Learning from missing data with the Latent Block Model

Gabriel Frisch, Jean-Benoist Leger, Yves Grandvalet

Learning from missing data with the Latent Block Model

Gabriel Frisch[1], Jean-Benoist Leger[1], and Yves Grandvalet[1]

[1] *Université de technologie de Compiègne, CNRS, Heudiasyc (Heuristics and Diagnosis of Complex Systems), CS 60 319 - 60 203 Compiègne Cedex*

2020, October

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 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.

1 Introduction

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

In Section 2 we review 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.

1.1 Related Works

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). In addition to that, our missing data model can be easily re-used for any other statistical co-clustering model as it is 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 missingness model of

type MNAR. 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. The MNAR missingness is modeled by 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 considered in the Matrix Factorization framework. Steck [2010] derives a weighted MF model and optimizes the parameters based on a metric that is robust to MNAR data. Hernández-Lobato et al. [2014] use a double probabilistic MF 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.

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 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 $\boldsymbol{X}$ 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_1 row clusters and k_2 column clusters, the reordering reveals $k_1 \times k_2$ 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 $\boldsymbol{X}$ can also be interpreted as the biadjacency matrix of a bipartite graph, whose two sets of vertices corresponds 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 $\boldsymbol{X}$, two partitions are defined by the latent variables $\boldsymbol{Y}$ and $\boldsymbol{Z}$, with $\boldsymbol{Y}$ being the $n_1 \times k_1$ indicator matrix of the latent row

clusters ($Y_{iq} = 1$ if row i belongs to group q and $Y_{iq} = 0$ otherwise), and $\boldsymbol{Z}$ being the $n_2 \times k_2$ indicator matrix of the latent column cluster. The group indicator of row i will be denoted $\boldsymbol{Y}_i$, and similarly, the group indicator of column j will be denoted $\boldsymbol{Z}_j$. The LBM makes several assumptions on the dependencies:

Independent rows and column clusters The latent variables $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are *a priori* independent.

$$p(\boldsymbol{Y}, \boldsymbol{Z}) = p(\boldsymbol{Y})p(\boldsymbol{Z}) \ .$$

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

Independent and identically distributed row clusters The latent variables $\boldsymbol{Y}$ are independent and follow a multinomial distribution $\mathcal{M}(1; \boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{k_1})$ is the mixing proportions of rows:

$$p(\boldsymbol{Y}; \boldsymbol{\alpha}) = \prod_i p(\boldsymbol{Y}_i; \boldsymbol{\alpha})$$
$$p(Y_{iq} = 1; \boldsymbol{\alpha}) = \alpha_q \ ,$$

with $\boldsymbol{\alpha} \in S_{(k_1 - 1)} = \{\boldsymbol{\alpha} \in \mathbb{R}_+^{k_1} | \sum_q \alpha_q = 1\}$.

Independent and identically distributed column clusters Likewise, the latent variables $\boldsymbol{Z}$ are independent and follow a multinomial distribution $\mathcal{M}(1; \boldsymbol{\beta})$, where $\boldsymbol{\beta} = (\beta_1, ..., \beta_{k_2})$ is the mixing proportions of columns:

$$p(\boldsymbol{Z}; \boldsymbol{\beta}) = \prod_j p(\boldsymbol{Z}_j; \boldsymbol{\beta})$$
$$p(Z_{jl} = 1; \boldsymbol{\beta}) = \beta_l \ ,$$

with $\boldsymbol{\beta} \in S_{(k_2 - 1)}$.

Given row and column clusters, independent and identically distributed block entries Given the row and colum clusters $(\boldsymbol{Y}, \boldsymbol{Z})$, the entries X_{ij} are independent and follow a Bernoulli distribution of parameter $\boldsymbol{\pi} = (\pi_{ql}; q = 1, ..., k_1; l = 1, ..., k_2)$: all elements of a block follow the same probability distribution.

$$p(\boldsymbol{X} | \boldsymbol{Y}, \boldsymbol{Z}; \boldsymbol{\pi}) = \prod_{ij} p(X_{ij} | \boldsymbol{Y}_i, \boldsymbol{Z}_j; \boldsymbol{\pi})$$
$$p(X_{ij} = 1 | Y_{iq} Z_{jl} = 1; \boldsymbol{\pi}) = \pi_{ql} \ .$$

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

$$p(\boldsymbol{X}; \theta) = \sum_{(\boldsymbol{Y}, \boldsymbol{Z}) \in I \times J} \left(\prod_{i,q} {\alpha_q}^{Y_{iq}} \right) \left(\prod_{j,l} {\beta_l}^{Z_{jl}} \right) \left(\prod_{i,j,q,l} \phi(X_{ij}; \pi_{ql})^{Y_{iq} Z_{jl}} \right) \ ,$$

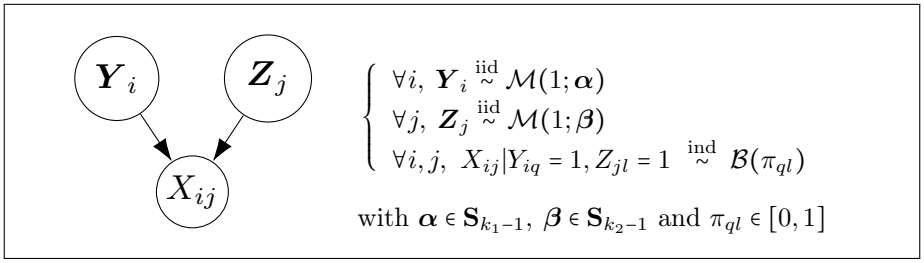


Figure 1: Summary of the standard Latent Block Model with binary data.

where $\phi(X_{ij};\pi_{ql}) = \pi_{ql}^{X_{ij}}(1-\pi_{ql})^{1-X_{ij}}$ 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_1 (resp. k_2) groups.

3 Extension to Informative Missing Data

The standard Latent Block Model does not accommodate missing observations, that is, the data matrix $\boldsymbol{X}$ 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, $\boldsymbol{X}^{(\text{o})}$ will denote the "partially observed" data matrix, with missing entries, whereas $\boldsymbol{X}$ denotes the "full" (unobserved) data matrix, without missing entries. The partially observed matrix $\boldsymbol{X}^{(\text{o})}$ is identical to the full matrix $\boldsymbol{X}$ except for the missing entries; $\boldsymbol{X}^{(\text{o})}$ takes its values in $\{0,1,\text{NA}\}$, where NA denotes a missing value. It will be convenient to introduce a binary mask matrix $\boldsymbol{M}$ that indicates the *non-missing* entries of $\boldsymbol{X}^{(\text{o})}$: if $M_{ij}=0$, then $X_{ij}^{(\text{o})} = \text{NA}$.

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 missing, we will model a real variable that defines the log-odds of this probability. This log-odds will be called here the "propensity" to be missing.

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 $\boldsymbol{X}^{(\text{o})}$ 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.

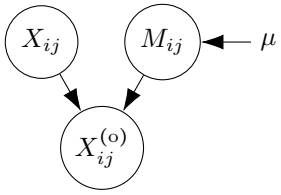


Figure 2: Graphical representation of the MCAR model. The partially observed entry $X_{ij}^{(o)}$ is generated by the corresponding entries of (i) the full matrix X_{ij} and (ii) the binary mask M_{ij}. The binary mask $\boldsymbol{M}$ does not depend on $\boldsymbol{X}$ and is defined here from a single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

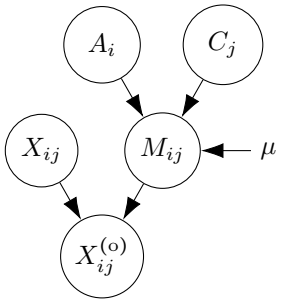


Figure 3: Graphical representation of the MAR model. The partially observed entry $X_{ij}^{(o)}$ is generated by the corresponding entries of (i) the full matrix X_{ij} and (ii) the binary mask M_{ij}. The binary mask $\boldsymbol{M}$ does not depend on $\boldsymbol{X}$ and is and is defined by a global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

Missing At Random (MAR) Missingness depends on the observed data, but not on the unobserved data. 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 $\boldsymbol{A}$, and another one for every column, denoted $\boldsymbol{C}$. 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, $\boldsymbol{B}$ and $\boldsymbol{D}$, adjust the propensity of missingness according to the actual value of X_{ij}. The graphical representation of this model is shown in Figure 4.

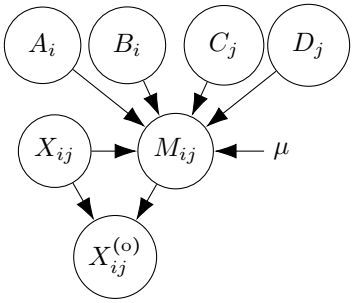


Figure 4: Graphical representation of the MNAR model. The partially observed entry $X_{ij}^{(o)}$ is generated by the corresponding entries of (i) the full matrix X_{ij} and (ii) the binary mask M_{ij}. The binary mask $\boldsymbol{M}$ depends on $\boldsymbol{X}$ and is defined by a global effect parameter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

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

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

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

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

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

$$\forall i, j \quad M_{ij} | A_i, B_i, C_j, D_j, X_{ij} \overset{\text{ind}}{\sim} \mathcal{B}(\text{expit}\,(P_{ij})) , \tag{2}$$

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.

This model 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 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 $\boldsymbol{X}$ and the mask matrix $\boldsymbol{M}$, all the elements of the observed matrix $\boldsymbol{X}^{(\mathrm{o})}$ are independent and identically distributed:

$$\left(X_{ij}^{(\mathrm{o})} \middle| X_{ij}, M_{ij}\right) = \begin{cases} X_{ij} & \text{if} \quad M_{ij} = 1 \\ \mathrm{NA} & \text{if} \quad M_{ij} = 0 \end{cases} . \tag{3}$$

Figure 5 summarizes the LBM extented to MNAR data.

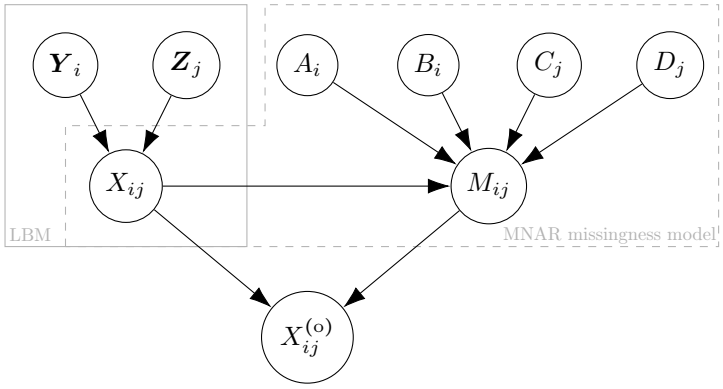
$\boldsymbol{X}^{(\mathrm{o})}$ taking its values in $(0, 1, \mathrm{NA})$, the same model can be rewritten with a Categorial distribution using directly the latent variables of both models:

$$\begin{aligned} \forall i,\ \boldsymbol{Y}_i &\overset{\text{iid}}{\sim} \mathcal{M}(1; \boldsymbol{\alpha}) \\ \forall j,\ \boldsymbol{Z}_j &\overset{\text{iid}}{\sim} \mathcal{M}(1; \boldsymbol{\beta}) \\ \forall i,\ A_i &\overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_A^2\right) \\ \forall i,\ B_i &\overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_B^2\right) \\ \forall j,\ C_j &\overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_C^2\right) \\ \forall j,\ D_j &\overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_D^2\right) \\ \forall i, j,\ X_{ij}^{(\mathrm{o})} \middle| Y_{iq} = 1, Z_{jl} = 1, A_i, B_i, C_j, D_j &\overset{\text{ind}}{\sim} \mathrm{Cat}\left(\begin{bmatrix} 0 \\ 1 \\ \mathrm{NA} \end{bmatrix}, \begin{bmatrix} p_0 \\ p_1 \\ 1 - p_0 - p_1 \end{bmatrix}\right) \end{aligned} \tag{4}$$

with

$$p_0 = (1 - \pi_{ql}) \operatorname{expit}(\mu + A_i - B_i + C_j - D_j) \tag{5}$$

$$p_1 = \pi_{ql} \operatorname{expit}(\mu + A_i + B_i + C_j + D_j) . \tag{6}$$



Latent Block Model

$$\boldsymbol{Y}_i \overset{\text{iid}}{\sim} \mathcal{M}(1;\boldsymbol{\alpha}), \qquad \boldsymbol{\alpha} \in \mathbf{S}_{k_1-1}$$

$$\boldsymbol{Z}_j \overset{\text{iid}}{\sim} \mathcal{M}(1;\boldsymbol{\beta}), \qquad \boldsymbol{\beta} \in \mathbf{S}_{k_2-1}$$

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

MNAR model

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

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

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

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

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

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

Observations are generated according to:

$$\left(X_{ij}^{(\mathrm{o})}\middle|X_{ij},M_{ij}\right) = \begin{cases} X_{ij} & \text{if} \quad M_{ij}=1 \\ \text{NA} & \text{if} \quad M_{ij}=0 \end{cases}$$

Figure 5: Graphical view and summary of the Latent Block Model extended to MNAR missingness process. The observed data $X_{ij}^{(\mathrm{o})}$ 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 $\boldsymbol{X}$ and the mask matrix $\boldsymbol{M}$ 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 (Equation 4). We can split our random variables into two sets: the set of unobserved latent variables and the set of observed variables consisting of $\boldsymbol{X}^{(\mathrm{o})}$ only. An observation of $\boldsymbol{X}^{(\mathrm{o})}$ only is called the incomplete data, and an observation of $\boldsymbol{X}^{(\mathrm{o})}$ together with the latent variables $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$ is called the complete data. Given the incomplete data, our objective is to infer the model parameters θ via maximum likelihood $\widehat{\theta} = \arg\max_\theta p(\boldsymbol{X}^{(\mathrm{o})};\theta)$. We resort to the Expectation Maximization (EM) algorithm to maximize $p(\boldsymbol{X}^{(\mathrm{o})};\theta)$ without explicitely 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 $\boldsymbol{X}^{(\mathrm{o})}$:

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

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 $p(\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D}|\boldsymbol{X}^{(\mathrm{o})})$ 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 $\boldsymbol{Y}$, $\boldsymbol{Z}$, $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$ and optimize the following lower bound on the log-likelihood of the incomplete data:

$$\mathcal{J}(q,\theta) = \log p\Big(\boldsymbol{X}^{(\mathrm{o})};\theta\Big) - KL\Big(q(\cdot) \parallel p(\cdot|\boldsymbol{X}^{(\mathrm{o})};\theta)\Big) \ , \tag{7}$$

where KL stands for the Kullback-Leibler divergence and $q(\cdot)$ denotes the variational distribution over the latent variables $\boldsymbol{Y}$, $\boldsymbol{Z}$, $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$. 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 $q(\cdot) = p(\cdot|\boldsymbol{X}^{(\mathrm{o})};\theta)$. Thus, maximizing the criterion $\mathcal{J}$ is equivalent to minimizing the discrepancy between $q(\cdot)$ and $p(\cdot|\boldsymbol{X}^{(\mathrm{o})};\theta)$, as measured by the Kullback divergence, and is also equivalent to maximizing the likelihood. The minimization of this Kullback divergence requires 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:

$$\mathcal{J}(q,\theta) = \mathcal{H}(q) + \mathbb{E}_q\Big[\log p\Big(\boldsymbol{X}^{(\mathrm{o})},\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D};\theta\Big)\Big] \ , \tag{8}$$

where $\mathcal{H}(q)$ is the entropy of the variational distribution and $\mathbb{E}_q$ is the expectation with respect to the variational distribution. The criteria $\mathcal{J}$ can become tractable if an exploration of a subspace, noted q_γ, of the latent distributions is made. However, this solution comes with the cost that the maximum found, will be a lower bound of the initial criteria:

$$\mathcal{J}(q,\theta) \geq \mathcal{J}(q_\gamma,\theta) \tag{9}$$

$\mathcal{J}(q_\gamma,\theta)$ is also 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 a feasible computation of the criterion. We choose to consider the following posterior shapes on the latent variables:

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

We also impose the conditional independence of the latent variables to get a feasible computation of the entropy and of the negative "energy" (Equation 8) under q_γ. This conditional independence is widely known as the "mean field

Algorithm 1: Variational Expectation Maximization algorithm.

Input: observed data $\boldsymbol{X}^{(o)}$, k_1 and k_2 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 $\mathcal{J}\left(\gamma^{(t)}, \theta\right)$:

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

end

Result: θ and γ: model and variational parameters

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

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

$$\mathcal{J}(\gamma, \theta) = \mathcal{H}(q_\gamma) + \mathbb{E}_{q_\gamma}\left[\log p\left(\boldsymbol{X}^{(o)}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}; \theta\right)\right] \tag{10}$$

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

$$\widehat{\theta} = \arg\max_{\theta} \left(\max_{\gamma} \mathcal{J}(\gamma, \theta)\right) \ . \tag{11}$$

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 additive across independent variables:

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

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

$$\begin{aligned}\mathbb{E}_{q_\gamma}\Big[\log p\Big(\boldsymbol{X}^{(\mathrm{o})}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}\Big)\Big] = &\ \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{Y})] \\ &+ \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{Z})] + \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{A})] + \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{B})] \\ + \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{C})] + \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{D})] + \mathbb{E}_{q_\gamma}&\Big[\log p\Big(\boldsymbol{X}^{(\mathrm{o})} \Big| \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}\Big)\Big] \ . \quad (12)\end{aligned}$$

Despite the variational approximation, the expectation of the complete log-likelihood (12) can not be exactly computed as its last term involves an expectation under q_γ of nonlinear functions:

$$\begin{aligned}\mathbb{E}_{q_\gamma}\Big[\log p\Big(\boldsymbol{X}^{(\mathrm{o})} \Big| \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}\Big)\Big] = &\sum_{ijql: X_{ij}^{(\mathrm{o})} = 0} \tau_{iq}^{(Y)} \tau_{jl}^{(Z)} \mathbb{E}_{q_\gamma}[\log(p_0)] \\ + \sum_{ijql: X_{ij}^{(\mathrm{o})} = 1} \tau_{iq}^{(Y)} \tau_{jl}^{(Z)} \mathbb{E}_{q_\gamma}[\log(p_1)] &+ \sum_{ijql: X_{ij}^{(\mathrm{o})} = \mathrm{NA}} \tau_{iq}^{(Y)} \tau_{jl}^{(Z)} \mathbb{E}_{q_\gamma}[\log(1 - p_0 - p_1)] \ , \quad (13)\end{aligned}$$

with p_0 and p_1 defined in Equations (5)–(6).

These expectations can be approximated by the delta method [Wasserman, 2004, p. 79]. 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 between a maximization with respect to the variational parameters γ and a maximization w.r.t 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 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 into 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 pursue 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].

For the Stochastic Block Model, a close relative of the Latent Block Model for graphs, Rohe et al. [2011] prove the consistency of spectral clustering to identify the parameters of the Stochastic Block Model. Following this idea, we use a double spectral clustering (with absolute eigenvalues of the Laplacian as Rohe et al. [2011]) on rows and columns on similarity matrices, 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 can not 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_1 row classes and k_2 column classes, the criterion reads

$$\begin{aligned} ICL(k_1, k_2) &= \log p(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \\ &= \log \int p(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}|\theta; k_1, k_2) p(\theta; k_1, k_2) \mathrm{d}\theta \ , \end{aligned} \tag{14}$$

with $p(\theta; k_1, k_2)$ the prior distribution of parameters. By taking into account the latent variables $\boldsymbol{Y}, \boldsymbol{Z}$, ICL is a clustering-oriented criterion , whereas BIC or AIC are driven by the faithfulness to the distribution of $\boldsymbol{X}$ [Biernacki et al., 1998].

For the LBM with MNAR missingness, ICL requires priors on the parameters of the missingness model. We chose independent and non informative InverseGamma(β, β) distribution (where β tends to zero) 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 parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ of mixing proportions of classes. ICL reads

$$\begin{aligned} ICL(k_1, k_2) &= \log p(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}) \\ &= \log \int p(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D} | \theta; k_1, k_2) p(\theta; k_1, k_2) \mathrm{d}\theta \end{aligned} \tag{15}$$

Proposition 1. *The ICL criterion for the LBM extended to the MNAR missingness process presented in Section 3.2 has the following asymptotic form for a data matrix of size $n_1 \times n_2$:*

$$\begin{aligned} ICL(k_1, k_2) = \max_{\theta} \; & \log p\Big(\boldsymbol{X}^{(o)}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}; \theta\Big) \\ & - \frac{k_1 k_2}{2} \log(n_1 n_2) - \frac{k_1 - 1}{2} \log(n_1) - \frac{k_2 - 1}{2} \log(n_2) \\ & + n_1 \log(2\pi) - \log(n_1) + n_2 \log(2\pi) - \log(n_2) \\ & + o(\log n_1) + o(\log n_2) \quad . \end{aligned}$$

See proof in Appendix C.

Since the maximum of the complete log-likelihood required to calculate the ICL is not available, in practice it is replaced by the lower bound provided by the variational approximation (see equation 10). An ICL criterion for the LBM with MAR missing data can be constructed in the same way, allowing for comparison with the MNAR model (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 Assessing the difficulty of a co-clustering task

In co-clustering, several loss functions are suited for measuring the discrepancy between the underlying classes ($\boldsymbol{Y}$, $\boldsymbol{Z}$) and some predictions ($\widehat{\boldsymbol{Y}}$, $\widehat{\boldsymbol{Z}}$). 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:

$$l_{item}(\boldsymbol{Y}, \boldsymbol{Z}, \widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{Z}}) = \overbrace{l_{row}(\boldsymbol{Y}, \widehat{\boldsymbol{Y}})}^{1 - \frac{1}{n_1} \sum_i \delta_{Y_i, \widehat{Y}_i}} + \overbrace{l_{col}(\boldsymbol{Z}, \widehat{\boldsymbol{Z}})}^{1 - \frac{1}{n_2} \sum_j \delta_{Z_j, \widehat{Z}_j}} - l_{row}(\boldsymbol{Y}, \widehat{\boldsymbol{Y}})\, l_{col}(\boldsymbol{Z}, \widehat{\boldsymbol{Z}})$$

where δ is the Kronecker delta.

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 alter its difficulty because the dimensions of the spaces where the clustering is performed are expanded. The duality between the rows and the columns imply 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 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*:

$$r_{item}(\widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{Z}}) = \mathbb{E}\Big[\, l_{item}(\boldsymbol{Y}, \boldsymbol{Z}, \widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{Z}}) \Big| \boldsymbol{X}^{(\mathrm{o})} \Big] \ , \tag{16}$$

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

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

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

$$\boldsymbol{\alpha} = \boldsymbol{\beta} = \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\pi} = \begin{pmatrix} \epsilon & \epsilon & 1-\epsilon \\ \epsilon & 1-\epsilon & 1-\epsilon \\ 1-\epsilon & 1-\epsilon & \epsilon \end{pmatrix} \ , \tag{17}$$

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

$$\mu = 1, \quad \sigma_A^2 = 1, \quad \sigma_B^2 = 1, \quad \sigma_C^2 = 1, \quad \sigma_D^2 = 1 \ , \tag{18}$$

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

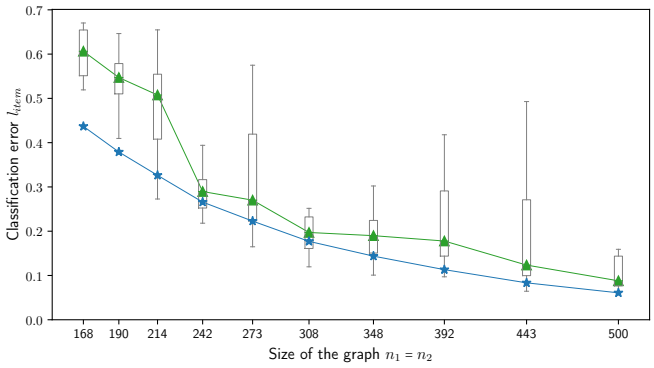


Figure 6: Classification error with respect to the size of the data matrix (lower is better); ⋆ is the median of the conditional Bayes risk; ▲ is the median prediction error obtained by our algorithm.

6.2 Analyzing the classification of the proposed inference

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 (17) 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 $(\boldsymbol{Y}, \boldsymbol{Z})$ 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.

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 show consistent estimations of the model parameters.

6.3 Analyzing the benefit of a MNAR model versus a 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 $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ of the LBM are fixed as in (17), 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 all set to one which gives an average

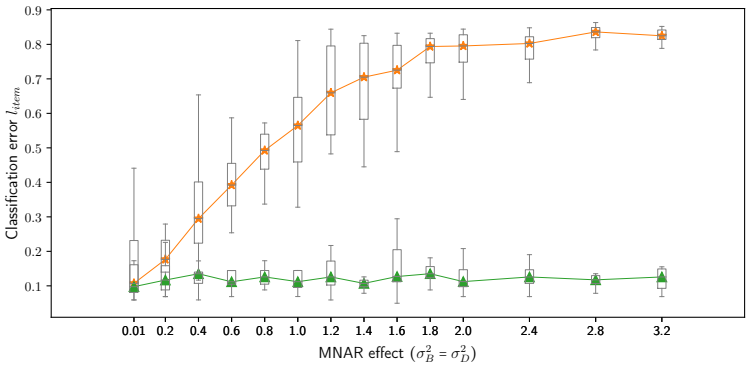


Figure 7: Classification error with respect to an increase of the MNAR effect (lower is better); ⋆ is the median prediction error obtained with the MAR model; ▲ is the median prediction error obtained with the MNAR model.

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. For each data matrix, we train the LBM with either the MAR or the MNAR model. This process is repeated 20 times, starting from the generation of a new fully observed data matrix.

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.

6.4 Analyzing the ability of the model selection criterion to select the adequate number of classes

We reuse the parameters (17) and (18) to analyze the behavior of the 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 (16) 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

[1] With equal class proportions, the expected classification error of a random allocation is $\frac{k_1-1}{k_1} + \frac{k_2-1}{k_2} - \frac{k_1-1}{k_1}\frac{k_2-1}{k_2}$, that is, 0.89 here where $k_1 = k_2 = 3$.

			$r_{item}(\widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{Z}}) = 5\%$				$r_{item}(\widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{Z}}) = 12\%$				$r_{item}(\widehat{\boldsymbol{Y}}, \widehat{\boldsymbol{Z}}) = 20\%$			
			k_2				k_2				k_2			
			2	3	4	5	2	3	4	5	2	3	4	5
$n_1 = n_2 = 30$	k_1	2	9	2			13	3			14	2		
		3	1	[7]			2	[2]			2	[1]		
		4												
		5				1								1
$n_1 = n_2 = 40$	k_1	2	4	2			17	1			17	1		
		3		[14]			1	[1]			1	[1]		
		4												
		5												
$n_1 = n_2 = 50$	k_1	2	1				11	2			15	1	2	
		3	2	[17]				[7]			1	[1]		
		4												
		5												
$n_1 = n_2 = 75$	k_1	2	1				9				13	2		
		3	1	[17]			1	[9]				[4]		
		4					1							
		5	1											1
$n_1 = n_2 = 100$	k_1	2					2				11			
		3		[19]				[18]			1	[7]	1	
		4												
		5				1								
$n_1 = n_2 = 150$	k_1	2	1					1			5	1		
		3		[14]	2		1	[18]			1	[11]		
		4		1										
		5				2								2

Figure 8: Count number of (k_1, k_2) models selected by the ICL criterion among 20 data matrices for different difficulties, as measured by the conditional Bayes risk, and different matrix sizes. All matrices are generated with the same number of row and column classes: $k_1 = k_2 = 3$.

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).

6.5 Analysing the ability of the model selection criterion to select 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_1, k_2) 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

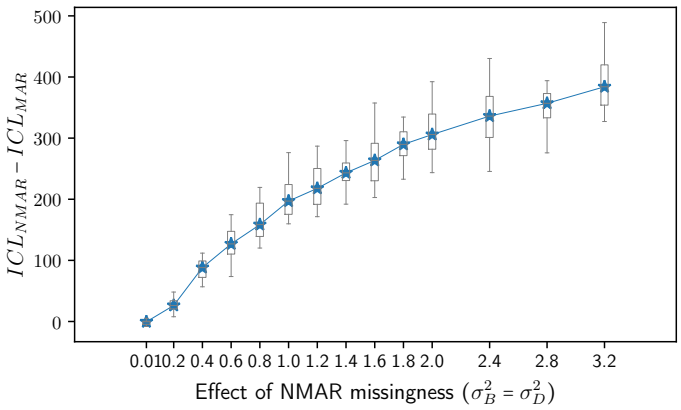


Figure 9: Difference in ICL between the MAR and MNAR models with respect to an increase of the MNAR effect, ⋆ is the median. The MNAR model is selected when the difference in ICL is positive.

experiments with the lowest MNAR effect ($\sigma_B^2 = \sigma_D^2 = 0.01$), where the prediction performances are almost identical (see Figure 7), with a median difference in ICL of -0.51 (the MAR model is chosen 13 times over the 20 repetitions).

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 577 French members of parliament (MPs) for the procedural motions and amendments for the 15th legislature (June 2017). For each text, 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 texts 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 text. 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 less that 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 frequency are detailed in Figure 10.

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

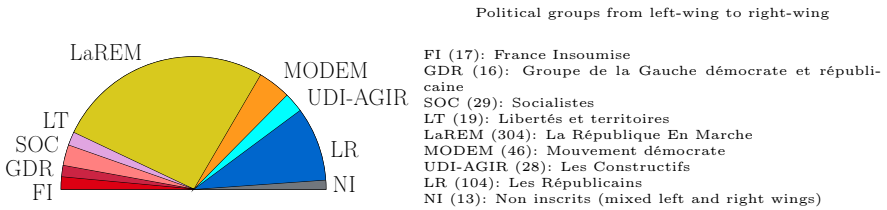


Figure 10: Hemicycle of the political groups of the French National Assembly

The ICL criterion, used to select both the numbers of classes and the type of missingness, favors a MNAR missingness with $k_1 = 14$ MP classes and $k_2 = 14$ text classes against a MAR model with 19 MP classes 23 text 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 orientation: 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 known to be political allies. Classes of texts can be analyzed with the available metadata. A bipartite opposition system appears from classes A and C. Texts from class A are 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. Texts from class C are mainly 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 $\boldsymbol{A}$ and $\boldsymbol{B}$ characterize the propensity of MPs to cast a vote. Figure 12 displays the scatter plot of $\nu_i^{(A)}$ and $\nu_i^{(B)}$, 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 $\boldsymbol{\nu}^{(A)}$ corresponding to a higher propensity to vote, and the ordinate $\boldsymbol{\nu}^{(B)}$ represents the additional effect of casting a vote when supporting the text. The membership of MPs to their political group is indicated by the plotting symbol.

We see two obvious clusters separated by the vertical axis $\boldsymbol{\nu}^{(B)}$: the bottom cluster is essentially formed by MPs from the LaREM and MODEM political

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

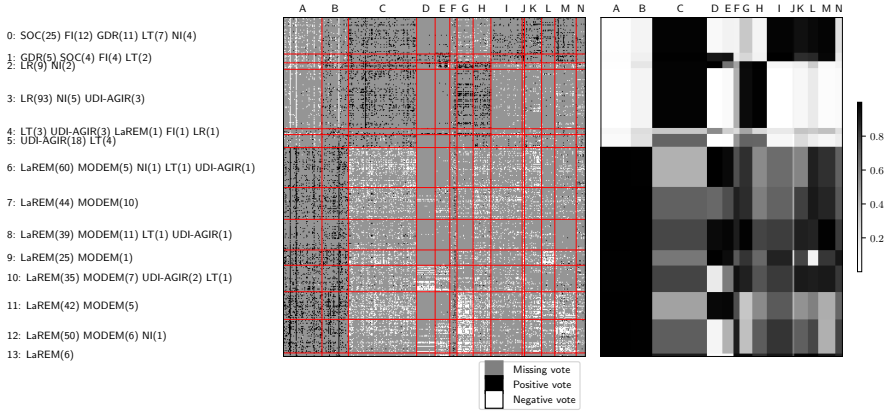


Figure 11: Left: matrix of votes reordered according to the row and column classes, for the MNAR LBM model selected by ICL, with 14 MP classes and 14 text classes. The red lines delineate class boundaries. The counts of MPs belonging to their political groups in each MP class is given on the left. Right: summary of the inferred opinions (expressed or not) for all classes of texts and MPs, as given by the estimated probability to support a text in each block of the reordered matrix.

groups, which support the government, whereas the top cluster is formed by the opposition political groups. The $\boldsymbol{\nu}^{(B)}$ 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 $\boldsymbol{\nu}^{(B)}$, 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 texts 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 manages 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

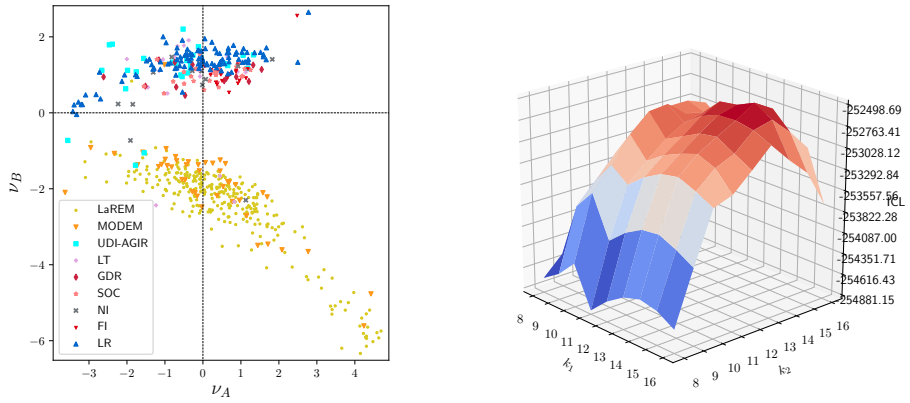


Figure 12: Left: maximum *a posteriori* estimates of the MPs propensities ($\nu_i^{(A)}$, $\nu_i^{(B)}$), with their political group memberships. $\nu_i^{(A)}$ drives the MAR effect and $\nu_i^{(B)}$ drives the MNAR one. Right: ICL curve. Maximum is reached for k_1=14 and k_2=14

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 the Delta method 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.

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.

References

Jean-Patrick Baudry and Gilles Celeux. EM for mixtures. *Statistics and Computing*, 25(4):713–726, 2015. doi: 10.1007/s11222-015-9561-x.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated classification likelihood. Technical Report RR-3521, INRIA, October 1998. URL https://hal.inria.fr/inria-00073163.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41: 561–575, 01 2003. doi: 10.1016/S0167-9473(02)00163-9.

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

Marco Corneli, Charles Bouveyron, and Pierre Latouche. Co-clustering of ordinal data via latent continuous random variables and not missing at random entries. *Journal of Computational and Graphical Statistics*, 2020. doi: 10.1080/10618600.2020.1739533. URL https://hal.archives-ouvertes.fr/hal-01978174.

Inderjit S. Dhillon. 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*, pages 269–274, aug 2001.

Inderjit S. Dhillon, Subramanyam Mallela, and Dharmendra S. Modha. Information-theoretic co-clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 89—-98, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581137370. doi: 10.1145/956750.956764. URL https://doi.org/10.1145/956750.956764.

Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 2006, pages 126–135, 01 2006. doi: 10.1145/1150402.1150420.

Thomas George and Srujana Merugu. A scalable collaborative filtering framework based on co-clustering. In *Fifth IEEE International Conference on Data Mining (ICDM)*, 12 2005. ISBN 0-7695-2278-5. doi: 10.1109/ICDM.2005.14.

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

José Miguel Hernández-Lobato, Neil Houlsby, and Zoubin Ghahramani. Probabilistic matrix factorization with non-random missing data. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1512–1520, 2014.

Tommi S. Jaakkola. Tutorial on variational approximation methods. In Manfred Opper and David Saad, editors, *Advanced Mean Field Methods: Theory and Practice*, pages 129–159. MIT Press, 2000.

Julien Jacques and Christophe Biernacki. Model-Based Co-clustering for Ordinal Data. *Computational Statistics & Data Analysis*, 123:101–115, July 2018. doi: 10.1016/j.csda.2018.01.014. URL https://hal.inria.fr/hal-01448299.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183—-233, November 1999. ISSN 0885-6125. doi: 10.1023/A:1007665907178. URL https://doi.org/10.1023/A:1007665907178.

Christine Keribin, Vincent Brault, Gilles Celeux, and Gérard Govaert. Model selection for the binary latent block model. In *Proceedings of COMPSTAT*, 08 2012.

Christine Keribin, Vincent Brault, Gilles Celeux, and Gérard Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6):1201–1216, 2015.

Yong-Deok Kim and Seungjin Choi. Bayesian binomial mixture model for collaborative prediction with non-random missing data. In *Eighth ACM Conference on Recommender Systems (RecSys)*, page 201–208, 2014.

Yuval Kluger, Ronen Basri, Joseph Chang, and Mark Gerstein. Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Research*, 13:703–716, 05 2003. doi: 10.1101/gr.648603.

Lazhar Labiod and Mohamed Nadif. Co-clustering for binary and categorical data with maximum modularity. In *11th IEEE International Conference on Data Mining (ICDM)*, pages 1140–1145, 2011. doi: 10.1109/ICDM.2011.37.

Pierre Latouche, Etienne Birmelé, and Christophe Ambroise. Overlapping stochastic block models with application to the French political blogosphere. *The Annals of Applied Statistics*, 5(1):309–336, Mar 2011. ISSN 1932-6157. doi: 10.1214/10-aoas382. URL http://dx.doi.org/10.1214/10-AOAS382.

Roderick J. A. Little and Donald B. Rubin. Introduction. In *Statistical Analysis with Missing Data*, chapter 1, pages 1–23. John Wiley & Sons, 1986. ISBN 9781119013563.

Aurore Lomet. *Sélection de modèle pour la classification croisée de données continues*. PhD thesis, Université de technologie de Compiègne, 2012. 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.

Aurore Lomet, Gérard Govaert, and Yves Grandvalet. Design of artificial data tables for co-clustering analysis. Technical report, Université de technologie de Compiègne, France, 2012.

Benjamin M. Marlin, Richard S. Zemel, Sam T. Roweis, and Malcolm Slaney. Collaborative filtering and the missing at random assumption. In *Twenty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 267–275, 2007.

Benjamin M. Marlin, Richard S. Zemel, Sam T. Roweis, and Malcolm Slaney. Recommender systems, missing data and statistical model estimation. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2686–2691, 2011.

Mohamed Nadif and Gérard Govaert. Latent block model for contingency table. *Communications in Statistics—Theory and Methods*, 39(3):416–425, 01 2010. doi: 10.1080/03610920903140197.

Evangelos E. Papalexakis, Nikolaos Sidiropoulos, and Rasmus Bro. From k-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors. *IEEE Transactions on Signal Processing*, 61(2):493–506, January 2013. ISSN 1053-587X. doi: 10.1109/TSP.2012.2225052.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Beatriz Pontes, Raúl Giráldez, and Jesús S. Aguilar-Ruiz. Biclustering on expression data: A review. *Journal of Biomedical Informatics*, 57:163–180, 2015. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2015.06.028. URL http://www.sciencedirect.com/science/article/pii/S1532046415001380.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915,

Aug 2011. ISSN 0090-5364. doi: 10.1214/11-aos887. URL http://dx.doi.org/10.1214/11-AOS887.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. ISSN 00063444. URL http://www.jstor.org/stable/2335739.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 1670–1679, 2016. URL http://proceedings.mlr.press/v48/schnabel16.html.

Margot Selosse, Julien Jacques, and Christophe Biernacki. Model-based co-clustering for mixed type data. *Computational Statistics & Data Analysis*, 144:106866, 2020a. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2019.106866. URL http://www.sciencedirect.com/science/article/pii/S016794731930221X.

Margot Selosse, Julien Jacques, and Christophe Biernacki. Textual data summarization using the self-organized co-clustering model. *Pattern Recognition*, 103:107315, 2020b. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2020.107315. URL http://www.sciencedirect.com/science/article/pii/S0031320320301199.

Hanhuai Shan and Arindam Banerjee. Bayesian co-clustering. In *2008 Eighth IEEE International Conference on Data Mining*, pages 530–539, 2008.

Emilie Shireman, Douglas Steinley, and Michael Brusco. Examining the effect of initialization strategies on the performance of Gaussian mixture modeling. *Behavior Research Methods*, 49, 12 2015. doi: 10.3758/s13428-015-0697-6.

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

Timothée Tabouy, Pierre Barbillon, and Julien Chiquet. Variational inference for stochastic block models from sampled data. *Journal of the American Statistical Association*, 115(529):455–466, 2020.

Diego P Vázquez, Nico Blüthgen, Luciano Cagnolo, and Natacha P Chacoff. Uniting pattern and process in plant–animal mutualistic networks: a review. *Annals of botany*, 103(9):1445–1457, 2009.

Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.

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

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

The criterion to be optimized is :

$$\mathcal{J}(q_\gamma, \theta) = \mathcal{H}(q_\gamma) + \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{X}^{(\mathrm{o})}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}; \theta)] \ , \tag{19}$$

where θ is the list of all model parameters: $\theta = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}, \mu, \sigma_A^2, \sigma_B^2, \sigma_C^2, \sigma_D^2)$.

We restrict the form of the variational distribution q_γ to get a fully factorized form:

$$q_\gamma = \prod_{i=1}^{n_1} \mathcal{M}\Big(1; \boldsymbol{\tau}_i^{(Y)}\Big) \times \prod_{j=1}^{n_2} \mathcal{M}\Big(1; \boldsymbol{\tau}_j^{(Z)}\Big) \times \prod_{i=1}^{n_1} \mathcal{N}\Big(\nu_i^{(A)}, \rho_i^{(A)}\Big) \times \prod_{i=1}^{n_1} \mathcal{N}\Big(\nu_i^{(B)}, \rho_i^{(B)}\Big) \times \prod_{j=1}^{n_2} \mathcal{N}\Big(\nu_j^{(C)}, \rho_j^{(C)}\Big) \times \prod_{j=1}^{n_2} \mathcal{N}\Big(\nu_j^{(D)}, \rho_j^{(D)}\Big) \ ,$$

where γ denotes the list of parameters of the variational distribution: $\gamma = (\boldsymbol{\tau}^{(Y)}, \boldsymbol{\tau}^{(Z)}, \boldsymbol{\nu}^{(A)}, \boldsymbol{\rho}^{(A)}, \boldsymbol{\nu}^{(B)}, \boldsymbol{\rho}^{(B)}, \boldsymbol{\nu}^{(C)}, \boldsymbol{\rho}^{(C)}, \boldsymbol{\nu}^{(D)}, \boldsymbol{\rho}^{(D)})$. The entropy is additive across independant variables, so we get:

$$\mathcal{H}(q_\gamma) = -\sum_{iq} \tau_{iq}^{(Y)} \log \tau_{iq}^{(Y)} - \sum_{jl} \tau_{jl}^{(Z)} \log \tau_{jl}^{(Z)} + (n_1 + n_2)(\log(2\pi) + 1) + \frac{1}{2} \sum_i \Big(\log \rho_i^{(A)} + \log \rho_i^{(B)} \Big) + \frac{1}{2} \sum_j \Big(\log \rho_j^{(C)} + \log \rho_j^{(D)} \Big) \ .$$

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

$$\mathbb{E}_{q_\gamma}\Big[\log p\Big(\boldsymbol{X}^{(\mathrm{o})}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}\Big)\Big] = \mathbb{E}_{q_\gamma}\Big[\log p\Big(\boldsymbol{X}^{(\mathrm{o})} \Big| \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}\Big)\Big] + \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{Y})] + \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{Z})] + \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{A})] + \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{B})] + \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{C})] + \mathbb{E}_{q_\gamma}[\log p(\boldsymbol{D})] \ ,$$

with the following terms:

$$\begin{aligned}
\mathbb{E}_{q_\gamma}[\log p(\boldsymbol{Y})] &= \sum_{iq} \mathbb{E}_{q_\gamma} Y_{iq} \log \alpha_q = \sum_{iq} \tau_{iq}^{(Y)} \log \alpha_q \\
\mathbb{E}_{q_\gamma}[\log p(\boldsymbol{Z})] &= \sum_{jl} \mathbb{E}_{q_\gamma} Z_{jl} \log \beta_l = \sum_{jl} \tau_{jl}^{(Z)} \log \beta_l \\
\mathbb{E}_{q_\gamma}[\log p(\boldsymbol{A})] &= -\frac{n_1}{2} \log 2\pi - \frac{n_1}{2} \log \sigma_A^2 - \frac{1}{2\sigma_A^2} \sum_i \mathbb{E}_{q_\gamma} A_i^2 \\
&= -\frac{n_1}{2} \log 2\pi - \frac{n_1}{2} \log \sigma_A^2 - \frac{1}{2\sigma_A^2} \sum_i \left(\left(\nu_i^{(A)}\right)^2 + \rho_i^{(A)} \right) \\
\mathbb{E}_{q_\gamma}[\log p(\boldsymbol{B})] &= -\frac{n_1}{2} \log 2\pi - \frac{n_1}{2} \log \sigma_B^2 - \frac{1}{2\sigma_B^2} \sum_i \left(\left(\nu_i^{(B)}\right)^2 + \rho_i^{(B)} \right) \\
\mathbb{E}_{q_\gamma}[\log p(\boldsymbol{C})] &= -\frac{n_2}{2} \log 2\pi - \frac{n_2}{2} \log \sigma_C^2 - \frac{1}{2\sigma_C^2} \sum_j \left(\left(\nu_j^{(C)}\right)^2 + \rho_j^{(C)} \right) \\
\mathbb{E}_{q_\gamma}[\log p(\boldsymbol{D})] &= -\frac{n_2}{2} \log 2\pi - \frac{n_2}{2} \log \sigma_D^2 - \frac{1}{2\sigma_D^2} \sum_j \left(\left(\nu_j^{(D)}\right)^2 + \rho_j^{(D)} \right)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_{q_\gamma}\left[\log p\left(\boldsymbol{X}^{(\mathrm{o})} \middle| \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D} \right)\right] &= \sum_{ql,ij:X_{ij}^{(\mathrm{o})}=1} \tau_{iq}^{(Y)} \tau_{jl}^{(Z)} \mathbb{E}_{q_\gamma}[\log p_1] \\
&+ \sum_{ql,ij:X_{ij}^{(\mathrm{o})}=0} \tau_{iq}^{(Y)} \tau_{jl}^{(Z)} \mathbb{E}_{q_\gamma}[\log p_0] + \sum_{ql,ij:X_{ij}^{(\mathrm{o})}=\mathrm{NA}} \tau_{iq}^{(Y)} \tau_{jl}^{(Z)} \mathbb{E}_{q_\gamma}[\log (1 - p_0 - p_1)] \quad , \qquad (20)
\end{aligned}$$

with p_0 and p_1 defined in Equations (5)–(6).

Equation (20) involves the computation of the expectations of the following nonlinear functions:

$$\begin{aligned}
f_1(x, y) &= \log\left(\pi_{ql} \operatorname{expit}(\mu + x + y)\right) \\
f_0(x, y) &= \log\left((1 - \pi_{ql}) \operatorname{expit}(\mu + x - y)\right) \\
f_{\mathrm{NA}}(x, y) &= \log\left(1 - \pi_{ql} \operatorname{expit}(\mu + x + y) - (1 - \pi_{ql}) \operatorname{expit}(\mu + x - y)\right) \quad .
\end{aligned}$$

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

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

which yields in our case:

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

The criterion is now fully computable.

B Initialization of the VEM algorithm with spectral clustering.

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

Input:

- observed data $\boldsymbol{X}^{(o)}$
- k_1 and k_2 number of row groups and column groups

Function `SpectralClustering`(W *adjacency matrix,* k *number of clusters*)**:**

- Define $\boldsymbol{D}$ the diagonal matrix, element of $\mathbb{R}^{n \times n}$: $D_{ii} = \sum_q W_{iq}$
- Define $\boldsymbol{L} = \boldsymbol{D}^{-1/2} \boldsymbol{W} \boldsymbol{D}^{-1/2}$
- Find the eigenvectors corresponding to the k eigenvalues of $\boldsymbol{L}$ that are largest in absolute value. Form the matrix $\boldsymbol{U} = [U_1, ..., U_k] \in \mathbb{R}^{n \times k}$ concatenating the eigenvectors into columns.

Return results of k-means with k clusters on $\boldsymbol{U}$.

begin

- Build $\boldsymbol{Y}$ the $n_1 \times k_1$ indicator matrix of the row cluster memberships with results of $SpectralClustering(\boldsymbol{X}\boldsymbol{X}^T, k_1)$
- Build $\boldsymbol{Z}$ the $n_2 \times k_2$ indicator matrix of the column cluster memberships with results of $SpectralClustering(\boldsymbol{X}^T\boldsymbol{X}, k_2)$.
- $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\pi}$ estimated from $\boldsymbol{Y}$ and $\boldsymbol{Z}$
- μ initialized such as $\text{expit}(\mu)$ is the global missingness rate
- σ_A^2, σ_B^2, σ_C^2 and σ_D^2 sampled from $U_{]0,1]}$

end

Result:

- $\theta = \left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}, \mu, \sigma_A^2, \sigma_B^2, \sigma_C^2, \sigma_D^2\right)$ the model parameters
- $\boldsymbol{Y}$ and $\boldsymbol{Z}$ the row and column cluster memberships

C Asymptotic form of the Integrated Completed Likelihood

C.1 ICL of the MNAR model

The ICL criterion of the LBM extended to the MNAR missingness process presented in Section 3.2 has the following asymptotic form for n_1 and n_2:

$$\begin{aligned} ICL(k_1,k_2) =& \max_{\boldsymbol{\theta}} \log p\Big(\boldsymbol{X}^{(\mathrm{o})},\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D};\boldsymbol{\theta}\Big) - \frac{k_1 k_2}{2}\log(n_1 n_2) \\ & - \frac{k_1-1}{2}\log(n_1) - \frac{k_2-1}{2}\log(n_2) \\ & + n_1\log(2\pi) - \log(n_1) + n_2\log(2\pi) - \log(n_2) \\ & + o(\log n_1) + o(\log n_2) \ . \end{aligned} \tag{21}$$

Proof. With independent latent variables and independent priors on the parameters, the ICL criterion reads

$$\begin{aligned} ICL &= \log p\Big(\boldsymbol{X}^{(\mathrm{o})},\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D}\Big) \\ &= \log \int p\Big(\boldsymbol{X}^{(\mathrm{o})},\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D}\Big|\boldsymbol{\theta}\Big)p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} \\ &= \log \int p\Big(\boldsymbol{X}^{(\mathrm{o})}\Big|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D},\boldsymbol{\pi}\Big)p(\boldsymbol{\pi})p(\mu)\mathrm{d}\boldsymbol{\pi}\mathrm{d}\mu \\ &\quad + \log \int p(\boldsymbol{Y}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})\mathrm{d}\boldsymbol{\alpha} + \log \int p(\boldsymbol{Z}|\boldsymbol{\beta})p(\boldsymbol{\beta})\mathrm{d}\boldsymbol{\beta} \\ &\quad + \log \int p\big(\boldsymbol{A}\big|\sigma_A^2\big)p\big(\sigma_A^2\big)\mathrm{d}\sigma_A^2 + \log \int p\big(\boldsymbol{B}\big|\sigma_B^2\big)p\big(\sigma_B^2\big)\mathrm{d}\sigma_B^2 \\ &\quad + \log \int p\big(\boldsymbol{C}\big|\sigma_C^2\big)p\big(\sigma_C^2\big)\mathrm{d}\sigma_C^2 + \log \int p\big(\boldsymbol{D}\big|\sigma_D^2\big)p\big(\sigma_D^2\big)\mathrm{d}\sigma_D^2 \ . \end{aligned} \tag{22}$$

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 $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$:

$$\begin{aligned} \log p(\boldsymbol{Y}) &= \log \int p(\boldsymbol{Y}|\boldsymbol{\alpha})p(\boldsymbol{\alpha};a)\mathrm{d}\boldsymbol{\alpha} \\ &= \prod_i \log \int \prod_q (\alpha_q)^{Y_{iq}} \frac{1}{\mathcal{B}(a)} \prod_{iq} (\alpha_q)^{a-1}\mathrm{d}\boldsymbol{\alpha} \\ &= \log \mathcal{B}(a+\sum_i \boldsymbol{Y}_i) - \log \mathcal{B}(a) \\ &= \sum_q \log\Gamma(Y_{\cdot q}+a) + \log\Gamma(k_1 a) - \log\Gamma(n_1+k_1 a) - k_1\log\Gamma(a) \ , \end{aligned}$$

where $Y_{\cdot q} = \sum_i Y_{iq}$. The Stirling approximation $\log\Gamma(x) = x\log x - x - \frac{1}{2}\log x +$

$o(\log x)$ leads to the following asymptotic development of $\log p(\boldsymbol{Y})$:

$$\begin{aligned}\log p(\boldsymbol{Y}) &= \sum_q \log\Gamma(Y_{:q}+a) - \log\Gamma(n_1+k_1a) + o(\log n_1)\\ &= \sum_q Y_{:q}\log Y_{:q} - n_1 - \frac{1}{2}n_1\\ &\quad - \left(n_1\log n_1 + k_1 a\log n_1 - n_1 - \frac{1}{2}\log n_1\right) + o(\log n_1)\ .\end{aligned}$$

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

$$\begin{aligned}\log p(\boldsymbol{Y}) &= \sum_q Y_{:q}\log(\frac{1}{n_1}Y_{:q}) - \frac{k_1-1}{2}\log n_1 + o(\log n_1)\\ &= \max_{\boldsymbol{\alpha}} \log p(\boldsymbol{Y};\boldsymbol{\alpha}) - \frac{k_1-1}{2}\log n_1 + o(\log n_1)\ .\end{aligned} \tag{23}$$

Similarly, we get:

$$\begin{aligned}\log p(\boldsymbol{Z}) &= \sum_l \log\Gamma(Z_{:l}+a) + \log\Gamma(k_2a) - \log\Gamma(n_2+k_2a) - k_2\log\Gamma(a)\\ &= \max_{\boldsymbol{\beta}} \log p(\boldsymbol{Z};\boldsymbol{\beta}) - \frac{k_2-1}{2}\log n_2 + o(\log n_2)\ ,\end{aligned} \tag{24}$$

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

We set non-informative InverseGamma(β, β) distributions (as β tends to zero) as priors on σ_A^2, σ_B^2, σ_C^2 and σ_D^2:

$$\begin{aligned}\log p(\boldsymbol{A}) &= \int p\big(\boldsymbol{A}|\sigma_A^2\big)p\big(\sigma_A^2;\beta\big)\,\mathrm{d}\sigma_A^2\\ &= \prod_i \log\int \left(2\sigma_A^2\right)^{-\frac{n_1}{2}}\exp\left(-\frac{\sum A_i^2}{2\sigma_A^2}\right)\frac{\beta^\beta}{\Gamma(\beta)}\exp\left(-\frac{\beta}{\sigma_A^2}\right)\left(\sigma_A^2\right)^{-\beta-1}\mathrm{d}\sigma_A^2\\ &= \log\frac{\beta^\beta}{\Gamma(\beta)}2^{\left(-\frac{n_1}{2}\right)}\int \sigma_A^{2\left(-\frac{n_1}{2}-\beta-1\right)}\exp\left(-\frac{2\beta+\sum A_i^2}{2}\cdot\frac{1}{\sigma_A^2}\right)\mathrm{d}\sigma_A^2\\ &= \log\frac{\beta^\beta}{\Gamma(\beta)}2^\beta\left(2\beta+\sum A_i\right)^{\left(-\frac{n_1}{2}-\beta\right)}\Gamma(\frac{n_1}{2}+\beta)\ .\end{aligned}$$

To consider a non-informative InverseGamma(β, β) distribution, we realize a first order Taylor development as β tends to 0:

$$\log p(\boldsymbol{A}) \approx \log\Gamma(\frac{n_1}{2}) + \log\beta - \frac{n_1}{2}\log(\sum A_i^2)\ .$$

Using the Stirling approximation of $\Gamma(x)$ we get the following asymptotic development of $\log p(\boldsymbol{A})$:

$$\begin{aligned}\log p(\boldsymbol{A}) &= \frac{n_1}{2}\log n_1 - \frac{n_1}{2} - \frac{1}{2}\log n_1 - \frac{n_1}{2}\log\sum A_i^2 + o(\log n_1)\\ &= \max_{\sigma_A^2}\log p\big(\boldsymbol{A};\sigma_A^2\big) + \frac{n_1}{2}\log(2\pi) - \frac{1}{2}\log n_1 + o(\log n_1)\ .\end{aligned} \tag{25}$$

Similarly, we get:

$$\begin{aligned}
\log p(\boldsymbol{B}) &= \max_{\sigma_B^2} \log p\left(\boldsymbol{B}; \sigma_B^2\right) + \frac{n_1}{2}\log(2\pi) - \frac{1}{2}\log n_1 + o(\log n_1) \\
\log p(\boldsymbol{C}) &= \max_{\sigma_C^2} \log p\left(\boldsymbol{C}; \sigma_C^2\right) + \frac{n_2}{2}\log(2\pi) - \frac{1}{2}\log n_2 + o(\log n_2) \\
\log p(\boldsymbol{D}) &= \max_{\sigma_D^2} \log p\left(\boldsymbol{D}; \sigma_D^2\right) + \frac{n_2}{2}\log(2\pi) - \frac{1}{2}\log n_2 + o(\log n_2) \quad .
\end{aligned} \tag{26}$$

Using the standard BIC approximation, we have

$$\begin{aligned}
\log p\left(\boldsymbol{X}^{(\mathrm{o})} | \boldsymbol{Y}, \boldsymbol{Z}\right) &= \log \int p\left(\boldsymbol{X}^{(\mathrm{o})} | \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\pi}\right) p(\boldsymbol{\pi}) p(\mu) \mathrm{d}\boldsymbol{\pi} \mathrm{d}\mu \\
&= \max_{\boldsymbol{\pi}} \log p\left(\boldsymbol{X}^{(\mathrm{o})} | \boldsymbol{Y}, \boldsymbol{Z}; \boldsymbol{\pi}, \mu\right) + \frac{k_1 k_2}{2}\log(n_1 n_2) \\
&\quad + o(\log n_1) + o(\log n_2) \quad .
\end{aligned} \tag{27}$$

The ICL criterion 21 is directly derived from equations 22, 23, 24, 25, 26 and 27.

□

C.2 ICL of the LBM with MAR data

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

Latent Block Model

$$\begin{aligned}
&Y_i \overset{\text{iid}}{\sim} \mathcal{M}(1; \boldsymbol{\alpha}), && \boldsymbol{\alpha} \in \mathbf{S}_{k_1 - 1} \\
&Z_j \overset{\text{iid}}{\sim} \mathcal{M}(1; \boldsymbol{\beta}), && \boldsymbol{\beta} \in \mathbf{S}_{k_2 - 1} \\
&(X_{ij} | Y_i = q, Z_j = l) \overset{\text{ind}}{\sim} \mathcal{B}(\pi_{ql}), && \pi_{ql} \in [0, 1]
\end{aligned}$$

MAR data model

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

Observations are generated according to:

$$\left(X_{ij}^{(\mathrm{o})} \middle| X_{ij}, M_{ij}\right) = \begin{cases} X_{ij} & \text{if} \quad M_{ij} = 1 \\ \text{NA} & \text{if} \quad M_{ij} = 0 \end{cases}$$

The ICL of this model has the following asymptotic form for n_1 and n_2:

$$\begin{aligned} ICL(k_1,k_2) = & \max_{\boldsymbol{\theta}} \log p\Big(\boldsymbol{X}^{(\mathrm{o})},\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{A},\boldsymbol{C};\boldsymbol{\theta}\Big) - \frac{k_1 k_2}{2}\log(n_1 n_2) \\ & - \frac{k_1-1}{2}\log(n_1) - \frac{k_2-1}{2}\log(n_2) \\ & + \frac{1}{2}\big(n_1\log(2\pi) - \log(n_1) + n_2\log(2\pi) - \log(n_2)\big) \\ & + o(\log n_1) + o(\log n_2) \ . \end{aligned} \tag{28}$$

D Supplemental figures for estimations

This section provides additional experimental results that show 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 13 displays the maximum absolute error made on the parameters $\boldsymbol{\pi}$ of the Bernoulli distributions that model the probability of $\boldsymbol{X}$ conditionally to the row and column classes. This error decreases as the size of the data matrices grows, which is consistent with our expectations.

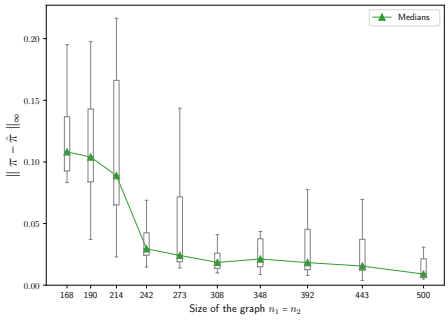


Figure 13: Maximum error between the true ($\boldsymbol{\pi}$) and the estimated ($\hat{\boldsymbol{\pi}}$) probabilities associated to the blocks of the data matrix $\boldsymbol{X}$ as a function of its size.

Figure 14 displays the mean squared error (MSE) between the generated and estimated values of the latent variables $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$ 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 $\boldsymbol{A}$ and $\boldsymbol{C}$ are comparable as well as the curves of the variables $\boldsymbol{B}$ and $\boldsymbol{D}$. This is expected as the data matrices are generated with symmetric characters in rows and columns.

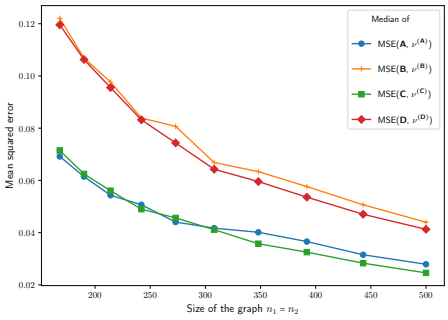


Figure 14: Mean squared error of the maximum *a posteriori* estimates of the latent variables $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$ governing the propensity of missingness.

Figure 15 compares the estimated values of $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$ 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 positives values.

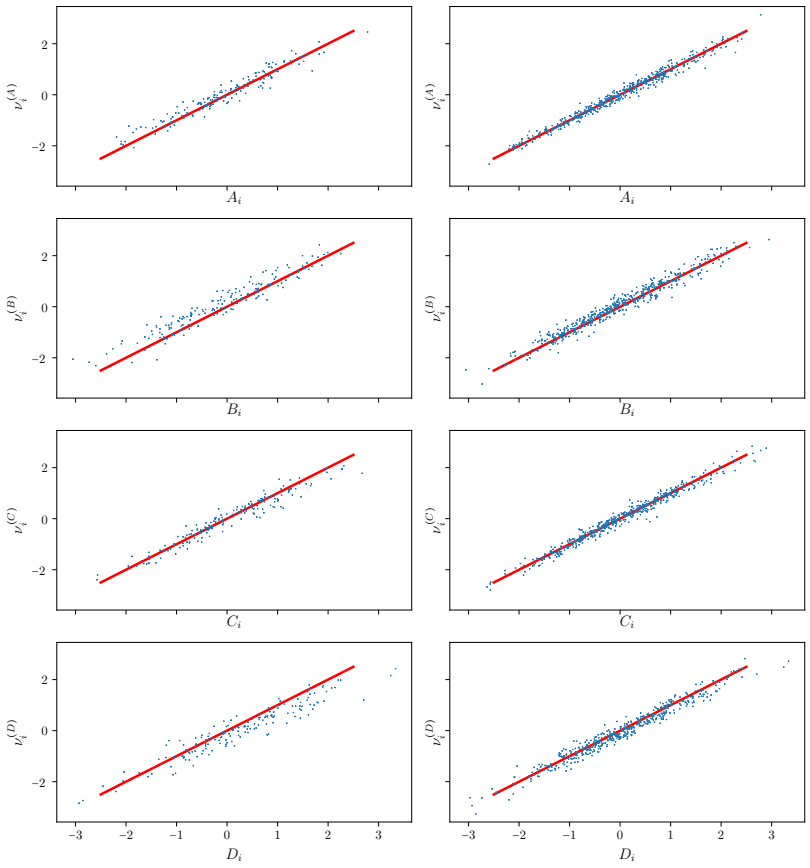


Figure 15: Maximum *a posteriori* estimates of the latent variables governing the propensity of missingness versus their true generated values.Left: $n_1 = n_2 = 168$ and the conditionnal Bayes risk is 0.44; right: $n_1 = n_2 = 500$ and the conditionnal Bayes risk is 0.05. The identity line is drawn in red for reference.

E Supplemental figures for the French national assembly votes analysis

Figure 16 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 the political groups supporting the government. The opposition systems appear clearly: on the texts from classes A and E, the votes contrast the membership to the opposition parties versus the governmental alliance, whereas on texts from classes C and D, they

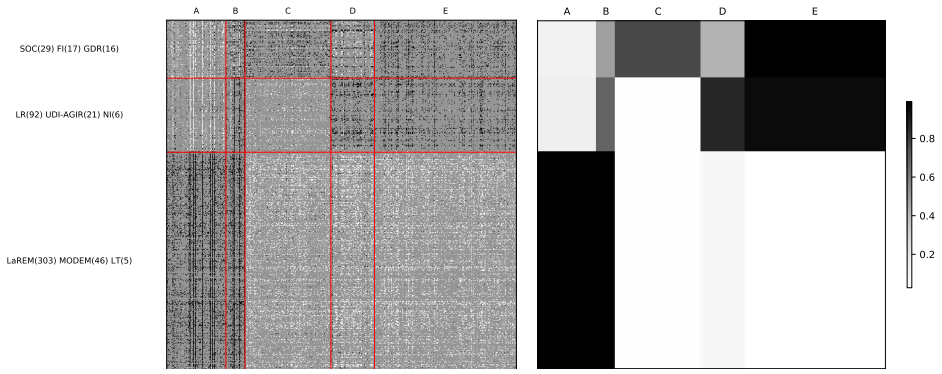


Figure 16: Left: matrix of votes reordered according to the row and column classes, for the MNAR LBM with 3 MP classes and 5 text classes. The red lines delineate class boundaries. The counts of MPs belonging to the three most represented political groups in each MP class is given on the left. Right: summary of the inferred opinions (expressed or not) for all classes of texts and MPs, as given by the estimated probability to support a text in each block of the reordered matrix.

separate the left-wing from the right-wing oppositions. Class B gathers various texts 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 text propensities to be voted upon and to be positively perceived by nonvoters. These propensities are encoded in the values of the latent variables $\boldsymbol{C}$ and $\boldsymbol{D}$. Figure 17 displays the scatter plot of $\nu_j^{(C)}$ and $\nu_j^{(D)}$, the maximum *a posteriori* estimates of C_j and D_j under the variational distribution, for all voted texts. The abscissa $\boldsymbol{\nu}^{(C)}$ reflects the mobilization on the texts, with higher mobilization for higher values, and the ordinate $\boldsymbol{\nu}^{(D)}$ represents the additional effect of mobilizing specifically supporting voters. The fourteen-cluster membership of texts (there is no obvious relevant classification for texts) is indicated by the plotting symbol.

Some relationship between missingness and membership to text classes emerge from this plot. A first cluster of text appears in the positive quadrant, with texts mainly proposed by the government, categorized in text classes A and B. A second cluster, smaller, on the upper left, is mainly formed by texts categorized in class D, voted positively by few voters. All the texts 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 texts, that would have a tendency to be voted negatively by nonvoters. These texts 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 texts, on the right-hand side, is made of texts

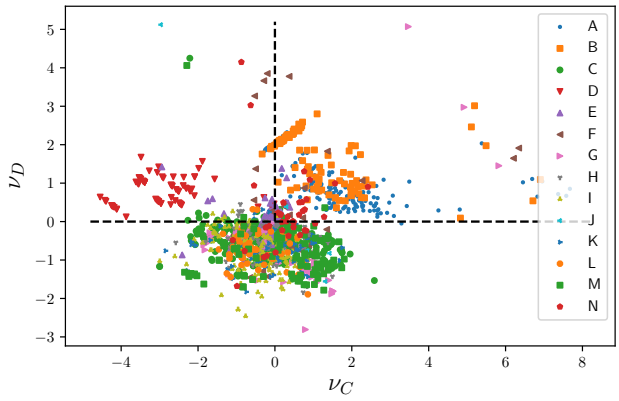


Figure 17: maximum *a posteriori* estimates of the text propensities $(\nu_j^{(C)}, \nu_j^{(D)})$, with their clustering class memberships. $\nu_j^{(C)}$ drives the MAR effect and $\nu_j^{(D)}$ drives the MNAR one.

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