Learning from Missing Data with the Binary Latent Block Model

Samara Ndiaye, Aymane Masrour

University Paris Saclay

December 3, 2024

Outline

Introduction

- **Unsupervised learning** uncovers hidden structures in data.
- **Co-clustering**: Simultaneously clusters rows and columns to identify local patterns.
- Missing data is prevalent in real-world datasets and can be **informative**, especially when Missing Not At Random (MNAR).
- This work focuses on incorporating **Missing Not At Random (MNAR)** data into co-clustering models.
- Give an interpretation of the behaviour of non-voters

Types of Missing Data

- **MCAR (Missing Completely At Random)**: Independent of data values.
- **MAR (Missing At Random)**: Dependent on observed data.
- **MNAR (Missing Not At Random)**: Dependent on missing data values.

Challenge with MNAR Data

Ignoring MNAR mechanisms can bias clustering results.

Contribution

Develop a co-clustering model for **binary dataset**, that leverages missingness as informative data (MNAR), to interpret both observed and missing patterns.

Binary Latent Block Model (LBM)

- Partitions rows and columns into clusters to reveal block structures.
    - $\forall i \; Y_i \sim_{iid} \mathcal{M}(1, \alpha)$.
    - $\forall j \; Z_j \sim_{iid} \mathcal{M}(1, \beta)$.
    - $Y \perp Z$ ie $P(Y, Z) = P(Y)P(Z)$
- Within-cluster entries follow a Bernoulli distribution: $\forall i, j \; X_{ij} \mid Y_{ik} = 1, Z_{jl} = 1 \sim_{iid} \mathcal{B}(\pi_{kl})$
- **Parameters**:
    - α, β: Mixture proportions for rows and columns respectively .
    - π_{kl}: Probability $X_{ij} \in$ Block (k, l) $\forall$ i ,j.
- **Joint probability**:

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

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

Missingness Mechanism

- $X^{(o)}$ observed matrix with NAN ,X the partially observed matrix
- Missingness is represented as a binary mask M:

$$M_{ij} = \begin{cases} 1 & \text{if observed} \\ 0 & \text{if missing} \end{cases}$$

- **Nested** MNAR Missingness(dependence on missing value)
  - Propensity to be missing

$$P_{ij} = \log \frac{P(M_{ij} = 1)}{P(M_{ij} = 0)} = \mu + A_i + C_j + (-1)^{1-X_{ij}}(B_i + D_j)$$

  - $A_i \sim \mathcal{N}(0, \sigma_A^2)$ and $C_j \sim \mathcal{N}(0, \sigma_C^2)$:global propensity to be missing
  - $B_i \sim \mathcal{N}(0, \sigma_B^2)$ and $D_j \sim \mathcal{N}(0, \sigma_D^2)$: propensity to be missing given the missing value

Incorporating MNAR into the LBM

- X^o **distribution can be written without the mask**
  - **Given all the latent varibales**

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

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

$$\theta = (\alpha, \beta, \pi, \mu, \sigma_A^2, \sigma_B^2, \sigma_C^2, \sigma_D^2)$$

- **How to perform Inference ?**

Inference Challenges

- The goal is to estimate parameters θ by maximizing the observed data likelihood:

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

- Direct optimization computationally intractable due to:
  - High-dimensional integrals.
  - Latent variables with complex dependencies.
- The Expectation-Maximization (EM) algorithm addresses this challenge by:
  - Iteratively estimating latent variables (E-step).
  - Updating parameters based on these estimates (M-step).
- E-step intractable due to posterior $P(Y_i, Z_j, A_i, B_i, C_j, D_j | X^o, \theta)$

Motivation for Variational EM

- To address intractable expectations, we introduce the Evidence Lower Bound (ELBO):

  $$\log P(X^{(o)};\theta) = \mathcal{L}(q,\theta) + \mathrm{KL}(q||P(Y,Z,A,B,C,D \mid X^{(o)};\theta)),$$

  where:
  - $\mathcal{L}(q,\theta)$: Evidence Lower Bound.
  - $\mathrm{KL}(q||P)$: Kullback-Leibler divergence between the variational distribution q and the true posterior.
- **Key Insight**:
  - Maximizing $\mathcal{L}(q,\theta)$ indirectly maximizes $\log P(X^{(o)};\theta)$.

ELBO: Mathematical Expression

- The ELBO is expressed as:

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

  where:
  - First term: Expectation of the joint log-likelihood under q.
  - Second term: Negative entropy of the variational distribution q.
- By maximizing $\mathcal{L}(q,\theta)$ on q we:
  - Approximate the true posterior $P(Y, Z, A, B, C, D \mid X^{(o)}; \theta)$.
  - Ensure tractability through assumptions about q.

Variational Approximation

- **Goal**: Replace the intractable posterior $P(Y, Z, A, B, C, D \mid X^{(o)}; \theta)$ with a simpler variational distribution q.
- Mean-Field Approximation: factorized form of q

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

- **Advantages**:
  - Simplifies inference by reducing dependencies among variables.
  - Enables efficient optimization using gradient-based methods.

Conditional Laws for Latent Variables

- The latent variables are governed by the following conditional laws, parameterized by γ:

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

- $Y_i \mid X^{(o)} \sim \mathcal{M}(1; \tau_i^{(Y)})$
- $Z_j \mid X^{(o)} \sim \mathcal{M}(1; \tau_j^{(Z)})$
- $A_i \mid X^{(o)} \sim \mathcal{N}(\nu_i^{(A)}, \rho_i^{(A)}), \quad B_i \mid X^{(o)} \sim \mathcal{N}(\nu_i^{(B)}, \rho_i^{(B)})$
- $C_j \mid X^{(o)} \sim \mathcal{N}(\nu_j^{(C)}, \rho_j^{(C)}), \quad D_j \mid X^{(o)} \sim \mathcal{N}(\nu_j^{(D)}, \rho_j^{(D)})$

Variational EM Framework

- The Variational EM algorithm alternates between two steps:
  - **VE-step (Variational E-step)**:
    - Update the variational parameters by maximizing the variational objective:

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

  - **M-step**:
    - Update the model parameters by maximizing the expected complete-data log-likelihood:

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

- This ensures convergence to a local optimum of the Evidence Lower Bound (ELBO).

Practical Implementation

- **Initialization**:
  - Spectral clustering for row and column clusters.
  - Random initialization for missingness parameters.
- **Optimization**:
  - Gradient-based methods (e.g., L-BFGS).
  - Taylor expansion for log-odds computations.
- Automatic differentiation tools (e.g., PyTorch's `Autograd`).

Model Selection via ICL

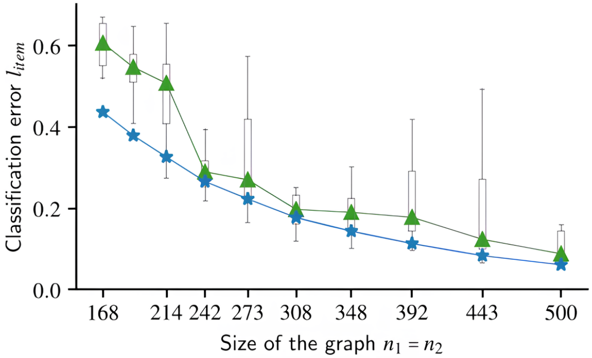
- Selects correct number of clusters (K, L).

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

- Selects appropriate missingness mechanism (MNAR vs MAR).
  - An ICL have also been developed for MAR and MNAR
  - Asymptotic ICL is higher for MNAR models when missingness is truly MNAR

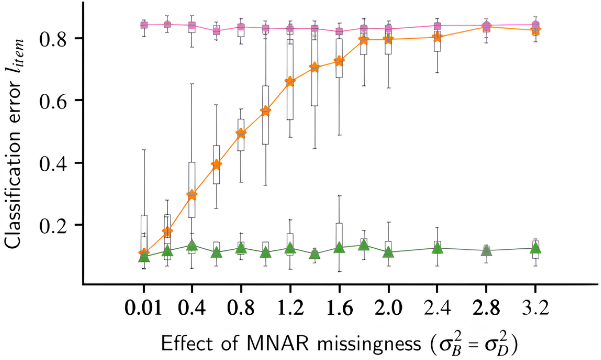
Effect of Matrix Size

- **Graphical Insight**: Classification error vs. matrix size.
- Smaller matrices: High error rates.
- Larger matrices: Error rates converge to Bayes risk due to richer data interactions.



Comparison of Missingness Models

- **Graphical Comparison**: Classification error vs. MNAR effect (σ_B^2, σ_D^2).
- MNAR model outperforms Categorical LBM and MAR models, especially under strong MNAR effects.



Dataset Overview

- **Dataset**: 576 Members of Parliament (MPs) voting on 1,256 ballots.
- **Vote Categories**:
  - "YES" (Positive vote)
  - "NO" (Negative vote)
  - "Missing" (Abstentions or absences)
- **Political Affiliations**:
  - **Right-wing**: Les Républicains (LR).
  - **Left-wing**: Socialist Party (SOC), France Insoumise (FI), etc.
  - **Center**: La République En Marche (LaREM) and Mouvement Démocrate (MODEM).
- **Government vs Opposition**: Center supports the government, while Right and Left typically form the opposition.

Expectations for Model Results

Row Clustering (MPs)

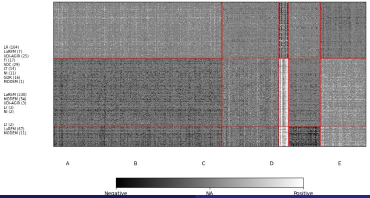
- Anticipated clusters reflect political affiliations: Right, Left, Center.
- Shared positions across ideological lines may create mixed clusters.

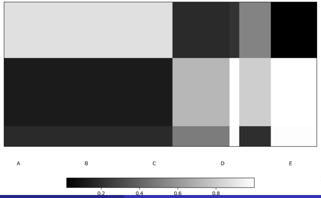
Column Clustering (Ballots)

- Expected clusters by topic: Immigration, economic policy, socio-political issues, and environmental matters.

Co-clustering results

- **MP Clustering**:
  - Group 1: Opposition MPs (SOC, LR, etc.).
  - Groups 2 & 3: Government-affiliated MPs (LaREM, MODEM).
  - Group 3 aligns with the opposition on some topics (e.g., topics B and D).
- **Ballot Clustering**:
  - Topics show clear divisions: ballots in A are proposed by the opposition , C and E proposed by the government
  - Topic D represents an area of overlap between opposition and government-affiliated MPs with an homogeneity in votes.



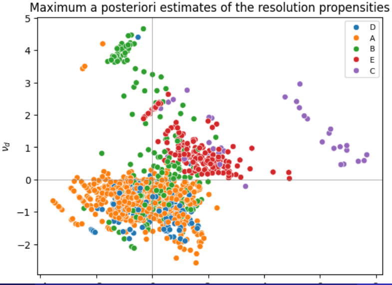


Interpretation of Non-Voters' Behavior

- **Graphical Analysis**:

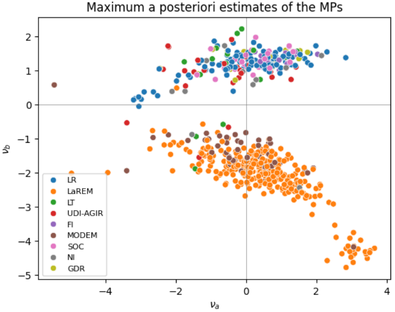
$$\nu_C \text{ vs. } \nu_D$$

- **Clusters of Ballots**:
  - Upper-right cluster: High propensity to be voted positively (government-proposed ballots like topics C and E).
  - Lower-left cluster: Low propensity to be voted (opposition votes on topics A, B, and D).



Non-Voters' Behavior (MPs)

- **MPs from the Opposition**:
  - Actively vote positively when supporting motions.
  - High attendance ensures minority voices are counted.
- **MPs from Government-Affiliated Parties**:
  - Sometimes abstain or are absent even when voting positively.
  - Reflects their majority position in parliament.



Future Directions

- Extend to non-binary and continuous data.
- Improve scalability with more efficient algorithms.
- Explore applications in recommender systems and bioinformatics.
- Research to proof conditions of identifiability

Conclusion

- The LBM-MNAR framework is robust for co-clustering tasks with missing data.
- Explicit modeling of missingness mechanisms ensures superior performance.
- ICL enhances utility, supporting accurate model selection.
- Applicability validated for both synthetic and real-world datasets.

Conclusion

- The LBM-MNAR framework is robust for co-clustering tasks with missing data.
- Explicit modeling of missingness mechanisms ensures superior performance.
- ICL enhances utility, supporting accurate model selection.
- Applicability validated for both synthetic and real-world datasets.

Thank You!

Questions?