k:}{X_{k:}}^T)\mathrm{diag}(\beta^\star\beta^{\star T})\mathrm{diag}(X_{k:}{X_{k:}}^T)$$

For T_{3a}, one can go further by using

$$\mathrm{diag}(X_{k:}X_{k:}^T)\beta^\star\beta^{\star T}\mathrm{diag}(X_{k:}X_{k:}^T) \leqslant \|\mathrm{diag}(X_{k:})\beta^\star\|^2 X_{k:}X_{k:}^T. \tag{D.14}$$

Let us prove that for all vector $v \neq 0$,

$$v^T(\|\mathrm{diag}(X_{k:})\beta^\star\|^2 X_{k:}X_{k:}^T - \mathrm{diag}(X_{k:}X_{k:}^T)\beta^\star\beta^{\star T}\mathrm{diag}(X_{k:}X_{k:}^T))v \geqslant 0, \text{ i.e.}$$
$$\underbrace{\sum_{j=1}^d \left(\left(\sum_{l=1}^d X_{il}^2\beta_{\star l}^2\right)X_{kj}^2 - X_{kj}^4\beta_j^{\star 2}\right)v_j^2 + 2\sum_{1\leqslant j<m\leqslant d}\left(\left(\sum_{l=1}^d X_{kl}^2\beta_l^{\star 2}\right)X_{kj}X_{km} - \beta_j^\star\beta_m^\star X_{kj}^2X_{km}^2\right)v_m v_j}_{=:Q} \geqslant 0$$

Indeed, $Q \geqslant \left(\sum_{j=1}^d \sqrt{\left(\sum_{l=1}^d X_{kl}^2\beta_l^{\star 2}\right)X_{kj}^2 - X_{kj}^4\beta_j^{\star 2}}v_j\right)^2 \geqslant 0$, since, looking at the term depending only on v_jv_m:

$$\left(\left(\sum_{l=1}^d X_{kl}^2\beta_l^{\star 2}\right)X_{kj}X_{km} - \beta_j^\star\beta_m^\star X_{kj}^2X_{km}^2\right)$$
$$\geqslant \sqrt{\left(\left(\sum_{l=1}^d X_{kl}^2\beta_l^{\star 2}\right)X_{kj}^2 - X_{kj}^4\beta_j^{\star 2}\right)\left(\left(\sum_{l=1}^d X_{kl}^2\beta_l^{\star 2}\right)X_{km}^2 - X_{km}^4\beta_m^{\star 2}\right)}$$

is equivalent to

$$\left(\sum_{l=1}^d X_{kl}^2\beta_l^{\star 2}\right)X_{kj}^4X_{km}^2\beta_j^{\star 2} + \left(\sum_{l=1}^d X_{kl}^2\beta_l^{\star 2}\right)X_{km}^4X_{kj}^2\beta_m^{\star 2} - 2\left(\sum_{l=1}^d X_{kl}^2\beta_l^{\star 2}\right)X_{kj}^3X_{km}^3\beta_j^\star\beta_m^\star \geqslant 0$$
$$\Leftrightarrow \left(\sqrt{\left(\sum_{l=1}^d X_{kl}^2\beta_l^{\star 2}\right)}X_{kj}^2X_{km}\beta_j^\star - \sqrt{\left(\sum_{l=1}^d X_{kl}^2\beta_l^{\star 2}\right)}X_{km}^2X_{kj}\beta_m^\star\right)^2 \geqslant 0$$

For T_{3b}, one can also dig deeper noting that

$$\mathrm{diag}(X_{k:}X_{k:}^T)\mathrm{diag}(\beta^\star\beta^{\star T})\mathrm{diag}(X_{k:}X_{k:}^T) \leqslant \|\mathrm{diag}(X_{k:})\beta^\star\|^2 X_{k:}X_{k:}^T. \tag{D.15}$$

For all vector $v \neq 0$, we aim at proving

$$v^T(\|\beta^{\star T}\mathrm{diag}(X_{k:})\|^2 X_{k:}X_{k:}^T - \mathrm{diag}(X_{k:}X_{k:}^T)\mathrm{diag}(\beta^\star\beta^{\star T})\mathrm{diag}(X_{k:}X_{k:}^T)))v \geqslant 0$$
$$\Leftrightarrow \underbrace{\sum_{j=1}^d \left(\left(\sum_{l=1}^d X_{kl}^2\beta_l^{\star 2}\right)X_{kj}^2 - X_{kj}^4\beta_j^{\star 2}\right)v_j^2 + 2\sum_{1\leqslant j<m\leqslant d}\left(\sum_{l=1}^d X_{kl}^2\beta_l^{\star 2}\right)X_{kj}X_{km}v_jv_m}_{=:Q'} \geqslant 0.$$

Indeed, $Q' \geqslant \left(\sum_{j=1}^d \sqrt{\left(\sum_{l=1}^d X_{kl}^2 \beta_l^{\star 2}\right) X_{kj}^2 - X_{kj}^4 \beta_j^{\star 2}} v_j\right)^2 \geqslant 0$ since

$$\begin{aligned}
&\left(\left(\sum_{l=1}^d X_{kl}^2 \beta_l^{\star 2}\right) X_{kj} X_{km}\right) \\
&\quad \geqslant \sqrt{\left(\left(\sum_{l=1}^d X_{kl}^2 \beta_l^{\star 2}\right) X_{kj}^2 - X_{kj}^4 \beta_j^{\star 2}\right)\left(\left(\sum_{l=1}^d X_{kl}^2 \beta_l^{\star 2}\right) X_{km}^2 - X_{km}^4 \beta_m^{\star 2}\right)} \\
&\iff \left(\sum_{l=1}^d X_{kl}^2 \beta_l^{\star 2}\right) X_{kj}^4 X_{km}^2 \beta_j^{\star 2} + \left(\sum_{l=1}^d X_{kl}^2 \beta_l^{\star 2}\right) X_{km}^4 X_{kj}^2 \beta_m^{\star 2} - X_{kj}^4 X_{km}^4 \beta_j^{\star 2} \beta_m^{\star 2} \geqslant 0
\end{aligned}$$

Combining (D.11), (D.14) and (D.15) lead to

$$\begin{aligned}
\mathbb{E}_{(X_{k:}, y_k)}\left[T_{3a}\right] &\leqslant \frac{(1-p_m)^2}{p_m^2} \|X_{k:}\|^2 \|\beta^\star\|^2 H \\
\mathbb{E}_{(X_{k:}, y_k)}\left[T_{3b}\right] &\leqslant \frac{(1-p_m)^3}{p_m^3} \|X_{k:}\|^2 \|\beta^\star\|^2 H
\end{aligned}$$

and to the final bound for T_3,

$$\mathbb{E}\left[T_3\right] \leqslant \frac{(1-p_m)^2}{p_m^2} \|X_{k:}\|^2 \|\beta^\star\|^2 H + \frac{(1-p_m)^3}{p_m^3} \|X_{k:}\|^2 \|\beta^\star\|^2 H. \tag{D.16}$$

Bound on $T_2 + T_2^T$. Firstly, focus on T_2:

$$\begin{aligned}
T_2 &= -(\tilde{X}_{k:}^T P^{-1} \beta^\star - y_k) P^{-1} \tilde{X}_{k:} \beta^{\star T} \mathrm{diag}(\tilde{X}_{k:} \tilde{X}_{k:}^T) P^{-2} (I - P) \\
&=: -(A - B),
\end{aligned}$$

where

$$\begin{aligned}
A &= P^{-1} \tilde{X}_{k:} \tilde{X}_{k:}^T P^{-1} \beta^\star \beta^{\star T} \mathrm{diag}(\tilde{X}_{k:} \tilde{X}_{k:}^T) P^{-2} (I - P) \\
B &= P^{-1} \tilde{X}_{k:} y_k \beta^{\star T} \mathrm{diag}(\tilde{X}_{k:} \tilde{X}_{k:}^T) P^{-2} (I - P).
\end{aligned}$$

Computation w.r.t. $\mathbb{E}_{D_{k:}}$. Term A can be split into three terms, denoting $\tilde{\mathbb{X}}_k := \tilde{X}_{k:} \tilde{X}_{k:}^T$

$$\begin{aligned}
A_1 &= P^{-1} \mathrm{diag}(\tilde{\mathbb{X}}_k) P^{-1} \beta^\star \beta^{\star T} \mathrm{diag}(\tilde{\mathbb{X}}_k) P^{-2} (I - P) \\
A_2 &= P^{-1} (\tilde{\mathbb{X}}_k - \mathrm{diag}(\tilde{\mathbb{X}}_k)) P^{-1} \mathrm{diag}(\beta^\star \beta^{\star T}) \mathrm{diag}(\tilde{\mathbb{X}}_k) P^{-2} (I - P) \\
A_3 &= P^{-1} (\tilde{\mathbb{X}}_k - \mathrm{diag}(\tilde{\mathbb{X}}_k)) P^{-1} (\beta^\star \beta^{\star T} - \mathrm{diag}(\beta^\star \beta^{\star T})) \mathrm{diag}(\tilde{\mathbb{X}}_k) P^{-2} (I - P).
\end{aligned}$$

Noting that

$$A_1 = P^{-2} \mathrm{diag}(\tilde{X}_{k:} \tilde{X}_{k:}^T) \beta^\star \beta^{\star T} \mathrm{diag}(\tilde{X}_{k:} \tilde{X}_{k:}^T) P^{-2} (I - P),$$

the expectation $\mathbb{E}_{D_{k:}}$ has already been computed in (D.13), so

$$\mathbb{E}_{D_{k:}}[A_1] = P^{-2}P\mathrm{diag}(X_{k:}{X_{k:}}^T)\beta^\star\beta^{\star T}\mathrm{diag}(X_{k:}{X_{k:}}^T)PP^{-2}(I-P) + P^{-2}(P-P^2)\mathrm{diag}(X_{k:}{X_{k:}}^T)\mathrm{diag}(\beta^\star\beta^{\star T})\mathrm{diag}(X_{k:}{X_{k:}}^T)P^{-2}(I-P). \quad \text{(D.17)}$$

As for A_2, making the structure of $P^{-1}(\tilde{X}_{k:}\tilde{X}_{k:}^T - \mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T))P^{-1}\mathrm{diag}(\beta^\star\beta^{\star T})\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)$ explicit,

$$A_2 = \begin{pmatrix} 0 & \frac{1}{p_1p_2}\delta_{k1}\delta_{k2}^3X_{k1}X_{k2}^3\beta_2^{\star 2} & \dots & \frac{1}{p_1p_d}\delta_{k1}\delta_{kd}^3X_{k1}X_{kd}^3\beta_d^{\star 2} \\ \frac{1}{p_1p_2}\delta_{k2}\delta_{k1}^3X_{k2}X_{k1}^3\beta_1^{\star 2} & 0 & & \\ & & \ddots & \\ \frac{1}{p_1p_d}\delta_{kd}\delta_{k1}^3X_{kd}X_{k1}^3\beta_1^{\star 2} & & & 0 \end{pmatrix},$$

one has

$$\mathbb{E}_{D_{k:}}[A_2] = (X_{k:}{X_{k:}}^T - \mathrm{diag}(X_{k:}{X_{k:}}^T))\mathrm{diag}(\beta^\star\beta^{\star T})\mathrm{diag}(X_{k:}{X_{k:}}^T)P^{-2}(I-P). \quad \text{(D.18)}$$

As for A_3, the term $P^{-1}(\tilde{X}_{k:}\tilde{X}_{k:}^T - \mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T))P^{-1}(\beta^\star\beta^{\star T} - \mathrm{diag}(\beta^\star\beta^{\star T}))\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)$ can be made explicit as

$$\begin{pmatrix} \sum_{l=2}^d \frac{1}{p_1p_l}\delta_{kl}X_{kl}\beta_l^\star\delta_{k1}X_{k1}^3\beta_1^\star & \sum_{l=3}^d \frac{1}{p_1p_l}\delta_{kl}X_{kl}\beta_l^\star\delta_{k1}\delta_{k2}^2X_{k1}X_{k2}^2\beta_2^\star & \dots & \sum_{l\neq 1,d} \frac{1}{p_1p_l}\delta_{kl}X_{kl}\beta_l^\star\delta_{k1}\delta_{kd}^2X_{k1}X_{kd}^2\beta_d^\star \\ & \ddots & & \\ & & \ddots & \\ & & & \sum_{l=1}^{d-1}\frac{1}{p_1p_d}\delta_{kl}X_{kl}\beta_l^\star\delta_{kd}X_{kd}^3\beta_d^\star \end{pmatrix}$$

which gives

$$\mathbb{E}_{D_{k:}}[A_3] = (X_{k:}X_{k:}^T - \mathrm{diag}(X_{k:}X_{k:}^T))(\beta^\star\beta^{\star T} - \mathrm{diag}(\beta^\star\beta^{\star T}))\mathrm{diag}(X_{k:}X_{k:}^T)PP^{-2}(I-P) + (I-P)\mathrm{diag}\left((X_{k:}X_{k:}^T - \mathrm{diag}(X_{k:}X_{k:}^T))(\beta^\star\beta^{\star T} - \mathrm{diag}(\beta^\star\beta^{\star T}))\mathrm{diag}(X_{k:}X_{k:}^T)\right)P^{-2}(I-P).$$

Noting the following,

$$\mathrm{diag}\left((\tilde{X}_{k:}\tilde{X}_{k:}^T - \mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T))(\beta^\star\beta^{\star T} - \mathrm{diag}(\beta^\star\beta^{\star T}))\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)\right) = \mathrm{diag}\left(\tilde{X}_{k:}\tilde{X}_{k:}^T\beta^\star\beta^{\star T}\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)\right) - \mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)\mathrm{diag}(\beta^\star\beta^{\star T})\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T),$$

one has

$$\begin{aligned}\mathbb{E}_{D_{k:}}[A_3] &= P^{-1}P(\tilde{X}_{k:}\tilde{X}_{k:}^T - \mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T))(\beta^\star\beta^{\star T} - \mathrm{diag}(\beta^\star\beta^{\star T}))\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)PP^{-2}(I-P) \\ &\quad + P^{-1}(P-P^2)\mathrm{diag}\left(\tilde{X}_{k:}\tilde{X}_{k:}^T\beta^\star\beta^{\star T}\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)\right)P^{-2}(I-P) \\ &\quad - P^{-1}(P-P^2)\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)\mathrm{diag}(\beta^\star\beta^{\star T})\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)P^{-2}(I-P) \quad \text{(D.19)}\end{aligned}$$

Term B can be made explicit as follows

$$P^{-1}\tilde{X}_{k:}\beta^{\star T}\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T) = \begin{pmatrix} \frac{1}{p_1}\beta_1^\star\delta_{i1}^3X_{i1}^3 & \frac{1}{p_1}\beta_1^\star\delta_{i1}^2\delta_{i2}X_{i1}^2X_{i2} \\ \frac{1}{p_2}\beta_2^\star\delta_{i2}^2X_{i2}^2\delta_{i1}X_{i1} & \frac{1}{p_2}\beta_2^\star\delta_{i2}^3X_{i2}^3 \\ & \ddots \end{pmatrix}$$

which implies

$$\begin{aligned}\mathbb{E}_{D_{k:}}[B] = y_kX_{k:}\beta^{\star T}\mathrm{diag}(X_{k:}X_{k:}{}^T)PP^{-2}(I-P) \\ + y_k(I-P)\mathrm{diag}(X_{k:}\beta^{\star T}\mathrm{diag}(X_{k:}X_{k:}{}^T))P^{-2}(I-P). \quad \text{(D.20)}\end{aligned}$$

Putting Equations (D.17), (D.18), (D.19) and (D.20) together,

$$\mathbb{E}\left[T_2 + T_2^T\right] = \mathbb{E}_{(X_{k:},y_k)}\left[T_{21} + T_{22} + T_{23} + T_{23}^T + T_{24} + T_{24}^T + T_{25}\right]$$

$$\begin{aligned}T_{21} &= -2(P^{-1}-I)\mathrm{diag}(X_{k:}X_{k:}{}^T)\beta^\star\beta^{\star T}\mathrm{diag}(X_{k:}X_{k:}{}^T)(P^{-1}-I)\\
T_{22} &= -2P^{-3}((I-P)(I-3P+2P^2)\mathrm{diag}(X_{k:}X_{k:}{}^T)\mathrm{diag}(\beta^\star\beta^{\star T})\mathrm{diag}(X_{k:}X_{k:}{}^T)\\
T_{23} &= -X_{k:}X_{k:}{}^T\mathrm{diag}(\beta^\star\beta^{\star T})\mathrm{diag}(X_{k:}X_{k:}{}^T)(P^{-2}(I-P)-P^{-1}(I-P))\\
T_{24} &= -(X_{k:}{}^T\beta^\star - y_k)X_{k:}\beta^{\star T}\mathrm{diag}(X_{k:}X_{k:}{}^T)P^{-1}(I-P)\\
T_{25} &= -2(X_{k:}{}^T\beta^\star - y_k)(I-P)\mathrm{diag}(X_{k:}\beta^{\star T}\mathrm{diag}(X_{k:}X_{k:}{}^T))P^{-2}(I-P),\end{aligned}$$

Computation w.r.t. $\mathbb{E}_{(X_{k:},y_k)}$. For T_{21}, it trivially holds that

$$-\mathrm{diag}(X_{k:}X_{k:}^T)\beta^\star\beta^{\star T}\mathrm{diag}(X_{k:}X_{k:}^T) \preccurlyeq 0. \quad \text{(D.21)}$$

Indeed, for all vector $v \neq 0$,

$$\sum_{j=1}^d X_{kj}^4\beta_j^{\star 2}v_j^2 + 2\sum_{1\leqslant j<m\leqslant d}\beta_j^\star\beta_m^\star X_{kj}^2X_{km}^2v_jv_m = \left(\sum_{j=1}^d X_{kj}^2\beta_j^\star v_j\right)^2 \geqslant 0.$$

Denoting the maximum of the coefficients of P as $p_M = \max_j p_j$, one has

$$\begin{aligned}T_{21} &\preccurlyeq -2\frac{(1-p_M)^2}{p_m^2}\mathrm{diag}(X_{k:}X_{k:}{}^T)\beta^\star\beta^{\star T}\mathrm{diag}(X_{k:}X_{k:}{}^T)\\
&\preccurlyeq 0 \qquad \text{(using (D.21)).}\end{aligned}$$

T_{22} is split into two terms,

$$\begin{aligned}T_{22a} &= -2P^{-3}((I-P)(I+2P^2))\mathrm{diag}(X_{k:}X_{k:}{}^T)\mathrm{diag}(\beta^\star\beta^{\star T})\mathrm{diag}(X_{k:}X_{k:}{}^T)\\
T_{22b} &= 6P^{-2}(I-P)\mathrm{diag}(X_{k:}X_{k:}{}^T)\mathrm{diag}(\beta^\star\beta^{\star T})\mathrm{diag}(X_{k:}X_{k:}{}^T)\end{aligned}$$

$$T_{22a} \preccurlyeq -2\frac{(1-p_M)(1+2p_M^2)}{p_m^3}\text{diag}(X_{k:}X_{k:}{}^T)\text{diag}(\beta^\star\beta^{\star T})\text{diag}(X_{k:}X_{k:}{}^T) \preccurlyeq 0,$$

since it is a diagonal matrix with only negative coefficients, and noting that $\frac{(1-p_M)(1+2p_M^2)}{p_m^3} > 0$. Then,

$$T_{22b} \preccurlyeq \frac{6(1-p_m)}{p_m^2}\text{diag}(X_{k:}X_{k:}{}^T)\text{diag}(\beta^\star\beta^{\star T})\text{diag}(X_{k:}X_{k:}{}^T)$$

which implies

$$\mathbb{E}_{(X_{k:},y_k)}\left[T_{22b}\right] \preccurlyeq \frac{6(1-p_m)}{p_m^2}\|X_{k:}\|^2\|\beta^\star\|^2 H$$

using (D.14) and (D.11).

As for $T_{23} + T_{23}^T$, note that

$$T_{23} + T_{23}^T \preccurlyeq -2\frac{(p_M-1)^2}{p_m^2}\big(X_{k:}X_{k:}{}^T\text{diag}(\beta^\star\beta^{\star T})\text{diag}(X_{k:}X_{k:}{}^T) + \text{diag}(\beta^\star\beta^{\star T})\text{diag}(X_{k:}X_{k:}{}^T)X_{k:}X_{k:}{}^T\big)$$

One prove that

$$-\big(X_{k:}X_{k:}{}^T\text{diag}(\beta^\star\beta^{\star T})\text{diag}(X_{k:}X_{k:}{}^T) + \text{diag}(X_{k:}X_{k:}{}^T)\text{diag}(\beta^\star\beta^{\star T})X_{k:}X_{k:}{}^T\big) \preccurlyeq -2\left(\min_{j=1,\dots,d}\beta_j^{\star 2}X_{kj}^2\right)X_{k:}X_{k:}^T \quad \text{(D.22)}$$

Indeed, denoting $m = \left(\min_{j=1,\dots,d}\beta_j^{\star 2}X_{kj}^2\right)$, one has

$$\begin{aligned}
&v^T\big(-2mX_{k:}X_{k:}^T + (X_{k:}X_{k:}{}^T\text{diag}(\beta^\star\beta^{\star T})\text{diag}(X_{k:}X_{k:}{}^T) \\
&\qquad + \text{diag}(X_{k:}X_{k:}{}^T)\text{diag}(\beta^\star\beta^{\star T})X_{k:}X_{k:}{}^T)\big)v \geqslant 0 \\
\Leftrightarrow &\sum_{j=1}^d\left(-2mX_{kj}^2 + 2X_{kj}^4\beta_j^{\star 2}\right)v_j^2 \\
&\qquad + 2\sum_{1\leqslant j<q\leqslant d}\left(-2mX_{kj}X_{kq} + X_{kj}^3X_{kq}\beta_j^{\star 2} + X_{kq}^3X_{kj}\beta_q^{\star 2}\right)v_jv_q \geqslant 0 \\
\Leftrightarrow &\sum_{j=1}^d\left(-2mX_{kj}^2 + 2X_{kj}^4\beta_j^{\star 2}\right)v_j^2 \\
&\qquad + 2\sum_{1\leqslant j<q\leqslant d}\sqrt{\left(-2mX_{kj}^2 + 2X_{kj}^4\beta_j^{\star 2}\right)\left(-2mX_{kq}^2 + 2X_{kq}^4\beta_q^{\star 2}\right)}v_jv_q \geqslant 0 \\
\Leftrightarrow &\left(\sum_{j=1}^d\sqrt{\left(-2mX_{kj}^2 + 2X_{kj}^4\beta_j^{\star 2}\right)}v_j\right)^2 \geqslant 0,
\end{aligned}$$

using that

$$\begin{aligned}&\left(-2mX_{kj}^2+2X_{kj}^4\beta_j^{\star 2}\right)\left(-2mX_{kq}^2+2X_{kq}^4\beta_q^{\star 2}\right)\\&\qquad\leqslant\left(-2mX_{kj}X_{kq}+X_{kj}^3X_{kq}\beta_j^{\star 2}+X_{kq}^3X_{kj}\beta_q^{\star 2}\right)^2\\&\Leftrightarrow\left(X_{kj}^3X_{kq}\beta_j^{\star 2}-X_{kq}^3X_{kj}\beta_q^{\star 2}\right)^2\geqslant 0\end{aligned}$$

Therefore

$$\mathbb{E}_{(X_{k:},y_k)}\left[T_{23}+T_{23}^T\right]\leqslant -2\frac{(p_M-1)^2}{p_m^2}\left(\min_{j=1,\dots,d}\beta_j^{\star 2}X_{kj}^2\right)H\leqslant 0,$$

since H is definite positive.

Finally one uses $(X_{k:}^T\beta^\star-y_k)=\epsilon_k$ to conclude by independence that $T_{24}=T_{25}=0$.

One gets

$$\mathbb{E}\left[T_2+T_2^T\right]\leqslant\frac{6(1-p_m)}{p_m^2}\|X_{k:}\|^2\|\beta^\star\|^2H. \tag{D.23}$$

Combining (D.12), (D.16) and (D.23) leads to the desired bound.

□

D.2.3 Proof of Lemma 4

Lemma 14. *For all $k\geqslant 0$, given the binary mask D, the adjusted gradient $\tilde{g}_k(\beta)$ is a.s. $L_{k,D}$-Lipschitz continuous, i.e. for all $u,v\in\mathbb{R}^d$,*

$$\|\tilde{g}_k(u)-\tilde{g}_k(v)\|\leqslant L_{k,D}\|u-v\|\ \ a.s..$$

Set

$$L:=\sup_{k,D}L_{k,D}\leqslant\frac{1}{p_m^2}\max_k\|X_{k:}\|^2\ \ a.s..$$

In addition, for all $k\geqslant 0$, $\tilde{g}_k(\beta)$ is almost surely co-coercive.

Proof. Note that

$$\begin{aligned}\|\tilde{g}_k(u)-\tilde{g}_k(v)\|&=\left\|\left(P^{-1}\tilde{X}_{k:}\tilde{X}_{k:}^TP^{-1}-(I-P)P^{-2}\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)\right)(u-v)\right\|\\&\leqslant\left\|\left(P^{-1}\tilde{X}_{k:}\tilde{X}_{k:}^TP^{-1}-(I-P)P^{-2}\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)\right)\right\|\|u-v\|\\&\leqslant\left\|\frac{1}{p_m^2}\left(\tilde{X}_{k:}\tilde{X}_{k:}^T-(1-p_m)\mathrm{diag}(\tilde{X}_{k:}\tilde{X}_{k:}^T)\right)\right\|\|u-v\|\\&\leqslant\frac{1}{p_m^2}\|\tilde{X}_{k:}\|^2\|u-v\|,\end{aligned}$$

where we have used the Weyl inequality in the last step.

One can thus choose $L_{k,D}=\frac{1}{p_m^2}\|\tilde{X}_{k:}\|^2$ and

$$L=\sup_{k,D}L_{k,D}\leqslant\frac{1}{p_m^2}\sup_k\|X_{k:}\|^2\leqslant\frac{1}{p_m^2}\max_k\|X_{k:}\|^2$$

Then, let us prove that the primitive of the adjusted gradient $\tilde{g}_k$ is convex. To do this, we check that the derivative of $\tilde{g}_k$ is definite positive:

$$\frac{\partial}{\partial\beta}\tilde{g}_k(\beta) = \frac{1}{p^2}\left(\tilde{X}_{k:}\tilde{X}_{k:}^T - (1-p)\mathrm{diag}\left(\tilde{X}_{k:}\tilde{X}_{k:}^T\right)\right)$$

since $\left(\tilde{X}_{k:}\tilde{X}_{k:}^T - (1-p)\mathrm{diag}\left(\tilde{X}_{k:}\tilde{X}_{k:}^T\right)\right)$ is positive semi-definite. Indeed,

$$\begin{aligned}
&v^T\left(\tilde{X}_{k:}\tilde{X}_{k:}^T - (1-p)\mathrm{diag}\left(\tilde{X}_{k:}\tilde{X}_{k:}^T\right)\right)v \geqslant 0\\
\Leftrightarrow &\sum_{j=1}^{d} p\tilde{X_{kj}}^2 v_j^2 + 2\sum_{1\leqslant j<l\leqslant d}\tilde{X}_{kj}\tilde{X}_{kl}v_jv_l \geqslant 0\\
\Leftrightarrow &\left(\sum_{j=1}^{d}\sqrt{p}\tilde{X}_{kj}v_j\right)^2 \geqslant 0,
\end{aligned}$$

using $p^2\left(\tilde{X}_{kj}\right)^2\left(\tilde{X}_{kj}\right)^2 \leqslant \left(\tilde{X}_{kj}\right)^2\left(\tilde{X}_{kl}\right)^2$ since $p \leqslant 1$. □

D.3 Proof of the theoretical convergence rate with estimated missing probabilities $(\hat{p}_j)_j$

In this section, we consider that we access $2n$ observations $(\tilde{X}'_{k:}, y'_k)_{1\leqslant k\leqslant n}$ and $(\tilde{X}_{k:}, y_k)_{1\leqslant k\leqslant n}$: we want to control the error of the estimator built with our Algorithm 2 using the second n observations with the proportions $\hat{p}$ *estimated* using the first n observations. In practice, it is likely that estimating the proportions on the same points used for running the algorithm would not hurt the performance. However, the proof requires the estimation of p and the stochastic gradient to be independent, we thus have to split the dataset. As we aim at proving that the convergence speed remains of $O(1/n)$, the induced multiplicative factor 2 on n will not modify the order of the convergence rate.

More precisely, we consider the proportions estimated on the points $(\tilde{X}'_{k:}, y'_k)_{1\leqslant k\leqslant n}$, for $1 \leqslant j \leqslant d$:

$$\hat{p}_j = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{X'_{kj}\neq \mathrm{NA}}. \tag{D.24}$$

Moreover, in the exceptional case that $\hat{p}_j = 0$ (which would correspond to a feature that is *never present* in the first half of the dataset, and thus would probably be discarded in practice), we correct the estimated proportion to n^{-1}. That is $\hat{p}_j = \max(n^{-1}, \frac{1}{n}\sum_{k=1}^{n}\mathbb{1}_{X'_{kj}\neq\mathrm{NA}})$. We do so only to ensure that $\hat{p}_j > 0$ which is necessary in the algorithm.

We then build the sequence $(\hat{\beta}_k)_{k\geqslant 0}$ of iterates constructed with an estimated value of the missing probabilities $\hat{p} = (\hat{p}_j)_{1\leqslant j\leqslant d} \in \mathbb{R}^d$ as follows:

$$\begin{cases} \hat{\beta}_0 = \beta_0 \\ \hat{\beta}_k = h_k(\hat{\beta}_{k-1}, \hat{p}) := \hat{\beta}_{k-1} - \alpha\tilde{g}_k(\hat{\beta}_{k-1}, \hat{p}),\ k > 0 \end{cases} \tag{D.25}$$

where

$$\tilde{g}_k(\beta_k, \hat{p}) := \hat{P}^{-1}\tilde{X}_{k:}\left(\tilde{X}_{k:}^T\hat{P}^{-1}\beta_k - y_k\right) - (\mathrm{I} - \hat{P})\hat{P}^{-2}\mathrm{diag}\left(\tilde{X}_{k:}\tilde{X}_{k:}^T\right)\beta_k$$

with $\hat{P} = \mathrm{diag}((\hat{p}_j)_{1\leqslant j\leqslant d})$.

We denote the averaged iterates of $(\hat{\beta}_k)_{k\geqslant 0}$ by $(\bar{\hat{\beta}}_k)_{k\geqslant 0}$ such that $\bar{\hat{\beta}}_k = \frac{1}{k+1}\sum_{\ell=0}^{k}\hat{\beta}_\ell$.

Theorem 1 (Convergence rate with estimated missing probabilities)**.** *Assume that the missing probabilities $(\hat{p}_j)_{j=1,\dots,d}$ are estimated as in Equation (D.24) using $(\tilde{X}'_{k:}, y'_k)_{1\leqslant k\leqslant n}$ and $\hat{\beta}_k$ given in Equation (D.25) is constructed using $(\tilde{X}_{k:}, y_k)_{1\leqslant k\leqslant n}$.*

There exists an event $A_n = \{\forall j \in \{1,\dots,d\}, \hat{p}_j > p_j/2\}$ with high probability $\mathbb{P}(A_n) \geqslant 1 - de^{-np_m/8}$. For any constant step-size $\alpha \leqslant \frac{1}{2L}$, Algorithm 2 ensures that, for any $n \geqslant 1$,

$$\mathbb{E}\left[\|\bar{\hat{\beta}}_n - \beta_\star\|_{H^{1/2}}^2 | A_n\right] \leqslant 2\underbrace{\mathbb{E}\left[\|\bar{\beta}_n - \beta_\star\|_{H^{1/2}}^2\right]}_{\text{Bounded by Theorem 17}} + \underbrace{\frac{2^6}{p_m^6}\frac{1}{\gamma\mu}CL\frac{5d}{n} + \frac{2^8}{p_m^6}\frac{1}{\gamma\mu}CL\frac{d(1-p_m)^{2n}}{n^2}}_{\text{Residual term due to the estimation of } \hat{p}},$$

where $p_m = \min_{j=1,\dots d} p_j$, $\gamma = \alpha\left(1 - \frac{\alpha L}{2}\right)$ and $C = \left(1 + \frac{1}{\gamma\mu}\right)\alpha^2 d C_{\mathrm{obs}}$, where L is given in Equation (4.6) and C_{obs} is such that $\mathbb{E}[\|\tilde{X}_{k:}\|^4(|\tilde{X}_{k:}^T\hat{\beta}_{k-1}| + |y_k|)^4] \leqslant {C_{\mathrm{obs}}}^2, \forall k \geqslant 0$, and where we denote $\|v\|_{H^{1/2}}^2 = \|H^{1/2}v\|_2^2$ and μ the smallest eigenvalue of $H = \mathbb{E}_{(X_{k:},y_k)}[X_{k:}X_{k:}^T]$ which is assumed to be positive.

In addition, one has the following unconditional result, for any $n \geqslant 1$,

$$\begin{aligned} R(\bar{\hat{\beta}}_n) - R(\beta_\star) &= \frac{1}{2}\mathbb{E}\left[\|\bar{\hat{\beta}}_n - \beta_\star\|_{H^{1/2}}^2\right] \\ &\leqslant \underbrace{\mathbb{E}\left[\|\bar{\beta}_n - \beta_\star\|_{H^{1/2}}^2\right]}_{\text{Bounded by Theorem 17}} + \underbrace{\frac{1}{\gamma\mu}CL\left(\frac{2^4}{p_m^6}\frac{5d}{n} + \frac{2^6}{p_m^6}\frac{d(1-p_m)^{2n}}{n^2} + n^6\sqrt{d}e^{-np_m/16}\right)}_{\text{Residual term due to the estimation of } \hat{p}}. \end{aligned}$$

Proof. The probability of A_n is given by Lemma 16. Let us first remark

$$\begin{aligned} \mathbb{E}\left[\|\bar{\hat{\beta}}_n - \beta^\star\|_{H^{1/2}}^2\right] &= \mathbb{E}\left[\|\bar{\hat{\beta}}_n - \bar{\beta}_n + \bar{\beta}_n - \beta_\star\|_{H^{1/2}}^2\right] \\ &\leqslant 2\left(\mathbb{E}\left[\|\bar{\hat{\beta}}_n - \bar{\beta}_n\|_{H^{1/2}}^2\right] + \mathbb{E}\left[\|\bar{\beta}_n - \beta_\star\|_{H^{1/2}}^2\right]\right). \end{aligned}$$

Note that for the first expectation in the last inequality, the randomness comes from the estimated proportions $(\hat{p}_j)_j$ and from the samples $(\tilde{X}_{k:}, y_k)_{1\leqslant k\leqslant n}$, whereas for the second expectation, the randomness is only due to $(\tilde{X}_{k:}, y_k)_{1\leqslant k\leqslant n}$. We then combine Theorem 17 and Lemma 17:

- Theorem 17 (and more precisely Remark 18) gives the bound for $\mathbb{E}\left[\|\bar{\beta}_n - \beta_\star\|_{H^{1/2}}^2\right]$. Note that for the conditional result, $\mathbb{E}\left[\|\bar{\beta}_n - \beta_\star\|_{H^{1/2}}^2 | A_n\right] = \mathbb{E}\left[\|\bar{\beta}_n - \beta_\star\|_{H^{1/2}}^2\right]$, because $\bar{\beta}_n \perp\!\!\!\perp A_n$.

- One has

$$\begin{aligned}\mathbb{E}\left[\|\bar{\hat{\beta}}_n - \bar{\beta}_n\|^2_{H^{1/2}}\right] &\leqslant L\mathbb{E}\left[\|\bar{\hat{\beta}}_n - \bar{\beta}_n\|^2\right] \\ &= L\mathbb{E}\left[\|(n+1)^{-1}\sum_{k=0}^{n}\hat{\beta}_k - \beta_k\|^2\right] \\ &\leqslant L(n+1)^{-1}\sum_{k=0}^{n}\mathbb{E}\left[\|\hat{\beta}_k - \beta_k\|^2\right]\end{aligned}$$

  The result follows by using Lemma 17.

□

We first prove the following key Lemma, that will be the main element in the proof of Lemma 17.

Lemma 15. *For all $k \geqslant 0$, one has*

$$\mathbb{E}\left[\|\hat{\beta}_k - \beta_k\|^2\right] \leqslant \frac{1}{\gamma\mu}C\left(\mathbb{E}\left[\frac{\|\hat{p}-p\|^4}{\min_{j=1,\dots,d}(p_j,\hat{p}_j)^{12}}\Big|\hat{p}\right]\right)^{1/2},$$

with $\gamma = \alpha\left(1-\frac{\alpha L}{2}\right)$ and $C = \left(1+\frac{1}{\gamma\mu}\right)\alpha^2 d C_{\text{obs}}$, where L is given in Equation (4.6) and C_{obs} is such that $\left(\mathbb{E}[\|\tilde{X}_{k:}\|^4(|\tilde{X}_{k:}^T\hat{\beta}_{k-1}| + |y_k|)^4]\right)^{1/2} \leqslant C_{\text{obs}}$, for all $k \geqslant 0$.

Proof. Let us denote $\delta_k^2 := \|\hat{\beta}_k - \beta_k\|^2 = \|h_k(\hat{\beta}_{k-1},\hat{p}) - h_k(\beta_{k-1},p)\|^2$. We first remark that

$$\mathbb{E}[\delta_k^2] = \mathbb{E}[\mathbb{E}[\delta_k^2|\hat{p}]],$$

so that we bound the conditional expectation $\mathbb{E}[\delta_k^2|\hat{p}]$. In the following, to control the deviation of $\hat{\beta}_k$ to β_k, we use

1. the deviation resulting from the use of $\hat{p}$ instead of p to construct $\hat{\beta}_k$ (term 1 in (D.26)),
2. the deviation resulting from the use of $\hat{\beta}_{k-1}$ as a support point instead of β_{k-1} (term 2 in (D.26)).

To do so, we introduce a "ghost" sequence (never computed) $h_k(\hat{\beta}_{k-1},p)$. Noting that for any $\eta > 0$, and any $a, b \in \mathbb{R}^d$, we have $||a+b||^2 \leqslant (1+\eta)||a||^2 + (1+\eta^{-1})||b||^2$, we have:

$$\begin{aligned}\mathbb{E}[\delta_k^2|\hat{p}] &= \mathbb{E}\left[\|h_k(\hat{\beta}_{k-1},\hat{p}) - h_k(\hat{\beta}_{k-1},p) + h_k(\hat{\beta}_{k-1},p) - h_k(\beta_{k-1},p)\|^2|\hat{p}\right] \\ &\leqslant (1+\frac{1}{\eta})\underbrace{\mathbb{E}\left[\|h_k(\hat{\beta}_{k-1},\hat{p}) - h_k(\hat{\beta}_{k-1},p)\|^2|\hat{p}\right]}_{\text{term 1}} + (1+\eta)\underbrace{\mathbb{E}\left[\|h_k(\hat{\beta}_{k-1},p) - h_k(\beta_{k-1},p)\|^2|\hat{p}\right]}_{\text{term 2}}.\end{aligned} \tag{D.26}$$

We control both terms separately.

Control of term 1. Almost surely, we have the following

$$\begin{aligned}\|h_k(\hat{\beta}_{k-1},\hat{p}) - h_k(\hat{\beta}_{k-1},p)\|^2 &= \|\hat{\beta}_{k-1} - \alpha\tilde{g}_k(\hat{\beta}_{k-1},\hat{p}) - \hat{\beta}_{k-1} + \alpha\tilde{g}_k(\hat{\beta}_{k-1},p)\|^2 \\ &= \alpha^2\|\tilde{g}_k(\hat{\beta}_{k-1},p) - \tilde{g}_k(\hat{\beta}_{k-1},\hat{p})\|^2 \\ &= \alpha^2\sum_{j=1}^{d}(\tilde{g}_{kj}(\hat{\beta}_{k-1},p) - \tilde{g}_{kj}(\hat{\beta}_{k-1},\hat{p}))^2,\end{aligned}$$

where $\tilde{g}_{kj}(\hat{\beta}_{k-1},p)$ denotes the j-th component of the vector $\tilde{g}_k(\hat{\beta}_{k-1},p) \in \mathbb{R}^d$. We introduce the function $\psi_{kj} : [0,1] \to \mathbb{R}$ such that $\psi_{kj}(t) = \tilde{g}_{kj}(\hat{\beta}_{k-1}, p + t(\hat{p}-p))$. By the mean value theorem, one has

$$\begin{aligned}\|h_k(\hat{\beta}_{k-1},\hat{p}) - h_k(\hat{\beta}_{k-1},p)\|^2 &= \alpha^2\sum_{j=1}^{d}(\psi_{kj}(1) - \psi_{kj}(0))^2 \\ &\leqslant \alpha^2\sum_{j=1}^{d}\sup_{t\in[0,1]}(\psi'_{kj}(t))^2 \qquad \text{(D.27)}\end{aligned}$$

Yet, $\psi'_{kj}(t) = \left\langle \nabla\tilde{g}_{kj}(\hat{\beta}_{k-1}, p + t(\hat{p}-p)), \hat{p}-p \right\rangle$. Using the Cauchy–Schwarz inequality and denoting $p_t = p + t(\hat{p}-p)$, one obtains

$$\alpha^2\sum_{i=1}^{d}\sup_{t\in[0,1]}(\psi'_{kj}(t))^2 \leqslant \alpha^2\sum_{j=1}^{d}\sup_{t\in[0,1]}\|\nabla\tilde{g}_{kj}(\hat{\beta}_{k-1},p_t)\|^2\|\hat{p}-p\|^2 \qquad \text{(D.28)}$$

Recall that $p_t = ((p_t)_1,\dots,(p_t)_d)^T \in \mathbb{R}^d$. Using the form of the debiased gradient given in Remark 16, one has for $1 \leqslant j \leqslant d$:

$$\tilde{g}_{kj}(\hat{\beta}_{k-1},p_t) = \underbrace{\left(\begin{pmatrix} \frac{1}{(p_t)_1(p_t)_j} & \cdots & \underbrace{\frac{1}{(p_t)_j}}_{\text{jth position}} & \cdots & \frac{1}{(p_t)_d(p_t)_j}\end{pmatrix} \odot \tilde{X}_{kj}\tilde{X}_{k:}^T\right)\hat{\beta}_{k-1}}_{\text{denoted } \tilde{g}^1_{kj}(\hat{\beta}_{k-1},p_t)} + \underbrace{\begin{pmatrix}\frac{1}{(p_t)_1} & \cdots \frac{1}{(p_t)_d}\end{pmatrix} \odot \tilde{X}_{kj}y_k}_{\text{denoted } \tilde{g}^2_{kj}(\hat{\beta}_{k-1},p_t)}$$

One has $\|\nabla\tilde{g}_{kj}(\hat{\beta}_{k-1},p_t)\|^2 = \sum_{\ell=1}^{d}\left(\frac{\partial\tilde{g}_{kj}}{\partial((p_t)_\ell)}(\hat{\beta}_{k-1},p_t)\right)^2$ with

$$\frac{\partial\tilde{g}^1_{kj}}{\partial((p_t)_\ell)}(\hat{\beta}_{k-1},p_t) = \begin{cases} \left(\frac{-1}{(p_t)_j^2}\begin{pmatrix}\frac{1}{(p_t)_1} & \cdots & \underbrace{1}_{j\text{th position}} & \cdots & \frac{1}{(p_t)_d}\end{pmatrix} \odot \tilde{X}_{kj}\tilde{X}_{k:}^T\right)\hat{\beta}_{k-1} & \text{if } \ell = j \\ \left(\frac{1}{(p_t)_j}\begin{pmatrix}0 & \cdots & \underbrace{\frac{-1}{(p_t)_\ell^2}}_{\ell\text{th position}} & \cdots & 0\end{pmatrix} \odot \tilde{X}_{kj}\tilde{X}_{k:}^T\right)\hat{\beta}_{k-1} & \text{otherwise}\end{cases}$$

and

$$\frac{\partial \tilde{g}_{kj}^2}{\partial((p_t)_\ell)}(\hat{\beta}_{k-1}, p_t) = \begin{pmatrix} 0 & \dots & \underbrace{\frac{-1}{(p_t)_\ell^2}}_{\ell\text{th position}} & \dots & 0 \end{pmatrix} \odot \tilde{X}_{kj} y_k$$

Therefore, $\forall \ell \in \{0, \dots, d\}$,

$$\left(\frac{\partial \tilde{g}_{kj}}{\partial((p_t)_\ell)}(\hat{\beta}_{k-1}, p_t)\right)^2 \leqslant \frac{1}{(p_t)_{\min}^6}(\tilde{X}_{kj}(|\tilde{X}_{k:}^T \hat{\beta}_{k-1}| + |y_k|))^2$$

with $(p_t)_{\min} = \min_{j=1,\dots,d}(p_t)_j$ which leads to

$$\|\nabla \tilde{g}_{kj}(\hat{\beta}_{k-1}, p_t)\|^2 = \sum_{\ell=1}^{d}\left(\frac{\partial \tilde{g}_{kj}}{\partial((p_t)_\ell)}(\hat{\beta}_{k-1}, p_t)\right)^2 \leqslant \frac{1}{(p_t)_{\min}^6}\|\tilde{X}_{k:}\|^2(|\tilde{X}_{k:}^T \hat{\beta}_{k-1}| + |y_k|)^2.$$

One obtains, plugging the equation above into Equations (D.27) and (D.28):

$$\|h_k(\hat{\beta}_{k-1}, \hat{p}) - h_k(\hat{\beta}_{k-1}, p)\|^2 \leqslant \alpha^2 d \sup_{t\in[0,1]} \frac{1}{(p_t)_{\min}^6}\|\tilde{X}_{k:}\|^2(|\tilde{X}_{k:}^T \hat{\beta}_{k-1}| + |y_k|)^2\|\hat{p} - p\|^2$$

and finally, taking expectation conditionally to $\hat{p}$:

$$\mathbb{E}[\|h_k(\hat{\beta}_{k-1}, \hat{p}) - h_k(\hat{\beta}_{k-1}, p)\|^2|\hat{p}] \leqslant \alpha^2 d\mathbb{E}\left[\|\tilde{X}_{k:}\|^2(\tilde{X}_{k:}^T \hat{\beta}_{k-1} + y_k)^2 \sup_{t\in[0,1]} \frac{\|\hat{p} - p\|^2}{(p_t)_{\min}^6}\Big|\hat{p}\right]$$

Assuming that $\left(\mathbb{E}[\|\tilde{X}_{k:}\|^4(|\tilde{X}_{k:}^T \hat{\beta}_{k-1}| + |y_k|)^4]\right)^{1/2} \leqslant C_{\text{obs}}$, one has, by Cauchy Schwartz,

$$\begin{aligned}\mathbb{E}[\|h_k(\hat{\beta}_{k-1}, \hat{p}) - h_k(\hat{\beta}_{k-1}, p)\|^2|\hat{p}] &\leqslant \alpha^2 d C_{\text{obs}}\mathbb{E}^{1/2}\left[\sup_{t\in[0,1]} \frac{\|\hat{p} - p\|^4}{(p_t)_{\min}^{12}}\Big|\hat{p}\right] \\ &\leqslant \alpha^2 d C_{\text{obs}}\mathbb{E}^{1/2}\left[\frac{\|\hat{p} - p\|^4}{\min_{j=1,\dots,d}(p_j, \hat{p}_j)^{12}}\Big|\hat{p}\right]. \qquad \text{(D.29)}\end{aligned}$$

Control of term 2. We now control the part of the distance coming form the fact that the true-iterate $h_k(\beta_{k-1}, p)$ and ghost-iterate $h_k(\hat{\beta}_{k-1}, p)$ updates are computed at two different points $\hat{\beta}_{k-1}$ and β_{k-1}:

$$\begin{aligned}\|h_k(\hat{\beta}_{k-1}, p) - h_k(\beta_{k-1}, p)\|^2 &= \|\hat{\beta}_{k-1} - \beta_{k-1} - \alpha(\tilde{g}_k(\hat{\beta}_{k-1}, \hat{p}) - \tilde{g}_k(\beta_{k-1}, p))\|^2 \\ &\leqslant \|\hat{\beta}_{k-1} - \beta_{k-1}\|^2 - 2\alpha\left\langle \hat{\beta}_{k-1} - \beta_{k-1}, \tilde{g}_k(\hat{\beta}_{k-1}, \hat{p}) - \tilde{g}_k(\beta_{k-1}, p)\right\rangle \\ &\quad + \alpha^2\|\tilde{g}_k(\hat{\beta}_{k-1}, \hat{p}) - \tilde{g}_k(\beta_{k-1}, p)\|^2\end{aligned}$$

Using Lemma 4 which gives the co-coercivity of the debiased gradient, one obtains

$$\alpha^2\|\tilde{g}_k(\hat{\beta}_{k-1}, p) - \tilde{g}_k(\beta_{k-1}, p)\|^2 \leqslant \alpha^2 L\left\langle \hat{\beta}_{k-1} - \beta_{k-1}, \tilde{g}_k(\hat{\beta}_{k-1}, p) - \tilde{g}_k(\beta_{k-1}, p)\right\rangle.$$

It implies that

$$\|h_k(\hat{\beta}_{k-1}, p) - h_k(\beta_{k-1}, p)\|^2 \leqslant \|\hat{\beta}_{k-1} - \beta_{k-1}\|^2 - 2\alpha\left(1 - \frac{\alpha L}{2}\right)\left\langle \hat{\beta}_{k-1} - \beta_{k-1}, \tilde{g}_k(\hat{\beta}_{k-1}, p) - \tilde{g}_k(\beta_{k-1}, p)\right\rangle$$

Denoting $\gamma = \alpha\left(1 - \frac{\alpha L}{2}\right)$, one has

$$\begin{aligned}\mathbb{E}\left[\|h_k(\hat{\beta}_{k-1}, p) - h_k(\beta_{k-1}, p)\|^2 | \hat{p}\right] \leqslant & \mathbb{E}\left[\|\hat{\beta}_{k-1} - \beta_{k-1}\|^2 | \hat{p}\right] \\ & - 2\gamma\mathbb{E}\left[\left\langle \hat{\beta}_{k-1} - \beta_{k-1}, \tilde{g}_k(\hat{\beta}_{k-1}, p) - \tilde{g}_k(\beta_{k-1}, p)\right\rangle | \hat{p}\right] \\ \leqslant & \mathbb{E}\left[\|\hat{\beta}_{k-1} - \beta_{k-1}\|^2 | \hat{p}\right] \\ & - 2\gamma\left\langle \mathbb{E}\left[\hat{\beta}_{k-1} - \beta_{k-1}\right], \nabla R(\hat{\beta}_{k-1}) - \nabla R(\beta_{k-1})\right\rangle,\end{aligned}$$

using that $\mathbb{E}[\tilde{g}_k(\hat{\beta}_{k-1}, p)|\hat{p}] = \nabla R(\hat{\beta}_{k-1})$ because $\hat{\beta}_{k-1}$ is constructed with n observations independent of the ones using for computing $\hat{p}$ (it implies $\tilde{g}_k(\cdot, p) \perp\!\!\!\perp \hat{p}$).

Using the μ-strong convexity of R, one has.

$$\mathbb{E}\left[\|h_k(\hat{\beta}_{k-1}, p) - h_k(\beta_{k-1}, p)\|^2 | \hat{p}\right] \leqslant (1 - 2\gamma\mu)\mathbb{E}\left[\|\hat{\beta}_{k-1} - \beta_{k-1}\|^2 | \hat{p}\right] \tag{D.30}$$

Conclusion. Choosing $\eta = \gamma\mu$ in Equation (D.26) with Equation (D.29) and Equation (D.30) leads to:

$$\mathbb{E}[\delta_k^2] \leqslant \left(1 + \frac{1}{\gamma\mu}\right)\alpha^2 d C_{\text{obs}} \mathbb{E}^{1/2}\left[\frac{\|\hat{p} - p\|^4}{\min_{j=1,\dots,d}(p_j, \hat{p}_j)^{12}} \Big| \hat{p}\right] + (1 - \gamma\mu)\mathbb{E}[\delta_{k-1}^2 | \hat{p}].$$

Denoting $C := \left(1 + \frac{1}{\gamma\mu}\right)\alpha^2 d C_{\text{obs}}$,

$$\begin{aligned}\mathbb{E}[\delta_{k+1}^2 | \hat{p}] &\leqslant (1 - \gamma\mu)^k \mathbb{E}[\delta_0^2 | \hat{p}] + C\sum_{i=0}^{k}(1 - \gamma\mu)^{k-i}\mathbb{E}^{1/2}\left[\frac{\|\hat{p} - p\|^4}{\min_{j=1,\dots,d}(p_j, \hat{p}_j)^{12}} \Big| \hat{p}\right] \\ &= (1 - \gamma\mu)^k \mathbb{E}[\delta_0^2] + C\frac{1 - (1 - \gamma\mu)^k}{\gamma\mu}\mathbb{E}^{1/2}\left[\frac{\|\hat{p} - p\|^4}{\min_{j=1,\dots,d}(p_j, \hat{p}_j)^{12}} \Big| \hat{p}\right] \\ &\leqslant \frac{1}{\gamma\mu} C \mathbb{E}^{1/2}\left[\frac{\|\hat{p} - p\|^4}{\min_{j=1,\dots,d}(p_j, \hat{p}_j)^{12}} \Big| \hat{p}\right], \end{aligned} \tag{D.31}$$

where in the last inequality we used that $\mathbb{E}[\delta_0^2 | \hat{p}] = 0$.

□

Lemma 16. *Let $A_n = \{\forall j \in \{1, \dots, d\}, \hat{p}_j > p_j/2\}$ be the event where the missing probabilities are not under-estimated by a factor of at least two. The probability of this event is such that*

$$\mathbb{P}(A_n) \geqslant 1 - d e^{-n p_m/8},$$

where $p_m = \min_{j=1,\dots d}\ p_j$.

Proof. We use the multiplicative Chernoff-Hoeffding inequality: if $X_1, \dots, X_n$ are i.i.d. variables such that $\mathbb{E}[\sum_{i=1}^n X_i] = \mathbb{E}[X] = \mu$, one has

$$\mathbb{P}(X \leqslant (1-\delta)\mu) \leqslant e^{-\delta^2\mu/2},\ 0 \leqslant \delta \leqslant 1.$$

Fix $j \in \{1, \dots, d\}$. Choosing $\delta = 1/2$ and applying the Chernoff-Hoeffding inequality to $n\hat{p}_j = \sum_{i=1}^n \delta_{ij}$ with $\delta_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{B}(p_j)$, one has

$$\mathbb{P}(n\hat{p}_j \leqslant np_j/2) \leqslant e^{-np_j/8}$$

implying that $\mathbb{P}(\hat{p}_j \leqslant p_j/2) \leqslant e^{-np_j/8}$.

Finally,

$$\begin{aligned}
\mathbb{P}(A_n) &= 1 - \mathbb{P}(\exists j \in \{1, \dots, d\}, \hat{p}_j \leqslant p_j/2) \\
&\geqslant 1 - \sum_{j=1}^d \mathbb{P}(\hat{p}_j \leqslant p_j/2) \\
&\geqslant 1 - \sum_{j=1}^d e^{-np_j/8} \geqslant 1 - de^{-np_m/8}.
\end{aligned}$$

□

Lemma 17. *Let $A_n = \{\forall j \in \{1, \dots, d\}, \hat{p}_j > p_j/2\}$. For any $k \geqslant 0$,*

$$\mathbb{E}\left[\|\hat{\beta}_k - \beta_k\|^2 | A_n\right] \leqslant \frac{2^5}{p_m^6}\frac{1}{\gamma\mu}C\frac{d}{n} + \frac{2^7}{p_m^6}\frac{1}{\gamma\mu}C\frac{d(1-p_m)^{2n}}{n^2},$$

with $p_m = \min_{j=1,\dots d}\ p_j$, $\gamma = \alpha\left(1 - \frac{\alpha L}{2}\right)$ and $C = \left(1 + \frac{1}{\gamma\mu}\right)\alpha^2 d C_{\text{obs}}$, where L is given in Equation (4.6) and C_{obs} is such that $\left(\mathbb{E}[\|\tilde{X}_{k:}\|^4(|\tilde{X}_{k:}^T\hat{\beta}_{k-1}| + |y_k|)^4]\right)^{1/2} \leqslant C_{\text{obs}}, \forall k \geqslant 0$.

In addition, $\forall k \geqslant 0$,

$$\mathbb{E}\left[\|\hat{\beta}_k - \beta_k\|^2\right] \leqslant \frac{1}{\gamma\mu}C\left(\frac{2^4}{p_m^6}\frac{5d}{n} + \frac{2^6}{p_m^6}\frac{d(1-p_m)^{2n}}{n^2} + n^6\sqrt{d}e^{-np_m/16}\right).$$

Proof. We start by proving the result conditional to the event A_n. We recall that

$$\mathbb{E}[\|\hat{\beta}_k - \beta_k\|^2 | A_n] = \frac{\mathbb{E}[\|\hat{\beta}_k - \beta_k\|^2 \mathbb{1}_{A_n}]}{\mathbb{P}(A_n)}$$

Using Lemma 16, one has $\mathbb{P}(A_n) \geqslant 1/2$. Thus,

$$\mathbb{E}[\|\hat{\beta}_k - \beta_k\|^2 | A_n] \leqslant 2\mathbb{E}[\|\hat{\beta}_k - \beta_k\|^2 \mathbb{1}_{A_n}].$$

By Lemma 15, it leads to

$$\mathbb{E}[\|\hat{\beta}_k - \beta_k\|^2 | A_n] \leqslant \frac{2}{\gamma\mu}C\left(\mathbb{E}\left[\frac{\|\hat{p} - p\|^4}{\min_{j=1,\dots,d}(p_j, \hat{p}_j)^{12}}\Big|\hat{p}\right]\right)^{1/2}$$

Yet, almost surely $\frac{\mathbb{1}_{A_n}}{\min_{j=1,\dots,d}(p_j,\hat{p}_j)^6} \leqslant \frac{2^6}{p_m^6}$ given that on the event A_n, $\forall j = 1,\dots,d$, $\hat{p}_j > p_j/2$. We thus have

$$\mathbb{E}[\|\hat{\beta}_k - \beta_k\|^2 | A_n] \leqslant \frac{2^7}{p_m^6}\frac{1}{\gamma\mu} C \mathbb{E}^{1/2}\left[\|\hat{p}-p\|^4\right]$$

Moreover, $\mathbb{E}\left[\|\hat{p}-p\|^4\right] = \mathbb{E}(\sum_{j=1}^d (\hat{p}_j - p_j)^2)^2 = (\sum_{j=1}^d \mathbb{E}(\hat{p}_j - p_j)^4) + \sum_{j,j'=1, j\neq j'}^d (\mathbb{E}(\hat{p}_j - p_j)^2)(\mathbb{E}(\hat{p}_{j'} - p_{j'})^2)$, by independence of the estimation of each coordinate.

We thus have to compute the 4-th order moment and the quadratic error of $\hat{p}_j$.

First, $\mathbb{E}(\hat{p}_j - p_j)^2) = \mathrm{Var}(\hat{p}_j) + \mathrm{Bias}(\hat{p}_j)^2$, with $\mathrm{Var}(\hat{p}_j) = \mathrm{Var}(\frac{1}{n}\sum_{k=1}^n p_j(1-p_j)) = \frac{1}{n}(p_j(1-p_j)) \leqslant \frac{1}{4n}$.

By denoting $\hat{p}_j^{nc} = \frac{1}{n}\sum_{k=1}^n \mathbb{1}_{X'_{kj}\neq \mathrm{NA}}$, one has $\mathbb{E}[\hat{p}_j^{nc}] = p_j$ and thus

$$\mathrm{Bias}(\hat{p}_j) = \mathbb{E}[\hat{p}_j - p_j] = \mathbb{E}[\hat{p}_j - \hat{p}_j^{nc}] = \mathbb{E}\left[\frac{1}{n}\mathbb{1}_{\hat{p}_j^{nc}\neq 0}\right] = \frac{1}{n}(1-p_j)^n \leqslant \frac{1}{n}(1-p_m)^n.$$

Thus $\sum_{j,j'=1, j\neq j'}^d (\mathbb{E}(\hat{p}_j - p_j)^2)(\mathbb{E}(\hat{p}_{j'} - p_{j'})^2) \leqslant d^2\left(\frac{1}{4n} + \left(\frac{1}{n}(1-p_m)^n\right)^2\right)^2$.

On the other hand, for the 4-th order moment, $\mathbb{E}((\hat{p}_j - p_j)^4) = \frac{1}{n^4}(n\mu_{4,p_j} + 3n(n-1)p_j^2(1-p_j)^2)$, with $\mu_{4,p_j} = p_j(1-p_j)^4 + p_j^4(1-p_j) \leqslant 1/12$ (this is the classical computation of the 4-th moment of a binomial random variable). Overall $\sum_{j=1}^d \mathbb{E}((\hat{p}_j - p_j)^4) \leqslant d(\frac{1}{12n^3} + \frac{3}{16n^2}) \leqslant \frac{d}{n^2} \leqslant \frac{d^2}{n^2}$.

Combining the second order and fourth order terms , $\mathbb{E}\left[\|\hat{p}-p\|^4\right] \leqslant \frac{d^2}{n^2} + d^2\left(\frac{5}{4n} + \left(\frac{1}{n}(1-p_m)^n\right)^2\right)^2 \leqslant d^2\left(\frac{5}{4n} + \left(\frac{1}{n}(1-p_m)^n\right)^2\right)^2$.

Therefore,

$$(\mathbb{E}\left[\|\hat{p}-p\|^4\right])^{1/2} \leqslant d\left(\frac{5}{4n} + \left(\frac{1}{n}(1-p_m)^n\right)^2\right) \tag{D.32}$$

which implies

$$\mathbb{E}[\|\hat{\beta}_k - \beta_k\|^2 | A_n] \leqslant \frac{2^5}{p_m^6}\frac{5}{\gamma\mu} C\frac{d}{n} + \frac{2^7}{p_m^6}\frac{1}{\gamma\mu} C\frac{d(1-p_m)^{2n}}{n^2}.$$

For the unconditional result, one splits the term to control on the event A_n and A_n^c, i.e.

$$\mathbb{E}^{\frac{1}{2}}\left[\frac{\|\hat{p}-p\|^4}{\min_{j=1,\dots,d}(p_j,\hat{p}_j)^{12}}\right] \leqslant \mathbb{E}^{\frac{1}{2}}\left[\frac{\|\hat{p}-p\|^4\mathbb{1}_{A_n}}{\min_{j=1,\dots,d}(p_j,\hat{p}_j)^{12}}\right] + \mathbb{E}^{\frac{1}{2}}\left[\frac{\|\hat{p}-p\|^4\mathbb{1}_{A_n^c}}{\min_{j=1,\dots,d}(p_j,\hat{p}_j)^{12}}\right]. \tag{D.33}$$

On the event A_n, one has $\hat{p}_j \geqslant p_j/2$ which leads to $\frac{1}{\min_{j=1,\dots,d}(p_j,\hat{p}_j)^6} \leqslant \frac{2^6}{p_j^6} \leqslant \frac{2^6}{p_m^6}$. Using Equation (D.32), one has

$$\mathbb{E}^{1/2}\left[\frac{\|\hat{p}-p\|^4}{\min_{j=1,\dots,d}(p_j,\hat{p}_j)^{12}}\mathbb{1}_{A_n}\right] \leqslant \frac{2^4}{p_m^6}\frac{5d}{n} + \frac{2^6}{p_m^6}\frac{d(1-p_m)^{2n}}{n^2}. \tag{D.34}$$

On the event A_n^c, one has $\hat{p}_j \leqslant p_j/2 \leqslant p_j$.

$$\mathbb{E}^{1/2}\left[\frac{\|\hat{p}-p\|^4}{\min_{j=1,\dots,d}(p_j,\hat{p}_j)^{12}}\mathbb{1}_{A_n^c}\right] \leqslant \mathbb{E}^{1/2}\left[\frac{1}{\hat{p}_j^{12}}\mathbb{1}_{A_n^c}\right]$$

As we have chosen to assign the minimal value n^{-1} to $\hat{p}_j$ in the rare event that the empirical proportion was 0, we have the lower bound $\hat{p}_j \geqslant \frac{1}{n}$ which implies

$$\mathbb{E}^{1/2}\left[\frac{\|\hat{p}-p\|^4}{\min_{j=1,\dots,d}(p_j,\hat{p}_j)^{12}}\mathbb{1}_{A_n^c}\right] \leqslant \mathbb{E}^{1/2}\left[\frac{1}{\hat{p}_j^{12}}\mathbb{1}_{A_n^c}\right] \leqslant \sqrt{n^{12}\mathbb{P}(A_n^c)} = n^6\sqrt{d}e^{-np_m/16}, \quad \text{(D.35)}$$

using Lemma 16. Combining Equations (D.33) to (D.35), one obtains

$$\mathbb{E}^{1/2}\left[\frac{\|\hat{p}-p\|^4}{\min_{j=1,\dots,d}(p_j,\hat{p}_j)^{12}}\right] \leqslant \left(\frac{2^4}{p_m^6}\frac{5d}{n} + \frac{2^6}{p_m^6}\frac{d(1-p_m)^{2n}}{n^2} + n^6\sqrt{d}e^{-np_m/16}\right)$$

Finally, with the bound given in Lemma 15, one has

$$\mathbb{E}[\|\hat{\beta}_k - \beta_k\|^2] \leqslant \frac{1}{\gamma\mu}C\left(\frac{2^4}{p_m^6}\frac{5d}{n} + \frac{2^6}{p_m^6}\frac{d(1-p_m)^{2n}}{n^2} + n^6\sqrt{d}e^{-np_m/16}\right).$$

□

Extension of such a result to cases without strong convexity (or independently of μ) is an interesting open direction.

D.4 Add-on to Section 4.5: Lipschitz constant computation

The Lipschitz constant L given in (4.6) is either computed from the complete covariates (oracle estimate) $\hat{L}_n^{\mathrm{OR}} = \frac{1}{p_m^2}\max_{1\leqslant k\leqslant n}\|X_{k:}\|^2$, or estimated from the incomplete data matrix, $\hat{L}_n^{\mathrm{NA}} = \frac{1}{\hat{p}_m^2}\max_{1\leqslant k\leqslant n}\frac{\|\tilde{X}_{k:}\|^2 d}{\sum_j D_{kj}}$, with $\hat{p}_m = \min_{1\leqslant j\leqslant d}\hat{p}_j$, and $\hat{p}_j = \frac{\sum_k D_{kj}}{n}$. In $\hat{L}_n^{\mathrm{NA}}$, the squared norm of each row $\|\tilde{X}_{k:}\|^2$ is divided by the proportion of observed values $\frac{d}{\sum_j D_{kj}}$. This way, the value of $\|\tilde{X}_{k:}\|^2$ is renormalized, by taking into account that some rows may contain more missing values than others. Note that theoretically the step size has to satisfy $\alpha \leqslant \frac{1}{2\hat{L}_n^{\mathrm{NA}}}$, thus $\hat{L}_n^{\mathrm{NA}}$ may be overestimated but should not be underestimated at the risk of instability in Algorithm 2. Figure D.1 shows that using a slightly overestimated Lipschitz constant estimate does not deteriorate the convergence obtained using the oracle estimate.

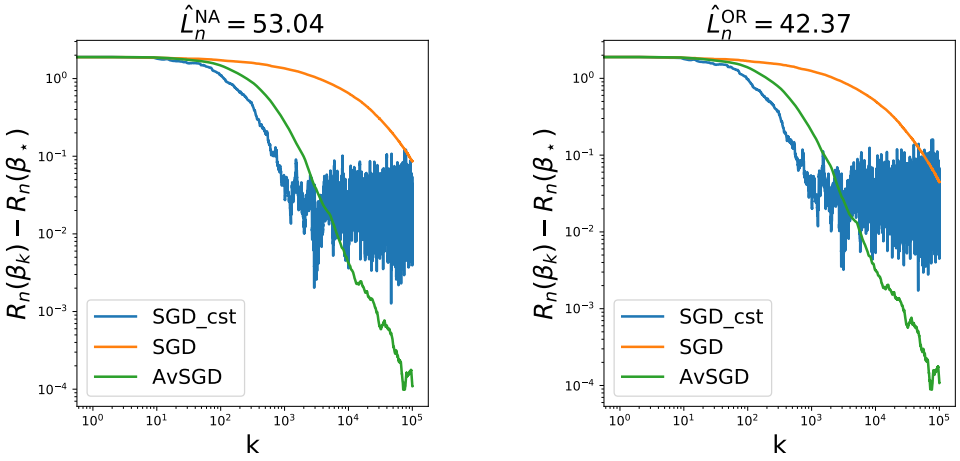


Figure D.1: Empirical excess risk ($R_n(\beta_k) - R_n(\beta^\star)$) given n for synthetic data ($n = 10^5$, $d = 10$) when there is 30% MCAR data, with 1 pass over the data and estimating the Lipschitz constant.

D.5 Add-on to Section 4.5: Handling polynomial missing features

The debiased averaged SGD algorithm proposed in Section 3.3.4 can be further extended to the case of polynomial features by using a different debiasing than in Equation (4.4).

For example, in dimension $d = 2$, with second-order polynomial features, the interaction effect of $X_{k1}X_{k2}$ and the effects of X_{k1}^2, X_{k2}^2 are accounted, so the augmented matrix design can be written as

$$(X_{:1}|X_{:2}|X_{:1}X_{:2}|X_{:1}^2|X_{:2}^2)^T.$$

Then, the "descent" direction at iteration k in Equation (4.4) should be chosen as

$$U^{\odot -1} \odot \tilde{X}_{k:}\tilde{X}_{k:}^T\beta_k - \mathrm{diag}(U)^{\odot -1} \odot \tilde{X}_{k:}y_k$$

where

$$U = \begin{pmatrix} p_1 & p_1p_2 & p_1p_2 & p_1 & p_1p_2 \\ p_1p_2 & p_2 & p_1p_2 & p_1p_2 & p_2 \\ p_1p_2 & p_1p_2 & p_1p_2 & p_1p_2 & p_1p_2 \\ p_1 & p_1p_2 & p_1p_2 & p_1 & p_1p_2 \\ p_1p_2 & p_2 & p_1p_2 & p_1p_2 & p_2 \end{pmatrix},$$

and $\mathrm{diag}(U)$ denotes the vector formed by the diagonal coefficients of U and $U^{\odot -1}$ stands for the matrix formed of the inverse coefficients of U.

Synthetic data. Considering a second-order model, we simulate data according to $y = (X_{:1}X_{:2}|X_{:1}^2|X_{:2}^2)^T\beta^\star + \epsilon$. An additional experiment is given in Figure D.2 in Section D.4,

illustrating that Algorithm 2 still achieves a rate of $\mathcal{O}\left(\frac{1}{n}\right)$ while dealing with polynomial features of degree 2.

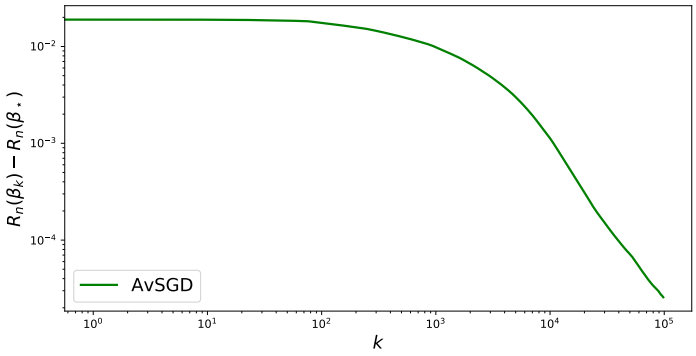


Figure D.2: Empirical excess risk ($R_n(\beta_k) - R_n(\beta^\star)$) given n for synthetic data ($n = 10^5$, $d = 10$) when the model accounts mixed effects.

Real dataset. About large-scale setting there is no computational barrier to apply the proposed method in high dimension, as the computational cost is similar to standard SGD strategies without missing data. These are computationally cheap at each iteration and particularly relevant on large datasets. In this section, we propose to run the proposed algorithm on the superconductivity dataset as in Section 4.5.3. 30% of missing values are uniformly introduced in the initial 81 features, with $n = 21263$. However, here we consider polynomial features of order 2, which increases the initial dimension 81 to 3400.

The empirical proportions of missing values for each variable in the resulting dataset are represented on Figure D.3, and the observed convergence rate for one pass on the data is displayed in Figure D.4. With the same numerical complexity, Algorithm 2 performs as well as an averaged SGD strategy run on the complete observations, whereas a standard SGD strategy run on imputed-by-0 data saturates far from the optimum.

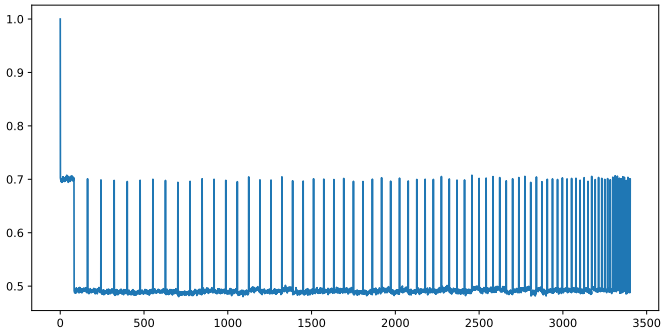


Figure D.3: Proportion of missing values for the polynomial features of degree 2 on the superconductivity dataset, when the initial missingness proportion on the raw features is 30%.

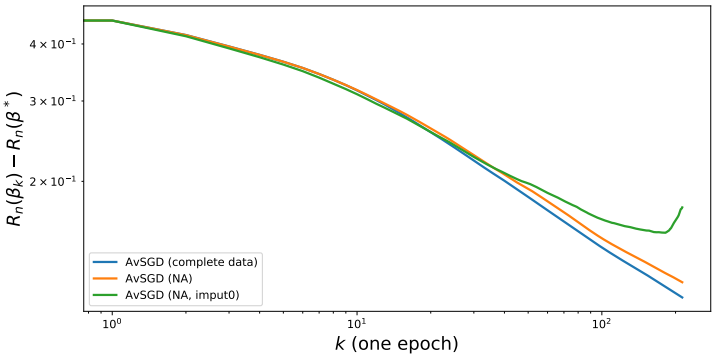


Figure D.4: Empirical excess risk $(R_n(\beta_k) - R_n(\beta^\star))$ given n for the superconductivity dataset ($n = 21263$) (containing 81 initial features) and $d = 3403$ with polynomial features of degree 2. Three different algorithms are compared: an averaged SGD on complete data (blue), the proposed debiased averaged SGD Algorithm 2 (orange) and an averaged SGD run on imputed-by-0 data without any debiasing (green).

D.6 Add-on to Section 4.5: Description of the Traumabase® data variables

The variables of the TraumaBase dataset which are used in experiments are the following:

- *Lactate*: The conjugate base of lactic acid.

- *Delta.Hemo*: The difference between the homoglobin on arrival at hospital and that in the ambulance.
- *VE*: A volume expander is a type of intravenous therapy that has the function of providing volume for the circulatory system.
- *RBC*: A binary index which indicates whether the transfusion of Red Blood Cells Concentrates is performed.
- *SI*: Shock index indicates level of occult shock based on heart rate (HR) and systolic blood pressure (SBP). $SI = \frac{HR}{SBP}$. Evaluated on arrival at hospital.
- *HR*: Heart rate measured on arrival of hospital.
- *Age*: Age.

Appendix E

Appendix of Chapter 5

E.1 Proof of Proposition 21

Proof of Proposition 21. We denote by $(\tilde{c}_1, \ldots, \tilde{c}_n)$ the patterns of missing data associated to the observed data $\tilde{y}^{\text{obs}}$. It is thus the concatenation $\tilde{c}_i = (c_i, \mathbf{0}_d)$ of c_i with the zero vector $\mathbf{0}_d = (0, \ldots, 0)$ of length d. Since all c_i values are observed in $\tilde{y}_i^{\text{obs}}$, it is the reason why the last d values in $\tilde{c}_i$ are fixed to zero. Then, the MAR assumption indicates that $\mathbb{P}(\tilde{c}_i \mid \tilde{y}_i, z_i; \lambda) = \mathbb{P}(\tilde{c}_i \mid \tilde{y}_i^{\text{obs}}; \lambda)$, with λ the related parameter. Consequently, similarly to (5.3) and using the i.i.d. assumption of all uplets $(\tilde{y}_i, z_i, \tilde{c}_i)$, the whole likelihood can be decomposed into two likelihoods

$$L(\pi, \theta, \psi, \lambda; \tilde{y}_i^{\text{obs}}, \tilde{c}_i) = L(\pi, \theta, \psi; \tilde{y}_i^{\text{obs}}) \times L(\lambda; \tilde{c}_i \mid \tilde{y}_i^{\text{obs}}). \tag{E.1}$$

Providing that (θ, ψ) and λ are functionally independent (ignorability of the MAR mechanism), the maximum likelihood estimate of (π, θ, ψ) is obtained by maximizing only $L(\pi, \theta, \psi; \tilde{y}_i^{\text{obs}})$, and does not depend on $L(\lambda; \tilde{c}_i \mid \tilde{y}_i^{\text{obs}})$. Finally, by using (5.16), the observed likelihood $L(\pi, \theta, \psi; \tilde{y}_i^{\text{obs}})$ is

$$\begin{aligned} L(\pi, \theta, \psi; \tilde{y}_i^{\text{obs}}) &= \sum_{k=1}^{K} \pi_k f_k(y_i^{\text{obs}}; \theta_k) \prod_{j=1}^{d} \rho(\alpha_{kj})^{c_{ij}} \rho(\alpha_{kj})^{(1-c_{ij})} && \text{(E.2)} \\ &= \sum_{k=1}^{K} \pi_k f_k(y_i^{\text{obs}}; \theta_k) \prod_{j=1}^{d} \mathbb{P}(c_{ij} \mid z_{ik} = 1; \psi). && \text{(E.3)} \end{aligned}$$

As $\mathbb{P}(c_{ij} \mid z_{ik} = 1; \psi)$ corresponds to the MNARz definition (5.12), the observed likelihood $L(\pi, \theta, \psi; \tilde{y}_i^{\text{obs}})$ is equal to the full observed likelihood $L(\pi, \theta, \psi; y_i^{\text{obs}}, c_i)$ associated to the MNARz model,

$$L(\pi, \theta, \psi; y_i^{\text{obs}}, c_i) = \sum_{k=1}^{K} \pi_k f_k(y_i^{\text{obs}}; \theta_k) \prod_{j=1}^{d} \mathbb{P}(c_{ij} \mid z_{ik} = 1; \psi).$$

□

E.2 Identifiability

E.2.1 Continuous and count data

Proof of Proposition 22. Suppose there exist two sets of parameters $\{\theta, \psi\}$ and $\{\theta', \psi'\}$ which have the same observed distribution, i.e. $f(y_i^{\text{obs}}, c_i; \theta, \psi) = f(y_i^{\text{obs}}, c_i; \theta', \psi')$. More precisely, one has

$$\begin{aligned}\forall y_i \in \mathbb{R}^d, \forall c_i \in \{0,1\}^d, \sum_{k=1}^{K} \int_{\mathcal{Y}_i^{\text{mis}}} \pi_k f_k(y_i; \theta_k) \prod_{j=1}^{d} \rho(\alpha_{kj} + \beta_{kj} y_{ij})^{c_{ij}} [1 - \rho(\alpha_{kj} + \beta_{kj} y_{ij})]^{1-c_{ij}} dy \\ = \sum_{k=1}^{K'} \int_{\mathcal{Y}_i^{\text{mis}}} \pi'_k f_k(y_i; \theta'_k) \prod_{j=1}^{d} \rho((\alpha')_{kj} + (\beta')_{kj} y_{ij})^{c_{ij}} [1 - \rho((\alpha')_{kj} + (\beta')_{kj} y_{ij})]^{1-c_{ij}} dy\end{aligned}$$

Let us consider the case when $c_{ij} = 1$ for all $j = 1, \ldots, d$. One has

$$\sum_{k=1}^{K} \pi_k f_k(y_i; \theta_k) \prod_{j=1}^{d} \rho(\alpha_{kj} + \beta_{kj} y_{ij}) = \sum_{k=1}^{K'} \pi'_k f_k(y_i; \theta'_k) \prod_{j=1}^{d} \rho((\alpha')_{kj} + (\beta')_{kj} y_{ij}).$$

By using the identifiability of the marginal mixture, one obtains $\theta_k = \theta'_k$. In addition, integrating out over all the elements but the j-th element, one has for all $y_{ij} \in \mathbb{R}$,

$$\sum_{k=1}^{K} \pi_k f_{kj}(y_{ij}; \theta_{kj}) \rho(\alpha_{kj} + \beta_{kj} y_{ij}) = \sum_{k=1}^{K'} \pi'_k f_{kj}(y_{ij}; \theta_{kj}) \rho((\alpha')_{kj} + (\beta')_{kj} y_{ij}).$$

In the sequel, we use the same reasoning of Theorem 2 in (Teicher, 1963).

Let us denote $F_k(y_{ij}) = f_{kj}(y_{ij}; \theta_{kj}) \rho(\alpha_{kj} + \beta_{kj} y_{ij})$ and $F'_k(y_{ij}) = f_{kj}(y_{ij}; \theta_{kj}) \rho((\alpha')_{kj} + (\beta')_{kj} y_{ij})$. Without loss of generality, assume that $F_k \prec F_l$ and $F'_k \prec F'_l$ for $k < l$. If $F_1 \neq F'_1$, we assume also without loss of generality that $F_1 \preceq F'_1$. Then, $F_1 \prec F'_k$ for $1 \leqslant k \leqslant K'$. For $u \in T_1$, where $T_1 = S_{F_1} \cap \{u : F_1(u) \neq 0\}$ is the domain of definition of F_1 such that $f_{1j}(u; \theta_{1j}) \rho(\alpha_{1j} + \beta_{1j} u) \neq 0$, one has

$$\pi_1 + \sum_{k=1}^{K} \pi_k \frac{F_k(u)}{F_1(u)} = \sum_{k=1}^{K'} \pi'_k \frac{F'_k(u)}{F_1(u)}$$

Letting $u \to +\infty$, $\pi_1 = 0$ which is in contradiction with the mixture model (where $\pi_k > 0$, $\forall k = 1, \ldots, K$. It implies that $F_1 = F'_1$. For any $u \in T_1$, one has

$$\pi_1 + \sum_{k=2}^{K} \pi_k \frac{F_k(u)}{F_1(u)} = \pi'_1 + \sum_{k=2}^{K'} \pi'_k \frac{F'_k(u)}{F_1(u)}$$

Letting again $u \to +\infty$, one obtains $\pi_1 = \pi'_1$ and $\sum_{k=2}^{K} \pi_k \frac{F_k(u)}{F_1(u)} = \sum_{k=2}^{K'} \pi'_k \frac{F'_k(u)}{F_1(u)}$. We repeat this argument to conclude that $F_k = F'_k$ and $\pi_k = \pi'_k$ for $k = 1, \ldots, \min\{K, K'\}$. Finally, if

$K \neq K'$, say $K > K'$, $\sum_{k=K'+1}^{K} \pi_k F_k(u) = 0$ implies $\pi_k = 0$ for $K'+1 \leqslant k \leqslant K$ which is in contradiction with the definition of the mixture model. Thus $K = K'$. Note that $F_k = F'_k$ implies that $\rho(\alpha_{kj} + \beta_{kj} y_{ij}) = \rho((\alpha')_{kj} + (\beta')_{kj} y_{ij})$ and thus $\alpha_{kj} = (\alpha')_{kj}$ and $\beta_{kj} = (\beta')_{kj}$, since ρ is an injective function. Indeed, ρ is assumed to be strictly monotone. □

Proof of Corollary 1. We use Proposition 22. We fix j. By abuse of notation, α_k, β_k, μ_k and σ_k correspond to the parameters α_{kj}, β_{kj}, μ_{kj} and Σ_{kj} of the variable j. Let us first consider the missing scenarios (5.6), (5.8) and (5.11) for which $\beta_k \neq \beta_{k+1}$. To obtain the total ordering, we need to prove that

$$\lim_{u \to +\infty} E_u = \frac{(1 + e^{-\alpha_k - \beta_k u}) e^{-\frac{(u-\mu_{k+1})^2}{2\sigma_{k+1}^2}}}{(1 + e^{-\alpha_{k+1} - \beta_{k+1} u}) e^{-\frac{(u-\mu_k)^2}{2\sigma_k^2}}} \frac{\sigma_k}{\sigma_{k+1}} = 0.$$

- If $\sigma_k^2 > \sigma_{k+1}^2$, $\lim_{u \to +\infty} E_u = \lim_{u \to +\infty} \exp -\frac{1}{2}(\frac{1}{\sigma_{k+1}^2} - \frac{1}{\sigma_k^2})u^2 = 0$.
- If $\sigma_k^2 = \sigma_{k+1}^2$, one has $\lim_{u \to +\infty} E_u = \lim_{u \to +\infty} \exp\left((\mu_k - \beta_k) - (\mu_{k+1} - \beta_{k+1})\right)u = 0$ discarding the case where $(\mu_k - \beta_k) - (\mu_{k+1} - \beta_{k+1}) = 0$ and assuming without loss of generality that $(\mu_k - \beta_k) > (\mu_{k+1} - \beta_{k+1})$. The set of nonidenfiable parameters is $\{\mu_k, \beta_k, \mu_{k+1}, \beta_{k+1} \text{ s.t.} (\mu_k - \beta_k) - (\mu_{k+1} - \beta_{k+1}) = 0\}_{k=1,\dots,K}$ and is of Lebesque zero measure.

Finally, for the missing scenarios (5.12) and (5.13), note that $\beta_k = \beta_{k+1} = 0$. For the missing scenarios (5.7), (5.9) and (5.10), one has $\beta_k = \beta_{k+1}$. Following the same reasoning as above, in the case where $\sigma_{k+1}^2 = \sigma_{k+1}^2$, one obtains the set of nonindentifiable parameters such that $\mu_k = \mu_{k+1}$, which is empty since $\mu_k \neq \mu_{k+1}$ if $\sigma_k^2 = \sigma_{k+1}^2$. □

Example 4 (Gaussian + Probit)**.** *Let us consider that ρ is the probit function and f_k (respectively f_{k+1}) the Gaussian density with parameters (μ_k, σ_k) (respectively $(\mu_{k+1}, \sigma_{k+1})$). Suppose without loss of generality that $\beta_k \geqslant \beta_{k+1}$. One want to prove that*

$$\lim_{u \to +\infty} E_u := \lim_{u \to +\infty} \frac{\int_{-\infty}^{\alpha_{k+1} + \beta_{k+1} u} e^{-t^2/2} dt}{\int_{-\infty}^{\alpha_k + \beta_k u} e^{-t^2/2} dt} \frac{\sigma_k \exp -\frac{(u-\mu_{k+1})^2}{2\sigma_{k+1}^2}}{\sigma_{k+1} \exp -\frac{(u-\mu_k)^2}{2\sigma_k^2}} = 0$$

Let us denote $\phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-t^2/2} dt$. One has

$$\lim_{u \to +\infty} \phi(u) = \begin{cases} 1 \text{ if } u > 0 \\ 1/2 \text{ if } u = 0 \\ 0 \text{ if } u < 0 \end{cases} \tag{E.4}$$

- *If $\beta_{k+1} > 0$ (and $\beta_k > 0$):*

$$\lim_{u \to +\infty} E_u = \lim_{u \to +\infty} \exp - \left(u^2 \left(\frac{1}{2\sigma_{k+1}^2} - \frac{1}{2\sigma_k^2} \right) + u \left(\frac{\mu_k}{\sigma_k} - \frac{\mu_{k+1}}{\sigma_{k+1}} \right) \right) = 0$$

assuming without loss of generality that $\sigma_k^2 > \sigma_{k+1}^2$ or $\mu_k > \mu_{k+1}$ if $\sigma_k^2 = \sigma_{k+1}^2$.

- *If $\beta_{k+1} \leqslant 0$ (and $\beta_k \geqslant 0$):*

$$\lim_{u \to +\infty} E_u = 0$$

since

$$\lim_{u \to +\infty} \exp - \left(u^2 \left(\frac{1}{2\sigma_{k+1}^2} - \frac{1}{2\sigma_k^2} \right) + u \left(\frac{\mu_k}{\sigma_k} - \frac{\mu_{k+1}}{\sigma_{k+1}} \right) \right) = 0$$

and

$$\lim_{u \to +\infty} \frac{\int_{-\infty}^{\alpha_{k+1}+\beta_{k+1}u} e^{-t^2/2} dt}{\int_{-\infty}^{\alpha_k+\beta_k u} e^{-t^2/2} dt} = \begin{cases} 0 \text{ if } \beta_{k+1} < 0 \\ 1/2 \text{ if } \beta_{k+1} = 0 \text{ and } \beta_k > 0 \\ 1 \text{ if } \beta_{k+1} = 0 \text{ and } \beta_k = 0 \end{cases} \tag{E.5}$$

- *If $\beta_{k+1} < 0$ and $\beta_k < 0$: One uses the upper and lower bounds for the probit function.*

$$\frac{1}{-t + \sqrt{t^2 + 4}} < \sqrt{\frac{\pi}{2}} \exp \frac{t^2}{2} \phi(t) < \frac{1}{-t + \sqrt{t^2 + 8/\pi}},$$

i.e. $\phi(t) < \sqrt{\frac{2}{\pi}} \frac{1}{-t+\sqrt{t^2+8/\pi}} \exp - \frac{t^2}{2}$ and $\frac{1}{\phi(t)} < (-t + \sqrt{t^2+4}) \sqrt{\frac{\pi}{2}} \exp \frac{t^2}{2}$ Thus, noting that $\lim_{u \to +\infty} \phi(\alpha_{k+1} + \beta_{k+1} u) = \lim_{u \to +\infty} \phi(\beta_{k+1} u)$,

$$\frac{\int_{-\infty}^{\alpha_{k+1}+\beta_{k+1}u} e^{-t^2/2} dt}{\int_{-\infty}^{\alpha_k+\beta_k u} e^{-t^2/2} dt} \underset{u \to +\infty}{=} \frac{\phi(\beta_{k+1} u)}{\phi(\beta_k u)} \underset{u \to +\infty}{<} \exp \left(\left(\frac{\beta_k^2}{2} - \frac{\beta_{k+1}^2}{2} \right) u^2 \right) \tag{E.6}$$

As $\beta_{k+1} \leqslant \beta_k < 0$, one has $\beta_k^2/2 - \beta_{k+1}^2/2 < 0$ and it implies

$$\lim_{u \to +\infty} \frac{\int_{-\infty}^{\alpha_{k+1}+\beta_{k+1}u} e^{-t^2/2} dt}{\int_{-\infty}^{\alpha_k+\beta_k u} e^{-t^2/2} dt} = 0.$$

Given that

$$\lim_{u \to +\infty} \exp - \left(u^2 \left(\frac{1}{2\sigma_{k+1}^2} - \frac{1}{2\sigma_k^2} \right) + u \left(\frac{\mu_k}{\sigma_k} - \frac{\mu_{k+1}}{\sigma_{k+1}} \right) \right) = 0,$$

assuming without loss of generality that $\sigma_k^2 > \sigma_{k+1}^2$ or $\mu_k > \mu_{k+1}$ if $\sigma_k^2 = \sigma_{k+1}^2$, one has

$$\lim_{u \to +\infty} E_u = 0.$$

Example 5 (Poisson + Probit)**.** *Considering that ρ is the probit function and f_k (respectively f_{k+1}) the Poisson distribution with parameters λ_k (respectively λ_{k+1}). Suppose without loss of generality that $\beta_k > \beta_{k+1}$ and $\lambda_k > \lambda_{k+1}$. One want to prove*

$$\lim_{u \to +\infty} E_u := \lim_{u \to +\infty} \frac{\int_{-\infty}^{\alpha_{k+1}+\beta_{k+1}u} e^{-t^2/2} dt}{\int_{-\infty}^{\alpha_k+\beta_k u} e^{-t^2/2} dt} \frac{\lambda_{k+1}^u e^{-\lambda_{k+1}}}{\lambda_k^u e^{-\lambda_k}} = 0.$$

- *If $\beta_{k+1} > 0$ (and $\beta_k > 0$): using (E.4), one has*

$$\lim_{u\to+\infty} E_u = \lim_{u\to+\infty} \exp u \ln \frac{\lambda_{k+1}}{\lambda_k} = 0.$$

- *If $\beta_{k+1} \leqslant 0$ (and $\beta_k \geqslant 0$): one has*

$$\lim_{u\to+\infty} E_u = 0.$$

using

$$\lim_{u\to+\infty} \exp u \ln \frac{\lambda_{k+1}}{\lambda_k} = 0$$

and (E.5) for the missing distribution part.

- *If $\beta_{k+1} < 0$ and $\beta_k < 0$: using (E.6), one obtains*

$$\lim_{u\to+\infty} E_u < \lim_{u\to+\infty} \exp\left(\left(\frac{\beta_k^2}{2} - \frac{\beta_{k+1}^2}{2}\right) u^2\right) \exp u \ln \frac{\lambda_{k+1}}{\lambda_k} = 0,$$

because $\beta_k^2/2 - \beta_{k+1}^2/2 < 0$.

Example 6 (Poisson + Logistic)**.** *Considering that ρ is the logistic function and f_k (respectively f_{k+1}) the Poisson distribution with parameters λ_k (respectively λ_{k+1}). One want to prove that*

$$\lim_{u\to+\infty} E_u = \lim_{u\to+\infty} \frac{1 + e^{-\alpha_k - \beta_k u}}{1 + e^{-\alpha_{k+1} - \beta_{k+1} u}} \exp u \ln \frac{\lambda_{k+1}}{\lambda_k} = 0.$$

Assume that $\lambda_k > \lambda_{k+1}$ without loss of generality.

- *For the missing scenarios (5.6), (5.8) and (5.11) for which $\beta_k \neq \beta_{k+1}$, one obtains the generic identifiability where the set of non-identifiable parameters is $\{\alpha_k, \beta_k, \lambda_k \text{ s.t.} (\ln \lambda_k - \beta_k) - (\ln \lambda_{k+1} - \beta_{k+1}) = 0\}_{k=1,\dots,K}$ and is of Lebesque zero measure.*

- *For the missing scenarios (5.12) and (5.13), note that $\beta_k = \beta_{k+1} = 0$. For the missing scenarios (5.7), (5.9) and (5.10), one has $\beta_k = \beta_{k+1}$. It implies that idenfiability holds since*

$$\lim_{u\to+\infty} E_u = \lim_{u\to+\infty} \exp u \ln \frac{\lambda_{k+1}}{\lambda_k} = 0.$$

E.2.2 Categorical data

Proof of Proposition 23. Let us first consider the case where $\beta_{kj} = (0, \dots, 0) \in \mathbb{R}^{\ell_j}, \forall k = 1, \dots, K, \forall j = 1, \dots, d$. Suppose there exists two sets of parameters $\{\theta, \psi\}$ and $\{\theta', \psi'\}$ which

have the same observed distribution.

$$\forall y_i \in \mathbb{R}^d, \forall c_i \in \{0,1\}^d, \sum_{k=1}^{K} \int_{y_i^{\text{mis}}} \pi_k f_k(y_i;\theta_k) \prod_{j=1}^{d} \rho(\alpha_{kj})^{c_{ij}} [1-\rho(\alpha_{kj})]^{1-c_{ij}} dy$$
$$= \sum_{k=1}^{K'} \int_{y_i^{\text{mis}}} \pi'_k f_k(y_i;\theta'_k) \prod_{j=1}^{d} \rho((\alpha')_{kj})^{c_{ij}} [1-\rho(\alpha'_{kj})]^{1-c_{ij}} dy$$

Identifiability of ψ This implies that the marginal distributions of the pattern of missing data for the two sets of parameters ψ and ψ' are equal.

$$\sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} \rho(\alpha_{kj})^{c_{ij}} [1-\rho(\alpha_{kj})]^{1-c_{ij}} = \sum_{k=1}^{K'} \pi'_k \prod_{j=1}^{d} \rho(\alpha'_{kj})^{c_{ij}} [1-\rho(\alpha'_{kj})]^{1-c_{ij}}$$

One recognizes the finite mixture of K different Bernoulli products with d components and with parameters $(\rho(\alpha_{k1}),\dots,\rho(\alpha_{kd})))_{k=1,\dots,K}$ and $(\rho(\alpha'_{k1}),\dots,\rho(\alpha'_{kd})))_{k=1,\dots,K}$. The generic identifiability up to a label swapping of these parameters is given by Corollary 5 in Allman et al. (2009). As the function ρ is strictly monotone, the equality $\rho(\alpha_{kj}) = \rho(\alpha'_{kj})$ implies $\alpha_{kj} = \alpha'_{kj}$.

Identifiability of θ Let us consider the case where all the elements of y_i are observed, i.e. $c_{ij} = 1, \forall j = 1,\dots,d$. One has

$$\sum_{k=1}^{K} \pi_k f_k(y_i;\theta_k) \prod_{j=1}^{d} \rho(\alpha_{kj}) = \sum_{k=1}^{K'} \pi'_k f_k(y_i;\theta'_k) \prod_{j=1}^{d} \rho(\alpha'_{kj}),$$

i.e. by independence to the group membership,

$$\sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} f_{kj}(y_{ij};\theta_{kj})\rho(\alpha_{kj}) = \sum_{k=1}^{K'} \pi'_k \prod_{j=1}^{d} f_{kj}(y_{ij};\theta'_{kj})\rho(\alpha'_{kj}),$$
$$\Leftrightarrow \sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} \rho(\alpha_{kj}) \prod_{h=1}^{\ell_j} (\theta^h_{kj})^{y^h_{ij}} = \sum_{k=1}^{K'} \pi'_k \prod_{j=1}^{d} \rho(\alpha_{kj}) \prod_{h=1}^{\ell_j} ((\theta'_{kj})^h)^{y^h_{ij}}.$$

We recognize the finite mixture of K multinomial distributions with d components for $y_{ij} = (y^1_{ij},\dots,y^{\ell_j}_{ij}), j = 1,\dots,d$ with paramaters $(\theta_{kj}) = (\theta^1_{kj},\dots,\theta^{\ell_j}_{kj}), j = 1,\dots,d$ and proportions $(\pi_k \prod_{j=1}^{d} \rho(\alpha_{kj}))_{k=1,\dots,K}$. We can thus apply Theorem 4 (Allman et al., 2009) with the model $\mathcal{M}(K; l_1,\dots,l_d)$ which gives the generic identifiability of the model paramaters up to a label swapping, i.e.

$$\forall k, \forall j,\ \theta^h_{kj} = (\theta'_{kj})^h$$
$$\forall k,\ \pi_k \prod_{j=1}^{d} \rho(\alpha_{kj}) = \pi'_k \prod_{j=1}^{d} \rho(\alpha'_{kj})$$

The second equality implies $\pi_k = \pi'_k$ using the generic identifiability of $\rho(\alpha_{kj}), \forall k, \forall j$ stated above. If $K \neq K'$, say $K > K'$, $\sum_{k=K'+1}^{K} \pi_k \prod_{j=1}^{d} \rho(\alpha_{kj}) \prod_{h=1}^{\ell_j} (\theta_{kj}^h)^{y_{ij}^h} = 0$ implies $\pi_k = 0$ for $K' + 1 \leqslant k \leqslant K$.

We consider now the missing scenarios for which $\beta_{kj} \neq 0$. The identifiability does not hold. We can present a counter-example. The set of parameters $\psi = \{\alpha = (1, \ldots, 1), \beta = (1, \ldots, 1)\}$ has the same observed distribution than another set of parameters $\psi' = \{\alpha' = (0, \ldots, 0), \beta' = (2, \ldots, 2)\}$. Indeed, in the case where $y_{ij} = (1, \ldots, 1)$, $\rho(\alpha_{kj} + \beta_{kj} y_{ij}) = \rho(\alpha'_{kj} + \beta'_{kj} y_{ij})$. □

E.3 Detailed algorithms

The algorithms for the different missing scenarios and type of data are given. In particular, for continuous data, we derive the formulae assuming Gaussian data.

E.3.1 EM algorithm

The EM algorithm consists on two steps iteratively proceeded: the E-step and M-step. For the E-step, one has

$$
\begin{aligned}
&Q(\pi, \theta, \psi; \pi^r, \theta^r, \psi^r) \\
&= \mathbb{E}[\ell_{\text{comp}}(\pi, \theta, \psi; y, z, c) | y_i^{\text{obs}}, c_i; \pi^r, \theta^r, \psi^r] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}\left[\log(\pi_k f_k(y_i; \theta) \mathbb{P}(c_i \mid y_i, z_{ik} = 1; \psi)) \mid y_i^{\text{obs}}, c_i; \pi^r, \theta^r, \psi^r\right] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \int_{\mathcal{Y}_i^{\text{mis}}} \log(\pi_k f_k(y_i; \theta) \mathbb{P}(c_i \mid y_i, z_{ik} = 1; \psi)) \mathbb{P}(y_i^{\text{mis}}, z_{ik} = 1 | y_i^{\text{obs}}, c_i; \pi^r, \theta^r, \psi^r) dy_i^{\text{mis}} \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} (\tau_{ik})^r \int_{\mathcal{Y}_i^{\text{mis}}} \log(\pi_k f_k(y_i; \theta) \mathbb{P}(c_i \mid y_i, z_{ik} = 1; \psi)) \mathbb{P}(y_i^{\text{mis}} | y_i^{\text{obs}}, z_{ik} = 1, c_i; \pi^r, \theta^r, \psi^r) dy_i^{\text{mis}}
\end{aligned}
$$

using $\mathbb{P}(y_i^{\text{mis}}, z_{ik} = 1 | y_i^{\text{obs}}, c_i; \pi^r, \theta^r, \psi^r) = (\tau_{ik})^r \mathbb{P}(y_i^{\text{mis}} | y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r)$ with $(\tau_{ik})^r = \mathbb{P}(z_{ik} = 1 \mid y_i^{\text{obs}}, c_i; \pi^r, \theta^r, \psi^r)$ in the last step.

It leads to the decomposition in (5.21) recalled here

$$
Q(\pi, \theta, \psi; \pi^r, \theta^r, \psi^r) = \sum_{i=1}^{n} \sum_{k=1}^{K} (\tau_{ik})^r \log(\pi_k) + (\tau_{ik})^r E_{iy}^r(\theta) + (\tau_{ik})^r E_{ic}^r(\psi).
$$

The terms involved in this decomposition, given in (5.24), (5.25) and (5.26), are now detailed.

(a) the expectation of the data mixture part over the missing values given the available information (i.e. the observed data and the indicator pattern), the class membership and the current value of the parameters:

$$
E_{iy}^r(\theta) = \mathbb{E}\left[\log(f_k(y_i; \theta_k)) \mid y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r\right],
$$

(b) the expectation of the missing mechanism part over the missing values given the available information, the class membership and the current value of the parameters:

$$E_{ic}^{r}(\psi) = \mathbb{E}\left[\log(\mathbb{P}(c_i \mid y_i, z_{ik} = 1; \psi)) \mid y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r\right],$$

(c) the conditional probability for an observation i to belong to the class k given the available information and the current value of the parameters:

$$(\tau_{ik})^r = \mathbb{P}(z_{ik} = 1 \mid y_i^{\text{obs}}, c_i; \pi^r, \theta^r, \psi^r).$$

Terms (a) and (b) require to integrate over the distribution $\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r)$. For Term (a), one has

$$\begin{aligned}\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r) &= \frac{\mathbb{P}(y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r)}{\mathbb{P}(y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r)} \\ &= \frac{\mathbb{P}(c_i \mid y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik} = 1; \psi^r)\mathbb{P}(y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik} = 1; \theta^r)}{\int_{\mathcal{Y}_i^{\text{mis}}} \mathbb{P}(c_i \mid y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik} = 1; \psi^r)\mathbb{P}(y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik} = 1; \theta^r) dy_i^{\text{mis}}}\end{aligned} \tag{E.7}$$

Term (c) corresponds to the conditional probability for an observation i to arise from the kth mixture component with the current values of the model parameter. More particularly, one has

$$\begin{aligned}(\tau_{ik})^r &= \frac{\mathbb{P}(z_{ik} = 1, y_i^{\text{obs}}, c_i; \pi^r, \theta^r, \psi^r)}{\mathbb{P}(y_i^{\text{obs}}, c_i; \pi^r, \theta^r, \psi^r)} \\ &= \frac{\mathbb{P}(z_{ik} = 1, y_i^{\text{obs}}, c_i; \pi^r, \theta^r, \psi^r)}{\sum_{h=1}^{K} \mathbb{P}(z_{ih} = 1, y_i^{\text{obs}}, c_i; \pi^r, \theta^r, \psi^r)} \\ &= \frac{\mathbb{P}(z_{ik} = 1; \pi^r)\mathbb{P}(y_i^{\text{obs}} \mid z_{ik} = 1; \theta_k^r)\mathbb{P}(c_i \mid y_i^{\text{obs}}, z_{ik} = 1; \theta^r, \psi^r)}{\sum_{h=1}^{K} \mathbb{P}(z_{ih} = 1; \pi^r)\mathbb{P}(y_i^{\text{obs}} \mid z_{ih} = 1; \theta_h^r)\mathbb{P}(c_i \mid y_i^{\text{obs}}, z_{ih} = 1; \theta^r, \psi^r)} \\ &= \frac{\pi_k^r f_k(y_i^{\text{obs}}; \theta_k^r)\mathbb{P}(c_i \mid y_i^{\text{obs}}, z_{ik} = 1; \theta^r, \psi^r)}{\sum_{h=1}^{K} \pi_h^r f_h(y_i^{\text{obs}}; \theta_h^r)\mathbb{P}(c_i \mid y_i^{\text{obs}}, z_{ih} = 1; \theta^r, \psi^r)}\end{aligned} \tag{E.8}$$

The quantity that can cause numerical difficulties is the probability $\mathbb{P}(c_i \mid y_i^{\text{obs}}, z_{ik} = 1; \theta^r, \psi^r)$.

E.3.1.1 Gaussian mixture for continuous data

The pdf $f_k(y_i; \theta) = \phi(y_i; \mu_k, \Sigma_k)$ is assumed to be a Gaussian distribution with mean vector μ_k and covariance matrix Σ_k. First, let us detail the terms of the E-step. Term (a) is written as follows:

$$\begin{aligned}\mathbb{E}\left[\log(\phi(y_i; \mu_k, \Sigma_k)) \mid y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r\right] &= -\frac{1}{2}\left[n\log(2\pi) + \log((\mid \Sigma_k \mid))\right] \\ &\quad - \frac{1}{2}\mathbb{E}\left[(y_i - \mu_k)^T(\Sigma_k)^{-1}(y_i - \mu_k) \mid y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r\right].\end{aligned}$$

This last term could be expressed using the commutativity and linearity of the trace function:

$$\mathbb{E}\left[(y_i-\mu_k)^T(\Sigma_k)^{-1}(y_i-\mu_k) \mid y_i^{\text{obs}}, z_{ik}=1, c_i; \theta^r, \psi^r\right] \\ = \text{tr}(\mathbb{E}\left[(y_i-\mu_k)(y_i-\mu_k)^T \mid y_i^{\text{obs}}, z_{ik}=1, c_i; \theta^r, \psi^r\right](\Sigma_k)^{-1}).$$

Finally note that only $\mathbb{E}\left[(y_i-\mu_k)(y_i-\mu_k)^T \mid y_i^{\text{obs}}, z_{ik}=1, c_i; \theta^r, \psi^r\right]$ has to be calculated.

MNARz and MNARz^j models For the MNARz and MNARz^j models, the effect of the missingness is only due to the class membership. Term (a) is the same for both models but (b) and (c) differ. Let us first detail these terms.

- For Term (a), note that

$$\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1, c_i; \theta^r, \psi^r) = \mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1; \theta^r),$$

  which makes the computation easy. Indeed, using (E.7),

$$\begin{aligned}
&\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1, c_i; \theta^r, \psi^r) \\
&= \frac{\prod_{j=1}^d \rho(\alpha_{kj}^r)^{c_{ij}}(1-\rho(\alpha_{kj}^r))^{1-c_{ij}}\mathbb{P}(y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}=1; \theta^r)}{\int_{y_i^{\text{mis}}}\prod_{j=1}^d \rho(\alpha_{kj}^r)^{c_{ij}}(1-\rho(\alpha_{kj}^r))^{1-c_{ij}}\mathbb{P}(y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}=1; \theta^r)dy_i^{\text{mis}}} \\
&= \frac{\mathbb{P}(y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}=1; \theta^r)}{\int_{y_i^{\text{mis}}}\mathbb{P}(y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}=1; \theta^r)dy_i^{\text{mis}}} = \mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1; \theta^r),
\end{aligned}$$

  since $\prod_{j=1}^d \rho(\alpha_{kj}^r)^{c_{ij}}(1-\rho(\alpha_{kj}^r))^{1-c_{ij}}$ does not depend on y_i^{mis} and is simplified with the numerator. The law of $(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1)$ is Gaussian. Noting that

$$\begin{aligned}
&(y_i \mid z_{ik}=1; \theta^r) \\
&= \left(\begin{pmatrix} y_i^{\text{obs}} \\ y_i^{\text{mis}} \end{pmatrix} \mid z_{ik}=1; \theta^r\right) \sim \mathcal{N}\left(\begin{pmatrix} (\mu_{ik}^{\text{obs}})^r \\ (\mu_{ik}^{\text{mis}})^r \end{pmatrix}, \begin{pmatrix} (\Sigma_{ik}^{\text{obs,obs}})^r & (\Sigma_{ik}^{\text{obs,mis}})^r \\ (\Sigma_{ik}^{\text{mis,obs}})^r & (\Sigma_{ik}^{\text{mis,mis}})^r \end{pmatrix}\right),
\end{aligned}$$

  one obtains

$$\left(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1; \theta^r\right) \sim \mathcal{N}\left((\tilde{\mu}_{ik}^{\text{mis}})^r, (\tilde{\Sigma}_{ik}^{\text{mis}})^r\right). \tag{E.9}$$

  with $(\tilde{\mu}_{ik}^{\text{mis}})^r$ and $(\tilde{\Sigma}_{ik}^{\text{mis}})^r$ the standard expression of the mean vector and covariance matrix of a conditional Gaussian distribution (see for instance Anderson (2003)) detailed as follows

$$(\tilde{\mu}_{ik}^{\text{mis}})^r = (\mu_{ik}^{\text{mis}})^r + (\Sigma_{ik}^{\text{mis,obs}})^r\left((\Sigma_{ik}^{\text{obs,obs}})^r\right)^{-1}\left(y_i^{\text{obs}} - (\mu_{ik}^{\text{obs}})^r\right), \tag{E.10}$$

$$(\tilde{\Sigma}_{ik}^{\text{mis}})^r = (\Sigma_{ik}^{\text{mis,mis}})^r - (\Sigma_{ik}^{\text{mis,obs}})^r\left((\Sigma_{ik}^{\text{obs,obs}})^r\right)^{-1}(\Sigma_{ik}^{\text{obs,mis}})^r. \tag{E.11}$$

Note also that we have

$$(y_i-\mu_k)(y_i-\mu_k)^T=\begin{pmatrix}(y_i^{\text{obs}}-\mu_{ik}^{\text{obs}})^T(y_i^{\text{obs}}-\mu_{ik}^{\text{obs}}) & (y_i^{\text{obs}}-\mu_{ik}^{\text{obs}})^T(y_i^{\text{mis}}-\mu_{ik}^{\text{mis}})\\ (y_i^{\text{mis}}-\mu_{ik}^{\text{mis}})^T(y_i^{\text{obs}}-\mu_{ik}^{\text{obs}}) & (y_i^{\text{mis}}-\mu_{ik}^{\text{mis}})^T(y_i^{\text{mis}}-\mu_{ik}^{\text{mis}})\end{pmatrix}.$$

Therefore, the expected value of each block for the current parameter value is

$$\begin{aligned}
&\mathbb{E}\left[(y_i^{\text{obs}}-\mu_{ik}^{\text{obs}})^T(y_i^{\text{obs}}-\mu_{ik}^{\text{obs}})\mid y_i^{\text{obs}},z_{ik}=1;\theta^r\right]=(y_i^{\text{obs}}-\mu_{ik}^{\text{obs}})^T(y_i^{\text{obs}}-\mu_{ik}^{\text{obs}})\\
&\mathbb{E}\left[(y_i^{\text{obs}}-\mu_{ik}^{\text{obs}})^T(y_i^{\text{mis}}-\mu_{ik}^{\text{mis}})\mid y_i^{\text{obs}},z_{ik}=1;\theta^r\right]=(y_i^{\text{obs}}-\mu_{ik}^{\text{obs}})^T((\tilde{\mu}_{ik}^{\text{mis}})^r-\mu_{ik}^{\text{mis}})\\
&\mathbb{E}\left[(y_i^{\text{mis}}-\mu_{ik}^{\text{mis}})^T(y_i^{\text{mis}}-\mu_{ik}^{\text{mis}})\mid y_i^{\text{obs}},z_{ik}=1;\theta^r\right]=((\tilde{\mu}_{ik}^{\text{mis}})^r-\mu_{ik}^{\text{mis}})^T((\tilde{\mu}_{ik}^{\text{mis}})^r-\mu_{ik}^{\text{mis}})+(\tilde{\Sigma}_{ik}^{\text{mis}})^r
\end{aligned}$$

- For Term (b), $\mathbb{P}(c_i\mid y_i,z_{ik}=1;\psi)$ is independent of y, which implies

$$\begin{aligned}
E_{ic}^r(\psi)&=\log(\mathbb{P}(c_i\mid z_{ik}=1;\psi))\\
&=\begin{cases}\sum_{j=1}^d c_{ij}\log\rho(\alpha_k)+(1-c_{ij})\log(1-\rho(\alpha_k)) & (\text{MNAR}z)\\ \sum_{j=1}^d c_{ij}\log\rho(\alpha_{kj})+(1-c_{ij})\log(1-\rho(\alpha_{kj})) & (\text{MNAR}z^j)\end{cases}
\end{aligned}\tag{E.12}$$

- For Term (c), one first remark that

$$\begin{aligned}
&\mathbb{P}(c_i\mid y_i^{\text{obs}},z_{ik}=1;\theta^r,\psi^r)\\
&\quad=\prod_{j=1}^d\mathbb{P}(c_{ij}=1\mid y_i^{\text{obs}},z_{ik}=1;\theta^r,\psi^r)^{c_{ij}}\mathbb{P}(c_{ij}=0\mid y_i^{\text{obs}},z_{ik}=1;\theta^r,\psi^r)^{1-c_{ij}}.
\end{aligned}$$

In particular, for MNARz and MNARz^j, by independence of y, one has

$$\mathbb{P}(c_{ij}=1\mid y_i^{\text{obs}},z_{ik}=1;\theta^r,\psi^r)=\mathbb{P}(c_{ij}=1\mid z_{ik}=1;\theta^r,\psi^r)=\begin{cases}\rho(\alpha_k) & (\text{MNAR}z)\\ \rho(\alpha_{kj}) & (\text{MNAR}z^j)\end{cases}$$

Using (5.26), one obtains

$$(\tau_{ik})^r=\begin{cases}\frac{\pi_k^r\phi(y_i^{\text{obs}};(\mu_{ik}^{\text{obs}})^r,(\Sigma_{ik}^{\text{obs,obs}})^r)\prod_{j=1}^d\rho(\alpha_k^r)^{c_{ij}}(1-\rho(\alpha_k^r))^{1-c_{ij}}}{\sum_{h=1}^K\pi_h^r\phi(y_i^{\text{obs}};(\mu_{ih}^{\text{obs}})^r,(\Sigma_{ih}^{\text{obs,obs}})^r)\prod_{j=1}^d\rho(\alpha_k^r)^{c_{ij}}(1-\rho(\alpha_k^r))^{1-c_{ij}}} & (\text{MNAR}z)\\ \frac{\pi_k^r\phi(y_i^{\text{obs}};(\mu_{ik}^{\text{obs}})^r,(\Sigma_{ik}^{\text{obs,obs}})^r)\prod_{j=1}^d\rho(\alpha_{kj}^r)^{c_{ij}}(1-\rho(\alpha_{kj}^r))^{1-c_{ij}}}{\sum_{h=1}^K\pi_h^r\phi(y_i^{\text{obs}};(\mu_{ih}^{\text{obs}})^r,(\Sigma_{ih}^{\text{obs,obs}})^r)\prod_{j=1}^d\rho(\alpha_{kj}^r)^{c_{ij}}(1-\rho(\alpha_{kj}^r))^{1-c_{ij}}} & (\text{MNAR}z^j)\end{cases}\tag{E.13}$$

If ρ is the logistic distribution, the expression can be written more simply

$$(\tau_{ik})^r\propto\pi_k^r\phi(y_i^{\text{obs}};\theta_k^r)\prod_{j=1}^d(1+\exp(-\delta_{ij}\alpha_{kj}^r))^{-1}\text{ where }\delta_{ij}=\begin{cases}1 & \text{if } c_{ij}=1\\ -1 & \text{otherwise.}\end{cases}$$

Finally, the E-step and the M-step can be sketched as follows in the Gaussian mixture case.

E-step The E-step for Term (a) consists of computing for $k = 1, \dots, K$ and $i = 1, \dots, n$

$$
\begin{aligned}
(\tilde{\mu}_{ik}^{\text{mis}})^r &= (\mu_{ik}^{\text{mis}})^r + (\Sigma_{ik}^{\text{mis,obs}})^r \left((\Sigma_{ik}^{\text{obs,obs}})^r\right)^{-1} \left(y_i^{\text{obs}} - (\mu_{ik}^{\text{obs}})^r\right) \\
(\tilde{\Sigma}_{ik}^{\text{mis}})^r &= (\Sigma_{ik}^{\text{mis,mis}})^r - (\Sigma_{ik}^{\text{mis,obs}})^r \left((\Sigma_{ik}^{\text{obs,obs}})^r\right)^{-1} (\Sigma_{ik}^{\text{obs,mis}})^r \\
(\tilde{y}_{i,k})^r &= (y_i^{\text{obs}}, (\tilde{\mu}_{ik}^{\text{mis}})^r) \\
\tilde{\Sigma}_{ik}^r &= \begin{pmatrix} 0_i^{\text{obs,obs}} & 0_i^{\text{obs,mis}} \\ 0_i^{\text{mis,obs}} & (\tilde{\Sigma}_{ik}^{\text{mis}})^r \end{pmatrix}
\end{aligned}
$$

Note that whenever the mixture covariance matrices are supposed diagonal then $(\tilde{\Sigma}_{ik}^{\text{mis}})^r$ is also a diagonal matrix. Term (c) also requires the computation of $(\tau_{ik})^r$ given in (E.13) for $k = 1, \dots, K$ and $i = 1, \dots, n$.

M-step The maximization of $Q_y(\pi, \theta; \pi^r, \theta^r)$ given in (5.22) leads to, for $k = 1, \dots, K$,

$$
\begin{aligned}
\pi_k^{r+1} &= \frac{1}{n} \sum_{i=1}^{n} (\tau_{ik})^r \\
\mu_k^{r+1} &= \frac{\sum_{i=1}^{n} (\tau_{ik})^r (\tilde{y}_{k,i})^r}{\sum_{i=1}^{n} (\tau_{ik})^r} \\
\Sigma_k^{r+1} &= \frac{\sum_{i=1}^{n} \left[(\tau_{ik})^r \left((\tilde{y}_{i,k})^r - \mu_k^{r+1})((\tilde{y}_{i,k})^r - \mu_k^{r+1})^T + \tilde{\Sigma}_{ik}^r \right) \right]}{\sum_{i=1}^{n} (\tau_{ik})^r}
\end{aligned}
$$

Then, the maximization of the function $Q_c(\psi; \psi^r)$ in ψ can be performed using a Newton Raphson algorithm. For $k = 1, \dots, K$, it remains to fit a generalized linear model with the binomial link function for the matrix $(\mathcal{J}_k^{\text{MNAR}z})^{r+1}$ (if the model is MNARz) or for the matrices $(\mathcal{J}_{kj}^{\text{MNAR}z^j})_{j=1,\dots,d}^{r+1}$ (for the MNARz model) and by giving $(\tau_{ik})^r$ as prior weights to fit the process.

$$
(\mathcal{J}_k^{\text{MNAR}z})^{r+1} = \left[\begin{array}{c|c} c_{\cdot 1} & 1 \\ \vdots & \vdots \\ c_{\cdot d} & 1 \end{array}\right]. \tag{E.14}
$$

$$
(\mathcal{J}_{kj}^{\text{MNAR}z^j})^{r+1} = \left[c_{\cdot j} \mid 1\right] \tag{E.15}
$$

MNAR$y*$ models For missing scenarios which model the effect of the missingness depending on the variable, the computations are more difficult.

- Because of the dependence of y, $\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r) = \mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik} =$

$1;\theta^r)$ does not hold anymore. Here, one has

$$\begin{aligned}
&\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1, c_i; \theta^r, \psi^r) \\
&= \frac{\prod_{h=1}^d \rho(\alpha_{kh}^r + \beta_{kh}^r y_{ih}^{\text{mis}})^{c_{ih}} (1-\rho(\alpha_{kh}^r + \beta_{kh}^r y_{ih}^{\text{obs}}))^{1-c_{ih}} \mathbb{P}(y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}=1; \theta^r)}{\int_{\mathcal{Y}_i^{\text{mis}}} \prod_{h=1}^d \rho(\alpha_{kh}^r + \beta_{kh}^r y_{ih}^{\text{mis}})^{c_{ih}} (1-\rho(\alpha_{kh}^r + \beta_{kh}^r y_{ih}^{\text{obs}}))^{1-c_{ih}} \mathbb{P}(y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}=1; \theta^r) dy_i^{\text{mis}}} \\
&= \frac{\prod_{h,c_{ih}=1} \rho(\alpha_{kh}^r + \beta_{kh}^r y_{ih}^{\text{mis}}) \mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1; \theta^r)}{\int_{\mathcal{Y}_i^{\text{mis}}} \prod_{h,c_{ih}=1} \rho(\alpha_{kh}^r + \beta_{kh}^r y_{ih}^{\text{mis}}) \mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1; \theta^r) dy_i^{\text{mis}}}.
\end{aligned} \tag{E.16}$$

which implies that Term (a) requires difficult computations if this distribution is not classical.

- For Term (b), it is the same problem, since $\mathbb{P}(c_i \mid y_i, z_{ik}=1;\psi)$ is no longer independent of y, then $E_{ic}^r(\psi)$ requires a specific numerical integration. Using (E.16),

$$\begin{aligned}
E_{ic}^r(\psi) =& \mathbb{E}\left[\log\left(\prod_{j=1}^d \rho(\alpha_{kj}+\beta_{kj}y_{ij})^{c_{ij}}(1-\rho(\alpha_{kj}+\beta_{kj}y_{ij}))^{1-c_{ij}}\right) | y_i^{\text{obs}}, z_{ik}=1, c_i; \theta^r, \psi^r\right] \\
&= \sum_{j=1}^d c_{ij} \int_{\mathcal{Y}_{ij}^{\text{mis}}} \log(\rho(\alpha_{kj}+\beta_{kj}y_{ij}^{\text{mis}})) \mathbb{P}(y_{ij}^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1, c_i; \theta^r, \psi^r) dy_{ij}^{\text{mis}} \\
&\qquad + (1-c_{ij}) \log(1-\rho(\alpha_{kj}+\beta_{kj}y_{ij}^{\text{obs}}))
\end{aligned}$$

where

$$\begin{aligned}
&\mathbb{P}(y_{ij}^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1, c_i; \theta^r, \psi^r) \\
&= \frac{\rho(\alpha_{kj}^r + \beta_{kj}^r y_{ij}^{\text{mis}})^{c_{ij}} (1-\rho(\alpha_{kj}^r + \beta_{kj}^r y_{ij}^{\text{mis}}))^{1-c_{ij}} \mathbb{P}(y_{ij}^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1; \theta^r)}{\int_{\mathcal{Y}_{ij}^{\text{mis}}} \rho(\alpha_{kj}^r + \beta_{kj}^r y_{ij}^{\text{mis}})^{c_{ij}} (1-\rho(\alpha_{kj}^r + \beta_{kj}^r y_{ij}^{\text{mis}}))^{1-c_{ij}} \mathbb{P}(y_{ij}^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1; \theta^r) dy_{ij}^{\text{mis}}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
E_{ic}^r(\psi) = \sum_{j=1}^d c_{ij} \int_{\mathcal{Y}_{ij}^{\text{mis}}} \log(\rho(\alpha_{kj}+\beta_{kj}y_{ij}^{\text{mis}})) \frac{\rho(\alpha_{kj}^r + \beta_{kj}^r y_{ij}^{\text{mis}})^{c_{ij}} \mathbb{P}(y_{ij}^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1; \theta^r)}{\int_{\mathcal{Y}_{ij}^{\text{mis}}} \rho(\alpha_{kj}^r + \beta_{kj}^r x)^{c_{ij}} \mathbb{P}(x \mid y_i^{\text{obs}}, z_{ik}=1; \theta^r) dx} dy_{ij}^{\text{mis}} \\
+ (1-c_{ij}) \log(1-\rho(\alpha_{kj}+\beta_{kj}y_{ij}^{\text{obs}}))
\end{aligned}$$

- There is no closed-form expression for Term (c).

$$\begin{aligned}
&\mathbb{P}(c_{ij} \mid y_i^{\text{obs}}, z_{ik}=1; \theta^r, \psi^r) \\
&= \int_{\mathcal{Y}_{ij}^{\text{mis}}} \mathbb{P}(c_{ij} \mid y_i^{\text{obs}}, y_{ij}^{\text{mis}}, z_{ik}=1; \psi^r) \mathbb{P}(y_{ij}^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1) dy_{ij}^{\text{mis}} \\
&= c_{ij} \int_{-\infty}^{+\infty} \rho(\alpha_{kj}^r + \beta_{kj}^r y_{ij}^{\text{mis}}) \phi(y_{ij}^{\text{mis}}; (\tilde{\mu}_{ik}^{\text{mis}})_j^r, (\tilde{\Sigma}_{ik}^{\text{mis}})_{jj}^r) dy_{ij}^{\text{mis}} + (1-c_{ij})(1-\rho(\alpha_{kj}^r + \beta_{kj}^r y_{ij}^{\text{obs}}))
\end{aligned} \tag{E.17}$$

Using (E.8), the probabilities $(\tau_{ik})^r$ can be deduced from Equation (E.17).

Let us detail the difficulties for two particular cases, if ρ is logistic or probit.

- ρ **is logistic**: Equation (E.16) leads to none classical distribution because

$$\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1, c_i; \theta^r, \psi^r) \propto \prod_{h, c_{ih}=1} \frac{1}{\exp(-(\alpha_{kh}^r + \beta_{kh}^r y_{ih}^{\text{mis}}))} \phi(y_i^{\text{mis}}; (\tilde{\mu}_{ik}^{\text{mis}})^r, (\tilde{\Sigma}_{ik}^{\text{mis}})^r)$$

  Term (b) is

$$E_{ic}^r(\psi) \propto \sum_{j=1}^d c_{ij} \int_{\mathcal{Y}_{ij}^{\text{mis}}} \frac{-\log(1+\exp(-(\alpha_{kj}+\beta_{kj} y_{ij}^{\text{mis}})))}{1+\exp(-(\alpha_{kj}^r + \beta_{kj}^r y_{ij}^{\text{mis}}))} \phi(y_{ij}^{\text{mis}}; (\tilde{\mu}_{ik}^{\text{mis}})_j^r, (\tilde{\Sigma}_{ik}^{\text{mis}})_{jj}^r) dy_{ij}^{\text{mis}} - (1-c_{ij}) \log(1+\exp(\alpha_{kj}+\beta_{kj} y_{ij}^{\text{obs}})),$$

  which amounts to compute the Gaussian moment of $\frac{\log(1+\exp(-u))}{1+\exp(-u)}$, but it has no closed form to our knowledge.

  Finally, Equation (E.17) does not have a closed form either because it requires the computation of

$$\int_{-\infty}^{+\infty} \frac{1}{1+\exp(-(\alpha_{kj}^r + \beta_{kj}^r y_{ij}^{\text{mis}}))} \phi(y_{ij}^{\text{mis}}; (\tilde{\mu}_{ik}^{\text{mis}})_j^r, (\tilde{\Sigma}_{ik}^{\text{mis}})_{jj}^r) dy_{ij}^{\text{mis}},$$

  i.e. the computation of the Gaussian moment of $\frac{1}{1+\exp(-u)}$.

- ρ **is Probit**: One can prove (presented in Appended E.3.2.1) that the conditional distribution $(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1, c_i)$ is a truncated Gaussian, which makes possible the computation of Term (a). Term (b) has no closed form to our knowledge

$$E_{ic}^r(\psi) \propto \sum_{j=1}^d c_{ij} \int_{\mathcal{Y}_{ij}^{\text{mis}}} \frac{\log\left(\int_{-\infty}^{\alpha_{kj}+\beta_{kj} y_{ij}^{\text{mis}}} e^{-t^2} dt\right)}{1+\exp(\alpha_{kj}^r + \beta_{kj}^r y_{ij}^{\text{mis}})} \phi(y_{ij}^{\text{mis}}; (\tilde{\mu}_{ik}^{\text{mis}})_j^r, (\tilde{\Sigma}_{ik}^{\text{mis}})_{jj}^r) dy_{ij}^{\text{mis}} - (1-c_{ij}) \log\left(1 - \int_{-\infty}^{\alpha_{kj}+\beta_{kj} y_{ij}^{\text{obs}}} e^{-t^2} dt\right).$$

  Equation (E.17) does not have a closed form either because it requires the computation of

$$\int_{-\infty}^{+\infty} \left(\int_{-\infty}^{\alpha_{kj}+\beta_{kj} y_{ij}^{\text{mis}}} e^{-t^2} dt\right) \phi(y_{ij}^{\text{mis}}; (\tilde{\mu}_{ik}^{\text{mis}})_j^r, (\tilde{\Sigma}_{ik}^{\text{mis}})_{jj}^r) dy_{ij}^{\text{mis}}.$$

E.3.1.2 Latent class model for categorical data

For categorical data, we have $\phi(y_i; \theta_k) = \prod_{j=1}^d \phi(y_{ij}; \theta_{kj}) = \prod_{j=1}^d \prod_{\ell=1}^{\ell_j} (\theta_{kj}^\ell)^{y_{ij}^\ell}$.

MNARz and MNARz^j models Term (a) is

$$\mathbb{E}\left[\log(\phi(y_i;p_k)) \mid y_i^{\text{obs}}, z_{ik}=1, c_i;\theta^r,\psi^r\right] = \sum_{j,c_{ij}=0}\sum_{\ell=1}^{\ell_j} y_{ij}^{\ell} + \sum_{j,c_{ij}=1}\sum_{\ell=1}^{\ell_j}\log(\theta_{kj}^{y_{ij}^{\ell}}) \tag{E.18}$$

Term (b) is the same as in the Gaussian case given in (E.12). Finally, the EM algorithm can be summarized as follows

E step: For $k=1,\dots,K$ and $i=,\dots,n$, compute

$$\tau_{ik}^r = \frac{\pi_k^r \prod_{j,c_{ij}=0}\prod_{\ell=1}^{\ell_j}(\theta_{kj}^{\ell})^{y_{ij}^{\ell}}\prod_{j=1}^{d}\rho(\alpha_{kj})}{\sum_{h=1}^{K}\pi_h^r\prod_{j,c_{ij}=0}\prod_{\ell=1}^{\ell_j}(\theta_{hj}^{\ell})^{y_{ij}^{\ell}}\prod_{j=1}^{d}\rho(\alpha_{hj})}$$
$$(\tilde{y}_{ij,k}^{\ell})^r = c_{ij}(\theta_{kj}^{\ell})^r + (1-c_{ij})y_{ij}^{\ell},\ \forall j=1,\dots,d,\forall\ell=1,\dots,\ell_j$$

M step: The maximization of $Q_y(\pi,\theta;\pi^r,\theta^r)$ over (π,θ) leads to, for $k=1,\dots,K$,

$$\pi_k^{r+1} = \frac{1}{n}\sum_{i=1}^{n}(\tau_{ik})^r$$
$$(\theta_{kj}^{\ell})^{r+1} = \frac{\sum_{i=1}^{n}(\tau_{ik})^r(\tilde{y}_{ij,k}^{\ell})^r}{\sum_{i=1}^{n}(\tau_{ik})^r},\ \forall j=1,\dots,d,\forall\ell=1,\dots,\ell_j$$

The M-step for ψ consists of performing a GLM with a binomial link and has already been given in detail in Appendix E.3.1.1 (see (E.29) and (E.30)).

E.3.1.3 Combining Gaussian mixture and latent class model for mixed data

If the data are mixed (continuous and categorical), the formulas can be extended straightforwardly if the continuous and the categorical variables are assumed to be independent knowing the latent clusters.

E.3.2 SEM algorithm

The SEM algorithm consists on two steps iteratively proceeded as presented in Section 5.5.2. The key issue is to draw the missing data $(y_i^{\text{mis}})^{r+1}$ and z_i^{r+1} according to their current conditional distribution $\mathbb{P}(y_i^{\text{mis}}, z_i \mid y_i^{\text{obs}}, c_i;\pi^r,\theta^r,\psi^r)$. By convenience, we use a Gibbs sampling and simulate two easier probabilities recalled here

$$z_i^{r+1} \sim \mathbb{P}(\cdot \mid y_i^r, c_i;\pi^r,\theta^r,\psi^r) \quad \text{and} \quad (y_i^{\text{mis}})^{r+1} \sim \mathbb{P}(\cdot \mid y_i^{\text{obs}}, z_i^{r+1}, c_i;\theta^r,\psi^r),$$

where $y_i^r = (y_i^{\text{obs}}, (y_i^{\text{mis}})^r)$. For the latter distribution, the membership k of z_i^{r+1} is drawn from the multinomial distribution with probabilities $(\mathbb{P}(z_{ik}=1 \mid y_i^r, c_i;\theta^r,\psi^r))_{k=1,\dots,K}$ detailed as

follows

$$\begin{aligned}\mathbb{P}(z_{ik}=1 \mid y_i^r, c_i; \pi^r, \theta^r, \psi^r) &= \frac{\mathbb{P}(z_{ik}=1, y_i^r, c_i; \pi^r, \theta^r, \psi^r)}{\mathbb{P}(y_i^r, c_i; \pi^r, \theta^r, \psi^r)} \\ &= \frac{\mathbb{P}(c_i \mid y_i^r, z_{ik}=1; \psi^r)\mathbb{P}(y_i^r \mid z_{ik}=1; \theta^r)\mathbb{P}(z_{ik}=1; \pi^r)}{\sum_{h=1}^K \mathbb{P}(c_i \mid y_i^r, z_{ih}=1; \psi^r)\mathbb{P}(y_i^r \mid z_{ih}=1; \theta^r)\mathbb{P}(z_{ih}=1; \pi^r)} \\ &= \frac{\mathbb{P}(c_i \mid y_i^r, z_{ik}=1; \psi^r)\mathbb{P}(y_i^r \mid z_{ik}=1; \theta^r)\pi_k^r}{\sum_{h=1}^K \mathbb{P}(c_i \mid y_i^r, z_{ih}=1; \psi^r)\mathbb{P}(y_i^r \mid z_{ih}=1; \theta^r)\pi_h^r} \end{aligned} \tag{E.19}$$

The conditional distribution of $((y_i^{\text{mis}})^{r+1} \mid y_i^{\text{obs}}, z_{ik}^{r+1}=1, c_i)$ has already been detailed in Equation (E.16) and recalled here

$$\begin{aligned}&\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1}=1, c_i; \theta^r, \psi^r) \\ &= \frac{\prod_{j,c_{ij}=1} \mathbb{P}(c_{ij}=1 \mid y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}^{r+1}=1; \psi^r)\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1}=1; \theta^r)}{\int_{\mathcal{Y}_i^{\text{mis}}} \prod_{j,c_{ij}=1} \mathbb{P}(c_{ij}=1 \mid y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}^{r+1}=1; \psi^r)\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1}=1; \theta^r) dy_i^{\text{mis}}}\end{aligned} \tag{E.20}$$

E.3.2.1 Gaussian mixture for continuous data

First note that the probabilities of the multinomial distribution for drawing z_i^{r+1} given in (E.19) can be easily computed for all cases.

$$\begin{aligned}&\mathbb{P}(z_{ik}=1 \mid y_i^r, c_i; \pi^r, \theta^r, \psi^r) \\ &= \frac{\prod_{j=1}^d \mathbb{P}(c_{ij}=1 \mid y_i^r, z_{ik}^r=1; \psi^r)^{c_{ij}} \mathbb{P}(c_{ij}=0 \mid y_i^r, z_{ik}^r=1; \psi^r)^{1-c_{ij}} \phi(y_i^r; \mu_k^r, \Sigma_k^r)\pi_k^r}{\sum_{h=1}^K \prod_{j=1}^d \mathbb{P}(c_{ij}=1 \mid y_i^r, z_{ih}^r=1; \psi^r)^{c_{ij}} \mathbb{P}(c_{ij}=0 \mid y_i^r, z_{ih}^r=1; \psi^r)^{1-c_{ij}} \phi(y_i^r; \mu_h^r, \Sigma_h^r)\pi_h^r},\end{aligned}$$

where $\phi(y_i; \mu_k, \Sigma_k)$ is assumed to be a Gaussian distribution with mean vector μ_k and covariance matrix Σ_k, and $\mathbb{P}(c_{ij}=1 \mid y_i^r, z_{ih}^r=1; \psi^r)$ is specified depending the MNAR model and the distribution ρ. The only difficulty of the SE-step is thus to draw from the distribution $(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1}=1, c_i)$.

In the sequel, we detail the distribution $(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1}=1, c_i)$ and the M-step for ψ depending the MNAR model.

MNAR$y*$ models The conditional distribution $(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1}=1, c_i)$ depends on the distribution ρ at hand. For the MNAR$y*$ models, we will consider two classical distributions for ρ: the logistic function and probit one.

- **Logistic distribution**: For the logistic function, the distribution given in (E.20) is not classical and drawing y_i^{mis} from it seems complicated. Indeed, one has

$$\begin{aligned}&\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1}=1, c_i; \theta^r, \psi^r) \\ &\qquad \propto \prod_{j=1, c_{ij}=1} \frac{1}{1+\exp(\alpha_{kj}^r + \beta_{kj}^r y_{ij}^{\text{mis}})} \phi(y_i^{\text{mis}}; (\tilde{\mu}_{ik}^{\text{mis}})^r, (\tilde{\Sigma}_{ik}^{\text{mis}})^r),\end{aligned}$$

where $(\tilde{\mu}_{ik}^{\text{mis}})^r$ and $(\tilde{\Sigma}_{ik}^{\text{mis}})^r$ are given in (E.10) and (E.11). We could use the Sampling Importance Resampling (SIR) algorithm which simulates a realization of $(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1} = 1, c_i)$ with a known instrumental distribution (for example: $(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1} = 1)$) and includes a re-sampling step. However, this algorithm may be computationnaly costly.

- **Probit distribution**: For the probit function, the distribution in (E.20) can be made explicit by using a latent variable L_i.

  More particularly, let L_i such that $L_i = \alpha_k^r + \beta_k^r y_i + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0_d, I_{d\times d})$. Then, c_i can be viewed as an indicator for whether this latent variable is positive, i.e. for all $j = 1, \ldots, d$,

$$c_{ij} = \begin{cases} 1 & \text{if } L_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{E.21}$$

  Thus, indeed to draw $(y_i^{\text{mis}})^{r+1}$ and z_i^{r+1} according to $\mathbb{P}(y_i^{\text{mis}}, z_i \mid y_i^{\text{obs}}, c_i; \pi^r, \theta^r, \psi^r)$, we draw L_i^{r+1}, $(y_i^{\text{mis}})^{r+1}$ and z_i^{r+1} according to $\mathbb{P}(L_i, y_i^{\text{mis}}, z_i \mid y_i^{\text{obs}}, c_i; \pi^r, \theta^r, \psi^r)$ by using a Gibbs sampling.

  – First, we have to draw L_i^{r+1} according to $\mathbb{P}(. \mid y_i^r, z_{ik}^r = 1, c_i; \psi^r)$. One has

$$\begin{aligned}
\mathbb{P}(L_i \mid y_i^r, z_{ik}^r = 1, c_i; \psi_r) &\propto \mathbb{P}(L_i, c_i \mid y_i^r, z_{ik}^r = 1; \psi_r) \\
&\propto \mathbb{P}(c_i \mid L_i^{r+1}, y_i^r, z_{ik}^r = 1; \psi^r)\mathbb{P}(L_i^{r+1} \mid y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}^r = 1; \psi^r) \\
&\overset{(i)}{\propto} \mathbb{P}(c_i \mid L_i^{r+1}; \psi^r)\mathbb{P}(L_i^{r+1} \mid y_i^r, z_{ik}^r = 1; \psi^r) \\
&\overset{(ii)}{=} \mathbb{1}_{\{c_i=1\}\cap\{L_i^{r+1}>0\}}\mathbb{P}(L_i^{r+1} \mid y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}^r = 1; \psi^r)
\end{aligned}$$

  where we use that L_i^{r+1} is a function of $y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik} = 1$ in step (i). Step (ii) is obtained by using (E.21). By abuse of notation, $\{c_i = 1\} \cap \{L_i^{r+1} > 0\}$ means that for all $j = 1, \ldots, d$, $\{c_{ij} = 1\} \cap \{L_{ij}^{r+1} > 0\}$. Finally the conditional distribution $(L_i \mid y_i^r, z_{ik}^r = 1, c_i)$ is a multivariate truncated Gaussian distribution denoted as $\mathcal{N}_t$, as detailed here

$$(L_i \mid y_i^r, z_{ik}^r = 1, c_i) \sim \mathcal{N}_t(\alpha_k^r + \beta_k^r y_i, I_{d\times d}; a, b), \tag{E.22}$$

  with $a \in \mathbb{R}^d$ and $b \in \mathbb{R}^d$ the lower and upper bounds such that for all $j = 1, \ldots, d$,

$$a_j = \begin{cases} 0 & \text{if } c_{ij} = 1, \\ -\infty & \text{otherwise.} \end{cases}$$

$$b_j = \begin{cases} +\infty & \text{if } c_{ij} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

– Secondly, we draw the membership k of z_i^{r+1} from the multinomial distribution with probabilities, for all $k = 1, \dots, K$ detailed as follows

$$\begin{aligned}
&\mathbb{P}(z_{ik} = 1 \mid L_i^{r+1}, y_i^r, c_i; \pi^r, \theta^r, \psi^r) \\
&= \frac{\mathbb{P}(z_{ik} = 1, L_i^{r+1}, y_i^r, c_i; \pi^r, \theta^r, \psi^r)}{\sum_{k=1}^K \mathbb{P}(z_{ik} = 1, L_i^{r+1}, y_i^r, c_i; \pi^r, \theta^r, \psi^r)} \\
&= \frac{\mathbb{P}(L_i^{r+1} | z_{ik} = 1, y_i^r, c_i; \psi^r)\mathbb{P}(z_{ik} = 1, y_i^r, c_i; \pi^r, \theta^r, \psi^r)}{\sum_{k=1}^K \mathbb{P}(L_i^{r+1} | z_{ik} = 1, y_i^r, c_i; \psi^r)\mathbb{P}(z_{ik} = 1, y_i^r, c_i; \pi^r, \theta^r, \psi^r)}
\end{aligned}$$

The part involving $\mathbb{P}(z_{ik} = 1, y_i^r, c_i; \pi^r, \theta^r, \psi^r)$ is given in (E.19) and $\mathbb{P}(L_i^{r+1} | z_{ik} = 1, y_i^r, c_i; \psi^r)$ is only the density of the multivariate truncated Gaussian distribution described in (E.22) evaluated in L_i^{r+1}.

– Finally, y_i^{r+1} is drawn according to $\mathbb{P}(.|L_i^{r+1}, z_{ik}^{r+1} = 1, y_i^{\text{obs}}, c_i; \theta^r, \psi^r)$. One has

$$\begin{aligned}
&\mathbb{P}(y_i^{\text{mis}} \mid L_i^{r+1}, z_{ik}^{r+1} = 1, y_i^{\text{obs}}, c_i; \theta^r, \psi^r) \\
&\propto \mathbb{P}(c_i, L_i^{r+1} \mid y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}^{r+1} = 1; \psi^r)\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1} = 1; \theta^r) \\
&\propto \mathbb{P}(c_i \mid L_i^{r+1}, y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}^{r+1} = 1; \psi^r)\mathbb{P}(L_i^{r+1} \mid y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}^{r+1} = 1; \psi^r)\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1} = 1; \theta^r) \\
&\propto \mathbb{P}(c_i \mid L_i^{r+1}; \psi^r)\mathbb{P}(L_i^{r+1} \mid y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}^{r+1} = 1; \psi^r)\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1} = 1; \theta^r) \\
&\propto \mathbb{P}(L_i^{r+1} \mid y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}^{r+1} = 1; \psi^r)\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1} = 1; \theta^r),
\end{aligned}$$

Yet, one has

$$\begin{aligned}
\mathbb{P}(L_i^{r+1} \mid y_i^{\text{mis}}, y_i^{\text{obs}}, z_{ik}^{r+1} = 1; \psi^r) &\propto e^{-\frac{1}{2}\left[(L_i^{r+1} - (\alpha_k^r + \beta_k^r y_i^{\text{mis}}))^T (L_i^{r+1} - (\alpha_k^r + \beta_k^r y_i))\right]} \\
\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1} = 1; \theta^r) &\propto e^{-\frac{1}{2}\left[(y_i^{\text{mis}} - (\tilde{\mu}_{ik}^{\text{mis}})^{r+1})^T ((\tilde{\Sigma}_{ik}^{\text{mis}})^r)^{-1} (y_i^{\text{mis}} - (\tilde{\mu}_{ik}^{\text{mis}})^{r+1})\right]},
\end{aligned}$$

with $\tilde{\mu}_{ik}^{\text{mis}}$ and $(\tilde{\Sigma}_{ik}^{\text{mis}})^r$ given in (E.9).
Finally combining these two eqution one obtains

$$\left(y_i^{\text{mis}} \mid L_i^{r+1}, z_{ik}^{r+1} = 1, y_i^{\text{obs}}, c_i\right) \sim \mathcal{N}\left(\mu_{ik}^{\text{SEM}}, \Sigma_{ik}^{\text{SEM}}\right),$$

where

$$\Sigma_{ik}^{\text{SEM}} = \left(((\tilde{\Sigma}_{ik}^{\text{mis}})^r)^{-1} + ((\beta_k^{\text{mis}})^r)^T (\beta_k^{\text{mis}})^r\right)^{-1},$$

$$\mu_{ik}^{\text{SEM}} = \Sigma_{ik}^{\text{SEM}} \left[((\tilde{\Sigma}_{ik}^{\text{mis}})^r)^{-1} \tilde{\mu}_{ik}^{\text{mis}} + ((\beta_k^{\text{mis}})^r)^T (L_i^{\text{mis}})^{r+1} - ((\beta_k^{\text{mis}})^r)^T (\alpha_k^{\text{mis}})^r\right],$$

with $(\beta_k^{\text{mis}})^r$ (resp. $(L_i^{\text{mis}})^{r+1}$ and $(\alpha_k^{\text{mis}})^r$) the vector β_k (resp. $(L_i)^{r+1}$ and $(\alpha_k)^r$) restricted to the coordinates $j \in \mathcal{Y}_i^{\text{mis}}$.

Finally, for fully describing the SEM-algorithm given in Algorithm 4, in the M-step, ψ^r is computed using a GLM with a binomial link function for a matrix depending on the MNAR model. In particular,

- For MNARy, the coefficient obtained with a GLM for the matrix $(\mathcal{H}_j^{\text{MNAR}y})^{r+1}$ are α_0 and $\beta_1^{r+1}, \ldots, \beta_d^{r+1}$, with

$$(\mathcal{H}^{\text{MNAR}y})^{r+1} = \left[\begin{array}{c|cccc} c_{.1} & 1 & y_{.1}^{r+1} & 0 & \ldots & 0 \\ c_{.2} & 1 & 0 & y_{.2}^{r+1} & \ldots & 0 \\ \vdots & \vdots & & \ddots & \ddots & \\ c_{.d} & 1 & 0 & 0 & \ldots & y_{.d}^{r+1} \end{array}\right]. \tag{E.23}$$

- For MNARy^k, the coefficient obtained with a GLM for the matrix $(\mathcal{H}_{kj}^{\text{MNAR}y^k})^{r+1}$ is α_0 and $\beta_{11}^{r+1}, \ldots, \beta_{K1}^{r+1}, \ldots, \beta_{Kd}^{r+1}$ with

$$(\mathcal{H}_{kj}^{\text{MNAR}y^k})^{r+1} = \left[\begin{array}{c|cccc} (c_{u1})_{u,z_{u1}^{r+1}=1} & 1 & (y_{u1}^{r+1})_{u,z_{u1}^{r+1}=1} & 0 & \ldots & 0 \\ \vdots & \vdots & & & \ddots & \vdots \\ (c_{u1})_{u,z_{uK}^{r+1}=1} & 1 & 0 & (y_{u1}^{r+1})_{u,z_{uK}^{r+1}=1} & & 0 \\ \vdots & \vdots & \vdots & & \ddots & \\ (c_{ud})_{u,z_{uK}^{r+1}=1} & 1 & 0 & 0 & & (y_{ud}^{r+1})_{u,z_{uK}^{r+1}=1} \end{array}\right]. \tag{E.24}$$

- For MNARyz, the coefficients obtained with a GLM for the matrix $(\mathcal{H}^{\text{MNAR}yz})^{r+1}$ are $\beta_1^{r+1}, \ldots, \beta_d^{r+1}$ and $\alpha_1^{r+1}, \ldots, \alpha_K^{r+1}$, with

$$(\mathcal{H}^{\text{MNAR}yz})^{r+1} = \left[\begin{array}{c|ccccccc} c_{.1} & y_{.1}^{r+1} & 0 & \ldots & 0 & z_{.1}^{r+1} & \ldots & z_{.K}^{r+1} \\ c_{.2} & 0 & y_{.2}^{r+1} & \ldots & 0 & z_{.1}^{r+1} & \ldots & z_{.K}^{r+1} \\ \vdots & & \ddots & \ddots & & \vdots & \vdots & \vdots \\ c_{.d} & 0 & 0 & \ldots & y_{.d}^{r+1} & z_{.1}^{r+1} & \ldots & z_{.K}^{r+1} \end{array}\right]. \tag{E.25}$$

- For MNARyz^j, the coefficients obtained with a GLM for the matrix $(\mathcal{H}_j^{\text{MNAR}yz^j})^{r+1}$ are $\beta_j^{r+1}, \alpha_{1j}^{r+1}, \ldots, \alpha_{Kj}^{r+1}$, with

$$(\mathcal{H}_j^{\text{MNAR}yz^j})^{r+1} = \left[c_{.j} \;\middle|\; y_{.j}^{r+1} \quad z_{.1}^{r+1} \quad \ldots \quad z_{.K}^{r+1}\right]. \tag{E.26}$$

- For MNARy^kz, the coefficients obtained with a GLM for the matrix $(\mathcal{H}_k^{\text{MNAR}y^kz})^{r+1}$ are $\beta_{k1}^{r+1}, \ldots, \beta_{kd}^{r+1}, \alpha_k^{r+1}$, with

$$(\mathcal{H}_k^{\text{MNAR}y^kz})^{r+1} = \left[\begin{array}{c|cccc} c_{u1} & y_{u1}^{r+1} & 0 & \ldots & 0 & 1 \\ c_{u2} & 0 & y_{u2}^{r+1} & \ldots & 0 & 1 \\ \vdots & & \ddots & \ddots & & 1 \\ c_{ud} & 0 & 0 & \ldots & y_{ud}^{r+1} & 1 \end{array}\right]_{u,z_{uk}^{r+1}=1}. \tag{E.27}$$

- For MNAR$y^k z^j$, the coefficients obtained with a GLM for the matrix $(\mathcal{H}_{kj}^{\text{MNAR}y^k z^j})^{r+1}$ are β_{kj}, α_{kj}, with

$$(\mathcal{H}_{kj}^{\text{MNAR}y^k z^j})^{r+1} = \left[c_{uj} \;\middle|\; y_{uj}^{r+1} \quad 1\right]_{u, z_{uk}^{r+1}=1} \tag{E.28}$$

- For MNARz, the coefficients obtained with a GLM for the matrix $(\mathcal{H}^{\text{MNAR}z})^{r+1}$ are $\alpha_1, \ldots, \alpha_K$, with

$$(\mathcal{H}^{\text{MNAR}z})^{r+1} = \left[\begin{array}{c|ccc} c_{\cdot 1} & z_{\cdot 1} & \ldots & z_{\cdot K} \\ \vdots & \vdots & \vdots & \vdots \\ c_{\cdot d} & z_{\cdot 1} & \ldots & z_{\cdot K} \end{array}\right] = \left[\begin{array}{c|ccc} c_{11} & z_{11}^{r+1} & \ldots & z_{1K}^{r+1} \\ \vdots & \vdots & \vdots & \vdots \\ c_{n1} & z_{n1}^{r+1} & \ldots & z_{nK}^{r+1} \\ \vdots & \vdots & \vdots & \vdots \\ c_{1d} & z_{11}^{r+1} & \ldots & z_{1K}^{r+1} \\ \vdots & \vdots & \vdots & \vdots \\ c_{nd} & z_{n1}^{r+1} & \ldots & z_{nK}^{r+1} \end{array}\right]. \tag{E.29}$$

- For MNARz^j, the coefficients obtained with a GLM for the matrix $(\mathcal{H}_j^{\text{MNAR}z^j})^{r+1}$ are $\alpha_{1j}, \ldots, \alpha_{Kj}$, with

$$(\mathcal{H}_j^{\text{MNAR}z^j})^{r+1} = \left[c_{\cdot j} \;\middle|\; z_{\cdot 1}^{r+1} \quad \ldots \quad z_{\cdot K}^{r+1}\right] \tag{E.30}$$

MNARz and MNARz^j models For the MNARz and MNARz^j models, the effect of the missingness is only due to the class membership. We have already proved in Appendix E.3.1.1 that

$$\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_i^r, c_i; \theta^r, \psi^r) = \mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_i^r; \theta^r),$$

and that this conditional distribution is Gaussian given in (E.9). The M-step for ψ has been specified in the previous paragraph with (E.29) and (E.30).

E.3.2.2 Latent class model for categorical data

For categorical data, we have $\phi(y_i; \theta_k) = \prod_{j=1}^d \phi(y_{ij}; \theta_{kj}) = \prod_{j=1}^d \prod_{\ell=1}^{\ell_j} (\theta_{kj}^\ell)^{y_{ij}^\ell}$.

MNARz and MNARz^j models For drawing from the conditional distribution $(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1} = 1)$, by independence of the features conditionally to the membership, we can draw for $j = 1, \ldots, d$ $y_{ij}^{\text{mis}} = ((y_{ij}^{\text{mis}})^1, \ldots, (y_{ij}^{\text{mis}})^{l_j})$ from the conditional distribution $(y_{ij}^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{r+1} = 1)$. This latter is a multinomial distribution with probabilities $(\theta_{kj}^\ell)_{\ell=1,\ldots,\ell_j}$.

Bibliography

N. Abiri, B. Linse, P. Edén, and M. Ohlsson. Establishing strong imputation performance of a denoising autoencoder in a wide range of missing data problems. *Neurocomputing*, 365: 137–146, 2019.

E. S. Allman, C. Matias, J. A. Rhodes, et al. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

T. W. Anderson. *An Introduction to Multivariate Statistical sAnalysis, 3rd edition*. Wiley, 2003.

R. R. Andridge and R. J. Little. A review of hot deck imputation for survey non-response. *International statistical review*, 78(1):40–64, 2010.

V. Audigier, F. Husson, and J. Josse. A principal component method to impute missing values for mixed data. *Advances in Data Analysis and Classification*, 10(1):5–26, 2016a.

V. Audigier, F. Husson, and J. Josse. Multiple imputation for continuous variables using a bayesian principal component analysis. *Journal of Statistical Computation and Simulation*, 86(11):2140–2156, 2016b. doi: 10.1080/00949655.2015.1104683. URL https://doi.org/10.1080/00949655.2015.1104683.

F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in neural information processing systems*, pages 773–781, 2013.

S. G. Baker and N. M. Laird. Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of the American Statistical association*, 83 (401):62–69, 1988.

J. D. Banfield and A. E. Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.

J.-P. Baudry et al. Estimation and model selection for model-based clustering with the conditional classification likelihood. *Electronic journal of statistics*, 9(1):1041–1077, 2015.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

R. M. Bell and Y. Koren. Lessons from the netflix prize challenge. *Acm Sigkdd Explorations Newsletter*, 9(2):75–79, 2007.

J. Bennett, S. Lanning, et al. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. Citeseer, 2007.

C. Beunckens, G. Molenberghs, G. Verbeke, and C. Mallinckrodt. A latent-class mixture model for incomplete longitudinal gaussian data. *Biometrics*, 64(1):96–105, 2008.

R. Bhattacharya, R. Nabi, I. Shpitser, and J. M. Robins. Identification in missing data models represented by directed acyclic graphs. In *Uncertainty in Artificial Intelligence*, pages 1149–1158. PMLR, 2020.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:719–725, 2000.

F. Biessmann, D. Salinas, S. Schelter, P. Schmidt, and D. Lange. "deep" learning for missing value imputationin tables with non-numerical data. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pages 2017–2025, 2018. ISBN 978-1-4503-6014-2. doi: 10.1145/3269206.3272005. URL http://doi.acm.org/10.1145/3269206.3272005.

M. Bogdan, E. Van Den Berg, C. Sabatti, W. Su, and E. J. Candès. Slope—adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3):1103, 2015.

C. Bouveyron and C. Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78, 2014.

C. Bouveyron, S. Girard, and C. Schmid. High-dimensional data clustering. *Computational Statistics & Data Analysis*, 52(1):502–519, 2007.

C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery. *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press, 2019.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

S. v. Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

T. Cai and W.-X. Zhou. A max-norm constrained minimization approach to 1-bit matrix completion. *The Journal of Machine Learning Research*, 14(1):3619–3647, 2013.

E. J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6): 925–936, 2010.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

E. J. Candès, C. A. Sing-Long, and J. D. Trzasko. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Transactions on Signal Processing*, 61(19): 4643–4657, 2013.

O. Cappé and E. Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3): 593–613, 2009.

J. Carpenter and M. Kenward. *Multiple Imputation and its Application*. John Wiley & Sons, Dec. 2012.

G. Celeux and J. Diebolt. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2: 73–82, 1985.

G. Celeux, D. Chauveau, and J. Diebolt. Stochastic versions of the em algorithm: an experimental study in the mixture case. *Journal of statistical computation and simulation*, 55(4):287–314, 1996.

G. Celeux, S. Frühwirth-Schnatter, and C. P. Robert. Models selection for mixture models. In S. Frühwirth-Schnatter, G. Celeux, , and C. P. Robert, editors, *Handbook of Mixture Analysis*, pages 117–154. CRC Press, 2019.

Y. Chen and C. Caramanis. Noisy and missing data regression: Distribution-oblivious support recovery. In *International Conference on Machine Learning*, pages 383–391, 2013.

Y. Chen, C. Caramanis, and S. Mannor. Robust sparse regression under adversarial corruption. In *International Conference on Machine Learning*, pages 774–782. PMLR, 2013.

X. Cheng, D. Cook, and H. Hofmann. Visually exploring missing values in multivariable data using a graphical user interface. *Journal of statistical software*, 68(1):1–23, 2015.

J. T. Chi, E. C. Chi, and R. G. Baraniuk. k-pod: A method for k-means clustering of missing data. *The American Statistician*, 70(1):91–99, 2016.

N. R. Council et al. Principles and methods of sensitivity analyses. In *The Prevention and Treatment of Missing Data in Clinical Trials*. National Academies Press (US), 2010.

D. R. Cox and D. V. Hinkley. *Theoretical statistics*. CRC Press, 1979.

A. Creemers, N. Hens, M. Aerts, G. Molenberghs, G. Verbeke, and M. G. Kenward. A sensitivity analysis for shared-parameter models for incomplete longitudinal outcomes. *Biometrical Journal*, 52(1):111–125, 2010.

A. S. Dalalyan and P. Thompson. Outlier-robust estimation of a sparse linear model using ℓ_1-penalized huber's m-estimator. *arXiv preprint arXiv:1904.06288*, 2019.

A. Datta, H. Zou, et al. Cocolasso for high-dimensional error-in-variables regression. *Annals of Statistics*, 45(6):2400–2426, 2017.

I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1):1–38, 2010. ISSN 1097-0312. doi: 10.1002/cpa.20303. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20303.

M. D. R. De Chaumaray and M. Marbac. Clustering data with nonignorable missingness using semi-parametric mixture models. *arXiv preprint arXiv:2009.07662*, 2020.

B. Delyon, M. Lavielle, E. Moulines, et al. Convergence of a stochastic approximation version of the em algorithm. *The Annals of Statistics*, 27(1):94–128, 1999.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

P. Descloux and S. Sardy. Model selection with lasso-zero: adding straw to the haystack to better find needles. *Journal of Computational and Graphical Statistics*, pages 1–29, 2020.

A. Dieuleveut, N. Flammarion, and F. Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1): 3520–3570, 2017.

A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size SGD and markov chains. *Ann. Statist.*, 48, 2020. doi: 10.1214/19-AOS1850. URL https://doi.org/10.1214/19-AOS1850.

D. Dua and C. Graff. UCI machine learning repository, 2019. URL http://archive.ics.uci.edu/ml.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

X. d'Haultfoeuille. A new instrumental method for dealing with endogenous selection. *Journal of Econometrics*, 154(1):1–15, 2010.

C. K. Enders. *Applied missing data analysis*. Guilford press, 2010.

N. Erler. Bayesian imputation of missing covariates. 2019.

N. S. Erler, D. Rizopoulos, and E. M. Lesaffre. Jointai: joint analysis and imputation of incomplete data in r. *arXiv preprint arXiv:1907.10867*, 2019.

S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Number 3. Birkhäuser Basel, 2013. URL http://www.ams.org/bull/2017-54-01/S0273-0979-2016-01546-1/.

G. Frisch, J.-B. Léger, and Y. Grandvalet. Learning from missing data with the latent block model. *arXiv preprint arXiv:2010.12222*, 2020.

R. I. Garcia, J. G. Ibrahim, and H. Zhu. Variable selection for regression models with missing data. *Statistica Sinica*, 20(1):149, 2010.

M. Gavish and D. L. Donoho. Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, 63(4):2137–2152, 2017.

A. Gelman and J. Hill. Opening windows to the black box. *Journal of Statistical Software*, 40, 2011.

J. Geweke, M. Keane, and D. Runkle. Alternative computational approaches to inference in the multinomial probit model. *The review of economics and statistics*, pages 609–632, 1994.

C. Giacobino, S. Sardy, J. Diaz-Rodriguez, and N. Hengartner. Quantile universal threshold. *Electronic Journal of Statistics*, 11(2):4701–4722, 2017. ISSN 1935-7524. doi: 10.1214/17-EJS1366. URL https://projecteuclid.org/euclid.ejs/1511492459.

W. R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):337–348, 1992.

R. J. Glynn, N. M. Laird, and D. B. Rubin. Selection modeling versus mixture modeling with nonignorable nonresponse. In *Drawing inferences from self-selected samples*, pages 115–142. Springer, 1986.

L. Gondara and K. Wang. Mida: Multiple imputation using denoising autoencoders. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 260–272. Springer, 2018.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

N. J. Gordon, D. J. Salmond, and A. F. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE proceedings F (radar and signal processing)*, volume 140, pages 107–113. IET, 1993.

J. W. Graham. Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60:549–576, 2009.

J. W. Graham, S. M. Hofer, and A. M. Piccinin. Analysis with missing data in drug prevention research. *NIDA research monograph*, 142:13–13, 1994.

M. Gürbüzbalaban, A. Ozdaglar, and P. Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, pages 1–36, 2015.

M. Hahsler. recommenderlab: A framework for developing and testing recommendation algorithms. Technical report, 2015.

S. R. Hamada, T. Gauss, F.-X. Duchateau, J. Truchot, A. Harrois, M. Raux, J. Duranteau, J. Mantz, and C. Paugam-Burtz. Evaluation of the performance of french physician-staffed emergency medical service in the triage of major trauma patients. *Journal of Trauma and Acute Care Surgery*, 76(6):1476–1483, 2014.

S. R. Hamada, T. Gauss, J. Pann, M. Dünser, M. Leone, and J. Duranteau. European trauma guideline compliance assessment: the etrauss study. *Critical care*, 19(1):1–8, 2015.

O. Harel and J. L. Schafer. Partial and latent ignorability in missing-data problems. *Biometrika*, 96(1):37–50, 2009.

T. Hastie. Ridge regularization: An essential concept in data science. *Technometrics*, 62(4): 426–433, 2020.

T. Hastie and R. Mazumder. *softImpute: Matrix Completion via Iterative Soft-Thresholded SVD*, 2015. URL https://CRAN.R-project.org/package=softImpute. R package version 1.4.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

T. Hastie, R. Mazumder, J. D. Lee, and R. Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402, 2015.

S. I. Hay, A. A. Abajobir, K. H. Abate, C. Abbafati, K. M. Abbas, F. Abd-Allah, R. S. Abdulkader, A. M. Abdulle, T. A. Abebo, S. F. Abera, et al. Global, regional, and national disability-adjusted life-years (dalys) for 333 diseases and injuries and healthy life expectancy (hale) for 195 countries and territories, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet*, 390(10100):1260–1344, 2017.

E. Hazan and S. Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 421–436, 2011.

J. J. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4*, pages 475–492. NBER, 1976.

J. J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

J. M. Hernández-Lobato, N. Houlsby, and Z. Ghahramani. Probabilistic matrix factorization with non-random missing data. In *International Conference on Machine Learning*, pages 1512–1520, 2014.

J. Honaker, G. King, M. Blackwell, et al. Amelia ii: A program for missing data. *Journal of statistical software*, 45(7):1–47, 2011.

L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

L. Hunt and M. Jorgensen. Mixture model clustering for mixed data with missing information. *Computational Statistics and Data Analysis*, 41:429–440, 2003.

J. G. Ibrahim. Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85(411):765–769, 1990.

J. G. Ibrahim, S. R. Lipsitz, and M.-H. Chen. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):173–190, 1999.

J. G. Ibrahim, M.-H. Chen, and S. R. Lipsitz. Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, 88(2):551–564, 2001.

R. Ihaka. R: Past and future history. *Computing Science and Statistics*, 392396, 1998.

A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values. *The Journal of Machine Learning Research*, 11:1957–2000, 2010.

S. Ilya, M. Karthika, and P. Judea. Missing data as a causal and probabilistic problem in proceedings of the thirty first conference on uncertainty in artificial intelligence (uai-15), 2015.

N. B. Ipsen, P.-A. Mattei, and J. Frellsen. not-miwae: Deep generative modelling with missing not at random data. *arXiv preprint arXiv:2006.12871*, 2020.

P. Jain, D. Nagaraj, and P. Netrapalli. SGD without Replacement: Sharper Rates for General Smooth Convex Functions. *arXiv e-prints*, art. arXiv:1903.01463, Mar 2019.

W. Jiang, M. Bogdan, J. Josse, B. Miasojedow, V. Rockova, and T. Group. Adaptive bayesian slope–high-dimensional model selection with missing values. *arXiv preprint arXiv:1909.06631*, 2019.

W. Jiang, J. Josse, M. Lavielle, and T. Group. Logistic regression with missing covariates—parameter estimation, model selection and prediction within a joint-modeling framework. *Computational Statistics & Data Analysis*, 145:106907, 2020.

I. T. Jolliffe. Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer, 1986.

M. P. Jones. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American statistical association*, 91(433):222–230, 1996.

M. Jorgensen and L. Hunt. Mixture model clustering of data sets with categorical and continuous variables. In *Proceedings of the Conference ISIS*, volume 96, pages 375–384. World Scientific, 1996.

J. Josse and F. Husson. Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis*, 56(6): 1869–1879, 2012.

J. Josse and J. P. Reiter. Introduction to the special section on missing data. *Statistical Science*, 33(2):139–141, 2018.

J. Josse, J. Pagès, and F. Husson. Testing the significance of the rv coefficient. *Computational Statistics & Data Analysis*, 53(1):82–91, 2008.

J. Josse, F. Husson, et al. missmda: a package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31, 2016a.

J. Josse, S. Sardy, and S. Wager. denoiser: A package for low rank matrix estimation. *Journal of Statistical Software*, 2016b.

J. Josse, N. Prost, E. Scornet, and G. Varoquaux. On the consistency of supervised learning with missing values. *arXiv preprint arXiv:1902.06931*, 2019.

A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

N. Kallus, X. Mao, and M. Udell. Causal inference with noisy and missing covariates via matrix factorization. *arXiv preprint arXiv:1806.00811*, 2018.

M. G. Kenward. Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statistics in medicine*, 17(23):2723–2732, 1998.

J. K. Kim and J. Shao. *Statistical methods for handling incomplete data*. CRC press, 2013.

N. Kishore Kumar and J. Schneider. Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra*, 65(11):2212–2244, 2017.

V. Koltchinskii, K. Lounici, A. B. Tsybakov, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

A. Kowarik and M. Templ. Imputation with the R package VIM. *Journal of Statistical Software*, 74(7):1–16, 2016. doi: 10.18637/jss.v074.i07.

J. Kuha, M. Katsikatsou, and I. Moustaki. Latent variable modelling with non-ignorable item nonresponse: multigroup response propensity models for cross-national analysis. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 181(4):1169–1192, 2018.

J. N. Laska, M. A. Davenport, and R. G. Baraniuk. Exact signal recovery from sparsely corrupted measurements through the Pursuit of Justice. In *2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers*, pages 1556–1560, 2009. doi: 10.1109/ACSSC.2009.5470141.

B. Laurent and P. Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1015957395. URL https://projecteuclid.org/euclid.aos/1015957395.

M. Le Morvan, J. Josse, T. Moreau, E. Scornet, and G. Varoquaux. Neumiss networks: differential programming for supervised learning with missing values. In *Advances in Neural Information Processing Systems 33*, 2020a.

M. Le Morvan, N. Prost, J. Josse, E. Scornet, and G. Varoquaux. Linear predictor on linearly-generated data with missing values: non consistency and solutions. In *International Conference on Artificial Intelligence and Statistics*, pages 3165–3174. PMLR, 2020b.

J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

B. Leurent, M. Gomes, R. Faria, S. Morris, R. Grieve, and J. R. Carpenter. Sensitivity analysis for not-at-random missing data in trial-based cost-effectiveness analysis: a tutorial. *PharmacoEconomics*, 36(8):889–901, 2018.

X. Li. Compressed Sensing and Matrix Completion with Constant Proportion of Corruptions. *Constructive Approximation*, 37(1):73–99, 2013. ISSN 1432-0940. doi: 10.1007/s00365-012-9176-9. URL https://doi.org/10.1007/s00365-012-9176-9.

Z. Li, F. Wu, and J. Wright. On the systematic measurement matrix for compressed sensing in the presence of gross errors. In *2010 Data Compression Conference*, pages 356–365. IEEE, 2010.

R. J. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

R. J. Little and D. B. Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

L. T. Liu, E. Dobriban, A. Singer, et al. *e* pca: High dimensional exponential family pca. *The Annals of Applied Statistics*, 12(4):2121–2150, 2018.

Y. Liu, Y. Wang, Y. Feng, and M. M. Wall. Variable selection and prediction with incomplete high-dimensional data. *The annals of applied statistics*, 10(1):418, 2016.

Y. Liu, P. Li, and J. Qin. Full-semiparametric-likelihood-based inference for non-ignorable missing data. *arXiv preprint arXiv:1908.01260*, 2019.

P.-L. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. In *Advances in Neural Information Processing Systems*, pages 2726–2734, 2011.

P.-L. Loh and M. J. Wainwright. High-Dimensional Regression with Noisy and Missing Data: Provable Guarantees with Nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012. ISSN 0090-5364. URL http://www.jstor.org/stable/41713688.

A. Ma and D. Needell. Stochastic gradient descent for linear systems with missing data. *Numerical Mathematics: Theory, Methods and Applications*, 12(1):1–20, 2018. ISSN 2079-7338. doi: https://doi.org/10.4208/nmtma.OA-2018-0066. URL http://global-sci.org/intro/article_detail/nmtma/12689.html.

W. Ma and G. H. Chen. Missing not at random in matrix completion: The effectiveness of estimating missingness probabilities under a low nuclear norm assumption. In *Advances in Neural Information Processing Systems*, pages 14871–14880, 2019.

S. Makridakis, E. Spiliotis, and V. Assimakopoulos. The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808, 2018.

M. Marbac, C. Biernacki, and V. Vandewalle. Model-based clustering of gaussian copulas for mixed data. *Communications in Statistics-Theory and Methods*, 46(23):11635–11656, 2017.

B. M. Marlin and R. S. Zemel. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the third ACM conference on Recommender systems*, pages 5–12, 2009.

P.-A. Mattei and J. Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International Conference on Machine Learning*, pages 4413–4423. PMLR, 2019.

I. Mayer, E. Sverdrup, T. Gauss, J.-D. Moyer, S. Wager, and J. Josse. Doubly robust treatment effect estimation with missing attributes. *Ann. Appl. Statist.*, 14(3):1409–1431, 2020. ISSN 1932-6157. doi: 10.1214/20-AOAS1356.

R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.

G. J. McLachlan and K. E. Basford. *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York, 1988.

G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

D. McParland and I. C. Gormley. Model based clustering for mixed data: clustmd. *Advances in Data Analysis and Classification*, 10(2):155–169, 2016.

X.-L. Meng and D. B. Rubin. Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association*, 86(416):899–909, 1991.

X.-L. Meng and D. B. Rubin. Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

W. Miao and E. T. Tchetgen. Identification and inference with nonignorable missing covariate data. *Statistica Sinica*, 28(4):2049, 2018.

W. Miao and E. J. Tchetgen Tchetgen. On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, 103(2):475–482, 2016.

W. Miao, L. Liu, E. T. Tchetgen, and Z. Geng. Identification, doubly robust estimation, and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *arXiv preprint arXiv:1509.02556*, 2015.

W. Miao, P. Ding, and Z. Geng. Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 111(516): 1673–1683, 2016.

K. Mohan. On handling self-masking and other hard missing data problems. 2018.

K. Mohan and J. Pearl. Graphical models for recovering probabilistic and causal queries from missing data. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/31839b036f63806cba3f47b93af8ccb5-Paper.pdf.

K. Mohan and J. Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, pages 1–42, 2021.

K. Mohan, J. Pearl, and J. Tian. Graphical models for inference with missing data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/0ff8033cf9437c213ee13937b1c4c455-Paper.pdf.

K. Mohan, F. Thoemmes, and J. Pearl. Estimation with incomplete data: The linear case. In *IJCAI*, pages 5082–5088, 2018.

G. Molenberghs, C. Beunckens, C. Sotto, and M. G. Kenward. Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society B*, 70:371–388, 2008.

G. Molenberghs, G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke. *Handbook of missing data methodology*. CRC Press, 2014.

K. Morikawa, J. K. Kim, and Y. Kano. Semiparametric maximum likelihood estimation with data missing not at random. *Canadian Journal of Statistics*, 45(4):393–409, 2017.

S. Moritz and T. Bartz-Beielstein. imputeTS: Time Series Missing Value Imputation in R. *The R Journal*, 9(1):207–218, 2017. doi: 10.32614/RJ-2017-009.

E. Moulines and F. R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

J. S. Murray and J. P. Reiter. Multiple imputation of missing categorical and continuous values via bayesian mixture models with local dependence. *Journal of the American Statistical Association*, 111(516):1466–1479, 2016.

J. S. Murray et al. Multiple imputation: a review of practical and theoretical findings. *Statistical Science*, 33(2):142–159, 2018.

B. Muzellec, J. Josse, C. Boyer, and M. Cuturi. Missing data imputation using optimal transport. In *International Conference on Machine Learning*, pages 7130–7140. PMLR, 2020.

R. Nabi, R. Bhattacharya, and I. Shpitser. Full law identification in graphical models of missing data: Completeness results. In *International Conference on Machine Learning*, pages 7153–7163. PMLR, 2020.

D. Needell, R. Ward, and N. Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in neural information processing systems*, pages 1017–1025, 2014.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

N. H. Nguyen and T. D. Tran. Robust Lasso With Missing and Grossly Corrupted Observations. *IEEE Transactions on Information Theory*, 59(4):2036–2058, 2013a. ISSN 0018-9448. doi: 10.1109/TIT.2012.2232347.

N. H. Nguyen and T. D. Tran. Exact Recoverability From Dense Corrupted Observations via ℓ_1-Minimization. *IEEE Transactions on Information Theory*, 59(4):2017–2035, 2013b. ISSN 0018-9448. doi: 10.1109/TIT.2013.2240435.

S. F. Nielsen et al. The stochastic em algorithm: estimation and asymptotic results. *Bernoulli*, 6(3):457–489, 2000.

A. Novo and J. Schafer. Norm: analysis of multivariate normal datasets with missing values. r package version 1.0-9.5. *Vienna, Austria: R Foundation for Statistical Computing*, 2013.

A. Oliver, A. Odena, C. Raffel, E. D. Cubuk, and I. J. Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *arXiv preprint arXiv:1804.09170*, 2018.

J. Pearl. Causality: models, reasoning, and inference. *Econometric Theory*, 19(675-685):46, 2003.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

M. Ramoni, P. Sebastiani, and P. Cohen. Bayesian clustering by dynamics. *Machine learning*, 47(1):91–121, 2002.

H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

G. Robin. *Low-rank methods for heterogeneous and multi-source data*. PhD thesis, Université Paris-Saclay (ComUE), 2019.

G. Robin, O. Klopp, J. Josse, É. Moulines, and R. Tibshirani. Main effects and interactions in mixed and incomplete data frames. *Journal of the American Statistical Association*, 115(531):1292–1303, 2020.

M. Rosenbaum, A. B. Tsybakov, et al. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.

M. Rosenbaum, A. B. Tsybakov, et al. Improved matrix uncertainty selector. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 276–290. Institute of Mathematical Statistics, 2013.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

D. B. Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. *arXiv:1003.2990 [math]*, 2010. URL http://arxiv.org/abs/1003.2990. arXiv: 1003.2990.

M. Sadinle and J. P. Reiter. Itemwise conditionally independent nonresponse modelling for incomplete multivariate data. *Biometrika*, 104(1):207–220, 2017.

M. Sadinle and J. P. Reiter. Sequentially additive nonignorable missing data modelling using auxiliary marginal information. *Biometrika*, 106(4):889–911, 2019.

S. Sardy, P. Tseng, and A. G. Bruce. Robust wavelet denoising. *IEEE Transactions on Signal Processing*, 49:1146–1152, 2001.

J. L. Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.

J. L. Schafer and J. W. Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

S. Seaman, J. Galati, D. Jackson, and J. Carlin. What is meant by" missing at random"? *Statistical Science*, pages 257–268, 2013.

S. R. Seaman and I. R. White. Review of inverse probability weighting for dealing with missing data. *Statistical methods in medical research*, 22(3):278–295, 2013.

A. Serafini, T. B. Murphy, and L. Scrucca. Handling missing data in model-based clustering. *arXiv preprint arXiv:2006.02954*, 2020.

O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

J. Shao and L. Wang. Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, 103(1):175–187, 2016.

I. Shpitser. Consistent estimation of functions of data missing non-monotonically and not at random. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3152–3160. Citeseer, 2016.

A. Sportisse, C. Boyer, and J. Josse. Imputation and low-rank estimation with missing not at random data. *Statistics and Computing*, 30(6):1629–1643, 2020.

StataCorp. *Stata Statistical Software: Release 16*. StataCorp LLC., College Station, TX, 2019.

D. J. Stekhoven and P. Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

T. Tabouy, P. Barbillon, and J. Chiquet. Variational inference for stochastic block models from sampled data. *Journal of the American Statistical Association*, 115(529):455–466, 2020.

F. Tang and H. Ishwaran. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):363–377, 2017.

G. Tang, R. J. Little, and T. E. Raghunathan. Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika*, 90(4):747–764, 2003.

N. Tang and Y. Ju. Statistical inference for nonignorable missing-data problems: a selective review. *Statistical Theory and Related Fields*, 2(2):105–133, 2018.

N. Tang, P. Zhao, and H. Zhu. Empirical likelihood for estimating equations with nonignorably missing data. *Statistica Sinica*, 24(2):723, 2014.

M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540, 1987.

P. Tardivel and M. Bogdan. On the sign recovery by lasso, thresholded lasso and thresholded basis pursuit denoising. 2019.

H. Teicher. Identifiability of finite mixtures. *The annals of Mathematical statistics*, pages 1265–1269, 1963.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

N. Tierney, D. Cook, M. McBain, and C. Fay. *naniar: Data Structures, Summaries, and Visualisations for Missing Data*. URL https://github.com/njtierney/naniar. R package version 0.2.0.

N. J. Tierney and D. H. Cook. Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. *arXiv preprint arXiv:1809.02264*, 2018.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

A. B. Tsybakov. Optimal rates of aggregation. In *Learning theory and kernel machines*, pages 303–313. Springer, 2003.

B. Twala, M. Jones, and D. J. Hand. Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, 29(7):950–956, 2008.

M. Udell and A. Townsend. Nice latent variable models have log-rank. *ArXiv*, abs/1705.07474, 2017. URL http://arxiv.org/abs/1705.07474.

M. Udell and A. Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019.

M. Udell, C. Horn, R. Zadeh, S. Boyd, et al. Generalized low rank models. *Foundations and Trends® in Machine Learning*, 9(1):1–118, 2016.

S. Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.

G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

M. Verbanck, J. Josse, and F. Husson. Regularised pca to denoise and visualise data. *Statistics and Computing*, 25(2):471–486, 2015.

S. Wang, J. Shao, and J. K. Kim. An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, pages 1097–1116, 2014.

X. Wang, R. Zhang, Y. Sun, and J. Qi. Doubly robust joint learning for recommendation on data missing not at random. In *International Conference on Machine Learning*, pages 6638–6647, 2019.

J. Wright and Y. Ma. Dense Error Correction Via ℓ_1-Minimization. *IEEE Transactions on Information Theory*, 56(7):3540–3560, 2010. ISSN 0018-9448. doi: 10.1109/TIT.2010.2048473.

J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust Face Recognition via Sparse Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009. ISSN 0162-8828. doi: 10.1109/TPAMI.2008.79.

Y. Xie, A. Presmanes Hill, and A. Thomas. *blogdown: Creating Websites with R Markdow*. The R Series. Chapman and Hall/CRC, 2017. ISBN 978-0815363729.

Y. Xiong and D.-Y. Yeung. Time series clustering with arma mixtures. *Pattern Recognition*, 37(8):1675–1689, 2004.

S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, pages 209–214, 1968.

C. Yang, Y. Akimoto, D. W. Kim, and M. Udell. Oboe: Collaborative filtering for automl initialization. *arXiv preprint arXiv:1808.03233*, 2018.

J. Yi, J. Lee, K. J. Kim, S. J. Hwang, and E. Yang. Why not to use zero imputation? correcting sparsity bias in training neural networks. *arXiv preprint arXiv:1906.00150*, 2019.

J. Yoon, J. Jordon, and M. Schaar. Gain: Missing data imputation using generative adversarial nets. In *International Conference on Machine Learning*, pages 5689–5698. PMLR, 2018.

J. You, X. Ma, D. Y. Ding, M. Kochenderfer, and J. Leskovec. Handling missing data with graph representation learning. *arXiv preprint arXiv:2010.16418*, 2020.

S. Zhang. Nearest neighbor selection for iteratively knn imputation. *Journal of Systems and Software*, 85(11):2541–2552, 2012.

J. Zhao and Y. Ma. A versatile estimation procedure without estimating the nonignorable missingness mechanism. *Journal of the American Statistical Association*, pages 1–44, 2021.

J. Zhao and J. Shao. Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data. *Journal of the American Statistical Association*, 110(512): 1577–1590, 2015.

S. Zhong and J. Ghosh. A unified framework for model-based clustering. *Journal of machine learning research*, 4(Nov):1001–1037, 2003.

D. L. Zhu and P. Marcotte. Co-coercivity and its role in the convergence of iterative schemes for solving variational inequalities. *SIAM Journal on Optimization*, 6(3):714–726, 1996.

Z. Zhu, T. Wang, and R. J. Samworth. High-dimensional principal component analysis with heterogeneous missingness. *arXiv preprint arXiv:1906.12125*, 2019.