Negative binomial latent block model with generalized constraints*

R. Priam†

November 19, 2024

Abstract

Constrained latent block models (LBM) are presented in this communication for contingency matrices. With count data, several distributions extend the Poisson one with fitting improvements for excess of zeros among the treated issues. Herein, it is proposed to consider such distributions with LBM and fitting algorithms which are suitable for large matrices such as a textual ones. Generalized constraints are introduced or extended to LBM in order to infer automatically the parameters of new models for co-clustering and visualization, with sparsity and latent reduced vectors. The negative binomial and Hurdle Poisson distribution are compared to the usual Poisson one. The proposed new models and algorithms are validated in the numerical experiments with several simulated and real textual contingency tables for co-clustering and non linear projection of the rows and columns.

Keywords: latent block model, contingency table, lasso, pca

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. 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 [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 latent block mixture model and algorithms are introduced, with many variant now published in the literature. See for instance also [16] for the statistical theory of the model with exponential family (efd) cell distribution and [17] for its variational inference and model selection.

Despite these previous works, many questions like robustness, missing data, bayesian variational inference or censoring among others have been very few or not at all answered or even studied yet. For the question concerning the occurrences of the terms from textual data, the counts are expected to have varying mean-variance ratios and at least not behave like in a Poisson distribution. Hence herein, it is proposed to consider a co-clustering model with alternatives to the Poisson distribution within the blocks. Alternative distributions to the Poisson one are not in the statistical exponential family (efd) such as a more general estimation needs to be involved here in comparison to previous generalized cases (see [18, 19] with a efd suitable for most usual cases). Note that in co-clustering of textual data, it may be preferred to add contraints to the parameters for sparse parameters, see [20, 21] for pioneer research for count data, and [22] more recently. Hence this framework is discussed herein in order to make it more pratical and avoid a different implementation for each dataset. Herein there is not just a change of the cell distributions in LBM but the update of the model and of the training algorithm is also addressed.

The plan of the paper is as follows. Section 1 is the introduction to the purpose. Section 2 presents the former model and its inference via a variational expectation-maximization algorithm. Section 3 reviews and compares different distributions from the literature for the cells. Section 4 proposes new algorithms for co-clustering and visualization of contigency tables, with generalized constraints via a new approach related to lasso regression. Section 5 is for the experiments validating the proposal. Section 6 concludes with perspectives to the reader.

*This document is a draft with undergoing writing, hence the text and the numerical results are provisory and are expected to change in a next version with updated algorithms and more experimental results on constrained visualization.

†rpriam@gmail.com

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.

Model and Loglikelihood

Within the context of the classical mixture model, a partition of I into g clusters is represented by a binary classification matrix $\mathbf{z}$. Just as I is partitioned into g clusters, columns can be partitioned into m clusters by a binary classification matrix $\mathbf{w}$. Hence $z_{ik} = 1$ indicates the component of the row i while $w_{j\ell} = 1$ indicates the component of the column j, with:

$$\begin{array}{rcl} \mathbf{z} & = & (z_{ik})_{n\times g} \text{ such that } z_{ik} \in \{0,1\} \text{ and } \sum_{k=1}^{g} z_{ik} = 1 \\ \mathbf{w} & = & (w_{j\ell})_{d\times m} \text{ such that } w_{j\ell} \in \{0,1\} \text{ and } \sum_{\ell=1}^{m} w_{i\ell} = 1\,. \end{array}$$

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 associated to $\mathbf{w}$ 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 $\mathbb{R}$ (or a subset) 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 from the log-likelihood:

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

The case of probability mass functions for positive integers x_{ij} in contingency tables is discussed next section.

Objective function and optimization

For this model even with the next constraints, one wants 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.

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,

$$Q_{(\mathbf{c},\mathbf{d})}(\mathbf{z},\mathbf{w}) = \prod_{i,k} (c_{ik})^{z_{ik}} \prod_{j,\ell} \left(d_{j\ell}\right)^{w_{j\ell}}\,.$$

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:

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

  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:

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

  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,

$$x_{k\ell}^{(t)} = \sum_{i,j} c_{ik}^{(t)} d_{j\ell}^{(t)} x_{ij}\,, \mu_k^{(t)} = \sum_i c_{ik}^{(t)} \mu_i\,, \nu_\ell^{(t)} = \sum_j d_{j\ell}^{(t)} \nu_j\,.$$

  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.

	Distribution	Parameters	$E_{k\ell}$	$V_{k\ell}$
1.	$\mathcal{P}$	$\alpha_{k\ell}$	$\lambda_{k\ell}^{ij}$	$\lambda_{k\ell}^{ij}$
2.	$\mathcal{NB}$	$\kappa_{k\ell}, \alpha_{k\ell}$	$\lambda_{k\ell}^{ij}$	$\lambda_{k\ell}^{ij}\left(1+\frac{\lambda_{k\ell}^{ij}}{\kappa_{k\ell}}\right)$
3.	$\mathcal{GP}$	$\kappa_{k\ell}, \alpha_{k\ell}$	$\frac{\lambda_{k\ell}^{ij}}{1-\kappa_{k\ell}}$	$\frac{\lambda_{k\ell}^{ij}}{(1-\kappa_{k\ell})^3}$
4.	$\mathcal{COM}$-$\mathcal{P}$	$\kappa_{k\ell}, \alpha_{k\ell}$	$\sum_{o=0}^{\infty}\frac{o\left(\lambda_{k\ell}^{ij}\right)^o}{(o!)^{\kappa_{k\ell}}Z(\lambda_{k\ell}^{ij},\kappa_{k\ell})} \sim \lambda_{k\ell}^{ij\frac{1}{\kappa_{k\ell}}}$	$\sum_{o=0}^{\infty}\frac{o^2\left(\lambda_{k\ell}^{ij}\right)^o}{(o!)^{\kappa_{k\ell}}Z(\lambda_{k\ell}^{ij},\kappa_{k\ell})} \sim \frac{1}{\kappa_{k\ell}}\lambda_{k\ell}^{ij\frac{1}{\kappa_{k\ell}}}$
5.	$\mathcal{H}$-$\mathcal{P}$	$\alpha_{k\ell}$	$\frac{(1-p_{k\ell})\lambda_{k\ell}^{ij}}{1-\exp\left(-\lambda_{k\ell}^{ij}\right)}$	$\frac{(1-p_{k\ell})\lambda_{k\ell}^{ij}}{1-\exp\left(-\lambda_{k\ell}^{ij}\right)}\left[\lambda_{k\ell}^{ij}+1-\frac{(1-p_{k\ell})\lambda_{k\ell}^{ij}}{1-\exp\left(-\lambda_{k\ell}^{ij}\right)}\right]$
6.	$\mathcal{H}$-$\mathcal{NB}$	$\kappa_{k\ell}, \alpha_{k\ell}$	$\frac{(1-p_{k\ell})\lambda_{k\ell}^{ij}}{1-\varphi(0;\theta_{k\ell})}$	$\frac{(1-p_{k\ell})\lambda_{k\ell}^{ij}}{1-\varphi_{NB}(0;\theta_{k\ell})}\left[\lambda_{k\ell}^{ij}-\frac{(1-p_{k\ell})\lambda_{k\ell}^{ij}}{1-\varphi_{NB}(0;\theta_{k\ell})}+\left(1+\frac{\lambda_{k\ell}^{ij}}{\kappa_{k\ell}}\right)\right]$
7.	$\mathcal{ZIP}$-$\mathcal{P}$	$\alpha_{k\ell}, p_{k\ell}$	$(1-p_{k\ell})\lambda_{k\ell}^{ij}$	$(1-p_{k\ell})\left(1+p_{k\ell}\lambda_{k\ell}^{ij}\right)\lambda_{k\ell}^{ij}$
8.	$\mathcal{ZIP}$-$\mathcal{NB}$	$\kappa_{k\ell}, \alpha_{k\ell}, p_{k\ell}$	$(1-p_{k\ell})\lambda_{k\ell}^{ij}$	$(1-p_{k\ell})\left(1+(\kappa_{k\ell}+p_{k\ell})\lambda_{k\ell}^{ij}\right)\lambda_{k\ell}^{ij}$

Table 1: First moments, Expectations E, and Variances V for the Poisson p.m.f. and related ones.

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. A variant algorithm is the classifying one or BCEM which prefers binarys quantities $z_{ik}^{(t)}, w_{j\ell}^{(t)}$ instead of the fuzzy probabilities $c_{ik}^{(t)}, d_{j\ell}^{(t)}$ for a hard clustering at each iterations while avoiding the need for the loop at the E-step, as denoted a C-step when choosing the current labels maximizing the direct posterior probabilities from previous parameters.

3 Distributional cell modeling

For count data, a probability mass distribution function (p.m.f.) is the Poisson one but alternatives may be considered in order to outperform the fit in the case when the Poisson one is not enough. Among them, the negative binomial one is often studied in biostatistics but is only for over-dispersed data hence alternatives ones may be considered together for further flexibility. It is also discussed the expectation and the variance for their property on the dispersion. Hence, in this section, it is considered for the cells diverse p.m.f.s which are all related to the Poisson one.

Poisson and related mass distributions for discrete cells

The **Poisson p.m.f.** denoted $\mathcal{P}$ is with parameter $\theta_{k\ell} = \alpha_{k\ell}$ for the expectations in [23]. 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 mass function φ for the block $k\ell$ is defined with:

$$\varphi_P(x_{ij};\theta_{k\ell}) = \frac{\lambda_{k\ell}^{ij x_{ij}}\exp\left(-\lambda_{k\ell}^{ij}\right)}{x_{ij}!}, x_{ij} = 0,1,2,3,\cdots.$$

Note that the expectation $E_{k\ell}$ and variance $V_{k\ell}$ for the r.v. in the cells as recalled in Table 1 are such that:

$$f_{k\ell} = \frac{E_{k\ell}}{V_{k\ell}} = 1 \text{ for all } (k,\ell) \text{ if the distribution is } \mathcal{P}\left(\lambda_{k\ell}^{ij}\right).$$

The alternative distributions below are able to correct for this ratio when greater of larger than one, depending of each one when it is for under-dispertion ($f_{k\ell} > 1$) or over-dispertion ($f_{k\ell} < 1$), as observed according to the expression of their statistics. In the case of this current distribution, the solution for α at the maximization step can be written,

$$\alpha_{k\ell}^{(t)} = \frac{x_{k,\ell}^{(t)}}{\mu_k^{(t)}\nu_\ell^{(t)}},$$

as a scalar for each cell or in a vectorial notation for all cells together otherwise. As observed during numerical experiments, without any constraint, at the end of the fit, one always get that:

$$\sum_{k,\ell}\hat{\alpha}_{k\ell}\hat{\mu}_k\hat{\nu}_\ell = \sum_{i,j}x_{ij}.$$

This is a nice property because the quantities $\hat{\alpha}_{k\ell}$ are summarizing.

The **Negative Binomial p.m.f.** denoted $\mathcal{NB}$ updates the Poisson one for over-dispertion with parameters $\theta_{k\ell} = (\kappa_{k\ell}, \alpha_{k\ell})$ via a Poisson-gamma mixture. When $\Gamma(\cdot)$ is the gamma function and,

$C_{k\ell}^{\kappa_{k\ell}} = \frac{\Gamma(\kappa_{k\ell}+x_{ij})}{\Gamma(\kappa_{k\ell})\Gamma(x_{ij}+1)}$, for $x_{ij} = 0,1,2,3,\cdots$, with:

$$\varphi_{NB}(x_{ij};\theta_{k\ell}) = C_{k\ell}^{\kappa_{k\ell}} \left[\frac{\kappa_{k\ell}}{\kappa_{k\ell}+\lambda_{k\ell}^{ij}}\right]^{\kappa_{k\ell}} \left[1-\frac{\kappa_{k\ell}}{\kappa_{k\ell}+\lambda_{k\ell}^{ij}}\right]^{x_{ij}}.$$

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.

The **Generalized Poisson p.m.f.** denoted $\mathcal{GP}$ is with parameters $\theta_{k\ell} = (\kappa_{k\ell}, \alpha k\ell)$, this distribution is found in [24] with the parameters $\theta_{k\ell} = (\kappa_{k\ell}, \alpha_{k\ell})$ where $\kappa_{k\ell} > 0$ and $|\lambda_{k\ell}^{ij}| < 1$, and for $x_{ij} = 0,1,2,3,\cdots$, and,

$$\varphi_{GP}(x_{ij};\theta_{k\ell}) = \frac{\kappa_{k\ell}(\kappa_{k\ell}+x_{ij}\lambda_{k\ell}^{ij})^{x_{ij}-1}\exp\left(-x_{ij}\lambda_{k\ell}^{ij}-\kappa_{k\ell}\right)}{x_{ij}!}.$$

A restricted variant was proposed in [25] by re writing the function. As able to underdispersion and overdispersion, it is an appearling alternative to the Poisson one.

The **Conway-Maxwell-Poisson (COM-Poisson) p.m.f.** denoted $\mathcal{COM}$-$\mathcal{P}$, [26] is defined with a normalization constant, $Z(\lambda,\kappa) = \sum_{o=0}^{\infty}\lambda^o/(o!)^{\kappa}$, with the parameters $\theta_{k\ell} = (\kappa_{k\ell}, \alpha_{k\ell})$ and,

$$\varphi_{COMP}(x_{ij};\theta_{k\ell}) = \frac{\lambda_{k\ell}^{ij\,x_{ij}}}{(x_{ij}!)^{\kappa_{k\ell}} Z(\lambda_{k\ell}^{ij},\kappa_{k\ell})}, x_{ij} = 0,1,2,3,\cdots.$$

Usually, some approximations are preferred for inference in order to get rid of the sum in the normalizing factor for faster and more stable inference, in particular for observed large values. Overdispersion and underdispersion for this distribution is discussed in [27], with in particular for large values of $\lambda_{k\ell}^{ij}$, an approximate expectation and variance directly related to Poisson ones. An extension to the COM-Negative Binomial p.m.f. was proposed in [28] for further flexibility. Able of underdispersion and overdispersion, it can replace the Poisson distribution but at the cost or trickier inference than some other ones.

The **ZIP p.m.f.** denoted $\mathcal{ZIP}$-$\mathcal{D}$, are generic and defined after existing p.m.f. $\mathcal{D}$ by focusing only on zeros. By adapting to the latent block models, these p.m.f. are writen with parameters $p_{k\ell}$ for the probability of zeros in cell $k\ell$ and,

$$\varphi_{ZIP}(x_{ij};\theta_{k\ell}) = \begin{cases} p_{k\ell}+(1-p_{k\ell})\tilde{\varphi}(0;\theta_{k\ell}) & ,x_{ij}=0 \\ (1-p_{k\ell})\tilde{\varphi}(x_{ij};\theta_{k\ell}) & ,x_{ij}=1,2,3,\cdots. \end{cases}$$

The zero-inflated generalized Poisson regression model is discussed in [29, 30] after that case for the Poisson and Binomial distributions were proposed in the literature. Extensions exist with for instance inflation for 1 for instance, while particular distributions for $\tilde{\varphi}(.;.)$ were proposed by diverse researches. This distribution is only for overdispersion as observed with the expression of the variance which can only be greater or equal to the expectation but not lower.

The **Hurdle p.m.f.**, denoted $\mathcal{H}$-$\mathcal{D}$, are generic too by removing the zeros and re-normalizing an existing p.m.f. $\mathcal{D}$, instead of modeling the zeros as coming from a bernoulli's law previously. These p.m.f. are writen with parameters $p_{k\ell}$ for the probability of non zeros in cell $k\ell$ and,

$$\varphi_{H_1^{\infty}}(x_{ij};\theta_{k\ell}) = \begin{cases} (1-p_{k\ell}) & ,x_{ij}=0 \\ p_{k\ell}\dfrac{\tilde{\varphi}(x_{ij};\theta_{k\ell})}{1-\tilde{\varphi}(0;\theta_{k\ell})} & ,x_{ij}=1,2,3,\cdots. \end{cases}$$

For inference, a way around this additional probability $p_{k\ell}$ is to keep only the non zeros values, which is different from the just previous generic distribution with a full model on the two possible states as both cases are linked to the wanted parameters. This distribution is only for overdispersion as observed with the expression of the variance and it is more general than the previous generic one just above. For all these reasons, this is why only this second generic distribution is involved later in the experiments, while the first one is left as a perspective for further analysis.

These last two variant distributions are often met in the literature as they are suitable to improve any existing p.m.f. $\tilde{\varphi}$ -such as above just before- for variance issues coming from the zeros or other count values, this makes them appealing for textual contigency tables. There exist many other models for count data which are not given here and as they are less usual but may be relevant to improve further the fitting. For comparing the variances and expectations from these different distribution for cell modeling, the statistics from the literature and given in Table 1 for completness.

Block distributions for real data

In order to check the cell distribution within the blocks, their empirical distributions are plot and compared to the proposed Poisson-related distributions from above just before. For instance, when checking the graphic of the barplot from the counts x_{ij} for the usual dataset CLASSIC3 the zeros may be too many such that a suitable modeling of this inflating behavior is expected to improve the clustering. It may also noticed that large count values are met in a few cells, such that truncating for $x_{ij} > M$, with herein $M = 10$, needs to be investigated, this is not rare that there was a removal of higher occurrences in textual contingency tables in previous works from the literature or when building the dataset. This also suggests an alternative truncated distributions, denoted $\mathcal{H}_1^M$-$\mathcal{D}$, with more constraints than for the Hurdle one, and with only possible values $x_{ij} \in \{1,..,M\}$ in a censoring way. After the truncation:

$$\varphi_{H_1^M}(x_{ij};\theta_{k\ell}) = \begin{cases} (1-p_{k\ell}) & ,x_{ij}=0 \\ p_{k\ell}\dfrac{\tilde{\varphi}(x_{ij};\theta_{k\ell})}{\sum\limits_{o\in\{1,..,M\}}\tilde{\varphi}(o;\theta_{k\ell})} & ,x_{ij}=1,\cdots M. \end{cases}$$

This kind of truncation or distribution is expected to improve the inference when computing gradient in nonlinear optimization, or at least to bring better understanding of the data, because a too high occurrence is rare and not relevant. Another related concern is that the size of the block $n_{k\ell}$ are so large that indeed large

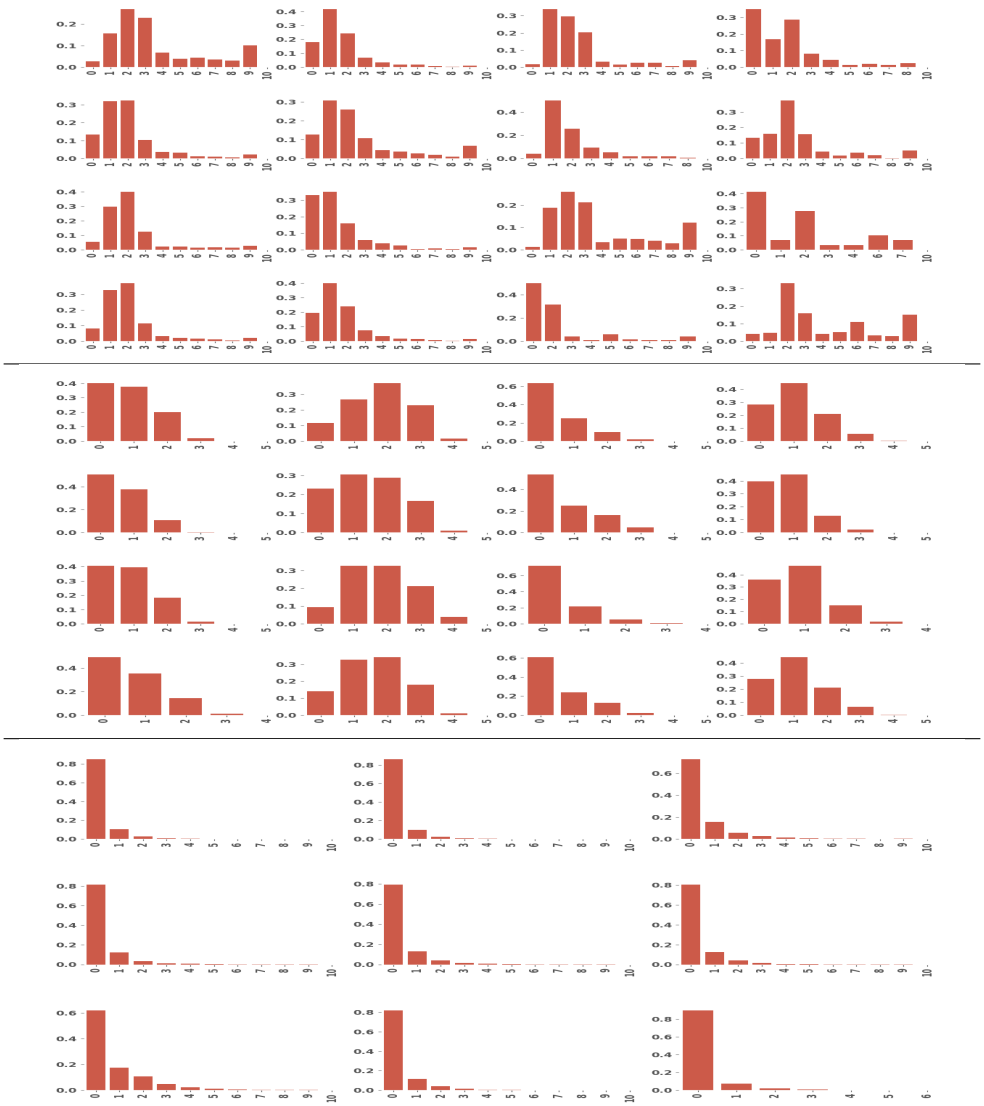

Figure 1: Exemples of barplot from the counts in the blocks from BCEM for PLBM with the datasets CSTR, WEBKB4 and CLASSIC3.

counts should be likely observed in such a sample. An ultimate purpose for these empirical counts is a comparison with the theoritical ones but as the block sizes $n_{k\ell}$ are large only the shapes are meaningful. The Hurdle variant looks like enough here when the probabilities for large cells are already small. Next section is dedicated to LBM with variants of Poisson distributions, say the negative binomial one and no ZIP trick but Hurdle one.

4 Inference of parameters

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 in PLBM. The estimation of the parameters from any distribution on the cells but without the zero-inflated variants are also discussed for more generality, as a nonlinear optimization problem.

Extensions to other distributions and constraints

In the case of nonlinear optimization, let keep t for as the iterations number in the BEM or BCEM algorithm. At the M-step, $\mathbf{H}_{\theta^{(t)}}$ and $\mathbf{G}_{\theta^{(t)}}$ denote respectively the Hessian matrix and gradient vector, both from parameters of previous step with first and second derivarives of $\tilde{Q}_{LBM}(\theta,\theta^{(t)})$ w.r. θ. One needs some suitable algorithm such as Newton-Raphson one:

$$\begin{aligned}
\theta^{(t+1)} &= \theta^{(t)} - \mathbf{H}^{-1}_{\theta^{(t)}}\mathbf{G}_{\theta^{(t)}} \\
\mathbf{G}_{\theta^{(t)}} &= \nabla_\theta \tilde{Q}_{LBM}(\theta,\theta^{(t)})\Big|_{\theta^{(t)}} \\
\mathbf{H}_{\theta^{(t)}} &= \nabla_\theta \nabla^T_\theta \tilde{Q}_{LBM}(\theta,\theta^{(t)})\Big|_{\theta^{(t)}} .
\end{aligned}$$

Considering the expression of $\tilde{Q}_{LBM}$, and an eventual rewriting the parameters in a block $\theta_{k\ell} = \phi(a_{k\ell})$ where $a_{k\ell}$ is a scalar or vector while $\phi()$ is a transformation, this leads to prefer for the derivatives w.r.t. $a_{k\ell}$ instead, with:

$$\begin{aligned}
\mathbf{G}_{a^{(t)}_{k\ell}} &= \sum_{i,j} c^{(t)}_{ik} d^{(t)}_{j\ell} \frac{\nabla_{a_{k\ell}} \varphi(x_{ij};\theta_{k\ell})}{\varphi(x_{ij};\theta_{k\ell})} \\
\mathbf{H}_{a^{(t)}_{k\ell}} &= \sum_{i,j} c^{(t)}_{ik} d^{(t)}_{j\ell} \frac{\nabla_{a_{k\ell}} \nabla^T_{a_{k\ell}} \varphi(x_{ij};\theta_{k\ell}) - \frac{\nabla_{a_{k\ell}} \varphi(x_{ij};\theta_{k\ell}) \nabla^T_{a_{k\ell}} \varphi(x_{ij};\theta_{k\ell})}{\varphi(x_{ij};\theta_{k\ell})}}{\varphi(x_{ij};\theta_{k\ell})} .
\end{aligned}$$

For the three selected cases,

- With a Poisson distribution, let write $\theta_{k\ell}$ as an exponential function or eventually a sigmoid one. When denoting, $x_{k\ell}^{(t)} = \sum_{i,j} c_{ik}^{(t)} d_{j\ell}^{(t)} x_{ij}$, $\mu_k^{(t)} = \sum_i c_{ik}^{(t)} \mu_i$ and $\nu_\ell^{(t)} = \sum_j d_{j\ell}^{(t)} \nu_j$, it may be writen directly $\mathbf{G}_{a_{k\ell}}^{(t)} = 0$ as the transformation is not required:

$$\begin{aligned} \mathbf{G}_{a_{k\ell}}^{(t)} &= x_{k\ell}^{(t)} \frac{\phi'(a_{k\ell})}{\phi(a_{k\ell})} - \mu_k^{(t)} \nu_\ell^{(t)} \phi'(a_{k\ell}) \\ {a_{k\ell}}^{(t)} &= \phi^{-1}\left(\frac{x_{k\ell}^{(t)}}{\mu_k^{(t)} \nu_\ell^{(t)}} \right). \end{aligned}$$

- With a Hurdle Poisson distribution, let write $\theta_{k\ell}$ as an exponential function or eventually a sigmoid one, then:

$$\begin{aligned} \mathbf{G}_{a_{k\ell}}^{(t)} &= \left(\frac{x_{k\ell}^{(t)}}{\phi(a_{k\ell})} - \mu_k^{(t)} \nu_\ell^{(t)} \right) \phi'(a_{k\ell}) + b_{k\ell}^{(t)} \\ b_{k\ell}^{(t)} &= -\phi'(a_{k\ell}) \sum_{i,j} c_{ik}^{(t)} d_{j\ell}^{(t)} \frac{\mu_i \nu_j}{1 - e^{-\mu_i \nu_j \phi(a_{k\ell})}}. \end{aligned}$$

- With a Negative Binomial distribution, let write $\theta_{k\ell}$ as an exponential function or eventually a sigmoid one, and when all $\kappa_{k\ell}$ are kept common equal to one unique value κ for instance.

Extension to constraints and visualizations

A reason of studying co-clustering for contingency table is the visualization of textual data by extending on the methods presented herein by including more constraints. This justifies a clustering instead of reduction as a preamble, before a projective parameterization. More generally, for constrained models, one may try to add some lasso or norm penalty to $\tilde{Q}_{LBM}(\theta, \theta^{(t)})$ with the penalization $-\lambda_P \sum_{k\ell} \Upsilon(\alpha_{k\ell})$ where λ_P is well chosen while $\Upsilon(\alpha_{k\ell}) = |\alpha_{k\ell}| \approx \alpha_{k\ell}^2 / \widetilde{(\alpha_{k\ell})}$ or $\Upsilon(\alpha_{k\ell}) = \log(\alpha_{k\ell})$. If this leads to closed-form solutions for Poisson with quadratic approximation of the penalization, this is not a selection which is involved in this version but separated penalizations.

Additional constraints are induced by latent vectors $\xi_k^T \xi_\ell$ as in a pca model but in all cases other constraints may be required in order to insure better properties such as orthogonality. For the nonlinear mapping, $\alpha_{k\ell}$ or c_{ik} is a function of two vectors ξ_k and ξ^ℓ for a reduction of the two spaces of the contingency matrix. Herein the focus in on principal component analysis via LBM. This is writen:

$$\alpha_{k\ell} = \frac{1}{1 + \exp(-\xi_k^T \xi_\ell)}.$$

Numerical nonlinear optimizations are known to required projected gradients for instance when the parameters are not completely free, see the experiments for more details.

5 Experimental and numerical results

For studying further the statistics of the counts within the block, the empirical means and standard-deviations are first compared in order to check if the Poisson distribution is relevant or not, and after the likelihoods from several distribution are computed for a co-clustering and compared in order to check further which fit looks the best. After this first stage, the selected distributions are tested for co-clustering in order to check if the clustering is improved or not. These distributions are also involved for visualization with further constraints at the end.

Datasets

For these experiments, three datasets are selected from the usual ones in the literature, after trucation for the higher counts. In order to insure that same dataset is used later, the total sum of the counts is also given.

Name	n	p	$x_{\bullet\bullet}$	sparsity (%)	g
CSTR	475	1000	59090	96.63	4
WEBKB4	4199	1000	459028	94.14	4
CLASSIC3	3891	4303	255892	98.95	3

Numerical results

Graphically, it is observed three different general shapes for the three datasets, somewhar similar from a cell to another. See Figure 2 for examples of barplots after a co-clustering with the usual distribution and $m = g$ without any constraint.

for these datasets, in the case of co-clustering (with 120 fits while keeping the best $\tilde{Q}$) there is not much improvement of the clustering or even the criterion even if a small increased is observed, this suggests to keep $m = g$ for these datasets as proposed in the literature, at least for clustering. Ideally, the number of clusters along the columns may be chosen automatically. Note also that for projection it may be learnt from the literature to ask for more clusters than clustering for a non linear projection but a same number of clusters as clustering in a linear projection such that mixture of principal components.

The empirical means $\bar{x}_{k\ell}$ and variances $v_{k\ell}$ are computed in each cell for each dataset such that, it is obtained, an observed ratio $\hat{f}_{k\ell} = \bar{x}_{k\ell} / v_{k\ell}$ which are given in the Table 2 below.

Name	size	mean	std	min	max
CSTR	4×4	0.23	0.04	0.17	0.30
WEBKB4	4×4	0.48	0.08	0.37	0.65
CLASSIC3	4×4	0.57	0.12	0.38	0.69

Table 2: Statistics from $\hat{f}_{k\ell}$

The table shows that for the three datasets, the counts are overdispersed, here without removing the zeros: with a ratio from a half to a quarter. The distribution for the Poisson, Hurdle Poisson, Negative binomial and Generalized Poisson are also fitted in each cell, after a co-clustering in order to compare the bic and aic. This is summarized with means and standard-deviations for each dataset and each distribution.

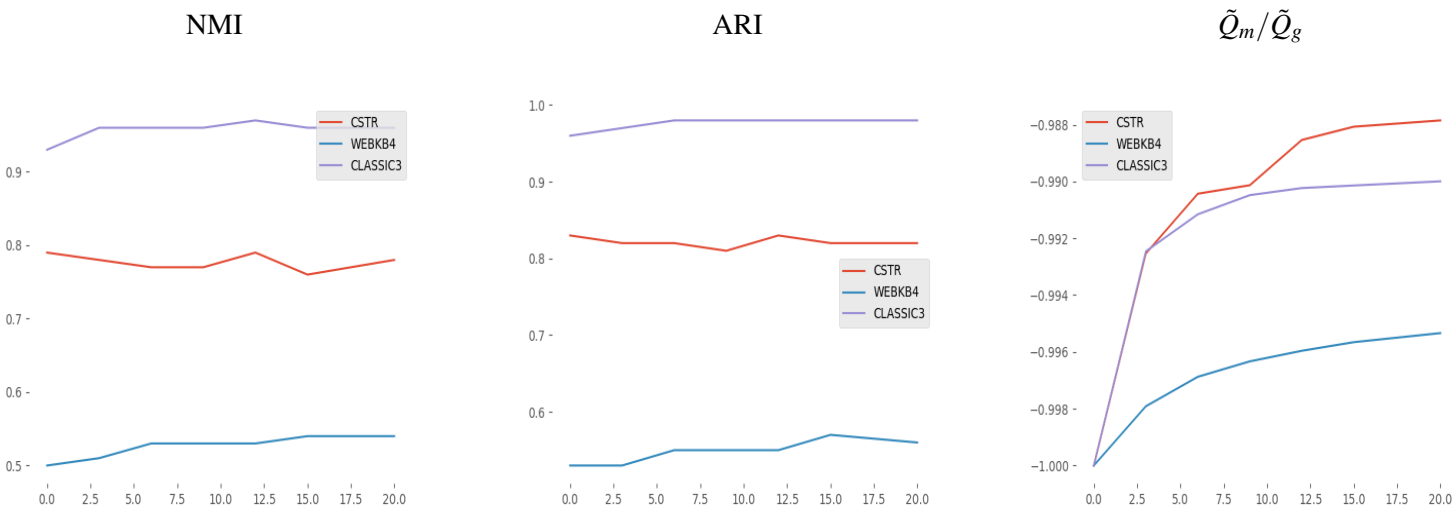

Figure 2: Exemples of nmi, ari and $\tilde{Q}$ for g constant and different values of m from g to $g+20$ after fitting with BCEM for PLBM.

6 Discussion and perspectives

Herein it is explained how generalized constraints may be added to a latent block model for visualization purposes. 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] [32] with a bayesian approach for the inference and not for textual data hence was not considered further, neither for visualization.

Appendix

Surrogate objective function for NB

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 [33] from the logit modeling may be relevant. This is because it is rewriten the mass function as:

$$\varphi(x_{ij};\theta_{k\ell}) = \frac{\Gamma(\kappa_{k\ell}+x_{ij})}{\Gamma(\kappa_{k\ell})x_{ij}!} e^{-\kappa_{k\ell}\lambda_{k\ell}^{ij}-\kappa_{k\ell}\gamma_{ij}^{k\ell}} \sigma(\lambda_{k\ell}^{ij}+\gamma_{ij}^{k\ell})^{x_{ij}+\kappa_{k\ell}}.$$

By convexity, the sigmoid is known to be bounded:

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

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,

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

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:

$$\begin{aligned} \sigma_{ij}^{k\ell} &= \sigma(\varepsilon_{ij}) \\ a_{ij}^{k\ell} &= \lambda_{k\ell}^{ij}+\gamma_{ij}^{k\ell}, \end{aligned}$$

the criterion to optimize may be written:

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

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.

Maximization step for NB

The optimization is then performed as follows:

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

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

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

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:

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

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.

Maximization step for large sparse matrix for NB

In the case of large matrix, with just a simple laptop, it may be relevant to be cautious with the expression of the optimized parameters, if some variant like Hurdle one is not involved. In particular, the zero must be take care separately here as we explain in this part. By this way, the new expression for $\alpha_{k\ell}^{(t)}$, $c_{ik}^{(t)}$, and $d_{j\ell}^{(t)}$ becomes more practical for large sparse matrices.

- **Rewritting the sums** To begin with a direct solution, it is rewritten the parameter for the expectations as follows:

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

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 and untractable to compute exactly for large contingency tables with an usual computer. 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 BCEM 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:

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

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 [34] 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.

- **Rewriting with sampling** An alternative is more appealing in order to handle more cases, such as fuzzy classifying vectors. In the clustering expectation maximization, most of the probabilities c_{ik} and $d_{j\ell}$ are expected to be small while only a few values for each row and column emerge as the corresponding solution to the clustering and are near one. Thus, we call $U_+^{k\ell}$ for this set of pairs (i,j) and $U_-^{k\ell}$ for the other remaining set of pairs such as if U is the whole set of pairs of size $n \times d$, then we have $U^{k\ell} = U_+^{k\ell} \cup U_-^{k\ell}$. Hence, this leads to write $\Lambda_{k\ell}^{+(t)}$ as the new sparse matrix where the pairs (i,j) have their probability enough large to belong to the a block, for instance for a threshold or a minimum rank (such as the two first larger ones). The new computation for the quantities

are rewritten:

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

Here, the vector $c_i^{(t)}$ and $d_j^{(t)}$ are defined with the components $c_{ik}^{(t)}$ and $d_{j\ell}^{(t)}$ with censoring the smaller probabilities which depend on the pairs already handled in the central matrix. The remaining terms are for the smaller probabilities which are estimated via an usual survey approach for a total after sampling in $U_-^{k\ell}$ an enough large set of pairs.

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.

[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, 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] V. Brault, C. Keribin, G. Celeux, and G. Govaert, "Estimation and selection for the latent block model on categorical data," vol. 25, pp. 1–16, 06 2014.

[18] R. Priam, M. Nadif, and G. Govaert, "Generalized topographic block model," *Neurocomputing*, vol. 173, pp. 442–449, 2016.

[19] R. Priam and M. Nadif, "Data visualization via latent variables and mixture models: a brief survey," *Pattern Analysis and Applications*, vol. 19, no. 3, pp. 807–819, Aug 2016.

[20] 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.

[21] 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.

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

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

[24] P. C. Consul and G. C. Jain, "A generalization of the poisson distribution," *Technometrics*, vol. 15, no. 4, pp. 791–799, 1973.

[25] F. Famoye, "Restricted generalized poisson regression model," *Communications in Statistics - Theory and Methods*, vol. 22, no. 5, pp. 1335–1354, 1993.

[26] K. F. Sellers, S. Borle, and G. Shmueli, "The com-poisson model for count data: a survey of methods and applications," *Applied Stochastic Models in Business and Industry*, vol. 28, no. 2, pp. 104–116, 2012.

[27] F. Daly and R. Gaunt, "The conway-maxwell-poisson distribution: Distributional theory and approximation," *Latin American journal of probability and mathematical statistics*, vol. 13, 07 2016.

[28] H. Zhang, K. Tan, and B. Li, "Com-negative binomial distribution: modeling overdispersion and ultrahigh zero-inflated count data," *Frontiers of Mathematics in China*, vol. 13, no. 4, pp. 967–998, Aug 2018.

[29] F. Famoye and K. Singh, "Zero-inflated generalized poisson regression model with an application to domestic violence data," *Journal of Data Science*, vol. 4, pp. 117–130, 01 2006.

[30] F. Famoye and J. S. Preisser, "Marginalized zero-inflated generalized poisson regression," *Journal of Applied Statistics*, vol. 45, no. 7, pp. 1247–1259, 2018.

[31] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[32] 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.

[33] 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.

[34] 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.