Model-based Clustering with Missing Not At Random Data

Aude Sportisse, Matthieu Marbac, Fabien Laporte, Gilles Celeux, Claire Boyer, Julie Josse, Christophe Biernacki

Model-based Clustering with Missing Not At Random Data

Aude Sportisse, Matthieu Marbac, Fabien Laporte, Gilles Celeux, Claire Boyer, Julie Josse, Christophe Biernacki

Abstract

Model-based unsupervised learning, as any learning task, stalls as soon as missing data occurs. This is even more true when the missing data are informative, or said missing not at random (MNAR). In this paper, we propose model-based clustering algorithms designed to handle very general types of missing data, including MNAR data. To do so, we introduce a mixture model for different types of data (continuous, count, categorical and mixed) to jointly model the data distribution and the MNAR mechanism, remaining vigilant to the relative degrees of freedom of each. Several MNAR models are discussed, for which the cause of the missingness can depend on both the values of the missing variable themselves and on the class membership. However, we focus on a specific MNAR model, called MNAR$\boldsymbol{z}$, for which the missingness only depends on the class membership. We first underline its ease of estimation, by showing that the statistical inference can be carried out on the data matrix concatenated with the missing mask considering finally a standard MAR mechanism. Consequently, we propose to perform clustering using the Expectation Maximization algorithm, specially developed for this simplified reinterpretation. Finally, we assess the numerical performances of the proposed methods on synthetic data and on the real medical registry TraumaBase® as well.

Keywords: Model-based Clustering, Informative Missing Values, EM and Stochastic EM Algorithms, Medical Data

1 Introduction

Clustering remains a crucial tool for the comprehensive analysis of large datasets, providing a concise summary through the grouping of observations. Notably, the model-based paradigm [1, 2] facilitates clustering by yielding interpretable models that enhance our understanding of the relationships between the formed clusters and the features at play. This parametric framework exhibits flexibility in handling high-dimensionality problems [3, 4], mixed datasets [5], and even time series and dependent

data [6, 7]. However, the challenge accompanying this multifaceted model-based clustering lies in the requisite modeling efforts to design mixture models tailored to the data structure.

In large-scale data analysis, the problem of missing data is ubiquitous, data collection being never perfect (*e.g.* machines which fail, non-responses in a study). Classical approaches for dealing with missing data consist of working on a complete dataset [8], either by using only complete individuals, or by imputing missing values. However, both methods can cause huge problems in the analysis, either by reducing too drastically the dataset to a possibly biased subsample, or by distorting the distribution of the completed samples, respectively. Furthermore, it's crucial to note that both of these strategies are essentially preprocessing steps and are not specifically tailored for the final clustering task. Alternatively, one can explore likelihood-based approaches, employing methods like Expectation Maximization (EM) algorithms [9]. In this paper, we adopt such an approach to enable model-based clustering to effectively handle informative missing data in an efficient manner.

We assume that the missing data are missing not at random (MNAR) [10–12]. More specifically, we consider that the cause of the missingness can be explained by the membership to a class, which is not observed, and we call this specific MNAR model MNARz. As the MNAR mechanism is neither ignorable for the density estimation (parameters estimation), nor for the clustering (partition estimation), dealing with such data does require the specific modeling effort for the distribution of the missing-data pattern, indicating where are the missing values in the data. An example of MNAR data includes clinical data collected in emergency situations, where doctors may choose to treat patients suffering from severe trauma before taking measurements. This intrinsically leads to more missing data in this class.

Related works on clustering despite missing values

In order to handle missing values in a model-based clustering framework, Hunt and Jorgensen [13] implement the standard EM algorithm [9] based on the observed likelihood and Serafini et al. [14] propose to perform multiple imputations (with Monte Carlo methods) in the E-step. However, both works only consider MCAR data, when the cause of the missingness is completely independent from the data values.

Different clustering methods have been developed to deal with MNAR mechanisms. In a partition-based framework, Chi et al. [15] propose an extension of k-means clustering for missing data, called k-Pod, without requiring the missing-data pattern to be modelled. However, like k-means clustering, the k-Pod algorithm relies on strong assumptions as equal proportions between clusters. Du Roy De Chaumaray and Marbac [16] perform clustering via a semiparametric mixture model using the pattern-mixture approach to formulate the joint distribution, which makes the method not suitable for estimating the density parameters or imputing missing values. For longitudinal data, Beunckens et al. [17] and Kuha et al. [18] jointly model the measurements and the dropout process by using an extension of the shared-parameter model.

Contributions

We first present the considered MNARz mechanism, in which the missingness depends on class membership, within the context of unsupervised classification based on mixture models for different types of data (continuous, count, categorical, and mixed). We then demonstrate that, under MNARz, statistical inference can be conducted on the augmented matrix formed from the concatenation of the data matrix and the missing-data pattern (binary mask indicating where are the missing values) by considering a missing at random (MAR) mechanism instead. In other words, the cause of the missingness can be explained by the observed variables, making it ignorable. This offers a significant advantage, as there is no need to model the missing-data mechanism in such cases. This also provides theoretical insights into an approach commonly used in practice without theoretical foundations, wherein working on the augmented data matrix under a MAR assumption is usually suggested to facilitate efficient learning, despite a more complex underlying missing mechanism [19]. We then propose an EM algorithm, which has been implemented and is available for reproducibility[1]. More general MNAR models are also discussed, for which the cause of the missingness can depend on the values of the missing variable themselves, as well as identifiability and estimation strategies. However, they are empirically shown to be well apprehended by the more simple MNARz model. Finally, we assess the numerical performance of our method on synthetic data and the real medical registry TraumaBase®.

2 Missing-data in model-based clustering

Set the dataset $Y = (\mathbf{y}_1|\dots|\mathbf{y}_n)^T$ consisting of n individuals, where each observation $\mathbf{y}_i = (y_{i1}, \dots, y_{id})^T$ belongs to a space $\mathcal{Y}$, depending on the type of data, defined by d features. The pattern of missing data is denoted by $C = (\mathbf{c}_1|\dots|\mathbf{c}_n)^T \in \{0,1\}^{n\times d}$, with $\mathbf{c}_i = (\mathbf{c}_{i1}, \dots, \mathbf{c}_{id})^T \in \{0,1\}^d$: $\mathbf{c}_{ij} = 1$ indicates that the value y_{ij} is missing and $\mathbf{c}_{ij} = 0$ otherwise. The values of the observed (resp. missing) variables for individual i are denoted by $\mathbf{y}_i^{\text{obs}}$ (resp. $\mathbf{y}_i^{\text{mis}}$). The objective of clustering is to estimate an unknown partition $Z = (\mathbf{z}_1|\dots|\mathbf{z}_n)^T \in \{0,1\}^{n\times K}$ that groups the full dataset Y into K classes, with $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})^T \in \{0,1\}^K$ and where $z_{ik} = 1$ if $\mathbf{y}_i$ belongs to cluster k, $z_{ik} = 0$ otherwise. Consequently, in a clustering context, the missing data are not only the values $\mathbf{y}_i^{\text{mis}}$ but also the partition labels $\mathbf{z}_i$.

2.1 Mixture models

Mixture models allow for clustering by modeling the distribution of the observed data $(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i)$. Assuming an underlying mixture model with K components, the probability distribution function (pdf) of the couple $(\mathbf{y}_i, \mathbf{c}_i)$ reads as

$$f(\mathbf{y}_i, \mathbf{c}_i; \theta) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i; \lambda_k) f_k(\mathbf{c}_i \mid \mathbf{y}_i; \psi_k), \tag{1}$$

[1]The code is available on https://github.com/AudeSportisse/Clustering-MNAR/tree/main.

where $\theta = (\gamma, \psi)$ gathers all the model parameters, $\gamma = (\pi, \lambda)$ groups the parameters related to the marginal distribution of $\mathbf{y}_i$, $\pi = (\pi_1, \ldots, \pi_K)$ is the vector of proportions with $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0$ for all $k \in \{1, \ldots, K\}$. Given $\lambda = (\lambda_1, \ldots, \lambda_K)$, $f_k(\cdot\,; \lambda_k)$ is the pdf of the k-th component parameterized by λ_k, $\psi = (\psi_1, \ldots, \psi_K)$ groups the parameters of the missingness mechanisms and $f_k(\mathbf{c}_i \mid \mathbf{y}_i; \psi_k)$ is the pdf related to the missingness mechanism under component k (*i.e.*, $f_k(\mathbf{c}_i \mid \mathbf{y}_i; \psi_k) = f(\mathbf{c}_i \mid \mathbf{y}_i, z_{ik} = 1; \psi_k)$). In many cases, the parameter ψ is interpreted as a nuisance parameter. However, when the mechanism is not ignorable, i.e. can not be ignored when performing inferences for λ, we need to consider the whole parameter θ to achieve clustering, since the pdf of the observed data is

$$f(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta) = \int f(\mathbf{y}_i, \mathbf{c}_i; \theta) d\mathbf{y}_i^{\text{miss}}. \tag{2}$$

Different types of pdf $f_k(\cdot\,; \lambda_k)$ can be considered, depending on the types of features at hand. Thus, if $\mathbf{y}_i$ is a vector of continuous variables, the pdf of a d-variate Gaussian distribution [1, 20] can be considered for $f_k(\mathbf{y}_i; \lambda_k)$ and thus λ_k groups the mean vector and the covariance matrix. Moreover, if some components of $\mathbf{y}_i$ are discrete or categorical, the latent class model (see [21, 22]) defining $f_k(\mathbf{y}_i; \lambda_k) = \prod_{j=1}^{d} f_{kj}(y_{ij}; \lambda_{kj})$ can be used, with $\lambda_k = (\lambda_{k1}, \ldots, \lambda_{kd})$. In such a case, f_{kj} could be the pdf of a Poisson (resp. multinomial) distribution with parameter λ_{kj} if y_{ij} is an integer (resp. categorical) variable. The choice of the modeling for the missingness mechanism (*i.e.*, the distribution $f_k(\mathbf{c}_i \mid \mathbf{y}_i; \psi_k)$) is discussed in the following.

To formulate the joint distribution of the data and the missing-data pattern, we consider in this paper the selection model [23], which factorizes it into the product of the marginal data density and the missing-data mechanism (1). This approach has the great advantage of allowing imputation of the missing values and density estimation throughout the parameter estimation of the mixture model. Another approach, called the pattern-mixture model [24], can be used, involving the product of the marginal density of C and the conditional density of Y given C; it has been considered by [25] for a clustering purpose.

2.2 The MNARz model

To handle MNAR data in selection models, the distribution of the missing-data pattern given the data and the partition should be specified. We consider that the elements of $\mathbf{c}_i$ are conditionally independent given $(\mathbf{y}_i, \mathbf{z}_i)$. By the categorical nature of the mask $\mathbf{c}_i$, this independence assumption is a quite natural hypothesis in the context of clustering [15, 16].

In the MNARz model, we consider that the only effect of missingness is on the class membership k, being the same for all variables. More specifically, the conditional distribution of c_{ij} given $(\mathbf{y}_i, \mathbf{z}_i)$ is assumed to be a generalized linear model with link

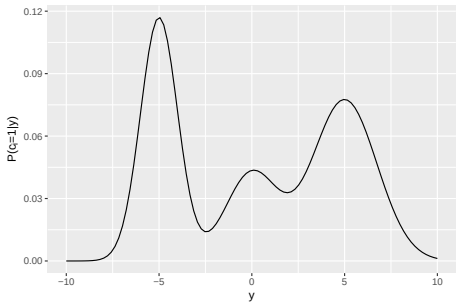


Fig. 1: Illustration of dependency between $\mathbf{c}_i$ and $\mathbf{y}_i$ in a MNARz model by drawing $\mathbb{P}(c_i \mid \mathbf{y}_i; \pi, \lambda, \psi)$ for a three component univariate Gaussian model with mixing proportions $\pi_1 = \pi_2 = 0.3$ and $\pi_3 = 0.4$, with centers $\mu_1 = \mu_3 = -5$ and $\mu_2 = 0$, and with variances $\sigma_k^2 = k$ ($k \in \{1, 2, 3\}$). The MNARz parameters are fixed to $\alpha_1 = 2$, $\alpha_2 = 0$ and $\alpha_3 = 1$.

function ρ, so that finally

$$f_k(\mathbf{c}_i \mid \mathbf{y}_i; \psi_k) = \prod_{j=1}^{d} \rho(\alpha_k)^{c_{ij}} \left(1 - \rho(\alpha_k)\right)^{1-c_{ij}}, \tag{3}$$

where $\psi_k = \alpha_k$, in this case. The MNARz model is the simplest of the MNAR models we can propose (see Section 3.3 for more general ones). Roughly speaking, this model assumes that the proportion of missing values can vary among the clusters. Although MNARz does not directly involve $\mathbf{y}_i$ in its ground definition (3), it does not mean that the pattern $\mathbf{c}_i$ does not depend on $\mathbf{y}_i$ since $\mathbf{z}_i$ depends itself on $\mathbf{y}_i$; see Figure 1 for an illustration.

3 Proposal

3.1 Reinterpretation of the MNARz model as a MAR strategy

Interestingly, the MNARz model can be turned into a MAR-like one by working on the augmented matrix formed from the concatenation of the data matrix and the missing-data pattern. This is the purpose of the next theorem, proven in Appendix A.

Theorem 1. *Consider the augmented dataset* $(\tilde{\mathbf{y}}_1^{\text{obs}}, \ldots, \tilde{\mathbf{y}}_n^{\text{obs}})$, $\tilde{\mathbf{y}}_i^{\text{obs}} = (\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i)$ *for* $i \in \{1, \ldots, n\}$. *Assume that all* $\tilde{\mathbf{y}}_i^{\text{obs}}$ *arise i.i.d. from the mixture model with a MAR mechanism*

$$\tilde{f}(\tilde{\mathbf{y}}_i^{\text{obs}}; \theta) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i^{\text{obs}}; \lambda_k) \prod_{j=1}^{d} \rho(\alpha_k)^{c_{ij}} \left(1 - \rho(\alpha_k)\right)^{1-c_{ij}}. \tag{4}$$

Then for fixed parameter $\theta = (\pi, \lambda, \psi)$, the mixture model for $\tilde{\mathbf{y}}_i^{\text{obs}}$ is the same as the distribution for $\mathbf{y}_i^{\text{obs}}$ with the mixture model (1) under the MNARz mechanism (3).

Theorem 1 implies that the maximum likelihood estimate of θ is the same considering $\tilde{\mathbf{y}}_i^{\text{obs}}$ under the MAR assumption and $\mathbf{y}_i^{\text{obs}}$ under the MNARz assumption (3). This implies that if the mechanism is MNARz, an (EM) algorithm designed for MAR data can be used on the augmented data set instead, capitalizing on efficient implementations dedicated to such a well-studied setting (see Section 4). In fact, Theorem 1 is the first theoretical result in unsupervised learning in line with the intuition developed in [19] for supervised learning and in [26] for estimation in low-rank models, that working with MAR strategies on the data set augmented by the missing pattern can actually tackle certain types of MNAR settings.

Furthermore, the identifiability of the model parameters when considering MNARz data follows directly from this reinterpretation as a MAR model. Indeed, identifiability in the complete case implies identifiability when MCAR or MAR values occur.

3.2 Associated EM algorithm

Assuming identifiability, we estimate parameters via likelihood maximization using the EM algorithm specifically designed for Gaussian, Poisson, multinomial and mixed data with MNAR data. Details of the algorithm are given in Appendix B. Assuming that the number K of clusters is known (its choice in practice is discussed in Section 4) and that the samples $(y_i, z_i, c_i)_{i=1,\dots,n}$ are i.i.d., the complete-data log-likelihood can be written as

$$\ell_{\text{comp}}(\theta; Y, Z, C) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \left(\pi_k f_k(\mathbf{y}_i; \lambda_k) f_k(\mathbf{c}_i \mid \mathbf{y}_i; \psi_k) \right). \tag{5}$$

The EM algorithm [9] is an iterative algorithm that permits to maximize the likelihood function under missingness. Initialized at the point $\theta^{[0]}$, its iteration $[r]$ consists, at the E-step, in computing the expectation of the complete-data log-likelihood $Q(\theta; \theta^{[r-1]}) = \mathbb{E}_{\theta^{[r-1]}} \left[\ell_{\text{comp}}(\theta; Y, Z, C) \mid \mathbf{Y}^{\text{obs}}, \mathbf{C} \right]$, then, at the M-step, updating the parameters by maximizing this function $\theta^{[r]} = \arg\max_{\theta} Q(\theta; \theta^{[r-1]})$. Note that

$$Q(\theta; \theta^{[r-1]}) = \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik}(\theta^{[r-1]}) \left[\log(\pi_k) + \tau_y(\lambda_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]}) + \tau_c(\psi_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]}) \right],$$

$$\begin{aligned}
\text{where} \quad & t_{ik}(\theta^{[r-1]}) = \frac{1}{f(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})} \int \pi_k^{[r-1]} f_k(\mathbf{y}_i; \lambda_k^{[r-1]}) f_k(\mathbf{c}_i \mid \mathbf{y}_i; \psi_k^{[r-1]}) d\mathbf{y}_i^{\text{miss}}, \\
& \tau_y(\lambda_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]}) = \mathbb{E}_{\theta^{[r-1]}} \left[\log f_k(y_i; \lambda_k) \mid \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, z_{ik} = 1 \right], \\
& \tau_c(\psi_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]}) = \mathbb{E}_{\theta^{[r-1]}} \left[\log f_k(c_i \mid y_i; \psi_k) \mid \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, z_{ik} = 1 \right].
\end{aligned}$$

Thus, the iteration $[r]$ of the EM algorithm is defined by

- **E-step:** Computation of $t_{ik}(\theta^{[r-1]})$, $\tau_y(\lambda_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]})$ and $\tau_c(\psi_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]})$.
- **M-step:** Updating the parameters

$$\lambda_k^{[r]} = \arg\max_{\lambda_k} \sum_{i=1}^{n} t_{ik}(\theta^{[r-1]}) \tau_y(\lambda_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]})$$

$$\pi_k^{[r]} = \frac{1}{n} \sum_{i=1}^{n} t_{ik}(\theta^{[r-1]})$$

$$\psi_k^{[r]} = \arg\max_{\psi_k} \sum_{i=1}^{n} t_{ik}(\theta^{[r-1]}) \tau_c(\psi_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]}).$$

The E-step requires to be able to integrate the distribution of $\mathbf{y}_i^{\text{mis}}$ given $(\mathbf{y}_i^{\text{obs}}, z_{ik} = 1, \mathbf{c}_i)$ and the M-step requires to maximize the resulting function. This is straightforward under the MNARz model, because the effect of the missingness does not depend on $\mathbf{y}_i$ (see Appendix B.1 for computation details in the case of Gaussian or categorical data).

3.3 Beyond the MNARz model

In Section 2.2, we proposed the MNARz model in (3). A more general model, called MNAR$y^k z^j$, can be considered, when the effect of missingness is on both the class membership and the variable itself:

$$f_k(\mathbf{c}_i \mid \mathbf{y}_i; \psi_k) = \prod_{j=1}^{d} \left(\rho(\alpha_{kj} + \beta_{kj} y_{ij})\right)^{c_{ij}} \left(1 - \rho(\alpha_{kj} + \beta_{kj} y_{ij})\right)^{1-c_{ij}}, \tag{6}$$

where $\psi_k = (\alpha_{k1}, \beta_{k1}, \ldots, \alpha_{kK}, \beta_{kK})$. The parameter α_{kj} represents a mean effect of missingness on the k-th class membership for the variable j (note that within a same class k, α_{kj} is not necessarily equal to $\alpha_{kj'}$ for $j \neq j'$). The parameter β_{kj} represents the direct effect of missingness on the variable j which depends on the class k as well.

Simpler models can also be derived from (6) by imposing equal parameters either across the class membership, or across the variables likely to be missing. First, we introduce three models, with a lower complexity than (6), that still allow the probability of being missing to depend on both the variable itself and the class membership:

- MNARyz^j model: when $\beta_{1j} = \ldots = \beta_{Kj}$, $\forall j$, the effect of missingness on a variable is the same regardless of the class (while keeping different mean effects α_{kj} on the class membership).
- MNAR$y^k z$ model: when $\alpha_{k1} = \ldots = \alpha_{kd}$, $\forall k$, the missingness has a same mean effect on class membership shared by all variables (while allowing different self-masked and class-wise parameters β_{kj}).
- MNARyz model: when $\beta_{1j} = \ldots = \beta_{Kj}$, $\forall j$ and $\alpha_{k1} = \ldots = \alpha_{kd}$, $\forall k$, the effects on a particular variable and on the class membership can be respectively the same for all the classes and for all the variables.

Secondly, the probability to be missing can also depend only on the variable itself:

- MNARy model[2]: when $\alpha_{11} = \ldots = \alpha_{1d} = \alpha_{21} = \ldots = \alpha_{Kd}$ and $\beta_{1j} = \ldots = \beta_{Kj} \quad \forall j$, the only effect of missingness is thus on the variable j, being the same regardless of the class membership.
- MNARy^k model: when $\alpha_{11} = \ldots = \alpha_{1d} = \alpha_{21} = \ldots = \alpha_{Kd}$, the effect of missingness on the variable j depends on the class k.

Thirdly, the probability to be missing can also depend only on the class membership, so that the missingness is class-wise only. This is the case of the MNARz model given in 3, but we can also consider a slightly more general case:

- MNARz^j model: when $\beta_{kj} = 0, \ \forall(k,j)$, the effect of missingness on the class membership k is not the same for all the variables.

Finally, the simplest model is the missing completely at random (MCAR) one, characterized by no dependence on variables, neither on class membership, *i.e.*, each variable has the same probability of missing,

$$\text{MCAR:} \quad \beta_{kj} = 0, \ \forall(k,j) \text{ and } \alpha_{1j} = \ldots = \alpha_{Kj}, \ \forall j. \tag{7}$$

For each of these MNAR models, we have studied the identifiability and have proposed a specific algorithm (EM or Stochastic EM); all the details are given in the accompanying note available here. Note that for MNAR data, beyond the clustering task, the main challenge to overcome consists in proving the identifiability of the parameters of the data and the missing-data pattern distributions [28]. The identifiability study showed that the most general models lead to non-identifiable parameters for categorical data (but the identifiability holds only for the MCAR, MNARz and MNARz^j mechanisms).

Despite the possibility of defining a large number of MNAR models, we have chosen to focus on the MNARz mechanism, which is a good compromise, clearly outperforming methods that do not consider MNAR data, while limiting the computational cost of the estimation in regard of more general MNAR mechanisms. Moreover, the MNARz model is robust to model mispecification (see Section 4, Figure 3).

4 Numerical experiments on synthetic data

To assess the quality of the clustering, it is possible to use an information criterion such as the Bayesian Information Criterion (BIC) [29] or the Integrated Complete-data Likelihood (ICL) [30]. The BIC criterion is expected to select a relevant mixture model from a density estimation perspective, while the ICL is expected to select a relevant mixture model for a clustering purpose [31]. Thus, we consider the latter in the following. As the ICL involves an integral which is generally not explicit, we can use an approximate version [31] that we adapt with missing data. For a model $\mathcal{M}$ with $\nu_{\mathcal{M}}$ parameters, the ICL reads as

$$\text{ICL}(\mathcal{M}) = \ell(\hat{\theta}_{\mathcal{M}}; Y^{\text{obs}}, C) - \frac{\nu_{\mathcal{M}}}{2} \log n + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik}^{\text{MAP}}(\hat{\theta}_{\mathcal{M}}) \log(\mathbb{P}(z_{ik} = 1 | \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \hat{\theta}_{\mathcal{M}})),$$

[2] This is actually a particular case of MNAR mechanims, widely used in practice [27].

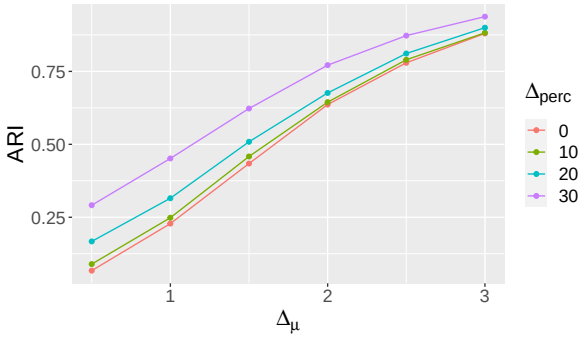


Fig. 2: Relative effect of both the separation strength Δ_μ of the mixture component and the MNAR evidence Δ_{perc} on theoretical ARI (*e.g.* if $\Delta_{\text{perc}} = 10\%$ (green line), the second class has 10% more missing values).

where $\hat{\theta}_{\mathcal{M}}$ is a maximum likelihood estimator, $\ell(\theta; Y^{\text{obs}}, C)$ is the observed log-likelihood, and with

$$z_{ik}^{\text{MAP}}(\theta) = \underset{k \in \{1,\dots,K\}}{\text{argmax}}\ \mathbb{P}(z_{ik} = 1 | \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta). \tag{8}$$

In addition, the Adjusted Rand Index (ARI) [32] can be computed between the true partition Z and the estimated one.

4.1 Leveraging from MNAR data in clustering

MNAR data are often considered as a real obstacle for statistical processing. Yet, the following numerical experiment illustrates that the MNAR mechanism may help performing the clustering task. Let us consider a bivariate isotropic Gaussian mixture model with two components and equal mixing proportions, under the MNARz mechanism (3) with a probit link function. The difference between the centers of both mixture components is taken as $\Delta_\mu = \mu_{21} - \mu_{11} = \mu_{22} - \mu_{12} \in \{0.5, 1, \dots, 3\}$. This cluster overlap controls the mixture separation, which can vary from a low separation ($\Delta_\mu = 0.5$) to a high separation ($\Delta_\mu = 3$). We also make the discrepancy between inter-cluster missing proportions $\Delta_{\text{perc}} = |\text{perc}_2 - \text{perc}_1|$, vary in $\{0, 0.1, 0.2, 0.3\}$[3] corresponds to emphasize the MNAR evidence: indeed, $\Delta_{\text{perc}} = 0$ corresponds to a MCAR model, whereas a high value of Δ_{perc} corresponds to a high difference of missing pattern proportions between clusters. For all possible values of $(\Delta_\mu, \Delta_{\text{perc}})$, 15% missing values are introduced. Figure 2 gives the theoretical ARI (*i.e.*, we compute the ARI with the theoretical parameters) as a function of the cluster overlap Δ_μ and the MNAR evidence Δ_{perc}. Although the good classification rate is mainly influenced by center separation Δ_μ, it also increases with the MNAR evidence Δ_{perc}. When classification is difficult because the mixture is not well separated ($\Delta_\mu = 0.5$), the fact that the data is MNAR helps clustering: the theoretical ARI for $\Delta_{\text{perc}} = 30$ (MNAR data)

[3]The value Δ_{perc} means that if the percentage of missing values in the first cluster is perc_1, the percentage of missing values in the second cluster is $\text{perc}_2 = (\text{perc}_1 + \Delta_{\text{perc}})$.

is significantly higher than the one for $\Delta_{\text{perc}} = 0$ (MCAR data). This toy example illustrates how clustering can leverage from MNAR values, rather generally considered a true hindrance for any statistical analysis.

4.2 Generic experiments

We consider a Gaussian mixture with three components having unequal proportions ($\pi_1 = 0.5$, $\pi_2 = \pi_3 = 0.25$) and independent variables:

$$\forall j \in \{1, \ldots, d\}, y_{ij} = \delta \sum_{k=1}^{3} \varphi_{kj} z_{ik} + \epsilon_{ij}, \tag{9}$$

with $\epsilon_{ij} \sim \mathcal{N}(0, 1)$ the noise term, $\varphi_k \in \{0, 1\}^d$ and $\delta > 0$. Thus, each entry y_{ij} follows a Gaussian distribution with variance 1 and mean $\delta \sum_{k=1}^{3} \varphi_{kj} z_{ik}$. The values of φ_{kj} are arbitrary chosen and highlight the interations between the variable j and the class membership k. This formulation allows to control, in any scenario, the theoretical rate of misclassification through the value of δ (and hence the theoretical ARI). We introduce missing values with a MNAR model (6), using a probit link function and control the rates of missingness through the value of ψ_k. For each experiment, the values of δ, ψ and ϕ are given in Appendix D. When not specified, the simulations have been performed for a theoretical rate of misclassification of 10% and a theoretical missing rate in the whole dataset of 30%.

Comparison of MNARz with other MNAR settings

We first vary the number of variables ($d = 3, 6, 9$) and consider $n = 100$ observations. The missing values are sequentially introduced with a MNAR setting. We compare the method considering the true mechanism (the one used to generate the missing values) with the EM algorithm for MCAR and MNARz values and the two-step heuristic based on `Mice`. This latter consists of first imputing the missing values using multiple imputations by chained equations [33] to get M completed datasets. Then, classical model-based clustering is performed on each completed dataset, for which the ARI is computed, Figure 3 shows the boxplot of the ARI for each scenario. First, the methods that consider a MNAR mechanism (MNAR$\star$) always outperform those that consider the MCAR mechanism and the two-step procedure based on `Mice`. Finally, the MNARz model remains a good compromise, clearly outperforming methods that do not consider MNAR data.

Moreover, in Appendix C, Figure C1 and C2 show the computation times for these numerical experiments; while the MNAR models considering that the probability of being missing depends on the variable itself are computationally very costly, MNARz model clearly limits the computational cost of the estimation.

Focus on the MNARz mechanism

Considering the setting (9) and under a MNARz mechanism, we then evaluate the impact of misspecification of the link function (Figure 4(a)), the misspecification of

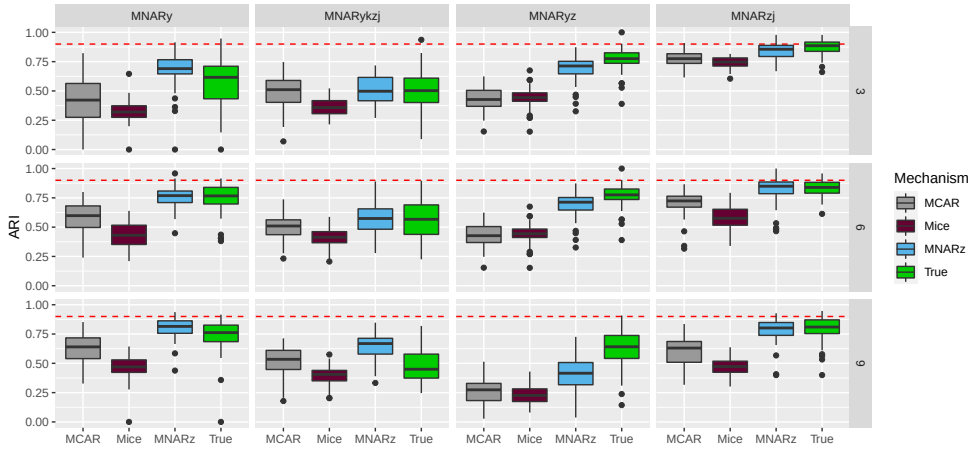
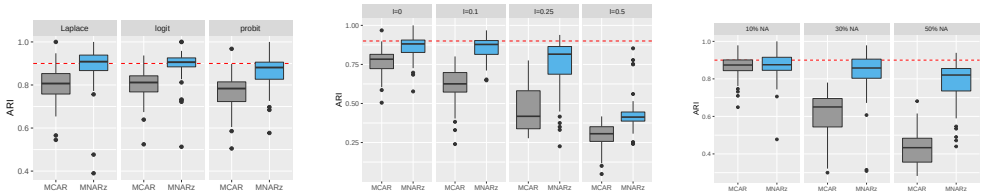


Fig. 3: Boxplot of the ARI obtained for 50 samples composed of $d = 3, 6, 9$ variables (rows) and $n = 100$ observations. Missing values are introduced with MNARy, MNAR$y^k z^j$ MNARyz or MNARz^j settings (columns). The boxplot in green is the one for the algorithm considering the true MNAR$\star$ setting; the boxplot in blue (resp. in gray) is the one for the EM algorithm considering the MNARz setting (resp. the MCAR setting); the boxplot in red in the two-step heuristic (`Mice`). The red dashed line indicates the theoretical ARI.

the data distribution (Figure 4(b)) and the percentage of missing values (Figure 4(c)) by comparing the ARI for the MNARz setting and the MCAR one.

In Figure 4(a), our algorithm always considering a probit function gives the best ARI (outperforming strategies assuming only MCAR data) regardless of the link function (Laplace distribution, logit, probit) used to introduce missing values under a MNARz model. This highlights the robustness of the MNARz setting to the link function. In Figure 4(b), we consider a three-component Gaussian mixture with non-diagonal covariance matrices. For each component, the diagonal terms of the covariance matrix are $\Sigma_{ii} = 1$ and the other terms $\Sigma_{ij} = \ell, i \neq j$, with $\ell \in \{0, 0.1, 0.25, 0.5\}$, while the algorithms assume $\ell = 0$. If the EM algorithm designed for MNARz data suffers from a huge deviation ($\ell = 0.5$) regarding the data distribution, it remains competitive for smaller ones ($\ell = 0.1, 0.25$). Finally, Figure 4(c) shows the boxplots of the ARI for 10%, 30% and 50% of missing values in the entire dataset. As the percentage of missing data increases, the gap between algorithms considering MCAR and MNARz data is widening, proving the relevancy of our algorithm even with high missing-data rates (50%). In Appendix C, we also provide the experiments for a theoretical rate of misclassification of 15%. Same conclusions hold.

When the number K of clusters is not known *a priori*, it can be automatically chosen using the ICL criterion: the idea is to run algorithms with several values for K ($K = 1, 2, 3, 4$ here), and to choose the model with the highest resulting ICL. To our knowledge, no method proposes an automatic choice of the number of clusters in unsupervised classification for the two-step heuristics, which is also a major drawback.



(a) (b) (c)

Fig. 4: Boxplot of the ARI obtained for 50 samples of dimension $d = 6$ variables. The missing values are introduced using a MNARz setting. The red dashed line indicates the theoretical ARI.

Table 1: Proportion of good selections of K ($K = 3$) using the ICL criterion for the EM algorithm, over 50 repetitions ($d = 6$).

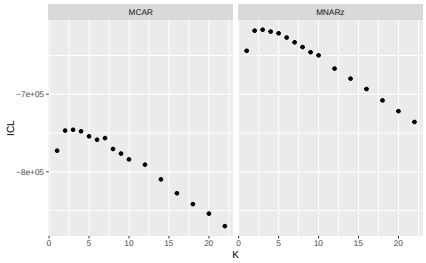
	MCAR		MNARz	
Sample size n	100	500	100	500
10 % NA	94%	100%	94%	100%
30 % NA	8%	96%	56%	100%
50 % NA	0%	0%	20%	98%

Table 1 gathers the percentages of times (over 50 repetitions) the correct number of classes ($K = 3$) is chosen by the ICL criterion for different missing-data rates (10, 30, 50%) and different sample sizes ($n = 100, 500$). In any case, the EM algorithm for MNARz data always outperforms the algorithm for MCAR data in terms of accurate model selection. The EM algorithm for MNARz data manages also to select the best model despite a high percentage of missing data (50%) provided that the sample size is large enough ($n = 500$).

We also illustrate in Appendix C (Figure C2) the findings of Theorem 1, by comparing our algorithm considering MCAR or MNARz data with the algorithm of the **RMixtComp** package [34] considering MCAR data and using the augmented data matrix $(Y|C)$. As expected, both approaches give similar results.

5 TraumaBase® dataset

In this section, we illustrate our approach on a public health application with the TraumaBase® Group (https://www.traumabase.eu/en_US) on the management of traumatized patients. This dataset contains 41 mixed variables (continuous, quantitative) on 8,248 polytraumatized patients who suffer from a major trauma (injuries from cycle or car accident). Data have been collected from 15 different hospitals. In this dataset, 11% of the data are missing and only 1.4% of the individuals are fully observed. More information on the variables can be found in Appendix E. The purpose of this real data analysis is twofold: (i) we want to know if considering the missingness process has an impact on the estimated partition, (ii) we compare our method with the classical imputation methods in Appendix E.



(a) ICL values for different K in the mixed case.

MCAR \ MNARz	Class 1	Class 2	Class 3
Class 1	**0.03**	0.47	0.61
Class 2	0.45	**0.05**	0.25
Class 3	0.63	0.23	**0.03**

(b) Total variation distance between the marginal distribution of the variable *Shock.index.ph* considering MNARz data and the one considering MCAR data.

MCAR \ MNARz	Class 1	Class 2	Class 3
Class 1	**2.43**	26.5	37.6
Class 2	26.2	**3.40**	20.1
Class 3	39.3	19.2	**2.05**

(c) Euclidean distance between conditional probabilities of the cluster memberships given observed values of the variable *Shock.index.ph*, considering MNARz or MCAR data (more details in Appendix E).

Fig. 5: Comparison of the MCAR and MNARz mechanism on Traumabase dataset.

After discussion with doctors, some variables can be considered to have MNAR values, such as *Shock.index.ph*, which denotes the ratio between heart rate and systolic arterial pressure. In fact, if this rate has a value that indicates that the patient's condition is critical, doctors cannot measure heart rate or systolic arterial pressure in emergency situations. Therefore, we expect that considering a MNAR mechanism can improve the classification.

We compare our algorithm designed for the MNARz data (3) and the MCAR data (7). Figure 5a presents the ICL values in the Traumabase dataset for different numbers of classes. If both algorithms select $K = 3$ number of classes, the ICL of the algorithm which considers MNARz data is nonetheless always significantly higher than that of the algorithm for MCAR data. Their corresponding ARI between classifications obtained assuming either MNARz or MCAR mechanisms is about 0.90. Thus, both partitions are close but not equal, which may reflect the influence of the mechanism. To deepen this issue, we focus on the variable *Shock.index.ph*. Table 5b and 5c compare the performances of the algorithm considering MNARz data with the one considering MCAR data in terms of modelling of the marginal distribution of *Shock.index.ph* and partition estimation. As the values can be compared only up to label swapping, we notice that the minimum values (on the diagonals) are significantly higher than zero, which indicates that there is an influence of the MNARz mechanism on the modeling of the data and on the classification rules.

In Appendix E, we assess also the results by using the function `catdes` of the R package FactoMineR [35] which allows to see how the cluster of the classification is described by the variables. The three groups described with our algorithm assuming MNARz data seem to be described by the same characteristics than those given by the doctors: the first group is formed by patients with a higher mortality rate and more severe injuries than the average population and the third group by a lower mortality rate and less serious injuries, whereas the second group may correspond to other cases. Note that this classification was done without using variables related to patient death and that it is quite striking to retrieve the same characteristics. This reinforces the

idea that the classification obtained makes sense and may provide other information than the one of the doctors, taking into account more variables.

6 Concluding remarks

This paper addresses model-based unsupervised learning when MNAR values occur. We propose to cluster individuals via an estimation of the mixture model parameters in play. A by-product of such an approach is that the missing values can be also imputed, once the distribution is estimated. To this end, we have proposed an approach which embeds MNAR data directly within model-based clustering algorithms, in particular the EM algorithm. We have discussed several possible MNAR specifications. However, the numerical experiment leads us to recommend using algorithms considering a simple missing-data mechanism, the MNARz mechanism, which models the probability of being missing only depending on the class membership. By its very simplicity, the latter is indeed able to straightforwardly deal with any kind of data. In addition to being interpretable (which is especially important for real applications), this MNARz mechanism can be apprehended as a MAR one on the augmented matrix $[Y|C]$, including the missing-data pattern C (Theorem 1). This echoes a widely-used approach in practice, not theoretically studied so far.

The seminal motivation of this work was clustering patients of the Traumabase dataset, in particular to assist doctors in their medical care. After a first conclusive application, there are still key challenges to make this work entirely applicable to real datasets. First, if our methodology can be applied to mixed data (categorical/quantitative), a straightforward extension of the proposed approach should be doable to handle variables that are not necessarily of the same type (MCAR, MAR and MNAR variables are indeed often coupled). Without any prior help from experts, this actually remains an open question to automatically evaluate the missing type of variables. Note however that one can arbitrate between the presented MNAR mechanisms using the ICL criterion, at the price of running multiple times the algorithm for the different MNAR scenarios.

Appendix A Proof of Theorem 1

Proof of Theorem 1. We denote by $(\tilde{\mathbf{c}}_1, \ldots, \tilde{\mathbf{c}}_n)$ the patterns of missing data associated to the observed data $\tilde{\mathbf{y}}^{\text{obs}}$. It is thus the concatenation $\tilde{\mathbf{c}}_i = (\mathbf{c}_i, \mathbf{0}_d)$ of $\mathbf{c}_i$ with the zero vector $\mathbf{0}_d = (0, \ldots, 0)$ of length d. Since all c_i values are observed in $\tilde{\mathbf{y}}_i^{\text{obs}}$, it is the reason why the last d values in $\tilde{\mathbf{c}}_i$ are fixed to zero. Then, the MAR assumption indicates that $\mathbb{P}(\tilde{\mathbf{c}}_i \mid \tilde{\mathbf{y}}_i, \mathbf{z}_i; \zeta) = \mathbb{P}(\tilde{\mathbf{c}}_i \mid \tilde{\mathbf{y}}_i^{\text{obs}}; \zeta)$, with ζ the related parameter. Consequently, using the MAR assumption and the i.i.d. assumption of all uplets $(\tilde{\mathbf{y}}_i, \mathbf{z}_i, \tilde{\mathbf{c}}_i)$, the whole likelihood can be decomposed into two likelihoods, one has

$$\begin{aligned}
L(\theta, \zeta; Y^{\tilde{\text{obs}}}, C) &= \prod_{i=1}^{n} \int f(\tilde{\mathbf{y}}_i, \tilde{\mathbf{c}}_i; \theta, \zeta) d\tilde{\mathbf{y}}_i^{\text{mis}} \\
&= \prod_{i=1}^{n} \int f(\tilde{\mathbf{y}}_i; \pi, \lambda, \psi) f(\tilde{\mathbf{c}}_i | \tilde{\mathbf{y}}_i; \zeta) d\tilde{\mathbf{y}}_i^{\text{mis}} \\
&= \prod_{i=1}^{n} \left[f(\tilde{\mathbf{c}}_i | \tilde{\mathbf{y}}_i^{\text{obs}}; \zeta) \times \int_{\mathcal{Y}_i^{\text{mis}}} f(\tilde{\mathbf{y}}_i; \pi, \lambda, \psi) d\tilde{\mathbf{y}}_i^{\text{mis}} \right] \\
&= \prod_{i=1}^{n} L(\zeta; \tilde{\mathbf{c}}_i \mid \tilde{\mathbf{y}}_i^{\text{obs}}) \times \prod_{i=1}^{n} L(\pi, \lambda, \psi; \tilde{\mathbf{y}}_i^{\text{obs}}).
\end{aligned}$$

$$L(\pi, \lambda, \psi, \zeta; \tilde{\mathbf{y}}_i^{\text{obs}}, \tilde{\mathbf{c}}_i) = L(\zeta; \tilde{\mathbf{c}}_i \mid \tilde{\mathbf{y}}_i^{\text{obs}}) \times L(\pi, \lambda, \psi; \tilde{\mathbf{y}}_i^{\text{obs}}). \tag{A1}$$

Providing that (π, λ, ψ) and ζ are functionally independent (ignorability of the MAR mechanism), the maximum likelihood estimate of $\theta = (\pi, \lambda, \psi)$ is obtained by maximizing only $L(\pi, \lambda, \psi; \tilde{\mathbf{y}}_i^{\text{obs}})$, and does not depend on $L(\zeta; \tilde{\mathbf{c}}_i \mid \tilde{\mathbf{y}}_i^{\text{obs}})$. Finally, by using (4), the observed likelihood $L(\pi, \lambda, \psi; \tilde{\mathbf{y}}_i^{\text{obs}})$ is

$$L(\pi, \lambda, \psi; \tilde{\mathbf{y}}_i^{\text{obs}}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i^{\text{obs}}; \lambda_k) \prod_{j=1}^{d} \rho(\alpha_k)^{c_{ij}} \rho(\alpha_k)^{(1-c_{ij})} \tag{A2}$$

$$= \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i^{\text{obs}}; \lambda_k) \prod_{j=1}^{d} f(c_{ij} \mid z_{ik} = 1; \psi). \tag{A3}$$

As $\mathbb{P}(c_{ij} \mid z_{ik} = 1; \psi)$ corresponds to the MNARz definition (3), the observed likelihood $L(\pi, \lambda, \psi; \tilde{\mathbf{y}}_i^{\text{obs}})$ is equal to the full observed likelihood $L(\pi, \lambda, \psi; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i)$ associated to the MNARz model,

$$L(\pi, \lambda, \psi; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i^{\text{obs}}; \lambda_k) \prod_{j=1}^{d} f(c_{ij} \mid z_{ik} = 1; \psi).$$

□

Appendix B Detail on EM algorithm

The EM algorithm consists on two steps iteratively proceeded: the E-step and M-step. For the E-step, one has

$$\begin{aligned} Q(\theta;\theta^{[r-1]}) &= \mathbb{E}[\ell_{\text{comp}}(\theta;Y,Z,C)|y_i^{\text{obs}},c_i;\theta^{[r-1]}] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\mathbb{E}\left[z_{ik}\log(\pi_k f_k(\mathbf{y}_i;\lambda)f(\mathbf{c}_i \mid \mathbf{y}_i,z_{ik}=1;\psi)) \mid y_i^{\text{obs}},c_i;\pi^{[r-1]},\lambda^{[r-1]},\psi^{[r-1]}\right] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}(\theta^{[r-1]})\mathbb{E}\left[\log(\pi_k f_k(\mathbf{y}_i;\lambda)f(\mathbf{c}_i \mid \mathbf{y}_i,z_{ik}=1;\psi)) \mid y_i^{\text{obs}},c_i,z_{ik}=1;\pi^{[r-1]},\lambda^{[r-1]},\psi^{[r-1]}\right] \end{aligned}$$

with $t_{ik}(\theta^{[r-1]}) = f(z_{ik}=1 \mid \mathbf{y}_i^{\text{obs}},\mathbf{c}_i;\theta^{[r-1]})$.

It leads to the decomposition

$$Q(\theta;\theta^{[r-1]}) = \sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}(\theta^{[r-1]})\left[\log(\pi_k) + \tau_y(\lambda_k;\mathbf{y}_i^{\text{obs}},\mathbf{c}_i,\theta^{[r-1]}) + \tau_c(\psi_k;\mathbf{y}_i^{\text{obs}},\mathbf{c}_i,\theta^{[r-1]})\right],$$

where the terms involved in this decomposition are now detailed.

(a) the expectation of the data mixture part over the missing values given the available information (*i.e.*, the observed data and the indicator pattern), the class membership and the current value of the parameters:

$$\tau_y(\lambda_k;\mathbf{y}_i^{\text{obs}},\mathbf{c}_i,\theta^{[r-1]}) = \mathbb{E}_{\theta^{[r-1]}}\left[\log f_k(y_i;\lambda_k) \mid \mathbf{y}_i^{\text{obs}},z_{ik}=1,\mathbf{c}_i\right],$$

(b) the expectation of the missing mechanism part over the missing values given the available information, the class membership and the current value of the parameters:

$$\tau_c(\psi_k;\mathbf{y}_i^{\text{obs}},\mathbf{c}_i,\theta^{[r-1]}) = \mathbb{E}_{\theta^{[r-1]}}\left[\log f_k(c_i \mid y_i;\psi_k) \mid \mathbf{y}_i^{\text{obs}},z_{ik}=1,\mathbf{c}_i\right].$$

(c) the conditional probability for an observation i to belong to the class k given the available information and the current value of the parameters:

$$t_{ik}(\theta^{[r-1]}) = f(z_{ik}=1 \mid \mathbf{y}_i^{\text{obs}},\mathbf{c}_i;\theta^{[r-1]}).$$

Terms (a) and (b) require to integrate over the distribution $f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}},z_{ik}=1,\mathbf{c}_i;\theta^{[r-1]})$. For Term (a), one has

$$\begin{aligned} &f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}},z_{ik}=1,\mathbf{c}_i;\theta^{[r-1]}) \\ &= \frac{f(\mathbf{y}_i^{\text{mis}},\mathbf{y}_i^{\text{obs}},z_{ik}=1,\mathbf{c}_i;\theta^{[r-1]})}{f(\mathbf{y}_i^{\text{obs}},z_{ik}=1,\mathbf{c}_i;\theta^{[r-1]})} \qquad \text{(B4)} \\ &= \frac{f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{mis}},\mathbf{y}_i^{\text{obs}},z_{ik}=1;\psi^{[r-1]})f(\mathbf{y}_i^{\text{mis}},\mathbf{y}_i^{\text{obs}},z_{ik}=1;\lambda^{[r-1]})}{\int_{\mathcal{Y}_i^{\text{mis}}} f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{mis}},\mathbf{y}_i^{\text{obs}},z_{ik}=1;\psi^{[r-1]})f(\mathbf{y}_i^{\text{mis}},\mathbf{y}_i^{\text{obs}},z_{ik}=1;\lambda^{[r-1]})d\mathbf{y}_i^{\text{mis}}}. \qquad \text{(B5)} \end{aligned}$$

Term (c) corresponds to the conditional probability for an observation i to arise from the kth mixture component with the current values of the model parameter. More particularly, one has

$$\begin{aligned} t_{ik}(\theta^{[r-1]}) &= \frac{f(z_{ik}=1, \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})}{f(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})} \\ &= \frac{f(z_{ik}=1, \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})}{\sum_{h=1}^{K} f(z_{ih}=1, \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})} \\ &= \frac{\pi_k^{[r-1]} f(\mathbf{y}_i^{\text{obs}} \mid z_{ik}=1; \lambda_k^{[r-1]}) f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]})}{\sum_{h=1}^{K} \pi_h^{[r-1]}) f(\mathbf{y}_i^{\text{obs}} \mid z_{ih}=1; \lambda_h^{[r-1]}) f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}, z_{ih}=1; \theta^{[r-1]})} \\ &= \frac{\pi_k^{[r-1]} f_k(\mathbf{y}_i^{\text{obs}}; \lambda_k^{[r-1]}) f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]})}{\sum_{h=1}^{K} \pi_h^{[r-1]} f_h(\mathbf{y}_i^{\text{obs}}; \lambda_h^{[r-1]}) f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}, z_{ih}=1; \theta^{[r-1]})} \end{aligned} \tag{B6}$$

B.1 Gaussian mixture for continuous data

The pdf $f_k(\mathbf{y}_i; \lambda) = \phi(\mathbf{y}_i; \mu_k, \Sigma_k)$ is assumed to be a Gaussian distribution with mean vector μ_k and covariance matrix Σ_k. First, let us detail the terms of the E-step. Term (a) is written as follows:

$$\begin{aligned} \mathbb{E}\left[\log(\phi(\mathbf{y}_i; \mu_k, \Sigma_k)) \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1, \mathbf{c}_i; \theta^{[r-1]}\right] &= -\frac{1}{2}\left[n\log(2\pi) + \log((\mid \Sigma_k \mid))\right] \\ &\quad - \frac{1}{2}\mathbb{E}\left[(\mathbf{y}_i - \mu_k)^T(\Sigma_k)^{-1}(\mathbf{y}_i - \mu_k) \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1, \mathbf{c}_i; \theta^{[r-1]}\right]. \end{aligned}$$

This last term could be expressed using the commutativity and linearity of the trace function:

$$\begin{aligned} &\mathbb{E}\left[(\mathbf{y}_i - \mu_k)^T(\Sigma_k)^{-1}(\mathbf{y}_i - \mu_k) \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1, \mathbf{c}_i; \theta^{[r-1]}\right] \\ &\qquad = \text{tr}(\mathbb{E}\left[(\mathbf{y}_i - \mu_k)(\mathbf{y}_i - \mu_k)^T \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1, c_i; \theta^{[r-1]}\right](\Sigma_k)^{-1}). \end{aligned}$$

Finally note that only $\mathbb{E}\left[(\mathbf{y}_i - \mu_k)(\mathbf{y}_i - \mu_k)^T \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1, \mathbf{c}_i; \theta^{[r-1]}\right]$ has to be calculated.

For the MNARz model, the effect of the missingness is only due to the class membership.

- For Term (a), note that

$$f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1, \mathbf{c}_i; \theta^{[r-1]}) = f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \lambda^{[r-1]}),$$

  which makes the computation easy. Indeed, using (B5),

$$f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1, \mathbf{c}_i; \theta^{[r-1]})$$

$$= \frac{\prod_{j=1}^{d} \rho(\alpha_k^{[r-1]})^{c_{ij}} (1-\rho(\alpha_k^{[r-1]}))^{1-c_{ij}} f(\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \lambda^{[r-1]})}{\int_{\mathcal{Y}_i^{\text{mis}}} \prod_{j=1}^{d} \rho(\alpha_k^{[r-1]})^{c_{ij}} (1-\rho(\alpha_k^{[r-1]}))^{1-c_{ij}} f(\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \lambda^{[r-1]}) d\mathbf{y}_i^{\text{mis}}}$$
$$= \frac{f(\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \lambda^{[r-1]})}{\int_{\mathcal{Y}_i^{\text{mis}}} f(\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \lambda^{[r-1]}) d\mathbf{y}_i^{\text{mis}}} = f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \lambda^{[r-1]}),$$

since $\prod_{j=1}^{d} \rho(\alpha_k^{[r-1]})^{c_{ij}} (1-\rho(\alpha_k^{[r-1]}))^{1-c_{ij}}$ does not depend on $\mathbf{y}_i^{\text{mis}}$ and is simplified with the numerator. The law of $(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1)$ is Gaussian (up to a reorganization of the variables associated to individual i). Noting that

$$\left(\mathbf{y}_i \mid z_{ik}=1; \lambda^{[r-1]}\right) = \left(\begin{pmatrix} \mathbf{y}_i^{\text{obs}} \\ \mathbf{y}_i^{\text{mis}} \end{pmatrix} \mid z_{ik}=1; \lambda^{[r-1]}\right)$$
$$\sim \mathcal{N}\left(\begin{pmatrix} (\mu_{ik}^{\text{obs}})^{[r-1]} \\ (\mu_{ik}^{\text{mis}})^{[r-1]} \end{pmatrix}, \begin{pmatrix} (\Sigma_{ik}^{\text{obs,obs}})^{[r-1]} & (\Sigma_{ik}^{\text{obs,mis}})^{[r-1]} \\ (\Sigma_{ik}^{\text{mis,obs}})^{[r-1]} & (\Sigma_{ik}^{\text{mis,mis}})^{[r-1]} \end{pmatrix}\right),$$

one obtains

$$\left(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \lambda^{[r-1]}\right) \sim \mathcal{N}\left((\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]}, (\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]}\right). \tag{B7}$$

with $(\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]}$ and $(\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]}$ the standard expression of the mean vector and covariance matrix of a conditional Gaussian distribution (see for instance [36]) detailed as follows

$$(\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]} = (\mu_{ik}^{\text{mis}})^{[r-1]} + (\Sigma_{ik}^{\text{mis,obs}})^{[r-1]} \left((\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}\right)^{-1} \left(\mathbf{y}_i^{\text{obs}} - (\mu_{ik}^{\text{obs}})^{[r-1]}\right), \tag{B8}$$

$$(\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]} = (\Sigma_{ik}^{\text{mis,mis}})^{[r-1]} - (\Sigma_{ik}^{\text{mis,obs}})^{[r-1]} \left((\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}\right)^{-1} (\Sigma_{ik}^{\text{obs,mis}})^{[r-1]}. \tag{B9}$$

Note also that we have

$$(\mathbf{y}_i - \mu_k)(\mathbf{y}_i - \mu_k)^T = \begin{pmatrix} (\mathbf{y}_i^{\text{obs}} - \mu_{ik}^{\text{obs}})^T (\mathbf{y}_i^{\text{obs}} - \mu_{ik}^{\text{obs}}) & (\mathbf{y}_i^{\text{obs}} - \mu_{ik}^{\text{obs}})^T (\mathbf{y}_i^{\text{mis}} - \mu_{ik}^{\text{mis}}) \\ (\mathbf{y}_i^{\text{mis}} - \mu_{ik}^{\text{mis}})^T (\mathbf{y}_i^{\text{obs}} - \mu_{ik}^{\text{obs}}) & (\mathbf{y}_i^{\text{mis}} - \mu_{ik}^{\text{mis}})^T (\mathbf{y}_i^{\text{mis}} - \mu_{ik}^{\text{mis}}) \end{pmatrix}.$$

Therefore, the expected value of each block for the current parameter value is

$$\mathbb{E}\left[(\mathbf{y}_i^{\text{obs}} - \mu_{ik}^{\text{obs}})^T (\mathbf{y}_i^{\text{obs}} - \mu_{ik}^{\text{obs}}) \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \lambda^{[r-1]}\right] = (\mathbf{y}_i^{\text{obs}} - \mu_{ik}^{\text{obs}})^T (\mathbf{y}_i^{\text{obs}} - \mu_{ik}^{\text{obs}})$$
$$\mathbb{E}\left[(\mathbf{y}_i^{\text{obs}} - \mu_{ik}^{\text{obs}})^T (\mathbf{y}_i^{\text{mis}} - \mu_{ik}^{\text{mis}}) \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \lambda^{[r-1]}\right] = (\mathbf{y}_i^{\text{obs}} - \mu_{ik}^{\text{obs}})^T ((\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]} - \mu_{ik}^{\text{mis}})$$
$$\mathbb{E}\left[(\mathbf{y}_i^{\text{mis}} - \mu_{ik}^{\text{mis}})^T (\mathbf{y}_i^{\text{mis}} - \mu_{ik}^{\text{mis}}) \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \lambda^{[r-1]}\right] = ((\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]} - \mu_{ik}^{\text{mis}})^T ((\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]} - \mu_{ik}^{\text{mis}}) + (\tilde{\Sigma}^{\text{mis}}_{ik}$$

- For Term (b), $f(\mathbf{c}_i \mid \mathbf{y}_i, z_{ik} = 1; \psi)$ is independent of $\mathbf{y}_i$, which implies

$$\log(f(\mathbf{c}_i \mid z_{ik} = 1; \psi)) = \sum_{j=1}^{d} c_{ij} \log \rho(\alpha_k) + (1 - c_{ij}) \log(1 - \rho(\alpha_k)) \tag{B10}$$

- For Term (c), one first remark that

$$\mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \theta^{[r-1]}) = \prod_{j=1}^{d} \mathbb{P}(c_{ij} = 1 \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \theta^{[r-1]})^{c_{ij}} \mathbb{P}(c_{ij} = 0 \mid y_i^{\text{obs}}, z_{ik} = 1; \theta^{[r-1]})^{1-c_{ij}}.$$

  In particular, for MNARz, by independence of $\mathbf{y}_i$, one has

$$\mathbb{P}(c_{ij} = 1 \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \theta^{[r-1]}) = \mathbb{P}(c_{ij} = 1 \mid z_{ik} = 1; \theta^{[r-1]}) = \rho(\alpha_k)$$

  Using (B6), one obtains

$$t_{ik}^{[r-1]}(\theta^{[r-1]}) = \frac{\pi_k^{[r-1]} \phi(\mathbf{y}_i^{\text{obs}}; (\mu_{ik}^{\text{obs}})^{[r-1]}, (\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}) \prod_{j=1}^{d} \rho(\alpha_k^{[r-1]})^{c_{ij}} (1 - \rho(\alpha_k^{[r-1]}))^{1-c_{ij}}}{\sum_{h=1}^{K} \pi_h^{[r-1]} \phi(\mathbf{y}_i^{\text{obs}}; (\mu_{ih}^{\text{obs}})^{[r-1]}, (\Sigma_{ih}^{\text{obs,obs}})^{[r-1]}) \prod_{j=1}^{d} \rho(\alpha_h^{[r-1]})^{c_{ij}} (1 - \rho(\alpha_h^{[r-1]}))^{1-c_{ij}}} \tag{B11}$$

If ρ is the logistic distribution, the expression can be written more simply

$$t_{ik}(\theta^{[r-1]}) \propto \pi_k^{[r-1]} \phi(\mathbf{y}_i^{\text{obs}}; \lambda_k^{[r-1]}) \prod_{j=1}^{d} \left(1 + \exp(-\delta_{ij} \alpha_k^{[r-1]})\right)^{-1} \text{ where } \delta_{ij} = \begin{cases} 1 & \text{if } c_{ij} = 1 \\ -1 & \text{otherwise.} \end{cases}$$

Finally, the E-step and the M-step can be sketched as follows in the Gaussian mixture case.

E-step The E-step for Term (a) consists of computing for $k = 1, \ldots, K$ and $i = 1, \ldots, n$

$$\begin{aligned}
(\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]} &= (\mu_{ik}^{\text{mis}})^{[r-1]} + (\Sigma_{ik}^{\text{mis,obs}})^{[r-1]} \left((\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}\right)^{-1} \left(\mathbf{y}_i^{\text{obs}} - (\mu_{ik}^{\text{obs}})^{[r-1]}\right) \\
(\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]} &= (\Sigma_{ik}^{\text{mis,mis}})^{[r-1]} - (\Sigma_{ik}^{\text{mis,obs}})^{[r-1]} \left((\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}\right)^{-1} (\Sigma_{ik}^{\text{obs,mis}})^{[r-1]} \\
(\tilde{y}_{i,k})^{[r-1]} &= (\mathbf{y}_i^{\text{obs}}, (\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]}) \\
\tilde{\Sigma}_{ik}^{[r-1]} &= \begin{pmatrix} 0_i^{\text{obs,obs}} & 0_i^{\text{obs,mis}} \\ 0_i^{\text{mis,obs}} & (\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]} \end{pmatrix}.
\end{aligned}$$

Note that whenever the mixture covariance matrices are supposed diagonal then $(\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]}$ is also a diagonal matrix. Term (c) also requires the computation of $t_{ik}(\theta^{[r-1]})$ given in (B11) for $k = 1, \ldots, K$ and $i = 1, \ldots, n$.

M-step The maximization of $Q(\theta;\theta^{[r-1]})$ over (π,λ) leads to, for $k=1,\ldots,K$,

$$\pi_k^{[r]} = \frac{1}{n}\sum_{i=1}^{n} t_{ik}(\theta^{[r-1]})$$
$$\mu_k^{[r]} = \frac{\sum_{i=1}^{n} t_{ik}(\theta^{[r-1]})(\tilde{y}_{k,i})^{[r-1]}}{\sum_{i=1}^{n} t_{ik}(\theta^{[r-1]})}$$
$$\Sigma_k^{[r]} = \frac{\sum_{i=1}^{n}\left[t_{ik}(\theta^{[r-1]})\left((\tilde{y}_{i,k})^{[r-1]}-\mu_k^r)((\tilde{y}_{i,k})^{[r-1]}-\mu_k^r)^T+\tilde{\Sigma}_{ik}^{[r-1]}\right)\right]}{\sum_{i=1}^{n} t_{ik}(\theta^{[r-1]})}.$$

Then, the maximization of $Q(\theta;\theta^{[r-1]})$ over ψ can be performed using a Newton Raphson algorithm. For $k=1,\ldots,K$, it remains to fit a generalized linear model with the binomial link function for the matrix $(\mathcal{J}_k^{\text{MNAR}z})^{[r]}$ and by giving $t_{ik}(\theta^{[r-1]})$ as prior weights to fit the process.

$$(\mathcal{J}_k^{\text{MNAR}z})^{[r]} = \begin{pmatrix} c_{.1} & 1 \\ \vdots & \vdots \\ c_{.d} & 1 \end{pmatrix}. \tag{B12}$$

The EM algorithm for the MNAR**z** model is described in Algorithm 1 for Gaussian mixture.

B.2 Latent class model for categorical data

For categorical data, we have $\phi(\mathbf{y}_i;\lambda_k)=\prod_{j=1}^{d}\phi(y_{ij};\lambda_{kj})=\prod_{j=1}^{d}\prod_{\ell=1}^{\ell_j}(\lambda_{kj}^{\ell})^{y_{ij}^{\ell}}$.

Term (a) is

$$\mathbb{E}\left[\log(\phi(\mathbf{y}_i;p_k))\mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1, \mathbf{c}_i;\lambda^{[r-1]}\right] = \sum_{j,c_{ij}=0}\sum_{\ell=1}^{\ell_j} y_{ij}^{\ell} + \sum_{j,c_{ij}=1}\sum_{\ell=1}^{\ell_j}\log(\lambda_{kj}^{y_{ij}^{\ell}}) \tag{B13}$$

Term (b) is the same as in the Gaussian case given in (B10). Finally, the EM algorithm can be summarized as follows

E step: For $k=1,\ldots,K$ and $i=1,\ldots,n$, compute

$$t_{ik}(\theta^{[r-1]}) = \frac{\pi_k^{[r-1]}\prod_{j,c_{ij}=0}\prod_{\ell=1}^{\ell_j}(\lambda_{kj}^{\ell})^{y_{ij}^{\ell}}\rho(\alpha_k)}{\sum_{h=1}^{K}\pi_h^{[r-1]}\prod_{j,c_{ij}=0}\prod_{\ell=1}^{\ell_j}(\lambda_{hj}^{\ell})^{y_{ij}^{\ell}}\rho(\alpha_h)}$$
$$(\tilde{y}_{ij,k}^{\ell})^{[r-1]} = c_{ij}(\theta_{kj}^{\ell})^{[r-1]} + (1-c_{ij})y_{ij}^{\ell},\ \forall j=1,\ldots,d, \forall \ell=1,\ldots,\ell_j.$$

Algorithm 1 EM algorithm for Gaussian mixture and MNARz model

Input: $Y \in \mathbb{R}^{n\times d}$ (matrix containing missing values), $K \geq 1$, $r_{\max}$.
Initialize π_k^0, μ_k^0, Σ_k^0 and ψ_k^0, for $k \in \{1, \ldots, K\}$.
for $r = 0$ **to** $r_{\max}$ **do**
  E-step:
  for $i = 1$ **to** n, $k = 1$ **to** K **do**
    $(\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]} = (\mu_{ik}^{\text{mis}})^{[r-1]} + (\Sigma_{ik}^{\text{mis,obs}})^{[r-1]} \left((\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}\right)^{-1} \left(\mathbf{y}_i^{\text{obs}} - (\mu_{ik}^{\text{obs}})^{[r-1]}\right)$.
    $(\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]} = (\Sigma_{ik}^{\text{mis,mis}})^{[r-1]} - (\Sigma_{ik}^{\text{mis,obs}})^{[r-1]} \left((\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}\right)^{-1} (\Sigma_{ik}^{\text{obs,mis}})^{[r-1]}$.
    $(\tilde{y}_{i,k})^{[r-1]} = (\mathbf{y}_i^{\text{obs}}, (\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]})$.
    $\tilde{\Sigma}_{ik}^{[r-1]} = \begin{pmatrix} 0_i^{\text{obs,obs}} & 0_i^{\text{obs,mis}} \\ 0_i^{\text{obs,mis}} & (\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]} \end{pmatrix}$, where $0_i^{\text{obs,obs}}$ and $0_i^{\text{obs,mis}}$ are the null matrix of size $n_i^{\text{obs}} \times n_i^{\text{obs}}$ and $n_i^{\text{obs}} \times n_i^{\text{mis}}$, with n_i^{obs} (resp. n_i^{miss}) the number of observed (reps. missing) variables for individual i.
    $t_{ik}(\theta^{[r-1]}) \propto \pi_k^{[r-1]} \phi(\mathbf{y}_i^{\text{obs}}; (\mu_{ik}^{\text{obs}})^{[r-1]}, (\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}) \prod_{j=1}^{d} \rho(\alpha_k^{[r-1]})^{c_{ij}} (1 - \rho(\alpha_k^{[r-1]}))^{1-c_{ij}}$
  end for
  M-step:
  for $k = 1$ **to** K **do**
    $\pi_k^{[r]} = \frac{1}{n}\sum_{i=1}^{n} t_{ik}(\theta^{[r-1]})$, $\mu_k^{[r]} = \frac{\sum_{i=1}^{n} t_{ik}(\theta^{[r-1]})(\tilde{y}_{k,i})^{[r-1]}}{\sum_{i=1}^{n} t_{ik}(\theta^{[r-1]})}$
    $\Sigma_k^{[r]} = \frac{\sum_{i=1}^{n}\left[t_{ik}(\theta^{[r-1]})\left((\tilde{y}_{i,k})^{[r-1]} - \mu_k^{[r]}\right)((\tilde{y}_{i,k})^{[r-1]} - \mu_k^{[r]})^T + \tilde{\Sigma}_{ik}^{[r-1]}\right)\right]}{\sum_{i=1}^{n} t_{ik}(\theta^{[r-1]})}$
    Let $\psi_k^{[r]}$ be the coefficients of a GLM with a binomial link function, by giving prior weights $t_{ik}(\theta^{[r-1]})$. In particular, the optimization problem is

$$(\mathcal{J}_k^{\text{MNAR}z})^{[r]} \psi_k^{[r]} = \log\left(\frac{1 - \mathbb{E}[\mathbf{c}|(\mathcal{J}_k^{\text{MNAR}z})^{[r]}])}{\mathbb{E}[\mathbf{c}|(\mathcal{J}_k^{\text{MNAR}z})^{[r]}]}\right),$$

    for a matrix $(\mathcal{J}_k^{\text{MNAR}z})^{[r]}$ given in (B12) and $\mathbf{c} = (\mathbf{c}_{\cdot 1}, \ldots, \mathbf{c}_{\cdot d})$ the concatenated missing data patterns for the variables $1, \ldots, d$.
  end for
end for

M step: The maximization of $Q(\theta; \theta^{[r-1]})$ over θ leads to, for $k = 1, \ldots, K$,

$$\pi_k^r = \frac{1}{n}\sum_{i=1}^{n} t_{ik}(\theta^{[r-1]})$$
$$(\theta_{kj}^{\ell})^r = \frac{\sum_{i=1}^{n} t_{ik}(\theta^{[r-1]})(\tilde{y}_{ij,k}^{\ell})^{[r-1]}}{\sum_{i=1}^{n} t_{ik}(\theta^{[r-1]})}, \ \forall j = 1, \ldots, d, \forall \ell = 1, \ldots, \ell_j.$$

d	φ
3	$\varphi_{11} = \varphi_{22} = \varphi_{33} = 1$
6	$\varphi_{11} = \varphi_{22} = \varphi_{33} = \varphi_{14} = \varphi_{36} = 1$
9	$\varphi_{11} = \varphi_{22} = \varphi_{33} = \varphi_{14} = \varphi_{36} = \varphi_{17} = \varphi_{27} = \varphi_{39} = 1$

Table D1: Choice of the values of φ and α for all the experiments of Section 4.2. Other values φ_{kj} are null.

The M-step for ψ consists of performing a GLM with a binomial link for the following matrix:

$$(\mathcal{H}^{\text{MNAR}z})^{[r]} = \left(\begin{array}{c|ccc} c_{\cdot 1} & z_{\cdot 1} & \dots & z_{\cdot K} \\ \vdots & \vdots & \vdots & \vdots \\ c_{\cdot d} & z_{\cdot 1} & \dots & z_{\cdot K} \end{array}\right) = \left(\begin{array}{c|ccc} c_{11} & z_{11}^{[r]} & \dots & z_{1K}^{[r]} \\ \vdots & \vdots & \vdots & \vdots \\ c_{n1} & z_{n1}^{[r]} & \dots & z_{nK}^{[r]} \\ \vdots & \vdots & \vdots & \vdots \\ c_{1d} & z_{11}^{[r]} & \dots & z_{1K}^{[r]} \\ \vdots & \vdots & \vdots & \vdots \\ c_{nd} & z_{n1}^{[r]} & \dots & z_{nK}^{[r]} \end{array}\right). \tag{B14}$$

B.3 Combining Gaussian mixture and latent class model for mixed data

If the data are mixed (continuous and categorical), the formulas can be extended straightforwardly if the continuous and the categorical variables are assumed to be independent knowing the latent clusters.

Appendix C Additional numerical experiments on synthetic data

Note that in Figure C2, the differences for $n = 100$ can be explained by the difference in initialization of the algorithms, which can play an important role for small sample sizes.

Appendix D Complements on generic experiments

This section gives the values of δ (see (9)) ψ (see (6)) and φ (see (9)) used during the different experiments. As explained in Section 4.2, their choice allows to control the rates of misclassification and missingness, as well as the interation between the variables and the class membership. To estimate these values, we have generated a large sample ($n = 10^5$) and compute the misclassification rate and the missingness rate for several values of δ and ψ and pick the ones which correspond to the setting of the experiment.

Appendix E Traumabase dataset

K	% NA	link	rate of misclassification	l	δ	α
3	30%	probit	90%	0	2.6	$(-1\ -0.3\ 0)$
3	30%	logit	90%	0	2.76	$(-1.5\ -0.8\ 0.1)$
3	30%	Laplace	90%	0	2.85	$(-1.1\ 0.3\ 0)$
3	30%	probit	85%	0	2.27	$(-1\ -0.3\ 0)$
3	30%	logit	85%	0	2.44	$(-1.5\ -0.8\ 0.1)$
3	30%	Laplace	85%	0	2.46	$(-1.1\ 0.3\ 0)$
3	30%	probit	90%	0.1	2.3	$(-1.16\ 0.3\ -0.42)$
3	30%	probit	90%	0.25	2.17	$(-1.16\ 0.3\ -0.4)$
3	30%	probit	90%	0.5	1.85	$(-1.16\ 0.3\ -0.4)$
3	30%	probit	85%	0.1	1.97	$(-1.16\ 0.3\ -0.42)$
3	30%	probit	85%	0.25	1.86	$(-1.16\ 0.3\ -0.4)$
3	30%	probit	85%	0.5	1.57	$(-1.16\ 0.3\ -0.4)$
3	10%	probit	90%	0	2.18	$(-1.65\ -1.2\ -0.9)$
3	50%	probit	90%	0	3.3	$(-0.55\ 0.25\ 1.7)$
3	10%	probit	85%	0	1.95	$(-1.65\ -1.2\ -0.9)$
3	50%	probit	85%	0	2.62	$(-0.55\ 0.25\ 1.7)$

Table D2: Choice of the values of δ and α for all the experiments of Section 4.2 and Appendix C for the MNARz mechanism. K denotes the number of class, the column denoted as % NA gives the rate of missingness, the column called link gives the link function of the missing-data mechanism used in the introduction of the missing values, l is the coefficient of correlation (anti-diagonal terms), δ is given in (9) and α in (6).

d	δ	α
3	20	$\begin{pmatrix} -0.4 & -0.65 & -0.65 \\ -1.1 & -1 & -1 \\ -0.6 & 0.4 & 0.4 \end{pmatrix}$
6	2.5	$\begin{pmatrix} -1.4 & -1.4 & -1.2 & -1.1 & -1 & -0.9 \\ -0.6 & 0.4 & 0.4 & 0.3 & 0.1 & 0.1 \\ -0.2 & -0.2 & -0.2 & -0.2 & -0.2 & -0.2 \end{pmatrix}$
9	1.78	$\begin{pmatrix} -0.5 & -0.65 & -0.65 & -1.1 & -1.7 & -1.7 & -1.4 & -1.4 & -1.4 \\ -0.6 & 0.4 & 0.4 & -0.2 & 0.3 & 0.4 & 0.3 & 0.3 & 0.3 \\ -0.4 & -0.4 & -0.4 & -0.4 & -0.4 & -0.4 & -0.4 & -0.4 & -0.4 \end{pmatrix}$

Table D3: Choice of the values of δ and α for all the experiments of Section 4.2 for the MNARz^j mechanism.

d	δ	α	β
3	3.5	-1.56	$(1.45\ 0.2\ -3)$
6	2.25	-0.7	$(-3\ 0.3\ -3\ -3\ -2\ 1)$
9	1.98	-0.68	$(0.5\ 0.1\ -1.2\ 0.4\ -0.1\ -1.3\ 0.3\ -0.1\ -1)$

Table D4: Choice of the values of δ, α and β for all the experiments of Section 4.2 for the MNARy mechanism.

d	δ	α	β
3	4.72	$(-1.2\ -0.8\ -0.5)$	$(-3\ 0.3\ 1)$
6	2.12	$(-1.35\ -0.29\ 0)$	$(-3\ 0.3\ -3\ -3\ -2\ 1)$
9	1.71	$(-1.34\ -0.34\ 0)$	$(-3\ 0.3\ -3\ -2.8\ -2\ 1\ 0.2\ 0.1\ 0.4)$

Table D5: Choice of the values of δ, α and β for all the experiments of Section 4.2 for the MNARyz mechanism.

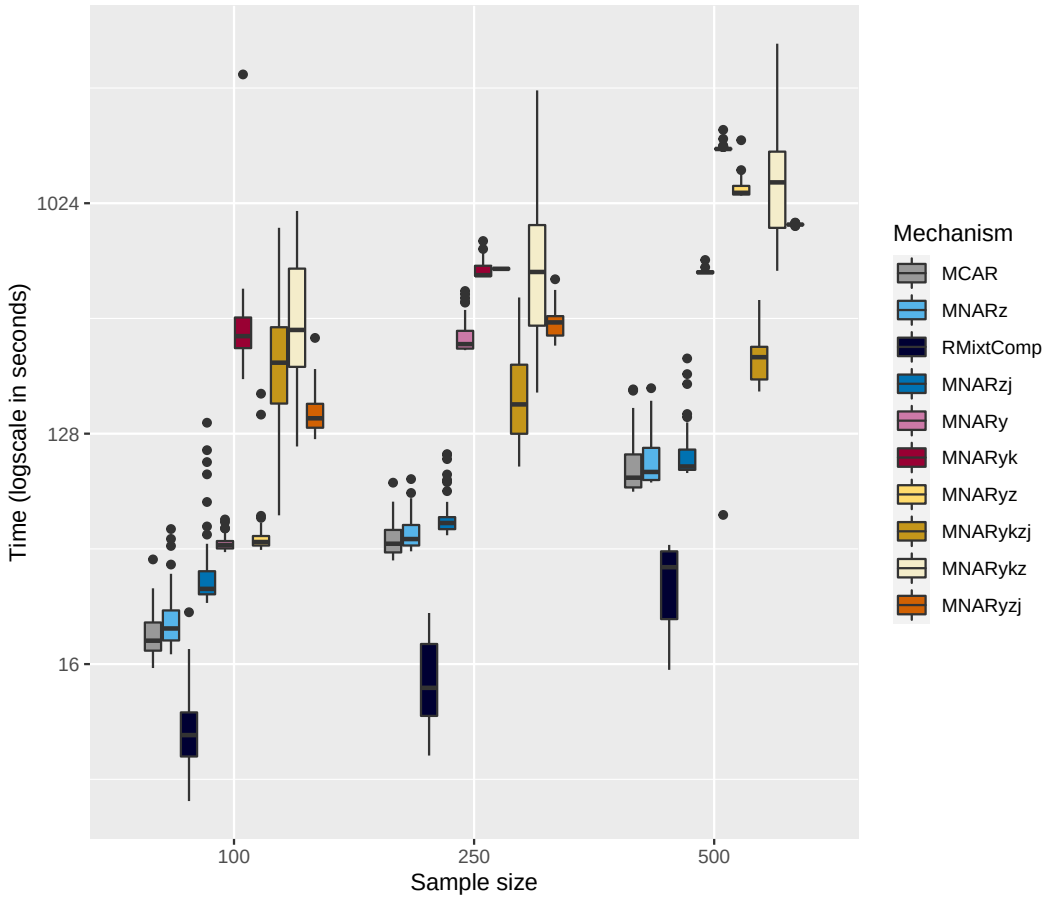


Fig. C1: Boxplot of the computational times (in seconds) obtained for 50 samples composed of $d = 6$ variables.

d	δ	α	β
3	2.55	$\begin{pmatrix} -1 & -0.95 & -0.9 \\ 0.75 & 0.7 & 0.8 \\ -0.2 & -0.2 & -0.2 \end{pmatrix}$	$\begin{pmatrix} -3 & 0.3 & -3 \\ 0.3 & -3 & 0.3 \\ -3 & 0.3 & -3 \end{pmatrix}$
6	1.96	$\begin{pmatrix} -1.2 & -1 & -0.9 & -0.9 & -0.7 & -0.8 \\ -0.6 & 0.4 & 0.4 & 0.3 & 0.1 & 0.1 \\ -0.4 & -0.4 & -0.4 & -0.4 & -0.4 & -0.4 \end{pmatrix}$	$\begin{pmatrix} -3 & 0.3 & -3 & -3 & -2 & 1 \\ 0.3 & -3 & 0.3 & -0.3 & -2 & 0.2 \\ -3 & 0.3 & -3 & -3 & -2 & 1 \end{pmatrix}$
9	1.45	$\begin{pmatrix} -1.4 & -1 & -1.1 & -1.1 & -0.9 & -0.8 & -1.2 & -1 & -1.1 \\ 0.3 & 0.5 & 0.2 & -0.6 & 0.4 & 0.4 & 0.3 & 0.1 & 0.1 \\ -0.4 & -0.4 & -0.4 & -0.4 & -0.4 & -0.4 & -0.4 & -0.4 & -0.4 \end{pmatrix}$	$\begin{pmatrix} -3 & 0.3 & -3 & -3 & -2 & 1 & -3 & 0.3 & 0.2 \\ 0.3 & -3 & 0.3 & -0.3 & -2 & 0.2 & 0.2 & 0.3 & -0.3 \\ -3 & 0.3 & -3 & -3 & -2 & 1 & -1 & -2 & -3 \end{pmatrix}$

Table D6: Choice of the values of δ, α and β for all the experiments of Section 4.2 for the MNAR$y^k z^j$ mechanism.

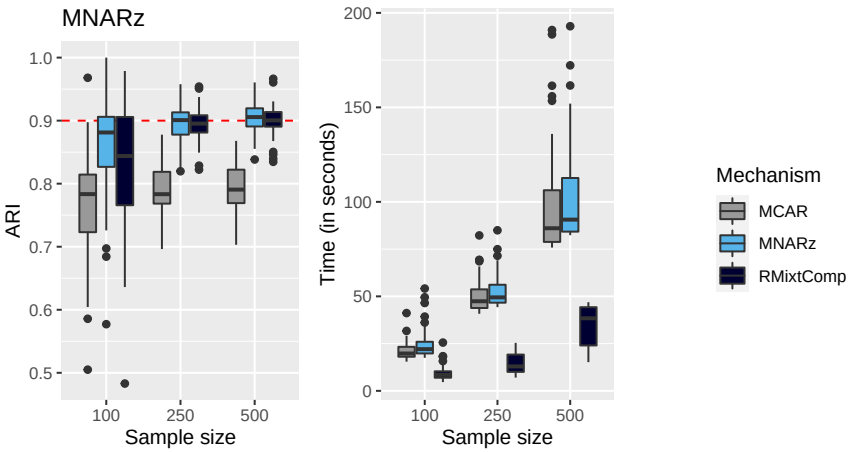


Fig. C2: Left graphic: Boxplot of the ARI obtained for 50 samples composed of $d = 6$ variables and $n \in \{100, 250, 500\}$. In grey: our EM implementation for MCAR data, in blue: our EM implementation for MNARz data, in green: SEM algorithm of **RMixtComp** for MCAR data using the augmented data matrix. Right graphic: associated computational times (in seconds).

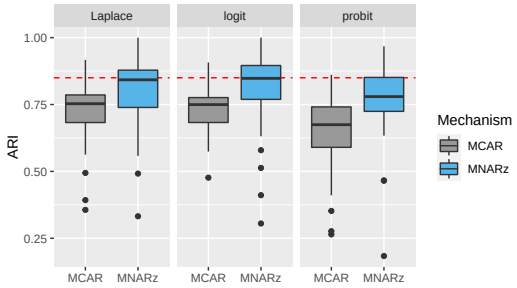


Fig. C3: Boxplot of the ARI obtained for 50 samples composed of $d = 6$ variables. The missing values are introduced using a MNARz setting. The misclassification rate is of 15%. Impact of the misspecification of the link function.

E.1 Impact of the MNARz process on the estimated partition

Table 5c gives the Euclidean distance between the conditional probabilities of the cluster memberships given the observed values of the variable *Shock.index.ph* obtained using the algorithm considering MNARz data and those obtained using the algorithm considering MCAR data. For clarity, the latter quantity is reported here,

$$\sqrt{\sum_{i=1}^{n}(\mathbb{P}(z_{ik}=1|y_{is}^{\text{obs}};\theta^{\text{MCAR}})-\mathbb{P}(z_{i\tilde{k}}=1|y_{is}^{\text{obs}};\theta^{\text{MNAR}}))^2}, \forall k, \tilde{k} \in \{1,2,3\}$$

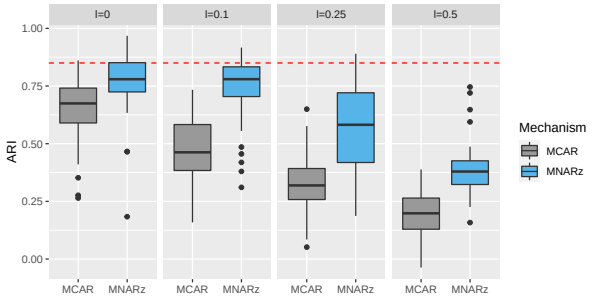


Fig. C4: Boxplot of the ARI obtained for 50 samples composed of $d = 6$ variables. The missing values are introduced using a MNARz setting. The misclassification rate is of 15%. Impact of the misspecification of the data distribution.

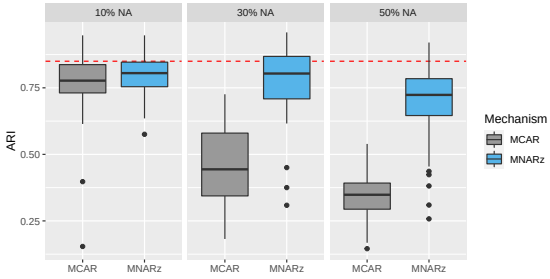


Fig. C5: Boxplot of the ARI obtained for 50 samples composed of $d = 6$ variables. The missing values are introduced using a MNARz setting. The misclassification rate is of 15%. Impact of the percentage of missing values.

with s the index of the variable *Shock.index.ph*, θ^{MCAR} (resp. θ^{MNAR}) the estimator returned by the algorithm considering MCAR data (resp. MNAR data).

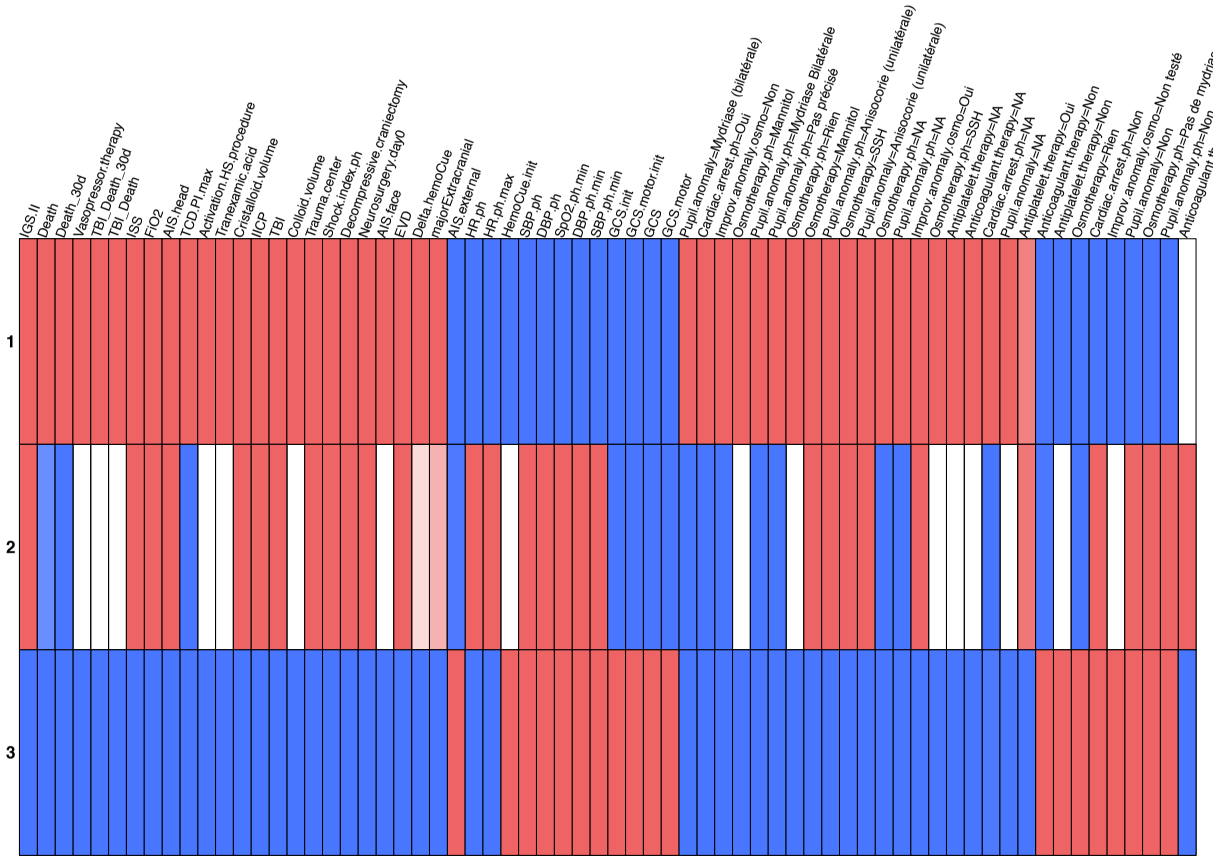


Fig. E6: Plot of the catdes output for the EM algorithm applied to the Traumabase dataset considering MNARz data. We consider $K = 3$ clusters. A cell colored in blue (resp. red) means that the average value of the variable in the given cluster is significantly lower (resp. higher) than in the overall data. In addition, the more transparent the cell is, the less significant the difference between the behavior of the variable in the given cluster and in the overall data is.

E.2 Imputation performances in the Traumabase dataset

We perform now simulations on the real dataset in order to be able to measure the quality of the imputation of our method compared to the multiple imputation [33] (Mice). We introduce some additional missing values in three quantitative variables (*TCD.PI.max, Shock.index.ph, FiO2*) by using the MNARz mechanism (3). The variables contain initially 51%, 31%, 7% and finally 63%, 50% and 32% missing values. The algorithm for continuous data specifically designed for MNARz data for $K = 3$ classes is compared with mean imputation and multiple imputation in terms of mean squared error (MSE). Denoting by $\hat{Y} \in \mathbb{R}^{n\times d}$ the imputed dataset and $\tilde{C} \in \mathbb{R}^{n\times d}$ the indicator pattern of missing data newly introduced, the mean squared error is given by

$$\mathbb{E}[(\hat{Y} - Y) \odot \tilde{C}]_F^2 / \mathbb{E}[Y \odot \tilde{C}]_F^2,$$

where $\odot$ is the Hadamard product and $\mathbb{E}[]_F^2 = \mathbb{E}[\|.\|_F^2]$ denotes the expectation of the Frobenius norm squared. In particular, to impute missing values using our clustering algorithm, we use the conditional expectation of the missing values given the observed ones, given that the data are assumed to be Gaussian and that all the parameters of the distribution are given by our algorithm. Imputation is carried out by taking the mean over 10^4 draws. In Figure E7, our clustering algorithm, designed for the MNAR setting, gives a significantly smaller error than other methods.

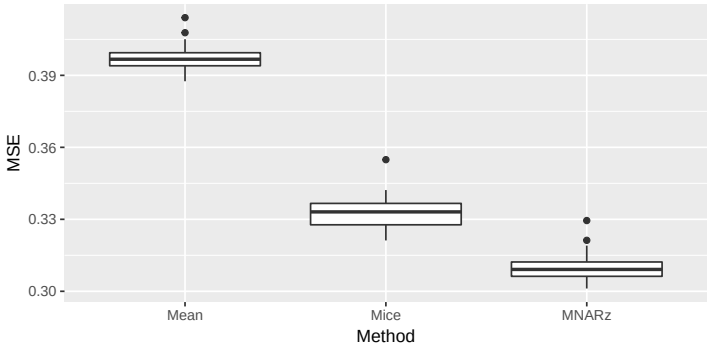


Fig. E7: Mean squared error of the imputation task for the Traumabase dataset.

E.3 Description of the variables in the Traumabase dataset

A description of the variables which are used in Section 5 is given. Figure E8 gives the percentage of missing values per variable. The indications given in parentheses ph (pre-hospital) and h (hospital) mean that the measures have been taken before the arrival at the hospital and at the hospital.

- *Trauma.center* (categorical, integers between 1 and 16, no missing values): name of the trauma center (ph & h).
- *Anticoagulant.therapy* (categorical, binary variable, 4.3% NA): oral anticoagulant therapy before the accident (ph).

- *Antiplatelet.therapy* (categorical, binary variable, 4.4% NA): anti-platelet therapy before the accident (ph).
- *GCS.init, GCS* (ordinal, integers between 3 and 15, 2% NA & 42% NA): Initial Glasgow Coma Scale (GCS) on arrival on scene of enhanced care team and on arrival at the hospital (GCS = 3: deep coma; GCS = 15: conscious and alert) (ph & h).
- *GCS.motor.init, GCS.motor* (ordinal, integers between 1 and 6, 7.6% NA & 43%): Initial Glasgow Coma Scale motor score (GCS.motor = 1: no response; GCS.motor = 6: obeys command/purposeful movement) (ph % h).
- *Pupil.anomaly.ph, Pupil.anomaly* (categorical, 3 categories: Non, Anisocoire (unilaterale), Mydriase Bilaterale, 2% NA & 1.7%): pupil dilation indicating brain herniation (ph & h).
- *Osmotherapy.ph, Osmotherapy* (categorical, 4 categories: Pas de mydriase, SSH, Mannitol, Rien, 1.7% NA and no missing values): administration of osmotherapy to alleviate compression of the brain (either Mannitol or hypertonic saline solution) (ph & h)
- *Improv.anomaly.osmo* (categorical, 3 categories: Non testé, Non, Oui, no missing values): change of pupil anomaly after ad- ministration of osmotherapy (ph).
- *Cardiac.arrest.ph* (categorical, binary variable, 2.3% NA): cardiac arrest during pre-hospital phase (ph).
- *SBP.ph, DBP.ph, HR.ph* (continuous, 29.3% NA & 29.6% NA & 29.5% NA): systolic and diastolic arterial pressure and heart rate during pre-hospital phase (ph).
- *SBP.ph.min, DBP.ph.min* (continuous, 12.8% NA & 13% NA): minimal systolic and diastolic arterial pressure during pre-hospital phase (ph).
- *HR.ph.max* (continuous, 13.7 % NA): maximal heart rate during pre-hospital phase (ph).
- *Cristalloid.volume* (continuous, positive values, 30% NA): total amount of prehospital adminis- tered cristalloid fluid resuscitation (volume expansion) (ph).
- *Colloid.volume* (continuous, positive values, 31.3% NA): total amount of prehospital administered colloid fluid resuscitation (volume expansion) (ph).
- *HemoCue.init* (continuous, 34.9% NA): prehospital capillary hemoglobin concentration (the lower, the more the patient is probably bleeding and in shock); hemoglobin is an oxygen carrier molecule in the blood (ph).
- *Delta.hemoCue* (continuous, 37.2% NA): difference of hemoglobin level between arrival at the hospital and arrival on the scene (h).
- *Vasopressor.therapy* (continuous, no missing values): treatment with catecholamines in case of physical or emotional stress increasing heart rate, blood pres- sure, breathing rate, muscle strength and mental alertness (ph).
- *SpO2.min* (continuous, 11.7% NA): peripheral oxygen saturation, measured by pulse oxymetry, to estimate oxygen content in the blood (95 to 100%: considered normal; inferior to 90% critical and associated with considerable trauma, danger and mortality) (ph).

- *TCD.PI.max* (continuous, 51.2% NA): pulsatility index (PI) measured by echodoppler sonographic examen of blood velocity in cerebral arteries (PI ¿ 1.2: indicates altered blood flow maybe due to traumatic brain injury) (h).
- *FiO2* (categorical, in $\{1, 2, 3, 4, 5\}$, 6.8% NA): inspired concentration of oxygen on ventilatory support (the higher the more critical; Ventilation = 0: no ventilatory support) (h).
- *Neurosurgery.day0* (categorical, binary variable, no missing values): neurosurgical intervention performed on day of admission (h).
- *IGS.II* (continuous, positive values, 2% NA): Simplified Acute Physiology Score (h).
- *Tranexomic.acid* (categorical, binary variable, no missing values): administration of the tranexomic acid (h).
- *TBI* (categorical, binary variable, no missing values): indicates if the patient suffers from a traumatic brain injury (h).
- *IICP* (categorical, binary variable, 70.9% NA): at least one episode of increased intracranial pressure; mainly in traumatic brain injury; usually associated with worse prognosis (h).
- *EVD* (categorical, binary variable, no missing values): external ventricular drainage (EVD); mean to drain cerebrospinal fluid to reduce intracranial pressure (h).
- *Decompressive.craniectomie* (categorical, binary variable, no missing values): surgical intervention to reduce intracranial hypertension (h).
- *Death* (categorical, binary variable, no missing values): death of the patient (h).
- *AIS.head, AIS.face* (ordinal, discrete, integers between 0 and 6 and 4 1.7% NA & 1.7% NA): Abbreviated Injury Score, describing and quantifying facial and head injuries (AIS = 0: no injury; the higher the more critical) (h).
- *AIS.external* (continuous, discrete, integers between 0 and 5, 1.7% NA): Abbreviated Injury Score for ex- ternal injuries, here it is assumed to be a proxy of information avail- able/visible during pre-hospital phase (ph/h).
- *ISS* (continuous, discrete, integers between 0 and 75, 1.6% NA): Injury Severity Score, sum of squares of top three AIS scores (h).
- *Activation.HS.procedure* (categorical, binary variable, 3.7% NA): activation of hemorragic shock procedure in case of HS suspicio (h).
- *TBI_Death* (categorical, binary variable, no missing values): death of the patients suffering from a traumatic brain injury (h).
- *TBI_Death_30d* (categorical, binary variable, no missing values): death of the patients suffering from a traumatic brain injury in the 30 days (h).
- *TBI_30d* (categorical, binary variable, no missing values): traumatic brain injury in the 30 days (h).
- *Death_30d* (categorical, binary variable, no missing values): death in the 30 days (h).
- *Shock.index.ph* (continuous, positive values, 30.5% NA): ratio of heart rate and systolic arterial pressure during pre-hospital phase (ph).
- *majorExtracranial* (categorical, binary variable, no missing values): major extracranial lesion (h).

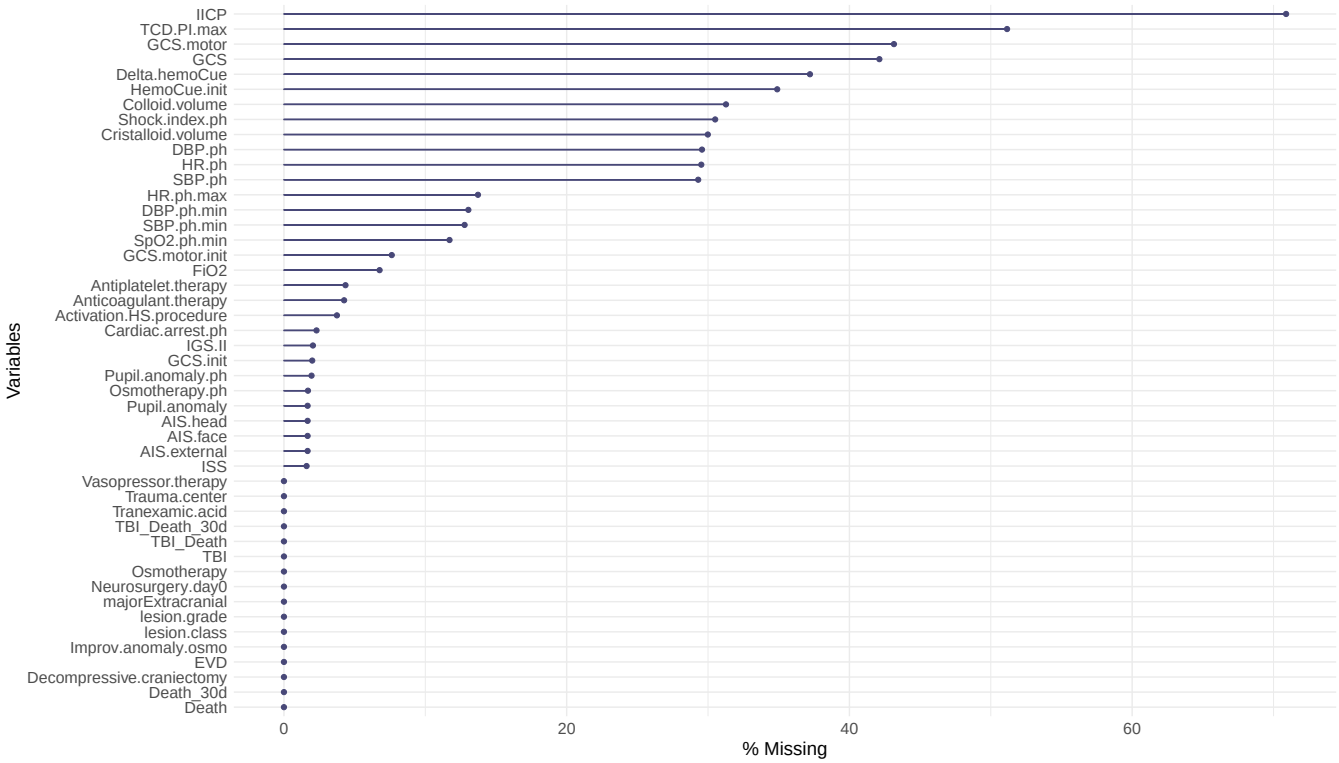


Fig. E8: Percentage of missing values per variable for the Traumabase dataset.

- *lesion.class* (no missing values): partition given by the doctors with $K = 4$ classes: axonal, extra, other, intra.
- *lesion.grade* (no missing values): partition given by the doctors with $K = 3$ classes: high, low, other.

References

[1] McLachlan, G.J., Basford, K.E.: Mixture Models: Inference and Applications to Clustering. M. Dekker New York, ??? (1988)

[2] Bouveyron, C., Celeux, G., Murphy, T.B., Raftery, A.E.: Model-based Clustering and Classification for Data Science: with Applications in R. Cambridge University Press, ??? (2019)

[3] Bouveyron, C., Girard, S., Schmid, C.: High-dimensional data clustering. Computational Statistics & Data Analysis **52**(1), 502–519 (2007)

[4] Bouveyron, C., Brunet-Saumard, C.: Model-based clustering of high-dimensional data: A review. Computational Statistics & Data Analysis **71**, 52–78 (2014)

[5] Marbac, M., Biernacki, C., Vandewalle, V.: Model-based clustering of gaussian copulas for mixed data. Communications in Statistics-Theory and Methods (2017)

[6] Ramoni, M., Sebastiani, P., Cohen, P.: Bayesian clustering by dynamics. Machine learning **47**(1), 91–121 (2002)

[7] Xiong, Y., Yeung, D.-Y.: Time series clustering with arma mixtures. Pattern Recognition **37**(8), 1675–1689 (2004)

[8] Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data, (2019)

[9] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society (1977)

[10] Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)

[11] Ibrahim, J.G., Chen, M.-H., Lipsitz, S.R.: Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. Biometrika (2001)

[12] Mohan, K., Thoemmes, F., Pearl, J.: Estimation with incomplete data: The linear case. In: IJCAI, pp. 5082–5088 (2018)

[13] Hunt, L., Jorgensen, M.: Mixture model clustering for mixed data with missing information. Computational Statistics and Data Analysis **41**, 429–440 (2003)

[14] Serafini, A., Murphy, T.B., Scrucca, L.: Handling missing data in model-based clustering. arXiv preprint (2020)

[15] Chi, J.T., Chi, E.C., Baraniuk, R.G.: k-pod: A method for k-means clustering of missing data. The American Statistician **70**(1), 91–99 (2016)

[16] Du Roy De Chaumaray, M., Marbac, M.: Clustering data with nonignorable missingness using semi-parametric mixture models. arXiv preprint (2020)

[17] Beunckens, C., Molenberghs, G., Verbeke, G., Mallinckrodt, C.: A latent-class mixture model for incomplete longitudinal gaussian data. Biometrics **64**(1), 96–105 (2008)

[18] Kuha, J., Katsikatsou, M., Moustaki, I.: Latent variable modelling with non-ignorable item nonresponse: multigroup response propensity models for cross-national analysis. Journal of the Royal Statistical Society. Series A: Statistics in Society (2018)

[19] Josse, J., Prost, N., Scornet, E., Varoquaux, G.: On the consistency of supervised learning with missing values. arXiv preprint arXiv:1902.06931 (2019)

[20] Banfield, J.D., Raftery, A.E.: Model-based gaussian and non-gaussian clustering. Biometrics, 803–821 (1993)

[21] Geweke, J., Keane, M., Runkle, D.: Alternative computational approaches to inference in the multinomial probit model. The review of economics and statistics (1994)

[22] McParland, D., Gormley, I.C.: Model based clustering for mixed data: clustmd. Advances in Data Analysis and Classification **10**(2), 155–169 (2016)

[23] Heckman, J.J.: Sample selection bias as a specification error. Econometrica (1979)

[24] Little, R.J.: Pattern-mixture models for multivariate incomplete data. JASA (1993)

[25] Chaumaray, M., Marbac, M.: Clustering data with non-ignorable missingness using semi-parametric mixture models assuming independence within components. Advances in Data Analysis and Classification, 1–42 (2023)

[26] Sportisse, A., Boyer, C., Josse, J.: Imputation and low-rank estimation with missing not at random data. Statistics and Computing **30**(6), 1629–1643 (2020)

[27] Mohan, K.: On handling self-masking and other hard missing data problems (2018)

[28] Molenberghs, G., Beunckens, C., Sotto, C., Kenward, M.G.: Every missingness not at random model has a missingness at random counterpart with equal fit. Journal of the Royal Statistical Society B **70**, 371–388 (2008)

[29] Schwarz, G.: Estimating the dimension of a model. Annals of Statistics **6**, 461–464 (1978)

[30] Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**, 719–725 (2000)

[31] Baudry, J.-P., *et al.*: Estimation and model selection for model-based clustering with the conditional classification likelihood. Electronic journal of statistics **9**(1), 1041–1077 (2015)

[32] Hubert, L., Arabie, P.: Comparing partitions. Journal of classification (1985)

[33] Buuren, S.v., Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in r. Journal of statistical software, 1–68 (2010)

[34] Biernacki, C., Deregnaucourt, T., Kubicki, V.: Model-based clustering with mixed/missing data using the new software mixtcomp. In: CMStatistics 2015 (ERCIM 2015) (2015)

[35] Lê, S., Josse, J., Husson, F.: FactoMineR: An R package for multivariate analysis. Journal of Statistical Software **25**(1), 1–18 (2008)

[36] Anderson, T.W.: An Introduction to Multivariate Statistical sAnalysis. Wiley, ??? (2003)