Clustering on incomplete data: direct method or multiple imputation?

V. Audigier, F. Sadou Zouleya

CNAM, CEDRIC-MSDMA, Paris
ENSAE, Paris

54emes journées de statistique, July 6th, 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

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

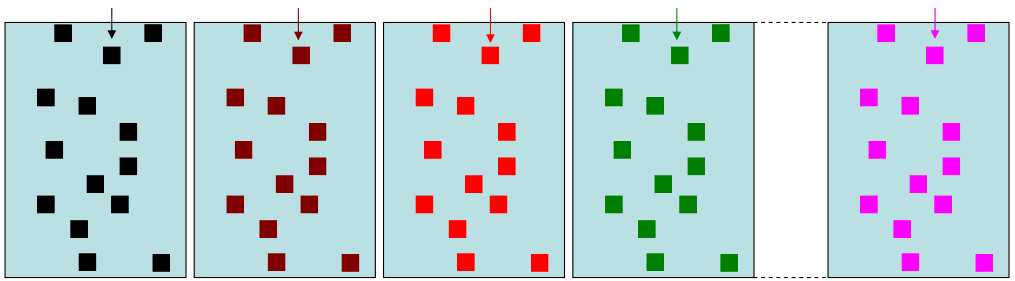
Advanced methods

- **Multiple Imputation (MI)**: addressing missing values by imputation and then performing cluster analysis
- **Direct methods**: optimising criterion based on observed values only (ex: likelihood, least squares criterion, ...)

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

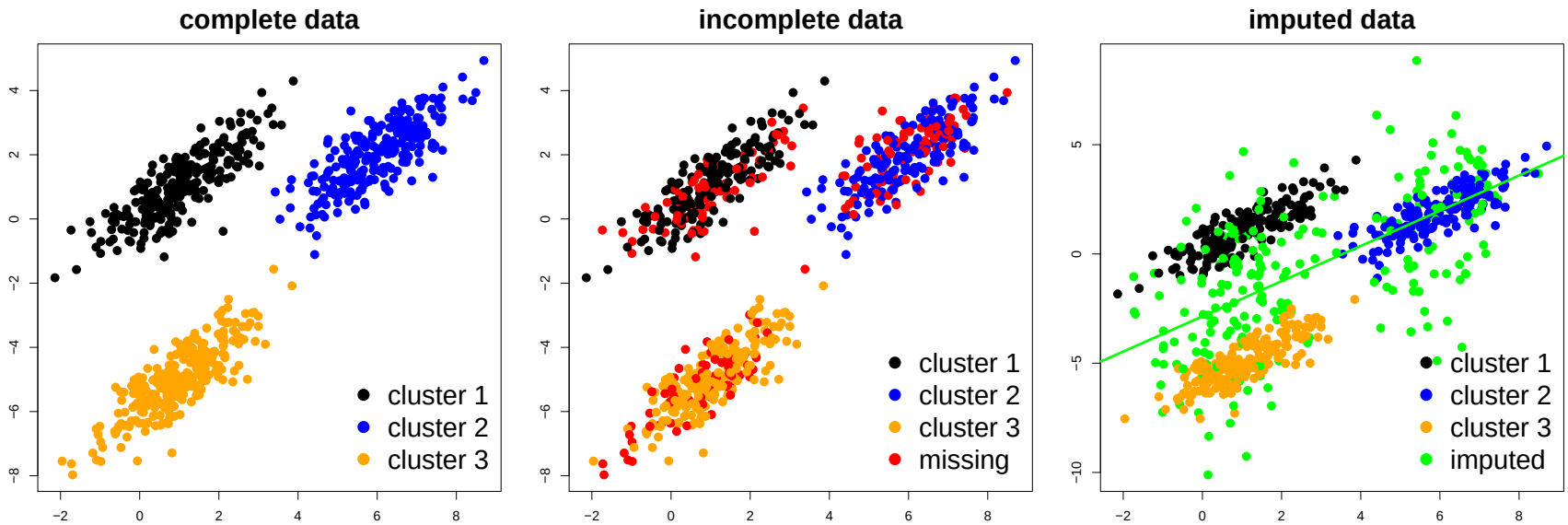
$$P\left(X^{miss} | X^{obs}, \zeta_1\right) \quad \ldots \quad \ldots \quad \ldots \quad P\left(X^{miss} | X^{obs}, \zeta_M\right)$$



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

3. Combine the results to obtain one partition Ψ and one associated instability V

Imputation step: the issue



MI methods for clustering

The imputation model needs to account for the underlying assumption related to the cluster analysis

- variable per variable imputation: FCS-homo, FCS-hetero (Audigier et al., 2021)
- joint imputation based on mixture: JM-GL (Schafer, 1997), JM-DP (Kim et al., 2014)

DP mixture of products of multivariate normal distributions

$$\mu_k|\Sigma_k \sim \mathcal{N}\left(\mu_0, h^{-1}\Sigma_k\right) \quad \Sigma_k \sim \mathcal{W}^{-1}\left(df, G\right)$$
$$\text{with } G = (g_1, ..., g_p) \;\; g_j \sim \mathcal{G}\left(a_0, b_0\right)$$
$$\pi_k = v_k \prod_{\ell<k} (1 - v_\ell)$$

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

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 $\left(X_c, \tilde{X}_c\right)_{1 \leq c \leq C}$ from X
- perform cluster analysis from $\left(X_c, \tilde{X}_c\right)_{1 \leq c \leq C}$ 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

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

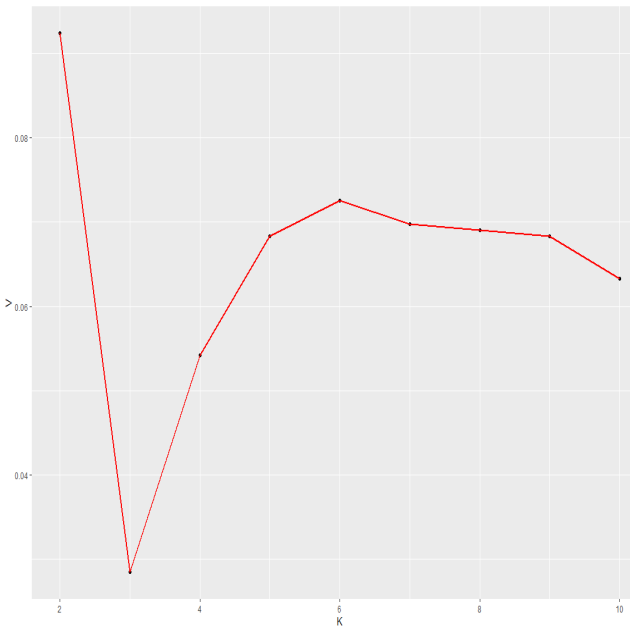


Figure: Instability (V) according to the number of clusters (K)

Pooling

Partitions pooling from $(\Psi_m)_{1\leq m\leq M}$

Faucheux et al. (2020); Bruckers et al. (2017); Basagana et al. (2013); Audigier and Niang (2022) proposed consensus clustering methods

Theoretically, NMF based methods are appealing

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

$$\underset{\mathrm{U}\in\mathcal{U}}{argmin} \parallel \mathrm{U} - \bar{\mathrm{U}} \parallel_F^2$$

The solution of this optimization problem can be obtained using NMF (Li et al., 2007).

Instabilities pooling from $(V_m)_{1\leq m\leq M}$

Audigier and Niang (2022) proposed

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

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

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

- $\mathcal{H}$ set of membership matrices $(n \times K)$, $B_{K\times p}$ matrix of centers coordinates
- $\parallel . \parallel_F$ Frobenius norm
- $\Omega \subset \{1,\ldots,n\} \times \{1,\ldots,p\} \rightarrow$ 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 $\Gamma = (\gamma_{ki})_{\substack{1 \le k \le K \\ 1 \le i \le n}}$ degrees of membership; α fuzzification parameter

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

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

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

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

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

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
mice	1077	77	Real	77	8	0.17	132	150

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

Simulation design

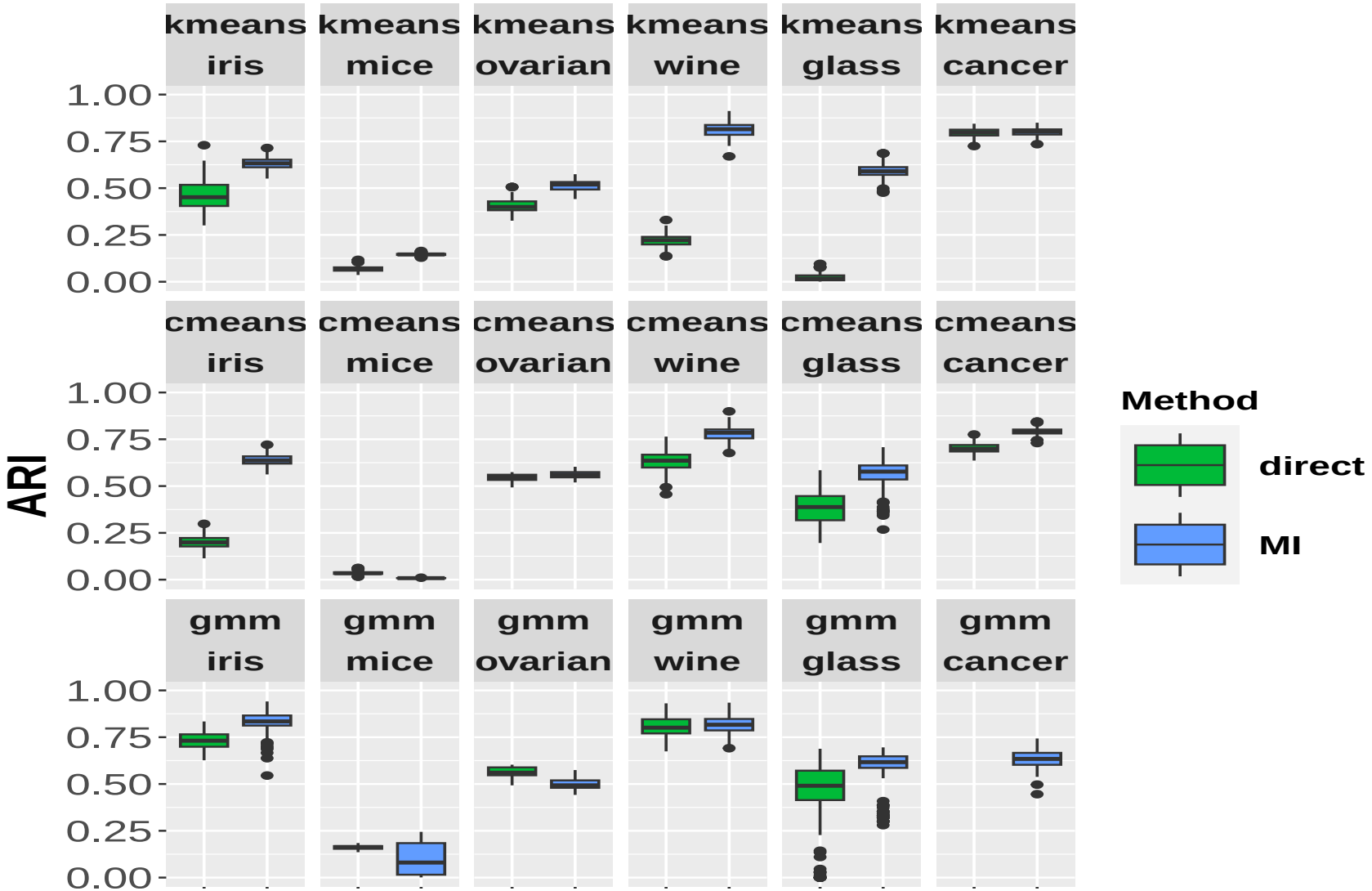
Data sets generation

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

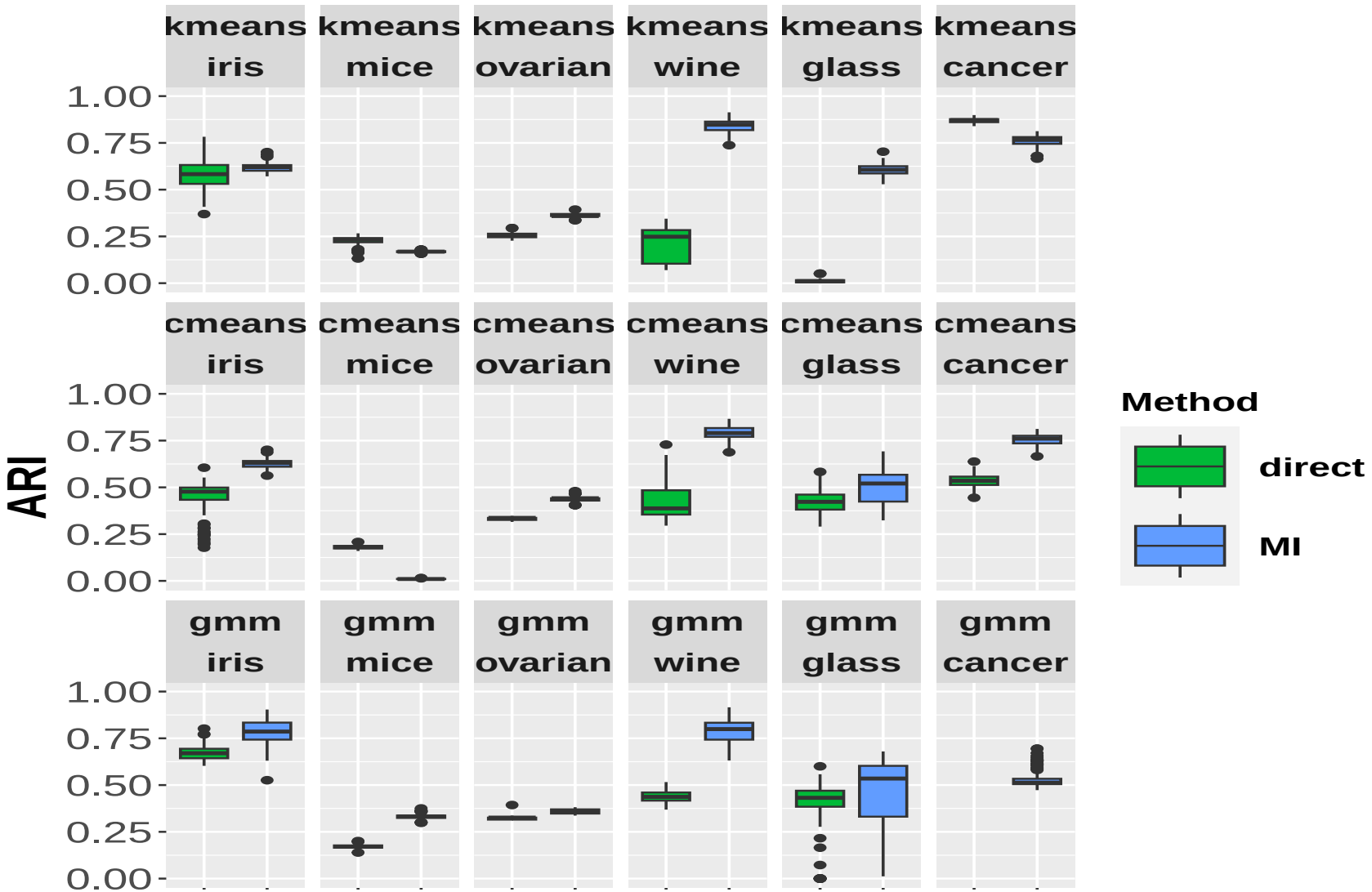
Data sets analysis

- cluster analysis by k-means, fuzzy c-means, GMM
- missing values are addressed by direct method or MI (JM-DP, $M = 20$)

Real data, MCAR



Real data, MAR



Summary

Based on **real datasets**

- MI generally outperforms direct methods for kmeans and fuzzy cmeans
- Under a MAR mechanism, MI outperforms direct method for GMM

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

- 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

Conclusion

This study shows

- MI looks like a **competitive method** for addressing missing values in clustering (for model-based or distance-based methods)
- **Clustering** is a case where direct and MI method provide different results

In practice

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

Some perspectives

- Theoretical comparison of both approaches
- Developing indices for clustering with missing values

References I

V. Audigier and N. Niang. Clustering with missing data: which equivalent for Rubin's rules? *Advances in Data Analysis and Classification*, 2022.

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

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`.

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.

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