Model based clustering with Missing Not At Random data

Maasai Seminar

Joint work with:
Matthieu Marbac (Ensai Rennes),
Claire Boyer (Sorbonne Université),
Christophe Biernacki (Inria Lille),
Julie Josse (Inria Antenne Montpellier),
Fabien Laporte (UCO), Gilles Celeux (Université Paris Saclay)

October 21, 2022

Outline

Missing values are everywhere

- Growing masses of data, multiplication of sources
  ⇒ 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	
⋮	⋮	⋮	⋮	⋮	

250 clinical variables (heterogeneous)

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

Missing values are everywhere

- Growing masses of data, multiplication of sources
  ⇒ `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.
- **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

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

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~~

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

- Loss of information.

  Traumabase®: only 5% of the rows are kept.

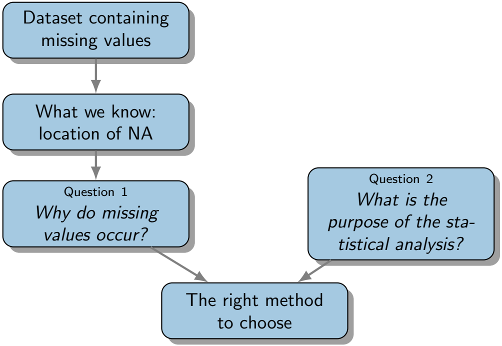
- Bias in the analysis.

  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 notations

- $Y = \{y_1 | \dots | y_n\}^T$: full dataset with n individuals
- Continuous, categorical or mixed data.
- $C = \{c_1 | \dots | c_n\}^T \in \{0,1\}^{n \times d}$: pattern of missing data for the full dataset

$$c_{ij} = 1 \Leftrightarrow y_{ij} \text{ is missing}$$

- y_i^{obs}: the observed variables values for individual i
- y_i^{mis}: the missing variables values for individual i

Missing-data mechanism (Rubin, 1976)

Data values (Y) ⟷ Link? ⟷ Missing values (C)

$$f(c|y;\psi), \psi \in \Omega_\psi$$

Missing Completely At Random (MCAR)

$$f(c|y;\psi) = f(c;\psi)$$

MCAR

Machines fail,
Doctors forget to fill the form

Missing At Random (MAR)

y^{obs}: observed component of y.

$$f(c|y;\psi) = f(c|y^{\text{obs}};\psi)$$

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
65
59
62
NA

"underlying" values:

HR
65
59
62
84

Ignorable *vs.* non ignorable mechanism

- Parametric estimation: model the joint distribution (Y, C) parametrized by $\gamma, \psi \in \Omega_{\gamma,\psi}$.
- Likelihood-approach: maximizing the full observed likelihood.

$$
\begin{aligned}
L_{\text{full,obs}}(\gamma, \psi; y^{\text{obs}}, c) &= \int L_{\text{full}}(\gamma, \psi; y, c) dy^{\text{mis}} \\
&= \int f(y; \gamma) f(c|y; \psi) dy^{\text{mis}} \\
&= f(c|y^{\text{obs}}; \psi) \int f(y; \gamma) dy^{\text{mis}} \qquad \text{M(C)AR mecha.} \\
&\propto L_{\text{ign}}(\gamma; y^{\text{obs}}) = \int f(y; \gamma) dy^{\text{mis}}
\end{aligned}
$$

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

MNAR: one should consider the mechanism.

Focus on MNAR mechanism

We should consider (Y, C) (not-ignorable mechanism).

The main MNAR specifications

- selection model [Heckman, 1979]:

$$f(y, c; \gamma, \psi) = f(y; \gamma) f(c|y; \psi)$$

- pattern-mixture model [Little, 1993]:

$$f(y, c; \xi, \varphi) = f(c; \xi) f(y|c; \varphi)$$

Q: *How to choose the MNAR specification ?*

- Estimate the parameters of the data distribution: **selection models.**
- Distribution is not the same for the observed data and the missing data: pattern-mixture models.

Focus on MNAR mechanism

We should prove the identifiability of the parameters.

Identifiability issue in the MNAR case Credit: Ilya Shpitser

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

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

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

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

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

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

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

Outline

Our goals

- MNAR mechanism.
- Selection model: $f(y, c; \gamma, \psi) = f(y; \gamma) f(c|y; \psi)$.

Embed missing data management into the analysis to:

- Perform clustering: identify relevant groups of individuals.
- Estimate the parameters of the data distribution.
- (Impute missing values.)

Clustering: model-based approach

- Partition with K clusters: $Z = (z_1 | \dots | z_n)^T \in \{0,1\}^{n \times K}$, with $z_{ik} = 1$ if y_i belongs to cluster k.

Mixture model

$$f(y_i; \gamma) = \sum_{k=1}^{K} \overbrace{\pi_k}^{=\mathbb{P}(z_{ik}=1)} \underbrace{f_k(y_i; \lambda_k)}_{\text{pdf in the cluster } k}$$

Clustering: model-based approach

- Partition with K clusters: $Z = (z_1|\dots|z_n)^T \in \{0,1\}^{n\times K}$, with $z_{ik} = 1$ if y_i belongs to cluster k.

Mixture model

$$f(y_i;\gamma) = \sum_{k=1}^{K} \overbrace{\pi_k}^{=\mathbb{P}(z_{ik}=1)} \underbrace{f_k(y_i;\lambda_k)}_{\text{pdf in the cluster } k}$$

- Missing data in Y.

Mixture model with missing data

$$f(y_i, c_i;\theta) = \sum_{k=1}^{K} \pi_k f_k(y_i;\lambda_k) f_k(c_i \mid y_i;\psi_k),$$

A zoology of MNAR models in clustering

Q: *Which distribution* $f_k(c_i \mid y_i; \psi_k)$ *to propose in this clustering context?*

$$f_k(c_i \mid y_i; \psi_k) = \prod_{j=1}^{d} \left(\rho(\alpha_{kj} + \beta_{kj} y_{ij})\right)^{c_{ij}} \left(1 - \rho(\alpha_{kj} + \beta_{kj} y_{ij})\right)^{1-c_{ij}},$$

where $\psi_k = (\alpha_{k1}, \beta_{k1}, \ldots, \alpha_{kK}, \beta_{kK})$ and ρ is a link function.

How to understand this distribution?

- α_{kj}: the missingness depends on the class membership k, not the same effect for every variable.
- β_{kj}: the missingness depends on the value itself (y_{ij}), not the same effect for each cluster.
- Simplest model:

$$\text{MCAR:} \quad \beta_{kj} = 0, \ \forall (k, j) \text{ and } \alpha_{1j} = \ldots = \alpha_{Kj}, \ \forall j.$$

A zoology of MNAR models in clustering

Parcimonious models: the probability of being missing depend

- **on both the variable and the class membership:**

$$\text{MNAR}yz^j: \quad \beta_{1j} = \ldots = \beta_{Kj},\ \forall j.$$
$$\text{MNAR}y^kz: \quad \alpha_{kj} = \ldots = \alpha_{k1},\ \forall k.$$
$$\text{MNAR}yz: \quad \beta_{1j} = \ldots = \beta_{Kj},\ \forall j \text{ and } \alpha_{kj} = \ldots = \alpha_{k1},\ \forall k.$$

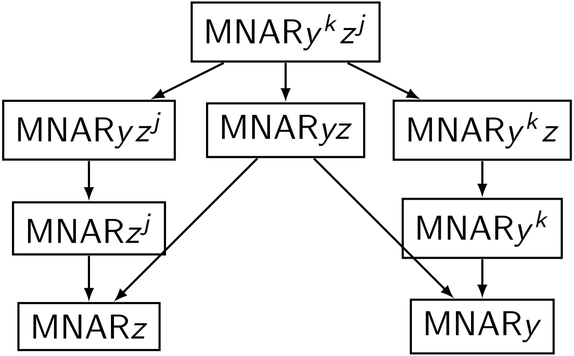
- **only on the variable itself:**

$$\text{MNAR}y: \quad \alpha_{11} = \ldots = \alpha_{1d} = \alpha_{21} = \ldots = \alpha_{Kd} \text{ and } \beta_{1j} = \ldots = \beta_{Kj}, \forall j.$$
$$\text{MNAR}y^k: \quad \alpha_{11} = \ldots = \alpha_{1d} = \alpha_{21} = \ldots = \alpha_{Kd}.$$

- **only on the class membership:**

$$\text{MNAR}z: \quad \beta_{kj} = 0,\ \forall (k,j) \text{ and } \alpha_{kj} = \ldots = \alpha_{kd}, \forall k.$$
$$\text{MNAR}z^j: \quad \beta_{kj} = 0,\ \forall (k,j).$$

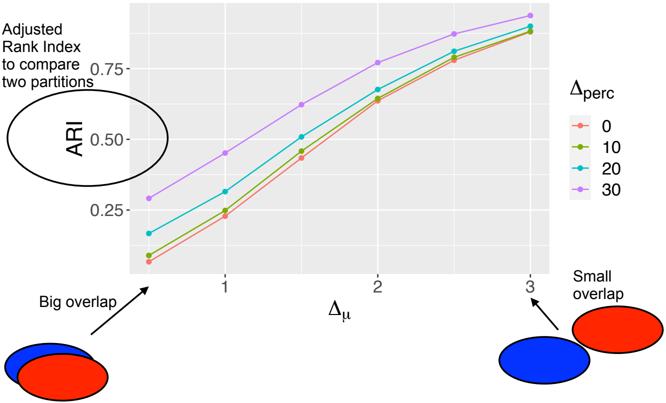
Proposed zoology of MNAR models in clustering



MNARz from every angle

(1) C gives information on partition Z

- MNARz 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.



MNARz from every angle

(2) MNARz (and MNARz^j) models interpreted as MAR

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

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

Proposition 1: in terms of maximum likelihood

MLE associated to $\tilde{Y}^{\text{obs}}$ under MAR model
$\Leftrightarrow$ MLE associated to Y^{obs} under MNARz/MNARz^j models.

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 $\sum_{k=1}^{K} \pi_k f_k(y_i; \theta_k)$ 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 or MNARz^j is identifiable.

✗ The mixture model with any MNAR$y*$ is not identifiable.

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

Identifiability up to a label swapping.

Outline

EM algorithm

Initialized at the point $\theta^{[0]} = (\pi^{[0]}, \lambda^{[0]}, \psi^{[0]})$, the iteration $[r]$ of the EM algorithm consists in performing two steps:

- E-step : compute the expectation of the complete-data log-likelihood $Q(\theta; \theta^{[r-1]}) = \mathbb{E}_{\theta^{[r-1]}} \left[\ell_{\text{comp}}(\theta; Y, Z, C) \mid Y^{\text{obs}}, C \right]$, $\ell_{\text{comp}}(\theta; Y, Z, C) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k f_k(y_i; \lambda_k) f_k(c_i \mid y_i; \psi_k) \right)$.
- M-step : update the parameters by maximizing this function $\theta^{[r]} = \text{argmax}_\theta Q(\theta; \theta^{[r-1]})$.

EM algorithm: feasible computations?

One has: $Q(\theta; \theta^{[r-1]}) =$

$$\sum_i \sum_k t_{ik}(\theta^{[r-1]}) \left[\log(\pi_k) + \underbrace{\tau_y(\lambda_k; y_i^{\text{obs}}, c_i, \theta^{[r-1]})}_{=\mathbb{E}_{\theta^{[r-1]}}\left[\ln f_k(y_i;\lambda_k) | y_i^{\text{obs}}, c_i, z_{ik}=1\right]} + \overbrace{\tau_c(\psi_k; y_i^{\text{obs}}, c_i, \theta^{[r-1]})}^{\mathbb{E}_{\theta^{[r-1]}}\left[\ln f_k(c_i|y_i;\phi_k) | y_i^{\text{obs}}, c_i, z_{ik}=1\right]} \right]$$

with $t_{ik}(\theta^{[r-1]}) = \mathbb{P}(z_{ik} = 1 | y_i^{\text{obs}}, c_i)$.

- Law of y_i^{mis} given $(y_i^{\text{obs}}, z_{ik} = 1, c_i)$?
- Computation of the expectation over this law of $f_k(c_i \mid y_i; \phi_k)$?

EM algorithm: feasible computations?

MNARz, MNARzj: needs some computations but still simple.

$$f_k(c_i \mid y_i; \psi_k) = \rho(\alpha_{kj}) \qquad (\perp\!\!\!\perp Y)$$

- $\Rightarrow \mathcal{L}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik} = 1, c_i) = \mathcal{L}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik} = 1)$
- EM algorithm for Gaussian data,
- EM for categorical data.

MNAR$y*$: needs approximations

$$f_k(c_i \mid y_i; \psi_k) = \rho(\alpha_{kj} + \beta_{kj} y_{ij}) \qquad (\text{not} \perp\!\!\!\perp Y)$$

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

SEM algorithm for MNAR$y*$

SEM easier? random drawing instead of expectation

- SE-step: draw the missing data
  $((y_i^{\text{mis}})^{[r]}, z_i^{[r]}) \sim (. \mid y_i^{\text{obs}}, c_i; \pi^{[r-1]}, \lambda^{[r-1]}, \psi^{[r-1]})$
  - $(y_i^{\text{mis}})^{[r]} \sim (\cdot \mid y_i^{\text{obs}}, z_i^{[r-1]}, c_i; \lambda^{[r-1]}, \psi^{[r-1]})$:
  - $z_i^{[r]} \sim (\cdot \mid y_i^{[r]}, c_i; \pi^{[r-1]}, \lambda^{[r-1]}, \psi^{[r]})$: draw the membership k of $z_i^{[r]}$ from the **multinomial distribution**

  Let $Y^{[r]} = (y_1^{[r]}|\ldots|y_n^{[r]})$, $Z^{[r]} = (z_1^{[r]}|\ldots|z_n^{[r]})$ be the imputed matrix and the partition.
- M-step: for $k = 1, \ldots, K$, compute $\pi_k^{[r]}$, $\lambda_k^{[r]}, \psi_k^{[r]}$.

$(\cdot \mid y_i^{\text{obs}}, z_i^{[r-1]}, c_i; \lambda^{[r-1]}, \psi^{[r-1]})$?

- not classical if ρ is Logit
- truncated Gaussian distribution if ρ is Probit

Summary of the algorithms

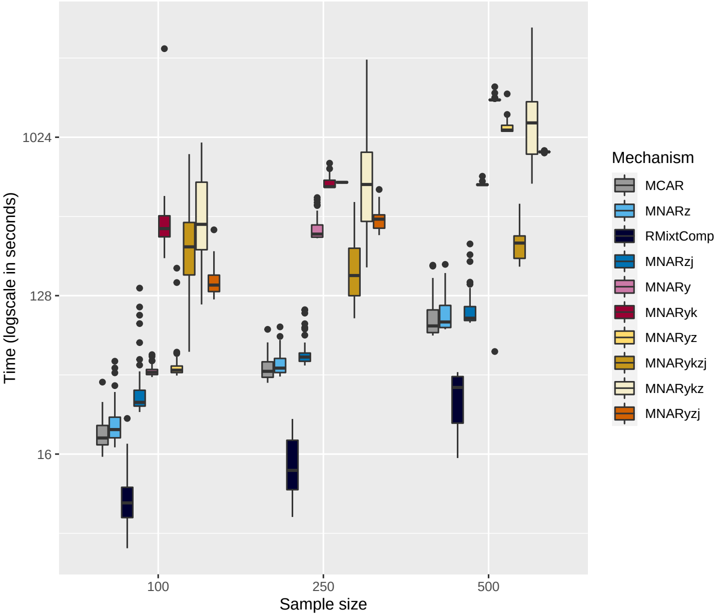
	EM		SEM	
	Gaussian	Categorical	Gaussian	Categorical
MNARz MNARz^j	✓	✓	✓	✓
MNAR$y*$	no closed form	not ident.	✓ (Probit)	not ident.

Outline

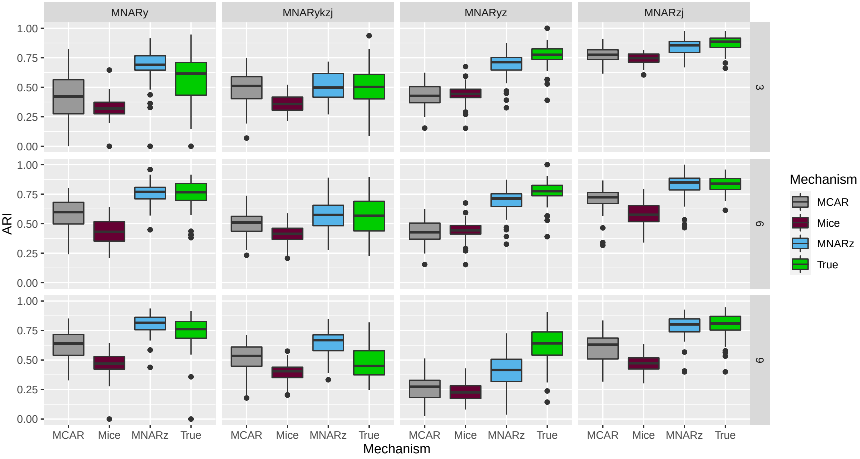
Setting

- Gaussian mixture with three components having unequal proportions (1 = 0.5, 2 = 3 = 0.25), independent variables.
- Control the rates of misclassification (10%) and missingness (30%): we fix them equal for each scenario.

Computation time

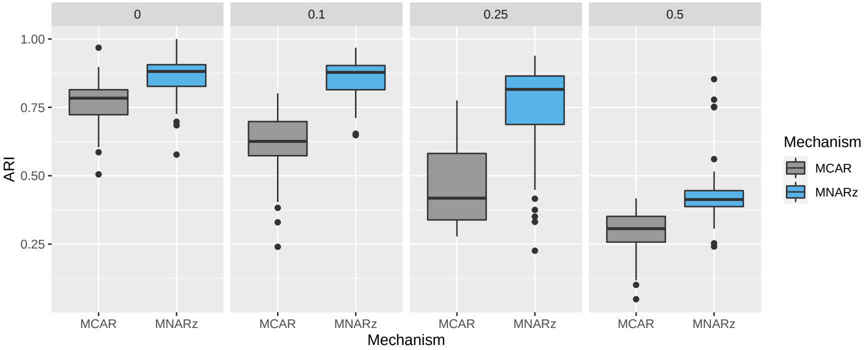


MNARz vs other MNAR model



MNARz: robustness to the misspecification of the data distribution

- Three-components Gaussian mixture with non-diagonal covariance matrices: $\Sigma_{ij} = \ell, i \neq j$, with $\ell \in \{0, 0.1, 0.25, 0.5\}$
- Algorithm assumes $\ell = 0$.

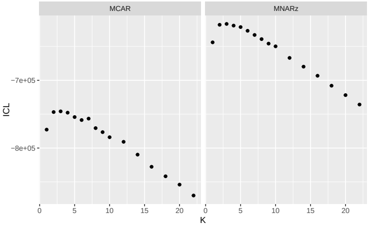


Results on real data

41 mixed variables containing missing values assumed to be MNARz The variables related to the patient death are not taken into account.

Can the MNAR mechanism improve the classification ? Is there an influence of the mechanism ?

- Same number of clusters selected by the ICL criterion;
- For $K = 3$, ARI between the classifications obtained assuming MNARz and MCAR $= 0.9$;



Results on rela data

$$\sqrt{\sum_{i=1}^{n}(\mathbb{P}(z_{ik}=1|y_{is}^{\text{obs}};\theta^{\text{MCAR}})-\mathbb{P}(z_{i\tilde{k}}=1|y_{is}^{\text{obs}};\theta^{\text{MNAR}}))^2}, \forall k,\tilde{k}\in\{1,2,3\}$$

MCAR \ MNARz	Class 1	Class 2	Class 3
Class 1	**2.43**	26.5	37.6
Class 2	26.2	**3.40**	20.1
Class 3	39.3	19.2	**2.05**

Table: Euclidean distance between the conditional probabilities of the cluster memberships given the observed values of the variable Shock.index.ph in the Traumabase dataset, obtained using the algorithm considering MNARz data, and the ones obtained with the algorithm considering MCAR data.

Outline

Conclusion

Summary

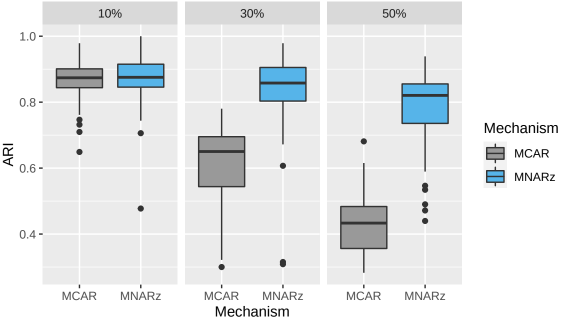
- Interest to put a model on c
- Interest of the simple but meaningful model MNARz
- Trade-off between biased mixture model and biased missingness mechanism.

Ongoing works

- Implement the proposed models/algo. in the Mixmod software[a]

[a] http://www.mixmod.org

MNARz: robustness to the NA% and choice of K



	MCAR	MNARz
10% NA	94%	94%
30% NA	8%	56%
50% NA	0%	20%

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.

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

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

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.

References II

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