Model-based clustering with missing not at random data. Missing mechanism

Fabien Laporte, Christophe Biernacki, Gilles Celeux, Julie Josse

Model-based clustering with missing not at random data.

Fabien Laporte [1,2], Christophe Biernacki [2], Gilles Celeux,[3] and Julie Josse,[1,4]
[1]CMAP, UMR 0320 / UMR 87641 Centre de Mathématiques Appliquées, 91128 Palaiseau, France
[2]MODAL project team, INRIA Lille, 59650 Villeneuve d'Ascq, France
[3]INRIA Saclay - Île-de-France, 91120 Palaiseau, France
[4] XPOP project team, INRIA Saclay - Île-de-France, 91120 Palaiseau, France

Summary

Since the 90s, model-based clustering is largely used to classify data. Nowadays, with the increase of available data, missing values are more frequent. We defend the need to embed the missingness mechanism directly within the clustering modeling step. There exist three types of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In all situations, logistic regression is proposed as a natural and flexible candidate model. In this unified context, standard model selection criteria can be used to select between such different missing data mechanisms, simultaneously with the number of clusters. Practical interest of our proposal is illustrated on data derived from medical studies suffering from many missing data.

Dataset Example

$$\mathbf{y} = \begin{pmatrix} 18 & ? & blue & 175 \\ 25 & 63 & green & 162 \\ ? & 56 & brown & 164 \\ 22 & 79 & ? & 182 \end{pmatrix}$$

$$\mathbf{c} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$\mathbf{o}_1 = (1, 3, 4)$
$\mathbf{y}_1^{\mathbf{o}_1} = (18, blue, 175)$
$\mathbf{m}_1 = (2)$
$\mathbf{y}^{\mathbf{o}} = \{\mathbf{y}_1^{\mathbf{o}_1}, \ldots, \mathbf{y}_4^{\mathbf{o}_4}\}$
$\mathbf{y}^{\mathbf{m}} = \{\mathbf{y}_1^{\mathbf{m}_1}, \ldots, \mathbf{y}_4^{\mathbf{m}_4}\}$

Missing mechanism (1)

Missing Completely at random (MCAR):
$P(c_i|\mathbf{Y}) = P(c_i)$

Missing at random (MAR):
$P(c_i|\mathbf{Y}) = P(c_i|\mathbf{Y}^o)$

Missing not at random (MNAR):
Not MCAR nor MAR

References

[1] Little, R.J.A. and Rubin, D.B. Statistical analysis with missing data. 1986. *John Wiley & Sons, Inc. New York, NY, USA*, 1986

Model

Model-based clustering:

$$P(y_i; \theta) = \sum_{k=1}^{K} \pi_k \phi_k(y_i)$$

where $\star$ $\pi_k = P(z_i^k = 1)$,

$\star$ ϕ_k are gaussian or multinomial distributions, with parameters μ_k and Σ_k or p_k.

$\star$ $\theta = (\pi_1, \ldots, \pi_K, (\mu_1, \Sigma_1)/p_1, \ldots, (\mu_K, \Sigma_K)/p_K)$

Missing mechanism models:

$\text{logit}(P(c_i^j|y_i, z_i; \psi)) = \alpha_0$, $\psi = \alpha_0$ (MCAR)
$\text{logit}(P(c_i^j|y_i, z_i = 1; \psi)) = \alpha_0 + \beta z_i$, $\psi = (\alpha_0, \beta_1, \ldots, \beta_K)$ (MNAR**z**)
$\text{logit}(P(c_i^j|y_i, z_i; \psi)) = \alpha_0 + \alpha_j y_i^j + \beta z_i$, $\psi = (\alpha_0, \ldots, \alpha_d, \beta_1, \ldots, \beta_K)$ (MNAR**yz**)

Inference: EM-like algorithm

Separability of the update:

$$Q_{Y,C,Z}(\theta, \psi|\theta^{(t)}, \psi^{(t)}) = Q_{Y,Z}(\theta|\theta^{(t)}, \psi^{(t)}) + Q_C(\psi|\theta^{(t)}, \psi^{(t)})$$

$\Longrightarrow$ Maximize on θ and ψ separatly.

E step simple for MCAR and MNARz:

$$\tau_i^{k,(t)} \propto \pi_k \phi_k(y_i^{o_i}) P(c_i|y_i^{o_i}, z_i; \psi^{(t)})$$

Simpler way: SEM algorithm

Link with Classical Methods

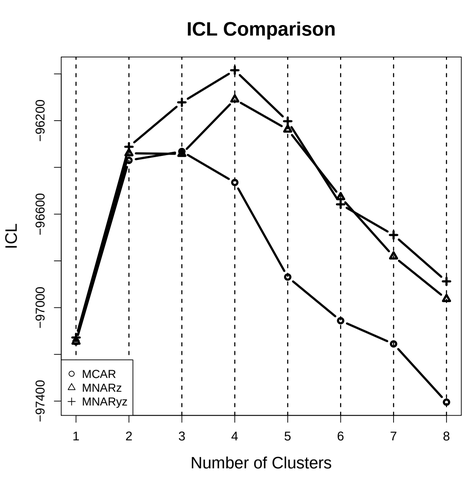
All available cases:
Conditional independence of the variables knowing the cluster $\Longleftrightarrow$ AAC to update θ in each step

Concatenation:
MBC on Y with a MNAR**z** mechanism $\Longleftrightarrow$ MBC on $(Y|C)$ with a MAR mechanism

Application

Dataset:
Number of individuals: $n = 5\ 146$
Number of feature: $d = 7$
Percentage of missing data: 6.4%

Results:
MCAR and MNARz are equivalent until K = 3
MNARz and MNARyz clearly indicate presence of an additional cluster (K = 4)



Conclusion

The method is implemented in the **R** Package "Relatedness", that performs inference on phased/unphased genotypic data, and accounts for a potential population structure. Core functions are developed in C++ and parallelized to speed-up inference on large panels. More information is available in *F. Laporte, C. Charcosset and T. Mary-Huard, Estimation of the relatedness coefficients from biallelic markers, application in plant mating designs, Biometrics in press.*