Learning from missing data with the LBM

Advanced unsupervised learning

Laura FUENTES, Angel REYERO

Université Paris-Saclay

November 30, 2023

Outline

Index

Motivation

Clustering

Trauma.center	Heart rate	Death	Anticoagulant. therapy	Glascow score
Pitie-Salpêtrière	88	0	No	3
Beaujon	103	0	Yes	5
Bicêtre	90	0	Yes	6
Bicêtre	89	0	No	4
Lille	62	0	Yes	6
Lille	98	0	No	5
⋮	⋮	⋮	⋮	⋮

Motivation

High-dimensionality ⇒**Co-Clustering**

Trauma.center	Heart rate	Death	Anticoagulant. therapy	Glascow score	...
Pitie-Salpêtrière	88	0	No	3	...
Beaujon	103	0	Yes	5	...
Bicêtre	90	0	Yes	6	...
Bicêtre	89	0	No	4	...
Lille	62	0	Yes	6	...
Lille	98	0	No	5	...
⋮	⋮	⋮	⋮	⋮	...

Motivation

Missing values (Not Available (NA))

Trauma.center	Heart rate	Death	Anticoagulant. therapy	Glascow score	...
Pitie-Salpêtrière	88	0	No	3	...
Beaujon	103	0	NA	5	...
Bicêtre	NA	0	Yes	6	...
Bicêtre	NA	0	No	NA	...
Lille	62	0	Yes	6	...
Lille	NA	0	No	NA	...
⋮	⋮	⋮	⋮	⋮	...

Binary data context

In [1], they focus exclusively on a binary matrix with missing values.

$X^{\text{obs}} =$

	1	2	...	n_2
1	0	NA	...	1
2	1	0	...	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n_1	NA	1	...	NA

Objective: Find the K row and L column clusters.

Index

Latent block model (LBM)

- Y: Row cluster latent variable (of size $n_1 \times K$).
- Z: Column cluster latent variable (of size $n_2 \times L$).

Latent block model (LBM)

- Y: Row cluster latent variable (of size $n_1 \times K$).
- Z: Column cluster latent variable (of size $n_2 \times L$).

Assumptions:

- $Y \perp\!\!\!\perp Z$

Latent block model (LBM)

- Y**:** Row cluster latent variable (of size $n_1 \times K$).
- Z**:** Column cluster latent variable (of size $n_2 \times L$).

Assumptions:

- $Y \perp\!\!\!\perp Z$
- $\forall i_1, i_2 \in \{1, \ldots, n_1\} \quad Y_{i_1} \perp\!\!\!\perp Y_{i_2} \quad \& \quad Y_i \sim \mathcal{M}(1; \alpha)$ for $\alpha \in \mathbb{R}_+^{n_1}$ such that $\sum_k \alpha_k = 1$.

Latent block model (LBM)

- Y: Row cluster latent variable (of size $n_1 \times K$).
- Z: Column cluster latent variable (of size $n_2 \times L$).

Assumptions:

- $Y \perp\!\!\!\perp Z$
- $\forall i_1, i_2 \in \{1, \ldots, n_1\} \quad Y_{i_1} \perp\!\!\!\perp Y_{i_2} \quad \& \quad Y_i \sim \mathcal{M}(1; \alpha)$ for $\alpha \in \mathbb{R}_+^{n_1}$ such that $\sum_k \alpha_k = 1$.
- $\forall j_1, j_2 \in \{1, \ldots, n_2\} \quad Z_{j_1} \perp\!\!\!\perp Z_{j_2} \quad \& \quad Z_j \sim \mathcal{M}(1; \beta)$ for $\beta \in \mathbb{R}_+^{n_2}$ such that $\sum_l \beta_l = 1$.

Latent block model (LBM)

- Y**:** Row cluster latent variable (of size $n_1 \times K$).
- Z**:** Column cluster latent variable (of size $n_2 \times L$).

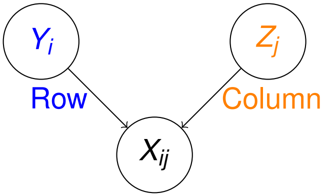
Assumptions:

- $Y \perp\!\!\!\perp Z$
- $\forall i_1, i_2 \in \{1, \ldots, n_1\} \quad Y_{i_1} \perp\!\!\!\perp Y_{i_2} \quad \& \quad Y_i \sim \mathcal{M}(1; \alpha)$ for $\alpha \in \mathbb{R}_+^{n_1}$ such that $\sum_k \alpha_k = 1$.
- $\forall j_1, j_2 \in \{1, \ldots, n_2\} \quad Z_{j_1} \perp\!\!\!\perp Z_{j_2} \quad \& \quad Z_j \sim \mathcal{M}(1; \beta)$ for $\beta \in \mathbb{R}_+^{n_2}$ such that $\sum_l \beta_l = 1$.
- For $\pi = (\pi_{kl}; k \in [K], l \in [L])$, we have

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

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

Latent block model (LBM)



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

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

Missingness model

We have the incomplete matrix X^{obs} and the mask matrix M where if $M_{ij} = 0$, then $X^{\text{obs}}_{ij} = \texttt{NA}$.

Missingness model

We have the incomplete matrix X^{obs} and the mask matrix M where if $M_{ij} = 0$, then $X_{ij}^{\mathrm{obs}} = \texttt{NA}$.

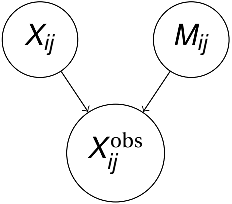


Figure 2: Latent variables of the missingness model.

Missingness model

We have the incomplete matrix X^{obs} and the mask matrix M where if $M_{ij} = 0$, then $X_{ij}^{\text{obs}} = \texttt{NA}$.

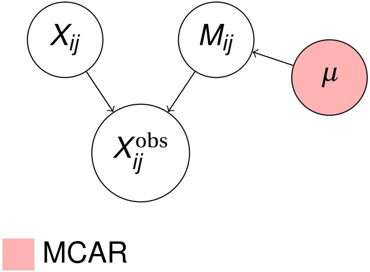


Figure 2: Latent variables of the missingness model.

Missingness model

We have the incomplete matrix X^{obs} and the mask matrix M where if $M_{ij} = 0$, then $X_{ij}^{\text{obs}} = \texttt{NA}$.

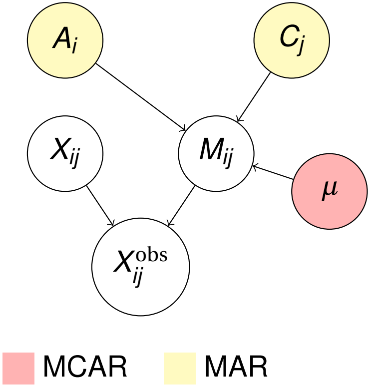


Figure 2: Latent variables of the missingness model.

Missingness model

We have the incomplete matrix X^{obs} and the mask matrix M where if $M_{ij} = 0$, then $X_{ij}^{\text{obs}} = \texttt{NA}$.

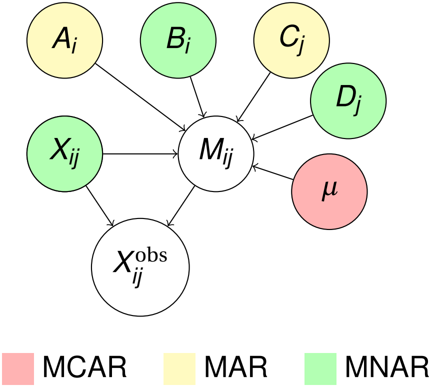


Figure 2: Latent variables of the missingness model.

Missingness model

$$\forall i, j, \; M_{ij} \sim \mathscr{B}\left(\text{expit}(P_{ij})\right),$$

MCAR

independent from the other and where

$$P_{ij} := \begin{cases} \mu & \text{if } X_{ij} = 1 \\ \mu & \text{if } X_{ij} = 0, \end{cases}$$

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

Missingness model

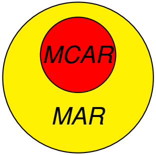
$$\begin{cases} \forall i, \quad A_i \sim \mathcal{N}(0, \sigma_A^2) \\ \forall j, \quad C_j \sim \mathcal{N}(0, \sigma_C^2) \end{cases}$$

$$\forall i, j, \; M_{ij} | A_i, \quad C_j \quad \sim \mathcal{B}\left(\operatorname{expit}(P_{ij})\right),$$

independent from the other and where

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

where $\operatorname{expit}(x) = \frac{1}{1+\exp(-x)}$.



Missingness model

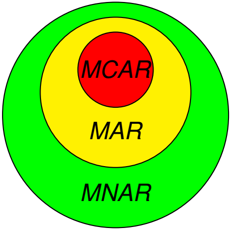
$$\begin{cases} \forall i, \quad A_i \sim \mathcal{N}(0, \sigma_A^2) \quad \& \quad B_i \sim \mathcal{N}(0, \sigma_B^2) \\ \forall j, \quad C_j \sim \mathcal{N}(0, \sigma_C^2) \quad \& \quad D_j \sim \mathcal{N}(0, \sigma_D^2) \end{cases}$$

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

independent from the other and where

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

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



Summary of the model

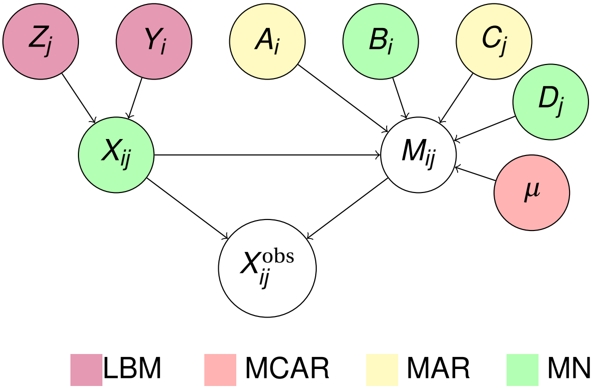


Figure 3: LBM adapted to the MNAR mechanism.

The latent variables are $\theta = (\alpha, \beta, \pi, \mu, \sigma_A^2, \sigma_B^2, \sigma_C^2, \sigma_D^2)$.

The final model

Rewriting we have

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

where

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

and

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

Index

Variational expectation maximization (VEM)

The *free energy*

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

Variational expectation maximization (VEM)

The *free energy*

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

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

$$+$$

$$\text{Mean} \quad \text{field} \quad \text{approximation.}$$

Variational expectation maximization (VEM)

Variational distribution:

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

Variational parameters:

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

VEM algorithm

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.

1 Initialize the parameters.

2 **while** *not stopping criterion satisfied* **do**

3 **VE-step:** we update the variational parameters:

$$\gamma^{t+1} \in \underset{\gamma}{\operatorname{argmax}} \mathcal{J}(q_\gamma, \theta^t).$$

M-step: we update the model parameters:

$$\theta^{t+1} \in \underset{\theta}{\operatorname{argmax}} \mathcal{J}(q_{\gamma^{t+1}}, \theta).$$

Integrated completed likelihood (ICL)

Log-integrated completed likelihood:

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

Integrated completed likelihood (ICL)

Log-integrated completed likelihood:

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

Asymptotic approximation:

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

Integrated completed likelihood (ICL)

Log-integrated completed likelihood:

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

Asymptotic approximation:

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

Practical approximation:

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

Index

Model computation: Maximization

VEM computation: 2 alternate maximizations

- **VE-Step:** Optimization with respect to variational distribution q_γ

$$\underset{\gamma}{\operatorname{argmax}} \mathscr{J}(\mathrm{q}_\gamma, \theta)$$

- **M-Step:** Optimization with respect to model parameters θ

$$\underset{\theta}{\operatorname{argmax}} \mathscr{J}(\mathrm{q}_\gamma, \theta)$$

Model computation: Maximization

VEM computation: 2 alternate maximizations

- **VE-Step:** Optimization with respect to variational distribution q_γ

$$\underset{\gamma}{\operatorname{argmax}} \mathscr{J}(\mathrm{q}_\gamma, \theta)$$

- **M-Step:** Optimization with respect to model parameters θ

$$\underset{\theta}{\operatorname{argmax}} \mathscr{J}(\mathrm{q}_\gamma, \theta)$$

No formal and explicit solutions: **L-BFGS optimization algorithm**

- Compute the gradients: computationally intense
- Autograd submodule from PyTorch: GPU capabilities

Results: Synthetic data

Objective: ensures certainty in the methodology employed to adapt to the underlying model

Data generation: various **sizes** and **difficulty levels** from a LBM with a MNAR missingness model

- Parameters incorporating 35% rate of global missingness
- Process repeated 20 times: variability parameter initialization

Synthetic data: Class prediction

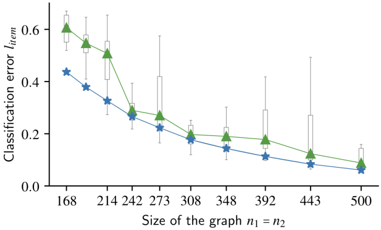


Figure 4: Comparison of classification errors concerning the size of the data matrix

⋆ The conditional Bayes risk,
▲ Results from the paper

- l_{item}: measures discrepancy among row and column clusters
- Task difficulty decreasement: increasing size of $n_1 = n_2$
- Better class prediction for larger datasets (lower l_{item})

Synthetic data: Missingness model

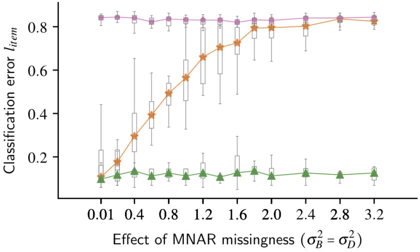


Figure 5: Classification errors as the MNAR effect intensifies

• Categorical LBM, ★ MAR model, ▲ MNAR model.

- MAR performance declines with increasing MNAR effect
- MNAR consistent classification error
- Importance to account for informative missingness

Synthetic data: Missingness model

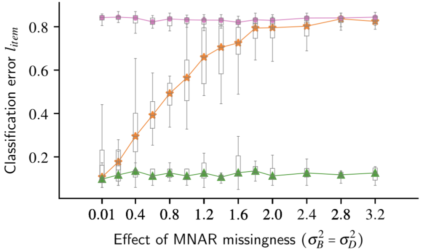


Figure 5: Classification errors as the MNAR effect intensifies

• Categorical LBM, ★ MAR model,
▲ MNAR model.

- MAR performance declines with increasing MNAR effect
- MNAR consistent classification error
- Importance to account for informative missingness

Model selection: Estimation ICL for each missingness mechanism (K, L known): MNAR consistently selected

Results: Real data

Objective: assess adaptability and flexibility of the assumed underlying model.

Results: Real data

Objective: assess adaptability and flexibility of the assumed underlying model.
3 datasets:

- **'votes'** (576×1256)
  - 1: Positive
  - -1: Negative
  - 0: `NA`/abstention (89%)

Results: Real data

Objective: assess adaptability and flexibility of the assumed underlying model.
3 datasets:

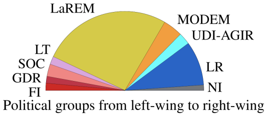
- **'votes'** (576×1256)
  - 1: Positive
  - -1: Negative
  - 0: NA/abstention (89%)
- **'texts':** ballots (1256 columns)
  - Amends, demandors, date

Results: Real data

Objective: assess adaptability and flexibility of the assumed underlying model.

3 datasets:

- **'votes'** (576 × 1256)
  - 1: Positive
  - -1: Negative
  - 0: NA/abstention (89%)
- **'texts':** ballots (1256 columns)
  - Amends, demandors, date
- **'deputes':** MPs (576 rows)
  - Names, political group etc.
  - Majority: Centrist MPs ('LaREM', 'MODEM')



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)

Hemicycle of the French National Assembly political groups

Real data: vote repartition

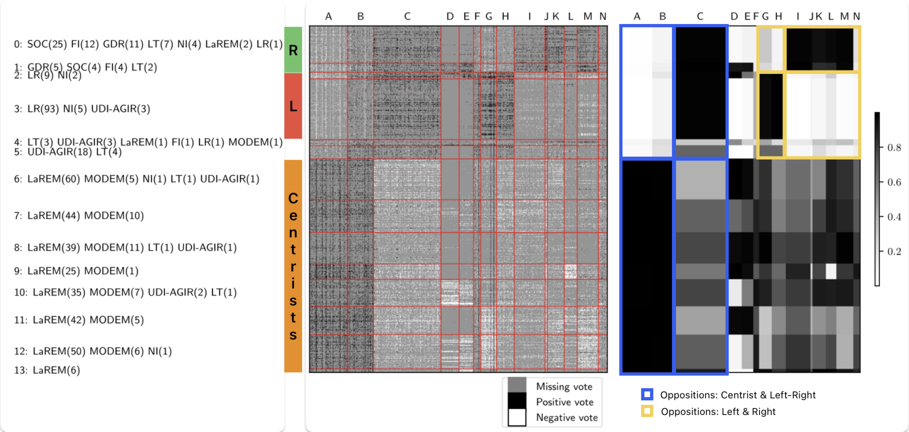
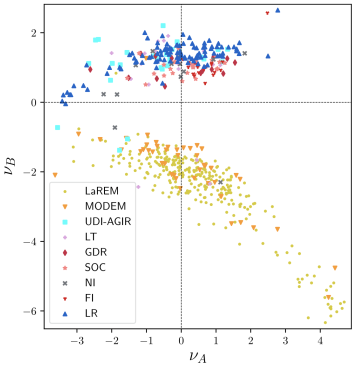


Figure 7: Reordered opinions according to row and column clusters

Real data: propensity to vote



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

- A: propensity to vote (MAR effect)
- B: additional effect of casting a vote when supporting the resolution (MNAR)
- ν_B: Discrimination of two clusters (centrists, left-right)

Index

Conclusion

- Scarcity of co-clustering methods for informative missingness
- Flexible missingness model for binary LBM
- Model estimation through a VEM approach
- Model selection criterion based on ICL
- Challenges: local minima convergence in VEM
- Future work:
  - Adapt the algorithm to ordinary data types
  - Formal identifiability proof

Index

References I

[1] Gabriel Frisch, Jean-Benoist Leger, and Yves Grandvalet. "Learning from missing data with the binary latent block model". In: *Statistics and Computing* 32.1 (2022), p. 9.

[2] Aurore Lomet, Gérard Govaert, and Yves Grandvalet. "Design of artificial data tables for co-clustering analysis". In: *Universit de Technologie de Compigne, France* (2012).

Index

Model estimation:

The observed log-likelihood can be rewritten as:

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

where *free energy* $\mathcal{J}$ given by

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

Classification error

Measure of discrepancy:

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

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

Conditioned Bayes risk

Conditioned Bayes risk on observed data matrices [[2]]:

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

- Control difficulty of clustering on simulated data matrices
- Tackle variability across risk on simulated data matrices

As the term $p(Y,Z|X^{\mathrm{obs}})$ is intractable, they compute the expectation as the average of a Gibbs sampler of $(Y,Z|X^{\mathrm{obs}})$.

Difference in ICL Figure

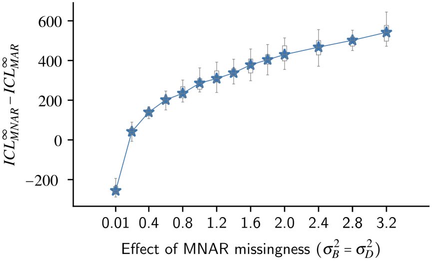


Figure 9: Difference in ICL between MAR and MNAR with respect to an increase in the MNAR effect

Where $\star$ is the median and MNAR model is selected when the ICL difference is positive.

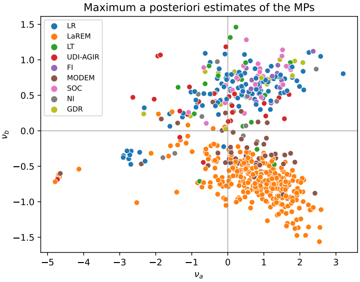
Number of row and column clusters figure

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

Figure 10: Number of (K, L) models selected by the asymptotic ICL among 20 trials on data matrices of different sizes and difficulties.

All matrices are generated with the same number of row and column classes: K = L = 3.

Our implementations



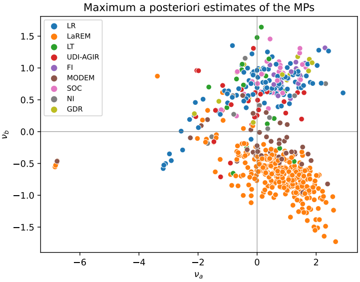


Figure 11: Maximum a posteriori estimates of the MPs propensities from our implementation for $K = 3$ and $L = 5$

Our implementations

Figure 12: Reordered opinions according to row and column clusters from our implementation $K = 3$ and $L = 5$