Learning from missing data with the binary latent block model

Gabriel Frisch, Jean-Benoist Leger, Yves Grandvalet

Statistics and Computing manuscript No.
(will be inserted by the editor)

Learning from missing data with the binary Latent Block Model

Gabriel Frisch · **Jean-Benoist Leger** · **Yves Grandvalet**

Received: date / Accepted: date

Abstract Missing data can be informative. Ignoring this information can lead to misleading conclusions when the data model does not allow information to be extracted from the missing data. We propose a co-clustering model, based on the binary Latent Block Model, that aims to take advantage of this nonignorable nonresponses, also known as Missing Not At Random data (MNAR). A variational expectation-maximization algorithm is derived to perform inference and a model selection criterion is presented. We assess the proposed approach on a simulation study, before using our model on the voting records from the lower house of the French Parliament, where our analysis brings out relevant groups of MPs and texts, together with a sensible interpretation of the behavior of non-voters.

Keywords Latent Block Model · MNAR · variational inference · co-clustering · missing data

1 Introduction

Co-clustering simultaneously groups the rows and the columns of a data matrix. Co-clustering has found applications in many areas such as genomic analysis (Pontes et al. 2015; Kluger et al. 2003), text analysis (Dhillon et al. 2003; Selosse et al. 2020b), collaborative filtering (George and Merugu 2005; Shan and Banerjee 2008), or political analysis (Latouche et al. 2011; Wyse and Friel 2012). Co-clustering methods can be divided into categories such as, but not limited to, spectral methods (Dhillon 2001; Kluger et al. 2003), mutual information methods (Dhillon et al. 2003), modularity based methods (Labiod and Nadif 2011), non negative matrix tri-factorization (Ding et al. 2006) or model-based methods. Among the model-based methods, the Latent Block Model (Govaert and Nadif 2008; Nadif and Govaert 2010; Lomet 2012; Keribin et al. 2015) relies on mixtures, assuming that the observations are generated from finite mixture components in rows and columns.

Most standard methods of clustering or co-clustering presuppose complete information and cannot be applied with missing data, or may provide misleading conclusions when missingness is informative. A careful examination of the data generating process is necessary for the processing of missing values, which requires identifying the type of missingness (Rubin 1976): Missing Completely At Random (MCAR) refers to the mechanism in which the probability of being missing does not depend on the variable of interest or any other observed variable; whereas in Missing At Random (MAR) the probability of being missing depends on some observed data but is still independent from the non-observed data; and finally Missing Not At Random (MNAR) refers to the mechanism in which the probability of being missing depends on the actual value of the missing data. Under the MAR hypothesis, no information on the generation of data can be extracted from its absence, but under a MNAR assumption, this absence is informative, and ignoring this information in likelihood-based imputation methods may lead to strong biases in estimation (Little and Rubin 1986). Missing Not At Random is also known as non-ignorable missingness, in opposition to the ignorable missingness of MCAR and MAR settings, as the absence of data is assumed to convey some information.

In this paper, we aim at clustering the rows and columns of a binary data matrix whose entries are missing not at random. Equivalently, we consider the clustering of the vertices of a bipartite graph whose edges are missing not at random. For this purpose, we introduce a co-clustering model

Gabriel Frisch · Jean-Benoist Leger · Yves Grandvalet
Université de technologie de Compiègne, CNRS, Heudiasyc
CS 60 319 - 60 203 Compiègne Cedex, France
E-mail: firstname.lastname@utc.fr

that combines a MNAR missingness model with the Latent Block Model (LBM).

Up to our knowledge, all existing co-clustering methods consider that missing data is either MCAR or MAR (Selosse et al. 2020a; Jacques and Biernacki 2018; Papalexakis et al. 2013), except one proposed by Corneli et al. (2020) used to co-cluster ordinal data. Their model is very parsimonious as it assumes that both data and missingness are only dependent on the row and column clusters. In this setting, they are able to consider MNAR data even if they suppose that missingness depends indirectly from the value of the data. The model we propose is less parsimonious, thus more flexible, as it supposes that missingness depends both on the value of the data and on the row and column indexes (not only on their respective cluster indexes). We exemplify our missing data model on the Latent Block Model for binary data; it can be easily reused for other probabilistic co-clustering models, as it is only weakly coupled to the generative model of the full data matrix.

In the simple clustering framework, few mixture models handling MNAR data have been proposed. Marlin et al. (2011) combine a multinomial mixture clustering model, used as a complete data model, with a MNAR-type missingness model. They propose two versions of their missingness model. The first one, called CPT-v, models the data observation probability depending only on the underlying value of the data. The second one, called Logit-vd, allows the probability of a data entry to be missing to depend both on the value of the underlying data and the characteristics of the column, giving more flexibility to the model. Our missingness model respects the symmetry of the co-clustering problem by depending identically on the characteristics of the row and column. Kim and Choi (2014) propose Bayesian-BM/OR, a simple mixture model of binomials in a Bayesian formalism. Their MNAR-type model is based on three factors, related to the row, the column and the data value, all three being modeled by Bernoulli variables combined together by a "or" logical operator. The choice of this missingness model is motivated by algorithmic considerations that are not relevant for co-clustering models. Tabouy et al. (2020), in a graph perspective, deal with nonobserved dyads during the sampling of a network and consecutive issues in the inference of the stochastic block model. They propose three different MNAR sampling designs in which observing dyads depends either on their underlying value, or on the class or on the degree of the nodes. The Stochastic Block Model, though similar from the Latent Block Model we use, is not usable for co-clustering purposes.

Also related to missing data but not to clustering, MNAR is also investigated in matrix factorization. Steck (2010) derives a weighted matrix factorization model and optimizes the parameters based on a metric that is robust to MNAR data. Hernández-Lobato et al. (2014) use a double probabilistic matrix factorization model; one is for the complete data and one for the missing data, where users and items propensities are both modeled with low rank matrices. Schnabel et al. (2016) propose an empirical risk minimization framework to derive a propensity scored matrix factorization method that can account for selection bias.

We present in Section 2 the Latent Block Model introduced by Govaert and Nadif (2008). In Section 3, we introduce our model, a LBM extended to a MNAR missingness process, and propose, in Section 4, a variational EM algorithm to infer its parameters. We also introduce, in Section 5, an Integrated Completed Likelihood (ICL) criterion to tackle model selection. We then conduct experiments on synthetic datasets in Section 6 to show that the overall approach is relevant to co-cluster MNAR data. Finally, an analysis of the voting records of the lower house of the French Parliament is presented in Section 7. The source code and the dataset of the voting records are provided for reproducibility purposes at `https://github.com/gfrisch/LBM-MNAR`.

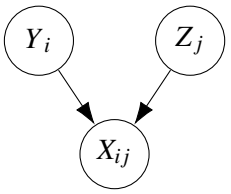
2 The Latent Block Model

The Latent Block Model (LBM) is a *co-clustering* model that classifies jointly the rows and the columns of a binary data matrix (Govaert and Nadif 2008). This probabilistic generative model assumes a double partition on the rows and the columns of a $(n_1 \times n_2)$ data matrix X that corresponds to a strong structure of the matrix in homogeneous blocks. This structure is unveiled by reordering the rows and columns according to their respective cluster index; for K row clusters and L column clusters, the reordering reveals $K \times L$ homogeneous blocks in the data matrix. Note that we adopt here the original view where the data matrix is interpreted as a data table. The binary matrix X can also be interpreted as the biadjacency matrix of a bipartite graph, whose two sets of vertices correspond to the rows and columns of the data matrix. In this interpretation, $X_{ij} = 1$ if an edge is present between "row node" i and "column node" j, and $X_{ij} = 0$ otherwise.

For the $(n_1 \times n_2)$ data matrix X, two partitions are defined by the latent variables Y and Z, with Y being the $n_1 \times K$ indicator matrix of the latent row clusters ($Y_{ik} = 1$ if row i belongs to group k and $Y_{ik} = 0$ otherwise), and Z being the $n_2 \times L$ indicator matrix of the latent column cluster. The group indicator of row i will be denoted Y_i, and similarly, the group indicator of column j will be denoted Z_j. The LBM makes several assumptions on the dependencies:

Independent rows and column clusters The latent variables Y and Z are *a priori* independent.

$$p(Y,Z) = p(Y)p(Z) \quad .$$



$$\begin{cases} \forall i,\ Y_i \overset{\text{iid}}{\sim} \mathcal{M}(1;\alpha) \\ \forall j,\ Z_j \overset{\text{iid}}{\sim} \mathcal{M}(1;\beta) \\ \forall i,j,\ X_{ij} \big| Y_{ik}=1, Z_{jl}=1 \overset{\text{ind}}{\sim} \mathcal{B}(\pi_{kl}) \end{cases}$$

with $\alpha \in \mathbf{S}_{K-1}$, $\beta \in \mathbf{S}_{L-1}$ and $\pi_{kl} \in [0,1]$

Fig. 1 Summary of the standard Latent Block Model with binary data.

Note that *a priori* independence does not imply *a posteriori* independence: given the data matrix X, the two partitions are (hopefully) not independent.

Independent and identically distributed row clusters The latent variables Y are independent and follow a multinomial distribution $\mathcal{M}(1;\alpha)$, where $\alpha = (\alpha_1,...,\alpha_K)$ contains the mixing proportions of rows:

$$p(Y;\alpha) = \prod_i p(Y_i;\alpha)$$
$$p(Y_{ik}=1;\alpha) = \alpha_k \ ,$$
with $\alpha \in \mathbf{S}_{(K-1)} = \{\alpha \in \mathbb{R}_+^K | \sum_k \alpha_k = 1\}$.

Independent and identically distributed column clusters Likewise, the latent variables Z are independent and follow a multinomial distribution $\mathcal{M}(1;\beta)$, where $\beta = (\beta_1,...,\beta_L)$ contains the mixing proportions of columns:

$$p(Z;\beta) = \prod_j p(Z_j;\beta)$$
$$p(Z_{jl}=1;\beta) = \beta_l \ ,$$
with $\beta \in \mathbf{S}_{(L-1)}$.

Given row and column clusters, independent and identically distributed block entries Given the row and colum clusters (Y,Z), the entries X_{ij} are independent and follow a Bernoulli distribution of parameter $\pi = (\pi_{kl}; k=1,...,K; l=1,...,L)$: all elements of a block follow the same probability distribution.

$$p(X|Y,Z;\pi) = \prod_{ij} p\left(X_{ij}\middle|Y_i,Z_j;\pi\right)$$
$$p\left(X_{ij}=1\middle|Y_{ik}Z_{jl}=1;\pi\right) = \pi_{kl} \ .$$

To summarize, the parameters of the LBM are $\theta = (\alpha, \beta, \pi)$ and the probability mass function of X can be written as:

$$p(X;\theta) = \sum_{(YZ)\in I\times J} \left(\prod_{ik} \alpha_k^{Y_{ik}}\right)\left(\prod_{jl} \beta_l^{Z_{jl}}\right)\left(\prod_{ijkl} \phi(X_{ij};\pi_{kl})^{Y_{ik}Z_{jl}}\right),$$

where $\phi(X_{ij};\pi_{kl}) = \pi_{kl}^{X_{ij}}(1-\pi_{kl})^{1-X_{ij}}$ is the mass function of a Bernoulli variable and where I (resp. J) denotes the set of all possible partitions of rows (resp. columns) into K (resp. L) groups.

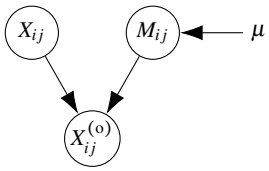


Fig. 2 Graphical representation of the MCAR model. The partially observed entry $X_{ij}^{(o)}$ is generated by the corresponding entries of the full matrix X_{ij} and the binary mask M_{ij}. The binary mask M does not depend on X and its distribution is defined by a single global effect parameter μ.

3 Extension to Informative Missing Data

The standard Latent Block Model does not accommodate missing observations, that is, the data matrix X is fully observed. This section introduces our missingness model, which will be coupled to the LBM, thereby enabling to process missing data.

We start by introducing some notation: from now on, $X^{(o)}$ will denote the "partially observed" data matrix, with missing entries, whereas X denotes the "full" (unobserved) data matrix, without missing entries. The partially observed matrix $X^{(o)}$ is identical to the full matrix X except for the missing entries; $X^{(o)}$ takes its values in $\{0,1,\text{NA}\}$, where NA denotes a missing value. It will be convenient to introduce a binary mask matrix M that indicates the *non-missing* entries of $X^{(o)}$: if $M_{ij}=0$, then $X_{ij}^{(o)} = \text{NA}$.

3.1 Models of Missingness

The three main types of missingness are Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR). We propose here a model for each missingness type. Instead of directly modeling the probability of being observed, we will model a real variable P_{ij} that defines the log-odds of this probability. This log-odds will be called here the "propensity" to be observed:

$$\forall i,j \quad P_{ij} = \log\left(\frac{p(M_{ij}=1)}{p(M_{ij}=0)}\right) \ .$$

Missing Completely At Random (MCAR) Missingness does not depend on data, whether observed or not. A simple model of missingness is obtained by assuming that every entry of $X^{(o)}$ has the same propensity of being missing. This is modeled by a single propensity parameter μ. The graphical representation of this model is shown in Figure 2.

Missing At Random (MAR) Missingness depends on the observed data $X_{ij}^{(o)} \in \{0,1,\text{NA}\}$ for all (i,j), but not on the unobserved data, that is, any X_{ij} corresponding to $X_{ij}^{(o)} = \text{NA}$.

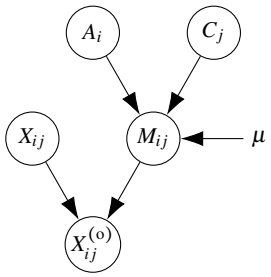


Fig. 3 Graphical representation of the MAR model. The partially observed entry $X_{ij}^{(o)}$ is generated by the corresponding entries of the full matrix X_{ij} and the binary mask M_{ij}. The binary mask M does not depend on X and its distribution is defined by a global effect parameter μ and two latent variables A and C that enable deviations from μ.

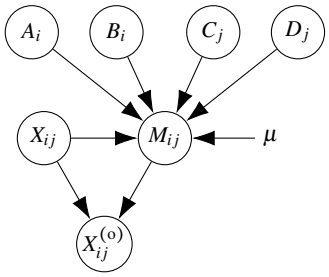


Fig. 4 Graphical representation of the MNAR model. The partially observed entry $X_{ij}^{(o)}$ is generated by the corresponding entries of the full matrix X_{ij} and the binary mask M_{ij}. The binary mask M depends on X and its distribution is also defined by a global effect parameter μ, two latent variables A and C that enable deviations from μ, and two latent variables B and D that drive the deviations from the MAR model.

The previous missingness model can be enlarged by allowing the propensity of missingness to depend on the row and column indexes. To do so, we can introduce a latent variable for every row, denoted A, and another one for every column, denoted C. For the sake of simplicity, all latent variables A_i and C_j are assumed independent. They allow deviations from the global propensity μ. The graphical representation of this model is shown in Figure 3.

Missing Not At Random (MNAR) Missingness here depends on unobserved data: the probability of observing the entries of the matrix depends on their values, whether observed or not. We equip the previous model with two additional latent variables to adapt the propensity of each entry of the data matrix to the unobserved data, that is, to X_{ij}. These new row and column latent variables, B and D, adjust the propensity of missingness according to the actual value of X_{ij}. The graphical representation of this model is shown in Figure 4.

We model the latent variables A, B, C, and D with Gaussian distributions centered at zero with free variances σ_A^2, σ_B^2, σ_C^2, and σ_D^2, respectively:

$$\begin{cases} \forall i, & A_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_A^2\right), & B_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_B^2\right) \\ \forall j, & C_j \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_C^2\right), & D_j \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_D^2\right) \end{cases} .$$

The global parameter μ and the latent variables define the propensity of missingness, that is, the log-odds of being missing as follows:

$$\forall i, j \quad P_{ij} = \begin{cases} \mu + A_i + B_i + C_j + D_j & \text{if} \quad X_{ij} = 1 \\ \mu + A_i - B_i + C_j - D_j & \text{if} \quad X_{ij} = 0 \end{cases} .$$

Then, given this propensity, every element M_{ij} of the mask matrix is independent and follows a Bernoulli distribution:

$$\forall i, j \quad M_{ij} \big| A_i, B_i, C_j, D_j, X_{ij} \overset{\text{ind}}{\sim} \mathcal{B}\left(\text{expit}\left(P_{ij}\right)\right) ,$$

with $\text{expit}(x) = 1/(1+\exp(-x))$.

Note that, if we omit the latent variables B_i and D_j, the missingness model follows the MAR assumption since P_{ij}, and thus M_{ij}, is then independent of X_{ij}. If we also omit the latent variables A_i and C_j, the missingness model follows the MCAR assumption: the missingness models are nested.

These models of missingness can be used for several applications. One of these, collaborative filtering, uses the history of user ratings to build a recommendation system. For this application, an MCAR modeling means that the probability of observing a rating for a particular item does not depend on the user nor the item; an MAR modeling means that missingness can depend on the user or the item; for example, some people give their opinion more often than others. The MAR simplifying assumption is often used in collaborative filtering. However, Marlin et al. (2007) show that there is often a dependency between the rating frequency and the underlying preference level, lending support to the hypothesis that ratings are generated by a MNAR process, where missingness depends on the actual rating that would be given. Some people give their opinion more often when they are satisfied and other ones when they are dissatisfied. Most collaborative filtering methods do not have a principled method for extracting information from missing data, which can lead to strong biases in estimations that may in turn drastically affect predictions (Hernández-Lobato et al. 2014). Our missingness model allows one to account for the users' propensity to give their opinion, and for the items' propensity to be rated, that is, their notoriety. These propensities could also reflect exogenous factors such as price; for example, more expensive items could be evaluated more often.

3.2 LBM with MNAR data

We extend the standard LBM using the previous modeling to MNAR data. Given the full matrix X and the mask matrix M, all the elements of the observed matrix $X^{(o)}$ are generated as follows:

$$X_{ij}^{(o)} = \begin{cases} X_{ij} & \text{if} \quad M_{ij} = 1 \\ \text{NA} & \text{if} \quad M_{ij} = 0 \end{cases} .$$

Figure 5 summarizes the LBM extented to MNAR data. $X^{(o)}$ taking its values in $(0,1,\text{NA})$, the same model can be rewritten without M_{ij}, thanks to a categorial distribution (that is, a multinomial distribution with one trial) using directly the latent variables of the missingness model:

$$\begin{aligned}
&\forall i,\, Y_i \overset{\text{iid}}{\sim} \mathcal{M}(1;\alpha) && \forall j,\, Z_j \overset{\text{iid}}{\sim} \mathcal{M}(1;\beta)\\
&\forall i,\, A_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0,\sigma_A^2\right) && \forall j,\, C_j \overset{\text{iid}}{\sim} \mathcal{N}\left(0,\sigma_C^2\right)\\
&\forall i,\, B_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0,\sigma_B^2\right) && \forall j,\, D_j \overset{\text{iid}}{\sim} \mathcal{N}\left(0,\sigma_D^2\right)
\end{aligned} \tag{1}$$

$$X_{ij}^{(o)} \Big| Y_{ik}=1, Z_{jl}=1, A_i, B_i, C_j, D_j \overset{\text{ind}}{\sim} \text{cat}\left(\begin{bmatrix}0\\1\\\text{NA}\end{bmatrix}, \begin{bmatrix}p_0\\p_1\\1-p_0-p_1\end{bmatrix}\right)$$

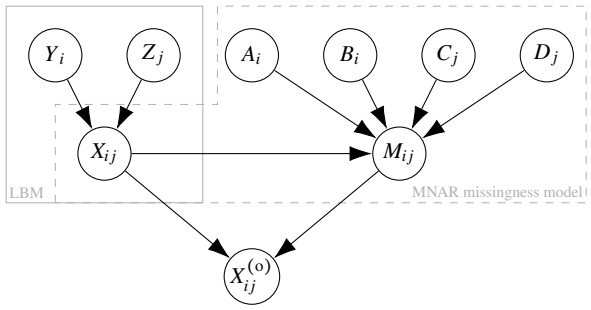
with

$$p_0 = (1-\pi_{kl})\,\text{expit}\left(\mu + A_i - B_i + C_j - D_j\right) \tag{2}$$

$$p_1 = \pi_{kl}\,\text{expit}\left(\mu + A_i + B_i + C_j + D_j\right)\ . \tag{3}$$

The parameters of the LBM with MNAR data are $\theta = (\alpha, \beta, \pi, \mu, \sigma_A^2, \sigma_B^2, \sigma_C^2, \sigma_D^2)$ with size $K+L+K\times L+5$.

Keribin et al. (2015) proved under mild assumptions that for a probability distribution $p_\theta(X)$ of a binary Latent Block Model, there exists a unique set of parameters $\theta = (\pi, \alpha, \beta)$, up to a permutation of row and column labels. Our missingness model belongs to the linear mixed effects model family whose identifiability depends on the covariance structures. Wang (2013) derived conditions of identifiability for the covariance parameters in a linear mixed effect model and study some commonly used covariance structures. We did not reach a proof of identifiability for the extended LBM with MNAR data. However, the stability of the inference shown by experiments on synthetic data (Section 6 and Annex D) suggests that there may be conditions under which the joint model would be identifiable.



Latent Block Model

$$Y_i \overset{\text{iid}}{\sim} \mathcal{M}(1;\alpha), \qquad \alpha \in \mathbf{S}_{K-1}$$

$$Z_j \overset{\text{iid}}{\sim} \mathcal{M}(1;\beta), \qquad \beta \in \mathbf{S}_{L-1}$$

$$\left(X_{ij} | Y_{ik}=1, Z_{jl}=1\right) \overset{\text{ind}}{\sim} \mathcal{B}(\pi_{kl}), \qquad \pi_{kl} \in (0,1)$$

MNAR model

$$A_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0,\sigma_A^2\right), \qquad \sigma_A^2 \in \mathbb{R}_+^*$$

$$B_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0,\sigma_B^2\right), \qquad \sigma_B^2 \in \mathbb{R}_+^*$$

$$C_j \overset{\text{iid}}{\sim} \mathcal{N}\left(0,\sigma_C^2\right), \qquad \sigma_C^2 \in \mathbb{R}_+^*$$

$$D_j \overset{\text{iid}}{\sim} \mathcal{N}\left(0,\sigma_D^2\right), \qquad \sigma_D^2 \in \mathbb{R}_+^*$$

$$\left(M_{ij} | A_i, B_i, C_j, D_j, X_{ij}=1\right) \overset{\text{ind}}{\sim} \mathcal{B}(\text{expit}\left(\mu + A_i + B_i + C_j + D_j\right))$$

$$\left(M_{ij} | A_i, B_i, C_j, D_j, X_{ij}=0\right) \overset{\text{ind}}{\sim} \mathcal{B}(\text{expit}\left(\mu + A_i - B_i + C_j - D_j\right))$$

Observations are generated according to:

$$X_{ij}^{(o)} = \begin{cases} X_{ij} & \text{if} \quad M_{ij}=1 \\ \text{NA} & \text{if} \quad M_{ij}=0 \end{cases}$$

Fig. 5 Graphical view and summary of the Latent Block Model extended to MNAR missingness process. The observed data $X_{ij}^{(o)}$ is generated by the necessary information carried by the class and propensity of row i and by the class and propensity of the column j.

4 Inference in the extented LBM

The dependency between the full data matrix X and the mask matrix M requires a joint inference of the LBM with the MNAR model. As the standard maximum likelihood approach cannot be applied directly, we adopt a strategy based on a variational EM.

During inference, we use the reformulation of Equation (1). We can split our random variables into two sets: the set of unobserved latent variables and the set of observed variables consisting of $X^{(o)}$ only. An observation of $X^{(o)}$ only is called the incomplete data, and an observation of $X^{(o)}$ together with the latent variables A, B, C, D, Y and Z is called the complete data. Given the incomplete data, our objective is to infer the model parameters θ via maximum likelihood $\hat{\theta} = \arg\max_\theta p(X^{(o)};\theta)$.

This likelihood of the incomplete data should be obtained by marginalization over all latent variables $p(X^{(o)};\theta) = \sum_{YZ}\int_{ABCD} p(X^{(o)},Y,Z,A,B,C,D;\theta)$, which is rapidly untractable as it involves an exponentially growing sum over all possible values of the latent variables.

We resort to the Expectation Maximization (EM) algorithm to maximize $p(X^{(o)};\theta)$ without explicitly calculating it. The EM algorithm iteratively applies the two following steps:

E-step Expectation step: from the current estimate $\theta^{(t)}$ of θ, compute the criterion $\mathcal{Q}(\theta|\theta^{(t)})$ defined as the expectation of the complete log-likelihood, conditionally on the observations $X^{(o)}$:

$$\mathcal{Q}(\theta|\theta^{(t)}) = \mathbb{E}_{Y,Z,A,B,C,D|X^{(o)},\theta^{(t)}}\left[\log p\left(X^{(o)},Y,Z,A,B,C,D;\theta\right)\right].$$

M-step Maximization step: find the parameters that maximize $\mathcal{Q}(\theta|\theta^{(t)})$.

$$\theta^{(t+1)} = \arg\max_{\theta} \mathcal{Q}(\theta|\theta^{(t)}) \ .$$

The computation of the complete log-likelihood at the E-step requires the posterior distribution of the latent variables $p(Y,Z,A,B,C,D|X^{(o)})$ which is intractable, because the search space of the latent variables is combinatorially too large. This problem is well known in the context of co-clustering; for the Latent Block Model, Celeux and Diebolt (1985); Keribin et al. (2015) propose a stochastic E-step with Monte Carlo sampling, but this strategy is not suited to large-scale problems. We follow the original strategy proposed by Govaert and Nadif (2008), which relies on a variational formulation of the problem, since it is more efficient in high dimension.

4.1 Variational EM

The variational EM (VEM) (Jordan et al. 1999; Jaakkola 2000) introduces $q(\cdot)$, a parametric inference distribution defined over the latent variables Y, Z, A, B, C, D and optimizes the following lower bound on the log-likelihood of the incomplete data:

$$\mathcal{J}(q,\theta) = \log p\left(X^{(o)};\theta\right) - KL\left(q(\cdot) \parallel p(\cdot|X^{(o)};\theta)\right) \ ,$$

where KL stands for the Kullback-Leibler divergence and $q(\cdot)$ denotes the variational distribution over the latent variables Y, Z, A, B, C, D. It can be shown that $\mathcal{J}(q,\theta)$ is a concave function of the variational distribution q and that its maximum is reached for $q(\cdot) = p(\cdot|X^{(o)};\theta)$. Thus, maximizing the criterion $\mathcal{J}$ is equivalent to minimizing the discrepancy between $q(\cdot)$ and $p(\cdot|X^{(o)};\theta)$, as measured by the KL-divergence, and is also equivalent to maximizing the likelihood. The minimization of this KL-divergence requires one to explore the whole space of latent distributions; the difficulty of the problem is equivalent, in terms of complexity, to the initial problem.

The criterion $\mathcal{J}(q,\theta)$ can also be expressed as the sum of a negative "energy" and the entropy of q hence its name "negative variational free energy" in analogy with the thermodynamic free energy:

$$\mathcal{J}(q,\theta) = \mathcal{H}(q) + \mathbb{E}_q\left[\log p\left(X^{(o)},Y,Z,A,B,C,D;\theta\right)\right] \ , \quad (4)$$

where $\mathcal{H}(q)$ is the entropy of the variational distribution and $\mathbb{E}_q$ is the expectation with respect to the variational distribution. This criterion can become tractable by restricting the search space of variational distributions to a subspace; the maximum found is then a lower bound of the initial criterion. The distributions in this subspace are denoted q_γ and $\mathcal{J}\left(q_\gamma,\theta\right)$ is known as the "Evidence Lower BOund" (ELBO) emphasizing the lower bound property on the evidence of the data.

A wise choice of the restriction on the variational distribution leads to a feasible computation of the criterion. We choose to consider the following posterior shapes on the latent variables:

$$\begin{aligned}
\forall i \quad & Y_i|X^{(o)} \underset{q_\gamma}{\sim} \mathcal{M}\left(1;\tau_i^{(Y)}\right) \\
\forall j \quad & Z_j|X^{(o)} \underset{q_\gamma}{\sim} \mathcal{M}\left(1;\tau_j^{(Z)}\right) \\
\forall i \quad & A_i|X^{(o)} \underset{q_\gamma}{\sim} \mathcal{N}\left(\nu_i^{(A)},\rho_i^{(A)}\right) \\
\forall i \quad & B_i|X^{(o)} \underset{q_\gamma}{\sim} \mathcal{N}\left(\nu_i^{(B)},\rho_i^{(B)}\right) \\
\forall j \quad & C_j|X^{(o)} \underset{q_\gamma}{\sim} \mathcal{N}\left(\nu_j^{(C)},\rho_j^{(C)}\right) \\
\forall j \quad & D_j|X^{(o)} \underset{q_\gamma}{\sim} \mathcal{N}\left(\nu_j^{(D)},\rho_j^{(D)}\right) \ .
\end{aligned}$$

We also impose the conditional independence of the latent variables to get a feasible computation of the entropy and of the negative "energy" (Equation 4) under q_γ. This conditional independence is widely known as the "mean field approximation" (Parisi 1988). We finally get the following fully factorized shape:

$$\begin{aligned}
q_\gamma = & \prod_{i=1}^{n_1} \mathcal{M}\left(1;\tau_i^{(Y)}\right) \times \prod_{j=1}^{n_2} \mathcal{M}\left(1;\tau_j^{(Z)}\right) \\
& \times \prod_{i=1}^{n_1} \mathcal{N}\left(\nu_i^{(A)},\rho_i^{(A)}\right) \times \prod_{i=1}^{n_1} \mathcal{N}\left(\nu_i^{(B)},\rho_i^{(B)}\right) \\
& \times \prod_{j=1}^{n_2} \mathcal{N}\left(\nu_j^{(C)},\rho_j^{(C)}\right) \times \prod_{j=1}^{n_2} \mathcal{N}\left(\nu_j^{(D)},\rho_j^{(D)}\right) \ ,
\end{aligned}$$

where $\gamma = (\tau^{(Y)},\tau^{(Z)},\nu^{(A)},\rho^{(A)},\nu^{(B)},\rho^{(B)},\nu^{(C)},\rho^{(C)},\nu^{(D)},\rho^{(D)})$ denotes the parameters' concatenation of the restricted variational distribution q_γ.

The new criterion $\mathcal{J}(\gamma,\theta)$ that we want to optimize from now on is:

$$\mathcal{J}(\gamma,\theta) = \mathcal{H}(q_\gamma) + \mathbb{E}_{q_\gamma}\left[\log p\left(X^{(o)},Y,Z,A,B,C,D;\theta\right)\right], \quad (5)$$

and the initial estimates of the model parameters $\hat{\theta}$ are inferred as:

$$\hat{\theta} = \arg\max_{\theta} \left(\max_{\gamma} \mathcal{J}(\gamma,\theta)\right) \ .$$

This double maximization is realized with an iterative strategy and can be seen as an extension of the EM algorithm. The two steps are described in Algorithm 1.

4.2 Computation of the variational criterion

The restriction on the space of the variational distribution simplifies the computation of $\mathcal{H}(q_\gamma)$ as entropy is addi-

Algorithm 1: Variational Expectation Maximization algorithm

Input: observed data $X^{(o)}$, K and L number of row groups and column groups ;
Initialize $\gamma^{(0)}$ and $\theta^{(0)}$;
while *not convergence of criterion* $\mathcal{J}$ **do**
  VE-step: find the variational parameters $\gamma^{(t+1)}$ that optimize $\mathcal{J}\left(\gamma, \theta^{(t)}\right)$

$$\gamma^{(t+1)} = \arg\max_{\gamma} \mathcal{J}\left(\gamma, \theta^{(t)}\right)$$

  M-step: find the model parameters $\theta^{(t+1)}$ that optimize $\mathcal{J}\left(\gamma^{(t+1)}, \theta\right)$:

$$\theta^{(t+1)} = \arg\max_{\theta} \mathcal{J}\left(\gamma^{(t+1)}, \theta\right)$$

end
Result: θ and γ: model and variational parameters

tive on independent variables:

$$\begin{aligned}\mathcal{H}(q_\gamma) = &- \sum_{ik} \tau_{ik}^{(Y)} \log \tau_{ik}^{(Y)} - \sum_{jl} \tau_{jl}^{(Z)} \log \tau_{jl}^{(Z)} \\ &+ \frac{1}{2}\sum_{i} \log\left(2\pi e \rho_i^{(A)}\right) + \frac{1}{2}\sum_{i} \log\left(2\pi e \rho_i^{(B)}\right) \\ &+ \frac{1}{2}\sum_{j} \log\left(2\pi e \rho_j^{(C)}\right) + \frac{1}{2}\sum_{j} \log\left(2\pi e \rho_j^{(D)}\right) \ .\end{aligned}$$

The independence of latent variables allows one to rewrite the expectation of the complete log-likelihood as:

$$\begin{aligned}\mathbb{E}_{q_\gamma}\left[\log p\left(X^{(o)}, Y, Z, A, B, C, D\right)\right] = &\ \mathbb{E}_{q_\gamma}[\log p(Y)] \\ &+ \mathbb{E}_{q_\gamma}[\log p(Z)] + \mathbb{E}_{q_\gamma}[\log p(A)] + \mathbb{E}_{q_\gamma}[\log p(B)] \\ &+ \mathbb{E}_{q_\gamma}[\log p(C)] + \mathbb{E}_{q_\gamma}[\log p(D)] \\ &+ \mathbb{E}_{q_\gamma}\left[\log p\left(X^{(o)} \middle| Y, Z, A, B, C, D\right)\right] \ .\end{aligned} \tag{6}$$

Despite the variational approximation, the expectation of the complete log-likelihood (6) cannot be exactly computed as its last term involves an expectation under q_γ of nonlinear functions:

$$\begin{aligned}\mathbb{E}_{q_\gamma}\left[\log p\left(X^{(o)} \middle| Y, Z, A, B, C, D\right)\right] = &\\ \sum_{ijkl: X_{ij}^{(o)}=0} &\tau_{ik}^{(Y)} \tau_{jl}^{(Z)} \mathbb{E}_{q_\gamma}[\log(p_0)] \\ + \sum_{ijkl: X_{ij}^{(o)}=1} &\tau_{ik}^{(Y)} \tau_{jl}^{(Z)} \mathbb{E}_{q_\gamma}[\log(p_1)] \\ + \sum_{ijkl: X_{ij}^{(o)}=\mathrm{NA}} &\tau_{ik}^{(Y)} \tau_{jl}^{(Z)} \mathbb{E}_{q_\gamma}[\log(1-p_0-p_1)] \ ,\end{aligned} \tag{7, 8}$$

with p_0 and p_1 defined in Equations (2)–(3).

These expectations can be approximated by Taylor expansions assuming a small variance of the Gaussian variational variables. This method has similarities with the delta method (Wasserman 2004, p. 79) with normal asymptotics and variances tending to zero. Using a first order Taylor expansion would lead to a criterion without maximum, so we use a second order Taylor expansion. The full expression of the criterion is given in Appendix A.

4.3 Maximization of the variational criterion

The VEM Algorithm 1 alternates maximizations with respect to the variational parameters γ and with respect to the model parameters θ. For our model, there is no explicit solution for the two maximizations of the criterion $\mathcal{J}(\gamma, \theta)$, which are carried out by the Limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) algorithm. We used automatic differentiation to compute the gradients needed for L-BFGS and for the Taylor series used in the variational criterion. We chose the Autograd library from HIPS and the submodule Autograd from PyTorch (Paszke et al. 2019). These libraries rely on a reverse accumulation computational graph to compute exact gradients. Their high efficiency, even with large graphs, thanks to GPU acceleration, makes them particularly well adapted for the VEM algorithm.

4.4 Initialization

VEM does not ensure convergence towards a global optimum. The EM-like algorithms are known to be sensitive to the initialization, particularly when applied to models with discrete latent space, and may get stuck in unsatisfactory local maxima (Biernacki et al. 2003; Baudry and Celeux 2015).

A simple solution consists in training for a few iterations from several random initializations, and pursuing optimization with the solution with highest value of the variational criterion (see, e.g., small EM for mixtures Baudry and Celeux 2015). This exploration strategy spends a great deal of computing resources to bring out only a few good estimates. Another solution is to rely on simpler clustering methods, such as k-means or spectral clustering, to initialize the algorithm (Shireman et al. 2015).

The parameters of the Stochastic Block Model, a close relative of the Latent Block Model for graphs, can be consistently identified by spectral clustering (Rohe et al. 2011). Following this idea, we use a double spectral clustering (with absolute eigenvalues of the Laplacian as in Rohe et al. 2011) on rows and columns on the similarity matrices XX^T and X^TX, to initialize our algorithm. Although this method is not designed for MNAR data, it can be expected to provide

a satisfying initialization of the Latent Block Model if the missingness is not predominant. The parameters of our missingness model cannot be initialized with this procedure; they are randomly initialized. The overall initialization procedure is described in Appendix B.

5 Model selection

5.1 Integrated Completed Likelihood criterion (ICL)

ICL, inspired by the Bayesian Information Criterion, was originally proposed to select a relevant number of classes for mixture models (Biernacki et al. 1998). It was extended to select an appropriate number of (row and column) clusters in the standard Latent Block Model (Keribin et al. 2012): for K row classes and L column classes, the criterion is defined as

$$\log \int p(X,Y,Z|\theta;K,L)p(\theta;K,L)\mathrm{d}\theta \quad ,$$

with $p(\theta;K,L)$ the prior distribution of parameters. By taking into account the latent variables Y,Z, ICL is a clustering-oriented criterion, whereas BIC or AIC are driven by the faithfulness to the distribution of X (Biernacki et al. 1998).

For the LBM with MNAR missingness, ICL requires priors on the parameters of the missingness model. We chose independent InverseGamma(1, 1) distributions for the parameters σ_A^2, σ_B^2, σ_C^2 and σ_D^2. As in Keribin et al. (2012), we use non-informative Dirichlet distribution priors on the α and β parameters of class mixing proportions.

Proposition 1 *The asymptotic ICL criterion,*

$$\begin{aligned} ICL^{\infty}(K,L) = & \max_{\theta,Y,Z,A,B,C,D} \log p\left(X^{(\mathrm{o})},Y,Z,A,B,C,D;\theta\right) \\ & - \frac{K-1}{2}\log(n_1) - \frac{L-1}{2}\log(n_2) \\ & - \frac{KL+1}{2}\log(n_1 n_2) - \log(n_1 n_2) \quad , \end{aligned}$$

for the LBM with the MNAR model of Section 3.2 is, up to an irrelevant constant, an asymptotic expansion of the log integrated completed likelihood

$$\log \int p(X,Y,Z,A,B,C,D|\theta;K,L)p(\theta;K,L)\mathrm{d}\theta \quad .$$

See proof in Appendix C.

As seen in Section 4, the maximized completed log-likelihood required for the asymptotic ICL cannot be calculated; in practice we use the expectation of the completed log-likelihood under the variational posterior, computed by the difference between the lower bound provided by the variational approximation and the entropy of the variational distribution (see Equation 5). The asymptotic ICL is thus approximated by:

$$\begin{aligned} & \mathcal{J}\left(\hat{\gamma},\hat{\theta}\right) - \mathcal{H}(q_{\hat{\gamma}}) - \frac{K-1}{2}\log(n_1) - \frac{L-1}{2}\log(n_2) \\ & \qquad - \frac{KL+1}{2}\log(n_1 n_2) - \log(n_1 n_2) \quad , \end{aligned}$$

where $\left(\hat{\gamma},\hat{\theta}\right) = \arg\max_{\gamma,\theta} \mathcal{J}(\gamma,\theta)$.

An asymptotic ICL criterion for the LBM with MAR data can be constructed in the same way, allowing for comparison with the MNAR model as the models are nested (see details in Appendix C).

6 Experiments on simulated data

Simulated data brings all the elements to assess clustering algorithms in controlled settings. Using controlled datasets provides the means to properly test the ability of an algorithm to recover the known underlying structure.

6.1 Difficulty of a co-clustering task

In co-clustering, several loss functions are suited for measuring the discrepancy between the underlying classes (Y,Z) and some predictions $(\widehat{Y},\widehat{Z})$. For our experiments, we will use the measure defined by Govaert and Nadif (2008), that is, the ratio of misclassified entries in the data matrix:

$$l_{item}(Y,Z,\widehat{Y},\widehat{Z}) = 1 - \max_{t\in\Omega_1, s\in\Omega_2} \frac{1}{n_1 n_2} \sum_{ijkl} Y_{ik}\widehat{Y}_{it(k)}Z_{jl}\widehat{Z}_{js(l)} \quad ,$$

where Ω_1 (resp. Ω_2) is the set of all possible permutations of $\{1,\ldots,K\}$ (resp. $\{1,\ldots,L\}$), introduced to take into account the fact that cluster indexes are known only up to a permutation.

In standard clustering, the difficulty of a task is often assessed by its Bayes risk, that is, by the minimum of the expectation of the loss function, which is typically approximated by Monte Carlo on simulated data. Co-clustering poses specific difficulties. Adding more rows or more columns alters its difficulty because the dimensions of the spaces where the clustering is performed are expanded. The duality between the rows and the columns implies that the size of the matrix is a characteristic of a co-clustering problem. In other words, given a fixed generative distribution, as the matrix size increases, the difficulty of the task decreases, in contrast to simple clustering, where the difficulty, as measured by the Bayes risk, remains constant when more examples (that is, rows) are added.

A simple Monte Carlo approximation of the risk consists in averaging over many statistical units. In simple clustering, this means generating a great number of rows in a data matrix. In co-clustering, the statistical unit is the whole matrix,

implying that a Monte Carlo approximation of the risk is obtained by generating a great number of data matrices, which then involves a great computational time. Furthermore, estimating the Bayes risk from a single data matrix is very inconstant; the risk may be very different between two data matrices of the same size generated from the same distribution. Hence the usual notion of Bayes risk is not appropriate for co-clustering. Lomet et al. (2012) argue that conditioning the Bayes risk on the observed matrix is more appropriate. They give a protocol to simulate data matrices in which the difficulty of the clustering task is controlled by the following *conditional Bayes risk*:

$$r_{item}(\widehat{Y},\widehat{Z}) = \mathbb{E}\Big[\, l_{item}(Y,Z,\widehat{Y},\widehat{Z}) \Big| X^{(o)} \Big] \;, \tag{9}$$

where the expectation is taken over Y,Z only and $\widehat{Y},\widehat{Z}$ are the clusterings returned by the *conditional Bayes classifier*, that is, the maximum *a posteriori*:

$$(\widehat{Y},\widehat{Z}) = \underset{Y,Z}{\arg\min}\, r_{item}(Y,Z) = \underset{Y,Z}{\arg\max} \sum_{ij} p\Big(Y_i, Z_j \Big| X^{(o)}\Big) \;.$$

The expectation (9) involves the non tractable posterior of the latent variables $p(Y,Z|X^{(o)})$ (see Section 4). The expectation is approximated by an average obtained from a Gibbs sampler of $(Y,Z|X^{(o)})$.

Lomet et al. (2012) released data sets, with different sizes and difficulties, simulated from the Latent Block Model. Using their protocol, we generated new data according the LBM with a MNAR missingness process. Data sets are generated according to the LBM with three row and column classes, with parameters

$$\alpha = \beta = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} \qquad \text{and} \qquad \pi = \begin{pmatrix} \varepsilon & \varepsilon & 1-\varepsilon \\ \varepsilon & 1-\varepsilon & 1-\varepsilon \\ 1-\varepsilon & 1-\varepsilon & \varepsilon \end{pmatrix} \;, \tag{10}$$

where ε defines the difficulty of the clustering task. The parameters of the MNAR process are

$$\mu = 1, \quad \sigma_A^2 = 1, \quad \sigma_B^2 = 1, \quad \sigma_C^2 = 1, \quad \sigma_D^2 = 1 \;, \tag{11}$$

which gives an average proportion of 35% of missing values.

6.2 Class prediction

We test here the ability of the proposed inference scheme to recover row and column classes. To conduct the experiments, we generate an initial data matrix of size $n_1 = n_2 = 500$ with a conditional Bayes risk of 5% set by choosing ε (10) by trial and error. The size of this matrix is then progressively reduced, removing rows and columns, to increase the difficulty of the classification task. The conditional Bayes risk is re-estimated on each sub-matrix to provide a reference. Our algorithm is then run on these data matrices using 20 initializations for each run, as described in Section 4.4. We then predict the row and column classes (Y,Z) with their maximum *a posteriori* estimators on the variational distribution. This whole process is repeated 20 times, leading to the results presented in Figure 6.

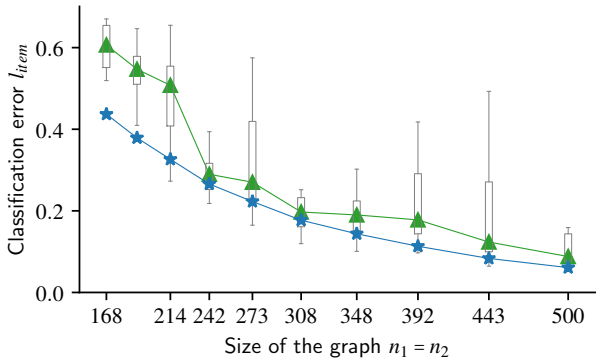


Fig. 6 Classification error with respect to the size of the data matrix (lower is better); ⋆ is the median of the conditional Bayes risk; ▲ is the median prediction error obtained by our algorithm.

As expected, the conditional Bayes risk decreases as the data matrices grow. The predictions returned by our algorithm follow the same pattern, with a diminishing gap to the conditional Bayes risk as the data matrices grow, which is consistent with our expectations. Appendix D provides additional experimental results that suggest consistent estimations of the model parameters.

6.3 MNAR versus MAR model for MNAR data

The importance of using the right missingness model is tested by comparing the classifications returned by an LBM with and without an MNAR model. A data set is generated according to the LBM with MNAR values where the parameters α, β and π of the LBM are fixed as in (10), and ε is chosen in order to get a conditional Bayes risk of 12%, for data matrices of size $n_1 = n_2 = 100$; the MNAR model parameters μ, σ_A^2 and σ_C^2 are all set to one which gives an average proportion of 35% of missing values. Several data matrices are generated using these parameters while varying the value of the σ_B^2 and σ_D^2 parameters that govern the MNAR effects; these variations do not affect the conditional Bayes risk nor the proportion of missing values as latent variables B and D follow Gaussian distributions centered at zero (see Figure 5). For each data matrix, we train the LBM with either the MAR or the MNAR model. We also train a categorical LBM (Keribin et al. 2015) considering missing values to be a level of the categorical distribution using the "blockcluster" package (Bhatia et al. 2014). This process is repeated 20 times, starting from the generation of a new fully observed data matrix.

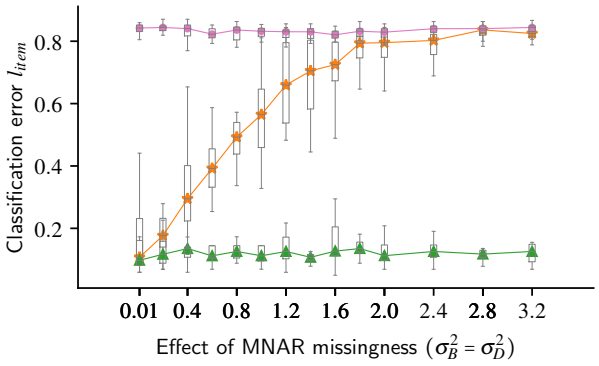


Fig. 7 Classification error with respect to an increase of the MNAR effect (lower is better); ⋆ is the median prediction error obtained with the MAR model; ▲ is the median prediction error obtained with the MNAR model; ● is the median prediction error obtained with the categorical LBM.

The median of the classification errors l_{item} are presented in Figure 7 as a function of the MNAR effect. They are essentially constant and close to the conditional Bayes risk for the LBM with the MNAR model, whereas the LBM with the MAR model is badly affected by MNAR data, eventually leading to a classification close to a totally random allocation[1]. Ignoring the nature of the missingness process leads here to strong biases in estimation that in turn drastically affect classification. Thankfully, the ICL criterion may be of great help to select the right missingness model as shown in Section 6.5. The classification errors obtained with the categorical LBM are steadily close to the one of the totally random allocation. Explaining missingness by cluster membership is not relevant here and impairs the fit of the model. For the same reason, similarly poor performances (not shown here) were obtained using the ordinal LBM with MNAR missingness of Corneli et al. (2020), in which data and missingness depend on the same row and column clusters.

6.4 Selecting the number of classes

We reuse the parameters (10) and (11) to analyze the behavior of the asymptotic ICL criterion. We consider different sizes of data matrices, between (30,30) and (150,150), with varying difficulty for each matrix size, with a conditional Bayes risk (9) of respectively 5%, 12% and 20%

The results in Figure 8 show that, as expected, the ICL criterion tends to select more often the right number of classes as the data matrices get larger and also when classes are more separated. We also observe that the ICL criterion tends to be conservative for small data matrices, by underestimating the number of classes. It could come to the fact that the size of the matrix is not large enough to consider the asymptotic approximation as valid and/or it could come from the approximations used to compute the log-likelihood $\mathcal{J}$ (variational restriction and delta method).

[1] With equal class proportions, the expected classification error of a random allocation is $\frac{K-1}{K} + \frac{L-1}{L} - \frac{K-1}{K}\frac{L-1}{L}$, that is, 0.89 here where $K = L = 3$.

			$r_{item}(\widehat{Y},\widehat{Z}) = 5\%$				$r_{item}(\widehat{Y},\widehat{Z}) = 12\%$				$r_{item}(\widehat{Y},\widehat{Z}) = 20\%$			
			L				L				L			
			2	3	4	5	2	3	4	5	2	3	4	5
$n_1 = n_2 = 30$	K	2	4	3	1		7	5			10	3		
		3	3	**9**			5	**1**	1		5	**1**		
		4					1					1		
		5												
$n_1 = n_2 = 40$	K	2	3	4			10	4	1		12	2	1	
		3		**12**				**5**			4	**1**		
		4		1										
		5												
$n_1 = n_2 = 50$	K	2		1			6	2			15	1	2	
		3	2	**16**				**11**			1	**0**		
		4		1			1				1			
		5												
$n_1 = n_2 = 75$	K	2												
		3	3				10	1			16			
		4		**16**				**8**				**4**		
		5	1				1							
$n_1 = n_2 = 100$	K	2												
		3					6				17			
		4		**20**				**14**				**2**	1	
		5												
$n_1 = n_2 = 150$	K	2												
		3		1			4	1			15		1	
		4		**18**				**15**				**4**		
		5		1										

Fig. 8 Number of (K,L) models selected by the asymptotic ICL criterion among 20 trials on data matrices of different sizes and difficulties, as measured by the conditional Bayes risk. All matrices are generated with the same number of row and column classes: $K = L = 3$.

For further experiments on the asymptotic behaviour of ICL of the Latent Block Model and comparisons with its exact expression, we refer the reader to Keribin et al. (2012) and Keribin et al. (2015).

6.5 Selecting the adequate missingness model

We use the models fitted in Section 6.3 to analyze the ability of the ICL criterion to select the right missingness model (MNAR or MAR). The difference in ICL between the MAR and MNAR models is computed for each data matrix, assuming that the right numbers of classes (K,L) are known.

The results, presented in Figure 9, show that ICL rightfully opts for the MNAR model almost everywhere, demonstrating the ability of this criterion to select the adequate missingness model. The MAR model is only chosen for some experiments with the lowest MNAR effect ($\sigma_B^2 = \sigma_D^2 = 0.01$), where the prediction performances are almost identical (see Figure 7).

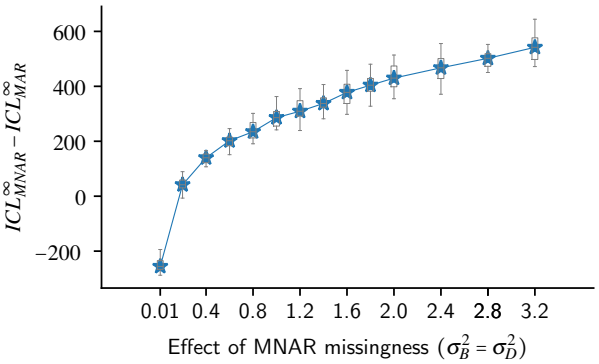
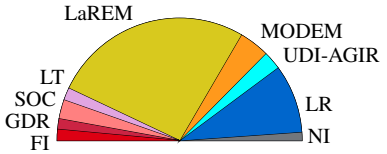


Fig. 9 Difference in ICL between the MAR and MNAR models with respect to an increase of the MNAR effect, where ⋆ is the median. The MNAR model is selected when the difference in ICL is positive.



FI (17): France Insoumise
GDR (16): Groupe de la Gauche démocrate et républicaine
SOC (29): Socialistes
LT (19): Libertés et territoires
LaREM (304): La République En Marche
MODEM (46): Mouvement démocrate
UDI-AGIR (28): Les Constructifs
LR (104): Les Républicains
NI (13): Non inscrits (mixed left and right wings)

Fig. 10 Hemicycle of the political groups of the French National Assembly

7 Experiments on real data

We consider voting records[2] from the lower house of the French Parliament (*Assemblée Nationale*). This dataset gathers the results of the 1256 ballots of year 2018 of the 576 French members of parliament (MPs) for the procedural motions and amendments for the 15th legislature (June 2017). For each ballot, the vote of each MP is recorded as a 4-level categorical response: "yes", 'no", "abstained" or "absent". Using our model, we bring out some relevant groups of ballots and MPs, as well as some structure in the behavior of nonvoters.

We gather the data in a matrix where each row represents an MP and each column represents a ballot. To use our model, we reduced the 4 response levels to 3 ("yes", "no", "missing") assuming that merging the "abstained" and "absent" categories would not affect much the underlying missingness process ("abstained" votes represent about 4% of the expressed votes, "missing" responses represent 85% of all votes).

At the lower house of French Parliament, MPs may group together according to their political affinities. Groups with fewer than 15 members or MPs who choose to be independent are gathered under the "Non inscrits" (NI) label, giving a heterogeneous range of political hues inside it. The names of the groups and their cardinalities are detailed in Figure 10.

The ICL criterion, used to select both the numbers of classes and the type of missingness, favors a MNAR missingness with $K = 14$ MP classes and $L = 14$ ballot classes (see Figure 13) against a MAR model with 19 MP classes 23 ballot classes. The reordered data matrix derived from this block clustering is displayed in Figure 11. Fewer classes lead to over-aggregated components hiding the subtleties of the network, but since they still correspond to well-identified groups and are more friendly to visual analysis, we provide them as additional material in Appendix E.

In Figure 11, classes of MPs are coherent to their political affiliation: class 0 and 1 are mainly made up of left-wing MPs from the groups SOC, FI, GDR, LT, classes 2 and 3 are mainly made up of right-wing MPs from LR and the classes from 6 to 13 are mainly made up of centrist MPs from LaREM and MODEM who are political allies. Classes of ballots can be analyzed with the available metadata. A bipartite opposition system appears from classes A and C. Ballots from class A refer to the original articles of law proposed by the government and are unsurprisingly voted positively by the MPs classes from 6 to 13 as they are from the same political mould as the French government. Ballots from class C mainly refer to amendments proposed by minority and are voted positively by both the left wing (class 0 and 1) and the right wing (classes 2 and 3) and negatively by the MPs supporting the government (classes 6 to 13). The left and right wings are yet divided by usual issues such as immigration regulation amendments gathered in classes G and M or general economic matters gathered in classes H and I.

In our model, the latent variables A and B characterize the propensity of MPs to cast a vote. Figure 12 displays the scatter plot of $v_i^{(A)}$ and $v_i^{(B)}$, the maximum *a posteriori* estimates of A_i and B_i for all MPs under the variational distribution. The abscissa represents the propensity to vote[3], with higher values of $v^{(A)}$ corresponding to a higher propensity to vote, and the ordinate $v^{(B)}$ represents the additional effect of casting a vote when approving the resolution. The membership of MPs to their political group is indicated by the plotting symbol.

We see two obvious clusters separated by the vertical axis $v^{(B)}$: the bottom cluster is essentially formed by MPs

[2] Votes from the French National Assembly are available from http://data.assemblee-nationale.fr/travaux-parlementaires/votes.

[3] More rigorously, the abscissa represents the *global deviation from the average* propensity to vote.

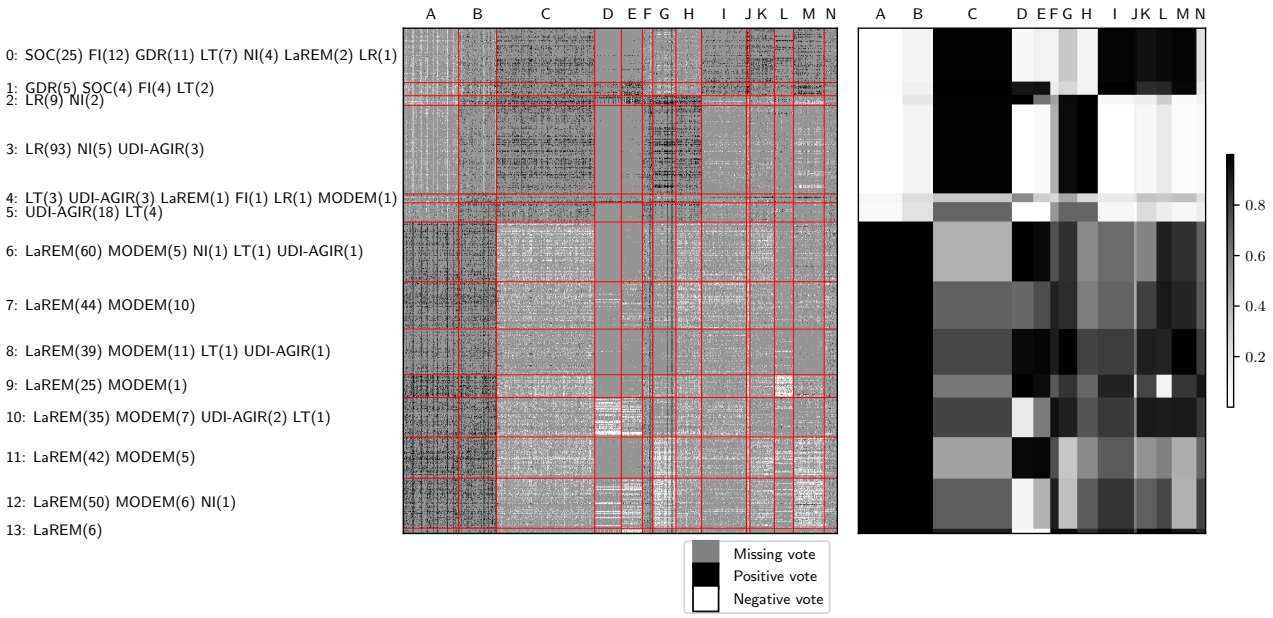


Fig. 11 Left: matrix of votes reordered according to the row and column classes, for the MNAR LBM model selected by ICL, with 14 MP classes and 14 ballot classes. The red lines delineate class boundaries. The breakdown of political groups in each cluster of MPs is given on the left. Right: summary of the inferred opinions (expressed or not) for all classes of ballots and MPs, as given by the estimated probability π_{kl} to approve a resolution in each block of the reordered matrix.

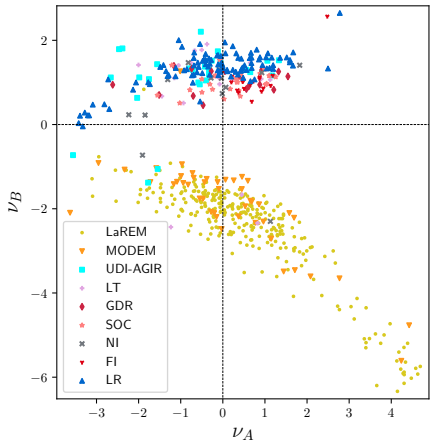


Fig. 12 Maximum *a posteriori* estimates of the MPs propensities $(\nu_i^{(A)}, \nu_i^{(B)})$, with their political group memberships. $\nu_i^{(A)}$ drives the MAR effect and $\nu_i^{(B)}$ drives the MNAR one.

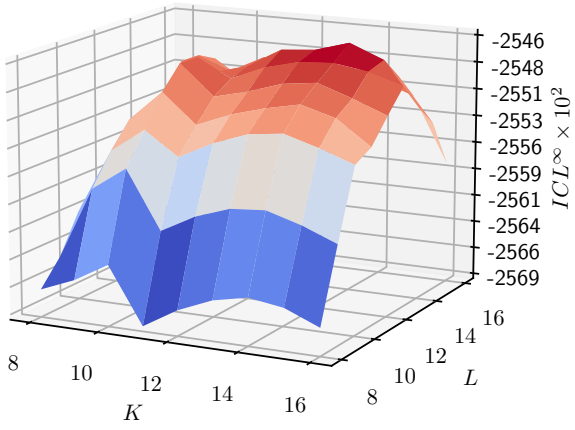


Fig. 13 Asymptotic ICL curve. Maximum is reached for K=14 and L=14

from the LaREM and MODEM political groups, which support the government, whereas the top cluster is formed by the opposition political groups. The $\nu^{(B)}$ estimates for the opposition cluster are positive, meaning that these MPs come to parliament to vote positively. This behavior is not surprising because the MPs of the opposition parties are outnumbered by the MPs supporting the government, so they must be diligent if they want their tabled motion or amendment passed. The dependency between the political groups and the MNAR effect encoded in the estimates $\nu^{(B)}$, which is confirmed by an ANOVA test (with a p-value smaller than numerical error), supports that the missingness patterns captured by our model are relevant for the problem at hand. A similar analysis is developed on ballots in Appendix E.

8 Conclusion

In many estimation problems, the absence of data conveys some information on the underlying phenomenon that should be exploited for its modeling. We propose a co-clustering model that accounts for this absence of data; it aims at retrieving groups of rows and columns based on the complete

data matrix instead of considering only the partitioning of the observed data matrix. This model consists of two building blocks: a co-clustering model (Latent Block Model) of the full data matrix, and a missingness model that explains the censoring that produces the observed data matrix. This missingness model preserves the symmetry of the co-clustering model by allowing two MNAR effects, one on the rows and the other on the columns. The overall model of the observed data matrix results from the combination of the model of the complete data matrix with the missingness model.

We used variational techniques and Taylor series to obtain a tractable approximation of the lower bound of the observed log-likelihood. We proposed a model selection criterion to select both the number of classes and the type of missingness (MAR versus MNAR).

Our experiments on synthetic datasets show that ignoring an informative missingness can lead to catastrophic co-clustering estimates, supporting the value of using expressive missingness models on such type of data. We also illustrate the use of our model on a real-world case where the missingness model provides an interesting basis for analyzing and interpreting the motivations of nonvoters. These experiments can be reproduced using the source code and the dataset available at `https://github.com/gfrisch/LBM-MNAR`.

Our model should also be useful in other fields such as in ecology, where the probability of observing interaction between species derives from some factors that also explain the true interactions (Vázquez et al. 2009), or in collaborative filtering, where the probability of observing a rating depends on the actual rating that would be given by the user (Marlin et al. 2007). In the latter application, the data sizes generally encountered in recommendation would require computational improvements in inference. Another useful future work is to extend our model to non-binary data.

Acknowledgements

We sincerely thank the anonymous reviewers whose critical and very thorough reading of the manuscript led to substantial improvements.

Appendix A Computing the criterion $\mathcal{J}(q_\gamma, \theta)$

The criterion to be optimized is :

$$\mathcal{J}(q_\gamma, \theta) = \mathcal{H}(q_\gamma) + \mathbb{E}_{q_\gamma}[\log p(X^{(o)}, Y, Z, A, B, C, D; \theta)] \ ,$$

where θ is the list of all model parameters: $\theta = (\alpha, \beta, \pi, \mu, \sigma_A^2, \sigma_B^2, \sigma_C^2, \sigma_D^2)$. We restrict the form of the variational distribution q_γ to get a fully factorized form:

$$\begin{aligned} q_\gamma = &\prod_i \mathcal{M}\left(1; \tau_i^{(Y)}\right) \times \prod_j \mathcal{M}\left(1; \tau_j^{(Z)}\right) \\ &\times \prod_i \mathcal{N}\left(\nu_i^{(A)}, \rho_i^{(A)}\right) \times \prod_i \mathcal{N}\left(\nu_i^{(B)}, \rho_i^{(B)}\right) \\ &\times \prod_j \mathcal{N}\left(\nu_j^{(C)}, \rho_j^{(C)}\right) \times \prod_j \mathcal{N}\left(\nu_j^{(D)}, \rho_j^{(D)}\right) \ , \end{aligned}$$

where γ denotes the list of parameters of the distribution: $\gamma = (\tau^{(Y)}, \tau^{(Z)}, \nu^{(A)}, \rho^{(A)}, \nu^{(B)}, \rho^{(B)}, \nu^{(C)}, \rho^{(C)}, \nu^{(D)}, \rho^{(D)})$.

The entropy is additive across independent variables, so we get:

$$\begin{aligned} \mathcal{H}(q_\gamma) = &- \sum_{ik} \tau_{ik}^{(Y)} \log \tau_{ik}^{(Y)} - \sum_{jl} \tau_{jl}^{(Z)} \log \tau_{jl}^{(Z)} \\ &+ (n_1 + n_2)(\log(2\pi) + 1) \\ &+ \frac{1}{2} \sum_i \left(\log \rho_i^{(A)} + \log \rho_i^{(B)}\right) + \frac{1}{2} \sum_j \left(\log \rho_j^{(C)} + \log \rho_j^{(D)}\right) \ . \end{aligned}$$

The independence of the latent variables allows one to rewrite the expectation of the complete log-likelihood as:

$$\begin{aligned} \mathbb{E}_{q_\gamma}\left[\log p\left(X^{(o)}, Y, Z, A, B, C, D\right)\right] &= \mathbb{E}_{q_\gamma}[\log p(Y)] \\ &+ \mathbb{E}_{q_\gamma}[\log p(Z)] + \mathbb{E}_{q_\gamma}[\log p(A)] + \mathbb{E}_{q_\gamma}[\log p(B)] \\ &+ \mathbb{E}_{q_\gamma}[\log p(C)] + \mathbb{E}_{q_\gamma}[\log p(D)] \\ &+ \mathbb{E}_{q_\gamma}\left[\log p\left(X^{(o)} \middle| Y, Z, A, B, C, D\right)\right] \ , \end{aligned}$$

with the following terms:

$$\mathbb{E}_{q_\gamma}[\log p(Y)] = \sum_{ik} \mathbb{E}_{q_\gamma} Y_{ik} \log \alpha_k = \sum_{ik} \tau_{ik}^{(Y)} \log \alpha_k$$

$$\mathbb{E}_{q_\gamma}[\log p(Z)] = \sum_{jl} \mathbb{E}_{q_\gamma} Z_{jl} \log \beta_l = \sum_{jl} \tau_{jl}^{(Z)} \log \beta_l$$

$$\begin{aligned} \mathbb{E}_{q_\gamma}[\log p(A)] &= -\frac{n_1}{2} \log 2\pi - \frac{n_1}{2} \log \sigma_A^2 - \frac{1}{2\sigma_A^2} \sum_i \mathbb{E}_{q_\gamma} A_i^2 \\ &= -\frac{n_1}{2} \log 2\pi - \frac{n_1}{2} \log \sigma_A^2 \\ &\quad - \frac{1}{2\sigma_A^2} \sum_i \left(\left(\nu_i^{(A)}\right)^2 + \rho_i^{(A)}\right) \end{aligned}$$

$$\begin{aligned} \mathbb{E}_{q_\gamma}[\log p(B)] &= -\frac{n_1}{2} \log 2\pi - \frac{n_1}{2} \log \sigma_B^2 \\ &\quad - \frac{1}{2\sigma_B^2} \sum_i \left(\left(\nu_i^{(B)}\right)^2 + \rho_i^{(B)}\right) \end{aligned}$$

$$\begin{aligned} \mathbb{E}_{q_\gamma}[\log p(C)] &= -\frac{n_2}{2} \log 2\pi - \frac{n_2}{2} \log \sigma_C^2 \\ &\quad - \frac{1}{2\sigma_C^2} \sum_j \left(\left(\nu_j^{(C)}\right)^2 + \rho_j^{(C)}\right) \end{aligned}$$

$$\begin{aligned} \mathbb{E}_{q_\gamma}[\log p(D)] &= -\frac{n_2}{2} \log 2\pi - \frac{n_2}{2} \log \sigma_D^2 \\ &\quad - \frac{1}{2\sigma_D^2} \sum_j \left(\left(\nu_j^{(D)}\right)^2 + \rho_j^{(D)}\right) \end{aligned}$$

$$\mathbb{E}_{q_\gamma}\Big[\log p\Big(X^{(\mathrm{o})}\Big|Y,Z,A,B,C,D\Big)\Big] = \sum_{kl,ij:X_{ij}^{(\mathrm{o})}=1} \tau_{ik}^{(Y)}\tau_{jl}^{(Z)}\mathbb{E}_{q_\gamma}[\log p_1]$$
$$+ \sum_{kl,ij:X_{ij}^{(\mathrm{o})}=0} \tau_{ik}^{(Y)}\tau_{jl}^{(Z)}\mathbb{E}_{q_\gamma}[\log p_0]$$
$$+ \sum_{kl,ij:X_{ij}^{(\mathrm{o})}=\mathrm{NA}} \tau_{ik}^{(Y)}\tau_{jl}^{(Z)}\mathbb{E}_{q_\gamma}[\log(1-p_0-p_1)] \ , \qquad (12)$$

with p_0 and p_1 defined in Equations (2)–(3).

Equation (12) involves the computation of the expectations of the following nonlinear functions:

$$f_1(x,y) = \log(\pi_{kl}\,\mathrm{expit}(\mu+x+y))$$
$$f_0(x,y) = \log((1-\pi_{kl})\,\mathrm{expit}(\mu+x-y))$$
$$f_{\mathrm{NA}}(x,y) = \log\Big(1-\pi_{kl}\,\mathrm{expit}(\mu+x+y) - (1-\pi_{kl})\,\mathrm{expit}(\mu+x-y)\Big) \ .$$

The approximation of these expectations given by the second-order Taylor series with independent random variables X and Y reads:

$$\mathbb{E}[f(X,Y)] \approx f(\mathbb{E}X,\mathbb{E}Y) + \frac{1}{2}\,var(X)\,\frac{\partial^2 f(\mathbb{E}[X],\mathbb{E}[Y])}{\partial(X)^2} + \frac{1}{2}\,var(Y)\,\frac{\partial^2 f(\mathbb{E}[X],\mathbb{E}[Y])}{\partial(Y)^2} \ ,$$

which yields in our case:

$$\mathbb{E}_{q_\gamma}\Big[f\Big(A_i+C_j,\,B_i+D_j\Big)\Big] \approx f\Big(\nu_i^{(A)}+\nu_j^{(C)},\,\nu_i^{(B)}+\nu_j^{(D)}\Big) + \frac{1}{2}\Big(\rho_i^{(A)}+\rho_j^{(C)}\Big)\,\frac{\partial^2 f\Big(\nu_i^{(A)}+\nu_j^{(C)},\,\nu_i^{(B)}+\nu_j^{(D)}\Big)}{\partial\Big(\nu_i^{(A)}+\nu_j^{(C)}\Big)^2} + \frac{1}{2}\Big(\rho_i^{(B)}+\rho_j^{(D)}\Big)\,\frac{\partial^2 f\Big(\nu_i^{(A)}+\nu_j^{(C)},\,\nu_i^{(B)}+\nu_j^{(D)}\Big)}{\partial\Big(\nu_i^{(B)}+\nu_j^{(D)}\Big)^2} \ .$$

The criterion is now fully computable.

Appendix B Initialization of the VEM algorithm with spectral clustering: Algorithm 2.

Algorithm 2: Initialization of the VEM algorithm with spectral clustering.

Input:
observed data $X^{(o)}$
K and L number of row groups and column groups
Function `SpectralClustering`(W *adjacency matrix,* k *number of clusters*)**:**
Define $D \in \mathbb{R}^{n\times n}$ the diagonal matrix with $D_{ii} = \sum_k W_{ik}$
Define $L = D^{-1/2} W D^{-1/2}$
Form the matrix $U = [U_1, ..., U_k] \in \mathbb{R}^{n\times k}$, where U_ℓ is the eigenvector of L with the ℓth largest eigenvalue in absolute value.
Return results of k-means with k clusters on U.
begin
Build Y the $n_1 \times K$ indicator matrix of the row cluster memberships with $SpectralClustering(XX^T, K)$
Build Z the $n_2 \times L$ indicator matrix of the column cluster memberships with $SpectralClustering(X^T X, L)$.
α, β and π are estimated from Y and Z
μ is initialized such as expit(μ) is the global missingness rate
σ_A^2, σ_B^2, σ_C^2 and σ_D^2 are sampled from $U_{]0,1]}$
end
Result:
$\theta = \left(\alpha, \beta, \pi, \mu, \sigma_A^2, \sigma_B^2, \sigma_C^2, \sigma_D^2\right)$ the model parameters
Y and Z the row and column cluster memberships

Appendix C Asymptotic form of the Integrated Completed Likelihood

C.1 ICL of the MNAR model

The asymptotic ICL criterion,

$$\begin{aligned} ICL^{\infty}(K,L) = &\max_{\theta, Y, Z, A, B, C, D} \log p\left(X^{(o)}, Y, Z, A, B, C, D; \theta\right) \\ &- \frac{K-1}{2}\log(n_1) - \frac{L-1}{2}\log(n_2) \\ &- \frac{KL}{2}\log(n_1 n_2) - \log(n_1 n_2) \ , \end{aligned}$$

for the LBM with the MNAR model of Section 3.2 is, up to an irrelevant constant, an asymptotic expansion of the log integrated completed likelihood

$$\log \int p(X, Y, Z, A, B, C, D | \theta; K, L) p(\theta; K, L) \mathrm{d}\theta \ .$$

Proof With independent latent variables and independent priors on the parameters, the log integrated completed likelihood reads

$$\begin{aligned} &\log \int p\left(X^{(o)}, Y, Z, A, B, C, D \middle| \theta\right) p(\theta) \mathrm{d}\theta \\ &= \log \int p\left(X^{(o)} \middle| Y, Z, A, B, C, D, \pi, \mu\right) p(\pi) p(\mu) \mathrm{d}\pi \mathrm{d}\mu \quad (13) \\ &+ \log \int p(Y|\alpha) p(\alpha) \mathrm{d}\alpha + \log \int p(Z|\beta) p(\beta) \mathrm{d}\beta \\ &+ \log \int p\left(A \middle| \sigma_A^2\right) p\left(\sigma_A^2\right) \mathrm{d}\sigma_A^2 + \log \int p\left(B \middle| \sigma_B^2\right) p\left(\sigma_B^2\right) \mathrm{d}\sigma_B^2 \\ &+ \log \int p\left(C \middle| \sigma_C^2\right) p\left(\sigma_C^2\right) \mathrm{d}\sigma_C^2 + \log \int p\left(D \middle| \sigma_D^2\right) p\left(\sigma_D^2\right) \mathrm{d}\sigma_D^2. \end{aligned}$$

As in the ICL developed by Keribin et al. (2012) for the standard LBM, we set non-informative Dirichlet distribution $\mathcal{D}(a, ..., a)$ priors on α and β:

$$\begin{aligned} \log p(Y) &= \log \int p(Y|\alpha) p(\alpha; a) \mathrm{d}\alpha \\ &= \log \int \prod_{ik} (\alpha_k)^{Y_{ik}} \frac{1}{\mathcal{B}(a)} \prod_{ik} (\alpha_k)^{a-1} \mathrm{d}\alpha \\ &= \log \mathcal{B}(a + \sum_i Y_i) - \log \mathcal{B}(a) \\ &= \sum_k \log\Gamma(Y_{:k} + a) + \log\Gamma(Ka) - \log\Gamma(n_1 + Ka) \\ &\quad - K \log\Gamma(a) \ , \end{aligned}$$

where $Y_{:k} = \sum_i Y_{ik}$. The Stirling expansion $\log\Gamma(x) = x\log x - x - \frac{1}{2}\log x + o(\log x)$ leads to the following asymptotic development of $\log p(Y)$:

$$\begin{aligned} \log p(Y) &= \sum_k \log\Gamma(Y_{:k} + a) - \log\Gamma(n_1 + Ka) + o(\log n_1) \\ &= \sum_k Y_{:k} \log Y_{:k} - n_1 - \frac{1}{2} n_1 \\ &\quad - \left(n_1 \log n_1 + Ka \log n_1 - n_1 - \frac{1}{2}\log n_1\right) + o(\log n_1). \end{aligned}$$

With the non-informative Jeffrey prior $a = \frac{1}{2}$, this gives:

$$\begin{aligned} \log p(Y) &= \sum_k Y_{:k} \log\left(\frac{1}{n_1} Y_{:k}\right) - \frac{K-1}{2}\log n_1 + o(\log n_1) \\ &= \max_{\alpha} \log p(Y; \alpha) - \frac{K-1}{2}\log n_1 + o(\log n_1). \quad (14) \end{aligned}$$

Similarly, we get:

$$\begin{aligned} \log p(Z) &= \sum_l \log\Gamma(Z_{:l} + a) + \log\Gamma(La) - \log\Gamma(n_2 + La) \\ &\quad - L \log\Gamma(a) \\ &= \max_{\beta} \log p(Z; \beta) - \frac{L-1}{2}\log n_2 + o(\log n_2) \ , \quad (15) \end{aligned}$$

where $Z_{:l} = \sum_j Z_{jl}$.

An InverseGamma(ξ, ξ) prior is set on σ_A^2:

$$
\begin{aligned}
p(A) &= \int p\left(A|\sigma_A^2\right) p\left(\sigma_A^2;\xi\right) \mathrm{d}\sigma_A^2 \\
&= \int \left(2\pi\sigma_A^2\right)^{-\frac{n_1}{2}} \exp\left(-\frac{\sum_i A_i^2}{2\sigma_A^2}\right) \\
&\qquad \frac{\xi^\xi}{\Gamma(\xi)} \exp\left(-\frac{\xi}{\sigma_A^2}\right) \left(\sigma_A^2\right)^{-\xi-1} \mathrm{d}\sigma_A^2 \\
&= \frac{\xi^\xi}{\Gamma(\xi)} (2\pi)^{\left(-\frac{n_1}{2}\right)} \\
&\quad \int \sigma_A^{2\left(-\frac{n_1}{2}-\xi-1\right)} \exp\left(-\frac{2\xi+\sum_i A_i^2}{2} \cdot \frac{1}{\sigma_A^2}\right) \mathrm{d}\sigma_A^2 \\
&= \frac{\xi^\xi}{\Gamma(\xi)} 2^\xi \pi^{-\frac{n_1}{2}} \left(2\xi + \sum_i A_i^2\right)^{\left(-\frac{n_1}{2}-\xi\right)} \Gamma\left(\frac{n_1}{2}+\xi\right) \ .
\end{aligned}
$$

Setting ξ to 1, one gets:

$$
p(A) = 2\,\pi^{-\frac{n_1}{2}} \left(2 + \sum_i A_i^2\right)^{-\frac{n_1}{2}-1} \Gamma\left(\frac{n_1}{2}+1\right) \ ,
$$

therefore,

$$
\begin{aligned}
\log p(A) &= \log 2 - \frac{n_1}{2}\log\pi + \log\Gamma\left(\frac{n_1}{2}+1\right) \\
&\quad - \left(\frac{n_1}{2}+1\right)\log\left(2+\sum_i A_i^2\right) \\
&= \log 2 - \frac{n_1}{2}\log\pi + \log\Gamma\left(\frac{n_1}{2}+1\right) - \left(\frac{n_1}{2}+1\right)\log n_1 \\
&\quad - \left(\frac{n_1}{2}+1\right)\log\left(\frac{2}{n_1} + \frac{1}{n_1}\sum_i A_i^2\right) \ .
\end{aligned}
$$

Using a Taylor expansion on the last term with $n_1 \to +\infty$ and using the fact that $\frac{1}{n_1}\sum_i A_i^2$ tends to a constant, we obtain:

$$
\begin{aligned}
\log p(A) &= \log 2 - \frac{n_1}{2}\log\pi + \log\Gamma\left(\frac{n_1}{2}+1\right) - \left(\frac{n_1}{2}+1\right)\log n_1 \\
&\quad - \left(\frac{n_1}{2}+1\right)\log\left(\frac{1}{n_1}\sum_i A_i^2\right)\left(1 + O\left(\frac{2}{n_1}\right)\right) \\
&= \log 2 - \frac{n_1}{2}\log\pi + \log\Gamma\left(\frac{n_1}{2}+1\right) - \left(\frac{n_1}{2}+1\right)\log n_1 \\
&\quad - \left(\frac{n_1}{2}+1\right)\log\left(\frac{1}{n_1}\sum_i A_i^2\right) + O(1) \\
&= \log 2 - \frac{n_1}{2}\log\pi + \log\Gamma\left(\frac{n_1}{2}+1\right) - \left(\frac{n_1}{2}+1\right)\log n_1 \\
&\quad - \frac{n_1}{2}\log\left(\frac{1}{n_1}\sum_i A_i^2\right) + O(1)
\end{aligned}
$$

Using the property of the gamma function $\Gamma(x+1) = x\Gamma(x)$ and applying the Stirling expansion of $\log\Gamma(x)$, we get for $n_1 \to +\infty$:

$$
\begin{aligned}
\log p(A) &= \log 2 - \frac{n_1}{2}\log\pi + \log\frac{n_1}{2} + \frac{n_1}{2}\log\frac{n_1}{2} - \frac{n_1}{2} \\
&\quad - \frac{1}{2}\log\frac{n_1}{2} - \left(\frac{n_1}{2}+1\right)\log n_1 \\
&\quad - \frac{n_1}{2}\log\left(\frac{1}{n_1}\sum_i A_i^2\right) + o(\log n_1) \\
&= -\frac{n_1}{2}\log(2\pi) - \frac{n_1}{2} - \frac{n_1}{2}\log\left(\frac{1}{n_1}\sum_i A_i^2\right) \\
&\quad - \frac{1}{2}\log n_1 + o(\log n_1) \\
&= \max_{\sigma_A^2} \log p\left(A;\sigma_A^2\right) - \frac{1}{2}\log n_1 + o(\log n_1) \ . \qquad (16)
\end{aligned}
$$

Similarly, with an identical prior on σ_B^2, σ_C^2 and σ_D^2 we get:

$$
\begin{aligned}
\log p(B) &= \max_{\sigma_B^2} \log p\left(B;\sigma_B^2\right) - \frac{1}{2}\log n_1 + o(\log n_1) \\
\log p(C) &= \max_{\sigma_C^2} \log p\left(C;\sigma_C^2\right) - \frac{1}{2}\log n_2 + o(\log n_2) \qquad (17) \\
\log p(D) &= \max_{\sigma_D^2} \log p\left(D;\sigma_D^2\right) - \frac{1}{2}\log n_2 + o(\log n_2) \ .
\end{aligned}
$$

Using a Laplace approximation as realized in the BIC, the penalty term differs from the categorical LBM (Keribin et al. 2015) as the levels of the distribution are linked (see Equations (2) and (3) from Section 3.2). We have:

$$
\begin{aligned}
&\log p\left(X^{(o)}|Y,Z,A,B,C,D\right) \qquad (18) \\
&\quad = \log \int p\left(X^{(o)}|Y,Z,A,B,C,D,\pi,\mu\right) p(\pi)p(\mu)\mathrm{d}\pi\mathrm{d}\mu \\
&\quad = \max_{\pi,\mu} \log p\left(X^{(o)}|Y,Z,A,B,C,D;\pi,\mu\right) \qquad (19) \\
&\quad + \frac{KL+1}{2}\log(n_1 n_2) + o(\log n_1) + o(\log n_2) \ ,
\end{aligned}
$$

as the number of free parameters in the conditional distribution of $X^{(o)}$ is $K \times L + 1$ which comes from the (K,L)-matrix of probabilities π and from μ, governing the global missingness rate. The ICL criterion (Proposition 1) is directly derived from Equations (13), (14), (15), (16), (17) and (19).

C.2 ICL of the LBM with MAR data

We consider the following LBM extended with the MAR missingness process:

Latent Block Model

$$Y_i \overset{\text{iid}}{\sim} \mathcal{M}(1;\alpha), \qquad \alpha \in \mathbf{S}_{K-1}$$
$$Z_j \overset{\text{iid}}{\sim} \mathcal{M}(1;\beta), \qquad \beta \in \mathbf{S}_{L-1}$$
$$\left(X_{ij} | Y_i = k, Z_j = l\right) \overset{\text{ind}}{\sim} \mathcal{B}(\pi_{kl}), \qquad \pi_{kl} \in [0,1]$$

MAR data model

$$A_i \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_A^2\right), \qquad \sigma_A^2 \in \mathbb{R}_+^*$$
$$C_j \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_C^2\right), \qquad \sigma_C^2 \in \mathbb{R}_+^*$$
$$\left(M_{ij} | A_i, C_j\right) \overset{\text{ind}}{\sim} \mathcal{B}\left(\text{expit}\left(\mu + A_i + C_j\right)\right)$$

Observations are generated according to:

$$X_{ij}^{(\text{o})} = \begin{cases} X_{ij} & \text{if} \quad M_{ij} = 1 \\ \text{NA} & \text{if} \quad M_{ij} = 0 \end{cases}$$

The asymptotic ICL of this model is:

$$ICL^{\infty}(K,L) = \max_{\theta,Y,Z,A,C} \log p\left(X^{(\text{o})}, Y, Z, A, C; \theta\right) \tag{20}$$
$$- \frac{KL+1}{2} \log\left(n_1 n_2\right)$$
$$- \frac{K-1}{2} \log\left(n_1\right) - \frac{L-1}{2} \log\left(n_2\right) \tag{21}$$
$$- \frac{1}{2} \log\left(n_1 n_2\right) \ .$$

Appendix D Supplemental figures for estimation

This section provides additional experimental results that suggest a consistent estimation of the model parameters. We reuse the data matrices generated by the LBM with missing data from Section 6.2. An initial data matrix of size $n_1 = n_2 = 500$ with a conditional Bayes risk of 5% was generated and progressively reduced, removing rows and columns, to increase the difficulty of the classification task.

Figure 14 displays the maximum absolute error made on the parameters π of the Bernoulli distributions that model the probability of X conditionally to the row and column classes. This error decreases as the size of the data matrices grows, which is consistent with our expectations.

Figure 15 displays the mean squared error (MSE) between the generated and estimated values of the latent variables A, B, C, D responsible for the individual variability of missingness. The estimated values are given by the maximum *a posteriori* of their corresponding variational distribution. The MSE curves of the variables A and C are comparable as well as the curves of the variables B and D. This is expected as the data matrices are generated with symmetric characters in rows and columns.

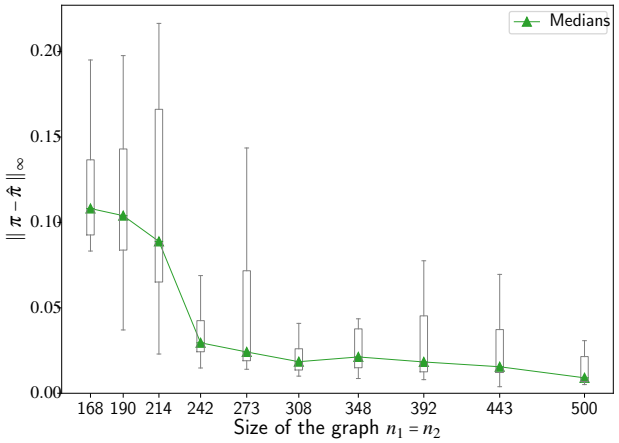


Fig. 14 Maximum error between the true (π) and the estimated ($\hat{\pi}$) probabilities associated to the blocks of the data matrix X as a function of its size.

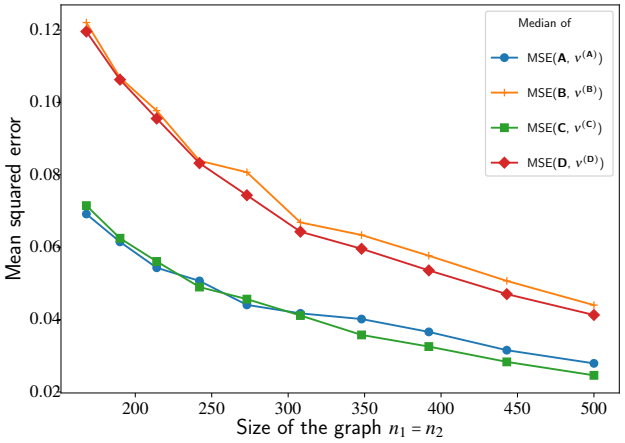


Fig. 15 Mean squared error of the maximum *a posteriori* estimates of the latent variables A, B, C, D governing the propensity of missingness.

Figure 16 compares the estimated values of A, B, C and D to their true generated values for two different sizes of data matrices, all other parameters being equal. A linear trend is exhibited from these scatter plots showing a good aptitude of the proposed inference to recover extreme negative and positive values.

Appendix E Supplemental figures for the French national assembly votes analysis

Figure 17 displays the reordered matrix of votes derived from a block clustering with a small number of classes. Such a simplification may be helpful for identifying global trends. With this model, the three MP classes are broadly identified as gathering the right-wing (first class) and left-wing (second class) opposition parties, the last class being formed of

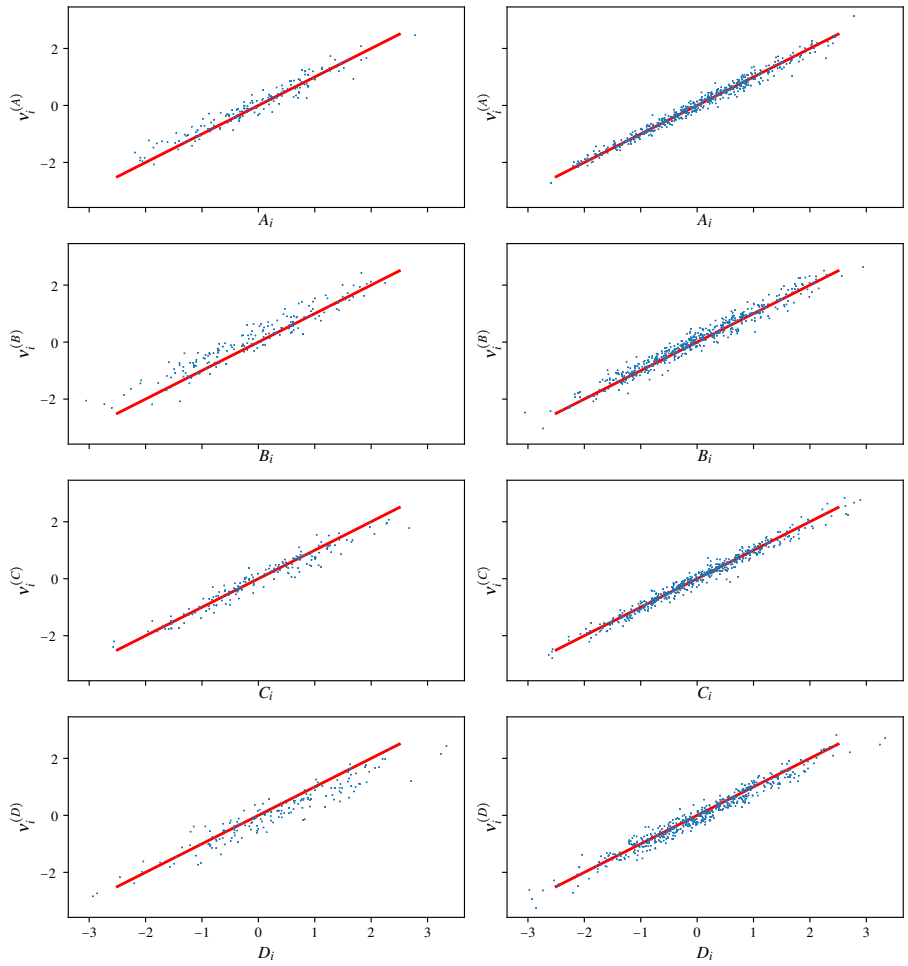


Fig. 16 Maximum *a posteriori* estimates of the latent variables governing the propensity of missingness versus their true generated values. Left: $n_1 = n_2 = 168$ and the conditional Bayes risk is 0.44; right: $n_1 = n_2 = 500$ and the conditional Bayes risk is 0.05. The identity line is drawn in red for reference.

the political groups supporting the government. The opposition systems appear clearly: on the ballots from classes A and E, the votes contrast the membership to the opposition parties versus the governmental alliance, whereas on the ballots from classes C and D, they separate the left-wing from the right-wing oppositions. Class B gathers various ballots on topics of rather general agreement pertaining to social or health matters.

Going back to the model selected by ICL described in Section 7, we analyze the resolution propensities to be voted upon and to be positively perceived by nonvoters. These propensities are encoded in the values of the latent variables C and D. Figure 18 displays the scatter plot of $v_j^{(C)}$ and $v_j^{(D)}$, the maximum *a posteriori* estimates of C_j and D_j under the variational distribution, for all ballots. The abscissa $v^{(C)}$ reflects the mobilization on the ballots, with higher mobilization for higher values, and the ordinate $v^{(D)}$ represents the additional effect of mobilizing specifically supporting voters. The fourteen-cluster membership of ballots (there is no obvious relevant classification for ballots) is indicated by the plotting symbol.

Some relationship between missingness and membership to ballot classes emerge from this plot. A first cluster of ballot appears in the positive quadrant, with propositions mainly proposed by the government, categorized in ballot classes A and B. A second cluster, smaller, on the upper left, is mainly formed by ballots categorized in class D, voted positively by few voters. All these propositions are related to the same law project regarding housing and were voted over a short period (06/03/2018 and 06/08/2018). The largest cluster, on the lower part of the graph, gathers most of the remaining ballots, that would have a tendency to be voted negatively by nonvoters. These propositions were proposed by either the right-wing or left-wing opposition, and get little support from a vast majority of MPs. Note also that the small group of highly voted propositions, on the right-hand side, is made of ballots belonging to six ballot classes.

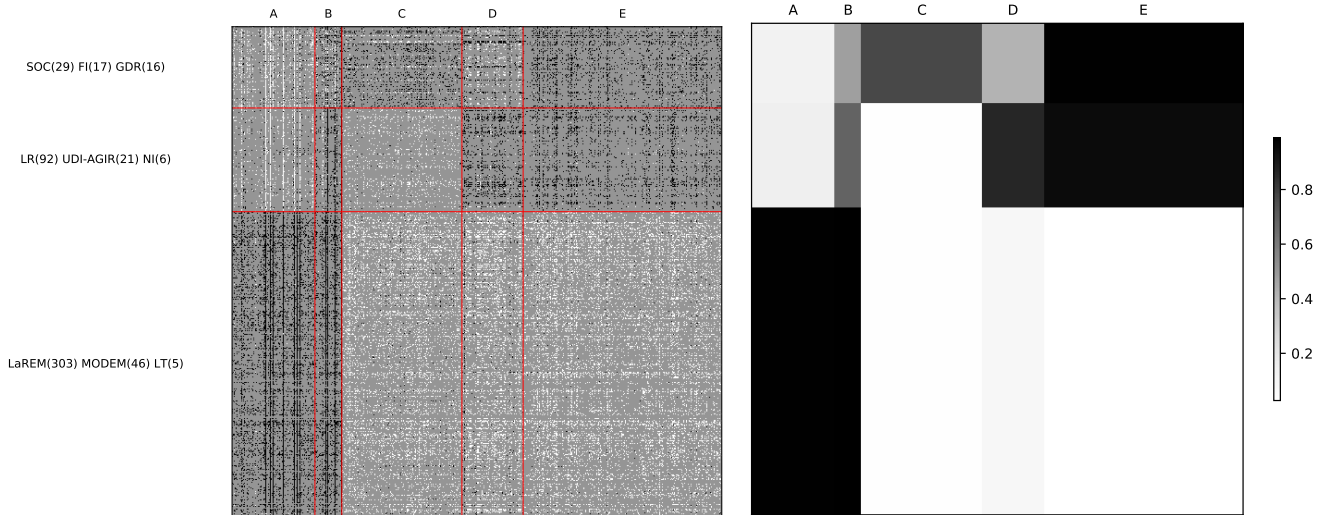


Fig. 17 Left: matrix of votes reordered according to the row and column classes, for the MNAR LBM with 3 MP classes and 5 ballot classes. The red lines delineate class boundaries. The breakdown of political groups in each cluster of MPs is given on the left. Right: summary of the inferred opinions (expressed or not) for all classes of ballots and MPs, as given by the estimated probability to approve a resolution in each block of the reordered matrix.

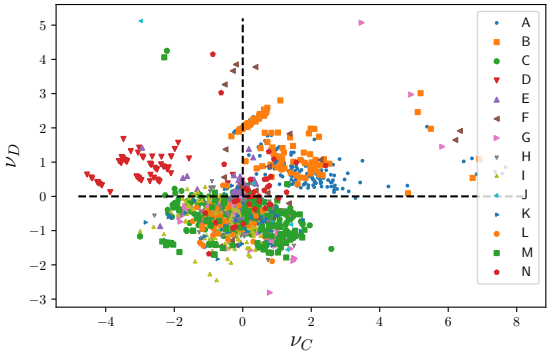


Fig. 18 maximum *a posteriori* estimates of the resolution propensities $(\nu_j^{(C)}, \nu_j^{(D)})$, with their clustering class memberships. $\nu_j^{(C)}$ drives the MAR effect and $\nu_j^{(D)}$ drives the MNAR one.

This reflects the fact that our model does not link the MNAR effect to the LBM memberships.

References

Baudry JP, Celeux G (2015) EM for mixtures. Statistics and Computing 25(4):713–726, DOI 10.1007/s11222-015-9561-x

Bhatia P, Iovleff S, Govaert G (2014) blockcluster: An R Package for Model Based Co-Clustering, URL https://hal.inria.fr/hal-01093554, working paper or preprint

Biernacki C, Celeux G, Govaert G (1998) Assessing a mixture model for clustering with the integrated classification likelihood. Tech. Rep. RR-3521, INRIA, URL https://hal.inria.fr/inria-00073163

Biernacki C, Celeux G, Govaert G (2003) Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Computational Statistics & Data Analysis 41:561–575, DOI 10.1016/S0167-9473(02)00163-9

Celeux G, Diebolt J (1985) The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Computational Statistics Quarterly 2:73–82

Corneli M, Bouveyron C, Latouche P (2020) Co-clustering of ordinal data via latent continuous random variables and not missing at random entries. Journal of Computational and Graphical Statistics DOI 10.1080/10618600.2020.1739533, URL https://hal.archives-ouvertes.fr/hal-01978174

Dhillon IS (2001) Co-clustering documents and words using bipartite spectral graph partitioning. In: Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 269–274

Dhillon IS, Mallela S, Modha DS (2003) Information-theoretic co-clustering. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, KDD '03, pp 89—-98, DOI 10.1145/956750.956764, URL https://doi.org/10.1145/956750.956764

Ding C, Li T, Peng W, Park H (2006) Orthogonal nonnegative matrix t-factorizations for clustering. In: Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol 2006, pp 126–135, DOI 10.1145/1150402.1150420

George T, Merugu S (2005) A scalable collaborative filtering framework based on co-clustering. In: Fifth IEEE International Conference on Data Mining (ICDM), DOI 10.1109/ICDM.2005.14

Govaert G, Nadif M (2008) Block clustering with Bernoulli mixture models: Comparison of different approaches. Computational Statistics & Data Analysis 52(6):3233–3245

Hernández-Lobato JM, Houlsby N, Ghahramani Z (2014) Probabilistic matrix factorization with non-random missing data. In: Proceedings of the 31st International Conference on Machine Learning, pp 1512–1520

Jaakkola TS (2000) Tutorial on variational approximation methods. In: Opper M, Saad D (eds) Advanced Mean Field Methods: Theory and Practice, MIT Press, pp 129–159

Jacques J, Biernacki C (2018) Model-Based Co-clustering for Ordinal Data. Computational Statistics & Data Analysis 123:101–115,

DOI 10.1016/j.csda.2018.01.014, URL https://hal.inria.fr/hal-01448299

Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK (1999) An introduction to variational methods for graphical models. Machine Learning 37(2):183—-233, DOI 10.1023/A:1007665907178, URL https://doi.org/10.1023/A:1007665907178

Keribin C, Brault V, Celeux G, Govaert G (2012) Model selection for the binary latent block model. In: Proceedings of COMPSTAT

Keribin C, Brault V, Celeux G, Govaert G (2015) Estimation and selection for the latent block model on categorical data. Statistics and Computing 25(6):1201–1216

Kim YD, Choi S (2014) Bayesian binomial mixture model for collaborative prediction with non-random missing data. In: Eighth ACM Conference on Recommender Systems (RecSys), p 201–208

Kluger Y, Basri R, Chang J, Gerstein M (2003) Spectral biclustering of microarray data: Coclustering genes and conditions. Genome Research 13:703–716, DOI 10.1101/gr.648603

Labiod L, Nadif M (2011) Co-clustering for binary and categorical data with maximum modularity. In: 11th IEEE International Conference on Data Mining (ICDM), pp 1140–1145, DOI 10.1109/ICDM.2011.37

Latouche P, Birmelé E, Ambroise C (2011) Overlapping stochastic block models with application to the French political blogosphere. The Annals of Applied Statistics 5(1):309–336, DOI 10.1214/10-aoas382, URL http://dx.doi.org/10.1214/10-AOAS382

Little RJA, Rubin DB (1986) Introduction. In: Statistical Analysis with Missing Data, John Wiley & Sons, chap 1, pp 1–23

Lomet A (2012) Sélection de modèle pour la classification croisée de données continues. PhD thesis, Université de technologie de Compiègne, URL http://www.theses.fr/2012COMP2041, thèse de doctorat dirigée par Govaert, Gérard et Grandvalet, Yves Technologies de l'information et des systèmes Compiègne 2012

Lomet A, Govaert G, Grandvalet Y (2012) Design of artificial data tables for co-clustering analysis. Tech. rep., Université de technologie de Compiègne, France

Marlin BM, Zemel RS, Roweis ST, Slaney M (2007) Collaborative filtering and the missing at random assumption. In: Twenty-Third Conference on Uncertainty in Artificial Intelligence (UAI), pp 267–275

Marlin BM, Zemel RS, Roweis ST, Slaney M (2011) Recommender systems, missing data and statistical model estimation. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI), pp 2686–2691

Nadif M, Govaert G (2010) Latent block model for contingency table. Communications in Statistics—Theory and Methods 39(3):416–425, DOI 10.1080/03610920903140197

Papalexakis EE, Sidiropoulos N, Bro R (2013) From k-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors. IEEE Transactions on Signal Processing 61(2):493–506, DOI 10.1109/TSP.2012.2225052

Parisi G (1988) Statistical field theory. Frontiers in Physics, Addison-Wesley, URL https://cds.cern.ch/record/111935

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Kopf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) Pytorch: An imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems 32, Curran Associates, Inc., pp 8024–8035

Pontes B, Giráldez R, Aguilar-Ruiz JS (2015) Biclustering on expression data: A review. Journal of Biomedical Informatics 57:163–180, DOI https://doi.org/10.1016/j.jbi.2015.06.028, URL http://www.sciencedirect.com/science/article/pii/S1532046415001380

Rohe K, Chatterjee S, Yu B (2011) Spectral clustering and the high-dimensional stochastic blockmodel. The Annals of Statistics 39(4):1878–1915, DOI 10.1214/11-aos887, URL http://dx.doi.org/10.1214/11-AOS887

Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592, URL http://www.jstor.org/stable/2335739

Schnabel T, Swaminathan A, Singh A, Chandak N, Joachims T (2016) Recommendations as treatments: Debiasing learning and evaluation. In: Proceedings of the 33nd International Conference on Machine Learning (ICML), pp 1670–1679, URL http://proceedings.mlr.press/v48/schnabel16.html

Selosse M, Jacques J, Biernacki C (2020a) Model-based co-clustering for mixed type data. Computational Statistics & Data Analysis 144:106866, DOI https://doi.org/10.1016/j.csda.2019.106866, URL http://www.sciencedirect.com/science/article/pii/S016794731930221X

Selosse M, Jacques J, Biernacki C (2020b) Textual data summarization using the self-organized co-clustering model. Pattern Recognition 103:107315, DOI https://doi.org/10.1016/j.patcog.2020.107315, URL http://www.sciencedirect.com/science/article/pii/S0031320320301199

Shan H, Banerjee A (2008) Bayesian co-clustering. In: 2008 Eighth IEEE International Conference on Data Mining, pp 530–539

Shireman E, Steinley D, Brusco M (2015) Examining the effect of initialization strategies on the performance of Gaussian mixture modeling. Behavior Research Methods 49, DOI 10.3758/s13428-015-0697-6

Steck H (2010) Training and testing of recommender systems on data missing not at random. In: 16th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pp 713–722

Tabouy T, Barbillon P, Chiquet J (2020) Variational inference for stochastic block models from sampled data. Journal of the American Statistical Association 115(529):455–466

Vázquez DP, Blüthgen N, Cagnolo L, Chacoff NP (2009) Uniting pattern and process in plant–animal mutualistic networks: a review. Annals of botany 103(9):1445–1457

Wang W (2013) Identifiability of linear mixed effects models. Electronic Journal of Statistics 7:244–263

Wasserman L (2004) All of Statistics: A Concise Course in Statistical Inference. Springer

Wyse J, Friel N (2012) Block clustering with collapsed latent block models. Statistics and Computing 22(2):415–428