Clustering with missing data: which imputation model for which cluster analysis method?

V. Audigier, N. Niang, M. Resche-Rigon

CNAM, CEDRIC-MSDMA, Paris

Midia, June 18th, 2021

Clustering

Data $Z = (z_{ij})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}}$ a continuous data set

Each individual i belongs to a unique cluster $w_i \in \{1, ..., K\}$.

Aim identify w_i for each i based on individual profiles $(z_i)_{(1 \leq i \leq n)}$

Methods

Distance-based

- k-means
- fuzzy C-means
- hierarchical clustering
- pam

Model-based

- gaussian mixture models
- mixture of multivariate t-distributions

Clustering with missing values

However, Z is frequently **incomplete**... $z_i = (z_i^{obs}, z_i^{miss})$

Ad-hoc methods

- complete cases analysis (CCA)
- removing incomplete variables
- single imputation (SI)

Direct methods

- k-means (Chi et al., 2016; Honda et al., 2011; Wagstaff, 2004)
- fuzzy C-means (Zhang et al., 2016; Hathaway and Bezdek, 2001)
- Gaussian mixture (Miao et al., 2016; Marbac et al., 2019; de Chaumaray and Marbac, 2020)

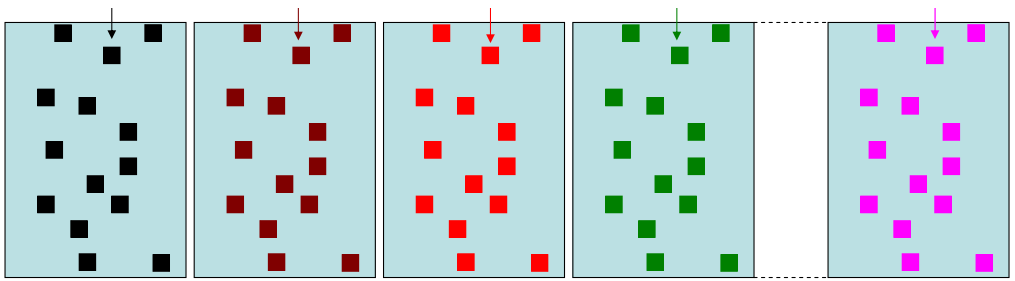
Multiple Imputation (MI)

- a popular method
- could be used for any clustering method

Multiple imputation (Rubin, 1987)

1. Generate a set of M parameters $(\zeta_m)_{1 \leq m \leq M}$ of an imputation model to generate M plausible imputed data sets

$$P\left(Z^{miss} | Z^{obs}, \zeta_1\right) \quad \ldots \quad \ldots \quad \ldots \quad P\left(Z^{miss} | Z^{obs}, \zeta_M\right)$$



2. Fit the analysis model on each imputed data set

3. Combine the results using Rubin's rules

$\Rightarrow$ **Provide estimation of the parameters and of their variability**

Partition pooling

Ψ_m the partition from (Z^{obs}, Z_m^{miss}), which average $\widehat{\Psi}$ for $(\Psi_m)_{1 \leq m \leq M}$?

Related works

- using stacking (Plaehn, 2019)
- using consensus clustering methods (Faucheux et al., 2020; Bruckers et al., 2017; Basagana et al., 2013; Audigier and Niang, 2020)

Theoretically, NMF based methods are appealing

- $(\mathrm{H}_m)_{1 \leq m \leq M}$ connectivity matrices associated to $(\Psi_m)_{1 \leq m \leq M}$
- $\mathrm{M} = \frac{1}{M} \sum_{m=1}^{M} \mathrm{H}_m$

$$argmin_{\mathrm{H}} \parallel \mathrm{M} - \mathrm{H} \parallel^2$$

The solution of this optimization problem can be obtained using non-negative matrix factorization (Li et al., 2007).

Instability

Fang and Wang (2012) assessed sampling instability in clustering

- generate C bootstrap pairs $\left(Z_c, \tilde{Z}_c\right)_{1 \leq c \leq C}$ from Z
- perform cluster analysis from $\left(Z_c, \tilde{Z}_c\right)_{1 \leq c \leq C}$ to obtain $\left(\Psi_c, \tilde{\Psi}_c\right)$
- classify individuals of Z from Ψ_c and $\tilde{\Psi}_c$ to obtain $\left(\Psi'_c, \tilde{\Psi}'_c\right)$
- the instability V is assessed by averaging the proportions of disagreements

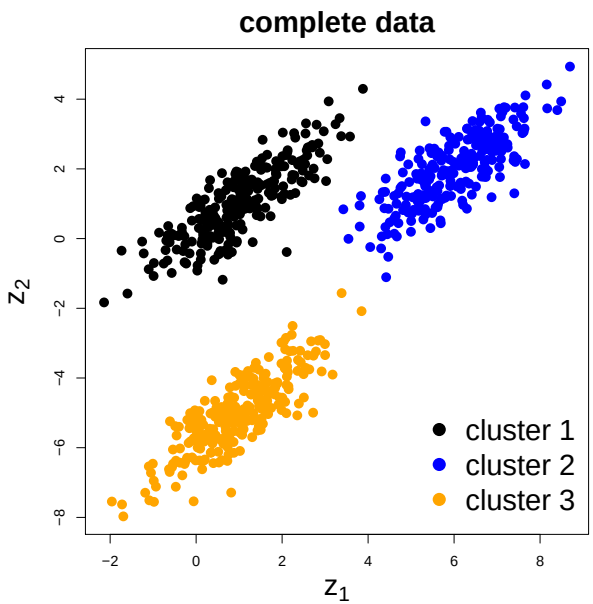
$$V = \frac{1}{C} \sum_{c=1}^{C} \delta(\Psi'_c, \tilde{\Psi}'_c)/n^2$$

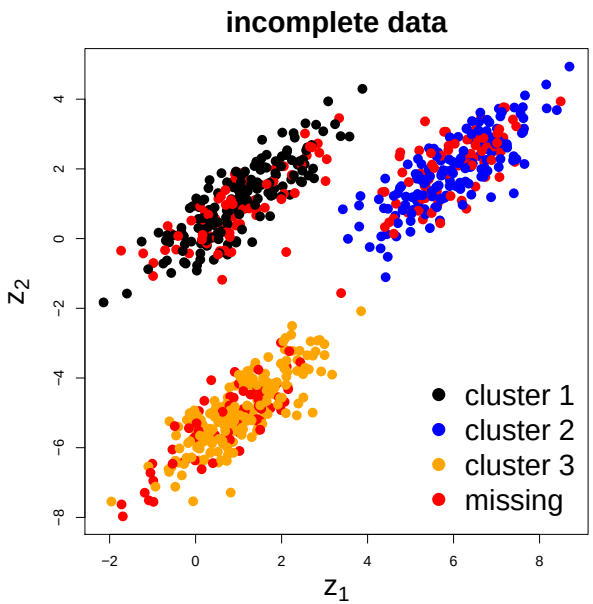
After MI, Audigier and Niang (2020) proposed

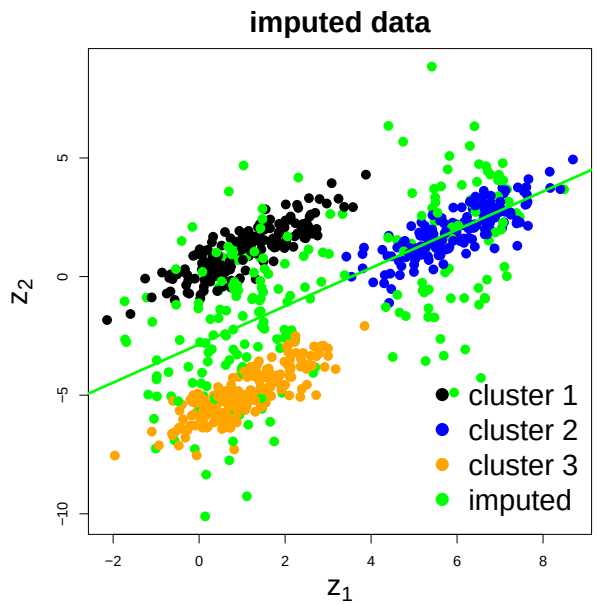
$$T = \frac{1}{M} \sum_{m=1}^{M} V_m + \frac{1}{M^2} \sum_{m=1}^{M} \sum_{m'=1}^{M} \delta\left(\Psi_m, \Psi_{m'}\right)/n^2$$

How to impute?

Example


complete data
cluster 1
cluster 2
cluster 3
z_1
z_2


incomplete data
cluster 1
cluster 2
cluster 3
missing
z_1
z_2


imputed data
cluster 1
cluster 2
cluster 3
imputed
z_1
z_2

Outline

MI for which cluster analysis method?

Cluster analysis using GMM is relevant to investigate congeniality

- clustering by GMM is a model-based clustering method

$$W \sim \mathcal{M}(1, \boldsymbol{\theta})$$
$$Z|W = w \sim \mathcal{N}_p(\mu_w, \boldsymbol{\Sigma}_w)$$

  $\mathcal{M}$ a multinomial distribution, $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)^t$ the vector of probabilities,

  $\psi = (\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K)$, $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K)$

- it generalizes other clustering methods
  - k-means (with $\Sigma_w = \sigma^2 \mathbb{I}$, $\sigma^2 \to 0$)
  - pam (with euclidean distance)
  - mixture of multivariate t-distributions (ddf $\to \infty$)

$\Rightarrow$ **Which MI methods are based on GMM?**

JM-DP (Kim et al., 2014)

Based on a Bayesian formulation

$$\mu_w|\Sigma_w \sim \mathcal{N}\left(\mu_0, h^{-1}\Sigma_w\right) \quad \Sigma_w \sim \mathcal{W}^{-1}\left(df, G\right)$$
$$\text{with } G = (g_1, ..., g_p) \;\; g_j \sim \mathcal{G}\left(a_0, b_0\right)$$
$$\theta_w = v_w \prod_{\ell<w} (1 - v_\ell)$$

$$\text{with } \begin{cases} v_w \sim \mathcal{B}eta\left(1, \alpha\right) \text{ and } \alpha \sim \mathcal{G}(a_\alpha, b_\alpha) \text{ for } w < K \\ v_K = 1 \end{cases}$$

- parameters: $\zeta = (\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$
- hyperparameters: h, μ_0, df, a_0, b_0, a_α, b_α

$(\zeta_m)_{1\leq m\leq M}$ is generated using a DA algorithm

Properties

JM-DP assumes

- no constraints on covariance matrices $(\Sigma_1, ..., \Sigma_K)$ (heteroscedasticity)
- the number of clusters is only bounded

Congenial with GMM?

- only with heteroscedastic GMM, conservative otherwise
- the number of clusters can be misspecified if n is small

R package DPImputeCont (Kim, 2020)

JM-GL (Schafer, 1997)

The gold-standard method to impute mixed data

$$W \sim \mathcal{M}(1, \boldsymbol{\theta})$$
$$Z|W = w \sim \mathcal{N}_p(\mu_w, \boldsymbol{\Sigma})$$

- Bayesian formulation (...)
- parameters: $\zeta = (\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$
- $(\zeta_m)_{1 \leq m \leq M}$ is generated using a DA algorithm

Properties

JM-GL assumes

- a constant covariance matrix $\boldsymbol{\Sigma}$ (homoscedasticity)
- a pre-defined number of clusters

Congenial with GMM?

- Congenial with homoscedastic GMM, potentially biased analysis otherwise

R package mix (Schafer, 2017)

DA for JM-GL

- initialize ζ by $\zeta^{(0)} = \widehat{\zeta}_{ML}$, *i.e.* the maximum likelihood estimator obtained by an EM algorithm
- for ℓ in $1, ..., L$

I-step for each i in $\{1, ..., n\}$

1. draw $w_i^{(\ell)}$ from $p\left(W | Z^{obs} = z_i^{obs}, \zeta = \zeta^{(\ell-1)}\right)$
2. draw $z_i^{miss\,(\ell)}$ from $p(Z^{miss} | W = w_i^{(\ell)}, Z^{obs} = z_i^{obs}, \mu = \boldsymbol{\mu}^{(\ell-1)}, \Sigma = \boldsymbol{\Sigma}^{(\ell-1)})$

P-step

1. draw $\theta^{(\ell)}$ from $p\left(\theta | W = w^{(\ell)}\right)$
2. draw $\Sigma^{(\ell)}$ from $p\left(\Sigma | Z = Z^{(\ell)}, W = w^{(\ell)}\right)$
3. draw $\mu^{(\ell)}$ from $p\left(\mu | \Sigma = \boldsymbol{\Sigma}^{(\ell)}, Z = Z^{(\ell)}, W = w^{(\ell)}\right)$

FCS-homo (Audigier et al., 2021)

Variable-by-variable imputation

- generating Z_i^{miss} given W is straightforward (Hughes et al., 2014)
- generating W given Z is more complicated...

$$p(W = w|Z) \text{ depends on } \boldsymbol{\theta} \text{ (unknown)}$$

1. Draw $\boldsymbol{\theta}^{(\ell)}$ from $p(\boldsymbol{\theta}|Z)$
   1. estimate $\zeta^\star$ using an EM algorithm
   2. draw W from $p(W|Z, \boldsymbol{\zeta} = \boldsymbol{\zeta}^\star)$
   3. generate $\boldsymbol{\theta}^{(\ell)}$ from $p(\boldsymbol{\theta}|W, Z)$
2. Draw $W^{(\ell)}$ from $p(W|Z, \boldsymbol{\theta}^{(\ell)}, \boldsymbol{\mu}^\star, \Sigma^\star)$

Properties

FCS-homo assumes

- homoscedastic regression models / GMM
- a pre-defined number of clusters

Congenial with GMM?

- only with homoscedastic GMM

Can be easily modified

- to account for heteroscedasticity (van Buuren, 2011)
- to improve sparsity (Zahid and Heumann, 2019)
- to use semi-parametric models (Morris et al., 2014)
- to address outliers (Templ et al., 2011)
- ...

Outline

Simulation design: data generation

Complete data generation

- A base-case configuration: GMM with $K = 3$ components

$$\begin{aligned} \mu_1 &= (0,0,0,0,\Delta,\Delta,0,\Delta^2) \\ \mu_2 &= (0,0,0,0,-\Delta,-\Delta,-\Delta,0) \\ \mu_3 &= (0,0,0,0,-\Delta,\Delta,\Delta,-\Delta^2) \end{aligned} \qquad \Sigma_w = \begin{pmatrix} I_4 & 0 \\ 0 & \begin{matrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{matrix} \end{pmatrix}$$

  $n_w = 250$ (for all w in $\{1,2,3\}$), $\Delta = 2$ $\rho = 0.3$

- 10 other configurations varying: the separability between clusters ,the number of clusters, the cluster size, the balance between clusters sizes and the heteroscedasticity.

Missing data generation

- MCAR: $Prob(r_{ij} = 0) = \tau$ $\quad \forall i,j$
- MAR 1: $Prob(r_{ij} = 0) = \Phi(a_\tau + z_{i1})$ $\quad \forall j \neq 1$
- MAR 2: $Prob(r_{ij} = 0) = \Phi(a_\tau + z_{i8})$ $\quad \forall j \neq 8$
- $\tau \in \{10\%, 25\%, 40\%\}$

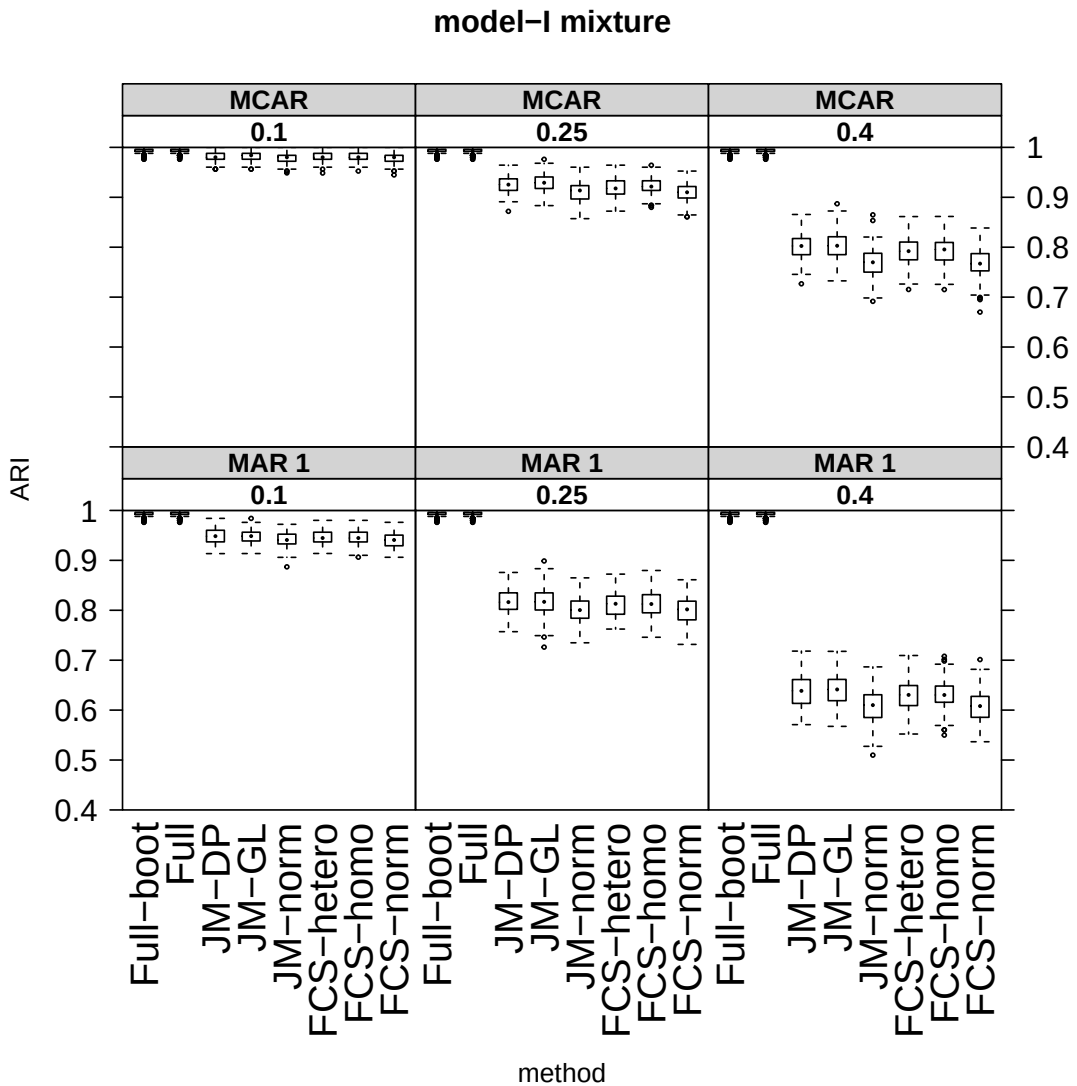
Simulation design: evaluation

For each incomplete data set (200 per configuration)

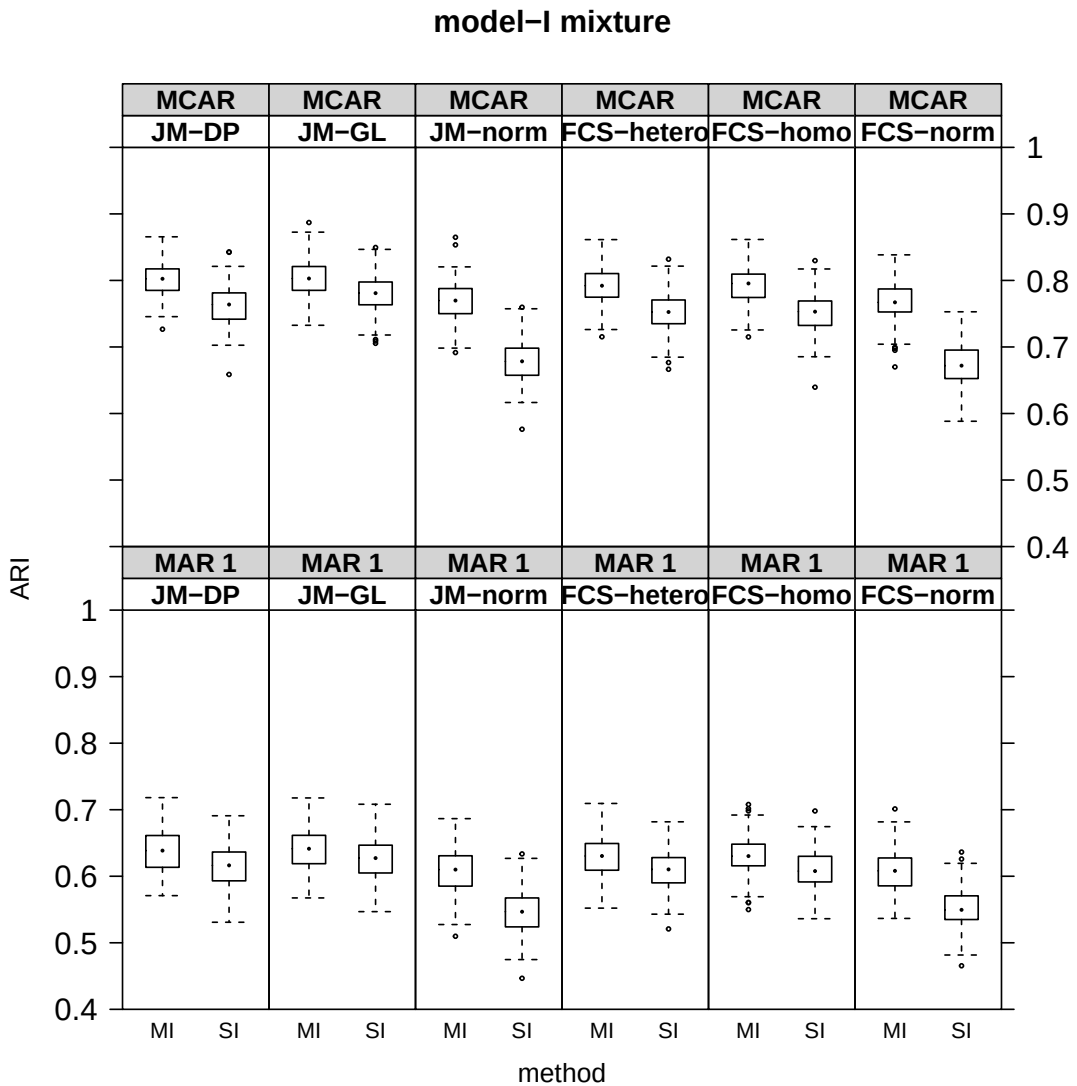
1. Imputation ($M = 20$)
   - FCS-homo
   - FCS-hetero
   - FCS-norm
   - JM-GL
   - JM-DP
   - JM-norm
2. Cluster analysis
   - clustering by GMM
   - k-means
   - pam
   - hierarchical clustering
3. Pooling by NMF

Criteria ARI and compared with SI / clustering on Full data (with or without bootstrap) (Dudoit and Fridly, 2003)

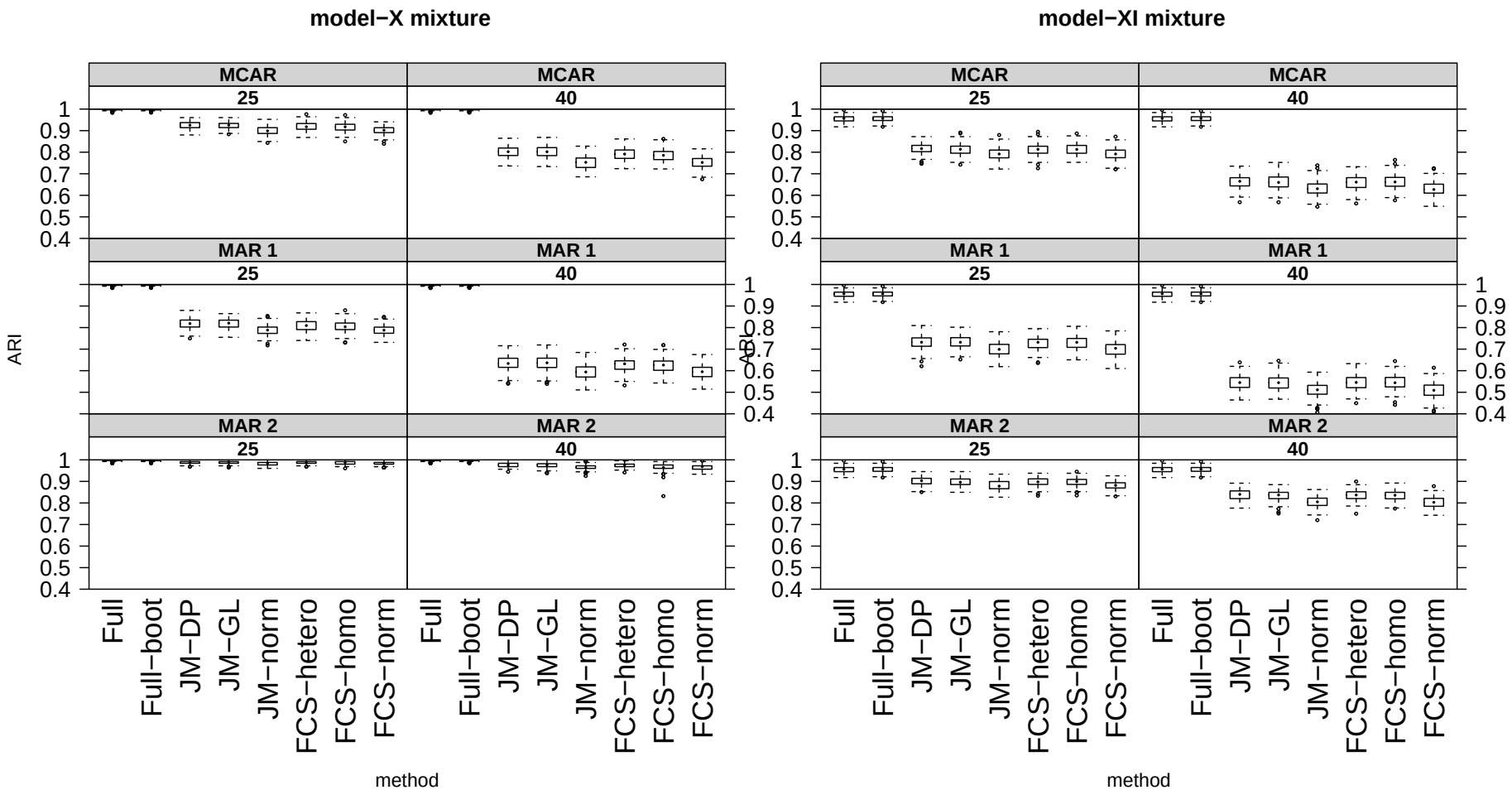
Results: base-case (1)



Results: base-case (2)

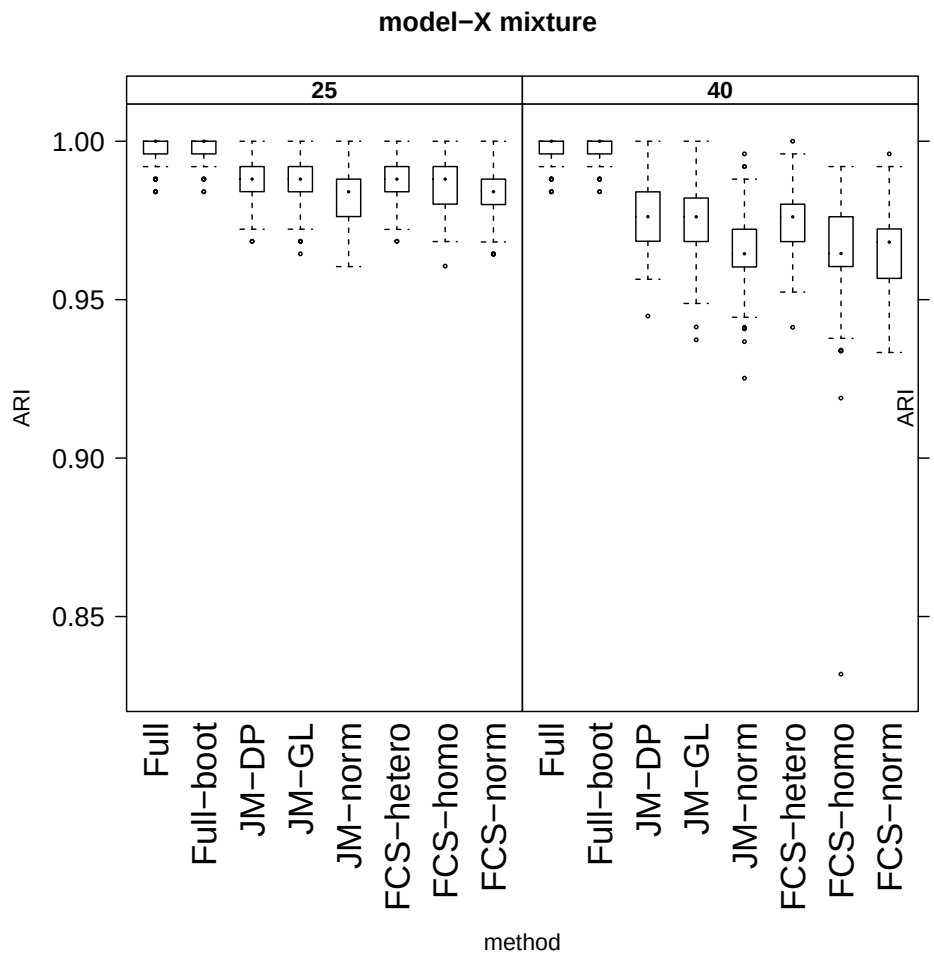


Results: heteroscedastic-case (1)

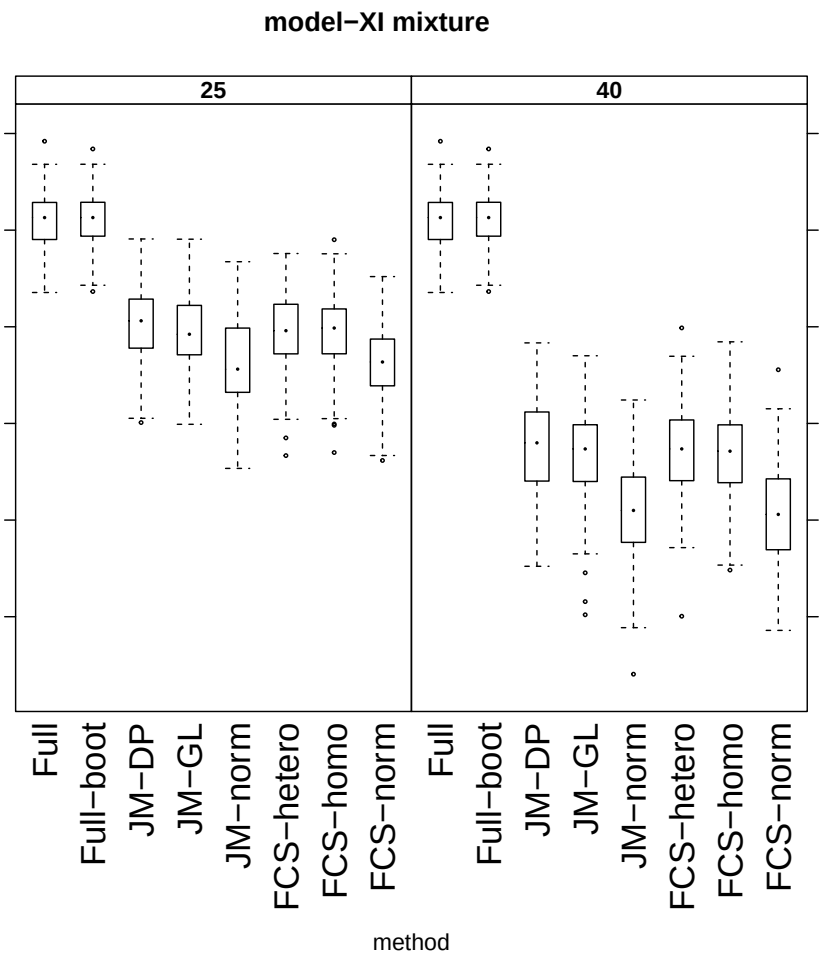


(a) model-X (b) model-XI

Results: heteroscedastic-case (2)



(a) model-X



(b) model-XI

Results: summary

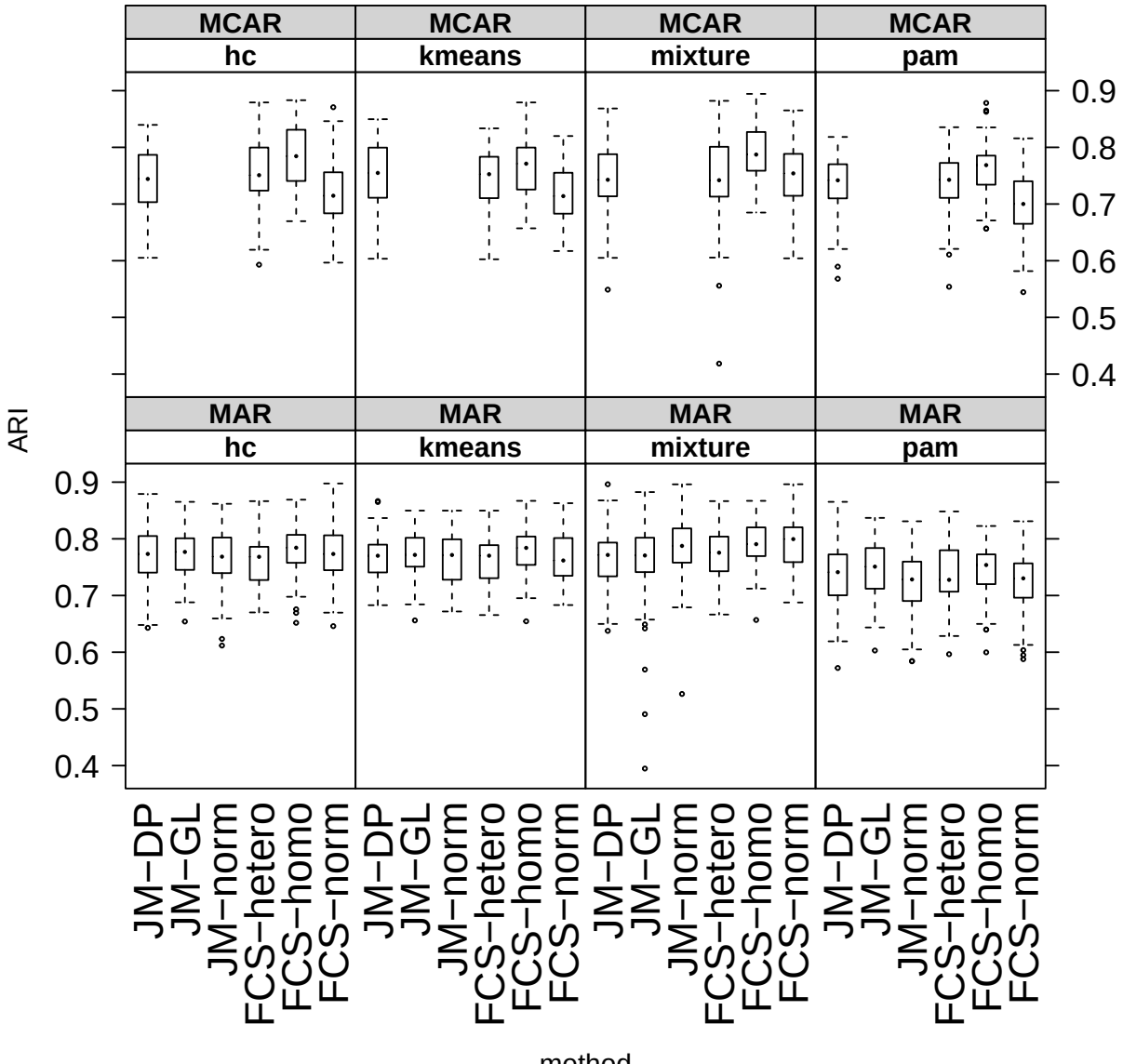
- smaller ARI when clusters are less separated
- ARI robust to the MI method used with 2 clusters
- ARI more stable with more individuals
- similar results with unbalanced data
- similar results with other cluster analysis methods

Wine data set (Asuncion and Newman, 2007)

- $n = 178$ Italian wines by $p = 13$ chemical descriptors
- $K = 3$ categories
- 40% missing values (MCAR or a MAR mechanism)
- 100 missing data patterns per mechanism

For each missing data pattern:

- variable selection to specify conditional imputation models in FCS (Bar-Hen and Audigier, 2021)
- cluster analysis by GMM (hetero), k-means, pam and hierarchical clustering



Conclusion

This study shows

- ignoring the underlying class structure in the imputation model increases bias
- with few missing values, MI by dedicated methods provides similar ARI than those observed for the full data case
- congenial imputation models are recommended, but the loss to use a heteroscedastic model instead of a homoscedastic model remains small (and vice versa)
- FCS MI is promising for real data

Note that

- the number of clusters can be easily estimated
- MI methods are available in the clusterMI R package

Some perspectives

- mixed data
- comparison with direct methods

References I

V. Audigier, N. Niang, and M. Resche-Rigon. Clustering with missing data: which imputation model for which cluster analysis method?, 2021. ArXiv preprint.

V. Audigier and N. Niang. Clustering with missing data: which equivalent for Rubin's rules?, 2020. ArXiv preprint.

S. Dudoit and J. Fridly. Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, pages 1090–1099, 2003.

Y. Fang and J. Wang. Selection of the number of clusters via the bootstrap method. *Comput. Stat. Data Anal.*, 56(3):468â477, 2012. ISSN 0167-9473. doi: 10.1016/j.csda.2011.09.003. URL `https://doi.org/10.1016/j.csda.2011.09.003`.

H. J. Kim, J. P. Reiter, Q. Wang, L. H. Cox, and A.F. Karr. Multiple imputation of missing or faulty values under linear constraints. *Journal of Business & Economic Statistics*, 32(3):375–386, 2014. doi: 10.1080/07350015.2014.885435. URL `https://doi.org/10.1080/07350015.2014.885435`.

T. Li, C. Ding, and M. I. Jordan. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, page 577â582, USA, 2007. IEEE Computer Society. ISBN 0769530184. doi: 10.1109/ICDM.2007.98.