Journal de la Société Française de Statistique
Vol. 156 No. 3 (2015)

Co-clustering through Latent Bloc Model: a Review

Titre: Une revue bibliographique de la classification croisée au travers du modèle des blocs latents

Vincent Brault[1] and Mahendra Mariadassou[2]

Abstract: We present here model-based co-clustering methods, with a focus on the latent block model (LBM). We introduce several specifications of the LBM (standard, sparse, Bayesian) and review some identifiability results. We show how the complex dependency structure prevents standard maximum likelihood estimation and present alternative and popular inference methods. Those estimation methods are based on a tractable approximation of the likelihood and rely on iterative procedures, which makes them difficult to analyze. We nevertheless present some asymptotic results for consistency. The results are partial as they rely on a reasonable but still unproved condition. Likewise, available model selection tools for choosing the number of groups in rows and columns are only valid up to a conjecture. We also briefly discuss non model-based co-clustering procedures. Finally, we show how LBM can be used for bipartite graph analysis and highlight throughout this review its connection to the Stochastic Block Model.

Résumé : Nous présentons ici les méthodes de co-clustering, avec une emphase sur les modèles à blocs latents (LBM) et les parallèles qui existent entre le LBM et le Modèle à Blocs Stochastiques (SBM), notamment pour l'analyse de graphes bipartites. Nous introduisons différentes variantes du LBM (standard, sparse, bayésien) et présentons des résultats d'identifiabilité. Nous montrons comment la structure de dépendance complexe induite par le LBM rend l'estimation des paramètres par maximum de vraisemblance impossible en pratique et passons en revue des méthodes d'inférence alternatives. Ces dernières sont basées sur des procédures itératives, combinées à des approximations faciles à maximiser de la vraisemblance, ce qui les rend malaisés à analyser théoriquement. Il existe néanmoins des résultats de consistence, partiels en ce qu'ils reposent sur une condition raisonnable mais encore non démontrée. De même, les outils de sélection de modèle actuellement disponibles pour choisir le nombre de cluster reposent sur une conjecture. Nous replacons brièvement LBM dans le contexte des méthodes de co-clustering qui ne s'appuient pas sur un modèle génératif, particulièrement celles basées sur la factorisation de matrices. Nous concluons avec une étude de cas qui illustre les avantages du co-clustering sur le clustering simple.

Keywords: Latent Variable model, Latent Block Model, Variational approximation, Model Selection, ICL, BIC, Bipartite Graphs
Mots-clés : Modèle à blocs latents, Modèle à variables latentes, Approximation variationnelle, Sélection de modèle, ICL, BIC, Graphes bipartites
AMS 2000 subject classifications: 62H30, 62-00, 62-07

1. Introduction

Cluster analysis is an essential tool of data science and comes in many flavors in fields as diverse as omics, computer vision, business analytics and more generally machine learning. In these contexts, data are recorded in a table where rows index samples, columns index features and table elements encode the relation between a sample and a feature. The focus of most clustering methods is

[1] AgroParisTech/UMR INRA MIA 518.
E-mail: vincent.brault@agroparistech.fr
[2] INRA, UR1404 Unité Mathématiques et Informatique Appliquées du Génome à l'Environnement, F78352 Jouy-en-Josas, France.
E-mail: mahendra.mariadassou@jouy.inra.fr

 ISSN: 2102-6238

to cluster either samples or variables. We focus here on so called co-clustering procedures that consider the table as a whole, instead of row-wise and column-wise, and simultaneously cluster samples and features into homogeneous sets. Co-clustering methods are usually faster and more accurate than separate clustering of the samples and the features (Govaert and Nadif, 2003). They usually also create easier to interpret groups. Examples and fields of application include, but are not limited to, the Netflix problem, metagenomics surveys and voting patterns. In the Netflix problem (Bennett and Lanning, 2007), co-clustering simultaneously clusters movies into genres and viewers into segments from a rating matrix. The goal is to understand users' preferences at an aggregated level and to recommend to viewers movies that they are likely to enjoy based on the movie genre and the viewer's segment. In metagenomics survey, co-clustering of samples and species from abundance data can help define environmental conditions and indicator species of those conditions (Aubert et al., 2014). In voting patterns, our worked-out example, co-clustering creates groups of voters with similar legislative agendas and groups of laws with similar legislative supports.

Co-clustering is also useful for data visualization when there is no natural order on samples and features. Reordering the margins of the table to make the homogeneous sets of samples and features contiguous induces a partition of the table into rectangle blocks (see Figure 1). Note that co-clustering techniques induce a crossed clustering of table elements which may be sub-optimal and inadequate if the underlying structure is not crossed (see Figure 2 for a graphic example). In this review, we are interested in model-based co-clustering and will focus on Latent Block Models (LBM) (Govaert and Nadif, 2013), a large class of probabilistic models where homogeneous sets and the resulting structure are encoded by latent variables. LBM can be defined on binary, Gaussian (Lomet, 2012), count (Govaert and Nadif, 2010), categorical (Keribin et al., 2014) and more generally any kind of valued data (Biernacki and Jacques, 2012).

Data tables can be studied as such but also in the context of graphs, where they can be viewed as adjacency matrices of bipartite graphs (Tanay et al., 2004), as illustrated in Figure 1. Nodes in bipartite graphs belong to one of two types, for example samples and features, and nodes from one type can only connect to nodes from the other type. This perspective casts co-clustering in terms of graphs analysis: co-clustering consists in clustering nodes into type-coherent groups. LBM are thus related to to Stochastic Block Models (SBM). SBM are traditionally used for graph analysis (Matias and Robin, 2014) but, just like LBM, can be viewed as models for finding structure in square data tables. This common framing explains why most theoretical results on LBM either originate from the SBM literature or can be adapted to SBM. However, crucial differences exist between LBM and SBM: the first one is the square versus rectangle nature of adjacency matrices (or data tables) which implies different asymptotics. The second and most important one is the nature of table margins: they represent the same objects in SBM (*e.g* graph nodes) but not in LBM (*e.g.* samples and features). Those differences explain why, although very similar from a theoretical point of view, LBM and SBM are used in different fields and for different applications.

The review is organized as follows, we present the probabilistic model of LBM in Section 2, known theoretical results in Section 3, discuss some issues and estimation procedures in Section 4, present a worked out example in Section 5 and conclude with some perspectives in Section 6. For the sake of clarity, we focus on the binary case (Govaert and Nadif, 2008; Keribin et al., 2012). Most definitions and results can be extended to the valued case.

 ISSN: 2102-6238

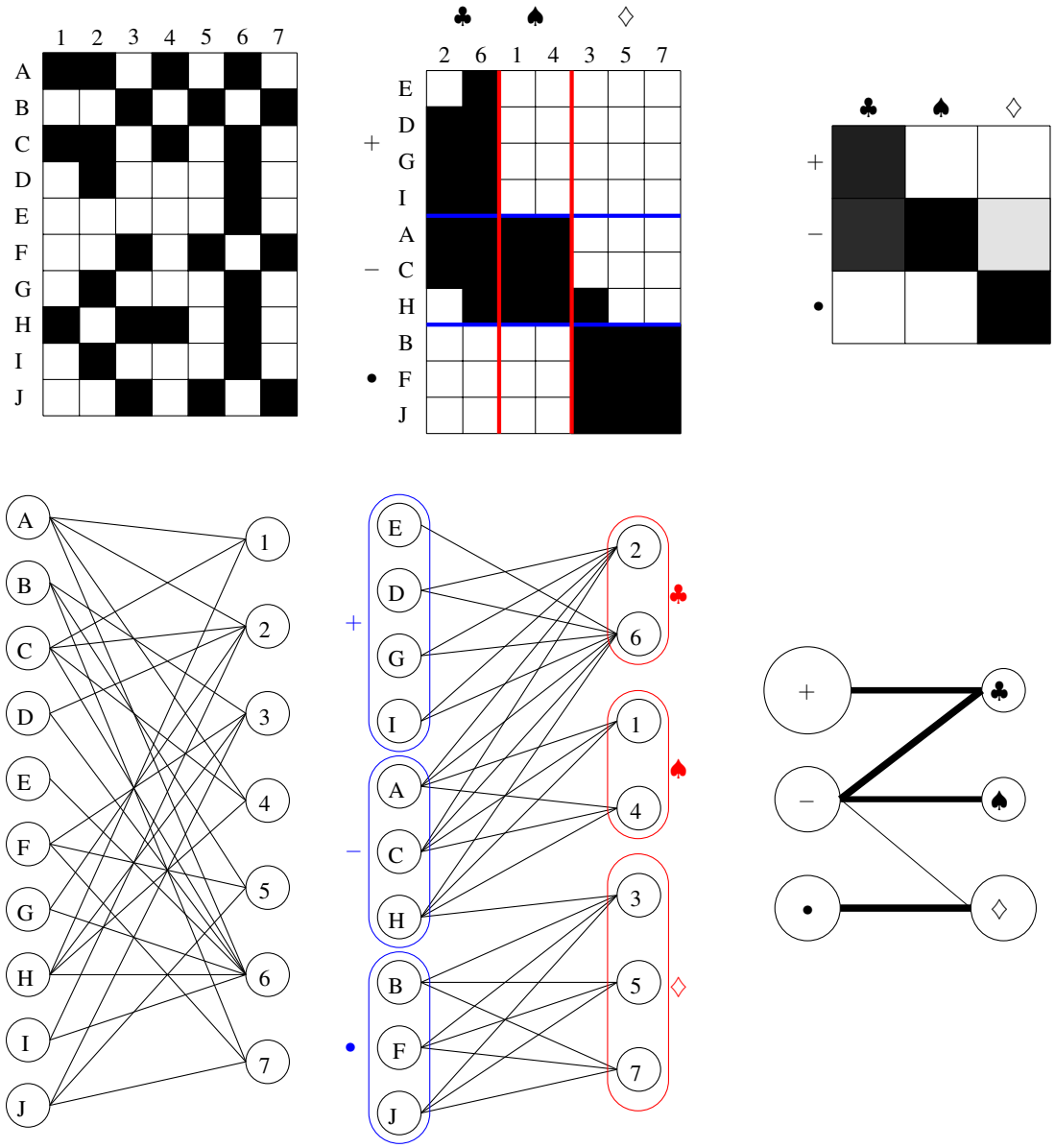

FIGURE 1. *Representation of LBM data table (top) and corresponding bi-partite graph (bottom). Left: Original data and graph. Middle: The array/graph is rearranged to sort rows/columns/nodes by group and highlight the underlying structure. Right: graphic representation of the LBM model parameters θ used to generate the data..*

 ISSN: 2102-6238

1	1	4	4	4	4
1	1	4	4	4	4
1	1	5	5	3	3
1	1	5	5	3	3
2	2	2	2	3	3
2	2	2	2	3	3

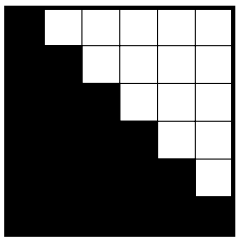

FIGURE 2. *Two examples of table structures that are not well captured by co-clustering techniques. Left: the natural structure of the table is captured by blocks indexed from 1 to 5, this structure can be fitted in a LBM model but only at the cost of redundant classes. Right: the table has a natural upper and lower triangular structure that is not block wise.*

2. Model and notations

2.1. Standard model

We consider a random binary matrix $\mathbf{x} = (x_{ij}; i = 1,\dots,n; j = 1,\dots,m)$ and interpret $x_{ij} = 1$ as a connection between row i and column j and consequently $x_{ij} = 0$ as a lack of connection. The distribution of $\mathbf{x}$ is specified by a latent structure on its rows and columns. Let $\mathbf{z} = (z_{iq}; i = 1,\dots,n; q = 1,\dots,Q)$ (resp. $\mathbf{w} = (w_{j\ell}; j = 1,\dots,m; \ell = 1,\dots,L)$) be the clustering of rows (resp. columns) into Q (resp. L) groups. $\mathbf{z}_i = (z_{iq})_q \in \{0,1\}^Q$ is the group indicator of row i, *i.e.* $\sum_q z_{iq} = 1$ and row i belongs to group q iff $z_{iq} = 1$. Similarly, $\mathbf{w}_j = (w_{j\ell})_\ell \in \{0,1\}^L$ is the group indicator of column j. $\mathbf{z}$ and $\mathbf{w}$ are the (usually unknown) latent structure.

Our model depends on the following parameters: two probability vectors $\alpha = (\alpha_1,\dots,\alpha_Q)$ and $\beta = (\beta_1,\dots,\beta_L)$ and an array of probabilities $\pi = (\pi_{q\ell})_{q=1\dots Q,\ell=1\dots L} \in (0,1)^{QL}$ collected in $\theta = (\alpha,\beta,\pi)$

We assume that:

– The $(\mathbf{z}_i)$ and the $(\mathbf{w}_j)$ are independent;
– The $(\mathbf{z}_i)$ are independent and identically distributed (i.i.d.) with $\mathbf{z}_i \sim \mathscr{M}(1;\alpha)$;
– The $(\mathbf{w}_j)$ are i.i.d. with $\mathbf{w}_j \sim \mathscr{M}(1;\beta)$;
– Conditionally on $\mathbf{z}$ and $\mathbf{w}$, the x_{ij} are independent with $x_{ij}|z_{iq}w_{j\ell} = 1 \sim \mathscr{B}(\pi_{q\ell})$

In other words, the rows and columns labels are independent, unlike in SBM where rows and columns stand for the same objects and share labels. α_q (resp. β_ℓ) is the proportion of rows (resp. columns) of type q (resp. ℓ) and $\pi_{q\ell}$ is the probability that a row of type q is connected to a column of type ℓ. Under these assumptions, the likelihood of $(\mathbf{z},\mathbf{w},\mathbf{x})$ can be factored as:

$$f(\mathbf{z},\mathbf{w},\mathbf{x};\theta) = f(\mathbf{z};\alpha)f(\mathbf{w};\beta)f(\mathbf{x};\mathbf{z},\mathbf{w},\pi) = \prod_{i,q}\alpha_q^{z_{iq}}\prod_{j,\ell}\beta_\ell^{w_{j\ell}}\prod_{i,j,q,\ell}\varphi(x_{ij};\pi_{q\ell})^{z_{iq}w_{j\ell}}$$

where $\varphi(x;\pi) = \pi^x(1-\pi)^{1-x}$ is the likelihood of a Bernoulli random variable. Note that this likelihood can be expressed in terms of exhaustive statistics as:

$$f(\mathbf{z},\mathbf{w},\mathbf{x};\theta) = \prod_q \alpha_q^{\mathbf{z}_{+q}}\prod_\ell \beta_\ell^{\mathbf{w}_{+\ell}}\prod_{q,\ell}\varphi((\mathbf{z}'\mathbf{x}\mathbf{w})_{q\ell},\mathbf{z}_{+q}\mathbf{w}_{+\ell};\pi_{q\ell}) \tag{1}$$

where $\mathbf{z}_{+q} = \sum_i z_{iq}$ is the count of row group q, $\mathbf{w}_{+\ell} = \sum_j w_{j\ell}$ is the count of column group q, $\mathbf{z}_{+q}\mathbf{w}_{+\ell}$ is the number of couples (i,j) of class (p,q), $(\mathbf{z}'\mathbf{x}\mathbf{w})_{q\ell} = \sum_{i,j} x_{ij}z_{iq}w_{j\ell}$ is the number

 ISSN: 2102-6238

of couples (i,j) of class (p,q) such that $x_{ij}=1$ and finally $\varphi(p,n;\pi)=\pi^p(1-\pi)^{n-p}$ is the likelihood of a binomial random variable.

2.2. Bayesian model

The principle of the Bayesian approach, in contrast to the frequentist point of view, is to suppose that model parameters are random variables with a prior distribution. For ease of computation, most authors choose conjugate priors (Raïffa and Schlaifer, 1961) for θ (Shan and Banerjee, 2008; Van Dijk et al., 2009; Wyse and Friel, 2012; Keribin et al., 2014). In our case, this means a Dirichlet distribution for proportions α and β. Note that Meeds and Roweis (2007) proposed a more general prior (Pitman-Yor prior).

For simplicity, the Dirichlet priors are centered around uniform proportions with dispersion parameters a_1 and a_2 (and often $a_1=a_2=a$):

$$\alpha\sim\mathscr{D}(a_1,\ldots,a_1)\text{ and }\beta\sim\mathscr{D}(a_2,\ldots,a_2)$$

and the π are assumed independent:

$$\forall(q,\ell),\ \pi_{q\ell}\sim\mathscr{B}e(b,b)\,.$$

Wyse and Friel (2012) use $a=1$ while Keribin et al. (2014) prefer $a=4$, since it decreases the frequency of empty classes. Both authors consider the model presented in section 2.1: all $\mathbf{z}_i$ (resp. $\mathbf{w}_j$) are sampled from $\mathscr{M}(1,\alpha)$ (resp. $\mathscr{M}(1,\beta)$). Other authors consider different formulations. For example, Shan and Banerjee (2008) create an additional sampling layer for $\mathbf{z}$ and $\mathbf{w}$: a different α_i is drawn for each row and $\mathbf{z}_i$ is then drawn from $\mathscr{M}(1,\alpha_i)$. The same applies for columns. This formulation drastically increases the dimension of the sampling space.

Another source of variation leading to different formulations lies in the nature of Q and L, the numbers of classes. Keribin et al. (2014) consider them as fixed parameters of the model (left half of Figure 3) whereas Wyse and Friel (2012); Van Dijk et al. (2009) consider it as a random variable and use a truncated Poisson distribution with mean λ as prior (right half of Figure 3).

The two approaches differ mostly in their method of choosing Q and Ł. The first performs one estimation per pair (Q,L) and then chooses among the pair using a model selection criterion (see section 3.5). The second uses only one run to estimate a posterior distribution for all parameters, including (Q,L). The unique run however takes longer to converge and the marginal posterior distribution of the pair is difficult to estimate (see section 3.3.3).

2.3. Sparse model

LBM models are mostly used in the context of big data where the density $\rho=\sum_{i,j}x_{ij}/mn$ of the table goes to 0 when $\min(n,m)$ goes to infinity, as the number of connections of a row scales sublinearly with the number m of columns. We therefore introduce a extension of LBM to sparse tables.

The sparse model behaves essentially like the standard LBM model with the following additional layer: each x_{ij} is independently randomly replaced by 0 with a probability $\rho_{m,n}$ that goes to

 ISSN: 2102-6238

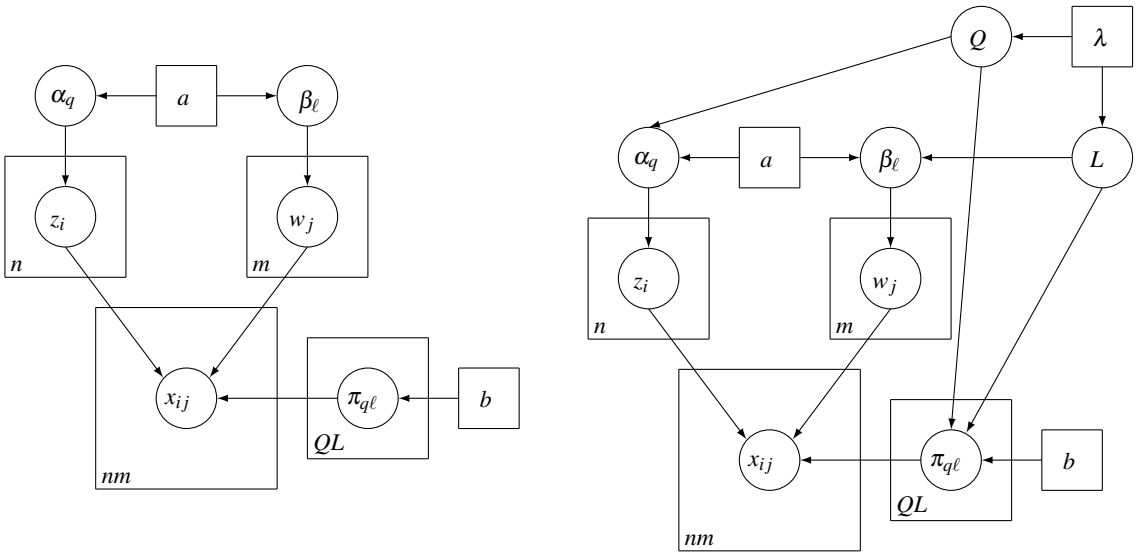

FIGURE 3. *Bayesian graphical model: Q and L can be either fixed (left) or random with a prior distribution (right).*

1 when $\min(n,m)$ goes to $+\infty$. Note that the sparse model induces a strong asymmetry between 0s and 1s: 1s are more informative than 0s.

Formally, the sparse model is a standard model with one additional parameter $\rho_{m,n}$ and a array of probabilities which now depends on (m,n) and is expressed as $\pi_{m,n} = \rho_{m,n}\pi$ for a fixed π. The likelihood is therefore:

$$f(\mathbf{z},\mathbf{w},\mathbf{x};\theta_{m,n}) = \prod_{i,q} \alpha_q^{z_{iq}} \prod_{j,\ell} \beta_\ell^{w_{j\ell}} \prod_{i,j,q,\ell} \varphi(x_{ij};\rho_{m,n}\pi_{q\ell})^{z_{iq}w_{j\ell}}.$$

where $\theta_{m,n} = (\alpha,\beta,\pi,\rho_{m,n})$. This parametrization retains the latent structure of the table and assumes that all probabilities go to 0 at the same rate. It is similar to the one proposed in Bickel and Chen (2009) for SBM.

2.4. *Group posterior distribution*

LBM in the context of model-based clustering are used with known $\mathbf{x}$ and unknown $\mathbf{z}$ and $\mathbf{w}$. We are therefore interested in the group posterior distribution of $(\mathbf{w},\mathbf{z})$ conditional on $\mathbf{x}$. The location and dispersion of this distribution are useful to understand how close the Maximum A Posteriori (MAP) label configuration is to the label configuration $(\mathbf{z}^\star,\mathbf{w}^\star)$ under which $\mathbf{x}$ was generated, refered to hereafter as the true configuration. We introduce here the log-likelihood ratio $\delta(\mathbf{w},\mathbf{z},\mathbf{w}^\star,\mathbf{z}^\star;\theta)$ of two configurations $(\mathbf{w},\mathbf{z})$ and $(\mathbf{w}^\star,\mathbf{z}^\star)$.

$$\delta(\mathbf{w},\mathbf{z},\mathbf{w}^\star,\mathbf{z}^\star;\theta) = \log(f(\mathbf{w},\mathbf{z}|\mathbf{x};\theta)) - \log(f(\mathbf{w}^\star,\mathbf{z}^\star|\mathbf{x};\theta)) \tag{2}$$

We will see in Section 3 that δ is close to its expected value. Furthermore, under technical assumptions detailed in Mariadassou and Matias (2015), $f(\mathbf{w},\mathbf{z}|\mathbf{x};\alpha,\beta,\pi)$ seen as a distribution

 ISSN: 2102-6238

over $\mathscr{Z} \times \mathscr{W} = \{1,\dots,Q\}^n \times \{1,\dots,L\}^m$, the set of all row and column configurations, is bounded from above and below by exponential-like distributions $Q(\mathbf{w},\mathbf{z}) \propto A\exp\left[-Cd^{\pi}((\mathbf{z},\mathbf{w}),(\mathbf{z}^{\star},\mathbf{w}^{\star}))\right]$ for some constants A, C and a semi-metric d^{π} defined on $(\mathscr{Z} \times \mathscr{W})^2$ defined in Section 3.4.

3. Theoretical results

Model-based co-clustering using LBM involves several steps ranging from appropriate model parametrization to effective co-clustering. The theoretical analysis of each of these steps comes with its own difficulties. We review here the results about identifiability, likelihood-based parameter estimation, clustering error and model selection, understood here as the choice of (Q,L).

3.1. Identifiability

Obviously LBM, as a mixture model, is not identifiable due to invariance to label permutation. This issue is irrelevant to maximum likelihood estimation and mostly affects Bayesian estimation by impeding the ability of a typical MCMC sampler to explore the configuration space. Standard techniques used in Bayesian estimation for mixture models (Frühwirth-Schnatter, 2006) can however be imported as such to mitigate this problem. Note also that identifiability *up to a permutation of the labels* is sufficient for clustering purposes as the labels themselves are of no interest.

Unfortunately, multivariate Bernoulli mixtures are generally not identifiable (Gyllenberg et al., 1994), regardless of invariance to relabelling. Allman et al. (2009) exhibited a sufficient condition for identifiability of Bernoulli mixtures that can not directly be applied to LBM. Building upon the results of Celisse et al. (2012) for SBM, Keribin et al. (2014) defined a set of sufficient conditions for binary LBM:

C_1: for all $1 \leq q \leq Q$, $\alpha_q > 0$ and the elements of vector $\tau = \pi\beta$ are distinct;
C_2: for all $1 \leq \ell \leq L$, $\beta_\ell > 0$ and the elements of vector $\sigma = \alpha'\pi$ are distinct;
C_3: $n \geq 2L-1$ and $m \geq 2Q-1$.

Under conditions (C_{1-3}), binary LBM is identifiable. The first half of condition C_1 is very natural: a group with proportion 0 is irrelevant and reveals a model that is overparametrized and therefore not identifiable. The second half is more interesting: τ_q is the scaled average number of 1s in a row from group q. C_1 therefore requires that each row group has a distinct average number of connections. Condition C_2 is symmetric to C_1 and requires that no column group is irrelevant and that each group has a unique average number of connections. Finally condition C_3 stems from the nature of the proof, which proceeds by building a system of algebraic equations in α_q, β_ℓ and $\pi_{q\ell}$ that must satisfied by a LBM model. This system has a unique solution as soon as condition C_3 holds. Note that in the special case of two groups in both rows and columns ($Q = L = 2$), conditions C_{1-2} are not necessary to ensure identifiability (Keribin et al., 2014).

Conditions (C_{1-2}) also appear in SBM and LBM-related works, most notably in Channarond et al. (2012); Brault (2014) where the authors leverage the uniqueness of (τ_q) to propose a fast and computation-light graph clustering technique. The principle is to first estimate the scaled number of 1s in each row and then to cluster rows based on these numbers. The estimates indeed converge towards distinct limits that are more separated with high values of m and large gaps between

 ISSN: 2102-6238

values of τ_q. The technique is less efficient that likelihood-based ones but offers a feasible and welcome alternative in the context of large graphs and large matrices.

3.2. Likelihood

We saw in section 2 that the classification likelihood of $(\mathbf{x},\mathbf{w},\mathbf{z};\theta)$ has a nice product form. The same does not hold for the likelihood $f(\mathbf{x};\theta)$ which involves a sum over a large set:

$$f(\mathbf{x};\theta)=\sum_{\mathbf{z},\mathbf{w}\in\mathscr{Z}\times\mathscr{W}}\prod_{i,q}\alpha_q^{z_{iq}}\prod_{j,\ell}\beta_\ell^{w_{j\ell}}\prod_{i,j,q,\ell}\varphi(x_{ij};\pi_{q\ell})^{z_{iq}w_{j\ell}}. \tag{3}$$

This combinatorial challenge is of course shared by many models with latent structure, including standard mixture models, but unlike other more structured problems the sum can not be simplified and would require unreasonable computing power for small data tables. The cardinal of $\mathscr{Z}\times\mathscr{W}$ is Q^nL^m. A square matrix of size $n=m=20$ with $Q=L=2$ groups on rows and columns requires a summation over 10^{12} unique configurations. At the rate of 100 000 configurations per second, computing the likelihood takes roughly 33 years (without mentioning numerical error problems).

Our inability to compute the likelihood means that methods based on the observed likelihood such as parameter estimation or model selection are untractable. We rely instead on EM-like procedures to maximize a tractable lower-bound of the observed likelihood.

3.3. Estimation

3.3.1. Variational EM

Consider the following functional of θ and a distribution $\mathbb{Q}$ over $\mathscr{Z}\times\mathscr{W}$:

$$F(\theta,\mathbb{Q})=\log f(\mathbf{x};\theta)-KL(\mathbb{Q},f(.|\mathbf{x};\theta))=\mathbb{E}_{\mathbb{Q}}[\log f(\mathbf{x},\mathbf{z},\mathbf{w};\theta)]+\mathscr{H}(\mathbb{Q}) \tag{4}$$

where KL and $\mathscr{H}$ are the Kullback-Leibler divergence and the Shannon entropy. It is clear from equation (4) that $F(\theta,\mathbb{Q})\leq\log f(\mathbf{x};\theta)$ with equality iff $\mathbb{Q}=f(.|\mathbf{x};\theta)$. The maximum likelihood estimate $\hat{\theta}_{ML}$ of θ is therefore:

$$\hat{\theta}_{ML}=\operatorname*{argmax}_\theta F(\theta,f(.|\mathbf{x};\theta))=\operatorname*{argmax}_\theta\{\operatorname*{argmax}_{\mathbb{Q}}F(\theta,\mathbb{Q})\}. \tag{5}$$

The joint maximization of F in $(\theta,\mathbb{Q})$ is equivalent to maximum likelihood estimation and therefore untractable. The classic Expectation-Maximization (EM) algorithm replaces the joint estimation by a coordinate-wise gradient descent where F is optimized iteratively and in turn in θ and $\mathbb{Q}$. Optimizing F in θ for fixed $\mathbb{Q}$ is relatively easy but the reverse is unfortunately not true. The optimal $\mathbb{Q}$ for a given θ is of course $\mathbb{Q}=f(.|\mathbf{x};\theta)$. Unfortunately and unlike more structured mixture models, the dependency graph of $(\mathbf{z},\mathbf{w})$ conditional on $\mathbf{x}$ is a complete bipartite graph: each $\mathbf{z}_i$ depends on every $\mathbf{w}_j$ and vice-versa. This means that computing the optimal $\mathbb{Q}$ is in general as hard as directly computing the likelihood since $\mathscr{H}(\mathbb{Q})$ and $f(\mathbf{x},\theta)$ involve a sum over

 ISSN: 2102-6238

the same set. An important exception where $\mathscr{H}(\mathbb{Q})$ is tractable is the special case where $\mathbb{Q}$ is a factor distribution, *i.e.* the $\mathbf{z}_i$ and $\mathbf{w}_j$ are independent, but not necessarily identically distributed, under $\mathbb{Q}$. In this special case, $\mathscr{H}(\mathbb{Q})$ involves a sum over $nQ+mL$ terms, instead of Q^nL^m in the general case.

We therefore consider the following estimator:

$$\hat{\theta}_{VEM} = \underset{\theta}{\operatorname{argmax}} \left\{ \underset{\{\mathbb{Q} \in \mathscr{I}\}}{\operatorname{argmax}} F(\theta, \mathbb{Q}) \right\} \tag{6}$$

where $\mathscr{I}$ is the set of factor distributions over $\mathscr{Z} \times \mathscr{W}$. VEM stands for Variational approximation EM. Approximation refers to the fact that during the E step, we optimize F over $\mathscr{I}$ instead of the general set of distributions over $\mathscr{Z} \times \mathscr{W}$, *i.e.* the procedure maximizes a (hopefully tight) lower-bound of the likelihood. Strictly speaking, variational is redundant as the EM procedure is already a variational one: it reframes the objective function $\log(f(\mathbf{x};\theta))$ as the maximum over $\mathbb{Q}$ of the functional $F(\theta,\mathbb{Q})$ and maximizes $\log(f(\mathbf{x};\theta))$ through F. This is not however the classic presentation of the EM algorithm (Dempster et al., 1977) and we stick to VEM for consistency with the literature. Note that a distribution $\mathbb{Q}$ in $\mathscr{I}$ is fully determined by the quantities $\tau_{iq} = \mathbb{Q}(z_{iq} = 1)$ and $\upsilon_{j\ell} = \mathbb{Q}(w_{j\ell} = 1)$. With these notations, the functional $F(\theta,\mathbb{Q})$ reduces to:

$$F(\theta,\mathbb{Q}) = \sum_{i,q} \tau_{iq} \log \alpha_q + \sum_{j,\ell} \upsilon_{j\ell} \log \beta_\ell + \sum_{i,j,q,\ell} \tau_{iq}\upsilon_{j\ell}[x_{ij} \log \pi_{q\ell} + (1-x_{ij}) \log(1-\pi_{q\ell})] \\ - \sum_{i,q} \tau_{iq} \log \tau_{iq} - \sum_{j,\ell} \upsilon_{j\ell} \log \upsilon_{j\ell}. \tag{7}$$

Joint optimization of F over θ and the (τ_{iq}) and $(\upsilon_{j\ell})$ is still a hard problem and we resort to an iterative coordinate-wise optimization procedure. We now detail each step of this procedure.

M step: Optimization of F in θ for fixed $\mathbb{Q}$ is easy as the entropy terms $\mathscr{H}(\mathbb{Q})$ does not involve θ and can be left out of the optimization. The remaining term $\mathbb{E}_{\mathbb{Q}}[\log f(\mathbf{x},\mathbf{z},\mathbf{w};\theta)]$ only involves first and second order moments of $\mathbb{Q}$ and straightforward computations give intuitive closed formula for elements of $\hat{\theta} = \underset{\theta}{\operatorname{argmax}} F(\theta,\mathbb{Q})$:

$$\hat{\alpha}_q = \frac{1}{n}\sum_i \tau_{iq}$$
$$\hat{\beta}_\ell = \frac{1}{m}\sum_j \upsilon_{j\ell}$$
$$\hat{\pi}_{q\ell} = \sum_{i,j} x_{ij}\tau_{iq}\upsilon_{j\ell} \Big/ \sum_{i,j} \tau_{iq}\upsilon_{j\ell}$$

VE step: Optimization of F in $\mathbb{Q}$ for fixed θ does not yield a closed formula for the (τ_{iq}) and $(\upsilon_{j\ell})$. Optimizing F with respect to these quantities and the constraints $(\tau_{iq})_q \in [0,1]^Q$ and $\sum_q \tau_{iq} = 1$ for all i, and similarly for $(\upsilon_{j\ell})$, yields coupled fixed point equations satisfied by

 ISSN: 2102-6238

$(\hat{\tau}_{iq})$ and $(\hat{\upsilon}_{j\ell})$:

$$\hat{\tau}_{iq} \propto \alpha_q \prod_{j,\ell} \varphi(x_{ij}; \pi_{q\ell})^{\hat{\upsilon}_{j\ell}} \quad \forall i, q$$
$$\hat{\upsilon}_{j\ell} \propto \beta_\ell \prod_{i,q} \varphi(x_{ij}; \pi_{q\ell})^{\hat{\tau}_{iq}} \quad \forall j, \ell$$

3.3.2. *Other approximation schemes*

The VEM procedure to find an estimate of θ was first proposed in Govaert and Nadif (2008), where the authors also showed that a standard *EM* procedure was untractable. The *VEM* procedure gives satisfying estimates but those estimates are highly dependent on a good initialization and have a marked tendency to produce empty clusters when plugged in a classification rule. Several alternative variants of the *EM* algorithm have been proposed:

- the *CEM* (Classification EM) algorithm (Govaert and Nadif, 2008) attempts to maximize the classification likelihood instead of the standard one, it shifts the focus of the procedure from parameter estimation to row and column clustering.
- the *SEM* (Stochastic EM) algorithm (Keribin et al., 2014) adds a stochastic step to the procedure: at the end of the VE step, each row i (resp. column j) is assigned to a group with probabilities proportional to τ_{iq} (resp $\upsilon_{j\ell}$). This stochastic step prevents the procedure from being stucked in local optima of the functional F.

These two examples are the closest to the likelihood-based VEM procedure we presented but by no means the only ones. A presentation and discussion of other algorithms, probabilistic, deterministic, based on the likelihood or other loss functions can be found in Brault and Lomet (2014).

3.3.3. *Bayesian algorithms*

Using the left hand side model of Figure 3 (Q and L fixed), Keribin et al. (2014) propose a Bayesian version of the *VEM* algorithm, called *V-Bayes*. V-Bayes is initialized by a *Gibbs* sampler to estimate the mode of the posterior law and use $a = 4$ for the Dirichlet priors of the group proportions. This choice limits the empirical tendency, observed in the frequentist setting, of the algorithm to provide empty clusters.

Using the right hand side model of Figure 3 (Q and L random), Wyse and Friel (2012) propose a collapsed sampler to estimate $\mathbf{z}$, $\mathbf{w}$, θ, Q and L. Empirical tests suggest that the algorithm provides satisfactory estimates but requires a large number of iterations to converge and tends to overestimate Q and L (see Keribin et al., 2014). In the same setting, with the same assumption, Van Dijk et al. (2009) use a *Gibbs* sampler and Shan and Banerjee (2008) a variational algorithm to estimate all the parameters.

 ISSN: 2102-6238

3.3.4. Classification rules

Parameter estimates $\hat{\theta}$ allow us to classify rows and columns using the Maximum a Posteriori (MAP) rule defined by:

$$\hat{z}_{iq} = \begin{cases} 1 & \text{if} \quad q = \text{argmax}_{q'} \, \mathbb{P}(z_{iq'} = 1|\mathbf{x}; \hat{\theta}) \\ 0 & \text{else} \end{cases}$$
$$\hat{w}_{j\ell} = \begin{cases} 1 & \text{if} \quad \ell = \text{argmax}_{\ell'} \, \mathbb{P}(w_{j\ell'} = 1|\mathbf{x}; \hat{\theta}) \\ 0 & \text{else} \end{cases}$$

$\mathbb{P}(z_{iq} = 1|\mathbf{x}; \hat{\theta})$ and $\mathbb{P}(w_{j\ell} = 1|\mathbf{x}; \hat{\theta})$ are in general as hard to compute as $f(\mathbf{x}; \hat{\theta})$ but variational procedures provide us with readily available estimates of those quantities in the form of $\hat{\tau}_{iq}$ and $\hat{\upsilon}_{j\ell}$. We therefore consider the following classification rule, which is a tractable approximation to the MAP rule:

$$\hat{z}_{iq} = \begin{cases} 1 & \text{if} \quad q = \text{argmax}_{q'} \, \tau_{iq'} \\ 0 & \text{else} \end{cases} \qquad \text{and} \qquad \hat{w}_{j\ell} = \begin{cases} 1 & \text{if} \quad \ell = \text{argmax}_{\ell'} \, \upsilon_{j\ell'} \\ 0 & \text{else} \end{cases}$$

3.3.5. Algorithmic complexity

Variational procedures fall in the large class of iterative procedures. As such, the complexity depends on an unknown number S_{EM} of iterations. Each M and VE step can nevertheless be analyzed independently. During the M step, $O(n)$ operations are required for each $\hat{\alpha}_q$, $O(m)$ for each $\hat{\beta}_l$ and $O(mn)$ for each $\hat{\pi}_{ql}$ resulting in a complexity of $O(mnQL)$ for this step. The VE step is itself iterative as the $(\hat{\tau}_{iq})$s and $(\hat{\upsilon}_{j\ell})$ are solutions of $nQ + mL$ coupled fixed-point equations. Each update of $\hat{\tau}_{iq}$ requires $O(mL)$ operations and each update of $\hat{\upsilon}_{j\ell}$ requires $O(nQ)$ operations. The total complexity of the VE step is therefore $O(mnQLS_{FP})$ where S_{FP} is an upper bound of the number of updates used to solve the fixed-point equations. VEM therefore has an overall $O(mnQLS_{EM}S_{FP})$ complexity with S_{EM} and S_{FP} stemming from iterative procedures. The Bayesian procedures based on a Gibbs sampler have similar complexity of $O(nmQLS_{Gibbs})$ where S_{Gibbs} depends on the nature of Q and L and the number of iterations before convergence.

3.4. Asymptotic behavior of LBM

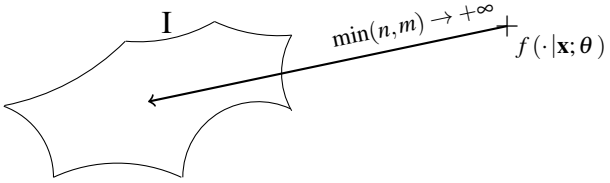

FIGURE 4. *$f(\cdot|\mathbf{x};\theta)$ converges with high probability to a product, and even a Dirac, distribution as* $\min(n,m)$ *goes to* $+\infty$.

 ISSN: 2102-6238

Rigorous analysis of variational approximation procedures are quite involved and theoretical results (Gunawardana and Byrne, 2005) show that the resulting estimates are biased except in degenerate cases when the approximation is in fact not an approximation. Numerical studies (Gazal et al., 2011) providing very accurate estimates for moderate values of $\min(n,m)$ suggested that LBM may be degenerate. This was first proved for SBM by Celisse et al. (2012) and later extended to LBM by Mariadassou and Matias (2015).

3.4.1. Group posterior distribution

Intuitively, the degeneracy arises from the group posterior distribution $f(.|\mathbf{x};\theta)$ converging to a Dirac mass on the true configuration (see Figure 4). Since a Dirac mass on $\mathscr{Z}\times\mathscr{W}$ is a factor distribution and belongs to $\mathscr{I}$, the approximation part of the VEM procedure is asymptotically exact.

The crux of the proof lies in bounding the likelihood ratio introduced in equation (2). Under technical conditions that exclude some probability matrices π and ill-behaved true configurations $(\mathbf{z}^\star,\mathbf{w}^\star)$, we can show that $\delta(\mathbf{w},\mathbf{z},\mathbf{w}^\star,\mathbf{z}^\star;\theta)$ is close to its deterministic expected value $\Delta(\mathbf{w},\mathbf{z},\mathbf{w}^\star,\mathbf{z}^\star;\theta)$ with high probability (whp).

Note that matrices π are excluded only if their coefficient satisfy some algebraic equations: excluded matrices exhibit a lot of coefficient symmetry and form a set of null Lebesgue measure in the matrix space $\mathscr{M}_{n,m}([0,1])$. Similarly, configurations $(\mathbf{z}^\star,\mathbf{w}^\star)$ are ill-behaved if some group counts $\mathbf{z}_{+q}$ (resp. $\mathbf{w}_{+\ell}$) are unexpectedly low, *i.e.* not of order $\Omega(n)$ (resp. $\Omega(m)$). The set of such configurations has an exponentially vanishing probability.

Furthermore, $\Delta(\mathbf{w},\mathbf{z},\mathbf{w}^\star,\mathbf{z}^\star;\theta)$ is bounded from above by:

$$\Delta(\mathbf{w},\mathbf{z},\mathbf{w}^\star,\mathbf{z}^\star;\theta)\leq(\|\mathbf{z}-\mathbf{z}^\star\|_1+\|\mathbf{w}-\mathbf{w}^\star\|_1)|\log\eta|-d^\pi(\mathbf{w},\mathbf{z},\mathbf{w}^\star,\mathbf{z}^\star)\kappa_{\min} \tag{8}$$

where $\eta=\min(\min_q\alpha_q,\min_\ell\beta_\ell)$ is the smallest group proportion, $\kappa_{\min}=\min_{\{q,\ell,q',\ell':\pi_{q\ell}\neq\pi_{q'\ell'}\}}KL(\pi_{q\ell},\pi_{q'\ell'})$ is the smallest positive Kullback-Leibler divergence between Bernoulli distributions with parameters in π and $d^\pi(\mathbf{w},\mathbf{z},\mathbf{w}^\star,\mathbf{z}^\star)=\sum_{i,j}\sum_{q,\ell}1\{\pi_{q\ell}\neq\pi_{q'\ell'}\}z_{iq'}w_{j\ell'}z^\star_{iq}w^\star_{j\ell}$ is a semi-metric indexed by π that counts the number of couples (i,j) with a different parameter $\pi_{q\ell}$ under configurations $(\mathbf{w},\mathbf{z})$ and $(\mathbf{w}^\star,\mathbf{z}^\star)$. d^π is maximal when each coefficient of π is unique and reduces then to $d^\pi(\mathbf{w},\mathbf{z},\mathbf{w}^\star,\mathbf{z}^\star)=n_1m+m_1n-n_1m_1$ where $n_1=\|\mathbf{z}-\mathbf{z}^\star\|_1$ and $m_1=\|\mathbf{w}-\mathbf{w}^\star\|_1$. For configurations $(\mathbf{w},\mathbf{z})$ away from $(\mathbf{w}^\star,\mathbf{z}^\star)$, a combinatorial arguments shows that $d^\pi(\mathbf{w},\mathbf{z},\mathbf{w}^\star,\mathbf{z}^\star)$ is bounded from below by $\eta^2(n_1m+m_1n)/8$ for any π that satisfies conditions C_{2-3}.

Returning to the likelihood ratio of one configuration against the true one, whp:

$$\frac{f(\mathbf{z},\mathbf{w}|\mathbf{x};\theta)}{f(\mathbf{z}^\star,\mathbf{w}^\star|\mathbf{x};\theta)}\leq\exp\left[-n_1\left(mc+\log\eta\right)-m_1\left(nc+\log\eta\right)\right] \tag{9}$$

where $c=\kappa_{\min}\eta^2/8$. This further results in

$$KL(f(.|\mathbf{x};\theta),\delta_{(\mathbf{w}^\star,\mathbf{z}^\star)})\leq a_{n,m}\exp(a_{n,m}) \tag{10}$$

where $a_{n,m}=n\exp(-c(m+\log\eta))+m\exp(-c(n+\log\eta))$. The smaller $a_{n,m}$, the tighter the lower bound in equation (4) and the more accurate the approximation of searching for $\mathbb{Q}\in\mathscr{I}$ in

 ISSN: 2102-6238

the VE step. This result also holds, with a smaller and less elegant rate c, when replacing the true parameter θ by a close value $\hat{\theta}$, for example an estimate. Asymptotic considerations show that $a_{n,m} \to 0$ with $\min(n,m)$ as soon as $\log m = o(n)$ and $\log n = o(m)$. This is also true for sparse graphs, as long as $\log m = o(n\rho_{m,n})$ and $\log n = o(\log m \rho_{m,n})$.

3.4.2. Consistency

No self contained consistency proof of variational procedures of parameter estimation exists for LBM. The closest is a consistency result for SBM (Celisse et al., 2012) but it is only a partial one as $\hat{\pi}$ and $\hat{\alpha}$ behave differently. $\hat{\pi}$ is consistent in general but α is only consistent if $\hat{\pi}$ converges quickly enough: namely $\|\hat{\pi} - \pi\|_1 = o(1/n)$. It is not yet known whether this speed condition is necessary or holds and therefore whether $\hat{\theta}_{VEM}$ is consistent in general.

The proof proceeds as follows: the authors first prove that the (untractable) maximum likelihood estimate $\hat{\theta}_{ML}$ of θ is consistent, under a similar speed condition for $\|\hat{\pi}_{ML} - \pi\|_1$. They then prove that the variational procedure is asymptotically equivalent to maximum likelihood estimation as $f(.|\mathbf{x};\theta)$ tends towards $\delta_{(\mathbf{z}^\star,\mathbf{w}^\star)}$. The same results have not been proved in the LBM framework but should also hold, albeit with different asymptotic rates.

Numeric results and simulation studies Gazal et al. (2011) suggest that the variational procedure is consistent and that $\hat{\pi}$ does satisfy $\|\hat{\pi} - \pi\|_1 = o(\log n/n)$ and therefore that $\hat{\theta}_{VEM}$ is consistent but no formal proof exists. Consequently, general consistency of the VEM for LBM remains an open question.

3.5. Model selection

Choosing a relevant number of clusters (Q,L) is of crucial importance in LBM, just like in most latent structure models. This model selection problem is difficult for several reasons. First, there are two numbers of clusters, Q and L, to select instead of a single one. Second, penalized likelihood criteria such as AIC or BIC are not directly available since computing the maximized likelihood is not feasible. Third, the definition and number of statistical units is not well defined in LBM (number of rows, number of columns, number of cells, ...).

Using the left hand side model of Figure 3, Keribin et al. (2014) proposed an exact form of the *Integrated Completed Likehood* (ICL) criterion that selects a model for the purpose of classification. For each pair (Q,L), a group label $(\widehat{\mathbf{z}},\widehat{\mathbf{w}})$ and parameter value $\widehat{\theta}$ are proposed (Keribin et al. (2014) suggests to take the ones produced by the *V-Bayes* algorithm) and the ICL criterion selects the couple (Q,L) maximizing:

$$\mathrm{ICL}(Q,L) = \log f\left(\mathbf{x},\widehat{\mathbf{z}},\widehat{\mathbf{w}};\widehat{\theta}\right).$$

Using the right hand side model of Figure 3, Wyse et al. (2014) also suggested ICL to select the number of component and built upon previous work done for SBM (Côme and Latouche, 2013) to propose a faster alternative to the original MCMC of Wyse and Friel (2012).

 ISSN: 2102-6238

Upon careful examination of its asymptotic form, Keribin et al. (2014) suggest that the BIC criterion can be formulated as:

$$\mathrm{BIC}(Q,L) = \max_{\theta} \log p(\mathbf{x};\theta) - \frac{Q-1}{2}\log n - \frac{L-1}{2}\log m - \frac{QL}{2}\log(nm). \tag{11}$$

The BIC criterion is known to be consistent for choosing the number of groups in classical mixtures. They suggest to approximate the log-likelihood maximum by its variational approximation. Using the results of Mariadassou and Matias (2015) and Celisse et al. (2012), they conjecture that the two criteria (ICL and BIC) are asymptotically equivalent for LBM and therefore that either both are consistent for choosing (Q,L) or none of them is.

4. Evaluation and alternatives

4.1. Evaluation of results

LBM models have been defined in this review in the context of co-clustering. As such, estimation procedures can and should be evaluated not only on the accuracy of $\hat{\theta}$ but also on the accuracy of $\hat{\mathbf{z}}$ and $\hat{\mathbf{w}}$. The focus can even be only on $\hat{\mathbf{z}}$ and $\hat{\mathbf{w}}$ if θ is regarded as a nuisance parameter. We consider here the MAP-like estimates $\hat{\mathbf{w}}$ and $\hat{\mathbf{z}}$ defined in Section 3.3.4. Since $\mathbf{z}$ and $\mathbf{w}$ are identifiable at most up to a permutation, we consider the following semi-norms:

$$\|\mathbf{z}' - \mathbf{z}\| = \inf_{\sigma \in S_Q} \frac{1}{2}\sum_{i,q} |z'_{i\sigma(q)} - z_{iq}|$$

$$\|\mathbf{w}' - \mathbf{w}\| = \inf_{\sigma \in S_L} \frac{1}{2}\sum_{j,\ell} |w'_{j\sigma(\ell)} - w_{j\ell}|$$

where S_n is the symmetric group of $\{1,\dots,n\}$. The asymptotic results presented in section 3.4 show that, if a good initial estimate of π is available, the VEM procedure and other approximation schemes produce an asymptotically perfect classification in the sense that

$$\|\mathbf{z}' - \mathbf{z}\| \to 0 \quad \text{and} \quad \|\mathbf{w}' - \mathbf{w}\| \to 0 \tag{12}$$

when n and m go to ∞ with the asymptotics specified in section 3.4. This result is quite powerful and can accommodate extension to sparse models. However, it is not obvious that this is the correct way to assess the asymptotic performance of a co-clustering procedure. Indeed, since $\mathbf{z}$ and $\mathbf{w}$ grow linearly with n and m, a pragmatic alternative to perfect classification would be low classification error rate, or formally:

$$\frac{\|\mathbf{z}' - \mathbf{z}\|}{n} = o(1) \quad \text{and} \quad \frac{\|\mathbf{w}' - \mathbf{w}\|}{m} = o(1). \tag{13}$$

Results in the SBM literature Rohe et al. (2011); Choi et al. (2012) and in the LBM literature Brault (2014) show that other procedures achieve this goal under both the standard and sparse models but can also accommodate other relaxations, such as growing number of row and columns groups, of order $Q_{n,m} = \Omega\left(\sqrt{m/\log n}\right)$ and $L_{n,m} = \Omega\left(\sqrt{n/\log m}\right)$.

 ISSN: 2102-6238

Apart from quantifying the accuracy of the result, assessing the difficulty of the co-clustering task for a given model or table is itself a difficult task. Asymptotic results show that the difficulty is dominated, in the worst case scenario, by $\alpha_{\min}\beta_{\min}\kappa_{\min}$. The worst case scenario is achieved when, for example, row groups are very unbalanced and have very similar probability vectors, differing only with respect to rare column groups. On the contrary, balanced groups with markedly different probability vectors constitute a favorable setting for co-clustering techniques. This characterization relies on the model parameters only and neglects effects due to data generation, with some draws being more favorable than others. Lomet et al. (2012) developed an interesting approach where difficulty is assessed *conditionally* to the draw of $\mathbf{x}$. Data sets are generated under a LBM model and authors compute the difficulty as an approximate Bayes classification error rate of data table items given by:

$$R(\mathbf{x},\mathbf{z},\mathbf{w}) = \inf_{(\sigma_1,\sigma_2)\in S_Q\times S_L} \frac{1}{2nm}\sum_{i,j} - \max\left(\sum_q |z_{iq} - \hat{z}_{i\sigma_1(q)}|, \sum_l |w_{j\ell} - \hat{w}_{j\sigma_2(l)}|\right)$$

where $(\mathbf{w},\mathbf{z})$ are the true labels and $(\hat{\mathbf{w}},\hat{\mathbf{z}}) = \operatorname{argmax} f(.|\mathbf{x};\theta)$ are MAP estimates of $(\mathbf{w},\mathbf{z})$. Since $(\hat{\mathbf{w}},\hat{\mathbf{z}})$ are hard to compute exactly, even when θ is known, the risk is computed on approximate versions, resulting from one of the estimation and classification procedures described in Section 3.3. Intuitively, $R(\mathbf{x},\mathbf{z},\mathbf{w})$ is the fraction of couples (i,j) in different classes (q,l) under classification $(\mathbf{w},\mathbf{z})$ and $(\hat{\mathbf{w}},\hat{\mathbf{z}})$ and corresponds to the lowest possible classification error rate. It is closely related to the previous semi-norms through:

$$R(\mathbf{x},\mathbf{z},\mathbf{w}) = \frac{1}{2mn}(m\|\mathbf{z}-\hat{\mathbf{z}}\| + n\|\mathbf{w}-\hat{\mathbf{w}}\| - \|\mathbf{z}-\hat{\mathbf{z}}\|\|\mathbf{w}-\hat{\mathbf{w}}\|).$$

Lomet et al. (2012) produced 72 artificial tables $\mathbf{x}$ with known labels $(\mathbf{z},\mathbf{w})$, sizes ranging from 50 to 500, number of groups ranging from 3 to 10 and error rates ranging from 5% to 20%. Our main complaint with the data set is that all tables are squares ($n = m$) and have equal number of row and column groups ($Q = L$) and therefore do not explore the variety of LBM settings. The dataset, available at `www.hds.utc.fr/coclustering/doku.php`, is nevertheless a convenient way to benchmark the performance of co-clustering techniques.

4.2. Deterministic methods

We presented in this review model-based procedures to recover the clusters. The probabilistic framework renders these procedures quite natural and enables one to derive careful analytic properties. Model-based procedures are however far from the only ones available to do co-clustering: they coexist with a wide variety of deterministic methods, most notably based on matrix factorization. We briefly discuss some popular ones:

- *"CRO" methods*, proposed by Govaert (1983) for binary data, search the blockwise partition of the table minimizing the loss of information between the data matrix and a simple blockwise summary, where each block is filled with its most abundant value (either 0 or 1).
- *Non-negative Matrix Factorization (NMF)* (Seung and Lee, 2001) searches for the decomposition of a non-negative matrix as a product of two arbitrary non-negative matrices:

$$\min_{A\in\mathbb{R}_+^{n\times d}, B\in\mathbb{R}_+^{d\times m}} \|\mathbf{x} - AB\|^2,$$

 ISSN: 2102-6238

where d is significantly smaller than both n and m. Each column of $\mathbf{x}$ is approximated by a linear combination of the columns of A, weighted by the component of the corresponding columns of B.

– *Non-negative Tri-Factorization (NTF)* takes NMF a step further and searches for a decomposition as the product of 3 matrices: a non-negative matrix $A = (A_{q\ell}; q = 1, \ldots, Q; \ell = 1, \ldots, L)$ (that plays a role similar to π) and two classification matrices $\mathbf{z}$ and $\mathbf{w}$:

$$\min_{\mathbf{z}\in\mathbb{R}_+^{n\times Q}, A\in\mathbb{R}_+^{Q\times L}, \mathbf{w}\in\mathbb{R}_+^{m\times L}} \left\|\mathbf{x} - \mathbf{z}A\mathbf{w}^T\right\|^2.$$

– *Non-negative Block Value Decomposition (NBVD)* (Long et al., 2005) is likewise based on an alternating minimization of the NTF criterion. Yoo and Choi (2010) additionally suggest to enforce orthogonal $\mathbf{z}$ and $\mathbf{w}$. This constraint allows a strict interpretation of the classification.

These four methods are not probabilistic in nature and no analytic property can be derived in general. They can nevertheless be studied in the LBM framework to assess their properties, such as classification error rates, on LBM generated data.

5. Case Study

To contrast the behavior of co-clustering and margin-wise clustering approaches, we applied them to the *UCI Congressional Voting Records* data set [1], already studied by Keribin et al. (2014) and Wyse and Friel (2012). It records the votes of the 435 members of the House of Representatives of the 98th congress of the United States of America on 16 different key issues. The votes consist of 'yes', 'no', 'abstained or absent' and are modeled as categorical data with 3 levels.

For the co-clustering approach, we used *Gibbs+V-Bayes* on categorical data with $(a,b) = (4,1)$ and selected the classes using *BIC*, as in Keribin et al. (2014). We obtained $(Q,L) = (4,6)$ classes, as represented in Figure 5 (left panel). Concerning the margin-wise clustering, we considered different models for the representatives and the issues. Each representative (row) $x_{i,\cdot}$ is drawn from a mixture of mutivariate multinomials:

$$x_{i,\cdot} \left| z_{iq} = 1 \sim \prod_{j=1}^{m} \mathscr{M}\left(1; \alpha_1^{j,q}, \alpha_2^{j,q}, \alpha_3^{j,q}\right)\right. .$$

where each of the 16 coordinates of the vector has its own set of parameters, so that each group of representatives is characterized by 32 parameters. By contrast, each issue (column) $x_{\cdot,j}$ was drawn from a mixture of mutivariate multinomials:

$$x_{\cdot,j} \left| z_{j\ell} = 1 \sim \prod_{i=1}^{n} \mathscr{M}\left(1; \alpha_1^{\ell}, \alpha_2^{\ell}, \alpha_3^{\ell}\right)\right. .$$

where the 435 coordinates share the same parameter α, so that each group of issues is characterized by 2 parameters. Using the *Expectation Maximisation* algorithm with several starting points and the *BIC* criterion, we obtain two groups of representatives as well as two groups of issues. The crossed classification induced by those groups is shown in Figure 5 (right panel).

[1] *Congressional Voting Records* data set is available from http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records.

 ISSN: 2102-6238

TABLE 1. *Repartition of Democrats and Republicans for each procedure: co-clustering (left) and simple classification (right).*

Co-clustering	Rep	Dem	Total
1	131	24	155
2	27	73	100
3	1	163	164
4	9	7	15

Representative clustering	Rep	Dem	Total
1	153	29	182
2	15	238	253

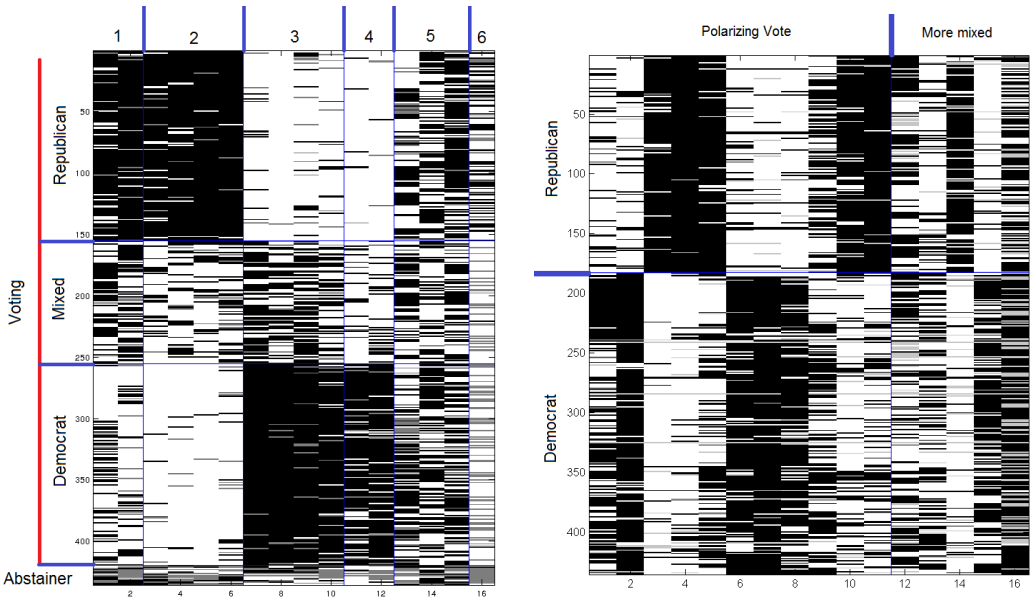

FIGURE 5. *Reordering of the voting matrix to highlight the groups: co-clustering (left) and simple classification (right). Black indicate a 'yes', white a 'no' and grey an abstention.*

The two groups of representatives found by the simple clustering approach mostly recover the Republican/Democrat split whereas co-clustering gives a more nuanced view of representatives, as shown in table 1, with an additional group of mixed voters and abstainers. Similarly, the two groups of issues recover polarizing and more consensual issues whereas co-clustering recovers issues supported by republicans, republicans and mixed voters, democrats, democrats and mixed voters, both parties and a last group with no support and a lot of abstention.

This example highlights the advantage of co-clustering over margin-wise clustering: LBM does tremendously better at capturing the underlying structure of the data and leads to groups that are both easy to interpret and go further in the description of the data. This relies on the ability of the LBM to capture and adapt to heterogeneity in both directions simultaneously whereas representatives and issues clustering can only account for heterogeneity in one dimension at the time.

 ISSN: 2102-6238

6. Discussion and perspectives

Latent Block Model provides a convenient probabilistic framework to model and analyze relational data between entities of different natures. The data are encoded in a table and the entities in the margins, rows and columns, of that table. The framework assumes a latent structure on both margins that induces a crossed partition of the table itself. The probabilistic nature of this framework suggests natural estimates of both the generating parameters θ of the model but also, and perhaps even more importantly, the latent structure $(\mathbf{z}, \mathbf{w})$ itself. The natural estimates are not tractable as such, due to the high dimension dependency structure induced by the latent structure on $f(.|\mathbf{x};\theta)$ but many approximation schemes, that effectively rely on tractable approximations of the likelihood, are available and constitute the cornerstone of iterative estimation procedures. The quality of those approximations is in general hard to assess but LBM constitute a peculiar and optimal playing field for variational methods as the approximation is asymptotically exact. As such, estimates of the latent structure and the parameters are asymptotically consistent, provided a good starting point for the iterative procedure is available. Model selection criteria to choose the number of groups for each margin compliment estimation procedures nicely to form a comprehensive package of results. The prowess of variational methods come at the cost of significant computational complexity. The complexity, of order $O(mnQLS_{EM}S_{FP})$, depend on the size of the model Q, L, the size of the data set m, n and the amount of allowed iterative steps for the inner (S_{FP}) and outer (S_{EM}) loops. Variational procedure are therefore limited to graphs with at most a few thousands elements in each margin. Note also that although variational procedures arise naturally in the LBM framework, non probabilistic methods are valid and can be rigorously analyzed to assess their performance (classification rate, complexity) on LBM generated data. Practical performance evaluation of an algorithm require artificial data sets covering favorable to adverse settings. Generating those data sets so that they meet a given "complexity" condition is itself a complex issue and currently done by trial and error.

This review focused on standard binary LBM but most of the results can be extended at no or little additional cost in several directions, some of them briefly discussed here. Sparse LBM intuitively mixes up 0 with missing values. It considers that only 1 (or more generally non null) values are informative and that their density in the table goes to 0 when $\min(n,m) \to +\infty$. Sparse LBM behaves qualitatively as standard LBM as long as the density is large compared to $\log(n)/m$ and $\log(m)/n$. Another extension considers growing number of groups, *i.e.* Q and L growing with n and m, although the asymptotic framework is rarely explicitly specified. LBM can be extended to valued tables by changing the density $\varphi(x;\pi)$. Identifiability conditions obviously depend on φ but the asymptotic results are retained for well-behaved φ, including most one-parameter exponential families.

Two other LBM extensions are noteworthy and call for further work. The first one concerns model selection and the crossed partition induced by the latent structure. The latent structure of the margins induces a crossed partition of the table elements into QL classes, each with its own parameter (or set of parameters) π_{ql}. Current model selection criteria are designed to select Q and L and do not penalize small differences between different π_{ql}, as seen in the structure of equation (11). Minimizing the number of different values taken by π_{ql} is a way to merge classes

 ISSN: 2102-6238

of the crossed partition and achieve a parsimonious description of π. It also allows LBM models to capture more efficiently block structure such as the one presented in Figure 2. The second one concerns extension of LBM to multidimensional arrays, as opposed to two-dimensional tables. This extension requires to leave behind the analogy behind LBM and bipartite graphs but allows the clustering of three-way (or more) tables on the basis of their margins. Examples of three-way tables are less frequent than their two-way counterparts but include viewers-movie-season table where the season can help distinguish between otherwise similar groups of viewers that consume certain genre throughout the year against groups that consume them seasonally, for example blockbusters in the summer and family movies around Christmas. We acknowledge that p dimensional arrays are quite uncommon and certainly require regularization of π, as outlined above, to ensure that the number of parameters to estimate remains manageable. We also argue that LBM, unlike SBM, are naturally suited to the analysis of multidimensional array.

References

Allman, E., Mattias, C., and Rhodes, J. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37:3099–3132.

Aubert, J., Ha, T., and MaryHuard, T. (2014). Modele à blocs latents pour l'analyse de données métagénomiques. In $46^{ème}$ *journées de Statistiques de la SFdS*.

Bennett, J. and Lanning, S. (2007). The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35.

Bickel, P. and Chen, A. (2009). A nonparametric view of network models and Newman-Girvan and other modularities. *PNAS*, 106(50):21068–21073.

Biernacki, C. and Jacques, J. (2012). Modele génératif pour données ordinales. In *44e Journées de Statistique, SFdS*, Bruxelles, Belgique.

Brault, V. (2014). *Estimation et sélection de modèles pour le modèle des blocs latents*. PhD thesis, Université Paris-Sud.

Brault, V. and Lomet, A. (2014). Revue bibliographique pour la classification croisée.

Celisse, A., Daudin, J.-J., and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899.

Channarond, A., Daudin, J.-J., and Robin, S. (2012). Classification and estimation in the stochastic blockmodel based on the empirical degrees. *Electronic Journal of Statistics*, 6:2574–2601.

Choi, D. S., Wolfe, P. J., and Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika*. in press.

Côme, E. and Latouche, P. (2013). Model selection and clustering in stochastic block models with the exact integrated complete data likelihood. *ArXiv e-prints*.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38. With discussion.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer.

Gazal, S., Daudin, J.-J., and Robin, S. (2011). Accuracy of variational estimates for random graph mixture models. *Journal of Statistical Computation and Simulation*, 0(0):1–14.

Govaert, G. (1983). *Classification croisée*. Thèse d'état, Université Pierre et Marie Curie.

Govaert, G. and Nadif, M. (2003). Clustering with block mixture models. *Pattern Recognition*, 36:463–473.

Govaert, G. and Nadif, M. (2008). Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics and Data Analysis*, 52:3233 – 3245.

Govaert, G. and Nadif, M. (2010). Latent block model for contingency table. *Communication in Statistics - Theory and Methods*, 39:416 – 425.

Govaert, G. and Nadif, M. (2013). *Co-Clustering*. ISTE Ltd and John Wiley & Sons, Inc.

Gunawardana, A. and Byrne, W. (2005). Convergence theorems for generalized alternating minimization procedures. *Journal of Machine Learning Research*, 6:2049–2073.

Gyllenberg, M., Koski, T., Reilink, E., and Verlann, M. (1994). Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, 31:542–548.

 ISSN: 2102-6238

Keribin, C., Brault, V., Celeux, G., and Govaert, G. (2012). Model selection for the binary latent block model. *Proceedings of COMPSTAT 2012*.

Keribin, C., Brault, V., Celeux, G., and Govaert, G. (2014). Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, pages 1–16.

Lomet, A. (2012). *Sélection de modèle pour la classification croisée de données continues.* PhD thesis, Université de Technologie de Compiègne.

Lomet, A., Govaert, G., and Grandvalet, Y. (2012). Design of artificial data tables for co-clustering analysis. Technical report, Université de Technologie de Compiègne, France.

Long, B., Zhang, Z. M., and Yu, P. S. (2005). Co-clustering by block value decomposition. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 635–640. ACM.

Mariadassou, M. and Matias, C. (2015). Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, 21:537–573.

Matias, C. and Robin, S. (2014). Modeling heterogeneity in random graphs: a selective review. *arXiv preprint arXiv:1402.4296*.

Meeds, E. and Roweis, S. (2007). Nonparametric Bayesian biclustering. Technical report, University of Toronto.

Raïffa, H. and Schlaifer, R. (1961). *Applied statistical decision theory*. Studies in managerial economics. Division of Research, Graduate School of Business Adminitration, Harvard University.

Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic block model. *Ann. Statist*, 39(4):1878–1915.

Seung, D. and Lee, L. (2001). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems 13*, pages 556–562.

Shan, H. and Banerjee, A. (2008). Bayesian co-clustering. In *Eighth IEEE International Conference on Data Mining, 2008. ICDM'08*, pages 530–539.

Tanay, A., Sharan, R., Kupiec, M., and Shamir, R. (2004). Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *PNAS*, 101(9):2981–2986.

Van Dijk, B., Van Rosmalen, J., and Paap, R. (2009). A Bayesian approach to two-mode clustering. Technical Report 2009-06, Econometric Institute.

Wyse, J. and Friel, N. (2012). Block clustering with collapsed latent block models. *Statistics and Computing*, 22:415–428.

Wyse, J., Friel, N., and Latouche, P. (2014). Inferring structure in bipartite networks using the latent block model and exact ICL. *ArXiv e-prints*.

Yoo, J. and Choi, S. (2010). Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on stiefel manifolds. *Information processing & management*, 46(5):559–570.

 ISSN: 2102-6238