Model-based Clustering with Missing Not At Random Data

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

Model-based Clustering with Missing Not At Random Data

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

February 13, 2023

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 degrees of freedom of each. Eight different MNAR models which depend on the class membership and/or on the values of the missing variables themselves are proposed. For a particular type of MNAR models, for which the missingness depends on the class membership, we show that the statistical inference can be carried out on the data matrix concatenated with the missing mask considering a MAR mechanism instead; this specifically underlines the versatility of the studied MNAR models. Then, we establish sufficient conditions for identifiability of parameters of both the data distribution and the mechanism. Regardless of the type of data and the mechanism, we propose to perform clustering using EM or stochastic EM algorithms specially developed for the purpose. 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, Identifiability, EM and Stochastic EM Algorithms, Medical Data.

1 Introduction

Clustering remains a pivotal tool for readable analysis of large datasets, offering a summary of datasets by grouping observations. In particular, the model-based paradigm (McLachlan and Basford, 1988; Bouveyron et al., 2019) allows to perform clustering, by providing interpretable models that are valuable to understand the connections between the constructed clusters and the features in play. This parametric framework provides a certain plasticity by handling high-dimensionality

problems (Bouveyron et al., 2007; Bouveyron and Brunet-Saumard, 2014), mixed datasets (Marbac et al., 2017), or even time series and dependent data (Ramoni et al., 2002; Xiong and Yeung, 2004). The counterpart to performing this multi-faceted model-based clustering is the modeling work involved to design mixture models appropriate 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 (Little and Rubin, 2019), 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. Note that both strategies are only preprocessing steps, not specifically designed for the final clustering task. Alternatively, one can consider likelihood approaches, using, for example, Expectation Maximization (EM) type algorithms (Dempster et al., 1977). We adopt such an approach in this paper, to make model-based clustering handle informative missing data in an efficient way.

1.1 MNAR data

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. $\mathbf{y}_i^{\text{mis}}$). In this paper, we assume the data missing not at random (MNAR) values (Rubin, 1976; Ibrahim et al., 2001; Mohan et al., 2018), *i.e.* the missing pattern $\mathbf{c}_i$ may depend on the missing values. An example include clinical data collected in emergency situations, where doctors may choose to treat patients before measuring heart rate: the missingness of heart rate depends on the missing heart rate itself. For such a setting, the observed data are therefore not representative of the population. 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 C (see Section 2.5)

There are mainly two approaches to formulate the joint distribution of the data and the missing-data pattern: (i) the selection model (Heckman, 1979) which factorizes it into the product of the marginal data density and the conditional density of the missing-data pattern given the data; (ii) the pattern-mixture model (Little, 1993) which uses the product of the marginal density of the missing-data pattern and the conditional density of the data given the missing-data pattern. In this paper, we adopt the selection model strategy, as it is more intuitive in our setting to model the distribution of the data (as usually done in parametric clustering approaches) and the cause of the lack according to the data. Although this point of view requires

to model the missing-data mechanism, it permits imputation of the missing values and density estimation throughout the parameter estimation of the mixture model.

1.2 Related works on clustering despite missing values

In order to handle missing values in a model-based clustering framework, Hunt and Jorgensen (2003) have implemented the standard EM algorithm (Dempster et al., 1977) based on the observed likelihood. More recently, Serafini et al. (2020) also propose an EM algorithm to estimate Gaussian mixture models in the presence of missing values by performing multiple imputations (with Monte Carlo methods) in the E-step. However, both works only consider M(C)AR data, when the missing pattern $\mathbf{c}_i$ cannot depend on the missing values.

Different clustering methods have been developed to deal with MNAR mechanisms. In a partition-based framework, Chi et al. (2016) 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 (2020) have proposed to perform clustering via a semi-parametric 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. (2008); Kuha et al. (2018) jointly model the measurements and the dropout process by using an extension of the shared-parameter model, which is a specific approach to deal with MNAR mechanisms, by assuming that both the data and the dropout process depend on shared latent variables. However, the MNAR model is restricted to the case where the missingness may depend on the latent variables but not on the missing variables themselves.

For MNAR data, beyond the clustering task, specifically in selection models, the main challenge to overcome consists in proving the identifiability of the parameters of the data and the missing-data pattern distributions. In particular, Molenberghs et al. (2008) prove that identifiability does not hold when the models are not fixed, *i.e.*, when there is no prior information on the type of distribution for the missing-data pattern. For fixed models, Miao et al. (2016) provide identifiability results of Gaussian mixture and t-mixture models with MNAR data. However, their identifiability results are restricted to specific missing scenarios in a univariate case (one variable), and no estimation strategy is proposed.

1.3 Contributions

We present and illustrate a relevant inventory of distributions for the MNAR missingness process in the context of unsupervised classification based on mixture models for different types of data (continuous, count, categorical and mixed). We then provide the identifiability of the mixture model parameters and missingness process parameters under certain conditions (including the data type and the link func-

tions governing the missingness mechanism distribution). This is a real issue in the context of MNAR data, as models often lead to unidentifiable parameters. When all variables are continuous or count, all models lead to identifiable parameters. In the categorical and mixed cases, only the models for which missingness depends uniquely on the class membership have identifiable parameters. These identifiability results represent a substantial extension of the work of Miao et al. (2016) to more complex missing scenario and to the multivariate case. For each model or submodel, an EM or Stochastic EM algorithm is proposed, implemented, and made available for reproducibility[1]. We also prove that, with respect to MNAR models for which missingness depends on class membership, statistical inference can be conducted on the augmented matrix $[Y, C]$ considering a missing at random (MAR) mechanism instead, *i.e.* when the missing pattern only depends on the observed values. The latter being ignorable, this is a real advantage, as the missing-data mechanism does not have to be modeled in such a case. This also gives theoretical insights about this approach, which is actually often used in practice without any theoretical foundation: working on the augmented data matrix under a MAR assumption is usually proposed to efficiently learn despite a more complex underlying missing mechanism (Josse et al., 2019).

2 Missing data in model-based clustering

2.1 Mixture models

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 $\mathbf{z}_i$.

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

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

where $\theta = (\gamma, \psi)$ gathers all the model parameters, $\gamma = (\pi, \lambda)$ groups the parameters related to the marginal distribution of $\mathbf{y}_i$, $\pi = (\pi_1,\dots,\pi_K)$ is the vector of proportions with $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k > 0$ for all $k \in \{1,\dots,K\}$. Given $\lambda = (\lambda_1,\dots,\lambda_K)$, $f_k(\cdot\,;\lambda_k)$ is the pdf of the k-th component parameterized by λ_k, $\psi = (\psi_1,\dots,\psi_K)$ 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

[1]The code is available on https://anonymous.4open.science/r/Clustering-MNAR-7E29.

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, 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 (McLachlan and Basford, 1988; Banfield and Raftery, 1993) 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 Geweke et al. (1994); McParland and Gormley (2016)) 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 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 next subsection discusses the choice of the modeling for the missingness mechanism (*i.e.*, the distribution $f_k(\mathbf{c}_i \mid \mathbf{y}_i; \psi_k)$).

2.2 Rationale for MNAR assumptions

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 the following assumptions:

1. The elements of $\mathbf{c}_i$ are conditionally independent given $(\mathbf{y}_i, \mathbf{z}_i)$.

2. The element c_{ij} is conditionally independent given $(\mathbf{y}_i, \mathbf{z}_i)$ from $y_{ij'}$ for $j \neq j'$.

By the categorical nature of the mask $\mathbf{c}_i$, the independence assumption **1.** is a quite natural hypothesis in the context of clustering (Du Roy De Chaumaray and Marbac, 2020; Chi et al., 2016). The independence assumption **2.** of the variables amounts to considering self-masked class-wise MNAR mechanisms for each variable: the missingness of the variable j may depend on its value itself (self-masked) and on the class membership (class-wise). Note that the self-masked feature, apart from limiting the number of parameters to be estimated, is now commonly met in the literature (Mohan, 2018; Sportisse et al., 2020; Le Morvan et al., 2020), and is able to retrieve some existing ad hoc MNAR procedures already used in machine learning community (see more details in Section 2.4, Theorem 2.1).

More specifically, the conditional distribution of c_{ij} given $(\mathbf{y}_i, \mathbf{z}_i)$ is assumed to be a (classical) generalized linear model with link function ρ, so that finally

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

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. This model is called MNAR$y^k z^j$ in the following.

Some variations include $f_k(\mathbf{c}_{ij} = 1 \mid \mathbf{y}_i; \psi_k) = \rho(\alpha_{kj} + \gamma_{kj'} y_{ij'})$, where the j'-th variable is always observed. In such a case, the missing values are missing at random (MAR) as only depending on observed variables. Identifiability guarantees and estimation are still valid in such a setting.

2.3 Variant MNAR modeling

Simpler models can be derived from (3) 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 (3), that still allow the probability of being missing to depend on both the variable itself and the class membership. For the so-called MNARyz^j model, 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), so that

$$\text{MNAR}yz^j: \quad \beta_{1j} = \ldots = \beta_{Kj}, \ \forall j. \tag{4}$$

For the MNAR$y^k z$ model, 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}):

$$\text{MNAR}y^k z: \quad \alpha_{k1} = \ldots = \alpha_{kd}, \ \forall k. \tag{5}$$

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, entailing the so-called MNARyz model:

$$\text{MNAR}yz: \quad \beta_{1j} = \ldots = \beta_{Kj}, \ \forall j \quad \text{and} \quad \alpha_{k1} = \ldots = \alpha_{kd}, \ \forall k. \tag{6}$$

Secondly, the probability to be missing can also depend only on the variable itself. This is actually a particular case of MNAR mechanims, widely used in practice (Mohan, 2018), that we call MNARy here. The only effect of missingness is thus on the variable j, being the same regardless of the class membership,

$$\text{MNAR}y: \quad \alpha_{11} = \ldots = \alpha_{1d} = \alpha_{21} = \ldots = \alpha_{Kd} \text{ and } \quad \beta_{1j} = \ldots = \beta_{Kj} \quad \forall j. \tag{7}$$

A slightly more general case can be considered by allowing the effect of missingness on the variable j to depend on the class k, as in the following MNARy^k model,

$$\text{MNAR}y^k: \quad \alpha_{11} = \ldots = \alpha_{1d} = \alpha_{21} = \ldots = \alpha_{Kd}. \tag{8}$$

Thirdly, the probability to be missing can also depend only on the class membership, so that the missingness is class-wise only. In the MNARz model, we

consider that the only effect of missingness is on the class membership k, being the same for all variables,

$$\text{MNAR}z: \quad \beta_{kj} = 0, \ \forall (k,j) \text{ and } \alpha_{k1} = \ldots = \alpha_{kd}, \forall k. \tag{9}$$

The MNARz^j model is a slightly more general case than the MNARz model, because the effect of missingness on the class membership k is not the same for all the variables,

$$\text{MNAR}z^j: \quad \beta_{kj} = 0, \ \forall (k,j). \tag{10}$$

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,

$$\text{MCAR}: \quad \beta_{kj} = 0, \ \forall (k,j) \text{ and } \alpha_{1j} = \ldots = \alpha_{Kj}, \ \forall j. \tag{11}$$

2.4 Analyzing the MNARz model at the light of an augmented MAR model

The MNARz model given in (9) is one of the simplest MNAR models previously listed. Strictly speaking, MNARz does not directly involve $\mathbf{y}_i$ in its ground definition (9), but the pattern $\mathbf{c}_i$ can be related to $\mathbf{y}_i$ through $\mathbf{z}_i$ so that the proportion of missing values can vary between clusters, see Figure 6 (see Appendix A. Interestingly, MNARz and MNARz^j can be turned into a MAR-like strategy by working on the concatenated dataset $\tilde{Y}^{\text{obs}} = (Y^{\text{obs}}|C)$. This is the purpose of the next theorem, proven in Appendix A.

Theorem 2.1. *Consider the 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*

$$\tilde{f}(\tilde{\mathbf{y}}_i^{\text{obs}}; \theta) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i^{\text{obs}}; \lambda_k) \prod_{j=1}^{d} \rho(\alpha_{kj})^{c_{ij}} \left(1 - \rho(\alpha_{kj})\right)^{1-c_{ij}} . \tag{12}$$

Then for fixed parameters (π, λ, ψ), *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 (9) or MNARz^j assumption (10).*

Theorem 2.1 implies that the maximum likelihood estimate of (π, λ, ψ) is the same considering $\tilde{\mathbf{y}}_i^{\text{obs}}$ under the MAR assumption and $\mathbf{y}_i^{\text{obs}}$ under the MNARz assumption (9) or MNARz^j assumption (10). This implies that if the mechanism is MNARz or MNARz^j, 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 2.1 is the first theoretical result in unsupervised learning in line with the intuition developed in (Josse et al., 2019) for supervised learning and in (Sportisse et al., 2020) 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.

2.5 Ignorability for clustering purpose?

MNAR is strictly speaking non-ignorable for model estimation purposes (Little and Rubin, 2019) but one could check whether it is ignorable for the clustering task of interest. A necessary and sufficient condition to have an ignorable missing process for clustering is that the distributions of $\mathbf{c}_i$ are equal among the mixture components. Therefore, the missingness process is said to be *ignorable for clustering* if

$$\forall(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i),\ \mathbb{P}(z_{ik} = 1|\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta) = \mathbb{P}(z_{ik} = 1|\mathbf{y}_i^{\text{obs}}; \theta).$$

This is equivalent to having $\frac{\pi_k f_k(\mathbf{y}_i^{\text{obs}};\lambda_k) f_k(\mathbf{c}_i|\mathbf{y}_i^{\text{obs}};\psi_k)}{\sum_{\ell=1}^K \pi_\ell f_\ell(\mathbf{y}_i^{\text{obs}};\lambda_\ell) f_\ell(\mathbf{c}_i|\mathbf{y}_i^{\text{obs}};\psi_\ell)} = \frac{\pi_k f_k(\mathbf{y}_i^{\text{obs}};\lambda_k)}{\sum_{\ell=1}^K \pi_\ell f_\ell(\mathbf{y}_i^{\text{obs}};\lambda_\ell)}$, for all $(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i)$.

The MCAR mechanism is trivially ignorable for clustering since $\mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i, z_{ik} = 1; \psi) = \mathbb{P}(\mathbf{c}_i; \psi)$. However, under the MNAR assumption, the missingness mechanism is no longer ignorable, even for clustering, and a specific estimation process for the vector parameter (π, θ, ψ) is needed. Obviously, it depends on the MNAR model at hand, *i.e.*, on the missing-pattern distribution $\mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i, \mathbf{z}_i; \psi)$.

2.6 Identifiability of the model parameters

The generic identifiability (Allman et al., 2009) of parameters for continuous, count, categorical, and mixed data (*i.e.*, when the set of unidentifiable parameters has a zero Lebesgue measure) is ensured by the following theorem. The latter requires assumptions that are properly defined in Appendix B, and just mentioned here. For continuous data, Assumptions **A1.** and **A2.** require that the parameters of the marginal mixture are identifiable and that a total ordering of the mixture densities holds. For categorical data, Assumption **A4.** requires the conditional independence of the features given the group membership and Assumption **A5.** links the dimension of the observations and the logarithm of the number of clusters. Finally, in both cases, Assumption **A3.** requires that the link function of the missing data mechanism is strictly monotone, but no assumption about its form (*e.g.* logit, probit) is made.

Theorem 2.2. *Define the conditions:*

- *C1 The variables correspond to continuous or count data, **A1.** and **A2.** hold true,*
- *C2 All the variables are categorical, **A4.** and **A5.** hold true and the mechanism is stated by (9), (10) or (11),*
- *C3 At least one variable is continuous or count data and has a marginal distribution that satisfies **A1.** and **A2.**, **A4.** holds true,*
- *C4 At least one variable is categorical and its associated mechanism is stated by (9), (10) or (11), **A4.** and **A5.** hold true.*

*Assume that Assumption **A3.** holds and that at least one of conditions C1-C4 is satisfied, then the parameters of the model in (2) are generically identifiable, up to label swapping.*

The proof is given in Appendix B. In the case of continuous and count variables, the proof follows the reasoning used by Teicher (1963, Theorem 2) which proves the identifiability of univariate finite mixtures. For categorical variables, the generic identifiability holds only for the MCAR, MNARz and MNARz^j mechanisms. The idea of the proof is to rewrite the observed likelihood as the finite mixture of K multinomial distributions, for which the identifiability is given by Corollary 5 of Allman et al. (2009). For MNAR$y\star$ mechanisms, the rewriting is impossible, because of the dependency on y of the mechanism. The identifiability of mixed data directly follows from the identifiability of continuous and categorical components.

3 Estimation of the proposed MNAR models

Assuming identifiability, we estimate parameters via likelihood maximization using EM and SEM algorithms specifically designed for Gaussian, Poisson, multinomial and mixed data with MNAR data. Details of the algorithms are given in Appendix C and D. 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

$$\ell_{\text{comp}}(\theta; \mathbf{Y}, \mathbf{Z}, \mathbf{C}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \left(\pi_k f_k(\mathbf{y}_i; \lambda_k) f_k(\mathbf{c}_i \mid \mathbf{y}_i; \psi_k) \right), \tag{13}$$

giving the starting point of the aforementioned algorithms.

3.1 The EM algorithm

The EM algorithm (Dempster et al., 1977) 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 $Q(\theta; \theta^{[r-1]}) = \mathbb{E}_{\theta^{[r-1]}} \left[\ell_{\text{comp}}(\theta; \mathbf{Y}, \mathbf{Z}, \mathbf{C}) \mid \mathbf{Y}^{\text{obs}}, \mathbf{C} \right]$, then, at the M-step, updating the parameters by maximizing this function $\theta^{[r]} = \arg\max_{\theta} Q(\theta; \theta^{[r-1]})$. Note that

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

where

$$t_{ik}(\theta^{[r-1]}) = \frac{1}{f(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})} \int \pi_k^{[r-1]} f_k(\mathbf{y}_i; \lambda_k^{[r-1]}) f_k(\mathbf{c}_i \mid \mathbf{y}_i; \psi_k^{[r-1]}) d\mathbf{y}_i^{\text{miss}},$$

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

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

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

- **E-step:** Computation of

$$t_{ik}(\theta^{[r-1]}), \tau_y(\lambda_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]} \text{ and } \tau_c(\psi_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]}).$$

- **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 MNARz and MNARz^j models, for which the effect of the missingness does not depend on $\mathbf{y}_i$ (see Appendix C.1 and C.2 for computation details in the case of Gaussian or categorical data). However, these steps become a delicate issue, when the missingness depends on variables y, such models being generically denoted by MNAR$y\star$ in the sequel. The computation of τ_y in particular remains feasible in some cases (for instance when the link function in (3) is probit), while to our knowledge, τ_c and t_{ik} do not admit closed forms.

3.2 The SEM algorithm for overpassing the EM's intractability

Some distributions entail untractable integrals at the E-step (*e.g.,* Gaussian components with MNAR$y\star$ mechanism defined with logit link). Then, the stochastic EM algorithm (Celeux and Diebolt, 1985) can circumvent this issue, by imputing missing values using Gibbs sampling instead of integrating over them. In addition, it has another advantage, unlike the EM algorithm, not to be necessarily trapped by the first encountered local maximum of the likelihood function in play (Celeux and Diebolt, 1985). The principle of the SEM algorithm is to involve a stochastic-E step (SE-step) instead of the traditional E-step of the EM algorihtm. The iteration

$[r]$ then becomes:

SE-step: Draw the missing data $(\mathbf{z}_i^{[r]}, \mathbf{y}_i^{\mathrm{mis}\,[\mathrm{r}]})$ according to their conditional distribution given the observed data $(\mathbf{y}_i^{\mathrm{obs}}, \mathbf{c}_i)$ and the current parameter $\theta^{[r-1]}$. As simulating according to this conditional distribution may be difficult, we simulate instead according to the following two conditional probabilities using a Gibbs sampler, by noting $\mathbf{y}_i^{[r]} = (\mathbf{y}_i^{\mathrm{obs}}, \mathbf{y}_i^{\mathrm{mis}[r]})$,

$$\mathbf{z}_i^{[r]} \sim \mathbf{z}_i \mid \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]} \quad \text{and} \quad \mathbf{y}_i^{\mathrm{mis}\,[\mathrm{r}]} \sim \mathbf{y}_i^{\mathrm{mis}} \mid \mathbf{y}_i^{\mathrm{obs}}, \mathbf{z}_i^{[r]}, \mathbf{c}_i; \theta^{[r-1]}. \quad (14)$$

M-step: Let $Y^{[r]} = (\mathbf{y}_1^{[r]} | \ldots | \mathbf{y}_n^{[r]})$ be the imputed matrix and let $Z^{[r]} = (\mathbf{z}_1^{[r]} | \ldots | \mathbf{z}_n^{[r]})$ be the current and corresponding partition. The parameter $\theta^{[r]}$ is computed using the maximum likelihood estimate in the complete case. For all $k \in \{1, \ldots, K\}$, the parameter $\pi_k^{[r]}$ is the proportion of rows of $Y^{[r]}$ belonging to class k. The parameter $\lambda_k^{[r]}$ is updated in a standard way, depending on the parametric mixture family in play. Finally, the parameter $\psi_k^{[r]}$ is the resulting coefficients of a GLM with a binomial link function, *cf* Appendix D for details.

In the SE-step, note that the sampling of $\mathbf{z}_i^{[r]}$ is performed by a multinomial distribution. The conditional distribution of $\mathbf{y}_i^{\mathrm{mis}}$ given $(\mathbf{y}_i^{\mathrm{obs}}, z_{ik}^{[r]} = 1, \mathbf{c}_i)$ parameterized by $\theta^{[r-1]}$ is

$$\begin{aligned} &f_k(\mathbf{y}_i^{\mathrm{mis}} \mid \mathbf{y}_i^{\mathrm{obs}}, \mathbf{c}_i; \theta^{[r-1]}) \\ &\quad = \frac{f_k(\mathbf{y}_i^{\mathrm{mis}} \mid \mathbf{y}_i^{\mathrm{obs}}; \theta^{[r-1]}) f_k(\mathbf{c}_i \mid \mathbf{y}_i^{\mathrm{mis}}, \mathbf{y}_i^{\mathrm{obs}}; \psi^{[r-1]})}{\int f_k(\mathbf{y}_i^{\mathrm{mis}} \mid \mathbf{y}_i^{\mathrm{obs}}; \theta^{[r-1]}) f_k(\mathbf{c}_i \mid \mathbf{y}_i^{\mathrm{mis}}, \mathbf{y}_i^{\mathrm{obs}}; \psi^{[r-1]}) d\mathbf{y}_i^{\mathrm{mis}}}. \end{aligned}$$

This distribution may not be classical in general. For example, for MNAR$y\star$ models, it is not explicit if the components are Gaussian and if the missing data distribution ρ is logistic (since the product of logistic and Gaussian distributions is not a standard law). Therefore, the SEM algorithm cannot be easily applied. However, if ρ is the probit function, we can make the distribution of interest explicit (it is a truncated Gaussian distribution when the variables are Gaussian). For MNARz and MNARz^j models, all the computations remain feasible. Table 1 summarizes the cases for which the EM or SEM algorithm is feasible.

	EM			SEM		
	Gaussian		Categorical	Gaussian		Categorical
	Appendix C.1		Appendix C.2	Appendix D.1		Appendix C.2
MNARz MNARz^j	✓		✓	✓		✓
	Probit	Logit		Probit	Logit	
MNAR$y\star$	no closed form	no closed form, optim. pb	not identifiable	✓	require algorithms as SIR (costly)	not identifiable

Table 1: Summary of the cases for which the EM and the SEM lead to feasible (or not feasible) computations. The symbol ✓ means that the computations are feasible (derived in Appendix C).

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) (Schwarz, 1978) or the Integrated Complete-data Likelihood (ICL) (Biernacki et al., 2000). 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 (Baudry et al., 2015). 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 (Baudry et al., 2015) that we adapt with missing data. For a model $\mathcal{M}$ with $\nu_{\mathcal{M}}$ parameters, the ICL reads as

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

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

$$\text{with} \quad z_{ik}^{\mathrm{MAP}}(\theta) = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}}\, \mathbb{P}(z_{ik}=1|\mathbf{y}_i^{\mathrm{obs}}, \mathbf{c}_i; \theta). \tag{15}$$

In addition, the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) can be computed between the true partition Z and the estimated one.

4.1 Illustration on how 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 (9) with a probit link function. The difference between the centers of both mixture components is taken as $\Delta_\mu = \mu_{21} - \mu_{11} = \mu_{22} - \mu_{12} \in \{0.5, 1, \ldots, 3\}$. 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 $\{0, 0.1, 0.2, 0.3\}$[2].

Increasing values of Δ_{perc} 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 1 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) 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.

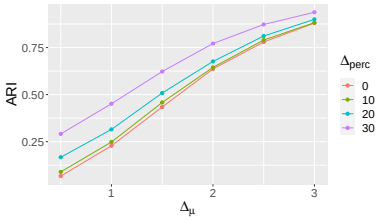


Figure 1: Relative effect of both the separation strength Δ_μ of the mixture component and the MNAR evidence Δ_{perc} on theoretical ARI (*e.g.* if $\Delta_{\text{perc}} = 10\%$ (green line), the second class has 10% more missing values).

4.2 Generic experiments

To perform clustering with missing data, we consider the following methods:

- the EM algorithm (for MCAR (11), MNARz (9) and MNARz^j (10) settings),
- the SEM algorithm (for MNARy (7), MNARy^k (8), MNAR$y^k z^j$ (3), MNARyz (6)),
- a two-step heuristics (denoted by `Mice` in the following) which consists of first imputing the missing values using multiple imputations by chained equations (Buuren and Groothuis-Oudshoorn, 2010) to get M completed datasets. Then, classical model-based clustering is performed on each completed dataset, for which the ARI is computed[3] and averaged.

[2] 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 $\text{perc}_2 = (\text{perc}_1 + \Delta_{\text{perc}})$.

[3] In simulations, we do not systematically consider this method, not specifically designed for clustering.

To compare the methods presented above, we consider a Gaussian mixture with three components having unequal proportions ($\pi_1 = 0.5$, $\pi_2 = \pi_3 = 0.25$) and independent variables:

$$\forall j \in \{1, \ldots, d\}, y_{ij} = \delta \sum_{k=1}^{3} \varphi_{kj} z_{ik} + \epsilon_{ij}, \tag{16}$$

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 (3), 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 G. All the simulations have been performed for a theoretical rate of misclassification of 10%[4] and a theoretical missing rate in the whole dataset of 30%.

For each MNAR setting, we assess the clustering performance through the consistency of the partition, by computing the ARI between the true partition Z and the estimated one, given by $\hat{Z}^{\text{MAP}} = \{z_{ik}^{\text{MAP}}(\hat{\theta})\}_{i,k} \in \mathbb{R}^{n \times K}$ in (15). We consider $d = 6$ variables and we vary the number of observations $n = 100, 250, 500$. In Figure 2, as expected, considering the mechanism always gives better results than using the MCAR model, especially for models with many parameters and larger sample sizes (as the MNARyz, MNAR$y^k z^j$, MNAR$y^k z$, MNARyz^j settings for $n = 250$ and $n = 500$). Finally, consistency seems satisfactory in each scenario, indicating that our tuning parameters for the algorithm (starting values, stopping rules) are quite suitable. The computation times for these numerical experiments (see Figure 7 and 8 in Appendix E) differ a lot between settings using the EM algorithm (MCAR, MNARz, MNARz^j) and the ones requiring the SEM algorithm (MNAR$y\star$), the latter being expansively time-consuming.

We also illustrate in Appendix E (Figure 8) the findings of Theorem 2.1, by comparing the EM algorithm coded by us considering MCAR or MNARz data with the SEM algorithm of the **RMixtComp** package (Biernacki et al., 2015) considering MCAR data and using the augmented data matrix $(Y|C)$. As expected, both approaches give similar results.

To evaluate the impact of the dimension, we 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`. Figure 3 shows the boxplot of the ARI for each scenario. First, the methods that consider

[4] In Appendix E, we also provide the experiments for a theoretical rate of misclassification of 15%. Same conclusions hold.

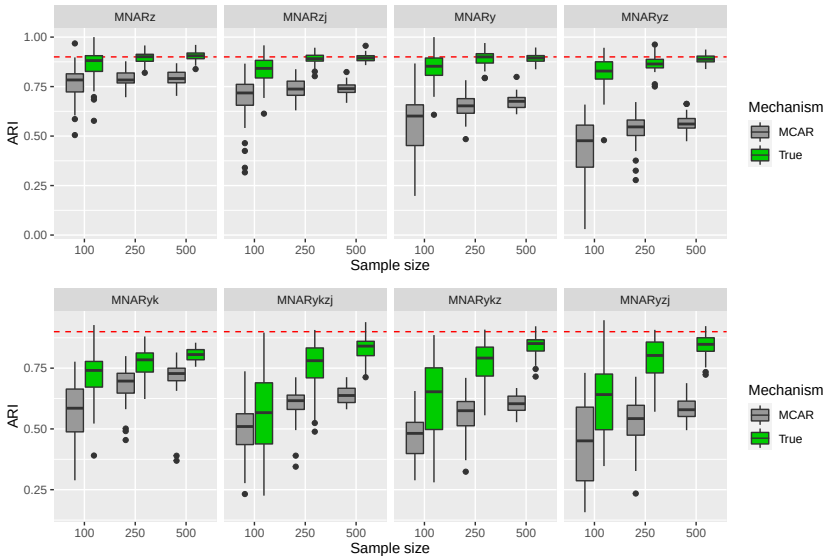


Figure 2: Boxplot of the ARI obtained for 50 samples composed of $d = 6$ variables. The sample size varies by $\{100,250,500\}$. The boxplots in green (True) correspond to the performance of the algorithm considering the MNAR setting matching the one that has been used for the missing value generation. The red dashed line indicates the theoretical ARI.

a MNAR mechanism (MNAR$\star$) always outperform those that consider the MCAR mechanism and the two-step procedure based on `Mice`. Note that comparing the MNARz setting with the real MNAR setting that generated the missing data is difficult, because it is not clear how much the MNARz setting deviates from the hypothesis (depending on the parameters chosen for the mechanism). However, the MNARz model remains 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.

4.3 Focus on the MNARz mechanism

Considering the same setting as in Section 4.2 and under a MNARz mechanism, we evaluate the impact of misspecification of the link function (Figure 4(a)), the misspecification of 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), the SEM 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

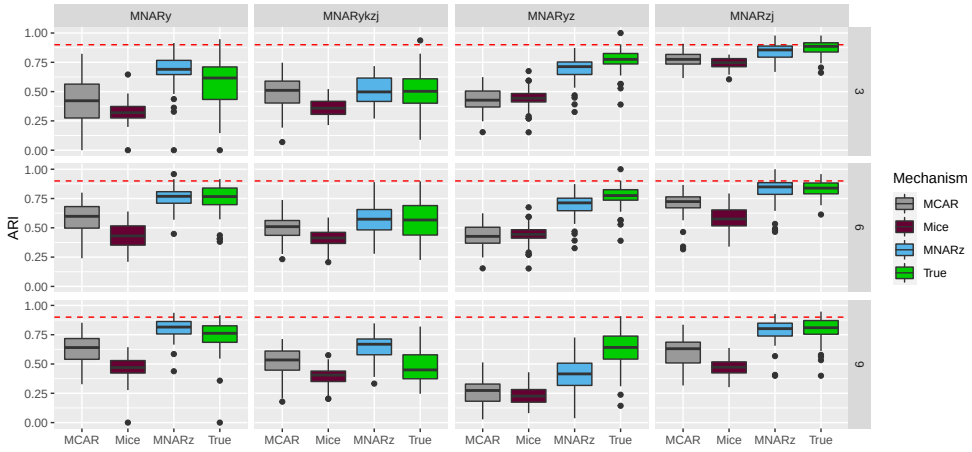


Figure 3: Boxplot of the ARI obtained for 50 samples composed of $d = 3, 6, 9$ variables (rows) and $n = 100$ observations. Missing values are introduced with MNARy, MNAR$y^k z^j$ MNARyz or MNARz^j settings (columns). The boxplot in green is the one for the algorithm considering the true MNAR$\star$ 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.

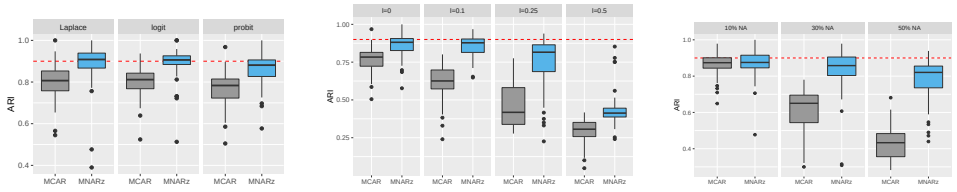
mixture with non-diagonal covariance matrices. For each component, the diagonal terms of the covariance matrix are $\Sigma_{ii} = 1$ and the other terms $\Sigma_{ij} = \ell, i \neq j$, with $\ell \in \{0, 0.1, 0.25, 0.5\}$, while the algorithms assume $\ell = 0$. If the EM algorithm designed for MNARz 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%).

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.

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

Therefore, we only compare the EM algorithm designed for the MNARz and MCAR data. Table 2 gathers the percentages



(a) (b) (c)

Figure 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.

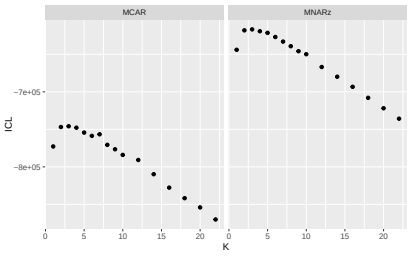
of times (over 50 repetitions) the correct number of classes ($K = 3$) is chosen by the ICL criterion for different missing-data rates (10, 30, 50%) and different sample sizes ($n = 100, 500$). In any case, the EM algorithm for MNARz data always outperforms the algorithm for MCAR data in terms of accurate 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 ($n = 500$).

5 Real medical 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 F. 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 F.

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 an MNAR mechanism can improve the classification.

We compare our algorithm designed for the MNARz data (9) and the MCAR data (11). 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 higher than that of the algorithm for MCAR data. Their corresponding ARI between classi-



(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 F).

Figure 5: Results on Traumabase dataset.

fications 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.

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 and SEM algorithms. This work also includes an exhaustive catalog of possible MNAR specifications. The identifiability study showed that the most general models lead to non-identifiable parameters for categorical data. This combined with the numerical experiments leads us to recommend using algorithms considering simple missing-data mechanisms, as 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 $[\mathbf{Y}|\mathbf{C}]$, including the missing-data pattern $\mathbf{C}$ (Theorem 2.1). 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. Therefore, without any insight on the MNAR type, we highly recommend to use the MNARz mechanism, its versatility having already been outlined.

References

E. S Allman, C. Matias, J. A Rhodes, et al. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37 (6A):3099–3132, 2009.

T. W. Anderson. *An Introduction to Multivariate Statistical sAnalysis*. Wiley, 2003.

J. D Banfield and A. E Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821, 1993.

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

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

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

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

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

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

C. Bouveyron, G. Celeux, T B. Murphy, and A. E Raftery. *Model-based clustering and classification for data science: with applications in R*. Cambridge University Press, 2019.

S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

G. Celeux and J. Diebolt. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 1985.

Jocelyn T Chi, Eric C Chi, and Richard G Baraniuk. k-pod: A method for k-means clustering of missing data. *The American Statistician*, 70(1):91–99, 2016.

A. P Dempster, N. M Laird, and D. B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 1977.

M. Du Roy De Chaumaray and M. Marbac. Clustering data with nonignorable missingness using semi-parametric mixture models. *arXiv preprint*, 2020.

J. Geweke, M. Keane, and D. Runkle. Alternative computational approaches to inference in the multinomial probit model. *The review of economics and statistics*, 1994.

J. J Heckman. Sample selection bias as a specification error. *Econometrica*, 1979.

L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 1985.

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

J. G Ibrahim, M.-H. Chen, and S. R Lipsitz. Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, 2001.

J. Josse, N. Prost, E. Scornet, and G. Varoquaux. On the consistency of supervised learning with missing values. *arXiv preprint arXiv:1902.06931*, 2019.

J. Kuha, M. Katsikatsou, and I. Moustaki. 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.

Marine Le Morvan, Julie Josse, Thomas Moreau, Erwan Scornet, and Gaël Varoquaux. Neumiss networks: differentiable programming for supervised learning with missing values. *Advances in Neural Information Processing Systems*, 33: 5980–5990, 2020.

R. JA Little. Pattern-mixture models for multivariate incomplete data. *JASA*, 1993.

R. JA Little and D. B Rubin. *Statistical analysis with missing data*. 2019.

M. Marbac, C. Biernacki, and V. Vandewalle. Model-based clustering of gaussian copulas for mixed data. *Communications in Statistics-Theory and Methods*, 2017.

G. J McLachlan and K. E Basford. *Mixture models: Inference and applications to clustering*. M. Dekker New York, 1988.

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

W. Miao, P. Ding, and Z. Geng. Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 2016.

K. Mohan. On handling self-masking and other hard missing data problems. 2018.

K. Mohan, F. Thoemmes, and J. Pearl. Estimation with incomplete data: The linear case. In *IJCAI*, pages 5082–5088, 2018.

G. Molenberghs, C. Beunckens, C. Sotto, and M. G. Kenward. 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.

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

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

A. Serafini, T. B. Murphy, and L. Scrucca. Handling missing data in model-based clustering. *arXiv preprint*, 2020.

Aude Sportisse, Claire Boyer, and Julie Josse. Imputation and low-rank estimation with missing not at random data. *Statistics and Computing*, 30(6):1629–1643, 2020.

H. Teicher. Identifiability of finite mixtures. *The annals of Mathematical statistics*, 1963.

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

S. J Yakowitz and J. D Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, pages 209–214, 1968.

Supplementary material

This file is a supplementary material. In Appendix A, Proposition 2.1 is proved. In Appendix B, the proof for Proposition 2.2 is given. The EM and SEM algorithms presented in Section 3 are detailed in Appendix C and D. In Appendix E, we add some numerical experiments on synthetic data. Appendix F gives more information on the variables of the real medical dataset. Appendix G gives the values of hyperparameters for the numerical experiments on synthetic data.

A Appendix 1: Proof of Proposition 2.1

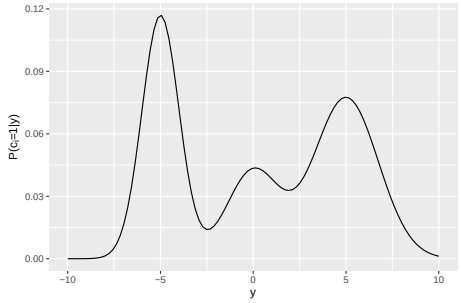


Figure 6: Illustration of the dependency between $\mathbf{c}_i$ and $\mathbf{y}_i$ in a MNARz model by drawing $\mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i; \pi, \lambda, \psi)$ for a three-component univariate Gaussian model with mixing proportions $\pi_1 = \pi_2 = 0.3$ and $\pi_3 = 0.4$, with centers $\mu_1 = \mu_3 = -5$ 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 $\alpha_3 = 1$.

Proof of Proposition 2.1. 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

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

$$L(\pi, \lambda, \psi, \zeta; \tilde{\mathbf{y}}_i^{\text{obs}}, \tilde{\mathbf{c}}_i) = L(\zeta; \tilde{\mathbf{c}}_i \mid \tilde{\mathbf{y}}_i^{\text{obs}}) \times L(\pi, \lambda, \psi; \tilde{\mathbf{y}}_i^{\text{obs}}). \tag{17}$$

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 (12), the observed likelihood $L(\pi, \lambda, \psi; \tilde{\mathbf{y}}_i^{\text{obs}})$ is

$$\begin{aligned} L(\pi, \lambda, \psi; \tilde{\mathbf{y}}_i^{\text{obs}}) &= \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i^{\text{obs}}; \lambda_k) \prod_{j=1}^{d} \rho(\alpha_{kj})^{c_{ij}} \rho(\alpha_{kj})^{(1-c_{ij})} &(18)\\ &= \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i^{\text{obs}}; \lambda_k) \prod_{j=1}^{d} f(c_{ij} \mid z_{ik} = 1; \psi). &(19) \end{aligned}$$

As $\mathbb{P}(c_{ij} \mid z_{ik} = 1; \psi)$ corresponds to the MNARz definition (9), 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,

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

□

B Appendix 2: Identifiability

B.1 Continuous and count data

A1. The parameters (π, λ) of the marginal mixture defined by the density $\sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i; \lambda_k)$ are identifiable;

A2. There exists a total ordering $\preceq$ of $\mathcal{F}_j \times \mathcal{R}$, for $j \in \{1, \ldots, d\}$ fixed, where $\mathcal{F}_j$ is the family of the data densities $\{f_{1j}, \ldots, f_{Kj}\}$ and $\mathcal{R}$ is the family of the mechanism densities $\{\rho_1, \ldots, \rho_K\} = \{\rho(.; \psi_1), \ldots, \rho(.; \psi_K)\}$, where ρ is the cumulative distribution function of any continuous distribution function and $(\psi_k)_{k \in \{1,\ldots,K\}}$ its parameter. The total ordering is such that $\forall k < \ell \in \{1, \ldots, K\}$, $\forall j \in \{1, \ldots, d\}$, $F_{kj} \preceq F_{\ell j}$ (denoting $F_{kj} = \rho_k f_{kj}$ and $F_{\ell j} = \rho_\ell f_{\ell j}$) implies $\lim_{u \to +\infty} \frac{\rho_\ell(u) f_{\ell j}(u)}{\rho_k(u) f_{kj}(u)} = 0$;

A3. The missing-data distribution ρ is assumed to be strictly monotone.

Assumption **A1.** means that the identifiability of the parameters (π, λ, ψ) of the model (2) requires the identifiability of the parameters (π, λ) of the marginal mixture of (Y, Z) (*i.e.*, considering the case without missing values). Some authors have already studied the identifiability of the mixture models, when no missing values in Y occur, especially Teicher (1963) for Gaussian mixtures (continuous variables) and Yakowitz and Spragins (1968) for Poisson mixtures (count variables). Assumption **A2.** is the core ingredient to prove the identifiability of the parameters and we illustrate it by considering concrete examples in the following. Note that under Assumption **A3.** the probit and the logistic functions may be considered, which are the most widely used for MNAR specifications.

Proof of Proposition 2.2, continuous case. Suppose there exists two sets of parameters $\{\gamma, \psi\}$ and $\{\gamma', \psi'\}$ which have the same observed distribution, *i.e.*, $f(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \gamma, \psi) = f(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \gamma', \psi')$. More precisely, one has

$$\begin{aligned}\forall \mathbf{y}_i \in \mathbb{R}^d, \forall \mathbf{c}_i \in \{0,1\}^d, \sum_{k=1}^{K} \int_{\mathcal{Y}_i^{\text{mis}}} \pi_k f_k(\mathbf{y}_i; \lambda_k) \prod_{j=1}^{d} \rho(\alpha_{kj} + \beta_{kj} y_{ij})^{c_{ij}} [1 - \rho(\alpha_{kj} + \beta_{kj} y_{ij})]^{1-c_{ij}} dy^{\text{mis}} \\ = \sum_{k=1}^{K'} \int_{\mathcal{Y}_i^{\text{mis}}} \pi'_k f_k(\mathbf{y}_i; \lambda'_k) \prod_{j=1}^{d} \rho((\alpha')_{kj} + (\beta')_{kj} y_{ij})^{c_{ij}} [1 - \rho((\alpha')_{kj} + (\beta')_{kj} y_{ij})]^{1-c_{ij}} dy^{\text{mis}}\end{aligned}$$

Let us consider the case when $c_{ij} = 0$ for all $j = 1, \ldots, d$. One has

$$\sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i; \lambda_k) \prod_{j=1}^{d} (1 - \rho(\alpha_{kj} + \beta_{kj} y_{ij})) = \sum_{k=1}^{K'} \pi'_k f_k(\mathbf{y}_i; \lambda'_k) \prod_{j=1}^{d} (1 - \rho((\alpha')_{kj} + (\beta')_{kj} y_{ij})).$$

By using the identifiability of the marginal mixture, one obtains $\lambda_k = \lambda'_k$ and $\pi_k \prod_{j=1}^{d} (1 - \rho(\alpha_{kj} + \beta_{kj} y_{ij})) = \pi'_k \prod_{j=1}^{d} (1 - \rho((\alpha')_{kj} + (\beta')_{kj} y_{ij}))$.

In the sequel, we use the same reasoning of Theorem 2 in (Teicher, 1963). Let us denote $F_k(y_{ij}) = f_{kj}(y_{ij}; \lambda_{kj}) \prod_{j=1}^{d} (1 - \rho(\alpha_{kj} + \beta_{kj} y_{ij}))$ and $F'_k(y_{ij}) = f_{kj}(y_{ij}; \lambda_{kj}) \prod_{j=1}^{d} (1 - \rho((\alpha')_{kj} + (\beta')_{kj} y_{ij}))$. Without loss of generality, assume that $F_k \prec F_l$ and $F'_k \prec F'_l$ for $k < l$. If $F_1 \neq F'_1$, we assume also without loss of generality that $F_1 \preceq F'_1$. Then, $F_1 \prec F'_k$ for $1 \leq k \leq K'$. For $u \in T_1$, where $T_1 = S_{F_1} \cap \{u : F_1(u) \neq 0\}$ is the domain of definition of F_1 such that $f_{1j}(u; \lambda_{1j}) \prod_{j=1}^{d} (1 - \rho(\alpha_{1j} + \beta_{1j} u)) \neq 0$, one has

$$\pi_1 + \sum_{k=1}^{K} \pi_k \frac{F_k(u)}{F_1(u)} = \sum_{k=1}^{K'} \pi'_k \frac{F'_k(u)}{F_1(u)}.$$

Letting $u \to +\infty$, $\pi_1 = 0$ which is in contradiction with the mixture model (where $\pi_k > 0$, $\forall k = 1, \ldots, K$. It implies that $F_1 = F'_1$. For any $u \in T_1$, one has

$$\pi_1 + \sum_{k=2}^{K} \pi_k \frac{F_k(u)}{F_1(u)} = \pi'_1 + \sum_{k=2}^{K'} \pi'_k \frac{F'_k(u)}{F_1(u)}.$$

Letting again $u \to +\infty$, one obtains $\pi_1 = \pi'_1$ and $\sum_{k=2}^{K} \pi_k \frac{F_k(u)}{F_1(u)} = \sum_{k=2}^{K'} \pi'_k \frac{F'_k(u)}{F_1(u)}$. We repeat this argument to conclude that $F_k = F'_k$ and $\pi_k = \pi'_k$ for $k = 1, \ldots, \min\{K, K'\}$. Finally, if $K \neq K'$, say $K > K'$, $\sum_{k=K'+1}^{K} \pi_k F_k(u) = 0$ implies $\pi_k = 0$ for $K' + 1 \leq k \leq K$ which is in contradiction with the definition of the mixture model. Thus $K = K'$.

Finally, $F_k = F'_k$ implies that $\prod_{j=1}^{d} (1 - \rho(\alpha_{kj} + \beta_{kj} y_{ij})) = \prod_{j=1}^{d} (1 - \rho((\alpha')_{kj} + (\beta')_{kj} y_{ij}))$. By integrating out over all the elements but the j-th element, one has for all $y_{ij} \in \mathbb{R}$, $\rho(\alpha_{kj} + \beta_{kj} y_{ij}) = \rho((\alpha')_{kj} + (\beta')_{kj} y_{ij})$. and $\alpha_{kj} = (\alpha')_{kj}$ and $\beta_{kj} = (\beta')_{kj}$, since ρ is an injective function. Indeed, ρ is assumed to be strictly monotone. □

B.2 On identifiability of the Gaussian mixture

Finite Gaussian mixtures are identifiable and, for any variable j, there is a total ordering defined by $\sigma_{kj}^2 > \sigma_{(k+1)j}^2$ and $\mu_{kj} > \mu_{(k+1)j}$ if $\sigma_{kj}^2 = \sigma_{(k+1)j}^2$, where μ_{kj} and σ_{kj}^2 are respectively the mean and the variance of variable j under component k. Example B.1 shows that the identifiability holds for Gaussian mixtures when there are missing values and that the distribution of the MNAR mechanism is a probit one.

Example B.1 (Gaussian + Probit)**.** *Let us consider that ρ is the probit function and f_k (respectively f_{k+1}) the Gaussian density with parameters (μ_k, σ_k) (respectively $(\mu_{k+1}, \sigma_{k+1})$). Suppose without loss of generality that $\beta_k \geq \beta_{k+1}$. One want to prove that*

$$\lim_{u\to+\infty} E_u := \lim_{u\to+\infty} \frac{\int_{-\infty}^{\alpha_{k+1}+\beta_{k+1}u} e^{-t^2/2}dt}{\int_{-\infty}^{\alpha_k+\beta_k u} e^{-t^2/2}dt} \frac{\sigma_k \exp -\frac{(u-\mu_{k+1})^2}{2\sigma_{k+1}^2}}{\sigma_{k+1}\exp -\frac{(u-\mu_k)^2}{2\sigma_k^2}} = 0$$

Let us denote $\phi(u) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u} e^{-t^2/2}dt$. One has

$$\lim_{u\to+\infty} \phi(u) = \begin{cases} 1 \text{ if } u > 0 \\ 1/2 \text{ if } u = 0 \\ 0 \text{ if } u < 0 \end{cases} \tag{20}$$

- *If $\beta_{k+1} > 0$ (and $\beta_k > 0$):*

$$\lim_{u\to+\infty} E_u = \lim_{u\to+\infty} \exp -\left(u^2\left(\frac{1}{2\sigma_{k+1}^2} - \frac{1}{2\sigma_k^2}\right) + u\left(\frac{\mu_k}{\sigma_k} - \frac{\mu_{k+1}}{\sigma_{k+1}}\right)\right) = 0.$$

  assuming without loss of generality that $\sigma_k^2 > \sigma_{k+1}^2$ or $\mu_k > \mu_{k+1}$ if $\sigma_k^2 = \sigma_{k+1}^2$.
- *If $\beta_{k+1} \leq 0$ (and $\beta_k \geq 0$):*

$$\lim_{u\to+\infty} E_u = 0$$

  since

$$\lim_{u\to+\infty} \exp -\left(u^2\left(\frac{1}{2\sigma_{k+1}^2} - \frac{1}{2\sigma_k^2}\right) + u\left(\frac{\mu_k}{\sigma_k} - \frac{\mu_{k+1}}{\sigma_{k+1}}\right)\right) = 0$$

  and

$$\lim_{u\to+\infty} \frac{\int_{-\infty}^{\alpha_{k+1}+\beta_{k+1}u} e^{-t^2/2}dt}{\int_{-\infty}^{\alpha_k+\beta_k u} e^{-t^2/2}dt} = \begin{cases} 0 \text{ if } \beta_{k+1} < 0 \\ 1/2 \text{ if } \beta_{k+1} = 0 \text{ and } \beta_k > 0 \\ 1 \text{ if } \beta_{k+1} = 0 \text{ and } \beta_k = 0. \end{cases} \tag{21}$$

- *If $\beta_{k+1} < 0$ and $\beta_k < 0$: One uses the upper and lower bounds for the probit function.*

$$\frac{1}{-t+\sqrt{t^2+4}} < \sqrt{\frac{\pi}{2}}\exp\frac{t^2}{2}\phi(t) < \frac{1}{-t+\sqrt{t^2+8/\pi}},$$

  i.e., $\phi(t) < \sqrt{\frac{2}{\pi}}\frac{1}{-t+\sqrt{t^2+8/\pi}}\exp -\frac{t^2}{2}$ and $\frac{1}{\phi(t)} < (-t+\sqrt{t^2+4})\sqrt{\frac{\pi}{2}}\exp\frac{t^2}{2}$ Thus, noting that $\lim_{u\to+\infty}\phi(\alpha_{k+1} + \beta_{k+1}u) = \lim_{u\to+\infty}\phi(\beta_{k+1}u)$,

$$\frac{\int_{-\infty}^{\alpha_{k+1}+\beta_{k+1}u} e^{-t^2/2}dt}{\int_{-\infty}^{\alpha_k+\beta_k u} e^{-t^2/2}dt} \underset{u\to+\infty}{=} \frac{\phi(\beta_{k+1}u)}{\phi(\beta_k u)} \underset{u\to+\infty}{<} \exp\left(\left(\frac{\beta_k^2}{2} - \frac{\beta_{k+1}^2}{2}\right)u^2\right). \tag{22}$$

  As $\beta_{k+1} \leq \beta_k < 0$, one has $\beta_k^2/2 - \beta_{k+1}^2/2 < 0$ and it implies

$$\lim_{u\to+\infty} \frac{\int_{-\infty}^{\alpha_{k+1}+\beta_{k+1}u} e^{-t^2/2}dt}{\int_{-\infty}^{\alpha_k+\beta_k u} e^{-t^2/2}dt} = 0.$$

Given that

$$\lim_{u\to+\infty} \exp - \left(u^2 \left(\frac{1}{2\sigma_{k+1}^2} - \frac{1}{2\sigma_k^2} \right) + u \left(\frac{\mu_k}{\sigma_k} - \frac{\mu_{k+1}}{\sigma_{k+1}} \right) \right) = 0,$$

assuming without loss of generality that $\sigma_k^2 > \sigma_{k+1}^2$ *or* $\mu_k > \mu_{k+1}$ *if* $\sigma_k^2 = \sigma_{k+1}^2$*, one has*

$$\lim_{u\to+\infty} E_u = 0.$$

This result has been already stated, in the case of univariate distributions, by Miao et al. (2016). In particular, the identifiability conditions in Miao et al. (2016) (conditions 1 and 2 of their paper) imply the existence of the total ordering defined in Assumption **A2.**. However, these conditions exclude the case of Gaussian mixture with a logistic missing-data distribution, which is very used in practice. In Corollary B.2, we therefore extend this result to the multivariate case with a logistic missing-data distribution.

Note first that with a logistic distribution, a total ordering cannot be defined. Indeed, for variable j, such an ordering cannot be defined if the two univariate variances are equal (*i.e.,* $\sigma_{kj}^2 = \sigma_{(k+1)j}^2$) and $\mu_{kj} - \beta_{kj} - \mu_{(k+1)j} + \beta_{(k+1)j} = 0$. However, for the specific case of Gaussian mixture where all the univariate variances are different between the components, then conditions of Proposition 2.2 hold true with a logistic missing-data distribution and so does its identifiability. In addition, for more parsimonious MNAR models for which the effect on the variable j does not depend on the class membership k (*i.e.*, $\beta_{kj} = \beta_{(k+1)j}$), the conditions of Proposition 2.2 hold true with a logistic missing-data distribution. Finally, as stated by Corollary B.2 below, the condition on the covariance matrices (including the case of homoscedastic Gaussian mixture) can be relaxed to obtain the generic identifiability of the model (*i.e.,* all not-identifiable parameter choices lie within a proper submanifold, and thus form a set of Lebesgue zero measure; Allman et al. (2009)).

Corollary B.2. *Assume that* $\sum_{k=1}^{K} \pi_k f_k(y_i; \lambda_k)$ *is a multivariate Gaussian mixture,* ρ *is the logistic function and that the missingness scenario is defined by* (3)*,* (5) *or* (8)*, then, the parameters* (π, λ, ψ) *of the model given by* (2) *are generically identifiable up to label swapping, i.e., all not-identifiable parameter choices lie within a proper submanifold, and thus form a set of Lebesgue zero measure.*

For the other MNAR models given in (4)*,* (6)*,* (7)*,* (9) *and* (10)*, the parameters* (π, λ, ψ) *of the model given by* (2) *are identifiable up to label swapping.*

Proof of Corollary B.2. We use Proposition 2.2. We fix j. By abuse of notation, α_k, β_k, μ_k and σ_k correspond to the parameters α_{kj}, β_{kj}, μ_{kj} and Σ_{kj} of the variable j. Let us first consider the missing scenarios (3), (5) and (8) for which $\beta_k \neq \beta_{k+1}$. To obtain the total ordering, we need to prove that

$$\lim_{u\to+\infty} E_u = \frac{(1 + e^{-\alpha_k - \beta_k u}) e^{-\frac{(u-\mu_{k+1})^2}{2\sigma_{k+1}^2}}}{(1 + e^{-\alpha_{k+1} - \beta_{k+1} u}) e^{-\frac{(u-\mu_k)^2}{2\sigma_k^2}}} \frac{\sigma_k}{\sigma_{k+1}} = 0.$$

- If $\sigma_k^2 > \sigma_{k+1}^2$, $\lim_{u\to+\infty} E_u = \lim_{u\to+\infty} \exp -\frac{1}{2}(\frac{1}{\sigma_{k+1}^2} - \frac{1}{\sigma_k^2})u^2 = 0$.
- If $\sigma_k^2 = \sigma_{k+1}^2$, one has $\lim_{u\to+\infty} E_u = \lim_{u\to+\infty} \exp\left((\mu_k - \beta_k) - (\mu_{k+1} - \beta_{k+1})\right)u = 0$ discarding the case where $(\mu_k - \beta_k) - (\mu_{k+1} - \beta_{k+1}) = 0$ and assuming without loss of generality that $(\mu_k - \beta_k) > (\mu_{k+1} - \beta_{k+1})$. The set of nonidenfiable parameters is $\{\mu_k, \beta_k, \mu_{k+1}, \beta_{k+1} \text{ s.t.} (\mu_k - \beta_k) - (\mu_{k+1} - \beta_{k+1}) = 0\}_{k=1,\dots,K}$ and is of Lebesque zero measure.

Finally, for the missing scenarios (9) and (10), note that $\beta_k = \beta_{k+1} = 0$. For the missing scenarios (4), (6) and (7), one has $\beta_k = \beta_{k+1}$. Following the same reasoning as above, in the case where $\sigma_{k+1}^2 = \sigma_{k+1}^2$, one obtains the set of nonindentifiable parameters such that $\mu_k = \mu_{k+1}$, which is empty since $\mu_k \neq \mu_{k+1}$ if $\sigma_k^2 = \sigma_{k+1}^2$. □

B.3 On identifiability of the Poisson mixture

Proposition B.2 can also be applied for variables with integer value (*i.e.*, count data), as shown below in Examples B.3 and B.4 for the Poisson mixture with probit or logistic missing-data distributions.

Example B.3 (Poisson + Probit)**.** *Considering that ρ is the probit function and f_k (respectively f_{k+1}) the Poisson distribution with parameters λ_k (respectively λ_{k+1}). Suppose without loss of generality that $\beta_k > \beta_{k+1}$ and $\lambda_k > \lambda_{k+1}$. One want to prove*

$$\lim_{u\to+\infty} E_u := \lim_{u\to+\infty} \frac{\int_{-\infty}^{\alpha_{k+1}+\beta_{k+1}u} e^{-t^2/2}dt}{\int_{-\infty}^{\alpha_k+\beta_k u} e^{-t^2/2}dt} \frac{\lambda_{k+1}^u e^{-\lambda_{k+1}}}{\lambda_k^u e^{-\lambda_k}} = 0.$$

- *If $\beta_{k+1} > 0$ (and $\beta_k > 0$): using (20), one has*

$$\lim_{u\to+\infty} E_u = \lim_{u\to+\infty} \exp u \log \frac{\lambda_{k+1}}{\lambda_k} = 0.$$

- *If $\beta_{k+1} \leq 0$ (and $\beta_k \geq 0$): one has*

$$\lim_{u\to+\infty} E_u = 0.$$

  using

$$\lim_{u\to+\infty} \exp u \log \frac{\lambda_{k+1}}{\lambda_k} = 0$$

  and (21) for the missing distribution part.

- *If $\beta_{k+1} < 0$ and $\beta_k < 0$: using (22), one obtains*

$$\lim_{u\to+\infty} E_u < \lim_{u\to+\infty} \exp\left(\left(\frac{\beta_k^2}{2} - \frac{\beta_{k+1}^2}{2}\right) u^2\right) \exp u \log \frac{\lambda_{k+1}}{\lambda_k} = 0,$$

  because $\beta_k^2/2 - \beta_{k+1}^2/2 < 0$.

Example B.4 (Poisson + Logistic)**.** *Considering that ρ is the logistic function and f_k (respectively f_{k+1}) the Poisson distribution with parameters λ_k (respectively λ_{k+1}). One want to prove that*

$$\lim_{u\to+\infty} E_u = \lim_{u\to+\infty} \frac{1 + e^{-\alpha_k - \beta_k u}}{1 + e^{-\alpha_{k+1} - \beta_{k+1} u}} \exp u \log \frac{\lambda_{k+1}}{\lambda_k} = 0.$$

Assume that $\lambda_k > \lambda_{k+1}$ without loss of generality.

- *For the missing scenarios (3), (5) and (8) for which $\beta_k \neq \beta_{k+1}$, one obtains the generic identifiability where the set of non-identifiable parameters is $\{\alpha_k, \beta_k, \lambda_k \text{ s.t.} (\log \lambda_k - \beta_k) - (\log \lambda_{k+1} - \beta_{k+1}) = 0\}_{k=1,\dots,K}$ and is of Lebesque zero measure.*
- *For the missing scenarios (9) and (10), note that $\beta_k = \beta_{k+1} = 0$. For the missing scenarios (4), (6) and (7), one has $\beta_k = \beta_{k+1}$. It implies that idenfiability holds since*

$$\lim_{u\to+\infty} E_u = \lim_{u\to+\infty} \exp u \log \frac{\lambda_{k+1}}{\lambda_k} = 0.$$

B.4 Categorical data

We assume the following:

A4. The feature are independently drawn conditionally to the group membership, *i.e.*,

$$f_k(\cdot;\theta_k) = \prod_{j=1}^{d} f_{kj}(\cdot\,;\theta_{kj}); \tag{23}$$

A5. The dimension d of the observations is related to the number K of clusters so that

$$d \geq 2\lceil \log_2 K \rceil + 1,$$

with $\lceil x \rceil$ the least integer greater than or equal to x.

Assumptions **A4.** and **A5.** are classical in the categorical case, even without missing values (Allman et al., 2009). Proposition 2.2 states that generic identifiability holds only for the MNARz and the MNARz^j missing scenarios and that the other missing scenarios lead to non-identifiable models. The proof uses Corollary 5 of Allman et al. (2009) which gives the identifiability of finite mixtures of Bernoulli products.

Proof of Proposition 2.2, categorical case. Let us first consider the case where $\beta_{kj} = (0,\ldots,0) \in \mathbb{R}^{\ell_j}, \forall k = 1,\ldots,K, \forall j = 1,\ldots,d$. Suppose there exists two sets of parameters $\{\gamma,\psi\}$ and $\{\gamma',\psi'\}$ which have the same observed distribution.

$$\forall \mathbf{y}_i \in \mathbb{R}^d, \forall \mathbf{c}_i \in \{0,1\}^d,\ \sum_{k=1}^{K} \int_{\mathcal{Y}_i^{\text{mis}}} \pi_k f_k(\mathbf{y}_i;\lambda_k) \prod_{j=1}^{d} \rho(\alpha_{kj})^{c_{ij}} [1-\rho(\alpha_{kj})]^{1-c_{ij}} dy^{\text{mis}}$$
$$= \sum_{k=1}^{K'} \int_{\mathcal{Y}_i^{\text{mis}}} \pi'_k f_k(\mathbf{y}_i;\lambda'_k) \prod_{j=1}^{d} \rho((\alpha')_{kj})^{c_{ij}} [1-\rho(\alpha'_{kj})]^{1-c_{ij}} dy^{\text{mis}}.$$

Let us consider the case where all the elements of $\mathbf{y}_i$ are observed, *i.e.*, $c_{ij} = 0, \forall j = 1,\ldots,d$. One has

$$\sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i;\lambda_k) \prod_{j=1}^{d} (1-\rho(\alpha_{kj})) = \sum_{k=1}^{K'} \pi'_k f_k(\mathbf{y}_i;\lambda'_k) \prod_{j=1}^{d} (1-\rho(\alpha'_{kj})),$$

i.e., by independence to the group membership,

$$\sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} f_{kj}(y_{ij};\lambda_{kj})(1-\rho(\alpha_{kj})) = \sum_{k=1}^{K'} \pi'_k \prod_{j=1}^{d} f_{kj}(y_{ij};\lambda'_{kj})(1-\rho(\alpha'_{kj})),$$
$$\Leftrightarrow \sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} (1-\rho(\alpha_{kj}))^{1-c_{ij}} \prod_{h=1}^{\ell_j} (\lambda_{kj}^h)^{y_{ij}^h} = \sum_{k=1}^{K'} \pi'_k \prod_{j=1}^{d} (1-\rho(\alpha_{kj}))^{1-c_{ij}} \prod_{h=1}^{\ell_j} ((\lambda'_{kj})^h)^{y_{ij}^h}.$$

We recognize the finite mixture of K multinomial distributions with d components for $w_{ij} = ((1-c_{ij}), y_{ij})$, $j = 1,\ldots,d$ with paramaters $(\lambda_{kj}) = ((1-\rho(\alpha_{kj})), \lambda_{kj}^1, \ldots, \lambda_{kj}^{\ell_j}), j = 1,\ldots,d$ and proportions π_k. We can thus apply

Theorem 4 (Allman et al., 2009) with the model $\mathcal{M}(K; \ell_1, \ldots, \ell_d)$ which gives the generic identifiability of the model paramaters up to a label swapping, *i.e.*,

$$\begin{aligned} \forall k, \forall j,\ \lambda_{kj}^h &= (\lambda'_{kj})^h \\ \forall k, \forall j,\ \rho(\alpha_{kj}) &= \rho(\alpha'_{kj}) \\ \forall k,\ \pi_k &= \pi'_k \end{aligned}$$

As the function ρ is strictly monotone, the equality $\rho(\alpha_{kj}) = \rho(\alpha'_{kj})$ implies $\alpha_{kj} = \alpha'_{kj}$. In addition, if $K \neq K'$, say $K > K'$, $\sum_{k=K'+1}^{K} \pi_k \prod_{j=1}^{d}(1-\rho(\alpha_{kj})) \prod_{h=1}^{\ell_j}(\lambda_{kj}^h)^{y_{ij}^h} = 0$ implies $\pi_k = 0$ for $K'+1 \leq k \leq K$.

We consider now the missing scenarios for which $\beta_{kj} \neq 0$. The identifiability does not hold. We can present a counter-example. The set of parameters $\psi = \{\alpha = (1, \ldots, 1), \beta = (1, \ldots, 1)\}$ has the same observed distribution than another set of parameters $\psi' = \{\alpha' = (0, \ldots, 0), \beta' = (2, \ldots, 2)\}$. Indeed, in the case where $y_{ij} = (1, \ldots, 1)$, $\rho(\alpha_{kj} + \beta_{kj} y_{ij}) = \rho(\alpha'_{kj} + \beta'_{kj} y_{ij})$.

□

C Appendix 3: Details on EM algorithm

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

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

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

$$\begin{aligned} & f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1, \mathbf{c}_i; \theta^{[r-1]}) \\ &= \frac{f(\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik} = 1, \mathbf{c}_i; \theta^{[r-1]})}{f(\mathbf{y}_i^{\text{obs}}, z_{ik} = 1, \mathbf{c}_i; \theta^{[r-1]})} \qquad (24) \\ &= \frac{f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \psi^{[r-1]}) f(\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \lambda^{[r-1]})}{\int_{\mathcal{Y}_i^{\text{mis}}} f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \psi^{[r-1]}) f(\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \lambda^{[r-1]}) d\mathbf{y}_i^{\text{mis}}}. \qquad (25) \end{aligned}$$

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

$$\begin{aligned} t_{ik}(\theta^{[r-1]}) &= \frac{f(z_{ik} = 1, \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})}{f(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})} \\ &= \frac{f(z_{ik} = 1, \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})}{\sum_{h=1}^{K} f(z_{ih} = 1, \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})} \\ &= \frac{\pi_k^{[r-1]} f(\mathbf{y}_i^{\text{obs}} \mid z_{ik} = 1; \lambda_k^{[r-1]}) f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \theta^{[r-1]})}{\sum_{h=1}^{K} \pi_h^{[r-1]}) f(\mathbf{y}_i^{\text{obs}} \mid z_{ih} = 1; \lambda_h^{[r-1]}) f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}, z_{ih} = 1; \theta^{[r-1]})} \\ &= \frac{\pi_k^{[r-1]} f_k(\mathbf{y}_i^{\text{obs}}; \lambda_k^{[r-1]}) f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \theta^{[r-1]})}{\sum_{h=1}^{K} \pi_h^{[r-1]} f_h(\mathbf{y}_i^{\text{obs}}; \lambda_h^{[r-1]}) f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}, z_{ih} = 1; \theta^{[r-1]})} \end{aligned} \qquad (26)$$

The quantity that can cause numerical difficulties is the probability $f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \theta^{[r-1]})$.

C.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 and MNARz^j models, the effect of the missingness is only due to the class membership. Term (a) is the same for both models but (b) and (c) differ. Let us first detail these terms.

- 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 (25),

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

since $\prod_{j=1}^d \rho(\alpha_{kj}^{[r-1]})^{c_{ij}} (1 - \rho(\alpha_{kj}^{[r-1]}))^{1-c_{ij}}$ 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

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

one obtains

$$\left(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \lambda^{[r-1]}\right) \sim \mathcal{N}\left((\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]}, (\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]}\right). \tag{27}$$

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 Anderson (2003)) detailed as follows

$$(\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]} = (\mu_{ik}^{\text{mis}})^{[r-1]} + (\Sigma_{ik}^{\text{mis,obs}})^{[r-1]} \left((\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}\right)^{-1} \left(\mathbf{y}_i^{\text{obs}} - (\mu_{ik}^{\text{obs}})^{[r-1]}\right), \tag{28}$$

$$(\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]} = (\Sigma_{ik}^{\text{mis,mis}})^{[r-1]} - (\Sigma_{ik}^{\text{mis,obs}})^{[r-1]} \left((\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}\right)^{-1} (\Sigma_{ik}^{\text{obs,mis}})^{[r-1]}. \tag{29}$$

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

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

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

$$\log(f(\mathbf{c}_i \mid z_{ik}=1;\psi)) = \begin{cases} \sum_{j=1}^d c_{ij}\log\rho(\alpha_k) + (1-c_{ij})\log(1-\rho(\alpha_k)) & \text{(MNAR}z\text{)} \\ \sum_{j=1}^d c_{ij}\log\rho(\alpha_{kj}) + (1-c_{ij})\log(1-\rho(\alpha_{kj})) & \text{(MNAR}z^j\text{)}. \end{cases} \tag{30}$$

- 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 and MNARz^j, by independence of $\mathbf{y}_i$, one has

$$\mathbb{P}(c_{ij}=1 \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1;\theta^{[r-1]}) = \mathbb{P}(c_{ij}=1 \mid z_{ik}=1;\theta^{[r-1]}) = \begin{cases} \rho(\alpha_k) & \text{(MNAR}z\text{)} \\ \rho(\alpha_{kj}) & \text{(MNAR}z^j\text{)}. \end{cases}$$

Using (26), one obtains

$$t_{ik}^{[r-1]}(\theta^{[r-1]}) = \begin{cases} \frac{\pi_k^{[r-1]}\phi(\mathbf{y}_i^{\text{obs}};(\mu_{ik}^{\text{obs}})^{[r-1]},(\Sigma_{ik}^{\text{obs,obs}})^{[r-1]})\prod_{j=1}^d \rho(\alpha_k^{[r-1]})^{c_{ij}}(1-\rho(\alpha_k^{[r-1]}))^{1-c_{ij}}}{\sum_{h=1}^K \pi_h^{[r-1]}\phi(\mathbf{y}_i^{\text{obs}};(\mu_{ih}^{\text{obs}})^{[r-1]},(\Sigma_{ih}^{\text{obs,obs}})^{[r-1]})\prod_{j=1}^d \rho(\alpha_k^{[r-1]})^{c_{ij}}(1-\rho(\alpha_k^{[r-1]}))^{1-c_{ij}}} & \text{(MNAR}z\text{)} \\ \frac{\pi_k^{[r-1]}\phi(\mathbf{y}_i^{\text{obs}};(\mu_{ik}^{\text{obs}})^{[r-1]},(\Sigma_{ik}^{\text{obs,obs}})^{[r-1]})\prod_{j=1}^d \rho(\alpha_{kj}^{[r-1]})^{c_{ij}}(1-\rho(\alpha_{kj}^{[r-1]}))^{1-c_{ij}}}{\sum_{h=1}^K \pi_h^{[r-1]}\phi(\mathbf{y}_i^{\text{obs}};(\mu_{ih}^{\text{obs}})^{[r-1]},(\Sigma_{ih}^{\text{obs,obs}})^{[r-1]})\prod_{j=1}^d \rho(\alpha_{kj}^{[r-1]})^{c_{ij}}(1-\rho(\alpha_{kj}^{[r-1]}))^{1-c_{ij}}} & \text{(MNAR}z^j\text{)}. \end{cases} \tag{31}$$

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

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

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 (31) 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 $k=1,\ldots,K$,

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

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 $(\mathcal{J}_k^{\text{MNAR}z})^{[r]}$ (if the model is MNARz) or for the matrices $(\mathcal{J}_{kj}^{\text{MNAR}z^j})_{j=1,\ldots,d}^{[r]}$ (for the MNARz model) and by giving $t_{ik}(\theta^{[r-1]})$ as prior weights to fit the process.

$$(\mathcal{J}_k^{\text{MNAR}z})^{[r]} = \begin{array}{cc} c_{\cdot 1} & 1 \\ \vdots & \vdots \\ c_{\cdot d} & 1 \end{array}. \tag{32}$$

$$(\mathcal{J}_{kj}^{\text{MNAR}z^j})^{[r]} = \begin{array}{cc} c_{\cdot j} & 1 \end{array} \tag{33}$$

The EM algorithm for the MNARz^j model is described in Algorithm 1 for Gaussian mixture.

For missing scenarios which model the effect of the missingness depending on the variable, the computations are more difficult.

- Because of the dependence of y, $f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1, \mathbf{c}_i; \theta^{[r-1]}) = f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]})$ does not hold anymore. Here, one has

$$\begin{aligned} &f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1, \mathbf{c}_i; \theta^{[r-1]}) \\ &= \frac{\prod_{h=1}^d \rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{mis}})^{c_{ih}} (1-\rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{obs}}))^{1-c_{ih}} f(\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]})}{\int_{\mathcal{Y}_i^{\text{mis}}} \prod_{h=1}^d \rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{mis}})^{c_{ih}} (1-\rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{obs}}))^{1-c_{ih}} f(\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]}) d\mathbf{y}_i^{\text{mis}}} \\ &= \frac{\prod_{h,c_{ih}=1} \rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{mis}}) f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]})}{\int_{\mathcal{Y}_i^{\text{mis}}} \prod_{h,c_{ih}=1} \rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{mis}}) f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]}) d\mathbf{y}_i^{\text{mis}}}. \end{aligned} \tag{34}$$

  which implies that Term (a) requires difficult computations if this distribution is not classical.

- For Term (b), it is the same problem, since $f(\mathbf{c}_i \mid \mathbf{y}_i, z_{ik}=1; \psi)$ is no longer independent of $\mathbf{y}$, then it requires a specific numerical integration. Using (34),

$$\begin{aligned} \tau_c(\psi_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]}) =& \mathbb{E}\left[\log\left(\prod_{j=1}^d \rho(\alpha_{kj} + \beta_{kj} y_{ij})^{c_{ij}} (1-\rho(\alpha_{kj} + \beta_{kj} y_{ij}))^{1-c_{ij}}\right) | \mathbf{y}_i^{\text{obs}}, z_{ik}=1, \mathbf{c}_i; \theta^{[r-1]}\right] \\ =& \sum_{j=1}^d c_{ij} \int_{\mathcal{Y}_{ij}^{\text{mis}}} \log(\rho(\alpha_{kj} + \beta_{kj} y_{ij}^{\text{mis}})) f(y_{ij}^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1, c_i; \theta^{[r-1]}) dy_{ij}^{\text{mis}} \\ &+ (1-c_{ij}) \log(1-\rho(\alpha_{kj} + \beta_{kj} y_{ij}^{\text{obs}})) \end{aligned}$$

  where

$$\begin{aligned} &f(y_{ij}^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1, \mathbf{c}_i; \theta^{[r-1]}) \\ &= \frac{\rho(\alpha_{kj}^{[r-1]} + \beta_{kj}^{[r-1]} y_{ij}^{\text{mis}})^{c_{ij}} (1-\rho(\alpha_{kj}^{[r-1]} + \beta_{kj}^{[r-1]} y_{ij}^{\text{mis}}))^{1-c_{ij}} f(y_{ij}^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]})}{\int_{\mathcal{Y}_{ij}^{\text{mis}}} \rho(\alpha_{kj}^{[r-1]} + \beta_{kj}^{[r-1]} y_{ij}^{\text{mis}})^{c_{ij}} (1-\rho(\alpha_{kj}^{[r-1]} + \beta_{kj}^{[r-1]} y_{ij}^{\text{mis}}))^{1-c_{ij}} f(y_{ij}^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]}) dy_{ij}^{\text{mis}}}. \end{aligned}$$

  Therefore,

$$\begin{aligned} &\tau_c(\psi_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]}) \\ &= \sum_{j=1}^d c_{ij} \int_{\mathcal{Y}_{ij}^{\text{mis}}} \log(\rho(\alpha_{kj} + \beta_{kj} y_{ij}^{\text{mis}})) \frac{\rho(\alpha_{kj}^{[r-1]} + \beta_{kj}^{[r-1]} y_{ij}^{\text{mis}})^{c_{ij}} f(y_{ij}^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]})}{\int_{\mathcal{Y}_{ij}^{\text{mis}}} \rho(\alpha_{kj}^{[r-1]} + \beta_{kj}^{[r-1]} x)^{c_{ij}} f(x \mid y_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]}) dx} dy_{ij}^{\text{mis}} \\ &\qquad + (1-c_{ij}) \log(1-\rho(\alpha_{kj} + \beta_{kj} y_{ij}^{\text{obs}})). \end{aligned}$$

Algorithm 1 EM algorithm for Gaussian mixture and MNARz^j 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.

$t_{ik}(\theta^{[r-1]}) \propto \pi_k^{[r-1]} \phi(\mathbf{y}_i^{\text{obs}}; (\mu_{ik}^{\text{obs}})^{[r-1]}, (\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}) \prod_{j=1}^{d} \rho(\alpha_{kj}^{[r-1]})^{c_{ij}} (1 - \rho(\alpha_{kj}^{[r-1]}))^{1-c_{ij}}$
end for
M-step:
for $k = 1$ **to** K **do**

$$\pi_k^{[r]} = \frac{1}{n}\sum_{i=1}^{n} t_{ik}(\theta^{[r-1]}), \qquad \mu_k^{[r]} = \frac{\sum_{i=1}^{n} t_{ik}(\theta^{[r-1]})(\tilde{y}_{k,i})^{[r-1]}}{\sum_{i=1}^{n} t_{ik}(\theta^{[r-1]})}$$

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

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, $\forall j \in \{1, \ldots, d\}$,

$$\mathcal{M}_{kj}\psi_k^{[r]} = \log\left(\frac{1 - \mathbb{E}[\mathbf{c}_{\cdot j}|\mathcal{M}_{kj}])}{\mathbb{E}[\mathbf{c}_{\cdot j}|\mathcal{M}_{kj}]}\right),$$

for a matrix $\mathcal{M}_{kj}$ depending on the MNAR model (see (32) and (33)) and $\mathbf{c}_{\cdot j}$ the missing data pattern for the variable j.
end for
end for

- There is no closed-form expression for Term (c).

$$
\begin{aligned}
&f(c_{ij} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]}) \\
&= \int_{\mathcal{Y}_{ij}^{\text{mis}}} f(c_{ij} \mid \mathbf{y}_i^{\text{obs}}, y_{ij}^{\text{mis}}, z_{ik}=1; \psi^{[r-1]}) f(\mathbf{y}_{ij}^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]}) dy_{ij}^{\text{mis}} \\
&= c_{ij} \int_{-\infty}^{+\infty} \rho(\alpha_{kj}^{[r-1]} + \beta_{kj}^{[r-1]} y_{ij}^{\text{mis}}) \phi(y_{ij}^{\text{mis}}; (\tilde{\mu}_{ik}^{\text{mis}})_j^{[r-1]}, (\tilde{\Sigma}_{ik}^{\text{mis}})_{jj}^{[r-1]}) dy_{ij}^{\text{mis}} + (1-c_{ij})(1-\rho(\alpha_{kj}^{[r-1]} + \beta_{kj}^{[r-1]} y_{ij}^{\text{obs}})).
\end{aligned} \tag{35}
$$

Using (26), the probabilities $t_{ik}(\theta^{[r-1]})$ can be deduced from Equation (35).

Let us detail the difficulties for two particular cases, if ρ is logistic or probit.

- ρ **is logistic**: Equation (34) leads to none classical distribution because

$$
\begin{aligned}
&f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1, c_i; \theta^{[r-1]}) \\
&\qquad \propto \prod_{h, c_{ih}=1} \frac{1}{\exp(-(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{mis}}))} \phi(y_i^{\text{mis}}; (\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]}, (\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]}).
\end{aligned}
$$

Term (b) is

$$
\begin{aligned}
&\tau_c(\psi_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]}) \\
&\qquad \propto \sum_{j=1}^{d} c_{ij} \int_{\mathcal{Y}_{ij}^{\text{mis}}} -\frac{\log(1+\exp(-(\alpha_{kj} + \beta_{kj} y_{ij}^{\text{mis}})))}{1+\exp(-(\alpha_{kj}^{[r-1]} + \beta_{kj}^{[r-1]} y_{ij}^{\text{mis}}))} \phi(y_{ij}^{\text{mis}}; (\tilde{\mu}_{ik}^{\text{mis}})_j^{[r-1]}, (\tilde{\Sigma}_{ik}^{\text{mis}})_{jj}^{[r-1]}) dy_{ij}^{\text{mis}} \\
&\qquad\qquad - (1-c_{ij}) \log(1+\exp(\alpha_{kj} + \beta_{kj} y_{ij}^{\text{obs}})),
\end{aligned}
$$

which amounts to compute the Gaussian moment of $\frac{\log(1+\exp(-u))}{1+\exp(-u)}$, but it has no closed form to our knowledge. Finally, Equation (35) does not have a closed form either because it requires the computation of

$$
\int_{-\infty}^{+\infty} \frac{1}{1+\exp(-(\alpha_{kj}^{[r-1]} + \beta_{kj}^{[r-1]} y_{ij}^{\text{mis}}))} \phi(y_{ij}^{\text{mis}}; (\tilde{\mu}_{ik}^{\text{mis}})_j^{[r-1]}, (\tilde{\Sigma}_{ik}^{\text{mis}})_{jj}^{[r-1]}) dy_{ij}^{\text{mis}},
$$

i.e., the computation of the Gaussian moment of $\frac{1}{1+\exp(-u)}$.

- ρ **is Probit**: One can prove (presented in Appended D.1) that the conditional distribution $(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1, \mathbf{c}_i)$ is a truncated Gaussian, which makes possible the computation of Term (a). Term (b) has no closed form to our knowledge

$$
\begin{aligned}
&\tau_c(\psi_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]}) \\
&\qquad \propto \sum_{j=1}^{d} c_{ij} \int_{\mathcal{Y}_{ij}^{\text{mis}}} \frac{\log\left(\int_{-\infty}^{\alpha_{kj} + \beta_{kj} y_{ij}^{\text{mis}}} e^{-t^2} dt\right)}{1+\exp(\alpha_{kj}^{[r-1]} + \beta_{kj}^{[r-1]} y_{ij}^{\text{mis}})} \phi(y_{ij}^{\text{mis}}; (\tilde{\mu}_{ik}^{\text{mis}})_j^{[r-1]}, (\tilde{\Sigma}_{ik}^{\text{mis}})_{jj}^{[r-1]}) dy_{ij}^{\text{mis}} \\
&\qquad\qquad - (1-c_{ij}) \log\left(1 - \int_{-\infty}^{\alpha_{kj} + \beta_{kj} y_{ij}^{\text{obs}}} e^{-t^2} dt\right),
\end{aligned}
$$

Equation (35) does not have a closed form either because it requires the computation of

$$
\int_{-\infty}^{+\infty} \left(\int_{-\infty}^{\alpha_{kj} + \beta_{kj} y_{ij}^{\text{mis}}} e^{-t^2} dt\right) \phi(y_{ij}^{\text{mis}}; (\tilde{\mu}_{ik}^{\text{mis}})_j^{[r-1]}, (\tilde{\Sigma}_{ik}^{\text{mis}})_{jj}^{[r-1]}) dy_{ij}^{\text{mis}}.
$$

C.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

$$\mathbb{E}\left[\log(\phi(\mathbf{y}_i;p_k))\mid\mathbf{y}_i^{\text{obs}},z_{ik}=1,\mathbf{c}_i;\lambda^{[r-1]}\right]=\sum_{j,c_{ij}=0}\sum_{\ell=1}^{\ell_j}y_{ij}^\ell+\sum_{j,c_{ij}=1}\sum_{\ell=1}^{\ell_j}\log(\lambda_{kj}^{y_{ij}^\ell}) \tag{36}$$

Term (b) is the same as in the Gaussian case given in (30). Finally, the EM algorithm can be summarized as follows **E step:** For $k=1,\dots,K$ and $i=,\dots,n$, compute

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

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

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

The M-step for ψ consists of performing a GLM with a binomial link and has already been given in detail in Appendix C.1 (see (50) and (51)).

C.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.

D Appendix 4: Details on SEM algorithm

The SEM algorithm consists on two steps iteratively proceeded as presented in Section 3.2. The key issue is to draw the missing data $(\mathbf{y}_i^{\text{mis}})^r$ and $\mathbf{z}_i^r$ according to their current conditional distribution $f(\mathbf{y}_i^{\text{mis}},\mathbf{z}_i\mid\mathbf{y}_i^{\text{obs}},\mathbf{c}_i;\pi^{[r-1]},\lambda^{[r-1]},\psi^{[r-1]})$. By convenience, we use a Gibbs sampling and simulate two easier probabilities recalled here

$$\mathbf{z}_i^{[r]}\sim f(\cdot\mid\mathbf{y}_i^{[r-1]},\mathbf{c}_i;\pi^{[r-1]},\lambda^{[r-1]},\psi^{[r-1]})\quad\text{and}\quad(\mathbf{y}_i^{\text{mis}})^{[r]}\sim f(\cdot\mid\mathbf{y}_i^{\text{obs}},\mathbf{z}_i^r,\mathbf{c}_i;\lambda^{[r-1]},\psi^{[r-1]}),$$

where $\mathbf{y}_i^{[r-1]} = (\mathbf{y}_i^{\text{obs}}, (\mathbf{y}_i^{\text{mis}})^{[r-1]})$. For the latter distribution, the membership k of $z_i^{[r]}$ is drawn from the multinomial distribution with probabilities $(\mathbb{P}(z_{ik} = 1 \mid \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \lambda^{[r-1]}, \psi^{[r-1]}))_{k=1,\dots,K}$ detailed as follows

$$\begin{aligned}
&\mathbb{P}(z_{ik} = 1 \mid \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]}) \\
&= \frac{\mathbb{P}(z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \pi^{[r-1]}, \lambda^{[r-1]}, \psi^{[r-1]})}{\mathbb{P}(\mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \pi^{[r-1]}, \lambda^{[r-1]}, \psi^{[r-1]})} \qquad (37) \\
&= \frac{\mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i^{[r-1]}, z_{ik} = 1; \psi^{[r-1]}) \mathbb{P}(\mathbf{y}_i^{[r-1]} \mid z_{ik} = 1; \lambda^{[r-1]}) \mathbb{P}(z_{ik} = 1; \pi^{[r-1]})}{\sum_{h=1}^K \mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i^{[r-1]}, z_{ih} = 1; \psi^{[r-1]}) \mathbb{P}(\mathbf{y}_i^{[r-1]} \mid z_{ih} = 1; \lambda^{[r-1]}) \mathbb{P}(z_{ih} = 1; \pi^{[r-1]})} \\
&= \frac{\mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i^{[r-1]}, z_{ik} = 1; \psi^{[r-1]}) \mathbb{P}(\mathbf{y}_i^{[r-1]} \mid z_{ik} = 1; \lambda^{[r-1]}) \pi_k^{[r-1]}}{\sum_{h=1}^K \mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i^{[r-1]}, z_{ih} = 1; \psi^{[r-1]}) \mathbb{P}(\mathbf{y}_i^{[r-1]} \mid z_{ih} = 1; \lambda^{[r-1]}) \pi_h^{[r-1]}}. \qquad (38)
\end{aligned}$$

The conditional distribution of $((\mathbf{y}_i^{\text{mis}})^{[r]} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1, \mathbf{c}_i)$ has already been detailed in Equation (34) and recalled here

$$\begin{aligned}
&f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1, \mathbf{c}_i; \theta^{[r-1]}) \\
&\qquad = \frac{\prod_{j, c_{ij}=1} f(c_{ij} = 1 \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \psi^{[r-1]}) f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \lambda^{[r-1]})}{\int_{\mathcal{Y}_i^{\text{mis}}} \prod_{j, c_{ij}=1} f(c_{ij} = 1 \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \psi^{[r-1]}) f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \lambda^{[r-1]}) d\mathbf{y}_i^{\text{mis}}}. \qquad (39)
\end{aligned}$$

D.1 Gaussian mixture for continuous data

First note that the probabilities of the multinomial distribution for drawing $z_i^{[r]}$ given in (38) can be easily computed for all cases.

$$\begin{aligned}
&\mathbb{P}(z_{ik} = 1 \mid \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]}) \\
&\qquad = \frac{\prod_{j=1}^d f(c_{ij} = 1 \mid \mathbf{y}_i^{[r-1]}, z_{ik}^{[r-1]} = 1; \psi^{[r-1]})^{c_{ij}} f(c_{ij} = 0 \mid \mathbf{y}_i^{[r-1]}, z_{ik}^{[r-1]} = 1; \psi^{[r-1]})^{1-c_{ij}} \phi(\mathbf{y}_i^{[r-1]}; \lambda_k^{[r-1]}) \pi_k^{[r-1]}}{\sum_{h=1}^K \prod_{j=1}^d f(c_{ij} = 1 \mid \mathbf{y}_i^{[r-1]}, z_{ih}^{[r-1]} = 1; \psi^{[r-1]})^{c_{ij}} f(c_{ij} = 0 \mid \mathbf{y}_i^{[r-1]}, z_{ih}^{[r-1]} = 1; \psi^{[r-1]})^{1-c_{ij}} \phi(\mathbf{y}_i^{[r-1]}; \lambda_h^{[r-1]}) \pi_h^{[r-1]}},
\end{aligned}$$

where $\phi(\mathbf{y}_i; \lambda_k) = \phi(\mathbf{y}_i; \mu_k, \Sigma_k)$ is assumed to be a Gaussian distribution with mean vector μ_k and covariance matrix Σ_k, and $f(c_{ij} = 1 \mid \mathbf{y}_i^{[r-1]}, z_{ih}^{[r-1]} = 1; \psi^{[r-1]})$ is specified depending the MNAR model and the distribution ρ. The only difficulty of the SE-step is thus to draw from the distribution $(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1, \mathbf{c}_i)$.

In the sequel, we detail the distribution $(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1, \mathbf{c}_i)$ and the M-step for ψ depending the MNAR model.

For MNAR$y\star$ models, the conditional distribution $(y_i^{\text{mis}} \mid y_i^{\text{obs}}, z_{ik}^{[r]} = 1, c_i)$ depends on the distribution ρ at hand. For the MNAR$y\star$ models, we will consider two classical distributions for ρ: the logistic function and probit one.

Logistic distribution: For the logistic function, the distribution given in (39) is not classical and drawing y_i^{mis} from it seems complicated. Indeed, one has

$$\begin{aligned}
&f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1, \mathbf{c}_i; \theta^{[r-1]}) \\
&\qquad \propto \prod_{j=1, c_{ij}=1} \frac{1}{1 + \exp(\alpha_{kj}^{[r-1]} + \beta_{kj}^{[r-1]} y_{ij}^{\text{mis}})} \phi(\mathbf{y}_i^{\text{mis}}; (\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]}, (\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]}),
\end{aligned}$$

where $(\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]}$ and $(\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]}$ are given in (28) and (29). We could use the Sampling Importance Resampling (SIR) algorithm which simulates a realization of $(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1, \mathbf{c}_i)$ with a known instrumental distribution (for example: $(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1)$) and includes a re-sampling step. However, this algorithm may be computationnaly costly.

Probit distribution: For the probit function, the distribution in (39) can be made explicit by using a latent variable $\mathbf{L}_i$.

More particularly, let $\mathbf{L}_i$ such that $\mathbf{L}_i = \alpha_k^{[r-1]} + \beta_k^{[r-1]}\mathbf{y}_i + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0_d, I_{d\times d})$. Then, $\mathbf{c}_i$ can be viewed as an indicator for whether this latent variable is positive, *i.e.*, for all $j = 1, \ldots, d$,

$$c_{ij} = \begin{cases} 1 & \text{if } L_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{40}$$

Thus, indeed to draw $(\mathbf{y}_i^{\text{mis}})^{[r]}$ and $\mathbf{z}_i^{[r]}$ according to $f(\mathbf{y}_i^{\text{mis}}, \mathbf{z}_i \mid \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})$, we draw $\mathbf{L}_i^{[r]}$, $(\mathbf{y}_i^{\text{mis}})^{[r]}$ and $\mathbf{z}_i^{[r]}$ according to $f(\mathbf{L}_i, \mathbf{y}_i^{\text{mis}}, \mathbf{z}_i \mid \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]}$ by using a Gibbs sampling.

First, we have to draw $\mathbf{L}_i^{[r]}$ according to $f(. \mid \mathbf{y}_i^{[r-1]}, z_{ik}^{[r-1]} = 1, \mathbf{c}_i; \psi^{[r-1]})$. One has

$$\begin{aligned} f(\mathbf{L}_i \mid \mathbf{y}_i^{[r-1]}, z_{ik}^{[r-1]} = 1, \mathbf{c}_i) & \\ &\propto f(\mathbf{L}_i, \mathbf{c}_i \mid \mathbf{y}_i^{[r-1]}, z_{ik}^{[r-1]} = 1; \psi^{[r-1]}) \\ &\propto f(\mathbf{c}_i \mid \mathbf{L}_i^{[r]}, \mathbf{y}_i^{[r-1]}, z_{ik}^{[r-1]} = 1; \psi^{[r-1]}) f(\mathbf{L}_i^{[r]} \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r-1]} = 1; \psi^{[r-1]}) \\ &\overset{(i)}{\propto} f(\mathbf{c}_i \mid \mathbf{L}_i^{[r]}; \psi^{[r-1]}) f(\mathbf{L}_i^{[r]} \mid \mathbf{y}_i^{[r-1]}, z_{ik}^{[r-1]} = 1; \psi^{[r-1]}) \\ &\overset{(ii)}{=} \mathbb{1}_{\{\mathbf{c}_i=1\}\cap\{\mathbf{L}_i^{[r]}>0\}} f(\mathbf{L}_i^{[r]} \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r-1]} = 1; \psi^{[r-1]}) \end{aligned}$$

where we use that $\mathbf{L}_i^{[r]}$ is a function of $\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik} = 1$ in step (i). Step (ii) is obtained by using (40). By abuse of notation, $\{\mathbf{c}_i = 1\} \cap \{\mathbf{L}_i^{[r]} > 0\}$ means that for all $j = 1, \ldots, d$, $\{c_{ij} = 1\} \cap \{L_{ij}^{[r]} > 0\}$. Finally the conditional distribution $(\mathbf{L}_i \mid \mathbf{y}_i^{[r-1]}, z_{ik}^{[r-1]} = 1, \mathbf{c}_i)$ is a multivariate truncated Gaussian distribution denoted as $\mathcal{N}_t$, as detailed here

$$(\mathbf{L}_i \mid \mathbf{y}_i^{[r-1]}, z_{ik}^{[r-1]} = 1, c_i) \sim \mathcal{N}_t(\alpha_k^{[r-1]} + \beta_k^{[r-1]}\mathbf{y}_i, I_{d\times d}; a, b), \tag{41}$$

with $a \in \mathbb{R}^d$ and $b \in \mathbb{R}^d$ the lower and upper bounds such that for all $j = 1, \ldots, d$,

$$a_j = \begin{cases} 0 & \text{if } c_{ij} = 1, \\ -\infty & \text{otherwise.} \end{cases}$$

$$b_j = \begin{cases} +\infty & \text{if } c_{ij} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Secondly, we draw the membership k of $\mathbf{z}_i^{[r]}$ from the multinomial distribution with probabilities, for all $k = 1, \ldots, K$ detailed as follows

$$\begin{aligned} \mathbb{P}(z_{ik} = 1 \mid \mathbf{L}_i^{[r]}, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]}) &= \frac{\mathbb{P}(z_{ik} = 1, \mathbf{L}_i^{[r]}, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]})}{\sum_{k=1}^K \mathbb{P}(z_{ik} = 1, \mathbf{L}_i^{[r]}, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]})} \\ &= \frac{f(\mathbf{L}_i^{[r]} | z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \psi^{[r-1]}) f(z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]})}{\sum_{k=1}^K f(\mathbf{L}_i^{[r]} | z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \psi^{[r-1]}) f(z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]})}. \end{aligned} \tag{42}$$

The part involving $f(z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]})$ is given in (38) and $f(\mathbf{L}_i^{[r]} | z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \psi^{[r-1]})$ is only the density of the multivariate truncated Gaussian distribution described in (41) evaluated in $L_i^{[r]}$.

Finally, $\mathbf{y}_i^{[r]}$ is drawn according to $f(.|\mathbf{L}_i^{[r]}, z_{ik}^{[r]} = 1, \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})$. One has

$$
\begin{aligned}
&f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{L}_i^{[r]}, z_{ik}^{[r]} = 1, \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]}) \\
&\propto f(\mathbf{c}_i, \mathbf{L}_i^{[r]} \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \psi^{[r-1]}) f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \theta^{[r-1]}) \\
&\propto f(\mathbf{c}_i \mid \mathbf{L}_i^{[r]}, \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \psi^{[r-1]}) f(\mathbf{L}_i^{[r]} \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \psi^{[r-1]}) f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \theta^{[r-1]}) \\
&\propto f(\mathbf{c}_i \mid \mathbf{L}_i^{[r]}; \psi^{[r-1]}) f(\mathbf{L}_i^{[r]} \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \psi^{[r-1]}) f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \theta^{[r-1]}) \\
&\propto f(\mathbf{L}_i^{[r]} \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \psi^{[r-1]}) f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \theta^{[r-1]}).
\end{aligned}
$$

Yet, one has

$$
\begin{aligned}
f(\mathbf{L}_i^{[r]} \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \psi^{[r-1]}) &\propto \exp\left(-\frac{1}{2}\left[(\mathbf{L}_i^{[r]} - (\alpha_k^{[r-1]} + \beta_k^{[r-1]}\mathbf{y}_i))^T (\mathbf{L}_i^{[r]} - (\alpha_k^{[r-1]} + \beta_k^{[r-1]}\mathbf{y}_i))\right]\right) \\
f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \theta^{[r-1]}) &\propto \exp\left(-\frac{1}{2}\left[(\mathbf{y}_i^{\text{mis}} - (\tilde{\mu}_{ik}^{\text{mis}})^{[r]})^T ((\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]})^{-1} (\mathbf{y}_i^{\text{mis}} - (\tilde{\mu}_{ik}^{\text{mis}})^{[r]})\right]\right),
\end{aligned}
$$

with $\tilde{\mu}_{ik}^{\text{mis}}$ and $(\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]}$ given in (27).

Finally combining these two equations one obtains

$$
\left(\mathbf{y}_i^{\text{mis}} \mid \mathbf{L}_i^{[r]}, z_{ik}^{[r]} = 1, \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i\right) \sim \mathcal{N}\left(\mu_{ik}^{\text{SEM}}, \Sigma_{ik}^{\text{SEM}}\right), \tag{43}
$$

where

$$
\begin{aligned}
\Sigma_{ik}^{\text{SEM}} &= \left((\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]})^{-1} + ((\beta_k^{\text{mis}})^{[r-1]})^T (\beta_k^{\text{mis}})^{[r-1]}\right)^{-1}, \\
\mu_{ik}^{\text{SEM}} &= \Sigma_{ik}^{\text{SEM}} \left[((\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]})^{-1} \tilde{\mu}_{ik}^{\text{mis}} + ((\beta_k^{\text{mis}})^{[r-1]})^T (\mathbf{L}_i^{\text{mis}})^{[r]} - ((\beta_k^{\text{mis}})^{[r-1]})^T (\alpha_k^{\text{mis}})^{[r-1]}\right],
\end{aligned}
$$

with $(\beta_k^{\text{mis}})^{[r-1]}$ (resp. $(\mathbf{L}_i^{\text{mis}})^{[r]}$ and $(\alpha_k^{\text{mis}})^{[r-1]}$) the vector β_k (resp. $(\mathbf{L}_i)^{[r]}$ and $(\alpha_k)^{[r-1]}$) restricted to the coordinates $j \in \mathcal{Y}_i^{\text{mis}}$.

Finally, for fully describing the SEM-algorithm, in the M-step, $\psi^{[r-1]}$ is computed using a GLM with a binomial link function for a matrix depending on the MNAR model. In particular,

- For MNARy, the coefficient obtained with a GLM for the matrix $(\mathcal{H}_j^{\text{MNAR}y})^{[r]}$ are α_0 and $\beta_1^{[r]}, \ldots, \beta_d^{[r]}$, with

$$
(\mathcal{H}^{\text{MNAR}y})^{[r]} = \left(\begin{array}{c|ccccc}
c_{.1} & 1 & y_{.1}^{[r]} & 0 & \ldots & 0 \\
c_{.2} & 1 & 0 & y_{.2}^{[r]} & \ldots & 0 \\
\vdots & \vdots & & \ddots & \ddots & \\
c_{.d} & 1 & 0 & 0 & \ldots & y_{.d}^{[r]}
\end{array}\right). \tag{44}
$$

- For MNARy^k, the coefficient obtained with a GLM for the matrix $(\mathcal{H}_{kj}^{\text{MNAR}y^k})^{[r]}$ is α_0 and $\beta_{11}^{[r]}, \ldots, \beta_{K1}^{[r]}, \ldots, \beta_{Kd}^{[r]}$ with

$$(\mathcal{H}_{kj}^{\text{MNAR}y^k})^{[r]} = \left|\begin{array}{c|cccc} (c_{u1})_{u,z_{u1}^{[r]}=1} & 1 & (y_{u1}^{[r]})_{u,z_{u1}^{[r]}=1} & 0 & \cdots & 0 \\ \vdots & \vdots & & & \ddots & \vdots \\ (c_{u1})_{u,z_{uK}^{[r]}=1} & 1 & 0 & (y_{u1}^{[r]})_{u,z_{uK}^{[r]}=1} & & 0 \\ \vdots & \vdots & \vdots & & \ddots & \\ (c_{ud})_{u,z_{uK}^{[r]}=1} & 1 & 0 & 0 & & (y_{ud}^{[r]})_{u,z_{uK}^{[r]}=1} \end{array}\right| . \tag{45}$$

- For MNARyz, the coefficients obtained with a GLM for the matrix $(\mathcal{H}^{\text{MNAR}yz})^{[r]}$ are $\beta_1^{[r]}, \ldots, \beta_d^{[r]}$ and $\alpha_1^{[r]}, \ldots, \alpha_K^{[r]}$, with

$$(\mathcal{H}^{\text{MNAR}yz})^{[r]} = \left|\begin{array}{c|cccccccc} c_{\cdot 1} & y_{\cdot 1}^{[r]} & 0 & \cdots & 0 & z_{\cdot 1}^{[r]} & \cdots & z_{\cdot K}^{[r]} \\ c_{\cdot 2} & 0 & y_{\cdot 2}^{[r]} & \cdots & 0 & z_{\cdot 1}^{[r]} & \cdots & z_{\cdot K}^{[r]} \\ \vdots & & \ddots & \ddots & & \vdots & \vdots & \vdots \\ c_{\cdot d} & 0 & 0 & \cdots & y_{\cdot d}^{[r]} & z_{\cdot 1}^{[r]} & \cdots & z_{\cdot K}^{[r]} \end{array}\right| . \tag{46}$$

- For MNARyz^j, the coefficients obtained with a GLM for the matrix $(\mathcal{H}_j^{\text{MNAR}yz^j})^{[r]}$ are $\beta_j^{[r]}, \alpha_{1j}^{[r]}, \ldots, \alpha_{Kj}^{[r]}$, with

$$(\mathcal{H}_j^{\text{MNAR}yz^j})^{[r]} = \left|\begin{array}{c|cccc} c_{\cdot j} & y_{\cdot j}^{[r]} & z_{\cdot 1}^{[r]} & \cdots & z_{\cdot K}^{[r]} \end{array}\right| . \tag{47}$$

- For MNARy^kz, the coefficients obtained with a GLM for the matrix $(\mathcal{H}_k^{\text{MNAR}y^kz})^{[r]}$ are $\beta_{k1}^{[r]}, \ldots, \beta_{kd}^{[r]}, \alpha_k^{[r]}$, with

$$(\mathcal{H}_k^{\text{MNAR}y^kz})^{[r]} = \left|\begin{array}{c|ccccc} (c_{u1})_{u,z_{uk}^{[r]}=1} & (y_{u1}^{[r]})_{u,z_{uk}^{[r]}=1} & 0 & \cdots & 0 & 1 \\ (c_{u2})_{u,z_{uk}^{[r]}=1} & 0 & (y_{u2}^{[r]})_{u,z_{uk}^{[r]}=1} & \cdots & 0 & 1 \\ \vdots & & & \ddots & \ddots & 1 \\ (c_{ud})_{u,z_{uk}^{[r]}=1} & 0 & 0 & \cdots & (y_{ud}^{[r]})_{u,z_{uk}^{[r]}=1} & 1 \end{array}\right| . \tag{48}$$

- For MNARy^kz^j, the coefficients obtained with a GLM for the matrix $(\mathcal{H}_{kj}^{\text{MNAR}y^kz^j})^{[r]}$ are β_{kj}, α_{kj}, with

$$(\mathcal{H}_{kj}^{\text{MNAR}y^kz^j})^{[r]} = \left|\begin{array}{c|cc} (c_{uj})_{u,z_{uk}^{[r]}=1} & (y_{uj}^{[r]})_{u,z_{uk}^{[r]}=1} & 1 \end{array}\right| \tag{49}$$

- For MNARz, the coefficients obtained with a GLM for the matrix $(\mathcal{H}^{\text{MNAR}z})^{[r]}$ are $\alpha_1, \ldots, \alpha_K$, with

$$(\mathcal{H}^{\text{MNAR}z})^{[r]} = \left|\begin{array}{c|ccc} c_{\cdot 1} & z_{\cdot 1} & \cdots & z_{\cdot K} \\ \vdots & \vdots & \vdots & \vdots \\ c_{\cdot d} & z_{\cdot 1} & \cdots & z_{\cdot K} \end{array}\right| = \left|\begin{array}{c|ccc} c_{11} & z_{11}^{[r]} & \cdots & z_{1K}^{[r]} \\ \vdots & \vdots & \vdots & \vdots \\ c_{n1} & z_{n1}^{[r]} & \cdots & z_{nK}^{[r]} \\ \vdots & \vdots & \vdots & \vdots \\ c_{1d} & z_{11}^{[r]} & \cdots & z_{1K}^{[r]} \\ \vdots & \vdots & \vdots & \vdots \\ c_{nd} & z_{n1}^{[r]} & \cdots & z_{nK}^{[r]} \end{array}\right| . \tag{50}$$

- For MNARz^j, the coefficients obtained with a GLM for the matrix $(\mathcal{H}_j^{\text{MNAR}z^j})^{[r]}$ are $\alpha_{1j}, \ldots, \alpha_{Kj}$, with

$$(\mathcal{H}_j^{\text{MNAR}z^j})^{[r]} = \begin{array}{c|ccc} c_{.j} & z_{.1}^{[r]} & \ldots & z_{.K}^{[r]} \end{array} \tag{51}$$

When ρ is the probit function, the SEM algorithm can be derived, see Algorithm 2. The initialization and the stopping criterion are discussed in Section 4.

Algorithm 2 SEM algorithm for Gaussian mixture, MNAR$y\star$ models, ρ is probit

Input: $Y \in \mathbb{R}^{n\times d}$ (matrix containing missing values), $K \geq 1$, $r_{\max}$.
Initialize Z^0, 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**
  SE-step:
  for $i = 1$ **to** n **do**
    Draw $(\mathbf{L}_i)^{[r]}$ from the multivariate truncated Gaussian distribution given in (41).
    Draw $\mathbf{z}_i^{[r]}$ from the multinomial distribution with probabilities detailed in (42).
    Draw $(\mathbf{y}_i^{\text{mis}})^{[r]}$ from the multivariate Gaussian distribution given in (43).
  end for
  Let $Y^{[r]} = (\mathbf{y}_1^{[r]}|\ldots|\mathbf{y}_n^{[r]})$ be the imputed matrix.
  Let $Z^{[r]} = (\mathbf{z}_1^{[r]}|\ldots|\mathbf{z}_n^{[r]})$ be the partition.
  M-step:
  for $k = 1$ **to** K **do**
    Let $\pi_k^{[r]}$ be the proportion of rows of $Y^{[r]}$ belonging class k.
    Let $\mu_k^{[r]}, \Sigma_k^{[r]}$ be the mean and covariance matrix of rows of $Y^{[r]}$ belonging to class k.
    Let $\psi_k^{[r]}$ be the resulted coefficients of a GLM with a binomial link function, *i.e.*, the optimization problem is $\forall j \in \{1, \ldots, d\}$,

$$\mathcal{M}_{kj}\psi_k^{[r]} = \log\left(\frac{1 - \mathbb{E}[\mathbf{c}_{.j}|\mathcal{M}_{kj}])}{\mathbb{E}[\mathbf{c}_{.j}|\mathcal{M}_{kj}]}\right),$$

    for a matrix $\mathcal{M}_{kj}$ depending on the MNAR model (see (44), (45), (46), (51), (49) and (50)) and $\mathbf{c}_{.j}$ the missing data pattern for the variable j.
  end for
end for

For the MNARz and MNARz^j models, the effect of the missingness is only due to the class membership. We have already proved in Appendix C.1 that

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

and that this conditional distribution is Gaussian given in (27). The M-step for ψ has been specified in the previous paragraph with (50) and (51).

D.2 Latent class model for categorical data

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

For drawing from the conditional distribution $(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1)$, by independence of the features conditionally to the membership, we can draw for $j = 1,\ldots,d$ $y_{ij}^{\text{mis}} = ((y_{ij}^{\text{mis}})^1,\ldots,(y_{ij}^{\text{mis}})^{l_j})$ from the conditional distribution $(y_{ij}^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1)$. This latter is a multinomial distribution with probabilities $(\lambda_{kj}^{\ell})_{\ell=1,\ldots,\ell_j}$.

E Appendix 5: Additional numerical experiments on synthetic data

Note that in Figure 8, 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.

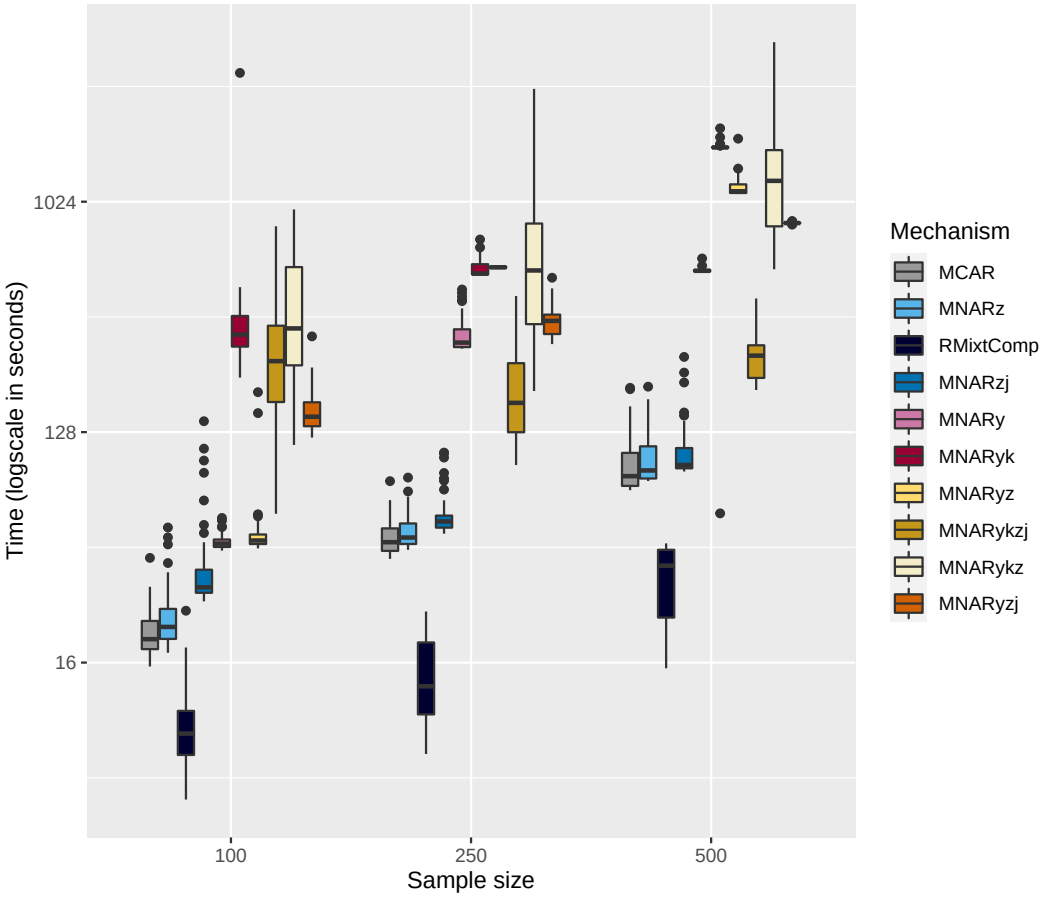


Figure 7: Boxplot of the computational times (in seconds) obtained for 50 samples composed of $d = 6$ variables.

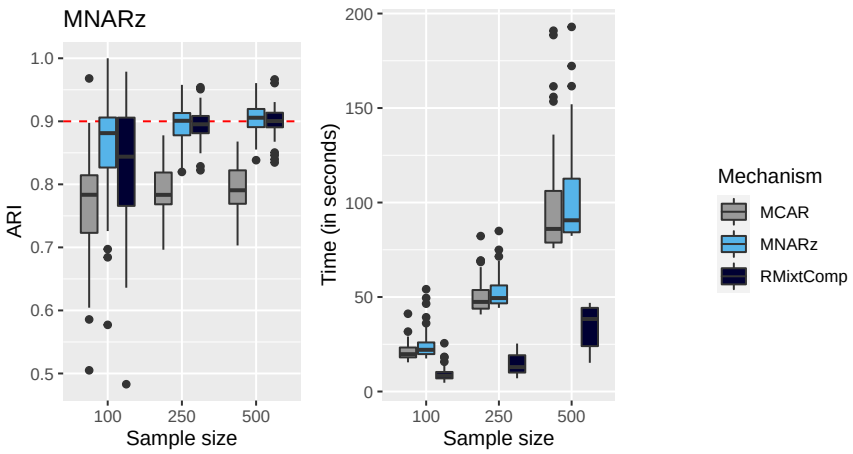
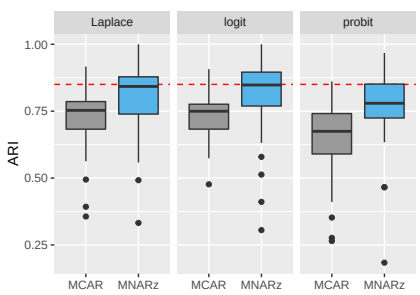
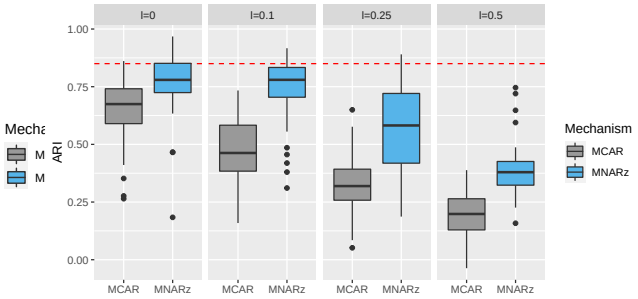


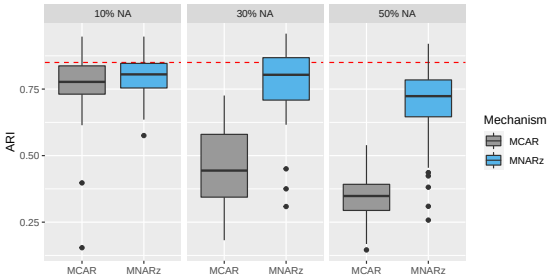
Figure 8: Left graphic: Boxplot of the ARI obtained for 50 samples composed of $d = 6$ variables and $n \in \{100, 250, 500\}$. In grey: our EM implementation for MCAR data, in blue: our EM implementation for MNARz data, in green: SEM algorithm of **RMixtComp** for MCAR data using the augmented data matrix. Right graphic: associated computational times (in seconds).



(a) Impact of the misspecification of the link function



(b) Impact of the misspecification of the data distribution



(c) Impact of the percentage of missing values

Figure 9: 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%.

F Appendix 6: Traumabase dataset

F.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,

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

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).

F.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 (Buuren and Groothuis-Oudshoorn, 2010) (Mice). We introduce some additional

missing values in three quantitative variables (*TCD.PI.max, Shock.index.ph, FiO2*) by using the MNARz mechanism (9). 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 10, our clustering algorithm, designed for the MNAR setting, gives a significantly smaller error than other methods.

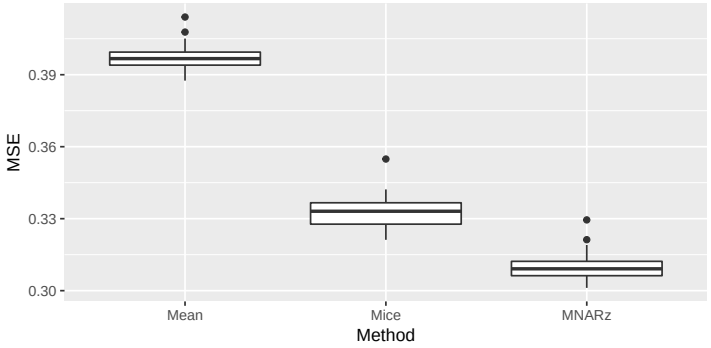


Figure 10: Mean squared error of the imputation task for the Traumabase dataset.

F.3 Description of the variables in the Traumabase dataset

A description of the variables which are used in Section 5 is given. Figure 11 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).
- *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.

G Appendix 7:Complements on generic experiments

This section gives the values of δ (see (16)) ψ (see (3)) and φ (see (16)) 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.

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

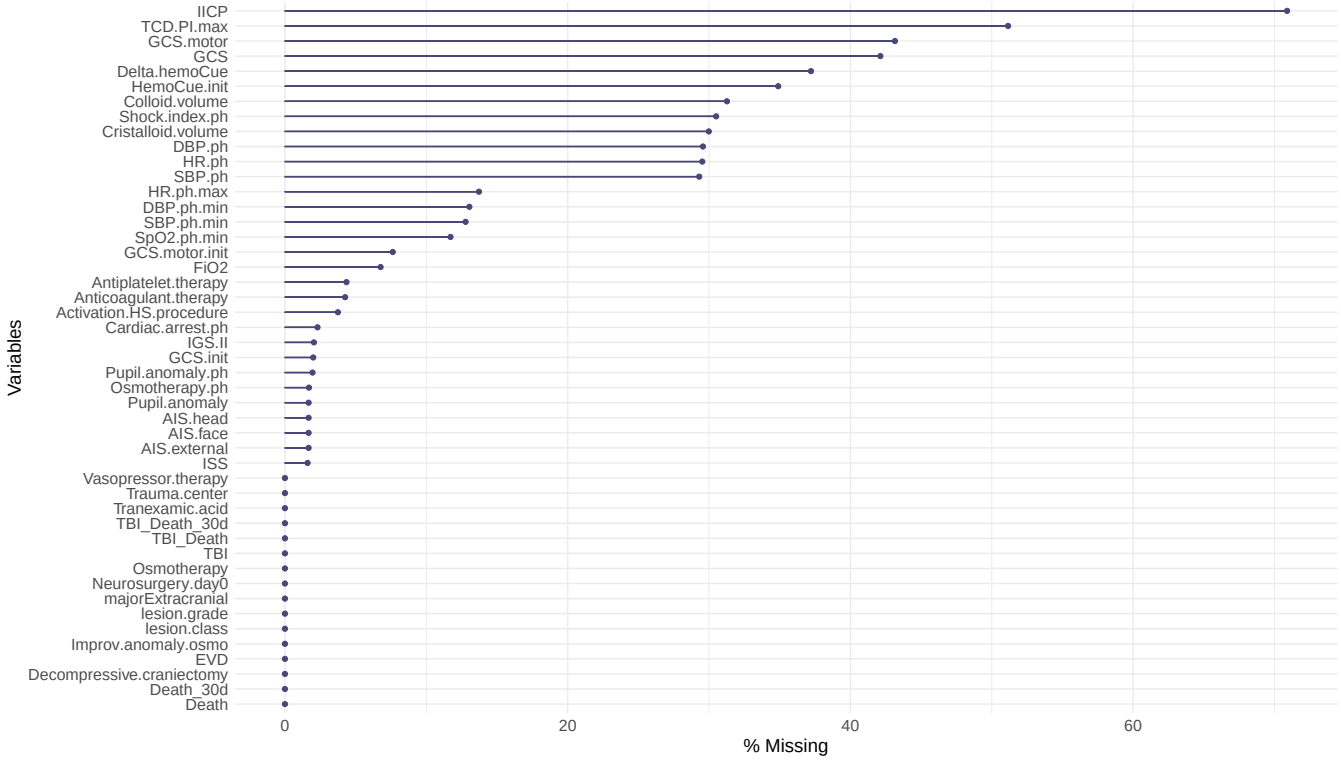


Figure 11: Percentage of missing values per variable for the Traumabase dataset.

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

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

Table 4: Choice of the values of δ and α for all the experiments of Section 4.2 and Appendix E 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 (16) and α in (3).

d	δ	α	β
3	3.5	-1.56	$\begin{pmatrix}1.45 & 0.2 & -3\end{pmatrix}$
6	2.25	-0.7	$\begin{pmatrix}-3 & 0.3 & -3 & -3 & -2 & 1\end{pmatrix}$
9	1.98	-0.68	$\begin{pmatrix}0.5 & 0.1 & -1.2 & 0.4 & -0.1 & -1.3 & 0.3 & -0.1 & -1\end{pmatrix}$

Table 6: 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	$\begin{pmatrix}-1.2 & -0.8 & -0.5\end{pmatrix}$	$\begin{pmatrix}-3 & 0.3 & 1\end{pmatrix}$
6	2.12	$\begin{pmatrix}-1.35 & -0.29 & 0\end{pmatrix}$	$\begin{pmatrix}-3 & 0.3 & -3 & -3 & -2 & 1\end{pmatrix}$
9	1.71	$\begin{pmatrix}-1.34 & -0.34 & 0\end{pmatrix}$	$\begin{pmatrix}-3 & 0.3 & -3 & -2.8 & -2 & 1 & 0.2 & 0.1 & 0.4\end{pmatrix}$

Table 7: Choice of the values of δ, α and β for all the experiments of Section 4.2 for the MNARyz mechanism.

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 8: 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.

d	δ	α	β
6	1.92	-0.75	$\begin{pmatrix}-3 & 0.3 & -3 & -3 & -2 & 1\\ 0.5 & -2 & 1 & 1 & 1 & 0.5\\ 1 & 1 & 0.5 & 0.5 & 0.5 & 2\end{pmatrix}$

Table 9: Choice of the values of δ, α and β for all the experiments of Section 4.2 for the MNARy^k mechanism.

d	δ	α	β
6	1.91	$\begin{pmatrix}-0.9 & -0.15 & 0\end{pmatrix}$	$\begin{pmatrix}-3 & 0.3 & -3 & -3 & -2 & 1\\ 0.3 & -3 & 0.3 & -0.3 & -2 & 0.2\\ -3 & 0.3 & -3 & -3 & -2 & 1\end{pmatrix}$

Table 10: Choice of the values of δ, α and β for all the experiments of Section 4.2 for the MNAR$y^k z$ mechanism.

d	δ	α	β
6	2.15	$\begin{pmatrix} -1.4 & -1.4 & -1.2 & -1.1 & -1 & -0.9 \\ -0.6 & 0.4 & 0.4 & 0.3 & 0.1 & 0.1 \\ -0.8 & -0.8 & 0.8 & -0.8 & -0.8 & 0.8 \end{pmatrix}$	$\begin{pmatrix} -3 & 0.3 & -3 & -3 & -2 & 1 \end{pmatrix}$

Table 11: Choice of the values of δ, α and β for all the experiments of Section 4.2 for the MNARyz^j mechanism.