Cluster analysis after multiple imputation

V. Audigier, N. Niang

CNAM, CEDRIC-MSDMA, Paris

ASMDA, June 1st, 2021

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)

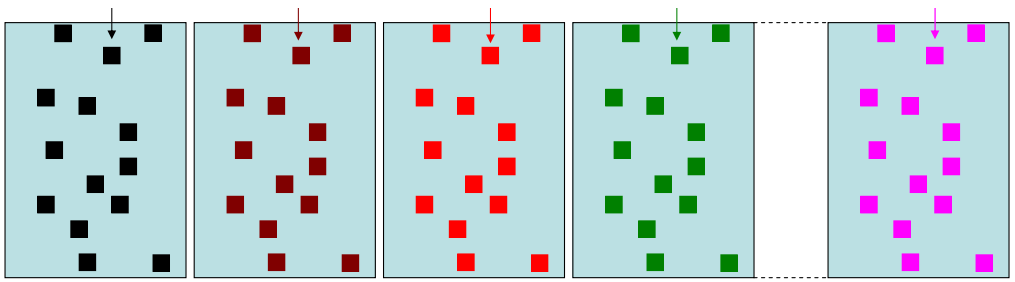
Multiple Imputation (MI)

- 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: $\hat{\psi}_m, \widehat{\text{Var}}\left(\hat{\psi}_m\right)$
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

1. How to "average" partitions?
2. 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)

However, no contribution for assessing a "variability" with missing values

Aim: highlighting rules for applying cluster analysis after MI

Partitions pooling

Ψ_m the partition from (Z^{obs}, Z_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

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

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

$$argmin_{\Psi \in \mathcal{P}_{n,K}} = \sum_{j=1}^{J} \delta^{\alpha}(\Psi, \Psi_j) \quad (1)$$

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

After MI

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

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 $\left(Z^{obs}, Z_m^{miss}\right)$, which average $\widehat{\Psi}$ for $(\Psi_m)_{1 \leq m \leq M}$?

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

- the expected mean

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

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

$$argmin_{\Psi \in \mathcal{P}_{n,K}} = \sum_{j=1}^{J} \delta^{\alpha}(\Psi, \Psi_j) \quad (1)$$

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

After MI

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

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

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

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 $(\mathbf{H}_m)_{1 \leq m \leq M}$ instead of partitions $(\Psi_m)_{1 \leq m \leq M}$

$$argmin_{\Psi} \sum_{m=1}^{M} \delta^{\alpha}(\Psi, \Psi_m) \Longleftrightarrow argmin_{\mathbf{H}} \parallel \mathbf{M} - \mathbf{H} \parallel^2$$

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

Properties

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

Instability with complete data

Which variability for Ψ?

From **complete data**, Fang and Wang (2012) assess instability in clustering

- generate C bootstrap pairs $\left(\mathbf{Z}_c, \tilde{\mathbf{Z}}_c\right)_{1 \leq c \leq C}$ from $\mathbf{Z}$
- perform cluster analysis from each pair to obtain a pair of partitions $\left(\Psi_c, \tilde{\Psi}_c\right)$
- classify individuals of $\mathbf{Z}$ from Ψ_c and $\tilde{\Psi}_c$ providing a pair of partitions $\left(\Psi'_c, \tilde{\Psi}'_c\right)$
- the instability V is assessed by averaging the proportions of disagreements

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

Instability after MI (Audigier and Niang, 2020)

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

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

- $T \in [0; 2]$
- T decreases when n increases
- T is robust to $M \geq 2$ (SI ignores the between instability)
- $T \approx V$ when the proportion of missing values tends to 0

Simulation design: data generation

Complete data generation

- gaussian mixture model with $K = 2$ components

$$\begin{array}{rcl} \mu_1 & = & (0,0,0,0,0,0,0,0,0,0) \\ \mu_2 & = & (0,0,0,0,0,2,2,2,2,2) \end{array} \qquad \Sigma = \begin{pmatrix} I_5 & & 0 \\ & \begin{matrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{matrix} & \\ 0 & & I_5 \end{pmatrix}$$

- $n \in \{25, 50, 100, 200\}$, $\rho \in \{0.3, 0.6\}$

Missing data generation

- MCAR: $Prob\,(missing) = \tau$
- MAR: $Prob\,(missing) = \Phi(a_\tau + z_{i1})$
- $\tau \in \{10\%, 30\%, 50\%\}$

Simulation design: evaluation

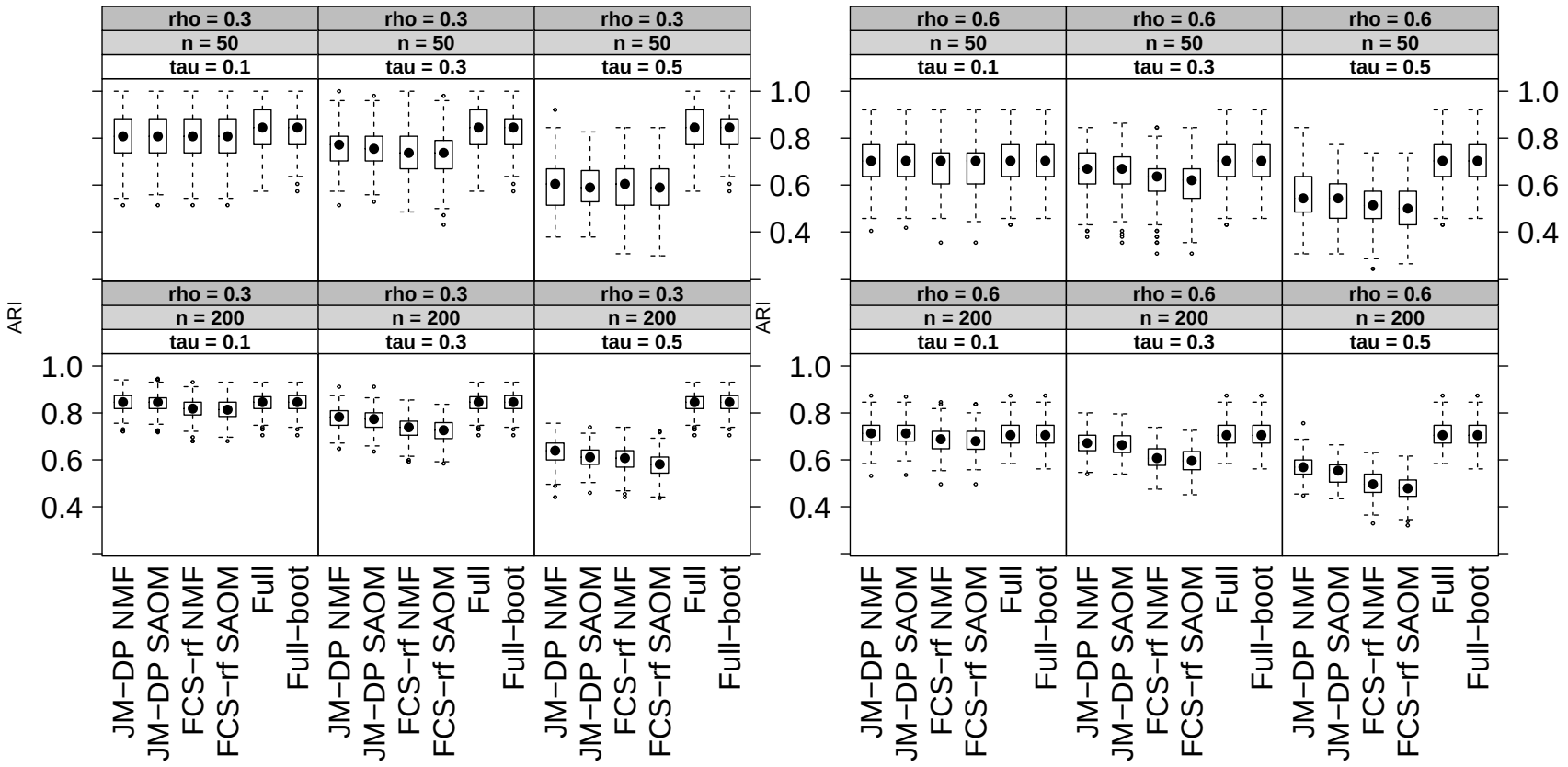
For each incomplete data set (200 per configuration)

- imputation methods
  - FCS-RF: imputation by random forest (Doove et al., 2014)
  - JM-DP: imputation based on a gaussian mixture model (Kim et al., 2014)
  - $M = 50$
- cluster analysis by kmeans
- pooling by SAOM/NMF, instability computed as proposed

Criteria

- accuracy: assessed using ARI and compared with kmeans on Full data and using bootstrap (Dudoit and Fridly, 2003)
- instability: comparison with CCA or Full data

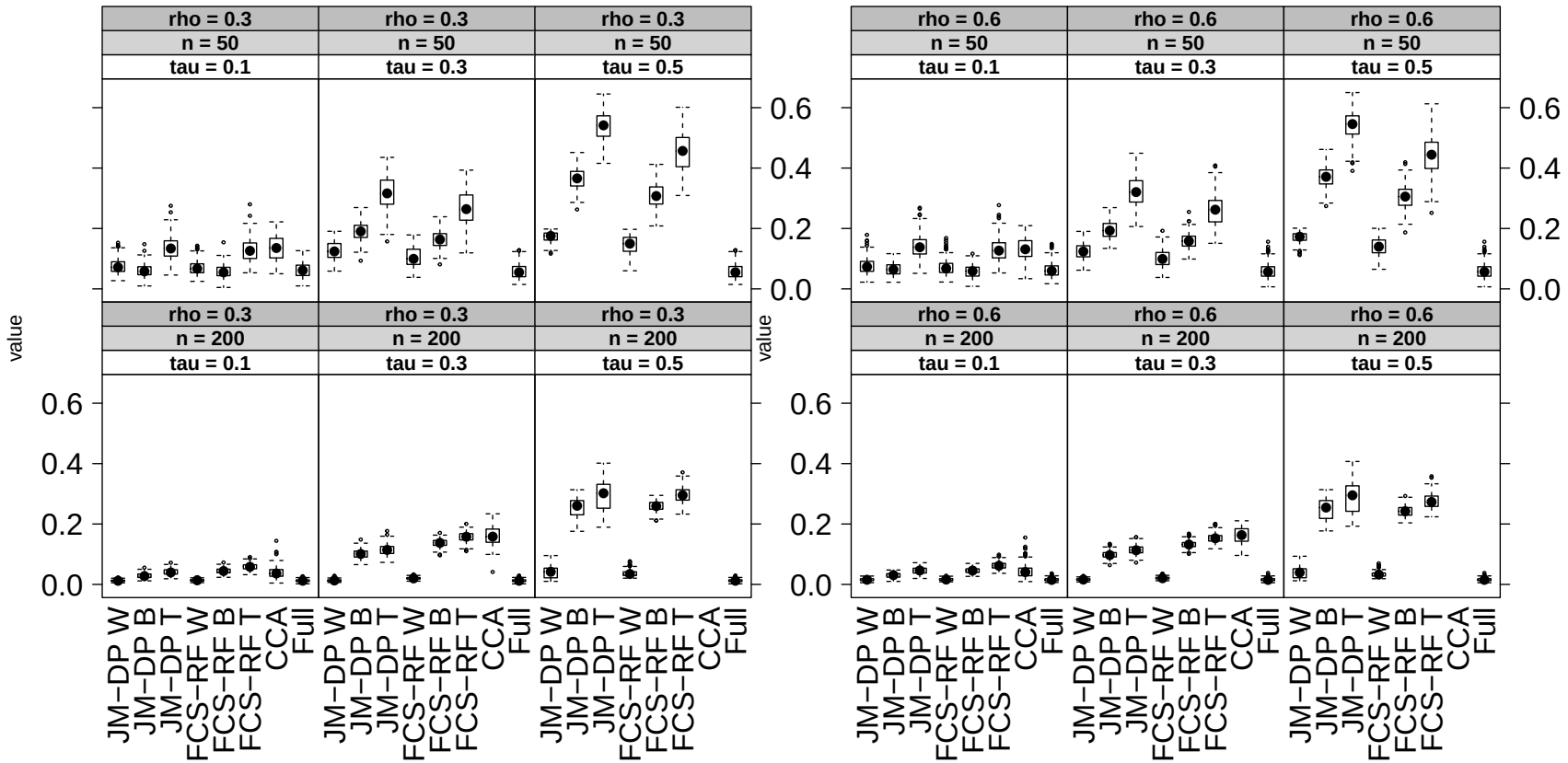
Results: pooling



(a) ARI distribution for $\rho = 0.3$ (b) ARI distribution for $\rho = 0.6$

Figure: Partition accuracy under the MCAR mechanism

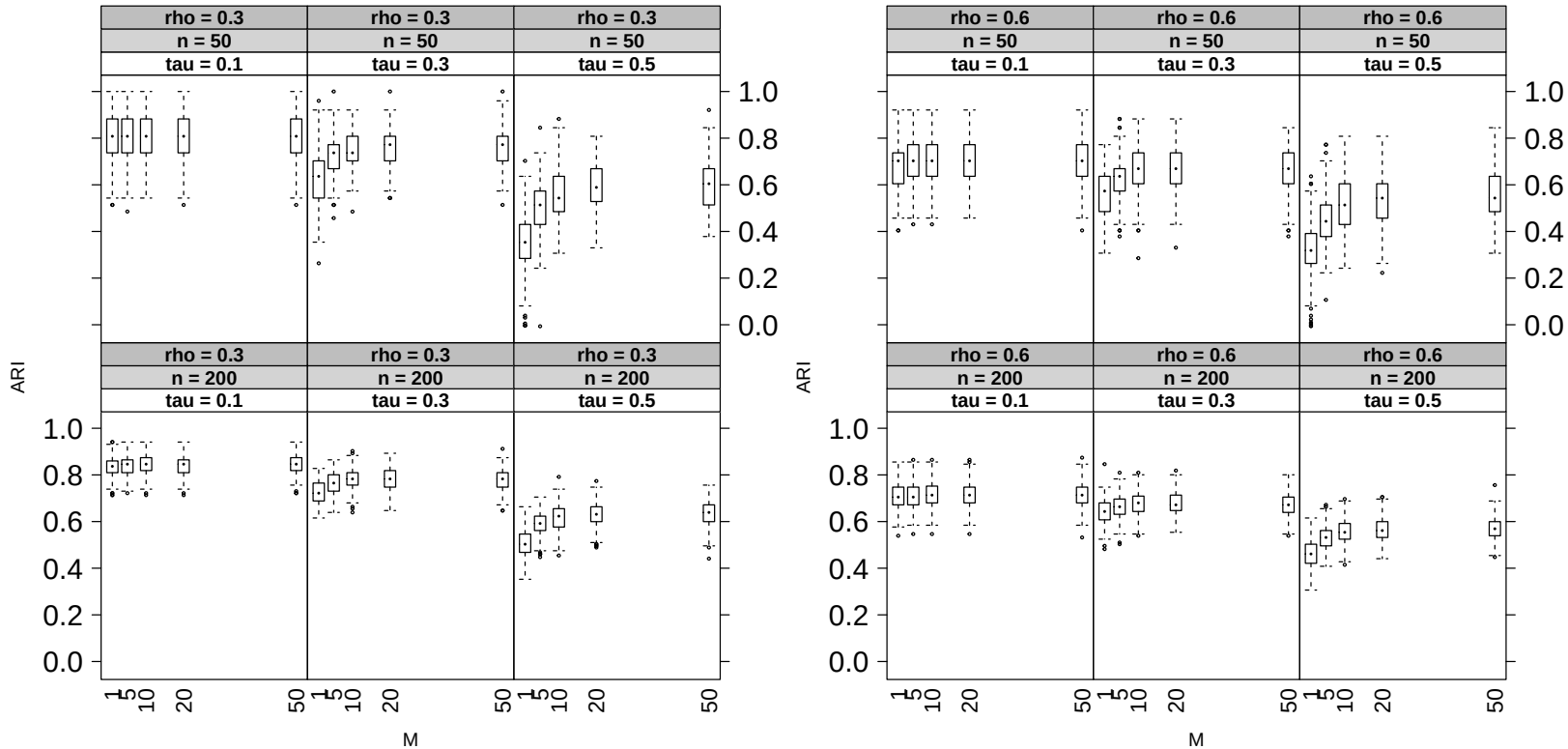
Results: instability



(a) $\rho = 0.3$ (b) $\rho = 0.6$

Figure: Partition instability under the MCAR mechanism

Results: influence of M



(a) ARI distribution for $\rho = 0.3$ (b) ARI distribution for $\rho = 0.6$

Figure: Accuracy of the clustering procedure according to M for JM-DP with NMF under a MCAR mechanism

Conclusion

Clustering after MI is challenging

- consensus by NMF based methods is theoretically appealing and outperforms Mirkin based methods
- the instability accounting for missing values can be assessed using bootstrap

MI is a valuable technique to tackle missing values in clustering

- can be used for any cluster analysis method
- it allows a way to choose the number of clusters with missing values

Available in the R package clusterMI

References I

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

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.

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.

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