Estimation and imputation in Probabilistic Principal Component Analysis with Missing Not At Random data

Aude Sportisse [1] Claire Boyer [1] Julie Josse [2]

[1]Sorbonne Université [2]INRIA

Proposal:

Handling **several MNAR variables (coupled with M(C)AR variables)** in PPCA model **without modeling the missing-data mechanism** and using **only the observed information**: **identifiability** and **estimation** of the model parameters and **imputation** of the missing values.

Missing data

- One of the ironies of working with Big Data is that missing data plays an ever more significant role.
- Three types of **missing-data mechanisms** [1]:
  - MCAR missing values does not depend on the data.
  - MAR missing values depends on the observed variables.
  - MNAR missing values depends on both observed and unobserved data such as its value itself.
- Most methods focus on the easiest M(C)AR data, here **focus on MNAR data**.

Existing works for MNAR data

- Modeling the MNAR mechanism [2, 3].
  ✗ Parametric assumption for the mechanism distribution.
  ✗ Computationally costly.
- Without modeling the mechanism and by only using all available observed cells [4, 5, 6].
  ✗ Restricted to simple linear models with few missing var.
- Most of the works consider **self-masked MNAR** variables: the missingness of a variable depends on the variable itself. E.g. the probability to have a missing value on income depends on the value of income (rich people less inclined to reveal their income).

References

[1] R. JA Little and D. B Rubin. *Statistical analysis with missing data.*

[2] J. G Ibrahim, S. R Lipsitz, and M-H Chen. Missing covariates in generalized linear models when the missing data mechanism is non-ignorable.

[3] A. Sportisse, C. Boyer, and J. Josse. Imputation and low-rank estimation with missing non at random data.

[4] W. Miao and E Tchetgen Tchetgen. Identification and inference with nonignorable missing covariate data.

[5] G. Tang, R. JA Little, and T. E Raghunathan. Analysis of multivariate missing data with nonignorable nonresponse.

[6] K. Mohan, F. Thoemmes, and J. Pearl. Estimation with incomplete data: The linear case.

[7] A. Ilin and T. Raiko. Practical approaches to principal component analysis in the presence of missing values.

[8] R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices.

Setting

- Data matrix $Y \in \mathbb{R}^{n\times p}$,
- **Coefficients matrix** $B \in \mathbb{R}^{r\times p}$ of rank $r < \min\{n, p\}$
- r **latent variables** grouped in $W \in \mathbb{R}^{n\times r}$,
- $\Omega \in \mathbb{R}^{n\times p}$ the missing-data pattern: $\Omega_{ij} = 1$ if Y_{ij} is observed, 0 otherwise.

$\dashrightarrow$ Access only to $Y \odot \Omega + \text{NA} \odot (1-\Omega)$

$$\begin{array}{ccccc} Y_{.1} & Y_{.2} & Y_{.3} & \dots & Y_{.p} \end{array}$$
$$\begin{pmatrix} 12 & 28 & 31 & \dots & \text{NA} \\ \text{NA} & 23 & 89 & \dots & 85 \\ 32 & 6 & 24 & \dots & \text{NA} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \text{NA} & 3 & 7 & \dots & 11 \end{pmatrix}$$

PPCA model

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

PPCA model identifiability with MNAR data

- Identifying the PPCA model $\Leftrightarrow$ identifying the missing-data mechanism.

Assumptions:

A01. d self-masked MNAR var. and $p-d$ other MCAR (or observed) var., F_m known strictly monotone with a finite support $\mathbb{P}(\Omega_{im} = 1|Y_{i.}) = F_m(\phi_m^0 + \phi_m^1 Y_{im})$, with $\phi_m = (\phi_m^1, \phi_m^2)$ the missing-data mechanism.

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

Proposition 1: idenfiability

- *Under **A01.** and **A02.**, the PPCA parameters (α, Σ) and the missing-data mechanism parameter ϕ are identifiable.*
- *Assuming that the noise level σ^2 is known, the coefficient matrix B is identifiable up to a row permutation.*

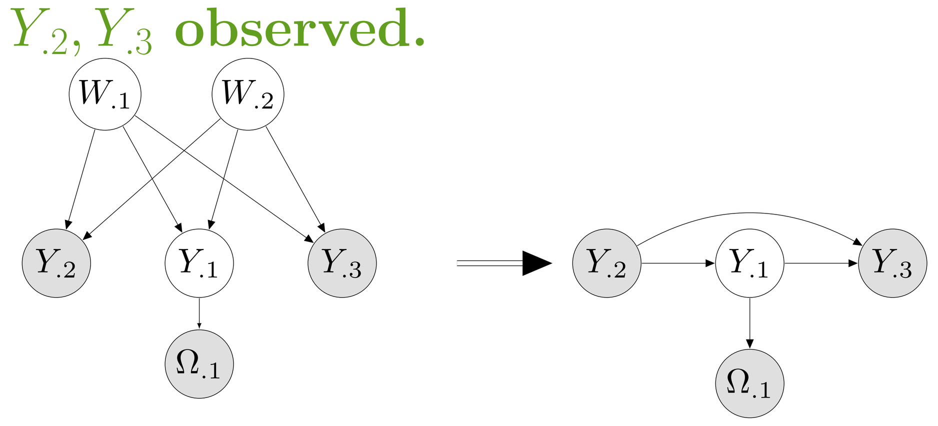
General MNAR setting for estimation/imputation

- r **pivot variables** indexed by $\mathcal{J}$ observed or MCAR.
- d **general MNAR** variables indexed by $\mathcal{M}$, missingness depends on all the variables except r ones. With $\bar{\mathcal{J}} = \{1,\dots,p\} \setminus \mathcal{J}$,

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

Estimation with MNAR data

Toy exemple: $p = 3$, $r = 2$, $Y_{.1}$ MNAR, $Y_{.2}, Y_{.3}$ observed.



- $(Y_{.1}\ Y_{.2}\ Y_{.3}) = \mathbf{1}\,(\alpha_1\ \alpha_2\ \alpha_3) + (W_{.1}\ W_{.2})\,B + \epsilon$.
- Assumption: **fully connected PPCA** i.e. any variable generated by all the latent variables $\Rightarrow$ linear links can be established.

Assumptions:

A1. $(B_{.1}\ B_{.2})$ is invertible ✓ fully connected PPCA

A2. $Y_{.2} \perp\!\!\!\perp \Omega_{.1}|Y_{.3}$ ✓ MNAR mechanism

A3. Consistent estimators for α_2 and α_3 ✓ observed variables

A4. Consistent estimators for $(\mathcal{B}^c_{2\to1,3[k]})_{k\in\{0,1,3\}}$ ✓ noise tends to zero

Proposition 4: mean estimation

- *Under **A1.** and **A2.**, one defines the estimator for the mean of $Y_{.1}$*
  $\hat{\alpha}_1 := \frac{\hat{\alpha}_2 - \hat{\mathcal{B}}^c_{2\to1,3[0]} - \hat{\mathcal{B}}^c_{2\to1,3[3]}\hat{\alpha}_3}{\hat{\mathcal{B}}^c_{2\to1,3[1]}}$, $(\mathcal{B}^c_{2\to1,3[k]})_{k\in\{0,1,3\}}$*: effects of $Y_{.2}$ on $Y_{.1}, Y_{.3}$ when $\Omega_{.1} = 1$.*
- *Under **A3.** and **A4.**, $\hat{\alpha}_1$ is a consistent estimator of α_1.*

- Same method for the variance and covariances, $\hat{\Sigma}$ estimates Σ.
- $\hat{\Sigma} - \sigma^2\text{Id}_{3\times3}$ estimates $B^TB \Rightarrow$ estimation of B (r, σ^2 known).

Imputation of the missing values

- 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}.

Practical implementation

- Pivot variable selection: with experts or var. with less %NA, bigger set ($> r$) and estimates aggregation.
- $(\mathcal{B}^c_{2\to1,3[k]})_{k\in\{0,1,3\}}$ 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.
- $\hat{\alpha}_2$ and $\hat{\alpha}_3$ are computed as empirical means of $Y_{.2}$ and $Y_{.3}$.
- CV strategy to estimate σ^2 and r.

Application to clinical data TraumaBase®

- **EMMAR**: EM algorithm to perform PPCA with MAR values [7].
- **SoftMAR**: matrix completion using iterative SVD algorithm for M(C)AR values [8].
- **MNARparam**: low-rank method for MNAR values (modeling the mechanism) [3].
- **Mean**: naive imputation by the mean.

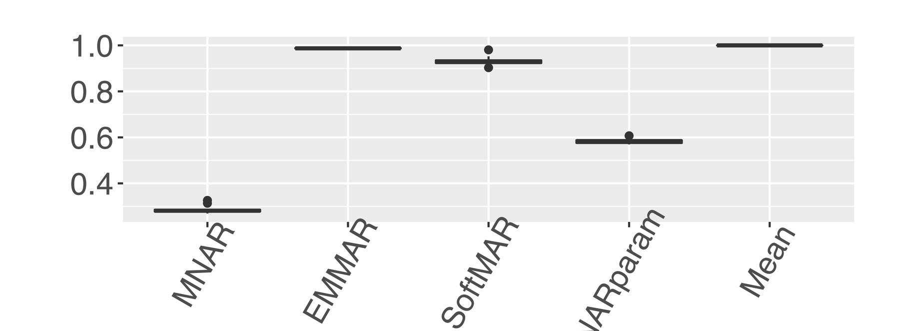


Figure 1: Our method **MNAR** compared with others methods in terms of imputation error.