# Learning from missing data with the binary latent block model

**Samara Ndiaye**[1] **Aymane Masrour**[1]

[1]Université Paris Saclay, France

## ABSTRACT

This report reviews the article "Learning from Missing Data with the Binary Latent Block Model" by Frisch, Leger, and Grandvalet. The paper addresses the challenge of co-clustering binary data matrices with missing entries that are Missing Not At Random (MNAR). It introduces an extension of the Latent Block Model (LBM) to incorporate MNAR mechanisms, leveraging a variational Expectation-Maximization (VEM) algorithm for parameter inference and proposing an Integrated Completed Likelihood (ICL) criterion for model selection. The approach is validated through simulation studies and applied to real-world data from the French Parliament's voting records, revealing interpretable patterns in MPs' voting behaviors and the structure of proposed resolutions. Our review situates the article within the broader literature on co-clustering and missing data, assesses the originality of its contributions, and examines its proofs, algorithms, and implementations.

**Keywords :** Informative Missing Data, Binary Data, Co-clustering, latent block model (LBM), variational expectation-maximization (VEM), integrated completed likelihood (ICL).

## INTRODUCTION

In unsupervised learning, the goal is to uncover hidden structures in data. Clustering allows for the grouping of individuals with similar characteristics, such as identifying customers with the same behavior in a market or segmenting different regions in an image. An extension of clustering is *co-clustering*, which simultaneously groups both rows and columns of a data matrix. The objective is to identify clusters along with the covariates that characterize them.

Co-clustering is a data analysis technique aimed at simultaneously clustering rows and columns of a data matrix, uncovering local patterns by focusing on submatrices within the data. This contrasts with standard clustering techniques that identify global patterns across all rows or columns and might be less effective in high-dimensional datasets. In particular, co-clustering can be viewed as a dimensionality reduction technique, as it reduces the complexity of the data by focusing on specific regions of interest. Additionally, this localized focus makes co-clustering particularly well-suited for sparse data. Many real-world datasets, such as user-item matrices in recommender systems, are sparse, with most entries being missing or zero. The ability of co-clustering to effectively handle such sparsity is highly relevant to the work at hand, which aims to provide a more robust way of dealing with missing values. Unlike traditional approaches that treat missing values as zero—potentially biasing clustering results—the model proposed in this article addresses missingness directly.

Co-clustering, also known as biclustering, was initially introduced by John A. Hartigan in 1972 and later generalized in 2000 by Cheng and Church, who applied it to gene expression studies [2]. Since then, co-clustering has been employed in numerous applications, including:

- **Text analysis**: Grouping documents with similar content and identifying the keywords that characterize them [3].
- **Recommender systems**: Clustering users and their preferences to predict products they may like.

Various co-clustering methods exist, including spectral co-clustering, matrix factorization techniques, and Latent Block Models (LBM). The LBM assumes a probabilistic mixture model for rows and columns, introducing two latent variables to encode cluster memberships for rows and columns. This article builds on the LBM framework with a Bernoulli mixture model (Govaert and Nadif, 2008 [4]) and provides a significant contribution by incorporating missing data into the model.

The authors address *Missing Not At Random* (MNAR) data, where missingness itself conveys information about the unobserved values. Missing data can be categorized into three types [5]:

- **MCAR (Missing Completely At Random)**: The probability of missingness is independent of both observed and unobserved data.
- **MAR (Missing At Random)**: The probability of missingness depends on observed data but not on the missing values themselves.
- **MNAR (Missing Not At Random)**: The probability of missingness depends on the missing values themselves.

In MNAR scenarios, ignoring the informativeness of missingness can lead to significant biases in estimation. For example, in a survey where participants report their income, missingness may depend on whether the income is unusually high or low. This dependence makes missingness informative, and failing to account for it can distort the analysis.

Most clustering models addressing missing data adopt simplifying assumptions. For instance, Sportisse and Celeux (2023) [6] introduced a MNARz model where

missingness depends solely on the latent variable determining cluster membership. In contrast, the MNAR model proposed in this article captures missingness as a function of both the data value and the row and column indices, providing a more comprehensive framework.
Another essential contribution of the article is the parameter estimation approach. The Latent Block Model parameters are typically inferred using variants of the Expectation-Maximization (EM) algorithm [1]:

- **SEM (Stochastic EM)**: The E-step is replaced with direct sampling of latent variables conditioned on the data using Gibbs sampling.
- **VEM (Variational EM)**: A tractable parametric distribution is used to approximate the posterior distribution, maximizing a lower bound of the complete-data likelihood. While this method restricts the search space and can yield suboptimal solutions, it enables efficient optimization in high-dimensional settings.

The authors adopt the VEM approach in this article, which will be discussed further in the following sections. For model selection, the Integrated Completed Likelihood (ICL) criterion was preferred for its ability to balance model complexity and fit.
This introduction sets the stage for a deeper exploration of the MNAR Latent Block Model and its application, as detailed in the subsequent sections.

## Notations

Let us introduce the main notations used in this paper. Rows (individuals) are indexed by $i \in [n_1]$, columns (variables) by $j \in [n_2]$, row clusters by $k \in [K]$, and column clusters by $l \in [L]$, where $[n]$ denotes the set $\{1, 2, \ldots, n\}$.

- $X$: Full binary data matrix of size $n_1 \times n_2$.
- $X^{(o)}$: Observed data matrix, taking values in $\{0, 1, \text{NA}\}$, where NA indicates missing entries.
- $M$: Binary mask matrix, where $M_{ij} = 1$ if $X_{ij}$ is observed and $M_{ij} = 0$ otherwise.
- $Y$: Indicator matrix for row clusters, where $Y_{ik} = 1$ if row $i$ belongs to cluster $k$, and 0 otherwise.
- $Z$: Indicator matrix for column clusters, where $Z_{jl} = 1$ if column $j$ belongs to cluster $l$, and 0 otherwise.
- $\alpha$: Mixture proportions for row clusters, with $\alpha_k$ representing the proportion for cluster $k$.
- $\beta$: Mixture proportions for column clusters, with $\beta_l$ representing the proportion for cluster $l$.
- $\pi_{kl}$: Probability of observing a 1 in the block $(k, l)$ under the Latent Block Model.
- $A_i, B_i$: Latent variables capturing row-level effects in the missingness model.
- $C_j, D_j$: Latent variables capturing column-level effects in the missingness model.
- $\mu$: Global propensity parameter for missingness.

## MODEL FRAMEWORK

The major contribution of this article is the integration of missing data information into the Latent Block Model (LBM). This innovation allows the model to leverage the informativeness of missing data, especially in cases where the missingness is not random. A key application highlighted in the paper involves analyzing a dataset from the French Parliament, where the model successfully interprets the behavior of non-respondents, providing a robust co-clustering framework.

### THE BINARY LATENT BLOCK MODEL

The Binary Latent Block Model assumes a probabilistic co-clustering framework with two partitions, one for the rows and another for the columns. The key hypotheses underlying the LBM are:

- Row cluster $Y_i$ and column cluster $Z_i$ are independent a priori.
- Row cluster memberships $Y_i$ are i.i.d. multinomial random variables, $Y_i \sim \mathcal{M}(1, \boldsymbol{\alpha})$.
- Column cluster memberships $Z_j$ are i.i.d. multinomial random variables, $Z_j \sim \mathcal{M}(1, \boldsymbol{\beta})$.
- The entries $X_{ij}$ are conditionally independent given the cluster memberships $(Y_{ik}, Z_{jl})$ and follow a Bernoulli distribution:

$$X_{ij} \mid Y_{ik} = 1, Z_{jl} = 1 \sim \mathcal{B}(\pi_{kl}).$$

The joint probability of the observed data $X$ and the latent variables $(Y, Z)$ given the parameters $\theta = (\alpha, \beta, \pi)$ can be expressed as:

$$P(X, Y, Z \mid \theta) = P(Y \mid \alpha) P(Z \mid \beta) P(X \mid Y, Z, \pi), \quad (1)$$

where:

$$P(Y \mid \alpha) = \prod_{i,k} \alpha_k^{Y_{ik}},$$
$$P(Z \mid \beta) = \prod_{j,l} \beta_l^{Z_{jl}},$$
$$P(X \mid Y, Z, \pi) = \prod_{i,j,k,l} \phi(\pi_{kl}, X_{ij})^{Y_{ik} Z_{jl}},$$

and $\phi(\pi_{kl}, X_{ij}) = \pi_{kl}^{X_{ij}} (1 - \pi_{kl})^{1 - X_{ij}}$ is the Bernoulli density.
Summing over all possible cluster memberships $(Y, Z)$, the marginal likelihood of $X$ is:

$$P(X \mid \theta) = \sum_{Y,Z} \prod_{i,k} \alpha_k^{Y_{ik}} \prod_{j,l} \beta_l^{Z_{jl}} \prod_{i,j,k,l} \phi(\pi_{kl}, X_{ij})^{Y_{ik} Z_{jl}}. \quad (2)$$

This marginalization over all possible cluster assignments makes exact inference computationally challenging, necessitating approximation methods such as Variational EM.

## EXTENDED LBM WITH MNAR DATA

The extended Latent Block Model (LBM) with MNAR data is a generative probabilistic framework designed to model a partially observed data matrix $X^{(o)}$ of size $n_1 \times n_2$. This extension builds upon the standard LBM by incorporating a missingness mechanism, enabling the model to account for both the observed data and the missing entries.

### Extending the LBM to Include Missing Data

In real-world scenarios, the data matrix $X$ is rarely fully observed. Instead, we observe $X^{(o)}$, a partially observed version of $X$, with missing entries denoted as NA. A binary mask matrix $M$ is introduced, where:

$$M_{ij} = \begin{cases} 1 & \text{if } X_{ij} \text{ is observed,} \\ 0 & \text{if } X_{ij} \text{ is missing.} \end{cases} \tag{3}$$

The generative process of the extended LBM with MNAR data involves two steps:

1. Generate a complete data matrix $X$ using the standard LBM.

2. Apply the mask matrix $M$ to $X$ to produce the observed matrix $X^{(o)}$:

$$X_{ij}^{(o)} = \begin{cases} X_{ij} & \text{if } M_{ij} = 1, \\ \text{NA} & \text{if } M_{ij} = 0. \end{cases} \tag{4}$$

### Modeling the Missingness Mechanism

The mask matrix $M$ is generated based on a log-odds propensity model that accounts for three types of missingness [5]:

- **MCAR (Missing Completely At Random)**: The probability of missingness is independent of both the observed and unobserved data. In this case, the log-odds is constant:

$$P_{ij} = \mu. \tag{5}$$

- **MAR (Missing At Random)**: The probability of missingness depends on observed data but not on the unobserved data. Latent variables $A_i$ and $C_j$ capture row and column effects:

$$P_{ij} = \mu + A_i + C_j. \tag{6}$$

- **MNAR (Missing Not At Random)**: The probability of missingness depends on both observed and unobserved data. Additional latent variables $B_i$ and $D_j$ capture dependencies on the data values:

$$P_{ij} = \begin{cases} \mu + A_i + B_i + C_j + D_j & \text{if } X_{ij} = 1, \\ \mu + A_i - B_i + C_j - D_j & \text{if } X_{ij} = 0. \end{cases} \tag{7}$$

The latent variables $A_i$, $B_i$, $C_j$, and $D_j$ are modeled as independent Gaussian random variables with mean 0 and variances $\sigma_A^2$, $\sigma_B^2$, $\sigma_C^2$, and $\sigma_D^2$, respectively.
The missingness mechanism can be summarized as:

$$M_{ij} \mid A_i, B_i, C_j, D_j, X_{ij} \sim \mathcal{B}\left(\text{expit}(P_{ij})\right), \tag{8}$$

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

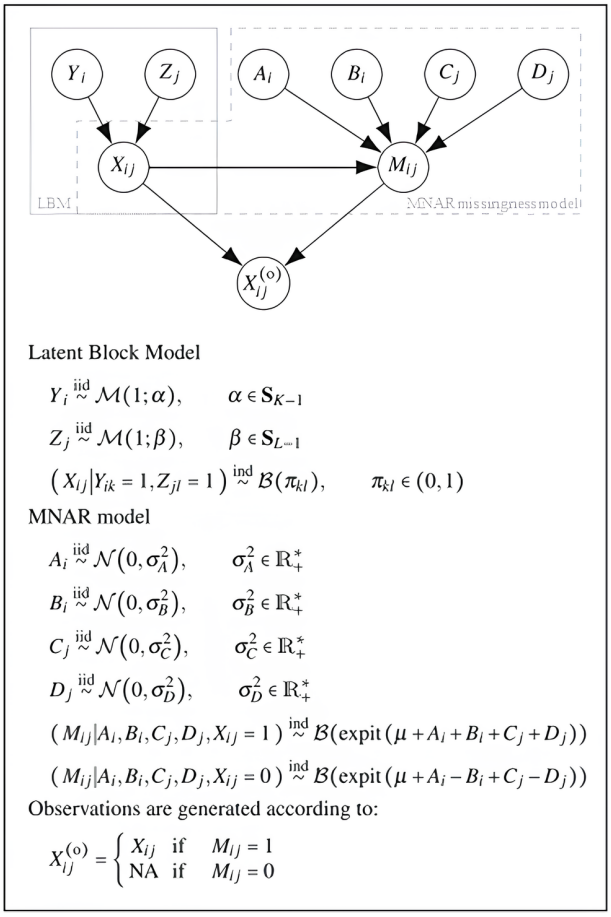


**Figure2 :** Summary of LBM with MNAR model

We can now formulate the LBM model with a MNAR missingness by only using the observed matrix $X^o$ and the latent variables.Knowing all the latent variables, $X^o$ follows a categorical distribution with 3 possibles values (0,1 ,NAN)

$$X_{ij}^o \mid Y_i, Z_j, A_i, B_i, C_j, D_j, \sim \mathcal{C}(\begin{bmatrix} 0 \\ 1 \\ NAN \end{bmatrix}, \begin{bmatrix} p_0 \\ p_1 \\ 1-p_0-p_1 \end{bmatrix})$$

Where

- $p_0 = (1-\pi_{kl})\text{expit}(\mu + A_i + C_j - B_i - D_j)$
- $p_1 = \pi_{kl}\text{expit}(\mu + A_i + C_j + B_i + D_j)$

The parameters to estimate for the LBM +MNAR model is now given by : $\theta = (\alpha, \beta, \pi, \mu, \sigma_A^2, \sigma_B^2, \sigma_C^2, \sigma_D^2)$

## MODEL INFERENCE

In this section, we describe the inference procedure for the Latent Block Model (LBM) under the MNAR (Missing Not At Random) assumption. The goal is to estimate the model parameters $\theta = (\alpha, \beta, \pi, \mu, \sigma_A^2, \sigma_B^2, \sigma_C^2, \sigma_D^2)$ using the observed data $X^{(o)}$, while accounting for the missing data $M$.

**Expectation-Maximization Framework**

We aim to maximize the marginal likelihood of the observed data:

$$\log p(X^{(o)}; \theta) = \log \sum_{Y,Z} \int_{ABCD} p(X^{(o)}, Y, Z, A, B, C, D; \theta)$$

Since direct maximization is intractable due to the latent variables $(Y, Z)$, we use the Expectation-Maximization (EM) algorithm. The EM alternates between:

- **E-step**: Compute the expected complete-data log-likelihood.

  $$Q(\theta \mid \theta_t) = \mathbb{E}\big[\log p(X^{(o)}, Y, Z, A, B, C, D; \theta)\big].$$

  where the expectation is under the conditional law of latent variables with respect to the observed data

  $$Y, Z, A, B, C, D \mid X^{(o)}, \theta_t$$

- **M-step**: Maximize this expectation with respect to $\theta$, updating the parameters sequentially:

  $$\theta_{t+1} = \arg\max_{\theta} Q(\theta \mid \theta_t).$$

**Variational Approximation**

Sampling directly from the conditional distribution of the latent variables can be computationally challenging. To address this, we introduce a variational approximation by considering $q(\cdot)$, a simpler and tractable parametric distribution designed to approximate the true posterior distribution. The goal is to optimize the parameters of $q(\cdot)$ by minimizing the Kullback-Leibler (KL) divergence between $q(\cdot)$ and the true posterior distribution which approximates the true posterior$p(Y, Z, A, B, C, D \mid X^{(o)}, \theta)$. .

**Evidence Lower Bound (ELBO)**

Decomposing the observed log-likelihood:

$$\log p(X^{(o)}; \theta) = \mathcal{J}(q, \theta) + \mathrm{KL}(q(\cdot) \parallel p(\cdot \mid X^{(o)}, \theta)),$$

where:

- $\mathcal{J}(q, \theta)$ is the ELBO given by :

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

- $\mathcal{H}(q)$: Entropy of $q(\cdot)$.
- $\mathrm{KL}(\cdot)$: Kullback-Leibler divergence between $q(\cdot)$ and the true posterior.

**Goal**: Maximize $\mathcal{J}(q, \theta)$, which indirectly maximizes $\log p(X^{(o)}; \theta)$.

**Factorized Variational Distribution**

To make optimization feasible, assume a mean-field approximation, i.e., independence among the latent variables:

$$q(Y, Z, A, B, C, D) = q(Y)q(Z)q(A)q(B)q(C)q(D).$$

as optimizing over the set of all distributions of this form is not feasible. Therefore, we need to constrain the selection.

**Constraints on $q(\cdot)$**

We introduce the distribution $q_\gamma$ defined using the following the following, distributions for each variable:

$$Y_i \mid X^{(o)} \sim \mathcal{M}(1; \tau_i^{(Y)}),$$

$$Z_j \mid X^{(o)} \sim \mathcal{M}(1; \tau_j^{(Z)}),$$

$$A_i \mid X^{(o)} \sim \mathcal{N}(\nu_i^{(A)}, \rho_i^{(A)}),$$

$$B_i \mid X^{(o)} \sim \mathcal{N}(\nu_i^{(B)}, \rho_i^{(B)}),$$

$$C_j \mid X^{(o)} \sim \mathcal{N}(\nu_j^{(C)}, \rho_j^{(C)}),$$

$$D_j \mid X^{(o)} \sim \mathcal{N}(\nu_j^{(D)}, \rho_j^{(D)}).$$

$$\gamma = (\tau_i^{(Y)}, \tau_j^{(Z)}, \nu_i^{(A)}, \rho_i^{(A)}, \nu_i^{(B)}, \rho_i^{(B)}, \nu_j^{(C)}, \rho_j^{(C)}, \nu_j^{(D)}, \rho_j^{(D)})$$

Defining the variational law following the previous descriptions allows for a simplified derivation of the ELBO term.

**Variational E-Step (VE-Step)**

We replace the expectation step in standard EM algorithm with a new step consisting in computing the expectation over a well chosen $q_\gamma(\cdot)$. Our aim is to choose $q_\gamma(\cdot)$ such that it minimizes the KL divergence:

$$\mathrm{KL}(q_\gamma(\cdot) \parallel p(\cdot \mid X^{(o)}, \theta)).$$

since this is fully minimized if $q_\gamma(\cdot) = p(\cdot \mid X^{(o)}, \theta)$.
In practice, this involves solving for the optimal $\tau, \nu, \rho$ for each latent variable's distribution. This insure that the considered law $q_\gamma(\cdot)$ is as close as possible in terms of KL divergence such that the ELBO calculated at this step is approximately close. Since $\log p(X^{(o)}; \theta)$ is independent of $\gamma$, this is equivalent to maximizing

$$\mathcal{J}(q_\gamma, \theta)$$

over the variational parameters $\gamma$. In the following we will focus on maximizing this quantity instead of minimizing the KL divergence.

**Maximization step (M-Step)**

Similar to EM algorithm we optimize parameters $\theta$ to maximize $\mathcal{J}(q_\gamma, \theta)$ given the parameters $\gamma$ obtained in the VE-step.

**Algorithm 1** VEM for LBM with MNAR

**Require:** The incomplete data $X^{\text{obs}}$ and the number of rows and columns clusters $K$ and $L$.
**Ensure:** The model $\theta$ and variational $\gamma$ parameters.
1: Initialize the parameters.
2: **while** not stopping criterion satisfied **do**
3: **VE-step**: Update the variational parameters:

$$\gamma_{t+1} \in \arg\max_{\gamma} \mathcal{J}(\gamma, \theta_t).$$

4: **M-step**: Update the model parameters:

$$\theta_{t+1} \in \arg\max_{\theta} \mathcal{J}(\gamma_{t+1}, \theta).$$

5: **end while**

**Practical Considerations**

Following the previous description of the VEM steps, it introduces a two step maximizations of the criterion $\mathcal{J}(q_\gamma, \theta)$ at each step $t$:

- Maximizing with respect to $\gamma$ given $\theta_t$ (**VE-Step**).
- Maximizing with respect to $\theta$ given $\gamma_{t+1}$(**M-Step**).

Given the absence of explicit solutions for these optimization problems, the L-BFGS optimization algorithm is employed. Gradients necessary for these computations are derived using PyTorch's `Autograd` module, which allows efficient handling of computational intensity and takes advantage of GPU capabilities.
For the computation of the criterion $\mathcal{J}(q_\gamma, \theta)$, approximations based on Taylor expansions were considered in the original article. Specifically, the softmax function used to define the propensity function was linearized up to the second order. This approximation simplifies the derivation, leveraging the expectation and variance of the latent variables. For detailed derivations, one can refer to the original article.

**Model Initialization**

The VEM algorithm is sensitive to initialization and may converge to local optima. To address this challenge, parameters related to the Stochastic Block Model (SBM), which is connected to the Latent Block Model (LBM) for graphs, are initialized using double spectral clustering on row and column similarity matrices ($XX^T$ and $X^TX$). This approach is effective when missing data is minimal.
However, this method cannot directly initialize parameters related to missingness. For these parameters, a random initialization strategy is used.

**Model Identifiability**

The identifiability of a model is very important because it allows to uniquely determine the values of the parameters given the observations . For the binary Latent Block Model Keribin and al [1] have shown that under the two following assumptions:

- $\forall k \in [K]\ \alpha_k > 0$ and all the components of the vector $\pi\alpha$ are distinct
- $\forall l \in [L]\ \beta_l > 0$ and all the components of the vector $\beta\pi$ are distinct

The Binary Latent Block Model is identifiable up to a permutation of row and column indexes. These conditions are not restrictive because the space of parameters $\theta = (\alpha, \beta; \pi)$ that does not verify these hypothesis has a Lebesgue measure of 0. For the Binary Latent Block model with the MNAR type of Missingness used in this paper the authors have not proved the Identifiability but pointed out the fact that their model belongs to the family of Linear Mixed Effect where the parameters $\theta = (\alpha, \beta, \pi, \mu, \sigma_A^2, \sigma_B^2, \sigma_C^2, \sigma_D^2)$ represent the global effect and the latent variables $Y, Z, A, B, C, D$ are the mixed effects .Some research in this direction may lead to the proof of the identifiability under some conditions of their model.

## Integrated Completed Likelihood Criterion (ICL)

Model selection in this context is challenging due to the need to compute the number of row and column groups. Standard approaches, such as AIC or BIC, are not applicable because calculating the maximized likelihood is infeasible. Fortunately, the Integrated Completed Likelihood (ICL) method can be extended and computed in this context.
In this section, we introduce the log-integrated completed likelihood, state a proposition regarding its asymptotic approximation, and provide a practical approximation.

**Log-Integrated Completed Likelihood**

The integrated completed likelihood for a given number of $K$ row clusters and $L$ column clusters is defined as:

$$\log \int p(X, Y, Z \mid \theta; K, L)\, p(\theta; K, L)\, d\theta,$$

where $p(\theta; K, L)$ is the prior distribution of the parameters. This criterion incorporates missing values and focuses on clustering from a probabilistic perspective.

**Prior Distributions**

For this method, the priors used include:

- Independent InverseGamma$(1, 1)$ distributions for $\sigma_A^2, \sigma_B^2, \sigma_C^2, \sigma_D^2$.
- Non-informative Dirichlet distribution priors for $\alpha$ and $\beta$.

**Asymptotic ICL**

**Proposition:** An asymptotic expansion of the log-integrated completed likelihood up to a constant is given by:

$$\begin{aligned}\mathrm{ICL}_\infty(K,L) = &\max_{\theta,Y,Z,A,B,C,D} \log p(X_{\mathrm{obs}},Y,Z,A,B,C,D;\theta)\\ &-\frac{K-1}{2}\log(n_1)-\frac{L-1}{2}\log(n_2)\\ &-\frac{KL+1}{2}\log(n_1n_2)-\log(n_1n_2)\end{aligned}$$

This result is derived using a Taylor expansion and Stirling's approximation. The detailed proof can be found in Appendix C of Frisch et al., 2022.

**Practical Approximation**

Since the first term in the expression is computationally intractable, we derive a practical criterion using the expectation of the log-likelihood under the variational posterior. Using the Evidence Lower Bound (ELBO) from Variational Expectation-Maximization (VEM), this expectation is computed as:

$$\begin{aligned}\mathrm{ICL}_{\mathrm{practical}} = &\mathcal{J}(q_{\gamma^*},\theta^*)-\mathcal{H}(q_{\gamma^*})\\ &-\frac{K-1}{2}\log(n_1)-\frac{L-1}{2}\log(n_2)\\ &-\frac{KL+1}{2}\log(n_1n_2)-\log(n_1n_2),\end{aligned}$$

where:

$$(q_{\gamma^*},\theta^*)\in\arg\max_{\gamma,\theta}\mathcal{J}(q_\gamma,\theta),$$

and $\mathcal{H}(q_{\gamma^*})$ represents the entropy of the variational distribution.

**ICL for MAR**

Using similar reasoning, the ICL for the MAR setting (as described in Section 2.2) can be expressed as:

$$\begin{aligned}\mathrm{ICL}_\infty^{\mathrm{MAR}}(K,L) = &\max_{\theta,Y,Z,A,C} \log p(X_{\mathrm{obs}},Y,Z,A,C;\theta)\\ &-\frac{K-1}{2}\log(n_1)-\frac{L-1}{2}\log(n_2)\\ &-\frac{KL+1}{2}\log(n_1n_2)-\log(n_1n_2).\end{aligned}$$

## Tests on synthetic data

The Latent Block Model (LBM) extended to handle Missing Not At Random (MNAR) data represents a significant step forward in co-clustering methodologies. This model explicitly integrates missingness mechanisms into its probabilistic framework, offering a robust solution for datasets where missing data carries informative patterns. This review focuses on the results obtained from synthetic data experiments, which are pivotal for validating the model's capacity to recover latent structures while addressing the challenges posed by MNAR missingness.

Synthetic data provides a controlled environment to evaluate the model's performance systematically. Through this, we explore key metrics, the impact of matrix size, and the comparison of different missingness models. Additionally, we analyze the role of the Integrated Completed Likelihood (ICL) criterion in selecting appropriate models, supported by detailed mathematical explanations and insightful interpretations.

**Data Generation**

The synthetic data generation process leverages the Latent Block Model with MNAR missingness to create controlled co-clustering scenarios. The model is parameterized as:

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

where:

- $\alpha,\beta$: Represent equal mixing proportions for the three row and column clusters.
- $\pi$: Block connectivity matrix, with $\epsilon$ controlling task difficulty by altering the overlap between clusters.
- MNAR parameters: Set to $\mu=1$, $\sigma_A^2=\sigma_B^2=\sigma_C^2=\sigma_D^2=1$, resulting in a global missingness rate of 35%.

The model ensures well-separated clusters under smaller $\epsilon$, while increasing $\epsilon$ adds ambiguity, testing the model's robustness in more challenging scenarios. Matrix sizes $(n_1,n_2)$ are varied to study the relationship between task difficulty and data availability.

**Performance Metrics**

To evaluate the model's ability to recover latent clusters, the discrepancy between true $(Y,Z)$ and predicted $(\hat{Y},\hat{Z})$ cluster assignments is measured using:

$$l_{\mathrm{item}}(Y,Z,\hat{Y},\hat{Z}) = 1-\max_{t\in\Omega_1,s\in\Omega_2}\frac{1}{n_1n_2}\sum_{ijkl}Y_{ik}\hat{Y}_{it(k)}Z_{jl}\hat{Z}_{js(l)},$$

where:

- $\Omega_1,\Omega_2$: Sets of all permutations of row and column clusters, accounting for label invariance in clustering.
- The summation: Captures the alignment between true and predicted labels over all matrix entries.

Additionally, the **Conditional Bayes Risk** ($r_{\mathrm{item}}$) evaluates performance relative to the observed data matrix:

$$r_{\mathrm{item}}(\hat{Y},\hat{Z}) = \mathbb{E}[l_{\mathrm{item}}(Y,Z,\hat{Y},\hat{Z})\mid X_{\mathrm{obs}}],$$

where the expectation is computed using a Gibbs sampler to approximate the intractable posterior distribution of $Y$ and $Z$.

## Results on Synthetic Data

### Effect of Matrix Size

Unlike standard clustering, where increasing data size will complicates tasks due to dimensionality, co-clustering benefits from larger matrices, as this allows for better representation of the data because the dimensions of the spaces where the clustering is performed are expanded. For an analogy with supervised learning, it is as if we are adding more features or projecting the data in a higher space where we have higher chances of better separability. In other words, higher row-column interactions provide richer information, simplifying the identification of latent structures.

**Key observations:**

- **Smaller Matrices**: Higher conditional Bayes risk reflects the inherent difficulty in identifying clusters with limited data.
- **Larger Matrices**: Classification errors diminish, converging to the conditional Bayes risk as the algorithm leverages increased information effectively.

### Comparison of Missingness Models

The experiments highlight the importance of modeling missingness correctly:

- **Categorical LBM**: Treats missing values as a separate category, leading to consistently poor performance due to an inability to extract meaningful patterns.
- **MAR Model**: Assumes missingness is unrelated to data values. Performance deteriorates as MNAR effects (e.g., $\sigma_B^2, \sigma_D^2$) intensify, resulting in errors akin to random allocation in extreme cases.
- **MNAR Model**: Incorporates the dependency of missingness on data values. This approach maintains stable classification errors close to the conditional Bayes risk, even under strong MNAR effects.

These results emphasize that failing to account for MNAR missingness introduces substantial biases in clustering, underscoring the MNAR model's robustness.

### Model Selection via ICL

The Integrated Completed Likelihood (ICL) criterion proves effective for model selection, particularly in identifying the correct missingness mechanism. The asymptotic ICL for the MNAR model is consistently higher than that for MAR or Categorical LBM, reflecting its superior adaptability.

The synthetic data results validate the LBM-MNAR framework as a powerful tool for co-clustering tasks with missing data. By explicitly modeling missingness mechanisms, it consistently outperforms alternatives, demonstrating robustness under diverse scenarios. The integration of the ICL criterion further enhances its practical utility, making it a reliable choice for both synthetic and real-world applications.

## Results on Real Data (Own implementation)

The French parliamentary dataset used to test our model comprises 576 individuals (Members of Parliament, MPs) voting on 1,256 ballots. During a vote, MPs can choose to vote positively ("YES"), negatively ("NO"), abstain, or be absent. For consistency with our model, abstentions and absences are grouped and referred to as "missing" (NAN). In the French political landscape, there are three main political affiliations:

- Right-wing: Represented primarily by Les Républicains (LR).
- Left-wing: Includes parties like the Socialist Party(SOC), La "France Insoumise" (FI), "Libertés et Territoires" (LT), and the "Gauche Démocrate et Républicaine" (GDR).
- Center: Composed of government-affiliated parties such as "La République En Marche" (LaREM) and "Mouvement Démocrate" (MODEM).

While the Center aligns with the government, the Right-wing and Left-wing parties typically form the opposition. However, the relationships between parties can be complex. For instance, MPs from "Les Républicains" may occasionally align with "Socialist Party" positions on economic issues, and vice versa.

### Expectations for Model Results

- **Row Clustering (MPs)**: We anticipate observing a division of MPs into clusters corresponding to political affiliations (Right, Left, Center). However, due to the complexity of inter-party relationships, some clusters may reflect shared positions across ideological lines, especially on specific issues.
- **Column Clustering (Ballots)** : We expect ballots to cluster by topic, such as immigration, economic policy, socio-political issues, and environmental matters.

For the sake of simplicity we choose a number of clusters on MPs equal to 3 and 5 on the ballots . This low number of clusters will give a global view on the political affiliations detected by the model and also the topics that clearly separate these groups. The missingness model chosen is the MNAR. We also decide to randomly initialize the parameters even though a spectral clustering initialization is recommended.

### Co-clustering Results

The results we have obtained are quite reasonable, given the number of iterations at which the training was stopped. The results show (Figure 1) three groups of MPs: the first group is composed primarily of members of the opposition (SOC, LR, etc.), while the second and third groups are composed of parties affiliated with the government (LaREM, MODEM). The third group sometimes tends to partially align with the opposition majority (LR "Les Républicains") on some topics, such as topics D and B (with ballots mainly proposed by the Left-wing), with a low probability of voting positively even though the government voted positively. We ob-

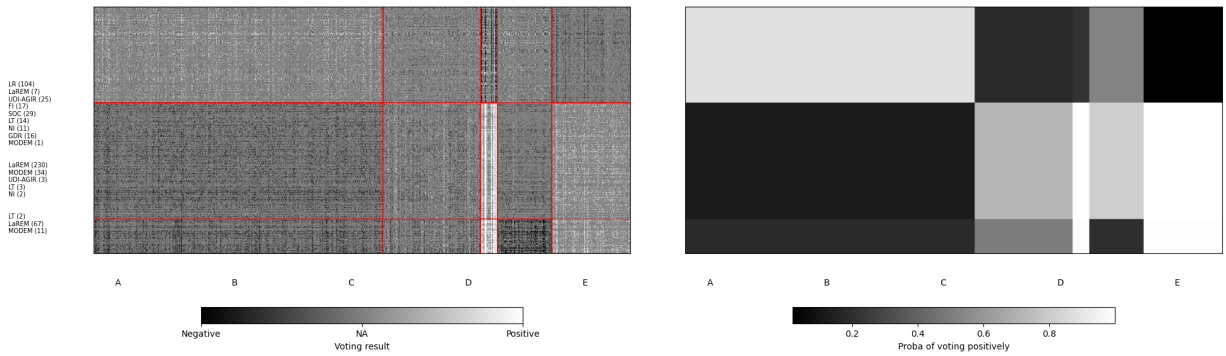


Figure 1: Co-clustering analysis of the French parliament data.

serve a clear opposition between groups 1 and 2 on the majority of topics, which is expected.For the ballots clustering we remark that ballots in topic A were proposed by the opposition and voted against by the government. Topics C and E were proposed and voted positively by the government.Ballots are hence mainly organized into the political affiliation of the requester.

**Interpretation of the non-voters' behavior**

In the graph below (Figure 2), we have represented $\nu_D$ against $\nu_C$ for all the ballots. We know that the value $\nu_C$ is directly linked to the global propensity of a ballot to be voted (either positively or negatively). The values of $\nu_D$ represent the propensity to be voted by its supporters,in other words, the capacity of the ballot to gather its supporters. We can separate two clusters: one in the upper-right and one in the lower-left side. The upper-right cluster groups the ballots with a high propensity to be voted and voted positively by their supporters. These ballots are from topics C and E, which were proposed by the government and massively voted by MPs affiliated with the government (Figure 1).They seems really important to the eye of the government. For most of the topics in A, D, and B (lower-left side), these ballots do not gather many supporters. This may indicate they are not of primary interest.

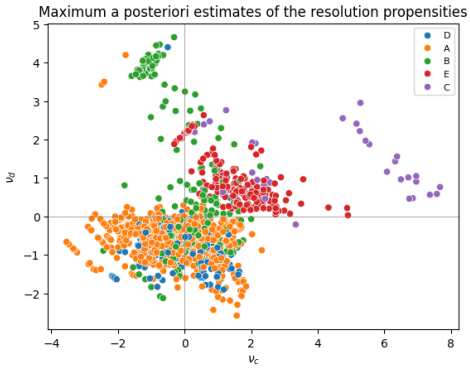


Figure 2: $\nu_D$ against $\nu_B$ for missingness interpretation in term of topic

We can give a similar type of interpretation for Figure 3. In this figure, we observe a clear separation into two clusters: one composed of MPs from the opposition and the other composed of MPs affiliated with the government. When MPs from the opposition are likely to vote positively, they actively come to vote in the parliament because they are the minority, and not voting may lead to a proposed law not being passed. This is not the case for LaREM and MODEM MPs, who are sometimes absent or abstain even when they intend to vote positively.

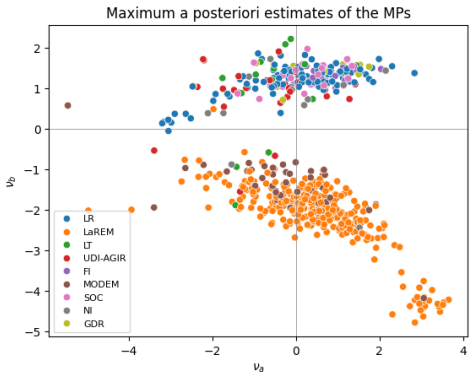


Figure 3: $\nu_B$ against $\nu_A$ for non voters behaviour interpretation

## Assessment of the Results

Although we have not obtained the exact same results as in the paper due to the premature stopping of the training process caused by numerical issues and the random initialization adopted, our results are quite reasonable and clearly interpretable. The MNAR missingness integration into the model is a very original and interesting idea because of the sensitive interpretation it provides for non-voters' behavior.

## Possible Follow-Up for the Article

The model proposed has original ideas that can be extended to handle continuous data or categorical data with more than two levels. Currently, the model has been developed only for binary data, which represents a minority of real-world datasets.
Apart from the French parliamentary data, the model has not been applied to other real-world datasets. We can explore its application in other domains, such as rec-

ommendation systems, to understand the relationship between non-responding customers and the underlying ratings they might want to give. This approach could provide better evaluations of a product even if the majority of customers have not responded.
The LBM with MNAR model presented here has some drawbacks that cannot be overlooked. One of them is the lack of proof in terms of model identifiability and consistency. Further research is needed to establish the conditions under which the parameters of the model are completely identifiable.

## Conclusion

In this paper, we have presented the article "Learning from Missing Data Using Binary Latent Block Model." We explained the framework, the inference method, the model selection criteria, and reproduced the results obtained on the parliamentary data.

Most often, missing values in machine learning are simply removed or replaced using basic imputation methods, such as K-Nearest Neighbors (K-NN). This paper provides another important perspective, enabling the exploitation of the information contained in the missing values when they are not Missing Completely at Random (MCAR). By doing so, the Binary Latent Block Model offers a structured approach to extracting meaningful insights from incomplete datasets, paving the way for more nuanced and effective handling of missing data scenarios.

## GitHub Implementation

The code and the notebooks used to reproduce the results can be found in the GitHub repository: https://github.com/Masrour-Aymane/LBM-with-MNAR/
In particular, the notebook `LBM_MNAR_review.ipynb` provides an in-depth analysis of the model and the obtained results.

## References


[1] Keribin Christine Biernacki Christophe Jacques Jérôme. "A Survey on Model-Based Co-Clustering: High Dimension and Estimation Challenges". In: *Journal of Classification* 40.2 (2023), pp. 332–381. DOI: 10.1007/s00357-023-09441-3. URL: https://link.springer.com/article/10.1007/s00357-023-09441-3.

[2] Ying Cheng and George M. Church. "Biclustering of Expression Data". In: *Proceedings of the International Conference on Intelligent Systems for Molecular Biology* 8 (2000), pp. 93–103.

[3] 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* (2001), pp. 269–274.

[4] Gérard Govaert and Mohamed Nadif. "Block clustering with Bernoulli mixture models: Comparison of different approaches". In: *Computational Statistics Data Analysis* 52.6 (2008), pp. 3233–3245.

[5] Donald B. Rubin. "Inference and Missing Data". In: *Biometrika* 63.3 (1976), pp. 581–592.

[6] Aude Sportisse et al. *Model-based Clustering with Missing Not At Random Data*. 2023. arXiv: 2112.10425 [stat.ML]. URL: https://arxiv.org/abs/2112.10425.