Multiple imputation for clustering on incomplete data

V. Audigier, N. Niang

CNAM, CEDRIC-MSDMA, Paris

Machine Learning Workshop, Centre Borelli, September 27th, 2023

Clustering

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

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

Aim identify $\mathcal{C}_i$ for each i based on individual profiles $(\mathrm{x}_i)_{1 \leq i \leq n}$

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, X is frequently **incomplete**... $x_i = (x_i^{obs}, x_i^{miss})$

Ad-hoc methods

- removing incomplete observations
- removing incomplete variables
- single imputation

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)

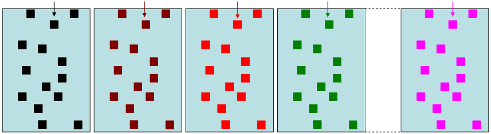
Multiple Imputation (MI)

- 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

$P\left(X^{miss}|X^{obs},\zeta_1\right)$ . . . . $P\left(X^{miss}|X^{obs},\zeta_M\right)$



2. Fit the analysis model on each imputed data set: $\hat{\psi}_m, \widehat{\text{Var}}\left(\hat{\psi}_m\right)$

3. Combine the results using Rubin's rules

   1. $\hat{\psi} = \frac{1}{M}\sum_{m=1}^{M}\hat{\psi}_m$
   2. $T = \frac{1}{M}\sum_{m=1}^{M}\widehat{\text{Var}}\left(\hat{\psi}_m\right) + \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\psi}_m - \hat{\psi}\right)^2$

Challenges in clustering

MI is not tailored for cluster analysis

- How to impute the incomplete data set?
- How to “average” partitions?
- How to assess a “variability” accounting for missing values?

Some works on the “averaging” step

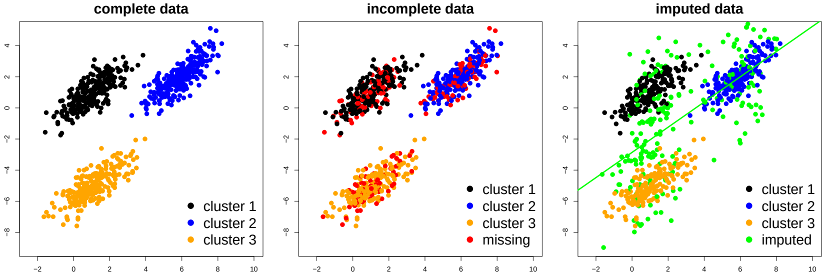
- by stacking (Plaehn, 2019)
- by using consensus clustering methods (Faucheux et al., 2020; Bruckers et al., 2017; Basagana et al., 2013)

Aim: highlighting how imputation, analysis and pooling steps should be carried out

Outline

1. Introduction
2. MI for clustering
   - Imputation
   - Analysis
   - Pooling
3. Direct Methods
   - k-POD
   - Optimal Completion Strategy
   - Ignorable GMM
4. Simulation study
   - Simulated data
   - Real data
5. Conclusion

Imputation model for clustering: the issue



JM-DP (Kim et al., 2014)

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

Based on a Bayesian formulation

$$\mu_k|\Sigma_k \sim \mathcal{N}\left(\mu_0, h^{-1}\Sigma_k\right) \quad \Sigma_k \sim \mathcal{W}^{-1}\left(df, G\right)$$

$$\text{with } diag(G) = (g_1, ..., g_p) \;\; g_j \sim \mathcal{G}\left(a_0, b_0\right)$$

$$\theta_k = v_k \prod_{\ell<k} (1 - v_\ell)$$

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

- 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 Data-Augmentation algorithm

Properties

JM-DP

- accounts for the heterogeneity
- accounts for the heteroscedasticity
- the number of clusters is only bounded

Modeling assumptions

- based on the normality assumption

R package **DPImputeCont** (Kim, 2020)

Other methods

Several relevant alternatives

- JM-GL (Schafer, 1997)
- FCS-homo, FCS-hetero (Audigier et al., 2021)

FCS methods can be easily modified

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

Available in the R package clusterMI (Audigier and Niang, 2023)

Analysis

Apply the cluster analysis to each imputed data set: Ψ_m, V_m

$\Psi_m \rightarrow$ kmeans, GMM, ... for V_m Fang and Wang (2012) proposed:

- generate C bootstrap pairs $\left(\mathrm{X}_c, \tilde{\mathrm{X}}_c\right)_{1\leq c\leq C}$ from X
- perform cluster analysis from $\left(\mathrm{X}_c, \tilde{\mathrm{X}}_c\right)_{1\leq c\leq C}$ to obtain $\left(\Psi_c, \tilde{\Psi}_c\right)$
- classify individuals of X 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

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

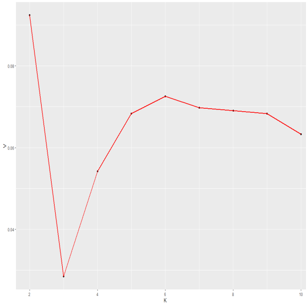


Figure: Instability (V) according to the number of clusters (K)

Partitions pooling

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

With **complete data** Jain (2017) extended to partitions

- the expected mean

$$\underset{\Psi\in\mathcal{P}_{n,K}}{arg\ min} \int_{\mathcal{P}_{n,K}} \delta^\alpha(\Psi^\star, \Psi)d\pi(\Psi^\star)$$

with δ a dissimilarity, $\alpha \in \mathbb{N}$, $\mathcal{P}_{n,K}$ the set of partitions of n observations in K clusters

- the mean estimate

$$\underset{\Psi\in\mathcal{P}_{n,K}}{arg\ min} \sum_{j=1}^{J} \delta^\alpha(\Psi, \Psi_j) \qquad (1)$$

with $(\Psi_j)_{1\leq j\leq J}$ a set of observed partitions

After MI

$$\widehat{\Psi} = \underset{\Psi\in\mathcal{P}_{n,K}}{arg\ min} \sum_{m=1}^{M} \delta^\alpha(\Psi, \Psi_m) \qquad \textit{(median partition problem)}$$

Properties

- Theoretically appealing, but solving (1) is highly challenging
- Iterative algorithms are required (Vega-Pons and Ruiz-Shulcloper, 2011)

Partitions pooling

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

With **complete data** Jain (2017) extended to partitions

- the expected mean

$$\underset{\Psi \in \mathcal{P}_{n,K}}{arg\ min} \int_{\mathcal{P}_{n,K}} \delta^\alpha(\Psi^\star, \Psi) d\pi(\Psi^\star)$$

with δ a dissimilarity, $\alpha \in \mathbb{N}$, $\mathcal{P}_{n,K}$ the set of partitions of n observations in K clusters

- the mean estimate

$$\underset{\Psi \in \mathcal{P}_{n,K}}{arg\ min} \sum_{j=1}^{J} \delta^\alpha(\Psi, \Psi_j) \qquad (1)$$

with $(\Psi_j)_{1 \leq j \leq J}$ a set of observed partitions

After MI

$$\widehat{\Psi} = \underset{\Psi \in \mathcal{P}_{n,K}}{arg\ min} \sum_{m=1}^{M} \delta^\alpha(\Psi, \Psi_m) \qquad \textit{(median partition problem)}$$

Properties

- Theoretically appealing, but solving (1) is highly challenging
- Iterative algorithms are required (Vega-Pons and Ruiz-Shulcloper, 2011)

Mirkin-based methods

δ chosen as the number of disagreements between partitions

$$\delta\left(\Psi, \Psi'\right) = \sum_{(i,i')} \delta_{ii'} \qquad \delta_{ii'} = \begin{cases} 1 & \text{if } i \text{ and } i' \text{ belong to the same cluster} \\ & \text{in one partition and not in the other} \\ 0 & \textit{otherwise} \end{cases}$$

Two methods can be exhibited

1. BOK: the space of solutions is constrained to $(\Psi_m)_{1\leq m\leq M}$ instead of $\mathcal{P}_{n,K}$
2. SAOM: the BOK solution is improved by using stochastic relabeling of individuals

Properties

- The error for the BOK solution does not exceed two times the error of the optimal partition (Filkov and Skiena, 2004)

$$\sum_{m=1}^{M} \delta(\Psi_{BOK}, \Psi_m) \leq 2 \sum_{m=1}^{M} \delta(\Psi_{opt}, \Psi_m)$$

- SAOM provides a better solution, but computationally intensive

NMF-based methods

Non negative matrix factorization is powerful method widely used for solving many optimization problems

Principle

- consider the Mirkin distance for δ
- rewrite the optimization problem in terms of connectivity matrices $(\mathrm{U}_m)_{1\leq m\leq M}$ instead of partitions $(\Psi_m)_{1\leq m\leq M}$

$$\underset{\Psi\in\mathcal{P}_n^K}{argmin}\sum_{m=1}^{M}\delta(\Psi,\Psi_m)\Longleftrightarrow\underset{\mathrm{U}\in\mathcal{U}}{argmin}\parallel\bar{\mathrm{U}}-\mathrm{U}\parallel_F^2$$

with $\bar{\mathrm{U}}=\frac{1}{M}\sum_{m=1}^{M}\mathrm{U}_m$

Properties

- can be solved using various algorithms (Lee and Seung, 2001; Li et al., 2007)
- monotone convergence

Instability after MI (Audigier and Niang, 2022)

Following Fang and Wang (2012), the **within instability** can be assessed by

$$\frac{1}{M}\sum_{m=1}^{M} V_m$$

while the **between instability** can be computed by averaging the proportions of disagreements

$$\frac{1}{M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M}\delta\left(\Psi_m,\Psi_{m'}\right)/n^2$$

the total instability T is

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

Properties

Based on a simulation study

- pooling partitions using NMF-based methods is less time consuming and more accurate than Mirkin-based methods
- a larger value for M improves the partition accuracy ($M \approx 20$)
- T provides an accurate estimate for the number of clusters with missing values

Outline

1. Introduction
2. MI for clustering
   - Imputation
   - Analysis
   - Pooling
3. Direct Methods
   - k-POD
   - Optimal Completion Strategy
   - Ignorable GMM
4. Simulation study
   - Simulated data
   - Real data
5. Conclusion

k-POD (Chi et al., 2016)

Chi et al. (2016) proposed a direct method for k-means clustering

k-means

$$\underset{\mathrm{A}\in\mathcal{H},\mathrm{B}}{\arg\min} \parallel X - \mathrm{AB} \parallel_F^2$$

k-POD

$$\underset{\mathrm{A}\in\mathcal{H},\mathrm{B}}{\arg\min} \parallel P_\Omega\left(X\right) - P_\Omega\left(\mathrm{AB}\right) \parallel_F^2$$

- $\mathcal{H}$ set of membership matrices $(n \times K)$, $B_{K\times p}$ matrix of centers coordinates
- $\parallel . \parallel_F$ Frobenius norm
- $\Omega \subset \{1, \dots, n\} \times \{1, \dots, p\}$ → subset of the indices for observed entries
- P_Ω projection operator so that $[P_\Omega(X)]_{ij} = \begin{cases} x_{ij} \text{ if } (i,j) \in \Omega \\ 0 \text{ otherwise} \end{cases}$

The criterion is optimised by alternating imputation by b_k and kmeans clustering

Available in the **kpodclustr** R package

FCM by optimal completion strategy (Hathaway and Bezdek, 2001)

fuzzy c-means

$$\underset{\Gamma,B}{argmin} \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ki}^{\alpha} \parallel x_i - b_k \parallel_2^2$$

fuzzy c-means OCS

$$\underset{\Gamma,B,X^{miss}}{argmin} \sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ki}^{\alpha} \parallel (x_i^{obs}, x_i^{miss}) - b_k \parallel_2^2$$

with $\Gamma = (\gamma_{ki})_{\substack{1 \leq k \leq K \\ 1 \leq i \leq n}}$ degrees of membership; α fuzzification parameter

The criterion is optimised by alternating FCM and imputation by weighted centers

$$\gamma_{ki}^{\alpha} \leftarrow 1 \Bigg/ \sum_{\ell=1}^{K} \left(\frac{\parallel x_i - b_k \parallel_2^2}{\parallel x_i - b_\ell \parallel_2^2} \right)^{\frac{1}{\alpha-1}}$$

$$b_{kj} \leftarrow \left(\sum_{i=1}^{n} \gamma_{ki}^{\alpha} x_{ij} \right) \Bigg/ \left(\sum_{i=1}^{n} \gamma_{ki}^{\alpha} \right)$$

$$x_{ij}^{miss} \leftarrow \left(\sum_{k=1}^{K} \gamma_{ki}^{\alpha} b_{kj} \right) \Bigg/ \left(\sum_{k=1}^{K} \gamma_{ki}^{\alpha} \right)$$

Ignorable-GMM (Marbac et al., 2019)

Gaussian mixture models (GMM)

$$f(x;\theta) = \sum_{k=1}^{K} \pi_k f_k(x;\theta_k) \qquad \theta = (\theta_k)_{1\leq k\leq K}, \theta_k = (\mu_k, \Sigma_k)$$

Log-likelihood GMM

$$\sum_{i=1}^{n} log \sum_{k=1}^{K} \pi_k \prod_{j=1}^{p} f_{kj}(x_{ij};\theta_{kj})$$

Log-likelihood ignorable-GMM

$$\sum_{i=1}^{n} log \sum_{k=1}^{K} \pi_k \prod_{j\in O_i} f_{kj}(x_{ij};\theta_{kj})$$

$O_i \subseteq 1,\ldots,p$ the subset of variables indices that are observed for individual i

The criterion is optimised by using an EM algorithm

Available in the **VarSelLCM** R package

Outline

1. Introduction
2. MI for clustering
   - Imputation
   - Analysis
   - Pooling
3. Direct Methods
   - k-POD
   - Optimal Completion Strategy
   - Ignorable GMM
4. Simulation study
   - Simulated data
   - Real data
5. Conclusion

Simulation design: data generation

Complete data generation

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

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

$n_k = 250$ (for all k 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: $Prob(r_{ij} = 0) = \tau \qquad \forall i, j$
- MAR: $Prob(r_{ij} = 0) = \Phi(a_\tau + x_{i1}) \qquad \forall j \neq 1$
- $\tau \in \{10\%, 25\%, 40\%\}$

Simulation design: evaluation

For each incomplete data set (200 per configuration)

Multiple Imputation

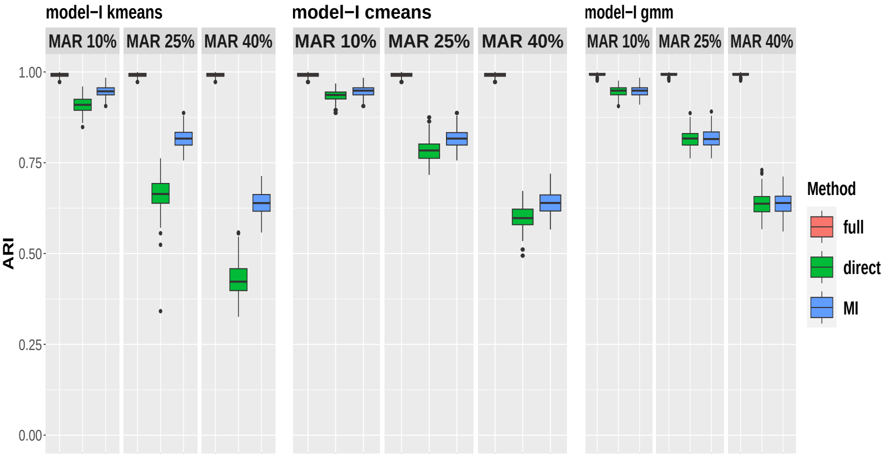
1. Imputation ($M = 20$, using JM-DP)
2. Cluster analysis
   - k-means
   - fuzzy c-means
   - clustering by GMM
3. Pooling by NMF

Direct Methods

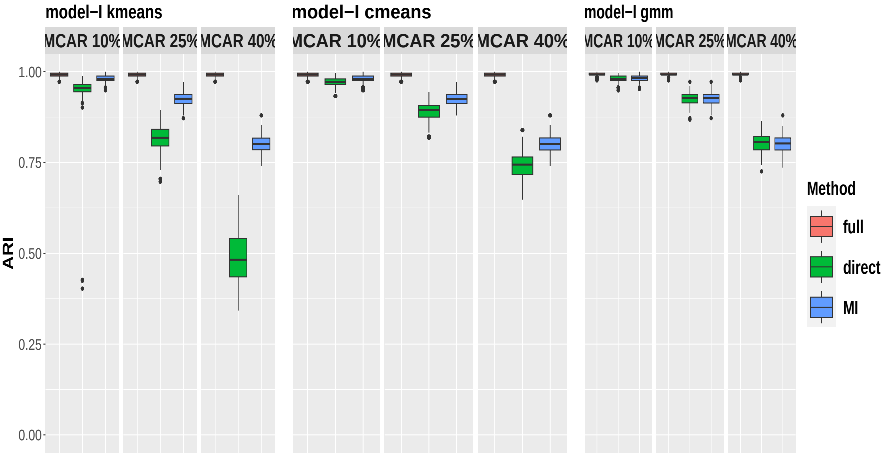
- k-POD (Chi et al., 2016) from the *kpodclustr* R package
- FCM by optimal completion strategy (Hathaway and Bezdek, 2001)
- Ignorable-GMM (Marbac et al., 2019) from the *VarSelLCM* R package

Criteria ARI, Full data

Results: base-case, MAR



Results: base-case, MCAR



Summary

Based on **simulated data** under mixture model distribution

- MI outperforms direct methods for kmeans and fuzzy c means
- MI and direct methods provide similar ARI for GMM
- Differences between MI and direct method highlighted for kmeans with more separated clusters
- Similar results by modifying the number of clusters, the cluster size, the balance between clusters sizes and the heteroscedasticity

Real data sets

Data

	Variables						Size of cluster	
	n	p	Type	Number for which Shapiro rejects normality	K	Silhouette Index	(min)	(max)
wine	178	13	Real	7	3	0.57	48	71
ovarian	216	100	Real	64	2	0.50	95	121
iris	150	4	Real	1	3	0.52	50	50
glass	214	9	Real	9	2	0.56	51	163
breast cancer	699	9	Discrete	9	2	0.59	241	458

- Gaussian assumption seems not observed
- p large
- n_k small compared to p
- partitions not obvious

Simulation design

Data sets generation

- 25% missing values (MCAR or a MAR mechanism)
- 200 missing data patterns per mechanism

Data sets analysis

- missing values are addressed by MI (FCS-homo, $M = 20$)
- cluster analysis by k-means, fuzzy c-means or GMM
- pooling using NMF

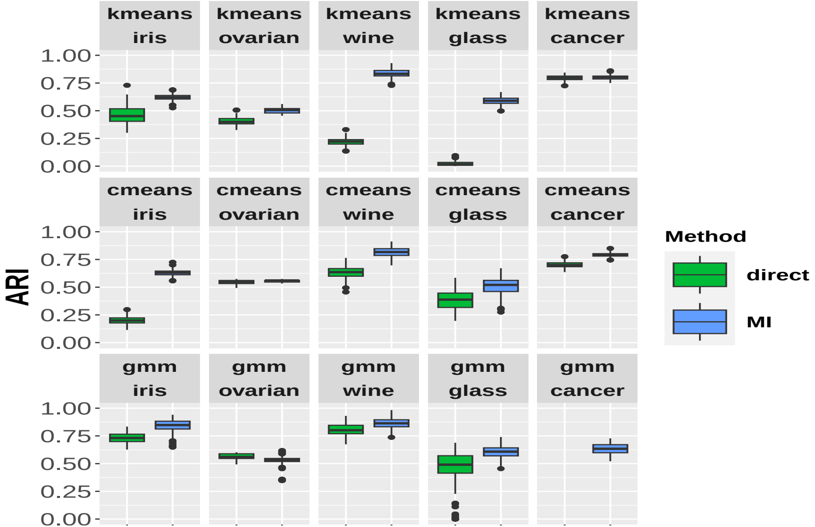
Evaluation

Competitive direct methods

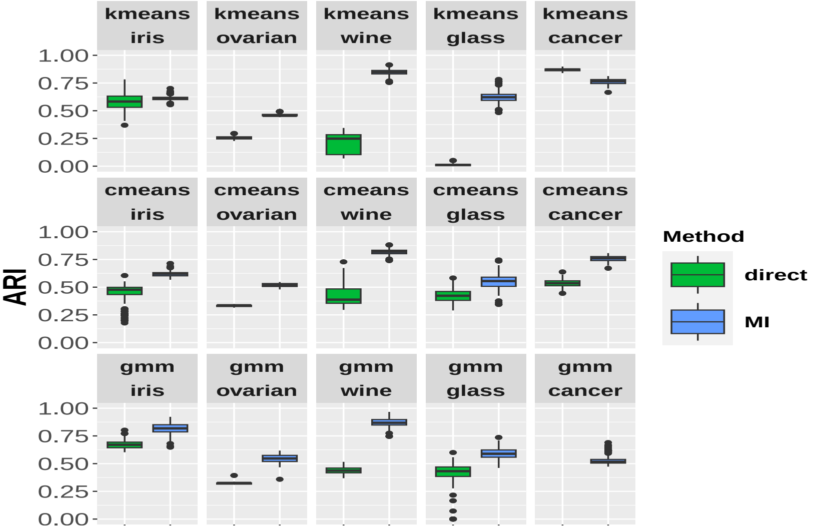
- k-POD (Chi et al., 2016) from the *kpodclustr* R package
- FCM by optimal completion strategy (Hathaway and Bezdek, 2001)
- Ignorable-GMM (Marbac et al., 2019) from the *VarSelLCM* R package

Criteria: Adjusted Rand Index

Real data, MCAR



Real data, MAR



Take-home message

MI is a **competitive method** for addressing missing values in clustering

- for model-based or distance-based methods
- good performances on real data

In practice

- A **suitable imputation model** is required
- A large value for M is recommended
- The number of clusters can be easily estimated
- MI method is available in the clusterMI R package

Some perspectives

- Addressing mixed data
- Developing indices for clustering with missing values

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 Wagstaff. 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

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

Vincent Audigier, Ndèye Niang, and Matthieu Resche-Rigon. Clustering with missing data: which imputation model for which cluster analysis method?, 2021.

F. M. Zahid and C. Heumann. Multiple imputation with sequential penalized regression. *Statistical Methods in Medical Research*, 28(5):1311–1327, 2019. doi: 10.1177/0962280218755574. PMID: 29451087.

Matthias Templ, Alexander Kowarik, and Peter Filzmoser. Iterative stepwise regression imputation using standard and robust methods. *Computational Statistics & Data Analysis*, 55(10):2793 – 2806, 2011. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2011.04.012. URL `http://www.sciencedirect.com/science/article/pii/S0167947311001411`.

References IV

Tim P Morris, Ian R White, and Patrick Royston. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC medical research methodology*, 14(1):75, 2014.

Vincent Audigier and Ndeye Niang. *clusterMI: Cluster analysis with missing values by multiple imputation*, 2023. URL `https://vincentaudigier.weebly.com/research.html`. R package version 0.0.17.

Y. Fang and J. Wang. Selection of the number of clusters via the bootstrap method. *Comput. Stat. Data Anal.*, 56(3):468–477, 2012. ISSN 0167-9473. doi: 10.1016/j.csda.2011.09.003. URL `https://doi.org/10.1016/j.csda.2011.09.003`.

B. J. Jain. Consistency of mean partitions in consensus clustering. *Pattern Recognition*, 71:26 – 35, 2017. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2017.04.021.

S. Vega-Pons and J. Ruiz-Shulcloper. A survey of clustering ensemble algorithms. *IJPRAI*, 25(3):337–372, 2011.

V. Filkov and S. Skiena. Integrating microarray data by consensus clustering. *International Journal on Artificial Intelligence Tools*, 13(04):863–880, 2004. doi: 10.1142/S0218213004001867.

References V

Daniel Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2001. URL https://proceedings.neurips.cc/paper/2000/file/f9d1152547c0bde01830b7e8bd60024c-Paper.pdf.

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.

Vincent Audigier and Ndèye Niang. Clustering with missing data: which equivalent for Rubin's rules? *Advances in Data Analysis and Classification*, September 2022. doi: 10.1007/s11634-022-00519-1. URL https://hal.science/hal-03766733.