Clustering Data with Nonignorable Missingness using Semi-Parametric Mixture Models

Marie Du Roy de Chaumaray and Matthieu Marbac

Univ. Rennes, Ensai, CNRS, CREST - UMR 9194, F-35000 Rennes, France

July 15, 2020

Main idea:
Use semiparametric mixture models for clustering (not for density estimation).

Data:
n subjects described by d continuous variables with nonignorable missingness
$\boldsymbol{z}_i$ indicates the subpopulation membership of subject i and is not observed
$r_{ij} = 1$ if $x_{ij} \in \mathbb{R}$ is observed and $r_{ij} = 0$ otherwise
$\boldsymbol{x}_i^{\text{obs}}$ denotes the observed values for subject i.

Assumptions:
The couples $(X_{ij}, R_{ij})^\top$ are conditionally independent given $\boldsymbol{Z}_i$.
No parametric assumptions on the distribution of $X_{ij} \mid \boldsymbol{Z}_i, R_{ij}$ but we need $d \geq 3$.

Model:
We use a pattern-mixture model for dealing with missingness and a semiparametric mixture for modeling the observed variables

$$f(\boldsymbol{x}_i^{\text{obs}}, \boldsymbol{r}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} \tau_{kj}^{r_{ij}} (1 - \tau_{kj})^{1-r_{ij}} p_{kj}^{r_{ij}}(x_{ij}),$$

We estimate the finite parameters (π_k and $\boldsymbol{\tau}_k$) and all the infinite parameters $p_{kj}(\cdot)$ (conditional density of X_{ij} given $Z_{ik} = 1$ and $R_{ij} = 1$) by a maximizing the smoothed likelihood via a MM algorithm (Levine *et al.*, Biometrika, 2011).

We generate 100 realizations $X_i \in \mathbb{R}^4$ where $\mathbb{P}(Z_i = 1) = 1/3$, $\mathbb{P}(Z_i = 2) = 2/3$,

$X_{ij} = \delta(Z_{i1} - Z_{i2}) + \varepsilon_{ij}$ and $\mathbb{P}(R_{ij} = 0 \mid X_{ij}, Z_i) = (1 + \exp(\gamma + \alpha(z_{i1} - z_{i2}) + \beta x_{ij}))^{-1}$,

where the noises ε_{ij} are independent from all the variables.

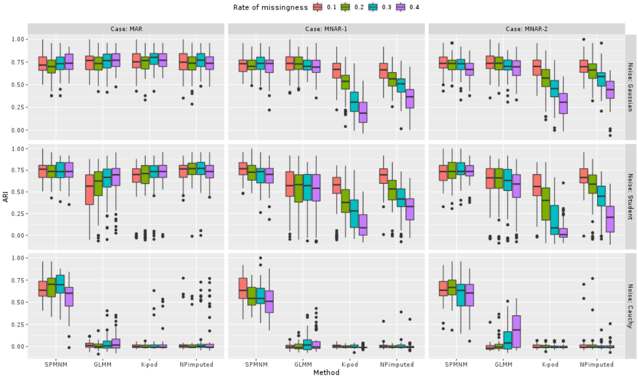

Figure: ARI obtained by SPMNM (proposed method), GLMM (Miao *et al.*, JASA, 2016), K-pod (Chi and Chi, Amer. Stat., 2016), NPimputed (*mixtools* and *missMDA*) on different scenario: MAR ($alpha = 0$, $\beta = 0$), MNAR-1 ($\alpha = 1$, $\beta = 0$), MNAR-2 ($\alpha = 0$, $\beta = 1$). γ and δ are used to define the rate of missingness and an error rate of 5%.

Results

- Lemma 1: If $d \geq 3$, the densities p_{kj} are linearly independent, $\pi_k > 0$ and $\tau_{kj} > 0$, then the model is identifiable, up to label swapping.
- Lemma 2: Let the assumptions of Lemma 1 hold true. Let $\boldsymbol{\theta}^{[r]}$ and $\boldsymbol{\theta}^{[r+1]}$ be the estimators obtained at iterations $[r]$ and $[r+1]$ respectively, we have, under mild assumptions, $\ell_n(\boldsymbol{\theta}^{[r]}) \leq \ell_n(\boldsymbol{\theta}^{[r+1]})$.
- Lemma 3: Let $\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta})$. If the assumptions of Lemma 2 hold true, the densities p_{kj}'s are three times continuously differentiable, $p'_{kj}/p_{kj} < \infty$, $p''_{kj}/p_{kj} < \infty$ and if the bandwidth $h \to 0$ when $n \to \infty$, then $\hat{\boldsymbol{\theta}}_n$ is consistent.

Ongoing research

- Bandwidth selection (experiments are done with $h = n^{-1/5}$).
- Number of components (Kasahara and Shimotsu, JRSS-B, 2014).