Estimation and imputation in Probabilitic Principal Component Analysis with Missing Not At Random Data

SIMPAS Group Meeting

Aude Sportisse [1,3] Claire Boyer [1,2] Julie Josse [3,4,5]

[1]Laboratoire de Probabilités Statistique et Modélisation, Sorbonne Université

[2]Département de Mathématiques et applications, Ecole Normale Supérieure

[3]Centre de Mathématiques Appliquées, Ecole Polytechnique

[4]XPOP, INRIA

[5]Visiting Researcher Google Brain

18th June 2020

Missing values are everywhere

- for many reasons: unanswered questions in a survey, lost data, damaged plants, machines that fail...

Traumabase: 15 000 patients/ 250 var/ 15 hospitals

Center	Age	Sex	Weight	Height	Heart rate	Lactates
Beaujon	54	m	85	NA	NA	NA
Lille	33	m	80	1.8	180	4.8
Pitie	26	m	NA	NA	NA	3.9
Beaujon	63	m	80	1.8	190	1.66
Pitie	30	w	NA	NA	NA	NA

NA: Not Available.

Classical definitions

- $Y \in \mathbb{R}^{n\times p}$ the data matrix,
- $\Omega \in \mathbb{R}^{n\times p}$ the missing-data pattern:

$$\Omega_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed,} \\ 0 & \text{otherwise,} \end{cases}$$

- **Issue: Cause of the missingness ?**
  [Rubin, 1976], [Little and Rubin, 2014]

Missing-data mechanism

ϕ: the unknown parameters of the missingness.

MCAR mechanism

The missingness does not depend on the variables.

$$p(\Omega | Y; \phi) = p(\Omega; \phi), \quad \forall Y, \phi.$$

MAR mechanism

The missingness depends on the observed variables.

$$p(\Omega | Y; \phi) = p(\Omega | Y_{\mathrm{obs}}; \phi), \quad \forall Y_{\mathrm{mis}}, \phi.$$

MNAR mechanism

Other case, i.e.

$$p(\Omega | Y; \phi) = p(\Omega | Y_{\mathrm{obs}}, Y_{\mathrm{mis}}; \phi), \quad \forall \phi.$$

Self-masked setting when the missingness of a variable depends on the variable itself:

$$p(\Omega_{.j} | Y; \phi) = p(\Omega_{.j} | Y_{.j}; \phi), \quad \forall \phi.$$

MNAR data are very frequent in practice...

Traumabase: 15 000 patients/ 250 var/ 15 hospitals

Center	Age	Sex	Weight	Height	Heart rate	Lactates
Beaujon	54	m	85	NA	NA	NA
Lille	33	m	80	1.8	180	4.8
Pitie	26	m	NA	NA	NA	3.9
Beaujon	63	m	80	1.8	190	1.66
Pitie	30	w	NA	NA	NA	NA

- MNAR case extremely frequent, such as the heart rate (HR).
  patient's condition critical
  → HR high or low
  → doctors would rather provide emergency care than measure HR.

...but hard to handle

- ✓ In the MAR setting, one can ignore the mechanism. Statistical inference is possible without modelling the missing-data mechanism distribution.
- ✗ In the MNAR setting, the observed variables are not representative of the population. One should consider the mechanism.

...but hard to handle

✓ In the MAR setting, one can ignore the mechanism. Statistical inference is possible without modelling the missing-data mechanism distribution.

✗ In the MNAR setting, the observed variables are not representative of the population. One should consider the mechanism.

- Estimation in the MNAR setting: we should take into account the mechanism explicitly (by modelling it) or implicitly.
- Identifiability in the MNAR setting: the law is identifiable only if the mechanism is identifiable.

Identifiability issue in the MNAR case

$Y^{\mathrm{NA}} = [1, \mathrm{NA}, 0, 1, \mathrm{NA}, 0]$.

- $\mathbb{P}(Y)$ is not identifiable without knowing $\mathbb{P}(\Omega|Y)$.
  - Y missing only if $Y = 1$. Thus, $Y = [1, 1, 0, 1, 1, 0]$ and $\mathbb{P}(Y) = 2/3$.
  - Y missing only if $Y = 0$. Thus, $Y = [1, 0, 0, 1, 0, 0]$ and $\mathbb{P}(Y) = 1/3$.
- 2 mechanisms yield to different data distribution.

Existing works for handling MNAR data (1)

How to estimate parameters of the data distribution in presence of MNAR data?

- By modeling the MNAR mechanism via a **Logistic Model** $\forall i \in [1, n]$, $\phi_j = (\phi_{1j}, \phi_{2j})$ denoting a parameter vector:

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

  and **using an EM algorithm** to estimate both parameters of the data and mechanism distributions.
  - in linear models [Ibrahim et al., 1999],
  - in low-rank models [S., Boyer, Josse 2018].

✓ Handling MNAR data.

✗ Often restricted to a limited number of MNAR variables.

✗ Parametric assumption for the mechanism distribution.

✗ Computationally costly.

Existing works for handling MNAR data (2)

How to estimate parameters of the data distribution in presence of MNAR data ?

- Without modeling the mechanism and by only using all available observed cells
  - for multivariate regression [Miao and Tchetgen, 2018, Tang et al., 2003],
  - in linear models, method based on graphical models [Mohan et al., 2018].

✓ Identifiability guarantees.

✓ No modelling for the mechanism.

✗ Restricted to simple models.

✗ The self-masked assumption may be strong.

Existing works for handling MNAR data (2)

How to estimate parameters of the data distribution in presence of MNAR data ?

- Without modeling the mechanism and by only using all available observed cells
  - for multivariate regression [Miao and Tchetgen, 2018, Tang et al., 2003],
  - in linear models, method based on graphical models [Mohan et al., 2018].

✓ Identifiability guarantees.

✓ No modelling for the mechanism.

✗ Restricted to simple models.

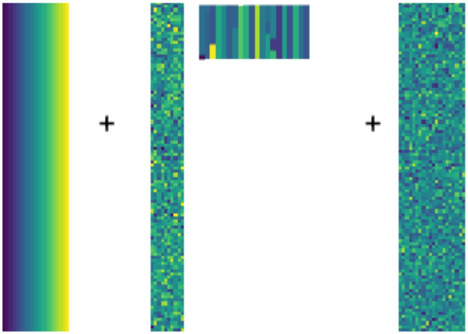
✗ The self-masked assumption may be strong.

$\Rightarrow$ **Proposal: Handling the MNAR data in probabilistic PCA model.**

Probabilistic Principal Component Analysis (PPCA) model

$Y \in \mathbb{R}^{n \times p}$ noisy realisation of the factorization of the coefficients matrix $B \in \mathbb{R}^{r \times p}$ and r latent variables grouped in the matrix $W \in \mathbb{R}^{n \times r}$:

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



Probabilistic Principal Component Analysis (PPCA) model

$Y \in \mathbb{R}^{n \times p}$ noisy realisation of the factorization of the coefficients matrix $B \in \mathbb{R}^{r \times p}$ and r latent variables grouped in the matrix $W \in \mathbb{R}^{n \times r}$:

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

where

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

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

Probabilistic Principal Component Analysis (PPCA) model

$Y \in \mathbb{R}^{n \times p}$ noisy realisation of the factorization of the coefficients matrix $B \in \mathbb{R}^{r \times p}$ and r latent variables grouped in the matrix $W \in \mathbb{R}^{n \times r}$:

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

⇢ Access only to the missing-data matrix $Y \odot \Omega + \mathrm{NA} \odot (1 - \Omega)$

$$Y = \begin{array}{c} \\ \\ \end{array}\begin{array}{cccccc} Y_{.\mathbf{1}} & Y_{.\mathbf{2}} & Y_{.\mathbf{3}} & Y_{.\mathbf{4}} & \dots & Y_{.p} \\ \end{array}$$

$$Y = \begin{pmatrix} 12 & 28 & 31 & 3 & \dots & \mathrm{NA} \\ \mathrm{NA} & 23 & 89 & 2 & \dots & 85 \\ 32 & 6 & 24 & \mathrm{NA} & \dots & \mathrm{NA} \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ \mathrm{NA} & 3 & 7 & \mathrm{NA} & \dots & 11 \end{pmatrix}$$

Proposal

Under the PPCA model $\forall i \in \{1, \ldots, n\}, \quad Y_{i.} \sim \mathcal{N}(\alpha, \Sigma), \Sigma = B^T B + \sigma^2 \mathrm{Id}_{p \times p}$,

- **Identifiability** of the PPCA parameters assuming self-masked MNAR,

Proposition: identifiability of the PPCA parameters

Consider that d variables are self-masked MNAR and $p - d$ variables are MCAR.

- The parameters (α, Σ) of the PPCA model and the mechanism parameter are identifiable.
- Assuming that the noise level σ^2 is known, the parameter B is identifiable up to a row permutation.

- Mean and covariance matrix estimation without explicitly modeling the MNAR mechanism and by only using all available observed cells **Consistency results** assuming general MNAR mechanism,
- Estimation of the loading matrix B,
- Imputation of the missing values in the data matrix Y.

Toy example

- $p = 3$, $r = 2$.
- $Y_{.1}$ is MNAR (self-masked in this case).
- As $r = 2$, it requires two **pivot variables**, say $Y_{.2}$ and $Y_{.3}$ **which are independent of the missing-data pattern** $\Omega_{.1}$.

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

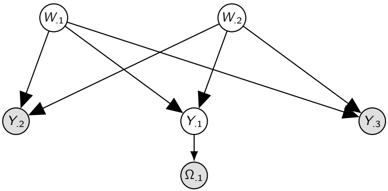
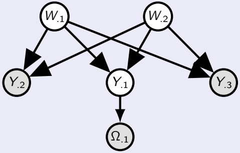


Figure: Graphical model for **"fully-connected"** PPCA model when $p = 3$, $r = 2$ and one variable $Y_{.1}$ is missing.

As any variable is generated by all the latent variables, linear links can be established.



Lemma exploiting the linear links between variables

$$\mathbf{Y}_{.2} = \mathcal{B}_{2\to 1,3[0]} + \mathcal{B}_{2\to 1,3[1]}\mathbf{Y}_{.1} + \mathcal{B}_{2\to 1,3[3]}\mathbf{Y}_{.3} + \zeta,$$

where

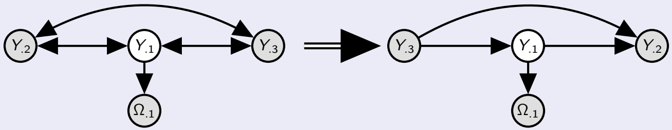
- $(\mathcal{B}_{2\to 1,3[k]})_{k\in\{0,1,3\}}$ denotes the coefficients standing for the effects of $Y_{.2}$ on $Y_{.1}$ and $Y_{.3}$.
- $(\mathcal{B}_{2\to 1,3[k]})_{k\in\{0,1,3\}}$ depends on the unknown matrix B and ζ the noise
- Note that $\mathbb{E}[\zeta | Y_{.1}, Y_{.3}] \neq 0$ (no exogeneity)

.

Mean estimation

Exploiting the linear links

As any variable is generated by all the latent variables, linear links can be established.



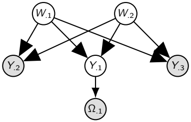
Lemma exploiting the linear links between variables

$$\mathbf{Y}_{.2} = \mathcal{B}_{2\to 1,3[0]} + \mathcal{B}_{2\to 1,3[1]}\mathbf{Y}_{.1} + \mathcal{B}_{2\to 1,3[3]}\mathbf{Y}_{.3} + \zeta,$$

where

- $(\mathcal{B}_{2\to 1,3[k]})_{k\in\{0,1,3\}}$ denotes the coefficients standing for the effects of $Y_{.2}$ on $Y_{.1}$ and $Y_{.3}$.
- $(\mathcal{B}_{2\to 1,3[k]})_{k\in\{0,1,3\}}$ depends on the unknown matrix B and ζ the noise
- Note that $\mathbb{E}[\zeta | Y_{.1}, Y_{.3}] \neq 0$ (no exogeneity)

.



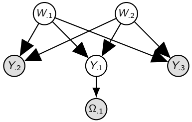
Mean estimation

Estimating in the complete-case analysis

Effects of $Y_{.2}$ on $Y_{.1}$ and $Y_{.3}$ in the complete case when $\Omega_{.1} = 1$:

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

$$\begin{array}{cc} & \begin{array}{ccc} Y_{.1} & Y_{.2} & Y_{.3} \end{array} \\ Y = & \begin{pmatrix} 12 & 28 & 31 \\ \cancel{\mathrm{NA}} & \cancel{23} & \cancel{89} \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \cancel{\mathrm{NA}} & \cancel{3} & \cancel{7} \end{pmatrix} \end{array}$$



Mean estimation

Estimating in the complete-case analysis

Effects of $Y_{.2}$ on $Y_{.1}$ and $Y_{.3}$ in the complete case when $\Omega_{.1} = 1$:

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

As $Y_{.2} \perp\!\!\!\perp \Omega_{.1} | Y_{.1}, Y_{.3}$, one has

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

Taking the expectation,

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

Mean formula

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

given that $\mathcal{B}^{c}_{2\to 1,3[1]} \neq 0$.

Consistency results

Definition of a mean estimator:

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

Consistency for the missing variable mean

Assume that:

- There exist consistent estimators for α_2 and α_3.
- There exist consistent estimators for $\mathcal{B}^c_{2\to 1,3[0]}$, $\mathcal{B}^c_{2\to 1,3[1]}$ and $\mathcal{B}^c_{2\to 1,3[3]}$.

Then, the estimator $\hat{\alpha}_1$ is consistent.

Estimation in practice

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

- $\hat{\alpha}_2$ and $\hat{\alpha}_3$ are computed as empirical quantities.
  - $\hat{\alpha}_2 = \bar{Y}_{.2}$
  - $\hat{\alpha}_3 = \bar{Y}_{.3}$

$$Y = \begin{array}{c} \begin{array}{ccc} Y_{.1} & Y_{.2} & Y_{.3} \end{array} \\ \begin{pmatrix} 12 & 28 & 31 \\ \mathrm{NA} & 23 & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \mathrm{NA} & 3 & 7 \end{pmatrix} \end{array}$$

Estimation in practice

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

- $\hat{\alpha}_2$ and $\hat{\alpha}_3$ are computed as empirical quantities.
  - $\hat{\alpha}_2 = \bar{Y}_{.2}$
  - $\hat{\alpha}_3 = \bar{Y}_{.3}$

$$\begin{array}{cc} & \begin{array}{ccc} Y_{.1} & Y_{.2} & Y_{.3} \end{array} \\ Y = & \begin{pmatrix} 12 & 28 & 31 \\ \mathrm{NA} & 23 & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \mathrm{NA} & 3 & 7 \end{pmatrix} \end{array}$$

- $\left(\mathcal{B}^c_{2\to 1,3[k]}\right)_{k\in\{0,1,3\}}$ estimated by the coefficients of the linear regression of $Y_{.2}$ on $Y_{.1}$ and $Y_{.3}$ using the rows where $Y_{.1}$ is observed.

$$\begin{array}{cc} & \begin{array}{ccc} Y_{.1} & Y_{.2} & Y_{.3} \end{array} \\ Y = & \begin{pmatrix} 12 & 28 & 31 \\ \cancel{\mathrm{NA}} & \cancel{23} & \cancel{89} \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \cancel{\mathrm{NA}} & \cancel{3} & \cancel{7} \end{pmatrix} \end{array}$$

Estimation of the loading matrix B

- Same methodology for the variance and covariances.
- Estimators obtained from the formulae:

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

.

Estimation of the loading matrix B

- Same methodology for the variance and covariances.
- Estimators obtained from the formulae:

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

.

- Assuming that σ^2 is known,

$$Y \sim \mathcal{N}\left(\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}, B^T B + \sigma^2 \mathrm{Id}\right) \Rightarrow \hat{\Sigma} - \sigma^2 \mathrm{Id}_{3\times 3} \text{ estimates } B^T B.$$

- Singular value decomposition:

$$\hat{\Sigma} - \sigma^2 \mathrm{Id}_{3\times 3} =: \hat{U}\hat{D}\hat{U}^T, \text{ with } \hat{U} = (\hat{u}_1 | \hat{u}_2 | \hat{u}_3).$$

- Assuming that $r = 2$,

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

Imputation of the missing values in Y

- Impute the missing values Y_{i1} for $i \in \{1, \dots, n\}$ such that $M_{i1} = 0$ **using the conditional expectation of** (Y_{i1}) **given** Y_{i2} **and** Y_{i3}.

$$Y = \begin{array}{c} \begin{array}{ccc} Y_{.1} & Y_{.2} & Y_{.3} \end{array} \\ \begin{pmatrix} 12 & 28 & 31 \\ \text{NA} & 23 & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \text{NA} & 3 & 7 \end{pmatrix} \end{array} \rightarrow Y = \begin{array}{c} \begin{array}{ccc} Y_{.1} & Y_{.2} & Y_{.3} \end{array} \\ \begin{pmatrix} 12 & 28 & 31 \\ 16 & 23 & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ 21 & 3 & 7 \end{pmatrix} \end{array}$$

The methodology is extended to the general case, for any data with p covariates, r latent variables and d missing variables.

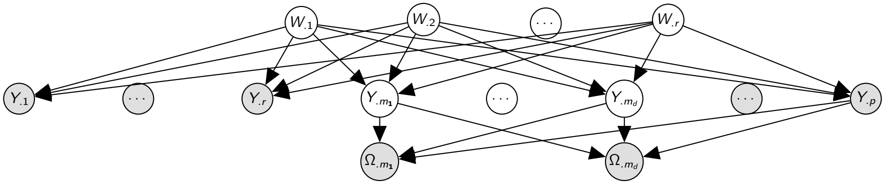
$$Y = \begin{array}{c} \begin{array}{ccccc} Y_{.1} & Y_{.2} & Y_{.3} & \dots & Y_{.p} \end{array} \\ \begin{pmatrix} \text{NA} & 23 & \text{NA} & \dots & 45 \\ 32 & \text{NA} & 24 & \dots & \text{NA} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{NA} & 3 & \text{NA} & \dots & 46 \end{pmatrix} \end{array} \rightarrow Y = \begin{array}{c} \begin{array}{ccccc} Y_{.1} & Y_{.2} & Y_{.3} & \dots & Y_{.p} \end{array} \\ \begin{pmatrix} 16 & 23 & 89 & \dots & 45 \\ 32 & 6 & 24 & \dots & 22 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 21 & 3 & 7 & \dots & 46 \end{pmatrix} \end{array}$$

Imputation of the missing values in Y

- Impute the missing values Y_{i1} for $i \in \{1, \dots, n\}$ such that $M_{i1} = 0$ **using the conditional expectation of** (Y_{i1}) **given** Y_{i2} **and** Y_{i3}.

$$
Y = \begin{array}{c} \begin{array}{ccc} Y_{.1} & Y_{.2} & Y_{.3} \end{array} \\ \begin{pmatrix} 12 & 28 & 31 \\ \color{cyan}{\mathrm{NA}} & 23 & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \color{cyan}{\mathrm{NA}} & 3 & 7 \end{pmatrix} \end{array} \rightarrow Y = \begin{array}{c} \begin{array}{ccc} Y_{.1} & Y_{.2} & Y_{.3} \end{array} \\ \begin{pmatrix} 12 & 28 & 31 \\ \color{cyan}{16} & 23 & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \color{cyan}{21} & 3 & 7 \end{pmatrix} \end{array}
$$

The methodology is extended to the general case, for any data with p covariates, r latent variables and d missing variables.



Measuring the performance

- estimation of B: RV coefficient (cosine between two subspaces).
- imputation of Y: $\|(\hat{Y} - Y) \odot (1 - \Omega)\|_F^2 / \|Y \odot (1 - \Omega)\|_F^2$.

Other methods

- **MAR** our method which has been adapted to handle MAR data (inspired by [Mohan et al., 2018] in linear models);
- **EMMAR**: EM algorithm to perform PPCA with MAR values [Ilin and Raiko, 2010];
- **SoftMAR**: matrix completion using iterative soft-thresholding singular value decomposition algorithm [Mazumder et al., 2010] relevant only for M(C)AR values;
- **MNARparam**: matrix completion technique modeling the MNAR mechanism with a parametric logistic model [Sportisse et al., 2018];
- **Del**: the naive listwise deletion method;
- **Mean**: the imputation by the mean.

$n = 1000,\ p = 20,\ r = 3,\ \sigma = 0.8$
10 MNAR variables with

$$\forall m \in [1:10],\ \mathbb{P}(\Omega_{.m} = 1|Y) = \mathbb{P}(\Omega_{.m} = 1|Y_{.m}, Y_{.k}, Y_{.\ell}),$$

where k and ℓ are indexes of MNAR variables randomly chosen such that $k \neq \ell \in [1:10] \backslash \{m\}$.

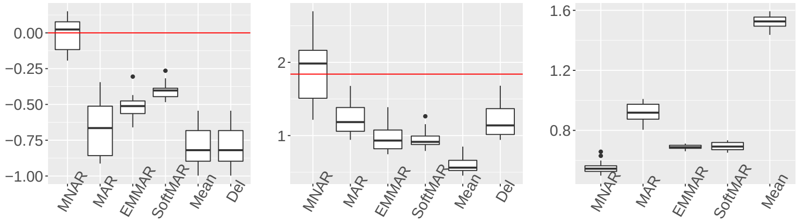


Figure: Mean estimation (left graphic), variance estimation (middle graphic) of one missing variable and prediction error (right graphic).

- Introduction additional MNAR values in the variable *HR.ph* using a logistic self-masked mechanism. Other variables considered as M(C)AR.

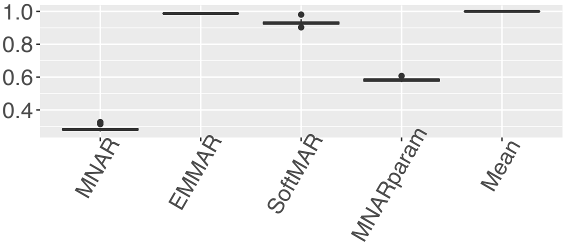
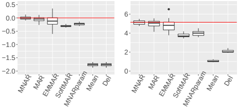
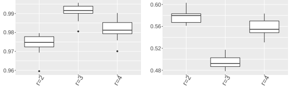
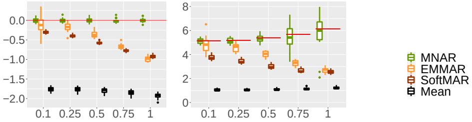


Figure: Comparison of the prediction error for the TraumaBase data.

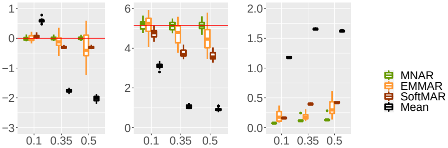
Numerical experiments
Others

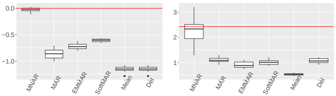






- Robustness to the percentage of missing values.
- **Robustness to the noise.**
- Model (PPCA) misspecification
- **Rank misspecification.**





Conclusion

Take-home messages

- MNAR is hard.
- Modeling the mechanism and using an EM algorithm is computationally costly.
- Our proposal: **new estimation and imputation method to perform PPCA with MNAR data**,
- **without any need of modeling the missing mechanism,**
- with strong theoretical guarantees as identifiability and consistency and efficient algorithm.

Perspectives

- Estimating the rank in the PPCA setting with MNAR data.
- Extension to the exponential family to process count data.

► Estimation and imputation in probabilistic principal component analysis withmissing not at random data. [Sportisse et al., 2020]

References I

Ibrahim, J. G., Lipsitz, S. R., and Chen, M.-H. (1999).
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.

Ilin, A. and Raiko, T. (2010).
Practical approaches to principal component analysis in the presence of missing values.
Journal of Machine Learning Research, 11(Jul):1957–2000.

Little, R. J. and Rubin, D. B. (2014).
Statistical analysis with missing data, volume 333.
John Wiley & Sons.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010).
Spectral regularization algorithms for learning large incomplete matrices.
Journal of machine learning research, 11(Aug):2287–2322.

Miao, W. and Tchetgen, E. T. (2018).
Identification and inference with nonignorable missing covariate data.
Statistica Sinica, 28(4):2049–2067.

References II

Mohan, K., Thoemmes, F., and Pearl, J. (2018).
Estimation with incomplete data: The linear case.
In *IJCAI*, pages 5082–5088.

Pearl, J. (2003).
Causality: models, reasoning, and inference.
Econometric Theory, 19(675-685):46.

Rubin, D. B. (1976).
Inference and missing data.
Biometrika, 63(3):581–592.

Sportisse, A., Boyer, C., and Josse, J. (2018).
Imputation and low-rank estimation with missing non at random data.
arXiv preprint arXiv:1812.11409.

Sportisse, A., Boyer, C., and Josse, J. (2020).
Estimation and imputation in probabilistic principal component analysis with missing not at random data.

Tang, G., Little, R. J., and Raghunathan, T. E. (2003).
Analysis of multivariate missing data with nonignorable nonresponse.
Biometrika, 90(4):747–764.

Missing-data mechanism

ϕ: the unknown parameters of the missingness.

$$\begin{array}{cc} Y_1 & Y_2 \\ \end{array}$$
$$\begin{pmatrix} 1 & 2 \\ 3 & 20 \\ 22 & 4 \end{pmatrix}$$

MCAR mechanism

The missingness does not depend on the variables.

$$p(\Omega|Y;\phi) = p(\Omega;\phi), \quad \forall Y, \phi.$$

MAR

$$\begin{pmatrix} 1 & 2 \\ \mathrm{NA} & 20 \\ 22 & 4 \end{pmatrix}$$

MAR mechanism

The missingness depends on the observed variables.

$$p(\Omega|Y;\phi) = p(\Omega|Y_{\mathrm{obs}};\phi), \quad \forall Y_{\mathrm{mis}}, \phi.$$

MNAR

$$\begin{pmatrix} 1 & 2 \\ 3 & 20 \\ \mathrm{NA} & 4 \end{pmatrix}$$

MNAR mechanism

Other case, i.e.

$$p(\Omega|Y;\phi) = p(\Omega|Y_{\mathrm{obs}}, Y_{\mathrm{mis}};\phi), \quad \forall \phi.$$

self-masked when the missingness of a variable depends on the variable itself.

...but hard to handle

- Likelihood approach: maximizing the joint log-likelihood,

$$\ell(\Theta, \phi; Y, M) = p(Y; \Theta)p(M|Y; \phi),$$

with Θ: unknown data distribution parameter.

- Missing data: maximizing the observed joint log-likelihood,

$$\ell(\Theta, \phi; Y_{\mathrm{obs}}, M) = \int \ell(\Theta, \phi; Y, M) dY_{\mathrm{mis}}.$$

✓ In the MAR setting, one can ignore the mechanism since

$$p(M|Y; \phi) = p(M|Y_{\mathrm{obs}}; \phi).$$

$$\Rightarrow \ell(\Theta, \phi; Y_{\mathrm{obs}}, M) \propto \ell(\Theta; Y_{\mathrm{obs}}) = \int \ell(\Theta; Y) dY_{\mathrm{mis}}.$$

✗ In the MNAR setting, one should consider the mechanism.

Existing works for handling MNAR data (1)

- Modeling the MNAR mechanism via a **Logistic Model**:
  $\forall i \in [1, n]$, $\phi_j = (\phi_{1j}, \phi_{2j})$ denoting a parameter vector:

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

  (self-masked MNAR mechanism)

- Maximizing the joint log-likelihood with the **EM algorithm** in linear models [Ibrahim et al., 1999] or in low-rank models [S., Boyer, Josse 2018].

  ✓ Handling MNAR data.
  ✗ Often restricted to a limited number of MNAR variables.
  ✗ Parametric assumption for the mechanism distribution.
  ✗ Computationally costly.

General setting

- d MNAR variables indexed by $\mathcal{M} := \{m_1, \ldots, m_d\} \subset \{1, \ldots, p\}$ (with $d < p$).
- Other variables are observed or M(C)AR.
- The distribution of the mechanism may depend on all variables (missing or observed) except r **pivot variables**, indexed by $\mathcal{J}$.

$$\forall m \in \mathcal{M}, \quad \mathbb{P}(\Omega_{.m} = 1 | Y) = \mathbb{P}(\Omega_{.m} = 1 | (Y_{.k})_{k \in \bar{\mathcal{J}}}),$$

$\bar{\mathcal{J}} = \{1, \ldots, p\} \backslash \mathcal{J}$.

Main assumptions

A1. $\forall m \in \mathcal{M}, \forall j \in \mathcal{J}, \quad \left(B_{.m} \quad (B_{.j'})_{j' \in \mathcal{J}_{-j}}\right)$ is invertible.

($\rightsquigarrow$ implies that any variable is generated by all the latent variables)

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

($\rightsquigarrow$ follows from the missing-data mechanism above)

Imputation of the missing values in Y

- Impute the missing values Y_{i1} for $i \in \{1, \dots, n\}$ such that $M_{i1} = 0$ **using the conditional expectation of** (Y_{i1}) **given** Y_{i2} **and** Y_{i3}.

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

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

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

$$Y = \begin{array}{c} \begin{array}{ccc} Y_{.1} & Y_{.2} & Y_{.3} \end{array} \\ \begin{pmatrix} 12 & 28 & 31 \\ \mathrm{NA} & 23 & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \mathrm{NA} & 3 & 7 \end{pmatrix} \end{array} \rightarrow Y = \begin{array}{c} \begin{array}{ccc} Y_{.1} & Y_{.2} & Y_{.3} \end{array} \\ \begin{pmatrix} 12 & 28 & 31 \\ 16 & 23 & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ 21 & 3 & 7 \end{pmatrix} \end{array}$$

Imputation of the missing values in Y

- Impute the missing values Y_{i1} for $i \in \{1, \ldots, n\}$ such that $M_{i1} = 0$ **using the conditional expectation of** (Y_{i1}) **given** Y_{i2} **and** Y_{i3}.

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

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

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

The methodology is extended to the general case, for any data with p covariates, r latent variables and d missing variables.

⋆ $r = 2$, $p = 10$, $n = 1000$, $\sigma = 0.1$, 7 self-masked MNAR variables.

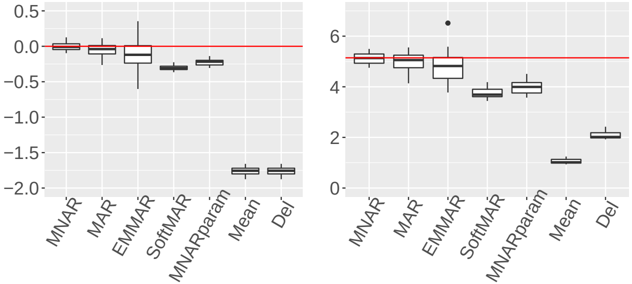
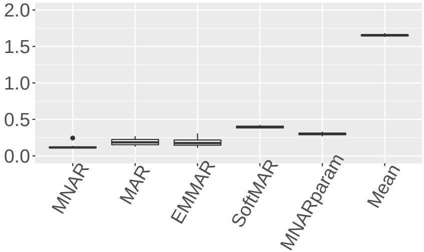


Figure: Mean and variance estimation for one MNAR variable.

⋆ $r = 2$, $p = 10$, $n = 1000$, $\sigma = 0.1$, 7 self-masked MNAR variables.



	RV
MNAR	0.997
MAR	0.993
EMMAR	0.988
SoftMAR	0.986
Mean	0.593

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

Exogeneity does not hold but does it have an impact on the results when the noise increases?
⋆ $r = 2$, $p = 10$, $n = 1000$, 7 self-masked MNAR variables.
For different values of the noise level.

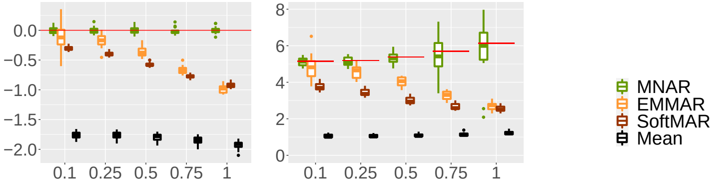


Figure: Mean and variance for one missing variable.

⋆ $r = 2$, $p = 10$, $n = 1000$, 7 self-masked MNAR variables.
For different values of the noise level.

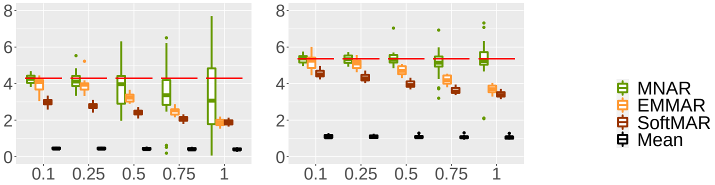


Figure: Covariances between a MNAR variable and a pivot one (left) and between two MNAR variables (right).

⋆ $r = 2$, $p = 10$, $n = 1000$, 7 self-masked MNAR variables.

For different values of the noise level.

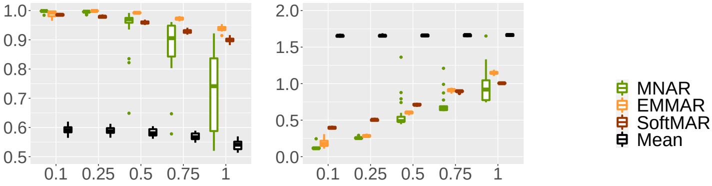


Figure: RV coefficient (left) and prediction error (right).

⋆ $r = 2$, $p = 10$, $n = 1000$, 7 self-masked MNAR variables.
For different values of the noise level.

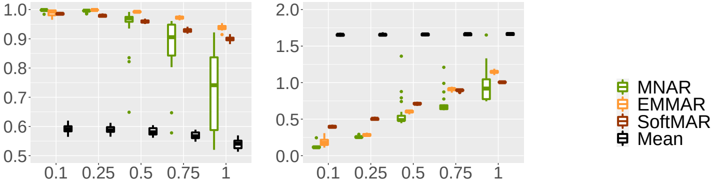


Figure: RV coefficient (left) and prediction error (right).

- ✗ Bias for the covariance MNAR/pivot variables.
- ✗ Loading matrix estimation detetoriates as the data gets noiser.
- ✓ Some estimations (mean, variance) are unbiased.
- ✓ In term of prediction error, it remains competitive.

⋆ $r = 3$, $p = 20$, $n = 1000$, $\sigma = 0.8$, 10 self-masked MNAR variables.

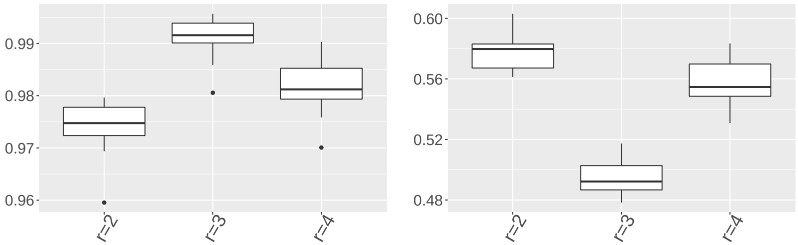


Figure: RV coefficients for the loading matrix (left graphic) and prediction error (right graphic) for different cases where the rank is either underestimated, well estimated or overestimated.

⋆ $r = 3$, $p = 20$, $n = 1000$, $\sigma = 0.8$, 10 self-masked MNAR variables.

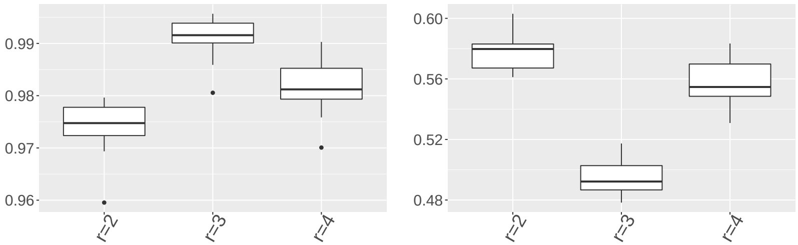


Figure: RV coefficients for the loading matrix (left graphic) and prediction error (right graphic) for different cases where the rank is either underestimated, well estimated or overestimated.

✓ Stability to rank misspecification.