Estimation and Selection for the Latent Block Model on Categorical Data

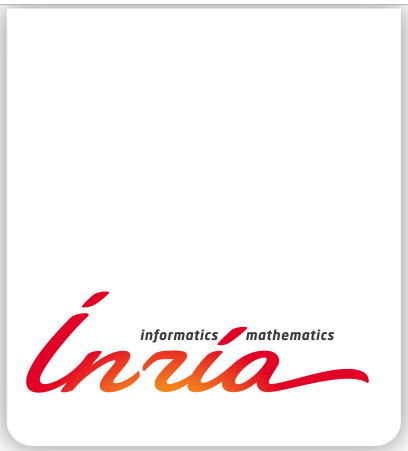
Christine Keribin, Vincent Brault, Gilles Celeux, Gérard Govaert

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 — 28 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 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 UMR 8628, Université Paris-Sud 11, F-91405 Orsay cedex
† INRIA Saclay Île de France Projet SELECT, Bat 425, Université Paris-Sud 11, F-91405 Orsay cedex
‡ UMR 7253, CNRS et Université de Technologie de Compiègne

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 of 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 (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 now described.

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 defined on $I \times J$ where $y_{ij} = h$, $1 \leq h \leq r$, r being the number of levels of the J variables. In the following, an alternative representation of the data set with binary indicator vectors will be often used for mathematical convenience: $\mathbf{y}_{ij} = (y_{ij}^h, h = 1, \ldots, r)$ with $y_{ij}^h = 1$ when $y_{ij} = h$ and $y_{ij}^h = 0$ otherwise.

It is assumed that there exists a partition into g clusters $\mathbf{z} = (z_{ik}; i = 1, \ldots, n; k = 1, \ldots g)$ on I and a partition $\mathbf{w} = (w_{j\ell}; j = 1, \ldots, d; \ell = 1, \ldots m)$ into m clusters on J, the z_{ik}s (resp. $w_{j\ell}$s) being 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

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

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 $(\rho_\ell = P(w_{j\ell} = 1), \ell = 1, \ldots, m)$ the mixing proportions of the mixture. Hence, the marginal density of $\mathbf{y}$ is a mixture density

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

$\mathcal{Z}$ and $\mathcal{W}$ denoting the sets of possible labels $\mathbf{z}$ for I and $\mathbf{w}$ for J.

The density of $\mathbf{y}$ parameterised by $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$, with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_m)$ and $\boldsymbol{\alpha} = (\alpha_{k\ell}; k = 1, \ldots g, \ell = 1, \ldots, m)$ is

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

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 (VEM) algorithm 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. Now, even if the parameter $\boldsymbol{\theta}$ is properly estimated, computing the likelihood of the LBM is challenging. 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, focusing on the LBM for categorical data, we show how it is possible to compute properly relevant model selection criteria. 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. Moreover, the VEM algorithm is presented. Section 3 is devoted to the presentation of a Stochastic version of the EM algorithm which avoids the variational approximation and is dramatically less sensitive to initial values than VEM. In Section 4, we show how Bayesian inference can be helpful to avoid empty clusters. Section 5 is concerned with model selection criteria. We first show how the integrated completed likelihood (ICL) of the LBM is closed form and we also derive the asymptotic approximations of the integrated likelihood (BIC) and the integrated completed likelihood (ICL-BIC). Section 6 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 and column clusters 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 Estimating the LBM through a variational EM algorithm

In this section, we first extend results on binary data such as presence-absence data (Govaert and Nadif, 2008; Keribin et al., 2012) to categorical data. Thus, a multinomial latent block model is considered: the conditional distribution of the outcome y_{ij} knowing the labels z_{ik} and $w_{j\ell}$ is a categorical distribution $\mathcal{M}(1, \alpha_{k\ell})$, of 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 $\mathbf{y}_{ij}$ 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

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

The $g + m + (r-1)gm - 2$ parameters to be estimated are $\boldsymbol{\pi}$, $\boldsymbol{\rho}$ and $\boldsymbol{\alpha}$.

Model identifiability Before considering the estimation problem, it is important to analyse the model generic 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 can not be applied to LBM. We set here a sufficient condition ensuring the identifiability of the Bernoulli LBM:

Theorem 1 *: With $\boldsymbol{\alpha}$, the matrix of the Bernoulli coefficients, $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$, the row and column mixing proportions of the mixture, assume that $n \geq 2m-1$ and $d \geq 2g-1$ and*

- *(H_1) for all $1 \leq k \leq g$, $\pi_k > 0$ and the coordinates of vector $\boldsymbol{\tau} = \boldsymbol{\alpha}\boldsymbol{\rho}$ are distinct,*
- *(H_2) for all $1 \leq \ell \leq m$, $\rho_\ell > 0$ and the coordinates of vector $\boldsymbol{\sigma} = \boldsymbol{\pi}'\boldsymbol{\alpha}$ are distinct (where $\boldsymbol{\pi}'$ is the transpose of $\boldsymbol{\pi}$),*

then, the binary LBM is identifiable.

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

H_1 and H_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 SBM, which is a "practical" identifiability, explaining why it works in the applications (Carreira-Perpiñàn and Renals, 2000). These assumptions 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) Assumptions H_1 and H_2 are the extension of an assumption made to ensure the identifiability of the Stochastic Block Model (Célisse et al., 2011), where $n = d$ and $\mathbf{z} = \mathbf{w}$.

It follows from assumption H_1 (resp. H_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 assumptions can be used to put a natural order on the mixture labels in a Bayesian setting. Notice that assumptions H_1 and H_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 assumptions are defined on vectors $\boldsymbol{\tau}^h = \boldsymbol{\alpha}^h\boldsymbol{\rho}$ and $\boldsymbol{\sigma}^h = \boldsymbol{\pi}'\boldsymbol{\alpha}^h$, with $h = 1, \ldots, r$ and $\boldsymbol{\alpha}^h = (\alpha^h_{kl})_{k=1,\ldots,g;\ell=1,\ldots,m}$.

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

$$\begin{aligned} L(\boldsymbol{\theta}) &= \sum_{(\mathbf{z},\mathbf{w}) \in \mathcal{Z} \times \mathcal{W}} p(\mathbf{z}; \boldsymbol{\theta}) p(\mathbf{w}; \boldsymbol{\theta}) f(\mathbf{y} | \mathbf{z}, \mathbf{w}; \boldsymbol{\theta}) \\ &= \sum_{(z,w) \in \mathcal{Z} \times \mathcal{W}} \left(\prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell,h} (\alpha^h_{k\ell})^{z_{ik} w_{j\ell} y^h_{ij}} \right). \end{aligned}$$

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, c denoting the iteration index,

$$e_{i,j,k,\ell}^{(c)} = P(z_{ik}w_{j\ell} = 1|\mathbf{y}; \boldsymbol{\theta}^{(c)}),$$

for $i = 1, \ldots, n, j = 1, \ldots, d, k = 1, \ldots, g$ and $\ell = 1, \ldots, m$, $\boldsymbol{\theta}^{(c)}$ being a current value of the parameter. Thus, the E step involves to compute too many terms that cannot be factorised as for a standard mixture, due to the dependence conditionally on the observations, of the row and column labels. For this very reason, 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 $q_{zw}^{(c)}(\mathbf{z}, \mathbf{w}) = q_z^{(c)}(\mathbf{z})q_w^{(c)}(\mathbf{w})$. To get simpler formulas, we will denote

$$s_{ik}^{(c)} = q_z^{(c)}(z_{ik} = 1)$$

and

$$t_{j\ell}^{(c)} = q_w^{(c)}(w_{j\ell} = 1).$$

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 log-likelihood 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 $\boldsymbol{\theta}^{(c)}$

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

   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 $s_{.k} = \sum_i s_{ik}$, $t_{.\ell} = \sum_j t_{j\ell}$, it leads to

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

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

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

This VEM algorithm has been proved to provide relevant estimates of the LBM model in different contexts (continuous or binary data matrices) (Govaert and Nadif, 2008). However, it presents several drawbacks illustrated in Section 6:

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

(ii) it has a marked tendency to provide spurious solutions related to empty clusters.

The algorithms we propose in this paper are aiming to answer those limitations.

3 The SEM-Gibbs algorithm

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 we propose is a simple adaptation to the LBM of the standard SEM algorithm of Celeux and Diebolt (1985). It is as follows for the categorical LBM.

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 $p(\mathbf{z}|\mathbf{y}, \mathbf{w}^{(c)}; \boldsymbol{\theta}^{(c)})$:

   $$p(z_i = k|\mathbf{y}_{i\cdot}, \mathbf{w}^{(c)}; \boldsymbol{\theta}^{(c)}) = \frac{\pi_k^{(c)} \psi_k(\mathbf{y}_{i\cdot}; \boldsymbol{\alpha}_{k\cdot}^{(c)})}{\sum_{k'} \pi_{k'}^{(c)} \psi_{k'}^{(c)}(\mathbf{y}_{i\cdot}; \boldsymbol{\alpha}_{k'\cdot}^{(c)})}, k = 1, \ldots, g$$

   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 $z_{.k} := \sum_i z_{ik}$ and $w_{.\ell} := \sum_j w_{j\ell}$, it leads to

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

and

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

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 $w_{j\ell}$. However, it is not the only difference between VEM and SEM-Gibbs: 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.

SEM-Gibbs is not increasing the loglikelihood at each iteration. It generates an irreducible Markov chain with a unique stationary distribution which is expected to be concentrated about 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. 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 spurious solutions and empty blocks. Finally, 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.

4 Bayesian inference

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 spurious solutions and thus to attenuate the "empty cluster" problem. In particular, for the categorical LBM, it is possible to consider non informative prior distribution for the model parameters (see Figure 1):

$$\boldsymbol{\pi} \sim \mathcal{D}(a, \ldots, a), \quad \boldsymbol{\rho} \sim \mathcal{D}(a, \ldots, a), \quad \boldsymbol{\alpha} \sim \prod_{k,\ell} \mathcal{D}(b, \ldots, b),$$

$\mathcal{D}(v, \ldots, v)$ denoting a Dirichlet distribution with parameter v. 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.

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

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

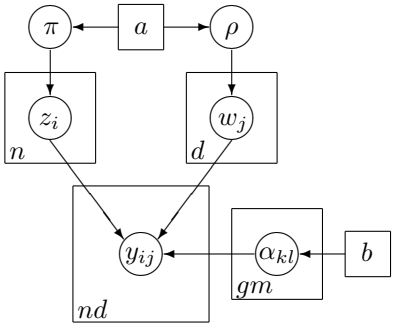


Figure 1: Bayesian latent block model.

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 straightfully 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)

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

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

The hyperparameters a and b are acting as regularisation parameters. In this perspective, the choices of a and b are important. It appears clearly from the updating equations of this M step that V-Bayes is useless for uniform priors ($a = b = 1$) and worse than useless for Jeffreys prior ($a = b = 1/2$). 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 Mergensen (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 6, 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 initiate V-Bayes with the solution derived from the Gibbs sampler that we now describe.

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 a $\mathcal{D}(a + z_{.1}^{(c+1)}, \ldots, a + z_{.g}^{(c+1)})$,
4. simulation of $\boldsymbol{\rho}^{(c+1)}$ according to a $\mathcal{D}(a + w_{.1}^{(c+1)}, \ldots, a + w_{.m}^{(c+1)})$,
5. simulation of $\boldsymbol{\alpha}_{k\ell}^{(c+1)}$ according to a $\mathcal{D}(b + {N_{k\ell}^1}^{(c+1)}, \ldots, b + {N_{k\ell}^r}^{(c+1)})$ 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, column (resp. row) labels are reordered according to the ascending values of $\boldsymbol{\tau}^h = \boldsymbol{\alpha}^h \boldsymbol{\rho}$ (resp. $\boldsymbol{\sigma}^h = \boldsymbol{\pi}' \boldsymbol{\alpha}^h$) coordinates for a given h, after each step 3 and 5 (resp. 4 and 5).

5 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. 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.

Fortunately, it is possible to compute the exact integrated completed loglikelihood (ICL) of the categorical LBM and an asymptotic approximation of the integrated likelihood could be derived from this ICL criterion.

5.1 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).

For the categorical LBM, proper non informative priors are available and ICL can be computed without requiring asymptotic approximations. Using the conjugacy 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 B for a detailed proof)

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

In practice, the missing labels $\mathbf{z}, \mathbf{w}$ are 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 ML estimate of the LBM parameter derived from the SEM-Gibbs algorithm followed by the VEM algorithm as described in Section 3.

5.2 Penalised information criteria

It could be of interest to analyse the behavior of standard information criteria as BIC and ICL-BIC derived from asymptotic approximations. Using the Stirling formula

$$\Gamma(z) \underset{+\infty}{\sim} z^{z-1/2}e^{-z}\sqrt{2\pi}$$

the asymptotic approximation of ICL as n and d tend to infinity is

$$\text{ICL-BIC}(g,m) = \max_{\boldsymbol{\theta}} \log p(\mathbf{y},\widehat{\mathbf{z}},\widehat{\mathbf{w}};\boldsymbol{\theta}) - \frac{g-1}{2}\log n - \frac{m-1}{2}\log d - \frac{gm(r-1)}{2}\log(nd) \quad (3)$$

This result is a generalisation to the categorical case of the result proved in Keribin et al. (2012) for the binary LBM. See Appendix C for details. Moreover BIC can be conjectured to have the following expression:

$$\mathrm{BIC}(g,m) = \max_{\boldsymbol{\theta}} \log p(\mathbf{y};\boldsymbol{\theta}) - \frac{g-1}{2}\log n - \frac{m-1}{2}\log d - \frac{gm(r-1)}{2}\log(nd). \quad (4)$$

As the maximised likelihood is unavailable for the LBM, it could be approximated by the maximised free energy $\mathcal{F}$. But there is no reason to replace any criterion when available with its asymptotic approximation. Thus, ICL can be preferred to ICL-BIC or BIC.

6 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

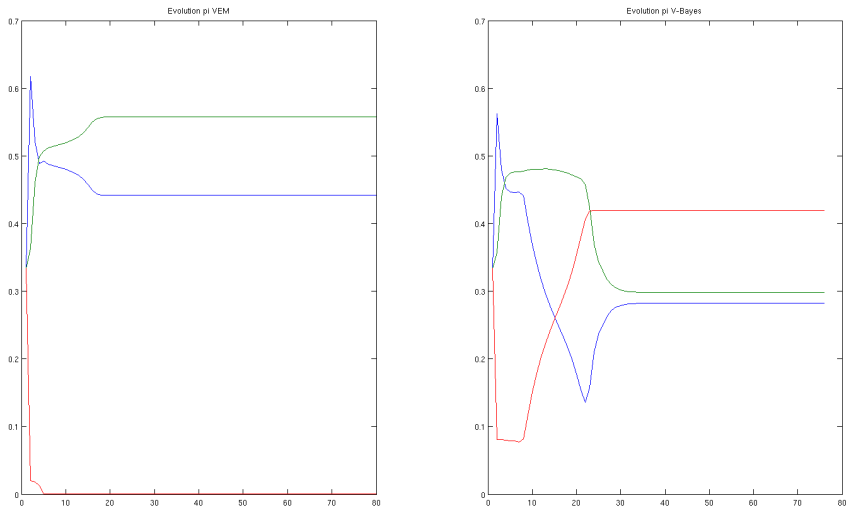


Figure 2: Evolution of $\hat{\boldsymbol{\pi}}$ for the VEM algorithm (on left) and the V-Bayes algorithm (on right)

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 compare the behaviour of SEM-Gibbs+VEM and Gibbs+V-Bayes algorithm and the influence of the hyperparameters a and b is analysed as the behaviour of the presented model selection criteria. Then, the LBM is experimented on a real categorical data set especially to assess the performances of the model selection criteria presented in Section 5.

Escaping from a spurious solution with V-Bayes A data set has been simulated according to the easy separated case of Lomet et al. (2012) and Lomet (2012) for the binary LBM model. The left part of Figure 2 displays the values of $\widehat{\boldsymbol{\pi}}$ along the iterations of the VEM algorithm while the right part of Figure 2 displays these for the V-Bayes algorithm with $(a, b) = (4, 1)$. Both algorithms started from the same position. VEM is rapidly trapped in a spurious maximum. But V-Bayes, for which the estimated proportions are not smaller than $\frac{a-1}{g(a-1)}$, escapes from this spurious solution and finally provides a satisfactory solution with the required number of clusters $g = 3$.

Comparing SEM-Gibbs+VEM and Gibbs+V-Bayes to avoid spurious solutions Monte Carlo experiments have been performed to assess the ability of Bayesian inference to avoid spurious solutions. Each considered LBM model, with $g = 5$ clusters on the rows and $m = 4$ clusters on the columns, has been replicated 500 times. The parameters of the first group of models, with equal proportions, are

$$\boldsymbol{\pi} = \begin{pmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \end{pmatrix}, \quad \boldsymbol{\rho} = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix} \text{ and } \boldsymbol{\alpha} = \begin{pmatrix} \varepsilon & \varepsilon & \varepsilon & \varepsilon \\ 1-\varepsilon & \varepsilon & \varepsilon & \varepsilon \\ 1-\varepsilon & 1-\varepsilon & \varepsilon & \varepsilon \\ 1-\varepsilon & 1-\varepsilon & 1-\varepsilon & \varepsilon \\ 1-\varepsilon & 1-\varepsilon & 1-\varepsilon & 1-\varepsilon \end{pmatrix}$$

where ε have been chosen according to the Govaert and Nadif (2008) procedure to get easily, moderately or hardly separated blocks. The second group of models differs by the following unequal mixing proportions:

$$\boldsymbol{\pi} = \begin{pmatrix} 0.1 \\ 0.15 \\ 0.2 \\ 0.25 \\ 0.3 \end{pmatrix} \text{ and } \boldsymbol{\rho} = \begin{pmatrix} 0.1 \\ 0.2 \\ 0.3 \\ 0.4 \end{pmatrix}.$$

For each model, two initial values have been chosen for the numbers of blocks:

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

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

For Bayesian inference, the hyperparameters a and b were chosen in $\{1, 4, 16\}^2 \backslash \{(1, 1)\}$. Tables 1-4 summarise the results. Notice that in these tables the algorithm SEM-Gibbs+VEM is associated by convention to the cell $a = 1$ and $b = 1$. It clearly appears that Bayesian inference with $a = 4$ or 16 and $b = 1$ is doing a good job to avoid spurious solutions. As expected, taking $a > 1$ is quite relevant in that purpose. On the contrary, it appears clearly that taking $b > 1$ is harmful.

	+			++			+++		
	b / a	1	4	16	1	4	16	1	4	16
(100,200)	1	5.6	38	65.6	7.2	20.6	56.8	13.8	11.8	48.6
	4	0.4	1	14.8	0.6	0.4	0.4	0.8	0.4	0.2
	16	0.2	0.2	1.8	0	0.4	0.2	0.2	0.2	0
(150,150)	1	5.4	16.2	47.4	5.6	8.4	36	7.4	5.2	39.2
	4	0	0	2	0.2	0	0	0.2	0	0.2
	16	0	0	0	0	0	0	0	0	0.2
(200,100)	1	3.2	12.2	49.2	5.6	4.2	34.8	10.4	6.6	44.4
	4	0.2	0	4.2	0	0.2	0.2	0	0	0
	16	0	0	0	0	0	0	0	0	0

Table 1: Percentage of spurious solutions, for data simulated with equal proportion models, get by SEM-Gibbs+VEM (cell $a = 1, b = 1$) and Gibbs+V-Bayes algorithms for 500 matrices in a easily separated +, moderately separated ++ and ill separated +++ cases, for three sample sizes starting with the true values for g and m.

Analysing the spurious 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 over sized and the algorithm provides the right or a more reasonable number of clusters. Tables 5 and 6 give the frequency of the final number of clusters got with SEM-Gibbs+VEM started with $g = 8$ and $m = 8$ for the data sets at hand. In each situation, the right number of clusters is chosen at most less than two percent of times.

Moreover, the SEM-Gibbs+VEM algorithm has been run on a binary matrix of sizes $n = 50$ and $d = 50$ simulated according to the protocol described in Lomet et al. (2012) and Lomet

	+			++			+++		
(100,200)	a \ b	1	4	16	1	4	16	1	4	16
	1	57.6	88.4	100	41.8	57.2	100	43.2	59	100
	4	0	0.8	100	0	0.8	98.2	1.8	3.6	93.2
	16	0	1.6	93.8	0.2	1	15	1.6	3	24.2
(150,150)	a \ b	1	4	16	1	4	16	1	4	16
	1	61.2	68.4	100	48.4	35.2	100	44	46	100
	4	0	0.4	100	0	0	84	0.8	1.4	77.2
	16	0	0	29.4	0	0	2	0.2	1.8	14.8
(200,100)	a \ b	1	4	16	1	4	16	1	4	16
	1	74.8	90	100	72	58.8	100	68.2	59.8	100
	4	0	0.6	100	0	0.2	99.2	0.8	1.8	95.6
	16	0	0.2	99.8	0	0	11.6	0.4	2	19.8

Table 2: Percentage of spurious solutions, for data simulated with equal proportion models, get by SEM-Gibbs+VEM (cell $a = 1, b = 1$) and Gibbs+V-Bayes algorithms for 500 matrices in the easily separated +, moderately separated ++ and ill separated +++ cases, for three sample sizes with starting values $g = 8$ and $m = 8$.

	+			++			+++		
(100,200)	a \ b	1	4	16	1	4	16	1	4	16
	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)	a \ b	1	4	16	1	4	16	1	4	16
	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)	a \ b	1	4	16	1	4	16	1	4	16
	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 3: Percentage of spurious solutions, for data simulated with unequal proportion models, get by SEM-Gibbs+VEM (cell $a = 1, b = 1$) and Gibbs+V-Bayes algorithms for 500 matrices in a easily separated +, moderately separated ++ and ill separated +++ cases, for three sample sizes starting with the true values for g and m.

(2012) to get highly separated clusters with $g = 10$ and $m = 10$. And, this algorithm was started with eight clusters for the rows and the columns. Although the numbers of requested clusters are smaller than the right values, the algorithm produces spurious solutions almost twenty percent of times (Table 7).

Analysing the behaviour of the model selection criteria The ICL, BIC and ICL-BIC criteria are studied under the unequal proportion simulation protocol: 50 data matrices are simulated and the Gibbs+V-Bayes(4,1) algorithm is run with g and m equal two to eight. Tables 8-13 summarise the number of models selected by each criterion. There is little difference between

	+			++			+++		
a \ b	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 4: Percentage of spurious solutions, for data simulated with unequal proportion models, get by SEM-Gibbs+VEM (cell $a = 1, b = 1$) and Gibbs+V-Bayes algorithms for 500 matrices in the easily separated +, moderately separated ++ and ill separated +++ cases, for three sample sizes starting with $g = 8$ and $m = 8$.

	+						++						+++					
	4	5	6	7	8		4	5	6	7	8		4	5	6	7	8	
4	0	0	0	0	0		4	0	0	0	1	0	4	0	0	1	0	0
100 5	[0]	0	1	1	1		5	[0]	0	0	0	1	5	[0]	4	1	0	0
200 6	0	16	19	15	19		6	0	8	9	4	12	6	1	7	10	8	6
7	1	5	20	34	57		7	0	3	19	24	34	7	0	8	15	18	34
8	0	5	25	69	212		8	0	3	17	75	291	8	0	4	26	73	284
	4	5	6	7	8			4	5	6	7	8		4	5	6	7	8
150 5	[0]	0	0	0	0		5	[0]	0	0	0	0	5	[1]	0	0	0	1
150 6	0	12	6	8	11		6	0	14	9	9	4	6	1	8	8	4	3
7	0	7	30	31	37		7	0	14	16	24	28	7	0	7	21	17	18
8	0	19	42	103	194		8	0	5	36	93	258	8	1	6	30	94	280
	4	5	6	7	8			3	4	5	6	7	8		4	5	6	7	8
4	0	0	0	0	0		4	1	0	0	0	0	0	4	0	0	0	0	0
200 5	[1]	1	1	0	0		5	0	[1]	2	0	0	0	5	[3]	2	0	0	0
100 6	2	13	10	11	5		6	0	0	19	9	4	3	6	2	16	13	4	1
7	0	24	31	28	12		7	0	0	15	34	24	8	7	1	22	37	23	6
8	0	32	85	118	126		8	0	0	34	79	127	140	8	1	30	68	112	159

Table 5: Number of clusters given, for data simulated with the equal proportion models, get by SEM-Gibbs+VEM for 500 matrices in the easily separated +, moderately separated ++ and ill separated +++ cases, starting with $g = 8$ and $m = 8$.

the three criteria. But the ICL criterion seems to provide a better estimation of the true number of clusters than BIC and ICL-BIC, especially for the smaller sizes of the data matrix. And, as expected, for large sizes the performance of the criteria improves and the difference between them decreases.

Congressional voting in US senate As Wyse and Friel (2012), we test our approach on UCI Congressional Voting data set recording the votes ('yes', 'no', 'abstained or absent') of 435 members of the 98th congress on 16 different key issues. The data set is available from http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records. This data set involves homogeneous 3-level categorical data. Since their method have been conceived for binary data, Wyse and Friel (2012) treated the absent and abstain votes as a 'no'. Acting in such a way, they lost potentially important information. The Gibbs+VEM algorithm has been run on the 3-level data with $a = 4$ and $b = 1$. In Figure 3 which summarises the results, the 'yes' is in green, the

+

		4	5	6	7	8
	4	0	0	0	0	0
100	5	[0]	3	2	1	0
200	6	2	16	15	13	16
	7	1	21	27	24	46
	8	0	18	47	73	175

		4	5	6	7	8
	4	1	0	0	0	0
150	5	[1]	1	1	2	0
150	6	0	15	20	9	5
	7	4	18	26	23	21
	8	0	28	63	108	154

		4	5	6	7	8
	4	0	0	0	0	0
200	5	[3]	4	0	1	0
100	6	5	21	9	7	0
	7	6	25	35	26	11
	8	7	60	82	105	93

++

		4	5	6	7	8
	4	0	0	0	1	0
	5	[5]	1	1	2	1
	6	2	13	13	6	12
	7	1	15	26	35	41
	8	0	3	24	71	227

		3	4	5	6	7	8
	4	0	0	0	0	0	0
	5	0	[4]	5	1	0	1
	6	1	1	11	9	6	1
	7	0	2	14	29	31	22
	8	0	1	17	39	76	229

		3	4	5	6	7	8
	4	1	0	0	0	0	0
	5	0	[8]	4	2	1	1
	6	0	9	13	15	8	6
	7	0	3	27	29	32	24
	8	0	2	20	57	77	161

+++

		4	5	6	7	8
	4	1	1	0	0	0
	5	[3]	2	1	3	4
	6	2	9	18	15	9
	7	2	5	13	25	37
	8	0	6	19	62	263

		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

		3	4	5	6	7	8
	4	0	0	1	0	0	0
	5	3	[8]	5	3	3	0
	6	1	13	19	11	4	0
	7	2	9	19	28	18	16
	8	0	6	18	64	90	159

Table 6: Number of provided clusters, for data simulated with the unequal proportion models, get by SEM-Gibbs+VEM for 500 matrices in the easily separated +, moderately separated ++ and ill separated +++ cases, starting with $g = 8$ and $m = 8$.

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

Table 7: 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 $m = 10$.

'no' in red and 'abstain or absent' in white. The ICL, BIC and ICL-BIC criteria are computed to determine the true number of classes.
The couple $(g = 5, m = 7)$ is chosen by ICL, while BIC and ICL-BIC selects $(g = 4, m = 6)$. These choices are quite different from the choices derived from the reversible jump sampler of Wyse and Friel (2012) which leads to prefer $g = 6$ or 7, and $m = 12$ or 13.

Row cluster composition by party is shown in Table 14. For each criterion, row cluster 1 is a republican cluster and row cluster 3 is a democrat cluster. Row cluster 2 is a mixed cluster. The main difference between ICL and the two other criteria concerns row clusters 4 and 5. With ICL, the six biggest abstainers are isolated. The rows and columns reorganisation derived from the block clustering for each criterion is displayed in Figure 3.

7 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 Variational EM or Variational Bayes algorithms the SEM-Gibbs algorithm and Gibbs sampling provide solutions not to 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 spurious solutions which jeopardizes maximum likelihood inference for the LBM. Moreover taking profit of the exhibited sufficient condition ensuring the identifiability of the

	+				++					+++			
	3	4	5	6		2	3	4	5		2	3	4
3					3	1				3	14	2	
4	18	8			4	1	46	1		4	4	30	
5	2	[17]	2		5		1	[0]		5			[0]
6		2			6					6			
7		1			7					7			

(100,200)

	3	4	5	6		2	3	4	5		2	3	4
3					3		1			3	4	10	
4	20	14			4	1	41	5		4		33	2
5	1	[6]	2		5			[2]		5		1	[0]
6		7			6					6			

(150,150)

	3	4	5	6		2	3	4	5		2	3	4
3					3		3			3	5	17	
4	15	13	5		4		32	13	1	4		27	1
5	2	[9]		3	5			[2]		5			[0]
6		3			6					6			

(200,100)

Table 8: The frequency of the selected models by the ICL criterion in the easily separated +, moderately separated ++ and ill separated + + + cases with unequal proportion for different matrix sizes.

	+				++					+++			
	3	4	5		2	3	4			2	3	4	
3					3		1			3	17	6	
4	29	10			4	1	47			4	3	24	
5		[10]			5		1	[0]		5			[0]
6		1			6					6			

(100,200)

	3	4	5		2	3	4			2	3	4	
3					3		4			3	10	17	
4	29	13			4		43	2		4		23	
5	2	[2]	1		5			[1]		5			[0]
6		3			6					6			

(150,150)

	3	4	5		2	3	4			2	3	4	
3					3		6			3	10	18	
4	30	8	1		4		40	3		4		22	
5	2	[7]			5		2	[0]		5			[0]
6		1			6					6			

(200,100)

Table 9: The frequency of the selected models by the BIC criterion in the easily separated +, moderately separated ++ and ill separated + + + cases with unequal proportion for different matrix sizes.

		+				++				+++		
		3	4	5		2	3	4		2	3	4
(100,200)	3				3	2	1		3	26	3	
	4	34	8		4	1	45		4	5	16	
	5	1	[7]		5		1	[0]	5			[0]
		3	4	5		2	3	4		2	3	4
(150,150)	3				3	2	5		3	22	13	
	4	33	13		4		42		4		14	1
	5	2	[1]		5			[1]	5			[0]
	6		1		6				6			
		3	4	5		2	3	4		2	3	4
(200,100)	3				3	1	10		3	14	20	
	4	31	10	2	4		35	5	4		16	
	5	2	[5]		5			[0]	5			[0]

Table 10: The frequency of the selected models by the BIC-ICL criterion in the easily separated +, moderately separated ++ and ill separated +++ cases with unequal proportion for different matrix sizes.

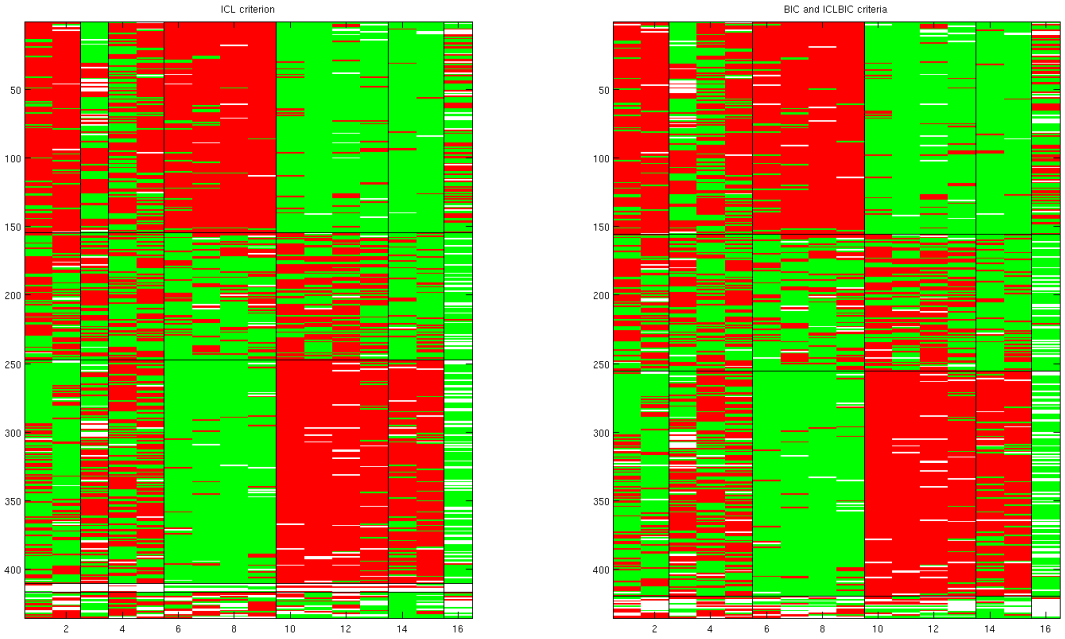


Figure 3: Reorganisation of the data by the ICL criterion (on left) and by the BIC and ICL-BIC criteria (on right).

LBM for categorical data, it is possible to define a natural order to relabel the rows and columns during Gibbs sampling in order to avoid the label switching problem in a proper way. On the

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

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 that the ICL criterion provides a consistent estimation of the number of clusters g and m for the LBM, contrary to what happens for the standard mixture model for which, under some conditions, BIC provides a consistent estimation of the number of mixture components, but not ICL. Such a result could be conjectured thanks to the work of Mariadassou and Matias (2012) who proved that, under some conditions, the conditional label distribution $p(\mathbf{z}, \mathbf{w}|\mathbf{y}; \boldsymbol{\theta})$ converges to a Dirac mass located at the actual labels of the true LBM as n and d tend to infinity. Otherwise, an alternative our approach is to use the collapsed reversible sampler

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

of Wyse and Friel (2012). But, this approach is by far more computational time demanding and we think that to get an honest estimate of g and m with a reversible sampler will require a formidable number of iterations. Moreover, Reversible jump MCMC are sensitive to the choices proposals to move around models and those choices remain difficult (Lee et al., 2009).

	+				++					+++				
		3	4	5		2	3	4	5		2	3	4	5
(100,200)	3				3	4	30			3	18			
	4				4		10	6		4	6	25		
	5	1	[34]	4	5			[0]		5		1	[0]	
	6	1	10		6					6				
(150,150)	3				3	2				3	27	1		
	4	1	1		4	3	41			4		22		
	5		[35]	3	5		0	[4]		5			[0]	
	6		10		6					6				
(200,100)	3				3	3	2			3	21	6		
	4	4	1		4		35	3		4		23		
	5		[41]	1	5			[7]		5			[0]	
	6		3		6					6				
(500,1000)	4				4					4		18		
	5		[38]		5			[44]		5		4	[28]	
	6		11		6			5		6				
	7		1		7			1		7				
(750,750)	3				3					3	1			
	4				4					4		14		
	5		[45]		5			[44]		5			[34]	
	6		5		6			6		6			1	
(1000,500)	4				4		2			4		12		
	5		[45]	4	5			[43]	5	5			[37]	1
	6		1		6					6				

Table 13: The frequency of the selected models by the ICL-BIC criterion in the easily separated +, moderately separated ++ and ill separated + + + cases with unequal proportion for different matrix sizes.

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

ICL	Rep	Dem	Total	BIC	Rep	Dem	Total
1	132	22	154	1	131	24	155
2	25	68	93	2	27	73	100
3	1	162	163	3	1	163	164
4	2	4	6	4	9	7	15
5	8	11	19				

Table 14: Repartition of democrats and republicans for each criteria.

function of matrix $\mathbf{y}$ having at least $2m-1$ rows and $2g-1$ columns. The proof is adapted from Célisse et al. (2011) who set up a similar result for the Stochastic Block Model. Notice first that $\boldsymbol{\tau} = \boldsymbol{\alpha}\boldsymbol{\rho}$ is the vector of the probability τ_k to have 1 in a cell of a row of given row class k:

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

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 Van der Monde, and hence invertible. Consider now u_i, the probability to have 1 on the i first cells of the first row of $\mathbf{y}$:

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

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 (Célisse et al., 2011) 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

$$\begin{aligned}
U_{ij} &= \mathbb{P}(y_{11} = 1, \ldots, y_{1i} = 1, y_{21} = 1, \ldots, y_{j1} = 1) \\
&= \sum_{k,\ell_1,\ell_2,\ldots,\ell_i,k_1,\ldots,k_j} \mathbb{P}(y_{11}|z_{1k} = 1, w_{1\ell} = 1)\mathbb{P}(z_{1k} = 1)\mathbb{P}(w_{1\ell} = 1) \times \\
&\quad \mathbb{P}(y_{12}|z_{1k} = 1, w_{2\ell_2} = 1)\mathbb{P}(w_{2\ell_2} = 1) \times \ldots \times \mathbb{P}(y_{1i}|z_{1k} = 1, w_{i\ell_i} = 1)\mathbb{P}(w_{i\ell_i} = 1) \times \\
&\quad \mathbb{P}(y_{21}|z_{2k_2} = 1, w_{1\ell} = 1)\mathbb{P}(z_{2k_2} = 1) \times \ldots \times \mathbb{P}(y_{j1}|z_{jk_j} = 1, w_{1\ell} = 1)\mathbb{P}(z_{jk_j} = 1) \\
&= \sum_{k,\ell} \mathbb{P}(y_{11}|z_{1k} = 1, w_{1\ell} = 1)\mathbb{P}(z_{1k} = 1)\mathbb{P}(w_{1\ell} = 1)\mathbb{P}(y_{1i}|z_{1k} = 1)^{i-1}\mathbb{P}(y_{j1}|w_{1\ell} = 1)^{j-1} \\
&= \sum_{k,\ell} \pi_k \tau_k^{i-1} \alpha_{k\ell} \rho_\ell \sigma_\ell^{j-1}.
\end{aligned}$$

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

$$U = R A_\pi \boldsymbol{\alpha} A_\rho S',$$

and it leads to define $\boldsymbol{\alpha} = A_\pi^{-1} R^{-1} U S^{'-1} A_\rho^{-1}$ 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 $\boldsymbol{\alpha}^h = (\alpha^h_{k\ell})_{k=1,\ldots,g;\ell=1,\ldots,m}$ is successively identified by considering $y_{ij} = \ell$, $i = 1, \ldots, m$, $j = 1, \ldots, g$.

B 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

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

Thus

$$\mathrm{ICL} = \log \mathrm{p}(\mathbf{z}) + \log \mathrm{p}(\mathbf{w}) + \log \mathrm{p}(\mathbf{y}|\mathbf{z}, \mathbf{w}). \tag{5}$$

Now, according to the LBM definition,

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

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

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

we have

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

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 $\mathrm{p}(\mathbf{z})$ directly follows from the Bayes theorem

$$\mathrm{p}(\mathbf{z}) = \frac{\Gamma(ag)}{\Gamma(a)^g} \frac{\prod_k \Gamma(z_{.k}+a)}{\Gamma(n+ag)}. \tag{6}$$

In the same manner,

$$\mathrm{p}(\mathbf{w}) = \frac{\Gamma(am)}{\Gamma(a)^m} \frac{\prod_\ell \Gamma(w_{.\ell}+a)}{\Gamma(d+am)}. \tag{7}$$

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

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

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

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

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,

$$\mathrm{p}(\mathbf{y}|\mathbf{z},\mathbf{w})=\prod_{k,\ell}\frac{\Gamma(rb)}{\Gamma(b)^{r}}\frac{\prod_{h}\Gamma(N_{k\ell}^{h}+b)}{\Gamma(z_{.k}w_{.\ell}+rb)}. \tag{8}$$

And, the ICL criterion, presented in Section 5.1, is straightforwardly derived from equations (5), (6), (7) and (8).

C Deriving ICL-BIC

The criterion ICL-BIC is an asymptotic approximation of ICL deduced from the Stirling approximation

$$\Gamma(z)\underset{+\infty}{\sim}z^{z-1/2}e^{-z}\sqrt{2\pi}.$$

Using this approximation and neglecting the terms not depending on n, we get

$$\begin{aligned}
\log p(\mathbf{z}) &= \log\Gamma(ag) - \log\Gamma(a)^g + \left(\sum_k \log\Gamma(z_{.k}+a)\right) - \log\Gamma(n+ag) \\
&= \sum_k \left[(z_{.k}+a-\tfrac{1}{2})\log(z_{.k}+a) - (z_{.k}+a)\right] \\
&\quad - \left[(n+ag-\tfrac{1}{2})\log(n+ag) - (n+ag)\right] + \mathcal{O}(1) \\
&= \left[\sum_k z_{.k}\log\left(z_{.k}\left(1+\tfrac{a}{z_{.k}}\right)\right)\right] + \left(a-\tfrac{1}{2}\right)\sum_k \log\left(z_{.k}\left(1+\tfrac{a}{z_{.k}}\right)\right) - n - ag \\
&\quad -n\log\left(n\left(1+\tfrac{ag}{n}\right)\right) - \left(ag-\tfrac{1}{2}\right)\log\left(n\left(1+\tfrac{ag}{n}\right)\right) + n + ag + \mathcal{O}(1) \\
&= \sum_k z_{.k}\log z_{.k} + \sum_k z_{.k}\times\frac{a}{z_{.k}} + \left(a-\tfrac{1}{2}\right)\sum_k \log z_{.k} + \left(a-\tfrac{1}{2}\right)\underbrace{\sum_k \frac{a}{z_{.k}}}_{=\mathcal{O}(1)} \\
&\quad -n\log n - n\times\frac{ag}{n} - \left(ag-\tfrac{1}{2}\right)\log n - \left(ag-\tfrac{1}{2}\right)\underbrace{\frac{ag}{n}}_{=\mathcal{O}(1)} + \mathcal{O}(1) \\
&= \sum_k z_{.k}\log z_{.k} + ag + \left(a-\tfrac{1}{2}\right)\sum_k \log z_{.k} \\
&\quad - \sum_k n\log z_{.k} - ag - g\left(a-\tfrac{1}{2}\right)\log n - \tfrac{g-1}{2}\log n + \mathcal{O}(1) \\
&= \sum_k z_{.k}\log\left(\frac{z_{.k}}{n}\right) + \left(a-\tfrac{1}{2}\right)\sum_k \underbrace{\log\left(\frac{z_{.k}}{n}\right)}_{=\mathcal{O}(1)} - \tfrac{g-1}{2}\log n + \mathcal{O}(1) \\
&\underset{+\infty}{\sim} \max_{\boldsymbol{\pi}} \log p(\mathbf{z}|\boldsymbol{\pi}) - \tfrac{g-1}{2}\log n. \qquad (9)
\end{aligned}$$

Similarly, the approximation on $\log p(\mathbf{w})$ is obtained:

$$\log p(\mathbf{w}) \underset{+\infty}{\sim} \max_{\rho} \log p(\mathbf{w}|\rho) - \tfrac{m-1}{2}\log d. \qquad (10)$$

And, using the standard BIC approximation we have

$$\int \mathrm{p}(\mathbf{y}|\mathbf{z},\mathbf{w},\boldsymbol{\alpha})\mathrm{p}(\boldsymbol{\alpha})d\boldsymbol{\alpha} \underset{+\infty}{\sim} \max_{\alpha} \log p(\mathbf{y}|\mathbf{z},\mathbf{w},\boldsymbol{\alpha}) - \tfrac{gm(r-1)}{2}\log(nd). \qquad (11)$$

Finally, the ICL-BIC criterion, presented in Section 5.2, is straightforwardly derived from equations (9), (10), and (11).

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.

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.

Célisse, A., Daudin, J.-J., and Latouche, P. (2011). Consistency of maximum-likelihood and variational estimators in the Stochastic Block Model. Technical report. http://128.84.158.119/abs/1105.3288v2.

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.

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

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.

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. (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.*

Lee, K., Marin, J.-M., Mengersen, K., and Robert, C. P. (2009). *Perspectives in Mathematical Sciences I, Probability and Statistics*, chapter Bayesian inference on mixtures of distributions, pages 165–202. World Scientific.

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. (2012). Convergence of the groups posterior distribution in latent or stochastic block models. Technical report. http://hal.archives-ouvertes.fr/hal-00713120/PDF/posterior_blockmodels_submit.pdf.

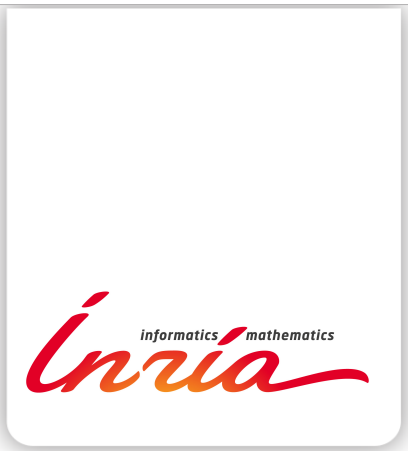
McLachlan, G. J. and Krishnan, T. (2008). *The EM algorithm and extensions.* Wiley series in probability and statistics. Wiley, 2nd edition.

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

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