

# Multiple imputation for clustering on incomplete data

**V. Audigier**, N. Niang

*CNAM, CEDRIC-MSDMA, Paris*

CMStatistics2023, December 18th, 2023

# Clustering

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

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

**Aim** identify  $C_i$  for each  $i$  using individual profiles  $(x_i)_{1 \leq i \leq n}$

## 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 (Wagstaff, 2004; Honda et al., 2011; Chi et al., 2016)
- fuzzy C-means (Hathaway and Bezdek, 2001; Zhang et al., 2016)
- Gaussian mixture (Miao et al., 2016; Marbac et al., 2019; McCaw et al., 2022)

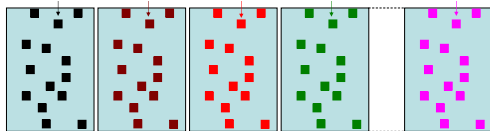
## Multiple Imputation (MI)

- 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

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



- 2 Apply the **cluster analysis** to each imputed data set:  $\Psi_m, V_m$
- 3 Combine the results to obtain one partition  $\Psi$  and one associated instability measure  $V$

# Outline

- ① Introduction
- ② Multiple imputation and clustering
  - Imputation
  - Analysis
  - Pooling
- ③ Simulation study
- ④ Conclusion

# Fully conditional specification

Two families of **imputation models**: JM and FCS

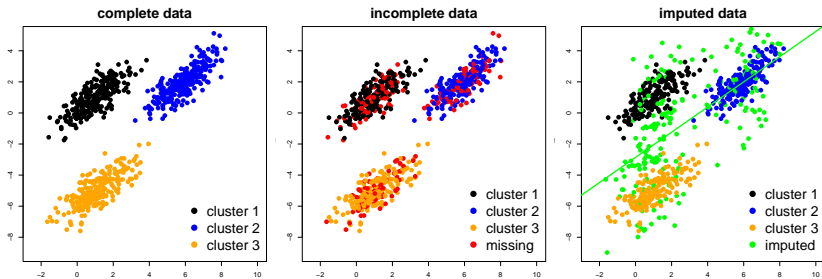
A conditional distribution is specified for each (incomplete) variable

$$\text{Ex : } P(X_j | X_{-j}; \zeta_j) = \mathcal{N}(X_{-j}\beta, \sigma^2) \quad \zeta_j = (\beta, \sigma)$$

To impute the  $m$ th data set

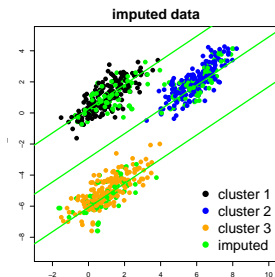
- initialize missing values of  $X$
- for  $j$  in  $1 \dots p$ 
  - a generate  $\zeta_j$  based on observed individuals on  $X_j$
  - b impute  $X_j^{miss}$  according to  $P(X_j | X_{-j}; \zeta_j)$
- repeat until convergence

# Imputation model for clustering: the issue



# FCS-homo (Audigier et al., 2021)

Addressing the issue by using regression models including the class variable  $W$  as explanatory variable



## FCS-homo

- generating  $X_j^{miss}$  given  $W$  is performed using regression models including an intercept specific to each cluster

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

- generating  $W$  given  $X$  by linear discriminant analysis

$$P(W = w | X; \zeta_w) \propto \exp(\delta_{w,x}) \quad \zeta_w = (\pi, \mu, \Sigma)$$



# Properties

## FCS-homo

- addresses the cluster structure
- assumes homoscedastic regression models and identical correlation between variables
- required a pre-defined number of clusters

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)
- ...

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

# Analysis

Apply the **cluster analysis** to each imputed data set:  $\Psi_m, V_m$

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

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

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

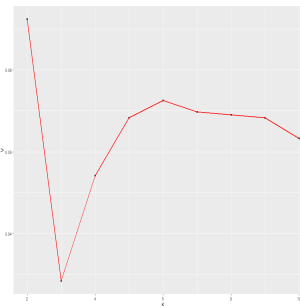


Figure: Instability ( $V$ ) according to the number of clusters ( $K$ )

# 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

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

$$\underset{U \in \mathcal{U}}{\operatorname{argmin}} \quad \| U - \bar{U} \|_F^2$$

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

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

# Outline

- ① Introduction
- ② Multiple imputation and clustering
  - Imputation
  - Analysis
  - Pooling
- ③ Simulation study
- ④ Conclusion

# Real data sets

## Data

	$n$	$p$	Variables			$K$	Silhouette Index	Size of cluster	
			Type	# of rejects in Shapiro test				(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

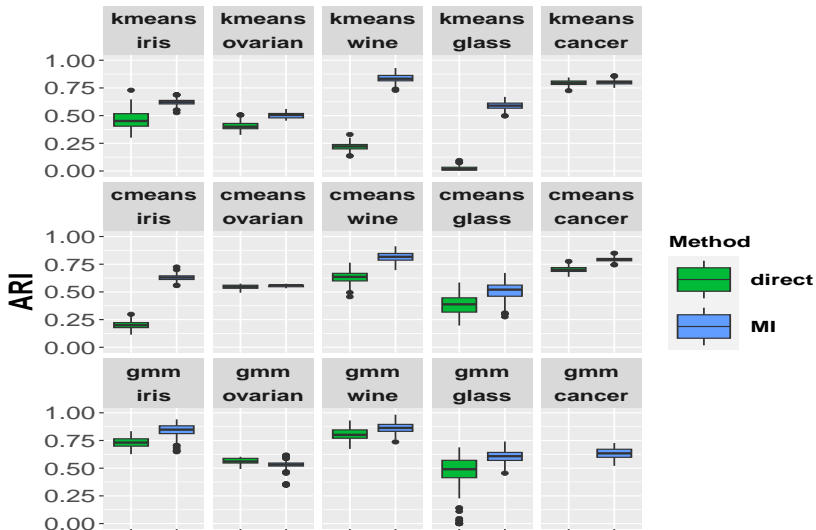
# Evaluation

## Competitive direct methods

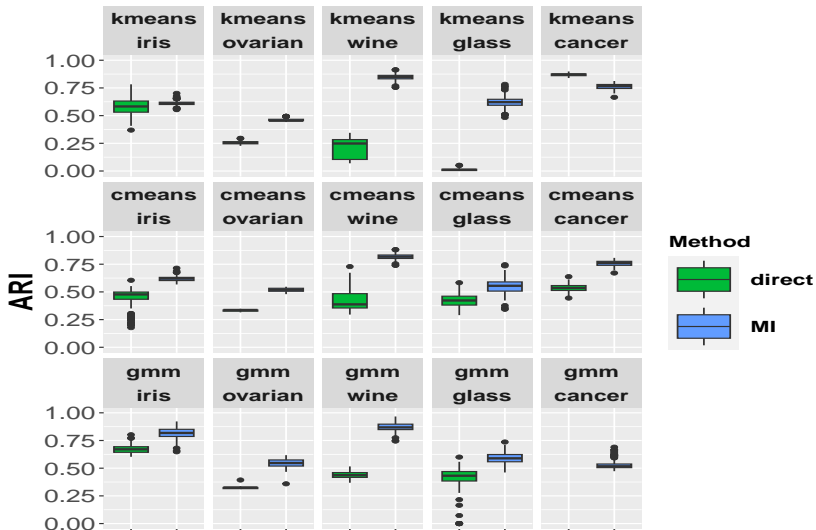
- **k-POD** (Chi et al., 2016, *The American Statistician*)  
implemented in the *kpodclustr* R package
- **FCM by optimal completion strategy** (Hathaway and Bezdek, 2001, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*)
- **Ignorable-GMM** (Marbac et al., 2019, *Journal of Classification*)  
implemented in the *VarSelLCM* R package

**Criteria:** Adjusted Rand Index

# Real data, MCAR



# Real data, MAR





# Simulated data

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
- Similar results by modifying the number of clusters, the cluster size, the balance between clusters sizes and the heteroscedasticity
- Differences between MI and direct method highlighted for kmeans with more separated clusters

# 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 **dedicated 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

- 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.