Handling heterogeneous and MNAR missing data in statistical learning frameworks:

imputation based on low-rank models
online linear regression with SGD,
and model-based clustering

PhD defense
Aude Sportisse (Sorbonne University)
Supervised by Claire Boyer and Julie Josse

June 29, 2021

Missing values are everywhere

- Growing masses of data, multiplication of sources
  ⇒ `Not Available` values (`NA`)
- Our public health application: the **Traumabase®** dataset.

250 clinical variables (heterogeneous)

Trauma.center	Heart rate	Death	Anticoagulant. therapy	Glascow score	...
Pitie-Salpêtrière	88	0	No	3	
Beaujon	103	0	NA	5	
Bicêtre	NA	0	Yes	6	
Bicêtre	NA	0	No	NA	
Lille	62	0	Yes	6	
Lille	NA	0	No	NA	
⋮	⋮	⋮	⋮	⋮	

1 patient; in total: **30 000 patients**

Missing values are everywhere

- Growing masses of data, multiplication of sources
  $\Rightarrow$ `Not Available` values (`NA`)
- Our public health application: the **Traumabase**® dataset.

Trauma.center	Heart rate	Death	Anticoagulant. therapy	Glascow score	...
Pitie-Salpêtrière	88	0	No	3	
Beaujon	103	0	NA	5	
Bicêtre	NA	0	Yes	6	
Bicêtre	NA	0	No	NA	
Lille	62	0	Yes	6	
Lille	NA	0	No	NA	
⋮	⋮	⋮	⋮	⋮	

23 different hospitals

Missing values are everywhere

Traumabase® dataset

- now **30 000** patients (begin of this PhD thesis: 10 000).
- **250** heterogeneous variables: continuous, categorical, ordinal,...
- **23** different hospitals
- **missing** values everywhere (1% to 90% NA in each variable).

- **Imputation:** provide a **complete dataset** to the doctors.
- **Estimation:** explain the level of platelet with pre-hospital characteristics.
- **Prediction:** predict the administration or not of the tranexomic acid.
- **Clustering:** identify relevant groups of patients sharing similarities.

Q: *How to deal with missing values?*

What we should not do

$$
\begin{pmatrix}
\text{Pitie-Salpêtrière} & 88 & 0 & \text{No} & 3 \\
\text{Beaujon} & 103 & 0 & \text{NA} & 5 \\
\text{Bicêtre} & \text{NA} & 0 & \text{Yes} & 6 \\
\text{Bicêtre} & \text{NA} & 0 & \text{No} & \text{NA} \\
\text{Lille} & 62 & 0 & \text{Yes} & 6 \\
\text{Lille} & \text{NA} & 0 & \text{No} & \text{NA}
\end{pmatrix}
\qquad
\begin{pmatrix}
\text{Pitie-Salpêtrière} & 88 & 0 & \text{No} & 3 \\
\text{Beaujon} & 103 & 0 & \text{NA} & 5 \\
\text{Bicêtre} & \text{NA} & 0 & \text{Yes} & 6 \\
\text{Bicêtre} & \text{NA} & 0 & \text{No} & \text{NA} \\
\text{Lille} & 62 & 0 & \text{Yes} & 6 \\
\text{Lille} & \text{NA} & 0 & \text{No} & \text{NA}
\end{pmatrix}
$$

Discarding individuals with missing values **is not** a solution

- Loss of information .

  Traumabase$^{\textcircled{R}}$: only 5% of the rows are kept.
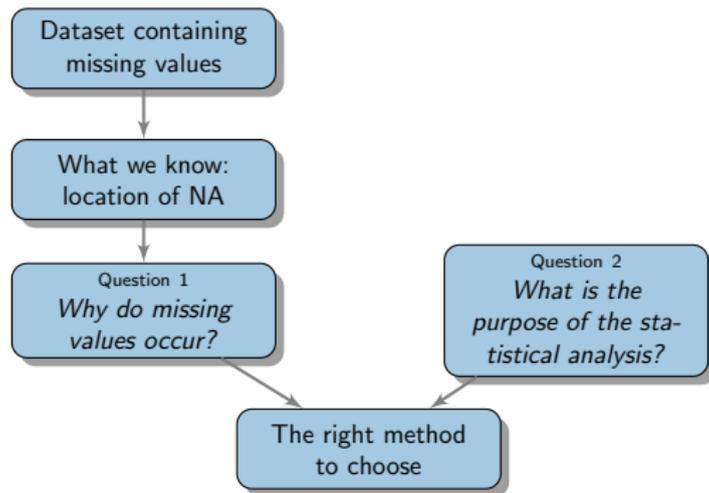
- Bias in the analysis .

  Kept observations: sub-population **not necessarily representative** of the overall population.

What we should do: handling missing values

The right method to choose

Q: *How to choose the right method to handle missing values?*

Imputation? Estimation? Prediction?

- The goal is **not necessarily** to obtain a complete dataset.
- A solution can be to **embed missing data management** into the statistical paradigm.

Missing-data pattern

- $X = (X_{1.}|\ldots|X_{n.})^T$ data sample of *n observations* , *d* variables
- $X_{i.} = (X_{i1}, \ldots, X_{id})^T \in \mathcal{X}$, with $\mathcal{X}$ *d* -dimensional features space
- $X_{i.}^{\mathrm{obs}}$ ($X_{i.}^{\mathrm{mis}}$): observed (missing) variables for the individual i.

Missing-data pattern

$M \in \{0, 1\}^{n \times d}$: indicates where are the missing values in X.

$$\forall i, j, \quad M_{ij} = \begin{cases} 1 & \text{if } X_{ij} \text{ is missing,} \\ 0 & \text{otherwise.} \end{cases}$$

Pitie-Salpêtrière	88	0	No	3		0	0	0	0	0
Beaujon	103	0	NA	5		0	0	0	1	0
Bicêtre	NA	0	Yes	6	$\rightarrow$	0	1	0	0	0
Bicêtre	NA	0	No	NA		0	1	0	0	1
Lille	62	0	Yes	6		0	0	0	0	0
Lille	NA	0	No	NA		0	1	0	0	1

We observe: $X \odot (1 - M)$, M and not X

Missing-data mechanism (Rubin, 1976)

Data values (X)	← Link? →	Missing values (M)
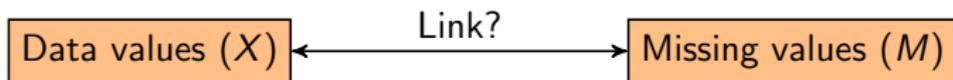
$$f_{M|X}(M|X; \phi), \phi \in \Omega_\phi$$

Missing Completely At Random (MCAR)

$$f_{M|X}(M|X; \phi) = f_M(M; \phi)$$

MCAR

Machines fail,
Doctors forget to fill the form

Missing At Random (MAR)

X^{obs}: observed component of X.

$$f_{M|X}(M|X; \phi) = f_{M|X^{\mathrm{obs}}}(M|X^{\mathrm{obs}}; \phi)$$

MAR

Aggregation of datasets

	HR	Death	A. therapy	GCS
Lille	65	0	Yes	6
Lille	59	0	No	4
Pitié	62	0	NA	6
Pitié	84	0	NA	5

Missing Not At Random (MNAR)

The MAR assumption does not hold.
The missingness can depend on the
missing data value itself.

MNAR

Emergency situations

HR		HR
65		65
59	"underlying" values:	59
62		62
NA		84

Key tools for missing-data analysis

- Parametric estimation: model the joint distribution (X, M) parametrized by $\theta, \phi \in \Omega_{\theta,\phi}$.
- Likelihood-approach: maximizing the full observed likelihood.

$$
\begin{aligned}
L_{\mathrm{full,obs}}(\theta, \phi; X^{\mathrm{obs}}, M) &= \int L_{\mathrm{full}}(\theta, \phi; X, M) dX^{\mathrm{mis}} \\
&= \int f(X; \theta) f(M|X; \phi) dX^{\mathrm{mis}} \\
&= f(M|X^{\mathrm{obs}}; \phi) \int f(X; \theta) dX^{\mathrm{mis}} \qquad \mathrm{M(C)AR\ mecha.} \\
&\propto L_{\mathrm{ign}}(\theta; X^{\mathrm{obs}}) = \int f(X; \theta) dX^{\mathrm{mis}}
\end{aligned}
$$

M(C)AR: one can ignore the mechanism.

MNAR: one should consider the mechanism.

Classical methods for M(C)AR data

- Most of the methods dedicated to MCAR.
  - EM algorithm for estimation [Dempster et al., 1977].
  - Multiple imputation for estimation and to get the variance of the estimates [Buuren and Groothuis-Oudshoorn, 2010].
  - Matrix completion [Hastie et al., 2015, Mattei and Frellsen, 2019].
- In this PhD thesis: focus on **MNAR**.

MNAR from every angle

We should consider (X, M) (not-ignorable mechanism).

The main MNAR specifications

- selection model [Heckman, 1979]:

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

- pattern-mixture model [Little, 1993]:

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

Q: *How to choose the MNAR specification ?*

- Estimate the parameters of the data distribution: selection models.
- Model the data distribution in the strata defined by different missing-data patterns: pattern-mixture models.

MNAR from every angle

We should consider (X, M) (not-ignorable mechanism).

The main MNAR specifications

- selection model [Heckman, 1979]:

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

- pattern-mixture model [Little, 1993]:

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

Q: *How to choose the MNAR specification ?*

- Estimate the parameters of the data distribution: **selection models.**
- Model the data distribution in the strata defined by different missing-data patterns: pattern-mixture models.

MNAR from every angle

We should prove the identifiability of the parameters.

Identifiability issue in the MNAR case Credit: Ilya Shpitser

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

- **Case 1:** X missing only if $X = 1$.

$$X = [1, 1, 0, 1, 1, 0], \ \mathbb{P}(X = 1) = 2/3.$$

- **Case 2:** X missing only if $X = 0$.

$$X = [1, 0, 0, 1, 0, 0], \ \mathbb{P}(X = 1) = 1/3.$$

$\Rightarrow$ We start from 2 equal observed distribution. It leads to different parameters of the data distribution $\mathbb{P}(X = 1)$.

Identifiability: the parameters of (X, M) are uniquely determined from available information $(X, M = 0)$.

MNAR from every angle

Specific methods should be used.

Existing methods for MNAR data

- Model the joint distribution (X, M) [Ibrahim et al., 1999].
  - Costly, done for few missing variables, specific missing-data mechanism.
- Semi-parametric models: model either X or $M|X$ [Tang and Ju, 2018]
  - For regression model when Y is missing and not X.
- Available-case analysis without modeling the missing-data mechanism [Mohan et al., 2018].
  - for linear regression.

$$X^{\mathrm{NA}} = \begin{pmatrix} 12 & 28 & \mathtt{NA} \\ 23 & \mathtt{NA} & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \mathtt{NA} & 3 & 7 \end{pmatrix}, \; X^{AC} = \begin{pmatrix} 12 & 28 & \mathtt{NA} \\ 23 & \mathtt{NA} & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \mathtt{NA} & 3 & 7 \end{pmatrix}$$

What this thesis is about

Handling MNAR data in low-rank models

- With fixed effects
  EM algorithm, MNAR
- With random effects
  available-case analysis, MNAR, identifiability

Handling missing data in statistical learning frameworks

- Online linear regression
  naive imputation + debiasing, SGD, heterogeneous MCAR
- Model-based clustering
  EM algorithms, MNAR, identifiability

R-miss-tastic: https://rmisstastic.netlify.app/

With Imke Mayer, Julie Josse, Nicholas Tierney and Nathalie Vialaneix.

- Main methods, references.
- Implementations (in R and python) for managing missing data, whether to impute, estimate or predict.

Outline

Low-rank model with fixed effects

- $X \in \mathbb{R}^{n \times d}$ noisy realisation of a low-rank matrix $\Theta \in \mathbb{R}^{n \times d}$:

$$X = \Theta + \epsilon, \text{ where } \begin{cases} \Theta \text{ with rank } r < \min\{n, d\}, \\ \epsilon_{ij} \overset{\perp\!\!\!\perp}{\sim} \mathcal{N}(0, \sigma^2), \forall i \in [1, n]. \end{cases}$$

- $X_{ij} \overset{\perp\!\!\!\perp}{\sim} \mathcal{N}(\Theta_{ij}, \sigma^2)$, σ^2 is assumed to be known.
- Access only to the missing-data matrix $X \odot (1 - M)$,

How to estimate Θ ? How to impute missing values ?

Low-rank model with fixed effects

- $X \in \mathbb{R}^{n \times d}$ noisy realisation of a low-rank matrix $\Theta \in \mathbb{R}^{n \times d}$:

$$X = \Theta + \epsilon, \text{ where } \begin{cases} \Theta \text{ with rank } r < \min\{n, d\}, \\ \epsilon_{ij} \overset{\perp}{\sim} \mathcal{N}(0, \sigma^2), \forall i \in [1, n]. \end{cases}$$

- $X_{ij} \overset{\perp}{\sim} \mathcal{N}(\Theta_{ij}, \sigma^2)$, σ^2 is assumed to be known.
- Access only to the missing-data matrix $X \odot (1 - M)$,

How to estimate Θ ? How to impute missing values ?

M(C)AR data: convex relaxation of the rank

$$\hat{\Theta} \in \operatorname{argmin}_{\Theta} \underbrace{\|(1 - M) \odot (X - \Theta)\|^2}_{\text{fits the data at best}} + \underbrace{\lambda \|\Theta\|_\star}_{\text{captures the low rank structure}},$$

- $\lambda \in \mathbb{R}$: regularization term.
- $\|\Theta\|_\star = \sum_{i=1}^{\operatorname{rank}\Theta} \sigma_i(\Theta)$, with $\sigma_i(\Theta)$ the singular values of Θ.
- Equivalence with the EM algorithm.

Method 1: modelling the mechanism

- self-masked **MNAR mechanism** (with a logit link)

$$f_{M|X}(M_{ij}|X_{ij}; \phi) = [(1 + e^{-\phi_{1j}(X_{ij}-\phi_{2j})})^{-1}]^{M_{ij}}$$
$$[1 - (1 + e^{-\phi_{1j}(X_{ij}-\phi_{2j})})^{-1}]^{(1-M_{ij})}.$$

- Maximize $L_{\mathrm{full,obs}}(\Theta, \phi; X^{\mathrm{obs}}, M) = \int f_X(X; \Theta) f_{M|X}(M|X; \phi) dX^{\mathrm{mis}}$

EM algorithm [S., Boyer, Josse 2020]

- **E-step:**

$$Q(\Theta, \phi|\Theta^r, \phi^r) = \mathbb{E}_{X^{\mathrm{mis}}} \left[L_{\mathrm{full}}(\Theta, \phi; X, M)|X^{\mathrm{obs}}, M; \Theta^r, \phi^r \right]$$

- **M-step:** $\Theta^{r+1}, \phi^{r+1} \in \mathrm{argmax}_{\Theta,\phi} \; Q(\Theta, \phi|\Theta^r, \phi^r) + \lambda\|\Theta\|_\star$

- E-step: Monte-Carlo approximation and SIR algorithm.
- M-step: Separability of Q:
  - Θ: `softImpute` [Hastie and Mazumder, 2015], FISTA
  - ϕ: Newton-Raphson algorithm.

  Handling MNAR data (under a self-masked logistic model) but
  computationally costly.

Method 2: implicitly modelling the mechanism

Add the mask !

$$
\begin{array}{cc}
X_1 & X_2 \\
\end{array}
$$

$$
\underbrace{\begin{pmatrix} 1 & 2 \\ 3 & \text{NA} \\ \text{NA} & 4 \end{pmatrix}}_{\text{To estimate } \Theta} \rightarrow
\underbrace{\begin{pmatrix} 1 & 2 & 0 & 0 \\ 3 & \text{NA} & 0 & 1 \\ \text{NA} & 4 & 1 & 0 \end{pmatrix}}_{\text{To estimate } \Xi}
$$

$$
\begin{array}{cccc}
X_1 & X_2 & M_1 & M_2 \\
\end{array}
$$

Solve the classical MAR optimization problem

$$
\hat{\Xi} \in \operatorname{argmin}_{\Xi} \tfrac{1}{2} \| \ [(1 - M) \odot X | M] \ - [M | 1] \odot \Xi \|_F^2 + \lambda \| \Xi \|_\star,
$$

- `softImpute`, FISTA.
- taking into account the mask binary type, with a Penalized Iteratively Reweighted Least Squares algorithm [Robin et al., 2020].

    Computationally efficient but no theoretical guaranties .

Results on real data

- $\simeq$ 3200 patients with brain trauma injury, 9 quantitative variables containing missing values are selected by doctors.
- Numerical comparison:
  - Methods which consider MAR data (in blue): the regularized iterative PCA and the matrix completion `softImpute` algorithms.
  - Method 1 by considering MNAR data (in red) with `softImpute` for the M-step.
  - Method 2 by adding the mask (in green) with the matrix completion `softImpute` algorithm and `mimi` which takes into account the binary type of the mask.

Imputation performances
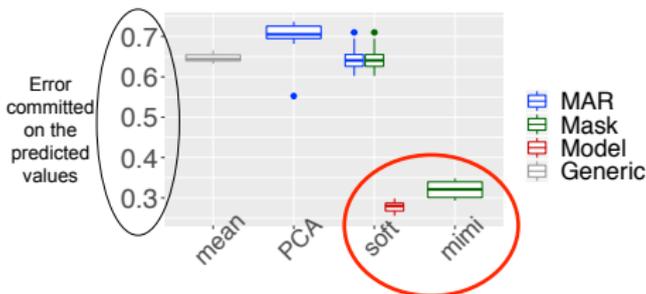
Perspectives

- 2 solutions with drawbacks (either computational or theoretical)
- Modelling the mechanism is costly.

Q: *Is there a solution for dealing with missing data in low-rank models, without modelling the mechanism and theoretically sound* ?

Graphical representation in a low-rank model?

Available-case analysis without modeling the missing-data mechanism, specifically in the linear regression [Mohan et al., 2018].

$$X^{\mathrm{NA}} = \begin{pmatrix} 12 & 28 & \mathrm{NA} \\ 23 & \mathrm{NA} & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \mathrm{NA} & 3 & 7 \end{pmatrix}, X^{AC} = \begin{pmatrix} 12 & 28 & \mathrm{NA} \\ 23 & \mathrm{NA} & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \mathrm{NA} & 3 & 7 \end{pmatrix}$$

Outline

Probabilistic Principal Component Analysis

- $X_{i.} = \alpha + B^T W_{i.}^T + \epsilon_{i.}$ with
  - the median effect: $\alpha \in \mathbb{R}^d$,
  - the loading matrix: $B \in \mathbb{R}^{r \times d}$ with a rank $r < \min\{n, d\}$,
  - the r latent variables $W_{i.} \sim \mathcal{N}(0_r, \mathrm{Id}_{r \times r})$,
  - the noise term $\epsilon_{i.} \sim \mathcal{N}(0_d, \sigma^2 \mathrm{Id}_{d \times d})$.
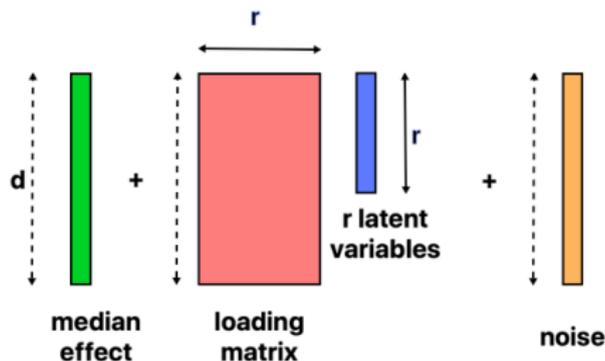
$$\Rightarrow X_{i.} \sim \mathcal{N}(\alpha, \Sigma), \ \Sigma = B^T B + \sigma^2 \mathrm{Id}_{d \times d}$$

- X contains several MNAR variables.

How to estimate α, Σ and B? How to impute missing values?

Estimation and Imputation in Probabilistic Principal Component Analysis with Missing Not At Random Data (NeurIPS 2020), Aude Sportisse, Claire Boyer, Julie Josse

Theoretical results

Assumptions for identifiability and consistency results:

- The mechanism of any MNAR variable $X_{\cdot m}$ can depend on all the variables except r called **the pivot variables**.
- The pivot variables are MCAR or observed.
- The missing-data patterns are independent given the data:

$$\forall (k, \ell) \in \{1, \ldots, d\}, k \neq \ell, \ M_{\cdot k} \perp\!\!\!\perp M_{:\ell} | Y$$
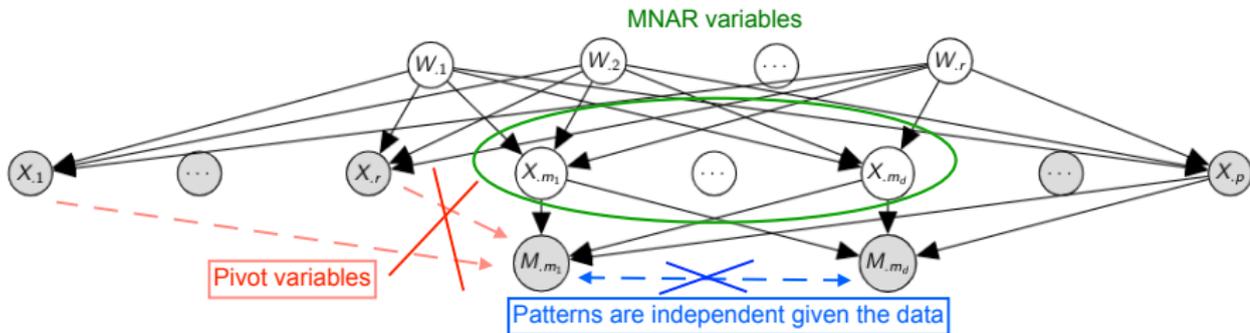
Theoretical results

Assumptions for identifiability and consistency results:

- The mechanism of any MNAR variable $X_{.m}$ can depend on all the variables except r called **the pivot variables**.
- The pivot variables are MCAR or observed.
- The missing-data patterns are independent given the data:

$$\forall (k, \ell) \in \{1, \ldots, d\}, k \neq \ell, \, M_{.k} \perp\!\!\!\perp M_{.\ell} | Y$$

- **Only for identifiability**: The MNAR variables are **self-masked**.

Proposition 1: identifiability [S., Boyer, Josse 2020]

- (α, Σ) are identifiable.
- the missing mechanism parameters are identifiable.
- B is identifiable up to a row permutation.

Toy example

- $d = 3$, $r = 2$.

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

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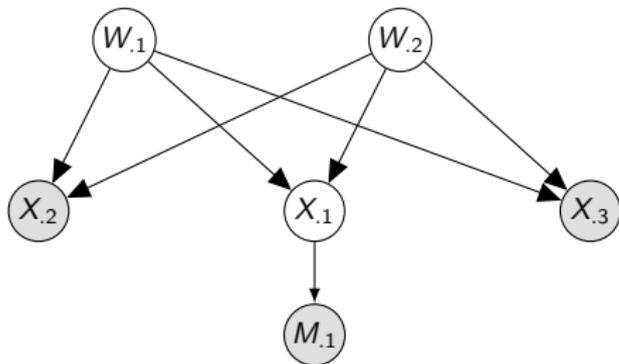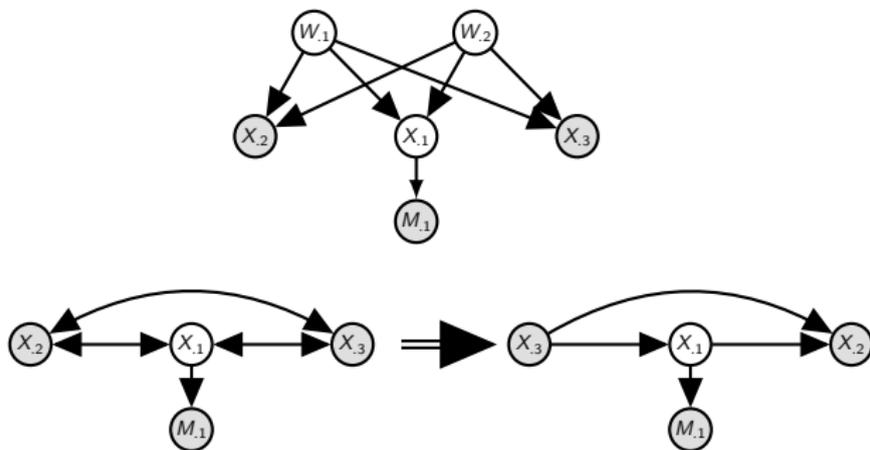
Toy example

- $d = 3$, $r = 2$.

$$\begin{pmatrix} X_{.1} & X_{.2} & X_{.3} \end{pmatrix} = 1 \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 \end{pmatrix} + \begin{pmatrix} W_{.1} & W_{.2} \end{pmatrix} B + \epsilon$$
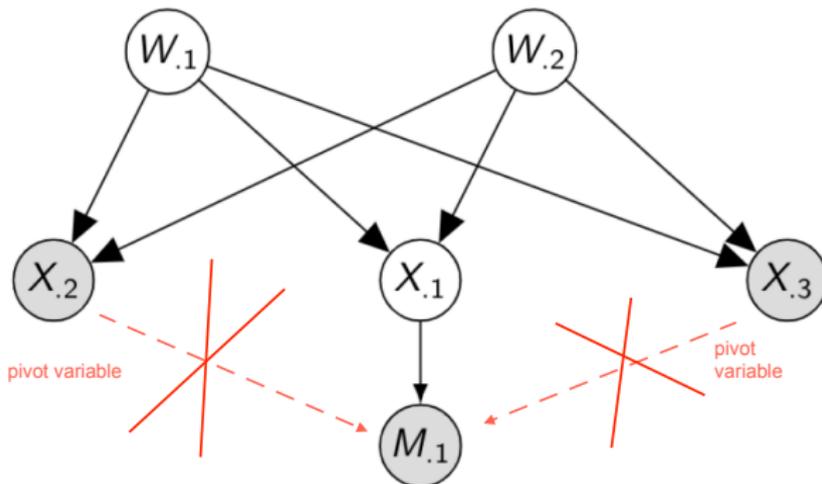
- $X_{.1}$ is MNAR (self-masked in this case).
- As $r = 2$, it requires two pivot variables, say $X_{.2}$ and $X_{.3}$ which are independent of the missing-data pattern $M_{.1}$.

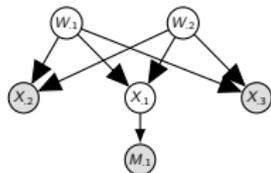Graphical model for **"fully-connected"** PPCA model
any variable is generated by all the latent variables.

Mean estimation

We can exploit the link between the variables.

$$X_{\cdot 2} = \underbrace{\mathcal{B}_{2 \to 1,3[0]}}_{\text{Mean effect of } X_{\cdot 2} \text{ on } X_{\cdot 1} \text{ and } X_{\cdot 3}} + \overbrace{\mathcal{B}_{2 \to 1,3[1]}}^{\text{Effect of } X_{\cdot 2} \text{ on } X_{\cdot 1}} X_{\cdot 1} + \underbrace{\mathcal{B}_{2 \to 1,3[3]}}_{\text{Effect of } X_{\cdot 2} \text{ on } X_{\cdot 3}} X_{\cdot 3} + \overbrace{\zeta}^{\text{noise}}$$

It is a linear **approximation**: $\mathbb{E}[\zeta | X_{\cdot 1}, X_{\cdot 3}] \neq 0$.

Mean estimation

Effects of $X_{.2}$ on $X_{.1}$ and $X_{.3}$ in the complete case when $M_{.1} = 0$:

$$(X_{.2})_{|M_{.1}=0} := \mathcal{B}^{c}_{2\to1,3[0]} + \mathcal{B}^{c}_{2\to1,3[1]}X_{.1} + \mathcal{B}^{c}_{2\to1,3[2]}X_{.3} + \zeta^{c},$$

As $X_{.2} \perp\!\!\!\perp M_{.1}|X_{.1}, X_{.3}$, one has

$$\mathbb{E}\left[X_{.2}|X_{.1}, X_{.3}, M_{.1} = 0\right] = \mathbb{E}\left[\mathcal{B}^{c}_{2\to1,3[0]} + \mathcal{B}^{c}_{2\to1,3[1]}X_{.1} + \mathcal{B}^{c}_{2\to1,3[3]}X_{.3}|X_{.1}, X_{.3}\right].$$

Taking the expectation,

$$\mathbb{E}\left[X_{.2}\right] = \mathcal{B}^{c}_{2\to1,3[0]} + \mathcal{B}^{c}_{2\to1,3[1]}\mathbb{E}\left[X_{.1}\right] + \mathcal{B}^{c}_{2\to1,3[3]}\mathbb{E}\left[X_{.3}\right].$$

Mean formula

$$\alpha_1 = \frac{\alpha_2 - \mathcal{B}^{c}_{2\to1,3[0]} - \mathcal{B}^{c}_{2\to1,3[3]}\alpha_3}{\mathcal{B}^{c}_{2\to1,3[1]}},$$

given that $\mathcal{B}^{c}_{2\to1,3[1]} \neq 0$.

Consistency results

Natural estimator for the mean:

$$\hat{\alpha}_1 := \frac{\hat{\alpha}_2 - \hat{\mathcal{B}}^c_{2\to1,3[0]} - \hat{\mathcal{B}}^c_{2\to1,3[3]}\hat{\alpha}_3}{\hat{\mathcal{B}}^c_{2\to1,3[1]}}.$$

Consistency for the mean of X_1

Assume that:

- There exist consistent estimators for α_2 and α_3.
- There exist consistent estimators for $\mathcal{B}^c_{2\to1,3[0]}$, $\mathcal{B}^c_{2\to1,3[1]}$ and $\mathcal{B}^c_{2\to1,3[3]}$.

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

Estimation in practice

Definition of a mean estimator:

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

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

$$X = \begin{pmatrix} X_{.1} & X_{.2} & X_{.3} \\ 12 & 28 & 31 \\ \text{NA} & 23 & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \text{NA} & 3 & 7 \end{pmatrix}$$

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

$$X = \begin{pmatrix} X_{.1} & X_{.2} & X_{.3} \\ 12 & 28 & 31 \\ \text{NA} & 23 & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \text{NA} & 3 & 7 \end{pmatrix}$$

Estimation of the loading matrix B

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

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

.
- Assuming that σ^2 is known,

$$X \sim \mathcal{N}\left(\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}, B^T B + \sigma^2 \mathrm{Id} \right) \Rightarrow \hat{\Sigma} - \sigma^2 \mathrm{Id}_{3\times3} \text{ estimates } B^T B.$$

- Singular value decomposition:

$$\hat{\Sigma} - \sigma^2 \mathrm{Id}_{3\times3} =: \hat{U}\hat{D}\hat{U}^T, \text{ with } \hat{U} = (\hat{u}_1|\hat{u}_2|\hat{u}_3).$$

- Assuming that $r = 2$,

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

Imputation of the missing values in X

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

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

The methodology is extended to the general case

for any continuous data

with p covariates, r latent variables and d missing variables.
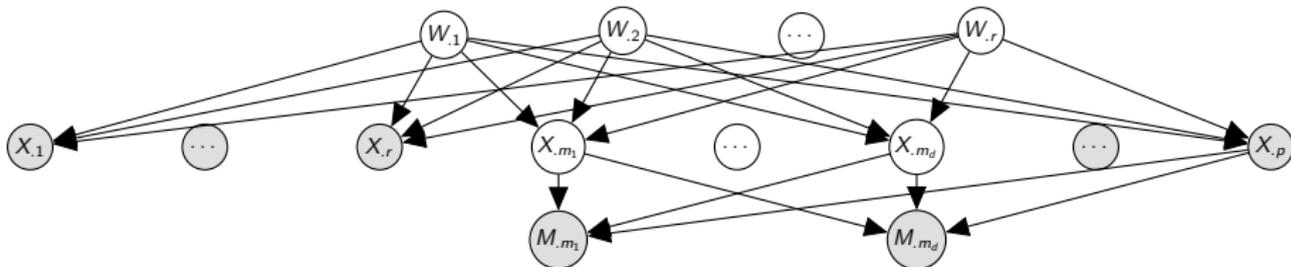
Results on real data

- $\simeq$ 3200 patients with brain trauma injury, 9 quantitative variables containing missing values are selected by doctors.
- Comparison with:
  - EMMAR: EM algorithm to perform PPCA with MAR data.
  - SoftMAR: matrix completion algorithm for MAR data, softImpute.
  - MNARparam: our method for low-rank models with fixed effect.
  - Mean: imputation by the mean.

Imputation performances
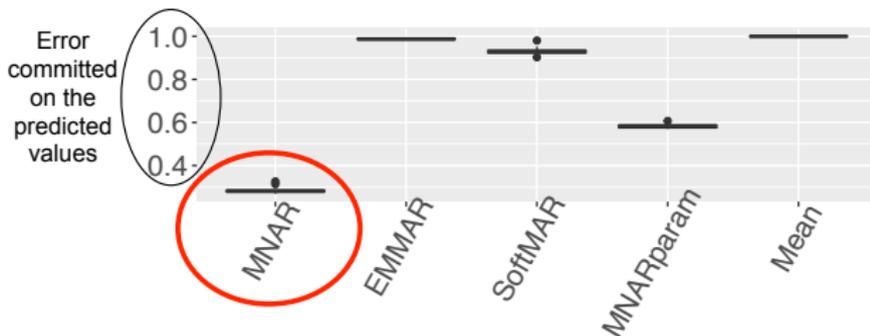
Outline

Linear regression model

Context
- **Large-scaling:** large number of observations, large d.
- **Online-setting**: the data come as it goes along.

- $(X_{i:}, y_i)_{i \geq 1} \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. observations

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

  parametrized by $\beta^\star \in \mathbb{R}^d$, with a noise term $\epsilon_i \in \mathbb{R}$.

- **Heterogeneous MCAR setting**: different missing probability for each covariate.

How to estimate $\beta^\star$?

Debiasing Stochastic Gradient Descent to handle missing values (NeurIPS 2020), Aude Sportisse, Claire Boyer, Aymeric Dieuleveut, Julie Josse

Stochastic Gradient Descent algorithm

Without missing values:

Optimization problem

- For $y_i = X_{i:}^T \beta^\star + \epsilon_i$, loss function: $f_i(\beta) = (\langle X_{i:}, \beta \rangle - y_i)^2 / 2$.
- **True** risk minimization:
$$\beta^\star = \operatorname{argmin}_{\beta \in \mathbb{R}^d} \left\{ R(\beta) := \mathbb{E}_{(X_{i:}, y_i)} \left[f_i(\beta) \right] \right\}$$

- **SGD**: using unbiased estimates of $\nabla R(\beta_{k-1})$.
$$\beta_k = \beta_{k-1} - \alpha g_k(\beta_{k-1})$$
where α is the step-size and $g_k(\beta_{k-1}) = \nabla f_k(\beta_{k-1})$.
$$\mathbb{E}\left[g_k(\beta_{k-1}) | \sigma(X_{1:}, y_1, \ldots, X_{k-1:}, y_{k-1}) \right] = \nabla R(\beta_{k-1}),$$

- **Averaged SGD**: using the Polyak-Ruppert averaged iterates.
$$\bar{\beta}_k = \frac{1}{k+1} \sum_{i=0}^{k} \beta_i$$

Large-data scaling and optimal convergence rate of $\mathcal{O}(k^{-1})$.

[Bach and Moulines, 2013]

Debiasing the gradient

With missing values:

Online-streaming: for a new observation $(X_{k:}^{\mathrm{NA}}, y_k)$

- **Imputing the missing values by 0.**

$$\tilde{X}_{k:} = X_{k:} \odot (1 - M_{k:}) \text{ imputed covariates}$$

- Using a **debiased gradient** for the **averaged SGD**:
  Find $\tilde{g}_k(\beta_k)$ such that

  $$\mathbb{E}\left[\tilde{g}_k(\beta_{k-1}) \mid \sigma(X_{1:}, y_1, M_{.1} \ldots, X_{k-1:}, y_{k-1}, M_{.k-1})\right] = \nabla R(\beta_{k-1})$$

Debiasing the gradient

Algorithm 1 Averaged SGD for Heterogeneous Missing Data

Input: data $\tilde{X}, y, \alpha$ (step size)

Initialize $\beta_0 = 0_d$.

Set $P = \text{diag} \left((p_j)_{j \in \{1, \ldots, d\}} \right) \in \mathbb{R}^{d \times d}$.

for $k = 1$ **to** n **do**

$\quad \tilde{g}_k(\beta_{k-1}) = P^{-1} \tilde{X}_{k:} \left(\tilde{X}_{k:}^T P^{-1} \beta_{k-1} - y_k \right) - (I - P) P^{-2} \text{diag} \left(\tilde{X}_{k:} \tilde{X}_{k:}^T \right) \beta_{k-1}$

$\quad \beta_k = \beta_{k-1} - \alpha \tilde{g}_k(\beta_{k-1})$

$\quad \bar{\beta}_k = \frac{1}{k+1} \sum_{i=0}^{k} \beta_i = \frac{k}{k+1} \bar{\beta}_{k-1} + \frac{1}{k+1} \beta_k$

end for

- $p = 1 \Rightarrow P^{-1} = I_d$ standard least squares stochastic algorithm.

- Computation cost for the gradient still low.

- Trivially extended to ridge regularization (no change for the gradient): $\min_{\beta \in \mathbb{R}^d} R(\beta) + \lambda \|\beta\|^2, \lambda > 0$

Theoretical results

Goal: establish a convergence rate

by controlling the noise introduced by NAs

Assumptions on the data: $(X_{k:}, y_k) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d., $\mathbb{E}[\|X_{k:}\|^2]$ and $\mathbb{E}[y_k^2]$ finite, $H := \mathbb{E}_{(X_{k:}, y_k)}[X_{k:} X_{k:}^T]$ invertible.

Lemmas 2, 3 [S., Boyer, Dieuleveut, Josse, 2020]

- The noise induced by the imputation by 0 is **structured**.
- $(\tilde{g}_k(\beta^\star))_k$ are **a.s. co-coercive**.

Theoretical results

Theorem: convergence rate of $\mathcal{O}(k^{-1})$, streaming setting

Assume for any i, $\|X_{i:}\| \leq \gamma$ almost surely for some $\gamma > 0$. For **any constant step-size** $\alpha \leq \frac{1}{2L}$, our algorithm ensures that, for any $k \geq 0$:

$$\mathbb{E}\left[R\left(\bar{\beta}_k\right) - R(\beta^\star)\right] \leq \frac{2}{k}\left(\underbrace{\sqrt{c(\beta^\star)d}}_{\text{variance term}} + \underbrace{\frac{\|\beta_0 - \beta^\star\|}{\sqrt{\alpha}}}_{\text{bias term}}\right)^2,$$

- $L := \sup_{k,D}$ Lipschitz constants of $\tilde{g}_k$
- $p_m = \min_{j=1,\ldots d} p_j$ minimal probability to be observed among the variables.

- $c(\beta^\star) = \underbrace{\overbrace{\frac{\text{Var}(\epsilon_k)}{p_m^2}}^{\text{classical term}} + \overbrace{\left(\frac{7(1-p_m)}{p_m^3}\right)\gamma^2\|\beta^\star\|^2}^{\text{multiplicative noise (due to naive imputation)}}}_{\text{increasing with the missing values rate}}$.

Theoretical results

Theorem 4 [S., Boyer, Dieuleveut, Josse 2020]

Assume for any i, $\|X_{i:}\| \leq \gamma$ almost surely for some $\gamma > 0$. For **any constant step-size** $\alpha \leq \frac{1}{2L}$, our algorithm ensures that, for any $k \geq 0$:

$$\mathbb{E}\left[R\left(\bar{\beta}_k\right) - R(\beta^\star)\right] \leq \frac{2}{k}\left(\underbrace{\sqrt{c(\beta^\star)d}}_{\text{variance term}} + \underbrace{\frac{\|\beta_0 - \beta^\star\|}{\sqrt{\alpha}}}_{\text{bias term}} \right)^2,$$
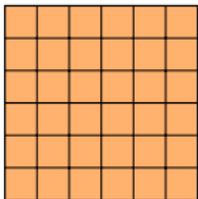
convergence rate of $\mathcal{O}(k^{-1})$

Optimal rate for least-squares regression.

Same bound as Bach and Moulines in the complete case.

What impact on missing values?

(1) Fewer complete observations is better than more incomplete ones: is it better to access 200 incomplete observations (with a probability 50% of observing) or to have 100 complete observations?
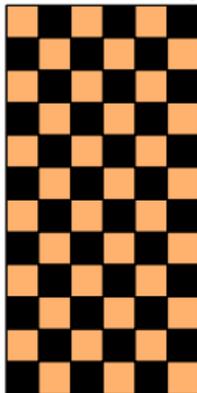
Variance bound scales as $\frac{\sigma^2 d}{kp}$ / Variance bound scales as $\frac{\sigma^2 d}{kp^2}$

The variance bound for 200 incomplete observations (with a probability 50% of observing) is twice as large as for 100 complete observations.

What impact on missing values ?

(2) We do better than discarding all observations which contain missing values:

$$X = \begin{pmatrix} X_1 & X_2 & X_3 \\ 12 & 28 & 31 \\ \text{NA} & 23 & 89 \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \text{NA} & 3 & 7 \end{pmatrix} \qquad X = \begin{pmatrix} X_1 & X_2 & X_3 \\ 12 & 28 & 31 \\ \overline{\text{NA}} & \overline{23} & \overline{89} \\ 32 & 6 & 24 \\ \vdots & \vdots & \vdots \\ \overline{\text{NA}} & \overline{3} & \overline{7} \end{pmatrix}$$

In the homogeneous case: our strategy has an

upper-bound p^{d-3} smaller than the lower bound of any algorithm relying only on the complete observations.

Results on real data

- Goal: model the level of platelet upon arrival at the hospital from the clinical data of 15785 patients.
- Explanatory variables selected by doctors: seven quantitative (missing) variables.
- Model estimation: do the effect of the variables on the platelet make sense ?
- Similar results than EM algorithm, the effects are in agreement with the doctors' opinion, except for HR and Δ.Hemo variables.

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

Outline

Mixture model-based clustering

- MNAR: model the joint distribution (X, M) as for low rank methods.
- Partition with K clusters: $Z = (Z_1 | \ldots | Z_n)^T \in \{0, 1\}^{n \times K}$
  - $Z_{ik} = 1$ if x_i belongs to cluster k.

$$f(X_i; \pi, \theta) = \sum_{k=1}^{K} \overbrace{\pi_k}^{=\mathbb{P}(Z_{ik}=1)} \underbrace{f_k(X_i; \theta_k)}_{\text{pdf in the cluster } k}$$

- We choose the selection models specification:

$$\mathbb{P}(X_i, M_i | Z_i) = \mathbb{P}(X_i | Z_i) \mathbb{P}(M_i | X_i, Z_i; \phi)$$

  - Identifiability.
  - Estimation of θ, π.

Proposed zoology of MNAR models in clustering

Conditional independence of the missing-data patterns.

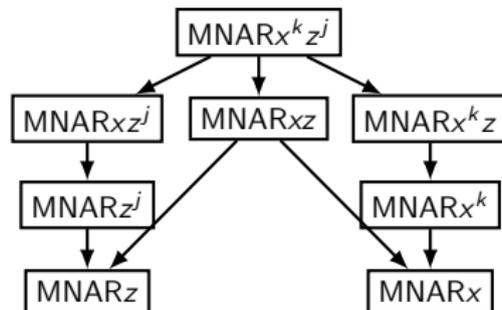$$\mathbb{P}(M_i|X_i, Z_{ik} = 1; \phi) = \prod_{j=1}^{d} \mathbb{P}(M_{ij}|X_i, Z_{ik} = 1; \phi)$$

$\boxed{\text{MNAR} x^k z^j}$ where

$$\mathbb{P}(M_{ij} = 1 \mid X_i, Z_{ik} = 1; \phi) = \rho(\phi^z{}_{kj} + \phi^x{}_{kj}X_{ij}),$$

with ρ: cdf of any continuous distribution (logit, probit)

- $\phi^z \in \mathbb{R}^{Kd}$: missingness depends on the class membership k, not the same effect for every variable.
- $\phi^x \in \mathbb{R}^{Kd}$: missingness depends on the value itself X_{ij}, not the same for each cluster.

Proposed zoology of MNAR models in clustering

- $\boxed{\text{MNAR}\,\boxed{x}}$ (self-masked): $\mathbb{P}(M_{ij} = 1 \mid X_i, Z_{ik} = 1; \phi) = \rho(\boxed{\phi^x}_j X_{ij})$.

- $\boxed{\text{MNAR}\,\boxed{z}}$: $\mathbb{P}(M_{ij} = 1 \mid X_i, Z_{ik} = 1; \phi) = \rho(\boxed{\phi^z}_k)$.
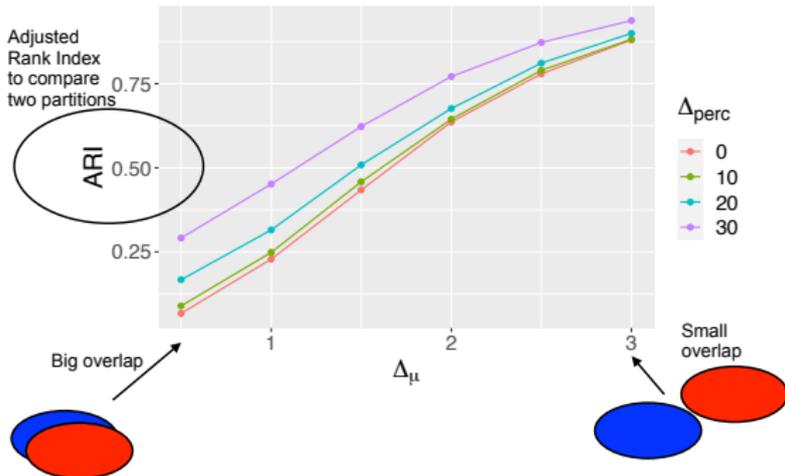
MNAR*z* from every angle

(1) M **depends on** X **through** Z

$$P(M_{ij} = 1 | X_i; \theta, \phi) = \sum_{k=1}^{K} P(M_{ij} = 1 | X_i, Z_{ik} = 1; \phi) P(Z_{ik} = 1 | X_i; \theta)$$

(2) M **gives information on partition** Z

- MNAR*z* model, Bivariate Gaussian model
- cluster overlap: $\Delta_\mu = |\mu_1 - \mu_2|$ varies.
- difference of percentage of NA between the 2 clusters: Δ_{perc} varies.

(3) MNAR*z* models interpreted as MAR

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

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

Proposition 1: in terms of maximum likelihood

MLE associated to $\tilde{X}^{\mathrm{obs}}$ under MAR model
$\Leftrightarrow$ MLE associated to X^{obs} under MNAR*z* model.

Identifiability results

Previous works: [Teicher, 1963], [Allman et al., 2009] (without NA), [Miao et al., 2016] (for MNAR data).

Proposition 2: identifiability for continuous and count data

Assume

1. The marginal mixture $\sum_{k=1}^{K} \pi_k f_k(x_i; \theta_k)$ is identifiable
2. There exists a total ordering $\preceq$ of $\mathcal{F}_j \times \mathcal{R}$, for $j \in \{1, \ldots, d\}$ fixed, where $\mathcal{F}_j = \{f_{1j}, \ldots, f_{Kj}\}$ and $\mathcal{R} = \{\rho_1, \ldots, \rho_K\}$.

The mixture model with any MNAR$*$ is identifiable.

Proposition 3: identifiability for categorical data

Assume $d_{\mathrm{cat}} \geq 2\lceil \log_2 K \rceil + 1$ and $f_k(\cdot; \theta_k) = \prod_{j=1}^{d} f_{kj}(\cdot; \theta_{kj})$
✓ The mixture model with MNARz is identifiable.
✗ The mixture model with any MNARx$*$ is not identifiable.

- For mixed data: result follows from Proposition 2 and 3.

Identifiability up to a label swapping.

EM algorithm: feasible computations ?

The **expected complete likelihood** knowing the observed data and a current value of the parameters is **decomposed into 2 parts**

$$Q(\theta, \phi, \pi; \theta^r, \phi^r, \pi^r) = \mathbb{E}[L_{\mathsf{full}}(\theta, \phi, \pi; X, Z, M)|X_i^{\mathrm{obs}}, M_i; \theta^r, \phi^r, \pi^r]$$

$\boxed{\text{MNAR}z}$: needs some computations but still simple.

$$\mathbb{P}(M_{ij} = 1 \mid X_i, Z_{ik} = 1; \phi) = \rho(\alpha_k) \qquad (\perp\!\!\!\perp X)$$

- EM algorithm for Gaussian data,
- EM for categorical data.

$\boxed{\text{MNAR}x*}$: needs approximations

$$\mathbb{P}(M_{ij} = 1 \mid X_i, Z_{ik} = 1; \phi) = \rho(\alpha_{kj} + \beta_{kj}X_{ij}) \qquad (\mathrm{not} \perp\!\!\!\perp X)$$

- $(x_i^{\mathrm{mis}} \mid x_i^{\mathrm{obs}}, z_{ik} = 1, M_i)$ not classical if Logit link.
- No closed forms.

SEM algorithm for MNAR$x*$

SEM easier? random drawing instead of expectation

- **SE-step:** draw the missing data

$$((X_i^{\mathrm{mis}})^{r+1}, Z_i^{r+1}) \sim (. \mid X_i^{\mathrm{obs}}, M_i; \theta^r, \phi^r, \pi^r)$$

- **M-step:** for $k = 1, \ldots, K$, compute $\pi_k^{r+1}, \mu_k^{r+1}, \Sigma_k^{r+1}, \phi^{r+1}$.
- Use of One-Gibbs and Probit link for the SE-step.

SEM algorithm for MNAR$x*$

SEM easier? random drawing instead of expectation

- **SE-step:** draw the missing data

$$((X_i^{\mathrm{mis}})^{r+1}, Z_i^{r+1}) \sim (. \mid X_i^{\mathrm{obs}}, M_i; \theta^r, \phi^r, \pi^r)$$

- **M-step:** for $k = 1, \ldots, K$, compute $\pi_k^{r+1}, \mu_k^{r+1}, \Sigma_k^{r+1}, \phi^{r+1}$.

- Use of One-Gibbs and Probit link for the SE-step.

	EM		SEM	
	Gaussian	Categorical	Gaussian	Categorical
MNARz MNARz^j	✓	✓	✓	✓
MNAR$x*$	no closed form	not ident.	✓ (Probit)	not ident.

Results on real data

- 41 mixed variables containing missing values assumed to be MNARz.
- Cluster the patients into 3 groups.
- Representation with FactoMineR [Husson et al., 2016].
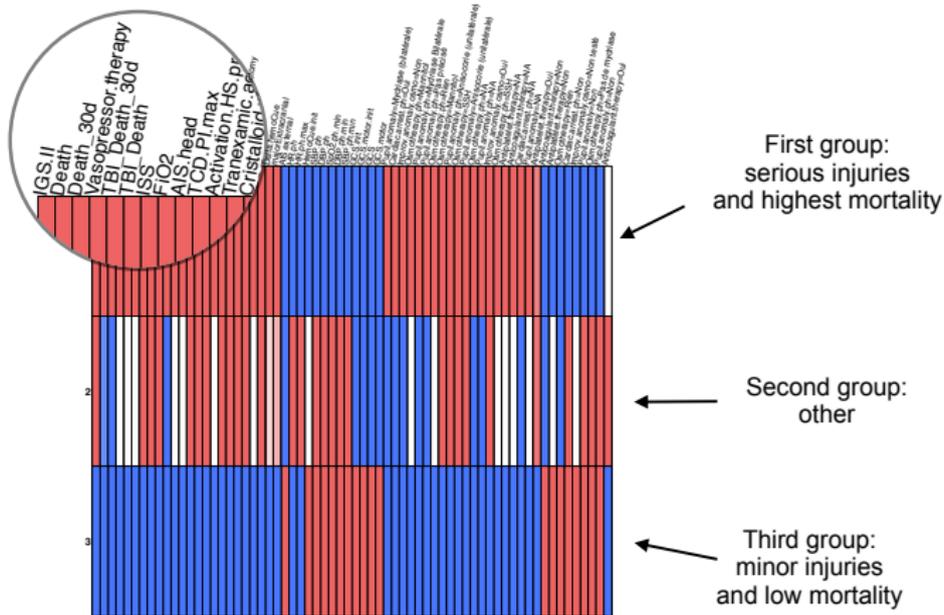
Same criteria as the groups made by the doctors.

Visit our website !

Imke Mayer, Julie Josse, Nicholas Tierney and Nathalie Vialaneix and **many other contributors**

Conclusion

- Goal: propose methods to handle heterogeneous and not-MCAR missing data **motivated by real-world problems**.

	Mechanism	Data type
Low rank model with fixed effect	self-masked MNAR	continuous
Low rank model with random effect	MNAR	continuous
Online linear regression SGD	heterogeneous MCAR	mixed
Model-based clustering	MNAR	mixed

Future work

- Put the methods into production: better implementations, R-packages, methods to automate the choice of hyperparameters.
- Semi-supervised models.

List of publications

- Imputation and low-rank estimation with Missing Not At Random data, A. Sportisse, C. Boyer, J. Josse, *Statistics & Computing, Springer*, 2020
- Estimation and Imputation in Probabilistic Principal Component Analysis with Missing Not At Random Data, A. Sportisse, C. Boyer, J. Josse, *Advances in Neural Information Processing Systems*, 2020
- Debiasing Stochastic Gradient Descent to handle missing values, A. Sportisse, C. Boyer, A. Dieuleveut, J. Josse, *Advances in Neural Information Processing Systems*, 2020

Submitted paper

- Robust Lasso-Zero for sparse corruption and model selection with missing covariates, led by Pascaline Descloux, and in collaboration with Claire Boyer, Julie Josse and Sylvain Sardy (submitted in 2020, in review)

Ongoing works

- Model-based Clustering with Missing Not At Random Data, initiated by Christophe Biernacki, Gilles Celeux, Julie Josse, Fabien Laporte, and reworked with Christophe Biernacki, Claire Boyer, Julie Josse and Matthieu Marbac
- R-miss-tastic: a unified platform for missing values methods and workflows, led by Imke Mayer, and in collaboration with Julie Josse, Nicholas Tierney, Nathalie Vialaneix

Thanks for your attention!

References I

Allman, E. S., Matias, C., Rhodes, J. A., et al. (2009).
Identifiability of parameters in latent structure models with many
observed variables.
The Annals of Statistics, 37(6A):3099–3132.

Bach, F. and Moulines, E. (2013).
Non-strongly-convex smooth stochastic approximation with
convergence rate o $(1/n)$.
In *Advances in neural information processing systems*, pages
773–781.

Buuren, S. v. and Groothuis-Oudshoorn, K. (2010).
mice: Multivariate imputation by chained equations in r.
Journal of statistical software, pages 1–68.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977).
Maximum likelihood from incomplete data via the em algorithm.
Journal of the royal statistical society. Series B (methodological),
pages 1–38.

References II

Hastie, T. and Mazumder, R. (2015).
softImpute: Matrix Completion via Iterative Soft-Thresholded SVD.
R package version 1.4.

Hastie, T., Mazumder, R., Lee, J. D., and Zadeh, R. (2015).
Matrix completion and low-rank svd via fast alternating least squares.
The Journal of Machine Learning Research, 16(1):3367–3402.

Heckman, J. J. (1979).
Sample selection bias as a specification error.
Econometrica: Journal of the econometric society, pages 153–161.

Honaker, J., King, G., Blackwell, M., et al. (2011).
Amelia ii: A program for missing data.
Journal of statistical software, 45(7):1–47.

Husson, F., Josse, J., Le, S., Mazet, J., and Husson, M. F. (2016).
Package 'factominer'.
An R package, 96:698.

References III

Ibrahim, J. G. (1990).
Incomplete data in generalized linear models.
Journal of the American Statistical Association, 85(411):765–769.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (1999).
Monte carlo em for missing covariates in parametric regression models.
Biometrics, 55(2):591–596.

Josse, J. and Husson, F. (2012).
Selecting the number of components in principal component analysis using cross-validation approximations.
Computational Statistics & Data Analysis, 56(6):1869–1879.

Josse, J., Husson, F., et al. (2016a).
missmda: a package for handling missing values in multivariate data analysis.
Journal of Statistical Software, 70(1):1–31.

References IV

Josse, J., Prost, N., Scornet, E., and Varoquaux, G. (2019).
On the consistency of supervised learning with missing values.
arXiv preprint arXiv:1902.06931.

Josse, J., Sardy, S., and Wager, S. (2016b).
denoiser: A package for low rank matrix estimation.
Journal of Statistical Software.

Little, R. J. (1993).
Pattern-mixture models for multivariate incomplete data.
Journal of the American Statistical Association, 88(421):125–134.

Loh, P.-L. and Wainwright, M. J. (2011).
High-dimensional regression with noisy and missing data: Provable
guarantees with non-convexity.
In *Advances in Neural Information Processing Systems*, pages
2726–2734.

References V

Ma, A. and Needell, D. (2018).
Stochastic gradient descent for linear systems with missing data.
Numerical Mathematics: Theory, Methods and Applications,
12(1):1–20.

Mattei, P.-A. and Frellsen, J. (2019).
Miwae: Deep generative modelling and imputation of incomplete
data sets.
In *International Conference on Machine Learning*, pages 4413–4423.
PMLR.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010).
Spectral regularization algorithms for learning large incomplete
matrices.
The Journal of Machine Learning Research, 11:2287–2322.

References VI

Miao, W., Ding, P., and Geng, Z. (2016).
Identifiability of normal and normal mixture models with nonignorable missing data.
Journal of the American Statistical Association, 111(516):1673–1683.

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

Robin, G., Klopp, O., Josse, J., Moulines, É., and Tibshirani, R. (2020).
Main effects and interactions in mixed and incomplete data frames.
Journal of the American Statistical Association, 115(531):1292–1303.

Tang, N. and Ju, Y. (2018).
Statistical inference for nonignorable missing-data problems: a selective review.
Statistical Theory and Related Fields, 2(2):105–133.

References VII

Teicher, H. (1963).
Identifiability of finite mixtures.
The annals of Mathematical statistics, pages 1265–1269.

Classical methods

The most popular

- **Mean imputation** :
  - Bad for estimation and imputation.
  - Good for prediction [Josse et al., 2019].

- **Model-based methods** : model for $(X^{\mathrm{obs}}, X^{\mathrm{mis}})$ or $(X^{\mathrm{obs}}|X^{\mathrm{mis}})$ e.g. for Gaussian data (`Amelia`), or nonparametric (`missForest`,`MIWAE`)

  [Honaker et al., 2011, **?**, Mattei and Frellsen, 2019]

- **Low-rank methods** : (`softImpute`, `imputePCA`)[Hastie et al., 2015, Josse et al., 2016a]
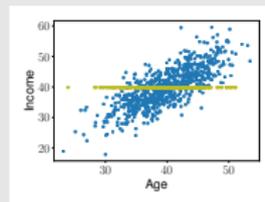
Multiple imputation to **reflect the variability**
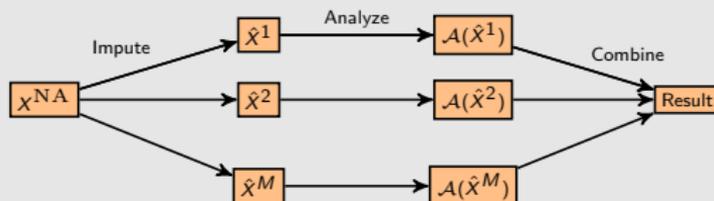
(`mice`) [Buuren and Groothuis-Oudshoorn, 2010]

Classical methods

EM algorithm written for M(C)AR data

- Estimate the parameter θ by modifying the estimation process.
- Particularly adapted for Gaussian data.

[Dempster et al., 1977, Ibrahim, 1990]

Naive imputation + debiasing

Goal: apply an algorithm A to the case with missing values.

- Naively impute the missing values, get $\tilde{X}$,
- Adapt algorithm A to account for the error and apply this debiased version to the complete dataset $\tilde{X}$.

For Lasso, SGD [Loh and Wainwright, 2011, Ma and Needell, 2018]

Classical methods

Method	Simple to implement	Imputation	Confidence intervals	Main drawbacks
Single imputation	✓	single	✗	biased estimates if too simple imputation
Multiple imputation	✓	multiple	✓	combining results can be delicate
EM	✗	not directly	can be obtained	specific algorithm for each statistical model
Naive imp. + debiasing	✓	not the goal	✗	debiasing each algorithm
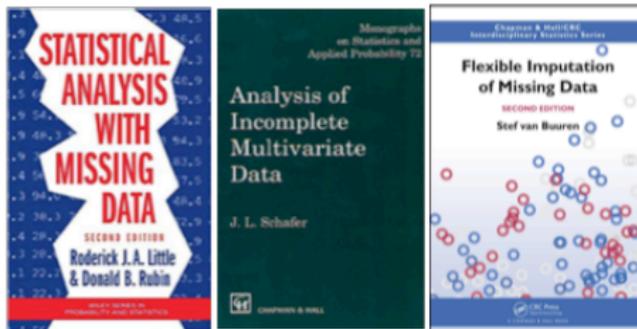
How to generate missing values?

- Why? For numerical experiments!
- Ambiguity on the missing-data mechanism definitions: *realised* or *everywhere* Seaman et al., 2013
- Two ways for generating M(N)AR missing values. For MAR:

Realised mechanism: the observations are not i.i.d. (not classical)

All variables can contain missing values!
We generate missing values in X_1 using a logistic model depending on the variables (X_2, X_3) (thus the missingness depends on the observed values). And do the same for X_2 and X_1.
See the implementation in R in R-miss-tastic

Everywhere mechanism: the observations are i.i.d. (more canonical)

In a dataset of 3 variables, we choose at least one variable which is **always observed**.
See the implementation in Python in R-miss-tastic

Still ambiguities to generate the missing values. Rows which contain only NA, ...

Low-rank model with fixed effects: identifiability?

- Identifiability of the parameter Θ?
- Result follows from Miao et al., Identifiability of normal and normal mixture models with nonignorable missing data, 2016?

Theorem 1 [Miao et al., 2016]

Under the following model:

- Gaussian data: $X \sim \mathcal{N}(\mu, \sigma^2)$,
- self-masked MNAR: $\mathbb{P}(M = 1 | X) = \rho(\phi_1 + \phi_2 X)$, with ρ the probit link.

We have the identifiability of the parameters $\mu, \sigma, \phi_1, \phi_2$.

- Ensure that it scales with the multidimensional case (several MNAR variables);
- In the paper, we have assumed a logit link: in this case, identifiability of the parameters if the sign of ϕ_2 is known;
- Condition: the left tail decay rate of F is not exponential, i.e. $\forall \delta > 0, \lim_{z \to -\infty} \frac{F(z)}{e^{-\delta z}} = 0 \text{ or } + \infty$.
- In practice: the logit link very closed to the probit link.

Low-rank model with fixed effects: E-step

We minimize the negative log-likelihood

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

$$C_1 = \log(f(X_{ij}, M_{ij}; \Theta_{ij}, \phi_j))$$

$$C_2 = \int \underbrace{\log(f(X_{ij}, M_{ij}; \Theta_{ij}, \phi_j))}_{\propto X_{ij}^2} \quad \underbrace{f(X_{ij} | M_{ij}; \hat{\Theta}_{ij}^{(t)}, \hat{\phi}_j^{(t)})}_{\propto \text{Gaussian distribution} \times \text{Logit distribution}} \quad dX_{ij}$$

- Consider Probit distribution? and use a latent variable (as for the clustering with MNAR data)?
- **Direct extension to the case where the entries of X are not independent? with more computations**

Low-rank model with fixed effects: Monte Carlo and SIR algorithms

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

$$v_{ij}^k = \left\{ \begin{array}{l} X_{ij} \text{ if } M_{ij} = 1, \\ z_{ij}^k \text{ otherwise,} \end{array} \right. \text{ with } z_{ij}^k \sim p\left(X_{ij}; \hat{\Theta}_{ij}^{(t)}\right) p\left(M_{ij} | X_{ij}; \hat{\phi}_j^{(t)}\right) = g(X_{ij}).$$

How to draw z_{ij}^k?

Algorithm 2 SIR

Sampling: a sample $x_1, \ldots, x_M \sim \mathcal{N}(\Theta_{ij}^{(t)}, \sigma^2)$.
Importance: compute the weights

$$\omega(x_m) = \frac{g(x_m)}{\varphi_{\Theta_{ij}^{(t)}, \sigma^2}(x_m)}, \text{ for } m = 1, \ldots, M,$$

with φ the density function of a Gaussian variable.
Resampling: draw z from the original sample $x_1, \ldots, x_M$ with probability proportional to $\omega(x_1), \ldots, \omega(x_M)$.

Low-rank model with fixed effects: computational aspects

Computational complexity of the algorithms

For 1 iteration and oracle parameter tuning

softImpute (SVD)	Our method 1 by modelling MNAR data
$\mathcal{O}\left((1-p_{\mathrm{NA}})ndr\right)$ [Mazumder et al., 2010]	$\mathcal{O}\left(\underbrace{N_{\mathrm{SIR}}p_{\mathrm{NA}}nd}_{E-step} + \underbrace{(1-p_{\mathrm{NA}})ndr}_{\texttt{softImpute}} + \underbrace{d^3 + nd^2}_{\mathrm{GLM}}\right)$

- p_{NA}: proportion of missing values
- r: rank of the low-rank matrix.
- N_{SIR}: number of SIR drawings.
- For us: complexity of GLM is problematic.
- In practice: if N_{SIR} is a big constant...
- Solution: implementation in C? Alternative algorithm for the E-step?

Low rank models: hyperparameters

- **Noise level:** use the residual sum of squares divided by the number of observations minus the number of estimated parameters as suggested by [Josse et al., 2016b], in complete case

$$\hat{\sigma}^2 = \frac{\|X - \sum_{l=1}^{r} u_l d_l v_l\|_2^2}{nd - nr - rd + r^2},$$

where u_l, v_l and d_l are the singular vectors and the singular values from the SVD of X.

- **Rank of X, r:** use a cross-validation for M(C)AR data [Josse and Husson, 2012].
- Regularization parameter (for the fixed effects): cross-validation for M(C)AR data.

Issue for the cross-validation with MNAR data

We want to choose λ the regularization parameter in

$$\hat{\Theta} \in \mathrm{argmin}_{\Theta} \|(1 - M) \odot (X - \Theta)\|^2 + \lambda \|\Theta\|_{\star},$$

- Gridsearch for λ: $[\lambda_1, \ldots, \lambda_L]$.
- For λ_l (do this several times for the same λ)
    - Introduce new missing data in X
    - Split your dataset in 2 datasets: $X^{(1)}$ and $X^{(2)}$.
    - Apply `softImpute` on $X^{(1)}$ with λ_1 and get $\hat{\Theta}^{(1)}$.
    - Impute $X^{(2)}$ with $\hat{\Theta}^{(1)}$ and compute the imputation error on $X^{(2)}$.

- It is costly.

- For M(N)AR data : introduce missing values which have the same mechanism than the true missing values is not an easy task.

Prediction task for the Traumabase dataset

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

By using random forest for the classification. Error corresponds to the validation error. AUC is the area under ROC; the accuracy (acc) is the number of true positive plus true negative divided by the total number of observations; the sensitivity (sen) is defined as the true positive rate; specificity (spe) as the true negative rate; the precision (pre) is the number of true positive over all positive predictions.

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

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

PPCA: the assumptions in practice

Jester dataset: 5000 users who rated jokes, with 27% of missing values.

A neutron walks into a bar and orders a drink. "How much do I owe you?" the neutron asks. The bartender replies, "for you, no charge."

1. Fully PPCA model: any user preference (variable) can be expressed as a linear combination of latent variables. The first latent variable opposes individuals who like jokes about physics but dislike jokes about sexuality, and conversely.

2. Mechanism assumption:
   - self-masked MNAR: users only rate jokes they like or dislike strongly or might be ashamed to assume their taste for sexual jokes.
   - pivot variables: a user's non-response for the sexual joke given all jokes may depend on the scores of the sexual and physical jokes but not on the scores of the musical and computer jokes.

3. How to select the r pivot variables? (MCAR or observed)
   - Naive solution: variables with the lowest missing rate.
   - Discuss with experts.
   - Select a bigger set and computing the final estimator with the median of the estimators over all possible combinations (costly). Cross-validation? (costly)

PPCA: no exogeneity

$$X_{.2} = \mathcal{B}_{2\to1,3[0]} + \mathcal{B}_{2\to1,3[1]}X_{.1} + \mathcal{B}_{2\to1,3[3]}X_{.3} + \zeta$$

$\mathbb{E}[\zeta|X_{.1}, X_{.3}] \neq 0$: the linear regression of $X_{.2}$ on $X_{.1}, X_{.3}$ gives biased estimates.

- In practice: it works well (simulations for different noise levels).
- How to handle a high noise level? Estimate the coefficients with other methods than linear regression.
- Instrumental variable regression (used for example in econometrics).
- The covariables are split in two parts:
  - one part which is not correlated to ζ (it is called the instrumental variable, which has to be correlated with the covariables),
  - one part which is correlated to $\zeta \to$ new noise.

Low rank methods: computational cost

Method	$r = 2, p = 10, n = 1000$ 35% MNAR values in 7 variables	$r = 5, p = 50, n = 1000$ 20% MNAR values in 20 variables
MNAR algebraic	0,1 s	11 min 48 s (1260 aggregations)
SoftMAR	5,5 s	28 s
EMMAR	50,8 s	2 min 9 s
Param	5 h 15 min	not evaluated

SGD: how to debiase the gradient

Our strategy

Online-streaming: for a new observation $(X_{k:}^{\mathrm{NA}}, y_k)$

- **Imputing the missing values by 0**.

$$\tilde{X}_{k:} = X_{k:}^{\mathrm{NA}} \odot M_{k:} = X_{k:} \odot M_{k:} \text{ imputed covariates}$$

- Using a **debiased gradient** for the **averaged SGD**:
  Find $\tilde{g}_k(\beta_k)$ such that $\mathbb{E}\left[\tilde{g}_k(\beta_{k-1}) \,|\, \mathcal{F}_{k-1}\right] = \nabla R(\beta_{k-1})$

- $\mathcal{F}_{k-1} = \sigma(X_{1:}, y_1, M_{1:} \ldots, X_{k-1:}, y_{k-1}, M_{k-1:})$
- $\nabla R(\beta_{k-1}) = \mathbb{E}_{(X_{k:}, y_k)}[X_{k:}(X_{k:}^T \beta_{k-1} - y_k)]$
- No access to $X_{k:}$, only to $\tilde{X}_{k:}$.
- Another source of randomness: $\mathbb{E} = \mathbb{E}_{(X_{k:}, y_k), M_{k:}} \overset{\mathrm{indep}}{=} \mathbb{E}_{(X_{k:}, y_k)} \mathbb{E}_{M_{k:}}$
- $\mathbb{E}_{M_{k:}} | \mathcal{F}_{k-1} \rightsquigarrow \mathbb{E}_{M_{k:}}$
  - Mask at step k independent from the previous constructed iterate.

SGD: how to debiase the gradient

$$\mathbb{E}_{M_{k:}}\left[\tilde{X}_{k:}\right] = \mathbb{E}_{M_{k:}}\left[\begin{pmatrix}\delta_{k1}X_{k1}\\ \vdots \\ \delta_{kd}X_{kd}\end{pmatrix}\right] = \begin{pmatrix}p_1 X_{k1}\\ \vdots \\ p_d X_{kd}\end{pmatrix}$$

Thus

$$\mathbb{E}_{M_{k:}}\left[P^{-1}\tilde{X}_{k:}\right] := \begin{pmatrix}p_1^{-1} & & \\ & \ddots & \\ & & p_d^{-1}\end{pmatrix}\begin{pmatrix}p_1 X_{k1}\\ \vdots \\ p_d X_{kd}\end{pmatrix} = X_{k:}$$

One obtains

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

SGD: technical lemmas

- Goal: establish a convergence rate.
- Assumptions on the data: $(X_{k:}, y_k) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d., $\mathbb{E}[\|X_{k:}\|^2]$ and $\mathbb{E}[y_k^2]$ finite, $H := \mathbb{E}_{(X_{k:}, y_k)}[X_{k:} X_{k:}^T]$ invertible.

Lemma: noise induced by the imputation by 0 is **structured**

$(\tilde{g}_k(\beta^\star))_k$ with $\beta^\star$ is $\mathcal{F}_k$−measurable and $\forall k \geq 0$,
- $\mathbb{E}[\tilde{g}_k(\beta^\star) \,|\, \mathcal{F}_{k-1}] = 0$ a.s.
- $\mathbb{E}[\|\tilde{g}_k(\beta^\star)\|^2 \,|\, \mathcal{F}_{k-1}]$ is a.s. finite.
- $\mathbb{E}[\tilde{g}_k(\beta^\star)\tilde{g}_k(\beta^\star)^T] \preccurlyeq C(\beta^\star) = c(\beta^\star)H$.

Lemma: $(\tilde{g}_k(\beta^\star))_k$ are **a.s. co-coercive**

For any k,
- $\tilde{g}_k$ is $L_{k,D}$-Lipschitz
- there exists a random primitive function $\tilde{f}_k$ which is a.s. convex

SGD: what impact of missing values ?

We do better than discarding all observations which contain missing values: Example in the homogeneous case with p the proportion of being observed.

- keeping only the complete observations, any algorithm:
  - number of complete observations $k_{co} \sim \mathcal{B}(k, p^d)$.
  - statistical lower bound: $\frac{\mathrm{Var}(\epsilon_k)d}{k_{co}}$.
  - in expectation, lower bound on the risk larger than $\frac{\mathrm{Var}(\epsilon_k)d}{kp^d}$.
- keeping all the observations, averaged SGD: upper bound
  $O\left(\frac{\mathrm{Var}(\epsilon_k)d}{kp^2} + \frac{C(X,\beta^\star)}{kp^3}\right)$.

Our strategy has an upper-bound p^{d-3} smaller than the lower bound of any algorithm relying only on the complete observations.

SGD: no result for empirical risk

Finite-sample setting: n is fixed

- **True risk**: same convergence rate holds for **only one epoch** (we can use only once each data).
  Otherwise: mask at step k independent from the previous constructed iterate $\Rightarrow$ bias in the gradient.

- **Empirical risk**: $\beta_\star^n = \operatorname{argmin}_{\beta \in \mathbb{R}^d} \left\{ R_n(\beta) := \frac{1}{n} \sum_{i=1}^n f_i(\beta) \right\}$
  How to choose the k-th obstervation ?
  - k uniformly at random $\Rightarrow$ we use a data several times.
  - k not chosen uniformly at random $\Rightarrow$ sampling not uniform and bias in the gradient.

Implications:

- No unbiased gradients for the empirical risk so far.
- Keep in mind: empirical risk is in any case not observed.

SGD: result with estimated missing probabilities

Finite-sample setting: n is fixed

- Algorithm and main result: requirement of $(p_j)_{j=1,\dots,d}$.
  $\rightarrow$ estimator $\bar{\beta}_k$
- In practice: estimated missing probabilities $(\hat{p}_j)_{j=1,\dots,d}$
  $\rightarrow$ estimator $\bar{\hat{\beta}}_k$. (finite-sample setting: first half of the data to evaluate $(\hat{p}_j)$, second half to build $\bar{\hat{\beta}}_k$).

Result with estimated missing probabilities (simplified version)

Under additional assumptions of **bounded iterates** and **strong convexity** of the risk, Algorithm 1 ensures that, for any $k \geq 0$:

$$\mathbb{E}\left[R(\bar{\hat{\beta}}_k) - R(\bar{\beta}_k) \right] = \mathcal{O}(1/kp_m^6),$$

with $p_m = \min_{j \in \{1,\dots,d\}} p_j$.

Comparison with related work

Comparison with Ma et Needell [Ma and Needell, 2018]:

- SGD with missing covariates for least-squares
- μ-strongly convex problem
- no averaged iterates

$\Rightarrow$ convergence rate of $\mathcal{O}(\frac{\log n}{\mu n})$.

- μ generally out of reach.
- only homogeneous MCAR data.
- main theorem mathematically invalid (empirical risk).

SGD: only one pass

- Only one pass!
  - $\mathcal{F}_{k-1} = \sigma(X_{1:}, y_1, M_{1:} \ldots, X_{k-1:}, y_{k-1}, M_{k-1:})$
  - $\nabla R(\beta_{k-1}) = \mathbb{E}_{(X_{k:}, y_k)}[X_{k:}(X_{k:}^T \beta_{k-1} - y_k)]$
  - No access to $X_{k:}$, only to $\tilde{X}_{k:}$.
  - Another source of randomness: $\mathbb{E} = \mathbb{E}_{(X_{k:}, y_k), M_{k:}} \overset{\text{indep}}{=} \mathbb{E}_{(X_{k:}, y_k)} \mathbb{E}_{M_{k:}}$
 - $\mathbb{E}_{M_{k:}} | \mathcal{F}_{k-1} \rightsquigarrow \mathbb{E}_{M_{k:}}$
    - Mask at step k independent from the previous constructed iterate.

- $X_{i:} \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$, where Σ with uniform random eigenvectors and decreasing eigenvalues, $\epsilon_i \sim \mathcal{N}(0, 1)$
- $y_i = X_{i:}\beta + \epsilon_i$, for β fixed
- $d = 10$, 30% missing values.

- **AvSGD** averaged iterates with a constant step size $\alpha = \frac{1}{2L}$.
- **SGD**[Ma and Needell, 2018] with iterates $\beta_{k+1} = \beta_k - \alpha_k \tilde{g}_{i_k}(\beta_k)$, and decreasing step size $\alpha_k = \frac{1}{\sqrt{k+1}}$.
- **SGD_cst** with a constant step size $\alpha = \frac{1}{2L}$.
- L is considered to be known.

SGD: only one pass

- **AvSGD** averaged iterates with a constant step size $\alpha = \frac{1}{2L}$.
- **SGD**[Ma and Needell, 2018] with iterates $\beta_{k+1} = \beta_k - \alpha_k \tilde{g}_{i_k}(\beta_k)$, and decreasing step size $\alpha_k = \frac{1}{\sqrt{k+1}}$.
- **SGD_cst** with a constant step size $\alpha = \frac{1}{2L}$.
- L is considered to be known.
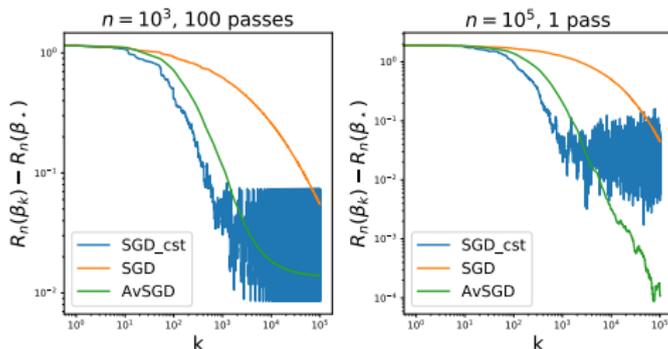
- Multiple passes (left): saturation.
- One pass (right): saturation for **SGD_cst**, $\mathcal{O}(n^{-1/2})$ for **SGD**, $\mathcal{O}(n^{-1})$ for **AvSGD**.

Advanced SGD: other mechanisms?

- For MNAR or MAR data?
  - $\mathcal{F}_{k-1} = \sigma(X_{1:}, y_1, M_1: \ldots, X_{k-1:}, y_{k-1}, M_{k-1:})$
  - $\nabla R(\beta_{k-1}) = \mathbb{E}_{(X_{k:}, y_k)}[X_{k:}(X_{k:}^T \beta_{k-1} - y_k)]$
  - No access to $X_{k:}$, only to $\tilde{X}_{k:}$.
  - Another source of randomness: $\mathbb{E} = \mathbb{E}_{(X_{k:}, y_k), M_{k:}} \underset{indep}{\times} \mathbb{E}_{(X_{k:}, y_k)} \mathbb{E}_{M_{k:}}$
  - $\mathbb{E}_{M_{k:}} | \mathcal{F}_{k-1} \rightsquigarrow \mathbb{E}_{M_{k:}}$
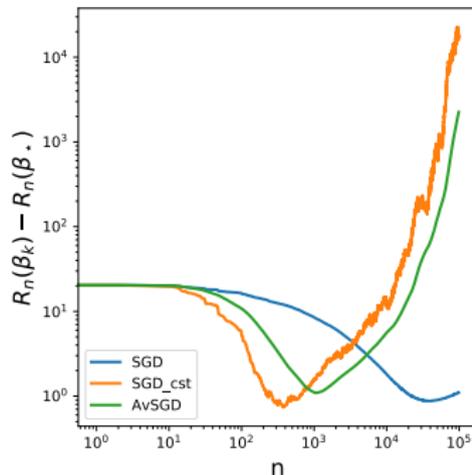   - Mask at step k independent from the previous constructed iterate.

Figure: 1 pass, assuming MAR data

Advanced SGD: other loss functions?

Logit loss-function ? **No solution yet.**

- $y_i \in \{1, -1\}$,
- Logit loss: $f_i(\beta) = \frac{1}{n} \sum_i \log(1 + \exp(-y_i X_i^T \beta))$
- Gradient: $\nabla f_i(\beta) = \frac{-y_i X_i}{1 + \exp(y_i X_i^T \beta)}$
- Approximation of the gradient $\frac{-y_i X_i}{1 + \exp(y_i X_i^T \beta)} \approx \frac{-y_i X_i}{2} + \frac{X_i^T \beta X_i}{4}$

Debiasing the gradient?

- Partially debiasing: $\frac{-y_i X_i}{p(1 + \exp(y_i X_i^T \beta))}$
- Debiasing the approximation of the gradient

Use of the algorithm of Bach and Moulines [Bach and Moulines, 2013] ?

$$\beta_k = \beta_{k-1} - \alpha(\nabla f_k(\bar{\beta}_{k-1}) + H_k(\bar{\beta}_{k-1})(\beta_{k-1} - \bar{\beta}_{k-1})$$

Advanced SGD: polynomial features

- We know how to debiase the gradient.
- Encouraging results on data.
- No theoretical results

$d = 2$. Accounting for the effects of X_{k1}^2, X_{k2}^2, $X_{k1}X_{k2}$.

- augmented design matrix: $(X_{:1}|X_{:2}|X_{:1}X_{:2}|X_{:1}^2|X_{:2}^2)^T$.
- Debiased gradient: $U^{\odot-1} \odot \tilde{X}_{k:}\tilde{X}_{k:}^T\beta_k - \mathrm{diag}(U)^{\odot-1} \odot \tilde{X}_{k:}y_k$

$$
U = \begin{pmatrix}
p_1 & p_1p_2 & p_1p_2 & p_1 & p_1p_2 \\
p_1p_2 & p_2 & p_1p_2 & p_1p_2 & p_2 \\
p_1p_2 & p_1p_2 & p_1p_2 & p_1p_2 & p_1p_2 \\
p_1 & p_1p_2 & p_1p_2 & p_1 & p_1p_2 \\
p_1p_2 & p_2 & p_1p_2 & p_1p_2 & p_2
\end{pmatrix},
$$

$U^{\odot-1}$: formed of the inverse coefficients of U.

Advanced SGD: polynomial features

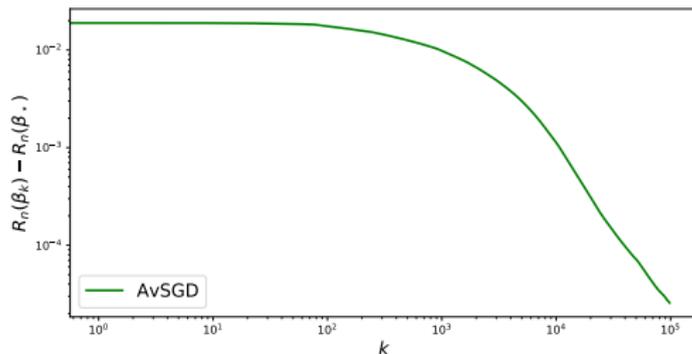$d = 2$. Accounting for the effects of X_{k1}^2, X_{k2}^2, $X_{k1}X_{k2}$.

Figure: Empirical excess risk $(R_n(\beta_k) - R_n(\beta^\star))$ given n for synthetic data $(n = 10^5,\ d = 10)$ when the model accounts mixed effects.

Advanced SGD: polynomial features

For real data (Superconductivity dataset) 3 algorithms to compare :

- the averaged SGD on complete data (blue)
- the proposed debiased averaged SGD (orange)
- the averaged SGD run on imputed-by-0 data without any debiasing (green)
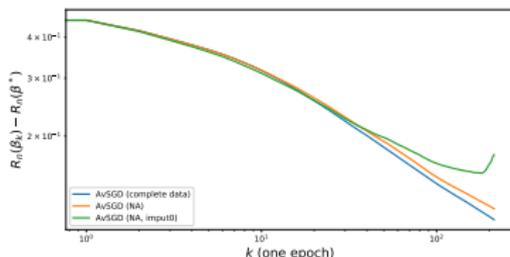

Figure: Empirical excess risk $(R_n(\beta_k) - R_n(\beta^\star))$ given n for the superconductivity dataset ($n = 21263$) (containing 81 initial features) and $d = 3403$ with polynomial features of degree 2.

Advanced SGD: missing-data patterns can be dependent

In our setting: independent missing-data patterns

$$M_{\cdot j} \perp\!\!\!\perp M_{\cdot j'}, j \neq j'$$

$$M = (\delta_{ij})_{1 \leq i \leq n, 1 \leq j \leq d} \quad \text{with} \quad \delta_{ij} \sim \mathcal{B}(p_j)$$

> **Dependent missing-data patterns**
>
> $$\tilde{g}_k(\beta) := (W \odot (\tilde{X}_{k:} \tilde{X}_{k:}^T))\beta - y_k P^{-1} \tilde{X}_{k:}$$
>
> with $W \in \mathbb{R}^{d \times d}$, and $W_{ij} := 1/\mathbb{E}[\delta_{ki}\delta_{kj}]$ for $1 \leq i, j \leq d$

Clustering: computations for the EM algorithm

$$Q(\theta, \phi, \pi; \theta^r, \phi^r, \pi^r) = Q_x(\theta, \pi; \theta^r, \phi^r, \pi^r) + Q_c(\phi; \theta^r, \phi^r, \pi^r)$$

$$Q_x(\theta, \pi; \theta^r, \phi^r, \pi^r) = \sum_{i=1}^{n} \sum_{k=1}^{K} (\tau_{ik})^r \log(\pi_k) + \sum_{i=1}^{n} \sum_{k=1}^{K} (\tau_{ik})^r E_{ix}^r(\theta)$$

$$Q_M(\phi; \theta^r, \phi^r, \pi^r) = \sum_{i=1}^{n} \sum_{k=1}^{K} (\tau_{ik})^r E_{iM}^r(\phi)$$

- Law of x_i^{mis} given $(x_i^{\mathrm{obs}}, z_{ik} = 1, M_i)$?
- Computation of the expectation over this law of $\log(\mathbb{P}(M_i \mid x_i, z_{ik} = 1; \phi))$ (for $E_{iM}^r(\phi)$)?
- $(\tau_{ik})^r$: Computation of $\mathbb{P}(M_i \mid x_i^{\mathrm{obs}}, z_{ik} = 1; \phi^r)$?

Clustering: EM algorithm for MNARz and MNARz^j models

MNARz, MNARz^j : needs some computations but still simple.

$$\mathbb{P}(M_{ij} = 1 \mid x_i, z_{ik} = 1; \phi) = \rho(\alpha_{kj}) \qquad (\perp\!\!\!\perp X)$$

Gaussian case for MNARz and MNARz^j

$\left(x_i^{\mathrm{mis}} \mid x_i^{\mathrm{obs}}, z_{ik} = 1; \theta^r\right) \sim \mathcal{N}\left((\tilde{\mu}_{ik}^{\mathrm{mis}})^r, (\tilde{\Sigma}_{ik}^{\mathrm{mis}})^r\right).$

- E-step : for $k = 1, \ldots, K$ and $i = 1, \ldots, n$, compute $(\tilde{\mu}_{ik}^{\mathrm{mis}})^r, (\tilde{\Sigma}_{ik}^{\mathrm{mis}})^r, (\tau_{ik})^r$.

- M-step : for $k = 1, \ldots, K$, compute $\pi_k^{r+1}, \mu_k^{r+1}, \Sigma_k^{r+1}$ For ϕ^{r+1}: maximization of $Q_M(\phi; \theta^r, \phi^r, \pi^r)$ over ϕ with a **Newton-Raphson algorithm** (classical procedure for link functions of interest)

An EM algorithm can also be easily derived for categorical data

Clustering: EM algorithm for MNARz and MNARz^j models

MNAR$x*$: needs approximations

$$\mathbb{P}(M_{ij} = 1 \mid x_i, z_{ik} = 1;) = \rho(\alpha_{kj} + \beta_{kj}x_{ij}) \qquad (\text{not} \perp\!\!\!\perp x)$$

Gaussian case for MNAR$x*$
- $(x_i^{\mathrm{mis}} \mid x_i^{\mathrm{obs}}, z_{ik} = 1, M_i)$:
  ✗ not classical if ρ is Logit, ✓ truncated Gaussian distribution if ρ is Probit
- No closed forms of $E_{iM}^r(\phi)$ and of $(\tau_{ik})^r$.

Clustering: SEM algorithm for MNAR$_{X*}$

Gaussian data:

- SE-step : draw the missing data
  $((x_i^{\mathrm{mis}})^{r+1}, z_i^{r+1}) \sim (. \mid x_i^{\mathrm{obs}}, M_i; \theta^r, \phi^r, \pi^r)$
  Use of **One**-Gibbs sampling :
  - $(x_i^{\mathrm{mis}})^{r+1} \sim (\cdot \mid x_i^{\mathrm{obs}}, z_i^r, c_i; \theta^r, \phi^r)$:
    ✗ not classical if ρ is Logit,
    ✓ truncated Gaussian distribution if ρ is Probit
  - $z_i^{r+1} \sim (\cdot \mid x_i^{r+1}, c_i; \theta^r, \phi^r, \pi^r)$: draw the membership k of z_i^{r+1} from the **multinomial distribution**

  Let $X^{r+1} = (x_1^{r+1} \mid \ldots \mid x_n^{r+1})$, $Z^{r+1} = (z_1^{r+1} \mid \ldots \mid z_n^{r+1})$ be the imputed matrix and the partition

- M-step : for $k = 1, \ldots, K$, compute $\pi_k^{r+1}, \mu_k^{r+1}, \Sigma_k^{r+1}, \phi^{r+1}$.

Clustering: identifiability for categorical data

f_k	Gaussian		Poisson	
ρ_k	Probit	Logit	Probit	Logit
MNAR$z^j x^k$				
MNAR$x^k z$	$\checkmark$	generic ident.	$\checkmark$	generic ident.
MNARx^k				
MNARxz^j				
MNARxz				
MNARx	$\checkmark$	$\checkmark$	$\checkmark$	$\checkmark$
MNARz				
MNARz^j				

Generic identifiability: all not-identifiable parameter choices lie within a proper subvariety, and thus form a set of Lebesgue zero measure

Computational comments for all the works

What is costly? MNAR!

- Low-rank model with fixed effects modelling the missing-data mechanism: Monte Carlo, SIR algorithm
- Low-rank model with random effects: number of *aggregations* for the combinations of the pivot variables $\Leftrightarrow$ number of linear regression to be performed
- SEM algorithm for MNAR$x*$: we use a One-Gibbs sampling, truncated Gaussian (difficulty of drawing)

Solutions?

- Consider the method adding the mask $\simeq$ same cost than MAR data.
- *simple* MNAR like MNARz for the model-based clustering.
- Better implementations.

Traumabase dataset