Dealing with missing data in model-based clustering through a MNAR model

Christophe Biernacki, Claire Boyer, Gilles Celeux, Julie Josse, Fabien Laporte, Matthieu Marbac Lourdelle, Aude Sportisse

▶ **To cite this version:**

Christophe Biernacki, Claire Boyer, Gilles Celeux, Julie Josse, Fabien Laporte, et al.. Dealing with missing data in model-based clustering through a MNAR model. The 14th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena, May 2021, Zakopane, Poland. hal-03505659

HAL Id: hal-03505659
https://hal.science/hal-03505659

Submitted on 31 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Dealing with missing data in model-based clustering through a MNAR model

Christophe Biernacki, Claire Boyer, Gilles Celeux, Julie Josse, Fabien Laporte, Matthieu Marbac Lourdelle, Aude Sportisse

THE 14TH PROFESSOR ALEKSANDER ZELIAS INTERNATIONAL CONFERENCE ON
MODELLING AND FORECASTING OF SOCIO-ECONOMIC PHENOMENA
ON-LINE CONFERENCE
MAY 11-13, 2021, ZAKOPANE, POLAND

Take home message

1. The missing data pattern may convey some information on clustering
2. Embed the missingness mechanism directly within the clustering modeling step

Outline

Missing data: an inevitable event

The larger the datasets, the more missing data may appear...

Two traditional solutions (for obtaining a filled dataset)

- Discard individuals with missing data: more variance or a biased subset
- Impute missing data: possible bias and underestimation of the variability

General guidelines

- Obtaining a complete dataset is not the final goal
- Missing data management should take into account the initial analysis target

Our analysis target: model-based clustering

Embed missing data management into this paradigm...

Missing data: notations

- $Y = \{y_1|\ldots|y_n\}^T$: full dataset with n individuals
- $y_i = (y_{i1}, \ldots, y_{id}) \in \mathcal{Y}$, depending on the data type: individual $i \in \{1, \ldots, n\}$
  - **continuous data**: $\mathcal{Y} = \mathbb{R}^d$
  - **categorical data**: $\mathcal{Y} = \{0,1\}^{\ell_1} \times \ldots \times \{0,1\}^{\ell_d}$ where ℓ_j is the number of levels for $y_{ij} = (y_{ij}^1, \ldots, y_{ij}^{\ell_j})$, where $y_{ij}^\ell = 1$ if y_{ij} takes the level ℓ, 0 otherwise.
  - **mixed data**: combination of continuous and categorical data.
- $C = \{c_1|\ldots|c_n\}^T \in \{0,1\}^{n \times d}$: pattern of missing data for the full dataset
- $c_i = (c_{i1}, \ldots, c_{id}) \in \{0,1\}^d$: pattern of missing data for individual $i \in \{1, \ldots, n\}$

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

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

Missing data: typology of the missing mechanisms

- Missing completely at random (MCAR):

$$\mathbb{P}(c|y;\psi) = \mathbb{P}(c;\psi) \quad \forall y$$

- Missing at random (MAR):

$$\mathbb{P}(c|y;\psi) = \mathbb{P}(c|y^{\text{obs}};\psi) \quad \forall y^{\text{mis}}$$

- Missing not at random (MNAR): the mechanism is not MCAR nor MAR

Example of MNAR data

The probability to have a missing value on income depends on the value of income (rich people less inclined to reveal their income).

Ignorable *vs.* non ignorable model

A missing mechanism is ignorable if likelihoods can be decomposed as

$$L(\theta, \psi; \underbrace{y^{\text{obs}}, c}_{\text{observed data}}) = L(\psi; c|y^{\text{obs}}) \times L(\theta; y^{\text{obs}})$$

Some simple algebra show that this occurs when missing mechanism is not MNAR

Inference of θ

"If the missing mechanism is ignorable then likelihood-based inferences for θ from $L(\theta; y^{\text{obs}})$ will be the same as likelihood based inference for θ from $L(\theta, \psi; y^{\text{obs}}, c)$." ([Little and Rubin, 2002] Section 6.2)

- M(C)AR is ignorable
- MNAR is not ignorable

Clustering: model-based approach

- Partition with K clusters: $Z = (z_1|\dots|z_n)^T \in \{0,1\}^{n\times K}$ where
  - $z_i = (z_{i1}, \dots, z_{iK}) \in \{0,1\}^K$
  - $z_{ik} = 1$ if y_i belongs to cluster k, $z_{ik} = 0$ otherwise
- Mixture model: $y_1, \dots, y_n$ are i.i.d. from the mixture

$$f(y_i; \pi, \theta) = \sum_{k=1}^{K} \pi_k f_k(y_i; \theta_k)$$

  - $\pi_k = P(z_{ik} = 1)$, $\pi = (\pi_1, \dots, \pi_K)$
  - $f_k(.; \theta_k)$: pdf of the k-th component parametrized by θ_k, $\theta = (\theta_1, \dots, \theta_K)$
    - **continuous data**: $f_k(.; \theta_k) = \phi(.; \mu_k, \Sigma_k)$ is the d-variate **Gaussian distribution** with mean vector μ_k and covariance matrix Σ_k
    - **categorical data**: the features are independent conditionally to the group membership i.e. $f_k(.; \theta_k) = \prod_{j=1}^{d} f_{kj}(.; \theta_{kj})$, where $f_{kj} = \prod_{\ell=1}^{\ell_j} (\theta_{kj}^{\ell})^{y_{ij}^{\ell}}$ is the **multinomial distribution** with $\theta_{kj} = (\theta_{kj}^{\ell} = \mathbb{P}(y_{ij}^{\ell} = 1 | z_{ik} = 1))_{\ell=1,\dots,\ell_j}$
    - **mixed data**: the features are independent conditionally to the group membership, $f_k(.; \theta_k)$ is the product of univariate Gaussian and multinomial distributions
    - Can also be extended to other cases (count data with Poisson distributions for instance)

Question we address in this work

Which distribution $\mathbb{P}(c|y, z; \psi)$ to propose in this clustering context?

Outline

Proposed zoology of MNAR models in clustering

$$\mathbb{P}(c_i|y_i, z_{ik} = 1; \psi) = \prod_{j=1}^{d} \mathbb{P}(c_{ij}|y_i, z_{ik} = 1; \psi)$$

- MNAR$y^k z^j$, with $\psi = (\alpha, \beta)$ where $\alpha = (\alpha_{11}, \ldots, \alpha_{1d}, \ldots, \alpha_{K1}, \ldots, \alpha_{Kd})^T \in \mathbb{R}^{Kd}$ and $\beta = (\beta_{11}, \ldots, \beta_{1d}, \ldots, \beta_{K1}, \ldots, \beta_{Kd})^T \in \mathbb{R}^{Kd}$

$$\mathbb{P}(c_{ij} = 1 \mid y_i, z_{ik} = 1; \psi) = \rho(\alpha_{kj} + \beta_{kj} y_{ij}),$$

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

- MNARyz, MNAR$y^k z$, MNARyz^j
- MNARy, MNARy^k

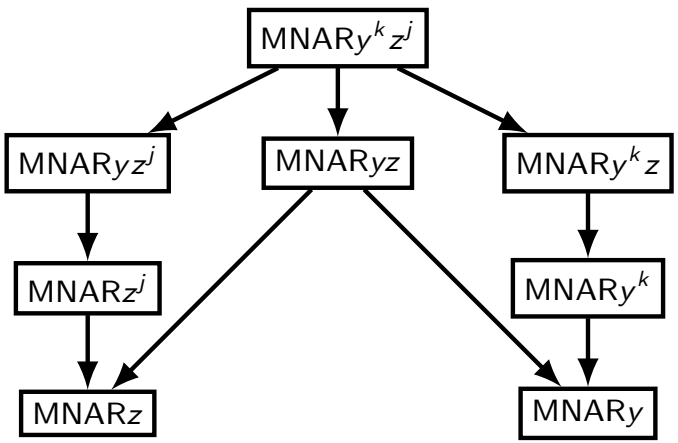
$\psi = (\beta_{11}, \ldots, \beta_{1d}, \ldots, \beta_{K1}, \ldots, \beta_{Kd})^T$

$$\mathbb{P}(c_{ij} = 1 \mid y_i, z_{ik} = 1; \psi) = \rho(\beta_{kj} y_{ij})$$

- MNARz, MNARz^j

$\psi = (\alpha_{11}, \ldots, \alpha_{1d}, \ldots, \alpha_{K1}, \ldots, \alpha_{Kd})^T$

$$\mathbb{P}(c_{ij} = 1 \mid y_i, z_{ik} = 1; \psi) = \rho(\alpha_{kj})$$



Overview of the proposed MNAR models

	Effect on the variable j		Effect on the class membership k		Nb parameters	
	Depends on j	Depends on k	Depends on j	Depends on k	Continuous	Categorical
MNAR$z^j y^k$	✓	✓	✓	✓	$2Kd$	$K(d + \sum_{j=1}^{d}(\ell_j - 1))$
MNARyz^j	✓	✗	✓	✓	$(K+1)d$	$Kd + \sum_{j=1}^{d}(\ell_j - 1)$
MNAR$y^k z$	✓	✓	✗	✓	$K(d+1)$	$K(1 + \sum_{j=1}^{d}(\ell_j - 1))$
MNARyz	✓	✗	✗	✓	$(K+d)$	$K + \sum_{j=1}^{d}(\ell_j - 1)$
MNARy	✓	✗	✗	✗	d	$\sum_{j=1}^{d}(\ell_j - 1)$
MNARy^k	✓	✓	✗	✗	Kd	$K\sum_{j=1}^{d}(\ell_j - 1)$
MNARz	✗	✗	✗	✓	K	K
MNARz^j	✗	✗	✓	✓	Kd	Kd

Terminology in the sequel:

- MNARz, MNARz^j: the only effect of missingness is on the class membership
- MNAR$y*$: all the other models which considers the effect of the missingness depending on the variable
- MNAR$*$: all the models

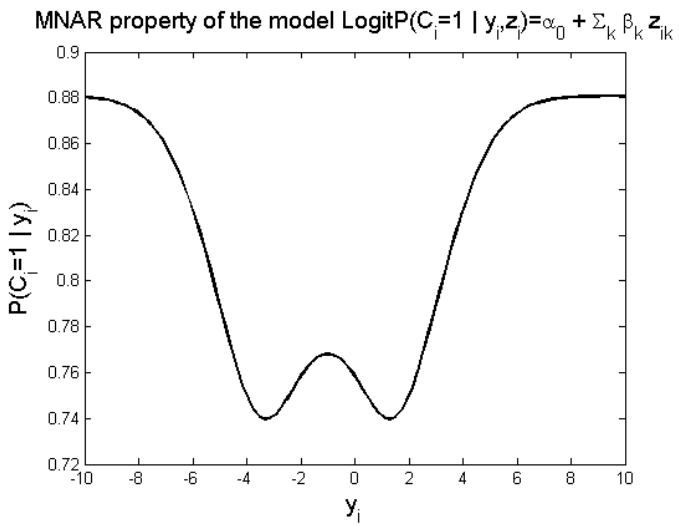
MNARz analysis: it depends on y through z!

$$P(c_{ij} = 1|y_i; \theta, \psi) = \sum_{k=1}^{K} P(c_{ij} = 1|y_i, z_{ik} = 1; \psi) P(z_{ik} = 1|y_i; \theta)$$

Example of a univariate Gaussian model with the three components

$$0.2N(\cdot; 0, 1) + 0.3N(\cdot; 1, 2) + 0.5N(\cdot; 2, 3)$$

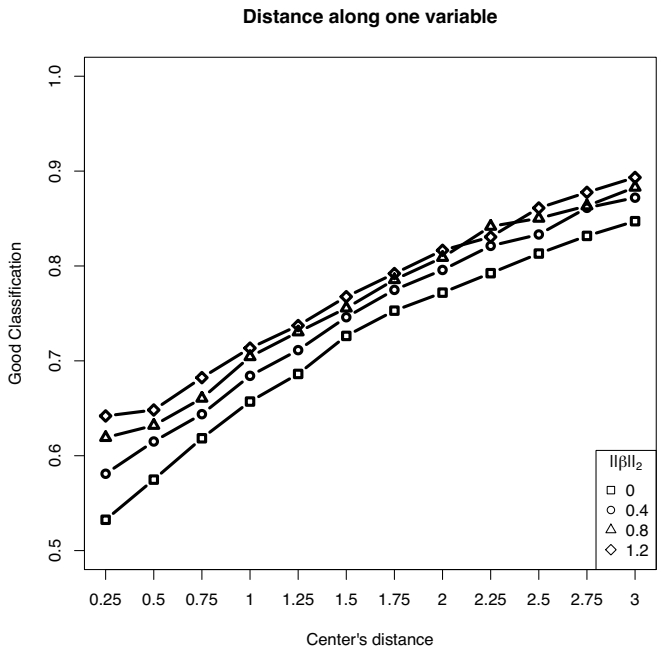
and with parameters of the logit expression: $\alpha_0 = 1, \beta_1 = 1, \beta_2 = -1, \beta_3 = 1$



MNARz analysis: pattern c gives information on partition z!

Draw Bayes error of a MNARz model with two components and 20% of missing data

$$\pi_k = 0.5,\ \|\mu_2 - \mu_1\| \text{ varies},\ \Sigma_1 = \Sigma_2 = \mathsf{I},\ |\beta_2 - \beta_1| \text{ varies}$$



Both μ_k and β_k act on the Bayes error

Reinterpretation of the MNARz and MNARz^j models as MAR

Commonly used in Machine Learning
[Jones, 1996, Little and Rubin, 2002, Josse et al., 2019]

Mixture model for Y^{obs} and Bernoulli distribution for C
$\Leftrightarrow$ MAR mixture model for $\tilde{Y}^{\text{obs}} = (Y^{\text{obs}}|C)$

For example,

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

then $\tilde{Y}^{\text{obs}}$ is expressed as

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

Proposition 1: in terms of maximum likelihood

The maximum likelihood estimate associated to the dataset $\tilde{Y}^{\text{obs}}$ under MAR model is the one associated to the dataset Y^{obs} under MNARz or MNARz^j models.
$\Rightarrow$ can be extended to other estimation strategies

Outline

Continuous and count data (1)

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

Identifiability for a mixture model with MNAR data
$\Leftrightarrow$ Mixture/MNAR parameters are uniquely determined from available information

Proposition 2: identifiability for continuous and count data

Assume that

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 $\mathcal{R} = \{\rho_1, \ldots, \rho_K\} = \{\rho(.; \psi_1), \ldots, \rho(.; \psi_K)\}$. The total ordering is s.t. $\forall k < \ell$, $F_k = \rho_k f_{kj} \preceq F_\ell = \rho_\ell f_{\ell j}$ implies

$$\lim_{u \to +\infty} \frac{\rho_\ell(u) f_{\ell j}(u)}{\rho_k(u) f_{kj}(u)} = 0$$

Then the mixture model with one of the MNAR$*$ mechanisms is identifiable up to label swapping

Continuous and cound data (2)

Is the total ordering checked for classical distributions ?

f_k	Gaussian		Poisson	
ρ_k	Probit	Logit	Probit	Logit
MNAR$z^j y^k$ MNAR$y^k z$ MNARy^k	✓	generic idenfiability	✓	generic idenfiability
MNARyz^j MNARyz MNARy MNARz MNARz^j	✓	✓	✓	✓

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

Categorical data

Previous work: [Allman et al., 2009] (without NA)
Recall that for categorical data: conditional independence of the features given the group membership holds i.e. $f_k(\cdot;\theta_k) = \prod_{j=1}^{d} f_{kj}(\cdot\,;\theta_{kj})$

Proposition 3: identifiability for categorical data

Assume that $d \geq 2\lceil \log_2 K \rceil + 1$ and $f_k(\cdot;\theta_k) = \prod_{j=1}^{d} f_{kj}(\cdot\,;\theta_{kj})$

✓ Then the mixture model with MNARz or MNARz^j mechanism is identifiable up to label swapping

✗ The mixture model with one of the MNAR$y*$ mechanisms is not identifiable

Mixed data

$f_k(\cdot;\theta_k) = \prod_{j=1}^{d} f_{kj}(\cdot\,;\theta_{kj})$, thus identifiability of mixed data directly follows from Proposition 2 for continuous variables and from Proposition 3 for categorical variables

Outline

EM algorithm: looks simple

The **expected complete log-likelihood** knowing the observed data and a current value of the parameters can be **decomposed into two parts**

$$\begin{aligned} Q(\theta, \psi, \pi; \theta^r, \psi^r, \pi^r) &= \mathbb{E}[\ell_{\text{comp}}(\theta, \psi, \pi; y, z, c) | y_i^{\text{obs}}, c_i; \theta^r, \psi^r, \pi^r] \\ &= Q_y(\theta, \pi; \theta^r, \psi^r, \pi^r) + Q_c(\psi; \theta^r, \psi^r, \pi^r) \end{aligned}$$

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

$$Q_c(\psi; \theta^r, \psi^r, \pi^r) = \sum_{i=1}^{n} \sum_{k=1}^{K} (\tau_{ik})^r E_{ic}^r(\psi)$$

where for $i = 1, \ldots, n$ and $k = 1, \ldots, K$,

$$\begin{aligned} E_{iy}^r(\theta) &= \mathbb{E}\left[\log(f_k(y_i; \theta_k)) \mid y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r\right] \\ E_{ic}^r(\psi) &= \mathbb{E}\left[\log(\mathbb{P}(c_i \mid y_i, z_{ik} = 1; \psi)) \mid y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r\right] \\ (\tau_{ik})^r &= \mathbb{P}(z_{ik} = 1 \mid y_i^{\text{obs}}, c_i; \theta^r, \psi^r, \pi^r) \propto \pi_k^r f_k(y_i^{\text{obs}}; \theta_k^r) \mathbb{P}(c_i \mid y_i^{\text{obs}}, z_{ik} = 1; \psi^r) \end{aligned}$$

EM algorithm for MNARz and MNARz^j

MNARz, MNARzj : needs some computations but still simple.

$$\mathbb{P}(c_{ij} = 1 \mid y_i, z_{ik} = 1; \psi) = \rho(\alpha_{kj}) \qquad (\textbf{independent of } y) \qquad (\Delta)$$

- Gaussian data : $(y_i | z_{ik} = 1; \theta^r) \sim \mathcal{N}(\mu_k, \Sigma_k)$
  - $\mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik} = 1, c_i; \theta^r, \psi^r) = \mathbb{P}(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik} = 1; \theta^r)$ **using** (Δ) and

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

    where $(\tilde{\mu}_{ik}^{\text{mis}})^r$ and $(\tilde{\Sigma}_{ik}^{\text{mis}})^r$ only depend on μ_k^r, Σ^r and y_i^{obs}
    $\Rightarrow E_{iy}^r(\theta) = \mathbb{E}\left[\log(f_k(y_i; \theta_k)) \mid y_i^{\text{obs}}, z_{ik} = 1, ; \theta^r\right]$ easy to compute (classical formulae)
  - **Using** (Δ)

$$E_{ic}^r(\psi) = \log(\mathbb{P}(c_i \mid z_{ik} = 1; \psi)) = \sum_{j=1}^{d} c_{ij} \log \rho(\alpha_{kj}) + (1 - c_{ij}) \log(1 - \rho(\alpha_{kj}))$$

  - **Using** (Δ)

$$(\tau_{ik})^r \propto \pi_k^r \phi(y_i^{\text{obs}}; (\mu_{ik}^{\text{obs}})^r, (\Sigma_{ik}^{\text{obs,obs}})^r) \prod_{j=1}^{d} \rho(\alpha_{kj}^r)^{c_{ij}} (1 - \rho(\alpha_{kj}^r))^{1 - c_{ij}}$$

EM algorithm for MNARz and MNARz^j

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

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

$$(\tilde{y}_{i,k})^r = (y_i^{\text{obs}}, (\tilde{\mu}_{ik}^{\text{mis}})^r) \qquad \tilde{\Sigma}_{ik}^r = \begin{pmatrix} 0_i^{\text{obs,obs}} & 0_i^{\text{obs,mis}} \\ 0_i^{\text{mis,obs}} & (\tilde{\Sigma}_{ik}^{\text{mis}})^r \end{pmatrix}$$

- M-step: for $k = 1, \ldots, K$, compute

$$\pi_k^{r+1} = \frac{1}{n}\sum_{i=1}^{n}(\tau_{ik})^r \qquad \mu_k^{r+1} = \frac{\sum_{i=1}^{n}(\tau_{ik})^r(\tilde{y}_{k,i})^r}{\sum_{i=1}^{n}(\tau_{ik})^r}$$

$$\Sigma_k^{r+1} = \frac{\sum_{i=1}^{n}\left[(\tau_{ik})^r\left((\tilde{y}_{i,k})^r - \mu_k^{r+1})((\tilde{y}_{i,k})^r - \mu_k^{r+1})^T + \tilde{\Sigma}_{ik}^r\right)\right]}{\sum_{i=1}^{n}(\tau_{ik})^r}$$

For ψ^{r+1}: maximization of $Q_c(\psi; \theta^r, \psi^r, \pi^r)$ over ψ with a **Newton-Raphson algorithm** (classical procedure for link functions of interest)

An EM algorithm can also be easily derived for categorical data

Not EM algorithm for MNAR$y*$

MNAR$y*$: needs approximations

$$\mathbb{P}(c_{ij} = 1 \mid y_i, z_{ik} = 1; \psi) = \rho(\alpha_{kj} + \beta_{kj} y_{ij}) \qquad (\textbf{not independent of } y)$$

- Gaussian data:
  - $(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik} = 1, c_i)$:
    ✗ not classical if ρ is Logit, ✓ truncated Gaussian distribution if ρ is Probit
  - No closed form of $E_{ic}^r(\psi)$ neither for Probit nor for Logit:

$$E_{ic}^r(\psi) = \sum_{j=1}^{d} c_{ij} \int_{\mathcal{Y}_{ij}^{\text{mis}}} \log(\rho(\alpha_{kj}+\beta_{kj} y_{ij}^{\text{mis}})) \frac{\rho(\alpha_{kj}^r + \beta_{kj}^r y_{ij}^{\text{mis}})^{c_{ij}} \mathbb{P}(y_{ij}^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik} = 1; \theta^r)}{\int_{\mathcal{Y}_{ij}^{\text{mis}}} \rho(\alpha_{kj}^r + \beta_{kj}^r x)^{c_{ij}} \mathbb{P}(x \mid y_i^{\text{obs}}, z_{ik} = 1; \theta^r) dx} dy_{ij}^{\text{mis}}$$
$$+ (1 - c_{ij}) \log(1 - \rho(\alpha_{kj} + \beta_{kj} y_{ij}^{obs}))$$

  ✗ not concave function if ρ is Logit
  - No closed form of $(\tau_{ik})^r$ neither for Probit nor for Logit

- In the Gaussian case, there is **no closed form** [Pirjol, 2013]
- SEM easier? random drawing instead of expectation

SEM algorithm for MNARy*

- SE-step: draw the missing data $((y_i^{\text{mis}})^{r+1}, z_i^{r+1}) \sim (. \mid y_i^{\text{obs}}, c_i; \theta^r, \psi^r, \pi^r)$

  Use of **One**-Gibbs sampling:

  - $(y_i^{\text{mis}})^{r+1} \sim (\cdot \mid y_i^{\text{obs}}, z_i^r, c_i; \theta^r, \psi^r)$:

    ✗ not classical if ρ is Logit, ✓ truncated Gaussian distribution if ρ is Probit
  - $z_i^{r+1} \sim (\cdot \mid y_i^{r+1}, c_i; \theta^r, \psi^r, \pi^r)$: draw the membership k of z_i^{r+1} from the **multinomial distribution** with probabilities

$$\mathbb{P}(z_{ik} = 1 | y_i^{r+1}, c_i; \theta^r, \psi^r, \pi^r) = \frac{\mathbb{P}(c_i | y_i^{r+1}, z_{ik} = 1; \psi^r)\mathbb{P}(y_i^{r+1} | z_{ik} = 1; \theta^r)\pi_k^r}{\sum_{h=1}^K \mathbb{P}(c_i | y_i^{r+1}, z_{ih} = 1; \psi^r)\mathbb{P}(y_i^{r+1} | z_{ih} = 1; \theta^r)\pi_h^r}$$

  Let $Y^{r+1} = (y_1^{r+1} | \ldots | y_n^{r+1})$, $Z^{r+1} = (z_1^{r+1} | \ldots | z_n^{r+1})$ be the imputed matrix and the partition

- M-step: for $k = 1, \ldots, K$, compute
  - π_k^{r+1} with the proportion of rows of Y^{r+1} belonging to class k
  - $\mu_k^{r+1}, \Sigma_k^{r+1}$ with the mean and covariance matrix of rows of Y^{r+1} belonging to class k
  - ψ^{r+1} with a Newton-Raphson algorithm

Summary for algorithms

	EM			SEM		
	Gaussian		Categorical	Gaussian		Categorical
MNARz MNARz^j	✓		✓	✓		✓
	Probit	Logit		Probit	Logit	
MNAR$y*$	no closed form	no closed form, optim. pb	not ident.	✓	require algorithms as SIR (costly)	not ident.

What about model selection?

Can select between MCAR and MNAR$*$ with any information criterion (BIC, ICL)

Even if the missing mechanism is ignorable for MCAR...

...need to model c to compare a MCAR and a MNAR model

CAUTION

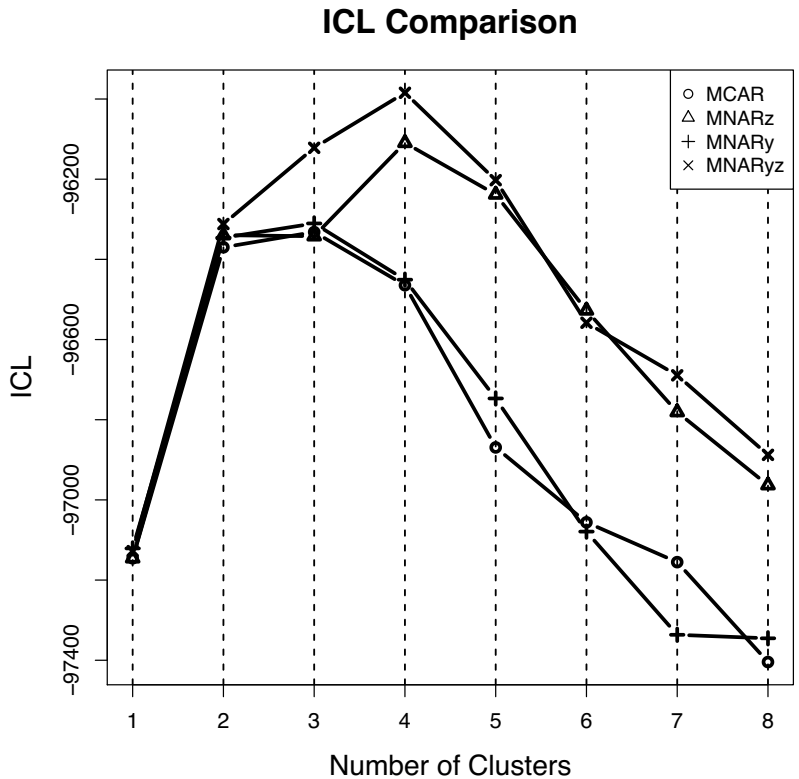
- It is just a selection between several proposed MNAR models
- It is not deciding if missingness procedure is "genererically" MNAR or not

Outline

Hospital Data: continuous features case

- Number of patients: $n = 5\ 146$
- Number of features: $d = 7$
  - Age
  - Size
  - Weight
  - Cardiac frequency
  - Hemoglobin concentration
  - Temperature
  - Minimum Diastolic and Systolic Blood Pressure
- Percentage of missing data: 6.4%

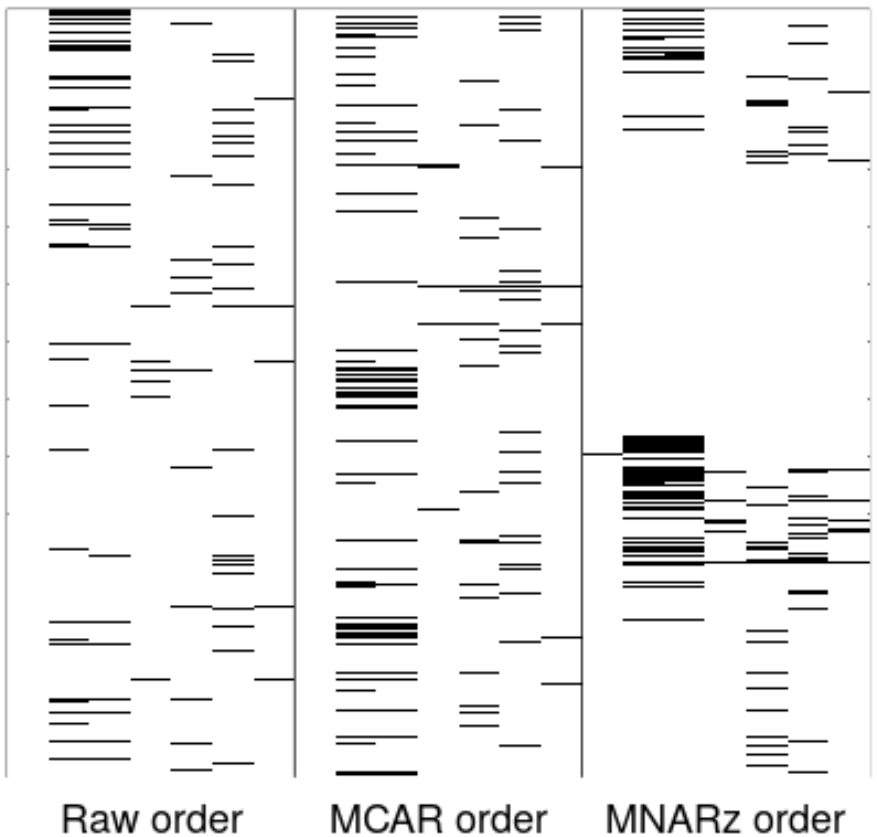
ICL comparison



- MCAR, MNARy and MNARz are equivalent until $K = 3$
- MNARz and MNARyz clearly indicate presence of an additional cluster ($K = 4$)

It seems to be an illustation of the effect of c through MNARz and MNARyz

Missing Pattern



It seems that MNAR*z* modelling leads to a missing free cluster

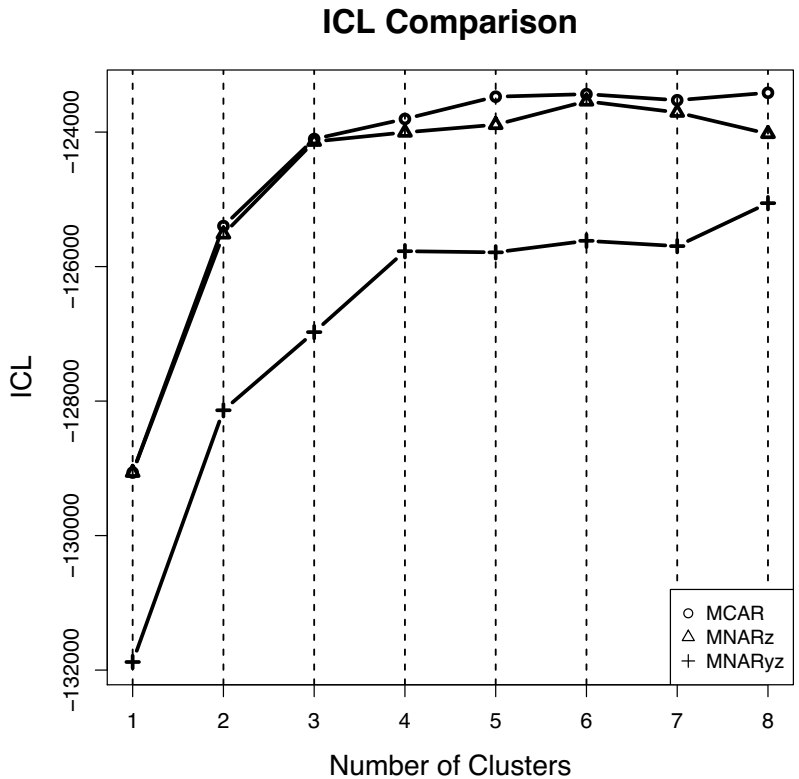
Hospital data: mixed features case

- Number of patients: $n = 5\ 146$
- Number of features: $d = 15$ (7 continuous and 8 categorical)
- Percentage of missing data: $\sim 4\%$

Model

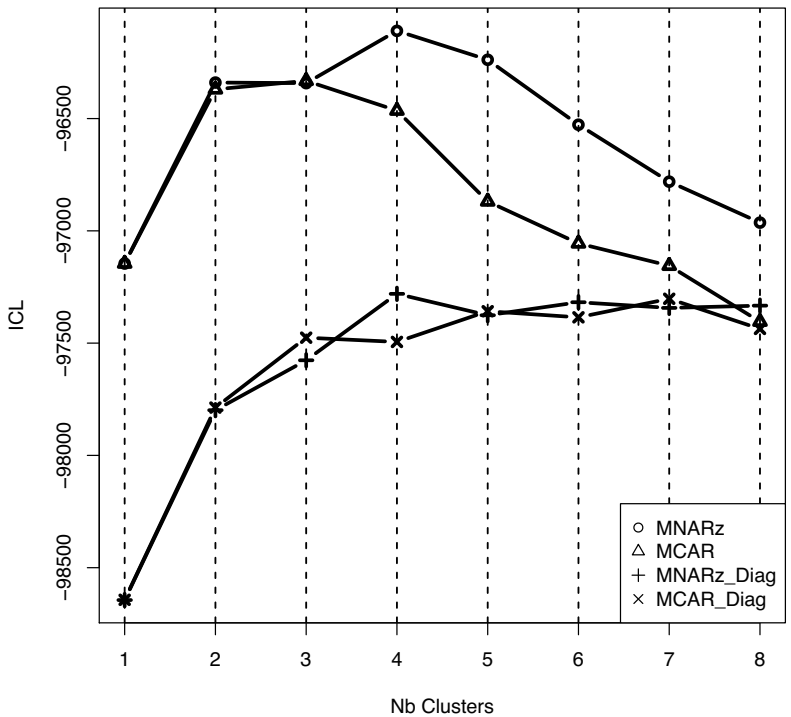
Conditional independence of the variables knowing the cluster

ICL comparison



- MCAR and MNARz are equivalent
- MNARyz seems really inappropriate
- Seems to miss the previous latent structure: requires a specific exploration...

Exploration 1: local independence is not relevant for this data set



Not accounting for possible conditional dependencies between the continuous variables is inappropriate for this dataset.

Exploration 2: mixture model bias *vs* missing model bias

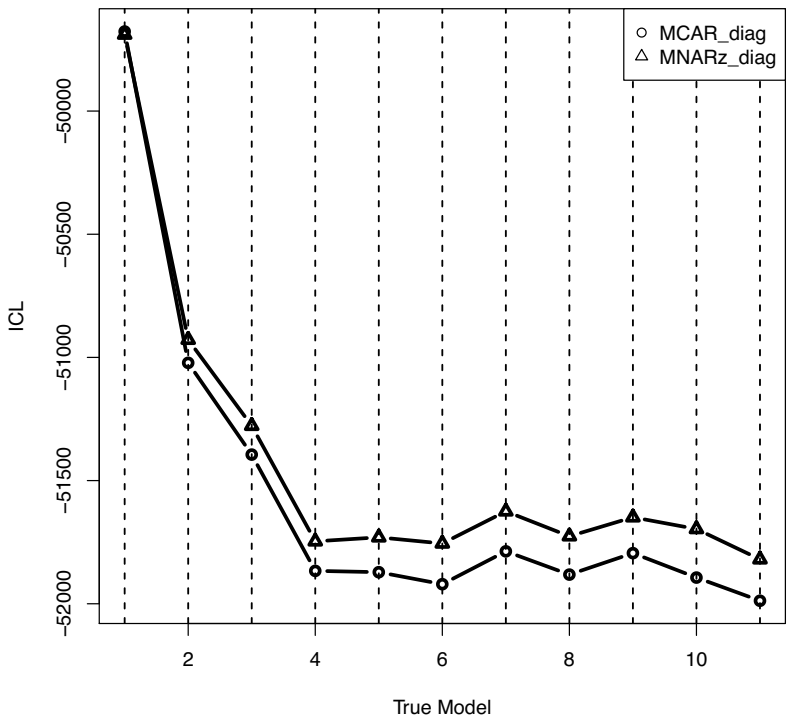
A simulated data set with the following parameters:

- Mixture model: varying from diagonal to non diagonal hypotheses
  - 2 clusters, dimension 7
  - 5 000 individuals
  - $\pi_1 = 0.3$, $\pi_2 = 0.7$, $\mu_1 = (0,0,0,0,0,0,0)$, $\mu_2 = (2,2,2,2,2,2,2)$
  - Covariance matrices: $r \in \{1, \ldots, 10, \infty\}$ ($r = \infty$ is the diagonal case)

$$\Sigma_k^{(r)} = \begin{pmatrix} 1 & 0.5^r & 0.25^r & \ldots & . & . & . \\ 0.5^r & 1 & 0.5^r & \ldots & . & . & . \\ 0.25^r & 0.5^r & 1 & \ldots & . & . & . \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ . & . & . & \ldots & 0.25^r & 0.5^r & 1 \end{pmatrix}$$

- Missingness model: MNAR hypothesis
  - Proportions of missingness: 0.001 for cluster 1 and 0.06 for cluster 2

Results from exploration 2



- The mixture model bias can not be compensated by the (unbiased) missing mechanism modeling
- It illustrates again the fact that information (on the latent partition) conveys by data is much more important than information conveys by the pattern

Outline

Summary

- Interest to put a model on c
- Interest of the simple but meaningful model MNARz
- Link between our models and usual methods

Ongoing works

- Deeper analysis of the previous results with doctors...
- Implement the proposed models/algo. in the Mixmod software[a]
- Address the trade-off between biased mixture model and biased missingness mechanism in particular for the mixed data case

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

References


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.


Jones, M. P. (1996).
Indicator and stratification methods for missing explanatory variables in multiple linear regression.
Journal of the American statistical association, 91(433):222–230.


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


Little, R. J. and Rubin, D. B. (2002).
Statistical Analysis with Missing Data.
Wiley.


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.


Pirjol, D. (2013).
The logistic-normal integral and its generalizations.
Journal of Computational and Applied Mathematics, 237(1):460–469.