Negative binomial latent block model with generalized constraints

Rodolphe Priam

Negative binomial latent block model with generalized constraints

R. Priam*

March 18, 2021

Abstract

The latent block model is a clustering method for rows and columns of a numerical matrix. Numerical experiments have shown that it is able to outperform other methods for several datasets when well estimated and well parameterized. For count data, the negative binomial distribution generalizes the poisson one with fitting improvements for excess of zeros. Herein, we propose to consider such distribution with the latent block model and a fitting algorithm which is suitable for a large sparse matrix such as a textual contingency table. Some constraints met in the literature are generalized via a regression setting in order to compute automatically the parameters.

Keywords: negative binomial distribution, latent block model, expectation maximization, regression

1 Introduction

Today, the data tables which are defined by a cross-product between two categorical variables are frequently met in data analysis. The purpose is to summarize in a comprehensible way their contents or retrieve in a fast way the main elements. These tables are called contingency tables, co-occurrence tables or count tables. They are processed for example in document retrieval [1], in clustering of texts [2] or in image segmentation. A cell contains the number of occurrences for a cross-category corresponding to a modality of each of the two variables. An example is the number of times a term occurs in a text, when the two variables are respectively a corpus and a vocabulary.

In such tables, I is the set of modalities for the rows (n categories) and J is in the same way defined for the columns (d categories). Generally a method which aims to analyse a table implies an evaluation of the relationships between the two variables I and J. For the analysis of such tables, correspondence analysis [3, 4] (CA) is an exploratory multivariate method which converts a data table into a particular type of graphical display. When the data matrix is large, a clustering [5] can give a quicker access to the data contents than a method for reducing the dimensionality of the features.

Some usual mass function considered in the literature are the Poisson one denoted $\mathcal{P}(.)$, the Restricted generalized Poisson one denoted $\mathcal{RGP}(.)$ and the Conway-Maxwell-Poisson (COM-Poisson) one denoted $\mathcal{CP}(.)$. when $Z(\lambda,\nu) = \sum_0^\infty \lambda^{x_{ij}}/(x_{ij}!)^\nu$ as a normalization constant, they are given in the following table:

	$P(x_{ij},\theta)$
$\mathcal{P}(.)$	$\frac{exp(-\lambda)(\lambda)^{x_{ij}}}{x_{ij}!}$
$\mathcal{RGP}(.)$	$\frac{\left(\frac{\lambda}{1+\kappa\lambda}\right)^{x_{ij}}(1+\kappa x_{ij})^{x_{ij}-1}}{x_{ij}!} e^{\left(\frac{-\lambda(1+\kappa x_{ij})}{1+\kappa\lambda}\right)}$
$\mathcal{CP}$	$\frac{\lambda^{x_{ij}}}{(x_{ij}!)^\nu Z(\lambda,\nu)}$

The COM-Poisson distribution is increasingly considerd in the literature in order to outperform the usual Poisson model for regression model but also other modeling not considered herein. NBM is often studied in biostatistics but is only for over-dispersed data hence, an alternative approach may be considered with more flexibility.

The family of co-clustering [6] methods makes possible to reveal the hidden association between the rows and the columns of a data table. A simultaneous partitioning treats symmetrically the table on the contrary to a clustering of just one dimension. Algorithms were introduced earlier in the literature in [6] and [7, 8]. There is also the information-theoretic co-clustering method

*rpriam@gmail.com

[9] and its generalization [10]. Other approaches are for instance the general method for prediction [11] or the non negative matricial decompositions [2, 12, 13]. In [14, 15], a block mixture model and its inference by a variational generalized EM algorithm are introduced, see also [16].

In this paper, it is proposed to consider a co-clustering model with an alternative to the Poisson distribution for the blocks. In co-clustering of textual data, it may be preferred to add contraints to the parameters, see [17, 18] for pioneer research for count data, and [19] more recently. Hence this framework is generalized herein in order to make it more pratical and avoid a different implementation for each constraint. The plan of the paper is as follows. Section 1 is the introduction to the purpose. Section 2 presents the proposed model and its inference via an expectation maximization. Section 3 presents an algorithm for the inference of the parameters by maximum likelihood via two approximations. Section 4 proposes generalized constraints via a new approach. Section 5 concludes with perspectives.

2 Co-clustering model

A brief review of the Block Latent Model (LBM) and its Poisson version (PLBM) are presented as the foundation of our proposal.

2.1 Model and Loglikelihood

Within the context of the classical mixture model, a partition of I into g clusters is represented by the binary classification matrix $\mathbf{z} = (z_{ik})_{n\times g}$ such that $\sum_{k=1}^{g} z_{ik} = 1$ and $z_{ik} = 1$ indicates the component of the row i. Just as I is partitioned into g clusters, columns can be partitioned into m clusters by the binary classification matrix $\mathbf{w} = (w_{j\ell})_{d\times m}$. If the most usual clustering methods deal with clustering of only the set I or eventually J, co-clustering is interested in the clustering of both. The $n\times d$ random variables generating the observed x_{ij} cells of the data matrix are assumed to be independent in LBM, once $\mathbf{z}$ and $\mathbf{w}$ are fixed. The set of all possible assignments $\mathbf{w}$ of J (resp. $\mathbf{z}$ of I) is denoted $\mathcal{W}$ (resp. $\mathcal{Z}$). The data table $\mathbf{x}$ is therefore a set of cells $(x_{11}, x_{12}, \ldots, x_{ij}, \ldots, x_{nd})$. The two sets of possible assignments $\mathbf{w}$ and $\mathbf{z}$ aggregate the cells of the matrix $\mathbf{x}$ into a number of contiguous, non-overlapping blocks. The following decomposition is obtained [15] by independence of $\mathbf{z}$ and $\mathbf{w}$, by summing over all the assignments $\mathcal{Z}\times\mathcal{W}$:

$$f_{LBM}(\mathbf{x};\theta) = \sum_{(\mathbf{z},\mathbf{w})} \prod_{i,k} p_k^{z_{ik}} \prod_{j,\ell} q_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi(x_{ij};\theta_{k\ell})^{z_{ik}w_{j\ell}} .$$

Here $\varphi(.;\theta_{k\ell})$ is a probability function defined on the real line $\mathbb{R}$ and $\{\theta_{k\ell}\}$ are unknown parameters. The vectors of the probabilities p_k and q_ℓ that a row and a column belong to the k^{th} component and to the ℓ^{th} component are respectively denoted $\mathbf{p} = (p_1, \ldots, p_g)$ and $\mathbf{q} = (q_1, \ldots, q_m)$. The set of parameters is denoted θ and is compound of $\mathbf{p}$ and $\mathbf{q}$ plus α which aggregates the $g\times m$ scalar $\alpha_{k\ell}$. The set of parameters θ of the model can be estimated by maximizing the log-likelihood:

$$L(\theta;\mathbf{x}) = \log f_{LBM}(\mathbf{x};\theta) .$$

A particular distribution is introduced next paragraph for contingency tables.

2.2 Objective function and optimization

For the proposed model with the introduced constraints, we aim to address the problem of the estimation of the parameters by a maximum likelihood (ML) approach such that:

$$\hat{\theta} = argmax_\theta L(\theta;\mathbf{x}) .$$

Let us focus on the estimation of a value of θ by the maximum likelihood approach associated to the block mixture model. For this model, the complete data are $(\mathbf{x},\mathbf{z},\mathbf{w})$ where the unobservable vectors $\mathbf{z}$ and $\mathbf{w}$ are the labels. The complete log-likelihood of $(\mathbf{x},\mathbf{z},\mathbf{w})$ leads to an algorithm EM where the E-step is intractable, thus a related algorithm has been proposed in the literature.

2.3 BEM algorithm

An approach based on a Generalized EM and a variational approximation by the product $c_{ik}^{(t)} d_{j\ell}^{(t)}$ has been proposed [15] previously in the literature, and named *Block EM* (BEM). Here it is denoted the variational probabilities c_{ik} such that $\sum_k c_{ik} = 1$, and also $d_{j\ell}$

such that $\sum_\ell d_{j\ell} = 1$. Their matricial representations are respectively $\mathbf{c} = (c_{ik})$, and $\mathbf{d} = (d_{j\ell})$, such that the variational distribution for the clustering is defined as $Q_{(\mathbf{c},\mathbf{d})}(\mathbf{z},\mathbf{w}) = \prod_{i,k}(c_{ik})^{z_{ik}} \prod_{j,\ell}(d_{j\ell})^{w_{j\ell}}$. Then, by the Jensen inequality a bound $\mathcal{F}(\mathbf{c},\mathbf{d};\theta)$ can be defined. The algorithm proceeds by defining a lower bound of the log-likelihood (see [15]) and repeats until convergence the two following steps:

E-step The posterior probabilities $\mathbf{e} = (\mathbf{c},\mathbf{d})$ are found at the current time (with the normalizing constraint to one). By maximizing $\mathcal{F}$ with respect to c_{ik} and $d_{j\ell}$. the resulting posterior probabilities are estimated with the dependent equations:

$$\begin{aligned} c_{ik}^{(t)} &\propto p_k^{(t)} \exp\left(\sum_{j,\ell} d_{j\ell}^{(t)} \log\varphi(x_{ij};\theta_{k\ell}^{(t)})\right), \\ d_{j\ell}^{(t)} &\propto q_\ell^{(t)} \exp\left(\sum_{i,k} c_{ik}^{(t)} \log\varphi(x_{ij};\theta_{k\ell}^{(t)})\right). \end{aligned}$$

Here the probabilities are hence obtained as a solution of the fixed point relations after initializing with previous values.

M-step A temporary value of the parameters is found at the new current time. By maximizing $\mathcal{F}$ with respect to θ, the objective function to maximize is:

$$\begin{aligned} \tilde{Q}_{LBM}(\theta,\theta^{(t)}) &= \textstyle\sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)} \log\varphi(x_{ij};\theta_{k\ell}) \\ &\quad + \textstyle\sum_{i,k} c_{ik}^{(t)} p_k + \sum_{j,\ell} d_{j\ell}^{(t)} q_\ell \,. \end{aligned}$$

Here, the posterior probabilities $c_{ik}^{(t)}$ and $d_{j\ell}^{(t)}$ are available from E-step, this results into the criterion also denoted $\tilde{Q}$. For $k = 1,...,g$ and $\ell = 1,...,m$, it is denoted the aggregated statistics, $y_{k\ell}^{(t)} = \sum_{i,j} c_{ik}^{(t)} d_{j\ell}^{(t)} x_{ij}$, $\mu_k^{(t)} = \sum_i c_{ik}^{(t)} \mu_i$, $\nu_\ell^{(t)} = \sum_j d_{j\ell}^{(t)} \nu_j$. Given $\theta^{(t)}$, the quantities $c_{ik}^{(t)}$ (resp. $d_{j\ell}^{(t)}$) are the posterior probabilities that a row (resp. a column) belongs to the block $k\ell$. When solving for maximizing (1), it is written:

$$\theta^{(t+1)} = argmax_\theta \tilde{Q}_{LBM}(\theta,\theta^{(t)}).$$

The solution for the mixing coefficients is also obtained as,

$$p_k^{(t+1)} = n^{-1} \sum_i c_{ik}^{(t)} \text{ and } q_\ell^{(t+1)} = d^{-1} \sum_j d_{j\ell}^{(t)},$$

respectively if they were not taken constant. Hence, the parameters are estimated by an iterative way, the BEM algorithm proceeds by an alternated maximization of $\tilde{Q}$ and converges to a final solution which maximizes (locally) the log-likelihood of the latent block model. A hat is added to each parameter or statistics which is estimated and found at the final stage of the inference algorithm.

2.4 Cell modeling

In this section, it is considered several distributions for the cells.

- **Poisson distribution**: This distribution $\mathcal{P}(.)$ is with parameter $\lambda_{k\ell}^{ij}$ for the expectations in [20]. This leads to assume that the observed values x_{ij} in a block $k\ell$ are drawn with:

  $$\lambda_{k\ell}^{ij} = \mu_i \nu_j \alpha_{k\ell},$$

  with $\theta_{k\ell} = \alpha_{k\ell}$, the effects $\mu = (\mu_1,\ldots,\mu_n)$ and $\nu = (\nu_1,\ldots,\nu_d)$. They are assumed equal to the following constant margin totals by rows and by columns, $\mu_i = \sum_j x_{ij}$ for $i \in I$ and $\nu_j = \sum_i x_{ij}$ for $j \in J$. Then the probability function φ for the block $k\ell$ is defined as follows:

  $$\varphi(x_{ij};\theta_{k\ell}) = \frac{exp(-\lambda_{k\ell}^{ij})(\lambda_{k\ell}^{ij})^{x_{ij}}}{x_{ij}!}.$$

  In the case of the Poisson distribution, the solution for α at the maximization step can be written $\alpha_{k\ell}^{(t)}$ as a scalar or in a matricial notation otherwise.

- **Negative binomial distribution**: For overdispertion, the negative binomial probability mass function correct the Poisson one. Here, $\mathcal{NB}(\kappa,\mu_i)$ denotes a negative binomial distribution with parameters,

  $$\theta_{k\ell} = (\kappa_{k\ell},\alpha_{k\ell}),$$

  with parameter $\kappa_{k\ell}$. When $\Gamma(\cdot)$ is the gamma function and,

  $$C_{k\ell}^{\kappa_{k\ell}} = \frac{\Gamma(\kappa_{k\ell}+x_{ij})}{\Gamma(\kappa_{k\ell})x_{ij}!},$$

the model is written for $i = 1, 2, \cdots, n$ with

$$\begin{aligned} & \varphi(x_{ij}; \theta_{k\ell}) \\ = \; & C_{k\ell}^{\kappa_{k\ell}} \left[\frac{\kappa_{k\ell}}{\kappa_{k\ell}+\lambda_{k\ell}^{ij}} \right]^{\kappa_{k\ell}} \left[\frac{\lambda_{k\ell}^{ij}}{\kappa_{k\ell}+\lambda_{k\ell}^{ij}} \right]^{x_{ij}} \\ = \; & \frac{\Gamma(\kappa_{k\ell}+x_{ij})}{\Gamma(\kappa_{k\ell})\, x_{ij}!} e^{-\kappa_{k\ell}\lambda_{k\ell}^{ij} - \kappa_{k\ell}\gamma_{ij}^{k\ell}} \sigma(\lambda_{k\ell}^{ij} + \gamma_{ij}^{k\ell})^{x_{ij}+\kappa_{k\ell}} \, . \end{aligned}$$

Here $\gamma_{ij}^{k\ell}$ denotes $\ln \kappa_{k\ell} + \gamma^{k\ell} + \gamma_{ij}$ where $\gamma^{k\ell}$ and γ_{ij} are eventual offsets. A limit of this model is to help only with overdispersion, but it is known to be able to reduce the problem with excess of zeros, while counting the Poisson model as a particular case.

Alternative distributions such as restricted generalized Poisson or Conway-Maxwell-Poisson (COM-Poisson) distributions are not considered here, neither zero-inflated models. Next the estimation of the paramters is discussed.

3 Inference

In the case of the negative binomial model, the non linearity asks for approximating the likelihood during the maximization step while the solution was directly in closed form for the poissonian one.

3.1 Surrogate objective function

In this subsection, it is considered the maximization step as follows, in order to find the new current value $\theta^{(t+1)}$ by the approximation. For a variational approach (VR), the function φ may be approximated for instance. For NBM, a bound on the sigmoid function [21] from the logit modeling may be relevant. By convexity, this is written:

$$\sigma(a) \quad \geq \quad \sigma(\varepsilon) \exp\left(\tfrac{1}{2}(a - \varepsilon) - \lambda(\varepsilon)(a^2 - \varepsilon^2) \right) \, ,$$

where $a \in \mathbb{R}$ while $\varepsilon \in \mathbb{R}$ is the variational parameter, and $\lambda(\varepsilon) = \frac{1}{4\varepsilon} tanh\left(\frac{\varepsilon}{2}\right)$. This induces that the parameter ε has to be estimated for maximizing the approximating function. Let's denote the vector from each variational parameter ε_{ij} from each cell,

$$\varepsilon \quad = \quad (\varepsilon_{11}, \varepsilon_{12}, \cdots, \varepsilon_{np})' \, .$$

This parameter is introduced next with the bound on the sigmoidal function in order to approximate the usual $\tilde{Q}$ function.

By using the bound on the sigmoid, and when denoting $\sigma_{ij}^{k\ell} = \sigma(\varepsilon_{ij})$ and,

$$a_{ij}^{k\ell} = \lambda_{k\ell}^{ij} + \gamma_{ij}^{k\ell} \, ,$$

the criterion to optimize may be written

$$\begin{aligned} & \tilde{Q}_{LBM}(\theta, \theta^{(t)}) \\ \geq \; & \textstyle\sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)} \log C_{k\ell}^{\kappa_{k\ell}} \\ - \; & \textstyle\sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)} \left\{ \kappa_{k\ell}\lambda_{k\ell}^{ij} + \kappa_{k\ell}\gamma_{ij}^{k\ell} - \log \sigma_{ij}^{k\ell} \right\} \\ + \; & \textstyle\sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)} \left(0.5(a_{ij}^{k\ell} - \varepsilon_{ij}^{k\ell}) \right) \\ - \; & \textstyle\sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)} \left(\lambda(\varepsilon_{ij}^{k\ell})((a_{ij}^{k\ell})^2 - (\varepsilon_{ij}^{k\ell})^2) \right) \\ = \; & \tilde{Q}_{LBM}(\theta, \varepsilon, \theta^{(t)}) \, . \end{aligned}$$

The variational approximation changes the maximization of a multidimensional nonlinear function into several simple univariate minimization problems and the maximization of a quadratic form which can be performed analytically. Note that this bound is enough small when the variational parameters are well chosen in order to be able to retrieve the curve of the sigmoid function in a vicinity of the current value of the parameters.

3.2 Maximization step

The optimization is then performed as follows:

- First, the new bounding objective function is optimized with respect to each ε_{ij}.

  $$\frac{\tilde{Q}_{LBM}(\theta,\varepsilon,\theta^{(t)})}{\partial \varepsilon_{ij}^{k\ell}} \propto \frac{\partial \lambda(\varepsilon_{ij}^{k\ell})}{\partial \varepsilon_{ij}^{k\ell}} ((a_{ij}^{k\ell})^2 - (\varepsilon_{ij}^{k\ell})^2) \, .$$

  The solution is finally obtained because in this case the first term is increased as a function of $\varepsilon_{ij}^{k\ell}$, and the variational approximation is symmetric, such that it can be written:

  $$\varepsilon_{ij}^{k\ell} = |a_{ij}^{k\ell}| \, .$$

- Second, knowing this value, we can maximize the objective function with respect to $\theta_{k\ell}$. This can be solved for $\alpha_{k\ell}$ in closed form with $\beta_{ij} = \mu_i \nu_j$, as the zero of:

  $$\begin{aligned} & \frac{\partial \tilde{Q}_{LBM}(\theta,\varepsilon,\theta^{(t)})}{\partial \alpha_{k\ell}} \\ = \; & \frac{\partial \sum_{i,j} c_{ik}^{(t)} d_{j\ell}^{(t)} \left(0.5(a_{ij}^{k\ell} - \lambda(\varepsilon_{ij}^{k\ell})(a_{ij}^{k\ell})^2 \right)}{\partial \alpha_{k\ell}} \\ = \; & \sum_{i,j} c_{ik}^{(t)} d_{j\ell}^{(t)} \left[\frac{\beta_{ij}}{2} - \lambda(\varepsilon_{ij}^{k\ell})(2\beta_{ij}^2 \alpha_{k\ell} + 2\beta_{ij}\gamma_{ij}^{k\ell}) \right] . \end{aligned}$$

Thus, this is written as follows:

$$\alpha_{k\ell}^{(t+1)} \quad = \quad \frac{\sum\limits_{i,j} c_{ik}^{(t)} d_{j\ell}^{(t)} \left[\beta_{ij}+4\lambda(\varepsilon_{ij}^{k\ell})\beta_{ij}\gamma_{ij}^{k\ell})\right]}{4\sum\limits_{i,j} c_{ik}^{(t)} d_{j\ell}^{(t)} \left[\lambda(\varepsilon_{ij}^{k\ell})\beta_{ij}^{2}\right]} .$$

As expected, the expression is positive because it is related to the expectations of count data. Hence the absolute value is not required when solving for $\varepsilon_{ij}^{k\ell}$ if $\gamma^{k\ell} > 0$ and $\gamma_{ij} > 0$ or even equal to zero as usually supposed.

Note that a simple expression for a current value of the parameter is available analytically in order to increase the objective function by updating a numerical value of the parameters. The Newton-Raphson algorithm is an alternative not considered herein, and left as a perspective.

3.3 CEM for large sparse matrices

In the case of large matrix, without a supercomputer and just a simple laptop, it may be relevant to be cautious with the expression of the optimized parameters. In particular, the zero must be take care separately here as we explain in this subsection. By this way, the expression for $\alpha_{k\ell}^{(t)}$, $c_{ik}^{(t)}$, and $d_{j\ell}^{(t)}$ becomes possible in practice for large sparse matrices. This is treated here as this may be useful in future for other related models.

To begin with, it is rewritten the parameter for the expectations as follows:

$$\alpha_{k\ell}^{(t+1)} \quad = \quad \frac{\mu_k^{(t)}\nu_\ell^{(t)}}{4\,\breve{\breve{\alpha}}_{k\ell}^{(t)}} + \frac{\breve{\alpha}_{k\ell}^{(t)}\ \ln\kappa_{k\ell}}{\breve{\breve{\alpha}}_{k\ell}^{(t)}} .$$

Here $\mu_k^{(t)} = \sum_i c_{ik}^{(t)}\mu_i\ \nu_\ell^{(t)} = \sum_j d_{j\ell}^{(t)}\nu_j$ as for the poissonian case, and it is supposed $\gamma^{k\ell} = 0$ and $\gamma_{ij} = 0$ without loss of generality. It may be noticed that $\varepsilon_{ij}^{k\ell} = \mu_i\nu_j\alpha_{k\ell}$, thus the matrix $\Lambda_{k\ell} = (\lambda(\varepsilon_{ij}^{k\ell})$ is dense for each (k,ℓ) hence it is not possible to compute for large contingency tables. One approach is proposed in order to reduce this limit to the use of the negative binomial distribution for co-clustering in the considered setting. In the classification expectation maximization (CEM) the smooth probabilities c_{ik} and $d_{j\ell}$ are replaced by binary variables $z_{ik}^{(t)}$ and $w_{j\ell}^{(t)}$ which have for value the indicator to belong to each corresponding cluster of rows or columns. With these quantites replacing $c_{ik}^{(t)}$ and $d_{j\ell}^{(t)}$, the sums involved are now not for the all set of rows and columns but only the rows and columns which belong to the block in stake. Hence, when $\Lambda_{k\ell}^{+(t)}$ is the matrix $\Lambda_{k\ell}$ made sparse by keeping only the cells with $c_{ik}^{(t)}d_{j\ell}^{(t)} \neq 0$, $z_i^{(t)}$ (resp. $w_j^{(t)}$) is the vector with the binary labels $z_{ik}^{(t)}$ (resp. $w_{j\ell}^{(t)}$), D_μ the diagonal matrix with not null elements μ_i, and D_ν the diagonal matrix with not null elements ν_j, the quantities above may rewritten:

$$\begin{aligned}
\breve{\alpha}_{k\ell}^{(t)} \quad &= \quad z_i^{(t)T} D_\mu \Lambda_{k\ell}^{+(t)} D_\nu w_j^{(t)} \\
\breve{\breve{\alpha}}_{k\ell}^{(t)} \quad &= \quad z_i^{(t)T} D_\mu^2 \Lambda_{k\ell}^{+(t)} D_\nu^2 w_j^{(t)} .
\end{aligned}$$

Note that for each block the central sparse matrix with the quantities $(\lambda(\varepsilon_{ij}^{k\ell})$ is also computed in a matricial way in order to avoid loops from programming langage such that with R or Octave. This induces roughly a burden of adding a complexity of $g \times m$ in the algorithm for the parameter estimation in comparison to the Poisson case, but which is kept dramatically less than multiply by $n \times d$ the complexity. This keeps to handle only sparse matrix or tall dense matrices without the bottleneck from large dense ones. Similarly, the estimation step is rewritten in a matricial way with the binary matrices $\mathbf{z}^{(t)} = (z_{ik}^{(t)})$ and $\mathbf{w}^{(t)} = (x_{j\ell}^{(t)})$ for reducing the numerical burden. For large matrices, the matrix aggregating the posterior probabilities are computed for subsets of rows and columns separately in order to avoid memory problem. Note also that a SEM-Gibbs algorithm introduced by [22] would lead to such discrete solution without much effort except the need for a fast implementation because it may be slow in practice with the generation of random variables, thus left as a perspective. Next it is discussed how update the structure of the parameters in order to look for a reduced rank matrix for the parameters.

3.4 Model selection

For choosing g and m, it was recently presented a criterion related to integrated Completed Loglikelihood (ICL) from usual mixture models, for count data. When denoting the complete likelihood $p(y,\mathbf{z},\mathbf{w};\theta)$ after maximization, this criterion may be written from [22], as after lengthy calculus, with r a number of parameters:

$$\begin{aligned} ICL_{\text{LBM}}(g,m) &= \log p(y,\hat{z},\hat{w};\hat{\theta}) \\ &- \log n^{\frac{g-1}{2}+\frac{gm(r-1)}{2}} \\ &- \log d^{\frac{m-1}{2}+\frac{gm(r-1)}{2}} . \end{aligned}$$

Note that on the contrary to the usual mixture model the criterion seems to be treated separately according to the distribution of the cell. The maximum is chosen as the best model in practice, by having g and m varying in an interval. During the experiments, it was observed that eventually the maximum likelihood may not always correspond to the maximum accuracy for some dataset, hence an alternative approach may be preferred sometimes.

4 Generalized constraints

Contraints for latent block clustering appeared after 2012 in the literature, this kind of approach is found in some other models such that gaussian mixture for the covariance (with often zeros off diagonal). For the models considered in the literature, this suggests to have some parameters which are identical, with $\alpha_{k\ell} = \tau$ for some common value $\tau > 0$. Note that eventually, it may be chosen several constants for different constraints but this seems not be tested currently. An alternative interpretation is that the parameters $\alpha_{k\ell}$ depend on a reduced space such that identifying this subspace allows to avoid overfitting for the clustering model. This becomes related to rank reduction in a non negative setting as explained in this section.

In order to induce the constraints for the parameters, we propose to rewrite $\alpha_{k\ell}$ with a regression, which leads to an automatic treatement during the inference. Note that this expression is mostly relevant when the M-step involves a quadratic form in order to solve analytically for a solution but non linear algorithm may be possible.

New parameters

The proposed model includes the constraints embedded in the center parameters as:

$$\alpha_{k\ell} = \tau^T \xi_{k\ell} .$$

Here, $\xi_{k\ell}$ are constant vector easy to define, depending on the given constraints and aggregated in a new matrix,

$$\Omega = [\xi_{11}|\xi_{22}|\cdots|\xi_{gm}] ,$$

It is also defined τ for a latent unknown vector with the unique scalar parameters in α, and related to regression coefficients: it contents is for the values of the common and the unique parameters. For instance, with the same examples as before, the matrix is written as follows.

Example 1 *(see [17, 18]). First for the matrix of parameters with constraints from the sparse latent block model is written:*

$$\alpha_{(1)} = \begin{pmatrix} \alpha_{11} & \tau & \tau \\ \tau & \alpha_{22} & \tau \\ \tau & \tau & \alpha_{33} \end{pmatrix} .$$

The corresponding matrix for the constraint is defined as follows,

$$\Omega_{(1)} = \begin{pmatrix} 1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1 \\ 0&1&1&1&0&1&1&1&0 \end{pmatrix}$$

□

Example 2. *With τ_1 and τ_2 two common parameters, the less constrained example is with the matrix of parameters,*

$$\alpha_{(1')} = \begin{pmatrix} \alpha_{11} & \tau_1 & \tau_1 \\ \tau_2 & \alpha_{22} & \tau_1 \\ \tau_2 & \tau_2 & \alpha_{33} \end{pmatrix}$$

This leads to write the new matrix and vector for the constraints as follows,

$$\Omega_{(1')} = \begin{pmatrix} 1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1 \\ 0&1&1&1&0&0&0&0&0 \\ 0&0&0&0&0&1&1&1&0 \end{pmatrix}$$

□

Optimization

With this new parameterization, the previous maximization step for updating $\alpha_{k\ell}$ at time (t) of the M-step is not anymore relevant. The optimization is for the regression coefficients τ, hence, the function to optimize is rewritten as follows, when writing cte for the constant non depending on the unknown parameters,

$$\begin{aligned} & \tilde{Q}_{LBM}(\theta,\varepsilon,\theta^{(t)}) \\ = \; & \sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)} 2\mu_i \nu_j \,(0.25 + \log \kappa_{k\ell})\, \tau^T \xi_{k\ell} + cte \\ - \; & \sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)} \mu_i^2 \nu_j^2 \lambda(\varepsilon_{ij}^{k\ell})\, \tau^T \xi_{k\ell} \xi_{k\ell}^T \tau . \end{aligned}$$

Thus the derivative leads to the solution after equating with zero,

$$\begin{aligned}&\frac{\partial\tilde{Q}_{LBM}(\theta,\varepsilon,\theta^{(t)})}{\partial\tau}\\ =\ & 2\sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)} \mu_i\nu_j\,(0.25+\log\kappa_{k\ell}\xi_{k\ell})\\ -\ & 2\sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)}\left[\mu_i^2\nu_j^2\lambda(\varepsilon_{ij}^{k\ell})\xi_{k\ell}\xi_{k\ell}^T\tau\right].\end{aligned}$$

A solution is thus written:

$$\begin{aligned}\tau^{(t+1)} =\ & \left[\sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)}\mu_i^2\nu_j^2\lambda(\varepsilon_{ij}^{k\ell})\xi_{k\ell}\xi_{k\ell}^T\right]^{-1}\\ & \left(\sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)}\mu_i\nu_j(0.25+\log\kappa_{k\ell})\xi_{k\ell}\right).\end{aligned}$$

But the constraint of positivity is required for τ which may not be true with this analytical expression. Note that for some other distribution, such as gaussian, or with a non linear transformation, such as sigmoidal, the positivity is not required. As for the scalar-wise expression, it may be preferred to change the posterior probabilities with the hard assignements to the clusters in order to avoid a large dense matrix for textual data for instance. The method is able to provide an analytical solution in closed form for any kind of linear constraint on the parameters. An explicit expression for τ is also available from the proposed approach when solving for the restricted regression problem.

Non negative restricted regression

The new criterion is rewritten as,

$$\begin{aligned}&\tilde{Q}_{LBM}(\theta,\Omega,\varepsilon,\theta^{(t)},\Omega^{(t)})+cte\\ =\ & -\sum_{i,j,k,\ell}\frac{c_{ik}^{(t)}d_{j\ell}^{(t)}}{\sqrt{\lambda(\varepsilon_{ij}^{k\ell})}^{-1}}\left[\mu_i\nu_j\xi_{k\ell}^T\tau-(0.25+\log\kappa_{k\ell})\right]^2.\end{aligned}$$

This underlines how the parameter quantity $\log\kappa_{k\ell}$ helps for a better fitting: by reducing the noise of this weighted regression problem. The constant 1 is such as the difference if two small or two large and must be corrected, which justifies the chosen distribution instead of the usual poissonian one.

To finish, this also suggests for an eventual analytical optimization of the approximated likelihood w.r.t. Ω in order to increase further the accuracy of the clustering, and separate at best the classes for visualization purposes as pioneering research made it possible:

$$\frac{\partial\tilde{Q}_{LBM}(\theta,\Omega,\varepsilon,\theta^{(t)},\Omega^{(t)})}{\partial\Omega}=0.$$

This illustrates further a generalized parameterization but only numerical results may validate this additional new model for real textual data, otherwise an alternative way may be required such as the not automatic one.

5 Discussion and perspectives

Herein it is explained how generalized constraints may be added to a latent block model via regressions, with α written as a function of the product of two matrices τ and Ω. The negative binomial distribution is considered as a way to improve the fitting in comparison to the usual Poisson one. Algorithms are proposed for the estimation with and without contraints from contingency table having eventually large dimensions.

To our knowledge this is a new approach for large contingency tables and their exploratory analysis. The only closely related model independently developped is for biological data[1] [23] with a bayesian approach for the inference and not for textual data hence was not considered further. A main perspective is the visualization of textual data by extending on the methods presented herein and including more constraints. This justifies this communication on clustering instead of reduction as a preamble, before a projective parameterization: for non linear mapping, $\alpha_{k\ell}$ is a function of two vectors ξ_k and ξ^ℓ or c_{ik} a function of $||\xi_{(i)}-\xi_k||^2$.

References

[1] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, pp. 391–407, 1990.

[2] T. Hofmann, "Probabilistic latent semantic analysis," *SIGIR'99*, pp. 50–57, 1999.

[1] https://cran.r-project.org/web/packages/cobiclust/index.html

[3] J. P. Benzecri, *L'analyse des données tome 1 et 2 : l'analyse des correspondances*. Paris:Dunod, 1980.

[4] M. Greenacre, *Theory and Applications of Correspondence Analysis*. London: Academic Press, 1983.

[5] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[6] J. A. Hartigan, *Clustering Algorithms*. New York: Wiley, 1975.

[7] G. Govaert, "Classification croisée," Thèse d'État, Université Paris 6, France, 1983.

[8] ——, "Simultaneous clustering of rows and columns," *Control and Cybernetics*, vol. 24, no. 4, pp. 437–458, 1995.

[9] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering," in *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003)*, 2003, pp. 89–98.

[10] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha, "A generalized maximum entropy approach to bregman co-clustering and matrix approximation," *J. Mach. Learn. Res.*, vol. 8, pp. 1919–1986, Dec. 2007.

[11] D. Agarwal and S. Merugu, "Predictive discrete latent factor models for large scale dyadic data," in *KDD*. ACM, 2007, pp. 26–35.

[12] C. Ding, T. Li, and W. Peng, "On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing," *Computational Statistics and Data Analysis*, vol. 52, pp. 3913–3927, 2008.

[13] D. Cai, X. Wang, and X. He, "Probabilistic dyadic data analysis with local and global consistency," in *ICML*, 2009.

[14] G. Govaert and M. Nadif, "Clustering with block mixture models," *Pattern Recognition*, vol. 36, pp. 463–473, 2003.

[15] ——, *Co-Clustering*. John Wiley & Sons, Ltd, 2013.

[16] V. Brault, C. Keribin, and M. Mariadassou, "Consistency and asymptotic normality of Latent Block Model estimators," *Electronic Journal of Statistics*, vol. 14, no. 1, pp. 1234 – 1268, 2020.

[17] M. Ailem, F. Role, and M. Nadif, "Sparse poisson latent block model for document clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 7, pp. 1563–1576, 2017.

[18] M. Ailem, F. Role, and M. Nadif, "Model-based co-clustering for the effective handling of sparse data," *Pattern Recognition*, vol. 72, pp. 108–122, 2017.

[19] M. Selosse, J. Jacques, and C. Biernacki, "Textual data summarization using the self-organized co-clustering model," *Pattern Recognition*, vol. 103, p. 107315, 2020.

[20] G. Govaert and M. Nadif, "Latent block model for contingency table," *Communications in Statistics-theory and Methods*, vol. 39, pp. 416–425, 2010.

[21] L. K. Saul, T. Jaakkola, and M. I. Jordan, "Mean field theory for sigmoid belief networks," *Journal of Artificial Intelligence Research*, vol. 4, pp. 61–76, 1996.

[22] C. Keribin, V. Brault, G. Celeux, and G. Govaert, "Estimation and selection for the latent block model on categorical data," *Statistics and Computing*, vol. 25, no. 6, pp. 1201–1216, Nov 2015.

[23] J. Aubert, S. Schbath, and S. Robin, "Model-based biclustering for overdispersed count data with application in microbial ecology," *Methods in Ecology and Evolution*, February 2021.