Model-based Clustering with Missing Not At Random Data

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

Model-based Clustering with Missing Not At Random Data

Aude Sportisse*,[1], Matthieu Marbac†,[2], Christophe Biernacki†,[3], Claire Boyer†,[4], Gilles Celeux,[5], Fabien Laporte,[6], and Julie Josse,[7]

[1]Inria Sophia Antipolis Sophia Antipolis, Université Côte d'Azur Nice, 3iA Côte d'Azur
[2]Ecole National de la Statistique et de l'Analyse de l'Information, Centre de Recherche en Economie et Statistique, CNRS, Université de Rennes 1
[3]Inria Centre de recherche Lille Nord Europe, Université Lille 1 Lille, CNRS
[4]LPSM, Sorbonne Université, Centre Inria de Paris
[5]Inria Paris Saclay, Université Paris-Sud
[6]Université Catholique de l'Ouest
[7]Inria Montpellier, Institut Desbrest de santé publique

May 13, 2022

Abstract

Traditional ways for handling missing values are not designed for the clustering purpose and they rarely apply to the general case, though frequent in practice, of Missing Not At Random (MNAR) values. This paper proposes to embed MNAR data directly within model-based clustering algorithms. 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. Eight different MNAR models are proposed, which may depend on the underlying (unknown) classes and/or the values of the missing variables themselves. We prove the identifiability of the parameters of both the data distribution and the mechanism, whatever the type of data and the mechanism, and propose an EM or Stochastic EM algorithm to estimate them. The code is available on `https://github.com/AudeSportisse/Clustering-MNAR`. We also prove that MNAR models for which the missingness depends on the class membership have the nice property that the statistical inference can be carried out on the data matrix concatenated with the mask by considering a MAR mechanism instead. Finally, we perform empirical evaluations for the proposed sub-models on synthetic data and we illustrate the relevance of our method on a medical register, the TraumaBase® dataset.

Keywords. Model-based Clustering, Missing Not At Random (MNAR) Data, Identifiability, EM and Stochastic EM Algorithms, Medical Data.

Contents

*Corresponding author: `aude.sportisse@inria.fr`
†Equal contribution.

1 Introduction

Clustering remains a pivotal tool for readable analysis of large datasets, offering a consistent summary of datasets by grouping individuals. In particular, the model-based paradigm [McLachlan and Basford, 1988, Zhong and Ghosh, 2003, 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 multifaceted 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, since the more data we have, the more missing values we can expect to have. 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. On the one hand, if we delete the individuals having missing values, the remaining observations can form a small sample subset which increases the variance of the estimates. Moreover, this subsample can be a biased subset of the population that lead to bias estimator when it is used for inference. On the other hand, if single imputation is used, the additional variability due to missing values is not taken account into subsequent analysis. Furthermore, neither of both strategies is specifically designed for the final clustering task. As an alternative, one can consider likelihood approaches, using, for instance, Expectation Maximization (EM) type algorithms [Dempster et al., 1977]. We detail such an approach in this paper and develop some clustering methods able to deal with informative missing data in an efficient way.

Notations and typology of the missing values mechanisms To correctly define the missing values mechanisms, some notations must be introduced. The full dataset consists of n individuals $Y = (\mathbf{y}_1|\dots|\mathbf{y}_n)^T$, 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 for the full dataset is denoted by $C = (\mathbf{c}_1|\dots|\mathbf{c}_n)^T \in \{0,1\}^{n\times d}$, $\mathbf{c}_i = (\mathbf{c}_{i1}, \dots, \mathbf{c}_{id})^T \in \{0,1\}^d$ being the indicator pattern of missing data for the individual $i \in \{1, \dots, n\}$: $\mathbf{c}_{ij} = 1$ indicates that the value y_{ij} is missing and $\mathbf{c}_{ij} = 0$ otherwise. The values of the observed variables for individual i are denoted by $\mathbf{y}_i^{\text{obs}}$. Similarly, the values of the missing variables for individual i are denoted by $\mathbf{y}_i^{\text{mis}}$. In addition, in a clustering context, the target 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 $\mathbf{z}_i$.

Rubin [1976] distinguishes three missing value mechanisms, namely Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing not at Random (MNAR). Missing data are MCAR when the missingness is independent of all the values, missing or not, and thus can be formalized by $\mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i, \mathbf{z}_i; \psi) = \mathbb{P}(\mathbf{c}_i; \psi)$, for all (missing or observed) values $(\mathbf{y}_i, \mathbf{z}_i)$, ψ generically designating a parameter of the multinomial pdf on $\mathbf{c}_i$. Missing data are MAR when missingness is independent of missing values, even if possibly depending on some (or all) observed values, which means that $\mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i, \mathbf{z}_i; \psi) = \mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}; \psi)$ for all missing values $(\mathbf{y}_i^{\text{mis}}, \mathbf{z}_i)$. The M(C)AR mechanisms are said to be ignorable, because estimating the parameter of the data distribution Y does

not require the modelisation of $\mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i, \mathbf{z}_i; \psi)$, considering that ψ is a nuisance parameter. Finally, MNAR corresponds to a missing-data mechanism that is not MCAR or MAR. For such missing data, the observed variables are not representative of the population. It is well known that the MNAR mechanism is nonignorable when the goal is to estimate the parameters of the mixture model [Little and Rubin, 2019]. The MNAR mechanism is actually also not-ignorable when the aim is to recover the partition of the data. Therefore, as the MNAR mechanism is neither ignorable for the density estimation, nor for the clustering, dealing with such data does require the specific modeling effort of $\mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i, \mathbf{z}_i; \psi)$.

MNAR data In this paper, the data are supposed to be MNAR, which is very frequent in practice [Ibrahim et al., 2001, Mohan et al., 2018]. Examples may include surveys where rich people would be less willing to disclose their income or clinical data collected in emergency situations, where doctors may choose to treat patients before measuring heart rate. In both cases, the missingness of income or heart rate depends on the missing values themself. The missing-data mechanism must generally be taken into account [Little and Rubin, 2019] by considering the joint distribution of the data and the missing-data pattern. 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 *i.e.* $\mathbb{P}(\mathbf{y}_i, \mathbf{c}_i|\mathbf{z}_i) = \mathbb{P}(\mathbf{y}_i|\mathbf{z}_i)\mathbb{P}(\mathbf{c}_i|\mathbf{y}_i, \mathbf{z}_i)$ (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 *i.e.* $\mathbb{P}(\mathbf{y}_i, \mathbf{c}_i|\mathbf{z}_i) = \mathbb{P}(\mathbf{c}_i|\mathbf{z}_i)\mathbb{P}(\mathbf{y}_i|\mathbf{c}_i, \mathbf{z}_i)$. 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.

Related works 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.

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 cannot identify difficult cluster structures, since it relies on strong assumptions as equal proportions between clusters. De Chaumaray and Marbac [2020] have proposed to perform clustering via a semiparametric mixture model using the pattern-mixture approach to formulate the joint distribution, which makes the method not suitable for estimating the density parameters or imputing missing values. For longitudinal data, Beunckens et al. [2008], Kuha et al. [2018] jointly model the measurements and the dropout process by using an extension of the shared-parameter model, which is specific approach to deal with MNAR mechanisms, by assuming that both the data and the dropout process depend on shared latent variables. They introduce for this a latent-class mixture model allowing classification of the subjects into latent

groups. 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, and specifically in selection models, the main challenge to overcome consists in proving the identifiability of the parameters of both 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. In this paper, their idenfiability results are extended to more complex missing scenario and to the multivariate case.

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 functions 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 on the class membership have identifiable parameters. For each model or submodel, an EM or Stochastic EM algorithm is proposed, implemented, and made available for reproducibility. 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 $[\mathbf{Y}, \mathbf{C}]$ considering the MAR mechanism instead; this is a real advantage, especially since the missing-data mechanism does not have to be modeled in this case. This also gives a theoretical interpretation of this approach often used in practice.

Outline of the paper The rest of the article is organized as follows. We introduce the model-based clustering in the presence of missing data in Section 2 and propose an exhaustive zoology of the possible MNAR specifications in this framework, for which the identifiability issue is addressed. We propose an estimation strategy in Section 3. Section 4 is devoted to numerical experiments with synthetic data in order to assess the performance of our methods. In Section 5, our method is finally illustrated on a public health application, the TraumaBase® dataset. Section 6 concludes this paper and provides some perspectives.

2 Missing data in model-based clustering

2.1 Mixture models

Mixture models permit to achieve the clustering aim by modeling the distribution of the observed data $(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i)$. This distribution can be obtained from the distribution of the couple $(\mathbf{y}_i, \mathbf{c}_i)$ that is supposed to be a mixture model with K components. Thus, using the model selection decomposition

for each component, the probability distribution function (pdf) of the couple $(\mathbf{y}_i, \mathbf{c}_i)$ is

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

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

One may expect that the individual data $\mathbf{Y}$ convey more information on the partition $\mathbf{Z}$ than the pattern $\mathbf{C}$ of missing data. Thus, it is hazardous to allow the missing data modeling to be more complex than the mixture model itself. Consequently, we assume that the elements of c_i are conditionally independent given $(\mathbf{y}_i, \mathbf{z}_i)$ and that the c_{ij} and $y_{ij'}$ are conditionally independent given $(y_{ij}, \mathbf{z}_i)$ for $j \neq j'$. This 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). In addition, the conditional distribution of c_{ij} given $(y_i, \mathbf{z}_i)$ is assumed to be a generalized linear model with link function ρ, so that finally

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

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 (hence the name of self-masked mechanisms) which depends on the class k as well.

The most general model that we consider is called MNAR$y^k z^j$ and is defined by (3). We now propose some gradual variants of this core model by decreasing their complexity, while always highlighting their associated interpretation.

2.2 Decreasing the complexity of the missingness mechanism models

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 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 MNARy^kz 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})

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

We can consider that the effects on a particular variable and on the class membership are 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; the missing mechanism therefore becomes self-masked only. 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, 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,

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

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 one with no dependence on the variables, neither on the class membership, which boils down to MCAR values, *i.e.* each value has the same probability to be missing,

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

2.3 About the specificity of some proposed models

The MNARz model given in (9) is the simplest of the MNAR models previously listed. Generally speaking, this model assumes that the proportion of missing values can vary among clusters. However, behind its apparent simplicity, it benefits from interesting properties. Although MNARz does

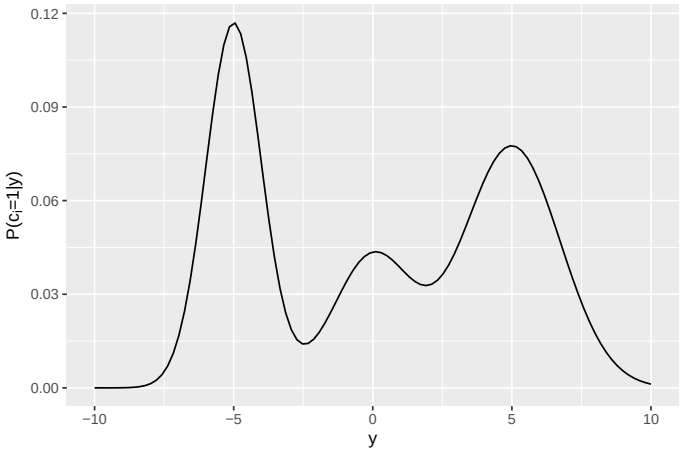

Figure 1: Illustration of the dependency between $\mathbf{c}_i$ and $\mathbf{y}_i$ in a MNARz model by drawing $\mathbb{P}(c_i \mid \mathbf{y}_i; \pi, \lambda, \psi)$ for a three-component univariate Gaussian model with mixing proportions $\pi_1 = \pi_2 = 0.3$ and $\pi_3 = 0.4$, with centers $\mu_1 =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

not directly involve $\mathbf{y}_i$ in its ground definition (9), the pattern $\mathbf{c}_i$ can be related to $\mathbf{y}_i$ through $\mathbf{z}_i$. This is illustrated in Figure 1.

Finally, it is important to mention that MNARz and MNARz^j can be turned into a MAR-like strategy, commonly used in the machine learning community [Josse et al., 2019], by working on the concatenated dataset $\tilde{\mathbf{Y}}^{\text{obs}} = (\mathbf{Y}^{\text{obs}}, \mathbf{C})$. In Proposition 1, we prove that the mixture model associated to this augmented dataset $\tilde{\mathbf{Y}}^{\text{obs}}$ with a MAR missing mechanism is equivalent to the mixture model for $\mathbf{Y}^{\text{obs}}$ given in (1) assuming a MNARz or MNARz^j model for $\mathbf{C}$. The proof of this proposition is given in Appendix A.

Proposition 1. *Consider the dataset* $(\tilde{\mathbf{y}}_1^{\text{obs}}, \ldots, \tilde{\mathbf{y}}_n^{\text{obs}})$ *such that* $\tilde{\mathbf{y}}_i^{\text{obs}} = (\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i)$ *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 than the distribution for* $\mathbf{y}_i^{\text{obs}}$ *with the mixture model (1) under the MNARz assumption (9) and MNAR*z^j *assumption (10).*

In particular, Proposition 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). As a matter of fact, Proposition 1 is the first theoretical result in unsupervised learning in line with the intuition, developed in Josse et al. [2019] for supervised learning, that working with MAR strategies on the data set augmented by the missing patterns can actually tackle certain types of MNAR settings.

2.4 Impact of the missingness mechanism on the clustering

MNAR is strictly speaking non-ignorable with the classical definition [Little and Rubin, 2019] but one could check whether it is ignorable for clustering, which is the 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. Thus, the missingness process is said to be *ignorable for clustering* if

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

This is equivalent to having

$$\forall(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i),\ \frac{\pi_k \int f_k(\mathbf{y}_i; \lambda_k) f_k(\mathbf{c}_i \mid \mathbf{y}_i; \psi_k) d\mathbf{y}_i^{\text{mis}}}{\int \sum_{\ell=1}^{K} \pi_\ell f_\ell(\mathbf{y}_i; \lambda_\ell) f_\ell(\mathbf{c}_i \mid \mathbf{y}_i; \psi_\ell) d\mathbf{y}_i^{\text{mis}}} = \frac{\pi_k f_k(\mathbf{y}_i^{\text{obs}}; \lambda_k)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\mathbf{y}_i^{\text{obs}}; \lambda_\ell)}.$$

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 $\mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i, \mathbf{z}_i; \psi)$.

2.5 Identifiability of the model parameters

This section gives sufficient conditions for the generic identifiability of the parameters for continuous, count, categorical, and mixed data (*i.e.,* the space where the parameters are not identifiable has a Lebesgue measure equal to zero; see Allman et al. [2009]).

Proposition 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 that 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 satisfy* ***A1.*** *and* ***A2.****,* ***A4.*** *hold 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 whole proof and assumptions are detailed in Appendix B. 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 minimum number of dimensions and the number of classes. Finally, in both cases, Assumption **A3.**, which requires that the missing data mechanism is strictly monotone, is made.

The proof for continuous and count variables 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 proof

uses Corollary 5 of Allman et al. [2009] which gives the identifiability of finite mixtures of Bernoulli products. The identifiability of mixed data directly follows from the identifiability of continuous and categorical components.

3 Estimation of the proposed MNAR models

The parameters identifiability makes the estimation procedure possible and sound. However, MNAR models are not ignorable and they require thus a specific inference procedure for estimating the parameters π, λ and ψ. This section gathers the description of the EM and SEM algorithms for Gaussian, Poisson, multinomial and mixed data with MNAR models in view of maximum likelihood estimation. Details of the algorithms are given in Appendix C. These iterative algorithms require to introduce the complete-data log-likelihood defined by

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

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

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

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

and

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

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

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

and

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

Note that the difficulty of the computation of quantities defined at the E-step as well as the difficulty of the maximization problem leading to $\lambda_k^{[r]}$ and $\psi_k^{[r]}$, depend on the MNAR model at hand. These steps are straightforward with the MNARz and MNARz^j models (see (9) and (10)) but more difficult with all the other MNAR models called in the sequel MNAR$y*$ (modelling the effect of missingness depending on y).

Note that computing $\tau_y(\lambda_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]})$ requires to integrate over $\mathbf{y}_i^{\text{mis}}$ by considering its conditional distribution given $(\mathbf{y}_i^{\text{obs}}, z_{ik} = 1, \mathbf{c}_i)$ and the parameter $\theta^{[r-1]}$. For MNARz and MNARz^j models, this conditional distribution is equal to the conditional distribution of $\mathbf{y}_i^{\text{mis}}$ by the dependence of y given $(\mathbf{y}_i^{\text{obs}}, z_{ik} = 1)$ and the parameter $\theta^{[r-1]}$. This makes $\tau_y(\lambda_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]})$ easy to be computed in most cases (see Appendix C.1.1 and C.1.2 for both Gaussian and categorical data). The MNAR$y*$ models consider the effect of the missingness depending on y and lead then to unfeasible computations. The conditional distribution of $\mathbf{y}_i^{\text{mis}}$ given $(\mathbf{y}_i^{\text{obs}}, z_{ik} = 1, \mathbf{c}_i)$ and the parameter $\theta^{[r-1]}$ is explicit in some cases (if the variables are Gaussian, it is a truncated Gaussian as shown in Appendix C.2.1) if the missing-data distribution ρ is probit but it is not a known distribution if ρ is logistic (it would require the use of sampling algorithm, as the Sampling Importance Resampling algorithm Gordon et al. [1993], that are time costly). However, to our knowledge, for both forms of missing-data distributions, $\tau_c(\psi_k; \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i, \theta^{[r-1]})$ and $t_{ik}(\theta^{[r-1]})$ do not have closed forms. In addition, the maximization problem leading to $\psi_k^{[r]}$ is a delicate issue because the function involved is not concave.

3.2 The SEM algorithm for overpassing the EM intractability's

Some distributions lead to computation of untractable integrals at the E-step (*e.g.,* gaussian components with MNAR$y*$ mechanism defined with logit link). In such case, the SEM algorithm [Celeux and Diebolt, 1985] could avoid this difficulty, by imputing missing values using a Gibbs sampling instead of integrating over them. In addition, it has another possible advantage over the EM algorithm since it is not trapped by the first local maximum encountered of the likelihood function [Celeux and Diebolt, 1985]. The SEM algorithm modifies the E-step of the EM algorihtm by considering a stochastic-E step (SE-step) while the M-step is unchanged. Thus, at the iteration $[r]$, the E-step is replaced by the following SE-step:

SE-step: Draw the missing data $(\mathbf{z}_i^{[r]}, \mathbf{y}_i^{\text{mis}\,[\text{r}]})$ according to their conditional distribution given the observed data $(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i)$ and the current parameter $\theta^{[r-1]}$. Since it is not convenient to simulate this conditional distribution, we simulate instead the following two easier conditional probabilities using a Gibbs sampling approach:

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

where $\mathbf{y}_i^{[r]} = (\mathbf{y}_i^{\text{obs}}, \mathbf{y}_i^{\text{mis}[r]})$.

The sampling of $\mathbf{z}_i^{[r]}$ is performed by a multinomial distribution whose probabilities of events are defined by $\pi_k^{[r-1]} f_k(\mathbf{y}_i; \lambda_k^{[r-1]}) f_k(\mathbf{c}_i \mid \mathbf{y}_i; \psi_k^{[r-1]}) / f(\mathbf{y}_i, \mathbf{c}_i; \theta^{[r-1]})$ for $k = 1, \ldots, K$.

Note that the conditional distribution of $\mathbf{y}_i^{\text{mis}}$ given $(\mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1, \mathbf{c}_i)$ parameterized by $\theta^{[r-1]}$ is defined by

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

This distribution may not be classical, in general. For the MNAR$y*$ models, the conditional distribution $\mathbf{y}_i^{\text{mis}}$ given $(\mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1, \mathbf{c}_i)$ 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. In Appendix C, Table 4 gives the cases for which the EM or SEM algorithm is feasible.

4 Numerical experiments on synthetic data

The performances of our methods are first illustrated by some numerical experiments on synthetic data. In Section 4.1, we show that clustering can leverage from MNAR missing values, by using the percentage of missing values per class. Section 4.2 focuses on the MNARz mechanism. Its robustness to misspecification of the link function, the data distribution, and the percentage of missing values, alternatively, is then addressed in Section 4.3.

Measuring the performance 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. 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 detail when missing data. For a model $\mathcal{M}$ with $\nu_{\mathcal{M}}$ parameters, the maximum likelihood estimators are denoted as $\hat{\theta}_{\mathcal{M}}$ and $\ell(\theta; \mathbf{Y}^{\text{obs}}, \mathbf{C})$ is the observed log-likelihood. One has

$$\text{ICL}(\mathcal{M}) = \ell(\hat{\theta}_{\mathcal{M}}; \mathbf{Y}^{\text{obs}}, \mathbf{C}) - \frac{\nu_{\mathcal{M}}}{2} \ln n + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik}^{\text{MAP}}(\hat{\theta}_{\mathcal{M}}) \log(\mathbb{P}(z_{ik} = 1 | \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \hat{\theta}_{\mathcal{M}})), \tag{15}$$

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

In addition, the Adjusted Rand Index (ARI) [Hubert and Arabie, 1985] can be computed between the true partition $\mathbf{Z}$ and the estimated one. Obviously other strategies are possible to select a useful mixture model (see Celeux et al. [2019]).

4.1 Leveraging from MNAR data in clustering illustration

MNAR data are often considered a real obstacle for statistical processing. Yet, this first numerical experiment illustrates that the MNAR mechanism may help performing the clustering task. Indeed, let us consider a bivariate isotropic Gaussian mixture model with two components and equal mixing proportions. The difference between the centers of both mixture components is taken as $\Delta_\mu = \mu_{21} - \mu_{11} = \mu_{22} - \mu_{12} \in \{0.5, 1, \dots, 3\}$. This cluster overlap controls the mixture separation, which can vary from a low separation ($\Delta_\mu = 0.5$) to a high separation ($\Delta_\mu = 3$). By considering the MNARz mechanism (9), one can play on the discrepancy between inter-cluster missing proportions $\Delta_{\text{perc}} =$

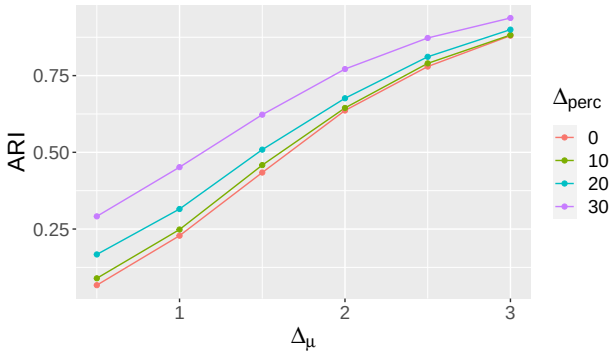

Figure 2: Relative effect of both the separation strength Δ_μ of the mixture component and the MNAR evidence Δ_{perc} on theoretical ARI. For example, if $\Delta_{\text{perc}} = 10\%$ (green line), it means that the second class has 10% more missing values than the first class.

$|\text{perc}_2 - \text{perc}_1|$, by making it varies in $\{0, 0.1, 0.2, 0.3\}$[1]. Increasing values of Δ_{perc} correspond to an increase in 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. Finally, 15% missing values are introduced regardless of MNAR evidence Δ_{perc} and mixture separation Δ_μ. Figure 2 gives the theoretical ARI (*i.e.* we compute the ARI with the theoretical parameters) as a function of the cluster overlap Δ_μ and the MNAR evidence Δ_{perc}. Although the good classification rate is mainly influenced by center separation Δ_μ, it also increases with the MNAR evidence Δ_{perc}. This toy example illustrates how clustering can leverage from MNAR missing values, generally considered a true hindrance for any statistical analysis.

4.2 Generic experiments

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

- the EM algorithm (designed for the MCAR, MNARz and MNARz^j settings, resp. defined in (11), (9) and (10)),
- the SEM algorithm (designed for MNARy, MNARy^k, MNAR$y^k z^j$, MNARyz, MNAR$y^k z$ and MNARyz^j settings, resp. defined in (7), (8), (3), (6), (5) and (4)),
- a two-step heuristics which consists of first imputing the missing values using multiple imputations by chained equations [Buuren and Groothuis-Oudshoorn, 2010] to get M completed datasets. This algorithm is called `Mice`. Then, classical model-based clustering is performed on each completed dataset, for which the performance is measured. The final performance of this method is computed with the mean. In simulations, we do not systematically consider this method, which is 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 such

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

that:

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

with $\epsilon_{ij} \sim \mathcal{N}(0, 1)$ the noise term, $\varphi_k \in \{0, 1\}^d$ and $\delta > 0$. We introduce missing values with a MNAR model (see (3)), using a probit link function and its associated parameters. The choice of the parameters ψ_k of the missing-data mechanism and δ allows to control the rates of misclassification and missingness (their values for each experiment are given in Appendix F). All the simulations have been performed for a theoretical rate of misclassification of 10% and a theoretical missing rate in the whole dataset of 30%.

Consistency of the estimators We first assess the consistency of the estimators for each of the MNAR settings. We consider $d = 6$ variables and we vary the number of observations $n = 100, 250, 500$. For $d = 6$, we fix $\varphi_{11} = \varphi_{22} = \varphi_{33} = \varphi_{14} = \varphi_{25} = \varphi_{36} = 1$ and the others $\varphi_{kj} = 0$. Figure 3 (left graphic) presents the boxplot of the ARI for each scenario. First, as expected, considering the mechanism always gives better results than using the MCAR model. This is especially true 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. Figure 4 focuses on the MNARz mechanism. We compare 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 $(\mathbf{Y}|\mathbf{C})$. As expected (see Proposition 1), the SEM algorithm designed for MNARz data and the one designed for MCAR data using the augmented data matrix give similar results. 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.

Computation time The computation times for these numerical experiments are given in Figure 5 and Figure 4 (right graphic), which focuses on the MNARz mechanism. There is a huge difference of computation times between the settings using the EM algorithm (MCAR, MNARz, MNARz^j) and the ones requiring the SEM algorithm, these latter being expansively time-consuming. We see that the SEM algorithm of **RMixtComp** is the fastest method, which makes Proposition 1 a key result. Indeed, it opens the way to the use of an already optimized (coded in C) algorithm.

Impact of the dimension In this experiment, we vary the number of variables ($d = 3$, $d = 6$ and $d = 9$) and consider $n = 100$ observations. For $d = 3$, we fix $\varphi_{11} = \varphi_{22} = \varphi_{33} = 1$ and the others $\varphi_{kj} = 0$. For $d = 6$, we retrieve the previous parameter setting dedicated to this case. For $d = 9$, we fix $\varphi_{11} = \varphi_{22} = \varphi_{33} = \varphi_{14} = \varphi_{25} = \varphi_{36} = \varphi_{17} = \varphi_{28} = \varphi_{39} = 1$ and the others $\varphi_{kj} = 0$. The missing values are sequentially introduced with a MNARz^j, MNARy, MNARy, MNARyz and MNAR$y^k z^j$ model. 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 (Mice). Figure 6 shows the boxplot of the ARI for each scenario. First, the methods considering an MNAR mechanism (MNAR$\star$) always outperform the competing methods, the one considering 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

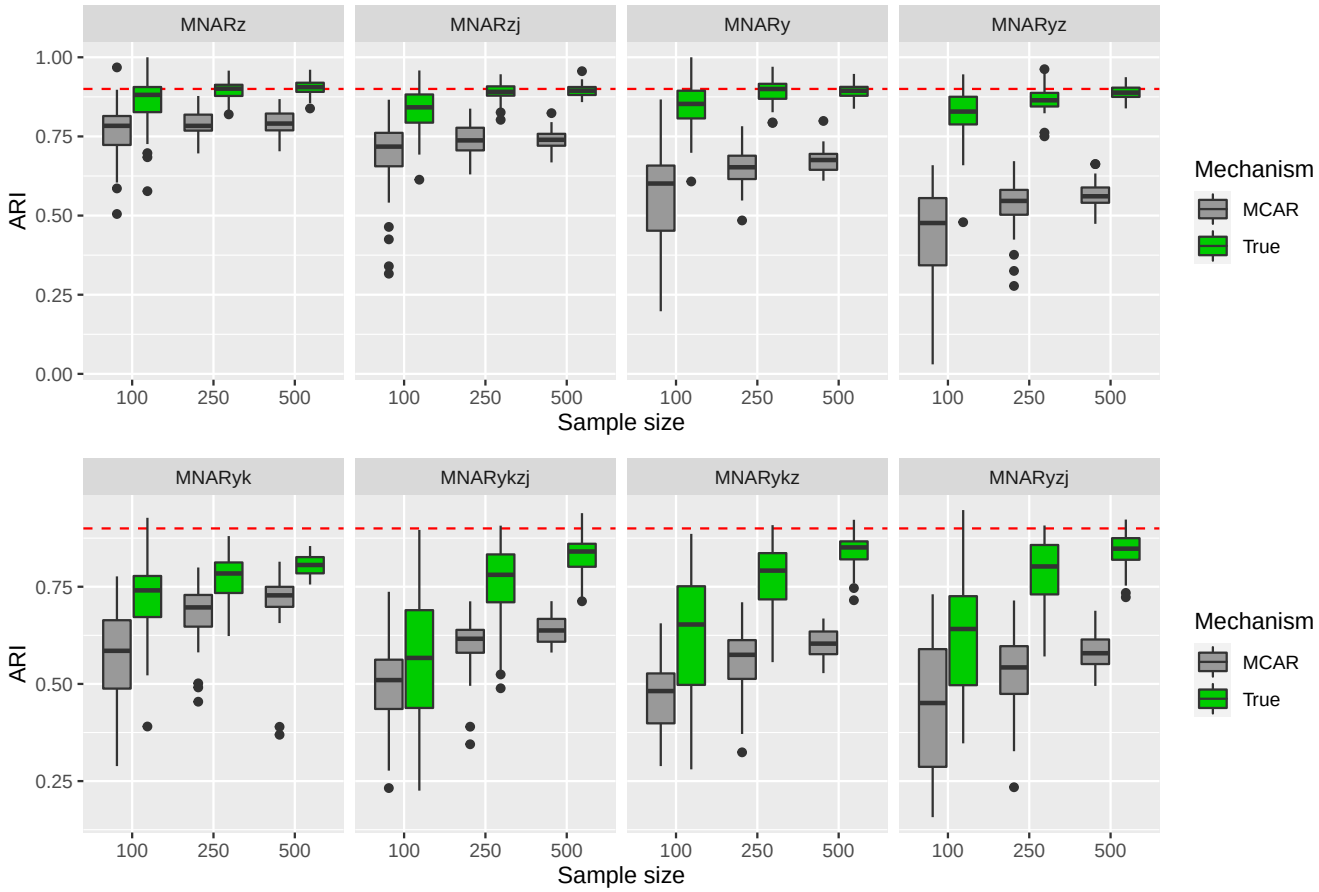

Figure 3: Boxplot of the ARI obtained for 50 samples composed of $d = 6$ variables with a misclassification rate of 10% and a missing-data rate of 30% in the whole dataset. The sample size varies by $\{100,250,500\}$. The missing values are introduced with a MNARy, MNARy^k MNARz or MNARz^j mechanism (top) and a MNAR$y^k z$, MNAR$y^k z^j$ MNARyz or MNARyz^j mechanism (bottom). The boxplot in green is the one for the algorithm considering the true MNAR$\star$ setting (noted "True" in the legend). The red line indicates the theoretical ARI.

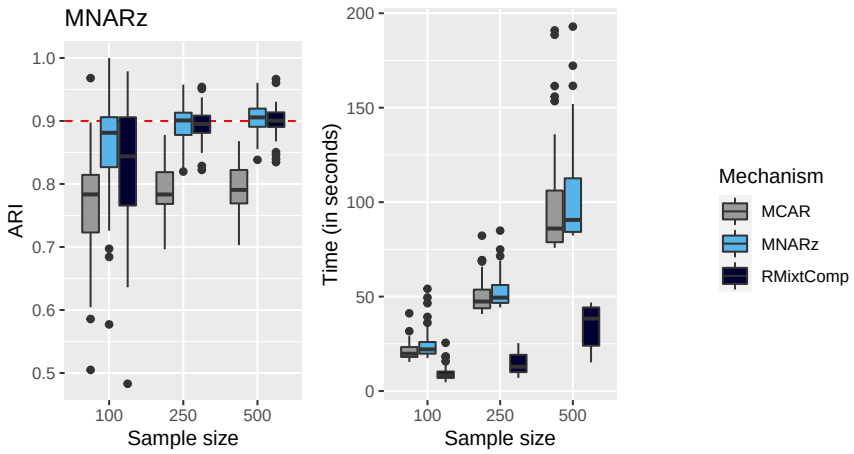

Figure 4: Left graphic: Boxplot of the ARI obtained for 50 samples composed of $d = 6$ variables with a misclassification rate of 10% and a missing-data rate of 30% in the whole dataset. Sample size varies in {100,250,500}. The missing values are introduced with a MNARz mechanism. ARI are showed in the left graphic and computational times (in seconds) in the right one (in grey: EM algorithm coded by us for MCAR data, in blue: EM algorithm coded by us for MNARz data, in green: SEM algorithm of **RMixtComp** for MCAR data using the augmented data matrix). The red line indicates the theoretical ARI. Right graphic: associated computational times (in seconds).

clear how much the MNARz setting deviates from the hypothesis (depending on the parameters chosen for the mechanism). However, the MNARz setting seems to be a compromise: It is clearly better than methods that do not consider MNAR data. It is sometimes not as good as the real setting that generated the missing data, but it allows one to overcome expansive computation time.

4.3 Focus on the MNARz mechanism

We consider the same setting as in Section 4.2 and focus now on the MNARz mechanism, because it is an interesting compromise between all the proposed MNAR mechanisms, as just discussed at the end of the previous section.

Impact of the misspecification of the link function Figure 7 shows the boxplots of the ARI for the MNARz setting and the MCAR one. The missing values are introduced using an MNARz model with different link functions (the Laplace density distribution, the logit link, and the probit link), whereas the estimation algorithm considers the probit one. The MNARz setting seems to be robust to the link function.

Impact of the misspecification of the data distribution Figure 8 shows the boxplots of the ARI for the MNARz setting and the MCAR one in another context. We consider a three-components Gaussian mixture with non-diagonal covariance matrices. For each component, the diagonal terms of the covariance matrix are $\Sigma_{ii} = 1$ and the other terms $\Sigma_{ij} = l, i \neq j$, with $l \in \{0, 0.1, 0.25, 0.5\}$, whereas the algorithms assume $l = 0$. Figure 8 shows the boxplot of the ARI for each scenario. Here, it is clear that the EM algorithm designed for MNARz data is not robust to a huge deviation ($l = 0.5$) from the hypothesis on the data distribution. For smaller deviations ($l = 0.1, 0.25$), the

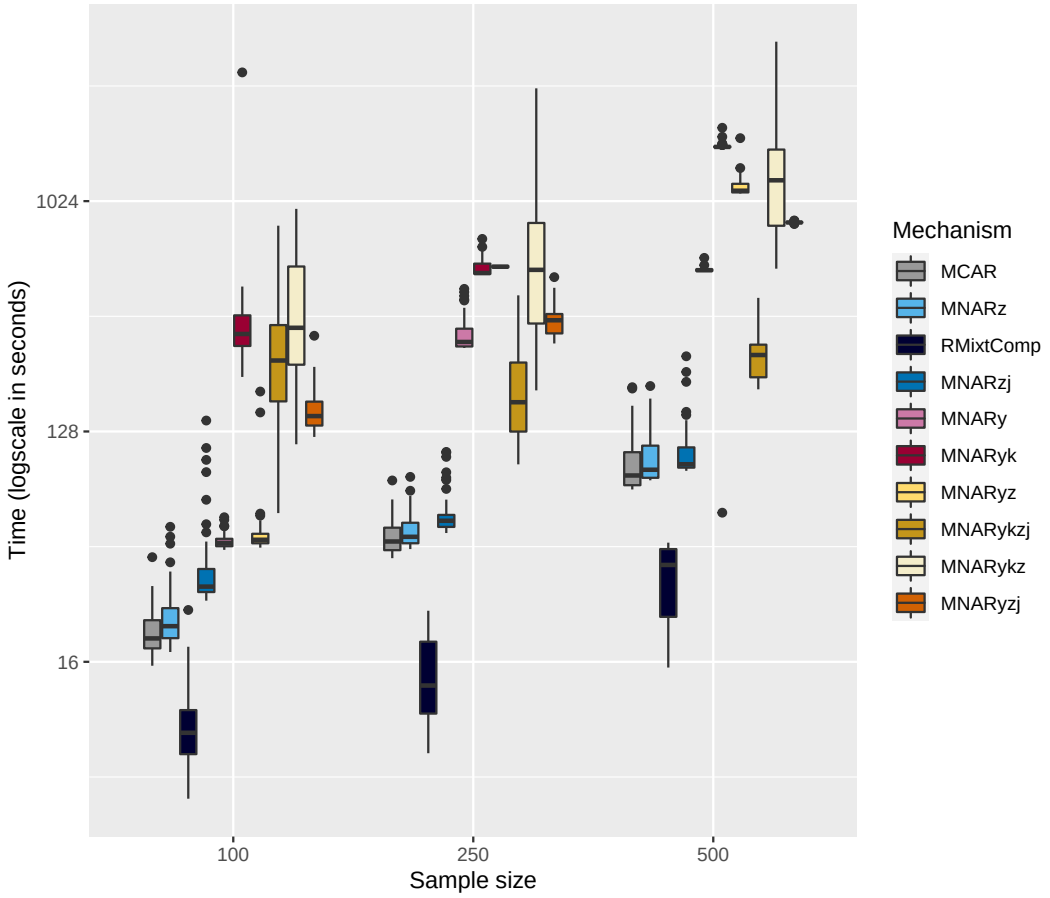

Figure 5: Boxplot of the computational times (in seconds) obtained for 50 samples composed of $d = 6$ variables with a misclassification rate of 10% and a missing-data rate of 30% in the whole dataset (see experiment on the consistency of the estimators illustrated by Figure 3).

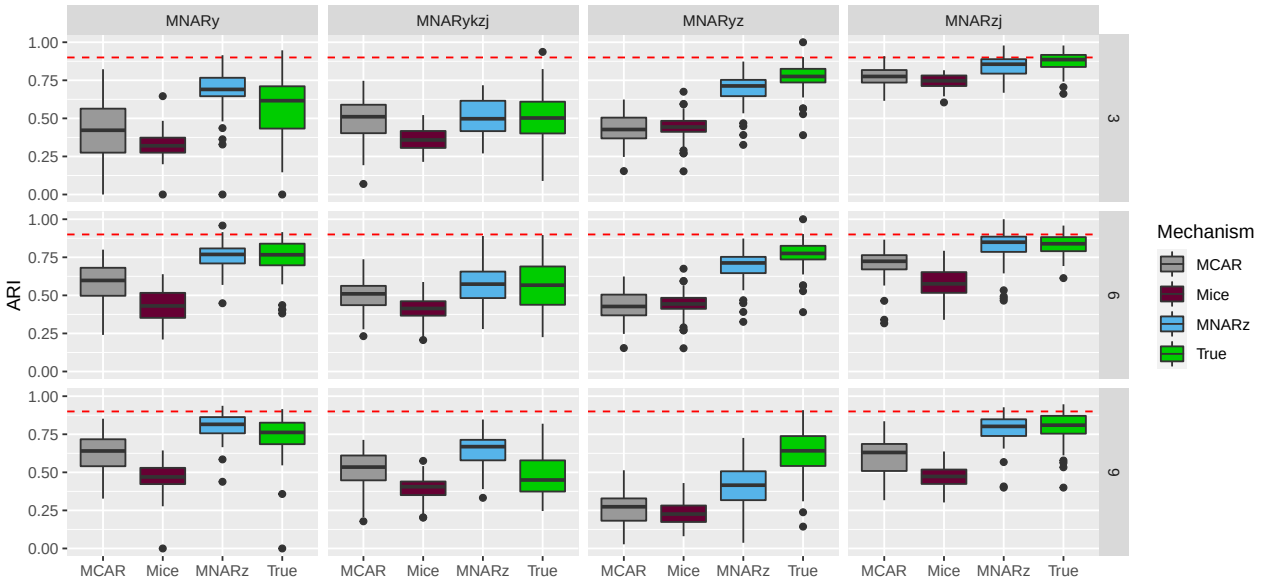

Figure 6: Boxplot of the ARI obtained for 50 samples composed of $d = 3, 6, 9$ variables and $n = 100$ observations with a misclassification rate of 10% and a missing-data rate of 30% in the whole dataset. Missing values are introduced with MNARy, MNAR$y^k z^j$ MNARyz or MNARz^j settings. The boxplot in green is the one for the algorithm considering the true MNAR$\star$ setting (noticed "True" in the legend); 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 line indicates the theoretical ARI.

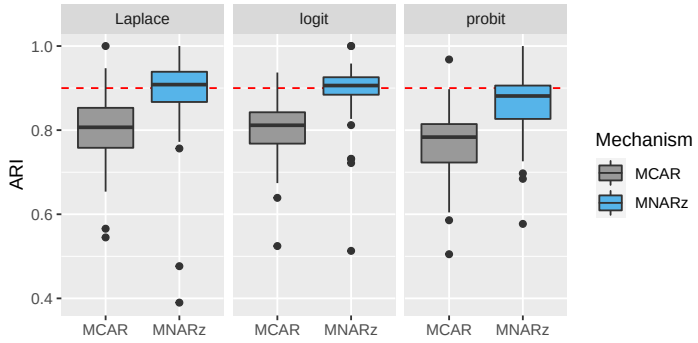

Figure 7: Boxplot of the ARI obtained for 50 samples composed of $d = 6$ variables with a misclassification rate of 10% and a missing-data rate of 30% in the whole dataset. The missing values are introduced using a MNARz setting with different link functions but only the probit link is involved for the estimation step. The red line indicates the theoretical ARI.

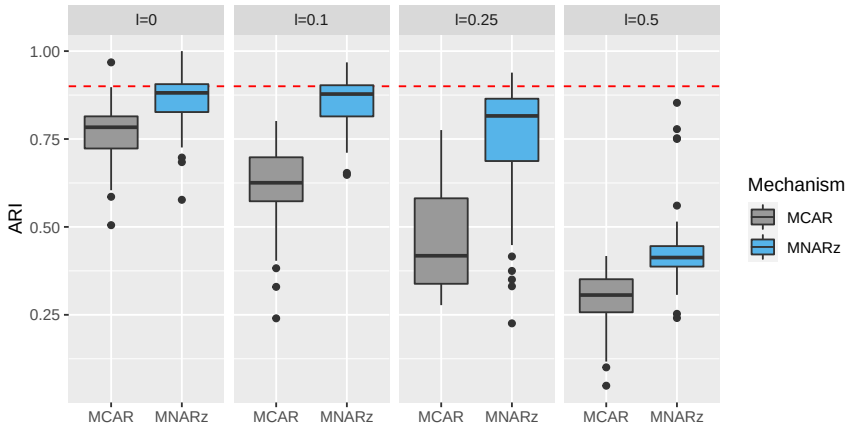

Figure 8: Boxplot of the ARI obtained for 50 samples composed of $d = 6$ variables with a misclassification rate of 10% and a missing-data rate of 30% in the whole dataset. The correlation coefficient of the covariance matrices in each component ($k = 1, 2, 3$) is $l = 0$, $l = 0.1$, $l = 0.25$ and $l = 0.5$, whereas the algorithms consider the diagonal case. The red line indicates the theoretical ARI.

results are still satisfactory and clearly outperform the ones given by the EM algorithm for the MCAR setting.

Impact of the percentage of missing values Figure 9 shows the boxplots of the ARI for the MNARz setting and the MCAR one for 10%, 30% and 50% of the missing-data rate in the entire dataset. As the percentage of missing data increases, the difference between the algorithms considering MCAR and MNARz data is greater. Even if the percentage of missing data has an impact on the algorithm considering MNARz data, it still gives results close to the theoretical ARI for a missing-data rate of 50%.

Choice of K The number of clusters was considered known until now, but it can be automatically chosen using the ICL criterion. The algorithms run with several values of the number of clusters $K = 1, 2, 3, 4$. The cluster number for the model with the highest ICL is then chosen. 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. Therefore, only the EM algorithm designed for MNARz and MCAR data can be compared. Table 5 gives the percentages of times 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 ($n = 100, 500$) for 50 repetitions. In any case, the EM algorithm for MNARz data selects the right number of classes more often. For $n = 100$, it clearly outperforms the results of the EM algorithm for MCAR data, but it has a poor percentage of selection of the right number of classes for a missing-data rate of 50%. For $n = 500$, the algorithm almost always selects the right number of classes.

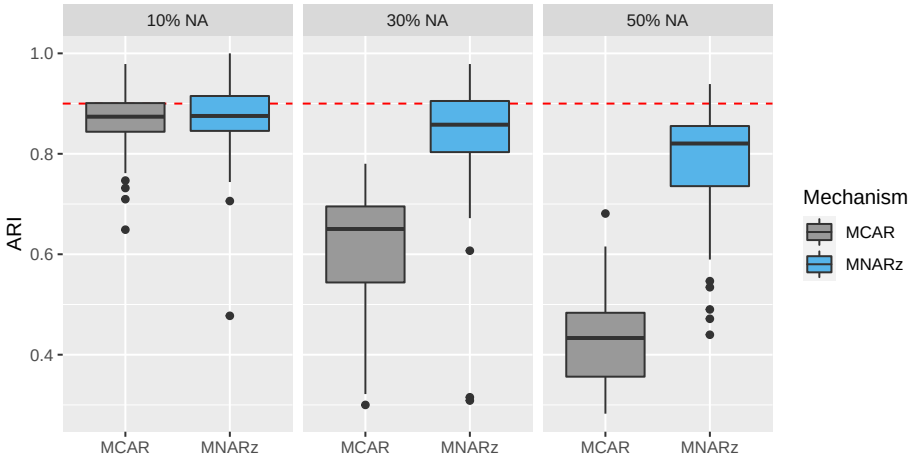

Figure 9: Boxplot of the ARI obtained for 50 samples composed of $d = 6$ variables with a misclassification rate of 10% and a missing-data rate of 10%, 30% and 50% in the whole dataset. The red line indicates the theoretical ARI.

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

Table 1: Choice of K using the ICL criterion for the EM algorithm considering MCAR data and MNARz data for a missing-data rate of 10%, 30% and 50%, $n = 100, 500$ and $d = 6$ variables. The percentages in the table indicate the number of times the correct number of classes ($K = 3$) is chosen by the ICL criterion.

Impact of the theoretical misclassification All previous experiments were conducted with a theoretical rate of misclassification of 10%. In Appendix D, we show the experiments of this section for a theoretical rate of misclassification of 15%. The same conclusions can be drawn.

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 32 quantitative and 15 categorical variables on measured on 8,248 polytraumatized patients who suffer from a major trauma (injuries from cycle or car accident). Data have been collected from 15 different hospitals. In this dataset, 11% of the data are missing and only 1,4% of the individuals are fully observed. More information on the variables can be found in Appendix E. The purpose of this real data analysis is twofold: (i) we first 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.

5.1 Classifications comparison

After discussion with doctors, some variables can be considered to have informative missing values, such as the variable *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.

In this section, the variables related to the patient death and also the hand-made classifications made by the doctors (one considers 3 groups, the other 4) are not taken into account for running the algorithms, as they were considered too informative for the classification. A total of 41 mixed variables (continuous, quantitative) can thus be used. We compare our algorithm designed for the MNARz data (9) and the MCAR data (11). Figure 10 presents the ICL values in the Traumabase dataset for different numbers of classes.

The algorithms designed for the MNARz data and MCAR data both select $K = 3$ for the number of classes. However, note that the ICL of the algorithm which considers MNARz data is always higher than the one of the algorithm for MCAR data. For $K = 3$ classes, the ARI between the classifications obtained assuming MNARz and MCAR mechanisms is 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*, which has been identified as MNAR by doctors. Table 2 gives the total variation distance between the marginal distribution of the variable *Shock.index.ph* obtained by the algorithm considering MNARz data and the one obtained by the algorithm considering MCAR data, and Table 3 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). In both cases, we can only compare the values up to label swapping. We notice that the minimum values (on the diagoals) are significantly higher than zero, which indicates that there is an influence of the MNARz mechanism both on the modeling of the data and on the classification rules.

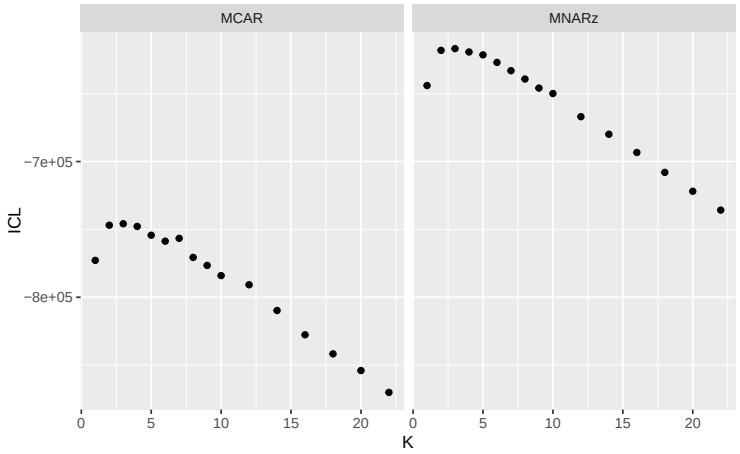

Figure 10: ICL values in the Traumabase dataset for different number of classes in the mixed case.

MNARz / MCAR	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**

Table 2: Total variation distance between the marginal distribution of the variable *Shock.index.ph* in the Traumabase dataset of the algorithm considering MNARz data and the one of the algorithm considering MCAR data.

MNARz / MCAR	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**

Table 3: Euclidean distance between the conditional probabilities of the cluster memberships given the observed values of the variable *Shock.index.ph* in the Traumabase dataset, obtained using the algorithm considering MNARz data, and the ones obtained with the algorithm considering MCAR data.

5.2 Imputation performances

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%,

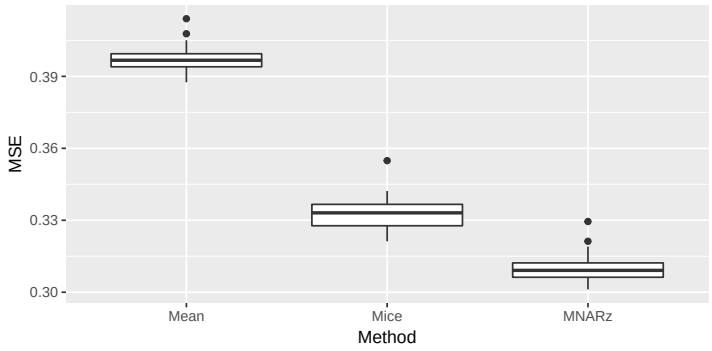

Figure 11: Mean squared error of the imputation task for the Traumabase dataset.

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 $\hat{\mathbf{Y}} \in \mathbb{R}^{n \times d}$ the imputed dataset and $\tilde{\mathbf{C}} \in \mathbb{R}^{n \times d}$ the indicator pattern of missing data newly introduced, the mean squared error is given by

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

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 11, our clustering algorithm, designed for the MNAR setting, gives a significantly smaller error than other methods.

6 Concluding remarks

This paper addresses unsupervised learning when MNAR values occur. The aim was two-fold: (i) to cluster individuals and (ii) to estimate the parameters of the distributions for each cluster (which can be used in turn to impute missing values). 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 their very simplicity, the model-based clustering algorithms considering the MNARz mechanism are indeed able to straightforwardly deal with any kind of data. Moreover, in Proposition 1, we have proven that the statistical inference may be conducted either on MNARz data on $\mathbf{Y}$ or on MAR data considering the augmented matrix $[\mathbf{Y}|\mathbf{C}]$, with $\mathbf{C}$ the missing-data pattern. It is worth noticing that this approach was widely used in practice but not theoretically studied. Finally, this mechanism has the advantage of being easily interpretable, which is especially important for real data applications.

The motivation of this work was the application on the Traumabase dataset, which can be extremely interesting, as it is genuinely useful to form groups of similarly-behaving patients for helping doctors in their decisions. To make this work entirely applicable to real datasets, there are still key challenges. First, in most datasets, the variables are not all of the same type (MCAR, MAR and MNAR variables are often coupled). There is probably no theoretical obstacle to consider such a case, but the implementation has not been done yet. There is no available method to decide whether a variable is M(C)AR or MNAR, this challenging question is out of the scope of this paper. On the other hand, we can choose between the different MNAR mechanisms proposed using the ICL criterion. However, since each algorithm has to be tested, this approach is not recommended due to time constraints. Therefore, by default, we recommend considering an MNARz mechanism.

Finally, note that our methodology can be applied in the case of mixed data (categorical/quantitative) by assuming that the features are independently drawn conditionally to the group membership.

7 Acknowledgments

The authors are thankful for fruitful discussion with Dr. Imke Maye, especially for the application on the real medical data. The work of Aude Sportisse has been supported by the French government (for Aude Sportisse), through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. The work of Julie Josse has been supported by ANR-16-IDEX-0006.

References

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

Theodore Wilbur Anderson. *An Introduction to Multivariate Statistical sAnalysis, 3rd edition.* Wiley, 2003.

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

Jean-Patrick 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.

Caroline Beunckens, Geert Molenberghs, Geert Verbeke, and Craig 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.

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

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

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

Charles Bouveyron, Gilles Celeux, T Brendan Murphy, and Adrian E Raftery. *Model-based clustering and classification for data science: with applications in R*, volume 50. 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*, 2:73–82, 1985.

G. Celeux, S. Frühwirth-Schnatter, and C. P. Robert. Models selection for mixture models. In S. Frühwirth-Schnatter, G. Celeux, , and C. P. Robert, editors, *Handbook of Mixture Analysis*, pages 117–154. CRC Press, 2019.

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.

Marie Du Roy De Chaumaray and Matthieu Marbac. Clustering data with nonignorable missingness using semi-parametric mixture models. *arXiv preprint arXiv:2009.07662*, 2020.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

John Geweke, Michael Keane, and David Runkle. Alternative computational approaches to inference in the multinomial probit model. *The review of economics and statistics*, pages 609–632, 1994.

Neil J Gordon, David J Salmond, and Adrian FM Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. In *IEE Proceedings F-radar and signal processing*, volume 140, pages 107–113. IET, 1993.

James J Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

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

Joseph G Ibrahim, Ming-Hui Chen, and Stuart R Lipsitz. Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, 88(2):551–564, 2001.

Julie Josse, Nicolas Prost, Erwan Scornet, and Gaël Varoquaux. On the consistency of supervised learning with missing values. *arXiv preprint arXiv:1902.06931*, 2019.

Jouni Kuha, Myrsini Katsikatsou, and Irini 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*, 181(4):1169–1192, 2018.

Roderick JA Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Matthieu Marbac, Christophe Biernacki, and Vincent Vandewalle. Model-based clustering of gaussian copulas for mixed data. *Communications in Statistics-Theory and Methods*, 46(23):11635–11656, 2017.

Geoffrey J McLachlan and Kaye E Basford. *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York, 1988.

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

Wang Miao, Peng Ding, and Zhi Geng. Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 111(516):1673–1683, 2016.

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

Karthika Mohan, Felix Thoemmes, and Judea 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.

Marco Ramoni, Paola Sebastiani, and Paul 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.

Alessio Serafini, Thomas Brendan Murphy, and Luca Scrucca. Handling missing data in model-based clustering. *arXiv preprint arXiv:2006.02954*, 2020.

Henry Teicher. Identifiability of finite mixtures. *The annals of Mathematical statistics*, pages 1265–1269, 1963.

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

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

Shi Zhong and Joydeep Ghosh. A unified framework for model-based clustering. *Journal of machine learning research*, 4(Nov):1001–1037, 2003.

A Proof of Proposition 1

Proof of Proposition 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

$$\begin{aligned}
L(\theta, \zeta; \tilde{\mathbf{Y}}^{\text{obs}}, \mathbf{C}) &= \int f(\tilde{\mathbf{y}}_i, \tilde{\mathbf{c}}_i; \theta, \zeta) d\tilde{\mathbf{y}}_i^{\text{mis}} \\
&= \int f(\tilde{\mathbf{y}}_i; \pi, \lambda, \psi) f(\tilde{\mathbf{c}}_i | \tilde{\mathbf{y}}_i; \zeta) d\tilde{\mathbf{y}}_i^{\text{mis}} \qquad (17) \\
&= \prod_{i=1}^{n} \left[f(\tilde{\mathbf{c}}_i | \tilde{\mathbf{y}}_i^{\text{obs}}; \zeta) \times \int_{\mathcal{Y}_i^{\text{mis}}} f(\tilde{\mathbf{y}}_i; \pi, \lambda, \psi) d\tilde{\mathbf{y}}_i^{\text{mis}} \right] \qquad (18) \\
&= \prod_{i=1}^{n} L(\zeta; \tilde{\mathbf{c}}_i \mid \tilde{\mathbf{y}}_i^{\text{obs}}) \times \prod_{i=1}^{n} L(\pi, \lambda, \psi; \tilde{\mathbf{y}}_i^{\text{obs}}). \qquad (19)
\end{aligned}$$

$$L(\pi, \lambda, \psi, \zeta; \tilde{\mathbf{y}}_i^{\text{obs}}, \tilde{\mathbf{c}}_i) = L(\zeta; \tilde{\mathbf{c}}_i \mid \tilde{\mathbf{y}}_i^{\text{obs}}) \times L(\pi, \lambda, \psi; \tilde{\mathbf{y}}_i^{\text{obs}}). \qquad (20)$$

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

$$\begin{aligned}
L(\pi, \lambda, \psi; \tilde{\mathbf{y}}_i^{\text{obs}}) &= \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i^{\text{obs}}; \lambda_k) \prod_{j=1}^{d} \rho(\alpha_{kj})^{c_{ij}} \rho(\alpha_{kj})^{(1-c_{ij})} \qquad (21) \\
&= \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i^{\text{obs}}; \lambda_k) \prod_{j=1}^{d} \mathbb{P}(c_{ij} \mid z_{ik} = 1; \psi). \qquad (22)
\end{aligned}$$

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,

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

□

B Identifiability

B.1 Continuous and count data

A1. The parameters (π, λ) of the marginal mixture defined by the density $\sum_{k=1}^{K} \pi_k f_k(y_i; \lambda_k)$ 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 $(\mathbf{Y}, \mathbf{Z})$ (*i.e.* considering the case without missing values). Some authors have already studied the identifiability of the mixture models, when no missing values in $\mathbf{Y}$ 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, continuous case. Suppose there exist 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

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

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

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

By using the identifiability of the marginal mixture, one obtains $\lambda_k = \lambda'_k$. In addition, integrating out over all the elements but the j-th element, one has for all $y_{ij} \in \mathbb{R}$,

$$\sum_{k=1}^{K} \pi_k f_{kj}(y_{ij}; \lambda_{kj}) \rho(\alpha_{kj} + \beta_{kj} y_{ij}) = \sum_{k=1}^{K'} \pi'_k f_{kj}(y_{ij}; \lambda_{kj}) \rho((\alpha')_{kj} + (\beta')_{kj} y_{ij}).$$

In the sequel, we use the same reasoning of Theorem 2 in [Teicher, 1963].

Let us denote $F_k(y_{ij}) = f_{kj}(y_{ij}; \lambda_{kj}) \rho(\alpha_{kj} + \beta_{kj} y_{ij})$ and $F'_k(y_{ij}) = f_{kj}(y_{ij}; \lambda_{kj}) \rho((\alpha')_{kj} + (\beta')_{kj} y_{ij})$. 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 $f_{1j}(u; \lambda_{1j}) \rho(\alpha_{1j} + \beta_{1j} u) \neq 0$, 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 $K = K'$. Note that $F_k = F_k'$ implies that $\rho(\alpha_{kj} + \beta_{kj} y_{ij}) = \rho((\alpha')_{kj} + (\beta')_{kj} y_{ij})$ and thus $\alpha_{kj} = (\alpha')_{kj}$ and $\beta_{kj} = (\beta')_{kj}$, since ρ is an injective function. Indeed, ρ is assumed to be strictly monotone. □

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 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 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

$$\lim_{u \to +\infty} \phi(u) = \begin{cases} 1 \text{ if } u > 0 \\ 1/2 \text{ if } u = 0 \\ 0 \text{ if } u < 0 \end{cases} \tag{23}$$

- *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

$$\lim_{u\to+\infty} \frac{\int_{-\infty}^{\alpha_{k+1}+\beta_{k+1}u} e^{-t^2/2}dt}{\int_{-\infty}^{\alpha_k+\beta_k u} e^{-t^2/2}dt} = \begin{cases} 0 \text{ if } \beta_{k+1} < 0 \\ 1/2 \text{ if } \beta_{k+1} = 0 \text{ and } \beta_k > 0 \\ 1 \text{ if } \beta_{k+1} = 0 \text{ and } \beta_k = 0 \end{cases} \tag{24}$$

- *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* $\lim_{u\to+\infty} \phi(\alpha_{k+1}+\beta_{k+1}u) = \lim_{u\to+\infty} \phi(\beta_{k+1}u)$,

$$\frac{\int_{-\infty}^{\alpha_{k+1}+\beta_{k+1}u} e^{-t^2/2}dt}{\int_{-\infty}^{\alpha_k+\beta_k u} e^{-t^2/2}dt} \underset{u\to+\infty}{=} \frac{\phi(\beta_{k+1}u)}{\phi(\beta_k u)} \underset{u\to+\infty}{<} \exp\left(\left(\frac{\beta_k^2}{2} - \frac{\beta_{k+1}^2}{2}\right)u^2\right) \tag{25}$$

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 excludes the case of Gaussian mixture with a logistic missing-data distribution, which is very used in practice. In Corollary 1, 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 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 hold true with a logistic missing-data distribution. Finally, as stated by Corollary 1 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 1. *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* <u>*generically identifiable*</u> *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* <u>*identifiable*</u> *up to label swapping.*

Proof of Corollary 1. We use Proposition 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 $\sigma_k^2 = \sigma_{k+1}^2$. □

On identifiability of the Poisson mixture Proposition 1 can also be applied for variables with integer value (*i.e.* count data), as shown below in Examples 2 and 3 for the Poisson mixture with probit or logistic missing-data distributions.

Example 2 (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* (23), *one has*

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

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

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

using

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

and (24) *for the missing distribution part.*

- *If* $\beta_{k+1} < 0$ *and* $\beta_k < 0$*: using* (25)*, one obtains*

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

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

Example 3 (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*

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

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* $\{\alpha_k, \beta_k, \lambda_k \text{ s.t.} (\ln \lambda_k - \beta_k) - (\ln \lambda_{k+1} - \beta_{k+1}) = 0\}_{k=1,\dots,K}$ *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*

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

B.2 Categorical data

We assume the following:

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

$$f_k(\cdot; \theta_k) = \prod_{j=1}^{d} f_{kj}(\cdot\,; \theta_{kj}); \tag{26}$$

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

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

Identifiability of ψ This implies that the marginal distributions of the pattern of missing data for the two sets of parameters ψ and ψ' are equal.

$$\sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} \rho(\alpha_{kj})^{c_{ij}} [1 - \rho(\alpha_{kj})]^{1-c_{ij}} = \sum_{k=1}^{K'} \pi'_k \prod_{j=1}^{d} \rho(\alpha'_{kj})^{c_{ij}} [1 - \rho(\alpha'_{kj})]^{1-c_{ij}}$$

One recognizes the finite mixture of K different Bernoulli products with d components and with parameters $(\rho(\alpha_{k1}), \dots, \rho(\alpha_{kd})))_{k=1,\dots,K}$ and $(\rho(\alpha'_{k1}), \dots, \rho(\alpha'_{kd})))_{k=1,\dots,K}$. The generic identifiability up to a label swapping of these parameters is given by Corollary 5 in Allman et al. [2009]. As the function ρ is strictly monotone, the equality $\rho(\alpha_{kj}) = \rho(\alpha'_{kj})$ implies $\alpha_{kj} = \alpha'_{kj}$.

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

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

i.e. by independence to the group membership,

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

We recognize the finite mixture of K multinomial distributions with d components for $y_{ij} = (y_{ij}^1, \dots, y_{ij}^{\ell_j}), j = 1, \dots, d$ with paramaters $(\lambda_{kj}) = (\lambda_{kj}^1, \dots, \lambda_{kj}^{\ell_j}), j = 1, \dots, d$ and proportions $(\pi_k \prod_{j=1}^{d} \rho(\alpha_{kj}))_{k=1,\dots,K}$. We can thus apply Theorem 4 [Allman et al., 2009] with the model $\mathcal{M}(K; \ell_1, \dots, \ell_d)$ which gives the generic identifiability of the model paramaters up to a label swapping, *i.e.*

$$\forall k, \forall j,\ \lambda_{kj}^h = (\lambda'_{kj})^h$$
$$\forall k,\ \pi_k \prod_{j=1}^{d} \rho(\alpha_{kj}) = \pi'_k \prod_{j=1}^{d} \rho(\alpha'_{kj})$$

The second equality implies $\pi_k = \pi'_k$ using the generic identifiability of $\rho(\alpha_{kj}), \forall k, \forall j$ stated above. If $K \neq K'$, say $K > K'$, $\sum_{k=K'+1}^{K} \pi_k \prod_{j=1}^{d} \rho(\alpha_{kj}) \prod_{h=1}^{\ell_j} (\lambda_{kj}^h)^{y_{ij}^h} = 0$ 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 Detailed algorithms

The algorithms for the different missing scenarios and type of data are given. In particular, for continuous data, we derive the formulae assuming Gaussian data.

C.1 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:

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

(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:

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

(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

$$\begin{aligned} f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1, \mathbf{c}_i; \theta^{[r-1]}) &= \frac{f(\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik} = 1, \mathbf{c}_i; \theta^{[r-1]})}{f(\mathbf{y}_i^{\text{obs}}, z_{ik} = 1, \mathbf{c}_i; \theta^{[r-1]})} \\ &= \frac{f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \psi^{[r-1]}) f(\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \lambda^{[r-1]})}{\int_{\mathcal{Y}_i^{\text{mis}}} f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \psi^{[r-1]}) f(\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \lambda^{[r-1]}) d\mathbf{y}_i^{\text{mis}}} \end{aligned} \tag{27}$$

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

$$\begin{aligned} t_{ik}(\theta^{[r-1]}) &= \frac{f(z_{ik} = 1, \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})}{f(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})} \\ &= \frac{f(z_{ik} = 1, \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})}{\sum_{h=1}^{K} f(z_{ih} = 1, \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})} \\ &= \frac{\pi_k^{[r-1]} f(\mathbf{y}_i^{\text{obs}} \mid z_{ik} = 1; \lambda_k^{[r-1]}) f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \theta^{[r-1]})}{\sum_{h=1}^{K} \pi_h^{[r-1]}) f(\mathbf{y}_i^{\text{obs}} \mid z_{ih} = 1; \lambda_h^{[r-1]}) f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}, z_{ih} = 1; \theta^{[r-1]})} \\ &= \frac{\pi_k^{[r-1]} f_k(\mathbf{y}_i^{\text{obs}}; \lambda_k^{[r-1]}) f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \theta^{[r-1]})}{\sum_{h=1}^{K} \pi_h^{[r-1]} f_h(\mathbf{y}_i^{\text{obs}}; \lambda_h^{[r-1]}) f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}, z_{ih} = 1; \theta^{[r-1]})} \end{aligned} \tag{28}$$

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.1.1 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.

MNARz and MNARz^j models 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 (27),

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

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

$$\left(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \lambda^{[r-1]}\right) \sim \mathcal{N}\left((\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]}, (\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]}\right). \tag{29}$$

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

$$(\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]} = (\mu_{ik}^{\text{mis}})^{[r-1]} + (\Sigma_{ik}^{\text{mis,obs}})^{[r-1]} \left((\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}\right)^{-1} \left(\mathbf{y}_i^{\text{obs}} - (\mu_{ik}^{\text{obs}})^{[r-1]}\right), \tag{30}$$

$$(\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]} = (\Sigma_{ik}^{\text{mis,mis}})^{[r-1]} - (\Sigma_{ik}^{\text{mis,obs}})^{[r-1]} \left((\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}\right)^{-1} (\Sigma_{ik}^{\text{obs,mis}})^{[r-1]}. \tag{31}$$

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

$$\log(f(\mathbf{c}_i \mid z_{ik}=1; \psi)) = \begin{cases} \sum_{j=1}^d c_{ij} \log \rho(\alpha_k) + (1-c_{ij}) \log(1-\rho(\alpha_k)) & (\text{MNAR}z) \\ \sum_{j=1}^d c_{ij} \log \rho(\alpha_{kj}) + (1-c_{ij}) \log(1-\rho(\alpha_{kj})) & (\text{MNAR}z^j) \end{cases} \tag{32}$$

- 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

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

  Using (28), one obtains

$$t_{ik}^{[r-1]}(\theta^{[r-1]}) = \left\{ \begin{array}{ll} \frac{\pi_k^{[r-1]} \phi(\mathbf{y}_i^{\text{obs}}; (\mu_{ik}^{\text{obs}})^{[r-1]}, (\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}) \prod_{j=1}^{d} \rho(\alpha_k^{[r-1]})^{c_{ij}} (1-\rho(\alpha_k^{[r-1]}))^{1-c_{ij}}}{\sum_{h=1}^{K} \pi_h^{[r-1]} \phi(\mathbf{y}_i^{\text{obs}}; (\mu_{ih}^{\text{obs}})^{[r-1]}, (\Sigma_{ih}^{\text{obs,obs}})^{[r-1]}) \prod_{j=1}^{d} \rho(\alpha_k^{[r-1]})^{c_{ij}} (1-\rho(\alpha_k^{[r-1]}))^{1-c_{ij}}} & (\text{MNAR}z) \\ \frac{\pi_k^{[r-1]} \phi(\mathbf{y}_i^{\text{obs}}; (\mu_{ik}^{\text{obs}})^{[r-1]}, (\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}) \prod_{j=1}^{d} \rho(\alpha_{kj}^{[r-1]})^{c_{ij}} (1-\rho(\alpha_{kj}^{[r-1]}))^{1-c_{ij}}}{\sum_{h=1}^{K} \pi_h^{[r-1]} \phi(\mathbf{y}_i^{\text{obs}}; (\mu_{ih}^{\text{obs}})^{[r-1]}, (\Sigma_{ih}^{\text{obs,obs}})^{[r-1]}) \prod_{j=1}^{d} \rho(\alpha_{kj}^{[r-1]})^{c_{ij}} (1-\rho(\alpha_{kj}^{[r-1]}))^{1-c_{ij}}} & (\text{MNAR}z^j) \end{array} \right. \tag{33}$$

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

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

  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 $i=1,\ldots,n$

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

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

$$(\mathcal{J}_k^{\text{MNAR}z})^{[r]} = \begin{array}{cc} c_{\cdot 1} & 1 \\ \vdots & \vdots \\ c_{\cdot d} & 1 \end{array}. \tag{34}$$

$$(\mathcal{J}_{kj}^{\text{MNAR}z^j})^{[r]} = \begin{array}{cc} c_{\cdot j} & 1 \end{array} \tag{35}$$

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

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

Input: $\mathbf{Y} \in \mathbb{R}^{n\times d}$ (matrix containing missing values), $K \geq 1$, $r_{\max}$.
Initialize π_k^0, μ_k^0, Σ_k^0 and ψ_k^0, for $k \in \{1,\dots,K\}$.
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**
$\pi_k^{[r]} = \frac{1}{n}\sum_{i=1}^n t_{ik}(\theta^{[r-1]})$, $\quad \mu_k^{[r]} = \frac{\sum_{i=1}^n t_{ik}(\theta^{[r-1]})(\tilde{y}_{k,i})^{[r-1]}}{\sum_{i=1}^n t_{ik}(\theta^{[r-1]})}$
$\Sigma_k^{[r]} = \frac{\sum_{i=1}^n \left[t_{ik}(\theta^{[r-1]})\left((\tilde{y}_{i,k})^{[r-1]} - \mu_k^{[r]}\right)\left((\tilde{y}_{i,k})^{[r-1]} - \mu_k^{[r]}\right)^T + \tilde{\Sigma}_{ik}^{[r-1]}\right)\right]}{\sum_{i=1}^n t_{ik}(\theta^{[r-1]})}$
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, $\forall j \in \{1,\dots,d\}$,

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

for a matrix $\mathcal{M}_{kj}$ depending on the MNAR model (see (34) and (35)) and $\mathbf{c}_{\cdot j}$ the missing data pattern for the variable j.
end for
end for

MNAR$y*$ models 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

$$
\begin{aligned}
&f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1, \mathbf{c}_i; \theta^{[r-1]}) \\
&= \frac{\prod_{h=1}^{d} \rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{mis}})^{c_{ih}} (1 - \rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{obs}}))^{1-c_{ih}} f(\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \theta^{[r-1]})}{\int_{\mathcal{Y}_i^{\text{mis}}} \prod_{h=1}^{d} \rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{mis}})^{c_{ih}} (1 - \rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{obs}}))^{1-c_{ih}} f(\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \theta^{[r-1]}) d\mathbf{y}_i^{\text{mis}}} \\
&= \frac{\prod_{h, c_{ih}=1} \rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{mis}}) f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \theta^{[r-1]})}{\int_{\mathcal{Y}_i^{\text{mis}}} \prod_{h, c_{ih}=1} \rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{mis}}) f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \theta^{[r-1]}) d\mathbf{y}_i^{\text{mis}}}. \qquad (36)
\end{aligned}
$$

  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 (36),

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

  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,

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

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

$$
\begin{aligned}
&f(c_{ij} \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \theta^{[r-1]}) \\
&= \int_{\mathcal{Y}_{ij}^{\text{mis}}} f(c_{ij} \mid \mathbf{y}_i^{\text{obs}}, y_{ij}^{\text{mis}}, z_{ik} = 1; \psi^{[r-1]}) f(\mathbf{y}_{ij}^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \theta^{[r-1]}) dy_{ij}^{\text{mis}} \\
&= c_{ij} \int_{-\infty}^{+\infty} \rho(\alpha_{kj}^{[r-1]} + \beta_{kj}^{[r-1]} y_{ij}^{\text{mis}}) \phi(y_{ij}^{\text{mis}}; (\tilde{\mu}_{ik}^{\text{mis}})_j^{[r-1]}, (\tilde{\Sigma}_{ik}^{\text{mis}})_{jj}^{[r-1]}) dy_{ij}^{\text{mis}} + (1 - c_{ij})(1 - \rho(\alpha_{kj}^{[r-1]} + \beta_{kj}^{[r-1]} y_{ij}^{\text{obs}})) \qquad (37)
\end{aligned}
$$

Using (28), the probabilities $t_{ik}(\theta^{[r-1]})$ can be deduced from Equation (37).

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

- ρ **is logistic**: Equation (36) leads to none classical distribution because

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

Term (b) is

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

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 (37) 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 C.2.1) 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

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

Equation (37) 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.1.2 Latent class model for categorical data

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

MNARz and MNARz^j models Term (a) is

$$\mathbb{E}\left[\log(\phi(\mathbf{y}_i;p_k)) \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1, \mathbf{c}_i; \lambda^{[r-1]}\right] = \sum_{j,c_{ij}=0}\sum_{\ell=1}^{\ell_j} y_{ij}^{\ell} + \sum_{j,c_{ij}=1}\sum_{\ell=1}^{\ell_j} \log(\lambda_{kj}^{y_{ij}^{\ell}}) \tag{38}$$

Term (b) is the same as in the Gaussian case given in (32). Finally, the EM algorithm can be summarized as follows

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

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

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

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

The M-step for ψ consists of performing a GLM with a binomial link and has already been given in detail in Appendix C.1.1 (see (51) and (52)).

C.1.3 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.

C.2 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 $(\mathbb{P}(z_{ik}=1 \mid \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \lambda^{[r-1]}, \psi^{[r-1]}))_{k=1,\ldots,K}$

detailed as follows

$$\begin{aligned}\mathbb{P}(z_{ik}=1 \mid \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]}) &= \frac{\mathbb{P}(z_{ik}=1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \pi^{[r-1]}, \lambda^{[r-1]}, \psi^{[r-1]})}{\mathbb{P}(\mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \pi^{[r-1]}, \lambda^{[r-1]}, \psi^{[r-1]})} \\ &= \frac{\mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i^{[r-1]}, z_{ik}=1; \psi^{[r-1]})\mathbb{P}(\mathbf{y}_i^{[r-1]} \mid z_{ik}=1; \lambda^{[r-1]})\mathbb{P}(z_{ik}=1; \pi^{[r-1]})}{\sum_{h=1}^K \mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i^{[r-1]}, z_{ih}=1; \psi^{[r-1]})\mathbb{P}(\mathbf{y}_i^{[r-1]} \mid z_{ih}=1; \lambda^{[r-1]})\mathbb{P}(z_{ih}=1; \pi^{[r-1]})} \\ &= \frac{\mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i^{[r-1]}, z_{ik}=1; \psi^{[r-1]})\mathbb{P}(\mathbf{y}_i^{[r-1]} \mid z_{ik}=1; \lambda^{[r-1]})\pi_k^{[r-1]}}{\sum_{h=1}^K \mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i^{[r-1]}, z_{ih}=1; \psi^{[r-1]})\mathbb{P}(\mathbf{y}_i^{[r-1]} \mid z_{ih}=1; \lambda^{[r-1]})\pi_h^{[r-1]}}\end{aligned} \tag{39}$$

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 (36) and recalled here

$$\begin{aligned}&f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]}=1, \mathbf{c}_i; \theta^{[r-1]}) \\ &\quad= \frac{\prod_{j,c_{ij}=1} f(c_{ij}=1 \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]}=1; \psi^{[r-1]}) f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]}=1; \lambda^{[r-1]})}{\int_{\mathcal{Y}_i^{\text{mis}}} \prod_{j,c_{ij}=1} f(c_{ij}=1 \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]}=1; \psi^{[r-1]}) f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]}=1; \lambda^{[r-1]}) d\mathbf{y}_i^{\text{mis}}}\end{aligned} \tag{40}$$

C.2.1 Gaussian mixture for continuous data

First note that the probabilities of the multinomial distribution for drawing $z_i^{[r]}$ given in (39) can be easily computed for all cases.

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

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.

MNAR$y*$ 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$y*$ 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 (40) is not classical and drawing y_i^{mis} from it seems complicated. Indeed, one has

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

where $(\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]}$ and $(\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]}$ are given in (30) and (31). 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 (40) 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 $j = 1, \ldots, d$,

$$c_{ij} = \begin{cases} 1 & \text{if } L_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{41}$$

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 (41). 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

$$(\mathbf{L}_i \mid \mathbf{y}_i^{[r-1]}, z_{ik}^{[r-1]} = 1, c_i) \sim \mathcal{N}_t(\alpha_k^{[r-1]} + \beta_k^{[r-1]}\mathbf{y}_i, I_{d\times d}; a, b), \tag{42}$$

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

$$\begin{aligned} \mathbb{P}(z_{ik} = 1 \mid \mathbf{L}_i^{[r]}, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]}) &= \frac{\mathbb{P}(z_{ik} = 1, \mathbf{L}_i^{[r]}, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]})}{\sum_{k=1}^K \mathbb{P}(z_{ik} = 1, \mathbf{L}_i^{[r]}, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]})} \\ &= \frac{f(\mathbf{L}_i^{[r]} | z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \psi^{[r-1]}) f(z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]})}{\sum_{k=1}^K f(\mathbf{L}_i^{[r]} | z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \psi^{[r-1]}) f(z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]})} \end{aligned} \tag{43}$$

The part involving $f(z_{ik}=1,\mathbf{y}_i^{[r-1]},\mathbf{c}_i;\theta^{[r-1]})$ is given in (39) 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 (42) 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 (29).

Finally combining these two equations one obtains

$$
\left(\mathbf{y}_i^{\text{mis}} \mid \mathbf{L}_i^{[r]},z_{ik}^{[r]}=1,\mathbf{y}_i^{\text{obs}},\mathbf{c}_i\right) \sim \mathcal{N}\left(\mu_{ik}^{\text{SEM}},\Sigma_{ik}^{\text{SEM}}\right), \tag{44}
$$

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

$$
(\mathcal{H}^{\text{MNAR}y})^{[r]} = \left[\begin{array}{c|ccccc} c_{.1} & 1 & y_{.1}^{[r]} & 0 & \ldots & 0\\ c_{.2} & 1 & 0 & y_{.2}^{[r]} & \ldots & 0\\ \vdots & \vdots & & \ddots & \ddots & \\ c_{.d} & 1 & 0 & 0 & \ldots & y_{.d}^{[r]} \end{array}\right]. \tag{45}
$$

- 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

$$(\mathcal{H}_{kj}^{\text{MNAR}y^k})^{[r]} = \left|\begin{array}{cc|cccc} (c_{u1})_{u,z_{u1}^{[r]}=1} & & 1 & (y_{u1}^{[r]})_{u,z_{u1}^{[r]}=1} & 0 & \ldots & 0 \\ \vdots & & \vdots & & & \ddots & \vdots \\ (c_{u1})_{u,z_{uK}^{[r]}=1} & & 1 & 0 & (y_{u1}^{[r]})_{u,z_{uK}^{[r]}=1} & & 0 \\ \vdots & & \vdots & \vdots & & \ddots & \\ (c_{ud})_{u,z_{uK}^{[r]}=1} & & 1 & 0 & 0 & & (y_{ud}^{[r]})_{u,z_{uK}^{[r]}=1} \end{array}\right| . \quad (46)$$

- 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

$$(\mathcal{H}^{\text{MNAR}yz})^{[r]} = \left|\begin{array}{c|cccccccc} c_{\cdot 1} & y_{\cdot 1}^{[r]} & 0 & \ldots & 0 & z_{\cdot 1}^{[r]} & \ldots & z_{\cdot K}^{[r]} \\ c_{\cdot 2} & 0 & y_{\cdot 2}^{[r]} & \ldots & 0 & z_{\cdot 1}^{[r]} & \ldots & z_{\cdot K}^{[r]} \\ \vdots & & \ddots & \ddots & & \vdots & \vdots & \vdots \\ c_{\cdot d} & 0 & 0 & \ldots & y_{\cdot d}^{[r]} & z_{\cdot 1}^{[r]} & \ldots & z_{\cdot K}^{[r]} \end{array}\right| . \quad (47)$$

- 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

$$(\mathcal{H}_j^{\text{MNAR}yz^j})^{[r]} = \left|\begin{array}{c|ccc} c_{\cdot j} & y_{\cdot j}^{[r]} & z_{\cdot 1}^{[r]} & \ldots & z_{\cdot K}^{[r]} \end{array}\right| . \quad (48)$$

- 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

$$(\mathcal{H}_k^{\text{MNAR}y^kz})^{[r]} = \left|\begin{array}{cc|cccc} (c_{u1})_{u,z_{uk}^{[r]}=1} & & (y_{u1}^{[r]})_{u,z_{uk}^{[r]}=1} & 0 & \ldots & 0 & 1 \\ (c_{u2})_{u,z_{uk}^{[r]}=1} & & 0 & (y_{u2}^{[r]})_{u,z_{uk}^{[r]}=1} & \ldots & 0 & 1 \\ \vdots & & & \ddots & \ddots & & 1 \\ (c_{ud})_{u,z_{uk}^{[r]}=1} & & 0 & 0 & \ldots & (y_{ud}^{[r]})_{u,z_{uk}^{[r]}=1} & 1 \end{array}\right| . \quad (49)$$

- 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

$$(\mathcal{H}_{kj}^{\text{MNAR}y^kz^j})^{[r]} = \left|\begin{array}{cc|cc} (c_{uj})_{u,z_{uk}^{[r]}=1} & & (y_{uj}^{[r]})_{u,z_{uk}^{[r]}=1} & 1 \end{array}\right| \quad (50)$$

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

with

$$(\mathcal{H}^{\text{MNAR}z})^{[r]} = \left[\begin{array}{c|ccc} c_{\cdot 1} & z_{\cdot 1} & \dots & z_{\cdot K} \\ \vdots & \vdots & \vdots & \vdots \\ c_{\cdot d} & z_{\cdot 1} & \dots & z_{\cdot K} \end{array}\right] = \left[\begin{array}{c|ccc} c_{11} & z_{11}^{[r]} & \dots & z_{1K}^{[r]} \\ \vdots & \vdots & \vdots & \vdots \\ c_{n1} & z_{n1}^{[r]} & \dots & z_{nK}^{[r]} \\ \vdots & \vdots & \vdots & \vdots \\ c_{1d} & z_{11}^{[r]} & \dots & z_{1K}^{[r]} \\ \vdots & \vdots & \vdots & \vdots \\ c_{nd} & z_{n1}^{[r]} & \dots & z_{nK}^{[r]} \end{array}\right]. \tag{51}$$

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

$$(\mathcal{H}_j^{\text{MNAR}z^j})^{[r]} = \left[\begin{array}{c|ccc} c_{\cdot j} & z_{\cdot 1}^{[r]} & \dots & z_{\cdot K}^{[r]} \end{array}\right] \tag{52}$$

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$y*$ models, ρ is probit

Input: $\mathbf{Y} \in \mathbb{R}^{n \times d}$ (matrix containing missing values), $K \geq 1$, $r_{\max}$.
Initialize $\mathbf{Z}^0$, π_k^0, μ_k^0, Σ_k^0 and ψ_k^0, for $k \in \{1, \dots, K\}$.
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 (42).
    Draw $\mathbf{z}_i^{[r]}$ from the multinomial distribution with probabilities detailed in (43).
    Draw $(\mathbf{y}_i^{\text{mis}})^{[r]}$ from the multivariate Gaussian distribution given in (44).
  end for
  Let $\mathbf{Y}^{[r]} = (\mathbf{y}_1^{[r]} | \dots | \mathbf{y}_n^{[r]})$ be the imputed matrix.
  Let $\mathbf{Z}^{[r]} = (\mathbf{z}_1^{[r]} | \dots | \mathbf{z}_n^{[r]})$ be the partition.
  M-step:
  for $k = 1$ **to** K **do**
    Let $\pi_k^{[r]}$ be the proportion of rows of $\mathbf{Y}^{[r]}$ belonging class k.
    Let $\mu_k^{[r]}, \Sigma_k^{[r]}$ be the mean and covariance matrix of rows of $\mathbf{Y}^{[r]}$ 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 $\forall j \in \{1, \dots, d\}$,

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

    for a matrix $\mathcal{M}_{kj}$ depending on the MNAR model (see (45),(46),(47),(52),(50) and (51)) and $\mathbf{c}_{\cdot j}$ the missing data pattern for the variable j.
  end for
end for

	EM			SEM		
	Gaussian		Categorical	Gaussian		Categorical
	Appendix C.1.1		Appendix C.1.2	Appendix C.2.1		Appendix C.1.2
MNARz MNARz^j	✓		✓	✓		✓
	Probit	Logit		Probit	Logit	
MNAR$y*$	no closed form	no closed form, optim. pb	not identifiable	✓	require algorithms as SIR (costly)	not identifiable

Table 4: Summary of the cases for which the EM and the SEM lead to feasable (or not feasable) computations. The symbol ✓ means that the computations are feasable and that they are derived in Appendix.

MNARz and MNARz^j models 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.1.1 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 (29). The M-step for ψ has been specified in the previous paragraph with (51) and (52).

C.2.2 Latent class model for categorical data

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

MNARz and MNARz^j models 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 $(\lambda_{kj}^{\ell})_{\ell=1,\ldots,\ell_j}$.

D Additional numerical experiments on synthetic data

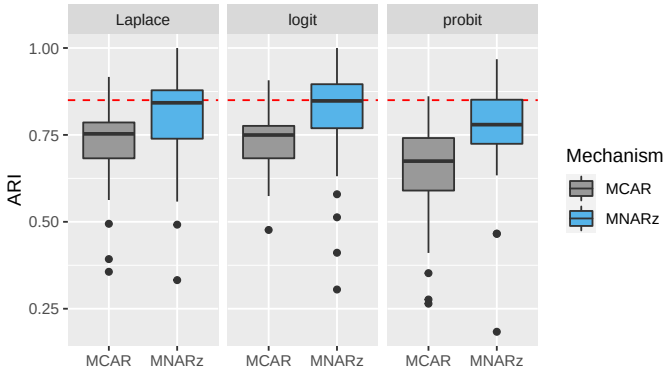

Figure 12: Boxplot of the ARI obtained for 50 samples composed of $d = 6$ variables with a misclassification rate of 15% and a missing-data rate of 30% in the whole dataset. The missing values are introduced using a MNARz setting with different link functions. The red line indicates the theoretical ARI.

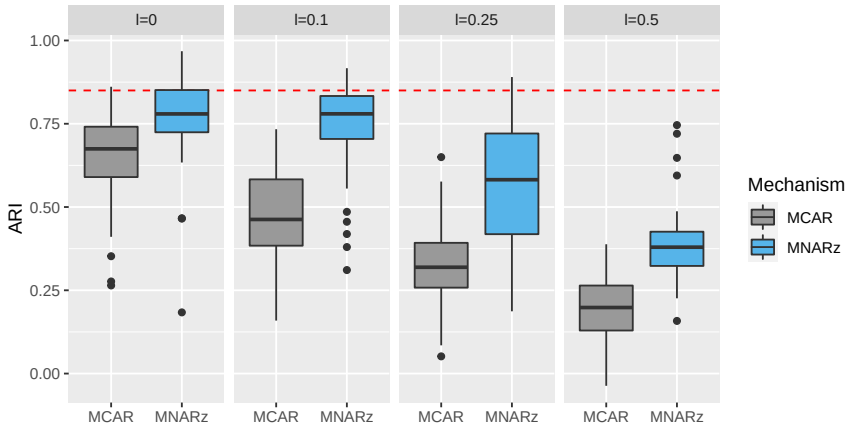

Figure 13: Boxplot of the ARI obtained for 50 samples composed of $d = 6$ variables with a misclassification rate of 15% and a missing-data rate of 30% in the whole dataset. The correlation coefficient of the covariance matrices in each component ($k = 1, 2, 3$) is $l = 0$, $l = 0.1$, $l = 0.25$ and $l = 0.5$, whereas the algorithms consider the diagonal case. The red line indicates the theoretical ARI.

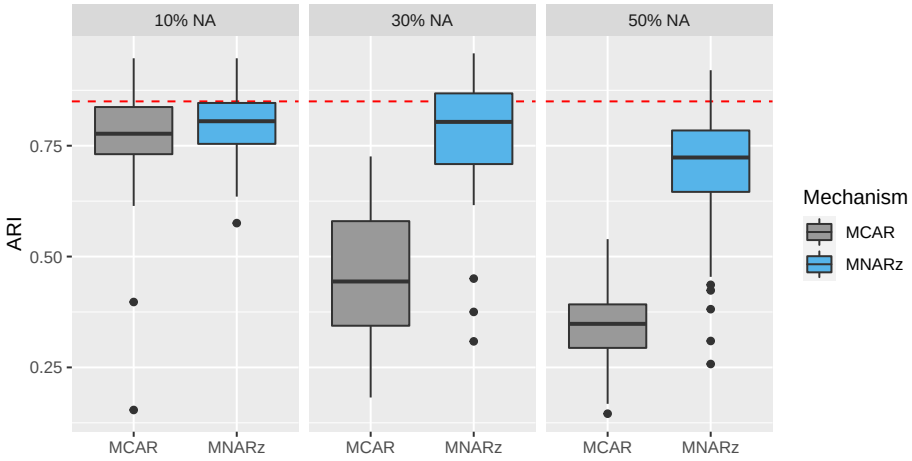

Figure 14: Boxplot of the ARI obtained for 50 samples composed of $d = 6$ variables with a misclassification rate of 15% and a missing-data rate of 10%, 30% and 50% in the whole dataset. The red line indicates the theoretical ARI.

E Description of the variables in the Traumabase dataset

A description of the variables which are used in Section 5 is given. Figure 15 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 administered 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.

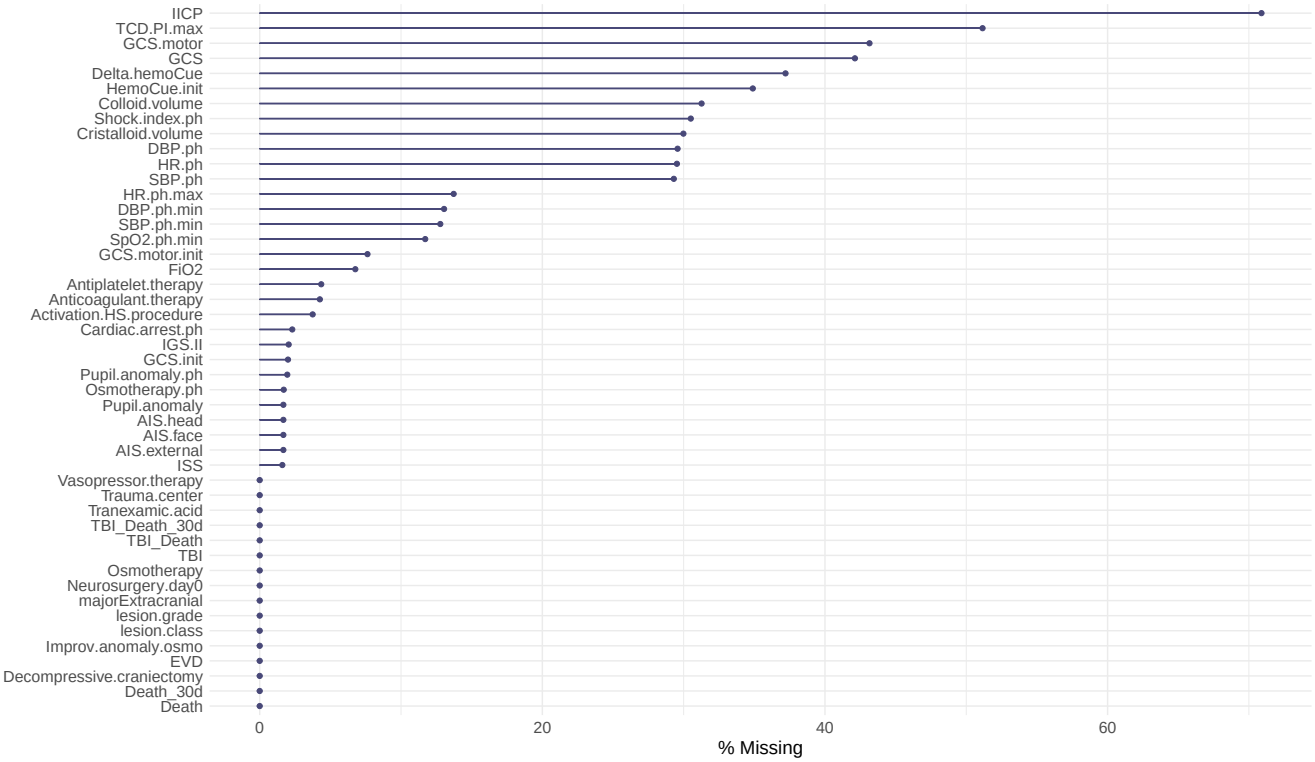

Figure 15: Percentage of missing values per variable for the Traumabase dataset.

F Complements on generic experiments

This section gives the values of δ (see (16)) and ψ (see 3) used during the different experiments. As exemplained in Section 4.2, their choice allows to control the rates of misclassification and missingness. 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.

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

d	δ	α	β
3	3.5	-1.56	$\begin{pmatrix} 1.45 & 0.2 & -3 \end{pmatrix}$
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 7: 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 8: 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 9: 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 10: 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 11: 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 12: Choice of the values of δ, α and β for all the experiments of Section 4.2 for the MNARyz^j mechanism.