Estimation with informative missing data in the low-rank model with random effects

Aude Sportisse, Claire Boyer, Julie Josse

Estimation with informative missing data in the low-rank model with random effects

Aude Sportisse, Claire Boyer, Julie Josse

June 4, 2019

Abstract

Matrix completion based on low-rank models is very popular and comes with powerful algorithms and theoretical guarantees. However, existing methods do not consider the case of values missing not at random (MNAR) which are widely encountered in practice. Considering a data matrix generated from a probabilistic principal component analysis (PPCA) model containing several MNAR variables, we propose estimators for the means, variances and covariances related to the MNAR missing variables and study their consistency. The proposed estimators present the advantage of being computed without explicitly modeling the MNAR mechanism and by only using observed data. In addition, we propose an imputation method of the data matrix and an estimation of the PPCA loading matrix. We compare our proposal with the classical methods used in low-rank models, as iterative methods based on singular value decomposition.

Keywords— graphical models, probabilistic principal component analysis, informative missing values, matrix completion, latent variables.

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 [18] introduced three missing-data mechanisms. The data are said (i) Missing Completely At Random (MCAR) if the probability of being missing for one observation is the same for all observations, (ii) Missing At Random (MAR) if the probability of being missing only depends on the value of observed variables, (iii) Missing Not At Random if the unavailability of the data depends on the values of other variables and its value itself. We focus on this later case, which is extremely frequent in practice. A classic example of MNAR data is surveys where rich people would be less willing to disclose their income.

When the data are MCAR or MAR, statistical inference is realized by ignoring the missing-data mechanism [12]. In the MNAR case, the observed variables are not representative of the population which leads to bias in the estimation. Consequently, it is usually necessary to take into account the specific distribution of the missing data. Often, the missing values mechanism distribution is assumed to be logistic (see for instance [7] in the case of parametric generalized linear models but also [7, 16, 20]). This is often associated with an important computational burden to perform inference.

Recently, for the specific case of linear model, Mohan et al. [15] proposed an approach based on graphical models to handle self-masked MNAR variable, *i.e.* where the unavailability of data only

depends on the value of the variable itself. In this context, they proved that the mean of the variable with missing values can be consistently estimated by only using the observed information and without explicitly modeling the missing values mechanism. In addition, they also proposed a method for estimating the variance of the variable.
In this work, we focus on low-rank models with MNAR data. Low-rank models has become very popular in recent years [11, 6], known as a very powerful solution for dealing with missing values [9] but their theoretical guarantees are only valid if the data are MCAR or MAR [10, 1]. [19] suggested a parametric approach for handling MNAR data in the low-rank fixed effect model also modeling the missing values mechanism with a logistic regression. Although this approach leads to accurate recovery of the low-rank structure and missing entries, it can be computationally expensive and relies on strong parametric assumptions.

Contributions. Assuming a probabilistic PCA (PPCA) model [21], we prove that the mean, the variance and the covariance of the missing variables can be consistently estimated in the MNAR case, without modeling the missing-data mechanism and by only using observed data. To our knowledge, this result is the first one to ensure a consistent estimate on the informative missing data in a low-rank model with random effects. In order to prove the consistency, two strategies are proposed: (i) the first one is made of algebraic arguments based on the linear models obtained from PPCA; (ii) the second one is inspired by Mohan et al. [15] using the graphical model associated with PPCA. Furthermore, in this same setting of MNAR missingness, we also suggest a strategy to estimate the coefficient matrix (loadings) of the PPCA model, still without any additional modelisation. This allows to apply PPCA even in this difficult setting. Finally, the estimated coefficient matrix can be used to impute the missing values. We compare our proposal (estimation of the mean/covariances and imputation) with classical methods, such as the iterative singular value decomposition algorithms which ignore the missing values mechanism [13], and the method based on modelling the MNAR mechanism by a logistic regression model in [19].

Model. A low-rank model is considered via the formalism of the PPCA with latent variables. Suppose that before the introduction of missing values, the data matrix $Y \in \mathbb{R}^{n\times p}$ is generated under a random effects model, i.e. it can be obtained by the factorization of the coefficients matrix $B \in \mathbb{R}^{r\times p}$ and r latent variables grouped in the matrix $W \in \mathbb{R}^{n\times r}$,

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

for σ^2 and r known. In the following, $Y_{.j}$ and $Y_{i.}$ respectively denote the column j and the row i of Y. Let us remark that the rows of Y are identically distributed,

$$\forall i \in \{1,\dots,n\}, \quad Y_{i.} \sim \mathcal{N}(\alpha, B^TB + \sigma^2\mathrm{Id}_{p\times p})$$

Let $\Omega \in \{0,1\}^{n\times p}$ denote the missing-data pattern as

$$\forall i \in \{1,\dots,n\},\ \forall j \in \{1,\dots,p\}, \quad \Omega_{ij} = \begin{cases} 0 & \text{if } Y_{ij} \text{ is missing,} \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

Whether for theoretical results or numerical experiments, we focus only on the hard setting of informative missing values under a self-masked MNAR mechanism, which definition is given hereafter.

Definition 1. *(Self-masked MNAR mechanism) Let* $Y = \begin{pmatrix} Y_{.1} & Y_{.2} & \dots & Y_{.p} \end{pmatrix}$ *be a matrix of size* $n \times p$*. For* $j \in \{1, \dots, p\}$*, a variable* $Y_{.j}$ *is subject to a self-masked MNAR mechanism if the probability for an observation of being missing only depends on its value itself*

$$\forall i \in \{1, \dots, n\}, \quad \mathbb{P}(\Omega_{ij} = 0 | Y_{i.}) = \mathbb{P}(\Omega_{ij} = 0 | Y_{ij}).$$

Organization of the paper. For the sake of clarity, Section 2 is dedicated to present the detailed methodology and results in small dimension with a data matrix containing only one self-masked MNAR missing variable: the consistency results of the mean, variance and covariances estimators are studiedn and used in a new method to impute the data matrix and to estimate the PPCA loading matrix. In Section 3, we present results in the general case for data matrices containing several MNAR missing variables for an arbitrary dimension. Section 4 is devoted to numerical experiments, illustrating the efficiency and robustness of the proposed estimators and the imputation method in practice.

2 A toy example in small dimension

For the sake of clarity, the proposed approach is detailed and illustrated in a small dimensional setting, in which $p = 3$, $r = 2$ and in which only one variable can be missing, fixed to be $Y_{.1}$, under a self-masked MNAR mechanism. The model then reads as

$$\begin{pmatrix} Y_{.1} & Y_{.2} & Y_{.3} \end{pmatrix} = \mathbf{1} \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 \end{pmatrix} + \begin{pmatrix} W_{.1} & W_{.2} \end{pmatrix} B + \epsilon, \tag{3}$$

with $B \in \mathbb{R}^{2\times 3}$ and $\epsilon \in \mathbb{R}^{n\times 3}$. Its graphical representation is given in Figure 1(a).
The goal is four-fold: (i) to estimate the mean of $Y_{.1}$, (ii) its variance and covariances, (iii) the coefficient matrix B and (iv) to impute missing entries of Y.

2.1 Mean estimation

The purpose of this part is to estimate the mean $\mathbb{E}[Y_{.1}]$ of the missing variable $Y_{.1}$, denoted by α_1. Using two different approaches (algebraic and graphical), an estimator $\hat{\alpha}_1$ is derived, and proven to be consistent under self-masked MNAR mechanism.

2.1.1 Algebraic approach

In this section, the details on an algebraic approach to derive a consistent estimator of the mean are given. In the interest of understanding, all the intermediate results are concisely proved. The starting point is to exploit the linear links between variables, as described in the following lemma.

Lemma 2. *Assume that the reduced matrix* $\begin{pmatrix} B_{.1} & B_{.3} \end{pmatrix}$ *of* B *has an inverse matrix denoted as* $B^{-13} \in \mathbb{R}^{2\times 2}$*. The PPCA model (3) leads to the following linear equation,*

$$Y_{.2} = \mathcal{B}_{2\to 1,3[0]} + \mathcal{B}_{2\to 1,3[1]} Y_{.1} + \mathcal{B}_{2\to 1,3[3]} Y_{.3} - \mathcal{B}_{2\to 1,3[1]} \epsilon_{.1} - \mathcal{B}_{2\to 1,3[3]} \epsilon_{.3} + \epsilon_{.2}, \tag{4}$$

where $\mathcal{B}_{2\to 1,3[0]}$, $\mathcal{B}_{2\to 1.3[1]}$ *and* $\mathcal{B}_{2\to 3.1[3]}$ *stand for the coefficients depending on* B,

$$\begin{aligned}
\mathcal{B}_{2\to 1,3[0]} &:= -(B_{11}^{-13}B_{12} + B_{12}^{-13}B_{22})\mathbf{1}\alpha_1 - (B_{21}^{-13}B_{12} + B_{22}^{-13}B_{22})\mathbf{1}\alpha_3 + \mathbf{1}\alpha_2,\\
\mathcal{B}_{2\to 1,3[1]} &:= B_{11}^{-13}B_{12} + B_{12}^{-13}B_{22},\\
\mathcal{B}_{2\to 1,3[3]} &:= B_{21}^{-13}B_{12} + B_{22}^{-13}B_{22}.
\end{aligned}$$

Proof. Equation (3) can be restricted to $Y_{\cdot 2}$ as

$$Y_{\cdot 2} = \mathbf{1}\alpha_2 + \begin{pmatrix} W_{\cdot 1} & W_{\cdot 2} \end{pmatrix} B_{\cdot 2} + \epsilon_{\cdot 2}, \tag{5}$$

and to $Y_{\cdot 1}$ and $Y_{\cdot 3}$ as

$$\begin{pmatrix} Y_{\cdot 1} & Y_{\cdot 3} \end{pmatrix} = \mathbf{1}\begin{pmatrix} \alpha_1 & \alpha_3 \end{pmatrix} + \begin{pmatrix} W_{\cdot 1} & W_{\cdot 2} \end{pmatrix}\begin{pmatrix} B_{\cdot 1} & B_{\cdot 3} \end{pmatrix} + \begin{pmatrix} \epsilon_{\cdot 1} & \epsilon_{\cdot 3} \end{pmatrix}.$$

Then, since the reduced matrix $\begin{pmatrix} B_{\cdot 1} & B_{\cdot 3} \end{pmatrix}$ is invertible, one has

$$\begin{pmatrix} W_{\cdot 1} & W_{\cdot 2} \end{pmatrix} = \left(\begin{pmatrix} Y_{\cdot 1} & Y_{\cdot 3} \end{pmatrix} - \mathbf{1}\begin{pmatrix} \alpha_1 & \alpha_3 \end{pmatrix} - \begin{pmatrix} \epsilon_{\cdot 1} & \epsilon_{\cdot 3} \end{pmatrix}\right)\begin{pmatrix} B_{\cdot 1} & B_{\cdot 3} \end{pmatrix}^{-1}. \tag{6}$$

Using (5) and (6), it gives

$$\begin{aligned}
Y_{\cdot 2} =& (B_{11}^{-13}B_{12} + B_{12}^{-13}B_{22})Y_{\cdot 1} + (B_{21}^{-13}B_{12} + B_{22}^{-13}B_{22})Y_{\cdot 3}\\
&- (B_{11}^{-13}B_{12} + B_{12}^{-13}B_{22})(\mathbf{1}\alpha_1 + \epsilon_{\cdot 1}) - (B_{21}^{-13}B_{12} + B_{22}^{-13}B_{22})(\mathbf{1}\alpha_3 + \epsilon_{\cdot 3})\\
&+ \epsilon_{\cdot 2} + \mathbf{1}\alpha_2,
\end{aligned}$$

which leads to the desired solution. □

Let us introduce some definitions specifying the coefficients of $Y_{\cdot 2}$ on $Y_{\cdot 1}$ and $Y_{\cdot 3}$ when $\Omega_{\cdot 1} = 1$, i.e. keeping only the observations i such as Y_{i1} is observed. It is referred to as the complete case in the following.

Definition 3 (Coefficients in the complete case)**.** *Let* $\mathcal{B}^c_{2\to 1,3[0]}$, $\mathcal{B}^c_{2\to 1,3[1]}$ *and* $\mathcal{B}^c_{2\to 1,3[3]}$ *be the coefficients standing for the effects of* $Y_{\cdot 2}$ *on* $Y_{\cdot 1}$ *and* $Y_{\cdot 3}$ *in the complete case, when* $\Omega_{\cdot 1} = 1$*, i.e.*

$$(Y_{\cdot 2}|\Omega_{\cdot 1} = 1) := \mathcal{B}^c_{2\to 1,3[0]} + \mathcal{B}^c_{2\to 1,3[1]}Y_{\cdot 1} + \mathcal{B}^c_{2\to 1,3[3]}Y_{\cdot 3} - \mathcal{B}^c_{2\to 1,3[1]}\epsilon_{\cdot 1} - \mathcal{B}^c_{2\to 1,3[3]}\epsilon_{\cdot 3} + \epsilon_{\cdot 2}. \tag{7}$$

Using Equation (7), an expression for the mean of the missing variable $Y_{\cdot 1}$ can be derived as given in the following proposition.

Proposition 4 (Mean formula in the toy example)**.** *Under the PPCA model* (3)*, assume that:*

A1. $\begin{pmatrix} B_{\cdot 1} & B_{\cdot 3} \end{pmatrix}$ *is an invertible matrix,*

A2. $Y_{\cdot 2} \perp\!\!\!\perp \Omega_{\cdot 1}|Y_{\cdot 1}, Y_{\cdot 3}$.

Assuming also that $\mathcal{B}^c_{2\to 1,3[1]}$ *is non-zero, one can derive that*

$$\alpha_1 = \frac{\alpha_2 - \mathcal{B}^c_{2\to 1,3[0]} - \mathcal{B}^c_{2\to 1,3[3]}\alpha_3}{\mathcal{B}^c_{2\to 1,3[1]}}, \tag{8}$$

where $\mathcal{B}^c_{2\to 1,3[0]}$, $\mathcal{B}^c_{2\to 1,3[1]}$ *and* $\mathcal{B}^c_{2\to 1,3[3]}$ *are given in Definition 3.*

Proof. Given that $\mathbb{E}[Y_{\cdot 2}] = \mathbb{E}[\mathbb{E}[Y_{\cdot 2}|Y_{\cdot 1}, Y_{\cdot 3}]]$, Assumption **A2.** implies

$$\mathbb{E}[Y_{\cdot 2}|Y_{\cdot 1}, Y_{\cdot 3}] = \mathbb{E}[Y_{\cdot 2}|Y_{\cdot 1}, Y_{\cdot 3}, \Omega_{\cdot 1} = 1].$$

Then, by Definition 3,

$$\begin{aligned}\mathbb{E}[Y_{\cdot 2}|Y_{\cdot 1}, Y_{\cdot 3}, \Omega_{\cdot 1} = 1] &= \mathbb{E}[\mathcal{B}^c_{2\to 1,3[0]} + \mathcal{B}^c_{2\to 1,3[1]} Y_{\cdot 1} + \mathcal{B}^c_{2\to 1,3[3]} Y_{\cdot 3} - \mathcal{B}^c_{2\to 1,3[1]}\epsilon_{\cdot 1} - \mathcal{B}^c_{2\to 1,3[3]}\epsilon_{\cdot 3} + \epsilon_{\cdot 2}|Y_{\cdot 1}, Y_{\cdot 3}] \\ &= \mathcal{B}^c_{2\to 1,3[0]} + \mathcal{B}^c_{2\to 1,3[1]}\mathbb{E}[Y_{\cdot 1}|Y_{\cdot 1}, Y_{\cdot 3}] + \mathcal{B}^c_{2\to 1,3[3]}\mathbb{E}[Y_{\cdot 3}|Y_{\cdot 1}, Y_{\cdot 3}] \\ &\quad - \mathcal{B}^c_{2\to 1,3[1]}\mathbb{E}[\epsilon_{\cdot 1}|Y_{\cdot 1}, Y_{\cdot 3}] - \mathcal{B}^c_{2\to 1,3[3]}\mathbb{E}[\epsilon_{\cdot 3}|Y_{\cdot 1}, Y_{\cdot 3}] + \mathbb{E}[\epsilon_{\cdot 2}|Y_{\cdot 1}, Y_{\cdot 3}].\end{aligned}$$

Thus, by taking the mean and given that $\mathbb{E}[\epsilon_{\cdot i}] = 0$ for $i = 1, 2, 3$, one has

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

leading to Equation (8), provided that $\mathcal{B}^c_{2\to 1,3[1]} \neq 0$. □

Note that Assumption **A2.** is verified under the self-masked MNAR mechanism, given in Definition 1. Equation (8) suggests a natural estimator of α_1 as given in the following definition.

Definition 5 (Mean estimator in the toy example)**.** *Denote $\hat{\alpha}_2$ and $\hat{\alpha}_3$ the empirical means of $Y_{\cdot 2}$ and $Y_{\cdot 3}$, estimators of α_2 and α_3 obtained using all the observations. Denote $\hat{\mathcal{B}}^c_{2\to 1,3[0]}$, $\hat{\mathcal{B}}^c_{2\to 1,3[1]}$ and $\hat{\mathcal{B}}^c_{2\to 1,3[3]}$ some estimators of $\mathcal{B}^c_{2\to 1,3[0]}$, $\mathcal{B}^c_{2\to 1,3[1]}$ and $\mathcal{B}^c_{2\to 1,3[3]}$ computed in the complete case. A natural estimator $\hat{\alpha}_1$ of α_1 is*

$$\hat{\alpha}_1 := \frac{\hat{\alpha}_2 - \hat{\mathcal{B}}^c_{2\to 1,3[0]} - \hat{\mathcal{B}}^c_{2\to 1,3[3]}\hat{\alpha}_3}{\hat{\mathcal{B}}^c_{2\to 1,3[1]}}. \tag{9}$$

Proposition 6 (Consistency for the missing variable mean in the toy example)**.** *Assume that:*

A3. *α_2 and α_3 are recoverable, i.e. there exist consistent estimators for both quantities,*

A4. *the coefficients $\mathcal{B}^c_{2\to 1,3[0]}$, $\mathcal{B}^c_{2\to 1,3[1]}$ and $\mathcal{B}^c_{2\to 1,3[3]}$ are recoverable.*

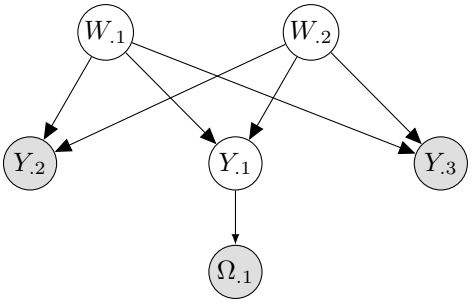
Then, the estimator $\hat{\alpha}_1$ of α_1 defined in Equation (9) is consistent.

The proof trivially follows from Equation (9) under **A3.** and **A4.**. Note that Equation (8), derived for PPCA, matches [15, Equation (8)], the latter being established for a linear model. To our knowledge, Equation (9) is the first proposition of a consistent estimator in a low-rank model under MNAR missing data (provided consistent estimators of the $\left(\mathcal{B}^c_{2\to 1,3[k]}\right)_{k\in\{0,1,3\}}$'s).

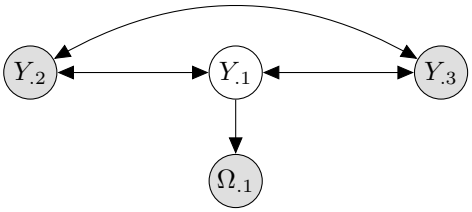
Remark 7. *In Lemma 2, an arbitrary choice has been made to derive an expression for $Y_{\cdot 2}$ according to $Y_{\cdot 1}$ and $Y_{\cdot 3}$. Since $Y_{\cdot 2}$ and $Y_{\cdot 3}$ are interchangeable, an expression of $Y_{\cdot 3}$ given $Y_{\cdot 2}$ and $Y_{\cdot 1}$ can also been obtained, and following a similar proof as in Proposition 4, another formula for α_1 can be derived as*

$$\alpha_1 = \frac{\alpha_3 - \mathcal{B}^c_{3\to 1,2[0]} - \mathcal{B}^c_{3\to 1,2[2]}\alpha_2}{\mathcal{B}^c_{3\to 1,2[1]}}.$$

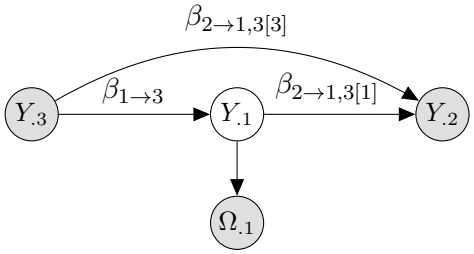
*However, it is not possible to use an expression of $Y_{\cdot 1}$ given $Y_{\cdot 2}$ and $Y_{\cdot 3}$ inasmuch as $Y_{\cdot 1}$ has self-masked MNAR missing values and therefore it does not comply with Assumption **A2.**. In Section 4, in practice, one will see that these formulae may lead to slightly different mean estimators; that is why an approach using aggregation between all possible mean estimators will be proposed.*



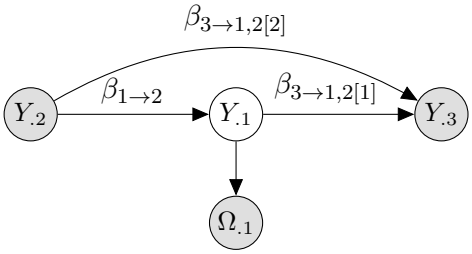
(a) Graphical model for PPCA.



(b) Graphical model where the bidirected edges encode the latent variables.



(c) Reduced graphical model.



(d) Reduced graphical model.

Figure 1: Graphical models for the toy example with one missing variable $Y_{\cdot 1}$, $p = 3$ and $r = 2$. (a) gives the graphical model associated with the PPCA model of Equation (3). In (b), the corresponding graphical model to the one in (a) is represented, in which latent variables have been replaced by bidirected edges; six reduced graphical models can be derived from it, for all possible arrow combinations. Only two of them are represented in (c) and (d). Therefore, one could derive the following implications between these graphical models (without the coefficients): 1(a) $\Rightarrow$ 1(b) $\Rightarrow$ (1(c) and 1(d)).

Estimation of the mean in practice from the algebraic approach. In practice, $\hat{\mathcal{B}}^c_{2\to 1,3[0]}$, $\hat{\mathcal{B}}^c_{2\to 1,3[1]}$ and $\hat{\mathcal{B}}^c_{2\to 1,3[3]}$ are estimated with the intercept and the coefficients of the linear regression of $Y_{\cdot 2}$ on $Y_{\cdot 1}$ and $Y_{\cdot 3}$ using the fully-observed data only. Even if exogeneity does not hold in Equation (7), since $\mathbb{E}[\epsilon_{\cdot 3}|Y_{\cdot 1}, Y_{\cdot 3}] \neq 0$ and $\mathbb{E}[\epsilon_{\cdot 1}|Y_{\cdot 1}, Y_{\cdot 3}] \neq 0$, it still leads to accurate estimation of α_1 in numerical experiments, as shown in Section 4.

2.1.2 Graphical approach

The graphical approach to construct an estimator of α_1 is based on the transformation illustrated in Figure 1 of the graphical model of PPCA as structural causal graphs, whose context is introduced in [17]. This latter framework allows to directly apply the results of Mohan et al. [15] who consider the associated (linear) structural causal equations under the exogeneity assumption with MNAR missing values for one variable.

More precisely, starting from Figure 1(a) one gets Figure 1(b) as $Y_{\cdot 1} \leftarrow W_{\cdot 1} \rightarrow Y_{\cdot 2}$ is equivalent to $Y_{\cdot 1} \leftrightarrow Y_{\cdot 2}$. Indeed, since $W_{\cdot 1}$ and $W_{\cdot 2}$ are latent variables, following the notation of Pearl [17, page 52], both unidirected edges $Y_{\cdot 1} \leftarrow W_{\cdot 1} \rightarrow Y_{\cdot 2}$ can be replaced by a bidirected edge $Y_{\cdot 1} \leftrightarrow Y_{\cdot 2}$. Then,

six reduced graphical models can be derived from Figure 1(b). Indeed, according to the rule proposed by Pearl [17, rule 1, page 147], a bidirected edge $Y_{\cdot 1} \leftrightarrow Y_{\cdot 2}$ can be interchanged 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 $Y_{\cdot 1}$ (by the d-criterion separation, see [17, page 17]). Figures 1(c) and 1(d) (without the coefficients) are two instances of the six possible graphs.

Then, assuming exogeneity, one can associate to Figure 1(c) the structural equation model given in the following lemma.

Lemma 8. *Assuming* $\mathbb{E}[\epsilon_{Y_{\cdot 2}}|Y_{\cdot 1}, Y_{\cdot 3}] = 0$*, the structural equation model associated with the graphical model in Figure 1(c) is*

$$Y_{\cdot 2} = \beta_{2\rightarrow 1,3[0]} + \beta_{2\rightarrow 1,3[1]} Y_{\cdot 1} + \beta_{2\rightarrow 1,3[3]} Y_{\cdot 3} + \epsilon_{Y_{\cdot 2}}, \tag{10}$$

where $\beta_{2\rightarrow 1,3[0]}$*,* $\beta_{2\rightarrow 1,3[1]}$ *and* $\beta_{2\rightarrow 1,3[3]}$ *are the intercept and the coefficients of the linear regression of* $Y_{\cdot 2}$ *on* $Y_{\cdot 1}$ *and* $Y_{\cdot 3}$*.*

Assuming **A2.** and using Figure 1(c) and Equation (10) allow to apply the results of Mohan et al. [15], that are summarized in Appendix II, to get an estimator for the mean of the first variable, i.e.

$$\hat{\alpha}_1 := \frac{\hat{\alpha}_2 - \hat{\beta}^c_{2\rightarrow 1,3[0]} - \hat{\beta}^c_{2\rightarrow 1,3[3]}\hat{\alpha}_3}{\hat{\beta}^c_{2\rightarrow 1,3[1]}}, \tag{11}$$

where $\hat{\beta}^c_{2\rightarrow 1,3[0]}$, $\hat{\beta}^c_{2\rightarrow 1,3[1]}$ and $\hat{\beta}^c_{2\rightarrow 1,3[3]}$ denote some estimators of $\beta^c_{2\rightarrow 1,3[0]}$, $\beta^c_{2\rightarrow 1,3[1]}$ and $\beta^c_{2\rightarrow 1,3[3]}$, the coefficients standing for the effects of $Y_{\cdot 2}$ on $Y_{\cdot 1}$ and $Y_{\cdot 3}$ in the complete case, when $\Omega_{\cdot 1} = 1$.

Remark 9. *Note that the graphical approach can be developed with another arbitrary choice of variables, for instance using Figure 1(d) instead of Figure 1(c). This other choice would have led to an expression of* $Y_{\cdot 3}$ *given* $Y_{\cdot 2}$ *and* $Y_{\cdot 1}$ *and by again applying the results of Mohan et al. [15], one could have derived that*

$$\hat{\alpha}_1 := \frac{\hat{\alpha}_3 - \hat{\beta}^c_{3\rightarrow 1,2[0]} - \hat{\beta}^c_{3\rightarrow 1,2[2]}\hat{\alpha}_2}{\hat{\beta}^c_{3\rightarrow 1,2[1]}}.$$

Algebraic vs. graphical approach. In both approaches, the PPCA model is translated into a linear model. However, both estimators in Equations (9) and (11) 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 $\hat{\mathcal{B}}^c_{2\rightarrow 1,3[0]}$, $\hat{\mathcal{B}}^c_{2\rightarrow 1,3[1]}$ and $\hat{\mathcal{B}}^c_{2\rightarrow 1,3[3]}$ 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. In practice, for both approaches, the same coefficients are naturally computed, *i.e.* $\hat{\beta}^c_{2\rightarrow 1,3[0]} = \hat{\mathcal{B}}^c_{2\rightarrow 1,3[0]}$, $\hat{\beta}^c_{2\rightarrow 1,3[1]} = \hat{\mathcal{B}}^c_{2\rightarrow 1,3[1]}$ and $\hat{\beta}^c_{2\rightarrow 1,3[3]} = \hat{\mathcal{B}}^c_{2\rightarrow 1,3[3]}$ which leads to the same computed estimators for the mean of $Y_{\cdot 1}$.

2.2 Variance and covariances estimation

In this section, we suggest estimators for the variance of $Y_{\cdot 1}$, for the covariance between $Y_{\cdot 1}$ and $Y_{\cdot 2}$ and for the covariance between $Y_{\cdot 1}$ and $Y_{\cdot 3}$. Their consistency can also be obtained using two arguments: algebraic and graphical ones.

2.2.1 Algebraic approach

Whereas only one linear equation between $Y_{\cdot 1}$, $Y_{\cdot 2}$ and $Y_{\cdot 3}$ was required to construct an estimator of the mean of $Y_{\cdot 1}$, two linearly independent equations between $Y_{\cdot 1}$, $Y_{\cdot 2}$ and $Y_{\cdot 3}$ are required to construct a variance and covariances estimator when the rank r is to 2. Therefore, the algebraic approach is based on the linear equation between $Y_{\cdot 2}$ and $(Y_{\cdot 1}, Y_{\cdot,3})$, given in Lemma 2, and on the one given hereafter between $Y_{\cdot 3}$ and $(Y_{\cdot 1}, Y_{\cdot,2})$.

Lemma 10. *Assume that the reduced matrix* $\begin{pmatrix} B_{\cdot 1} & B_{\cdot 2} \end{pmatrix}$ *of* B *has an inverse matrix denoted by* $B^{-12} \in \mathbb{R}^{2\times 2}$. *The PPCA model* (3) *leads to the following linear equation,*

$$Y_{\cdot 3} = \mathcal{B}_{3\to 1,2[0]} + \mathcal{B}_{3\to 1,2[1]} Y_{\cdot 1} + \mathcal{B}_{3\to 1,2[2]} Y_{\cdot 2} - \mathcal{B}_{3\to 1,2[1]} \epsilon_{\cdot 1} - \mathcal{B}_{3\to 1,2[2]} \epsilon_{\cdot 2} + \epsilon_{\cdot 3}, \tag{12}$$

where $\mathcal{B}_{3\to 1,2[0]}$, $\mathcal{B}_{3\to 1,2[1]}$ *and* $\mathcal{B}_{3\to 1,2[2]}$ *stand for the linear coefficients depending on* B,

$$\begin{aligned}
\mathcal{B}_{3\to 1,2[0]} &:= -(B_{11}^{-12} B_{13} + B_{12}^{-12} B_{23}) \mathbf{1}\alpha_1 - (B_{21}^{-12} B_{13} + B_{22}^{-12} B_{23}) \mathbf{1}\alpha_2 + \mathbf{1}\alpha_3, \\
\mathcal{B}_{3\to 1,2[1]} &:= B_{11}^{-12} B_{13} + B_{12}^{-12} B_{23}, \\
\mathcal{B}_{3\to 1,2[2]} &:= B_{21}^{-12} B_{13} + B_{22}^{-12} B_{23}.
\end{aligned}$$

Let us introduce some definitions specifying the coefficients of $Y_{\cdot 2}$ on $Y_{\cdot 1}$ and $Y_{\cdot 3}$ when $\Omega_{\cdot 1} = 1$, i.e. in the complete case setting.

Definition 11 (Coefficients in the complete case)**.** *Let* $\mathcal{B}^c_{3\to 1,2[0]}$, $\mathcal{B}^c_{3\to 1,2[1]}$ *and* $\mathcal{B}^c_{3\to 1,2[2]}$ *be the coefficients standing for the effects of* $Y_{\cdot 3}$ *on* $Y_{\cdot 1}$ *and* $Y_{\cdot 2}$ *in the complete case, when* $\Omega_{\cdot 1} = 1$, *i.e.*

$$(Y_{\cdot 3} | \Omega_{\cdot 1} = 1) := \mathcal{B}^c_{3\to 1,2[0]} + \mathcal{B}^c_{3\to 1,2[1]} Y_{\cdot 1} + \mathcal{B}^c_{3\to 1,2[2]} Y_{\cdot 2} - \mathcal{B}^c_{3\to 1,2[1]} \epsilon_{\cdot 1} - \mathcal{B}^c_{3\to 1,2[2]} \epsilon_{\cdot 2} + \epsilon_{\cdot 3}. \tag{13}$$

Combining (7) and (13), the following proposition gives formulae for the variance and the covariances of $Y_{\cdot 1}$.

Proposition 12 (Variance and covariances formulae in the toy example)**.** *Assume* ***A1.***, ***A2.*** *and that*

A6. $\begin{pmatrix} B_{\cdot 1} & B_{\cdot 2} \end{pmatrix}$ *is an invertible matrix,*

A7. $Y_{\cdot 3} \perp\!\!\!\perp \Omega_{\cdot 1} | Y_{\cdot 1}, Y_{\cdot 2}$.

The following matrix system holds,

$$M_1 X + o(\sigma^2) = M_2, \tag{14}$$

with

$$X = \begin{pmatrix} \mathrm{Var}(Y_{\cdot 1}) \\ \mathrm{Cov}(Y_{\cdot 2}, Y_{\cdot 1}) \\ \mathrm{Cov}(Y_{\cdot 3}, Y_{\cdot 1}) \end{pmatrix},$$

$$M_1 = \begin{pmatrix} (\mathcal{B}^c_{2\to 1,3[1]})^2 & 0 & 2\mathcal{B}^c_{2\to 1,3[1]}\mathcal{B}^c_{2\to 1,3[3]} \\ -\mathcal{B}^c_{2\to 1,3[1]} & 1 & -\mathcal{B}^c_{2\to 1,3[3]} \\ -\mathcal{B}^c_{3\to 1,2[1]} & -\mathcal{B}^c_{3\to 1,2[2]} & 1 \end{pmatrix},$$

$$M_2 = \begin{pmatrix} \mathrm{Var}(Y_{.2}) - Q^c - (\mathcal{B}^c_{2\to 1,3[3]})^2\mathrm{Var}(Y_{.3}) \\ \mathcal{B}^c_{2\to 1,3[1]}\mathbb{E}[Y_{.1}]^2 + \mathcal{B}^c_{2\to 1,3[0]}\mathbb{E}[Y_{.1}] + \mathcal{B}^c_{2\to 1,3[3]}\mathbb{E}[Y_{.3}]\mathbb{E}[Y_{.1}] - \mathbb{E}[Y_{.2}]\mathbb{E}[Y_{.1}] \\ \mathcal{B}^c_{2\to 1,3[1]}\mathbb{E}[Y_{.1}]^2 + \mathcal{B}^c_{3\to 1,2[0]}\mathbb{E}[Y_{.1}] + \mathcal{B}^c_{3\to 1,2[2]}\mathbb{E}[Y_{.2}]\mathbb{E}[Y_{.1}] - \mathbb{E}[Y_{.3}]\mathbb{E}[Y_{.1}] \end{pmatrix},$$

where $\mathcal{B}^c_{2\to 1,3[0]}$, $\mathcal{B}^c_{2\to 1,3[1]}$, $\mathcal{B}^c_{2\to 1,3[3]}$, $\mathcal{B}^c_{3\to 1,2[0]}$, $\mathcal{B}^c_{3\to 1,2[1]}$ *and* $\mathcal{B}^c_{3\to 1,2[2]}$ *are given in Definitions 3 and 11, and*

$$Q^c := \left(\mathrm{Var}(Y_{.2}) - \mathrm{Cov}([Y_{.1}, Y_{.3}], Y_{.2})\mathrm{Var}([Y_{.1}, Y_{.3}])\mathrm{Cov}([Y_{.1}, Y_{.3}], Y_{.2})^T \middle| \Omega_{.1} = 1\right), \tag{15}$$

with $(.|\Omega_{.1} = 1)$ *meaning that the quantities are computed for* $\Omega_{.1} = 1$.
Up to an $o(\sigma^2)$*-term, the variance and the covariances of* $Y_{.1}$ *can be obtained by solving the matrix system in* (14).

Proof. **About the variance.** The law of total variance reads as

$$\mathrm{Var}(Y_{.2}) = \mathbb{E}[\mathrm{Var}(Y_{.2}|Z)] + \mathrm{Var}(\mathbb{E}[Y_{.2}|Z]), \tag{16}$$

with $Z = (Y_{.1}, Y_{.3}, \epsilon_{.1}, \epsilon_{.3})$. As for the first term, using Assumption **A2.**, one has

$$\mathrm{Var}(Y_{.2}|Z) = \mathrm{Var}(Y_{.2}|Z, \Omega_{.1} = 1).$$

As the conditional variance for a Gaussian vector gives

$$\mathrm{Var}(Y_{.2}|Z) = \mathrm{Var}(Y_{.2}) - \mathrm{Cov}(Z, Y_{.2})\mathrm{Var}(Z)^{-1}\mathrm{Cov}(Z, Y_{.2})^T,$$

it implies that

$$\mathrm{Var}(Y_{.2}|Z, \Omega_{.1} = 1) = \left(\mathrm{Var}(Y_{.2}) - \mathrm{Cov}(Z, Y_{.2})\mathrm{Var}(Z)^{-1}\mathrm{Cov}(Z, Y_{.2})^T \middle| \Omega_{.1} = 1\right)$$

and then, as deterministic quantity,

$$\mathbb{E}[\mathrm{Var}(Y_{.2}|Z)] = \left(\mathrm{Var}(Y_{.2}) - \mathrm{Cov}(Z, Y_{.2})\mathrm{Var}(Z)^{-1}\mathrm{Cov}(Z, Y_{.2})^T \middle| \Omega_{.1} = 1\right).$$

Noting that $\mathrm{Cov}(\epsilon_{.1}, Y_{.2}) = \mathrm{Cov}(\epsilon_{.3}, Y_{.2}) = 0$, one has

$$\mathrm{Cov}(Z, Y_{.2})\mathrm{Var}(Z)^{-1}\mathrm{Cov}(Z, Y_{.2})^T = \mathrm{Cov}([Y_{.1}, Y_{.3}], Y_{.2})\mathrm{Var}([Y_{.1}, Y_{.3}])^{-1}\mathrm{Cov}([Y_{.1}, Y_{.3}], Y_{.2})^T$$

leading to

$$\mathbb{E}[\mathrm{Var}(Y_{.2}|Z)] = Q^c, \tag{17}$$

where Q^c is defined in (15). As for the second term of (16), remark that **A2.** implies that

$$\mathrm{Var}(\mathbb{E}[Y_{.2}|Z]) = \mathrm{Var}(\mathbb{E}[Y_{.2}|Z, \Omega_{.1} = 1]),$$

and by Definition 3,

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[Y_{.2}|Z,\Omega_{.1}=1]) &= \mathrm{Var}(\mathbb{E}[\mathcal{B}^c_{2\to1,3[0]}+\mathcal{B}^c_{2\to1,3[1]}Y_{.1}+\mathcal{B}^c_{2\to1,3[3]}Y_{.3}-\mathcal{B}^c_{2\to1,3[1]}\epsilon_{.1}-\mathcal{B}^c_{2\to1,3[3]}\epsilon_{.3}+\epsilon_{.2}|Z]\\ &= \mathrm{Var}(\mathcal{B}^c_{2\to1,3[0]}+\mathcal{B}^c_{2\to1,3[1]}Y_{.1}+\mathcal{B}^c_{2\to1,3[3]}Y_{.3}-\mathcal{B}^c_{2\to1,3[1]}\epsilon_{.1}-\mathcal{B}^c_{2\to1,3[3]}\epsilon_{.3}).\end{aligned}$$

Using $\mathrm{Var}(\epsilon_{.i})=\sigma^2$, $\mathrm{Cov}(\epsilon_{.i},Y_{.i})=\sigma^2, i\in\{1,3\}$ and $\mathrm{Cov}(\epsilon_{.i},Y_{.j})=0, i\neq j\in\{1,3\}^2$, one has

$$\mathrm{Var}(\mathbb{E}[Y_{.2}|Z,\Omega_{.1}=1]) = (\mathcal{B}^c_{2\to1,3[1]})^2\mathrm{Var}(Y_{.1})+(\mathcal{B}^c_{2\to1,3[3]})^2\mathrm{Var}(Y_{.3})+2\mathcal{B}^c_{2\to1,3[1]}\mathcal{B}^c_{2\to1,3[3]}\mathrm{Cov}(Y_{.1},Y_{.3}) \tag{18}$$

Combining (17) with (18), one get the following expression for the variance

$$\mathrm{Var}(Y_{.2}) = Q^c+(\mathcal{B}^c_{2\to1,3[1]})^2\mathrm{Var}(Y_{.1})+(\mathcal{B}^c_{2\to1,3[3]})^2\mathrm{Var}(Y_{.3})+2\mathcal{B}^c_{2\to1,3[1]}\mathcal{B}^c_{2\to1,3[3]}\mathrm{Cov}(Y_{.1},Y_{.3}) \tag{19}$$

About the covariances. Consider

$$\begin{aligned}\mathrm{Cov}(Y_{.2},Y_{.1}) &= \mathbb{E}[Y_{.2}Y_{.1}]-\mathbb{E}[Y_{.2}]\mathbb{E}[Y_{.1}] = \mathbb{E}[\mathbb{E}[Y_{.2}Y_{.1}|Z]]-\mathbb{E}[Y_{.2}]\mathbb{E}[Y_{.1}],\\ &= \mathbb{E}[Y_{.1}\mathbb{E}[Y_{.2}|Z]]-\mathbb{E}[Y_{.2}]\mathbb{E}[Y_{.1}].\end{aligned} \tag{20}$$

As for the first term in (20), one has

$$\begin{aligned}\mathbb{E}[Y_{.1}\mathbb{E}[Y_{.2}|Z]] =& \mathbb{E}[Y_{.1}\mathbb{E}[Y_{.2}|Z,\Omega_{.1}=1]] && \text{(using \textbf{A2.})}\\ =& \mathbb{E}[Y_{.1}\left(\mathcal{B}^c_{2\to1,3[0]}+\mathcal{B}^c_{2\to1,3[1]}Y_{.1}+\mathcal{B}^c_{2\to1,3[3]}Y_{.3}-\mathcal{B}^c_{2\to1,3[1]}\epsilon_{.1}-\mathcal{B}^c_{2\to1,3[3]}\epsilon_{.3}\right)]\\ =& \mathcal{B}^c_{2\to1,3[0]}\mathbb{E}[Y_{.1}]+\mathcal{B}^c_{2\to1,3[1]}\mathbb{E}[Y_{.1}^2]+\mathcal{B}^c_{2\to1,3[3]}\mathbb{E}[Y_{.1}Y_{.3}]-\sigma^2\mathcal{B}^c_{2\to1,3[1]},\end{aligned}$$

where in the last equality we used $\mathbb{E}[Y_{.1}\epsilon_{.1}]=\mathrm{Cov}(Y_{.1},\epsilon_{.1})=\sigma^2$ and $\mathbb{E}[Y_{.1}\epsilon_{.3}]=0$. By (20), one has

$$\begin{aligned}\mathrm{Cov}(Y_{.2},Y_{.1}) = \mathcal{B}^c_{2\to1,3[0]}\mathbb{E}[Y_{.1}]+\mathcal{B}^c_{2\to1,3[1]}(\mathrm{Var}(Y_{.1})+\mathbb{E}[Y_{.1}]^2)\\ +\mathcal{B}^c_{2\to1,3[3]}(\mathrm{Cov}(Y_{.1},Y_{.3})+\mathbb{E}[Y_{.3}]\mathbb{E}[Y_{.1}])-\mathbb{E}[Y_{.2}]\mathbb{E}[Y_{.1}]+o(\sigma^2).\end{aligned} \tag{21}$$

Similarly, by Assumption **A7.**,

$$\begin{aligned}\mathrm{Cov}(Y_{.3},Y_{.1}) = \mathcal{B}^c_{3\to1,2[0]}\mathbb{E}[Y_{.1}]+\mathcal{B}^c_{3\to1,2[1]}(\mathrm{Var}(Y_{.1})+\mathbb{E}[Y_{.1}]^2)\\ +\mathcal{B}^c_{3\to1,2[2]}(\mathrm{Cov}(Y_{.1},Y_{.2})+\mathbb{E}[Y_{.2}]\mathbb{E}[Y_{.1}])-\mathbb{E}[Y_{.3}]\mathbb{E}[Y_{.1}]+o(\sigma^2).\end{aligned} \tag{22}$$

Combining Equations (19), (21) and (22) forms the desired matrix system (14). □

By ignoring the $o(\sigma^2)$-term in (14), one can define estimators of the variance and covariances of the missing variable, as follows.

Definition 13 (Variance and covariances estimators in the toy example)**.** *Denote*

- $\widehat{\mathrm{Var}}(Y_{.2})$, $\widehat{\mathrm{Var}}(Y_{.3})$, $\widehat{\mathrm{Cov}}(Y_{.2},Y_{.3})$ *and* $\hat{Q}^c$, *some estimators of* $\mathrm{Var}(Y_{.2})$, $\mathrm{Var}(Y_{.3})$, $\mathrm{Cov}(Y_{.2},Y_{.3})$ *and* Q^c,
- $\hat{\mathcal{B}}^c_{3\to1,2[0]}$, $\hat{\mathcal{B}}^c_{3\to1,2[1]}$ *and* $\hat{\mathcal{B}}^c_{3\to1,2[2]}$, *some estimators of* $\mathcal{B}^c_{3\to1,2[0]}$, $\mathcal{B}^c_{3\to1,2[1]}$ *and* $\mathcal{B}^c_{3\to1,2[2]}$ *computed in the complete case.*

Let $\widehat{\mathrm{Var}}(Y_{.1})$, $\widehat{\mathrm{Cov}}(Y_{.2}, Y_{.1})$ *and* $\widehat{\mathrm{Cov}}(Y_{.3}, Y_{.1})$ *be estimators of* $\mathrm{Var}(Y_{.1})$, $\mathrm{Cov}(Y_{.2}, Y_{.1})$ *and* $\mathrm{Cov}(Y_{.3}, Y_{.1})$, *defined as follows*

$$\begin{pmatrix} \widehat{\mathrm{Var}}(Y_{.1}) \\ \widehat{\mathrm{Cov}}(Y_{.2}, Y_{.1}) \\ \widehat{\mathrm{Cov}}(Y_{.3}, Y_{.1}) \end{pmatrix} := \tag{23}$$

$$\begin{pmatrix} (\hat{\mathcal{B}}^c_{2\to 1,3[1]})^2 & 0 & 2\hat{\mathcal{B}}^c_{2\to 1,3[1]}\hat{\mathcal{B}}^c_{2\to 1,3[3]} \\ -\hat{\mathcal{B}}^c_{2\to 1,3[1]} & 1 & -\hat{\mathcal{B}}^c_{2\to 1,3[3]} \\ -\hat{\mathcal{B}}^c_{3\to 1,2[1]} & -\hat{\mathcal{B}}^c_{3\to 1,2[2]} & 1 \end{pmatrix}^{-1} \begin{pmatrix} \widehat{\mathrm{Var}}(Y_{.2}) - \hat{Q}^c - (\hat{\mathcal{B}}^c_{2\to 1,3[3]})^2 \widehat{\mathrm{Var}}(Y_{.3}) \\ \hat{\mathcal{B}}^c_{2\to 1,3[1]}\hat{\alpha}_1^2 + \hat{\mathcal{B}}^c_{2\to 1,3[0]}\hat{\alpha}_1 + \hat{\mathcal{B}}^c_{2\to 1,3[3]}\hat{\alpha}_3\hat{\alpha}_1 - \hat{\alpha}_2\hat{\alpha}_1 \\ \hat{\mathcal{B}}^c_{3\to 1,2[1]}\hat{\alpha}_1^2 + \mathcal{B}^c_{3\to 1,2[0]}\hat{\alpha}_1 + \hat{\mathcal{B}}^c_{3\to 1,2[2]}\hat{\alpha}_2\hat{\alpha}_1 - \hat{\alpha}_3\hat{\alpha}_1 \end{pmatrix},$$

provided that in the last expression, this matrix inverse exists.

Proposition 14 (Consistency for the variance and covariances in the toy example)**.** *Assume* ***A3.***, ***A4.*** *and that*

A9. $\mathrm{Var}(Y_{.2})$, $\mathrm{Var}(Y_{.3})$, $\mathrm{Cov}(Y_{.2}, Y_{.3})$ *and* Q^c *are recoverable,*

A10. *the coefficients* $\mathcal{B}^c_{3\to 1,2[0]}$, $\mathcal{B}^c_{3\to 1,2[1]}$ *and* $\mathcal{B}^c_{3\to 1,2[2]}$ *are recoverable.*

Then, the estimators $\widehat{\mathrm{Var}}(Y_{.1})$, $\widehat{\mathrm{Cov}}(Y_{.2}, Y_{.1})$ *and* $\widehat{\mathrm{Cov}}(Y_{.3}, Y_{.1})$ *of* $\mathrm{Var}(Y_{.1})$, $\mathrm{Cov}(Y_{.2}, Y_{.1})$ *and* $\mathrm{Cov}(Y_{.3}, Y_{.1})$ *defined by Equation* (23) *are consistent, when* σ^2 *tends to zero.*

The proof trivially follows from (23) under **A3.**, **A4.**, **A9.** and **A10.**.

Variance estimation in practice for the algebraic approach. On the one hand, as for the mean estimation, the non-exogeneity in (7) and (13) is ignored to provide the $\left(\mathcal{B}^c_{2\to 1,3[k]}\right)_{k\in\{0,1,3\}}$ and $\left(\mathcal{B}^c_{3\to 1,2[k]}\right)_{k\in\{0,1,2\}}$. Indeed, they are computed using the coefficients of the linear regression of $Y_{.i}$ on $(Y_{.j}, Y_{.k})$ in the complete case. On the other hand, $\widehat{\mathrm{Var}}(Y_{.2})$, $\widehat{\mathrm{Var}}(Y_{.3})$, $\widehat{\mathrm{Cov}}(Y_{.2}, Y_{.3})$ and $\hat{Q}^c$ are computed as empirical quantities, using all data for estimating the variances and the covariance and considering the complete case for Q. In addition, $\hat{\alpha}_1$ is given by (11).

2.2.2 Graphical approach

Whereas only one simplified graphical model between $Y_{.1}$, $Y_{.2}$ and $Y_{.3}$, displayed in Figure 1(c), was required to construct an estimator of the mean of $Y_{.1}$, two simplified graphical model between $Y_{.1}$, $Y_{.2}$ and $Y_{.3}$ are required to construct an estimator of the variance and covariances when the rank r is 2. Therefore, the graphical approach is based on the linear equation between $Y_{.2}$ and $(Y_{.1}, Y_{.,3})$, given in Equation (10) and based on Figure 1(c), and on the one given hereafter between $Y_{.3}$ and $(Y_{.1}, Y_{.,2})$, based on Figure 1(d).

Lemma 15. *Assuming* $\mathbb{E}[\epsilon_{Y_{.3}}|Y_{.1}, Y_{.2}] = 0$, *the structural equation model associated with the graphical model in Figure 1(d) is*

$$Y_{.3} = \beta_{3\to 1,2[0]} + \beta_{3\to 1,2[1]} Y_{.1} + \beta_{3\to 1,2[2]} Y_{.2} + \epsilon_{Y_{.3}}, \tag{24}$$

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* $Y_{.3}$ *on* $Y_{.1}$ *and* $Y_{.2}$ *in the complete case.*

Under the two equations (10) and (24), supposing that Assumptions **A2.** and **A7.** hold allows to apply the results of Mohan et al. [15] summarized in Appendix II which give the following estimates for the variance and the covariances of the first variable:

$$\widehat{\mathrm{Var}}(Y_{.1}) := \frac{\widehat{\mathrm{Var}}(Y_{.3})}{\hat{\beta}^c_{3\to 1}} \frac{1}{\hat{\beta}^c_{2\to 1,3[1]}} \left(\frac{\widehat{\mathrm{Cov}}(Y_{.2}, Y_{.3})}{\widehat{\mathrm{Var}}(Y_{.3})} - \hat{\beta}^c_{2\to 1,3[3]} \right), \tag{25}$$

$$\widehat{\mathrm{Cov}}(Y_{.1}, Y_{.2}) := \frac{1}{\hat{\beta}^c_{3\to 1,2[1]}} \left(\frac{\widehat{\mathrm{Cov}}(Y_{.2}, Y_{.3})}{\widehat{\mathrm{Var}}(Y_{.2})} - \hat{\beta}^c_{3\to 1,2[2]} \right) \widehat{\mathrm{Var}}(Y_{.2}), \tag{26}$$

$$\widehat{\mathrm{Cov}}(Y_{.1}, Y_{.3}) := \frac{1}{\hat{\beta}^c_{2\to 1,3[1]}} \left(\frac{\widehat{\mathrm{Cov}}(Y_{.2}, Y_{.3})}{\widehat{\mathrm{Var}}(Y_{.3})} - \hat{\beta}^c_{2\to 1,3[3]} \right) \widehat{\mathrm{Var}}(Y_{.3}), \tag{27}$$

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 $\beta^c_{3\to 1}$ standing for the effects of $Y_{.3}$ on $Y_{.1}$ and $Y_{.2}$ and $Y_{.3}$ on $Y_{.1}$ in the complete case, when $\Omega_{.1} = 1$.

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 Equation (23) with Equations (25), (26) and (27)). Indeed, the algebraic approach is based on the use of conditionality, whereas the graphical one relies on graphical results standing for the linear models when exogeneity holds. Moreover, one will see in Section 4 that the graphical approach is more stable (less variance in the results), probably as there is no matrix inversion but only standard divisions.

Remark 16. *As with the mean (see Remark 9), one could derived another expression of the variance of $Y_{.1}$ for instance using 1(d) instead of 1(c).*

2.3 Estimation of the loading matrix

Let us denote the covariance matrix estimator obtained with the algebraic approach by $\hat{\Sigma}$, i.e.

$$\hat{\Sigma} = \begin{pmatrix} \widehat{\mathrm{Var}}(Y_{.1}) & \widehat{\mathrm{Cov}}(Y_{.1}, Y_{.2}) & \widehat{\mathrm{Cov}}(Y_{.1}, Y_{.3}) \\ \widehat{\mathrm{Cov}}(Y_{.2}, Y_{.1}) & \widehat{\mathrm{Var}}(Y_{.2}) & \widehat{\mathrm{Cov}}(Y_{.2}, Y_{.3}) \\ \widehat{\mathrm{Cov}}(Y_{.3}, Y_{.1}) & \widehat{\mathrm{Cov}}(Y_{.3}, Y_{.2}) & \widehat{\mathrm{Var}}(Y_{.3}) \end{pmatrix}, \tag{28}$$

which entries in the first column and row are determined by Equation (23), and where estimated covariances and variances of fully observed variables are based on the regular empirical estimators. Note that a similar expression of $\hat{\Sigma}$ could be obtained with the graphical approach, using Equations (25), (26) and (27) instead.

Assuming that the level of noise σ^2 is known and that the rank r of the loading matrix is known, an estimator of the coefficient matrix can be derived based on the estimated covariance matrice $\hat{\Sigma}$. Indeed, note that by the intrinsic PPCA model,

$$Y \sim \mathcal{N}\left(\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}, B^T B + \sigma^2 \mathrm{Id}_{3\times 3} \right), \tag{29}$$

so that the matrix $\hat{\Sigma} - \sigma^2 \mathrm{Id}_{3\times 3}$ is an estimate of $B^T B$. An estimator of B can be thus defined using the singular value decomposition of $\hat{\Sigma} - \sigma^2 \mathrm{Id}_{3\times 3}$, this is the purpose of the following definition.

Definition 17 (Estimation of the loading matrix in the toy example)**.** *Given $\hat{\Sigma}$ in Equation (28), let the orthogonal matrix $\hat{U} \in \mathbb{R}^{3\times 3}$ and the diagonal matrix $\hat{D} = \mathrm{diag}(\hat{d}_1, \hat{d}_2, \hat{d}_3) \in \mathbb{R}^{3\times 3}$ with $d_1 \geq d_2 \geq d_3 \geq 0$, form the singular value decomposition of the following matrix*

$$\hat{\Sigma} - \sigma^2 \mathrm{Id}_{3\times 3} =: \hat{U}\hat{D}\hat{U}^T,$$

and denote by $\hat{u}_1, \hat{u}_2, \hat{u}_3$ the singular vectors of $\hat{\Sigma} - \sigma^2 \mathrm{Id}_{3\times 3}$, so $\hat{U} = (\hat{u}_1 | \hat{u}_2 | \hat{u}_3)$. Assuming that $r = 2$, an estimator $\hat{B}$ of B can be defined as follows

$$\hat{B} = \hat{D}_{|2}^{1/2} \hat{U}_{|2}^T = \begin{pmatrix} \sqrt{\hat{d}_1} & 0 \\ 0 & \sqrt{\hat{d}_2} \end{pmatrix} \begin{pmatrix} \hat{u}_1^T \\ \hat{u}_2^T \end{pmatrix}. \tag{30}$$

2.4 Imputation of the data matrix

In the previous sections, estimators for the mean, variance and covariances related to the missing variable have been proposed, as well as for the loading matrix B. All these estimators can be reused to impute missing values in the data matrix Y, using their estimated conditional expectation. Indeed, denoting $A = B^T B + \sigma^2 \mathrm{Id}_{3\times 3}$, for $i \in \{1, \ldots, n\}$, one has

$$\mathbb{E}[Y_{i1} | Y_{\cdot 2}, Y_{\cdot 3}] = \alpha_1 + \begin{pmatrix} A_{12} & A_{13} \end{pmatrix} \begin{pmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{pmatrix}^{-1} \left(\begin{pmatrix} Y_{i2} \\ Y_{i3} \end{pmatrix} - \begin{pmatrix} \alpha_2 \\ \alpha_3 \end{pmatrix} \right).$$

Assuming that the level of noise σ^2 is known, an imputation method dealing with MNAR missing values is described in the following algorithm.

Algorithm 1 Proposed method to impute missing values in the data matrix in the toy example setting, namely when $p = 3$, $r = 2$, and only one variable $Y_{\cdot 1}$ is likely to be missing.

Require: $r = 2$ and σ^2 known.

1: Evaluate $\hat{\alpha}_1$ an estimator of the missing variable mean given in (9).
2: Evaluate $\hat{\Sigma}$ an estimator of the covariance matrix using (28).
3: Compute the estimator $\hat{B}$ of the loading matrix, given in (30), with $r = 2$.
4: Compute

$$\hat{A} = \hat{B}^T \hat{B} + \sigma^2 \mathrm{Id}_{3\times 3}.$$

5: Impute the missing values (Y_{i1}) for $i \in \{1, \ldots, n\}$ such that $\Omega_{i1} = 0$ as follows

$$\hat{Y}_{i1} = \hat{\alpha}_1 + \begin{pmatrix} \hat{A}_{12} & \hat{A}_{13} \end{pmatrix} \begin{pmatrix} \hat{A}_{22} & \hat{A}_{23} \\ \hat{A}_{32} & \hat{A}_{33} \end{pmatrix}^{-1} \left(\begin{pmatrix} Y_{i2} \\ Y_{i3} \end{pmatrix} - \begin{pmatrix} \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{pmatrix} \right).$$

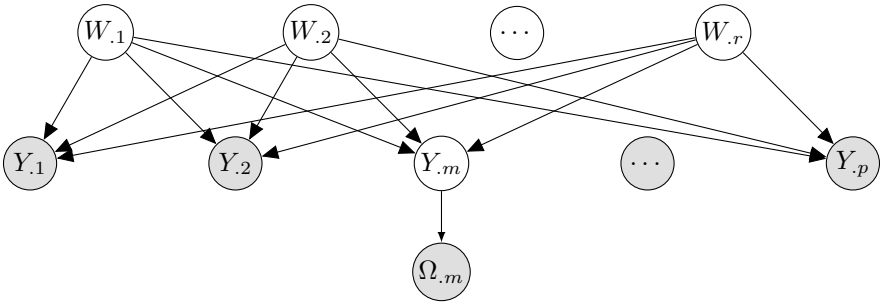


Figure 2: Graphical model for PPCA with p covariates, r latent variables and one self-masked MNAR missing variable $Y_{\cdot m}$.

3 Estimation and imputation under MNAR mechanism

In this section, still under the PPCA model given in (1) and recalled here

$$Y = \mathbf{1}\alpha + WB + \epsilon.,$$

the methodology presented in Section 2 is extended to the general case, for any data with p covariates, r latent variables and d missing variables denoted by $Y_{m_1}, \ldots, Y_{m_d}$ (with $d < p$). As visible in its corresponding graphical representation in Figure 2, we assume a "dense" or "fully-connected" PPCA model, meaning that all the latent variables are connected to each missing/observed variable.
In this section, for this general setting, the means of the MNAR variables of Y, their variances and associated covariances, are shown to be possibly consistently estimated, using a generalization of the method described for the toy example. In the same way, the loading matrix B is proposed to be estimated, and the data matrix Y can be imputed.
Since this section is a direct extension of the approach described in Section 2, we only present the final estimators in the general case, and intermediate steps to get them are enclosed in Appendix III.

Remark 18. *This section is only presented from the algebraic point of view. Indeed, in practice in the graphical point of view, the toy example with two latent variables is always considered, as the general setting can be reduced to it. This amounts to only using two observed variables to construct the mean, variance and covariances estimators of the missing variables, regardless of the data matrix dimensions and rank. This choice is explained in [15], the results being based on graphical models.*

3.1 Mean estimation

Based on a direct extension of Proposition 6, mean estimators of all MNAR missing variables can be obtained handling any value of r and p. In the toy example of Section 2, in order to estimate the mean of the missing variable $Y_{\cdot 1}$, two other (observed) variables $Y_{\cdot 2}$ and $Y_{\cdot 3}$ with recoverable means were needed. Dealing with arbitrary rank r and dimension p, the missing variables means can be estimated one by one, using r other variables with means that can be recovered (i.e. a fully observed variable or a MCAR missing variable for instance). That is why in the following definition, the mean estimator is made explicit for a particular missing variable. Note that this definition is a byproduct of Proposition 30 derived in the general case, being the counterpart of Proposition 6 in the toy example.

Definition 19 (Mean estimator)**.** *Consider the PPCA model given in* (1)*. An estimator of the mean of a MNAR variable $Y_{.m}$ is constructed using r variables $Y_{.j_1}, \ldots, Y_{.j_r}$ with means that can be recovered as*

$$\hat{\alpha}_m := \frac{\hat{\alpha}_{j_1} - \hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}[0]} - \sum_{k \in \mathcal{J}_{-j_1}} \hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}[k]} \hat{\alpha}_k}{\hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}[m]}}, \tag{31}$$

with $\mathcal{J} := \{j_1, j_2, \ldots, j_r\}$ and $\mathcal{J}_{-j_1} := \mathcal{J} \setminus \{j_1\} = \{j_2, \ldots, j_r\}$, and

- *$\hat{\alpha}_{j_1}, \ldots, \hat{\alpha}_{j_r}$, some estimators of $\alpha_{j_1}, \ldots, \alpha_{j_r}$, computed with the empirical mean,*
- *For $k \in \mathcal{J}_{-j_1}$, $\hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}[0]}$, $\hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}[m]}$ and $\hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}[k]}$ some estimators of $\mathcal{B}^c_{j_1 \to m, \mathcal{J}_{-j_1}[0]}$, $\mathcal{B}^c_{j_1 \to m, \mathcal{J}_{-j_1}[m]}$ and $\mathcal{B}^c_{j_1 \to m, \mathcal{J}_{-j_1}[k]}$ computed in the complete case, being the coefficients of the regression of $Y_{.j_1}$ on $(Y_{.m}, (Y_{.l})_{l \in \mathcal{J}_{-j_1}})$.*

Consistency of the mean estimator directly follows, provided consistent estimators of the $(\mathcal{B}^c_{i \to j,k})$'s. Note that with d MNAR variables, the mean of each one can be estimated using Definition 19, implying that the number d of MNAR variables should at most satisfy

$$d < p - r.$$

Note also that the need of r variables which means are recoverable can be restricted to a stronger assumption, such as the existence of r variables of Y completely observed. The choice of these r variables in practice is discussed in Section 4.

3.2 Variance and covariances estimation

Estimators of the variances and covariances of all MNAR missing variables are directly obtained by following the proof steps of the toy example and extending them to arbitrary dimension and rank. It is worth noting that two scenarios are possible: evaluating the covariance between a MNAR missing variable and a variable with recoverable mean/variance, or evaluating the covariance between two MNAR missing variables.

Variance and covariances between missing variables and recoverable-mean/variance ones. Despite the possibility of several MNAR missing variables, the study is conducted for a missing variable at once. For a given missing variable $Y_{.m}$, select r variables $Y_{.j_1}, \ldots, Y_{.j_r}$ with recoverable mean and variance. As previously, the starting point is to derive equations linking the variance and associated covariances between $Y_{.m}$ and these r selected variables. This is addressed in Proposition 31 of Appendix III dealing with general r and p (being the homologous result in Proposition 14, leading to the following estimators.

Definition 20 (Variance and covariances estimators)**.** *Consider the PPCA model given in* (1)*. An estimator of the variance and the covariances of a MNAR variable $Y_{.m}$ is constructed using r variables*

$Y_{.j_1}, \ldots, Y_{.j_r}$ *with means and variances that can be recovered as*

$$\begin{pmatrix} \widehat{\mathrm{Var}}(Y_{.m}) \\ \widehat{\mathrm{Cov}}(Y_{.m}, Y_{.j_1}) \\ \widehat{\mathrm{Cov}}(Y_{.m}, Y_{.j_2}) \\ \vdots \\ \widehat{\mathrm{Cov}}(Y_{.m}, Y_{.j_r}) \end{pmatrix} := (\hat{M}_1^\star)^{-1} \hat{M}_2^\star, \tag{32}$$

(provided that $(\hat{M}_1^\star)^{-1}$ *exists) with*

$$\hat{M}_1^\star = \begin{pmatrix} (\hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}[m]})^2 & 0 & 2\hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}[j_1]} \hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}[j_2]} & \cdots & 2\hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}[j_1]} \hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}[j_r]} \\ \hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}[m]} & 1 & -\hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}[j_2]} & \cdots & -\hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}[j_r]} \\ & & \ddots & & \\ & & & \ddots & \\ -\hat{\mathcal{B}}^c_{j_r \to m, \mathcal{J}_{-j_r}[m]} & -\hat{\mathcal{B}}^c_{j_r \to m, \mathcal{J}_{-j_r}[j_1]} & -\hat{\mathcal{B}}^c_{j_r \to m, \mathcal{J}_{-j_r}[j_2]} & \cdots & 1 \end{pmatrix},$$

$$\hat{M}_2^\star = \begin{pmatrix} \widehat{\mathrm{Var}}(Y_{.j_1}) - \hat{Q}^{\star c} - (\hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}[\mathcal{J}_{-j_1}]})^T \widehat{\mathrm{Var}}(Y_{\mathcal{J}_{-j_1}}) \hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}[\mathcal{J}_{-j_1}]} \\ (\hat{\mathcal{B}}^c_{j_1 \to m, \mathcal{J}_{-j_1}})^T \begin{pmatrix} 1 & \hat{\alpha}_m & \hat{\alpha}_{.j_1} & \ldots & \hat{\alpha}_{.j_r} \end{pmatrix}^T - \hat{\alpha}_{.j_1} \hat{\alpha}_{.m} \\ \vdots \\ (\hat{\mathcal{B}}^c_{j_r \to m, \mathcal{J}_{-j_r}})^T \begin{pmatrix} 1 & \hat{\alpha}_m & \hat{\alpha}_{.j_1} & \ldots & \hat{\alpha}_{.j_r} \end{pmatrix}^T - \hat{\alpha}_{.j_r} \hat{\alpha}_{.m} \end{pmatrix},$$

where

- $\widehat{\mathrm{Var}}(Y_{.k}), k \in \{j_1, \ldots, j_r\}$ *and* $\widehat{\mathrm{Cov}}(Y_{.k}, Y_{.l}), k \in \{j_1, \ldots, j_r\}$ *are the empirical variances and covariances of the* r *selected variables* $Y_{.j_1}, \ldots, Y_{.j_r}$,
- $\hat{\alpha}_{j_1}, \ldots, \hat{\alpha}_{j_r}$ *are the empirical means of* $Y_{.j_1}, \ldots, Y_{.j_r}$,
- $\hat{\alpha}_m$, *the estimator of the MNAR missing variable mean* α_m *given in Definition 19,*
- *For* $j \in \{j_1, \ldots, j_r\}$ $\hat{\mathcal{B}}^c_{j \to m, \mathcal{J}_{-j}[0]}$, *and* $\hat{\mathcal{B}}^c_{j \to m, \mathcal{J}_{-j}[m]}$ *are respectively estimators of* $\mathcal{B}^c_{j \to m, \mathcal{J}_{-j}[0]}$, *and* $\mathcal{B}^c_{j \to m, \mathcal{J}_{-j}[m]}$ *being the coefficients of the regression of* $Y_{.j}$ *on* $(Y_{.m}, (Y_{.l})_{l \in \mathcal{J}_{-j}})$, *computed in the complete case. Moreover,* $\hat{\mathcal{B}}_{j \to m, \mathcal{J}_{-j}}$ *denotes the entire estimated vector of coefficients of the regression of* $Y_{.j}$ *on* $(Y_{.m}, (Y_{.l})_{l \in \mathcal{J}_{-j}})$.

Consistency of the variance and covariances estimators in $\hat{X}$ directly follows from the consistency of the estimators $(\hat{\mathcal{B}}^c_{i \to j,k})$'s, when σ^2 tends to zero.

Remark 21 (About the covariances between the missing variable and all recoverable-mean/variance variables)**.** *Note that to estimate the variance of the missing variable, only* r *extra variables indexed by* $\{j_1, \ldots, j_r\}$ *in Definition 20, are required to solve (32). However, in order to evaluate* $\mathrm{Cov}(Y_{.m}, Y_{.j})$ *for all* j *such that* $Y_{.j}$ *is a variable with recoverable mean and variance and such that* $j \notin \{j_1, \ldots, j_r\}$, *one could simply loop on the other variables by applying Definition 20 to groups of variables of size* r.

Variance and covariances between two MNAR missing variables. In this part, the emphasis is put on estimating $\mathrm{Cov}(Y_{.m_1}, Y_{.m_2})$ for $Y_{.m_1}$ and $Y_{.m_2}$ two MNAR missing variables (not necessarily under the same missing mechanism). To do so, $r-1$ recoverable-mean/variance variables are needed (contrary to r as previously). Nevertheless, similar techniques are used based on regressions over $r+1$ variables, including the $r-1$ selected recoverable-mean/variance variables and the two MNAR missing variables. Note that the regressions are performed in the complete case for both missing variables, i.e. for individuals i such that $\Omega_{im_1} = 1$ and $\Omega_{im_2} = 1$.

Definition 22 (Covariance estimator between two missing variables)**.** *Let us denote $\mathcal{H} = \mathcal{J} \cup \{m_1, m_2\}$. Under the PPCA model given in (1), an estimator of the covariance between two MNAR missing variables, denoted $Y_{.m_1}$ and $Y_{.m_2}$, is constructed using $r-1$ variables $Y_{.j_1}, \ldots, Y_{.j_{r-1}}$ with means and variances that can be recovered as*

$$\hat{K}\widehat{\mathrm{Cov}}(Y_{.m_1}, Y_{.m_2}) = \widehat{\mathrm{Var}}(Y_{.j_1}) - \hat{Q}^{\star,c} - \sum_{k \in \mathcal{H}_{-j_1}} (\hat{\mathcal{B}}^c_{j_1 \to \mathcal{H}_{-j_1}[k]})^2 \widehat{\mathrm{Var}}(Y_{.k}) - \sum_{\substack{k \neq l \\ k \in \mathcal{H}_{-j_1}, l \in \mathcal{H}_{-(j_1,m_1,m_2)}}} 2\hat{\mathcal{B}}^c_{j_1 \to \mathcal{H}_{-j_1}[k]} \hat{\mathcal{B}}^c_{j_1 \to \mathcal{H}_{-j_1}[l]} \widehat{\mathrm{Cov}}(Y_{.k}, Y_{.l}),$$

(provided $\hat{K} \neq 0$) with $\hat{K} = 2\hat{\mathcal{B}}^c_{j_1 \to \mathcal{H}_{-j_1}[m_1]} \hat{\mathcal{B}}^c_{j_1 \to \mathcal{H}_{-j_1}[m_2]}$ and

$$\hat{Q}^{\star c} = \left(\widehat{\mathrm{Var}}(Y_{.j_1}) - \widehat{\mathrm{Cov}}(Z^\star, Y_{.j_1}) \widehat{\mathrm{Var}}(Z^\star) \widehat{\mathrm{Cov}}(Z^\star, Y_{.j_1})^T \big| \Omega_{.m_1} = 1, \Omega_{.m_2} = 1 \right),$$

where $Z^\star = [Y_{.m_1}, Y_{.m_2}, Y_{.j_2}, \ldots, Y_{.j_r}]$.

Then, this estimator is consistent, provided consistent estimators of the $(\mathcal{B}^c_{i \to j,k})$'s, the variance and covariances of the missing variables, i.e. $\mathrm{Var}(Y_{.m_1})$, $\mathrm{Var}(Y_{.m_2})$ and $\mathrm{Cov}(Y_{.l}, Y_{.m_k})$ for $l \in \mathcal{H}_{-j_1}, k \in \{1, 2\}$.

Covariance matrix estimator. Compiling all the previous estimators, one can form an estimator $\hat{\Sigma}$ for the covariance matrix as follows

$$\hat{\Sigma} = \left(\widehat{\mathrm{Cov}}(Y_{.j}, Y_{.k}) \right)_{j,k \in \{1,\ldots,p\}}. \tag{33}$$

where

- when $Y_{.j}$ is a MNAR missing variable and $Y_{.k}$ is a recoverable-mean/variance variable, $\widehat{\mathrm{Cov}}(Y_{.j}, Y_{.k})$ is given in Equation (32),
- when $Y_{.j}$ and $Y_{.k}$ are both MNAR missing variables, $\widehat{\mathrm{Cov}}(Y_{.j}, Y_{.k})$ is given in Equation (49),
- when $Y_{.j}$ and $Y_{.k}$ are both recoverable-mean/variance variables, $\widehat{\mathrm{Cov}}(Y_{.j}, Y_{.k})$ can be evaluated by the standard empirical covariance estimator.

3.3 Estimation of the loading matrix

Methods for the estimation of the PPCA loading matrix, presented for the toy example in Section 2, can be extended to arbitrary rank r and dimension p. Once the variances and covariances estimated, Assuming that the level of noise σ^2 is known and that the rank r of B is known, the following definition can be used to derive an estimator $\hat{B}$ of the loading matrix B: it is based on the singular value decomposition of the matrix $\hat{\Sigma} - \mathrm{Id}_{p\times p}$.

Definition 23 (Estimation of the loading matrix)**.** *Given the estimator $\hat{\Sigma}$ of the covariance matrix in (33), let the orthogonal matrix $\hat{U} \in \mathbb{R}^{p\times p}$ and the diagonal matrix $\hat{D} = \mathrm{diag}(\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_p) \in \mathbb{R}^{p\times p}$ with $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$ form the singular value decomposition of the following matrix*

$$\hat{\Sigma} - \sigma^2 \mathrm{Id}_{p\times p} =: \hat{U}\hat{D}\hat{U}^T,$$

and denote by $\hat{u}_1, \ldots, \hat{u}_p$ the singular vectors of $\hat{\Sigma} - \sigma^2\mathrm{Id}_{p\times p}$, so $\hat{U} = (\hat{u}_1|\ldots|\hat{u}_p)$. An estimator $\hat{B}$ of B can be defined using the r first singular vectors of the previous decomposition, such as

$$\hat{B} = \hat{D}_{|r}^{1/2}\hat{U}_{|r}^T = \begin{pmatrix} \sqrt{\hat{d}_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\hat{d}_r} \end{pmatrix} \begin{pmatrix} \hat{u}_1^T \\ \hline \vdots \\ \hline \hat{u}_r^T \end{pmatrix}. \tag{34}$$

In practice, one will see in Section 4 that the proposed estimator $\hat{B}$ leads to a good estimation of the coefficient matrix B. The method presented here thus makes it possible to empirically estimate the loading matrix within the MNAR setting, which is, to our knowledge, the first proposed approach to do so. For MCAR and MAR data, an Expectation-Maximization algorithm extended to the missing data case is usually applied to recover B [3, 2].

3.4 Imputation of the data matrix

In the previous sections, estimators for the mean, variance and covariances related to missing variables have been proposed, as well as for the loading matrix B. All these estimators can be reused to impute missing values in the data matrix Y, using their estimated conditional expectation, extending Algorithm 1 to arbitrary rank r and dimension p. For the sake of clarity, let us consider that the observed variables are $Y_{.1}, \ldots, Y_{.p-d}$ and the MNAR missing variables are $Y_{.p-d+1}, \ldots, Y_{.p}$. Denoting $A = B^TB + \sigma^2\mathrm{Id}_{p\times p}$, for $i \in \{1, \ldots, n\}$ and $j \in \{p-d+1, \ldots p\}$, one has

$$\mathbb{E}[Y_{ij}|Y_{.1}, \ldots, Y_{.p-d}] = \alpha_j + \begin{pmatrix} A_{j1} & \ldots & A_{j(p-d)} \end{pmatrix} \begin{pmatrix} A_{11} & \ldots & A_{1(p-d)} \\ & \ddots & \\ A_{(p-d)1} & \ldots & A_{(p-d)(p-d)} \end{pmatrix}^{-1} \left(\begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{i(p-d)} \end{pmatrix} - \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{p-d} \end{pmatrix} \right).$$

Assuming that the level of noise σ^2 is known, an imputation method dealing with MNAR missing values is described in Algorithm 2.
In practice, one will see in Section 4 that this imputation method gives good results and allows to impute a matrix under the PPCA model, which contains MNAR missing values.

Algorithm 2 Proposed method to impute missing values in the data matrix with d MNAR missing variables, denoted by $Y_{.p-d+1}, \dots, Y_{.p}$.

Require: r and σ^2 known.

1: **for** the MNAR missing variables indexed by $j \in \{p-d+1, \dots, p\}$ **do**

2: Evaluate $\hat{\alpha}_j$ the estimator of the missing variable mean given in (31) using r observed variables.

3: Evaluate $\widehat{\mathrm{Var}(Y_{.j})}$, and $\widehat{\mathrm{Cov}}(Y_{.j}, Y_{.k})$ with $k \in \{1, \dots, d-p\}$ using (32) based on r observed variables.

4: Evaluate $\widehat{\mathrm{Cov}}(Y_{.j}, Y_{.k})$ with $k \in \{p-d+1, \dots, j-1\}$ using Proposition 32 based on $r-1$ observed variables.

5: **end for**

6: Form $\hat{\Sigma}$ the estimator of the covariance matrix using the previous estimations and standard empirical estimators for variances and covariances between fully observed variables.

7: Compute the estimator $\hat{B}$ of the loading matrix, given in (34).

8: Compute

$$\hat{A} = \hat{B}^T \hat{B} + \sigma^2 \mathrm{Id}_{p \times p}.$$

9: Impute the missing values (Y_{ij}) for $i \in \{1, \dots, n\}$ such that $\Omega_{ij} = 0$ and $j \in \{p-d+1, \dots, p\}$ as follows

$$\hat{Y}_{ij} = \hat{\alpha}_j + \begin{pmatrix} \hat{A}_{j1} & \dots & \hat{A}_{j(p-d)} \end{pmatrix} \begin{pmatrix} \hat{A}_{11} & \dots & \hat{A}_{1(p-d)} \\ & \ddots & \\ \hat{A}_{(p-d)1} & \dots & \hat{A}_{(p-d)(p-d)} \end{pmatrix}^{-1} \left(\begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{i(p-d)} \end{pmatrix} - \begin{pmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_{p-d} \end{pmatrix} \right).$$

4 Numerical experiments

The data matrix Y is generated from a PPCA model given in Equation (1). The MNAR missing values are introduced using a logistic regression given by

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

for all $i \in \{1\dots,n\}$ and where for j indexing a missing variable, $\phi_j = (\phi_{1j}, \phi_{2j})$ denotes a parameter vector of the missingness distribution. The means, the variances, and covariances of the variables with missing values are estimated from the incomplete data. The PPCA loading matrix is also estimated and the data matrix is imputed. To do so, the following methods are compared, denoted by the following keywords:

(a) MNAR: both graphical and algebraic methods, developed in this paper which consider the MNAR feature of the missing mechanism but do not assume a parametric model for it,

- Graphical: refers to the graphical approach, in Equations (11), (25), and (26),
- Algebraic: refers to the algebraic approach, in Equations (9),(23) and (27);

(b) MAR: application to the PPCA model of the method suggested in [14, Theorems 1, 2, 3], the latter being designed to handle MAR missing values in linear models. See Appendix III for details;

(c) Mean: the imputation by the mean which consists in imputing the missing values by the mean of the variables computed over the observed entries. This can serve as a benchmark;

(d) Del: the listwise deletion method with consists in estimating the parameters with the fully-observed rows only.

In addition, two methods are also implemented, which are designed to handle fixed effects model, *i.e.* where the data $Y \in \mathbb{R}^{n\times p}$ is generated as a sum of a low-rank matrix $\Theta \in \mathbb{R}^{n\times p}$ (the rank r of Θ satisfies $r < \min\{n,p\}$) and a Gaussian noise matrix, i.e.

$$Y = \Theta + \epsilon. \tag{35}$$

These two methods are called

(e) SoftMAR: which minimizes the weighted least squares penalized by the nuclear norm [13] using the algorithm `softImpute` [5] which is appropriate under the MCAR or MAR assumption;

(f) Param: the parametric method suggested in [19] which parameterizes the MNAR mechanism using a logistic model. More particularly, in order to estimate Θ, this method minimizes the penalized negative joint-likelihood as follows

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

where $\|.\|_\star$ is the nuclear norm, known to be a convex relaxation of the rank penalty. It is achieved using a Monte-Carlo Expectation Maximization algorithm, which can be computationally expensive.

The results are presented for different numbers of observations and variables, percentages of missing values, ranks and noise levels.

Measuring the performance The methods are compared in terms of quality of estimation and imputation. For the loading matrix, the RV coefficient [8] between the estimate $\hat{B}$ and the true B is computed, being an extension of the correlation coefficient for random vectors, particularly well fitted to compare spanned subspaces. An RV coefficient close to one means high correlation between the image spaces of $\hat{B}$ and B. The quality of imputation is measured with the normalized prediction error given by

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

Selection of the hyperparameters In the parametric method (f), the level of noise σ^2 is known. Note also that both methods (e) and (f) require the regularization parameter λ to be tuned. The complete matrix Y is thus used to choose the optimal λ among some fixed grid $\mathcal{G} = \{\lambda_1, \ldots, \lambda_M\}$ by minimizing the true prediction error. On the other hand, Methods (a) and (b) assume the rank r and the level of noise σ^2 to be known to estimate the loading matrix and to impute the data matrix. However, note that only the knowledge of the rank r is required to estimate the means, the variances and the covariances of the missing variables.
Besides, Method (a) involves the selection of observed variables on which the regression will be performed. Two approaches are then proposed:

- aggregation: in which the final estimator is provided by computing the median of intermediate mean or variance estimators corresponding to every possible combinations of the observed variables; this kind of method will be denoted in light blue in the following boxplots.
- random: the final estimator is built upon only one choice of fully observed variables, uniformly randomly drawn among all combinations of observed variables. This method will be denoted in dark blue in the following boxplots.

4.1 Numerical experiments for the PPCA model

PPCA model generated from two latent variables. In this section, a data matrix of size $n = 1000$ and $p = 10$ is generated from two latent variables ($r = 2$) and with a noise level $\sigma = 0.1$. Seven MNAR missing variables $Y_{\cdot j}, j \in \{1, 2, 3, 4, 5, 9, 10\}$ are introduced and the logistic parameters choice leads to 35% of missing values in total.
For instance, for the first missing variable (without lack of generality for other missing variables), Figure 3 and 4 show that our approach (a) is the only one which gives unbiased estimators of the mean and variances of $Y_{\cdot 1}$. Recall that estimators of the mean in both graphical and algebraic methods (Equation (11) on the one hand, and Equation (9) on the other hand) are the same but estimators of the variance (Equations (25) and (26) versus Equation (23)) differ. Figure 3 shows that the dispersion of the boxplots for the algebraic method is slightly larger than the one of the boxplots for the graphical approach, which can be due to the instability of the matrix inversion in (23) to get an estimator of the variance. In addition, as expected, the aggregation option on Method (a) for both graphical and algebraic approaches improves on the random option, i.e. when only one random combination of variables is used to compute estimators. However, this latter, computationally faster, still provides an unbiased estimate for the mean and the variance and proves to outperform the MAR method (b), which

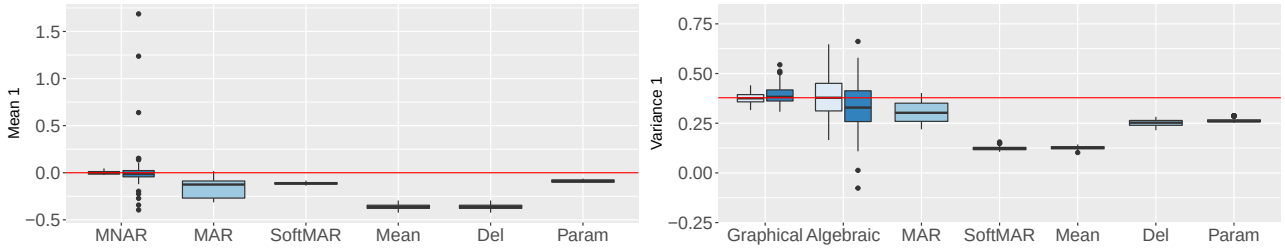


Figure 3: Mean and variance estimations of the missing variable $Y_{\cdot 1}$ when $r = 2$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

is the most competitive method ignoring the MNAR mechanism. As expected, Method (d) discarding individuals with missing variables provides biased estimate inasmuch as the observed sample is not representative of the population with MNAR data. The results obtained with the parametric method (f) are improved upon the benchmark mean imputation (c), and on Methods (d) and (e) as well, as it explicitly takes into account the MNAR nature of the missing entries. However, it still leads to biased estimates which can be explained by the fact that this method is developed under the fixed effect model given in (35), different from the random effects model of the PPCA. Note that similar results hold for the other six missing variables (see in Appendix IV, Figures 15, 16, 17, 18, 19 and 20).
In Figure 4, covariance estimations of $\mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot j}), j \in \{2, 3, 4\}$ are displayed. The MNAR method (a) combined with the algebraic approach provides unbiased estimates for all quantities. On the contrary, Method (a) combined with the graphical approach provides biased estimates of $\mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot j}), j \in \{3, 4\}$ and is in this particular case no longer competitive compared to the MAR method (b). As a matter of fact, the latter may be efficient, as it does not involve division by coefficients possibly close to 0 or matrix inversion.
Figure 5 shows that our method (a) considering the algebraic approach and the aggregation gives the best estimate of the loading matrix and the smallest imputation error. The graphical approach and the algebraic approach with a random choice of variables combination are no longer competitive, since outliers in estimates of means, variances and covariances have a significant impact in the estimation of B and the imputation.

PPCA model generated from three latent variables. Similar conclusions can be drawn when the rank is increased. A data matrix of size $n = 1000$ and $p = 10$ is generated from three latent variables ($r = 3$) with the same noise level $\sigma = 0.1$ and still seven MNAR missing variables $Y_{\cdot j}, j \in \{1, 2, 3, 4, 5, 9, 10\}$ introduced using a logistic model. Figures 6 and 7 show that our approach (a) remains the only one which gives unbiased estimators of the mean and variances of the first missing variable (and for the other ones as well, see Figures 21, 22, 23, 24, 25, 26 in Appendix IV). Moreover, in Figure 8, our method (a) considering the algebraic approach combined with aggregation still gives the best estimate of the loading matrix and the smallest imputation error, despite a larger dispersion in the boxplots compared to MAR and other methods. Nevertheless, the proposed method (a) combined

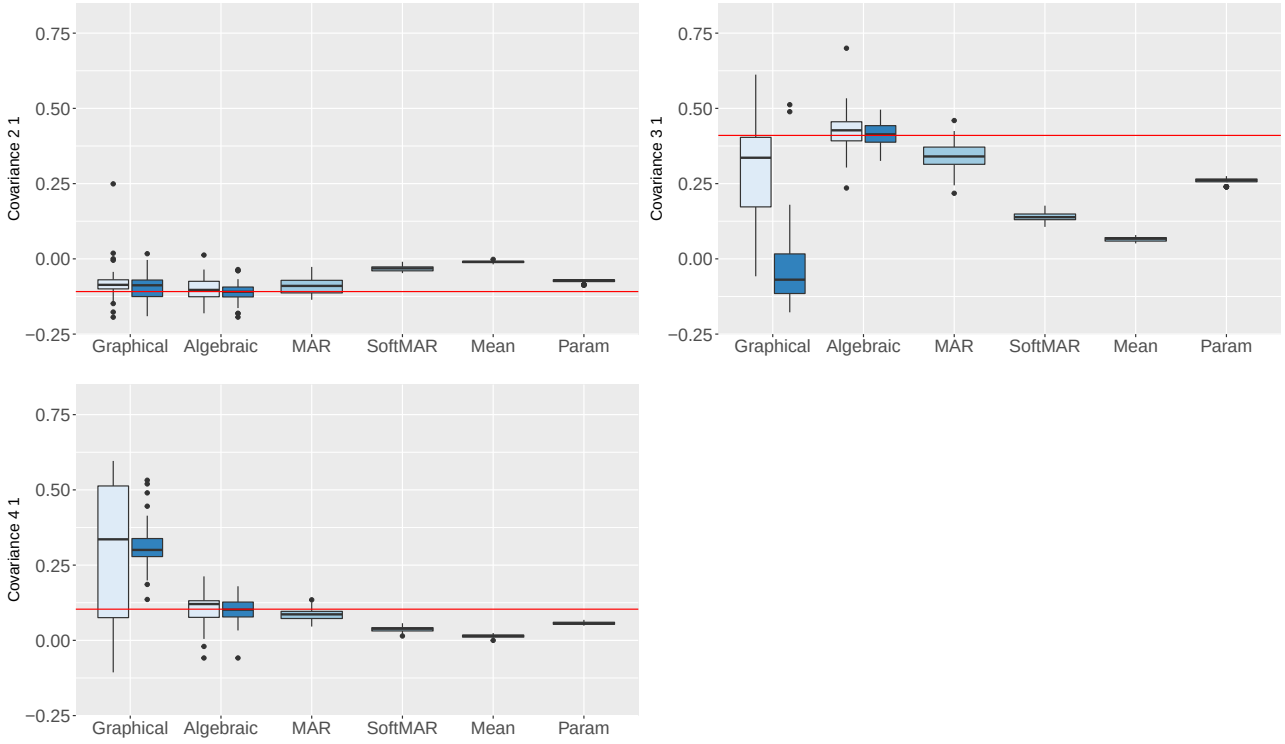


Figure 4: Covariances estimations of $\mathrm{Cov}(Y_{.1}, Y_{.j}), j \in \{2,3,4\}$ when $r = 2$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

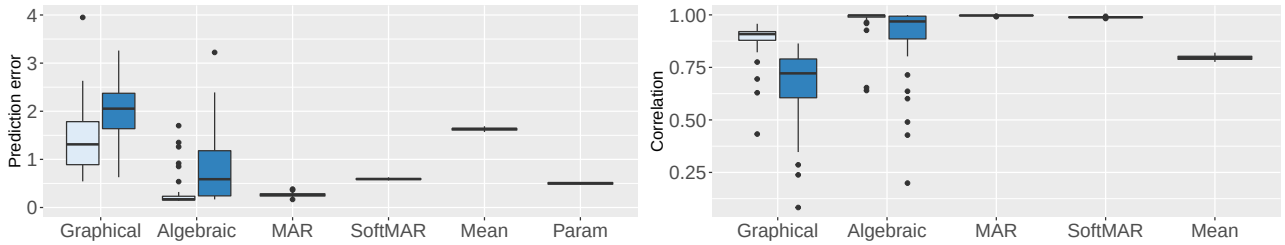


Figure 5: Prediction error in imputation and RV coefficient for the loading matrix when $r = 2$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination.

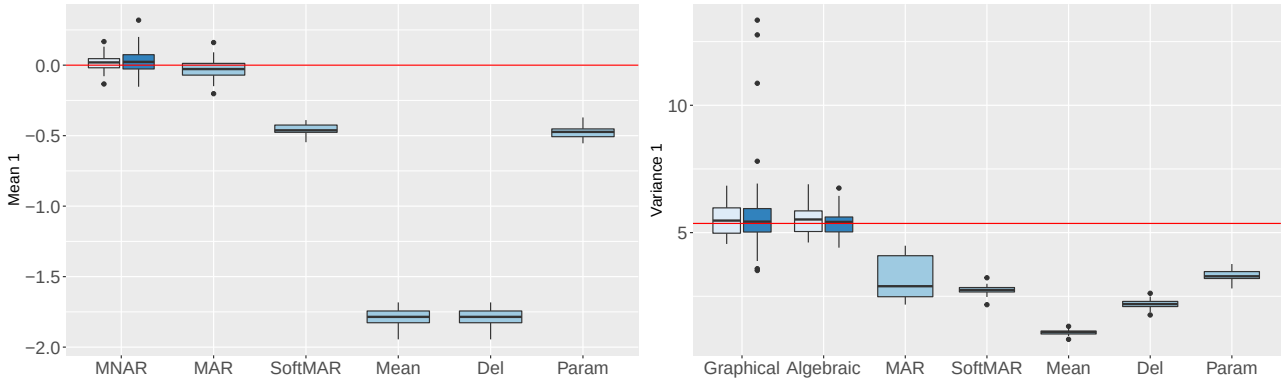


Figure 6: Mean and variance estimations of the missing variable $Y_{.1}$ when $r = 3$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

with the random choice performs poorly (either with the algebraic or the graphical approach) in terms of estimation of B and imputation, compared to the MAR (b) and the SoftMAR (e) methods.

4.2 Robustness to noise

Using the same setting as in PPCA model generated from three latent variables, the results are now presented for different noise levels $\sigma^2 = \{0.1, 0.3, 0.5, 0.7, 1\}$. Four methods are compared:

- Method (a) combined with the aggregation approach (i.e. aggregating estimators provided by all combinations of observed variables on which the regression will be performed), either relying on algebraic arguments, or graphical ones,
- Method (e) using softImpute and ignoring the MNAR mechanism and
- the naive one with mean imputation (c).

For instance, for the first missing variable, when increasing the noise level, the proposed estimators for Method (a) still considerably improve on the others in terms of quality of estimation for the mean in Figure 9 and the variance in Figure 10. As expected, the boxplots dispersion tends to increase with noise level. For all the other missing variables, the results are similar but not shown here due to space constraints.
Figure 11 and 12 show the correlation between the estimation of the loading matrix and the true one, as well as the prediction error. When the noise level increases, it is expected that the linear equations used at the start of the analysis, such as (4) for the toy example, will be less and less exogenous and that ignoring it in practice can be made to the detriment of performance. As expected, estimation deteriorates as the data gets noisier and then the loading matrix estimation and the prediction error get closer to the results of mean imputation. The proposed method yet remains competitive in regards

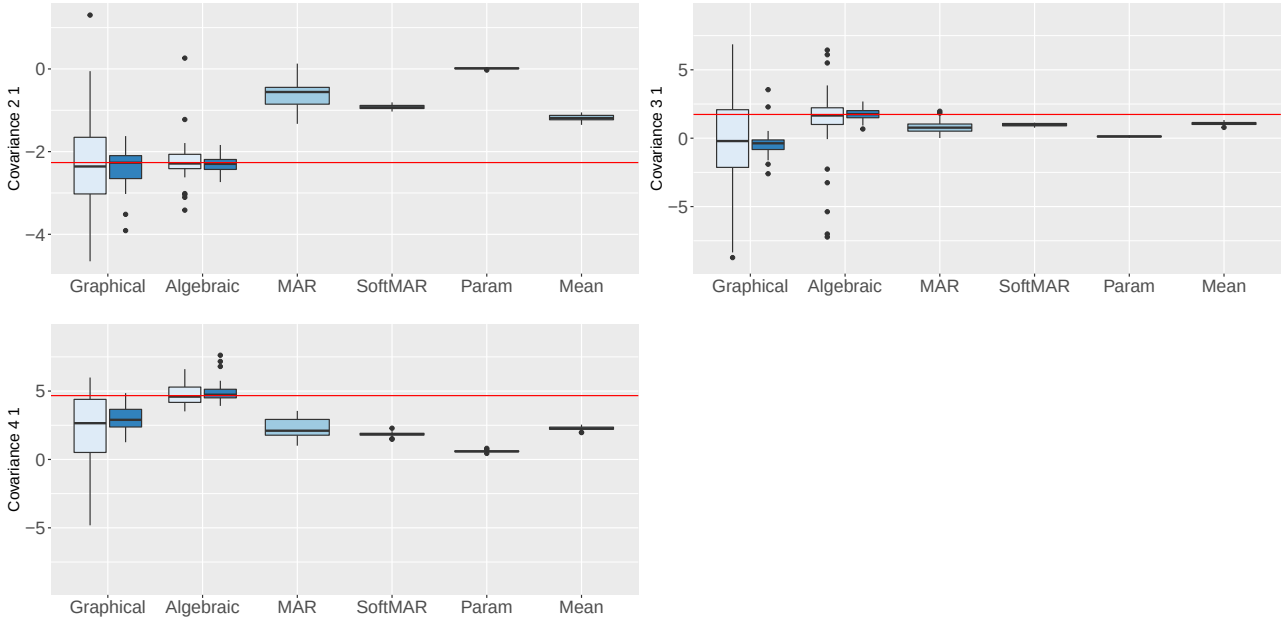


Figure 7: Covariances estimations of $\mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot j}), j \in \{2, 3, 4\}$ when $r = 3$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

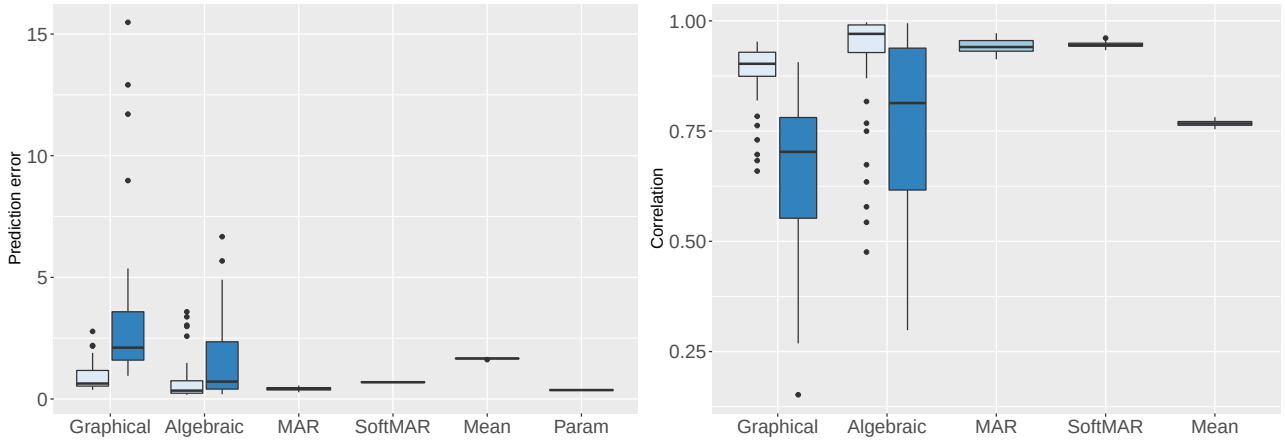


Figure 8: Prediction error for imputation and RV coefficient for the loading matrix when $r = 3$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination.

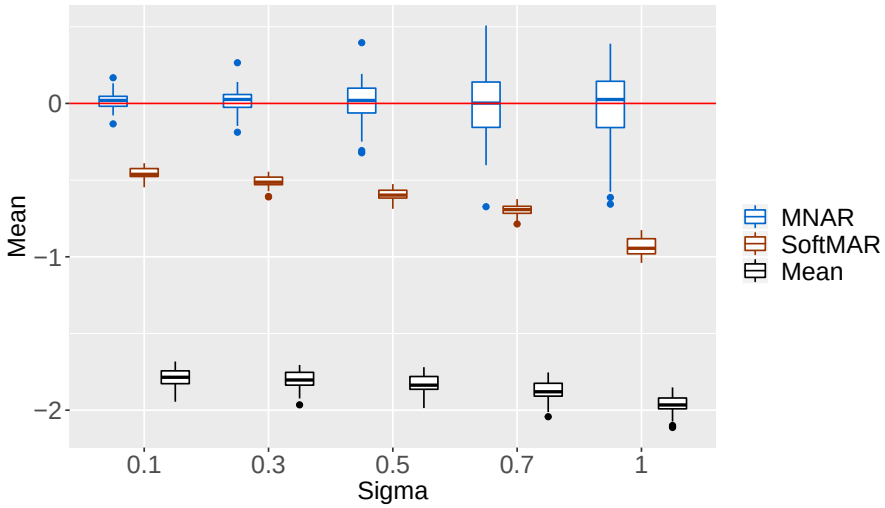


Figure 9: Mean estimation for different values of the level of noise when $r = 3$, $n = 1000$, $p = 10$ and 7 variables are missing leading to 35 % of MNAR values. The aggregation approach, which chooses the observed variables on which the regression will be formed by aggregating every possible combination, is used here for both graphical and algebraic methods. The red lines indicate the true values.

of the approach (e) until the level of noise reaches $\sigma = 1$ for the prediction error and until $\sigma = 0.5$ for the loading matrix estimation. Remark also that the graphical approach seems to have a smaller prediction error than the algebraic one (for $\sigma = 0.7$ and $\sigma = 1$) when the noise increases; it could be due to the growing impact of the matrix inversion in the algebraic approach.

4.3 Misspecification to the PPCA model

In this section, the methods stability to a wrong model specification is evaluated. The data matrix Y of size $n = 200$ and $p = 10$ is generated under the fixed effects model as (35) with a rank $r = 3$ (for Θ) and a noise level $\sigma = 0.1$. There again, seven MNAR missing variables $Y_{.j}, j \in \{1, 2, 3, 4, 5, 9, 10\}$ are introduced, resulting in 35% missing data in the whole matrix. For instance, for $Y_{.1}$ (similar results are obtained for the other missing variables), Figure 13 shows that, regarding the mean and variance estimations, our method (a) provides less biased estimates than the parametric one (f), while precisely dedicated to this specific setting.
Note that Method (f) based on the fixed effects model provides accurate estimation of the mean but the variance is slightly under-estimated, which is expected as the method imputes missing entries with $\hat{\Theta}$ and consequently the variability in the imputed data is smaller than the one in the observed data. As for the prediction performance, Figure 14 shows that our approach (a), using the aggregation option, does not remain competitive with Method (f). However, despite the model misspecification, it gives similar results as Method (e), which ignores the MNAR mechanism but is specially designed to handle fixed effect models. Our approach (a) with the random option is not competitive regarding to the prediction performance.

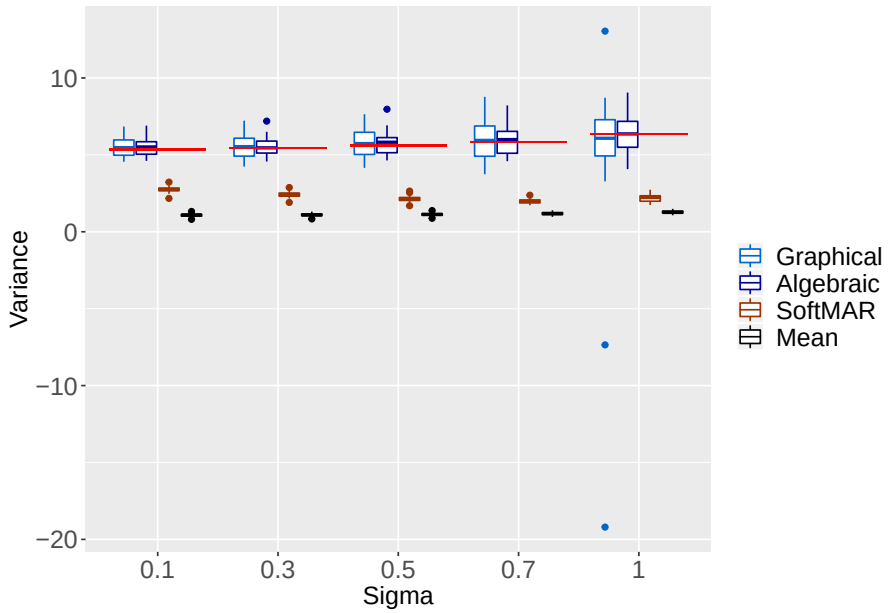


Figure 10: Variance estimations for different values of the level of noise when $r = 3$, $n = 1000$, $p = 10$ and 7 variables are missing leading to 35 % of MNAR values. The aggregation approach, which chooses the observed variables on which the regression will be formed by aggregating every possible combination, is used here for both graphical and algebraic methods. The red lines indicate the true values.

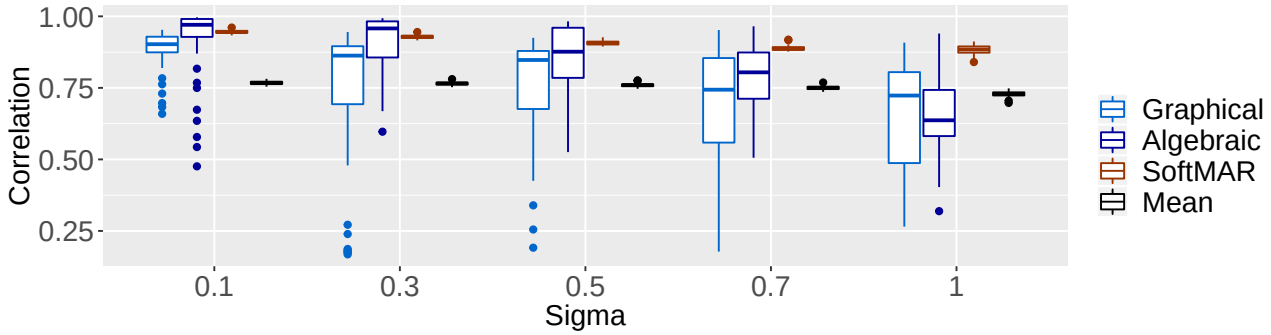


Figure 11: Correlation for different values of the level of noise when $r = 3$, $n = 1000$, $p = 10$ and 7 variables are missing leading to 35 % of MNAR values. The aggregation approach, which chooses the observed variables on which the regression will be formed by aggregating every possible combination, is used here for both graphical and algebraic methods.

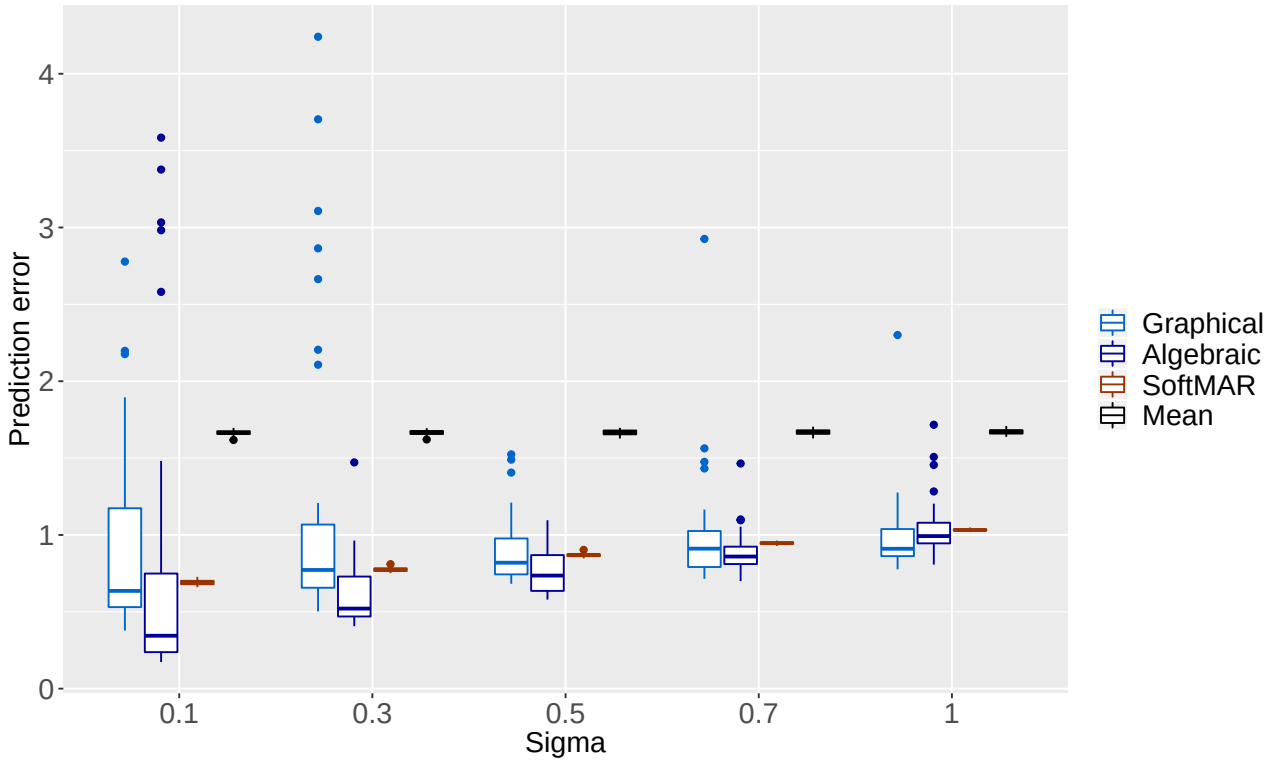


Figure 12: Prediction error for different values of the level of noise when $r = 3$, $n = 1000$, $p = 10$ and 7 variables are missing leading to 35 % of MNAR values. The aggregation approach, which chooses the observed variables on which the regression will be formed by aggregating every possible combination, is used here for both graphical and algebraic methods.

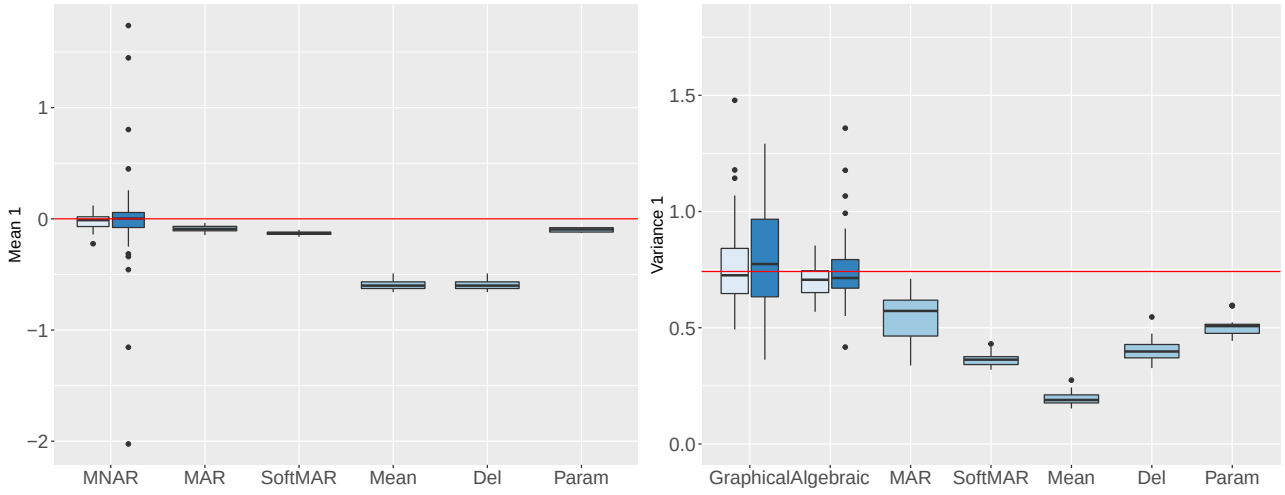


Figure 13: Mean and variance estimations for two different variables when data are generated under the fixed effects model given in (35), $r = 3$, $n = 200$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

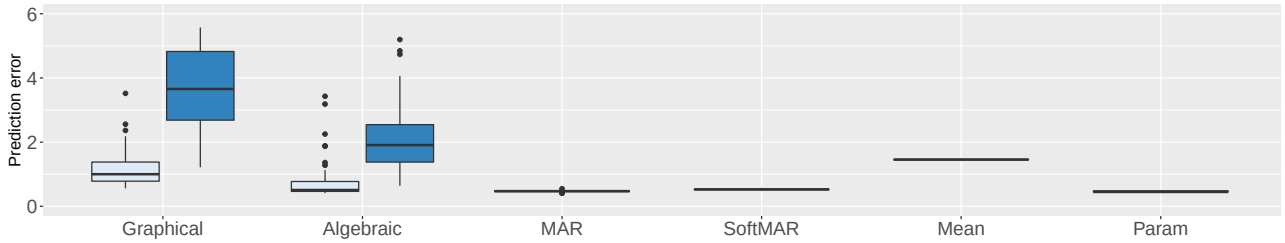


Figure 14: Prediction error when data are generated under the fixed effects model given in (35) $r = 3$, $n = 200$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination.

Conclusion

In this paper, we study estimation of mean, variance and covariances related to a self-masked MNAR missing variable in the context of the PPCA model. Despite the common belief of hardness for such MNAR missing values, information of interest can be retrieved by exploiting linear links between variables, which is particularly allowed by the PPCA model. This is at the core of the proposed estimators, enabled by a relatively simple technicality based only on linear regressions. As a matter of fact, the strength of such estimators is to be free from a specific modelling of the missing mechanism. In practice, the proposed estimators outperform standard estimators, generally designed for the MAR setting and by ignoring the MNAR missing mechanism.
The constructed estimators of the variances and covariances can be in turn used to estimate the loading matrix of the PPCA model and to impute missing entries in the data matrix. In simulations, this new method of imputation handling MNAR missing variables proves to be competitive in comparison to more involved techniques, such that parametric methods explicitly modelling the missing mechanism, which entails a computational burden.
Despite the non-exogeneity assumption in the theoretical framework, it seems that this assumption can be overlooked when it comes to numerical experiments. This is yet a clear limitation of the proposed method, that to our knowledge, is more tangible with the algebraic approach than the graphical one. Hence, it suggests some lines for theoretical improvement, specially when the noise level increases.
It should also be noted that the proposed method requires solving linear systems, which can be numerically instable: in simulation, one can see the presence of outliers in imputation generally due to outliers already present in the variance and covariances estimation. The robustification of the proposed approach is beyond the scope of this paper, that we think is already innovative enough, providing the first consistency results in presence of informative missing values in low-rank models.
As promising perspectives, this work could be extended to other variants of PPCA, such that the probabilistic Poisson PCA [4] covering the exponential family framework instead of the Gaussian one.

References

[1] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[2] John Canny. Collaborative filtering with privacy via factor analysis. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 238–245. ACM, 2002.

[3] Tao Chen, Elaine Martin, and Gary Montague. Robust probabilistic pca with missing data and contribution analysis for outlier detection. *Computational Statistics & Data Analysis*, 53(10): 3706–3716, 2009.

[4] Julien Chiquet, Mahendra Mariadassou, Stéphane Robin, et al. Variational inference for probabilistic poisson pca. *The Annals of Applied Statistics*, 12(4):2674–2698, 2018.

[5] Trevor Hastie and Rahul Mazumder. *softImpute: Matrix Completion via Iterative Soft-Thresholded SVD*, 2015. URL https://CRAN.R-project.org/package=softImpute. R package version 1.4.

[6] Trevor Hastie, Rahul Mazumder, Jason D Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1): 3367–3402, 2015.

[7] Joseph G Ibrahim, Stuart 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.

[8] Julie Josse, Jérome Pagès, and François Husson. Testing the significance of the rv coefficient. *Computational Statistics & Data Analysis*, 53(1):82–91, 2008.

[9] Julie Josse, Sylvain Sardy, and Stefan Wager. denoiser: A package for low rank matrix estimation. *Journal of Statistical Software*, 2016.

[10] Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078, 2010.

[11] N Kishore Kumar and Jan Schneider. Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra*, 65(11):2212–2244, 2017.

[12] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 333. John Wiley & Sons, 2014.

[13] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.

[14] Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. In *Advances in neural information processing systems*, pages 1277–1285, 2013.

[15] Karthika Mohan, Felix Thoemmes, and Judea Pearl. Estimation with incomplete data: The linear case. In *IJCAI*, pages 5082–5088, 2018.

[16] Kosuke Morikawa, Jae Kwang Kim, and Yutaka Kano. Semiparametric maximum likelihood estimation with data missing not at random. *Canadian Journal of Statistics*, 45(4):393–409, 2017.

[17] Judea Pearl. Causality: models, reasoning, and inference. *Econometric Theory*, 19(675-685):46, 2003.

[18] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[19] Aude Sportisse, Claire Boyer, and Julie Josse. Imputation and low-rank estimation with missing non at random data. *arXiv preprint arXiv:1812.11409*, 2018.

[20] Fei Tang and Hemant Ishwaran. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):363–377, 2017.

[21] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

I MAR formulae

The formulae are given in the toy example case (Section 2) for $p = 3$ and $r = 2$ with one missing at random variable $Y_{\cdot 1}$ and can be directly extended to any p and r. The following proposition is an extension of the results of Mohan et al. [15] (Theorem 1, 2, 3).

Proposition 24 (Expectation, variance and covariances formulae for a missing at random variable when $p = 3$ and $r = 2$)**.** *Under the PPCA model* (3)*, assume that:*

- $\begin{pmatrix} B_{\cdot 2} & B_{\cdot 3} \end{pmatrix}$ *is an invertible matrix,*
- $Y_{\cdot 1} \perp\!\!\!\perp \Omega_{\cdot 1} | Y_{\cdot 2}, Y_{\cdot 3}$.

One can derive

- *the mean of the missing variable*

$$\alpha_1 = \mathcal{B}^c_{1\to 2,3[0]} + \mathcal{B}^c_{1\to 2,3[2]}\alpha_2 + \mathcal{B}^c_{1\to 2,3[3]}\alpha_3,$$

- *the variance of the missing variable*

$$\mathrm{Var}(Y_{\cdot 1}) = Q^c_{\mathrm{MAR}} + (\mathcal{B}^c_{1\to 2,3[2]})^2\mathrm{Var}(Y_{\cdot 2}) + (\mathcal{B}^c_{1\to 2,3[3]})^2\mathrm{Var}(Y_{\cdot 3}) + 2\mathcal{B}^c_{1\to 2,3[2]}\mathcal{B}^c_{1\to 2,3[3]}\mathrm{Cov}(Y_{\cdot 2}, Y_{\cdot 3}),$$

with $Q^c_{\mathrm{MAR}} = \left(\mathrm{Var}(Y_{\cdot 1}) - \mathrm{Cov}([Y_{\cdot 2}, Y_{\cdot 3}], Y_{\cdot 1})\mathrm{Var}([Y_{\cdot 2}, Y_{\cdot 3}])\mathrm{Cov}([Y_{\cdot 2}, Y_{\cdot 3}], Y_{\cdot 1})^T \middle| \Omega_{\cdot 1} = 1\right)$,

- *the covariances associated to the missing variable*

$$\mathrm{Cov}(Y_{\cdot 2}, Y_{\cdot 1}) = \mathcal{B}^c_{1\to 2,3[0]}\mathbb{E}[Y_{\cdot 2}] + \mathcal{B}^c_{1\to 2,3[2]}(\mathrm{Var}(Y_{\cdot 2}) + \mathbb{E}[Y_{\cdot 2}]^2) + \mathcal{B}^c_{1\to 2,3[3]}(\mathrm{Cov}(Y_{\cdot 2}, Y_{\cdot 3}) + \mathbb{E}[Y_{\cdot 3}]\mathbb{E}[Y_{\cdot 2}]) - \mathbb{E}[Y_{\cdot 1}]\mathbb{E}[Y_{\cdot 2}] + o(\sigma^2$$

$$\mathrm{Cov}(Y_{\cdot 3}, Y_{\cdot 1}) = \mathcal{B}^c_{1\to 2,3[0]}\mathbb{E}[Y_{\cdot 3}] + \mathcal{B}^c_{1\to 2,3[3]}(\mathrm{Var}(Y_{\cdot 3}) + \mathbb{E}[Y_{\cdot 3}]^2) + \mathcal{B}^c_{1\to 2,3[2]}(\mathrm{Cov}(Y_{\cdot 3}, Y_{\cdot 2}) + \mathbb{E}[Y_{\cdot 2}]\mathbb{E}[Y_{\cdot 3}]) - \mathbb{E}[Y_{\cdot 1}]\mathbb{E}[Y_{\cdot 3}] + o(\sigma^2$$

where $\mathcal{B}^c_{1\to 2,3[0]}$, $\mathcal{B}^c_{1\to 2,3[2]}$ *and* $\mathcal{B}^c_{1\to 2,3[3]}$ *stand for the coefficients of* $Y_{\cdot 1}$ *on* $Y_{\cdot 1}$ *and* $Y_{\cdot 3}$ *when* $\Omega_{\cdot 1} = 1$*, associated with* $\mathcal{B}_{1\to 2,3[0]}$, $\mathcal{B}_{1\to 2,3[2]}$ *and* $\mathcal{B}_{1\to 2,3[3]}$ *depending on* B*,*

$$\begin{aligned}
\mathcal{B}_{1\to 2,3[0]} &:= -(B_{11}^{-23}B_{11} + B_{12}^{-23}B_{21})\mathbf{1}\alpha_2 - (B_{21}^{-23}B_{11} + B_{22}^{-23}B_{21})\mathbf{1}\alpha_3 + \mathbf{1}\alpha_1,\\
\mathcal{B}_{1\to 2,3[2]} &:= B_{11}^{-23}B_{11} + B_{12}^{-23}B_{21},\\
\mathcal{B}_{1\to 2,3[3]} &:= B_{21}^{-23}B_{11} + B_{22}^{-23}B_{21}.
\end{aligned}$$

In the same way as in the MNAR case detailed in Section 2, the formulae lead to natural estimates for the mean, the variance and the covariances of the missing at random variable.

II Results of Mohan et al. [15] for graphical approach in Section 2

The results and the proofs of Mohan et al. [15] are presented here for $p = 3$ and $r = 2$. Recall the preliminaries results.

Lemma 25 (Lemma 2 [15]). *Let us consider the m-graph G. The coefficient of the linear regression of $Y_{.j}$ on $Y_{.k}, k \neq j$, denoted as $\beta_{j\to k,k\neq j}$ is recoverable if $Y_{.j} \perp\!\!\!\perp \Omega | Y_{.k}, k \neq j$ and one has*

$$\beta_{j\to k,k\neq j} = \beta^c_{j\to k,k\neq j}.$$

Lemma 26 (Lemma 1). *[15]](Graphical approach for computing the covariance) Let G be a m-graph with k unblocked paths $p_1, \ldots, p_k$ between two variables $Y_{.\tau}$ and $Y_{.\delta}$. Let A_{p_i} be the ancestor of all notes on path p_i. Let the number of notes 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,

$$\beta_{Y\to X} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}, \tag{36}$$

where Y and X are two variables of a linear model.
A formula for the mean of the missing variable $Y_{.1}$ is derived as follows.

Proposition 27 (Expectation formula resulting from the graphical approach when $p = 3$ and $r = 2$). *The probabilistic model (3) is considered. Assuming **A2.** and $\beta^c_{2\to 1.3} \neq 0$, one has*

$$\alpha_1 = \frac{\alpha_2 - \beta^c_{2\to 1,3[0]} - \beta^c_{2\to 1,3[3]}\alpha_3}{\beta^c_{2\to 1,3[1]}}. \tag{37}$$

Proof. Indeed, one has:

$$\begin{aligned}
\mathbb{E}[Y_{.2}] &= \mathbb{E}[\mathbb{E}[Y_{.2}|Y_{.1}, Y_{.3}] \\
&= \mathbb{E}[\mathbb{E}[Y_{.2}|Y_{.1}, Y_{.3}, \Omega_{.1} = 1]] && \text{(by using A2.)} \\
&= \mathbb{E}[\mathbb{E}[\beta^c_{2\to 1,3[0]} + \beta^c_{2\to 1,3[1]} Y_{.1} + \beta^c_{2\to 3,1[3]} Y_{.3} + \epsilon_{Y_{.2}} | Y_{.1}, Y_{.3}]] \\
&= \beta^c_{2\to 1,3[0]} + \beta^c_{2\to 1,3[1]} \mathbb{E}[Y_{.1}] + \beta^c_{2\to 3,1[3]} \mathbb{E}[Y_{.3}],
\end{aligned}$$

which leads to the desired Equation (37), provided that $\beta^c_{2\to 1,3[1]} \neq 0$. □

The following proposition gives formulae for the variance and the covariances of the missing variable $Y_{.1}$. It is a slightly modification of the one in Mohan et al. [15], since one presents here a result whereas in Mohan et al. [15] only a method is given.

Proposition 28 (Variance and covariances formulae resulting from the graphical approach when $p = 3$ and $r = 2$). *Under the two equations (10) and (24), suppose that **A2.** and **A7.** hold. Assuming also that $\beta^c_{3\to 1} \neq 0$, $\beta^c_{2\to 1,3[1]} \neq 0$ and $\mathrm{Var}(Y_{.3}) \neq 0$, one can derive that*

$$\mathrm{Var}(Y_{.1}) = \frac{\mathrm{Var}(Y_{.3})}{\beta^c_{3\to 1}} \frac{1}{\beta^c_{2\to 1,3[1]}} \left(\frac{\mathrm{Cov}(Y_{.2}, Y_{.3})}{\mathrm{Var}(Y_{.3})} - \beta^c_{2\to 1,3[3]} \right), \tag{38}$$

with $\beta^c_{3\to 1}$ *the coefficient standing for the effects of* $Y_{\cdot 3}$ *on* $Y_{\cdot 1}$ *in the complete case and* $\beta^c_{2\to 1,3[1]}$ *and* $\beta^c_{2\to 1,3[3]}$ *introduced in Section 2.1.2. In addition, assuming* $\beta^c_{3\to 1,2[1]} \neq 0$ *and* $\mathrm{Var}(Y_{\cdot 2}) \neq 0$*, one has*

$$\mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot 2}) = \frac{1}{\beta^c_{3\to 1,2[1]}} \left(\frac{\mathrm{Cov}(Y_{\cdot 2}, Y_{\cdot 3})}{\mathrm{Var}(Y_{\cdot 2})} - \beta^c_{3\to 1,2[2]} \right) \mathrm{Var}(Y_{\cdot 2}), \tag{39}$$

$$\mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot 3}) = \frac{1}{\beta^c_{2\to 1,3[1]}} \left(\frac{\mathrm{Cov}(Y_{\cdot 2}, Y_{\cdot 3})}{\mathrm{Var}(Y_{\cdot 3})} - \beta^c_{2\to 1,3[3]} \right) \mathrm{Var}(Y_{\cdot 3}). \tag{40}$$

Proof. Using Equation (36),

$$\mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot 3}) = \mathrm{Var}(Y_{\cdot 1})\beta_{3\to 1},$$
$$\mathrm{Cov}(Y_{\cdot 3}, Y_{\cdot 1}) = \mathrm{Var}(Y_{\cdot 3})\beta_{1\to 3},$$

so

$$\mathrm{Var}(Y_{\cdot 1}) = \frac{\mathrm{Var}(Y_{\cdot 3})\beta_{1\to 3}}{\beta_{3\to 1}}.$$

Considering the graphical model in Figure 1(c),

$$\begin{aligned} \mathrm{Cov}(Y_{\cdot 2}, Y_{\cdot 3}) &= \beta_{2\to 1,3[1]}\beta_{1\to 3}\mathrm{Var}(Y_{\cdot 3}) + \beta_{2\to 1,3[3]}\mathrm{Var}(Y_{\cdot 3}) && \text{(by Lemma 26)} \\ \Rightarrow \beta_{1\to 3} &= \frac{1}{\beta_{2\to 1,3[1]}} \left(\frac{\mathrm{Cov}(Y_{\cdot 2}, Y_{\cdot 3})}{\mathrm{Var}(Y_{\cdot 3})} - \beta_{2\to 1,3[3]} \right) \\ \Rightarrow \beta_{1\to 3} &= \frac{1}{\beta^c_{2\to 1,3[1]}} \left(\frac{\mathrm{Cov}(Y_{\cdot 2}, Y_{\cdot 3})}{\mathrm{Var}(Y_{\cdot 3})} - \beta^c_{2\to 1,3[3]} \right) \end{aligned} \tag{41}$$

where the last implication is given by Lemma 25 and Assumption **A2.**, giving also

$$\beta_{3\to 1} = \beta^c_{3\to 1},$$

which concludes on Equation (38).
By (36), the covariances can be expressed in two different ways,

$$\mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot 2}) = \beta_{2\to 1}\mathrm{Var}(Y_{\cdot 1}) \quad \text{and} \quad \mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot 3}) = \beta_{3\to 1}\mathrm{Var}(Y_{\cdot 1}), \tag{42}$$
$$\mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot 2}) = \beta_{1\to 2}\mathrm{Var}(Y_{\cdot 2}) \quad \text{and} \quad \mathrm{Cov}(Y_{\cdot 1}, Y_{\cdot 3}) = \beta_{1\to 3}\mathrm{Var}(Y_{\cdot 3}). \tag{43}$$

In (42), the coefficients $\beta_{2\to 1}$ and $\beta_{3\to 1}$ can be estimated on the complete case using Lemma 25, but the variance of $Y_{\cdot 1}$ has still to be taken care of. Instead of potentially propagate error from (38), we propose to favor the expressions given in (43) to evaluate the covariances.
Focusing on (43), the coefficient $\beta_{1\to 3}$ is given in (41) and $\beta_{1\to 2}$ can be obtained using the same method, based on the reduced graphical model in Figure 1(d) (by Assumption **A7.**), so

$$\beta_{1\to 2} = \frac{1}{\beta^c_{3\to 1,2[1]}} \left(\frac{\mathrm{Cov}(Y_{\cdot 2}, Y_{\cdot 3})}{\mathrm{Var}(Y_{\cdot 2})} - \beta^c_{3\to 1,2[2]} \right).$$

Therefore, by plugging it in (43), Equations (39) and (40) are obtained. □

III Detailed results for Section 3

Consider any data matrix with p covariates and generated with a PPCA model with r latent variables $Y_{.j_1}, \ldots, Y_{.j_r}$ containing one variable missing not at random, denoted as $Y_{.m}$.
In the sequel, let us denote $\mathcal{J} := \{j_1, \ldots, j_r\}$ and $\mathcal{J}_{-k} := \{j_1, \ldots, j_r\} \backslash \{k\}$.

About the mean. Result on the consistency of a constructed mean estimator is first derived, by exploiting the linear links between variables, given in the following lemma.

Lemma 29. *Consider the model (1) and assume that* $\begin{pmatrix} B_{.m} & B_{.j_2} & B_{.j_2} & \ldots & B_{.j_r} \end{pmatrix}$ *has an inverse matrix denoted as* $B^{-1} \in \mathbb{R}^{r \times r}$. *One has*

$$Y_{.j_1} = \mathcal{B}_{\mathcal{B}_{j_1 \to m, \mathcal{J}_{-j_1}}[0]} + \sum_{k \in \mathcal{J}_{-j_1}} \mathcal{B}_{j_1 \to m, \mathcal{J}_{-j_1}}[k] Y_{.k} + \mathcal{B}_{j_1 \to m, \mathcal{J}_{-j_1}}[m] Y_{.m} - \sum_{k \in \mathcal{J}_{-j_1}} \mathcal{B}_{j_1 \to m, \mathcal{J}_{-j_1}}[k] \epsilon_{.k} - \mathcal{B}_{j_1 \to m, \mathcal{J}_{-j_1}}[m] \epsilon_{.m} + \epsilon_{.j_1}, \tag{44}$$

with:

$$\mathcal{B}_{j_1 \to m, \mathcal{J}_{-j_1}}[k] := B^{-1}_{mk} B_{j_1 m} + \cdots + B^{-1}_{j_r k} B_{j_1 j_r}, \tag{45}$$

$$\mathcal{B}_{j_1 \to m, \mathcal{J}_{-j_1}}[m] := B^{-1}_{mm} B_{j_1 m} + \cdots + B^{-1}_{j_r m} B_{j_1 j_r}, \tag{46}$$

$$\mathcal{B}_{j_1 \to m, \mathcal{J}_{-j_1}}[0] := \mathbf{1}\alpha_{r+1} - \sum_{k \in \mathcal{J}_{-j_1}} \mathcal{B}_{j_1 \to m, \mathcal{J}_{-j_1}}[k] \mathbf{1}\alpha_k - \mathcal{B}_{j_1 \to m, \mathcal{J}_{-j_1}}[m] \mathbf{1}\alpha_m. \tag{47}$$

Proof of Lemma 29. Without loss of generality, let us consider that $Y_{.1}$ has missing values ($m = 1$). Arbitrarily choosing $j_r = r + 1$ and $j_1 = 2, j_2 = 3, j_3 = 4, \ldots, j_{r-1} = r$, let us prove that $Y_{.r+1}$ is a linear combination of $Y_{.1}, Y_{.2}, \ldots, Y_{.r}$. Starting from (1) 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

$$\begin{pmatrix} Y_{.1} & \ldots & Y_{.r} \end{pmatrix} = \mathbf{1}\alpha_{|r} + \begin{pmatrix} W_{.1} & \ldots & W_{.r} \end{pmatrix} B_{|r} + \epsilon_{|r},$$

where $B_{|r} \in \mathbb{R}^{r \times r}$ denotes the reduced matrix of B in (1) keeping the first r variables of B. Similarly, $\alpha_{|r} \in \mathbb{R}^r$ and $\epsilon_{|r} \in \mathbb{R}^{r \times r}$ denote the reduced matrices of α and ϵ. B^{-r} denotes the inverse $B^{-1}_{|r}$ of $B_{|r}$, which exists since $B_{|r}$ has a full rank by assumption.
Then, one can derive that

$$\begin{pmatrix} W_{.1} & \ldots & W_{.r} \end{pmatrix} = \left(\begin{pmatrix} Y_{.1} & \ldots & Y_{.r} \end{pmatrix} - \mathbf{1}\alpha_{|r} - \epsilon_{|r} \right) B^{-r}.$$

The expression of $Y_{.r+1}$ as a function of the latent variables is

$$Y_{.r+1} = \mathbf{1}\alpha_{r+1} + B_{r+1.} \begin{pmatrix} W_{.1} & \ldots & W_{.r} \end{pmatrix} + \epsilon_{.r+1} = \mathbf{1}\alpha_{r+1} + B_{r+1.} \left(\begin{pmatrix} Y_{.1} \ldots & Y_{.r} \end{pmatrix} - \mathbf{1}\alpha_{|r} - \epsilon_{|r} \right) B^{-r} + \epsilon_{.r+1},$$

so that

$$Y_{.r+1} = \sum_{j=1}^{r} (B^{-r}_{1j} B_{(r+1)1} + \cdots + B^{-r}_{rj} B_{(r+1)r}) Y_{.j} - \sum_{j=1}^{r} (B^{-r}_{1j} B_{(r+1)1} + \cdots + B^{-r}_{rj} B_{(r+1)r})(\mathbf{1}\alpha_j + \epsilon_{.j}) + \epsilon_{.r+1} + \mathbf{1}\alpha_{r+1}.$$

Using the notations introduced in (45), (46) and (47), one has

$$\begin{aligned}
\forall j \in \{2,\dots,r\}, \mathcal{B}_{(r+1)\to 1,\mathcal{J}_{-(r+1)}[j]} &= B_{1j}^{-r}B_{(r+1)1} + \cdots + B_{rj}^{-r}B_{(r+1)r},\\
\mathcal{B}_{(r+1)\to 1,\mathcal{J}_{-(r+1)}[1]} &= B_{11}^{-r}B_{(r+1)1} + \cdots + B_{r1}^{-r}B_{(r+1)r},\\
\mathcal{B}_{(r+1)\to 1,\mathcal{J}_{-(r+1)}[0]} &= \mathbf{1}\alpha_{r+1} - \sum_{j=2}^{r}\mathcal{B}_{r+1)\to 1,\mathcal{J}_{-(r+1)}[j]}\mathbf{1}\alpha_j - \mathcal{B}_{(r+1)\to 1,\mathcal{J}_{-(r+1)}[1]}\mathbf{1}\alpha_1.
\end{aligned}$$

One obtain then the desired solution

$$\begin{aligned}
Y_{\cdot r+1} = \mathcal{B}_{(r+1)\to 1,\mathcal{J}_{-(r+1)}[0]} + \sum_{j=2}^{r}\mathcal{B}_{(r+1)\to 1,\mathcal{J}_{-(r+1)}[j]}Y_{\cdot j} + \mathcal{B}_{(r+1)\to 1,\mathcal{J}_{-(r+1)}[1]}Y_{\cdot 1}\\
- \sum_{j=2}^{r}\mathcal{B}_{(r+1)\to 1,\mathcal{J}_{-(r+1)}[j]}\epsilon_{\cdot j} - \mathcal{B}_{(r+1)\to 1,\mathcal{J}_{-(r+1)}[1]}\epsilon_{\cdot 1} + \epsilon_{\cdot r+1}.
\end{aligned}$$

□

An expression for the mean of the missing variable $Y_{\cdot m}$ is given in the following proposition.

Proposition 30 (Mean formula)**.** *Under the PPCA model (1), assume that it exists r variables $Y_{\cdot j_1},\dots,Y_{\cdot j_r}$ such that:*

A11. $\begin{pmatrix} B_{\cdot m} & B_{\cdot j_1} & B_{\cdot j_2} & \dots & B_{\cdot j_{r-1}}\end{pmatrix}$ *is an invertible matrix,*

A12. $Y_{\cdot j_1} \perp\!\!\!\perp \Omega_{\cdot m} | Y_{\cdot m}, Y_{\cdot j_1},\dots,Y_{\cdot j_r}$.

Assuming also that $\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[m]}$ is non-zero, one can derive that

$$\alpha_m = \frac{\alpha_{j_1} - \mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[0]} - \sum_{k\in\mathcal{J}_{-j_1}}\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[k]}\alpha_k}{\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[m]}}, \tag{48}$$

where for $k \in \mathcal{J}_{-j_1}$, $\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[0]}$, $\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[m]}$ and $\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[k]}$ are the coefficients standing for the effects of the regression of $Y_{\cdot j_1}$ on $(Y_{\cdot m},(Y_{\cdot k})_{k\in\mathcal{J}_{-j_1}})$ in the complete case, when $\Omega_{\cdot m}=1$.

The expression for the mean of the missing variable given by (48) leads to a natural estimator of the mean of $Y_{\cdot m}$ given in Definition 19.

Proof of Proposition 30. Without loss of generality, let us consider that $Y_{\cdot 1}$ has missing values ($m=1$). Arbitrarily choosing $j_1 = r+1$ and that $j_2 = 2, j_3 = 3, j_4 = 4,\dots,j_r = r$, let us prove that $Y_{\cdot r+1}$ is a linear combination of $Y_{\cdot 1}, Y_{\cdot 2},\dots,Y_{\cdot r}$.
Given that $\mathbb{E}[Y_{\cdot r+1}] = \mathbb{E}[\mathbb{E}[Y_{\cdot r+1}|Y_{\cdot 1},\dots,Y_{\cdot r}]]$, Assumption **A12.** leads to

$$\mathbb{E}[Y_{\cdot 1}|Y_{\cdot 1},\dots,Y_{\cdot r}] = \mathbb{E}[Y_{\cdot 1}|Y_{\cdot 1},\dots,Y_{\cdot r},\Omega_{\cdot 1}=1].$$

Then, by definition of $(\mathcal{B}^c_{i\to j,k})$'s,

$$\begin{aligned}
&\mathbb{E}[Y_{\cdot r+1}|Y_{\cdot 1},\dots,Y_{\cdot r},\Omega_{\cdot 1}=1]\\
&=\mathbb{E}\left[\mathcal{B}^c_{r+1\to 1,\mathcal{J}_{-(r+1)}[0]}+\sum_{k=1}^{r}\mathcal{B}^c_{r+1\to 1,\mathcal{J}_{-(r+1)}[k]}(Y_{\cdot k}-\epsilon_{\cdot k})+\epsilon_{\cdot r+1}\middle|Y_{\cdot 1},\dots,Y_{\cdot r}\right]\\
&=\mathcal{B}^c_{r+1\to 1,\mathcal{J}_{-(r+1)}[0]}+\sum_{k=1}^{r}\mathcal{B}^c_{r+1\to 1,\mathcal{J}_{-(r+1)}[k]}Y_{\cdot k}-\sum_{k=1}^{r}\mathcal{B}^c_{r+1\to 1,\mathcal{J}_{-(r+1)}[k]}\mathbb{E}[\epsilon_{\cdot k}|Y_{\cdot 1},\dots,Y_{\cdot r}]\middle|Y_{\cdot 1},\dots,Y_{\cdot r}].
\end{aligned}$$

Thus, by taking the mean and given that $\mathbb{E}[\epsilon_{\cdot i}]=0$ for $i=1,\dots,r$, one has

$$\mathbb{E}[Y_{\cdot r+1}]=\mathcal{B}^c_{r+1\to 1,\mathcal{J}_{-(r+1)}[0]}+\sum_{k=2}^{r}\mathcal{B}^c_{r+1\to 1,\mathcal{J}_{-(r+1)}[k]}\mathbb{E}[Y_{\cdot k}]+\mathcal{B}^c_{r+1\to 1,\mathcal{J}_{-(r+1)}[1]}\mathbb{E}[Y_{\cdot 1}],$$

implying Equation (48), provided that $\mathcal{B}^c_{r+1\to 1,\mathcal{J}_{-(r+1)}[1]}\neq 0$. □

About the variance and covariances. One construct now estimators of the variance and covariances, by exploiting all links between the variables $Y_{j_1},\dots,Y_{j_r}$ and $Y_{\cdot m}$ i.e. to write the r linear equations expressed $Y_{\cdot j}, j\in\mathcal{J}:=\{j_1,\dots,j_r\}$ according to the others variables. For $j\in\mathcal{J}$, the coefficients $Y_{\cdot j}$ are expressed according to $Y_{\cdot m}$ and $(Y_{\cdot l})_{l\in\mathcal{J}_{-j}}$, which implies thus r linear equations. This leads to the following proposition which gives the variance and covariances formulae.

Proposition 31 (Variance and covariances formulae)**.** *Under the PPCA model (1), assume that it exists r variables $Y_{\cdot j_1},\dots,Y_{\cdot j_r}$ such that Assumptions **A11.**, **A12.** are verified, as well as the following ones:*

A13. $\forall j\in\mathcal{J}$, $Y_{\cdot j}\perp\!\!\!\perp\Omega|Y_{\cdot m},Y_{\cdot k},k\in\mathcal{J}_{-j}$,

A14. $\forall j\in\mathcal{J}$, $\begin{pmatrix}B_{\cdot m} & (B_{\cdot l})_{l\in\mathcal{J}_{-j}}\end{pmatrix}$ *has an inverse matrix.*

One can derive that

$$M_1^\star X^\star+o(\sigma^2)=M_2^\star,\ \textit{with:}$$

$$M_1^\star=\begin{pmatrix}
(\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[m]})^2 & 0 & 2\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[j_1]}\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[j_2]} & \cdots & 2\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[j_1]}\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[j_r]}\\
\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[m]} & 1 & -\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[j_2]} & \cdots & -\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[j_r]}\\
 & & \ddots & & \\
 & & & \ddots & \\
-\mathcal{B}^c_{j_r\to m,\mathcal{J}_{-j_r}[m]} & -\mathcal{B}^c_{j_r\to m,\mathcal{J}_{-j_r}[j_1]} & -\mathcal{B}^c_{j_r\to m,\mathcal{J}_{-j_r}[j_2]} & \cdots & 1
\end{pmatrix},$$

$$X^\star=\begin{pmatrix}\mathrm{Var}(Y_{\cdot m})\\ \mathrm{Cov}(Y_{\cdot m},Y_{\cdot j_1})\\ \mathrm{Cov}(Y_{\cdot m},Y_{\cdot j_2})\\ \vdots\\ \mathrm{Cov}(Y_{\cdot m},Y_{\cdot j_r})\end{pmatrix},$$

$$M_2^\star = \begin{pmatrix} \mathrm{Var}(Y_{.j_1}) - Q^{\star c} - (\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[\mathcal{J}_{-j_1}]})^T \mathrm{Var}(Y_{\mathcal{J}_{-j_1}})\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}[\mathcal{J}_{-j_1}]} \\ (\mathcal{B}^c_{j_1\to m,\mathcal{J}_{-j_1}})^T \begin{pmatrix} 1 & \mathbb{E}[Y_{.m}] & \mathbb{E}[Y_{.j_1}] & \dots & \mathbb{E}[Y_{.j_r}] \end{pmatrix}^T - \mathbb{E}[Y_{.j_1}]\mathbb{E}[Y_{.m}] \\ \vdots \\ (\mathcal{B}^c_{j_r\to m,\mathcal{J}_{-j_r}})^T \begin{pmatrix} 1 & \mathbb{E}[Y_{.m}] & \mathbb{E}[Y_{.j_1}] & \dots & \mathbb{E}[Y_{.j_r}] \end{pmatrix}^T - \mathbb{E}[Y_{.j_r}]\mathbb{E}[Y_{.m}] \end{pmatrix},$$

where the coefficients for $k \in \mathcal{J}_{-j}$, $\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[0]}$, $\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[m]}$ *and* $\mathcal{B}^c_{j\to m,\mathcal{J}_{-j}[k]}$ *are the coefficients standing for the effects of* $Y_{.j}$ *on* $(Y_{.m},(Y_{.l})_{l\in\mathcal{J}_{-j}})$ *when* $\Omega_{.m}=1$ *and*

$$Q^{\star c} = \left(\mathrm{Var}(Y_{.j_1}) - \mathrm{Cov}(Z^\star, Y_{.j_1})\mathrm{Var}(Z^\star)\mathrm{Cov}(Z^\star, Y_{.j_1})^T \middle| \Omega_{.m}=1\right),$$

with $Z^\star = [Y_{.m}, Y_{.j_2}, \dots, Y_{.j_r}]$.

As precised in Definition 20, a natural estimators for the variance and covariances are then

$$\hat{X}^\star = (\hat{M}_1^\star)^{-1}\hat{M}_2^\star,$$

provided that $(\hat{M}_1^\star)^{-1}$ exists.

About covariance between two missing variables. To calculate the covariances between two missing variables, one uses the following proposition.

Proposition 32 (Covariance formula between two missing variables)**.** *Under the PPCA model given in* (1)*, assume that it exists* $r-1$ *variables* $Y_{.j_1},\dots,Y_{.j_{r-1}}$ *such that* ***A11.*** *and* ***A12.*** *are verified. Let us den ote* $\mathcal{H} = \mathcal{J}\cup\{m_1,m_2\}$*. A formula of the covariance between two MNAR missing variables is*

$$K\mathrm{Cov}(Y_{.m_1},Y_{.m_2}) = \mathrm{Var}(Y_{.j_1}) - Q^{\star,c} - \sum_{k\in\mathcal{H}_{-j_1}} (\mathcal{B}^c_{j_1\to\mathcal{H}_{-j_1}[k]})^2\mathrm{Var}(Y_{.k}) - \sum_{k\in\mathcal{H}_{-j_1},l\in\mathcal{H}_{-(j_1,m_1,m_2)}k\neq l} 2\mathcal{B}^c_{j_1\to\mathcal{H}_{-j_1}[k]}\mathcal{B}^c_{j_1\to\mathcal{H}_{-j_1}[l]}\mathrm{Cov}(Y_{.k},Y_{.l}), \quad (49)$$

with $K = 2\mathcal{B}^c_{j_1\to\mathcal{H}_{-j_1}[m_1]}\mathcal{B}^c_{j_1\to\mathcal{H}_{-j_1}[m_2]}$ *and*

$$Q^{\star c} = \left(\mathrm{Var}(Y_{.j_1}) - \mathrm{Cov}(Z^\star, Y_{.j_1})\mathrm{Var}(Z^\star)\mathrm{Cov}(Z^\star, Y_{.j_1})^T \middle| \Omega_{.m_1}=1,\Omega_{.m_2}=1\right),$$

where $Z^\star = [Y_{.m_1}, Y_{.m_2}, Y_{.j_2}, \dots, Y_{.j_{r-1}}]$.

If $K\neq 0$*, one derive the expression of* $Cov(Y_{.m_1},Y_{.m_2})$.

Equation (49) is at the origin of the covariance estimator between two missing variables proposed in Definition 22.

IV Complementary figures

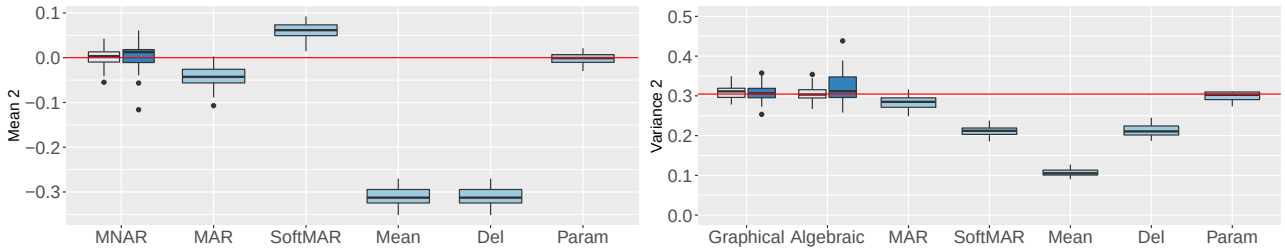


Figure 15: Mean and variance estimations of the missing variable $Y_{.2}$ when $r = 2$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

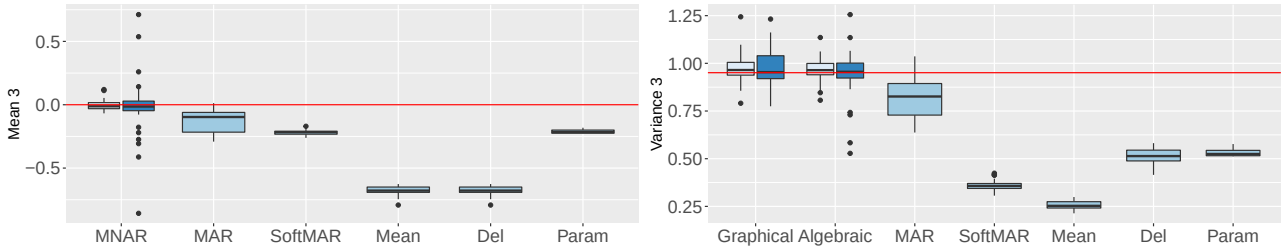


Figure 16: Mean and variance estimations of the missing variable $Y_{.3}$ when $r = 2$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

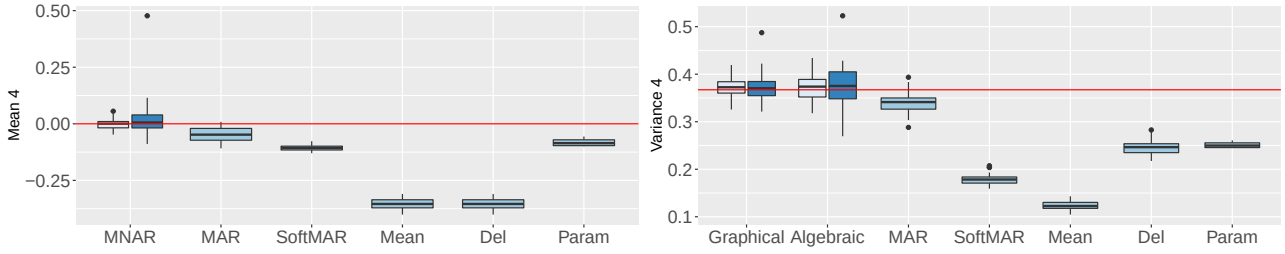


Figure 17: Mean and variance estimations of the missing variable $Y_{.4}$ when $r = 2$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

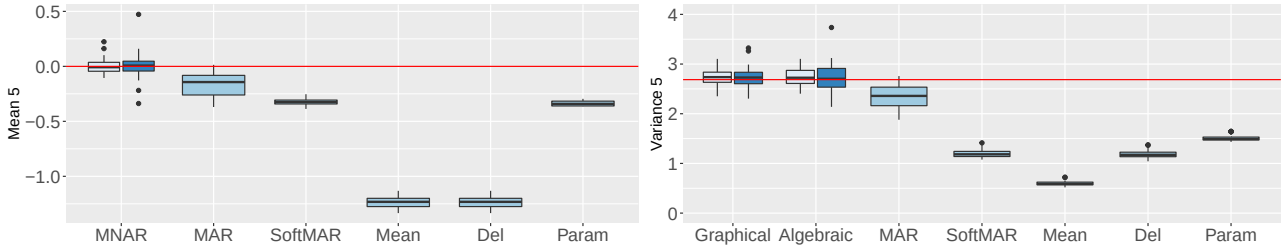


Figure 18: Mean and variance estimations of the missing variable $Y_{.5}$ when $r = 2$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

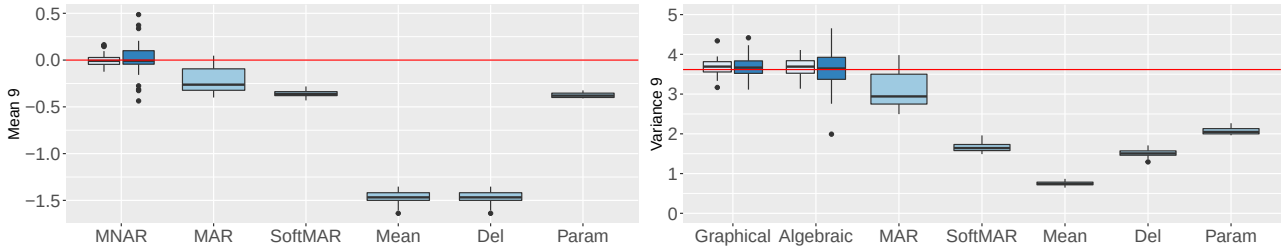


Figure 19: Mean and variance estimations of the missing variable $Y_{.9}$ when $r = 2$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

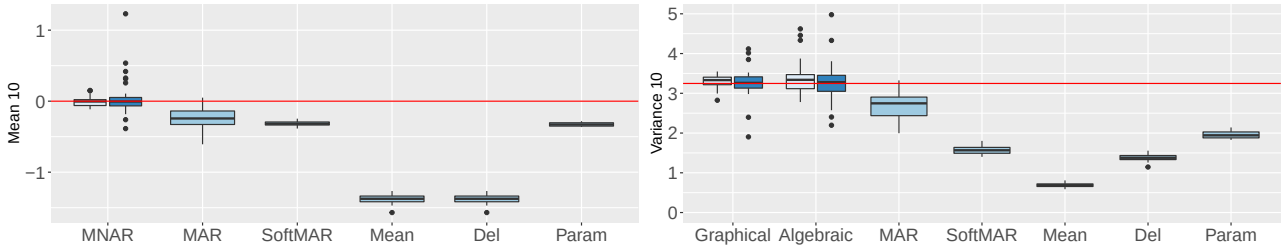


Figure 20: Mean and variance estimations of the missing variable $Y_{.10}$ when $r = 2$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

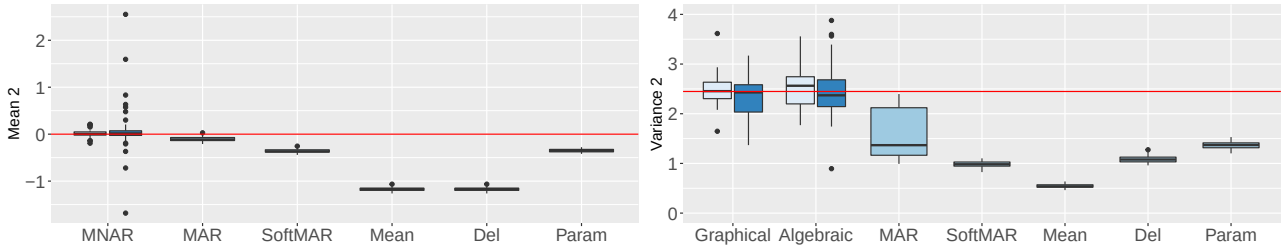


Figure 21: Mean and variance estimations of the missing variable $Y_{.2}$ when $r = 3$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

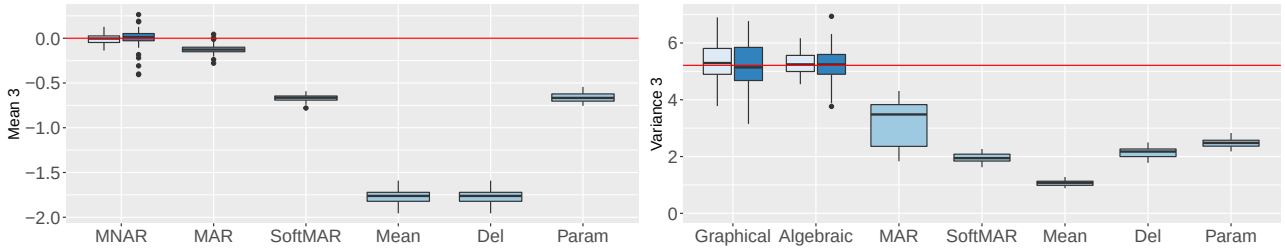


Figure 22: Mean and variance estimations of the missing variable $Y_{.3}$ when $r = 3$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

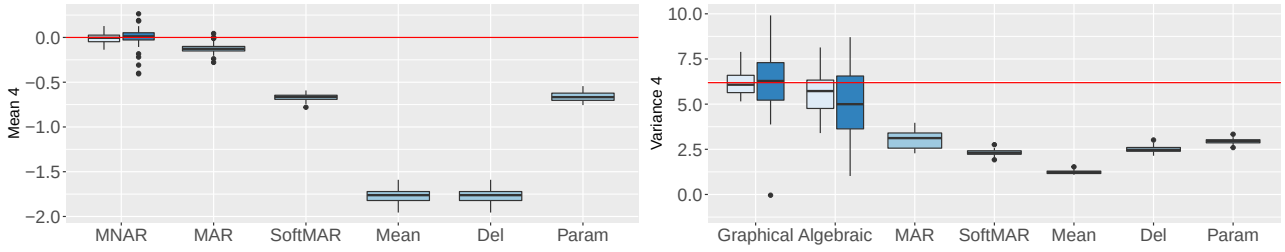


Figure 23: Mean and variance estimations of the missing variable $Y_{.4}$ when $r = 3$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

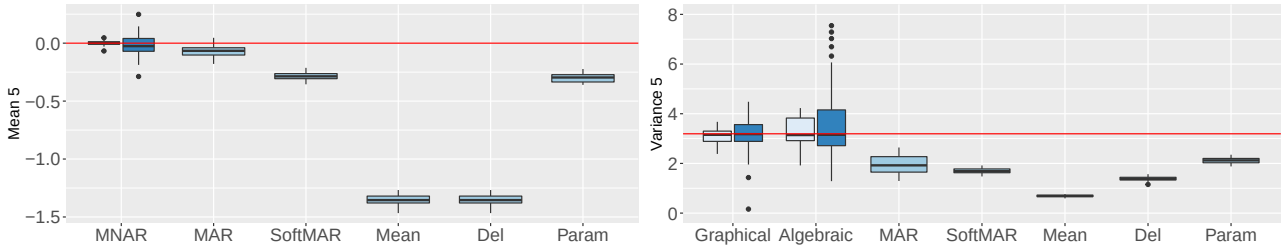


Figure 24: Mean and variance estimations of the missing variable $Y_{.5}$ when $r = 3$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

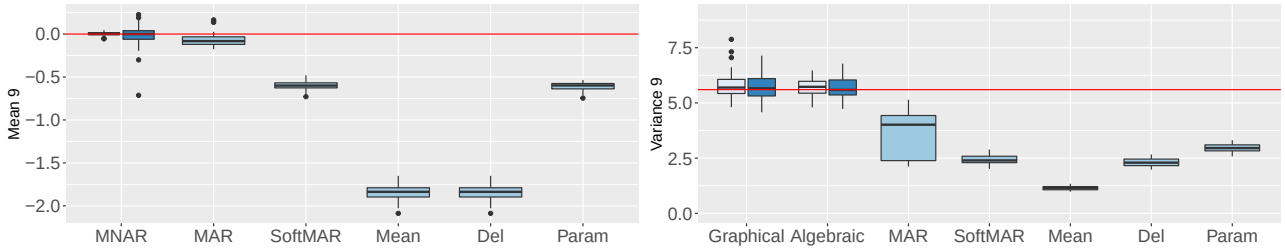


Figure 25: Mean and variance estimations of the missing variable $Y_{.9}$ when $r = 3$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

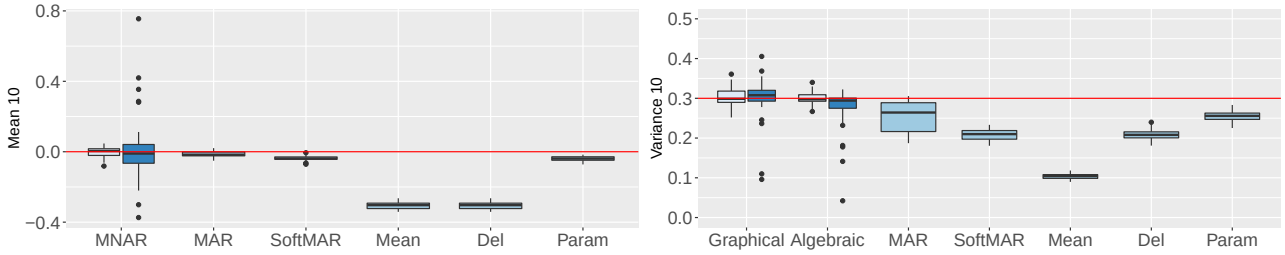


Figure 26: Mean and variance estimations of the missing variable $Y_{.10}$ when $r = 3$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

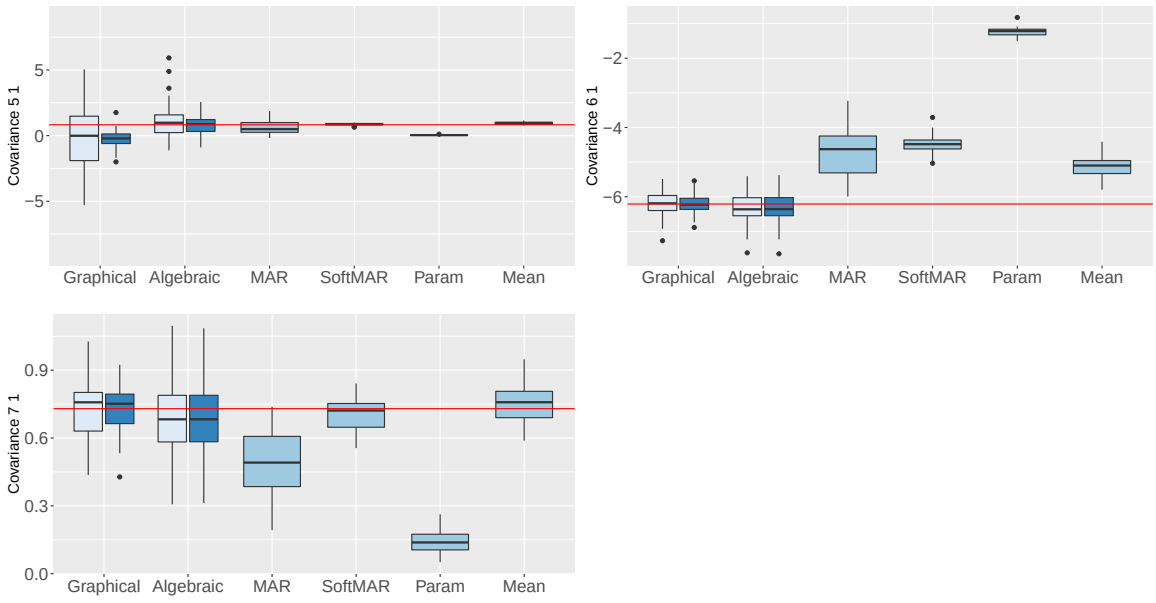


Figure 27: Covariances estimations of $\mathrm{Cov}(Y_{.1}, Y_{.j}), j \in \{5, 6, 7\}$ when $r = 2$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.

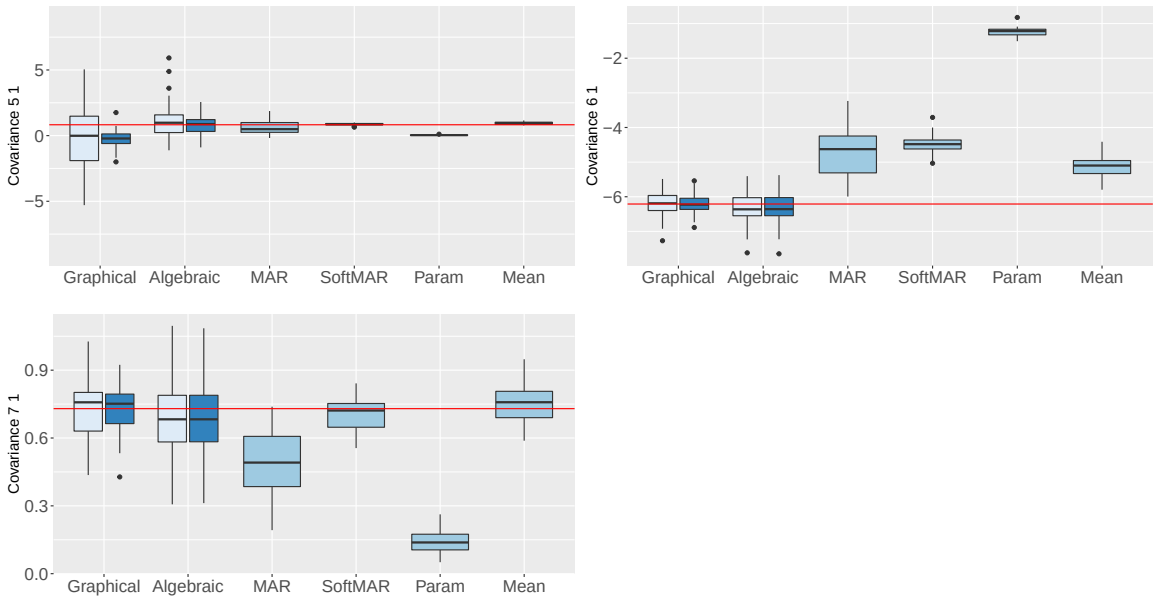


Figure 28: Covariances estimations of $\mathrm{Cov}(Y_{.1}, Y_{.j}), j \in \{8, 9, 10\}$ when $r = 2$, $n = 1000$, $p = 10$, $\sigma = 0.1$ and 7 variables are missing leading to 35 % of MNAR values. Light blue boxplots stand for the aggregation approach which chooses the observed variables on which the regression will be formed by aggregating every possible combination, dark blue boxplots represent the random approach which randomly selects a combination. The red lines indicate the true values.