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

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

CNAM, CEDRIC-MSDMA, Paris

Séminaire MAP5, June 4th, 2021

Clustering

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

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

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

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; Wagsta□, 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)

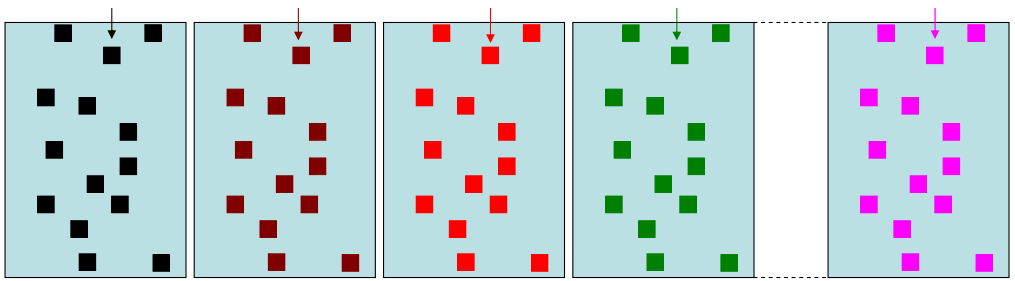
a popular method

could be used for any clustering method

Multiple imputation (Rubin, 1987)

1. Generate a set of M parameters $(\zeta_m)_{1 \quad m \quad M}$ 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

) 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 $(\Psi_m)_{1 \quad m \quad M}$?

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

$(\mathrm{H}_m)_{1 \quad m \quad M}$ connectivity matrices associated to $(\Psi_m)_{1 \quad m \quad M}$

$\mathrm{M} = \frac{1}{M} \sum_{m=1}^{M} \mathrm{H}_m$

$$argmin_{\mathrm{H}} \; k \; \mathrm{M} \quad \mathrm{H} \; k^2$$

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

Instability (Audigier and Niang, 2020)

Fang and Wang (2012) assessed sampling instability in clustering

generate C bootstrap pairs $\left(Z_c, \tilde{Z}_c\right)_{1 \; c \; C}$ from Z

perform cluster analysis from $\left(Z_c, \tilde{Z}_c\right)_{1 \; c \; C}$ to obtain $\left(\Psi_c, \tilde{\Psi}_c\right)$

classify individuals of Z from Ψ_c and $\tilde{\Psi}_c$ to obtain $\left(\Psi_c^0, \tilde{\Psi}^0{}_c\right)$

the instability V is assessed by averaging the proportions of disagreements

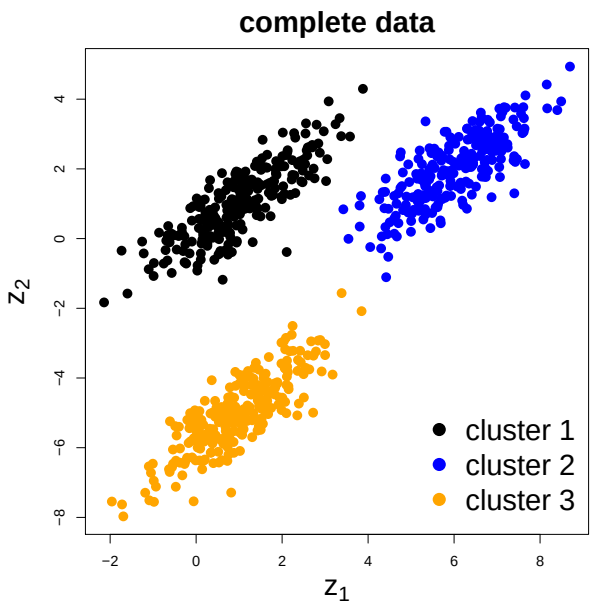
$$V = \frac{1}{C} \sum_{c=1}^{C} \delta(\Psi_c^0, \tilde{\Psi}^0{}_c)/n^2$$

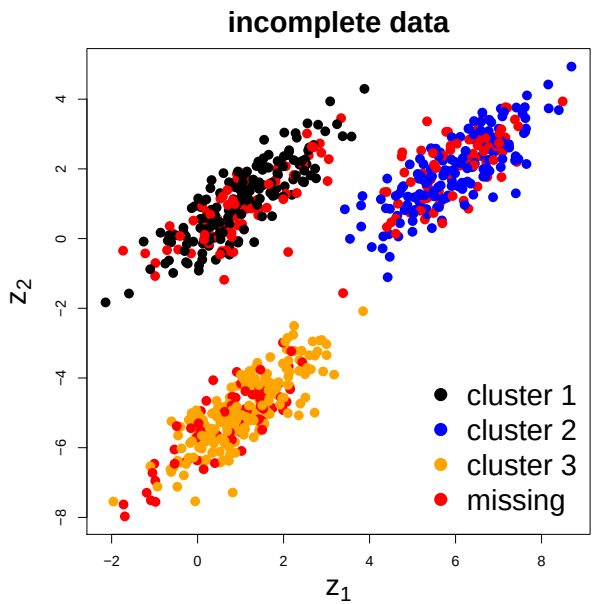
After MI, Audigier and Niang (2020) proposed

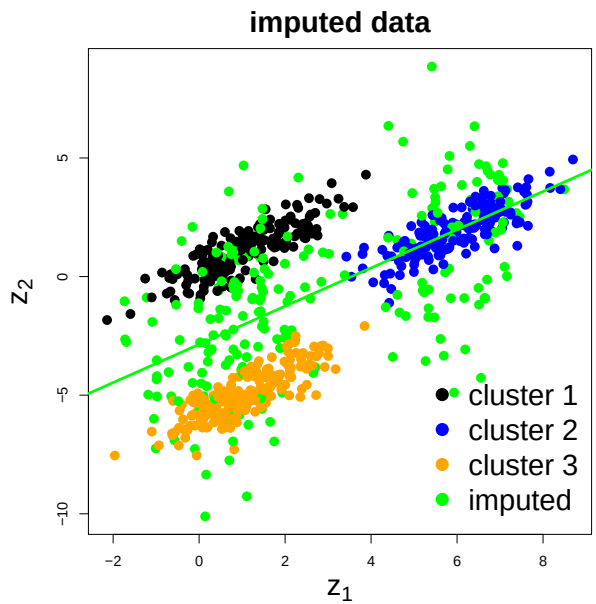
$$T = \frac{1}{M} \sum_{m=1}^{M} V_m + \frac{1}{M^2} \sum_{m=1}^{M} \sum_{m'=1}^{M} \delta\left(\Psi_m, \Psi_{m'}\right)/n^2$$

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

Proper imputation

Given an imputation model $P(Z;\zeta)$

Ex: $Z = (Z^{obs}, Z^{miss})$ $N_p(\mu, \Sigma)$, $\zeta = (\mu, \Sigma)$

Two phases for each data set m

a generate ζ_m from $P(\zeta | Z^{obs})$

b draw Z^{miss} from $P(Z^{miss} | Z^{obs}; \zeta_m)$

Multiple imputation is not M single imputations (improper)

Generation of $(\zeta_m)_{1 \quad m \quad M}$

Bayesian

- Prior distribution $p(\zeta)$
- Derive the posterior distribution $p(\zeta | Z^{obs})$
- (Data-Augmentation)
- Draw from $p(\zeta | Z^{obs})$ M times

Non-parametric Bootstrap

- Sampling observations with replacement M times
- Estimate ζ_m from each bootstrap sample (EM)

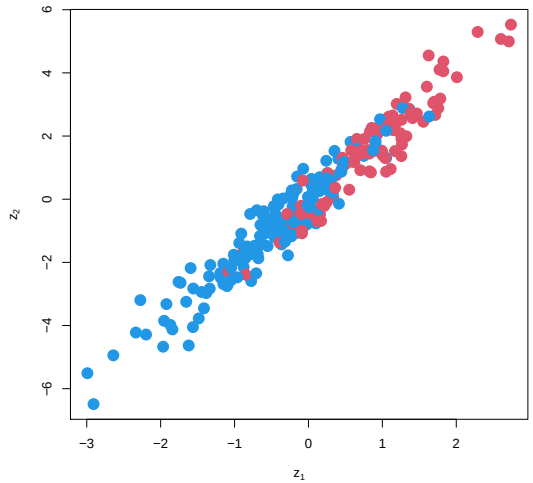
...

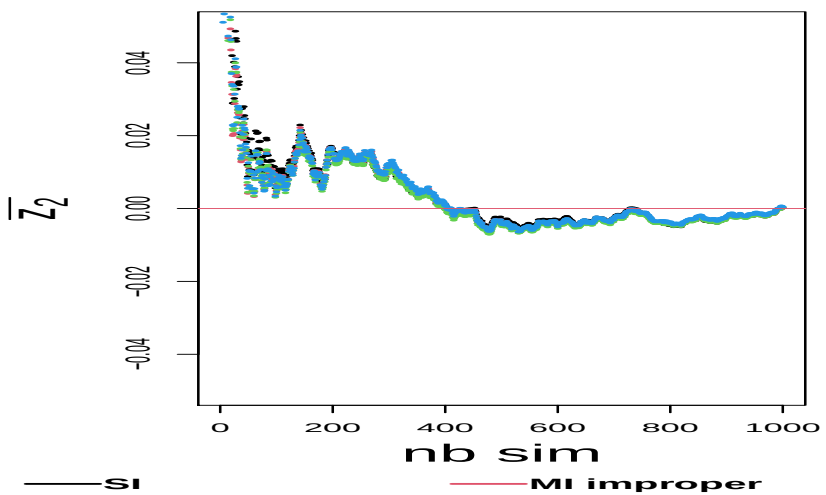
Example

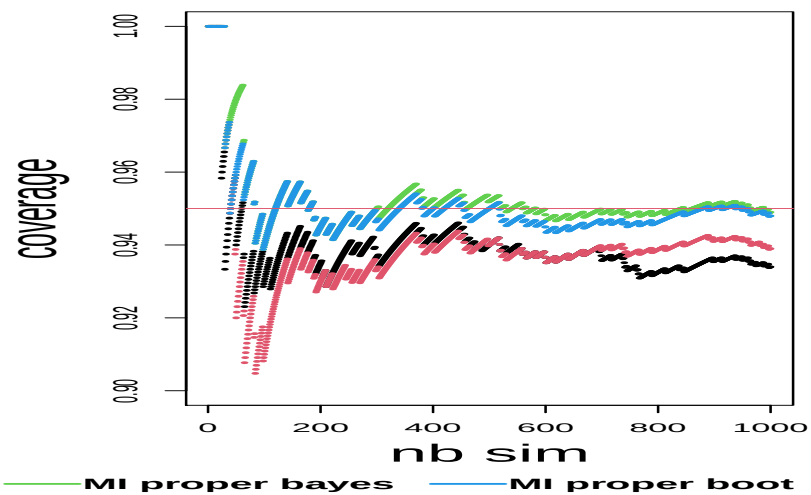
$n = 150$, $p = 2$

missing values on Z_2 (MAR)

parameter: $\mathrm{E}[Z_2]$







Some imputation models

Many imputation models based on the multivariate normal distribution

	Model	R package
Schafer (1997)	Gaussian model / Bayesian	norm
Honaker et al. (2011)	Gaussian model / Bootstrap	Amelia
Quartagno and Carpenter (2016)	multivariate LMM/ Bayesian	jomo
Schafer (1997)	multivariate LMM/ Bayesian	pan

However

- the Gaussian distribution can be inappropriate with a moderate/large number of variables
- models have many parameters leading to overfitting with a small number of observations

Fully conditional specification

Instead of specifying one joint distribution $P(Z;\zeta)$, a conditional distribution is specified for each (incomplete) variable $P\left(Z_j | Z_{-j};\zeta_j\right)$

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

To impute the mth data set

initialize z_i^{miss} for all i

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 pros and cons

Pros

- sparsity
- accounting for interactions effects
- addressing outliers
- semi or non-parametric models

Cons

- time consuming
- no theoretical guaranties (except in specific cases)
- checking convergence is not possible with a large number of variables

Relationship between JM and FCS

FCS is close to a Gibbs sampler, but

In a Gibbs sampler, a joint distribution $P(Z;\zeta)$ is specified and conditional models $P\left(Z_j jZ_{\ j};\zeta_j\right)$ $(1 \quad j \quad p)$ are accordingly chosen

In FCS, conditional models are specified and convergence to an unknown joint distribution is expected

) When conditional imputation models are specified from a joint model, FCS MI is equivalent to JM MI.

Ex: JM by multivariate gaussian model = FCS by linear regression (Hughes et al., 2014)

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***

Example (Schafer, 1997)

model 1: $Z_3 = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \varepsilon$ model 2: $Z_3 = \beta_0 + \beta_1 Z_1 + \varepsilon$

If the true model is (2)

(2) + (2) ! congenial

(1) + (2) ! uncongenial, but conservative

If the true model is (1)

(1) + (1) ! congenial

(2) + (1) ! uncongenial, biased

) The imputation model should not impose restrictions on unknown parameters

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

$$W \quad M(1, \boldsymbol{\theta})$$
$$Z|W = w \quad N_p(\mu_w, \boldsymbol{\Sigma}_w)$$

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 $\Sigma_w = \sigma^2 \mathbf{I}$, $\sigma^2 \to 0$)
- 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

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

$$\text{with } \begin{cases} v_w \quad Beta(1, \alpha) \text{ and } \alpha \quad G(a_\alpha, b_\alpha) \text{ for } w < K \\ v_K = 1 \end{cases}$$

where G corresponds to the gamma distribution. h, μ_0, df, a_0, b_0, a_α, b_α are the hyperparameters.

$(\zeta_m)_{1 \ m \ M}$ 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

$$\begin{array}{ll} W & M(1,\boldsymbol{\theta}) \\ Z|W=w & N_p(\mu_w,\boldsymbol{\Sigma}) \end{array}$$

$\zeta = (\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)^t$ the vector of proportions of each category, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K)^t$ the vector of means $\boldsymbol{\Sigma}$ the covariance matrix

Based on a Bayesian formulation, $(\zeta_m)_{1 \quad m \quad M}$ is generated using a DA algorithm

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 $f1, ..., ng$

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

P-step

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

Properties

JM-GL assumes

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

Congenial with GMM?

- Congenial with homoscedastic GMM, potentially biased otherwise

R package mix (Schafer, 2017)

FCS-homo

Variable-by-variable imputation

generating Z_i^{miss} given W is easy (Hughes et al., 2014)

generating W given Z is more complicated...

$p(W = wjZ)$ depends on θ (unknown)

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

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
- to improve sparsity
- to use semi-parametric models
- ...

Outline

Simulation design: data generation

Complete data generation

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

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

$n_w = 250$ (for all w in $f1,2,3g$), $= 2$ $\rho = 0.3$

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: $Prob(r_{ij} = 0) = \tau$

MAR 1: $Prob(r_{ij} = 0) = (a_\tau + z_{i1})$

MAR 2: $Prob(r_{ij} = 0) = (a_\tau + z_{i8})$

τ (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

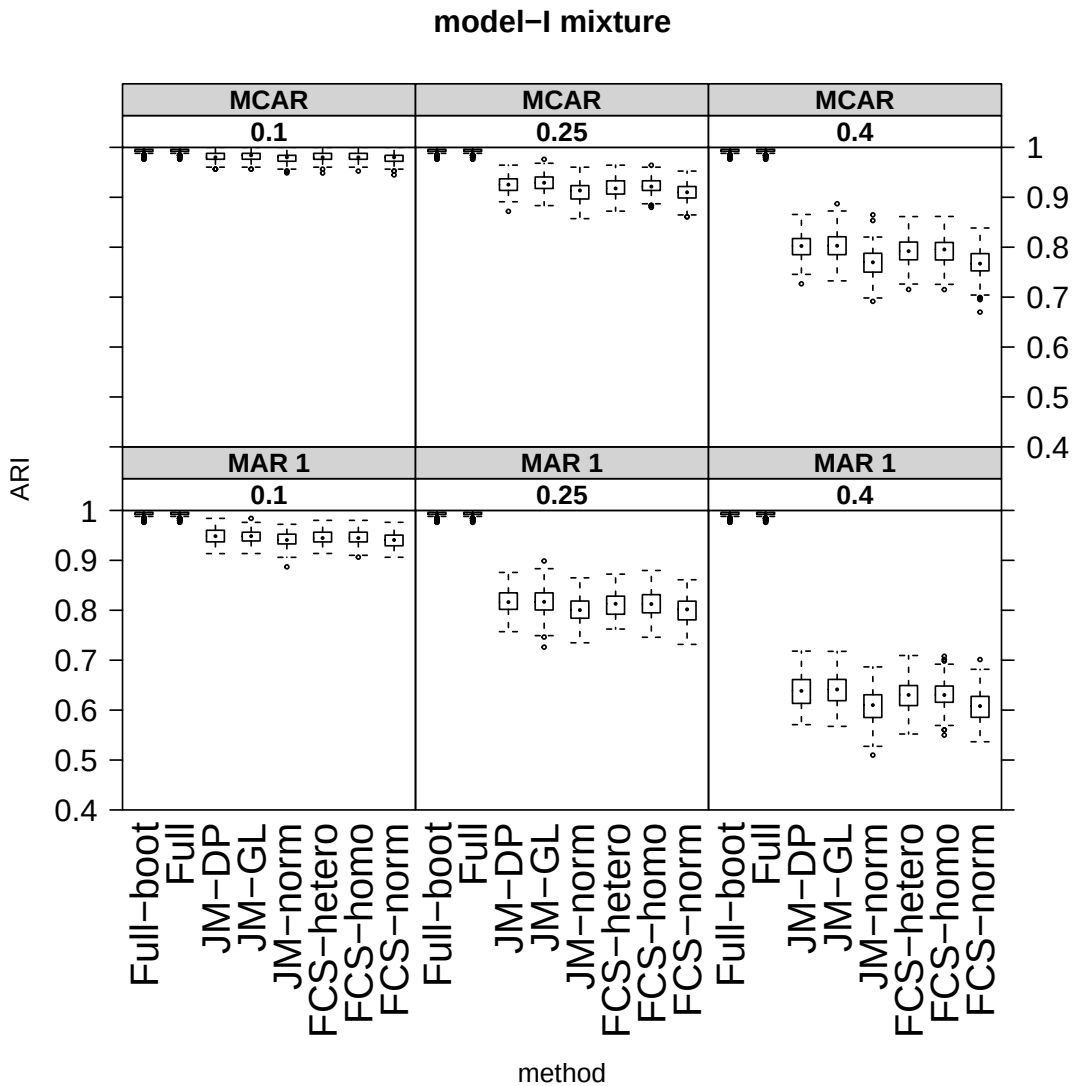
2. Cluster analysis

clustering by GMM	pam
k-means	hierarchical clustering

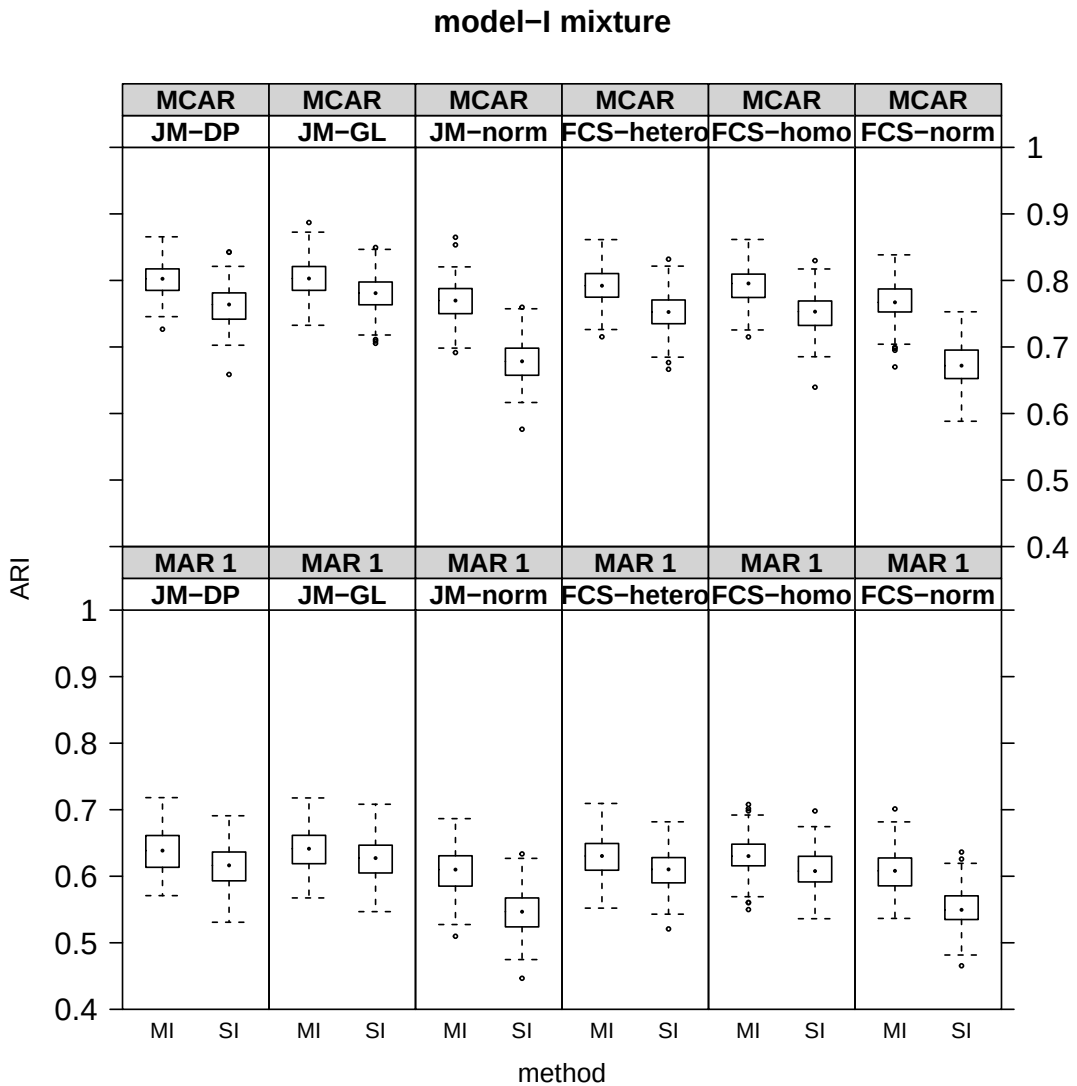
3. Pooling by NMF

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

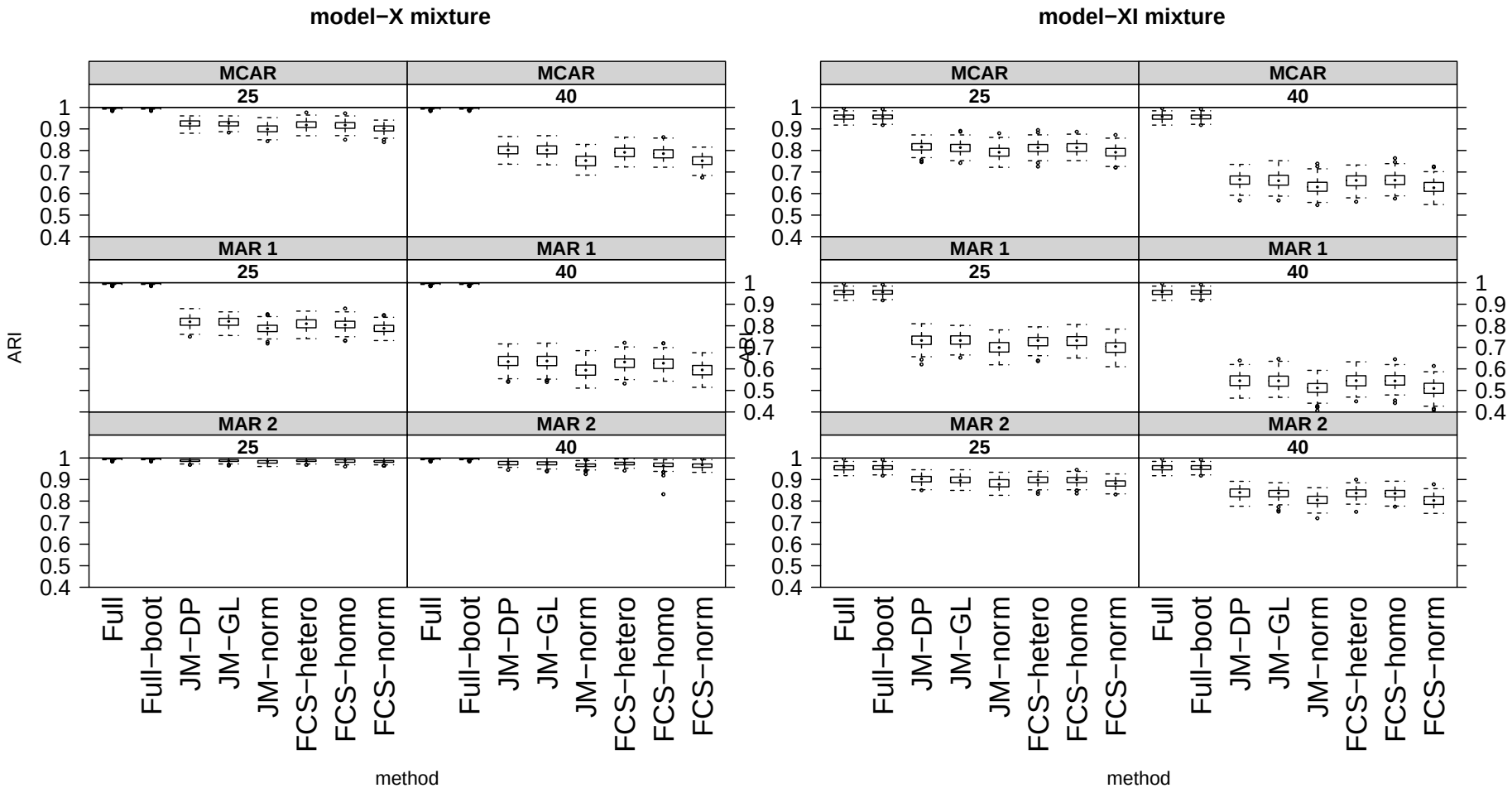
Results: base-case (1)



Results: base-case (2)

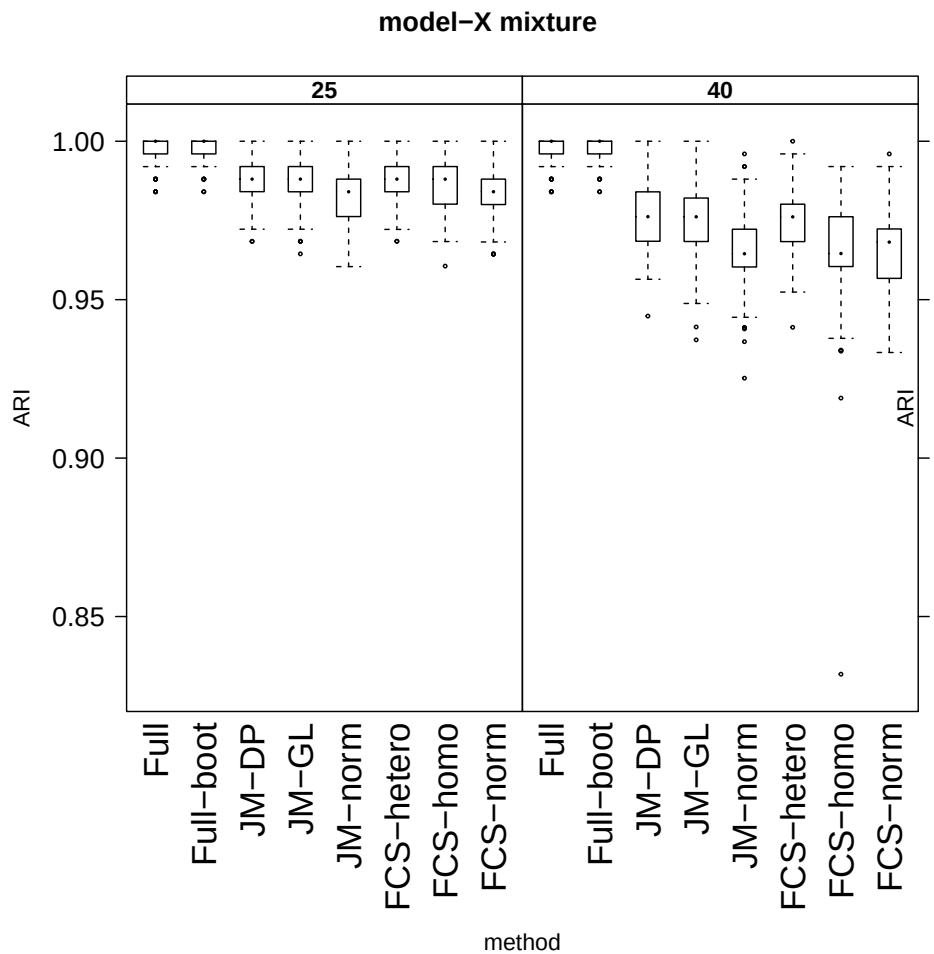


Results: heteroscedastic-case (1)

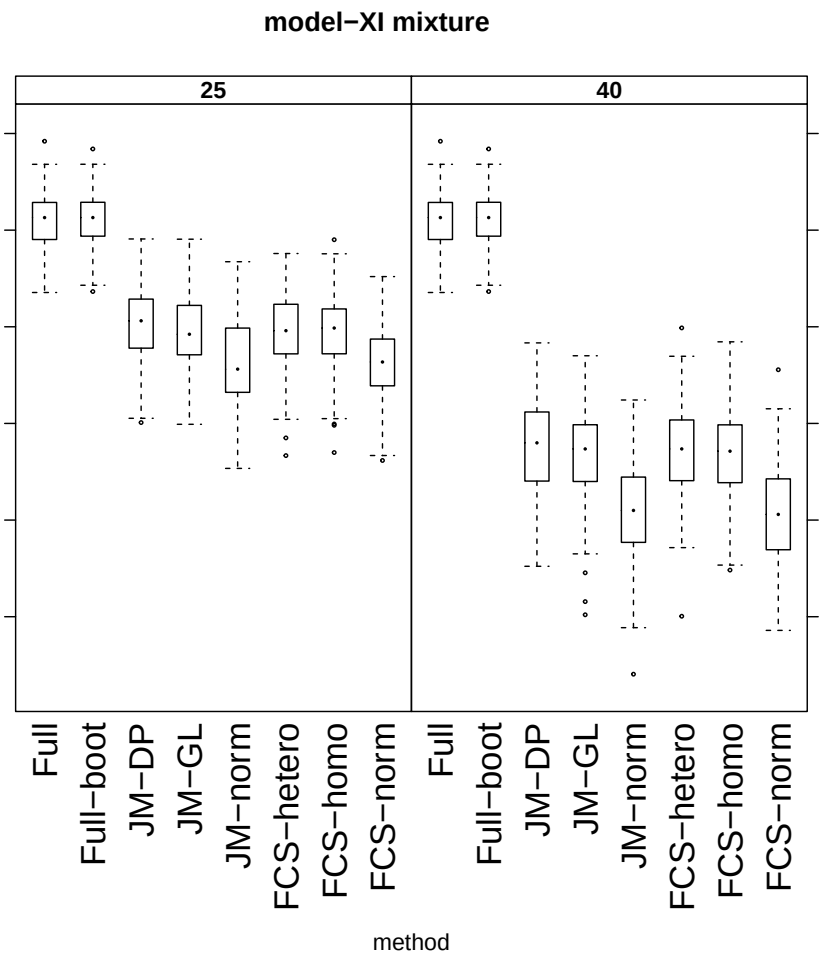


(a) model-X (b) model-XI

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

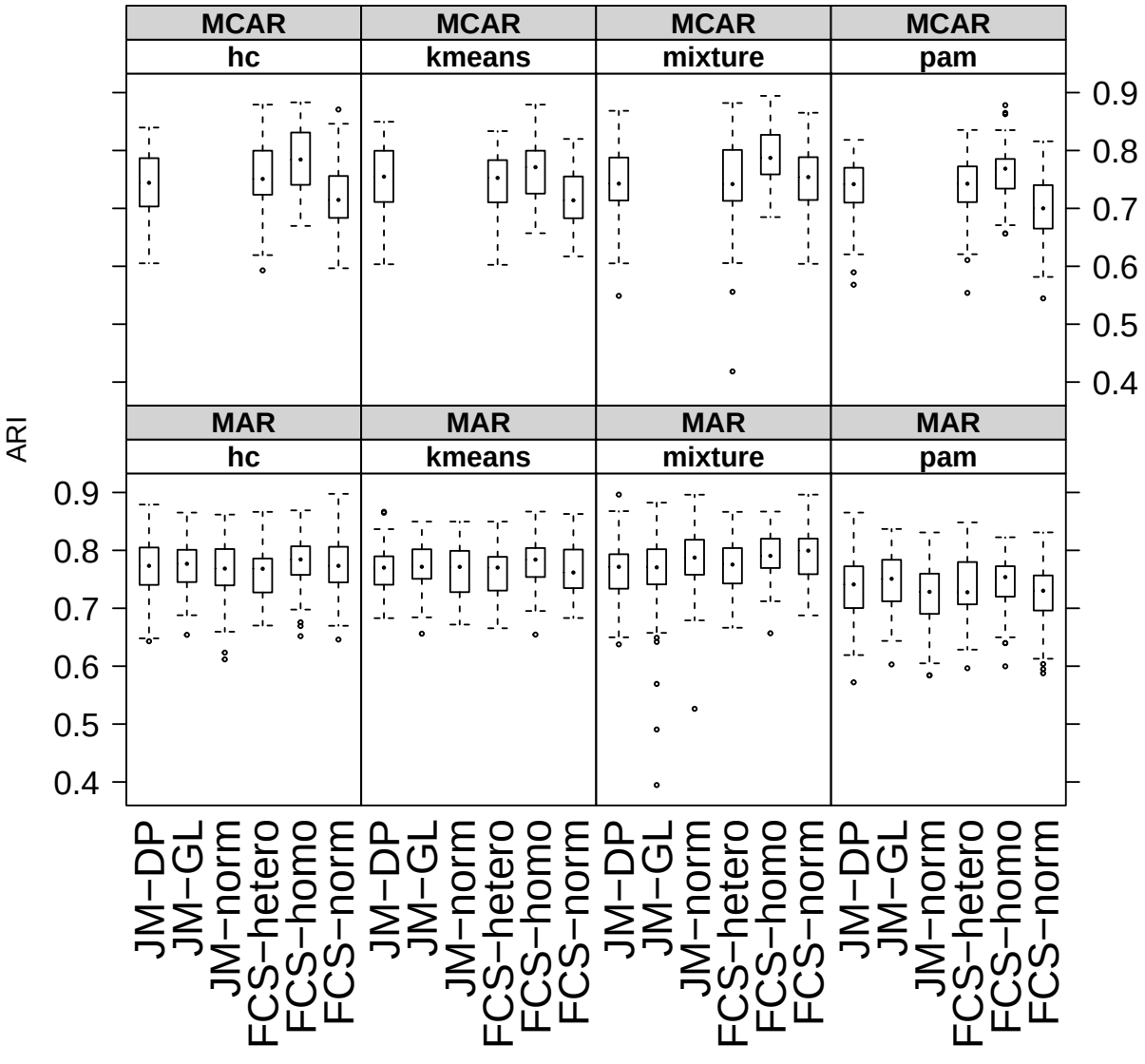
40% missing values (MCAR or a MAR mechanism)

100 missing data patterns per mechanism

For each missing data pattern:

variable selection to specify conditional imputation models in FCS (Bar-Hen and Audigier, 2020)

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 (or even better!) 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

Jocelyn T. Chi, Eric C. Chi, and Richard G. Baraniuk. k-pod: A method for k-means clustering of missing data. *The American Statistician*, 70(1):91–99, 2016. doi: 10.1080/00031305.2015.1086685.

K. Honda, R. Nonoguchi, A. Notsu, and H. Ichihashi. Pca-guided k-means clustering with incomplete data. In *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*, pages 1710–1714, 2011. doi: 10.1109/FUZZY.2011.6007312.

K Wagsta□. Clustering with missing values: No imputation required. In D. Banks, F. R. McMorris, P. Arabie, and W. Gaul, editors, *Classification, Clustering, and Data Mining Applications*, pages 649–658, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-642-17103-1.

Liyong Zhang, Wei Lu, Xiaodong Liu, Witold Pedrycz, and Chongquan Zhong. Fuzzy c-means clustering of incomplete data based on probabilistic information granules of missing values. *Knowledge-Based Systems*, 99:51–70, 2016.

R. J. Hathaway and J. C. Bezdek. Fuzzy c-means clustering of incomplete data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 31(5): 735–744, 2001. doi: 10.1109/3477.956035.

Wang Miao, Peng Ding, and Zhi Geng. Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 111(516):1673–1683, 2016. doi: 10.1080/01621459.2015.1105808.

References II

Matthieu Marbac, Mohammed Sedki, and Tienne Patin. Variable selection for mixed data clustering: application in human population genomics. *Journal of Classification*, pages 1–19, 2019.

Marie Du Roy de Chaumaray and Matthieu Marbac. Clustering data with nonignorable missingness using semi-parametric mixture models, 2020.

D. Rubin. *Multiple Imputation for Non-Response in Survey*. Wiley, New-York, 1987.

D. Plaehn. Revisiting french tomato data: Cluster analysis with incomplete data. *Food Quality and Preference*, 76:146 – 159, 2019. ISSN 0950-3293. doi: https://doi.org/10.1016/j.foodqual.2019.03.014.

L. Faucheux, M. Resche-Rigon, E. Curis, V. Soumelis, and S. Chevret. Clustering with missing and left-censored data: A simulation study comparing multiple-imputation-based procedures. *Biometrical Journal*, n/a(n/a), 2020. doi: 10.1002/bimj.201900366.

L. Bruckers, G. Molenberghs, and P. Dendale. Clustering multiply imputed multivariate high-dimensional longitudinal profiles. *Biometrical Journal*, 59(5): 998–1015, 2017. doi: 10.1002/bimj.201500027.

X. Basagana, J. Barrera-Gomez, M. Benet, J. M. Anto, and J. Garcia-Aymerich. A Framework for Multiple Imputation in Cluster Analysis. *American Journal of Epidemiology*, 177(7):718–725, 2013. doi: 10.1093/aje/kws289.

References III

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

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.

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.

J. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC, London, 1997.

James Honaker, Gary King, and Matthew Blackwell. Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47, 2011. URL http://www.jstatsoft.org/v45/i07/.

M. Quartagno and J. Carpenter. Multiple imputation for IPD meta-analysis: allowing for heterogeneity and studies with missing covariates. *Statistics in Medicine*, 35 (17):2938–2954, 2016. ISSN 1097-0258.

References IV

J. Schafer. Imputation of missing covariates under a multivariate linear mixed model. Technical report, Dept. of Statistics, The Pennsylvania State University, 1997.

R. A. Hughes, I. R. White, S. Seaman, J. Carpenter, K. Tilling, and J. Sterne. Joint modelling rationale for chained equations. *BMC Medical Research Methodology*, 14 (1):28, 2014.

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.

H. J Kim. *DPImputeCont*, 2020. URL https://github.com/hang-j-kim/DPImputeCont. R package version 1.2.2.

J. Schafer. *mix: Estimation/Multiple Imputation for Mixed Categorical and Continuous Data*, 2017. URL https://CRAN.R-project.org/package=mix. R package version 1.0-10.

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

Avner Bar-Hen and Vincent Audigier. An ensemble learning method for variable selection: application to high dimensional data and missing values, 2020.

A. Asuncion and D.J. Newman. UCI machine learning repository, 2007. URL http://archive.ics.uci.edu/ml.