Co-Clustering Binary Data Using Covariates

Serge Iovleff, Seydou Nourou, Cheikh Loucoubar

Co-Clustering Binary Data Using Covariates

Serge Iovleff & Seydou Nourou Syllla & Cheikh Loucoubar

September 10, 2018

Abstract

We present a novel co-clustering method using co-variates with application to genomic data.

Contents

1 Introduction

Classification is a method of data analysis that aims to group together a set of observations into homogeneous classes. Its aim is the automatic resolution of problems by decision-making based on the observations induced to the problems. Its main purpose is to define rules for classifying objects based on qualitative or quantitative variables characterizing these objects. It plays an increasingly important role in many scientific and technical fields. Clustering may be the most popular technique for data analysis in many disciplines.

Unlike classical clustering, which groups similar objects from a single collection of objects, coclustering or biclustering [MO04] aims at simultaneously grouping objects from two disjoint sets, thus revealing interactions between elements of two sets. In recent years, co-clustering has been increasingly used in many areas ranging from information retrieval, data mining, computer vision, biology, and so on. It is most often used with bipartite spectral graphing partitioning methods in the field of [Dhi01] extracting text data by simultaneously grouping documents and content (words) and analyzing huge corpora unlabeled documents [XZDZ10] to simultaneously understand aggregates of subsets of web users(sessions) and information from the page views. Co-clustering algorithms have also been developed for computer vision applications. it is used for grouping images by simultaneously grouping images with their low-level visual characteristics and for content-based image search [GQX05][RV09] [Qiu04].

2 Block mixture models

2.1 Classical latent block model

Let $\mathbf{x}$ be a data set doubly indexed by a set I with n elements (individuals) and a set J with m elements (variables). We represent a partition of I into g clusters by $\mathbf{z} = (z_{11}, \ldots, z_{ng})$ with $z_{ik} = 1$ if i belongs to cluster k and $z_{ik} = 0$ otherwise, $z_i = k$ if $z_{ik} = 1$ and we denote by $z_{.k} = \sum_i z_{ik}$ the cardinality of row cluster k. Similarly, we represent a partition of J into d clusters by $\mathbf{w} = (w_{11}, \ldots, w_{md})$ with $w_{j\ell} = 1$ if j belongs to cluster ℓ and $w_{j\ell} = 0$ otherwise, $w_j = \ell$ if $w_{j\ell} = 1$ and we denote $w_{.\ell} = \sum_j w_{j\ell}$ the cardinality of column cluster ℓ.

The block mixture model formulation is defined in [GN03] and [BIG14] (among others) by the following probability density function

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{\mathbf{u} \in \mathcal{U}} p(\mathbf{u}; \boldsymbol{\theta}) f(\mathbf{x}|\mathbf{u}; \boldsymbol{\theta})$$

where $\mathcal{U}$ denotes the set of all possible labellings of $I \times J$ and $\boldsymbol{\theta}$ contains all the unknown parameters of this model. By restricting this model to a set of labellings of $I \times J$ defined by a product of labellings of I and J, and further assuming that the labellings of I and J are independent of each other, one obtain the decomposition

$$f(\mathbf{x}; \boldsymbol{\theta}) = \sum_{(\mathbf{z}, \mathbf{w}) \in \mathcal{Z} \times \mathcal{W}} p(\mathbf{z}; \boldsymbol{\theta}) p(\mathbf{w}; \boldsymbol{\theta}) f(\mathbf{x}|\mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) \tag{1}$$

where $\mathcal{Z}$ and $\mathcal{W}$ denote the sets of all possible labellings $\mathbf{z}$ of I and $\mathbf{w}$ of J. Equation (1) define a *Latent Block Model*.

2.2 Latent block model for binary variables with co-variables: General formulation

From now, we assume that $\mathbf{x}$ is a binary data set. Let $\mathbf{y}$ represents a data-set (co-variables) of $\mathbb{R}^p$ indexed by I. In order to take into account this set of co-variables the classical block model formulation is extended to propose a block mixture model defined by the following probability density function

$$f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) = \sum_{(\mathbf{z}, \mathbf{w}) \in \mathcal{Z} \times \mathcal{W}} p(\mathbf{z}; \boldsymbol{\theta}) p(\mathbf{w}; \boldsymbol{\theta}) f(\mathbf{x}|\mathbf{y}, \mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) f(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}). \tag{2}$$

By extending the latent class principle of local independence to our block model, each data pair $(x_{ij}, \mathbf{y}_i)$ will be independent once z_i and w_j are fixed. Hence we have

$$f(\mathbf{x}, \mathbf{y}|\mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) = \prod_{i,j} f(x_{ij}, \mathbf{y}_i; \boldsymbol{\theta}).$$

We choose to model the dependency between x_{ij} and $\mathbf{y}_i$ using the canonical link for binary response data

$$f(x_{ij}|\mathbf{y}_i, \boldsymbol{\beta}_{z_i w_j}) = \text{logis}(\beta_{0,z_i w_j} + \boldsymbol{\beta}_{z_i w_j}^T \mathbf{y}_i)^{x_{ij}} \left(1 - \text{logis}(\beta_{0,z_i w_j} + \boldsymbol{\beta}_{z_i w_j}^T \mathbf{y}_i)\right)^{1 - x_{ij}} \tag{3}$$

with $(\beta_0, \boldsymbol{\beta}_{k,l}) \in \mathbb{R}^{p+1}$ and $\text{logis}(x) = e^x/(1 + e^x)$. Each data point $\mathbf{y}_i$ will be independent once z_i are fixed. In the examples presented in section 3, we choose

$$f(\mathbf{y}|\mathbf{z}; \boldsymbol{\theta}) = \prod_i \phi(\mathbf{y}_i; \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})$$

with ϕ denoting the multivariate Gaussian density in $\mathbb{R}^p$.

In order to simplify the notation, we add a constant coordinate 1 to vectors $\mathbf{y}_i$ and write $\boldsymbol{\beta}_{k,l}$ in the later rather than $(\beta_{0,k,l}, \boldsymbol{\beta}_{k,l})$.

The parameters are thus $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_d)$ are the vectors of probabilities π_k and ρ_ℓ that a row and a column belong to the kth row component and to the ℓth column component respectively, $\boldsymbol{\beta} = (\boldsymbol{\beta}_{kl})$ are the coefficients of the logistic function,

$\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the means and variances of the Gaussian density. Summarizing, we obtain the latent block mixture model with pdf

$$f(\mathbf{x},\mathbf{y}|\boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} \prod_{i,j} \pi_{z_i}\rho_{w_j} \operatorname{logis}(\mathbf{y}_i^T\boldsymbol{\beta}_{z_i w_j})^{x_{ij}} \left(1-\operatorname{logis}(\mathbf{y}_i^T\boldsymbol{\beta}_{z_i w_j})\right)^{1-x_{ij}} \phi(\mathbf{y}_i;\boldsymbol{\mu}_{z_i},\boldsymbol{\Sigma}_{z_i}). \quad (4)$$

Using above formulation, the randomized data generation process can be described by the four steps row labellings (R), column labellings (C), co-variable data generation (Y) and data generation (X) as follows:

(R) Generate the labellings $\mathbf{z} = (z_1,\ldots,z_n)$ according to the distribution $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_g)$.

(C) Generate the labellings $\mathbf{w} = (w_1,\ldots,w_m)$ according to the distribution $\boldsymbol{\rho} = (\rho_1,\ldots,\rho_d)$.

(Y) Generate for $i = 1,...,n$ vector $\mathbf{y}_i$ according to the Gaussian distribution $\mathcal{N}_p(\boldsymbol{\mu}_{z_i},\boldsymbol{\Sigma}_{z_i})$.

(X) Generate for $i = 1,...,n$ and $j = 1,...,m$ a value x_{ij} according to the Bernoulli distribution $f(x_{ij}|\mathbf{y}_i;\boldsymbol{\beta}_{z_i w_j})$ given in (3).

2.3 Model parameters estimation

The complete data is represented as a vector $(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{w})$ where unobservable vectors $\mathbf{z}$ and $\mathbf{w}$ are the labels. The log-likelihood to maximize is

$$l(\boldsymbol{\theta}) = \log f(\mathbf{x},\mathbf{y};\boldsymbol{\theta}) \quad (5)$$

and the double missing data structure, namely $\mathbf{z}$ and $\mathbf{w}$, makes statistical inference more difficult than usual. More precisely, if we try to use an EM algorithm as in standard mixture model [DLR97] the complete data log-likelihood is find to be

$$L_C(\mathbf{z},\mathbf{w},\boldsymbol{\theta}) = \sum_k z_{\cdot k}\log\pi_k + \sum_\ell w_{\cdot\ell}\log\rho_\ell + \sum_{i,j,k,\ell} z_{ik}w_{j\ell}\log f(x_{ij},\mathbf{y}_i;\boldsymbol{\theta}_{k\ell}). \quad (6)$$

The EM algorithm maximizes the log-likelihood $l(\boldsymbol{\theta})$ iteratively by maximizing the conditional expectation $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(c)})$ of the complete data log-likelihood given a previous current estimate $\boldsymbol{\theta}^{(c)}$ and $(\mathbf{x},\mathbf{y})$:

$$Q(\boldsymbol{\theta},\boldsymbol{\theta}^{(c)}) = \mathbb{E}\left[L_C(\mathbf{z},\mathbf{w},\theta)\,\middle|\,\mathbf{x},\mathbf{y},\boldsymbol{\theta}^{(c)}\right] = \sum_{i,k} t_{ik}^{(c)}\log\pi_k + \sum_{j,\ell} r_{j\ell}^{(c)}\log\rho_\ell + \sum_{i,j,k,\ell} e_{ikj\ell}^{(c)}\log f(x_{ij},\mathbf{y}_i;\boldsymbol{\theta}_{k\ell})$$

where

$$t_{ik}^{(c)} = P(z_{ik}=1|\mathbf{x},\mathbf{y},\boldsymbol{\theta}^{(c)}), \qquad r_{jl}^{(c)} = P(w_{j\ell}=1|\mathbf{x},\mathbf{y},\boldsymbol{\theta}^{(c)}), \qquad e_{ikj\ell}^{(c)} = P(z_{ik}w_{j\ell}=1|\mathbf{x},\mathbf{y},\boldsymbol{\theta}^{(c)})$$

Unfortunately, difficulties arise owing to the dependence structure in the model, in particular to determinate $e_{ikj\ell}^{(c)}$. The assumed independence of $\mathbf{z}$ and $\mathbf{w}$ in (1) is not preserved by the posterior probability.

To solve this problem an approximate solution is proposed in [GN03] using the [Hat86] and [NH98] interpretation of the VEM algorithm. Consider a family of probability distribution $q(z_{ik},w_{j\ell})$ verifying $q(z_{ik},w_{j\ell}) > 0$ and the relation $q(z_{ik},w_{j\ell}) = q(z_{ik})q(w_{j\ell})$, for all i,j,k,l. Set $t_{ik} = q(z_{ik})$ and $r_{jl} = q(w_{j\ell})$, $\mathbf{t} = (t_{ik})_{ik}$ for $i = 1,\ldots,n$, $k = 1,\ldots,g$ and $\mathbf{r} = (r_{jl})_{jl}$ for $j = 1,\ldots,m$ and $l = 1,\ldots,d$. Using the concavity of the log function, one shows easily that

$$l(\boldsymbol{\theta}) \geq \tilde{F}_C(\mathbf{t},\mathbf{r};\boldsymbol{\theta}) + KL(q(\mathbf{z},\mathbf{w}) \parallel p(\mathbf{z},\mathbf{w}|\mathbf{x},\mathbf{y},\boldsymbol{\theta})) \quad (7)$$

with $KL(q \parallel p)$ denoting the Kullback-Liebler divergence of distribution p and q,

$$\tilde{F}_C(\mathbf{t},\mathbf{r};\boldsymbol{\theta}) = \sum_k t_{\cdot k}\log\pi_k + \sum_\ell r_{\cdot\ell}\log\rho_l + \sum_{i,j,k,\ell} t_{ik}r_{j\ell}\log f(x_{ij},\mathbf{y}_i;\boldsymbol{\theta}_{k\ell}) + H(\mathbf{t}) + H(\mathbf{r}) \quad (8)$$

and $H(\mathbf{t})$, $H(\mathbf{r})$ denoting the entropy of $\mathbf{t}$ and $\mathbf{r}$, i.e.

$$H(\mathbf{t}) = \sum_{ik} t_{ik}\log t_{ik}, \qquad H(\mathbf{r}) = \sum_{jl} r_{jl}\log r_{jl}.$$

$\tilde{F}_C$ is called the free energy or the fuzzy criterion. As the Kullback-Liebler divergence is always positive, the fuzzy criterion is a lower bound of the log-likelihood and is use in replacement of it. Doing that, the maximization of the likelihood $l(\boldsymbol{\theta})$ is replaced by the following problem

$$\underset{\mathbf{t},\mathbf{r},\boldsymbol{\theta}}{\operatorname{argmax}}\, \tilde{F}_C(\mathbf{t},\mathbf{r},\boldsymbol{\theta}).$$

This maximization can be achieved using the BEM algorithm detailed in next section.

2.4 Block expectation maximization (BEM) Algorithm

The fuzzy clustering criterion given in (8) can be maximized using a variational EM algorithm (VEM). We here outline the various expressions evaluated during E and M steps.

E-Step: we compute either the values of $\mathbf{t}$ (respectively $\mathbf{r}$) with $\mathbf{r}$ (respectively $\mathbf{t}$) and $\boldsymbol{\theta}$ fixed. Details are given in appendix A.

M-Step: we calculate row proportions $\boldsymbol{\pi}$ and column proportions $\boldsymbol{\rho}$. The maximization of $\tilde{F}_C$ w.r.t. $\boldsymbol{\pi}$, and w.r.t $\boldsymbol{\rho}$, is obtained by maximizing $\sum_k t_{\cdot k} \log \pi_k$, and $\sum_\ell r_{\cdot \ell} \log \rho_\ell$ respectively, which leads to

$$\pi_k = \frac{t_{\cdot k}}{n} \quad \text{and} \quad \rho_\ell = \frac{r_{\cdot \ell}}{m}. \tag{9}$$

Also the estimate of model parameters $\boldsymbol{\beta}$ will be obtained by maximizing

$$\boldsymbol{\beta}_{kl} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \sum_{ij} t_{ik} r_{jl} \log f(x_{ij}|\mathbf{y}_i;\boldsymbol{\beta}), \quad k = 1,\ldots,g, \quad l = 1,\ldots,d. \tag{10}$$

(see appendix B for details) and the parameters of the Gaussian density by the usual formulas

$$\boldsymbol{\mu}_k = \frac{1}{t_{\cdot k}} \sum_i t_{ik} \mathbf{y}_i \qquad \text{and} \qquad \boldsymbol{\Sigma}_k = \frac{1}{t_{\cdot k}} \sum_i t_{ik} (\mathbf{y}_i - \boldsymbol{\mu}_k)(\mathbf{y}_i - \boldsymbol{\mu}_k)^T. \tag{11}$$

BEM algorithm: Using the **E** and **M** steps defined above, **BEM** algorithm can be enumerated as follows:

Initialization Set $\mathbf{t}^{(0)}, \mathbf{r}^{(0)}$ and $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\pi}^{(0)}, \boldsymbol{\rho}^{(0)}, \boldsymbol{\beta}^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)})$.

(a) Row-EStep Compute $\mathbf{t}^{(c+1)}$ using formula

$$t_{ik}^{(c+1)} = \frac{\pi_k^{(c)} \prod_{jl} \left(f(x_{ij}|\mathbf{y}_i;\boldsymbol{\beta}_{kl}^{(c)}) \phi(\mathbf{y}_i;\boldsymbol{\mu}_k^{(c)},\boldsymbol{\Sigma}_k^{(c)}) \right)^{r_{jl}^{(c)}}}{\sum_k \pi_k^{(c)} \prod_{jl} \left(f(x_{ij}|\mathbf{y}_i;\boldsymbol{\beta}_{kl}^{(c)}) \phi(\mathbf{y}_i;\boldsymbol{\mu}_k^{(c)},\boldsymbol{\Sigma}_k^{(c)}) \right)^{r_{jl}^{(c)}}}. \tag{12}$$

(b) Row-MStep Compute $\boldsymbol{\pi}^{(c+1)}$, $\boldsymbol{\mu}^{(c+1)}$, $\boldsymbol{\Sigma}^{(c+1)}$ using equations (9) and (11) and estimate $\boldsymbol{\beta}^{(c+1/2)}$ by solving maximization problem (10).

(c) Col-EStep Compute $\mathbf{r}^{(c+1)}$ using formula

$$r_{jl}^{(c+1)} = \frac{\rho_l^{(c)} \prod_{ik} f(x_{ij}|\mathbf{y}_i;\boldsymbol{\beta}_{kl}^{(c+1/2)})^{t_{ik}^{(c+1)}}}{\sum_l \rho_l^{(c)} \prod_{ik} f(x_{ij}|\mathbf{y}_i;\boldsymbol{\beta}_{kl}^{(c+1/2)})^{t_{ik}^{(c+1)}}}. \tag{13}$$

Observe that r_{jl} does not depend of the density of $\mathbf{y}$.

(d) Col-MStep Compute $\boldsymbol{\rho}^{(c+1)}$ using equations (9) and estimate $\boldsymbol{\beta}^{(c+1)}$ by solving maximization problem (10).

Iterate Iterate **(a)-(b)-(c)-(d)** until convergence.

2.5 Selecting the number of blocks

BIC is an information criterion defined as an asymptotic approximation of the logarithm of the integrated likelihood ([S+78]). The standard case leads to write BIC as a penalised maximum likelihood:

$$\text{BIC} = -2 \max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) + D \log(n)$$

where n is the number of statistical units and D the number of free parameters and $l(\boldsymbol{\theta})$ defined in (5). Unfortunately, this approximation cannot be used for LBM, due to the dependency structure of the observations $(\mathbf{x}, \mathbf{y})$. However, a heuristic have been stated to define BIC in [KBC+12] and [KBCG15]. BIC-like approximations ICL lead to the following approximation as n and m tend to infinity

$$\text{BIC}(g,d) = -2 \max_{\boldsymbol{\theta}} \log f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}) + (g-1)\log n + \lambda \log n + (d-1)\log m + gd(p+1)\log(mn) \quad (14)$$

with λ the number of parameters of the $\mathbf{y}$ distribution. For LBM, the intractable likelihood $f(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})$ is replaced by the maximized free energy $\tilde{F}_C$ in (8) obtained by the BEM algorithm.

2.6 Measuring Influence of a Variable

Let j be fixed (a column of the matrix $\mathbf{x}$). We would like to measure the effect of the variable $\mathbf{x}^j = (x_{ij})_{i=1}^n$ on $\mathbf{y}$. It is possible to obtain a measure of this efffect by looking to the posterior probability of $\mathbf{y}$.

Lemma 1 *Let $(\mathbf{x}, \mathbf{z}, \mathbf{w})$ fixed. For $l = 1, \ldots, d$ let m_l denotes the number of columns with label l, i.e $m_l = \#\{w_{jl} = 1,\ j = 1, \ldots m\}$ and for a row i fixed let m_{il} denotes the number of elements such that $w_{jl} = 1$ and $x_{ij} = 1$, i.e. $m_{il} = \#\{w_{jl} x_{ij} = 1,\ j = 1, \ldots m\}$. The posterior probability of the co-variable $\mathbf{y}$ is*

$$\begin{aligned} f(\mathbf{y}|\mathbf{x}, \mathbf{z}, \mathbf{w}, \boldsymbol{\theta}) &\propto \prod_{i=1}^{n} \prod_{l=1}^{d} \pi_{z_i} \rho_l^{m_l} \, \text{logis}(\mathbf{y}_i^T \boldsymbol{\beta}_{z_i l})^{m_{il}} \left(1 - \text{logis}(\mathbf{y}_i^T \boldsymbol{\beta}_{z_i l})\right)^{m_l - m_{il}} \phi(\mathbf{y}_i; \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \\ &\propto \prod_{i=1}^{n} \pi_{z_i} \phi(\mathbf{y}_i; \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}) \prod_{l=1}^{d} \rho_l^{m_l} \frac{e^{m_{il} \mathbf{y}_i^T \boldsymbol{\beta}_{z_i l}}}{\left(1 + e^{\mathbf{y}_i^T \boldsymbol{\beta}_{z_i l}}\right)^{m_l}} \end{aligned} \quad (15)$$

Alternatively, for $k = 1, \ldots, g$, let n_k denotes the number of rows with label k, i.e. $n_k = \#\{z_{ik} = 1, i = 1, \ldots, n\}$. The posterior probability of the co-variable $\mathbf{y}$ is

$$f(\mathbf{y}|\mathbf{x}, \mathbf{z}, \mathbf{w}, \boldsymbol{\theta}) \propto \prod_{j=1}^{m} \rho_{w_j} \prod_{k=1}^{g} \pi_k^{n_j} \prod_{i:z_i=k} \text{logis}(\mathbf{y}_i^T \boldsymbol{\beta}_{k w_j})^{x_{ij}} \left(1 - \text{logis}(\mathbf{y}_i^T \boldsymbol{\beta}_{k w_j})\right)^{1 - x_{ij}} \phi(\mathbf{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \quad (16)$$

The proof of this lemma is straightforward and therefore omitted.

Assuming $\mathbf{z}$ and $\mathbf{w}$ known, we measure the influence of variable using its contribution to the posterior probability. Fixing j, taking the logarithm and eliminating terms independant of $\mathbf{x}^j$, we obtain the *influence measure criteria*

$$\begin{aligned} I(j) &= \log \rho_{w_j} + \sum_{i=1}^{n} x_{ij} \log \text{logis}(\mathbf{y}_i^T \boldsymbol{\beta}_{z_i w_j}) + \sum_{i=1}^{n} (1 - x_{ij}) \log \left(1 - \text{logis}(\mathbf{y}_i^T \boldsymbol{\beta}_{z_i w_j})\right) \\ &= \log \rho_{w_j} + \sum_{i=1}^{n} \left(x_{ij} \mathbf{y}_i^T \boldsymbol{\beta}_{z_i w_j} - \log(1 + \exp(\mathbf{y}_i^T . \boldsymbol{\beta}_{z_i w_j})) \right). \end{aligned} \quad (17)$$

Replacing the unknown labels w_j and z_i by there MAP estiamtors $\hat{w}_j$ and $\hat{z}_i$, we can sort the variables from the most to the less influential.

3 Examples

3.1 Simulated data

We compute 80 times the accuracy and the elapsed time of the model for various configurations of the parameter on a HP Zbook G3. The (averaged) computing time for different values of n is plotted in the figure below

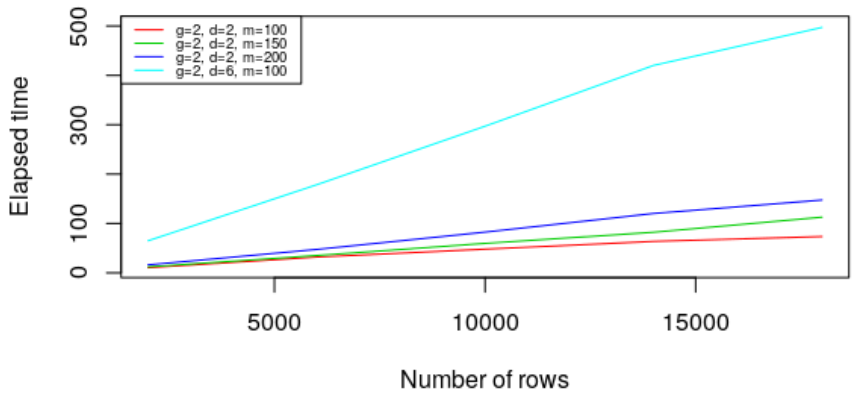

Figure 1: computational elapsed time for $n = 2000$, 6000, 10000, 14000 and 18000 (in minutes)

We can observe that as n grow the elapsed time grow linearly, but that the slope increases as d (the number of class in columns) is increased. The averaged well classified rate for these data is given in the table below with the standard deviation

d	m	n	Well classified rows	Well classified Columns
2	100	2000	0.9091250	0.9626562
2	100	6000	0.9027500	0.9444604
2	100	10000	0.9207500	0.9618788
2	100	14000	0.8928750	0.9451339
2	100	18000	0.8875000	0.9498715
2	150	2000	0.9030833	0.9773625
2	150	6000	0.9165000	0.9542896
2	150	10000	0.9358333	0.9732000
2	150	14000	0.9035833	0.9731152
2	150	18000	0.9408333	0.9771382
2	200	2000	0.9345000	0.9770188
2	200	6000	0.8940000	0.9568625
2	200	10000	0.9011875	0.9722437
2	200	14000	0.9170000	0.9663027
2	200	18000	0.9058125	0.9755889
6	100	2000	0.8535000	0.7367750
6	100	6000	0.8836250	0.7706229
6	100	10000	0.8920000	0.8069587
6	100	14000	0.8982500	0.7922634
6	100	18000	0.8565000	0.7899250

Table 1: Estimated Proportions of Well-Classified rows and columns for $g = 2$ and various configurations of d, m, n. Estimations were replicated 80 times.

3.2 Real Data Analysis

We consider the genetic data used by ([LGB+16]) and use this to compare the co-clustering with co-variable and the co-clustering without co-variable. The genetic data gave a 515721 SNPs and 455 individulas. The quantitative phenotype represents an individual falciparum attack(ipfa) for each individu([LGB+16]). The SNPs are coded in dominante.

The groups on the lines are divided in two parts: the susceptibility composed of a group of individuals with a positive IPFA and the resistant ones composed of a group of individuals with a negative IPFA value. We can take into account or not the mixture on the target variable in the proposed model.

In this part we are interested in the performance of the selection criterion BIC for choosing the number of partitions in columns. We have 515721 variables, we have evaluated from 2 to 30 partitions and the penalized BIC suggests 14 partitions in columns (cf figure 2).

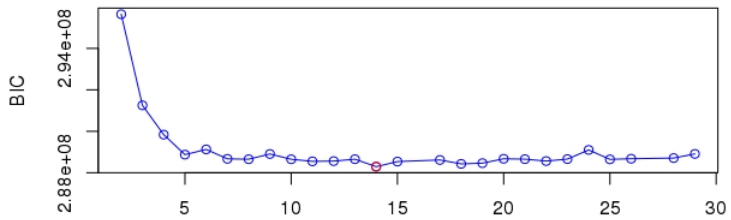

Figure 2: BIC Results

The proportion of mutations in each block are given below

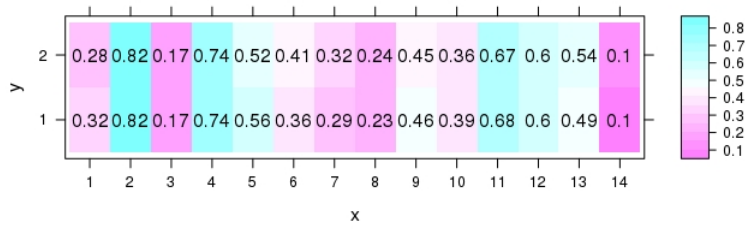

Figure 3: Percent of ones in each blocks

4 Conclusion

References

[BIG14] Parmeet Bhatia, Serge Iovleff, and Gérard Govaert. Blockcluster: An r package for model based co-clustering. 2014.

[Dhi01] 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*, KDD '01, pages 269–274, New York, NY, USA, 2001. ACM.

[DLR97] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data with the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1, 1997.

[GN03] Gérard Govaert and Mohamed Nadif. Clustering with block mixture models. *Pattern Recognition*, 36(2):463 – 473, 2003.

[GQX05] J. Guan, G. Qiu, and X.Y. Xue. Spectral images and features co-clustering with application to content-based image retrieval. In *Multimedia Signal Processing, 2005 IEEE 7th Workshop on*, pages 1–4. IEEE, 2005.

[Hat86] R.J. Hathaway. Another interpretation of the em algorithm for mixture distributions. *Statistics & Probability Letters*, 4(2):53–56, 1986.

[KBC+12] Christine Keribin, Vincent Brault, Gilles Celeux, Gérard Govaert, et al. Model selection for the binary latent block model. In *Proceedings of COMPSTAT*, volume 2012, 2012.

[KBCG15] Christine Keribin, Vincent Brault, Gilles Celeux, and Gérard Govaert. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6):1201–1216, 2015.

[LGB+16] Cheikh Loucoubar, Audrey V Grant, Jean-François Bureau, Isabelle Casademont, Ndjido Ardo Bar, Avner Bar-Hen, Mamadou Diop, Joseph Faye, Fatoumata Diene Sarr, Abdoulaye Badiane, et al. Detecting multi-way epistasis in family-based association studies. *Briefings in bioinformatics*, 18(3):394–402, 2016.

[MO04] Sara C Madeira and Arlindo L Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 1(1):24–45, 2004.

[NH98] R.M. Neal and G.E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. *NATO ASI SERIES D BEHAVIOURAL AND SOCIAL SCIENCES*, 89:355–370, 1998.

[Qiu04] G. Qiu. Image and feature co-clustering. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 991–994. IEEE, 2004.

[RV09] N. Rasiwasia and N. Vasconcelos. Holistic context modeling using semantic co-occurrences. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1889–1895. IEEE, 2009.

[S+78] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[XZDZ10] G. Xu, Y. Zong, P. Dolog, and Y. Zhang. Co-clustering analysis of weblogs using bipartite spectral projection approach. *Knowledge-Based and Intelligent Information and Engineering Systems*, pages 398–407, 2010.

A Computing the (rows and columns) E-Step

For the E-Step t_{ik} value maximize the fuzzy criterion given in equation (8). Derivative with respect to t_{ik} gives

$$\frac{\partial \tilde{F}_C(\mathbf{t}, \mathbf{r}; \boldsymbol{\theta})}{\partial t_{ik}} = \log \pi_k + \sum_{j,\ell} r_{j\ell} \log f_{k\ell}(x_{ij}, \mathbf{y}_i; \boldsymbol{\theta}) - \log t_{ik} - 1.$$

Equating this equation to zero, taking exponential and recalling that $\sum_k t_{ik} = 1$, we obtain that t_{ik} is updated as

$$t_{ik}^{(c+1)} = \frac{\pi_k^{(c)} \prod_{j,l} \left[f(x_{ij}, \mathbf{y}_i; \boldsymbol{\theta}^{(c)}) \right]^{r_{jl}^{(c)}}}{\sum_k \prod_{j,l} \left[f(x_{ij}, \mathbf{y}_i; \boldsymbol{\theta}^{(c)}) \right]^{r_{jl}^{(c)}}}.$$

For numerical reason, we prefer to compute the logarithm of this expression which is

$$\log(t_{ik}^{(c+1)}) \propto \log(\pi_k^{(c)}) + \sum_{j,l} r_{jl}^{(c)} \log f(x_{ij}, \mathbf{y}_i; \boldsymbol{\theta}^{(c)}).$$

Recall that (see equation 3)

$$\begin{aligned}\log f(x_{ij}|\mathbf{y}_i;\boldsymbol{\beta}_{kl}^{(c)}) &= x_{ij}\log(\text{logis}(\mathbf{y}_i^T\boldsymbol{\beta}_{kl}^{(c)})) + (1-x_{ij})\log(1-\text{logis}(\mathbf{y}_i^T\boldsymbol{\beta}_{kl}^{(c)}))\\ &= \log(1-\text{logis}(\mathbf{y}_i^T\boldsymbol{\beta}_{kl}^{(c)})) + x_{ij}\log\left(\frac{\text{logis}(\mathbf{y}_i^T\boldsymbol{\beta}_{kl}^{(c)})}{1-\text{logis}(\mathbf{y}_i^T.\boldsymbol{\beta}_{kl})}\right)\\ &= \log(1+\exp(\mathbf{y}_i^T\boldsymbol{\beta}_{kl}^{(c)})) + x_{ij}\mathbf{y}_i^T\boldsymbol{\beta}_{kl}^{(c)}\end{aligned}$$

giving

$$\log t_{ik}^{(c+1)} \propto \log\pi_k^{(c)} + \sum_{j,l} r_{jl}^{(c)} x_{ij}\mathbf{y}_i^T.\boldsymbol{\beta}_{kl}^{(c)} - \sum_l r_{.l}^{(c)}\log(1+e^{\mathbf{y}_i^T.\boldsymbol{\beta}_{kl}^{(c)}}) + m\ \log\phi(\mathbf{y}_i;\boldsymbol{\mu}_k^{(c)},\boldsymbol{\Sigma}_k^{(c)}).$$

Similar computation gives for r_{jl}

$$\log(r_{jl}^{(c+1)}) \propto \log\left(\rho_l^{(c)}\right) + \sum_{i,k} t_{ik}^{(c+1)}\left(x_{ij}\mathbf{y}_i^T\boldsymbol{\beta}_{kl}^{(c+1/2)} - \log\left(1+e^{\mathbf{y}_i^T.\boldsymbol{\beta}_{kl}^{(c+1/2)}}\right)\right).$$

Observe that the Gaussian distribution does not depend of j nor l. This term become constant when summing over i and k and disappears when r_{jl} values are normalized.

B Computing the M-Step

For the M-Step, we use a Newton-Raphson algorithm in order to solve the equation (10). For each pair (k,l) the function to maximize can be written

$$\ell_{k,l}(\boldsymbol{\beta}) = \sum_{i,j}\left(r_{jl}t_{ik}x_{ij}\mathbf{y}_i^T\boldsymbol{\beta} - r_{jl}t_{ik}\log(1+\exp(\mathbf{y}_i^T.\boldsymbol{\beta}))\right)$$

The first derivative with respect to the d-th coordinate β_d is

$$\frac{\partial\ell_{k,l}(\beta)}{\partial\beta_d} = \sum_{i,j}\left(r_{jl}t_{ik}x_{ij}y_{i,d} - r_{jl}t_{ik}y_{i,d}\frac{\exp(\mathbf{y}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{y}_i^T\boldsymbol{\beta})}\right)$$

giving the following expression for the gradient

$$\nabla_\beta\ell_{k,l}(\boldsymbol{\beta}) = Y^T D(X-\boldsymbol{\mu})$$

with $Y = [\mathbf{y}_i]_{i=1}^N$, $X = \left[\sum_j r_{jl}x_{ij}\right]_{i=1}^N$, $\boldsymbol{\mu} = \left[r_{.l}\frac{\exp(\mathbf{y}_i^T.\boldsymbol{\beta})}{1+\exp(\mathbf{y}_i^T\boldsymbol{\beta})}\right]_{i=1}^N$, $D = \text{diag}(t_{ik})_{i=1}^N$ The second derivative with respect to β_d and $\beta_{d'}$ is

$$\frac{\partial^2\ell_{k,l}(\boldsymbol{\beta})}{\partial\beta_d\partial\beta_{d'}} = -\sum_{i,j}\left(r_{jl}t_{ik}y_{i,d}y_{i,d'}\frac{\exp(\mathbf{y}_i^T\boldsymbol{\beta})}{(1+\exp(\mathbf{y}_i^T\boldsymbol{\beta}))^2}\right)$$

giving the following expression for the hessian

$$H_\beta = -Y^t DWY \qquad \text{with} \qquad W = \text{diag}\left(\frac{r_{.l}\ \exp(\mathbf{y}_i^T.\boldsymbol{\beta})}{(1+\exp(\mathbf{y}_i^T\boldsymbol{\beta}))^2}\right) = \text{diag}\left(r_{.l}\,\mu_i(1-\mu_i)\right)$$