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

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

CNAM, CEDRIC-MSDMA, Paris

IFCS, July 22th, 2022

Clustering

Data $\mathbf{Z} = (z_{ij})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq p}}$ 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 $(\mathbf{z}_i)_{1 \leq i \leq n}$

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, $\mathbf{Z}$ is frequently **incomplete**... $\mathbf{z}_i = (\mathbf{z}_i^{obs}, \mathbf{z}_i^{miss})$

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)

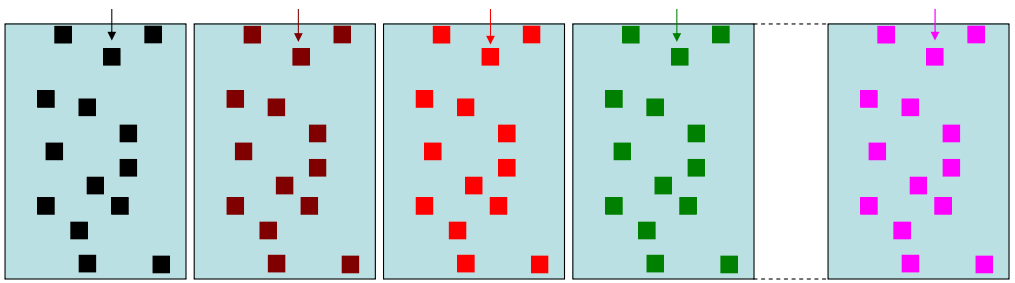
Multiple Imputation (MI)

- 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

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



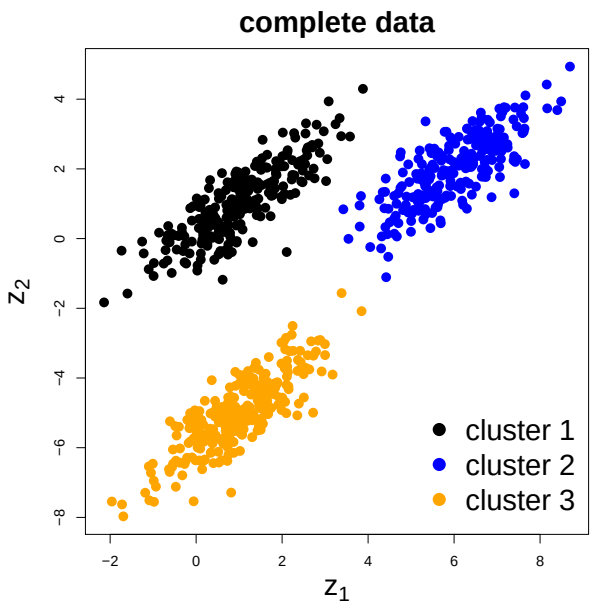
2. Fit the analysis model on each imputed data set

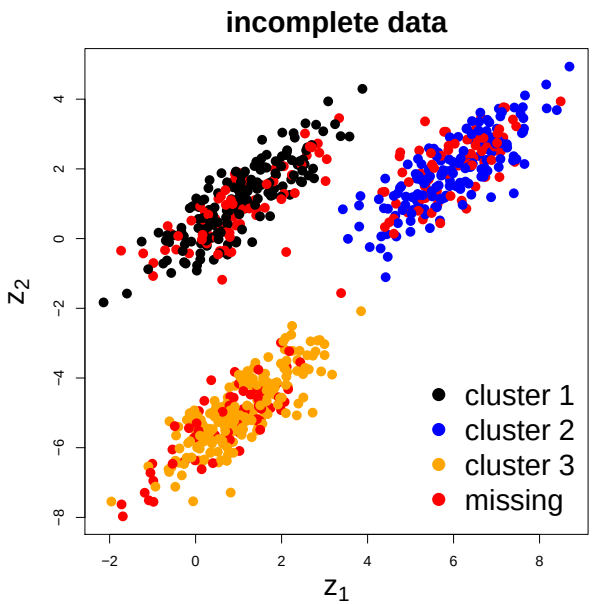
3. Combine the results using Rubin's rules

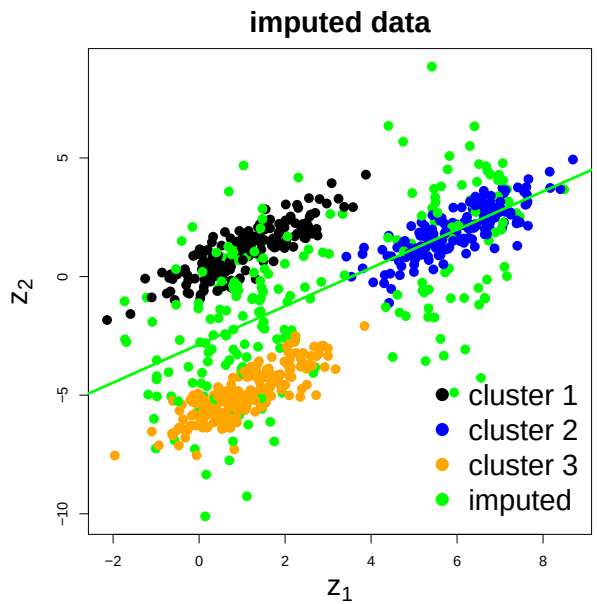
$\Rightarrow$ **Provide estimation of the parameters and of their variability**

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

Congeniality

- In an ideal world, the imputation model would be based on the true data model. Then, any analysis model could be applied
- In a real world, the true distribution for $P(Z;\zeta)$ is unknown... and the imputation model is usually misspecified
- To avoid a bias due to the imputation step, the imputation model should be in lines with the assumptions related to the analysis model

$\Rightarrow$ Both models should be ***congenial***

Outline

JM-DP (Kim et al., 2014)

Joint Modeling based on Dirichlet Process mixture of products of multivariate normal distributions

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

R package DPImputeCont (Kim, 2020)

JM-GL (Schafer, 1997)

Joint Modeling by General Location model

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

R package mix (Schafer, 2017)

Properties

JM-DP assumes

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

JM-GL assumes

- a constant covariance matrix Σ (homoscedasticity)
- a pre-defined number of clusters

Congenial with GMM?

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

- Congenial with homoscedastic GMM, potentially biased analysis otherwise

Fully conditional specification

A conditional distribution is specified for each (incomplete) variable $P\left(Z_j|Z_{-j};\zeta_j\right)$

Ex : $P\left(Z_j|Z_{-j};\zeta_j\right) = \mathcal{N}(Z_{-j}\beta, \sigma^2) \quad \zeta_j = (\beta, \sigma)$

To impute the mth data set

- initialize missing values of $\mathbf{Z}$
- for j in 1 ... p
  a. generate ζ_j based on observed individuals on Z_j
  b. impute Z_j^{miss} according to $P\left(Z_j|Z_{-j};\zeta_j\right)$
- repeat until convergence

FCS-homo (Audigier et al., 2021)

Variable-by-variable imputation

- generating Z_i^{miss} given W is performed using regression models including a intercept specific to each cluster.

  $P\left(Z_j|Z_{-j}, W; \zeta_j\right) = \mathcal{N}(Z_{-j}\beta + \mu_w, \sigma^2) \quad \zeta_j = (\beta, \sigma, \mu_w)$

- generating W given Z is more complicated...

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

1. Draw $\boldsymbol{\theta}^{(\ell)}$ from $p(\boldsymbol{\theta}|Z)$
   1. estimate $\zeta^\star$ using an EM algorithm
   2. draw W from $p(W|Z, \zeta = \zeta^\star)$
   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
- 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 address outliers (Templ et al., 2011)
- to use semi-parametric models (Morris et al., 2014)
- ...

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

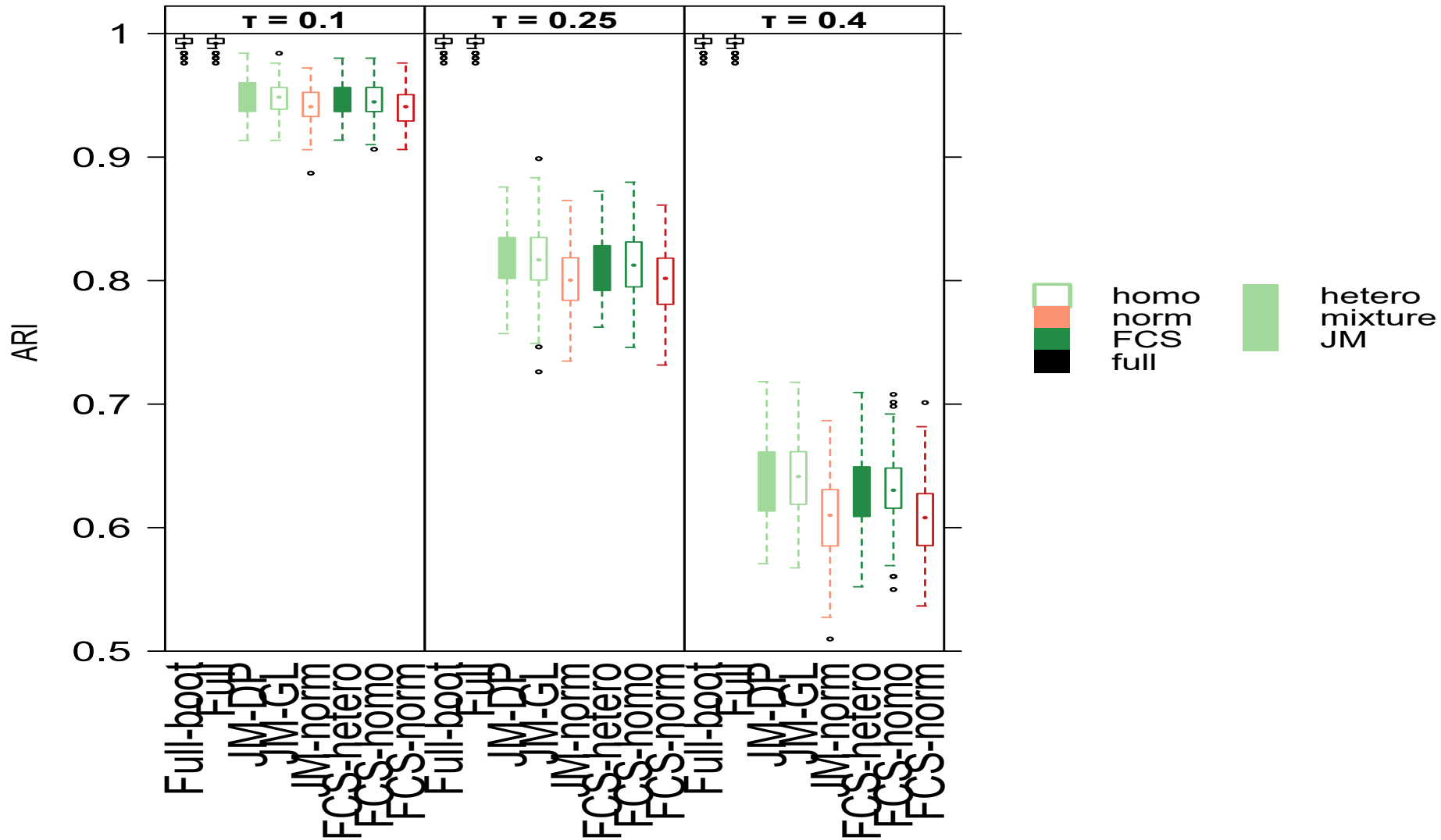
Simulation design: evaluation

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 clustering on Full data (with or without bootstrap) (Dudoit and Fridly, 2003)

Results: base-case, MAR 1, GMM



Results: summary

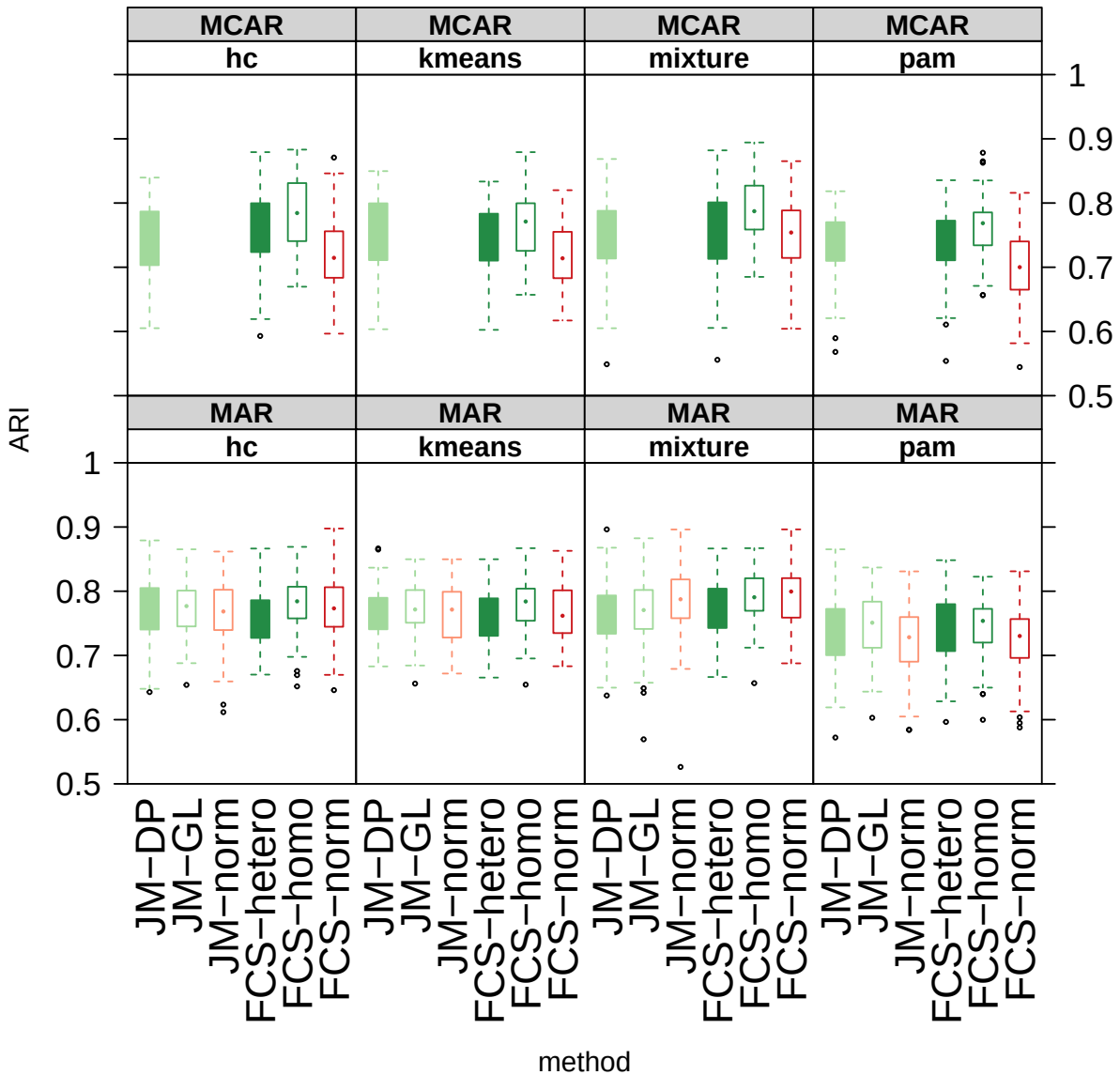
- higher ARI with FCS-hetero and JM-DP on heteroscedastic data
- smaller ARI when clusters are less separated
- ARI robust to the MI method used with 2 clusters
- similar results with unbalanced data
- similar results with other cluster analysis methods
- Interquartile range smaller for ARI with more individuals

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, 2022)
- 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
- 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):468âĂŞ477, 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 577-582, USA, 2007. IEEE Computer Society. ISBN 0769530184. doi: 10.1109/ICDM.2007.98.