Negative binomial latent block model with generalized constraints

R. Priam*

November 22, 2024

Abstract

Constrained latent block models (LBM) are proposed for contingency matrices herein. Several discrete distributions related to the usual Poisson one are compared for modeling the blocks in a co-clustering and a reduction of the rows and columns.

Keywords: latent block model, contingency table, 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. See also for instance [16] for a statistical theory of the model with exponential family (efd) cell distribution and [17] for its variational inference and model selection.

*rpriam@gmail.com

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. 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 [18] with a bayesian approach for the inference and not for textual data hence was not considered further, neither for visualization.

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 [19, 20] 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 [21, 22] for pioneer research for count data, and [23] more recently.

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. Section 5 is for the experiments validating the proposal. Section 6 concludes with future 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.

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:

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

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:

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

  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,

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

  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. 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.

	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.	$\mathcal{COM}$-$\mathcal{P}$	$\kappa_{k\ell}, \alpha_{k\ell}$	$\sum_{o=0}^{\infty} \frac{o\left(\lambda_{k\ell}^{ij}\right)^o}{(o!)^{\kappa_{k\ell}} Z(\lambda_{k\ell}^{ij}, \kappa_{k\ell})}$	$\sum_{o=0}^{\infty} \frac{o^2\left(\lambda_{k\ell}^{ij}\right)^o}{(o!)^{\kappa_{k\ell}} Z(\lambda_{k\ell}^{ij}, \kappa_{k\ell})}$
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 $E_{k\ell}$ and Variances $V_{k\ell}$, for the Poisson and related p.m.f.

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 [24]. 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,

$$\alpha_{k\ell}^{(t+1)} = \frac{x_{k,\ell}^{(t)}}{\mu_k^{(t)} \nu_\ell^{(t)}},$$

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:

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

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 the literature 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,

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

Several variants have been proposed in the literature, as for most of the other distributions in this section.

The **Conway-Maxwell-Poisson (COM-Poisson) p.m.f.** denoted $\mathcal{COM}$-$\mathcal{P}$, 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,

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

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 the literature, it can replace well the Poisson distribution at the cost of a cumbersome inference.

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 written with parameters $p_{k\ell}$ for the probability of excess 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 were discussed in the literature after Poisson and Binomial models. 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 directly modeling the excess of zeros as just above. These p.m.f. are written 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 distributions for cell modeling, the statistics from the literature are 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 distribution is more relevant after truncation (censoring) but large values for the cells are rarely observed such as the Hurdle variant is preferred next after. A related concern is that the sizes of the block $n_{k\ell}$ are so large that indeed large 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 very large only the shapes may be meaningful. Next section is dedicated to LBM with variants of Poisson distributions, say the negative binomial and no ZIP but Hurdle.

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.

Derivatives of $\tilde{Q}_{LBM}$

For nonlinear optimization, let keep t for as the iterations number in the BEM or BCEM algorithm. 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 first order derivatives w.r.t. $a_{k\ell}$ instead,

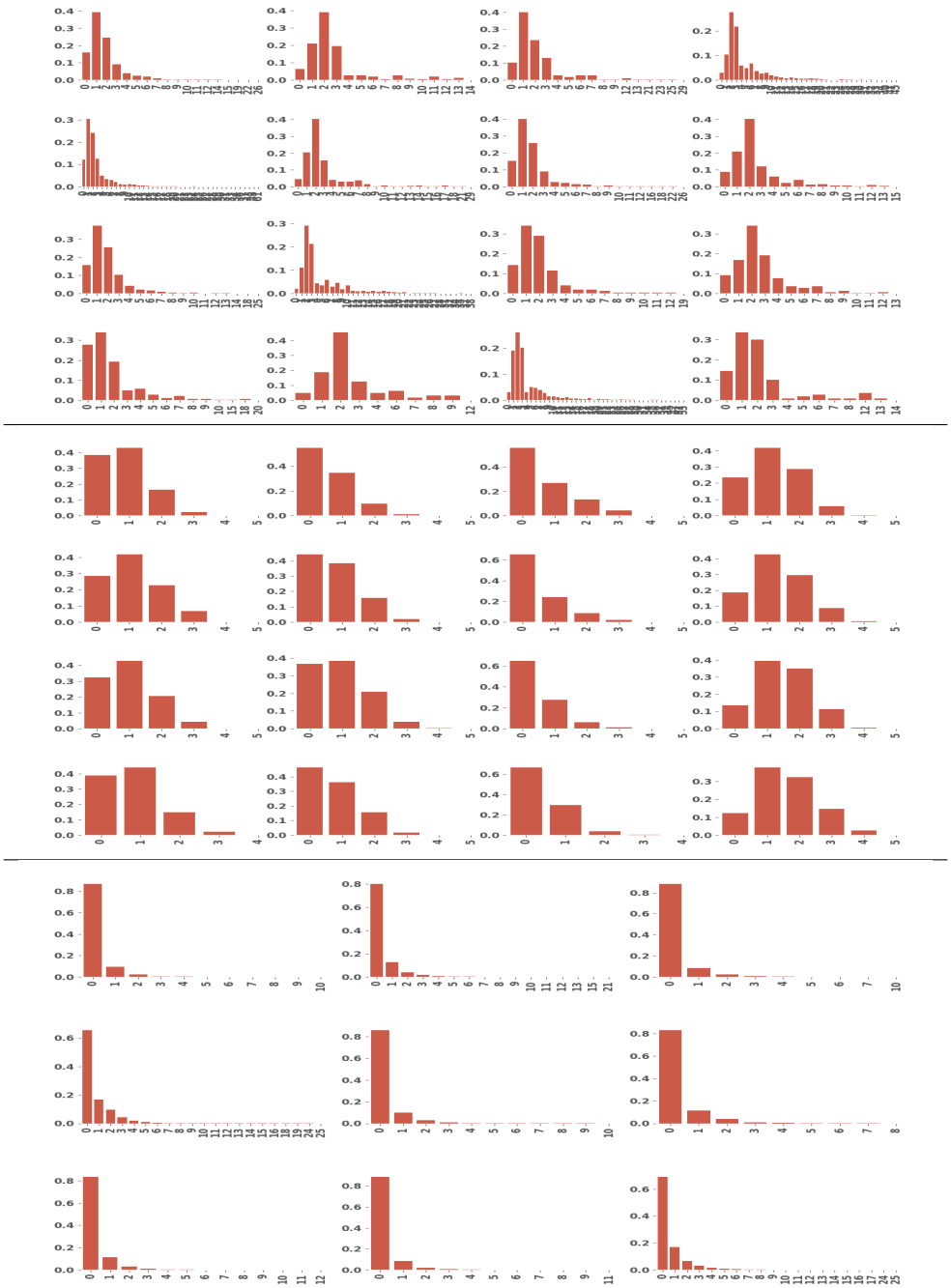

Figure 1: Exemples of barplot from the counts in the blocks from BCEM for PLBM with the datasets CSTR, WEBKB4 and CLASSIC3.

with:

$$\mathbf{G}_{a_{k\ell}^{(t)}} = \sum_{i,j} c_{ik}^{(t)} d_{j\ell}^{(t)} \frac{\nabla_{a_{k\ell}} \varphi(x_{ij};\theta_{k\ell})}{\varphi(x_{ij};\theta_{k\ell})}$$

$$\mathbf{H}_{a_{k\ell}^{(t)}} = \sum_{i,j} c_{ik}^{(t)} d_{j\ell}^{(t)} \frac{\nabla_{a_{k\ell}} \nabla_{a_{k\ell}}^T \varphi(x_{ij};\theta_{k\ell}) - \frac{\nabla_{a_{k\ell}} \varphi(x_{ij};\theta_{k\ell}) \nabla_{a_{k\ell}}^T \varphi(x_{ij};\theta_{k\ell})}{\varphi(x_{ij};\theta_{k\ell})}}{\varphi(x_{ij};\theta_{k\ell})}.$$

For the derivatives, one may get that:

- With a Poisson distribution, let write $\theta_{k\ell}$ as an exponential function or eventually a sigmoid one. When denoting, it may be written directly $\mathbf{G}_{a_{k\ell}}^{(t)} = 0$ as the transformation is not required:

$$\mathbf{G}_{a_{k\ell}}^{(t)} = x_{k\ell}^{(t)} \frac{\phi'(a_{k\ell})}{\phi(a_{k\ell})} - \mu_k^{(t)} \nu_\ell^{(t)} \phi'(a_{k\ell}).$$

This results into the solution given in the literature [15] and at previous section when reviewing PLBM.

- With a Hurdle Poisson distribution, let write $\theta_{k\ell}$ as an exponential function or eventually a sigmoid one, when $c_k^{(t)} = [c_{1k}^{(t)}, \cdots, c_{nk}^{(t)}]^T$, $d_\ell^{(t)} = [d_{1\ell}^{(t)}, \cdots, d_{d\ell}^{(t)}]^T$, $\mu_{1;n} = [\mu_1, \cdots, \mu_n]^T$, $\nu_{1;d} = [\nu_1, \cdots, \nu_d]^T$, and $\sigma_{k\ell}^-(u) = u/(1 - e^{-u\phi(a_{k\ell})})$

$$\frac{\mathbf{G}_{a_{k\ell}}^{(t)}}{\phi'(a_{k\ell})} = \left(\frac{x_{k\ell}^{(t)}}{\phi(a_{k\ell})} - \mu_k^{(t)} \nu_\ell^{(t)} \right) - c_k^{(t)T} \sigma_{k\ell}^-(\mu_{1;n} \nu_{1;d}^T) d_\ell^{(t)}.$$

- 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 in-

stance, while $\Lambda_{k\ell} = (\lambda_{k\ell}^{ij})_{ij}$, $\sigma_{k\ell}^{+}(u) = u/(\kappa_{k\ell} + u\phi(a_{k\ell}))$,

$$\frac{\mathbf{G}_{a_{k\ell}}^{(t)}}{\phi'(a_{k\ell})} = c_k^{(t)} \left[\kappa_{k\ell} + \mathbf{x}\right) \odot \sigma_{k\ell}^{+}\left(\mu_{1;n}\nu_{1;d}^T\right) + \mathbf{x} \oslash \Lambda_{k\ell}\right] d_\ell^{(t)}.$$

Here $\odot$ denotes the element-wise product, $\oslash$ the element-wise division, while the functions $\sigma_{k\ell}^{+}(.)$ and $\sigma_{k\ell}^{-}(.)$ are also element-wise resulting into matrices of same sizes.

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 other constraints may be wanted 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 written with a sigmoid $\sigma(u) = \frac{1}{1+e^{-u}}$, such that:

$$\alpha_{k\ell} = \sigma(\xi_k^T \xi_\ell).$$

The derivatives $\mathbf{G}_{\xi_k}^{(t)}$ and $\mathbf{G}_{\xi_{ell}}^{(t)}$ are found with $a_{k\ell} = \xi_k^T \xi_\ell$ by following the composite function chain rule:

$$\begin{aligned} \mathbf{G}_{\xi_k}^{(t)} &= \mathbf{G}_{a_{k\ell}}^{(t)} \frac{\partial a_{k\ell}}{\partial \xi_k}\Big|_{\xi_\ell^{(t)}} = \mathbf{G}_{a_{k\ell}}^{(t)} \xi_\ell^{(t)} \\ \mathbf{G}_{\xi_\ell}^{(t)} &= \mathbf{G}_{a_{k\ell}}^{(t)} \frac{\partial a_{k\ell}}{\partial \xi_\ell}\Big|_{\xi_k^{(t)}} = \mathbf{G}_{a_{k\ell}}^{(t)} \xi_k^{(t)}. \end{aligned}$$

Numerical nonlinear optimizations are known to required projected gradients for instance when the parameters are not completely free, see the experiments for more details. Other parameterizations are left as a perspective to a next communication if possible.

Implementations for M-step

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.t. θ. One needs some suitable algorithm such as follows.

- The usual algorithm for a fast optimization is the Newton-Raphson procedure with the Hessian matrix and gradient vector computed from the full dataset such that:

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

- The derivatives of the criterion and the variational posterior probabilities have a non sparse expression and require often herein all the cell indices from the data matrix, this suggests for an iterative optimization. A derivative $\tilde{G}$ w.r.t. only a part of the dataset and a stochastic (or mini-batch) algorithm, with a small parameter $\eta_{(t)}$, lead to:

$$\theta^{(t+1)} = \theta^{(t)} - \eta_{(t)} \tilde{\mathbf{G}}_{\theta^{(t)}}.$$

For an usual implementation and usual computer, sequential algorithms are more suitable for experiments with dense matrices otherwise approximations of the gradient is required via a sampling or a random projection for the vectors from the posterior. At least a computation from blocks of $\mu_{1;n}\nu_{1;d}^T$ allows to deal with a large dense matrix for the storage because high level languages or matricial libraries seem to not manage this kind of issue.

Alternative optimization for Negative Binomial p.m.f.

A closed-form solution is available via a variational bound (see for instance [25] for a related model), but this kind of approach was not found able to converge always faster than with a direct nonlinear optimization. This may be written as follow, from the mass function, its bound and the resulting update formula:

$$\varphi_{NB}(x_{ij}; \theta_{k\ell}) = C_{k\ell}^{\kappa_{k\ell}} e^{-\kappa_{k\ell}\lambda_{k\ell}^{ij} - \kappa_{k\ell}\gamma_{ij}^{k\ell}} \sigma(\lambda_{k\ell}^{ij} + \gamma_{ij}^{k\ell})^{x_{ij}+\kappa_{k\ell}}.$$

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. 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. With $\varepsilon_{ij}^{k\ell} = |a_{ij}^{k\ell}|$ where $a_{ij}^{k\ell} = \lambda_{k\ell}^{ij} + \gamma_{ij}^{k\ell}$, this results into the updates:

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

This induces to compute variational parameters $\varepsilon_{ij}^{k\ell}$ for each cells of the data matrix, which is very costly on an usual computer.

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

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

order to check if the Poisson distribution is relevant or not when compared to the selected distributions.

Datasets

For these experiments, three datasets are selected from the usual ones in the literature, after truncation 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

Table 2: Datasets.

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. 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 the observed ratios $\hat{f}_{k\ell} = \bar{x}_{k\ell}/v_{k\ell}$ which are given in the Table 3 below with the mean, standard-deviation, minimum and maximum from the $g\times$ values computed at each block for each dataset. The table shows that for the three datasets, the counts are over-dispersed, 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.

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	3×3	0.57	0.12	0.38	0.69

Table 3: Statistics from $\hat{f}_{k\ell}$.

6 Discussion and perspectives

Herein several constraints are added to a latent block model for clustering and reduction purposes. The negative binomial distribution and related ones are checked for improving the fitting of the blocks by comparing with the results from the usual Poisson one. Algorithms are proposed for the estimation of the parameters and the estimated labels from sparse contingency tables. A main perspective is an even more general model for co-clustering, reduction, factorization and visual data analysis for also other types of numerical matrices with diverses constraints of latent vectors.

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] 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.
[19] R. Priam, M. Nadif, and G. Govaert, "Generalized topographic block model," *Neurocomputing*, vol. 173, pp. 442–449, 2016.
[20] 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.
[21] 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.
[22] 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.
[23] M. Selosse, J. Jacques, and C. Biernacki, "Textual data summarization using the self-organized co-clustering model," *Pattern Recognition*, vol. 103, p. 107315, 2020.
[24] G. Govaert and M. Nadif, "Latent block model for contingency table," *Communications in Statistics-theory and Methods*, vol. 39, pp. 416–425, 2010.
[25] R. Priam, M. Nadif, and G. Govaert, "Topographic bernoulli block mixture mapping for binary tables," *Pattern Analysis and Applications*, vol. 17, pp. 839–847, 2014.