Model-based statistical learning: Co-clustering with the latent bloc model

LBM

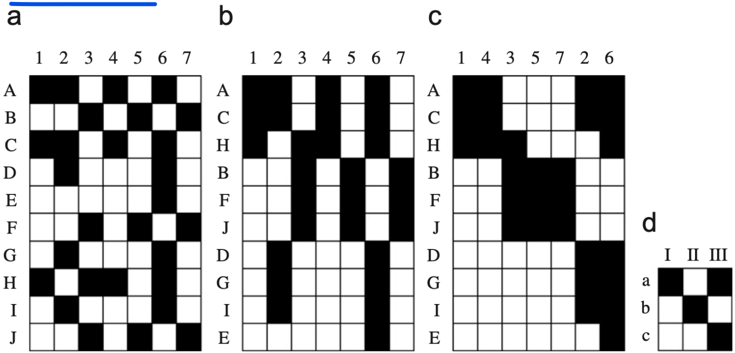
Vincent Vandewalle (vincent.vandewalle@univ-cotedazur.fr)

Msc 2 Data-Science & IA, 2024-2025

UNIVERSITÉ CÔTE D'AZUR

Introduction

Co-clustering aims at performing simultaneous clustering of both rows and columns:



Source: Christophe Biernacki, Julien Jacques, and Christine Keribin (2022). "A Survey on Model-Based Co-Clustering: High Dimension and Estimation Challenges". In

Bi-clustering, co-clustering and Latent Block Model (LBM)

- **Bi-clustering algorithms**: aim to detect homogeneous blocks within the data matrix which do not cover the entire matrix and which may overlap.
- **Co-clustering**: a specific bi-clustering model which assumes that all the individuals belong to one and only one row cluster, and *symmetrically* all the variables belong to only one column cluster.
- **Latent Block Model (LBM)**: LBM is a model for performing a model-based co-clustering

See Sara C Madeira and Arlindo L Oliveira (2004). "Biclustering algorithms for biological data analysis: a survey". In: *IEEE/ACM transactions on computational biology and bioinformatics* 1.1, pp. 24–45 for more details on bi-clustering algorithms.

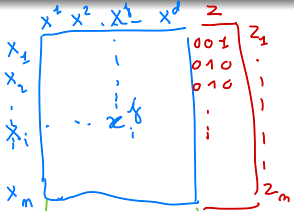
1. Recall the principle of model-based clustering
2. For what type of data is it designed? Any king, but need a model on $X|Z=k$
3. What is the link between the components of the mixture and the clusters? Each component of the mixture is interpreted a cluster
4. How to select the number of clusters? BIC
5. How can your compare two partitions when performing clustering?
6. Why using the rand index?
7. Why performing only clustering on rows, then on columns would not be sufficient to solve the co-clustering problem? Do the clustering of rows and columns simultaneously

ARI : Adjusted Rand Index (idea computed the percentage of concording pairs in the two clustering

Questions on Model-Based Clustering (MBC)

1. Recall the principle of model-based clustering Model the distribution of the data as a mixture of distributions.
2. For what type of data is it designed? Any kind of data as soon as we are able to propose a model for the class specific density.
3. What is the link between the component of the mixture and the clusters? Each component is interpreted as a cluster
4. How to select the number of clusters? It can be selected by AIC, BIC or ICL → Choose the number of clusters maximizing BIC criterion
5. How can your compare two partitions when performing clustering? By using the Adjusted Rand Index
6. Why using the rand index? It is invariant up to class permutation
7. Why performing only clustering on rows, then on columns would not be sufficient to solve the co-clustering problem? I allow to model the whole data matrix by a very sparse model.

The Latent Block Model (LBM) assumptions (1/2)



Data matrix $\mathbf{x}$ $(n \times d)$

- $\mathbf{x}_i$: the row/individual number i
- $\mathbf{x}^j$: the column/variable number j of $\mathbf{x}$
- x_i^j : variable j of individual i

nb of cluster in lines

Partition of the rows $\mathbf{z}$ $(n \times K)$

- $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$
- $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK}) \in \{0,1\}^K$
- $z_{ik} = 1$ if i belongs to row group k and 0 otherwise

Partition of the columns $\mathbf{w}$ $(d \times L)$

nb of cluster in columns

- $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_d)$
- $\mathbf{w}_i = (w_{j1}, \ldots, w_{jL}) \in \{0,1\}^L$
- $w_{j\ell} = 1$ if variable $\mathbf{x}^j$ belongs to column group ℓ and 0 otherwise

Main assumption: each point x_i^j is assumed to be independent given $\mathbf{z}_i$ and $\mathbf{w}_j$ (the knowledge of the block):

$$f(\mathbf{x}|\mathbf{z},\mathbf{w};\theta) = \prod_{k=1}^{K}\prod_{\ell=1}^{L}\prod_{i=1}^{n}\prod_{j=1}^{d} \underbrace{f(x_i^j;\alpha_{k\ell})}_{\text{pdf of } x_i^j \text{ in block } k\ell}{}^{z_{ik}w_{j\ell}}$$

$$z_{ik}w_{j\ell} = \begin{cases} 1 & \text{iff } z_{ik}=1 \text{ and } w_{j\ell}=1 \\ 0 & \text{otherwise} \end{cases}$$

with $f(\cdot;\alpha_{k\ell})$ the pdf associated to block $k\ell$ and parametrized by $\alpha_{k\ell}$.

Moreover independence is assumed between all $\mathbf{z}_i$ and $\mathbf{w}_j$:

$$f(\mathbf{z},\mathbf{w};\theta) = \underbrace{\overbrace{\prod_{i,k}}^{\prod_{i=1}\prod_{k=1}} \pi_k^{z_{ik}}}_{f(z,\theta)} \underbrace{\overbrace{\prod_{j,\ell}}^{\prod_{j=1}\prod_{\ell=1}} \rho_\ell^{w_{j\ell}}}_{f(w;\theta)}$$

π_k: proportion of cluster k in row

ρ_j: proportion of cluster j in column

with $\pi = (\pi_k)_k$ (the probabilities of each cluster in row), $\rho = (\rho_\ell)_\ell$ (the probabilities of each cluster in column). $\theta = (\pi, \rho, \alpha)$ groups all the parameters

Thus

$$f(\mathbf{x},\mathbf{z},\mathbf{w};\theta) = \overbrace{\prod_{i,k}\pi_k^{z_{ik}}}^{f(z;\theta)} \overbrace{\prod_{j,\ell}\rho_j^{w_{j\ell}}}^{f(w;\theta)} \overbrace{\prod_{i,j,k,\ell} f(x_i^j;\alpha_{k\ell})^{z_{ik}w_{j\ell}}}^{f(x|z,w;\theta)}$$

Marginalizing over $\mathbf{z}$ and $\mathbf{w}$ (since they are not observed in practice ...), the pdf of $\mathbf{x}$ is sum untractable

any distribution
eg. normal, multinomial

observed likelihood →

$$f(\mathbf{x};\theta) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} \overbrace{\prod_{i,k}\pi_k^{z_{ik}} \prod_{j,\ell}\rho_j^{w_{j\ell}} \prod_{i,j,k,\ell} f(x_i^j;\alpha_{k\ell})^{z_{ik}w_{j\ell}}}^{f(x,z,w;\theta)}$$

parameter specific to block $k\ell$

with $\mathcal{Z}$ (resp. $\mathcal{W}$) the set of all possible partitions of the rows (resp. the columns)

ex: x_i^j continuous $\alpha_{k\ell} = (\mu_{k\ell}, \sigma_{k\ell})$

depending the model of x_i^j

Choice of $f(\cdot;\alpha_{k\ell})$ according the type of data for x_i^j

- **Binary**: Bernoulli of parameter $\alpha_{k\ell}$
- **Categorical with r levels**: Multinomial distribution with parameters $\alpha_{k\ell} = (\alpha_{k\ell}^1, \dots, \alpha_{k\ell}^r)$ $\sum_{m=1}^{r} \alpha_{k\ell}^m = 1$
- **Count data**: Poisson distribution with parameter $\alpha_{k\ell}$
- **Continuous**: Normal distribution with parameters $\alpha_{k\ell} = (\mu_{k\ell}, \sigma_{k\ell}^2)$
- Can be extended to numerous other data types (ordinal, functional, textual, ...)

These models are very parsimonious even in high dimension!
ToDo : Count the number of parameters of the LBM for each data type

LBM estimation

The observed log-likelihood is defined as:

$$\ell(\theta;\mathbf{x}) = \log f(\mathbf{x};\theta) = \log\left(\sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}}\prod_{i,k}\pi_k^{z_{ik}}\prod_{j,\ell}\rho_j^{w_{j\ell}}\prod_{i,j,k,\ell} f(x_i^j;\alpha_{k\ell})^{z_{ik}w_{j\ell}}\right)$$

- $\ell(\theta;\mathbf{x})$ requires the computation of K^nL^d terms which correspond to all the possible configurations of unobserved labels $\mathbf{z}$ and $\mathbf{w}$!
- The problem is a missing data problem thus possible to use the EM algorithm

$Q(\theta;\theta')$ the expectation of the completed log-likelihood

- $\ell_c(\theta;\mathbf{x},\mathbf{z},\mathbf{w})$ the completed likelihood
- $Q(\theta,\theta') = \mathbb{E}(\ell_c(\theta;\mathbf{x},\mathbf{z},\mathbf{w})|\mathbf{x},\theta')$ the expectation of the completed log-likelihood given the current parameters θ'

EM algorithm starting from $\theta^{(0)}$ and loop until convergence

- Expectation (E) step: Computation of $Q(\theta;\theta')$
- Maximization (M) step: $\theta^{(q+1)} = \arg\max_\theta Q(\theta,\theta^{(q)})$

E step: computation of $Q(\theta, \theta^{(q)})$

The EM algorithm allows to increase the log-likelihood at each iteration: $\ell(\theta^{(q+1)} \geq \ell(\theta^{(q)})$ and thus to converge to a local maximum of the likelihood

$$\ell_c(\theta; \mathbf{x}, \mathbf{z}, \mathbf{w}) = \sum_k (\sum_i z_{ik}) \log \pi_k + \sum_\ell (\sum_j w_{j\ell}) \log \rho_\ell + \sum_{i,j,k,\ell} \log f(x_i^j; \alpha_{k\ell})$$

binary variable so it can take 0 or 1

Thus by taking the conditional expectation, we get:

$$\begin{aligned} Q(\theta, \theta^{(q)}) &= \sum_{i,k} p(z_{ik} = 1 | \mathbf{x}, \theta^{(q)}) \log \pi_k + \sum_{j,\ell} p(w_{j\ell} = 1 | \mathbf{x}, \theta^{(q)}) \log \rho_\ell \\ &+ \sum_{i,j,k,\ell} p(z_{ik} w_{j\ell} = 1 | \mathbf{x}; \theta^{(q)}) \log f(x_i^j; \alpha_{k\ell}) \end{aligned}$$

Let $s_{ik}^{(q)} = p(z_{ik} = 1 | \mathbf{x}; \theta^{(q)})$, $t_{j\ell}^{(q)} = p(w_{j\ell} = 1 | \mathbf{x}; \theta^{(q)})$ and $p(z_{ik} w_{j\ell} = 1 | \mathbf{x}; \theta^{(q)})$. All these computations are intractable due to dependence structure in the model.

Question: Assume that you would know these intractable quantities, how would perform the M-step?

Solution to the intractable E-step

- Variational approach: Constrain the joint probability to satisfy the relation

  posterior distribution $$p(\mathbf{z}, \mathbf{w}|\mathbf{x}; \theta) \approx p_z(\mathbf{z}|\mathbf{x}; \theta) p_w(\mathbf{w}|\mathbf{x}; \theta)$$

  where p_z and p_w are chosen to provide the closest approximation of $p(\mathbf{z}, \mathbf{w}|\mathbf{x}; \theta)$ while still being computable. The algorithm maximizes an evidence lower bound (ELBO)

  $$\ell(\theta; \mathbf{x}) \geq \mathcal{F}(\theta; \mathbf{x}) = \max_{p_z, p_w} (\ell_c(\theta; \mathbf{x}, \mathbf{z}, \mathbf{w}) - \log(p_z(\mathbf{z}) p_w(\mathbf{w}))$$

  this algorithm is called VEM as variational EM

sampling
- SEM algorithm : alternates the following steps: simulate $\mathbf{z}|\mathbf{x}, \mathbf{w}; \theta$ and then $\mathbf{w}|\mathbf{x}, \mathbf{z}; \theta$. Then update θ given the simulated classes $\mathbf{z}$ and $\mathbf{w}$

Estimating and evaluation of the rows and the columns clusters

Estimation

- VEM : based on $p_z(\mathbf{z}|\mathbf{x};\hat{\theta})$ and $p_w(\mathbf{w}|\mathbf{x};\hat{\theta})$ at the last iteration
- SEM: Based on sampling $(\mathbf{z},\mathbf{w})|\mathbf{x};\hat{\theta}$ by a Gibbs sampler, then estimate $(\hat{\mathbf{z}},\hat{\mathbf{w}})$ by the mode of the marginal sampled distribution.
- CEM: Based on an alternate optimization of the completed log-likelihood

Evaluation

- ARI: Adjusted Rand Rand Index / For the rows and columns respectively
- CARI: Co-clustering ARI developed for co-clustering

SEM-Gibbs algorithm

- Initialize the partitions in rows $\mathbf{z}^{(0)}$ and and in columns $\mathbf{w}^{(0)}$.
- For r in 1 to r^{max}
  - Compute $\theta^{(r)} = \operatorname{argmax}_\theta f(\mathbf{x}, \mathbf{z}^{(r-1)}, \mathbf{w}^{(r-1)}; \theta)$
  - Sample $\mathbf{z}^{(r)} \sim \mathbf{z}|\mathbf{x}, \mathbf{w}^{(r-1)}, \theta^{(r)}$
  - Sample $\mathbf{w}^{(r)} \sim \mathbf{w}|\mathbf{x}, \mathbf{z}^{(r)}, \theta^{(r)}$

This produce a sequence of parameter $\theta^{(0)}, \theta^{(1)}, \ldots$ converging in the neighbourhood of the MLE. A usual choice is to retain the last value $\hat{\theta} = \theta^{(r^{max})}$.

Estimation of $\hat{\mathbf{z}}$ and $\hat{\mathbf{w}}$

Given this fixed value of $\hat{\theta}$ it is possible to sample new values of $\mathbf{z}$ and $\mathbf{w}$ according to $p(\mathbf{z}, \mathbf{w}|\mathbf{x}; \hat{\theta})$ using the following Gibbs algorithm:

- $\mathbf{z}^{(r)} \sim \mathbf{z}|\mathbf{w}^{(r-1)}; \hat{\theta}$
- $\mathbf{w}^{(r)} \sim \mathbf{w}|\mathbf{z}^{(r)}; \hat{\theta}$

$\hat{\mathbf{z}}$ and $\hat{\mathbf{w}}$ are obtained by taking the mode of the sampled partitions

Details on the computation of $p(z_{ik} = 1|\mathbf{x}, \mathbf{w}; \theta)$

$$p(z_{ik} = 1|\mathbf{x}, \mathbf{w}; \theta) \propto f(\mathbf{x}, \mathbf{w}, z_{ik} = 1; \theta)$$

and

$$f(\mathbf{x}, \mathbf{w}, z_{ik} = 1; \theta) = p(z_{ik} = 1; \theta) p(\mathbf{w}; \theta) f(\mathbf{x}_i|\mathbf{w}, z_{ik} = 1; \theta) \times f(\mathbf{x}_{\{-i\}}|\mathbf{w}; \theta)$$

where $\mathbf{x}_{\{-i\}}$ denotes all the rows of $\mathbf{x}$ except row i. The last term does not depend on k, thus

$$p(z_{ik} = 1|\mathbf{x}, \mathbf{w}; \theta) \propto \alpha_k \left(\prod_{j,\ell} \rho_\ell^{w_{j\ell}} \right) \left(\prod_{j,\ell} f(x_i^j; \alpha_{k\ell})^{w_{j\ell}} \right) = \alpha_k \prod_{j,\ell} \rho_\ell^{w_{j\ell}} f(x_i^j; \alpha_{k\ell})^{w_{j\ell}}$$

And as a consequence

$$p(z_{ik} = 1|\mathbf{x}, \mathbf{w}; \theta) = \frac{\alpha_k \prod_{j,\ell} \rho_\ell^{w_{j\ell}} f(x_i^j; \alpha_{k\ell})^{w_{j\ell}}}{\sum_{k'=1}^K \alpha_{k'} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} f(x_i^j; \alpha_{k'\ell})^{w_{j\ell}}}$$

Thus the label of each row can be sample independently given the data of the row and the labels of all the columns.

Details on the VEM algorithm

Contrary to the SEM which is stochastic, the VEM algorithm is deterministic its tries to maximize the ELBO

$$\hat{\theta}_{VEM} = \arg\max_{\theta} \mathcal{F}(\theta; \mathbf{x}), \text{ and}$$

ELBO

$$\mathcal{F}(\theta; \mathbf{x}) = \max_{p_z, p_w} (\mathbb{E}_{\mathbf{z}\sim p_z, \mathbf{w}\sim p_w} [\ell_c(\theta; \mathbf{x}, \mathbf{z}, \mathbf{w}) - \log(p_z(\mathbf{z}) p_w(\mathbf{w}))])$$

$\ell(\theta)$

Thus the VEM algorithm performs an alternate optimization between θ and p_z, p_w:

- Update θ given $p_z^{(r-1)}$ and $p_w^{(r-1)}$: standard M-step

$$\theta^{(r)} = \arg\max_{\theta} \mathbb{E}_{p_z^{(r-1)}, p_w^{(r-1)}} [\log f(\mathbf{x}, \mathbf{z}, \mathbf{w}; \theta)]$$

- Update p_z, p_w given $\theta^{(r)}$: solve a coupled fixed point equation. Let $p_z(z_{ik} = 1) = \tau_{ik}$ and $p_w(w_{j\ell} = 1) = \nu_{j\ell}$

$$\tau_{ik} \propto \pi_k^{(r)} \prod_{j,\ell} f(x_i^j; \alpha_{k\ell}^{(r)})^{\nu_{j\ell}} \ \forall i, k \ \text{ and } \nu_{j\ell} \propto \rho_\ell^{(r)} \prod_{j,\ell} f(x_i^j; \alpha_{k\ell}^{(r)})^{\tau_{ik}} \ \forall j, \ell$$

Adjusted Rand Index (ARI)

Purpose

The Adjusted Rand Index (ARI) measures the similarity between two clusterings, correcting for chance. It is widely used to evaluate the quality of clustering results.

Rand Index (RI)

The Rand Index evaluates the agreement between two clusterings C_1 and C_2 by considering:

- a: Number of pairs of elements in the same cluster in both C_1 and C_2.
- b: Number of pairs of elements in different clusters in both C_1 and C_2.

The formula for the Rand Index is $\text{RI} = \frac{a+b}{\binom{n}{2}}$

n(n-1)/2

Adjusted Rand Index (ARI)

The ARI adjusts the Rand Index to account for the expected similarity due to chance:

$$\text{ARI} = \frac{\text{RI} - \mathbb{E}[\text{RI}]}{\max(\text{RI}) - \mathbb{E}[\text{RI}]}$$

- $\mathbb{E}[\text{RI}]$: Expected Rand Index for random clusterings.
- Range: -1 (disagreement) to 1 (perfect agreement), with 0 indicating random labeling.

Choice of the number of clusters

Since the computation of the observed likelihood is difficult, a solution is to use the ICL criterion to select K and L:

$$\mathsf{ICL}(K,L) = \log f(\mathbf{x}, \hat{\mathbf{z}}^{K,L}, \hat{\mathbf{w}}^{K,L}; \hat{\theta}^{K,L}) - \frac{\text{nb param}(K,L)}{2}\log(nm)$$

where ^{K,L} stands for the values estimated using K clusters in rows and L clusters in columns, and nb param(K,L) is the number of parameters for the model.