Learning from missing data with the binary latent block model

Laura Fuentes
Paris-Saclay University
Saclay, France
laura.fuentes-vicente@universite-paris-saclay.fr

Angel Reyero Lobo
Paris-Saclay University
Saclay, France
angel.reyero-lobo@universite-paris-saclay.fr

ABSTRACT

In the context of information extraction, clustering proves to be an effective method for regrouping similar individuals. However, given the high-dimensionality of current data, there is a need for an approach to also reduce the variables, giving rise to co-clustering. This method has been widely applied in real-life cases, such as collaborative filtering, which relies on considering other users' ratings of items to suggest similar items to individuals. Nevertheless, it is observed that not all users will rate all products, leading to missing cases.

Moreover, it has been shown that missingness is induced by the nature of the data; users tend to rate items they either like a lot or find dissatisfying. However, there is a scarcity of literature addressing informative missing values. To tackle this gap, Frisch et al., 2022 proposed an informative missingness model that can be integrated with the standard latent block model for co-clustering, particularly in the case of binary data. They also suggested estimating the latent variables through the variational expectation-maximization and introduced a model selection criterion. Finally, the proposed model was tested in a real-world scenario.

Keywords: Co-clustering, latent block model (LBM), variational expectation-maximization (VEM), missing values, missing not at random (MNAR), integrated completed likelihood (ICL).

Notations: Let us introduce some notations. First, we denote the interval $\{1, ..., h\}$ as $[h]$. Throughout the paper, index i will refer to the individuals, j to the variables, k to the row clusters and l to the column clusters. Their respective ranges are $[n_1]$, $[n_2]$, $[K]$ and $[L]$. Then, the binary data matrix X will be of size $n_1 \times n_2$. The partially observed data matrix will be denoted by X^{obs}, taking in each case values in $\{0, 1, \text{NA}\}$. M denotes the binary mask, where if $M_{ij} = 0$ then the i, j-th entry is missing $X_{ij}^{\text{obs}} = \text{NA}$.

1 INTRODUCTION

The primary objective of statistics is to summarize information to enhance understanding and facilitate better decision-making. Clustering is among the most widely used methods for achieving this, as it combines data reduction and information extraction. However, the choice of the correct representation is crucial, such as determining the appropriate number of clusters. To address this issue and take advantage from the benefits of statistics, model-based clustering (MBC) is considered a reference point. This approach captures the flexibility of mixture models, where each cluster is represented as a probability distribution.

Given the current prevalence of high-dimensional data sets, an extension of the clustering concept for covariates has been suggested. This extension is known as the *co-clustering* paradigm, providing a method for simultaneously clustering both rows and columns in the data matrix. This enables us to preserve interpretability in the reduced data, as both the meaning of individuals and covariates is retained. We may notice that the co-clustering model differs from bi-clustering as it does not allow overlaps between the clusters.

Similar to the model-based clustering approach, a model-based method is proposed for the co-clustering problem known as the latent block model (LBM) (refer to Biernacki et al., 2023), which provides a strong statistical foundation for estimation and model selection.

We observe that almost all clustering and co-clustering methods are not adapted to accommodate missing values. Nevertheless, in the era of big data, finding data sets with no missing values remains idealistic. To incorporate missing values into our model, it is necessary to examine the data-generating process to identify the type of missingness, as described in RUBIN, 1976.

Missingness can be independent of the data, as seen in cases of random sensor failures or forgetting to fill in a form, leading to missing completely at random (MCAR). In another scenario, missingness depends on the values observed in the data, as in a medical study where not all patients undergo all medical tests if the observed values are typical. This scenario is referred to as missing at random (MAR). Finally, there can be a situation where missingness is determined by the underlying values, such as reviewers expressing their opinion only when products are extremely good or bad in the case of collaborative filtering. This missingness mechanism is called missing not at random (MNAR), and not accounting for it can introduce bias in the estimation of underlying parameters.

MNAR gap in the models: In the context of Matrix Factorization, Hernandez-Lobato et al., 2014 proposed a probabilistic model that demonstrated the advantages of the MNAR setting over the previous MAR setting in collaborative filtering.

Regarding clustering, few methods have been proposed to accommodate the MNAR assumption. In Marlin et al., 2011, based on responses from a survey on an online radio service, a MNAR mechanism was proposed, significantly improving previous results obtained under the MAR assumption. The CPT-v captures the dependency of an item's probability of being rated on the user's rating for it with a value v. In Marlin et al., 2012, they propose the Logit-vd, a generalization of CPT-v that allows the missingness probability to vary among items. It includes two factors: one depending on the rating value v and another on each item d. Following a symmetric idea for co-clustering, Frisch et al., 2022 also includes the row.

Adaptations of co-clustering for missing data have been introduced progressively over the years, with a focus mainly on MCAR and MAR missing models. For example, Selosse et al., 2020 presents the multiple latent block model, which not only allows for different types of data but also accommodates missing values under the

MAR assumption. Only Corneli et al., 2020 deals with the MNAR assumption. They propose a latent Gaussian random variable to generate ordinal data using a threshold. Their missingness mechanism depends on the row and column cluster, indirectly on the missing value.

In contrast to the MNAR presented in Corneli et al., 2020, Frisch et al., 2022 proposes a model that explicitly depends on the column, row and the underlying value. The referred model will be presented in section 2.2.

Contributions: Frisch et al., 2022 presents an extension of the Latent Block Model (LBM) to address missing data on binary data matrices. The main contributions of this publication could be grouped in two parts. First, they propose a flexible extension of the LBM algorithm adapted to MNAR missing process. To deal with tractability issues concerning the complete likelihood, they implement a Variational Expectation Maximization algorithm (VEM) coupled to Taylor expansion to compute the lower bound of the observed likelihood. Secondly, they introduce an adapted version of the model selection criterion, ICL, to select adequate missing model and number of classes.

In the interest of comprehension, we have decomposed the code from Frisch et al., 2022 into multiple Jupyter notebooks. This allows for an easy understanding of each function and the general methodology. For further information about the notebook structure, we refer to appendix B. The code is publicly available at:

https://github.com/AngelReyero/LBM-MNAR.

In section 2, we describe the model designed to accommodate informative missing values as an extension of the standard LBM. In section 3, we provide details on how to estimate the model and introduce a model selection criterion. Finally, in section 4, we present experimental results showcasing the accuracy of the estimation model on synthetic data and demonstrate the flexibility of the model assumptions through a real-world data case.

2 MODEL

In this section, we outline the assumptions made about the underlying data. We begin by recalling the known latent block model in section 2.1, followed by an explanation of the assumptions regarding the missingness mechanism in section 2.2. Finally, we integrate and summarize the previous in section 2.3.

2.1 Latent block model (LBM)

The latent block model is a model-based approach to find K row and L column clusters in the $n_1 \times n_2$ data matrix X, so that after reordering, we obtain $K \times L$ homogeneous blocks. In this article, we recall that the data matrix is binary.

We use two latent matrices to determine the row and column clusters. The indicator matrix Y, of size $n_1 \times K$, is the latent variable where Y_{ik} is 1 if the i-th row belongs to the k-th cluster. Similarly, for columns, we use the latent variable Z of size $n_2 \times L$. We denote Y_i the row indicator for the i-th individual and Z_j the column indicator for the j-th covariate.

We make the following independence assumptions in this model:

Assumption 1 (Independent rows and columns clusters)**.** The latent variables Y and Z are independent:

$$Y \perp\!\!\!\perp Z.$$

We observe that this does not imply the independence conditioned on the data matrix X.

Assumption 2 (I.I.D. row clusters)**.** Each row cluster Y_i is independent from the rest and it follows a multinomial distribution of parameter $\alpha := (\alpha_1, \dots, \alpha_K)$, where each $\alpha_k > 0$ and $\sum \alpha_k = 1$, so that $\mathbb{P}(Y_{ik} = 1) = \alpha_k$.

$$\forall i_1, i_2 \in \{1, \dots, n_1\} \quad Y_{i_1} \perp\!\!\!\perp Y_{i_2} \qquad \& \qquad Y_{i_1} \sim \mathcal{M}(1; \alpha).$$

Simmilarly but on the columns:

Assumption 3 (I.I.D. column clusters)**.** Each column cluster Z_j is independent from the rest and it follows a multinomial distribution of parameter $\beta := (\beta_1, \dots, \beta_L)$, where each $\beta_l > 0$ and $\sum \beta_l = 1$ so that $\mathbb{P}(Z_{jl} = 1) = \beta_l$.

$$\forall j_1, j_2 \in \{1, \dots, n_2\} \quad Z_{j_1} \perp\!\!\!\perp Z_{j_2} \qquad \& \qquad Z_{j_2} \sim \mathcal{M}(1; \beta).$$

Assumption 4 (I.I.D. block entries given column and row clusters)**.** Given the row and the column cluster, each entry is independent and distributed as a Bernoulli of parameter $\pi = (\pi_{kl}; k \in [K], l \in [L])$.

$$\forall i_1, i_2 \in [n_1], j_1, j_2 \in [n_2], \quad X_{i_1 j_1} | Y_{i_1}, Z_{j_1} \perp\!\!\!\perp X_{i_2 j_2} | Y_{i_2}, Z_{j_2}$$

$$\mathbb{P}\left(X_{ij} = 1 \middle| Y_{ik} Z_{jl} = 1; \pi\right) = \pi_{kl}.$$

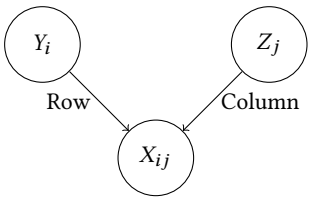
Then, we observe that the parameters of the model are $\theta := (\alpha, \beta, \pi)$, and that using the previous assumptions we can rewrite the distribution of X as

$$\begin{aligned}
f(X; \theta) &= \sum_{(Y,Z) \in I \times J} f(X|Y, Z; \theta) f(Y, Z; \theta) \\
&= \sum_{(Y,Z) \in I \times J} f(X|Y, Z; \theta) f(Y; \theta) f(Z; \theta) \\
&= \sum_{(Y,Z) \in I \times J} \prod_{ijkl} f\left(X_{ij} \middle| Y_{ik} Z_{kl} = 1; \pi\right)^{Y_{ik} Z_{jl}} \prod_{ik} \alpha_k^{Y_{ik}} \prod_{jl} \beta_l^{Z_j l} \\
&= \sum_{(Y,Z) \in I \times J} \prod_{ijkl} \left(\pi_{kl}^{X_{ij}} (1 - \pi_{kl})^{1 - X_{ij}}\right)^{Y_{ik} Z_{jl}} \prod_{ik} \alpha_k^{Y_{ik}} \prod_{jl} \beta_l^{Z_j l},
\end{aligned}$$

where I and J denote all the possible partitions of rows and columns into K and L groups respectively. It is summarized in the [Figure 1].

2.2 Missingness mechanism

In many co-clustering applications, such as collaborative filtering, which involves users providing ratings for a set of items and is frequently used in recommendation systems, it is common for users not to rate all the products. Consequently, dealing with missing values becomes a necessity. However, there is a scarcity of methods adapted to this type of data. Moreover, the scarcity is even more pronounced under the assumption of non-ignorable missing data, as is the case, for example, in collaborative filtering (refer to Marlin et al., 2012). In this section, we will present a missingness model that will be incorporated into the LBM, as discussed in section 2.3.



$$\forall i, Y_i \sim \mathcal{M}(1;\alpha)$$
$$\forall j, Z_j \sim \mathcal{M}(1;\beta)$$
$$\forall i,j, X_{ij}|Y_{ik}=1, Z_{jl}=1 \sim \mathcal{B}(\pi_{kl})$$

Figure 1: Assumptions of the latent block model (LBM).

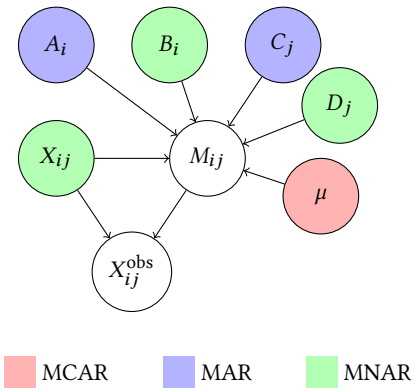


Figure 2: Latent variables of the missingness model.

We recall that X^{obs} denote the partially observed data matrix and M denotes the mask. If $M_{ij} = 0$ then the ij-th entry is not observed $X_{ij} = \text{NA}$.

In this section, we introduce a nested missingness model used in Frisch et al., 2022. This hierarchical model transitions from a completely missingness mechanism to a more nuanced non-ignorable one.

First, in the context of *Missing Completely At Random* (MCAR), a fixed parameter μ represents the missingness propensity for all cases. Then, to account for factors such as varying user response rates or the tendency to rate more expensive products, we extend the model to *Missing At Random* (MAR) by introducing latent variables A and C for rows and columns, respectively. Furthermore, to capture the idea that the probability of rating an article depends on various factors, including the current opinion of the article, we introduce two additional latent variables, B and D, for rows and columns. These variables are employed differently in the model based on the underlying value of X_{ij}, giving an *Missing Not At Random* (MNAR) model. Importantly, the flexibility provided by these variables has been demonstrated in both synthetic and real data experiments (refer to Frisch et al., 2022). This relations are graphically summarized in [Figure 2].

For simplicity, we assume the independence and Gaussianity of the missingness latent variables:

Assumption 5 (I.I.D. missingness latent variables)**.** All the A, B, C and D are independent and distributed as

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

Finally, we assume that the missingness of each case is independent of the rest and follows a Bernoulli distribution based on combinations of the preceding latent variables as follows:

Assumption 6 (Missingness distribution)**.** We have

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

independent from the other and where

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

and $\text{expit}(x) = \frac{1}{1+\exp(-x)}$.

With this model, we can consider not only the probability of an item being rated or users' propensity to express their opinion but also their underlying opinion on the product, influencing their decision to rate or not.

2.3 Combining the LBM with the missingness mechanism

We recall that X_{ij}^{obs} is either NA whenever $M_{ij} = 0$ and X_{ij} otherwise. Then, in each case, the observed value can be either 0, 1 or NA. Moreover, we note that it is possible to express the entire model using the preceding latent variables (Y, Z, A, B, C and D), eliminating the necessity of the mask matrix, through a categorical distribution:

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

where

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

and

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

In both terms, we note that the initial component arises from the probability assigned by the LBM to be either 0 or 1, while the second term accounts for the probability of missingness. Summarizing the parameters of our model, we have that the latent variables depend on $\theta = (\alpha, \beta, \pi, \mu, \sigma_A^2, \sigma_B^2, \sigma_C^2, \sigma_D^2)$, having a total of $K + L + K \times L + 5$.

3 MODEL ESTIMATION

As usual, to estimate the parameters of the model, our objective is to maximize the observed log-likelihood. To achieve this, we recognize that by marginalizing the latent variables, we obtain

$$p\left(X^{\text{obs}};\theta\right) = \sum_{Y,Z} \int_{ABCD} p\left(X^{\text{obs}}, Y, Z, A, B, C, D; \theta\right).$$

Unfortunately, this problem is intractable. As an alternative, we may use the EM algorithm to avoid the need for explicit computation. This algorithm iterates the following steps:

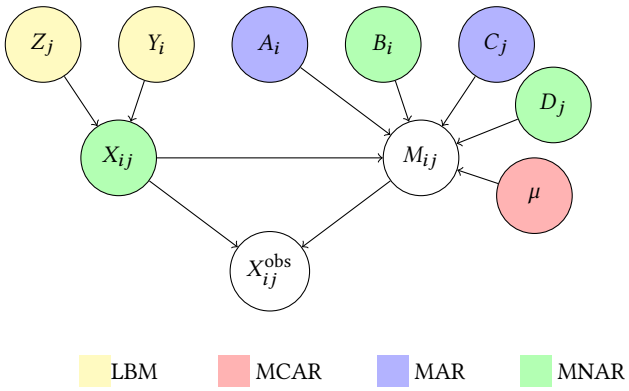


Figure 3: LBM adapted to the MNAR mechanism.

Expectation: It consists on computing

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

Maximization: We seek to find the parameters that maximize the previous expectation:

$$\theta^{t+1} = \text{argmax}\left(Q(\theta|\theta^t)\right).$$

However, the expectation step is computationally infeasible due to its complexity. Alternatives, such as Stochastic Expectation Maximization (SEM) (refer to Brault et al., 2014), which involves Monte Carlo sampling, are not suitable for our context due to scalability issues. In this article, we adopt a variational approach.

3.1 Variational expectation maximization (VEM)

We start by introducing $q(\cdot)$, which is the variational distribution over the latent variables A, B, C, D, Y and Z. Then, we remark that after adding and substracting the entropy of the variation distribution to the observed log-likelihood we have

$$\log p\left(X^{\text{obs}}; \theta\right) = \mathcal{J}(q, \theta) + \text{KL}\left(q(\cdot)\middle\|p(\cdot|X^{\text{obs}}; \theta)\right), \tag{4}$$

where KL accounts for the Kullback-Leibler divergence and $\mathcal{J}$ refers to the *free energy*, which is given by

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

where $\mathcal{H}$ refers to the entropy.

From (4), we observe that in order to maximize the $\mathcal{J}$ criterion, we need to minimise the discrepancy between $q(\cdot)$ is $p(\cdot|X^{\text{obs}}; \theta)$ given by the Kullback-Leibler divergence:

$$\text{KL}\left(q(\cdot)\middle\|p(\cdot|X^{\text{obs}}; \theta)\right).$$

Then, we would like the variational distribution as $p(\cdot|X^{\text{obs}}; \theta)$ almost everywhere. From this, given that the evidence is given by

$$\log p\left(X^{\text{obs}}; \theta\right),$$

we also call $\mathcal{J}(q, \theta)$ the Evidence Lower BOund (ELBO).

Unfortunately, optimizing this problem over the set of all distributions is not feasible. Therefore, we need to constrain the selection of the posterior distribution of the latent variables to an specific subset:

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

Moreover, using the *mean field approximation*, we assume independence among these distributions for feasibility. This allows us to obtain the following factorized version of the variational distribution:

$$\begin{aligned}
q_\gamma = & \prod_{i=1}^{n_1} \mathcal{M}\left(1; \tau_i^{(Y)}\right) \times \prod_{j=1}^{n_2} \mathcal{M}\left(1; \tau_j^{(Z)}\right) \\
& \times \prod_{i=1}^{n_1} \mathcal{N}\left(\nu_i^{(A)}, \rho_i^{(A)}\right) \times \prod_{i=1}^{n_1} \mathcal{N}\left(\nu_i^{(B)}, \rho_i^{(B)}\right) \\
& \times \prod_{j=1}^{n_2} \mathcal{N}\left(\nu_j^{(C)}, \rho_j^{(C)}\right) \times \prod_{j=1}^{n_2} \mathcal{N}\left(\nu_j^{(D)}, \rho_j^{(D)}\right),
\end{aligned} \tag{5}$$

where we denote q_γ the restriction of the variational distribution to the previous assumptions and its parameters

$$\gamma = \left(\tau^{(Y)}, \tau^{(Z)}, \nu^{(A)}, \rho^{(A)}, \nu^{(B)}, \rho^{(B)}, \nu^{(C)}, \rho^{(C)}, \nu^{(D)}, \rho^{(D)}\right).$$

Finally, we use the previous to construct a two-step iterative algorithm. It begins by estimating the parameters γ through the optimization of the free energy, constrained to the already explained posterior distributions given by:

$$\mathcal{J}(q_\gamma, \theta) = \mathcal{H}(q_\gamma) + \mathbb{E}_{q_\gamma}\left[\log p\left(X^{\text{obs}}, Y, Z, A, B, C, D\right)\right]. \tag{6}$$

Once computed, we take the maximization step for the parameters of the model:

$$\widehat{\theta} \in \underset{\theta}{\text{argmax}}\left(\max_{\gamma} \mathcal{J}(q_\gamma, \theta)\right).$$

This two-step iterative algorithm is shown in algorithm 1.

3.2 Computation of the variational criterion

We note from algorithm 1 that we need to compute the variational criterion $\mathcal{J}$. To do so, we are going to develop both terms shown in (6).

Entropy term: To solve this first term, we leverage the independence of the posterior variational distributions, yielding the factorized form as given in (5). Exploiting this factorization, we can then apply the additivity of entropy terms for independent variables.

Algorithm 1: VEM for LBM with MNAR

Data: The incomplete data X^{obs} and the number of rows and columns clusters K and L.
Result: The model θ and variational γ parameters.
Initialize the parameters.
while *not stopping criterion satisfied* **do**
  VE-step: we update the variational parameters:
$$\gamma^{t+1} \in \underset{\gamma}{\text{argmax}} \mathcal{J}(q_\gamma, \theta^t).$$
  M-step: we update the model parameters:
$$\theta^{t+1} \in \underset{\theta}{\text{argmax}} \mathcal{J}(q_{\gamma^{t+1}}, \theta).$$

Additionally, we utilize the entropy results derived from proposition A.1 and proposition A.2. Then, we have

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

Expectation over the variation distribution of the complete likelihood term: For the second term, we begin by applying the independence assumption on the latent variables within the missingness mechanism, as specified in assumption 5. Therefore we have

$$\begin{aligned}\mathbb{E}_{q_\gamma}\left[\log p\left(X^{\text{obs}}, Y, Z, A, B, C, D\right)\right] &= \mathbb{E}_{q_\gamma}\left[\log p(Y)\right] \\ &+ \mathbb{E}_{q_\gamma}\left[\log p(Z)\right] + \mathbb{E}_{q_\gamma}\left[\log p(A)\right] + \mathbb{E}_{q_\gamma}\left[\log p(B)\right] \\ &+ \mathbb{E}_{q_\gamma}\left[\log p(C)\right] + \mathbb{E}_{q_\gamma}\left[\log p(D)\right] \\ &+ \mathbb{E}_{q_\gamma}\left[\log p\left(X^{\text{obs}} \middle| Y, Z, A, B, C, D\right)\right]. \end{aligned} \tag{7}$$

To compute this term, we refer to the direct computation of all the terms in appendix A.1, except for the last one. By primarily using the expression provided in (1) to avoid the need for the mask matrix, we observe that the last term is given by:

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

where the p_0 ad p_1 are defined in (2) and (3). This term is not explicitly computable, then assuming a small variance term we will use the second-order Taylor series with independent random variables described in proposition A.3.

Maximization: As mentioned earlier, we confront two maximization steps for the criterion $\mathcal{J}(q_\gamma, \theta)$. The first involves optimizing with respect to γ (VE-Step) , and the second entails optimization with respect to θ (M-Step). Given the absence of a formal and explicit solution for these problems, an implementation of the L-BFGS optimization algorithm is employed. To compute the gradients essential for this optimization problems (approximated by the Taylor developments using proposition A.3), the Autograd submodule from PyTorch is deployed. Given the computational intensity of such calculus, the use of this tool is indispensable, particularly for its ability to leverage GPU capabilities.

Initialization: Addressing VEM algorithms presents a significant challenge in achieving convergence to global maxima and overcoming initialization dependence. In the paper, the authors tackle this problem by implementing the parameters of the Stochastic Block Model, linked to LBM for graphs. Such parameters are identified by using double spectral clustering on rows and columns on the similarity matrices XX^t and X^tX to initialize the algorithm. While this method may not be suited for MNAR data, its effectiveness is anticipated when missingness is not predominant. Due to the inability to directly initialize missingness parameters using this approach, the authors opt for a random initialization strategy.

3.3 Integrated completed likelihood criterion (ICL)

Model selection is challenging in this context due to the need to compute two different numbers of groups, and other approaches such as AIC or BIC are not applicable because calculating the maximized likelihood is not feasible. Fortunately, the method ICL extended to this context by Brault et al., 2014 is computable.

In this section, we first introduce the log-integrated completed likelihood. Next, we state a proposition regarding an asymptotic approximation of it, and finally, we provide a practical approximation.

The integrated completed likelihood for a given number of K row classes and L column classes is

$$\log \int p\left(X, Y, Z | \theta; K, L\right) p\left(\theta; K, L\right) d\theta, \tag{8}$$

where $p\left(\theta; K, L\right)$ is the prior distribution of the parameters. We note that as it takes the missing values into account, it is focused on a clustering point of view. Frisch et al., 2022 propose to take independent InverseGamma(1,1) distributions as priors for the $\sigma_A^2, \sigma_B^2, \sigma_C^2$ and σ_D^2 and for α and β the non-informative Dirichlet distribution priors as in Keribin et al., 2012.

Proposition 3.1 (Asymptotic ICL). *An asymptotic expansion of the log-integrated completed likelihood* (8) *up to a constant, is given by*

$$\begin{aligned}ICL^{\infty}(K, L) = &\max_{\theta, Y, Z, A, B, C, D} \log p\left(X^{\text{obs}}, Y, Z, A, B, C, D; \theta\right) \\ &- \frac{K-1}{2}\log(n_1) - \frac{L-1}{2}\log(n_2) \\ &- \frac{KL+1}{2}\log(n_1 n_2) - \log(n_1 n_2).\end{aligned} \tag{9}$$

This proof of this proposition mainly uses a Taylor expansion and an Stirling expansion, and it can be find in the Appendix C from paper Frisch et al., 2022.

Nevertheless, the first term cannot be computed. To derive a practical criterion, we will use the expectation of the log-likelihood under the variational posterior. Using (6), this expectation can be computed as the difference between the Evidence Lower Bound (ELBO) provided by the Variational Expectation-Maximization (VEM) $\mathcal{J}\left(q_{\widehat{\gamma}}, \widehat{\theta}\right)$ and the entropy of the variational distribution $q_{\widehat{\gamma}}$, where

$$\left(q_{\widehat{\gamma}}, \widehat{\theta}\right) \in \underset{\gamma,\theta}{\operatorname{argmax}} \mathcal{J}(q_\gamma, \theta).$$

This is expressed as:

$$\mathcal{J}\left(q_{\widehat{\gamma}}, \widehat{\theta}\right) - \mathcal{H}\left(q_{\widehat{\gamma}}\right) - \frac{K-1}{2}\log(n_1) - \frac{L-1}{2}\log(n_2) - \frac{KL+1}{2}\log(n_1 n_2) - \log(n_1 n_2).$$

Following the same reasoning we can construct an ICL for the MAR described by the section 2.2 using the following asymptotic development of the log-integrated completed likelihood:

$$ICL^{\infty}_{\text{MAR}}(K,L) = \max_{\theta,Y,Z,A,C} \log p\left(X^{\text{obs}}, Y, Z, A, C; \theta\right) - \frac{K-1}{2}\log(n_1) - \frac{L-1}{2}\log(n_2) - \frac{KL+1}{2}\log(n_1 n_2) - \log(n_1 n_2). \quad (10)$$

4 RESULTS

In this section, we begin by presenting the results obtained from synthetic data. This initial step ensures certainty in the methodology employed to adapt to the underlying model. Subsequently, we transition to a real-world case involving votes in the French parliament. This practical application enables us to assess the adaptability and flexibility of the assumed underlying model.

4.1 Results on synthetic data

Simulated data serves as a crucial tool for testing the algorithm's capability to generate consistent outcomes within controlled environments. However, when dealing with co-clustering problems, there is a necessity to redefine certain metrics used to assess performance. Initially, the indexes of row (resp. column) clusters are known up to a permutation [K] (resp. [L]). To account for this permutation space when measuring discrepancy, the following formula is employed:

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

where Ω_1 (resp. Ω_2) represents the set over all permutations of [K] (resp. [L]).

Moreover, the use of Bayes risk is not suitable in this context. The approximation of these parameters necessitates a Monte Carlo averaging across a substantial number of data matrices, rendering the process computationally expensive. In fact, estimating the Bayes risk on two data matrices generated with the same distribution may lead to very different results. The authors use the conditioned Bayes risk on observed data matrices presented in Lomet et al., 2012, to control difficulty of clustering on simulated data matrices and tackle such variability:

$$r_{item}(\widehat{Y}, \widehat{Z}) = \mathbb{E}[l_{item}(Y, Z, \widehat{Y}, \widehat{Z}) | X^{\text{obs}}]$$
$$(\widehat{Y}, \widehat{Z}) = \underset{Y,Z}{\operatorname{argmax}} \Sigma_{ij} p(Y_i, Z_j | X^{\text{obs}}).$$

As the term $p(Y, Z|X^{\text{obs}})$ is intractable, they compute the expectation as the average of a Gibbs sampler of $(Y, Z|X^{\text{obs}})$. Additionally, a notable distinction between standard clustering and co-clustering arises. While augmenting the size of a data matrix typically heightens the difficulty of standard clustering tasks, in the realm of co-clustering, an increase in data matrix size paradoxically diminishes task difficulty as it can be seen as having more *discriminating* information.

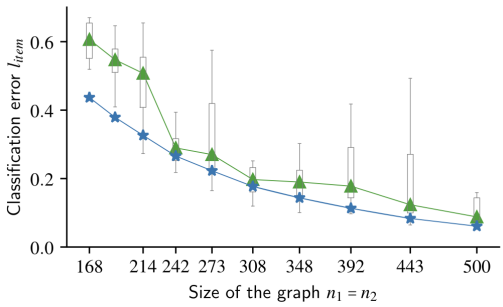


Figure 4: Comparison of classification errors concerning the size of the data matrix

The conditional Bayes risk is represented in blue, while the results from the paper are highlighted in green.

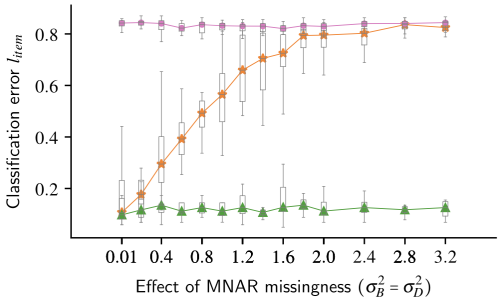


Figure 5: Analysis of classification errors as the MNAR missingness model effect intensifies

Categorical LBM is depicted in pink, the MAR model in orange, and the MNAR model in green.

Synthetic data generation: Synthetic data sets have been generated using various sizes and difficulty levels from a LBM with a MNAR missingness model. These data sets are produced for the

LBM, featuring three row and column classes:

$$\alpha = \beta = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}, \pi = \begin{bmatrix} \epsilon & \epsilon & 1-\epsilon \\ \epsilon & 1-\epsilon & 1-\epsilon \\ 1-\epsilon & 1-\epsilon & \epsilon \end{bmatrix},$$

where ϵ refers to the difficulty of the clustering task and $\mu = \sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2 = 1$ which incorporate a 35% rate of global missingness. In practice, the entire estimation process has been repeated 20 times in order to study the variability induced by the initializations.

Class prediction: To facilitate class prediction, they fine-tuned the parameter $\epsilon = 5\%$ through a trial-and-error process. Subsequently, an initial data matrix of size $n_1 = n_2 = 500$ was generated, and its size was progressively reduced to intensify the task difficulty. Indeed, as anticipated, more favorable outcomes in terms of the selection of the number of classes are achieved when utilizing larger matrices with well-separated classes, and Bayes risk decreases [Figure 4].

Missingness models: To compare the performance of missingness models, datasets were generated for LBM with $\epsilon = 12\%$, $n_1 = n_2 = 100$ and 35% of missing data. The values of σ_B^2 and σ_D^2 were varied to model the MNAR missingness effect. For each generated data matrix, LBM models were trained with Missing at Random (MAR), Missing Not at Random (MNAR), and Categorical LBM (refer to Keribin et al., 2015). The classification error l_{item} with respect to $\sigma_B^2 = \sigma_D^2$ illustrates that the Categorical Latent Block Model is unsuitable for handling missing values. Furthermore, the performance of the Missing at Random (MAR) model declines with an escalation in the effect of MNAR missingness model. In contrast, the MNAR model demonstrates effective adaptation, as evidenced by a consistent classification error [Figure 5]. This observation underscores the significance of accounting for informative missingness, as neglecting it introduces substantial biases into estimation.

In the pursuit of selecting an appropriate missingness model, the Integrated Completed Likelihood (ICL) criterion was employed. As discussed in section 3.3, it was possible to develop an asymptotic ICL for each missingness mechanism for which a practical criterion can be developed. Assuming known values of K and L, the results of both methods assessing missingness were estimated and compared. The MNAR model was consistently selected, exhibiting better ICL scores across various scenarios, emphasizing its appropriateness for capturing the underlying missingness patterns.

4.2 Real data experiments

The paper focuses on a set of three data matrices that characterize voting records from the lower house of the French parliament. The primary data set, labeled 'votes,' compiles the voting outcomes of 1256 ballots (columns) for 576 members of the parliament (MPs) (rows). Each vote is represented in a 3-level format, 1 for positive, -1 for negative, and 0 for missing values or abstention. Notably, level 0 accounts for 89% of the entire data set. Note that this does not comply with the assumption of non-predominance of MNAR data utilized for initialization. The 'deputes' data set, with size 576, contains detailed information about each MP, including their first name, family name, political group, and other relevant details. As we can observe in [Appendix: Figure 8], the majority of the hemicycle consists of centrist MPs (government) and right and left wing members represent a minority. Lastly, the 'texts' data set (size 1256) encompasses information related to voted ballots, such as the demandor and the name of amendments, providing additional context to the voting records.

They deploy the ICL criterion in order to select the number of row, column clusters and missingness model. The criterion selects $K = L = 14$ coupled to the MNAR model. The results displayed in the paper are consequently used for such a number of row and column cluster. Nevertheless, due to computation limitations, we implemented ourselves the reduced version proposed ($K = 3$ and $L = 5$).

In [Figure 6], the coherence of row clusters with political affiliations is evident. Examining row clusters, we observe centrist MPs predominantly in (clusters 6:13), right-wing MPs in (clusters 0, 1), and left-wing MPs in (clusters 2,3). Turning to column clusters, a notable distinction arises: centrist MPs consistently vote positively on original government demands (cluster A), while both right and left wings favor amendments proposed by minorities (cluster C). However, as anticipated, ideological divisions persist on various topics (clusters G:N). where a pronounced distinction between centrists and right-left MPs is highlighted in both cases. However, differentiation between right and left wings, and ballot clusters are less apparent. This observation could be attributed to the initialization dependence, which may lead to local maxima, or to the selection of K=3 and L=5, driven by computational limitations.

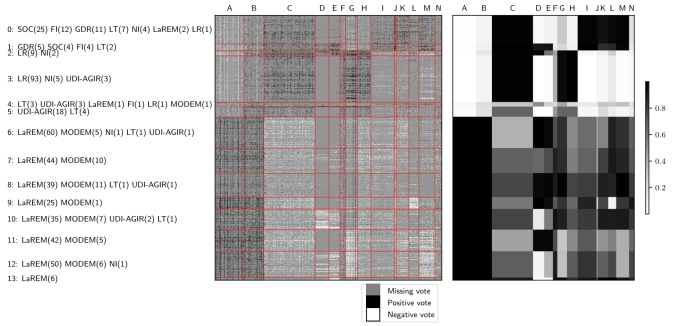


Figure 6: Reordered opinions according to row and column clusters

Left: Reordered votes matrix according to row and column clusters with K = L = 14. The red lines delineate cluster boundaries.
Right: Summary of the inferred opinions for ballots and MPs using the reordered version of π_{kl}.

In [Figure 7], the results showcase the maximum a posteriori estimates of MPs' propensities $(\nu_i^{(A)}, \nu_i^{(B)})$. In this visualization, the term 'A' represents the propensity to vote accounting for the MAR effect, while 'B' indicates the additional effect of casting a vote when supporting the resolution (MNAR effect). Notably, two distinct clusters emerge, particularly evident along the $\nu_i^{(B)}$ axis. The first cluster delineates the opposition, comprising MPs who

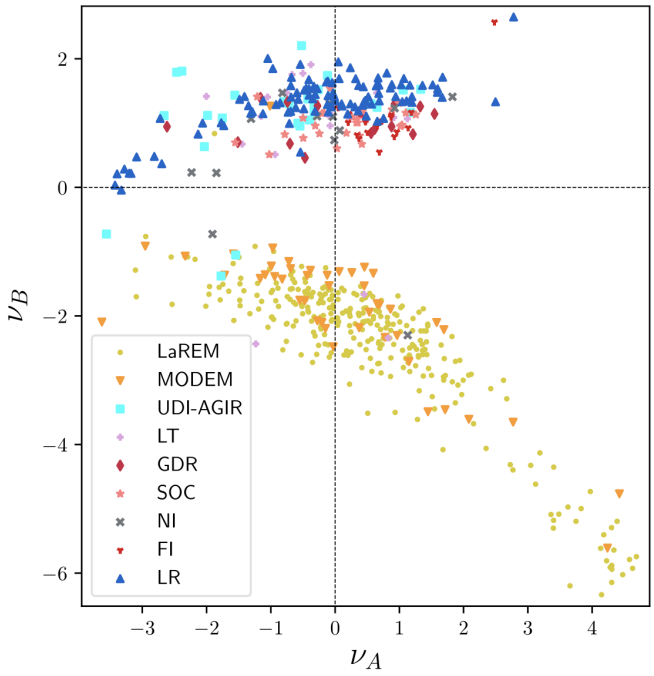


Figure 7: Maximum a posteriori estimates of the MPs propensities $(\nu_i^{(A)}, \nu_i^{(B)})$ for K=L=14

vote positively to advance the amendment, strategically aligning with the understanding that government supporters outnumber the opposition members. The second cluster includes members from 'LaREM' and 'MODEM,' indicating support for the government. This clear separation along the $\nu_i^{(B)}$ axis underscores the distinct voting behaviors and affiliations among the parliamentarians. Similar clusters are evident in the clusters depicted in [Appendix: Figure 9], reflecting our own implementation. The primary distinction lies in the propensities of the vote, $\nu_i^{(A)}$, which appear to exhibit a more pronounced spread.

5 CONCLUSION AND PERSPECTIVES

The primary objective of this article has been to identify the scarcity of co-clustering methods adapted to informative missingness. Additionally, the authors aim to demonstrate the advantages of employing such methods, proving their necessity in various contexts, such as collaborative filtering. Following a comprehensive literature review, the article delves into a detailed explanation of the methodology and assumptions outlined in Frisch et al., 2022, focusing specifically on the case of binary data.

We begin by introducing a flexible missingness model designed to complement the standard latent block model, nested within the missingness assumptions and capable of accommodating mechanisms for missing data not at random. Subsequently, we present the technical tools essential for estimating model parameters through a variational expectation-maximization approach. Moreover, we develop a model selection criterion based on the integrated completed likelihood.

Finally, we present experimental results, starting with a synthetic dataset that validates the estimation algorithm's ability to capture the underlying model. We extend our analysis to a real-world data case, demonstrating the model's flexibility in capturing missingness information.

Identifiability: While the estimation problem has been addressed in Frisch et al., 2022, it is also important to consider the identifiability of the model. Although experiments have indicated stable estimates, there is a lack of formal results in this regard. One potential approach to address this issue would be to expand upon the sufficient conditions of identifiability proposed for the binary LBM in Brault et al., 2014 to this MNAR model.

Data types: We have observed that, due to the relatively separation of this missingness model from the latent block model, there is potential to adapt the algorithm to data types more general than binary data. Multiple expansion avenues could be explored, such as adapting for ordinal data, functional data, or mixed data. To achieve this, we may refer to the data models explained in Biernacki et al., 2023.

Usability: We established a GitHub repository to facilitate a clear and systematic application of the model. Our efforts involved reproducing figures from the original article and incorporating a function to compute the ICL criterion eq. (9). However, we encountered various challenges during the code implementation process. The utilization of L-BFGS for training proved to be computationally intensive, necessitating access to a GPU. A noteworthy observation concerns the selection of parameters K and L. While the paper indicated that optimal values were $K = L = 14$, the absence of the corresponding code hindered our ability to find these optimal parameters to alternative datasets.

Local minima: Unfortunately, through extensive real-world data testing, it has come to our attention that distinguishing between MPs and ballots poses a considerable challenge. Although centrist MPs generally exhibit a tendency to coalesce, a notable trend emerges wherein right and left-wing affiliations are prone to being erroneously grouped together. This phenomenon may be attributed to a dependency on initial parameters, potentially resulting in convergence towards local minima. Another plausible explanation lies in the selection of parameter L. As previously observed, both left and right wing MPs often express affirmative votes for minority-requested amendments. A small L may be sensible to higher contrasts which may render more challenging to discern a significant difference between left and right-wing oppositions.

BIBLIOGRAPHY

Biernacki, C., Jacques, J., and Keribin, C. (2023). A survey on model-based co-clustering: high dimension and estimation challenges. *Journal of Classification*, 40(2):332–381.

Brault, V., Keribin, C., Celeux, G., and Govaert, G. (2014). Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25:1–16.

Corneli, M., Bouveyron, C., and Latouche, P. (2020). Co-clustering of ordinal data via latent continuous random variables and not missing at random entries. *Journal of Computational and Graphical Statistics*, 29(4):771–785.

Frisch, G., Leger, J.-B., and Grandvalet, Y. (2022). Learning from missing data with the binary latent block model. *Statistics and Computing*, 32(1):9.

Hernandez-Lobato, J. M., Houlsby, N., and Ghahramani, Z. (2014). Probabilistic matrix factorization with non-random missing data. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1512–1520, Bejing, China. PMLR.

Keribin, C., Brault, V., Celeux, G., and Govaert, G. (2012). Model selection for the binary latent block model. In *20th International Conference on Computational Statistics (COMPSTAT 2012)*, pages 379–390, Limassol, Cyprus.

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

Lomet, A., Govaert, G., and Grandvalet, Y. (2012). Design of artificial data tables for co-clustering analysis. *Universit de Technologie de Compigne, France.*

Marlin, B. M., Zemel, R. S., Roweis, S. T., and Slaney, M. (2011). Recommender systems, missing data and statistical model estimation. In *International Joint Conference on Artificial Intelligence.*

Marlin, B. M., Zemel, R. S., Roweis, S. T., and Slaney, M. (2012). Collaborative filtering and the missing at random assumption. *CoRR*, abs/1206.5267.

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

Selosse, M., Jacques, J., and Biernacki, C. (2020). Model-based co-clustering for mixed type data. *Computational Statistics Data Analysis*, 144:106866.

A FOUNDATIONAL MATHEMATICAL CONCEPTS

In this section, we present mathematical tools employed in the proofs throughout the article.

Proposition A.1 (Entropy of a Bernoulli). *Let $X \sim \mathcal{B}(\tau)$ be a Gaussian distribution. Then, its entropy is given by*

$$h(X) = -\tau\log(\tau).$$

Proof. We only need to notice that under the convention $0\log 0 = 0$ we have

$$h(X) = -\mathbb{E}\left[\log(p(X))\right] = -0\log 0 - \tau\log\tau.$$

□

Proposition A.2 (Entropy of a Gaussian). *Let $X \sim \mathcal{N}(\mu, \sigma)$ be a Gaussian distribution. Then, its entropy is given by*

$$h(X) = \frac{1}{2}\log\left(2\pi\sigma^2 e\right).$$

Proof.

$$\begin{aligned} h(X) &= -\mathbb{E}\left[\log p(X)\right] \\ &= -\mathbb{E}\left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)\right)\right] \\ &= \frac{1}{2}\log\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}\mathbb{E}\left[(X-\mu)^2\right] \\ &= \frac{1}{2}\log\left(2\pi\sigma^2 e\right). \end{aligned}$$

□

Now, we introduce a second-order Taylor series, which we utilize in the computation of the Evidence Lower BOund (ELBO), as demonstrated in section 3.2.

Proposition A.3. *For independent variables X and Y we have the following second-order Taylor serie approximation*

$$\begin{aligned} \mathbb{E}\left[f(X,Y)\right] \approx f(\mathbb{E}\left[X\right], \mathbb{E}\left[Y\right]) &+ \frac{1}{2}\mathrm{var}(X)\frac{\partial^2 f(\mathbb{E}\left[X\right], \mathbb{E}\left[Y\right])}{\partial x^2} \\ &+ \frac{1}{2}\mathrm{var}(Y)\frac{\partial^2 f(\mathbb{E}\left[X\right], \mathbb{E}\left[Y\right])}{\partial y^2}. \end{aligned}$$

A.1 Supplements on the computation of the variational criterion

In this section, we explicitly compute the terms in (7), excluding the last term, the development of which was previously explained in section 3.2.

We have for the Y latent variable

$$\mathbb{E}_{q_Y}\left[\log p(Y)\right] = \sum_{ik}\log(\alpha_k)\mathbb{E}_{q_Y}\left[Y_{ik}\right] = \sum_{ik}\log(\alpha_k)\tau_{ik}^{(Y)}.$$

Similarly, for the Z

$$\mathbb{E}_{q_Y}\left[\log p(Z)\right] = \sum_{jl}\log(\beta_l)\mathbb{E}_{q_Y}\left[Z_{jl}\right] = \sum_{jl}\log(\beta_l)\tau_{jl}^{(Z)}.$$

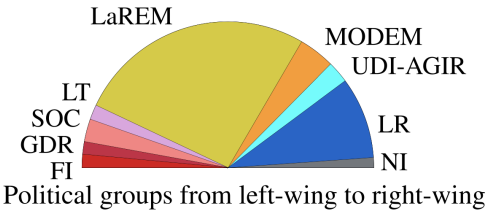
Now, following the same principle for the 4 Gaussian variables A, B, C and D we have

$$\begin{aligned} \mathbb{E}_{q_Y}\left[\log p(A)\right] &= -\frac{n_1}{2}\log(2\pi) - \frac{n_1}{2\pi}\log(\sigma_A^2) - \frac{1}{2\sigma_A^2}\sum_i \mathbb{E}_{q_Y}\left[A_i^2\right] \\ &= -\frac{n_1}{2}\log(2\pi) - \frac{n_1}{2\pi}\log(\sigma_A^2) \\ &\quad - \frac{1}{2\sigma_A^2}\sum_i\left(\left(\nu_i^{(A)}\right)^2 + \rho_i^{(A)}\right), \\ \mathbb{E}_{q_Y}\left[\log p(B)\right] &= -\frac{n_1}{2}\log(2\pi) - \frac{n_1}{2\pi}\log(\sigma_B^2) - \frac{1}{2\sigma_B^2}\sum_i \mathbb{E}_{q_Y}\left[B_i^2\right] \\ &= -\frac{n_1}{2}\log(2\pi) - \frac{n_1}{2\pi}\log(\sigma_B^2) \\ &\quad - \frac{1}{2\sigma_B^2}\sum_i\left(\left(\nu_i^{(B)}\right)^2 + \rho_i^{(B)}\right), \\ \mathbb{E}_{q_Y}\left[\log p(C)\right] &= -\frac{n_2}{2}\log(2\pi) - \frac{n_2}{2\pi}\log(\sigma_C^2) - \frac{1}{2\sigma_C^2}\sum_j \mathbb{E}_{q_Y}\left[C_j^2\right] \\ &= -\frac{n_2}{2}\log(2\pi) - \frac{n_2}{2\pi}\log(\sigma_C^2) \\ &\quad - \frac{1}{2\sigma_C^2}\sum_j\left(\left(\nu_j^{(C)}\right)^2 + \rho_j^{(C)}\right), \\ \mathbb{E}_{q_Y}\left[\log p(C)\right] &= -\frac{n_2}{2}\log(2\pi) - \frac{n_2}{2\pi}\log(\sigma_D^2) - \frac{1}{2\sigma_D^2}\sum_j \mathbb{E}_{q_Y}\left[D_j^2\right] \\ &= -\frac{n_2}{2}\log(2\pi) - \frac{n_2}{2\pi}\log(\sigma_D^2) \\ &\quad - \frac{1}{2\sigma_D^2}\sum_j\left(\left(\nu_j^{(D)}\right)^2 + \rho_j^{(D)}\right). \end{aligned}$$

B GITHUB IMPLEMENTATION

We have established a GitHub repository containing computational applications built upon the groundwork presented in [gfrisch/LBM-MNAR]. In this repository, we have curated five notebooks, each dedicated to exploring different aspects of the code version of the LBM for MNAR scenarios. These notebooks serve as comprehensive guides for users interested in understanding and utilizing the model implementation.

1.1-Dummy_training.ipynb: This notebook initiates the training process for the Variational Expectation-Maximization (VEM) model, as proposed in the referenced article. It provides a preliminary exploration with a focus on one iteration of the VEM algorithm,



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

Figure 8: Representation of the hemicycle of the French National Assembly political groups

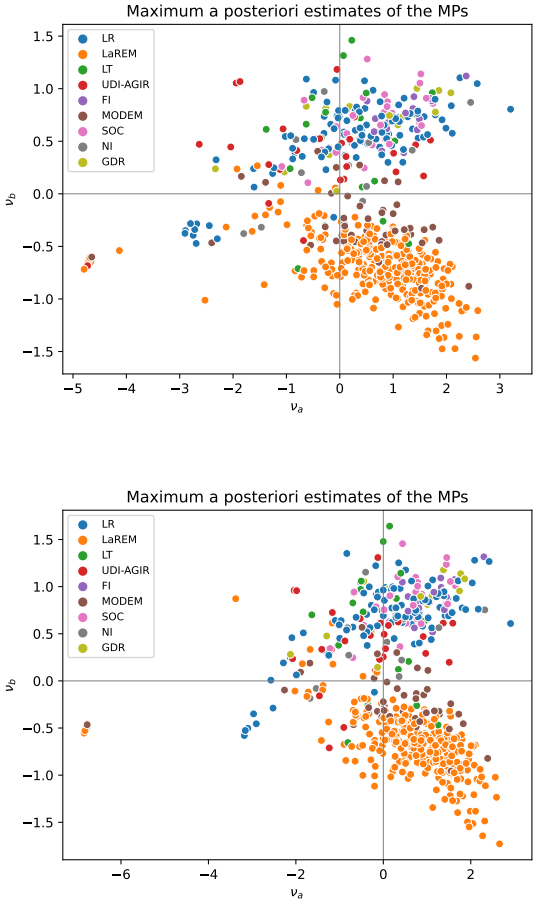


Figure 9: Maximum a posteriori estimates of the MPs propensities from our implementation for $K = 3$ and $L = 5$

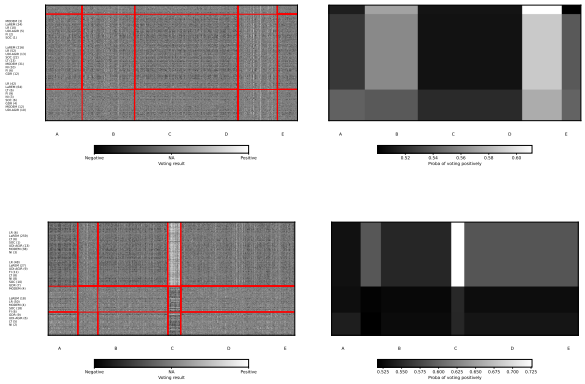


Figure 10: Reordered opinions according to row and column clusters from our implementation $K = 3$ and $L = 5$

offering insights into the model's early learning dynamics.

1.2-Model_LBM_MNAR.ipynb: Provides an overview of the computation of the criterion

2-Train.ipynb: Designed to train the entire model on the parliament dataset. Given the potential time-intensive nature of this procedure, we have saved the parameters in the file named: "trained_parameters.yaml", so that, computing this step is not required to continue exploring the notebooks.

3-Figure_creation.ipynb: Specifically crafted for creating figures 12, 17, and 18 from the article. Running this notebook does not necessitate the execution of the entire model, as parameters are loaded from the yaml file.

4-ICL.ipynb: Designed to compute the ICL criterion associated to the trained model from 2-Train.ipynb

Given the potential computational expense of training, we recommend utilizing a GPU. To specify the device, the device argument can be employed, with 'cuda' recommended for general (use or 'mps' for Mac). The default configuration sets the number of row classes to 3 and column classes to 5.

C AUTHOR'S CONTRIBUTIONS

- **Angel Reyero Lobo**
  - Introduction
  - Model
  - Model estimation
  - Conclusion
- **Laura Fuentes Vicente**
  - GitHub repository
  - Model estimation: Maximization and Initialization
  - Results