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

Christophe Biernacki, Gilles Celeux, Julie Josse, Fabien Laporte

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

Christophe Biernacki, Gilles Celeux, Julie Josse, Fabien Laporte

Final CRoNoS meeting and Workshop on Multivariate Data Analysis and Software
14-16 April 2019, Limassol, Cyprus

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: expect to add variance into analysis
- Impute missing data: expect to add bias modeling into analysis

General guidelines

- Obtaining a non-missing 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

- $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$: full dataset with n individuals
- $\mathbf{y}_i = (y_i^1, \ldots, y_i^d) \in \mathbb{R}^d$: full individual $i \in \{1, \ldots, n\}$
- $\mathbf{c} = \{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$: pattern of missing data for the full dataset
- $\mathbf{c}_i = (c_i^1, \ldots, c_i^d) \in \{0, 1\}^d$: pattern of missing data for individual $i \in \{1, \ldots, n\}$

$$c_i^j = 1 \Leftrightarrow y_i^j \text{ is missing}$$

- $o_i = \{j : c_i^j = 0\}$: the observed variables indexes for individual i
- $\mathbf{y}_i^{o_i}$: the observed variables values for individual i
- $\mathbf{y}^o = \{\mathbf{y}_1^{o_1}, \ldots, \mathbf{y}_n^{o_n}\}$: the observed values in $\mathbf{y}$
- $m_i = \{j : c_i^j = 1\}$: the missing variables indexes for individual i
- $\mathbf{y}_i^{m_i}$: the missing variables values for individual i
- $\mathbf{y}^m = \{\mathbf{y}_1^{m_1}, \ldots, \mathbf{y}_n^{m_n}\}$: the missing values in $\mathbf{y}$

$$\mathbf{y} = \{\mathbf{y}^o, \mathbf{y}^m\} \text{ is the full dataset with its observed and missing parts}$$

Missing data: typology of the missing mechanisms

- Missing completely at random (MCAR):

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

- Missing at random (MAR):

$$P(\mathbf{c}|\mathbf{y};\boldsymbol{\psi}) = P(\mathbf{c}|\mathbf{y}^{o};\boldsymbol{\psi}) \quad \forall \boldsymbol{y}^{m}$$

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

Clustering: model-based approach

- Partition with K clusters: $\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)$ where
  - $\mathbf{z}_i = (z_i^1, \dots, z_i^K) \in \{0,1\}^K$
  - $z_i^k = 1$ if $\mathbf{y}_i$ belongs to cluster k, $z_i^k = 0$ otherwise
- Gaussian mixture: $\mathbf{y}_1, \dots, \mathbf{y}_n$ are i.i.d. from the mixture

$$f(\mathbf{y}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \phi_k(\mathbf{y}_i; \theta_k)$$

where:

- $\pi_k = P(z_i^k = 1)$
- $\phi_k(.; \theta_k)$: d-variate Gaussian pdf with mean vector and covariance matrix $\theta_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, d-multinomial pdf with $\theta_k = p_k$ probabilities vector or mixed pdf.
- $\boldsymbol{\theta} = (\pi_1, \dots, \pi_K, \theta_1, \dots, \theta_K)$

Question we address in this work

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

Outline

Logistic model: a natural and flexible candidate

$$P(\mathbf{c}|\mathbf{y},\mathbf{z};\boldsymbol{\psi}) = \prod_{i=1}^{n}\prod_{j=1}^{d} P(c_i^j|\mathbf{y},\mathbf{z};\boldsymbol{\psi})$$

- MCAR, with $\boldsymbol{\psi} = \alpha_0$

$$\text{logit}(P(c_i^j = 1|\mathbf{y},\mathbf{z};\boldsymbol{\psi})) = \alpha_0$$

- MNAR**z** (MNAR**z**^{j}), with $\boldsymbol{\psi} = (\alpha_0, \beta_1^{1...d}, \ldots, \beta_K^{1...d})$

$$\text{logit}(P(c_i^j = 1|\mathbf{y},\mathbf{z};\boldsymbol{\psi})) = \alpha_0 + \sum_{k=1}^{K} \beta_k^j z_i^k$$

- MNAR**y**, with $\boldsymbol{\psi} = (\alpha_0, \alpha_1, \ldots, \alpha_d)$

$$\text{logit}(P(c_i^j = 1|\mathbf{y},\mathbf{z};\boldsymbol{\psi})) = \alpha_0 + \alpha_j y_i^j$$

- MNAR**yz**, with $\boldsymbol{\psi} = (\alpha_0, \alpha_1, \ldots, \alpha_d, \beta_1, \ldots, \beta_K)$

$$\text{logit}(P(c_i^j = 1|\mathbf{y},\mathbf{z};\boldsymbol{\psi})) = \alpha_0 + \alpha_j y_i^j + \sum_{k=1}^{K} \beta_k z_i^k$$

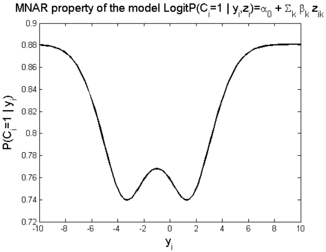
MNAR**z** analysis: it depends on **y** through **z**!

$$P(c_i^j = 1|\mathbf{y};\boldsymbol{\theta},\psi) = \sum_{k=1}^{K} P(c_i^j = 1|\mathbf{y},\mathbf{z};\psi)P(\mathbf{z}|\mathbf{y};\boldsymbol{\theta})$$

Example of a univariate Gaussian model with the three components

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

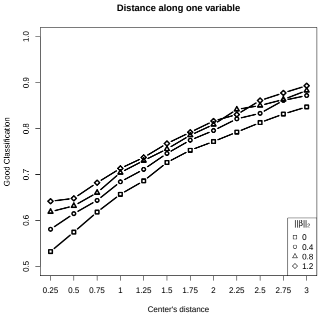
and with parameters of the logit expression: $\alpha_0 = 1, \beta_1 = 1, \beta_2 = -1, \beta_3 = 1$



MNAR**z** analysis: pattern **c** gives information on partition **z**!

Draw Bayes error of a MNAR**z** model with two components and 20% of missing data

$$\pi_k = 0.5,\ \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\| \text{ varies},\ \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I},\ |\beta_2 - \beta_1| \text{ varies}$$



Both $\boldsymbol{\mu}_k$ and β_k act on the Bayes error

Outline

Ignorable *vs.* non ignorable model

A missing mechanism is ignorable if likelihoods can be decomposed as

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

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

Inference of $\boldsymbol{\theta}$

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

- MCAR is ignorable
- MNAR**z**, MNAR**y** and MNAR**yz** are non ignorable

EM algorithm: looks simple

Decomposition of $Q(\boldsymbol{\theta}, \boldsymbol{\psi}; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})$

The expected complete log-likelihood conditional related to observed data is:

$$E\left[L(\boldsymbol{\theta}, \boldsymbol{\psi}; \mathbf{c}, \mathbf{y}, \mathbf{z}); \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}} | \mathbf{y}^o, \mathbf{c}\right] = Q_y(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) + Q_c(\boldsymbol{\psi}; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}})$$

$$Q_y(\boldsymbol{\theta}; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_i^k E\left[\log(\pi_k \phi_k(\mathbf{y}_i; \theta_k)) | \mathbf{y}_i^{o_i}, \mathbf{c}_i; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}\right]$$

$$Q_c(\boldsymbol{\psi}; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_i^k E\left[\log(P(\mathbf{c}_i | z_i^k = 1, \mathbf{y}_i; \boldsymbol{\theta}, \boldsymbol{\psi})) | \mathbf{y}_i^{o_i}, \mathbf{c}_i; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}\right]$$

$$\tau_i^k = P(z_i^k = 1 | \mathbf{c}_i, \mathbf{y}_i^{o_i}; \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\psi}}) = \frac{\widehat{\pi}_k \phi_k(\mathbf{y}_i^{o_i}; \theta_k^{o_i}) P(\mathbf{c}_i | z_i^k = 1, \mathbf{y}_i^{o_i}; \widehat{\boldsymbol{\psi}})}{\sum_{h=1}^{K} \widehat{\pi}_h \phi_h(\mathbf{y}_i^{o_i}) P(\mathbf{c}_i | z_i^h = 1, \mathbf{y}_i^{o_i}; \widehat{\boldsymbol{\psi}})}$$

EM and/or SEM algorithms

- MCAR (and also MAR...): classical formula! .... (EM , SEM)

$$\tau_i^k \propto \widehat{\pi}_k \phi_k(\mathbf{y}_i^{o_i}; \theta_k^{o_i})$$

- MNAR**z**: needs some new calculus but still simple .... (EM , SEM)

$$\tau_i^k \propto \widehat{\pi}_k \phi_k(\mathbf{y}_i^{o_i}; \theta_k^{o_i}) \prod_{j=1}^{d} (1 + \exp(-r_i^j \widehat{\beta}_k))^{-1} \text{ where } r_i^j = \begin{cases} 1 & \text{if } c_i^j = 1 \\ -1 & \text{otherwise} \end{cases}$$

- MNAR**y**: needs approximations .... (~~EM~~ , SEM)

$$P(c_i^j | \mathbf{y}_i^{o_i}, z_i^k = 1; \boldsymbol{\psi}) = \begin{cases} \int_{-\infty}^{+\infty} \frac{1}{1+\exp(-(\alpha_j y_i^j))} \phi_k(y_i^j; \theta_k^j) dy_i^j & \text{if } c_i^j = 1 \\ \frac{1}{1+\exp(\alpha_j y_i^j)} & \text{otherwise} \end{cases}$$

  - In the Gaussian case, there is **no closed form** [Pirjol, 2013] (same for MNAR**yz**)
  - But SEM is still simple in that case thanks to random drawing instead of expectation

Link with some usual procedures!

Concatenation [Jones, 1996]: model equivalence

$$\text{MNARz}^j(\mathbf{y}^o, \mathbf{c}) \iff \text{MCAR}\left(\quad (\mathbf{y}^o|\mathbf{c}) \quad \right)$$

$$\text{MNARz}(\mathbf{y}^o, \mathbf{c}) \iff \text{MCAR}\left(\quad \left(\mathbf{y}^o \middle| \left(\sum_{j=1}^{d} \mathbf{c}_{\cdot}^{j}\right)\right) \quad \right)$$

"All Available Cases" [Little and Rubin, 2002]: estimation equivalence

In case of conditional independence between variables, whatever MCAR or MNAR*:

$$\text{Classical (S)EM} \iff \text{(S)EM without estimating missing } \mathbf{y}^m$$

. . . an opportunity to reduce the computing time

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. . .

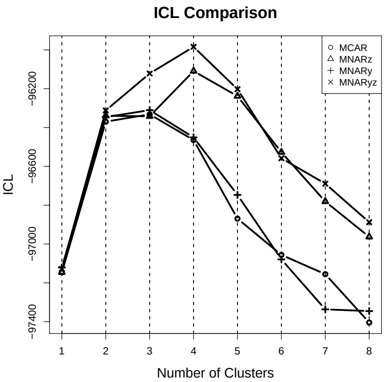
. . . need to model **c** to compare a MCAR and a MNAR model

Outline

Hospital Data

- 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%

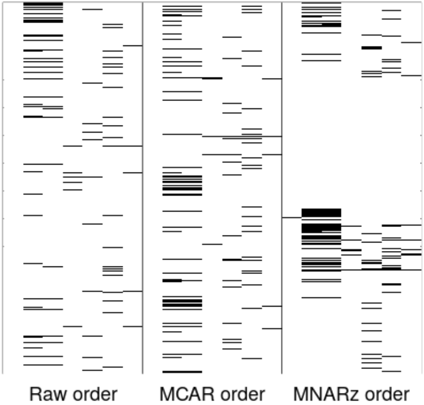
ICL comparison



- MCAR, MNAR**y** and MNAR**z** are equivalent until $K = 3$
- MNAR**z** and MNAR**yz** clearly indicate presence of an additional cluster ($K = 4$)

It seems to be an illustation of the effect of **c** through MNAR**z** and MNAR**yz**

Missing Pattern



It seems that MNARz modelling leads to a missing free cluster

Outline

Summary

- Interest to put a model on **c**
- Interest of the simple but meaningful model MNAR**z**
- 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]
- Use mixed data algorithms for medical study with the same MNAR* models

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

References

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.

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

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