Advances in Data Analysis and Classification (2023) 17:623–657
https://doi.org/10.1007/s11634-022-00519-1

REGULAR ARTICLE

Clustering with missing data: which equivalent for Rubin's rules?

Vincent Audigier[1] · **Ndèye Niang**[1]

Received: 29 July 2021 / Revised: 30 May 2022 / Accepted: 16 August 2022 /
Published online: 1 September 2022

Abstract
Multiple imputation (MI) is a popular method for dealing with missing values. However, the suitable way for applying clustering after MI remains unclear: how to pool partitions? How to assess the clustering instability when data are incomplete? By answering both questions, this paper proposed a complete view of clustering with missing data using MI. The problem of partitions pooling is here addressed using consensus clustering while, based on the bootstrap theory, we explain how to assess the instability related to observed and missing data. The new rules for pooling partitions and instability assessment are theoretically argued and extensively studied by simulation. Partitions pooling improves accuracy, while measuring instability with missing data enlarges the data analysis possibilities: it allows assessment of the dependence of the clustering to the imputation model, as well as a convenient way for choosing the number of clusters when data are incomplete, as illustrated on a real data set.

Keywords Clustering · Consensus clustering · Missing data · Multiple imputation · Rubin's rules · Uncertainty

Mathematics Subject Classification 62H30 · 62D10

1 Introduction

Clustering individuals is an essential task for data science. Clustering aims at partitioning a sample of individuals in several groups (clusters) so that individuals in a same cluster are similar from a multidimensional point of view, while individuals in

✉ Vincent Audigier
vincent.audigier@cnam.fr

Ndèye Niang
n-deye.niang_keita@cnam.fr

[1] CNAM, Laboratoire Cedric-MSDMA, 2 Rue Conte, 75003 Paris, France

separate clusters are different. $k-$means clustering (Forgy 1965), partitioning around medoids (Kaufman and Rousseeuw 1990), clustering by mixture models (McLachlan and Basford 1988) or hierarchical clustering (Ward 1963) are some popular methods for building this partition.

However, data are often incomplete and clustering algorithms cannot be directly applied on incomplete data. A common strategy to deal with missing values in data analysis consists in using multiple imputation (Rubin 1976, 1987; Schafer 1997). Multiple imputation (MI) consists of 3 steps: (1) the imputation of the data set according to an imputation model several times (2) the analysis of each imputed data set according to a substantive model (3) the pooling of the analysis results according to the Rubin's rules. Such methods are mainly used for inference in linear models (see Marshall et al. (2009) for other uses), but not for clustering. For instance, for applying a regression model on incomplete quantitative data: (1) data can be imputed according to a Gaussian model (Schafer 1997), (2) the regression model is fit on each imputed data set, leading to several estimates of the regression coefficients and their associated variance, (3) regression estimates are averaged and the associated variance is computed. Thus, despite missing values, MI yields a unique estimate of substantive model parameters and an uncertainty measure, which is expressed as a variance estimate. One major issue for applying clustering after MI is how to apply an equivalent of Rubin's rules in this context, *i.e.* how to apply step 3) ? Basagana et al. (2013), Faucheux et al. (2020), Bruckers et al. (2017) brought some answers in terms of partitions pooling, but the question of the uncertainty measure has not been discussed.

Uncertainty in clustering refers to *(in)stability* and results in various Voronoi tessellations of the metric space. These variations can cover many aspects: the used clustering algorithm, its initialization, the chosen number of clusters, etc. Here, we focus on another aspect which is the stability related to sampling. Note that we distinguish it to the probability in the assignment of an individual to a cluster (easily obtained for model-based clustering methods), which assesses the uncertainty in cluster assignment, but would remain even if the sample were fixed. See Dudoit and Fridly (2003) or Bruckers et al. (2017) for related works.

When data are complete, several resampling techniques have been proposed to assess the clustering stability (Hennig 2007). One main advantage of these methods consists in being relevant for both distance-based and model-based clustering methods. They are generally motivated by the determination of the number of clusters. The rationale is that a "too" large number of clusters should lead to a significant increase of instability. Jain and Moreau (1987), Wang (2010), Fang and Wang (2012), Mourer et al. (2020) proposed several approaches in this line. In particular, Wang (2010) proposed a measure of stability based on cross-validation. Authors demonstrate the asymptotic selection consistency of their procedure. However, the expected value of this measure is related to the number of individuals. Consequently, since by data splitting cross-validation reduces the sample size, the stability estimate could be biased. For this reason, Fang and Wang (2012) proposed an insightful bootstrap technique avoiding data splitting.

In this paper, we generally more focus on the pooling step, both in terms of partitions pooling and in terms of sample variability with incomplete data. The rest of the paper is as follows. Based on the literature on consensus clustering and stability assessment in

the context of complete data, we argue in Sect. 2 how to apply Rubin's rules after MI. Then in Sect. 3, our methodology is assessed by simulation. Finally, an application to a real data set is proposed to determine the number of clusters when data are incomplete.

2 Method

2.1 Notations

Following standard notations for incomplete data analysis, we denote $\mathbf{X} = (x_{i\ell})_{1\leq i\leq n, 1\leq \ell\leq p}$ the full data set and $\mathbf{R} = (r_{i\ell})_{1\leq i\leq n, 1\leq \ell\leq p}$ the missing data pattern, so that $r_{i\ell} = 0$ if $x_{i\ell}$ is missing and 1 if observed. X and R are the associated random variables. For a given observation i, the set of observed values is denoted x_i^{obs}, while the set of missing values is denoted x_i^{miss}, so that $x_i = \left(x_i^{obs}, x_i^{miss}\right)$. Note that this partition of variables is specific for each observation. Similarly, we denote X^{obs} and X^{miss} the observed and the missing part of X so that $X = \left(X^{obs}, X^{miss}\right)$. The distribution of random variable is denoted F. Next, we place ourselves under the standard missing at random (MAR) assumption, meaning R and X^{miss} are independent conditionally to X^{obs} Rubin (1976).

2.2 Rubin's rules

From a frequentist point of view, MI aims at estimating the expected mean and the expected variance of a statistic Q over the realizations of $\left(X^{obs}, R\right)$. For instance, Q could be the least squared estimator of regression coefficients in a linear model, or the estimator of correlation between variables, etc. Under the MAR assumption, consistent estimators can be obtained by ignoring the distribution of R. The associated estimates are obtained in three steps:

1. M values of X^{miss} are drawn from their predictive distribution $F_{X^{miss}|X^{obs}}$, leading to M imputed data sets $\left(\mathbf{X}^{obs}, \mathbf{X}_m^{miss}\right)_{1\leq m\leq M}$.
2. Q is evaluated on each one, providing a set of point estimates $\left(\hat{Q}_m\right)_{1\leq m\leq M}$ with $\hat{Q}_m = Q\left(\mathbf{X}^{obs}, \mathbf{X}_m^{miss}\right)$, as well as a set of estimated variances $(U_m)_{1\leq m\leq M}$.
3. These values are aggregated according to the Rubin's rules (Rubin 1976) leading to a unique point estimate $\bar{Q}$ and a unique variance estimate T as follows:

$$\bar{Q} = \frac{1}{M}\sum_{m=1}^{M}\widehat{Q}_m \tag{1}$$

$$T = \underbrace{\frac{1}{M}\sum_{m=1}^{M}U_m}_{\bar{U}} + \underbrace{\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{Q}_m - \bar{Q}\right)^2}_{B} \tag{2}$$

By the second step, we obtain M independent realizations of Q given X^{obs}. These realizations are centered around $Q\left(\mathbf{X}^{obs}, \mathbf{X}^{miss}\right)$ (in expectation), *i.e.* around the value of the statistic which would be observed if data were fully observed. Consequently, their average given by $\bar{Q}$ in the third step is an unbiased estimate of the expected value of Q over all observed samples. Thus, $\bar{Q}$ estimates

$$\mathbb{E}_{X^{obs}}\left[\mathbb{E}_X\left[Q\left(X^{obs}, X^{miss}\right)|X^{obs}\right]\right] \tag{3}$$

Similarly, $\bar{U} = \frac{1}{M}\sum_{m=1}^{M} U_m$ estimates the variance of Q over observed samples and $B = \frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{Q}_m - \bar{Q}\right)^2$ the additional variance related to missing values. Following an ANOVA decomposition, the total variance T is expressed as the sum of a within imputation variance ($\bar{U}$) and a between imputation variance (B), so that T estimates

$$\underbrace{\mathbb{E}_{X^{obs}}\left[Var_X\left[Q\left(X^{obs}, X^{miss}\right)|X^{obs}\right]\right]}_{within} + \underbrace{Var_{X^{obs}}\left[\mathbb{E}_X\left[Q\left(X^{obs}, X^{miss}\right)|X^{obs}\right]\right]}_{between} \tag{4}$$

Note that according to the values of $\left(\mathbf{X}_m^{miss}\right)_{1\leq m\leq M}$, $\bar{Q}$ randomly varies around its expectation given X^{obs}. When M is small, this additional variability cannot be ignored. For this reason, B is generally corrected to account for it. Such a correction is obtained by multiplying B by $(1 + 1/M)$ (Schafer 1997).

Compared to single imputation, MI accounts for the variability due to missing values thanks to the between variance B. The ratio B/T is helpful for interpretation since it assesses the robustness of the final analysis results to the imputation model (van Buuren 2018).

Challenges and motivations Rubin's rules (Eqs. 1 and 2) have been developed for statistic Q in $\mathbb{R}$ (Marshall et al. 2009). However, a clustering algorithm does not lie within this scope, since it can be expressed as a categorical variable Ψ with values in the set of partitions of n observations in K clusters at the most. Thus, this paper aims to develop new rules in the context of cluster analysis: a first rule to pool the M realizations of Ψ obtained from each imputed data set, as well as a second rule to compute a unique associated uncertainty measure which accounts for sample variability and missing values. Such an innovative methodology would offer a new way for applying any cluster analysis method on incomplete data.

2.3 Partitions pooling

In the context of clustering, partitions pooling refers to consensus clustering. Based on the literature in this research field when data are complete, we are going to propose an equivalent of the first rule (Eq. 1).

2.3.1 Consensus with complete data

Consensus clustering has been addressed by several authors since several decades (see e.g. Day (1986), Vega-Pons and Ruiz-Shulcloper (2011) for a survey). The main idea of consensus clustering is to agglomerate the separate partitions $\left(\Psi_j\right)_{1\leq j\leq J}$ (called *contributory partitions*) into a global partition that must be as similar as possible to the contributory partitions according to an index, like the Rand index defined as the proportion of agreements between partitions, *i.e.* $\frac{1}{n(n-1)}\sum_{(i,i')}\delta_{ii'}$ where $\delta_{ii'}$ is equal to 0 if individuals i and i' belong to the same cluster in one partition and not in the other; and $\delta_{ii'}$ is equal to 1 otherwise. These contributory partitions can be due to various algorithms, or several tuning parameters, several sets of features etc. Recently, Jain (2017) brought a strong theoretical framework to consensus clustering by seeing a consensus algorithm as an estimate of the partition minimizing the expected sum of the dissimilarity δ with all separate partitions (over all partitions). More precisely, the expected partition is defined as follows

$$argmin_{\Psi\in\mathcal{P}_{n,K}}\int_{\mathcal{P}_{n,K}}\delta^{\alpha}(\Psi^{\star},\Psi)d\pi(\Psi^{\star}) \tag{5}$$

where π is a probability distribution on the partition space $\mathcal{P}_{n,K}$ of n observations in K clusters at the most, δ a dissimilarity function and α a positive real. An estimator of (5) can be defined as the partition minimizing the loss function from the observed contributory partitions $\left(\Psi_j\right)_{1\leq j\leq J}$:

$$L(\Psi)=\sum_{j=1}^{J}\delta^{\alpha}(\Psi,\Psi_j) \tag{6}$$

By this theoretical framework, Jain (2017) extends the notion of mean (dedicated to real statistic) to the context of partitions.

However, because of the huge number of partitions, minimizing (6) is highly challenging. The literature mainly deals with α tuned to 1 or 2, referred as *median partition problem*. Vega-Pons and Ruiz-Shulcloper (2011) distinguished four families of methods for solving it: non-negative matrix factorization (NMF) based methods, Mirkin distance-based methods, Kernel methods and genetics algorithms. The first two have interesting theoretical properties.

NMF methods consists in rewriting the median partition problem as

$$argmin_{\mathbf{H}}\parallel\mathbf{M}-\mathbf{H}\parallel^2 \tag{7}$$

where $\parallel\ .\ \parallel$ denotes the Frobenius norm, $\mathbf{H}$ $(n\times n)$ denotes a connectivity matrix (meaning $\mathbf{H}=(h_{ii'})_{1\leq i,i'\leq n}$ with $h_{ii'}=1$ if the individuals i and i' are in a same cluster and $h_{ii'}=0$ otherwise) and $\mathbf{M}=\frac{1}{J}\sum_{j=1}^{J}\mathbf{H}_j$ denotes the mean of the connectivity matrices $\left(H_j\right)_{1\leq j\leq J}$ associated to each contributory partition Ψ_j $(1\leq j\leq J)$.

The constraint that $\mathbf{H}$ must be a connectivity matrix can be expressed as an optimization over the set of the orthogonal matrices (Li et al. 2007). The solution of this problem under orthogonality constraint corresponds to the partition minimizing the loss function given in Eq. (6). NMF is a powerful method widely used for solving many optimization problems (beyond the clustering framework) in several fields. One of the main theoretical strengths of the method is the monotone convergence of the optimization algorithm used.

Mirkin distance-based methods focus on minimizing the loss function (6) for δ chosen as the number of disagreements between partitions (called *Mirkin distance*). The Mirkin distance does not make the problem less complex, but it has been widely studied and benefits from theoretical results. For example, when the solution is restricted to the set of contributory partitions, the error cannot exceed two times the one obtained by the global optimum (Filkov and Skiena 2004).

Note that many other methods have been proposed to perform consensus clustering, like the popular Cluster based Similarity Partitioning Algorithm (CSPA), which consists in re-clustering the individuals from the average ($\mathbf{M}$) of the connectivity matrices associated to the contributory partitions. However, those methods cannot be expressed as a median partition problem and consequently cannot be justified from an inferential point of view. See Vega-Pons and Ruiz-Shulcloper (2011), Strehl et al. (2002) for a review.

2.3.2 Partitions pooling after MI

Partitions pooling after MI aims at aggregating several partitions varying by the imputed values only. As the first Rubin's rule is motivated by inferential argument, consensus clustering based on the median partition problem is theoretically appealing since it directly extends the notion of expected mean to the clustering framework, providing a straightforward application of the first Rubin's rule to clustering. Among the consensus methods based on the median partition problem, genetics algorithms do not offer theoretical guaranties, while kernel-based methods seem irrelevant in the context of MI. Indeed, since imputed values are generated independently from their predictive distribution, we assume all contributory partitions should have the same weight. Thus, only NMF-based methods and Mirkin distance-based methods provide a suitable way to pool partitions after MI. Formally, our equivalent of the first Rubin's rule for clustering after MI is as follows:

$$\bar{\Psi} = argmin_{\Psi} \sum_{m=1}^{M} \delta(\Psi, \Psi_m) \tag{8}$$

with Ψ_m is the partition obtained from $\left(\mathbf{X}^{obs}, \mathbf{X}_m^{miss}\right)$ and δ the Mirkin distance, or equivalently

$$argmin_{\mathbf{H}} \parallel \mathbf{M} - \mathbf{H} \parallel^2 \tag{9}$$

with $\mathbf{M} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{H}_m$ and $\mathbf{H}_m$ the connectivity matrix associated to Ψ_m.

Following Jain (2017), the obtained partition estimates the theoretical partition minimizing the expected sum of the dissimilarity δ with all separate partitions given X^{obs}.

Based on other techniques, clustering after MI has been previously investigated. Some authors have proposed stacking centroids or stacking imputed data sets for dealing with several imputed data sets (Plaehn 2019). We can note that stacking ignores the differences between imputed data sets. Among other methods, Basagana et al. (2013) proposed a general framework for consensus by previously investigating the choice of the number of partitions and the choice of the subset of variables retained for clustering. For a given number of clusters and a set of variables, consensus is performed by majority vote over the contributory partitions. More recently, Bruckers et al. (2017) proposed a consensus for functional data. For achieving this goal, they first identify the indicator matrices associated to each contributory partition and then, they look at the fuzzy matrix minimizing the Euclidean distance over the indicator matrices. The consensus partition is then obtained by majority vote (Dimitriadou et al. 2002). Like CSPA, this approach in two steps cannot be considered as based on the median partition problem (Vega-Pons and Ruiz-Shulcloper 2011). Lately, Faucheux et al. (2020) proposed consensus based on the MultiCons algorithm (Al-Najdi et al. 2016). The algorithm presents many advantages, in particular it allows a visualization of the hidden cluster structure in the data set, but it does not aim at minimizing the median partition problem (Al-Najdi et al. 2016, p. 16).

While these other methods yield a unique partition from several imputed data sets, this partition cannot be expressed as an estimator of a theoretical partition, contrary to those based on the median partition problem.

2.4 Instability pooling

Based on the literature in cluster stability when data are complete, we now propose an equivalent of the second rule (Eq. 2).

2.4.1 Instability with complete data

Assessing the instability in clustering is important for data analysis. For achieving this goal, resampling methods are appealing, especially when the clustering algorithm is distance-based, like k-means, k-medoids or hierarchical clustering. Wang (2010) and Fang and Wang (2012) proposed two ways for computing instability measure from any clustering algorithms. The first one is based on cross-validation, while the second is based on bootstrap. Since the cross-validation method tends to underestimate the instability (Wang 2010; Fang and Wang 2012), the bootstrap method appears like more relevant. The main idea consists in defining a theoretical distance (instability) δ between clusterings based on the sample distribution F_X. Then, this distribution is mimicked by bootstrap and the distance is evaluated from each bootstrap replicate. Finally, distances are aggregated by averaging. More precisely, the theoretical distance

δ_{F_X} between two clusterings Ψ and Ψ' is defined by

$$\delta_{F_X}(\Psi_j, \Psi_{j'}) = \mathbb{P}_{F_X}\{I\left(V_{\Psi_j}(X) = V_{\Psi_j}\left(X'\right)\right) + I\left(V_{\Psi_{j'}}(X) = V_{\Psi_{j'}}\left(X'\right)\right) = 1\} \tag{10}$$

where X and X' are independently drawn from the distribution F_X and $V_{\Psi_j}(X)$ is the Voronoi cell for X according to the partition given by Ψ_j. This distance measures the probability of disagreement between both clusterings.

Based on this definition, the instability of Ψ is defined as

$$\mathbb{E}_{X^n \sim F_X^n, \tilde{X}^n \sim F_X^n}\left[\delta_{F_X}\left(\Psi\left(X^n\right), \Psi\left(\tilde{X}^n\right)\right)\right] \tag{11}$$

i.e. as the expectation over all random samples of size n of the distances between partitions given by clustering trained on them.

Fang and Wang (2012) proposes an estimate of (11) by bootstrap: C bootstrap pairs $\left(\mathbf{X}_c, \tilde{\mathbf{X}}_c\right)_{1 \leq c \leq C}$ are drawn from the empirical distribution $\widehat{F}^n$. From each one, Ψ is evaluated. Both estimates are used to classify the individuals of $\mathbf{X}$ according to the Voronoi cells defined by each partition (e.g. by considering the closest centroid). Finally, the instability of the clustering is estimated by

$$U^{boot} = \frac{1}{C}\sum_{c=1}^{C} \delta_{\hat{F}^n}\left(\Psi\left(\mathbf{X}_c\right), \Psi\left(\tilde{\mathbf{X}}_c\right)\right) \tag{12}$$

with

$$\delta_{\hat{F}^n}\left(\Psi\left(\mathbf{X}_c\right), \Psi\left(\tilde{\mathbf{X}}_c\right)\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i'=1}^{n}\left|I\left(V_{\Psi(\mathbf{X}_c)}(x_i) = V_{\Psi(\mathbf{X}_c)}(x_{i'})\right) - I\left(V_{\Psi\left(\tilde{\mathbf{X}}_c\right)}(x_i) = V_{\Psi\left(\tilde{\mathbf{X}}_c\right)}(x_{i'})\right)\right| \tag{13}$$

Note that $\delta_{\hat{F}^n}$ corresponds to the proportion of disagreements between both partitions of $\mathbf{X}$ (up to the scalar factor $\frac{n}{n-1}$) and can be viewed as a normalized Mirkin distance. Fang and Wang (2012) indicate moderate values of C (20 or 50) are sufficient for a precise instability assessment. Furthermore, this instability can be assessed for distance-based or non-distance-based clustering algorithms.

2.4.2 Instability with incomplete data

Following previous developments for complete data, the instability with missing data can be defined as the expectation given in (11) over observed data. Following the second Rubin's rule, such an instability can be decomposed as the sum of a within

instability (Eq. 14) and a between instability (Eq. 15):

$$\mathbb{E}_{X^{obs},\tilde{X}^{obs}}\left[\mathbb{E}_{X,\tilde{X}}\left[\delta_{F_{X|X^{obs}}}\left(\Psi\left(X^{obs},X^{miss}\right),\Psi\left(\tilde{X}^{obs},\tilde{X}^{miss}\right)\right)|X^{obs}\right]\right] \quad (14)$$

$$\mathbb{E}_{X^{obs},\tilde{X}^{obs}}\left[\delta_{F_X}\left(\Psi\left(X^{obs},X^{miss}\right),\Psi\left(\tilde{X}^{obs},\tilde{X}^{miss}\right)\right)|X^{obs}\right] \quad (15)$$

Following Eq. (2), the within instability (Eq. 14) can be estimated from imputed M data sets $\left(\mathbf{X}^{obs},\mathbf{X}_m^{miss}\right)_{1\leq m\leq M}$ by:

$$\bar{U}=\frac{1}{M}\sum_{m=1}^{M}U_m^{boot} \quad (16)$$

where U_m^{boot} is the instability estimated from $\left(\mathbf{X}^{obs},\mathbf{X}_m^{miss}\right)$ according to Eq. (12) and the between instability (Eq. 15) can be estimated by:

$$B=\frac{1}{M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M}\delta_{\hat{F}_{X|X^{obs}}}\left(\Psi\left(\mathbf{X}^{obs},\mathbf{X}_m^{miss}\right),\Psi\left(\mathbf{X}^{obs},\mathbf{X}_{m'}^{miss}\right)\right) \quad (17)$$

where $\delta_{\hat{F}_{X|X^{obs}}}\left(\Psi\left(\mathbf{X}^{obs},\mathbf{X}_m^{miss}\right),\Psi\left(\mathbf{X}^{obs},\mathbf{X}_{m'}^{miss}\right)\right)$ corresponds to the proportion of disagreements between partitions obtained from imputed data sets m and m' as in Eq. (13). We note this expression does not depend on the mean partition, while the rule in Eq. (2) depends on the mean. For this reason, no correction for small values of M is required here.

The total instability is given by the sum of Eqs. (16) and (17):

$$T=\frac{1}{M}\sum_{m=1}^{M}U_m^{boot}+\frac{1}{M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M}\delta_{\hat{F}_{X|X^{obs}}}\left(\Psi\left(\mathbf{X}^{obs},\mathbf{X}_m^{miss}\right),\Psi\left(\mathbf{X}^{obs},\mathbf{X}_{m'}^{miss}\right)\right) \quad (18)$$

Note that without missing values, Eq. (18) is equivalent to the instability proposed in Fang and Wang (2012). More generally, T is positive and bounded by 2. The value of 0 is reached if $K=1$ or if the clustering is constant whatever the incomplete sample. The less stable the clustering is, the larger T will be. A large value of the between instability compared to the total one indicates a strong dependence of the clustering to the imputation model.

2.5 Summary

Based on above developments, the full procedure to perform cluster analysis after multiple imputation can be summarized as follows:

Imputation From an incomplete data set, generate M imputed data sets according to a predefined multiple imputation method
Analysis For m in $\{1\ldots M\}$:

1. build a partition Ψ_m from the m^{th} imputed data set
2. compute U_m^{boot} the associated instability:
   (a) by resampling individuals with replacement, generate C bootstrap pairs $\left(\mathbf{X}_c, \tilde{\mathbf{X}}_c\right)_{1\leq c\leq C}$ from the m^{th} imputed data set
   (b) for each bootstrap pair c in $\{1 \dots C\}$
      – perform cluster analysis from $\left(\mathbf{X}_c, \tilde{\mathbf{X}}_c\right)_{1\leq c\leq C}$ to obtain a pair of partitions $\left(\Psi_c, \tilde{\Psi}_c\right)$
      – classify the n individuals of the m^{th} imputed data set according to Ψ_c and $\tilde{\Psi}_c$ to obtain a pair of partitions $\left(\Psi_c', \tilde{\Psi}_c'\right)$
      – compute the proportion of disagreements between Ψ_c' and $\tilde{\Psi}_c'$ as $U_m^c = \frac{1}{n^2}\sum_{(i,i')}\delta_{i,i'}$ where $\delta_{ii'}$ is equal to 0 if individuals i and i' belong to the same cluster in one partition and not in the other; and $\delta_{ii'}$ is equal to 1 otherwise
   (c) compute the instability associated to Ψ_m by averaging $U_m^{boot} = \frac{1}{C}\sum_{c=1}^{C} U_m^c$

Pooling The set of partitions $(\Psi_m)_{1\leq m\leq M}$ and the set of associated instability estimates $\left(U_m^{boot}\right)_{1\leq m\leq M}$ are aggregated as follows:

First rule (partitions pooling) using NMF or Mirkin based methods, compute the consensus partition as

$$\bar{\Psi} = argmin_{\Psi} \sum_{m=1}^{M} \delta(\Psi, \Psi_m)$$

where $\delta(\Psi, \Psi_m)$ denotes the number of disagreements between partitions Ψ and Ψ_m (Mirkin distance)

Second rule (instability pooling) compute the total instability as:

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

Note that an implementation is provided through the R package clusterMI which is available at the web page of the first author.

3 Simulations

After proposing rules for pooling clusterings after MI, we highlight how the pooled results vary according to the data structure. Furthermore, we investigate their robustness to the number of imputed data sets M. The R code used for simulations is available on demand.

3.1 Simulation design

3.1.1 Data generation

Full data are simulated according to a p multivariate Gaussian mixture model with two mixture components:

$$X \sim \pi_1 \mathcal{N}_p\left(\mu_1, \Sigma\left(\rho\right)\right) + \pi_2 \mathcal{N}_p\left(\mu_2, \Sigma\left(\rho\right)\right)$$

where $p = 10$, $\pi_1 = \pi_2 = 1/2$, $\mu_1 = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$, $\mu_2 = (0, 0, 0, 0, 0, 2, 2, 2, 2, 2)$ and $\Sigma\left(\rho\right) = \begin{pmatrix} I_5 & 0 \\ 0 & \begin{matrix} 1 & \rho & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho & \rho \\ \rho & \rho & 1 & \rho & \rho \\ \rho & \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{matrix} \end{pmatrix}$. Several configurations are investigated by varying the number of individuals $n \in \{25, 50, 100, 200\}$, the correlation between variables $\rho \in \{0.3, 0.6\}$. For each configuration, $S = 200$ data sets are generated. For each data set, several missing data patterns are considered varying by the percentage of missing values $\tau \in \{0.1, 0.3, 0.5\}$ and their distribution: $Prob(r_{i\ell} = 0) = \tau$ for all $1 \leq i \leq n$ and $1 \leq \ell \leq p$ (MCAR mechanism) or $Prob(r_{i\ell} = 0) = \Phi(a_\tau + x_{i1})$ for all $1 \leq i \leq n$ and $2 \leq \ell \leq p$ with Φ the cumulative distribution function of the standard normal distribution (MAR mechanism) and a_τ a constant to control the percentage of missing values in expectation. Thus, 9600 incomplete data sets are investigated. Note the computational cost to investigate each one does not allow a larger number of replications.

3.1.2 Methods

Each incomplete data set is imputed according to two MI methods accounting for the structure of individuals (Schafer 2003)

- JM-DP: MI using a non-parametric extension of the mixture model namely the Dirichlet process mixture of products of multivariate normal distributions (Kim et al. 2014). The number of components is bounded by 5. The number of iterations for the burn-in period is tuned to 500 and the number of skipped iterations to keep one imputed data set after the burn-in period is tuned to 100.
- FCS-RF: MI by random forest (Doove et al. 2014). The number of iterations for the multivariate imputation by chained equations algorithm is tuned to 10.

For each method, the number of imputed data sets M varies in $\{1, 5, 10, 20, 50\}$. Note that the case $M = 1$ corresponds to single imputation. Then, k-means clustering is performed on each imputed data set (using 2 clusters, standardization of variables and 100 initializations) and partitions are pooled according to the proposed rules. More precisely, the mean partition is estimated using the NMF clustering based method (Eq. (8)) as proposed in Li et al. (2007) and also using a Mirkin distance-based method

(Eq. (9)) called Simulated-Annealing-One-element-Move (SAOM) (Filkov and Skiena 2004). Furthermore, the total instability is computed according to Eq. (18).

As benchmark, k-means clustering is also performed on the full data and on the complete cases. In addition, k-means through a bagging procedure based on bootstrap (Dudoit and Fridly 2003) is investigated on full data, while k-means through the k-pod algorithm (Chi et al. 2016) is investigated on incomplete data. This later algorithm overcomes missing values in k-means by optimizing the k-means criterion over observed values only. For achieving this goal, a majorization-minimization algorithm is used, consisting in alternating clustering of individuals (by k-means) and imputation of incomplete observations by the coordinates of their associated centroid.

3.1.3 Criteria

The accuracy of the consensus partition is assessed according to the mean (over the S generated data sets) of the adjusted rand index (ARI) (Hubert and Arabie 1985) between the consensus partition and the reference one known by simulation. The mean (over the S generated data sets) of the intra instability $\bar{U}$, the mean of the between variability B, and the mean of the total instability T are reported for each configuration.

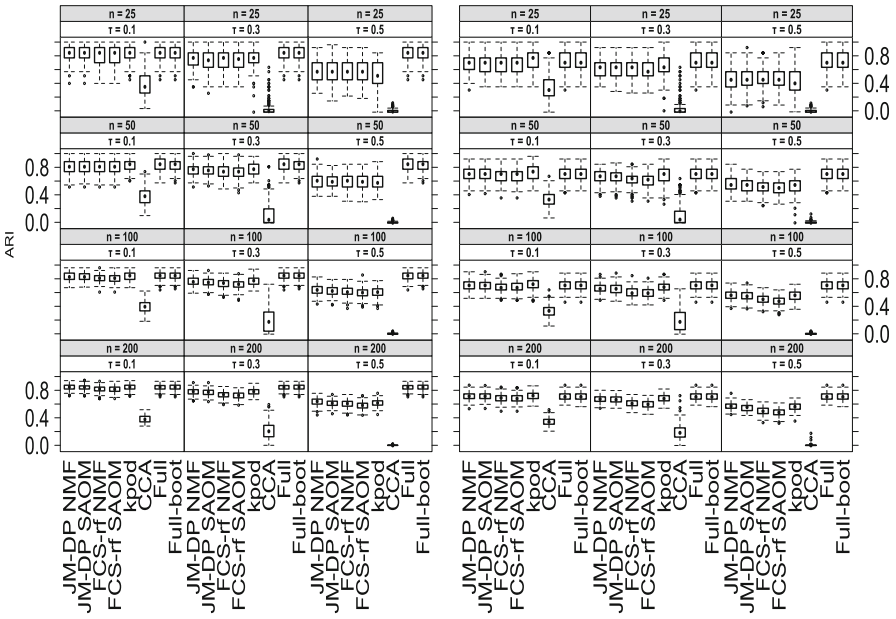
3.2 Results

3.2.1 Partitions pooling

Figures 1 and 2 summarize results over the $S = 200$ simulations for both mechanisms (averages and interquartile ranges are available in "Tables 3 and 4 in Appendix"). Performances for MI methods consider $M = 50$ imputed data sets, the influence of M is discussed in Sect. 3.2.3. Note that because complete-case analysis does not cluster all individuals (compared to MI methods or clustering on full data), the following process is applied for a fair comparison: first, complete cases are clustered using kmeans clustering. Then, based on their observed profile, each incomplete case is classified according to the closest centroid. If the data set did not contain complete cases, then a cluster would be assigned at random for each incomplete case. Thus, the resulting partition concerns all observations.

For both mechanisms, the ARI over the S data sets when the contributory partitions are pooled using NMF or when they are pooled using SAOM remain generally close. However, both MI methods show higher ARI values with NMF pooling when the number of individuals and the proportion of missing values increase.

As expected, the ARI obtained by MI is close to the one obtained without missing values (Full or Full-boot) when the proportion of missing values τ equals 0.1, and the difference is larger when this proportion increases. Furthermore, clustering after MI outperforms complete-cases analysis even if the proportion of missing values is small. Compared to the direct application of kmeans using the k-pod algorithm, similar ARI are observed for a MCAR mechanism, but MI outperforms under the MAR mechanism for a moderate ($\tau = 0.3$) or large ($\tau = 0.5$) proportion of missing values.



(a) ARI distributions when $\rho = 0.3$ **(b)** ARI distributions when $\rho = 0.6$

Fig. 1 Accuracy of the clustering procedure under a MCAR mechanism: distribution of the adjusted rand index over the $S = 200$ generated data sets varying by the number of individuals (n), the correlation between variables (ρ) and the proportion of missing values (τ) for various imputation methods (JM-DP or FCS-RF), various consensus methods (NMF or SAOM). For each case, clustering is performed using k-means clustering. As benchmark, ARI obtained by applying k-means on complete-cases (CCA), using k-pod algorithm (kpod), on full data (Full) or using a bagging procedure (Full-boot) are also reported

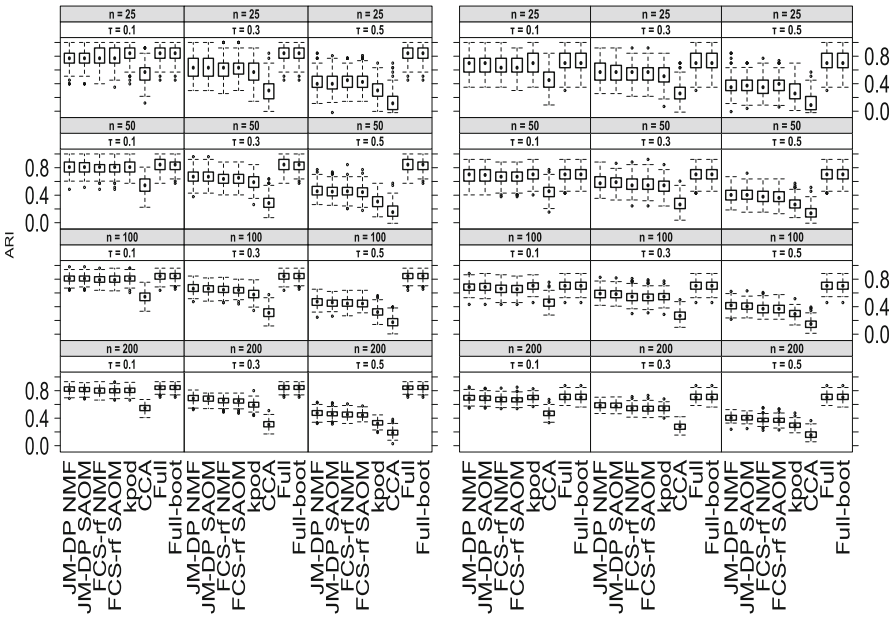
A large number of individuals slightly increases the ARI and decreases the interquartile range in MI, while it only decreases the interquartile range when data are full.

Finally, the ARI is usually higher when data are imputed using JM-DP than FCS-RF. It could be expected since JM-DP is based on an imputation model close to the model used for data generation, while FCS-RF is based on a non-parametric model.

3.2.2 Instability pooling

Table 1 gathers the average of within instability, between instability and total instability over the $S = 200$ data sets for configurations under a MCAR mechanism with $n = 50$ or $n = 200$ individuals (see "Table 6 in Appendix" for a MAR mechanism and Tables 5 for $n \in \{25, 100\}$).

As expected, the total instability is always higher by using MI ($M = 50$) instead of SI ($M = 1$) for all configurations. Furthermore, clustering instability is generally smaller when using MI instead of complete-case analysis and higher compared to clustering instability when data are full. We note the average between instability (B) tends to increase when the proportion of missing values increases. This behavior was



(a) ARI distributions when $\rho = 0.3$ **(b)** ARI distributions when $\rho = 0.6$

Fig. 2 Accuracy of the clustering procedure under a MAR mechanism: distribution of the adjusted rand index over the $S = 200$ generated data sets varying by the number of individuals (n), the correlation between variables (ρ) and the proportion of missing values (τ) for various imputation methods (JM-DP or FCS-RF), various consensus methods (NMF or SAOM). For each case, clustering is performed using k-means clustering. As benchmark, ARI obtained by applying k-means on complete-cases (CCA), using k-pod algorithm (kpod), on full data (Full) or using a bagging procedure (Full-boot) are also reported

expected. More surprisingly, the within instability ($\bar{U}$) is also increasing with the proportion of missing values, even if this increase remains relatively smaller. This is highlighted for small values of n. Such a behavior can be explained by overfitting of the imputation models. Indeed, FCS-RF as non-parametric imputation method requires a large number of observations, while JM-DP as complex model requires a large number of observations to fit accurately the data structure. For this reason, when the number of individuals is small, the imputed values are highly variable, yielding to an increase of the within instability. This behavior is more severe when the proportion of missing values is large. Note that instability using the k-pod algorithm is not considered since the method only returns a partition from incomplete data, but no instability measure.

3.2.3 Influence of *M*

As underlined in Sect. 2.4.2, the instability given by the second rule does not depend (in expectation) on the number of imputed data sets (if $M \geq 2$). As regard the partition accuracy, a large number of imputed data sets should bring the consensus partition closer to the partition obtained from full data. Figure 3 reports the influence of M on the ARI for MI using JM-DP and NMF consensus.

Table 1 Instability of the clustering procedure under a MCAR mechanism: average within-instability ($\bar{U}$) average between-instability (B) and average total instability (T) over the $S = 200$ generated data sets for various number of individuals (n), correlation between variables (ρ) and proportion of missing values (τ)

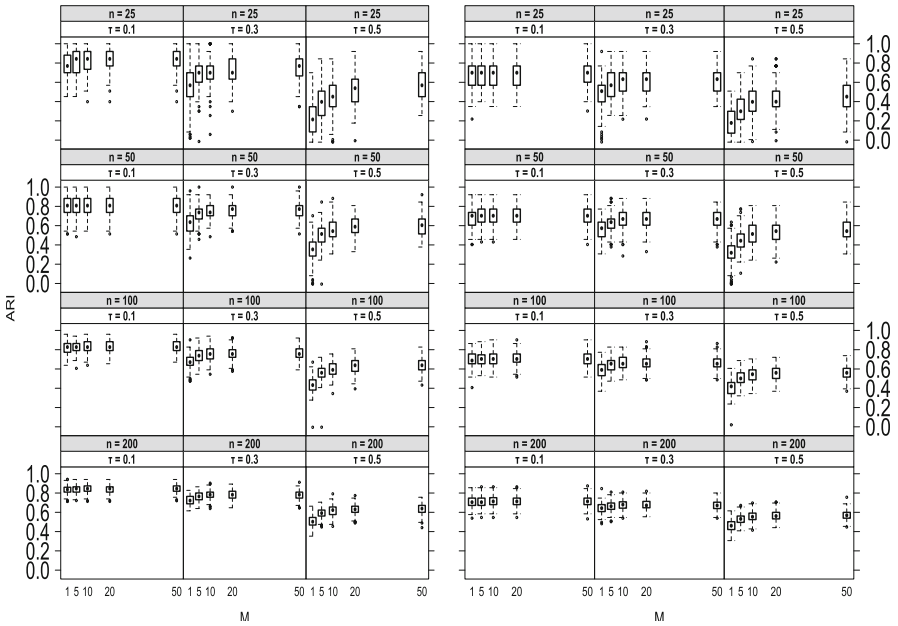
n	ρ	τ	M	JM-DP			FCS-RF			Full	CCA
				$\bar{U}$	B	T	$\bar{U}$	B	T	T	T
50	0.3	0.1	1	0.08	0.00	0.08	0.07	0.00	0.07	0.06	0.13
50	0.3	0.1	50	0.08	0.06	0.14	0.07	0.06	0.13	0.06	0.13
50	0.3	0.3	1	0.13	0.00	0.13	0.11	0.00	0.11	0.06	
50	0.3	0.3	50	0.13	0.19	0.32	0.10	0.16	0.27	0.06	
50	0.3	0.5	1	0.17	0.00	0.17	0.14	0.00	0.14	0.06	
50	0.3	0.5	50	0.17	0.37	0.54	0.15	0.31	0.45	0.06	
50	0.6	0.1	1	0.08	0.00	0.08	0.07	0.00	0.07	0.06	0.13
50	0.6	0.1	50	0.08	0.06	0.14	0.07	0.06	0.13	0.06	0.13
50	0.6	0.3	1	0.13	0.00	0.13	0.10	0.00	0.10	0.06	
50	0.6	0.3	50	0.12	0.20	0.32	0.10	0.16	0.26	0.06	
50	0.6	0.5	1	0.17	0.00	0.17	0.14	0.00	0.14	0.06	
50	0.6	0.5	50	0.17	0.37	0.54	0.14	0.30	0.44	0.06	
200	0.3	0.1	1	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.04
200	0.3	0.1	50	0.01	0.03	0.04	0.01	0.04	0.06	0.01	0.04
200	0.3	0.3	1	0.01	0.00	0.01	0.02	0.00	0.02	0.01	0.16
200	0.3	0.3	50	0.01	0.10	0.11	0.02	0.14	0.16	0.01	0.16
200	0.3	0.5	1	0.04	0.00	0.04	0.04	0.00	0.04	0.01	
200	0.3	0.5	50	0.04	0.25	0.29	0.04	0.26	0.30	0.01	
200	0.6	0.1	1	0.02	0.00	0.02	0.02	0.00	0.02	0.02	0.04
200	0.6	0.1	50	0.02	0.03	0.05	0.02	0.05	0.06	0.02	0.04
200	0.6	0.3	1	0.02	0.00	0.02	0.02	0.00	0.02	0.02	0.16
200	0.6	0.3	50	0.02	0.10	0.11	0.02	0.13	0.15	0.02	0.16
200	0.6	0.5	1	0.04	0.00	0.04	0.03	0.00	0.03	0.02	
200	0.6	0.5	50	0.04	0.25	0.29	0.03	0.24	0.28	0.02	

Two imputation methods are investigated (JM-DP or FCS-RF) using $M = 1$ or $M = 50$ imputed data sets. For each case, clustering is performed by k-means. As benchmark, ARI obtained by applying k-means clustering on full data (Full) or complete-case analysis (CCA) are also reported (not possible for a large proportion of missing values)

For all configurations, a large value of M tends to increase the ARI meaning a large number of imputed data sets tends to increase clustering accuracy. The increase is even more important as the proportion of missing values is large or as the number of individuals is small.

Similar results have been observed when data are imputed by FCS-RF ("Fig. 8 in Appendix") or when a MAR mechanism is considered (see "Fig. 9 in Appendix" for imputation by FCS-RF and Fig. 7 for imputation by JM-DP).

Figure 4 reports the influence of M on the instability for MI using JM-DP and NMF consensus.



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

Fig. 3 Accuracy of the clustering procedure according to M: adjusted rand index over the $S = 200$ generated data sets varying by the number of individuals (n), the correlation between variables (ρ) and the proportion of missing values (τ) generated under a MCAR mechanism. Data sets are imputed by JM-DP varying by the number of imputed data sets (M). For each data set, clustering is performed using k-means clustering and consensus clustering is performed using NMF

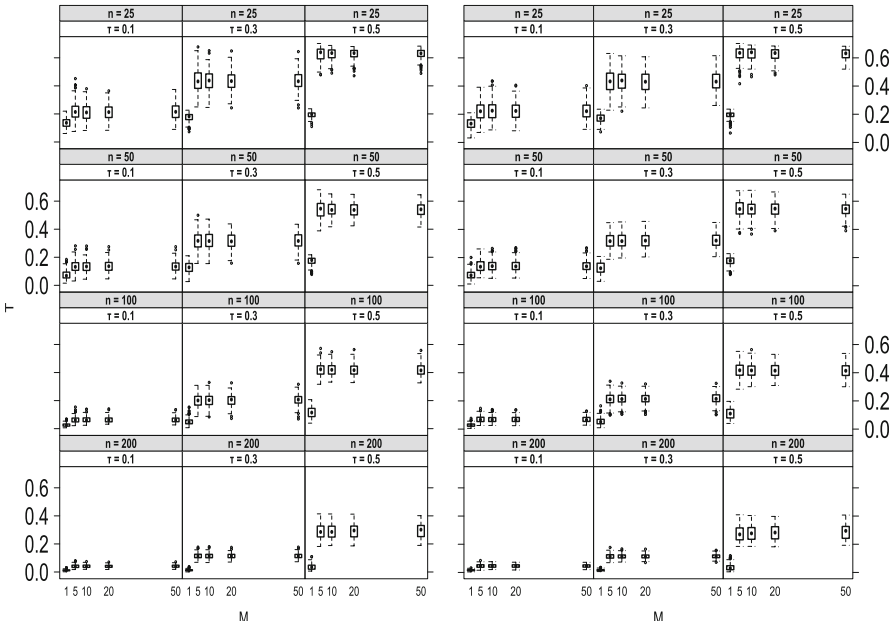
As expected, for $M \geq 2$ the instability is constant whatever the proportion of missing values, the number of individuals or the correlation between variables. Similar results are observed when data are generated under a MAR mechanism or when data are imputed by FCS-RF (see "Figs. 10, 11, 12 in Appendix").

3.3 Complement: number of clusters

As written in the introduction, having an instability measure with missing values can provide a way for estimating the number of clusters from incomplete data. To assess the second rule with regards to this goal, a complementary simulation study is conducted by considering a grid for the number of clusters. More precisely, after multiple imputation, k-means clustering is applied for K in $\{2, \ldots, 5\}$. By applying the second rule, a value of instability T_K is obtained from each value from the grid. The estimated number of clusters is given by

$$argmin_{K \in \{1,\ldots,K_{max}\}} T_K. \tag{19}$$

Results are reported in Table 2. The number of clusters is accurately estimated when performing imputation by JM-DP, but a small number of observations or a large pro-



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

Fig. 4 Instability according to M: total instability (T) over the $S = 200$ generated data sets varying by the number of individuals (n), the correlation between variables (ρ) and the proportion of missing values (τ) generated under a MCAR mechanism. Data sets are imputed by JM-DP varying by the number of imputed data sets (M). For each data set, clustering is performed using k-means clustering and consensus clustering is performed using NMF

portion of missing values leads to an upper bias. This behaviour is similar to complete case analysis (even if the method cannot always be applied). Results when using FCS-RF leads to $K = 4$ or $K = 5$ in most of the cases. Better performances of JM-DP compared to FCS-RF were expected since JM-DP is well tailored to account for the clustered structure of observations (Audigier et al. 2021).

4 Application

In addition to estimating the instability in clustering, the second rule can be used for tuning the number of clusters with missing values. The *animals* data set from the *cluster* R package (Maechler et al. 2019) is used as an example (Kaufman and Rousseeuw 1990). It describes 20 animals by six binary variables (warm-blooded, can fly, vertebrate, endangered, live in groups, have hair). Five individuals are incomplete (see "Table 7 in Appendix").

We propose to perform hierarchical clustering using the flexible UPGMA method (Belbin et al. 1992). Flexible UPGMA can be seen as a generalization of the average method which ensures the desirable monotonicity property of the algorithm. For achieving this goal, data are imputed $M = 50$ times according to a log-linear model

Table 2 Estimated numbers of clusters based on the second rule: frequencies of estimated values within a grid between 2 and 5 clusters over the $S = 200$ generated data sets for various number of individuals (n), correlation between variables (ρ), missing data mechanisms (MCAR and MAR) and proportion of missing values (τ)

n	ρ	mech	τ	JM-DP				FCS-RF				CCA				Full			
				2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
50	0.3	MCAR	0.1	39	161	0	0	0	4	46	150	52	4	13	131	192	0	0	8
50	0.3	MCAR	0.3	4	196	0	0	0	0	2	198	0	0	0	0	195	1	0	4
50	0.3	MCAR	0.5	0	178	0	22	0	0	5	195	0	0	0	0	195	1	0	4
50	0.3	MAR	0.1	29	171	0	0	0	2	56	142	156	1	2	41	192	1	0	7
50	0.3	MAR	0.3	0	200	0	0	0	0	50	150	37	12	10	139	194	0	0	6
50	0.3	MAR	0.5	0	166	1	33	0	0	61	139	1	0	0	3	195	1	0	4
50	0.6	MCAR	0.1	28	172	0	0	0	1	50	149	57	3	10	130	187	0	0	13
50	0.6	MCAR	0.3	1	199	0	0	0	0	2	198	0	0	0	0	192	0	0	8
50	0.6	MCAR	0.5	0	182	0	18	0	0	1	199	0	0	0	0	192	0	0	8
50	0.6	MAR	0.1	14	186	0	0	2	1	63	134	149	2	1	48	190	0	0	10
50	0.6	MAR	0.3	1	199	0	0	0	0	29	171	39	2	14	143	190	0	0	10
50	0.6	MAR	0.5	0	179	0	21	0	0	54	146	0	0	0	8	192	0	0	8
200	0.3	MCAR	0.1	200	0	0	0	2	1	148	49	198	0	0	2	200	0	0	0
200	0.3	MCAR	0.3	197	3	0	0	0	0	193	7	1	1	3	24	200	0	0	0
200	0.3	MCAR	0.5	94	106	0	0	0	0	101	99	0	0	0	0	200	0	0	0
200	0.3	MAR	0.1	199	1	0	0	2	2	133	63	200	0	0	0	200	0	0	0
200	0.3	MAR	0.3	195	5	0	0	0	0	173	27	193	0	1	6	200	0	0	0
200	0.3	MAR	0.5	120	80	0	0	0	0	172	28	96	1	9	94	200	0	0	0
200	0.6	MCAR	0.1	200	0	0	0	2	2	147	49	196	0	0	4	200	0	0	0

Table 2 continued

n	ρ	mech	τ	JM-DP				FCS-RF				CCA				Full			
				2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
200	0.6	MCAR	0.3	200	0	0	0	0	0	183	17	1	1	0	15	200	0	0	0
200	0.6	MCAR	0.5	90	110	0	0	0	0	86	114	0	0	0	0	200	0	0	0
200	0.6	MAR	0.1	199	1	0	0	3	1	125	71	200	0	0	0	200	0	0	0
200	0.6	MAR	0.3	200	0	0	0	0	0	151	49	186	0	0	14	200	0	0	0
200	0.6	MAR	0.5	139	61	0	0	0	0	151	49	87	1	4	108	200	0	0	0

The expected number of clusters is $K = 2$. Two imputation methods are investigated (JM-DP or FCS-RF) using $M = 50$ imputed data sets. For each case, clustering is performed by k-means. As benchmark, estimated numbers obtained by applying k-means clustering on full data (Full) or complete-case analysis (CCA) are also reported (not possible for a large proportion of missing values)

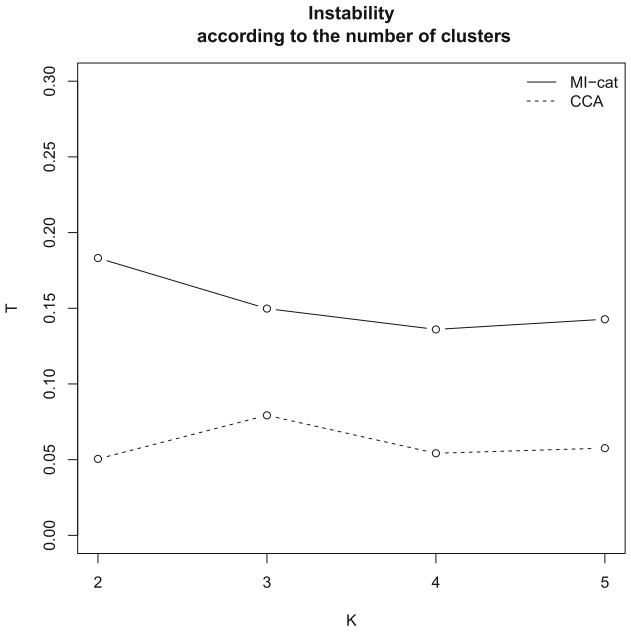


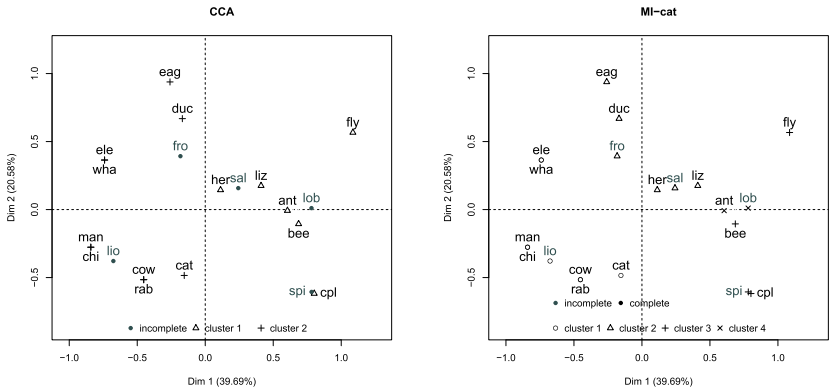
Fig. 5 Instability estimation in hierarchical clustering according to the number of clusters. Data are imputed using a log-linear model (MI-cat) with $M = 50$ imputed data sets. The instability with complete-case analysis (CCA) is also reported

(Schafer 1997), which is considered as the gold standard for binary variables. Then, hierarchical clustering is applied on each imputed data set for a given number of clusters K. Finally, the clustering instability T is assessed using the second rule. This process is repeated for K varying in $\{2, 3, 4, 5\}$. Results are presented in Fig. 5 and instability using only complete cases is also reported.

Using MI, the instability is the smallest for $K = 4$ clusters, suggesting a consensus clustering in 4 clusters. Results for complete-case analysis are less clear, but a partition in two clusters could be suggested. A larger number could be inappropriate compared to the number of complete cases (15).

Comparison between the consensus partition obtained by NMF in four clusters and the one obtained by hierarchical clustering on complete cases in two clusters are presented through a principal factor map (Fig. 6).

The partition obtained by complete-case analysis gathers mammals with birds in the first cluster, while insects and fishes are gathers in the second. On the opposite, consensus clustering after MI in four clusters isolates mammals in the first cluster, or insects in the fourth. Furthermore, it allows suitable clustering of incomplete individuals: lion among mammals, spider among insects, salmon and frog with herring, lobster with ant (both have the same observed profile *cf* "Table 7 in Appendix").



(a) Hierarchical clustering in two clusters using complete-case analysis. Incomplete individuals are colored in gray.

(b) Hierarchical clustering in four clusters using consensus clustering (NMF) after MI under log-linear model ($M = 50$).

Fig. 6 Animals data set: visualization of partitions through the principal factor map obtained using the iterative MCA algorithm (Josse et al. 2012)

5 Discussion

Multiple imputation is a widely used method for dealing with missing values. However, applying Rubin's rules with clustering remained unclear: 1) how to pool the partitions obtained from each imputed data set? In this paper, we argue to use median partition-based methods for pooling partitions. In particular, NMF methods are theoretically and computationally attractive for achieving this goal. 2) How to assess the instability of the clustering with missing values? Based on Fang and Wang (2012), we propose a new rule for assessing the stability with missing values. An associated R package entitled clusterMI is available at the web page of the first author

From a practical point of view, the first rule provides accurate clustering with missing values. Indeed, even without missing data, Dudoit and Fridly (2003) have shown bagging procedures based on bootstrap improve clustering accuracy. By generating M times the imputed values and aggregating the partitions obtained from each imputed data set, a similar improvement is observed with multiple imputation. It has been highlighted that the accuracy is sensitive to M, particularly when the number of individuals is small or the proportion of missing values is large. Simulations show $M = 50$ is generally enough, but M can be tuned by investigating the evolution of the pooled partition according to the number of imputed data sets.

The second rule allows calculation of an additional between instability B related to missing values. This instability has the advantage to be robust to the choice of M. Availability of a between instability is precious in practice for several uses. Firstly, it provides a new way for dealing with the number of clusters when data are incomplete. This is particularly useful for distance-based clustering methods like k-means or k-medoids. Secondly, the ratio B/T provides a new way to highlight how the partition is robust to the missing values (van Buuren 2018).

In this work, we assumed data were already imputed. It could be also interesting to investigate more deeply the suitable imputation method according to the clustering

method applied on each imputed data set. The topic is commonly discussed under the term *congeniality* (Meng 1994; Schafer 2003). Simulation results subtend a sensitivity to the imputation method both in terms of accuracy and instability which is substantiated by recent research (Audigier et al. 2021). We also assumed data were missing at random, but beyond the difficulty to impute non-missing at random data, the proposed rules could be directly applied.

Furthermore, we assumed all contributory partitions have the same number of clusters K, but the methodology can be directly applied for various number of clusters, like in hierarchical clustering where the number is generally unknown in advance. Indeed, NMF consensus clustering is essentially based on the average of the connectivity matrices associated to each contributory partition. Such an average can be obtained whatever the number of clusters since the dimensions of each connectivity matrix depends only on the number of individuals which is constant for all partitions. The only requirement is that the clustering method allows classification for new individuals. Even if this classification is always possible, certain classification methods are connected to certain clustering methods. For instance, classification using the closest centroid is suitable for k-means or k-medoids, while quadratic discriminant analysis could be more reliable for Gaussian mixtures. For other clustering methods, anyone of these classification methods could be used, but the robustness to the stability measurement should require more research.

Finally, several NMF-based methods are available for partitions pooling. In this paper, we focus on the multiplicative rules method as proposed in Li et al. (2007) which is the most common method, but which is not necessarily the most efficient. Among alternatives, alternating least squares algorithms are notably recommended for large scale data (Andrzej Cichocki and ichi Amari 2009).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

Partitions pooling

See Tables 3 and 4.

Table 3 Accuracy of the clustering procedure under a MCAR mechanism: average adjusted rand index (mean) and interquartile range (IQ) over the $S = 200$ generated data sets varying by the number of individuals (n), the correlation between variables (ρ) and the proportion of missing values (τ) for various imputation methods (JM-DP or FCS-RF), various consensus methods (NMF or SAOM)

n	ρ	τ	NMF				SAOM				kpod		CCA		Full		Full-boot	
			JM-DP		FCS-RF		JM-DP		FCS-RF									
			Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ
25	0.3	0.1	0.82	0.15	0.80	0.22	0.82	0.15	0.80	0.22	0.83	0.15	0.37	0.25	0.83	0.15	0.83	0.15
25	0.3	0.3	0.75	0.16	0.73	0.21	0.73	0.21	0.74	0.21	0.75	0.14	0.03	0.04	0.83	0.15	0.83	0.15
25	0.3	0.5	0.56	0.25	0.58	0.25	0.57	0.25	0.57	0.25	0.52	0.30	0.00	0.02	0.83	0.15	0.83	0.15
25	0.6	0.1	0.70	0.15	0.69	0.20	0.70	0.20	0.68	0.20	0.73	0.21	0.32	0.23	0.71	0.21	0.71	0.21
25	0.6	0.3	0.64	0.19	0.61	0.19	0.63	0.19	0.60	0.19	0.64	0.20	0.04	0.05	0.71	0.21	0.71	0.21
25	0.6	0.5	0.46	0.22	0.47	0.17	0.46	0.20	0.47	0.20	0.43	0.27	-0.00	0.02	0.71	0.21	0.71	0.21
50	0.3	0.1	0.81	0.15	0.80	0.15	0.81	0.15	0.80	0.15	0.83	0.11	0.38	0.17	0.84	0.15	0.83	0.11
50	0.3	0.3	0.76	0.11	0.74	0.14	0.75	0.11	0.73	0.12	0.78	0.14	0.13	0.20	0.84	0.15	0.83	0.11
50	0.3	0.5	0.61	0.16	0.60	0.16	0.60	0.13	0.59	0.16	0.59	0.16	-0.00	0.01	0.84	0.15	0.83	0.11
50	0.6	0.1	0.70	0.14	0.68	0.13	0.70	0.14	0.68	0.13	0.72	0.17	0.34	0.14	0.71	0.14	0.71	0.14
50	0.6	0.3	0.66	0.13	0.62	0.10	0.65	0.12	0.61	0.13	0.68	0.17	0.11	0.16	0.71	0.14	0.71	0.14
50	0.6	0.5	0.55	0.15	0.51	0.12	0.54	0.15	0.50	0.14	0.53	0.15	-0.00	0.01	0.71	0.14	0.71	0.14
100	0.3	0.1	0.83	0.09	0.81	0.07	0.83	0.08	0.81	0.08	0.84	0.09	0.39	0.12	0.84	0.07	0.84	0.07
100	0.3	0.3	0.76	0.09	0.73	0.09	0.75	0.08	0.72	0.08	0.77	0.09	0.21	0.28	0.84	0.07	0.84	0.07
100	0.3	0.5	0.64	0.10	0.61	0.08	0.62	0.10	0.59	0.09	0.60	0.09	0.00	0.01	0.84	0.08	0.84	0.07
100	0.6	0.1	0.70	0.10	0.68	0.08	0.70	0.09	0.68	0.09	0.72	0.10	0.33	0.11	0.70	0.10	0.70	0.10
100	0.6	0.3	0.66	0.08	0.60	0.10	0.65	0.10	0.59	0.10	0.68	0.08	0.20	0.25	0.70	0.10	0.70	0.10

Table 3 continued

n	ρ	τ	NMF				SAOM				kpod		CCA		Full		Full-boot	
			JM-DP		FCS-RF		JM-DP		FCS-RF									
			Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ
100	0.6	0.5	0.57	0.09	0.49	0.09	0.55	0.09	0.48	0.08	0.55	0.10	0.00	0.01	0.70	0.10	0.70	0.10
200	0.3	0.1	0.84	0.06	0.82	0.05	0.84	0.05	0.81	0.06	0.85	0.06	0.39	0.08	0.84	0.05	0.84	0.06
200	0.3	0.3	0.78	0.06	0.74	0.06	0.77	0.06	0.72	0.07	0.78	0.06	0.21	0.17	0.84	0.05	0.84	0.06
200	0.3	0.5	0.63	0.07	0.60	0.07	0.61	0.06	0.58	0.07	0.61	0.06	-0.00	0.00	0.84	0.05	0.84	0.06
200	0.6	0.1	0.71	0.07	0.69	0.07	0.71	0.07	0.68	0.08	0.72	0.08	0.34	0.07	0.71	0.08	0.71	0.08
200	0.6	0.3	0.67	0.07	0.61	0.07	0.67	0.07	0.60	0.08	0.68	0.07	0.20	0.13	0.71	0.08	0.71	0.08
200	0.6	0.5	0.57	0.06	0.50	0.08	0.55	0.07	0.48	0.07	0.56	0.08	0.00	0.00	0.71	0.08	0.71	0.08

For each case, clustering is performed using k-means clustering. As benchmark, ARI obtained by applying k-means on complete-cases (CCA), using k-pod algorithm (kpod), on full data (Full) or using a bagging procedure (Full-boot) are also reported

Table 4 Accuracy of the clustering procedure under a MAR mechanism: average adjusted rand index (mean) and interquartile range (IQ) over the $S = 200$ generated data sets varying by the number of individuals (n), the correlation between variables (ρ) and the proportion of missing values (τ) for various imputation methods (JM-DP or FCS-RF), various consensus methods (NMF or SAOM)

n	ρ	τ	NMF				SAOM				kpod		CCA		Full		Full-boot	
			JM-DP		FCS-RF		JM-DP		FCS-RF									
			Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ
25	0.3	0.1	0.79	0.14	0.79	0.22	0.79	0.14	0.79	0.22	0.81	0.15	0.54	0.18	0.83	0.15	0.83	0.15
25	0.3	0.3	0.64	0.26	0.64	0.19	0.64	0.26	0.64	0.16	0.56	0.25	0.31	0.22	0.83	0.15	0.83	0.15
25	0.3	0.5	0.42	0.16	0.43	0.16	0.42	0.19	0.44	0.16	0.31	0.18	0.15	0.19	0.83	0.15	0.83	0.15
25	0.6	0.1	0.68	0.20	0.67	0.20	0.68	0.20	0.66	0.20	0.70	0.27	0.47	0.22	0.71	0.21	0.71	0.21
25	0.6	0.3	0.58	0.25	0.55	0.18	0.57	0.21	0.55	0.18	0.53	0.19	0.27	0.17	0.71	0.21	0.71	0.21
25	0.6	0.5	0.38	0.15	0.38	0.19	0.39	0.15	0.37	0.15	0.28	0.22	0.12	0.19	0.71	0.21	0.71	0.21
50	0.3	0.1	0.80	0.15	0.79	0.11	0.80	0.13	0.79	0.11	0.80	0.15	0.54	0.18	0.84	0.15	0.83	0.11
50	0.3	0.3	0.67	0.13	0.64	0.13	0.67	0.13	0.64	0.13	0.59	0.16	0.29	0.13	0.84	0.15	0.83	0.11
50	0.3	0.5	0.47	0.12	0.46	0.11	0.46	0.12	0.45	0.13	0.31	0.14	0.17	0.15	0.84	0.15	0.83	0.11
50	0.6	0.1	0.68	0.17	0.66	0.13	0.68	0.17	0.66	0.13	0.69	0.14	0.46	0.13	0.71	0.14	0.71	0.14
50	0.6	0.3	0.59	0.16	0.56	0.15	0.58	0.14	0.56	0.15	0.54	0.15	0.28	0.15	0.71	0.14	0.71	0.14
50	0.6	0.5	0.40	0.13	0.38	0.15	0.40	0.13	0.38	0.14	0.28	0.11	0.15	0.12	0.71	0.14	0.71	0.14
100	0.3	0.1	0.81	0.07	0.79	0.08	0.81	0.07	0.79	0.09	0.80	0.07	0.55	0.11	0.84	0.08	0.84	0.07
100	0.3	0.3	0.67	0.10	0.65	0.09	0.66	0.09	0.64	0.08	0.58	0.11	0.32	0.11	0.84	0.08	0.84	0.07
100	0.3	0.5	0.47	0.10	0.45	0.09	0.46	0.09	0.45	0.10	0.32	0.09	0.18	0.10	0.84	0.07	0.84	0.07
100	0.6	0.1	0.68	0.09	0.66	0.10	0.68	0.10	0.66	0.10	0.70	0.09	0.46	0.10	0.70	0.10	0.70	0.10
100	0.6	0.3	0.59	0.11	0.54	0.10	0.58	0.10	0.54	0.10	0.55	0.09	0.27	0.10	0.70	0.10	0.70	0.10

Table 4 continued

n	ρ	τ	NMF				SAOM				kpod		CCA		Full		Full-boot	
			JM-DP		FCS-RF		JM-DP		FCS-RF									
			Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ	Mean	IQ
100	0.6	0.5	0.42	0.09	0.37	0.11	0.41	0.09	0.37	0.11	0.30	0.09	0.15	0.09	0.70	0.10	0.70	0.10
200	0.3	0.1	0.82	0.06	0.80	0.07	0.82	0.06	0.80	0.06	0.81	0.06	0.55	0.07	0.84	0.05	0.84	0.06
200	0.3	0.3	0.69	0.07	0.65	0.06	0.68	0.07	0.65	0.06	0.59	0.07	0.31	0.07	0.84	0.05	0.84	0.06
200	0.3	0.5	0.47	0.07	0.45	0.07	0.46	0.06	0.45	0.06	0.33	0.06	0.19	0.07	0.84	0.05	0.84	0.06
200	0.6	0.1	0.69	0.07	0.67	0.07	0.69	0.06	0.67	0.06	0.70	0.06	0.47	0.06	0.71	0.08	0.71	0.08
200	0.6	0.3	0.59	0.07	0.54	0.07	0.58	0.07	0.54	0.07	0.55	0.07	0.28	0.07	0.71	0.08	0.71	0.08
200	0.6	0.5	0.41	0.07	0.38	0.05	0.40	0.06	0.37	0.06	0.30	0.06	0.16	0.08	0.71	0.08	0.71	0.08

For each case, clustering is performed using k-means clustering. As benchmark, ARI obtained by applying k-means on complete-cases (CCA), using k-pod algorithm (kpod), on full data (Full) or using a bagging procedure (Full-boot) are also reported

Instability pooling

See Tables 5 and 6.

Table 5 Instability of the clustering procedure under a MCAR mechanism: average within-instability ($\bar{U}$) average between-instability (B) and average total instability (T) over the $S = 200$ generated data sets for various number of individuals (n), correlation between variables (ρ) and proportion of missing values (τ)

n	ρ	τ	M	JM-DP			FCS-RF			Full	CCA
				$\bar{U}$	B	T	$\bar{U}$	B	T	T	T
25	0.3	0.1	1	0.14	0.00	0.14	0.12	0.00	0.12	0.11	0.16
25	0.3	0.1	50	0.13	0.08	0.22	0.12	0.07	0.19	0.11	0.16
25	0.3	0.3	1	0.18	0.00	0.18	0.15	0.00	0.15	0.11	
25	0.3	0.3	50	0.17	0.26	0.43	0.15	0.20	0.35	0.11	
25	0.3	0.5	1	0.19	0.00	0.19	0.18	0.00	0.18	0.11	
25	0.3	0.5	50	0.19	0.43	0.63	0.18	0.37	0.55	0.11	
25	0.6	0.1	1	0.13	0.00	0.13	0.12	0.00	0.12	0.11	0.16
25	0.6	0.1	50	0.13	0.09	0.22	0.12	0.07	0.20	0.11	0.16
25	0.6	0.3	1	0.17	0.00	0.17	0.15	0.00	0.15	0.11	
25	0.6	0.3	50	0.17	0.26	0.44	0.15	0.20	0.34	0.11	
25	0.6	0.5	1	0.19	0.00	0.19	0.17	0.00	0.17	0.11	
25	0.6	0.5	50	0.19	0.43	0.63	0.17	0.37	0.54	0.11	
100	0.3	0.1	1	0.03	0.00	0.03	0.03	0.00	0.03	0.02	0.09
100	0.3	0.1	50	0.03	0.04	0.06	0.03	0.05	0.08	0.02	0.09
100	0.3	0.3	1	0.05	0.00	0.05	0.05	0.00	0.05	0.02	
100	0.3	0.3	50	0.05	0.15	0.20	0.04	0.15	0.19	0.02	
100	0.3	0.5	1	0.12	0.00	0.12	0.08	0.00	0.08	0.02	
100	0.3	0.5	50	0.11	0.30	0.42	0.08	0.28	0.36	0.02	
100	0.6	0.1	1	0.03	0.00	0.03	0.03	0.00	0.03	0.03	0.10
100	0.6	0.1	50	0.03	0.04	0.07	0.03	0.05	0.08	0.03	0.10
100	0.6	0.3	1	0.05	0.00	0.05	0.04	0.00	0.04	0.03	
100	0.6	0.3	50	0.06	0.16	0.22	0.05	0.14	0.19	0.03	
100	0.6	0.5	1	0.11	0.00	0.11	0.08	0.00	0.08	0.03	
100	0.6	0.5	50	0.11	0.31	0.41	0.08	0.26	0.34	0.03	

Two imputation methods are investigated (JM-DP or FCS-RF) using $M = 1$ or $M = 50$ imputed data sets. For each case, clustering is performed by k-means. As benchmark, ARI obtained by applying k-means clustering on full data (Full) or complete-case analysis (CCA) are also reported (not possible with a large proportion of missing values)

Table 6 Instability of the clustering procedure under a MAR mechanism: average within-instability ($\bar{U}$) average between-instability (B) and average total instability (T) over the $S = 200$ generated data sets for various number of individuals (n), correlation between variables (ρ) and proportion of missing values (τ)

n	ρ	τ	M	JM-DP			FCS-RF			Full	CCA
				$\bar{U}$	B	T	$\bar{U}$	B	T	T	T
(a) $n \in \{25, 50\}$											
25	0.3	0.1	1	0.14	0.00	0.14	0.12	0.00	0.12	0.11	0.14
25	0.3	0.1	50	0.13	0.09	0.22	0.12	0.08	0.20	0.11	0.14
25	0.3	0.3	1	0.17	0.00	0.17	0.15	0.00	0.15	0.11	0.16
25	0.3	0.3	50	0.17	0.27	0.44	0.15	0.22	0.37	0.11	0.16
25	0.3	0.5	1	0.19	0.00	0.19	0.17	0.00	0.17	0.11	
25	0.3	0.5	50	0.19	0.41	0.60	0.17	0.36	0.53	0.11	
25	0.6	0.1	1	0.13	0.00	0.13	0.13	0.00	0.13	0.11	0.13
25	0.6	0.1	50	0.13	0.10	0.23	0.12	0.08	0.20	0.11	0.13
25	0.6	0.3	1	0.17	0.00	0.17	0.15	0.00	0.15	0.11	0.15
25	0.6	0.3	50	0.17	0.26	0.42	0.15	0.21	0.35	0.11	0.15
25	0.6	0.5	1	0.19	0.00	0.19	0.17	0.00	0.17	0.11	
25	0.6	0.5	50	0.19	0.41	0.60	0.17	0.36	0.52	0.11	
50	0.3	0.1	1	0.08	0.00	0.08	0.07	0.00	0.07	0.06	0.09
50	0.3	0.1	50	0.08	0.07	0.15	0.07	0.07	0.14	0.06	0.09
50	0.3	0.3	1	0.12	0.00	0.12	0.10	0.00	0.10	0.06	0.14
50	0.3	0.3	50	0.12	0.22	0.34	0.10	0.19	0.30	0.06	0.14
50	0.3	0.5	1	0.17	0.00	0.17	0.14	0.00	0.14	0.06	0.15
50	0.3	0.5	50	0.17	0.36	0.53	0.14	0.33	0.47	0.06	0.15
50	0.6	0.1	1	0.08	0.00	0.08	0.07	0.00	0.07	0.06	0.09
50	0.6	0.1	50	0.08	0.08	0.15	0.07	0.06	0.14	0.06	0.09
50	0.6	0.3	1	0.12	0.00	0.12	0.10	0.00	0.10	0.06	0.14
50	0.6	0.3	50	0.12	0.22	0.34	0.10	0.19	0.29	0.06	0.14
50	0.6	0.5	1	0.16	0.00	0.16	0.14	0.00	0.14	0.06	0.16
50	0.6	0.5	50	0.16	0.36	0.53	0.14	0.32	0.46	0.06	0.16
(b) $n \in \{100, 200\}$											
100	0.3	0.1	1	0.03	0.00	0.03	0.03	0.00	0.03	0.02	0.04
100	0.3	0.1	50	0.03	0.05	0.07	0.03	0.06	0.09	0.02	0.04
100	0.3	0.3	1	0.05	0.00	0.05	0.05	0.00	0.05	0.02	0.09
100	0.3	0.3	50	0.05	0.19	0.24	0.05	0.18	0.23	0.02	0.09
100	0.3	0.5	1	0.11	0.00	0.11	0.08	0.00	0.08	0.02	0.15
100	0.3	0.5	50	0.11	0.33	0.43	0.08	0.31	0.39	0.02	0.15
100	0.6	0.1	1	0.03	0.00	0.03	0.03	0.00	0.03	0.03	0.04
100	0.6	0.1	50	0.03	0.05	0.08	0.03	0.06	0.09	0.03	0.04
100	0.6	0.3	1	0.05	0.00	0.05	0.05	0.00	0.05	0.03	0.10
100	0.6	0.3	50	0.05	0.19	0.24	0.05	0.17	0.21	0.03	0.10
100	0.6	0.5	1	0.10	0.00	0.10	0.08	0.00	0.08	0.03	0.15
100	0.6	0.5	50	0.10	0.33	0.43	0.08	0.30	0.37	0.03	0.15

Table 6 continued

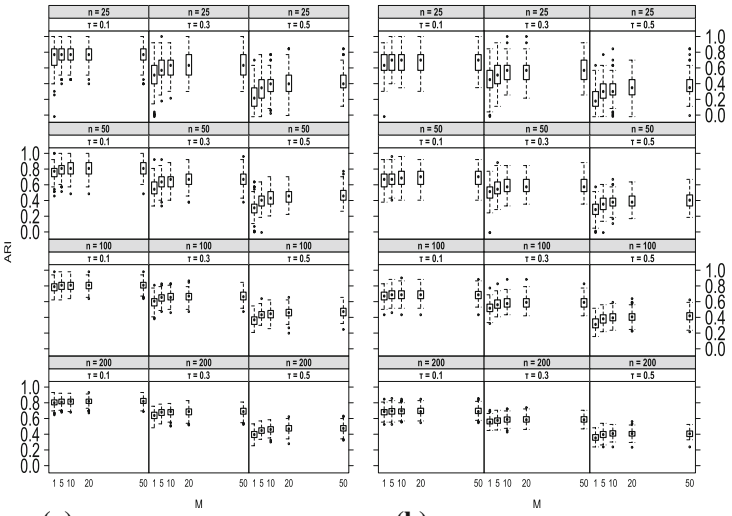
n	ρ	τ	M	JM-DP			FCS-RF			Full	CCA
				$\bar{U}$	B	T	$\bar{U}$	B	T	T	T
200	0.3	0.1	1	0.01	0.00	0.01	0.01	0.00	0.01	0.01	0.02
200	0.3	0.1	50	0.01	0.04	0.05	0.01	0.05	0.07	0.01	0.02
200	0.3	0.3	1	0.01	0.00	0.01	0.02	0.00	0.02	0.01	0.05
200	0.3	0.3	50	0.01	0.15	0.16	0.02	0.17	0.19	0.01	0.05
200	0.3	0.5	1	0.03	0.00	0.03	0.03	0.00	0.03	0.01	0.11
200	0.3	0.5	50	0.04	0.29	0.33	0.04	0.30	0.33	0.01	0.11
200	0.6	0.1	1	0.02	0.00	0.02	0.02	0.00	0.02	0.02	0.02
200	0.6	0.1	50	0.02	0.04	0.06	0.02	0.05	0.07	0.02	0.02
200	0.6	0.3	1	0.02	0.00	0.02	0.02	0.00	0.02	0.02	0.06
200	0.6	0.3	50	0.02	0.14	0.16	0.02	0.16	0.18	0.02	0.06
200	0.6	0.5	1	0.03	0.00	0.03	0.03	0.00	0.03	0.02	0.12
200	0.6	0.5	50	0.03	0.28	0.31	0.03	0.28	0.32	0.02	0.12

Two imputation methods are investigated (JM-DP or FCS-RF) using $M = 1$ or $M = 50$ imputed data sets. For each case, clustering is performed by k-means. As benchmark, ARI obtained by applying k-means clustering on full data (Full) or complete-case analysis (CCA) are also reported (not possible with a large proportion of missing values)

Influence of *M*

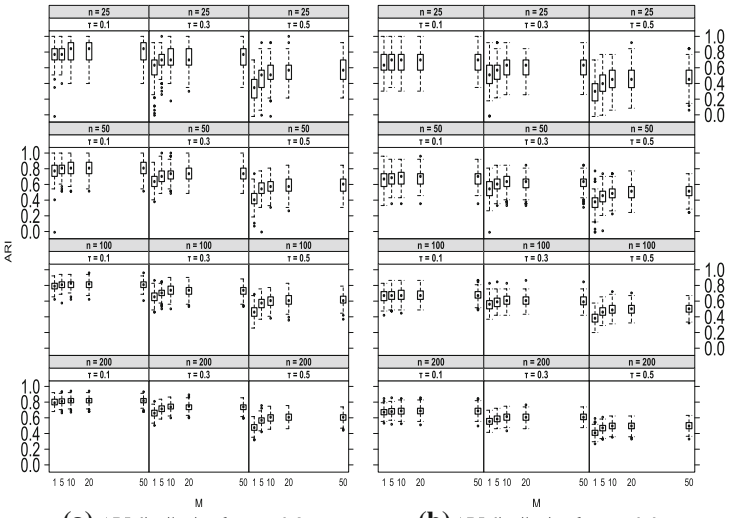
Accuracy

See Figs. 7, 8 and 9.



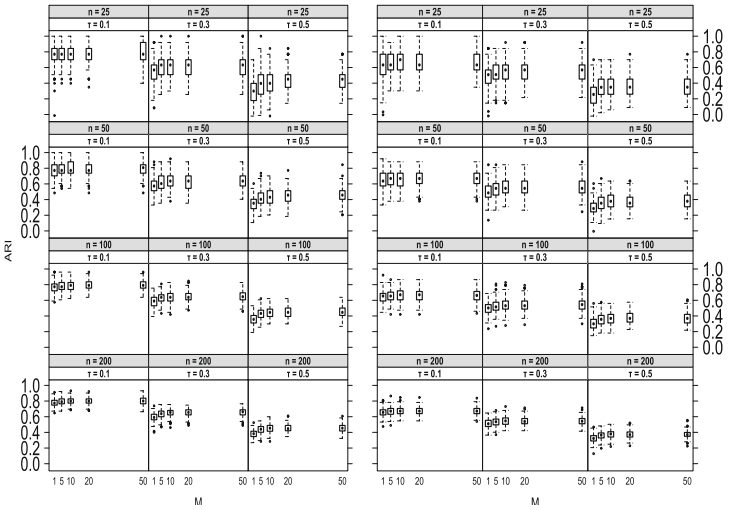
(a) ARI distribution for $\rho = 0.3$ **(b)** ARI distribution for $\rho = 0.6$

Fig. 7 Accuracy of the clustering procedure according to M: adjusted rand index over the $S = 200$ generated data sets varying by the number of individuals (n), the correlation between variables (ρ) and the proportion of missing values (τ) generated under a MAR mechanism. Data sets are imputed by JM-DP varying by the number of imputed data sets (M). For each data set, clustering is performed using k-means clustering and consensus clustering is performed using NMF



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

Fig. 8 Accuracy of the clustering procedure according to M: adjusted rand index over the $S = 200$ generated data sets varying by the number of individuals (n), the correlation between variables (ρ) and the proportion of missing values (τ) generated under a MCAR mechanism. Data sets are imputed by FCS-RF varying by the number of imputed data sets (M). For each data set, clustering is performed using k-means clustering and consensus clustering is performed using NMF

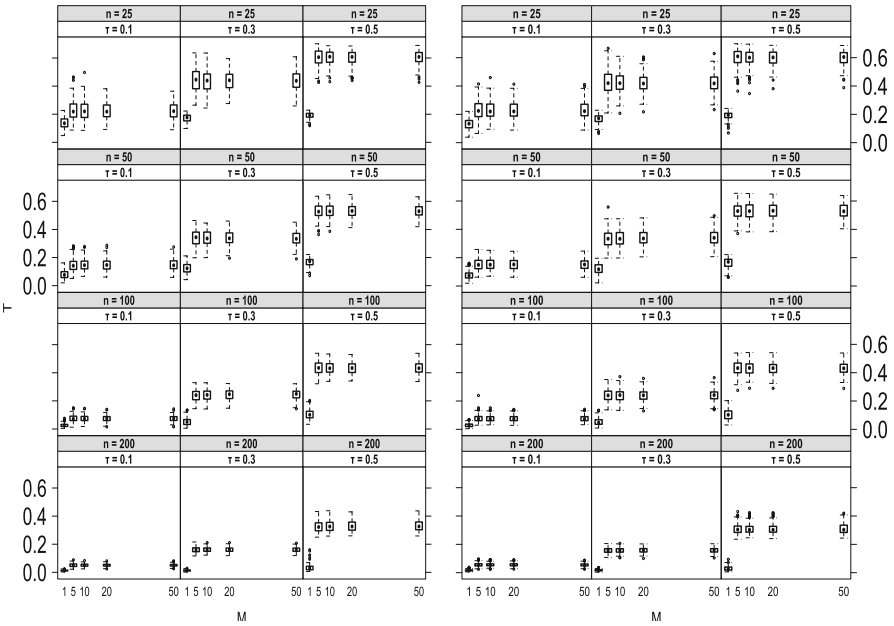


(a) ARI distribution for $\rho = 0.3$ **(b)** ARI distribution for $\rho = 0.6$

Fig. 9 Accuracy of the clustering procedure according to M: adjusted rand index over the $S = 200$ generated data sets varying by the number of individuals (n), the correlation between variables (ρ) and the proportion of missing values (τ) generated under a MAR mechanism. Data sets are imputed by FCS-RF varying by the number of imputed data sets (M). For each data set, clustering is performed using k-means clustering and consensus clustering is performed using NMF

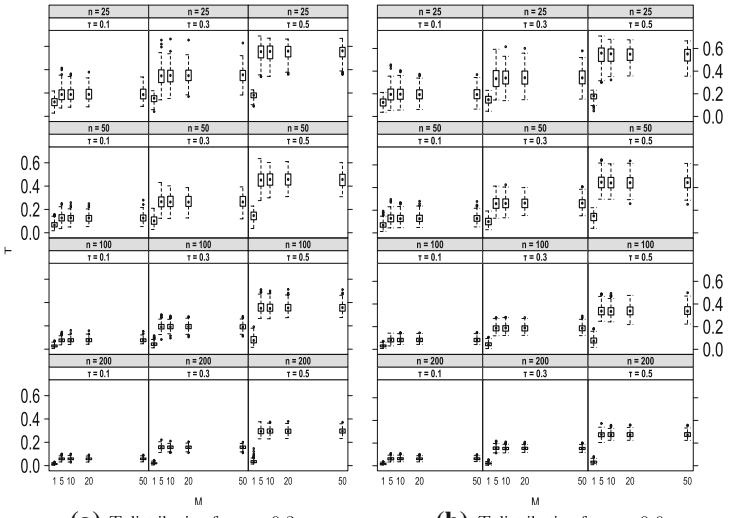
Instability

See Figs. 10, 11 and 12.



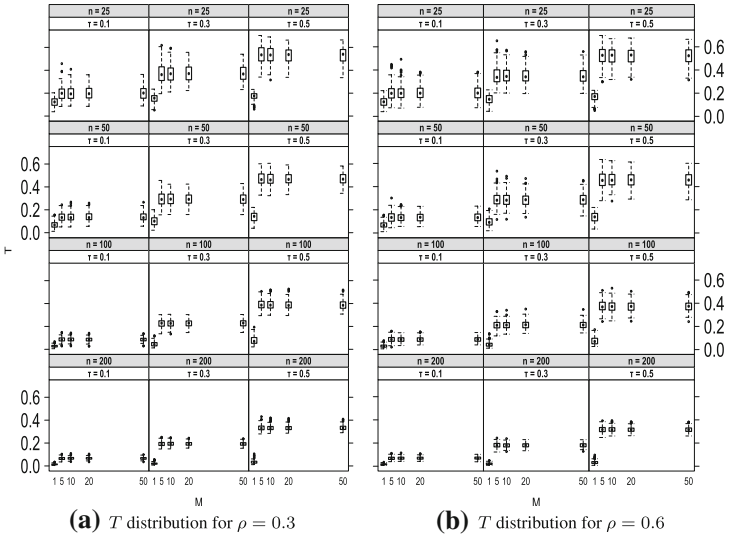
(a) T distribution for $\rho = 0.3$ **(b)** T distribution for $\rho = 0.6$

Fig. 10 Instability according to M: total instability (T) over the $S = 200$ generated data sets varying by the number of individuals (n), the correlation between variables (ρ) and the proportion of missing values (τ) generated under a MAR mechanism. Data sets are imputed by JM-DP varying by the number of imputed data sets (M). For each data set, clustering is performed using k-means clustering and consensus clustering is performed using NMF



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

Fig. 11 Instability according to M: total instability (T) over the $S = 200$ generated data sets varying by the number of individuals (n), the correlation between variables (ρ) and the proportion of missing values (τ) generated under a MCAR mechanism. Data sets are imputed by FCS-RF varying by the number of imputed data sets (M). For each data set, clustering is performed using k-means clustering and consensus clustering is performed using NMF



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

Fig. 12 Instability according to M: total instability (T) over the $S = 200$ generated data sets varying by the number of individuals (n), the correlation between variables (ρ) and the proportion of missing values (τ) generated under a MAR mechanism. Data sets are imputed by FCS-RF varying by the number of imputed data sets (M). For each data set, clustering is performed using k-means clustering and consensus clustering is performed using NMF

Application

See Table 7.

Table 7 Animals data set

	War	Fly	Ver	End	Gro	Hai
Ant	0	0	0	0	1	0
Bee	0	1	0	0	1	1
Cat	1	0	1	0	0	1
Cpl	0	0	0	0	0	1
Chi	1	0	1	1	1	1
Cow	1	0	1	0	1	1
Duc	1	1	1	0	1	0
Eag	1	1	1	1	0	0
Ele	1	0	1	1	1	0
Fly	0	1	0	0	0	0
Fro	0	0	1	1		0
Her	0	0	1	0	1	0
Lio	1	0	1		1	1
Liz	0	0	1	0	0	0
Lob	0	0	0	0		0
Man	1	0	1	1	1	1
Rab	1	0	1	0	1	1
Sal	0	0	1	0		0
Spi	0	0	0		0	1
Wha	1	0	1	1	1	0

References

Al-Najdi A, Pasquier N, Precioso F (2016) Frequent closed patterns based multiple consensus clustering. In: Rutkowski L, Korytkowski M, Scherer R, Tadeusiewicz R, Zadeh LA, Zurada JM (eds) Artificial intelligence and soft computing. Springer, Cham, pp 14–26

Andrzej CAHP, Zdunek R, ichi AS, (2009) Alternating least squares and related algorithms for NMF and SCA problems. Wiley, vol 4, pp 203–266. https://doi.org/10.1002/9780470747278.ch4, https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470747278.ch4

Audigier V, Niang N, Resche-Rigon M (2021) Clustering with missing data: Which imputation model for which cluster analysis method? arXiv:2106.04424

Basagana X, Barrera-Gomez J, Benet M, Anto JM, Garcia-Aymerich J (2013) A framework for multiple imputation in cluster analysis. Am J Epidemiol 177(7):718–725. https://doi.org/10.1093/aje/kws289

Belbin L, Faith DP, Milligan GW (1992) A comparison of two approaches to beta-flexible clustering. Multivar Behav Res 27(3):417–433. https://doi.org/10.1207/s15327906mbr2703_6

Bruckers L, Molenberghs G, Dendale P (2017) Clustering multiply imputed multivariate high-dimensional longitudinal profiles. Biomet J 59(5):998–1015. https://doi.org/10.1002/bimj.201500027

Chi JT, Chi EC, Baraniuk RG (2016) k-pod: a method for k-means clustering of missing data. Am Stat 70(1):91–99. https://doi.org/10.1080/00031305.2015.1086685

Day W (1986) Foreword: comparison and consensus of classifications. J Classif 3(2):183–185. https://doi.org/10.1007/BF01894187
Dimitriadou E, Weingessel A, Hornik K (2002) A combination scheme for fuzzy clustering. In: Pal NR, Sugeno M (eds) Advances in soft computing: AFSS 2002. Springer, Berlin, pp 332–338
Doove L, van Buuren S, Dusseldorp E (2014) Recursive partitioning for missing data imputation in the presence of interaction effects. Comput Stat Data Anal 72:92–104. https://doi.org/10.1016/j.csda.2013.10.025
Dudoit S, Fridly J (2003) Bagging to improve the accuracy of a clustering procedure. Bioinformatics:1090–1099
Fang Y, Wang J (2012) Selection of the number of clusters via the bootstrap method. Comput Stat Data Anal 56(3):468–477. https://doi.org/10.1016/j.csda.2011.09.003
Faucheux L, Resche-Rigon M, Curis E, Soumelis V, Chevret S (2020) Clustering with missing and left-censored data: a simulation study comparing multiple-imputation-based procedures. Biomet J. https://doi.org/10.1002/bimj.201900366
Filkov V, Skiena S (2004) Integrating microarray data by consensus clustering. Int J Artif Intell Tools 13(04):863–880. https://doi.org/10.1142/S0218213004001867
Forgy E (1965) Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics 21:768–780
Hennig C (2007) Cluster-wise assessment of cluster stability. Comput Stat Data Anal 52(1):258–271. https://doi.org/10.1016/j.csda.2006.11.025
Hubert L, Arabie P (1985) Comparing partitions. J Classif 2(1):193–218
Jain A, Moreau J (1987) Bootstrap technique in cluster analysis. Pattern Recogn 20(5):547–568. https://doi.org/10.1016/0031-3203(87)90081-1
Jain BJ (2017) Consistency of mean partitions in consensus clustering. Pattern Recogn 71:26–35. https://doi.org/10.1016/j.patcog.2017.04.021
Josse J, Chavent M, Liquet B, Husson F (2012) Handling missing values with regularized iterative multiple correspondence analysis. J Classif 29(1):91–116
Kaufman L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley
Kim HJ, Reiter JP, Wang Q, Cox LH, Karr A (2014) Multiple imputation of missing or faulty values under linear constraints. J Bus Econ Stat 32(3):375–386. https://doi.org/10.1080/07350015.2014.885435
Li T, Ding C, Jordan MI (2007) Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In: Proceedings of the 2007 seventh IEEE international conference on data mining, IEEE Computer Society, USA, ICDM '07, pp 577–582. https://doi.org/10.1109/ICDM.2007.98
Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2019) Cluster: cluster analysis basics and extensions
Marshall A, Altman D, Holder R, Royston P (2009) Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. BMC Med Res Methodol 9(1):57
McLachlan GJ, Basford KE (1988) Mixture models: inference and applications to clustering, vol 38. M. Dekker New York
Meng XL (1994) Multiple-imputation inferences with uncongenial sources of input (with discussion). Stat Sci 10:538–573
Mourer A, Forest F, Lebbah M, Azzag H, Lacaille J (2020) Selecting the number of clusters k with a stability trade-off: an internal validation criterion. arXiv:2006.08530
Plaehn D (2019) Revisiting french tomato data: cluster analysis with incomplete data. Food Qual Pref 76:146–159. https://doi.org/10.1016/j.foodqual.2019.03.014
Rubin D (1976) Inference and missing data. Biometrika 63:581–592
Rubin D (1987) Multiple imputation for non-response in survey. Wiley, New-York
Schafer J (1997) Analysis of incomplete multivariate data. Chapman & Hall/CRC, London
Schafer J (2003) Multiple imputation in multivariate problems when the imputation and analysis models differ. Statistica Neerlandica 57(1):19–35
Strehl A, Ghosh J, Cardie C (2002) Cluster ensembles: a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 3:583–617
van Buuren S (2018) Flexible imputation of missing data (Chapman & Hall/CRC Interdisciplinary Statistics). Chapman and Hall/CRC
Vega-Pons S, Ruiz-Shulcloper J (2011) A survey of clustering ensemble algorithms. IJPRAI 25(3):337–372

Wang J (2010) Consistent selection of the number of clusters via crossvalidation. Biometrika 97(4):893–904. https://doi.org/10.1093/biomet/asq061
Ward JHJ (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58(301):236–244. https://doi.org/10.1080/01621459.1963.10500845