Impact of Missing Data on Mixtures and Clustering

Christophe Biernacki

▶ **To cite this version:**

Christophe Biernacki. Impact of Missing Data on Mixtures and Clustering. CMStatistics 2022 - 15th International Conference of the ERCIM WG on Computational and Methodological Statistics, Dec 2022, London, United Kingdom. hal-03936786

HAL Id: hal-03936786
https://inria.hal.science/hal-03936786

Submitted on 12 Jan 2023

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.

Impact of Missing Data on Mixtures and Clustering

C. Biernacki

Laboratoire P. Painlevé, UMR CNRS 8524 & Université de Lille & Inria

(with C. Boyer, G. Celeux, J. Josse, F. Laporte, M. Marbac, A. Sportisse, V. Vandewalle)

CMStatistics 2022
17-19 December 2022, King's College London, UK

Take home message

Missing data may change preconceptions

- Mixtures: EM has unexpected behaviour concerning degeneracy dynamic
- Clustering: the missing data pattern may convey some information on partition

These topics are in the research agenda of Statisticians since:

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

Outline

Outline

Missing data: notations

- $y = \{y_1 | \ldots | y_n\}^T$: full dataset with n individuals
- $y_i = (y_{i1}, \ldots, y_{id}) \in \mathcal{Y} = \mathbb{R}^d$ with individual $i \in \{1, \ldots, n\}$
- $c = \{c_1 | \ldots | c_n\}^T \in \{0,1\}^{n \times d}$: 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 indiv. i (and $y^{\text{obs}} = \{y_1^{\text{obs}} | \ldots | y_n^{\text{obs}}\}^T$)
- y_i^{mis}: the missing variables values for individual i (and $y^{\text{mis}} = \{y_1^{\text{mis}} | \ldots | y_n^{\text{mis}}\}^T$)

Typology of the missingness mechanisms

- Missing completely at random (MCAR):

$$\mathrm{p}(c|y;\psi) = \mathrm{p}(c;\psi) \quad \forall y$$

- Missing at random (MAR):

$$\mathrm{p}(c|y;\psi) = \mathrm{p}(c|y^{\mathrm{obs}};\psi) \quad \forall y^{\mathrm{mis}}$$

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

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

Mixture model and clustering

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

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

  - $\pi_k = \mathrm{p}(z_{ik}=1)$, $\pi = (\pi_1,\ldots,\pi_K)$
  - $f_k(.;\lambda_k) = \phi(.;\mu_k,\Sigma_k)$ is the d-variate **Gaussian distribution** with mean vector μ_k and covariance matrix Σ_k
  - $\theta = (\ (\pi)_k\ ,(\lambda)_k\ )$ is the whole mixture parameter
- Clustering: MAP principle from the mixture output to estimate the partition

Outline

Degeneracy genesis: unbounded likelihood

- d-variate Gaussian mixture

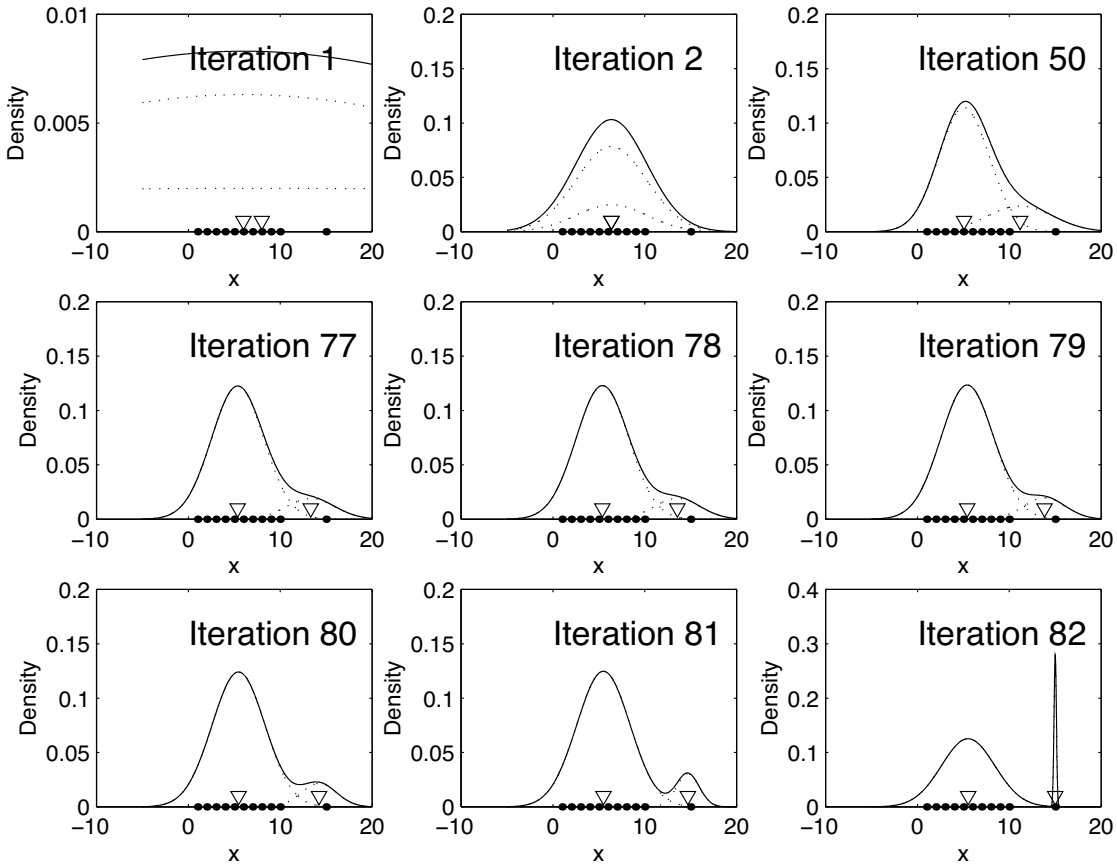
$$f(y_i;\theta) = \sum_{k=1}^{K} \pi_k \underbrace{\frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\Big(-\frac{1}{2}(y_i-\mu_k)^T\Sigma_k^{-1}(y_i-\mu_k)\Big)}_{\phi(y_i;\mu_k,\Sigma_k)}$$

- Sampling: $y_1,\ldots,y_n \overset{i.i.d.}{\sim} p(.;\theta)$ without any missing data ($y^{\text{obs}} = y$)
- Likelihood: $\ell(\theta;y) = \ln L(\theta;y) = \sum_{i=1}^{n} \ln f(y_i;\theta)$

$$\text{particular center } \mu_2 = y_i \quad \Rightarrow \quad \lim_{|\Sigma_2|\to 0} \ell(\theta;y) = +\infty$$

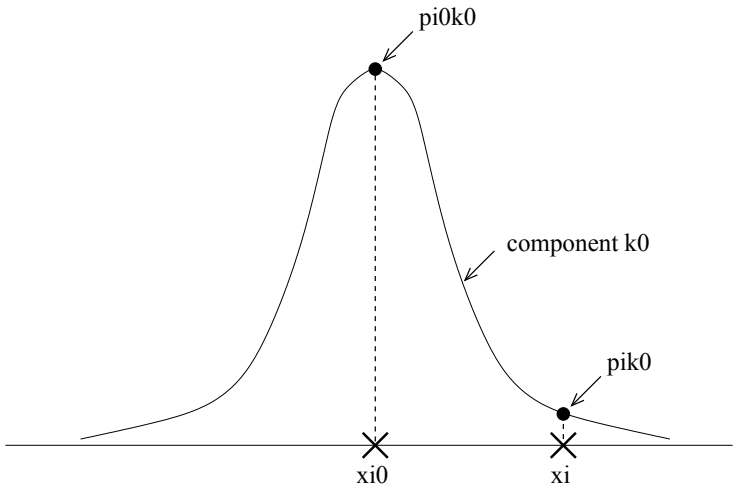
[Kiefer and Wolfowitz, 1956] [Day, 1969]

EM behaviour: illustration



- degeneracy may occur even when starting from large variances
- convergence can be slow when far from the degenerate limit
- convergence extremely fast near degeneracy

EM behaviour: results



$$u_0 = \left[\frac{1}{p_{i_0 k_0}}, \{p_{ik_0}\}_{i \neq i_0} \right]$$

degeneracy of component k_0 at y_{i_0}

$\Leftrightarrow$

$\|u_0\| \to 0$

[Biernacki and Chrétien, 2003]

[Ingrassia and Rocci, 2009]

Proposition 1: Existence of a bassin of attraction

$\exists \epsilon > 0$ *s.t. if* $\|u_0\| \leq \epsilon$ *then* $\|u_0^+\| = o\|u_0\|$ *with probability 1.*

Proposition 2: Speed towards degeneracy is exponential

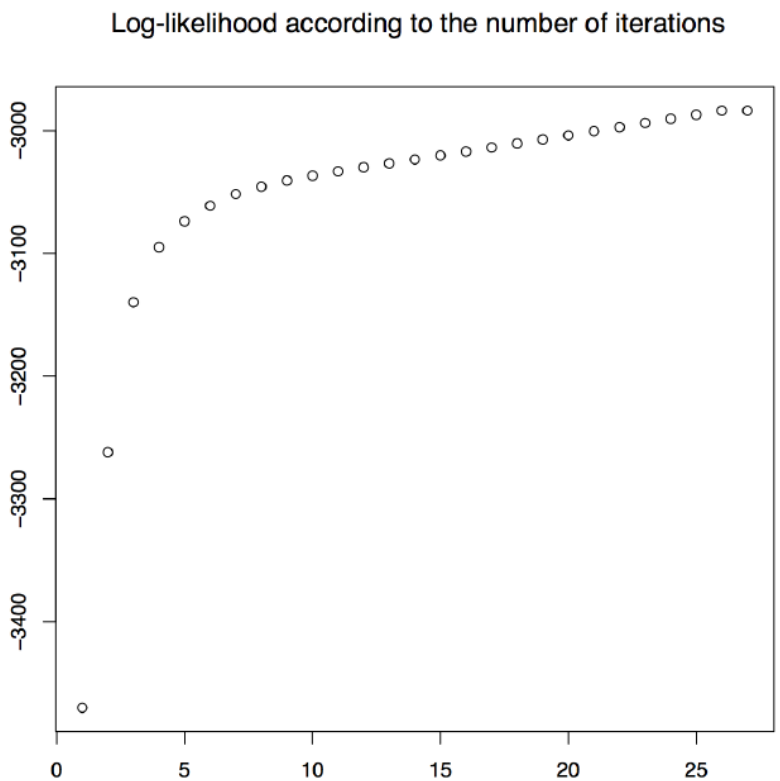
$\exists \epsilon > 0$, $\alpha > 0$ *and* $\beta > 0$ *s.t. if* $\|u_0\| \leq \epsilon$ *then, with probability 1,*

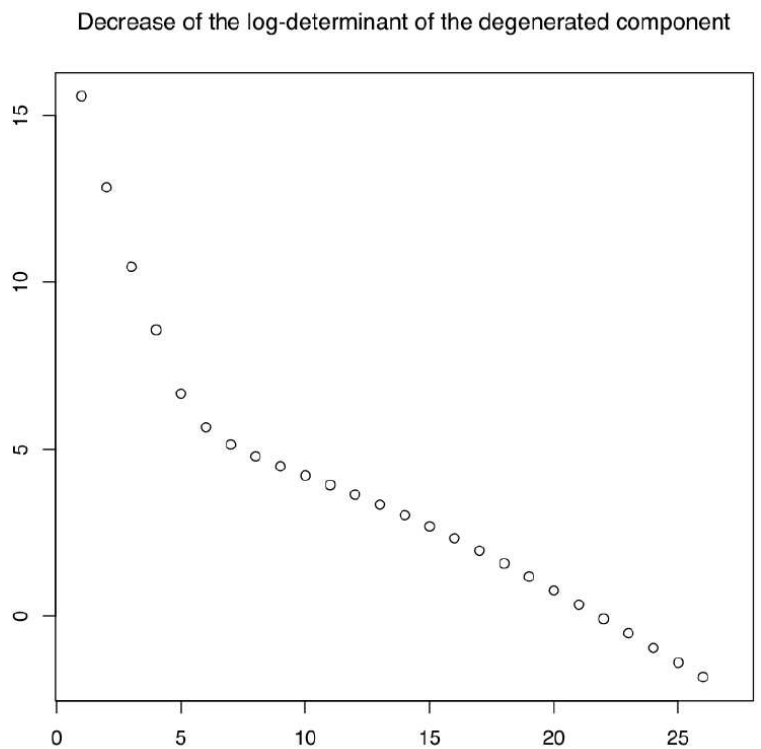
$$|\Sigma_{k_0}^+| \leq \alpha / |\Sigma_{k_0}| \cdot \exp\Big(- \beta / |\Sigma_{k_0}| \Big).$$

Outline

EM behaviour illustration

- Breast cancer tissue of the UCI database repository: 106 units, 9 variables.
- 10% of missing data randomly generated
- $K = 4$ clusters





Detail from the illustration

	1	2	3	4	5	6	7	8	9
1	211.00		0.09	30.75	151.98	4.94	14.27	27.24	217.13
2	196.86	0.02	0.09	28.59	82.06	2.87	7.97	27.66	200.75
3	144.00	0.12	0.05	19.65	70.43	3.58		7.57	160.37
4	172.52	0.13	0.04		192.22	5.12	19.32	32.19	174.93
5	121.00	0.17	0.09	24.44	144.47	5.91	22.02	10.59	141.77
6	223.00	0.12	0.08	33.10	197.01	5.95	30.45	12.96	252.48
7		0.17	0.23	34.22	94.35	2.76	31.28	13.88	180.61
8	303.00	0.06	0.04	22.57		4.54	21.83	5.72	321.65
9	250.00	0.09	0.09	29.64	180.76	6.10	26.14	13.96	280.12
10	391.00	0.06	0.01	35.78		7.41	22.13	28.11	400.99
11	176.00	0.09	0.08	20.59	79.71		18.23	9.58	191.99
12	145.00		0.11	21.22	82.46	3.89	20.30	6.17	162.51
13	124.13	0.13	0.11	20.59			18.46	9.12	134.89
14	103.00	0.16	0.29	23.75	78.26	3.29	22.32	8.12	124.98

Table : Data belonging to the degenerated component.

- Cvg. towards a degenerated component (no plateau of the log-likelihood)
- Degeneracy relatively slow: log-likelihood linear according to the nb of it.
- Number of points of the degenerated solution greater than the space dimension d (but the number of complete points lower than d)

Intermediate conclusion on missing data

Like the complete data y case

- Likelihood is unbounded
- EM can be attracted by degenerate solutions

Unlike the complete data y case

- Risk to consider a degenerated solution as valid
- Risk of losing a lot of time in useless iterations

Statisticians should be aware of such dangerous EM behaviour...

...since missing data are more and more frequent

Understanding degeneracy speed on a toy example

- Univariate framework, no mixture, only one observed data: y
- Maximum likelihood estimator (Unbounded likelihood!): $\mu = y$, $\sigma^2 = 0$
- Suppose equivalently that $n-1$ data are unobserved (unchanged likelihood)
- Here is one iteration of a (useless) EM algorithm (it. q)

$$\mu^{(q+1)} = \frac{(n-1)\mu^{(q)} + y}{n} \quad \text{and} \quad \sigma^{2(q+1)} = \frac{(n-1)\sigma^{2(q)} + (y - \mu^{(q+1)})^2}{n}$$

Linear grow of the log-likelihood (have a look also when n increases!)

$$\ell(\theta^{(q)}; x) \sim -0.5q \log \frac{n-1}{n}$$

Geometrical convergence rate towards 0 for the variance

$$\sigma^{2(q)} \sim \sigma^{2(0)} \left(\frac{n-1}{n}\right)^q$$

Influence of the missing data rate

% missing data	0	5	10	15	20	25	30
% deg.	16	4	12	11	46	51	100
Average nb of iterations before deg.	2	13	13	82	304	138	215

Table : Frequency and speed of degeneracy (deg.) according to the rate of missing data on the breast cancer data set.

When the rate of missing data increases:

- The rate of degeneracy increases
- The number of iterations before degeneracy seems to (globally) increase

Again, statisticians should be aware of such dangerous EM behaviour. . .

. . . since missing data are more and more frequent

Outline

Proposed MNAR models in clustering

Question we address now

Since MNAR is not ignorable, which distribution $\mathrm{p}(c|y, z; \psi)$ to propose?

Hypothesis 1: conditional independence

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

Hypothesis 2: linear function within canonical link functions ρ

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

- $\psi = (\alpha, \beta)$ where $\alpha = (\alpha_{11}, \ldots, \alpha_{1d}, \ldots, \alpha_{K1}, \ldots, \alpha_{Kd})^T \in \mathbb{R}^{Kd}$ and $\beta = (\beta_{11}, \ldots, \beta_{1d}, \ldots, \beta_{K1}, \ldots, \beta_{Kd})^T \in \mathbb{R}^{Kd}$
- ρ is the cdf of any continuous distribution (logit, probit)

A by-product zoology of 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
MNAR$z^j y^k$	✓	✓	✓	✓	$2Kd$
MNARyz^j	✓	✗	✓	✓	$(K+1)d$
MNAR$y^k z$	✓	✓	✗	✓	$K(d+1)$
MNARyz	✓	✗	✗	✓	$(K+d)$
MNARy	✓	✗	✗	✗	d
MNARy^k	✓	✓	✗	✗	Kd
MNARz	✗	✗	✗	✓	K
MNARz^j	✗	✗	✓	✓	Kd

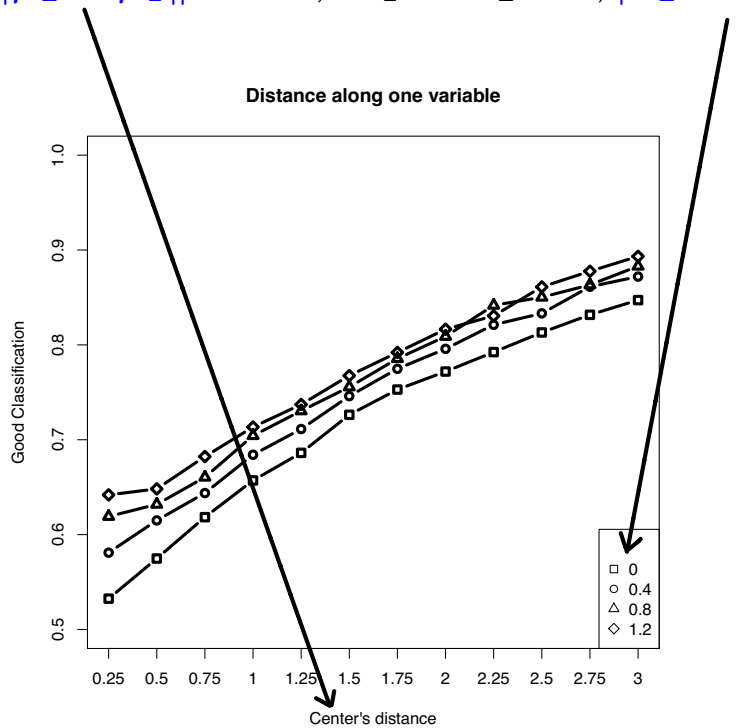
Remarks:

- MNAR$z^j y^k$ is the most complex model
- MNARz, MNARz^j: the only effect of missingness is on the class membership, $\psi = (\alpha_{11}, \ldots, \alpha_{1d}, \ldots, \alpha_{K1}, \ldots, \alpha_{Kd})^T$, $\mathrm{p}(c_{ij} = 1 \mid y_i, z_{ik} = 1; \psi) = \rho(\alpha_{kj})$
- MCAR is a specific and simple case

MNARz analysis: pattern c gives information on partition z!

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

$\pi_k = 0.5$, $\|\mu_2 - \mu_1\|$ varies, $\Sigma_1 = \Sigma_2 = \mathbf{I}$, $|\alpha_2 - \alpha_1|$ varies



Both μ_k and α_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 3: 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.

Identifiability

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

Proposition 4

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

All MNAR models are identifiable (or at least generically identifiable) for probit/logit

Estimation procedure overview

- Use EM or Stochastic EM (SEM) algorithms
- MNARz and MNARz^j: EM and SEM are very simple
- MNAR$y*$: the SE step requires a within Gibbs loop, sometimes involving itself a Sampling Importance Resampling (SIR)

	EM		SEM	
MNARz MNARz^j	✓		✓	
	Probit	Logit	Probit	Logit
MNAR$y*$	no closed form	no closed form, optim. pb	✓	require algorithms as SIR (costly)

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 "generically" MNAR or not

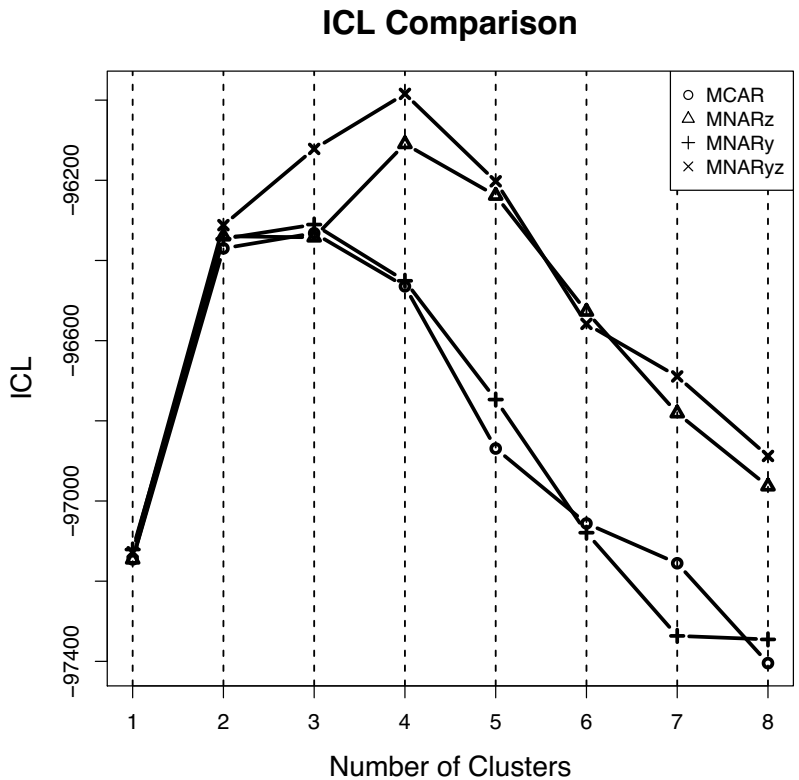
Outline

Hospital data description

- 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%

Doctors are convinced that their missing data are MNAR

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

Outline

To conclude

Summary

- Statisticians should properly consider the potential missing data impact both from an algorithmic and from a modeling point of view
- EM: be careful about degeneracy which seems to exacerbated/masqued
- MNAR: interest of the simple but meaningful model MNARz, link with usual methods

Ongoing works

- EM: propose mechanism to identify/discard degenerate runs
- MNAR: extend to categorical, count and mixed data

Thanks!