Estimation and Selection for the Latent Block Model on Categorical Data

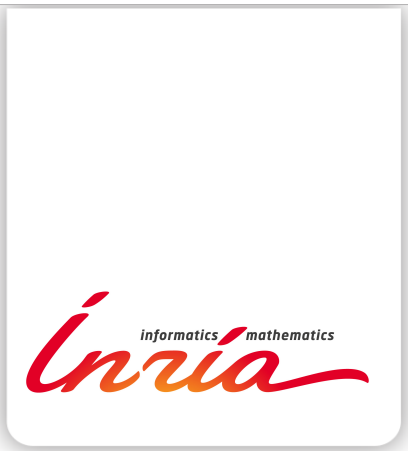
Christine Keribin, Vincent Brault, Gilles Celeux, Gérard Govaert

Estimation and Selection for the Latent Block Model on Categorical Data

Christine Keribin
, Vincent Brault , Gilles Celeux Gérard Govaert

RESEARCH REPORT
N° 8264
Mars 2013
Project-Team Select

ISSN 0249-6399 ISRN INRIA/RR--8264--FR+ENG



Estimation and Selection for the Latent Block Model on Categorical Data

Christine Keribin*†
, Vincent Brault * †, Gilles Celeux † Gérard Govaert‡

Project-Team Select

Research Report n° 8264 — Mars 2013 — 27 pages

Abstract: This paper is dealing with estimation and model selection in the Latent Block Model (LBM) for categorical data. First, after providing sufficient conditions ensuring the identifiability of this model, it generalises estimation procedures and model selection criteria derived for binary data. Secondly, it develops Bayesian inference through Gibbs sampling. And, with a well calibrated non informative prior distribution, Bayesian estimation is proved to avoid the traps encountered by the LBM with the maximum likelihood methodology. Then model selection criteria are presented. In particular an exact expression of the integrated completed likelihood (ICL) criterion requiring no asymptotic approximation is derived. Finally numerical experiments on both simulated and real data sets highlight the interest of the proposed estimation and model selection procedures.

Key-words: EM algorithm, Variational Approximation, Stochastic EM, Bayesian Inference, Gibbs Sampling, BIC Criterion, Integrated Completed Likelihood.

* Laboratoire de Mathématiques, Equipe Probabilités et Statistiques, UMR 8628,Bâtiment 425, Université Paris-Sud, F-91405 Orsay Cedex, France
† INRIA Saclay Île de France, Bâtiment 425, Université Paris-Sud, F-91405 Orsay Cedex, France
‡ U.T.C, U.M.R. 7253, C.N.R.S. Heudiasyc, Centre de Recherches de Royallieu, B.P. 20529, F-60205 Compiègne Cedex, France

RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

1 rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing
Campus de l'École Polytechnique
91120 Palaiseau

Estimation et sélection pour le modèle des blocs latents avec données catégorielles

Résumé : Cet article traite de l'estimation et de la sélection pour le modèle des blocs latents (LBM) avec données catégorielles. Nous commençons par donner des conditions suffisantes pour obtenir l'identifiabilité de ce modèle. Nous généralisons les procédures d'estimation et les critères de sélection obtenus dans le cadre binaire. Nous considérons l'inférence bayésienne à travers l'échantillonneur de Gibbs couplé avec une approche variationnelle : avec une distribution a priori non informative correctement calibrée, ces algorithmes évitent mieux les extrema locaux que la méthodologie fréquentiste. Nous présentons des critères de sélection de modèle et nous donnons une forme exacte non asymptotique pour le critère ICL. Les résultats obtenus sur des données simulées et réelles illustrent l'intérêt de notre procédure d'estimation et de sélection de modèle.

Mots-clés : Algorithme EM, approximation variationnelle, Stochastique EM, inférence bayésienne, échantillonneur de Gibbs, critère BIC, critère ICL.

1 Introduction

Block clustering methods are aiming to design in a same exercise a clustering of the rows and the columns of a large array of data. These methods could be expected to be useful to summarise large data sets by dramatically smaller data sets with the same structure. Since more and more huge data sets are available, more and more block clustering methods have been proposed. Many application fields as genomic (Jagalur et al., 2007) or recommendation system (Shan and Banerjee, 2008) are concerned with block clustering. In particular, block clustering methods have been developed to deal with binary data present in archaeology (Govaert, 1983) and sociology (Wyse and Friel, 2012). Madeira and Oliveira (2004) described an extensive list of block clustering methods. The checker board structure studied in this article has been considered from two point of views: determinist approaches (see for instance Banerjee et al. 2007; Govaert 1977) and model-based approaches. The model-based view has been considered through the maximum likelihood methodology (Govaert and Nadif, 2008) and through Bayesian inference (Meeds and Roweis, 2007; Wyse and Friel, 2012). Among them, the Latent Block Model (LBM) is attractive since it could lead to powerful representations. For instance, as illustrated in Govaert and Nadif (2008), summarising binary tables with grey summaries derived from the conditional probabilities of belonging to a block could be quite realistic and suggestive. Moreover, this model provides a probabilistic framework to choose a relevant block clustering.

The LBM is as follows. A population of n observations described with d categorical variables of the same nature with r levels is available. Saying that the categorical variables are of the same nature means that it is possible to code them in a same (and natural) way. This assumption is needed to ensure that decomposing the data set in a block structure is making sense. Let $\mathbf{y} = (y_{ij}, i = 1, \ldots, n; j = 1, \ldots, d)$ be the data matrix where $y_{ij} = h$, $1 \leq h \leq r$, r being the number of levels of the categorical variables.

It is assumed that there exists a partition into g row clusters $\mathbf{z} = (z_{ik}; i = 1, \ldots, n; k = 1, \ldots, g)$ and a partition into m column clusters $\mathbf{w} = (w_{j\ell}; j = 1, \ldots, d; \ell = 1, \ldots, m)$. The z_{ik}s (resp. $w_{j\ell}$s) are binary indicators of row i (resp. column j) belonging to row cluster k (resp. column cluster ℓ), such that the random variables y_{ij} are conditionally independent knowing $\mathbf{z}$ and $\mathbf{w}$ with parameterised density $\varphi(y_{ij}; \alpha_{k\ell})^{z_{ik}w_{j\ell}}$. Thus, the conditional density of $\mathbf{y}$ knowing $\mathbf{z}$ and $\mathbf{w}$ is

$$f(\mathbf{y}|\mathbf{z}, \mathbf{w}; \boldsymbol{\alpha}) = \prod_{i,j,k,\ell} \varphi(y_{ij}; \alpha_{k\ell})^{z_{ik}w_{j\ell}}$$

where $\boldsymbol{\alpha} = (\alpha_{k\ell}; k = 1, \ldots, g; \ell = 1, \ldots, m)$. Moreover, it is assumed that the row and column labels are independent: $p(\mathbf{z}, \mathbf{w}) = p(\mathbf{z})p(\mathbf{w})$ with $p(\mathbf{z}) = \prod_{ik} \pi_k^{z_{ik}}$ and $p(\mathbf{w}) = \prod_{j\ell} \rho_\ell^{w_{j\ell}}$, where $(\pi_k = P(z_{ik} = 1), k = 1, \ldots, g)$ and $(\rho_\ell = P(w_{j\ell} = 1), \ell = 1, \ldots, m)$ are the mixing proportions. Hence, the marginal density of $\mathbf{y}$ is a mixture density

$$f(\mathbf{y}; \boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w}) \in \mathcal{Z} \times \mathcal{W}} p(\mathbf{z}; \boldsymbol{\pi}) p(\mathbf{w}; \boldsymbol{\rho}) f(\mathbf{y}|\mathbf{z}, \mathbf{w}; \boldsymbol{\alpha}),$$

$\mathcal{Z}$ and $\mathcal{W}$ denoting the sets of possible row labels $\mathbf{z}$ and column labels $\mathbf{w}$, and $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$, with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_m)$. The density of $\mathbf{y}$ can be written as

$$f(\mathbf{y}; \boldsymbol{\theta}) = \sum_{\mathbf{z},\mathbf{w}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi(y_{ij}; \alpha_{k\ell})^{z_{ik}w_{j\ell}}. \quad (1)$$

The LBM involves a double missing data structure, namely $\mathbf{z}$ and $\mathbf{w}$, which makes statistical inference more difficult than for standard mixture model. Govaert and Nadif (2008) show that the

EM algorithm is not tractable and proposed a variational approximation of the EM algorithm (VEM) to derive the maximum likelihood (ML) estimate of $\boldsymbol{\theta}$. This VEM algorithm could give satisfactory estimates despite it is highly dependent of its initial values and has a marked tendency to provide empty clusters. Moreover, even if the parameter $\boldsymbol{\theta}$ is properly estimated, and for even moderate data matrices, it is unreasonable to compute the likelihood. As a consequence, computing penalised model selection criteria such as AIC or BIC is challenging as well. Moreover, these difficulties to compute relevant model selection criteria are increased by the fact that the LBM statistical units could be defined in several different ways.

The aim of this paper is to overcome all those limitations. First, we propose algorithms aiming to avoid the above mentioned drawbacks of the VEM algorithm. Secondly, we show how it is possible to compute properly relevant model selection criteria. This is presented for the LBM for categorical data, a natural extension of the LBM for binary data, but not developed as so far. The article is organised as follows. In Section 2, the LBM is detailed for categorical data and sufficient conditions ensuring the identifiability of this model are provided. Section 3 is devoted to the presentation of a stochastic algorithm aiming to avoid the variational approximation, and it shows how Bayesian inference can be helpful to avoid empty clusters. Section 4 is concerned with model selection criteria. We first show how the integrated completed likelihood (ICL) of the LBM is closed form and derive from ICL an approximation of BIC. Section 5 is devoted to numerical experiments on both simulated and real data sets to illustrate the behaviour of the proposed estimation algorithms and model selection criteria. A discussion section proposing some perspectives for future work ends this paper.

Notation To simplify the notation, the sums and products relative to rows, columns, row clusters, column clusters and levels will be subscripted respectively by the letters i, j, k, ℓ, h without indicating the limits of variation that will be implicit as in (1). So, for instance, the sum $\sum_i$ stands for $\sum_{i=1}^{n}$, $\sum_h$ stands for $\sum_{h=1}^{r}$, and $\prod_{i,j,k,\ell}$ stands for $\prod_{i=1}^{n}\prod_{j=1}^{d}\prod_{k=1}^{g}\prod_{\ell=1}^{m}$.

2 Model definition and identifiability

The LBM has already been defined on binary (Govaert and Nadif, 2008; Keribin et al., 2012), Gaussian (Lomet (2012)) and count data (Govaert and Nadif (2010)). We consider here the LBM for categorical data, where the conditional distribution $\varphi(y_{ij};\alpha_{kl})$ of the outcome y_{ij} knowing the labels z_{ik} and $w_{j\ell}$ is a multinomial distribution $\mathcal{M}(1;\alpha_{k\ell})$, with parameter $\alpha_{k\ell} = (\alpha_{k\ell}^h)_{h=1,\dots,r}$, where $\alpha_{k\ell}^h \in (0;1)$ and $\sum_h \alpha_{k\ell}^h = 1$. Using the binary indicator vector $\mathbf{y}_{ij} = (y_{ij}^1,\dots,y_{ij}^r)$, such that for every $h = 1,\dots,r$, $y_{ij}^h = 1$ when $y_{ij} = h$ and $y_{ij}^h = 0$ otherwise, the density per block is

$$\varphi(\mathbf{y}_{ij};\alpha_{k\ell}) = \prod_h (\alpha_{k\ell}^h)^{y_{ij}^h}$$

and the mixture density is

$$f(\mathbf{y};\boldsymbol{\theta}) = \sum_{\mathbf{z},\mathbf{w}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell,h} (\alpha_{k\ell}^h)^{z_{ik}w_{j\ell}y_{ij}^h}.$$

The parameter to be estimated is $\theta = (\boldsymbol{\pi},\ \boldsymbol{\rho},\ \boldsymbol{\alpha})$, with $g + m + (r-1)gm - 2$ independent components.

Before considering the estimation problem, it is important to analyse the model identifiability (Frühwirth-Schnatter 2006, pp. 21-23). Obviously, LBM, as a mixture model, is not identifiable due to invariance to relabelling the blocks, but it is of no importance when concerned with

maximum likelihood estimation. It will be necessary to go back to this issue when concerned with Bayesian estimation. Unfortunately, it is also well-known that simple multivariate Bernoulli mixtures are not identifiable (Gyllenberg et al., 1994), regardless of this invariance to relabelling. Allman et al. (2009) set a sufficient condition to their identifiability that cannot be easily extended to LBM. We define here a set of sufficient conditions ensuring the identifiability of the Bernoulli LBM:

Theorem 1 (Binary LBM identifiability) *: Consider the binary LBM. Let $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ be the row and column mixing proportions and $A = (\alpha_{kl})$ the $g \times m$ matrix of Bernoulli parameters. Define the following conditions:*

- *C_1: for all $1 \leq k \leq g$, $\pi_k > 0$ and the coordinates of vector $\boldsymbol{\tau} = A\boldsymbol{\rho}$ are distinct.*
- *C_2: for all $1 \leq \ell \leq m$, $\rho_\ell > 0$ and the coordinates of vector $\boldsymbol{\sigma} = \boldsymbol{\pi}'A$ are distinct (where $\boldsymbol{\pi}'$ is the transpose of $\boldsymbol{\pi}$).*

If conditions C_1 and C_2 hold, then the binary LBM is identifiable for $n \geq 2m-1$ and $d \geq 2g-1$.

Proof. The proof of this theorem is given in Appendix A.

Conditions C_1 and C_2 are not strongly restrictive since the set of vectors $\boldsymbol{\tau}$ and $\boldsymbol{\sigma}$ violating them is of Lebesgue measure 0. Therefore, Theorem 1 asserts the generic identifiability of LBM, which is a "practical" identifiability, explaining why it works in the applications (Carreira-Perpiñàn and Renals, 2000). These conditions could appear to be somewhat unnatural. However:

(i) It is not surprising that the number of row labels g (resp. column labels m) is constrained by the number of columns d (resp. rows n). In case of a simple finite mixture of g different Bernoulli products with d components, more clusters you define in the mixture, more components for the multivariate Bernoulli you need in order to ensure the identifiability: see also the following condition $d > 2\lceil \log_2 g \rceil + 1$ of Allman et al. (2009) for simple Bernoulli mixtures, where $\lceil . \rceil$ is the ceil function.

(ii) Conditions C_1 and C_2 are extensions of a condition set to ensure the identifiability of the Stochastic Block Model (Celisse et al., 2012), where $n = d$ and $\mathbf{z} = \mathbf{w}$.

It follows from conditions C_1 (resp. C_2) that the probabilities $\tau_k = P(y_{ij} = 1 | z_{ik} = 1)$ (resp. $\sigma_\ell = P(y_{ij} = 1 | w_{j\ell} = 1)$) to observe an event in a cell of a row of row class k (resp. in a cell of a column of column class ℓ) can be sorted in a strictly ascending order. Hence, contrary to what happens in the Gaussian mixture context, these conditions can be used to put a natural order on the mixture labels in a Bayesian setting. Notice that conditions C_1 and C_2 are sufficient for the identifiability, and can be proved to be not necessary for $g = 2$ and $m = 2$.

Theorem 1 is easily extended to the categorical case, where the conditions are defined on vectors $\boldsymbol{\tau}^h = A^h\boldsymbol{\rho}$ and $\boldsymbol{\sigma}^h = \boldsymbol{\pi}'A^h$, with $h = 1, \ldots, r-1$ and $A^h = (\alpha^h_{kl})_{k=1,\ldots,g;\ell=1,\ldots,m}$.

3 Model estimation

With g and m fixed, the likelihood of the model parameter $L(\boldsymbol{\theta}) = f(\mathbf{y}; \boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ is written

$$L(\boldsymbol{\theta}) = \sum_{\mathbf{z},\mathbf{w}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell,h} (\alpha^h_{k\ell})^{z_{ik} w_{j\ell} y^h_{ij}}.$$

Even for small tables, computing this likelihood (or its logarithm) is difficult. For instance, with a data matrix 20×20 with $g = 2$ and $m = 2$, it requires the calculation of $g^n \times m^d \approx 10^{12}$ terms. In the same manner, deriving the maximum likelihood estimator with the EM algorithm

is challenging. As a matter of fact, the E step requires the computation of the joint conditional distributions of the missing labels $P(z_{ik}w_{j\ell} = 1|\mathbf{y}; \boldsymbol{\theta}^{(c)})$ for $i = 1, \ldots, n, j = 1, \ldots, d, k = 1, \ldots, g$, $\ell = 1, \ldots, m$ with $\boldsymbol{\theta}^{(c)}$ being a current value of the parameter. Thus, the E step involves computing too many terms that cannot be factorised as for a standard mixture, due to the dependence of the row and column labels knowing the observations.

To tackle this problem, Govaert and Nadif (2008) proposed to use a variational approximation of the EM algorithm by imposing that the conditional joint distribution of the labels knowing the observations factorizes as $q^{(c)}_{zw}(\mathbf{z}, \mathbf{w}) = q^{(c)}_z(\mathbf{z})q^{(c)}_w(\mathbf{w})$. This VEM algorithm has been proved to provide relevant estimates of the LBM model in different contexts with continuous or binary data matrices (Govaert and Nadif, 2008), and is directly extended to categorical data, see Appendix B. However, it presents several drawbacks illustrated in Section 5:

(i) as most variational approximation algorithms, the VEM appears to be quite sensitive to starting values,

(ii) it has a marked tendency to provide solutions with empty clusters, i.e. with fewer clusters than required after a maximum a posteriori (MAP) classification rule,

(iii) it can only compute a lower bound of the maximum likelihood called free energy.

A possible way to attenuate the dependence of VEM to its initial values is to use stochastic versions of EM which are not stopping at the first encountered fixed point of EM (see McLachlan and Krishnan, 2008, chap. 6). The basic idea of these stochastic EM algorithms is to incorporate a stochastic step between the E and M steps where the missing data are simulated according to their conditional distribution knowing the observed data and a current estimate of the model parameters. For the LBM, it is not possible to simulate in a single exercise the missing labels $\mathbf{z}$ and $\mathbf{w}$ and a Gibbs sampling scheme is required to simulate the couple $(\mathbf{z}, \mathbf{w})$. The SEM-Gibbs algorithm for binary data (Keribin et al., 2012) is a simple adaptation to the LBM of the standard SEM algorithm of Celeux and Diebolt (1985). It is directly extended for the categorical LBM, see Appendix C.

While VEM is based on the variational approximation of the LBM, SEM-Gibbs uses no approximation, but runs a Gibbs sampler to simulate the unknown labels with their conditional distribution knowing the observations and a current estimation of the parameters. Hence, SEM-Gibbs is not increasing the loglikelihood at each iteration, but it generates an irreducible Markov chain with a unique stationary distribution which is expected to be concentrated around the ML parameter estimate. Thus a natural estimate of $\boldsymbol{\theta}$ derived from SEM-Gibbs is the mean $\bar{\boldsymbol{\theta}}$ of $(\boldsymbol{\theta}^{(c)}; c = B+1, \ldots, B+C)$ get after a burn-in period of length B. Moreover, as every stochastic algorithm, SEM-Gibbs is theoretically subject to label switching (see Frühwirth-Schnatter, 2006, Section 3.5.5). However this possible drawback of SEM-Gibbs does not occur in most practical situations. Numerical experiments presented in Keribin et al. (2012) for binary data show that SEM-Gibbs is by far less sensitive to starting values than VEM. Those results lead them to advocate initializing the VEM algorithm with the SEM-Gibbs mean parameter estimate $\bar{\boldsymbol{\theta}}$ to get a good approximation of the ML estimate for the latent block model.

If SEM-Gibbs can be expected to be insensitive to its initial values, there is no reason to think that it can be useful to avoid solutions with empty clusters. Bayesian inference in statistics can be regarded as a well-ground tool to regularize ML estimate in a poorly posed setting. In the LBM setting, Bayesian inference could be thought of as useful to avoid empty cluster solutions and thus to attenuate the "empty cluster" problem. In particular, for the categorical LBM, it is possible to consider proper and independent non informative prior distributions for the mixing

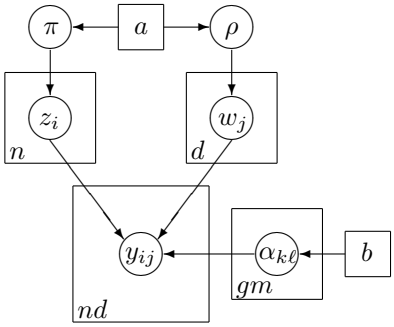


Figure 1: A graphical representation of the Bayesian latent block model.

proportions $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$, and for parameter $\boldsymbol{\alpha}$ as a product of $g \times m$ non informative priors on each multinomial parameter $\alpha_{k\ell} = (\alpha_{k\ell}^h)_{h=1,\dots,r}$ (see Figure 1):

$$\boldsymbol{\pi} \sim \mathcal{D}(a, \dots, a), \;\; \boldsymbol{\rho} \sim \mathcal{D}(a, \dots, a), \; \alpha_{k\ell} \sim \mathcal{D}(b, \dots, b),$$

$\mathcal{D}(v, \dots, v)$ denoting a Dirichlet distribution with parameter v. Note that Meeds and Roweis (2007) proposed a more general prior (Pitman-Yor prior).

Using Bayesian inference in a regularisation perspective, the model parameter can be estimated by maximising the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$, and it leads to the so-called MAP (Maximum A Posteriori) estimate

$$\widehat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{y}).$$

The Bayes formula

$$\log p(\boldsymbol{\theta}|\mathbf{y}) = \log p(\mathbf{y}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\mathbf{y})$$

allows to straightforwardly define an EM algorithm for the computation of the MAP estimate:

- the E Step relies on the computation of the conditional expectation of the complete log-likelihood $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(c)})$ as for the ML estimator:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(c)}) = \mathbb{E}\big[\log p(\mathbf{y}, \mathbf{z}, \mathbf{w}|\boldsymbol{\theta})|\mathbf{y}, \boldsymbol{\theta}^{(c)}\big],$$

- the M Step differs in that the objective function for the maximisation process is equal to the $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(c)})$ function augmented by the logarithm of the prior density

$$\boldsymbol{\theta}^{(c+1)} = \arg\max_{\boldsymbol{\theta}} \big(Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(c)}) + \log p(\boldsymbol{\theta})\big).$$

This M Step forces an increase in the log posterior function $p(\boldsymbol{\theta}|\mathbf{y})$ (McLachlan and Krishnan, 2008, chap. 6, p. 231). For LBM, the Bayesian approach combined with the variational approximation leads to the V-Bayes algorithm, with the following M step:

M step (V-Bayes algorithm: estimation of the posterior mode)

$$\pi_k^{(c+1)} = \frac{a - 1 + s_{.k}^{(c+1)}}{n + g(a-1)}, \quad \rho_\ell^{(c+1)} = \frac{a - 1 + t_{.\ell}^{(c+1)}}{d + m(a-1)}$$

$$\alpha_{k\ell}^{h\ (c+1)} = \frac{b - 1 + \sum_{i,j} s_{ik}^{(c+1)} t_{j\ell}^{(c+1)} y_{ij}^{h}}{r(b-1) + s_{.k}^{(c+1)} t_{.\ell}^{(c+1)}},$$

where $s_{ik}^{(c+1)}$ and $t_{j\ell}^{(c+1)}$ are the current conditional probabilities of the labels and $s_{.k}^{(c+1)}$ and $t_{.\ell}^{(c+1)}$, their sum over the lines and columns respectively, as defined in the VEM algorithm, see Appendix B. The hyperparameters a and b are acting as regularisation parameters. In this perspective, the choice for a and b is important. It appears clearly from the updating equations of this M step that V-Bayes collapses to the VEM algorithm with the uniform prior $a = b = 1$ and involves no regularisation. And, for Jeffreys prior ($a = b = 1/2$), V-Bayes has a tendency to provide more empty cluster solutions than VEM. Frühwirth-Schnatter (2011) shows the great influence of a for Bayesian inference in the Gaussian mixture context. Based on an asymptotic analysis by Rousseau and Mengersen (2011) and on a thorough finite sample analysis, she advocated to take $a = 4$ for moderate dimension ($g < 8$) and $a = 16$ for larger dimensions to avoid empty clusters.

In Section 5, we present numerical experiments highlighting the ability of the V-Bayes algorithm to avoid the "empty cluster" cases encountered with the VEM algorithm.

However, as VEM, a V-Bayes algorithm could be expected to be highly dependent of its initial values. Thus it could be of interest to initialise V-Bayes with the solution derived from the Gibbs sampler. Obviously, since Dirichlet prior distributions are conjugate priors for the multinomial distribution, full conditional posterior distributions of the LBM parameters are closed form and Gibbs sampling is easy to implement, in such way:

Gibbs sampling for the categorical LBM

1. simulation of $\mathbf{z}^{(c+1)}$ according to $p(\mathbf{z}|\mathbf{y}, \mathbf{w}^{(c)}; \boldsymbol{\theta}^{(c)})$ as in SEM-Gibbs
2. simulation of $\mathbf{w}^{(c+1)}$ according to $p(\mathbf{w}|\mathbf{y}, \mathbf{z}^{(c+1)}; \boldsymbol{\theta}^{(c)})$ as in SEM-Gibbs
3. simulation of $\boldsymbol{\pi}^{(c+1)}$ according to

   $$\mathcal{D}(a + z_{.1}^{(c+1)}, \ldots, a + z_{.g}^{(c+1)})$$

   where $z_{.k} = \sum_i z_{ik}$ denotes the number of lines in line cluster k
4. simulation of $\boldsymbol{\rho}^{(c+1)}$ according to

   $$\mathcal{D}(a + w_{.1}^{(c+1)}, \ldots, a + w_{.m}^{(c+1)})$$

   where $w_{.\ell} = \sum_j w_{j\ell}$ denotes the number of columns in column cluster ℓ
5. simulation of $\boldsymbol{\alpha}_{k\ell}^{(c+1)}$ according to

   $$\mathcal{D}(b + N_{k\ell}^{1\ (c+1)}, \ldots, b + N_{k\ell}^{r\ (c+1)})$$

   for $k = 1, \ldots, g; \ell = 1, \ldots, m$ and with

   $$N_{k\ell}^{h\ (c+1)} = \sum_{i,j} z_{ik}^{(c+1)} w_{j\ell}^{(c+1)} y_{ij}^{h}. \tag{2}$$

As Gibbs sampling explores the whole distribution, it is subject to label switching. This problem can be sensibly solved for the categorical LBM by using the identifiability conditions of Theorem 1: these conditions define a natural order of row and column labels except on a set of parameters of measure zero. Hence, row and column labels are post processed separately after the last iteration, by reordering them according to the ascending values of $\boldsymbol{\tau}^h = \boldsymbol{\alpha}^h \boldsymbol{\rho}$ and $\boldsymbol{\sigma}^h = \boldsymbol{\pi}' \boldsymbol{\alpha}^h$ coordinates respectively, and for a given h.

4 Model selection

Choosing relevant numbers of clusters in a latent block model is obviously of crucial importance. This model selection problem is difficult for several reasons. First there is a couple (g, m) of number of clusters to be selected instead of a single number. Secondly penalised likelihood criteria such as AIC or BIC are not directly available since computing the maximised likelihood is not feasible. Third determining the number of statistical units of a LBM could be questionable (number of rows, number of columns, number of cells, ...).

Fortunately, it is possible to compute the exact integrated completed loglikelihood (ICL) of the categorical LBM.

4.1 The Integrated Completed Loglikelihood (ICL)

ICL is the logarithm of the integrated completed likelihood

$$p(\mathbf{y}, \mathbf{z}, \mathbf{w} \mid g, m) = \int_{\Theta_{g,m}} p(\mathbf{y}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\theta}_{g,m}) p(\boldsymbol{\theta}_{g,m}) d\boldsymbol{\theta}_{g,m}$$

where the missing data $(\mathbf{z}, \mathbf{w})$ have to be replaced by some values closely related to the model at hand. By taking into account the missing data, ICL is focusing on the clustering view of the model. For this very reason, ICL could be expected to select a stable model allowing to partitioning the data with the greatest evidence (Biernacki et al., 2000).

As stated in Section 3, Dirichlet proper non informative priors are available for the categorical LBM and ICL can be computed without requiring asymptotic approximations. Using the conditional independence of the $\mathbf{z}$s and the $\mathbf{w}$s conditionally to $\boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}$, the integrated completed likelihood can be written

$$\text{ICL} = \log \mathrm{p}(\mathbf{z}) + \log \mathrm{p}(\mathbf{w}) + \log \mathrm{p}(\mathbf{y}|\mathbf{z}, \mathbf{w}). \tag{3}$$

Using now the conjugate properties of the prior Dirichlet distributions and the conditional independence of the y_{ij} knowing the latent vectors $\mathbf{z}$ and $\mathbf{w}$ and the LBM parameters, we get (see Appendix D for a detailed proof)

$$\begin{aligned}
\text{ICL}(g, m) \quad = \quad & \log\Gamma(ga) + \log\Gamma(ma) - (m+g)\log\Gamma(a) \\
& + mg(\log\Gamma(rb) - r\log\Gamma(b)) \\
& - \log\Gamma(n+ga) - \log\Gamma(d+ma) \\
& + \sum_k \log\Gamma(z_{\cdot k} + a) + \sum_\ell \log\Gamma(w_{\cdot \ell} + a) \\
& + \sum_{k,\ell} \Bigg[\Big(\sum_h \log\Gamma(N_{k\ell}^h + b) \Big) \\
& - \log\Gamma(z_{\cdot k} w_{\cdot \ell} + rb) \Bigg].
\end{aligned}$$

where $z_{\cdot k}$, $w_{\cdot \ell}$ and $N_{k\ell}^h$ are defined as for the Gibbs sampler (see Equation 2). In practice, the missing labels $\mathbf{z}, \mathbf{w}$ have to be chosen. Following Biernacki et al. (2000), they are replaced by

$$(\hat{\mathbf{z}}, \hat{\mathbf{w}}) = \arg\max_{(\mathbf{z},\mathbf{w})} p(\mathbf{z}, \mathbf{w}|\mathbf{y}; \hat{\boldsymbol{\theta}}),$$

$\hat{\boldsymbol{\theta}}$ being the estimate of the LBM parameter computed from the V-Bayes algorithm initialised by the Gibbs sampler as described in Section 3. The maximising partition $(\hat{\mathbf{z}}, \hat{\mathbf{w}})$ is obtained with a MAP rule after the last V-Bayes step.

4.2 Penalised information criterion

BIC is an information criterion defined as an asymptotic approximation of the logarithm of the integrated likelihood (Schwarz, 1978). The standard case leads to write BIC as a penalised maximum likelihood:

$$\mathrm{BIC} = \max_{\theta} \log(p(\mathbf{y};\theta)) - \frac{D}{2}\log(n),$$

where n is the number of statistical units and D the number of free parameters. Unfortunately, this approximation cannot be used for LBM, due to the dependency structure of the observations $\mathbf{y}$.

However, a heuristic can be stated to define BIC. BIC-like approximations of each term of ICL (Equation 3) lead to the following approximation as n and d tend to infinity:

$$\begin{aligned}\mathrm{ICL}(g,m) &\simeq \max_{\boldsymbol{\theta}} \log p(\mathbf{y},\widehat{\mathbf{z}},\widehat{\mathbf{w}};\boldsymbol{\theta}) \\ &\quad -\frac{g-1}{2}\log n - \frac{m-1}{2}\log d \\ &\quad -\frac{gm(r-1)}{2}\log(nd). \end{aligned} \tag{4}$$

Now, following Biernacki et al. (2000), the parameter maximizing the completed likelihood $\log p(\mathbf{y},\widehat{\mathbf{z}},\widehat{\mathbf{w}};\boldsymbol{\theta})$ is replaced by the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$:

$$\begin{aligned}\max_{\boldsymbol{\theta}} \log p(\mathbf{y},\widehat{\mathbf{z}},\widehat{\mathbf{w}};\boldsymbol{\theta}) &\simeq \log p(\mathbf{y},\widehat{\mathbf{z}},\widehat{\mathbf{w}};\widehat{\boldsymbol{\theta}}) \\ &\simeq \log p(\widehat{\mathbf{z}},\widehat{\mathbf{w}}|\mathbf{y};\widehat{\boldsymbol{\theta}}) + \log p(\mathbf{y};\widehat{\boldsymbol{\theta}}).\end{aligned}$$

Therefore ICL is the sum of the logarithm of the conditional distribution $\log p(\widehat{\mathbf{z}},\widehat{\mathbf{w}}|\mathbf{y};\widehat{\boldsymbol{\theta}})$, measuring the assignment confidence, and a penalised likelihood. The decomposition of ICL

$$\mathrm{ICL}(g,m) = \log p(\widehat{\mathbf{z}},\widehat{\mathbf{w}}|\mathbf{y};\widehat{\boldsymbol{\theta}}) + \mathrm{BIC}(g,m) \tag{5}$$

as a sum of an entropy term and BIC is classical (McLachlan and Peel (2000), 6.10.3) and, by analogy, we conjecture from (4) and (5) the following form for BIC after a straightforward factorisation of the penalty:

$$\begin{aligned}\mathrm{BIC}(g,m) &= \log p(\mathbf{y};\widehat{\boldsymbol{\theta}}) \\ &\quad -\frac{gm(r-1)+g-1}{2}\log n \\ &\quad -\frac{gm(r-1)+m-1}{2}\log d. \end{aligned} \tag{6}$$

Notice that the maximum likelihood is not available and is replaced by a lower bound computed by variational approximation. This expression of $\mathrm{BIC}(g,m)$ highlights two penalty terms, one for the columns and one for the rows: the statistical units can be seen to be the rows and columns respectively. The number of row (resp. column) free parameters takes into account the proportions $\boldsymbol{\pi}$ (resp. $\boldsymbol{\rho}$) and the parameter $\boldsymbol{\alpha}$ of the categorical conditional distribution.

BIC is known to be a consistent estimator of the order of mixture models with bounded log likelihoods (Keribin, 2000), whereas ICL is not (Baudry, 2009). However ICL can also be expected to be consistent for estimating the LBM order due to a specific behaviour of the conditional distribution of the labels $p(\mathbf{z},\mathbf{w}|\mathbf{y};\boldsymbol{\theta})$. Indeed, Mariadassou and Matias (2013) proved

that for a categorical LBM, this distribution concentrates on the actual configuration with large probability as soon as $\boldsymbol{\theta}$ is such that $\boldsymbol{\alpha}$ is not too far from the true value. As a consequence, relying on the MAP rule and with an estimator $\widehat{\boldsymbol{\theta}}$ such that $\widehat{\boldsymbol{\alpha}}$ is converging to the true parameter value, $\log p(\widehat{\mathbf{z}}, \widehat{\mathbf{w}}|\mathbf{y}; \widehat{\boldsymbol{\theta}})$ vanishes for true g and m. Moreover, this non positive term may add to BIC a supplementary penalty for incorrect g or m and hence leads ICL to be also consistent to estimate the order. Notice that this conjecture needs to assume the consistency of the maximum likelihood and the variational estimators for LBM. These results have not been proved yet for LBM as far as we know, and still remain an interesting challenge. But this conjecture is supported by the numerical experiments of the next section where ICL outperforms BIC for low n and d, and has the same behaviour as BIC for greater values.

5 Numerical experiments

Relevant numerical experiments on both simulated and real data sets supporting the claims of this paper are now presented. First we analyse the ability of the Bayesian inference through Gibbs sampling to avoid the tendency of the maximum likelihood methodology through VEM or SEM-Gibbs algorithms to provide empty clusters. After a simple illustration, Monte Carlo experiments on simulated data are performed (i) to compare the behaviour of SEM-Gibbs+VEM and Gibbs+V-Bayes algorithms (ii) to analyse the influence of the hyperparameters a and b (iii) to enlighten the behaviour of model selection criteria. Then, the LBM is experimented on a real categorical data set already used by Wyse and Friel (2012).

5.1 Escaping from an empty cluster solution with V-Bayes

This illustration is done with an easily separated data matrix produced by Lomet et al. (2012) and Lomet (2012) for the binary LBM model[1], with 50 rows, 50 columns, $(g, m) = (3, 3)$ clusters and equal mixing proportions. The left part of Figure 2 displays the three components of $\widehat{\boldsymbol{\pi}}$ along the iterations of the VEM algorithm while the right part of Figure 2 displays them for the V-Bayes algorithm with $(a, b) = (4, 1)$. Both algorithms started from the same position. VEM is rapidly trapped into an empty cluster solution. But V-Bayes, for which the estimated proportions are not smaller than $\frac{a-1}{g(a-1)}$, escapes from this empty cluster solution and finally provides a satisfactory solution with the real number of row clusters $g = 3$.

5.2 Experiments with simulated data

Monte Carlo experiments were performed to assess the ability of Bayesian inference to avoid empty cluster solutions. A binary LBM with $(g, m) = (5, 4)$ clusters was considered with following Bernoulli parameters

$$\boldsymbol{\alpha} = \begin{pmatrix} \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ 1-\varepsilon & \varepsilon & \varepsilon & \varepsilon \\ 1-\varepsilon & 1-\varepsilon & \varepsilon & \varepsilon \\ 1-\varepsilon & 1-\varepsilon & 1-\varepsilon & \varepsilon \\ 1-\varepsilon & 1-\varepsilon & 1-\varepsilon & 1-\varepsilon \end{pmatrix}$$

where ε allows to produce easily $(+)$, moderately $(++)$ or hardly $(+++)$ separated mixtures, see Govaert and Nadif (2008), and with unequal $\boldsymbol{\rho} = (0.1\ 0.2\ 0.3\ 0.4)$ and $\boldsymbol{\pi} = (0.1\ 0.15\ 0.2\ 0.25\ 0.3)$

[1]The data set is available from https://www.hds.utc.fr/coclustering/doku.php.

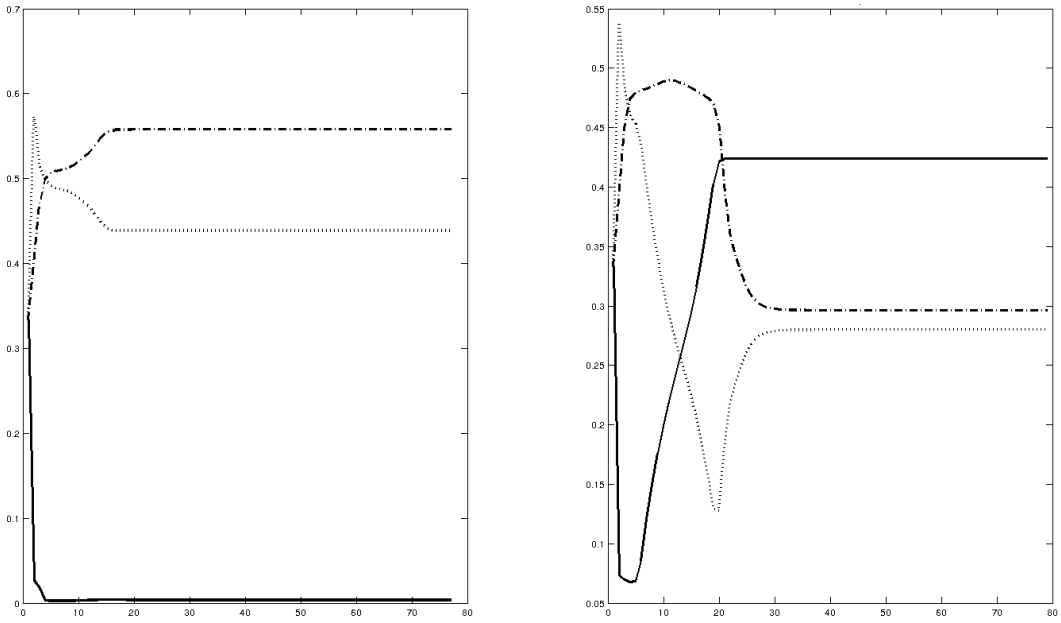


Figure 2: Evolution of $\hat{\boldsymbol{\pi}}$ for the VEM algorithm (on left) and the V-Bayes algorithm (on right)

or equal mixing proportions. As results for equal and unequal mixing proportions were quite analogous, they are not both shown in the following (see Keribin et al. (2013) for a complete display).

Comparing SEM-Gibbs+VEM and Gibbs+V-Bayes to avoid empty cluster solutions. For each case of separation, and each sample size $(n, d) = (100, 200)$, $(150, 150)$, $(200, 100)$, 500 data matrices were simulated. Both algorithms were initialized with two different numbers of clusters:

(1) the true number of clusters: $(g, m) = (5, 4)$,

(2) more than the true number of clusters: $(g, m) = (8, 8)$.

The hyperparameters a and b for Bayesian inference were chosen in $\{1, 4, 16\}^2 \backslash \{(1, 1)\}$. The results are summarised in Tables 1-2 for unequal mixing proportions. Case $(a, b) = (1, 1)$ corresponds to the SEM-Gibbs+VEM algorithm and can be compared with Gibbs+V-Bayes algorithm on higher a and b values. It clearly appears that Bayesian inference with $a = 4$ or 16 and $b = 1$ is doing a good job to avoid empty cluster solutions. As expected, taking $a > 1$ is quite relevant in that purpose. On the contrary, it appears clearly that taking $b > 1$ is harmful. As noticed by a reviewer, one possible reason for this could be that taking $b > 1$ means that the prior weight in the Dirichlet is more focused on the centre of the simplex i.e. all category weights are equal. In this case, $b > 1$ puts most prior weight on equal proportions for success and failure in the Bernoulli distribution. Such heavy prior weight could penalize against separation of the block clusters i.e. this puts more prior weight on saying there is no difference in the block probabilities (i.e. all α_{kl} close to 0.5), and so there could be more of a tendency for empty clusters and hence overall homogeneity in the solution.

	+			++			+++		
	b / a	1	4	16	1	4	16	1	4	16
(100,200)	1	5.4	48	83.2	7	23.2	73.4	16.8	18.8	74.8
	4	0.8	2.6	30.4	1.6	1.8	0.8	2.2	2.2	0.8
	16	1	0.8	9.4	0.6	1	1.2	1	1.2	2.6
(150,150)	1	7.6	29.2	71.8	13	11.4	62.6	17.6	8.4	65
	4	1	0	6.8	0.6	0.6	0.6	0.6	1.2	0.6
	16	0.2	0.6	0.2	0.2	1.4	0.8	0.4	0.8	0
(200,100)	1	13.2	20	71.8	11.8	18.4	67.6	18.2	7.8	69.6
	4	1	0.2	12	0.4	0.8	0.4	0.8	1	0.6
	16	0.6	0.4	0.6	0.4	0.8	0.2	0	0.4	0

Table 1: Percentage of empty cluster solutions, for data simulated with unequal proportion models, obtained by SEM-Gibbs+VEM (cell $a = 1, b = 1$) and Gibbs+V-Bayes algorithms, called with the true (g, m), for 500 matrices in a easily separated $+$, moderately separated $++$ and hardly separated $+++$ cases, for three sample sizes.

	+			++			+++		
	b / a	1	4	16	1	4	16	1	4	16
(100,200)	1	65	96.8	100	54.6	82.6	100	47.4	84.2	100
	4	1.2	5.8	100	1.8	3.6	98.2	3	10	97.6
	16	2.2	3.8	99	2.4	4.4	25.2	5.2	13.2	31
(150,150)	1	69.2	89	100	54.2	64.6	100	57.2	69.4	100
	4	0.4	1.2	100	0.2	1.4	91.6	2.2	5.8	86.8
	16	2.6	2.2	34	1.4	2.2	6	3	7.4	31.8
(200,100)	1	81.4	96	100	67.8	76.8	100	68.2	73	100
	4	1.4	2.6	100	1.2	2.8	99	2.2	5.8	94.2
	16	2.2	4.6	99.2	2.4	3	23	3.8	9	27.8

Table 2: Percentage of empty cluster solutions, for data simulated with unequal proportion models, obtained by SEM-Gibbs+VEM (cell $a = 1, b = 1$) and Gibbs+V-Bayes algorithms, called with $(g, m) = (8, 8)$, for 500 matrices in the easily separated $+$, moderately separated $++$ and hardly separated $+++$ cases, for three sample sizes.

Analysing the empty cluster solutions of SEM-Gibbs+ VEM. The fact that the SEM-Gibbs+VEM algorithm produces empty clusters can be thought of as beneficial: it could mean that the number of clusters has been oversized and the algorithm provides a more relevant number of clusters. To answer to this question, SEM-Gibbs+VEM was initialised with $(g, m) = (8, 8)$ and run on 500 data sets simulated under the same conditions than previously described. Table 3 reports the number of cluster finally obtained in the hardly separated cases with $n = 150$, $d = 150$ and unequal mixing proportions. Results for other cases are available in Keribin et al. (2013). In all situations, the right number of clusters is chosen less than two percent of times.

Moreover, the SEM-Gibbs+VEM algorithm produces empty clusters even in case of initial

g \ m	3	4	5	6	7	8
4	0	2	0	0	0	0
5	1	**4**	4	3	2	0
6	0	7	13	15	6	8
7	0	3	12	26	23	21
8	0	3	13	35	85	214

Table 3: Number of provided clusters, for data simulated with the unequal proportion models, obtained by SEM-Gibbs+VEM for 500 matrices in the hardly separated $+++$ case with $n = 150$ and $d = 150$, starting with $g = 8$ and $m = 8$.

numbers of clusters less than the real numbers. This algorithm was started with $(g, m) = (8, 8)$ on 50 simulated binary matrix of sizes $n = 50$ and $d = 50$ of hardly separated $(g, m) = (10, 10)$ clusters. Although the numbers of requested clusters are smaller than the right values, the algorithm produces empty cluster solutions almost twenty percent of times (Table 4).

g \ m	6	7	8
7	1	2	18
8	5	64	410

Table 4: Number of provided clusters by SEM-Gibbs+VEM for 500 random initialisations in the separated case for $g = 8$ and $m = 8$ whereas the right repartition is $g = 10$ and $m = 10$.

Analysing the behaviour of the model selection criteria. For each g and m varying from two to eight, the Gibbs+V-Bayes algorithm was run on 50 simulated data sets with number of clusters $(g, m) = (5, 4)$ and BIC and ICL criteria were computed. Tables 5 and 6 display how many times each (g, m) was selected in the equal mixture proportion case. The ICL criterion seems to provide a better estimation of the number of clusters than BIC, especially for the smaller sizes of the data matrix. Otherwise, both criteria tend to choose less clusters than there actually are in the cases where the true number of clusters is not correctly identified. This tendency appears also on numerical experiments not reported here but available in Keribin et al. (2013) for mixtures with the same α parameters, but unequal mixing proportions. It deserves the following comments.

First, it is worth noting that the tendency of ICL to choose fewer components than the true number of components of a mixture is well-known. The purpose of ICL is to assess the number of mixture components that leads to the best clustering and it can happen that the number of clusters is smaller than the number of mixture components. This is because a cluster may be better represented by a mixture of components than by a single component (see Biernacki et al. (2000) or Baudry et al. (2010)). For the LBM as mentioned in Section 4.2, when computing BIC, the maximum likelihood is approximated by the maximum free energy. The tendency of BIC to underestimate the true number of components could indicate that the difference between the maximum likelihood and the maximum free energy increases with the number of mixture components. As expected, for large sizes, the performances of both criteria improve and the difference between ICL and BIC decreases. This is in agreement with the conjecture presented

in Section 4.2.

	+				++					+++				
		4	5	6		2	3	4	5		2	3	4	5
(100,200)	3				3					3	7			
	4				4		12			4	3	33		
	5	[34]	4		5		8	[23]	2	5		7	[0]	
	6	11			6			4		6				
	7	1			7			1		7				
		4	5	6		2	3	4	5		2	3	4	5
(150,150)	3				3					3	11			
	4				4	1	18			4		32	1	
	5	[35]	3	1	5		2	[29]		5		1	[5]	
	6	11			6					6			1	
		4	5	6		2	3	4	5		2	3	4	5
(200,100)	3				3	1	1			3	5	2		
	4				4		16	5		4		36	4	
	5	[41]	4	1	5		1	[25]	1	5			[2]	1
	6	4			6					6				
		4	5	6		2	3	4	5		2	3	4	5
(500,1000)	4				4					4		10		
	5	[38]			5			[44]		5		2	[36]	
	6	11			6			5		6			1	
	7	1			7			1		7			1	
		4	5	6		2	3	4	5		2	3	4	5
(750,750)	4				4					4			8	
	5	[45]			5			[44]	3	5			[40]	
	6	5			6			3		6			2	
		4	5	6		2	3	4	5		2	3	4	5
(1000,500)	4				4		1			4		5		
	5	[45]	4		5			[43]	6	5			[43]	2
	6	1			6					6				

Table 5: The frequency of the models selected by the ICL criterion in the easily separated +, moderately separated ++ and ill separated +++ cases with equal proportion for different matrix sizes.

5.3 Congressional voting in US senate

The combination *Gibbs+V-Bayes* was tested on *UCI Congressional Voting Records* data set[2] recording the votes ('yes', 'no', 'abstained or absent') of 435 members of the 98th congress on 16 different key issues. This data set involves homogeneous 3-level categorical data.

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

	+				++					+++				
		3	4	5		2	3	4	5		2	3	4	5
(100,200)	3				3					3	12			
	4				4	1	30	1		4	4	34		
	5	2	[34]	5	5		12	[6]		5			[0]	
	6		9		6					6				
		3	4	5		2	3	4	5		2	3	4	5
(150,150)	3				3	1				3	22	2		
	4	1			4	2	42			4		25		
	5		[35]	2	5		1	[4]		5			[1]	
	6		12		6					6				
		3	4	5		2	3	4	5		2	3	4	5
(200,100)	3				3	3	1			3	13	5		
	4	4			4		36	2		4		32		
	5		[41]		5			[8]		5			[0]	
	6		5		6					6				
		3	4	5		2	3	4	5		2	3	4	5
(500,1000)	4				4					4		1		
	5		[38]	1	5			[45]		5		1	[45]	
	6		11		6			4		6			2	
	7				7			1		7			1	
		3	4	5		2	3	4	5		2	3	4	5
(750,750)	4				4					4		8		
	5		[45]		5			[44]		5			[39]	
	6		5		6			6		6			3	
		3	4	5		2	3	4	5		2	3	4	5
(1000,500)	4				4					4		4		
	5		[45]	2	5			[43]	1	5			[43]	1
	6		3		6			6		6			2	

Table 6: The frequency of the models selected by the BIC criterion in the easily separated +, moderately separated ++ and ill separated +++ cases with equal proportion for different matrix sizes.

The Gibbs+V-Bayes algorithm was run on the 3-level data with $(a, b) = (4, 1)$ and the ICL and BIC criteria were computed to select the number of clusters.
ICL selected $(g, m) = (5, 7)$ clusters, while BIC selected $(g, m) = (4, 6)$. The rows and columns reorganisation derived from the block clustering for each criterion is displayed in Figure 3 where 'yes' is coloured in black, 'no' in white and 'abstained or absent' in grey.

Row cluster composition by Party is shown in Table 7. Roughly speaking, the first three clusters selected with both criteria are the same. Row cluster 1 is mainly Republican while row cluster 3 is mainly Democrat. Row cluster 2 is mixed. Cluster 4 selected with BIC is split into two clusters (4 and 5) in the model selected with ICL (the six bigger abstainers are isolated in cluster 5).

ICL	Rep	Dem	Total
1	132	22	154
2	25	68	93
3	1	162	163
4	2	4	6
5	8	11	19

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

Table 7: Repartition of Democrats and Republicans for each criterion.

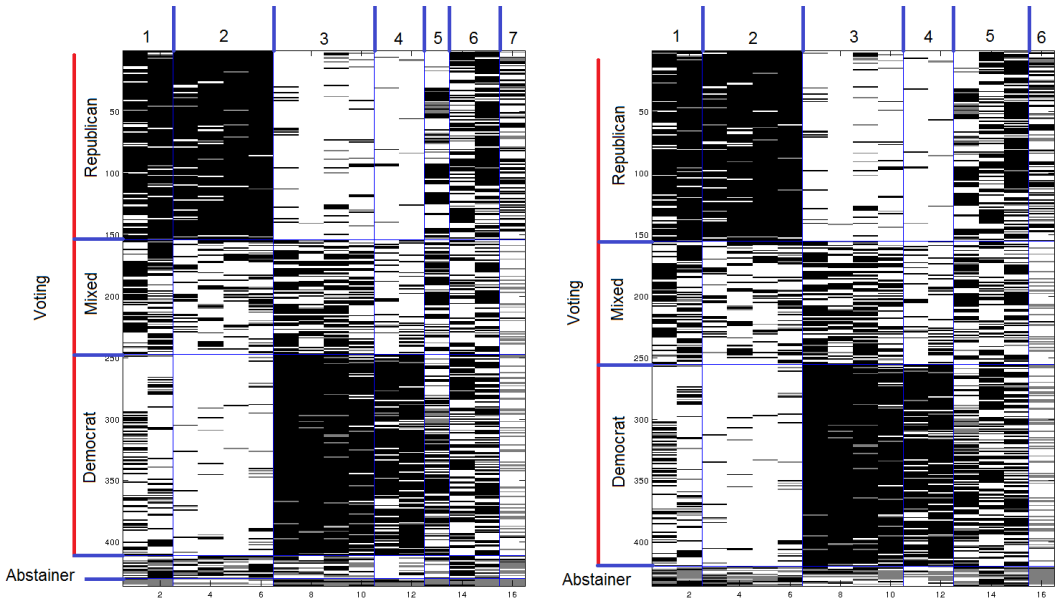


Figure 3: Reorganisation of the data by the ICL criterion (on left) and by the BIC criterion (on right).

The first four column clusters selected with both criteria are identical. Column clusters 1 and 2 contain the issues on which Republicans voted 'yes' and Democrats 'no'. These two column clusters only differ for the vote of congress members belonging to row cluster 2. In contrast, column clusters 3 and 4 contain the issues on which Republicans voted 'no' and Democrats 'yes'. Moreover, both criteria isolated in a specific cluster the issue characterised by a high level of abstainers. Column cluster 5 selected with BIC is split into two column clusters (5 and 6) in the model selected with ICL, due to one issue more subject to abstention.

This data set was already studied by Wyse and Friel (2012) where the 'no' and 'abstained or absent' were aggregated in order to work on binary data. Assuming that g and m were random, Wyse and Friel used a collapsed sampler which marginalises over the model parameters with uniform prior (i.e. $(a, b) = (1, 1)$ in Figure 1) aiming to compute the distribution of the most visited models and the maximum a posteriori cluster membership $(\mathbf{z}, \mathbf{w})$. This maximisation is

analogous to maximising the ICL criterion because :

$$\begin{aligned} \log p(\mathbf{z}, \mathbf{w}, g, m|\mathbf{y}) &= \text{ICL}(\mathbf{z}, \mathbf{w}, g, m) \\ &\quad + \log p(g, m) - \log p(\mathbf{y}). \end{aligned} \tag{7}$$

Since the prior on g and m was taken to be uniform, the two algorithms should give the same optimal quadruplet $(\mathbf{z}, \mathbf{w}, g, m)$. Wyse and Friel gave the a posteriori distribution for (g, m) but did not provide the maximum a posteriori cluster membership $(\mathbf{z}, \mathbf{w})$. We computed it by running their collapsed sampler with 110 000 iterations, 10 000 of them as burn-in iterations[3]. This led to $(g, m) = (6, 11)$ clusters (which is not in their 60% confidence level) with a corresponding ICL $= -3565$. Replicating 30 times gave variations from 10 to 14 column clusters, with 6 row clusters each time, and corresponding ICLs were computed. The best ICL value was -3554 for $(g, m) = (6, 13)$. Running Gibbs+V-Bayes on the same binary data set and the same priors selected $(g, m) = (5, 13)$ clusters with corresponding ICL$= -3553$. Notice that ICL(g, m) is quite flat around its maximum. The results are similar, requiring six times more iterations for the collapsed sampler.

6 Discussion

In this paper, we show that Bayesian inference through Gibbs sampling is beneficial to produce regularised estimates of the LBM for categorical data. A sufficient solution to ensure the identifiability of the LBM for categorical data has been provided. Contrary to the VEM or V-Bayes algorithms, the SEM-Gibbs algorithm and Gibbs sampling provide solutions not highly dependent of the starting values. Restricting attention to point estimation, Bayesian inference through Gibbs sampling from non informative priors produces regularised parameter estimation. The solution providing the posterior mode of the parameters is not suffering the label switching problem and can be used as initial values of the V-Bayes algorithm, and thus avoid the empty cluster solutions which jeopardizes maximum likelihood inference for the LBM. Moreover, taking profit of the exhibited sufficient condition ensuring the identifiability of the LBM for categorical data, it is possible to define a natural order to post process the rows and columns after Gibbs sampling in order to deal with the label switching problem in a proper way. On the other hand, we select a latent block model using the integrated completed likelihood ICL which can be computed without requiring asymptotic approximations. A good perspective for future work would be to prove the conjecture stated in Section 4.2: namely ICL criterion provides a consistent estimation of the number of clusters (g, m) for the LBM, contrary to the standard mixture model for which ICL is not consistent. Otherwise, an alternative is to use the collapsed sampler of Wyse and Friel (2012). Maximising the posterior collapsed is closely related to maximising ICL for flat (g, m) priors (see equation 7). In our experiments, ICL appears to be efficient to find the $(\mathbf{z}, \mathbf{w}, g, m)$ maximising the posterior $\log p(\mathbf{z}, \mathbf{w}, g, m|\mathbf{y})$ and less computationally demanding than the collapsed sampler. But, the latter is able to explore the uncertainty in the joint support of all models and corresponding missing labels. A perspective would be to analyse the variability of ICL with a random sampling of $(\mathbf{z}, \mathbf{w})$ according to the distribution $p(\mathbf{z}, \mathbf{w}; \widehat{\theta})$. Preliminary numerical experiments on simulated data show that the variability of ICL depends on (g, m) and is clearly more important for g or m greater than the real values. It could be an interesting subject for further research.

[3]The code of Wyse and Friel is available at http://www.ucd.ie/statdept/jwyse.

Acknowledgements The authors thank the reviewers and the Associate Editor for their very helpful comments which have greatly improved this paper. C. K. has been supported by LMH (Hadamard Mathematics Labex), backed by the foundation FMJH (Fondation Mathématique Jacques Hadamard).

Appendices

A Proof of Theorem 1

The identifiability of a parametric model requires that for any two different values $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$ in the parameter space, the corresponding probability distribution $\mathbb{P}_{\boldsymbol{\theta}}$ and $\mathbb{P}_{\boldsymbol{\theta}'}$ are different. We prove that, under assumptions of Theorem 1, there exists a unique parameter $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$, up to a permutation of row and column labels, corresponding to $\mathbb{P}(\mathbf{y})$ the probability distribution function of matrix $\mathbf{y}$ having at least $2m-1$ rows and $2g-1$ columns. The proof is adapted from Celisse et al. (2012) who set up a similar result for the Stochastic Block Model. Notice first that $\boldsymbol{\tau} = A\boldsymbol{\rho}$ is the vector of the probability τ_k to have 1 in a cell of a row of given row class k:

$$\tau_k = \mathbb{P}(y_{ij} = 1 | z_{ik} = 1) = \sum_{\ell} \rho_\ell \alpha_{k\ell} = (A\boldsymbol{\rho})_k.$$

As all the coordinates of $\boldsymbol{\tau}$ are distinct, R defined as the $g \times g$ matrix of coefficients $R_{ik} = (\tau_k)^i$, for $0 \leq i < g$ and $1 \leq k \leq g$, is Vandermonde, and hence invertible. Consider now u_i, the probability to have 1 on the i first cells of the first row of $\mathbf{y}$:

$$\begin{aligned}
u_i &= \mathbb{P}(y_{11} = 1, \ldots, y_{1i} = 1) \\
&= \sum_{k,\ell_1,\ldots,\ell_i} \mathbb{P}(z_{1k} = 1) \times \\
&\quad \prod_{\substack{j=i \\ j=1}} \big(\mathbb{P}(y_{1j} | z_{1k} = 1, w_{j\ell_j} = 1)\mathbb{P}(w_{j\ell_j} = 1)\big) \\
&= \sum_{k} \mathbb{P}(z_{1k} = 1) \times \\
&\quad \sum_{\ell_1} \mathbb{P}(y_{11} | z_{1k} = 1, w_{1\ell_1} = 1)\mathbb{P}(w_{1\ell_1} = 1) \times \\
&\quad \ldots \times \sum_{\ell_i} \mathbb{P}(y_{1i} | z_{1k} = 1, w_{i\ell_i} = 1)\mathbb{P}(w_{i\ell_i} = 1) \\
&= \mathbb{P}(y_{11} = 1, \ldots, y_{1i} = 1) = \sum_{k} \pi_k (\tau_k)^i
\end{aligned}$$

With a given $\mathbb{P}(\mathbf{y})$, $u_1, \ldots, u_{2g-1}$ are known, and we denote $u_0 = 1$. Let now M be the $(g+1) \times g$ matrix defined by $M_{ij} = u_{i+j-2}$, for all $1 \leq i \leq g+1$ and $1 \leq j \leq g$ and define M_i as the square matrix obtained by removing the row i from M. The coefficients of M are

$$M_{ij} = u_{i+j-2} = \sum_{1 \leq k \leq g} \tau_k^{i-1} \pi_k \tau_k^{j-1}.$$

We can write, with $A_\pi = \mathrm{Diag}(\boldsymbol{\pi})$

$$M_g = R A_\pi R'.$$

Now, R, unknown at this stage, can be retrieved by noticing that the coefficients of $\boldsymbol{\tau}$ are the roots of the following polynomial (Celisse et al., 2012) of degree g

$$B(x) = \sum_{k=0}^{g} (-1)^{k+g} D_k x^k,$$

where $D_k = \det M_k$ and $D_g \neq 0$ as M_g is a product of invertible matrices. Hence, it is possible to determine $\boldsymbol{\tau}$, and R is now known. Consequently, $\boldsymbol{\pi}$ is defined in a unique manner by $A_\pi = R^{-1} M_g R^{'-1}$.
In the same way, $\boldsymbol{\rho}$ is defined in a unique manner by considering the probabilities σ_ℓ to have 1 in a column of column class ℓ and the probabilities v_j to have 1 on the j first cells of the first column of $\mathbf{y}$

$$\sigma_\ell = \mathbb{P}(y_{ij} = 1 | w_{j\ell} = 1) \text{ and } v_j = \mathbb{P}(y_{11} = 1, \ldots, y_{j1} = 1)$$

To determine $\boldsymbol{\alpha}$, we introduce for $1 \leq i \leq g$ and $1 \leq j \leq m$ the probabilities U_{ij} to have 1 in each i first cells of the first row and in each j first cells of the first column

$$\begin{aligned}
U_{ij} &= \mathbb{P}(y_{11} = 1, \ldots, y_{1i} = 1, y_{21} = 1, \ldots, y_{j1} = 1) \\
&= \sum_{\substack{k,\ell_1,\ell_2,\ldots,\ell_i, \\ k_1,\ldots,k_j}} \mathbb{P}(y_{11} | z_{1k} = 1, w_{1\ell} = 1) \times \\
&\quad \mathbb{P}(z_{1k} = 1)\mathbb{P}(w_{1\ell} = 1) \times \\
&\quad \prod_{\lambda=2}^{i} \mathbb{P}(y_{1\lambda} | z_{1k} = 1, w_{\lambda\ell_\lambda} = 1)\mathbb{P}(w_{\lambda\ell_\lambda} = 1) \times \\
&\quad \prod_{\eta=2}^{j} \mathbb{P}(y_{\eta 1} | z_{\eta k_\eta} = 1, w_{1\ell} = 1)\mathbb{P}(z_{\eta k_\eta} = 1) \\
&= \sum_{k,\ell} \mathbb{P}(y_{11} | z_{1k} = 1, w_{1\ell} = 1)\mathbb{P}(z_{1k} = 1) \times \\
&\quad \mathbb{P}(w_{1\ell} = 1)\mathbb{P}(y_{1i} | z_{1k} = 1)^{i-1}\mathbb{P}(y_{j1} | w_{1\ell} = 1)^{j-1} \\
&= \sum_{k,\ell} \pi_k \tau_k^{i-1} \alpha_{k\ell} \rho_\ell \sigma_\ell^{j-1}.
\end{aligned}$$

These probabilities are known and we can write, with $S_{j\ell} = (\sigma_\ell)^{j-1}$, $j = 1, \ldots m$, $\ell = 1, \ldots, m$

$$U = R A_\pi A A_\rho S',$$

and it leads to define $A = A_\pi^{-1} R^{-1} U S^{'-1} A_\rho^{-1}$, and hence $\boldsymbol{\alpha}$, in a unique manner.

The proof is straightforwardly extended to the categorical case. The identification of $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ are obtained by considering the probabilities of $y_{ij} = 1$ where 1 is the first outcome of the multinomial distribution. Then, each $A^h = (\alpha^h_{k\ell})_{k=1,\ldots,g;\ell=1,\ldots,m}$ is successively identified by considering $y_{ij} = \ell$, $i = 1, \ldots, m$, $j = 1, \ldots, g$.

B VEM algorithm for categorical data

Govaert and Nadif (2008) proposed to use, for the binary case, a variational approximation of the EM algorithm by imposing that the joint distribution of the labels takes the form $q^{(c)}_{zw}(\mathbf{z}, \mathbf{w}) = q^{(c)}_z(\mathbf{z}) q^{(c)}_w(\mathbf{w})$. To get simpler formulas, we will denote

$$s^{(c)}_{ik} = q^{(c)}_z(z_{ik} = 1)$$

and

$$t^{(c)}_{j\ell} = q^{(c)}_w(w_{j\ell} = 1).$$

Using the variational approximation, the maximisation of the loglikelihood is replaced by the maximisation of the free energy

$$\mathcal{F}(q_{zw}, \boldsymbol{\theta}) = \mathbb{E}_{q_z q_w}\left[\log \frac{p(\mathbf{y}, \mathbf{z}, \mathbf{w}; \boldsymbol{\theta})}{q_z(\mathbf{z}) q_w(\mathbf{w})}\right]$$

alternatively in q_z, q_w and $\boldsymbol{\theta}$ (see Keribin, 2010). The difference between the maximum loglikelihood and the maximum free energy is the Kullback divergence $KL(q_{zw}||p(\mathbf{z}, \mathbf{w}|\mathbf{y}; \boldsymbol{\theta})) = \mathbb{E}_{q_{zw}}\left[\log \frac{q_{zw}(\mathbf{z},\mathbf{w})}{p(\mathbf{z},\mathbf{w}|\mathbf{y};\boldsymbol{\theta})}\right]$. This can be extended to the categorical LBM, and leads to the following Variational EM (VEM) algorithm:

E step It consists of maximising the free energy in q_z and q_w, and it leads to:

1. Computing $s_{ik}^{(c+1)}$ with fixed $w_{jl}^{(c)}$ and $\boldsymbol{\theta}^{(c)}$

$$s_{ik}^{(c+1)} = \frac{\pi_k^{(c)} \psi_k(\mathbf{y}_{i\cdot}; \boldsymbol{\alpha}_{k\cdot}^{(c)})}{\sum_{k'} \pi_{k'}^{(c)} \psi_{k'}^{(c)}(\mathbf{y}_{i\cdot}; \boldsymbol{\alpha}_{k'\cdot}^{(c)})}, k = 1, \ldots, g$$

where $\mathbf{y}_{i\cdot}$ denotes the row i of the matrix $\mathbf{y}$, $\boldsymbol{\alpha}_{k\cdot} = (\alpha_{k1}, \ldots, \alpha_{km})$, and

$$\psi_k(\mathbf{y}_{i\cdot}; \boldsymbol{\alpha}_{k\cdot}^{(c)}) = \prod_{\ell,h} (\alpha_{k\ell}^{h\ (c)})^{\sum_j t_{j\ell}^{(c)} y_{ij}^h}$$

2. Computing $t_{jl}^{(c+1)}$ with fixed $s_{ik}^{(c+1)}$ and $\boldsymbol{\theta}^{(c)}$

$$t_{j\ell}^{(c+1)} = \frac{\rho_\ell^{(c)} \phi_\ell(\mathbf{y}_{\cdot j}; \boldsymbol{\alpha}_{\cdot\ell}^{(c)})}{\sum_{\ell'} \rho_{\ell'}^{(c)} \phi_{\ell'}(\mathbf{y}_{\cdot j}; \boldsymbol{\alpha}_{\cdot\ell'}^{(c)})}, \ell = 1, \ldots, m$$

where $\mathbf{y}_{\cdot j}$ denotes the column j of the matrix $\mathbf{y}$, $\boldsymbol{\alpha}_{\cdot\ell} = (\alpha_{1\ell}, \ldots, \alpha_{g\ell})$ and

$$\phi_\ell(\mathbf{y}_{\cdot j}; \boldsymbol{\alpha}_{\cdot\ell}^{(c)}) = \prod_{k,h} (\alpha_{k\ell}^{h\ (c)})^{\sum_i s_{ik}^{(c+1)} y_{ij}^h}.$$

M step Updating $\boldsymbol{\theta}^{(c+1)}$. Denoting $s_{\cdot k}^{(c+1)} = \sum_i s_{ik}^{(c+1)}$, $t_{\cdot\ell}^{(c+1)} = \sum_j t_{j\ell}^{(c+1)}$, it leads to

$$\pi_k^{(c+1)} = \frac{s_{\cdot k}^{(c+1)}}{n},$$

$$\rho_\ell^{(c+1)} = \frac{t_{\cdot\ell}^{(c+1)}}{d},$$

$$\alpha_{k\ell}^{h\ (c+1)} = \frac{\sum_{i,j} s_{ik}^{(c+1)} t_{j\ell}^{(c+1)} y_{ij}^h}{s_{\cdot k}^{(c+1)} t_{\cdot\ell}^{(c+1)}}.$$

C SEM algorithm for categorical data

E and S step

1. computation of $p(\mathbf{z}|\mathbf{y}, \mathbf{w}^{(c)}; \boldsymbol{\theta}^{(c)})$, then simulation of $\mathbf{z}^{(c+1)}$ according to $p(\mathbf{z}|\mathbf{y}, \mathbf{w}^{(c)}; \boldsymbol{\theta}^{(c)})$:

$$p(z_i = k|\mathbf{y}_{i\cdot}, \mathbf{w}^{(c)}; \boldsymbol{\theta}^{(c)}) = \frac{\pi_k^{(c)} \psi_k(\mathbf{y}_{i\cdot}; \boldsymbol{\alpha}_{k\cdot}^{(c)})}{\sum_{k'} \pi_{k'}^{(c)} \psi_{k'}^{(c)}(\mathbf{y}_{i\cdot}; \boldsymbol{\alpha}_{k'\cdot}^{(c)})},$$

for $k = 1, \ldots, g$ with

$$\psi_k(\mathbf{y}_{i\cdot}; \boldsymbol{\alpha}_{k\cdot}) = \prod_{\ell,h} (\alpha_{k\ell}^{h\ (c)})^{\sum_j w_{j\ell}^{(c)} y_{ij}^h}$$

2. computation of $p(\mathbf{w}|\mathbf{y}, \mathbf{z}^{(c+1)}; \boldsymbol{\theta}^{(c)})$, then simulation of $\mathbf{w}^{(c+1)}$ according to $p(\mathbf{w}|\mathbf{y}, \mathbf{z}^{(c+1)}; \boldsymbol{\theta}^{(c)})$.

M step Denoting $z_{\cdot k} := \sum_i z_{ik}$ and $w_{\cdot \ell} := \sum_j w_{j\ell}$, it leads to

$$\pi_k^{(c+1)} = \frac{z_{\cdot k}^{(c+1)}}{n}, \rho_\ell^{(c+1)} = \frac{w_{\cdot \ell}^{(c+1)}}{d}$$

and

$$\alpha_{k\ell}^{h\ (c+1)} = \frac{\sum_{i,j} z_{ik}^{(c+1)} w_{j\ell}^{(c+1)} y_{ij}^h}{z_{\cdot k}^{(c+1)} w_{\cdot \ell}^{(c+1)}}.$$

Note that the formulae of VEM and SEM-Gibbs are essentially the same, except that the probabilities s_{ik} and $t_{j\ell}$ are replaced with binary indicator values z_{ik} and $w_{j\ell}$.

D Computing ICL

In this section, the exact ICL expression is derived for categorical data. Using the conditional independence of the $\mathbf{z}$s and the $\mathbf{w}$s conditionally to $\boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}$, the integrated completed likelihood can be written

$$\begin{aligned}
\mathrm{p}(\mathbf{y}, \mathbf{z}, \mathbf{w}) &= \int \mathrm{p}(\mathbf{y}, \mathbf{z}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}) \mathrm{p}(\boldsymbol{\alpha}) \mathrm{p}(\boldsymbol{\pi}) \mathrm{p}(\boldsymbol{\rho}) d\boldsymbol{\alpha} d\boldsymbol{\pi} d\boldsymbol{\rho} \\
&= \int \mathrm{p}(\mathbf{y}|\mathbf{z}, \mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}) \mathrm{p}(\mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}) \times \\
&\quad \mathrm{p}(\boldsymbol{\alpha}) \mathrm{p}(\boldsymbol{\pi}) \mathrm{p}(\boldsymbol{\rho}) d\boldsymbol{\alpha} d\boldsymbol{\pi} d\boldsymbol{\rho} \\
&= \int \mathrm{p}(\mathbf{y}|\mathbf{z}, \mathbf{w}, \boldsymbol{\alpha}) \mathrm{p}(\boldsymbol{\alpha}) d\boldsymbol{\alpha} \times \\
&\quad \int \mathrm{p}(\mathbf{z}|\boldsymbol{\pi}) \mathrm{p}(\boldsymbol{\pi}) d\boldsymbol{\pi} \int \mathrm{p}(\mathbf{w}|\boldsymbol{\rho}) \mathrm{p}(\boldsymbol{\rho}) d\boldsymbol{\rho} \\
&= \mathrm{p}(\mathbf{z}) \mathrm{p}(\mathbf{w}) \mathrm{p}(\mathbf{y}|\mathbf{z}, \mathbf{w}).
\end{aligned}$$

Thus

$$\text{ICL} = \log \text{p}(\mathbf{z}) + \log \text{p}(\mathbf{w}) + \log \text{p}(\mathbf{y}|\mathbf{z}, \mathbf{w}).$$

Now, according to the LBM definition,

$$\text{p}(\mathbf{z}|\boldsymbol{\pi}) = \prod_{i,k} \pi_k^{z_{ik}}.$$

Since the prior distribution of $\boldsymbol{\pi}$ is the Dirichlet distribution $\mathcal{D}(a, \ldots, a)$

$$\text{p}(\boldsymbol{\pi}) = \frac{\Gamma(ga)}{\Gamma(a)^g} \prod_k \pi^{a-1},$$

we have

$$\text{p}(\boldsymbol{\pi}|\mathbf{z}) = \frac{\text{p}(\mathbf{z}|\boldsymbol{\pi})\text{p}(\boldsymbol{\pi})}{\int \text{p}(\mathbf{z}|\boldsymbol{\pi})\text{p}(\boldsymbol{\pi})} \propto \prod_k \pi_k^{z_{.k}+a-1}.$$

We recognise a non-normalised Dirichlet distribution $\mathcal{D}(z_{.1} + a, \ldots, z_{.g} + a)$ with the normalising factor

$$\frac{\Gamma(n + ga)}{\prod_k \Gamma(z_{.k} + a)}.$$

The expression of $\text{p}(\mathbf{z})$ directly follows from the Bayes theorem

$$\text{p}(\mathbf{z}) = \frac{\Gamma(ag)}{\Gamma(a)^g} \frac{\prod_k \Gamma(z_{.k} + a)}{\Gamma(n + ag)}. \tag{8}$$

In the same manner,

$$\text{p}(\mathbf{w}) = \frac{\Gamma(am)}{\Gamma(a)^m} \frac{\prod_\ell \Gamma(w_{.\ell} + a)}{\Gamma(d + am)}. \tag{9}$$

We now turn to the computation of $\text{p}(\mathbf{y}|\mathbf{z}, \mathbf{w})$. Since $\boldsymbol{\alpha}$ and $(\mathbf{z}, \mathbf{w})$ are independent, we have

$$\text{p}(\boldsymbol{\alpha}|\mathbf{y}, \mathbf{z}, \mathbf{w}) = \frac{\text{p}(\mathbf{y}|\mathbf{z}, \mathbf{w}, \boldsymbol{\alpha})\text{p}(\boldsymbol{\alpha})}{\int \text{p}(\mathbf{y}|\mathbf{z}, \mathbf{w}, \boldsymbol{\alpha})\text{p}(\boldsymbol{\alpha}) d\boldsymbol{\alpha}},$$

and, using the conditional independence of y_{ij} knowing $\mathbf{z}$, $\mathbf{w}$ and $\boldsymbol{\alpha}$,

$$\begin{aligned}
\text{p}(\mathbf{y}|\mathbf{z}, \mathbf{w}, \boldsymbol{\alpha}) &= \prod_{i,j,k,\ell} \left(\prod_h (\alpha_{k\ell}^h)^{y_{ij}^h} \right)^{z_{ik} w_{j\ell}} \\
&= \prod_{k,\ell} \left(\prod_h (\alpha_{k\ell}^h)^{\sum_{i,j} z_{ik} w_{j\ell} y_{ij}^h} \right) \\
&= \prod_{k,\ell} \left(\prod_h (\alpha_{k\ell}^h)^{N_{k\ell}^h} \right).
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathrm{p}(\boldsymbol{\alpha}|\mathbf{y},\mathbf{z},\mathbf{w}) &\propto \prod_{k,\ell}\left(\mathrm{p}(\alpha_{k\ell})\prod_h(\alpha_{k\ell}^h)^{N_{k\ell}^h}\right)\\
&\propto \prod_{k,\ell}\left[\left(\prod_h(\alpha_{k\ell}^h)^{b-1}\right)\left(\prod_h(\alpha_{k\ell}^h)^{N_{k\ell}^h}\right)\right]\\
&\propto \prod_{k,\ell}\left(\prod_h(\alpha_{k\ell}^h)^{b+N_{k\ell}^h-1}\right).
\end{aligned}$$

because $\mathrm{p}(\alpha_{k\ell})$ is the density of a Dirichlet distribution $\mathcal{D}(b,\ldots,b)$. Each term $k\ell$, is the density of a non-normalised Dirichlet distribution $\mathcal{D}(b+N_{k\ell}^1,\ldots,b+N_{k\ell}^r)$ with the normalising factor

$$\frac{\Gamma(z_{.k}w_{.\ell}+rb)}{\prod_h\Gamma(N_{k\ell}^h+b)}.$$

Thus, by the Bayes formula,

$$\mathrm{p}(\mathbf{y}|\mathbf{z},\mathbf{w})=\prod_{k,\ell}\frac{\Gamma(rb)}{\Gamma(b)^r}\frac{\prod_h\Gamma(N_{k\ell}^h+b)}{\Gamma(z_{.k}w_{.\ell}+rb)}. \tag{10}$$

And, the ICL criterion, presented in Section 4.1, is straightforwardly derived from equations (3), (8), (9) and (10).

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.

Banerjee, A., Dhillon, I., Ghosh, J., Merugu, S., and Modha, D. S. (2007). A generalized maximum entropy approach to bregman co-clustering and matrix approximation. *Journal of Maching Learning Research*, 8:1919–1986.

Baudry, J.-P. (2009). *Sélection de modèle pour la classification non supervisée. Choix du nombre de classes.* PhD thesis, Université Paris Sud.

Baudry, J.-P., Raftery, A., Celeux, G., Lo, K., and Gottardo, R. (2010). Combining Mixture Components for Clustering. *Journal of Computational and Graphical Statistics*, 19:332–353.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:719–725.

Biernacki, C., Celeux, G., and Govaert, G. (2010). Exact and monte carlo calculations of integrated likelihoods for the latent class model. *Journal of Statistical Planning and Inference*, 140:2991–3002.

Carreira-Perpiñàn, M. and Renals, S. (2000). Practical identifiability of finite mixtures of multivariate bernoulli distributions. *Neural Computation*, 12:141–152.

Celeux, G. and Diebolt, J. (1985). Stochastic versions of the em algorithm. *Computational Statistics Quaterly*, 2:73–82.

Celisse, A., Daudin, J.-J., and Latouche, P. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Statist.*, 6:1847–1899.

Daudin, J.-J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, 18:173–183.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer series in statistics. Springer.

Frühwirth-Schnatter, S. (2011). *Mixtures : estimation and applications*, chapter Dealing with label switching under model uncertainty, pages 193–218. Wiley.

Govaert, G. (1977). Algorithme de classification d'un tableau de contingence. In *First international symposium on data analysis and informatics*, pages 487–500, Versailles. INRIA.

Govaert, G. (1983). *Classification croisée*. PhD thesis, Université Paris 6, France.

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.

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.

Jagalur, M., Pal, C., Learned-Miller, E., Zoeller, R. T., and Kulp, D. (2007). Analyzing in situ gene expression in the mouse brain with image registration, feature extraction and block clustering. *BMC Bioinformatics*, 8:S5.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya Series A*, 62:49–66.

Keribin, C. (2010). Méthodes bayésiennes variationnelles: concepts et applications en neuroimagerie. *Journal de la Société Française de Statistique*, 151:107–131.

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. (2013). Estimation and Selection for the Latent Block Model on Categorical Data. Rapport de recherche RR-8264, INRIA.

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). Un protocole de simulation de données pour la classification croisée. In *44ème journées de statistique*, Bruxelles.

Madeira, S. C. and Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1:24–45.

Mariadassou, M. and Matias, C. (2013). Convergence of the groups posterior distribution in latent or stochastic block models. *arXiv preprint arXiv:1206.7101v2*.

McLachlan, G. J. and Krishnan, T. (2008). *The EM algorithm and extensions*. Wiley, Nex York, 2nd edition.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, Nex York.

Meeds, E. and Roweis, S. (2007). Nonparametric bayesian biclustering. Technical Report UTML TR 2007-001, Department of Computer Science, University of Toronto.

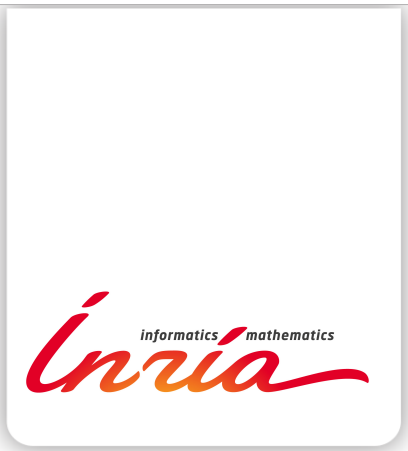
Nobile, A. and Fearnside, A. T. (2007). Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing*, 17:147–162.

Rousseau, J. and Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted models. *Journal of the Royal Statistical Society*, 73:689–710.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

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

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



RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

1 rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing
Campus de l'École Polytechnique
91120 Palaiseau

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399