Accompanying note : Model-based Clustering with Missing Not At Random Data

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

Accompanying note: Model-based Clustering with Missing Not At Random Data

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

December 21, 2023

Contents

1 Organization of this accompanying note

This document is the accompanying note of the main paper (Sportisse et al., 2023). We assume the data missing not at random (MNAR) values (Rubin, 1976; Ibrahim et al., 2001; Mohan et al., 2018), *i.e.* the effect of missingness depends on on the missing values themselves. An example includes clinical data collected in emergency situations, where doctors may choose to treat patients before measuring heart rate: the missingness of heart rate depends on the missing heart rate itself. For such a setting, the observed data are therefore not representative of the population. The main paper focuses on the specific MNARz setting, for which the only effect of missingness is on the class membership; in this document, we give some details for other MNAR settings.

1.1 Contributions

In Section 2, 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). In Section 3, we 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 uniquely on the class membership have identifiable parameters. These identifiability results represent a substantial extension of the work of Miao et al. (2016) to more complex missing scenario and to the multivariate case. For each model or submodel, an EM or Stochastic EM algorithm is proposed in Section 4, implemented, and made available for reproducibility[1]. Finally, we propose a numerical experiment in Section 5 to assess the clustering performance in each proposed MNAR setting.

1.2 Notations

Set the dataset $Y = (\mathbf{y}_1|\dots|\mathbf{y}_n)^T$ consisting of n individuals, where each observation $\mathbf{y}_i = (y_{i1}, \dots, y_{id})^T$ belongs to a space $\mathcal{Y}$, depending on the type of data, defined by d features. The pattern of missing data is denoted by $C = (\mathbf{c}_1|\dots|\mathbf{c}_n)^T \in \{0,1\}^{n\times d}$, with $\mathbf{c}_i = (\mathbf{c}_{i1}, \dots, \mathbf{c}_{id})^T \in \{0,1\}^d$: $\mathbf{c}_{ij} = 1$ indicates that the value y_{ij} is missing and $\mathbf{c}_{ij} = 0$ otherwise. The values of the observed (resp. missing) variables for individual i are denoted by $\mathbf{y}_i^{\text{obs}}$ (resp. $\mathbf{y}_i^{\text{mis}}$). The objective of clustering is to estimate an unknown partition $Z = (\mathbf{z}_1|\dots|\mathbf{z}_n)^T \in \{0,1\}^{n\times K}$ that groups the full dataset Y into K classes, with $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})^T \in \{0,1\}^K$

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

and where $z_{ik} = 1$ if $\mathbf{y}_i$ belongs to cluster k, $z_{ik} = 0$ otherwise. Consequently, in a clustering context, the missing data are not only the values $\mathbf{y}_i^{\text{mis}}$ but also the partition labels $\mathbf{z}_i$.

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

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

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

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

Different types of pdf $f_k(\cdot\,; \lambda_k)$ can be considered, depending on the types of features at hand. Thus, if $\mathbf{y}_i$ is a vector of continuous variables, the pdf of a d-variate Gaussian distribution (McLachlan and Basford, 1988; Banfield and Raftery, 1993) can be considered for $f_k(\mathbf{y}_i; \lambda_k)$ and thus λ_k groups the mean vector and the covariance matrix. Moreover, if some components of $\mathbf{y}_i$ are discrete or categorical, the latent class model (see Geweke et al. (1994); McParland and Gormley (2016)) defining $f_k(\mathbf{y}_i; \lambda_k) = \prod_{j=1}^{d} f_{kj}(y_{ij}; \lambda_{kj})$ can be used, with $\lambda_k = (\lambda_{k1}, \ldots, \lambda_{kd})$. In such case, f_{kj} could be the pdf of a Poisson (resp. multinomial) distribution with parameter λ_{kj} if y_{ij} is an integer (resp. categorical) variable. The next subsection discusses the choice of the modeling for the missingness mechanism (*i.e.*, the distribution $f_k(\mathbf{c}_i \mid \mathbf{y}_i; \psi_k)$).

2 Variant MNAR modeling

2.1 Rationale for MNAR assumptions

To handle MNAR data in selection models, the distribution of the missing-data pattern given the data and the partition should be specified. We consider the following assumptions:

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

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

By the categorical nature of the mask $\mathbf{c}_i$, the independence assumption **1.** is a quite natural hypothesis in the context of clustering (Du Roy De Chaumaray and Marbac, 2020; Chi et al., 2016). The independence assumption **2.** of the variables amounts to considering self-masked class-wise MNAR mechanisms for each variable: the missingness of the variable j may depend on its value itself (self-masked) and on the class membership (class-wise). Note that the self-masked feature, apart from limiting the number of parameters to be estimated, is now commonly met in the literature (Mohan, 2018; Sportisse et al., 2020; Le Morvan et al., 2020), and is able to retrieve some existing ad hoc MNAR procedures already used in machine learning community (see more details in (Sportisse et al., 2023, Theorem 1)).

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

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

where $\psi_k = (\alpha_{k1}, \beta_{k1}, \ldots, \alpha_{kK}, \beta_{kK})$.

The parameter α_{kj} represents a mean effect of missingness on the k-th class membership for the variable j (note that within a same class k, α_{kj} is not necessarily equal to $\alpha_{kj'}$ for $j \neq j'$). The parameter β_{kj} represents the direct effect of missingness on the variable j which depends on the class k as well. This model is called MNAR$y^k z^j$ in the following.

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

2.2 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 so-called MNARyz^j model, the effect of missingness on a variable is the same regardless of the class (while keeping different mean effects α_{kj} on the class membership), so that

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

For the MNAR$y^k z$ model, the missingness has a same mean effect on class membership shared by all variables (while allowing different self-masked and class-wise parameters β_{kj}):

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

The effects on a particular variable and on the class membership can be respectively the same for all the classes and for all the variables, entailing the so-called MNARyz model:

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

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

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

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

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

Thirdly, the probability to be missing can also depend only on the class membership, so that the missingness is class-wise only. In the MNARz model, we consider that the only effect of missingness is on the class membership k, being the same for all variables,

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

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

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

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

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

3 Identifiability of the model parameters

The generic identifiability (Allman et al., 2009) of parameters for continuous, count, categorical, and mixed data (*i.e.*, when the set of unidentifiable parameters has a zero Lebesgue measure) is ensured by the following theorem. We consider the following assumptions:

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

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

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

$$f_k(\cdot;\lambda_k) = \prod_{j=1}^{d} f_{kj}(\cdot\,;\lambda_{kj}); \tag{12}$$

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.

Assumption **A1.** means that the identifiability of the parameters (π, λ, ψ) of the model (2) requires the identifiability of the parameters (π, λ) of the marginal mixture of (Y, Z) (*i.e.*, considering the case without missing values). Some authors have already studied the identifiability of the mixture models, when no missing values in Y occur, especially Teicher (1963) for Gaussian mixtures (continuous variables) and Yakowitz and Spragins (1968) for Poisson mixtures (count variables). Assumption **A2.** is the core ingredient to prove the identifiability of the parameters, requiring that a total ordering of the mixture densities holds. We illustrate it by considering concrete examples in Appendix A. Note that under Assumption **A3.** requires that the link function of the missing data mechanism is strictly monotone, but no assumption about its form (*e.g.* logit, probit) is made. Assumption **A4.** requires the conditional independence of the features given the group membership and Assumption **A5.** links the dimension of the observations and the logarithm of the number of clusters. Both assumptions **A4.** and **A5.** are classical in the categorical case, even without missing values (Allman et al., 2009).

Theorem 3.1. *Define the conditions:*

C1 The variables correspond to continuous or count data, ***A1.*** *and* ***A2.*** *hold true,*

C2 All the variables are categorical, ***A4.*** *and* ***A5.*** *hold true and the mechanism is stated by* (9)*,* (10) *or* (11)*,*

C3 At least one variable is continuous or count data and has a marginal distribution that satisfies ***A1.*** *and* ***A2.****,* ***A4.*** *holds true,*

C4 At least one variable is categorical and its associated mechanism is stated by (9)*,* (10) *or* (11)*,* ***A4.*** *and* ***A5.*** *hold true.*

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

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

4 Estimation of the proposed MNAR models

Assuming identifiability, we estimate parameters via likelihood maximization using EM and SEM algorithms specifically designed for Gaussian, Poisson, multinomial and mixed data with MNAR data.

For the MNAR**z** mechanism, an EM algorithm an be derived (see the main paper (Sportisse et al., 2023)), because the effect of the missingness does not depend on $\mathbf{y}_i$. As the MNARz^j mechanism has the same property, an EM algorithm can also be derived (see Appendix B). However, the EM algorithm becomes untractable when the missingness depends on variables y, such models being generically denoted by MNAR$y\star$ in the sequel. In particular, some distributions entail untractable integrals at the E-step (*e.g.,* Gaussian components with MNAR$y\star$ mechanism defined with logit link, see Appendix B for more details).

The stochastic EM algorithm (Celeux and Diebolt, 1985) can overpass the EM's intractability, by imputing missing values using Gibbs sampling instead of integrating over them. In addition, it has another advantage, unlike the EM algorithm, not to be necessarily trapped by the first encountered local maximum of the likelihood function in play (Celeux and Diebolt, 1985). The principle of the SEM algorithm is to involve a stochastic-E step (SE-step) instead of the traditional E-step of the EM algorihtm. Details of the algorithm is given in Appendix C. The iteration $[r]$ then becomes:

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

pler, by noting $\mathbf{y}_i^{[r]} = (\mathbf{y}_i^{\text{obs}}, \mathbf{y}_i^{\text{mis}[r]})$,

$$\mathbf{z}_i^{[r]} \sim \mathbf{z}_i \mid \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]} \quad \text{and} \quad \mathbf{y}_i^{\text{mis [r]}} \sim \mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, \mathbf{z}_i^{[r]}, \mathbf{c}_i; \theta^{[r-1]}. \quad (13)$$

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

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

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

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

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

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

5 Implementation and numerical experiments

The SEM algorithm has been implemented for each MNAR setting for Gaussian data and is available on https://anonymous.4open.science/r/Clustering-MNAR-7E29.

For each MNAR setting, we assess the clustering performance through the consistency of the partition, by computing the ARI between the true partition Z and the estimated one, given by $\hat{Z}^{\text{MAP}} = \{z_{ik}^{\text{MAP}}(\hat{\theta})\}_{i,k} \in \mathbb{R}^{n \times K}$ as follows

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

We consider thus the following methods:

- the EM algorithm (for MCAR (11), MNARz (9) and MNARz^j (10)),
- the SEM algorithm (for MNARy (7), MNARy^k (8), MNAR$y^k z^j$ (3), MNARyz (6))

The data are generated using a Gaussian mixture with three components having unequal proportions ($\pi_1 = 0.5$, $\pi_2 = \pi_3 = 0.25$) and independent variables:

$$\forall j \in \{1, \dots, d\}, y_{ij} = \delta \sum_{k=1}^{3} \varphi_{kj} z_{ik} + \epsilon_{ij}, \tag{14}$$

with $\epsilon_{ij} \sim \mathcal{N}(0, 1)$ the noise term, $\varphi_k \in \{0, 1\}^d$ and $\delta > 0$. We consider $d = 6$ variables and we vary the number of observations $n = 100, 250, 500$. In Figure 1, as expected, considering the mechanism always gives better results than using the MCAR model, especially for models with many parameters and larger sample sizes (as the MNARyz, MNAR$y^k z^j$, MNAR$y^k z$, MNARyz^j settings for $n = 250$ and $n = 500$). Finally, consistency seems satisfactory in each scenario, indicating that our tuning parameters for the algorithm (starting values, stopping rules) are quite suitable.

In the main paper (Sportisse et al., 2023), we compare the MNARz setting with other ones, and discuss the computational cost of the estimation for the different MNAR settings.

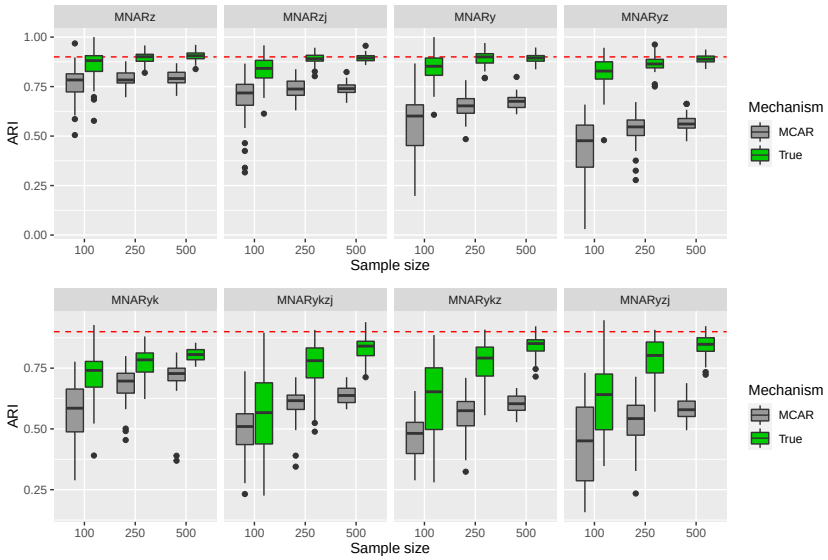

Figure 1: Boxplot of the ARI obtained for 50 samples composed of $d = 6$ variables. The sample size varies by $\{100,250,500\}$. The boxplots in green (True) correspond to the performance of the algorithm considering the MNAR setting matching the one that has been used for the missing value generation. The red dashed line indicates the theoretical ARI. This experiment been performed for a theoretical rate of misclassification of 10% and a theoretical missing rate in the whole dataset of 30%.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Aude Sportisse, Matthieu Marbac, Christophe Biernacki, Claire Boyer, Gilles Celeux, Julie Josse, and Fabien Laporte. Model-based clustering with missing not at random data. *arXiv preprint arXiv:2112.10425*, 2023.

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

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

Supplementary material

This file is a supplementary material. In Appendix A, the proof for Proposition 3.1 is given. Some complements on the EM algorithm are given in Appendix B. The SEM algorithm presented in Section 4 is detailed in Appendix C. . Appendix D gives the values of hyperparameters for the numerical experiments on synthetic data.

A Appendix 1: Identifiability

A.1 Continuous and count data

Proof of Proposition 3.1, continuous case. Suppose there exists two sets of parameters $\{\gamma, \psi\}$ and $\{\gamma', \psi'\}$ which have the same observed distribution, *i.e.*, $f(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \gamma, \psi) = f(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \gamma', \psi')$. More precisely, one has

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

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

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

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

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

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

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

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

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

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

A.2 On identifiability of the Gaussian mixture

Finite Gaussian mixtures are identifiable and, for any variable j, there is a total ordering defined by $\sigma^2_{kj} > \sigma^2_{(k+1)j}$ and $\mu_{kj} > \mu_{(k+1)j}$ if $\sigma^2_{kj} = \sigma^2_{(k+1)j}$, where μ_{kj} and σ^2_{kj} are respectively the mean and the variance of variable j under component k. Example A.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 A.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*

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

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

$$\lim_{u\to+\infty} \phi(u) = \begin{cases} 1 \text{ if } u > 0 \\ 1/2 \text{ if } u = 0 \\ 0 \text{ if } u < 0 \end{cases} \tag{15}$$

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

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

  assuming without loss of generality that $\sigma^2_k > \sigma^2_{k+1}$ or $\mu_k > \mu_{k+1}$ if $\sigma^2_k = \sigma^2_{k+1}$.

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

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

  since

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

  and

$$\lim_{u\to+\infty} \frac{\int_{-\infty}^{\alpha_{k+1}+\beta_{k+1}u} e^{-t^2/2}dt}{\int_{-\infty}^{\alpha_k+\beta_k u} e^{-t^2/2}dt} = \begin{cases} 0 \text{ if } \beta_{k+1} < 0 \\ 1/2 \text{ if } \beta_{k+1} = 0 \text{ and } \beta_k > 0 \\ 1 \text{ if } \beta_{k+1} = 0 \text{ and } \beta_k = 0. \end{cases} \tag{16}$$

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

$$\frac{\int_{-\infty}^{\alpha_{k+1}+\beta_{k+1}u} e^{-t^2/2}dt}{\int_{-\infty}^{\alpha_k+\beta_k u} e^{-t^2/2}dt} \underset{u\to+\infty}{=} \frac{\phi(\beta_{k+1}u)}{\phi(\beta_k u)} \underset{u\to+\infty}{<} \exp\left(\left(\frac{\beta_k^2}{2} - \frac{\beta_{k+1}^2}{2}\right)u^2\right). \tag{17}$$

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

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

Given that

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

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

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

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

Note first that with a logistic distribution, a total ordering cannot be defined. Indeed, for variable j, such an ordering cannot be defined if the two univariate variances are equal (*i.e.,* $\sigma_{kj}^2 = \sigma_{(k+1)j}^2$) and $\mu_{kj} - \beta_{kj} - \mu_{(k+1)j} + \beta_{(k+1)j} = 0$. However, for the specific case of Gaussian mixture where all the univariate variances are different between the components, then conditions of Proposition 3.1 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 3.1 hold true with a logistic missing-data distribution. Finally, as stated by Corollary A.2 below, the condition on the covariance matrices (including the case of homoscedastic Gaussian mixture) can be relaxed to obtain the generic identifiability of the model (*i.e.,* all not-identifiable parameter choices lie within a proper submanifold, and thus form a set of Lebesgue zero measure; Allman et al. (2009)).

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

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

Proof of Corollary A.2. We use Proposition 3.1. 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 $\sigma_k^2 = \sigma_{k+1}^2$. □

A.3 On identifiability of the Poisson mixture

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

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

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

- *If $\beta_{k+1} > 0$ (and $\beta_k > 0$): using (15), one has*

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

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

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

  using

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

  and (16) for the missing distribution part.

- *If $\beta_{k+1} < 0$ and $\beta_k < 0$: using (17), one obtains*

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

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

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

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

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

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

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

A.4 Categorical data

Proposition 3.1 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 3.1, 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.

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

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

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

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

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

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

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

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

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

□

B Appendix 2: Details on EM algorithm

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

$$Q(\theta; \theta^{[r-1]}) = \mathbb{E}[\ell_{\text{comp}}(\theta; \mathbf{y}, \mathbf{z}, \mathbf{c}) | y_i^{\text{obs}}, c_i; \theta^{[r-1]}],$$

where

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

It leads to the decomposition

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

with $t_{ik}(\theta^{[r-1]}) = f(z_{ik} = 1 \mid \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})$. The terms involved in this decomposition are now detailed.

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

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

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

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

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

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

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

$$\begin{aligned} t_{ik}(\theta^{[r-1]}) &= \frac{f(z_{ik} = 1, \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})}{f(\mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \theta^{[r-1]})} \\ &= \frac{\pi_k^{[r-1]} f_k(\mathbf{y}_i^{\text{obs}}; \lambda_k^{[r-1]}) f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \theta^{[r-1]})}{\sum_{h=1}^K \pi_h^{[r-1]} f_h(\mathbf{y}_i^{\text{obs}}; \lambda_h^{[r-1]}) f(\mathbf{c}_i \mid \mathbf{y}_i^{\text{obs}}, z_{ih} = 1; \theta^{[r-1]})} \end{aligned} \tag{18}$$

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

Computations for the MNARz^j model For the MNARz and MNARz^j models, the effect of the missingness is only due to the class membership. The EM algorithm for the MNARz model is detailed in the main paper (Sportisse et al., 2023). Term (a) is the same for both MNARz and MNARz^j models but (b) and (c) differ. For Term (b), $f(\mathbf{c}_i \mid \mathbf{y}_i, z_{ik} = 1; \psi)$ is independent of $\mathbf{y}_i$, which implies

$$\log(f(\mathbf{c}_i \mid z_{ik} = 1; \psi)) = \sum_{j=1}^d c_{ij} \log \rho(\alpha_{kj}) + (1 - c_{ij}) \log(1 - \rho(\alpha_{kj})). \tag{19}$$

For Term (c), one remark that

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

and that, by independence of $\mathbf{y}_i$, one has

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

Using (18), one obtains

$$t_{ik}^{[r-1]}(\theta^{[r-1]}) = \frac{\pi_k^{[r-1]} \phi(\mathbf{y}_i^{\text{obs}}; (\mu_{ik}^{\text{obs}})^{[r-1]}, (\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}) \prod_{j=1}^d \rho(\alpha_{kj}^{[r-1]})^{c_{ij}} (1 - \rho(\alpha_{kj}^{[r-1]}))^{1-c_{ij}}}{\sum_{h=1}^K \pi_h^{[r-1]} \phi(\mathbf{y}_i^{\text{obs}}; (\mu_{ih}^{\text{obs}})^{[r-1]}, (\Sigma_{ih}^{\text{obs,obs}})^{[r-1]}) \prod_{j=1}^d \rho(\alpha_{hj}^{[r-1]})^{c_{ij}} (1 - \rho(\alpha_{hj}^{[r-1]}))^{1-c_{ij}}} \qquad \text{(MNAR}z^j\text{)}. \tag{20}$$

The E-step is derived in the same way as for the MNAR**z** model with these terms. The M-step for ψ consists of performing a GLM with a binomial link for the matrices $(\mathcal{J}_{kj}^{\text{MNAR}z^j})_{j=1,\ldots,d}^{[r]}$ and by giving $t_{ik}(\theta^{[r-1]})$ as prior weights to fit the process.

$$(\mathcal{J}_{kj}^{\text{MNAR}z^j})^{[r]} = \quad c_{\cdot j} \quad 1 \tag{21}$$

Intractability of the EM algorithm for MNAR$y\star$ 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

$$
\begin{aligned}
&f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1, \mathbf{c}_i; \theta^{[r-1]}) \\
&= \frac{\prod_{h=1}^{d} \rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{mis}})^{c_{ih}} (1-\rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{obs}}))^{1-c_{ih}} f(\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]})}{\int_{\mathcal{Y}_i^{\text{mis}}} \prod_{h=1}^{d} \rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{mis}})^{c_{ih}} (1-\rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{obs}}))^{1-c_{ih}} f(\mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]}) d\mathbf{y}_i^{\text{mis}}} \\
&= \frac{\prod_{h, c_{ih}=1} \rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{mis}}) f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]})}{\int_{\mathcal{Y}_i^{\text{mis}}} \prod_{h, c_{ih}=1} \rho(\alpha_{kh}^{[r-1]} + \beta_{kh}^{[r-1]} y_{ih}^{\text{mis}}) f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]}) d\mathbf{y}_i^{\text{mis}}}.
\end{aligned} \tag{22}
$$

  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 (22),

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

  where

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

  Therefore,

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

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

$$
\begin{aligned}
&f(c_{ij} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]}) \\
&= \int_{\mathcal{Y}_{ij}^{\text{mis}}} f(c_{ij} \mid \mathbf{y}_i^{\text{obs}}, y_{ij}^{\text{mis}}, z_{ik}=1; \psi^{[r-1]}) f(\mathbf{y}_{ij}^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}=1; \theta^{[r-1]}) dy_{ij}^{\text{mis}} \\
&= c_{ij} \int_{-\infty}^{+\infty} \rho(\alpha_{kj}^{[r-1]} + \beta_{kj}^{[r-1]} y_{ij}^{\text{mis}}) \phi(y_{ij}^{\text{mis}}; (\tilde{\mu}_{ik}^{\text{mis}})_j^{[r-1]}, (\tilde{\Sigma}_{ik}^{\text{mis}})_{jj}^{[r-1]}) dy_{ij}^{\text{mis}} + (1-c_{ij})(1-\rho(\alpha_{kj}^{[r-1]} + \beta_{kj}^{[r-1]} y_{ij}^{\text{obs}})).
\end{aligned} \tag{23}
$$

  Using (18), the probabilities $t_{ik}(\theta^{[r-1]})$ can be deduced from Equation (23).

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

- ρ **is logistic**: Equation (22) 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

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

  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 (23) 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) 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

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

  Equation (23) 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 Appendix 3: Details on SEM algorithm

The SEM algorithm consists on two steps iteratively proceeded as presented in Section 4. The key issue is to draw the missing data $(\mathbf{y}_i^{\text{mis}})^r$ and $\mathbf{z}_i^r$ according to their current conditional distribution $f(\mathbf{y}_i^{\text{mis}}, \mathbf{z}_i \mid \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i; \pi^{[r-1]}, \lambda^{[r-1]}, \psi^{[r-1]})$. By convenience, we use a Gibbs sampling and simulate two easier probabilities recalled here

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

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

$$\begin{aligned}
&\mathbb{P}(z_{ik} = 1 \mid \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]}) \\
&= \frac{\mathbb{P}(z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \pi^{[r-1]}, \lambda^{[r-1]}, \psi^{[r-1]})}{\mathbb{P}(\mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \pi^{[r-1]}, \lambda^{[r-1]}, \psi^{[r-1]})} \qquad (24) \\
&= \frac{\mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i^{[r-1]}, z_{ik} = 1; \psi^{[r-1]})\mathbb{P}(\mathbf{y}_i^{[r-1]} \mid z_{ik} = 1; \lambda^{[r-1]})\mathbb{P}(z_{ik} = 1; \pi^{[r-1]})}{\sum_{h=1}^K \mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i^{[r-1]}, z_{ih} = 1; \psi^{[r-1]})\mathbb{P}(\mathbf{y}_i^{[r-1]} \mid z_{ih} = 1; \lambda^{[r-1]})\mathbb{P}(z_{ih} = 1; \pi^{[r-1]})} \\
&= \frac{\mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i^{[r-1]}, z_{ik} = 1; \psi^{[r-1]})\mathbb{P}(\mathbf{y}_i^{[r-1]} \mid z_{ik} = 1; \lambda^{[r-1]})\pi_k^{[r-1]}}{\sum_{h=1}^K \mathbb{P}(\mathbf{c}_i \mid \mathbf{y}_i^{[r-1]}, z_{ih} = 1; \psi^{[r-1]})\mathbb{P}(\mathbf{y}_i^{[r-1]} \mid z_{ih} = 1; \lambda^{[r-1]})\pi_h^{[r-1]}}. \qquad (25)
\end{aligned}$$

The conditional distribution of $((\mathbf{y}_i^{\text{mis}})^{[r]} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1, \mathbf{c}_i)$ has already been detailed in Equation (22) and recalled here

$$\begin{aligned}
&f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1, \mathbf{c}_i; \theta^{[r-1]}) \\
&\qquad = \frac{\prod_{j,c_{ij}=1} f(c_{ij} = 1 \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \psi^{[r-1]}) f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \lambda^{[r-1]})}{\int_{\mathcal{Y}_i^{\text{mis}}} \prod_{j,c_{ij}=1} f(c_{ij} = 1 \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \psi^{[r-1]}) f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \lambda^{[r-1]}) d\mathbf{y}_i^{\text{mis}}}. \qquad (26)
\end{aligned}$$

Gaussian mixture for continuous data First note that the probabilities of the multinomial distribution for drawing $z_i^{[r]}$ given in (25) can be easily computed for all cases.

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

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

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

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

Logistic distribution: For the logistic function, the distribution given in (26) is not classical and drawing y_i^{mis} from it seems complicated. Indeed, one has

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

where $(\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]}$ and $(\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]}$ are given in (33) and (34). 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 (26) 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 $j = 1, \ldots, d$,

$$c_{ij} = \begin{cases} 1 & \text{if } L_{ij} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{27}$$

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 (27). 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

$$(\mathbf{L}_i \mid \mathbf{y}_i^{[r-1]}, z_{ik}^{[r-1]} = 1, c_i) \sim \mathcal{N}_t(\alpha_k^{[r-1]} + \beta_k^{[r-1]}\mathbf{y}_i, I_{d\times d}; a, b), \tag{28}$$

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

$$\begin{aligned} \mathbb{P}(z_{ik} = 1 \mid \mathbf{L}_i^{[r]}, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]}) &= \frac{\mathbb{P}(z_{ik} = 1, \mathbf{L}_i^{[r]}, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]})}{\sum_{k=1}^K \mathbb{P}(z_{ik} = 1, \mathbf{L}_i^{[r]}, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]})} \\ &= \frac{f(\mathbf{L}_i^{[r]} | z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \psi^{[r-1]}) f(z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]})}{\sum_{k=1}^K f(\mathbf{L}_i^{[r]} | z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \psi^{[r-1]}) f(z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]})}. \end{aligned} \tag{29}$$

The part involving $f(z_{ik} = 1, \mathbf{y}_i^{[r-1]}, \mathbf{c}_i; \theta^{[r-1]})$ is given in (25) 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 (28) 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

$$
f(\mathbf{L}_i^{[r]} \mid \mathbf{y}_i^{\text{mis}}, \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \psi^{[r-1]}) \propto \exp\left(-\frac{1}{2}\left[(\mathbf{L}_i^{[r]} - (\alpha_k^{[r-1]} + \beta_k^{[r-1]}\mathbf{y}_i))^T(\mathbf{L}_i^{[r]} - (\alpha_k^{[r-1]} + \beta_k^{[r-1]}\mathbf{y}_i))\right]\right) \tag{30}
$$

$$
f(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik}^{[r]} = 1; \theta^{[r-1]}) \propto \exp\left(-\frac{1}{2}\left[(\mathbf{y}_i^{\text{mis}} - (\tilde{\mu}_{ik}^{\text{mis}})^{[r]})^T((\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]})^{-1}(\mathbf{y}_i^{\text{mis}} - (\tilde{\mu}_{ik}^{\text{mis}})^{[r]})\right]\right), \tag{31}
$$

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)). In particular, one has in this case:

$$
\left(\mathbf{y}_i^{\text{mis}} \mid \mathbf{y}_i^{\text{obs}}, z_{ik} = 1; \lambda^{[r-1]}\right) \sim \mathcal{N}\left((\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]}, (\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]}\right). \tag{32}
$$

with:

$$
(\tilde{\mu}_{ik}^{\text{mis}})^{[r-1]} = (\mu_{ik}^{\text{mis}})^{[r-1]} + (\Sigma_{ik}^{\text{mis,obs}})^{[r-1]}\left((\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}\right)^{-1}\left(\mathbf{y}_i^{\text{obs}} - (\mu_{ik}^{\text{obs}})^{[r-1]}\right), \tag{33}
$$

$$
(\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]} = (\Sigma_{ik}^{\text{mis,mis}})^{[r-1]} - (\Sigma_{ik}^{\text{mis,obs}})^{[r-1]}\left((\Sigma_{ik}^{\text{obs,obs}})^{[r-1]}\right)^{-1}(\Sigma_{ik}^{\text{obs,mis}})^{[r-1]}. \tag{34}
$$

Finally combining the two equations (30) and (31) one obtains

$$
\left(\mathbf{y}_i^{\text{mis}} \mid \mathbf{L}_i^{[r]}, z_{ik}^{[r]} = 1, \mathbf{y}_i^{\text{obs}}, \mathbf{c}_i\right) \sim \mathcal{N}\left(\mu_{ik}^{\text{SEM}}, \Sigma_{ik}^{\text{SEM}}\right), \tag{35}
$$

where

$$
\Sigma_{ik}^{\text{SEM}} = \left((\tilde{\Sigma}_{ik}^{\text{mis}})^{[r-1]})^{-1} + ((\beta_k^{\text{mis}})^{[r-1]})^T(\beta_k^{\text{mis}})^{[r-1]}\right)^{-1},
$$

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

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

$$
(\mathcal{H}^{\text{MNAR}y})^{[r]} = \left[\begin{array}{c|ccccc} c_{.1} & 1 & y_{.1}^{[r]} & 0 & \ldots & 0 \\ c_{.2} & 1 & 0 & y_{.2}^{[r]} & \ldots & 0 \\ \vdots & \vdots & & \ddots & \ddots & \\ c_{.d} & 1 & 0 & 0 & \ldots & y_{.d}^{[r]} \end{array}\right]. \tag{36}
$$

- 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

$$(\mathcal{H}_k^{\text{MNAR}y^k})^{[r]} = \left| \begin{array}{c|cccc} (c_{u1})_{u,z_{u1}^{[r]}=1} & 1 & (y_{u1}^{[r]})_{u,z_{u1}^{[r]}=1} & 0 & \ldots & 0 \\ \vdots & \vdots & & & \ddots & \vdots \\ (c_{u1})_{u,z_{uK}^{[r]}=1} & 1 & 0 & (y_{u1}^{[r]})_{u,z_{uK}^{[r]}=1} & & 0 \\ \vdots & \vdots & \vdots & & \ddots & \\ (c_{ud})_{u,z_{uK}^{[r]}=1} & 1 & 0 & 0 & & (y_{ud}^{[r]})_{u,z_{uK}^{[r]}=1} \end{array} \right| . \tag{37}$$

- 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

$$(\mathcal{H}^{\text{MNAR}yz})^{[r]} = \left| \begin{array}{c|cccccccc} c_{.1} & y_{.1}^{[r]} & 0 & \ldots & 0 & z_{.1}^{[r]} & \ldots & z_{.K}^{[r]} \\ c_{.2} & 0 & y_{.2}^{[r]} & \ldots & 0 & z_{.1}^{[r]} & \ldots & z_{.K}^{[r]} \\ \vdots & & \ddots & \ddots & & \vdots & \vdots & \vdots \\ c_{.d} & 0 & 0 & \ldots & y_{.d}^{[r]} & z_{.1}^{[r]} & \ldots & z_{.K}^{[r]} \end{array} \right| . \tag{38}$$

- 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

$$(\mathcal{H}_j^{\text{MNAR}yz^j})^{[r]} = \left| \begin{array}{c|ccc} c_{.j} & y_{.j}^{[r]} & z_{.1}^{[r]} & \ldots & z_{.K}^{[r]} \end{array} \right| . \tag{39}$$

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

$$(\mathcal{H}_k^{\text{MNAR}y^kz})^{[r]} = \left| \begin{array}{c|cccccc} (c_{u1})_{u,z_{uk}^{[r]}=1} & (y_{u1}^{[r]})_{u,z_{uk}^{[r]}=1} & 0 & \ldots & 0 & 1 \\ (c_{u2})_{u,z_{uk}^{[r]}=1} & 0 & (y_{u2}^{[r]})_{u,z_{uk}^{[r]}=1} & \ldots & 0 & 1 \\ \vdots & & \ddots & \ddots & & 1 \\ (c_{ud})_{u,z_{uk}^{[r]}=1} & 0 & 0 & \ldots & (y_{ud}^{[r]})_{u,z_{uk}^{[r]}=1} & 1 \end{array} \right| . \tag{40}$$

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

$$(\mathcal{H}_{kj}^{\text{MNAR}y^kz^j})^{[r]} = \left| \begin{array}{c|cc} (c_{uj})_{u,z_{uk}^{[r]}=1} & (y_{uj}^{[r]})_{u,z_{uk}^{[r]}=1} & 1 \end{array} \right| \tag{41}$$

When ρ is the probit function, the SEM algorithm can be derived, see Algorithm 1.

Remark C.1 (SEM for MNARz and MNARz^j mechanisms)**.** *A SEM algorithm can also be derived for these two mechanisms. For continuous data, we can prove 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 (the SE step is then just a draw from this law). The M-step for ψ are given in the following.

Algorithm 1 SEM algorithm for Gaussian mixture, MNAR$y\star$ models, ρ is probit

Input: $Y \in \mathbb{R}^{n\times d}$ (matrix containing missing values), $K \geq 1$, $r_{\max}$.
Initialize Z^0, π_k^0, μ_k^0, Σ_k^0 and ψ_k^0, for $k \in \{1, \ldots, K\}$.
for $r = 0$ **to** $r_{\max}$ **do**
  SE-step:
  for $i = 1$ **to** n **do**
    Draw $(\mathbf{L}_i)^{[r]}$ from the multivariate truncated Gaussian distribution given in (28).
    Draw $\mathbf{z}_i^{[r]}$ from the multinomial distribution with probabilities detailed in (29).
    Draw $(\mathbf{y}_i^{\text{mis}})^{[r]}$ from the multivariate Gaussian distribution given in (35).
  end for
  Let $Y^{[r]} = (\mathbf{y}_1^{[r]}| \ldots |\mathbf{y}_n^{[r]})$ be the imputed matrix.
  Let $Z^{[r]} = (\mathbf{z}_1^{[r]}| \ldots |\mathbf{z}_n^{[r]})$ be the partition.
  M-step:
  for $k = 1$ **to** K **do**
    Let $\pi_k^{[r]}$ be the proportion of rows of $Y^{[r]}$ belonging class k.
    Let $\mu_k^{[r]}, \Sigma_k^{[r]}$ be the mean and covariance matrix of rows of $Y^{[r]}$ belonging to class k.
    Let $\psi_k^{[r]}$ be the resulted coefficients of a GLM with a binomial link function, *i.e.*, the optimization problem is $\forall j \in \{1, \ldots, d\}$,

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

    for a matrix $\mathcal{M}_{kj}$ depending on the MNAR model (see (36), (37), (38), (43), (41) and (42)) and $\mathbf{c}_{.j}$ the missing data pattern for the variable j.
  end for
end for

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

$$(\mathcal{H}^{MNARz})^{[r]} = \left(\begin{array}{c|ccc} c_{.1} & z_{.1} & \ldots & z_{.K} \\ \vdots & \vdots & \vdots & \vdots \\ c_{.d} & z_{.1} & \ldots & z_{.K} \end{array}\right) = \left(\begin{array}{c|ccc} c_{11} & z_{11}^{[r]} & \ldots & z_{1K}^{[r]} \\ \vdots & \vdots & \vdots & \vdots \\ c_{n1} & z_{n1}^{[r]} & \ldots & z_{nK}^{[r]} \\ \vdots & \vdots & \vdots & \vdots \\ c_{1d} & z_{11}^{[r]} & \ldots & z_{1K}^{[r]} \\ \vdots & \vdots & \vdots & \vdots \\ c_{nd} & z_{n1}^{[r]} & \ldots & z_{nK}^{[r]} \end{array}\right). \tag{42}$$

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

$$(\mathcal{H}_j^{MNARz^j})^{[r]} = \left(\begin{array}{c|ccc} c_{.j} & z_{.1}^{[r]} & \ldots & z_{.K}^{[r]} \end{array}\right) \tag{43}$$

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

D Appendix 4: Complements on the numerical experiment

This section gives the values of δ (see (14)) ψ (see (3)) and φ (see (14)) used during the numerical experiments. For most MNAR models, he values of these parameters are available in the main paper (Sportisse et al., 2023).

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 2: Choice of the values of δ, α and β for all the experiments of Section 5 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 3: Choice of the values of δ, α and β for all the experiments of Section 5 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 4: Choice of the values of δ, α and β for all the experiments of Section 5 for the MNARyz^j mechanism.