Noname manuscript No.
(will be inserted by the editor)

Fast and Consistent Algorithm for the Latent Block Model

Vincent Brault · Antoine Channarond

Received: date / Accepted: date

Abstract The latent block model is used to simultaneously rank the rows and columns of a matrix to reveal a block structure. The algorithms used for estimation are often time consuming. However, recent work shows that the log-likelihood ratios are equivalent under the complete and observed (with unknown labels) models and the groups posterior distribution to converge as the size of the data increases to a Dirac mass located at the actual groups configuration. Based on these observations, the algorithm *Largest Gaps* is proposed in this paper to perform clustering using only the marginals of the matrix, when the number of blocks is very small with respect to the size of the whole matrix in the case of binary data. In addition, a model selection method is incorporated with a proof of its consistency. Thus, this paper shows that studying simplistic configurations (few blocks compared to the size of the matrix or very contrasting blocks) with complex algorithms is useless since the marginals already give very good parameter and classification estimates.

Thanks to Stéphane Robin and the reviewers for them suggestions. This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) funded by the French program Investissement d'avenir and partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003). All the computations presented in this paper were performed using the GRICAD infrastructure (https://gricad.univ-grenoble-alpes.fr), which is supported by Grenoble research communities.

Vincent Brault
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France
E-mail: vincent.brault@univ-grenoble-alpes.fr

Antoine Channarond
UMR6085 CNRS, Laboratoire de Mathématiques Raphaël Salem, Université de Rouen Normandie, 76800 Saint-Étienne-du-Rouvray, France
E-mail: antoine.channarond@univ-rouen.fr

Keywords Latent Block Model · Largest Gaps Algorithm · Model Selection · Data analysis

Mathematics Subject Classification (2020)
62H30 · 62-07 · 60K35

1 Introduction

Block clustering methods aim at clustering rows and columns of a matrix simultaneously to form homogeneous blocks. There are a lot of applications of this method: genomics [12, 14], recommender systems [2, 22], archeology [9], sociology [11, 17, 25] or network [1, 4] for example. Among the methods proposed to solve this question, the Latent Block Model (LBM) [10] provides a chessboard structure induced by the classification of the rows and the classification of the columns. In this model, it is assumed that a sample of n individuals is collected, which contains the observation of d binary variables of the same nature. Saying that the binary variables are of the same nature means that it is possible to encode them in the same (and natural) way. This assumption is needed to ensure that decomposing the dataset in a block structure makes sense and [18] shows the equivalence between searching for a classification in the LBM and solving an optimal transport problem. In the case where the rows and columns represent the same individuals, we generally speak instead of *Stochastic Block Model* [1, 4, 24]. However, in the case where the matrices are not symmetric, [15] shows that it may be more interesting to use the LBM model rather than the SBM model. Although this model is very similar, we will not discuss it in this article.

Given the number of blocks and in order to estimate the parameters, [10] suggest using a variational algorithm, [16] propose an adaptation of the Stochastic Expectation Maximisation introduced by [6] in the mixture case, [17] studied a Bayesian version of these two algorithms and [25] propose a Bayesian algorithm including the estimation of the number of blocks. However, each of these iterative algorithms has a complexity at least $\mathcal{O}\left(ndN_{Block}N_{Iter}\right)$ where N_{Block} is the number of blocks and N_{Iter} is the number of iterations necessary for the convergence of the algorithm. Moreover, the procedures are often associated with a model selection criterion requiring the computation of maximum likelihood estimators for all combinations of the expected number of blocks. However, the theoretical results obtained show that the distributions of the estimators (see [5, 7, 19]) are asymptotically trivial: the log-likelihood ratios are equivalent under the complete and observed (with unknown labels) models and the groups posterior distribution to converge as the size of the data increases to a Dirac mass located at the actual groups configuration.

In this article, we propose an adaptation of the *Largest Gaps* (LG) algorithm introduced by [8] in the *Stochastic Block Model* with a low complexity $\mathcal{O}(nd)$ (Section 3) and present the conditions to obtain asymptotically a good estimation (Section 4). We then prove that it provides a consistent procedure for all inference tasks inherent in LBM: unsupervised classification and estimation of the parameters and a selection procedure for the number of blocks (the last unknown theoretical point) is also proposed and shown to be consistent (Section 5). For ease of reading, the proofs of the results are postponed to the appendices (Sections A, B and C). These theoretical results are also illustrated on simulated data (Section 6).

By proving the consistency of the LG algorithm, some features of the asymptotic regime of the LBM will be highlighted, in particular, the concentration of some marginal distributions of the model. The secondary objective of the paper is to discuss as a conclusion (Section 7) the consequence of this when fitting the LBM to large matrices; in particular, this leads to take a step back regarding the relevance of the model and the estimates when the number of blocks is small with respect to the number of cells of the matrix.

2 Notations and model

The binary Latent Block Model (LBM) is as follows. Let $\mathbf{x} = (x_{ij})_{i=1,\dots,n;j=1,\dots,d}$ be the data matrix where $x_{ij} \in \{0,1\}$; observation of a random variable $\mathbf{X}$. It is assumed that there exists a partition into g row clusters $\mathbf{z} = (z_{ik})_{i=1,\dots,n;k=1,\dots,g}$ and a partition into m column clusters $\mathbf{w} = (w_{j\ell})_{j=1,\dots,d;\ell=1,\dots,m}$. 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 X_{ij} are independent conditionally on $\mathbf{z}$ and $\mathbf{w}$ with parametric density $\varphi(\cdot;\alpha_{k\ell})^{z_{ik}w_{j\ell}}$, where $\alpha_{k\ell}$ is the parameter of the conditional density of the data given $Z_{ik} = 1$ and $W_{j\ell} = 1$. As the binary case is studied, the density is defined for each $x \in \{0,1\}$ and $\alpha \in [0;1]$ by

$$\varphi(x;\alpha) = \alpha^x (1-\alpha)^{1-x}.$$

Thus, the density of $\mathbf{X}$ conditionally on $\mathbf{z}$ and $\mathbf{w}$ is defined for each $\mathbf{x}$ by

$$\begin{aligned} f(\mathbf{x}|\mathbf{z},\mathbf{w};\boldsymbol{\alpha}) &= \prod_{i=1}^{n}\prod_{j=1}^{d}\prod_{k=1}^{g}\prod_{\ell=1}^{m}\varphi(x_{ij};\alpha_{k\ell})^{z_{ik}w_{j\ell}} \\ &=: \prod_{i,j,k,\ell}\varphi(x_{ij};\alpha_{k\ell})^{z_{ik}w_{j\ell}} \end{aligned}$$

where $\boldsymbol{\alpha} = (\alpha_{k\ell})_{k=1,\dots,g;\ell=1,\dots,m}$.

Moreover, it is assumed that the row and column labels $\mathbf{z}$ and $\mathbf{w}$ are the observations of two independent variables $\mathbf{Z}$ and $\mathbf{W}$: $p(\mathbf{z},\mathbf{w};\boldsymbol{\pi},\boldsymbol{\rho}) = p(\mathbf{z};\boldsymbol{\pi})p(\mathbf{w};\boldsymbol{\rho})$ with $p(\mathbf{z};\boldsymbol{\pi}) = \prod_{i,k}\pi_k^{z_{ik}} = \prod_k \pi_k^{z_{+k}}$ and $p(\mathbf{w};\boldsymbol{\rho}) = \prod_{j,\ell}\rho_\ell^{w_{j\ell}} = \prod_\ell \rho_\ell^{w_{+\ell}}$, where $(\pi_k = \mathbb{P}(Z_{ik} = 1), k = 1,\dots,g)$ and $(\rho_\ell = \mathbb{P}(W_{j\ell} = 1), \ell = 1,\dots,m)$ are the mixing proportions and $z_{+k} = \sum_{i=1}^n z_{ik}$ (resp. $w_{+\ell} = \sum_{j=1}^d w_{j\ell}$) represents the number of rows (resp. columns) in the class k (resp. ℓ). Hence, the density of $\mathbf{X}$ is defined for every $\mathbf{x}$ by

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

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

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

To estimate both the classification and the parameters, many algorithms exist (for example [10], [17] or [25]) but each of these algorithms has a complexity larger than $\mathcal{O}(ndgmN_{Iter})$ where N_{Iter} is the number of iterations of the algorithm. This makes their use on large matrices difficult.

In the Stochastic Block Model (SBM), rows and columns are associated with the same individuals, which allows to represent a graph, whereas LBM allows to represent bipartite graphs. [8] suggested a fast algorithm, called *LG* for *Lagest Gaps*, based on a marginal of the matrix $\mathbf{x}$, the degrees.

3 Algorithm *Largest Gaps*

Before the introduction of the classification algorithm *Largest Gaps* (*LG*), let us recall the main idea, inspired by [8].

3.1 Main ideas

Conditionally on $Z_{ik} = 1$, meaning that row i is in class k, the probability that the variable X_{ij} equals 1 for any $j \in \{1, \dots, d\}$ is

$$\begin{aligned}\mathbb{P}(X_{ij} = 1 | Z_{ik} = 1) &= \sum_{\ell=1}^{m} \mathbb{P}(X_{ij} = 1 | Z_{ik} = 1, W_{j\ell} = 1)\, \mathbb{P}(W_{j\ell} = 1 | Z_{ik} = 1) \\ &= \sum_{\ell=1}^{m} \mathbb{P}(X_{ij} = 1 | Z_{ik} = 1, W_{j\ell} = 1)\, \mathbb{P}(W_{j\ell} = 1) \\ &= \sum_{\ell=1}^{m} \alpha_{k\ell}\rho_{\ell} =: \tau_k. \end{aligned} \tag{1}$$

In particular, conditionally on $Z_{ik} = 1$, variables of row i are independent and identically distributed Bernoulli variables with parameter τ_k and the sum of the cells of any row i, denoted by X_{i+}, is hence a binomial distributed variable $\mathcal{B}in(d, \tau_k)$.

As a consequence of the subgaussian concentration property of binomial distributions, variables $\overline{X_{i\bullet}} = \frac{X_{i+}}{d}$ fastly concentrates around the mean associated with its own class when d tend to infinity. The point is here: if moreover these means $\tau_1, \dots, \tau_g$ are pairwise distinct, the set of the variables $\{\overline{X_{i\bullet}}; i = 1, \dots, n\}$ asymptotically splits into clusters, separated by large gaps, and which exactly correspond to the clusters of the model, for d large enough. In the whole paper, $\tau_1, \dots, \tau_g$ will be assumed to be pairwise distinct (Assumption (I), see discussion in Section 4).

The middle right picture of Figure 1 shows the histogram of the variable set $\{\overline{X_{i\bullet}}; i = 1, \dots, n\}$ for a matrix simulated under LBM with five clusters. The five clusters of rows can be seen, as well as the four large gaps which separate them. The middle left picture of Figure 1 is a representation of the vector ($\overline{X_{i\bullet}}$ sorted in ascending order) and the bottom left picture of Figure 1, the size of the gaps between two consecutive sorted values.

To classify the rows, the idea is to identify the gaps between the clusters, which are expected to be asymptotically larger. Indeed, the internal gaps (those between two rows of the same cluster) vanish when d tends to infinity due to concentration, while the external gaps (those between two rows of distinct clusters)

do not, since the means are assumed to be distinct. Symmetrically the same holds for the columns.

There are several strategies to identify the large gaps. In their article, [8] assume that the number Q of clusters (the same for the rows and the columns) is known and partition the population into Q clusters by finding the $Q-1$ largest gaps. In order to choose Q, a model selection procedure can be firstly done separately, before the classification. The strategy chosen in this article consists in thresholding the gaps, to distinguish outer gaps from inner ones. It advantageously yields both the clusters and the numbers of clusters in only one pass.

The choice of thresholds is critical. Let's comment on the beautiful configuration of Figure 1, where the asymptotic regime has been reached, to present the key issues. If the thresholds are too large (greater than 0.06 on that example), the thresholding step will select only some of the four outer deviations and will not distinguish the five clusters from each other. Conversely, if they are too small (less than 0.01), the thresholding step will select all four outer gaps, but also some undesirable inner gaps, and the algorithm will wrongly separate some clusters.

3.2 Algorithm

The algorithm *Largest Gaps*[1] is given in Algorithm 1 and an illustration is provided in Figure 1. The principle is to calculate the means of the values of the cells of each row and each column, to order them and to calculate the differences between two consecutive values. Once this step is done, the clusters are formed assuming that each difference greater than the threshold implies a change of cluster. In the sequel, the estimators provided by the algorithm are denoted by $\widehat{\mathbf{Z}}$, $\widehat{\mathbf{W}}$ and $\widehat{\boldsymbol{\theta}}$.

Estimator of $\boldsymbol{\theta}^\star$. For the rest of the article, $\boldsymbol{\theta}^\star = (\boldsymbol{\pi}^\star, \boldsymbol{\rho}^\star, \boldsymbol{\alpha}^\star)$ represents the parameters to be estimated (i.e. the parameters that were used to simulate the data). In the algorithm 1, the estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^\star$ is based on $\widehat{\mathbf{Z}}$ and $\widehat{\mathbf{W}}$. $\widehat{\pi_k}$ (resp. $\widehat{\rho_\ell}$) is the proportion of class k (resp. ℓ) in the partition $\widehat{z}$ (resp. $\widehat{w}$) and the estimator $\widehat{\boldsymbol{\alpha}}$ is for all $(k,\ell) \in \{1,\dots,\widehat{g}\} \times \{1,\dots,\widehat{m}\}$:

$$\widehat{\alpha_{k\ell}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} \widehat{z_{ik}}\,\widehat{w_{j\ell}}\,x_{ij}}{\widehat{z_{+k}}\,\widehat{w_{+\ell}}}.$$

Complexity of the algorithm. On algorithm 1, the complexities of each step are added in blue at the end of the line. In the end, the LG algorithm has a complexity of $\mathcal{O}\left(\max\left(nd\left[\widehat{g}+\widehat{m}\right], n\log n, d\log d\right)\right)$. As will be seen in the section 5.3, $\log n$ is required to be much smaller than d and $\log d$ much smaller than n. In this case, the complexity is $\mathcal{O}\left(nd\left[\widehat{g}+\widehat{m}\right]\right)$.
Moreover, $\sum_{i=2}^{n} G_i = \overline{X_{\bullet(n)}} - \overline{X_{\bullet(1)}}$ being smaller than 1 and for all $k \in \{1,\dots,\widehat{g}-1\}$, $G_{(i_k)}$ being greater than S_g then, in the worst case, $\widehat{g}$ is smaller than $1/S_g + 1$. As a conclusion, the complexity is $\mathcal{O}\left(nd\left[1/S_g + 1/S_m\right]\right)$ and, if the classification only is processed, the complexity is $\mathcal{O}(nd)$.

[1] An implementation in the language `R` is available on the following Gitlab: https://gricad-gitlab.univ-grenoble-alpes.fr/braultv/largest-gaps

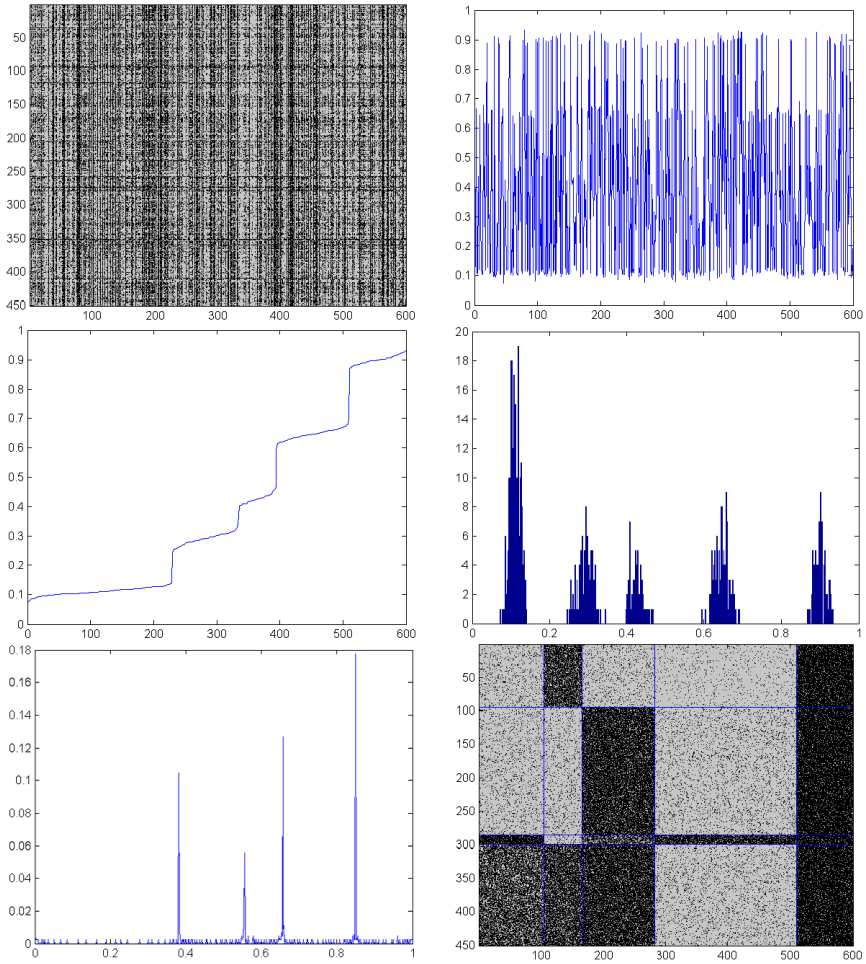

Fig. 1 Top-left: Initial matrix. Top-right: Example of a vector $\left(\overline{X_{\bullet(1)}}, \ldots, \overline{X_{\bullet(d)}}\right)$. Middle-left: representation of the vector $\left(\overline{X_{\bullet(1)}}, \ldots, \overline{X_{\bullet(d)}}\right)$ sorted in increasing order. Middle-right: histograms of $\left(\overline{X_{\bullet(1)}}, \ldots, \overline{X_{\bullet(d)}}\right)$. Bottom-left: representation of the vector of gaps $(G_2, \ldots, G_d)$ where for all $j \in \{2, \ldots, d\}$, $G_j = \overline{X_{\bullet(j)}} - \overline{X_{\bullet(j-1)}}$. Bottom-right: reorganized matrix.

4 Identifiability of the model: assumptions on the model parameters

In this section, the conditions for the consistency of the estimators are explained. For the rest of the article, $g^\star$ and $m^\star$ represent the true number of clusters in rows and columns and, as mentioned in the section 3.2, $\boldsymbol{\theta}^\star = (\boldsymbol{\pi}^\star, \boldsymbol{\rho}^\star, \boldsymbol{\alpha}^\star)$ corresponds to the true parameters. Moreover, $\mathbf{z}^\star$ and $\mathbf{w}^\star$ are the unobserved partition matrices, resulting from the laws $\mathcal{M}(1; \boldsymbol{\pi}^\star)$ and $\mathcal{M}(1; \boldsymbol{\rho}^\star)$, used to simulate the data.

Input: data matrix $\mathbf{x}$, threshold for row S_g and for column S_m.

// Computation of gaps
for $i \in \{1, \ldots, n\}$ **do**
  Computation of $\overline{X_{i\bullet}} = \frac{x_{i+}}{d}$.
// $\mathcal{O}(nd)$

Ascending sort of $(\overline{X_{(1)\bullet}}, \ldots, \overline{X_{(n)\bullet}})$. // $\mathcal{O}(n \log n)$

for $i \in \{2, \ldots, n\}$ **do**
  Computation of the gaps $G_i = \overline{X_{(i)\bullet}} - \overline{X_{(i-1)\bullet}}$.
// $\mathcal{O}(n)$

// Computation of $\widehat{g}$
Selection of $i_1 < \ldots < i_{\widehat{g}-1}$ such that $(G_{i_1}, \ldots, G_{i_{\widehat{g}-1}})$ are every greater than S_g.
// $\mathcal{O}(n)$

// Computation of $\widehat{\mathbf{Z}}$
for $i \in \{(1), \ldots, (n)\}$ **do**
  Definition of $\widehat{z}_{(i)k} = 1$ if and only if $(i_{k-1}) < (i) \leq (i_k)$ with $i_0 = 0$ and $i_{\widehat{g}} = n$.
// $\mathcal{O}(n)$

// Computation of $\widehat{m}$ and $\widehat{\mathbf{W}}$
Do the same on the columns. // $\mathcal{O}(\max(dn, d \log d))$

// Computation of $\widehat{\boldsymbol{\theta}}$
for $k \in \{1, \ldots, \widehat{g}\}$ **do**
  Computation of $\widehat{\pi_k} = \frac{\widehat{z}_{+k}}{n}$.
// $\mathcal{O}(n/S_g)$

for $\ell \in \{1, \ldots, \widehat{m}\}$ **do**
  Computation of $\widehat{\rho_\ell} = \frac{\widehat{w}_{+\ell}}{d}$.
// $\mathcal{O}(d/S_m)$

Computation of $\widehat{\boldsymbol{\alpha}} = \left(\widehat{\mathbf{Z}}\right)^T \mathbf{x} \widehat{\mathbf{W}} / \left[\widehat{\pi_k} \left(\widehat{\rho_\ell}\right)^T\right] \times nd$. // $\mathcal{O}\left(nd\left[1/S_g + 1/S_m\right]\right)$

Output: Numbers of clusters $\widehat{g}$ and $\widehat{m}$, matrices $\widehat{\mathbf{Z}}$ and $\widehat{\mathbf{W}}$ and parameter $\widehat{\boldsymbol{\theta}}$.

Algorithm 1: Algorithm *Largest Gaps*. The complexity of each step is highlighted in blue.

Notations 1 *Key model parameters:*
Let us define $\pi^\star_{\min}$ *and* $\rho^\star_{\min}$ *the smallest class proportions:*

$$\pi^\star_{\min} = \min_{1 \leq k \leq g^\star} \pi^\star_k \text{ and } \rho^\star_{\min} = \min_{1 \leq \ell \leq m^\star} \rho^\star_\ell$$

and the smallest distance between any two conditional expectations of the normalized degrees (called model smallest gaps):

$$\delta_{\boldsymbol{\pi}^\star} = \min_{1 \leq k \neq k' \leq g^\star} \left|\tau^\star_k - \tau^\star_{k'}\right| \text{ and } \delta_{\boldsymbol{\rho}^\star} = \min_{1 \leq \ell \neq \ell' \leq m^\star} \left|\xi^\star_\ell - \xi^\star_{\ell'}\right|$$

where $\boldsymbol{\tau}^\star = \boldsymbol{\rho}^\star \boldsymbol{\alpha}^{\star T}$ *and* $\boldsymbol{\xi}^\star = \boldsymbol{\pi}^\star \boldsymbol{\alpha}^\star$ *are the proportions of the binomial distributions defined in Equation* (1).

Cluster identifiability by the *LG* algorithm is possible under the following sufficient conditions:

Assumptions 1 ***(I)*** *The proportion of each row class (respectively column class) is non-negative, and conditional expected degrees of row (resp. column) clusters* $(\tau_k)_{1\leq k\leq g^\star}$ *(resp.* $(\rho_\ell)_{1\leq \ell\leq m^\star}$*) are all distinct, which respectively amounts to:*

$$\pi^\star_{\min}\rho^\star_{\min} > 0 \text{ and } \delta_{\boldsymbol{\pi}^\star}\delta_{\boldsymbol{\rho}^\star} > 0. \tag{I}$$

These assumptions will be always made in the sequel of this article. They are equivalent to the sufficient conditions of identifiability of LBM given in [17] but stronger than the conditions of [5] or the identifiability conditions in the case where $(g, m) = (2, 2)$ in the [17].

The first assumption, $\pi^\star_{\min}\rho^\star_{\min} > 0$, is classical in mixture models (for example, see [5, 13, 20]); without this assumption, the probability of having the right number of clusters is null. Thus the mixture model would be degenerated, as clusters with proportion zero would be always empty and the number of actually present clusters would be improper.

The second one, $\delta_{\boldsymbol{\pi}^\star}\delta_{\boldsymbol{\rho}^\star} > 0$, ensures that not only the distribution of $\mathbf{x}$ is identifiable, but also that of $(X_{i+})_{1\leq i\leq n}$, which is a marginal distribution of $\mathbf{x}$. This is critical in order to recover all clusters with the LG algorithm, which actually infers the clusters only from the set of variables $\{\overline{X_{i\cdot}}; i = 1, \ldots, n\}$. More precisely, it ensures that the distribution of X_{i+} is a proper mixture with the same number of clusters as the distribution of X_{ij}. For example, Assumption (I) excludes the following typical case (yet identifiable according to [17]):

$$g = 2,\ m = 2,\ \boldsymbol{\pi} = (1/2, 1/2)$$
$$\text{and } \boldsymbol{\alpha} = \begin{pmatrix} a & b \\ b & a \end{pmatrix} \text{ where } a, b \in]0, 1[.$$

Indeed, it gives $\tau_1 = \tau_2 = \frac{a+b}{2}$, and hence $\delta_{\boldsymbol{\pi}^\star} = 0$. Conditionally on either $z_{i1} = 1$ or $z_{i2} = 1$, the distribution of X_{i+} is the same: $\mathcal{B}in\left(d, \frac{a+b}{2}\right)$, and therefore the two clusters cannot be distinguished just with the vector $(X_{i+})_{1\leq i\leq n}$. The distribution of X_{i+} is not a proper mixture with two distinct clusters, it is a simple binomial distribution. Thus with our approach this model would be confused with the model $g = 1,\ \boldsymbol{\pi} = 1,\ \boldsymbol{\alpha} = \frac{a+b}{2}$. Note that the set of parameters such that $\delta_{\boldsymbol{\pi}^\star}\delta_{\boldsymbol{\rho}^\star} = 0$ has zero Lebesgue measure.

5 Consistency

This section presents the main result (Theorem 2), namely the consistency of estimators. Here, consistency means that the numbers of clusters and classifications are correct and that the distance between the estimates and the model parameters is smaller at any $t > 0$ with a probability tending towards one, when the size of the matrix (n, d) tends towards infinity. Before stating this theorem, we introduce some notations, especially related to the label switching problem.

5.1 Distance on the parameters and the label switching issue

For any two parameters $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ with (g, m) clusters and $\boldsymbol{\theta}' = (\boldsymbol{\pi}', \boldsymbol{\rho}', \boldsymbol{\alpha}')$, with (g', m') clusters, we define their distance as follows:

$$d^{\infty}\left(\boldsymbol{\theta}, \boldsymbol{\theta}'\right) = \begin{cases} \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_{\infty} \text{ if } g = g',\ m = m' \\ +\infty \text{ otherwise,} \end{cases}$$

where $\|\cdot\|_{\infty}$ denotes the norm defined for any $\mathbf{y} \in \mathbb{R}^g$ by $\|\mathbf{y}\|_{\infty} = \max_{1 \leq k \leq g} |y_k|$.

We assume that two matrices $\mathbf{z}, \mathbf{z}' \in \mathcal{M}_{n \times g}(\{0,1\})$ are equivalent, denoted $\mathbf{z} \equiv_{\mathcal{Z}} \mathbf{z}'$, if there exists a permutation $s \in \mathfrak{S}(\{1, \ldots, g\})$ such that for all $(i,k) \in \{1, \ldots, n\} \times \{1, \ldots, g\}$, $z'_{i,s(k)} = z_{ik}$. By convention, we assume that two matrices with different numbers of columns are not equivalent. We introduce the similar notation $\equiv_{\mathcal{W}}$ for the matrix $\mathbf{w}$.
For all parameter $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ with (g, m) clusters and for all permutions $(s,t) \in \mathfrak{S}(\{1, \ldots, g\}) \times \mathfrak{S}(\{1, \ldots, m\})$, we denote $\boldsymbol{\theta}^{s,t} = (\boldsymbol{\pi}^s, \boldsymbol{\rho}^t, \boldsymbol{\alpha}^{s,t})$, by:

$$\boldsymbol{\pi}^s = \left(\pi_{s(1)}, \ldots, \pi_{s(g)}\right), \quad \boldsymbol{\rho}^t = \left(\rho_{t(1)}, \ldots, \rho_{t(m)}\right)$$
$$\text{and } \boldsymbol{\alpha}^{s,t} = \left(\alpha_{s(1),t(1)}, \alpha_{s(1),t(2)}, \ldots, \alpha_{s(1),t(m)}, \alpha_{s(2),t(1)}, \ldots, \alpha_{s(g),t(m)}\right).$$

Like all mixture models, the LBM is affected by the label switching problem: clusters are defined up to a permutation. The algorithm can therefore find the right clusters but at a permutation of the labels. This also affects the parameters of the model, because the order of their coordinates depends on the labeling. The comparison of two classifications, and two parameter estimates, must then be done carefully: the distance between two parameter estimates must be calculated after permutation of their coordinates, using the permutation transforming the label allocation of the classification algorithm into the original label allocation of the model. Moreover, such a permutation exists and is unique when the classification is correct, i.e. when $\widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^{\star}$ (respectively $\widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^{\star}$). This permutation will thus be noted $s_{\mathcal{Z}}$ (resp. $t_{\mathcal{W}}$) on the event $\{\widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^{\star}\}$ (resp. $\widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^{\star}$). Thus, the consistency of the parameter estimators amounts to proving that the following quantity disappears in probability when (n, d) tends to infinity:

$$\forall t > 0, \mathbb{P}\left[d^{\infty}\left(\widehat{\boldsymbol{\theta}}^{s_{\mathcal{Z}}, t_{\mathcal{W}}}, \boldsymbol{\theta}^{\star}\right) > t \middle| \widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^{\star}, \widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^{\star}\right] \underset{n,d \to +\infty}{\longrightarrow} 0.$$

Outside of the event $\{\widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^{\star}\}$ (resp. $\widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^{\star}$), $s_{\mathcal{Z}}$ (resp. $t_{\mathcal{W}}$) will be defined as any arbitrary permutation in $\mathfrak{S}(\{1, \ldots, \widehat{g}\})$ (resp. $\mathfrak{S}(\{1, \ldots, \widehat{m}\})$), the identity for instance.

5.2 A non-asymptotic upper bound

This paragraph presents the main technical tool for obtaining consistency: a non-asymptotic upper bound from which the consistency results will be derived.

Theorem 1 (Concentration inequality) *Under identifiability assumptions (I), and if* $S_g \in]0, \delta_{\boldsymbol{\pi}^\star}[$ *and* $S_m \in]0, \delta_{\boldsymbol{\rho}^\star}[$ *for* n, d *large enough, then we have for all* $t > 0$*:*

$$\begin{aligned}
&\mathbb{P}\left(\widehat{g} \neq g^\star \text{ or } \widehat{m} \neq m^\star \text{ or } \widehat{\mathbf{Z}} \not\equiv_{\mathcal{Z}} \mathbf{z}^\star \text{ or } \widehat{\mathbf{W}} \not\equiv_{\mathcal{W}} \mathbf{w}^\star \text{ or } d^\infty\left(\widehat{\boldsymbol{\theta}}^{s_{\mathcal{Z}}, t_{\mathcal{W}}}, \boldsymbol{\theta}^\star\right) > t\right) \\
&\leq 2n \exp\left(-\frac{d}{2}\min(\delta_{\boldsymbol{\pi}^\star} - S_g, S_g)^2\right) + g^\star\left(1 - \pi^\star_{\min}\right)^n \\
&\quad + 2d \exp\left(-\frac{n}{2}\min(\delta_{\boldsymbol{\rho}^\star} - S_m, S_m)^2\right) + m^\star\left(1 - \rho^\star_{\min}\right)^d \\
&\quad + 2g^\star m^\star\left[1 - \pi^\star_{\min}\rho^\star_{\min}\left(1 - e^{-2t^2}\right)\right]^{nd} + 2g^\star e^{-2nt^2} + 2m^\star e^{-2dt^2}.
\end{aligned}$$

The proof (in Appendix A) is made in two steps, emphasizing the originality of the method in comparison with EM-like algorithms: here the classification is completely done first, and parameters are then estimated afterwards. Thus an upper bound on classifications and selection of class numbers will be first established (Proposition 1), and secondly an upper bound on the parameter estimators, given that both classifications and class numbers are right (Proposition 2).

If n and d increase at the same speed, the larger the size, the smaller the bound. We find the importance that S_g and S_m are small enough to detect the real jumps but not so small that there are only them. Moreover, the larger $\pi^\star_{\min}$ and $\rho^\star_{\min}$ are, the more likely it is to have at least one representative of each cluster.

In addition to showing the consistency of the *LG* algorithm's estimators (see section 5.3), the bound of theorem 1 quantifies how (very) easy a configuration is: given a configuration $(\boldsymbol{\theta}, n, d)$, it is possible to estimate the probability that the algorithm *Largest Gaps* finds the right configuration. In the case of a configuration obtained by another algorithm, this bound can be calculated and if it is (very) small, it means that a study of the marginals would have been sufficient; if the *Largest Gaps* algorithm does not find the same results at all, a doubt can be cast on the results (estimation or relevance of the use of a LBM) by precaution.

Remark 1 In the proof, the Hoeffding inequality is used. In the case where the parameters are very close to 0 or 1, Bernstein's inequality allows to improve the bound. This is particularly interesting if the parameters evolve with n and d (see Section 5.4).

5.3 Consistency of the estimators

The following theorem provides sufficient assumptions on the threshold sequences $(S_g^{n,d}, S_m^{n,d})_{n,d}$ to ensure the consistency of the inference method based on the *LG* algorithm when n, d tend to infinity. Note that this result is therefore asymptotic only, and it does not provide any guarantee for fixed (n, d). Nevertheless the rates can be used as a suggestion to choose the thresholds, as even though it must be carefully used, because it is anyway impossible to know whether the asymptotic regime is reached or not.

Theorem 2 (Consistency) *Under identifiability assumptions* (I) *and the following assumptions:*

$$\begin{gathered} S_g^{n,d} \underset{n,d\to+\infty}{\longrightarrow} 0, \quad S_m^{n,d} \underset{n,d\to+\infty}{\longrightarrow} 0, \\ \underline{\lim}_{n,d\to+\infty} S_g^{n,d}\sqrt{\frac{d}{\log n}} > \sqrt{2} \quad \textit{and} \quad \underline{\lim}_{n,d\to+\infty} S_m^{n,d}\sqrt{\frac{n}{\log d}} > \sqrt{2}, \end{gathered} \tag{TA}$$

where $\underline{\lim}$ *is the lower limit; then classifications, model selection and estimators are consistent, that is, for all* $t > 0$*:*

$$\begin{aligned} &\mathbb{P}\left(\widehat{g} \neq g^\star \text{ or } \widehat{m} \neq m^\star \text{ or } \widehat{\mathbf{Z}} \not\equiv_{\mathcal{Z}} \mathbf{z}^\star \text{ or } \widehat{\mathbf{W}} \not\equiv_{\mathcal{W}} \mathbf{w}^\star \right. \\ &\left. \text{or } d^\infty\left(\widehat{\boldsymbol{\theta}}^{s_{\mathcal{Z}},t_{\mathcal{W}}}, \boldsymbol{\theta}^\star\right) > t\right) \underset{n,d\to+\infty}{\longrightarrow} 0. \end{aligned}$$

The proof of this result is available in Appendix B.

The *LG* algorithm has two input parameters, the (S_g, S_m) thresholds, which must be set correctly to discover all clusters. Recall that the purpose of gap thresholding is to distinguish between external and internal gaps (see comments in section 3.1). First, the thresholds must be smaller than the smallest gaps in the model, $S_g^{n,d} < \delta_{\boldsymbol{\pi}^\star}$ and $S_m^{n,d} < \delta_{\boldsymbol{\rho}^\star}$, for n and d sufficiently large; otherwise, some clusters will consist of mixed clusters. Since $\delta_{\boldsymbol{\pi}^\star}$ and $\delta_{\boldsymbol{\rho}^\star}$ are not known a priori, we assume that the thresholds decrease with respect to n and d, to ensure that they are asymptotically small enough. On the other hand, if the threshold sequences decrease too fast, the thresholds will be asymptotically too small and at least one class will be split into several clusters by the algorithm. More technically, the convergence rate given in the theorem guarantees that the upper bound of the theorem 1 tends to 0 which implies consistency. If the sequences disappear faster, coherence is no longer guaranteed.

Consistency can thus be obtained whatever the sequences $S_g^{n,d}$ and $S_m^{n,d}$ taken provided that they verify the conditions (TA); nevertheless, if we couple with the condition that $S_g^{n,d} < \delta_{\boldsymbol{\pi}^\star}$ and $S_m^{n,d} < \delta_{\boldsymbol{\rho}^\star}$ of Theorem 1, we see that it is preferable that the thresholds decrease rapidly. For example, the threshold $S_g^{n,d} = \sqrt{2\log(n)/d}(1+\varepsilon)$ with $\varepsilon > 0$ can be used. Moreover, once a first configuration has been found, it is possible to estimate $\delta_{\boldsymbol{\pi}^\star}$ and $\delta_{\boldsymbol{\rho}^\star}$ and to re-estimate the partitions with the thresholds $\widehat{\delta_{\boldsymbol{\pi}^\star}}/2$ and $\widehat{\delta_{\boldsymbol{\rho}^\star}}/2$ in order to see if a better configuration is obtained.

Remark 2 The assumption (TA) of the theorem implies that $n/\log d$ and $d/\log n$ tend to $+\infty$; this assumption is found in the results on the consistency of the maximum likelihood estimator (see [5]). Therefore, $\mathbf{x}$ is allowed to have an oblong shape. For example, $d = n^\gamma$ with $\gamma > 0$ satisfies the assumption.

Remark 3 The roles of the rows and columns are symmetric: the result remains true if the transpose of $\mathbf{X}$ is studied rather than the matrix $\mathbf{X}$.

5.4 Consistency when model parameters vary

Finally varying model parameters are also considered in this paragraph. Indeed when n and d increase, it can be reasonable to assume that new clusters arise:

in this paragraph $g^{\star n,d}$ and $m^{\star n,d}$ are hence assumed to be growing to infinity when n, d tend to $+\infty$. The consequence of this assumption is the convergence to zero of both the proportions of the smallest clusters and the model smallest gaps: $\pi^\star_{\min}, \rho^\star_{\min}, \delta_{\boldsymbol{\pi}^\star}, \delta_{\boldsymbol{\rho}^\star}$ tend to 0 when n, d tend to infinity. For example, since the parameters $\boldsymbol{\tau}^\star = (\tau^\star_k)_{1\leq k\leq g^\star}$ are probabilities, the following inequalities are obtained:

$$(g^\star - 1)\delta_{\boldsymbol{\pi}^\star} \leq \sum_{k=1}^{g^\star - 1} (\tau^\star_{(k+1)} - \tau^\star_{(k)}) = \tau^\star_{(g^\star)} - \tau^\star_{(1)} \leq 1$$

$$\text{and } (m^\star - 1)\delta_{\boldsymbol{\rho}^\star} \leq \sum_{l=1}^{m^\star - 1} (\xi^\star_{(l+1)} - \xi^\star_{(l)}) \leq 1,$$

where $\tau^\star_{(1)} < \cdots < \tau^\star_{(g^\star)}$ (resp. $\xi^\star_{(1)} < \cdots < \xi^\star_{(m^\star)}$) are the $(\tau^\star_k)_{1\leq k\leq g^\star}$ (resp. $(\xi^\star_\ell)_{1\leq l\leq m^\star}$) sorted in increasing order. In the other side, $g^\star$ is bounded by $1/\pi^\star_{\min}$ by the following inequality:

$$g^\star \pi^\star_{\min} \leq \sum_{k=1}^{g^\star} \pi^\star_k = 1.$$

In particular, it is possible that one of these parameters tend to zero while the number of groups does not change (for example, in the case of increasingly sparse matrices). In this framework, admissible convergence rates of the model parameters are provided, as well as the corresponding admissible convergence rates of the thresholds. It thus tells how robust the consistency is.

Theorem 3 (Consistency in sparse case) *Under the assumptions of the previous theorem ((I) and (TA)) and the following additional assumptions :*

– *Assumptions on $\delta^{n,d}_{\boldsymbol{\pi}^\star}$ and $S^{n,d}_g$ (resp. $\delta^{n,d}_{\boldsymbol{\rho}^\star}$ and $S^{n,d}_m$):*

$$\varliminf_{n,d\to+\infty} \frac{\delta^{n,d}_{\boldsymbol{\pi}^\star}}{S^{n,d}_g} > 2, \quad \textit{and} \quad \varliminf_{n,d\to+\infty} \frac{\delta^{n,d}_{\boldsymbol{\rho}^\star}}{S^{n,d}_m} > 2. \tag{MA.1}$$

– *Assumptions on $\pi^{\star\, n,d}_{\min}$ and $\rho^{\star\, n,d}_{\min}$:*

$$n\pi^{\star\, n,d}_{\min} \underset{n,d\to+\infty}{\longrightarrow} +\infty \quad \textit{and} \quad d\rho^{\star\, n,d}_{\min} \underset{n,d\to+\infty}{\longrightarrow} +\infty \tag{MA.2}$$

then classifications, model selection and estimators are also consistent.

The proof of this result is available in Appendix C.

Theorem 3 shows that the estimates remain consistent even with sparse matrices. On the other hand, for an optimal result and if a number of non-zero cells is fixed, the estimation will be all the easier if the latter are well distributed between the classes, for example, in a staircase shape.

Remark 4 As $\pi^{\star\, n,d}_{\min} \leq 1/g^{\star n,d}$ and $\rho^{\star\, n,d}_{\min} \leq 1/m^{\star n,d}$, the assumptions (MA.2) imply that

$$\frac{nd}{N^{\star\, nd}_{\text{block}}} \underset{n,d\to+\infty}{\longrightarrow} +\infty$$

where $N^{\star\, nd}_{\text{block}} = g^{\star n,d} m^{\star n,d}$ is the number of blocks. In particular, the number of blocks can therefore increase with n and d but not too quickly.

6 Simulations

In this section, simulations to test the quality of the bounds obtained in the previous theorems and to compare the computing times with classical procedures are presented[2].

6.1 Estimation of the number of clusters

We use an experimental design to illustrate the results of Theorem 2. As the number of row (resp column) clusters is the basis of the other estimations, this is the only parameter studied in this section. The experimental design is defined with $g^\star = 5$ and $m^\star = 4$ and the following parameters

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

with $\varepsilon \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. For $\varepsilon = 0.05$, the values of the Bernoulli parameters are 0.05 (resp. 0.95) and the associated blocks contain essentially white (resp. black) cases. At the opposite, for $\varepsilon = 0.3$, the block are more homogeneous and more difficult to distinguish. For the class proportions, we suppose two possibilities

- Balanced proportions:

$$\boldsymbol{\pi}^\star = \begin{pmatrix} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\rho}^\star = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}$$

  with the following parameters

$$\pi^\star_{\min} = 0.2 \text{ and } \delta_{\boldsymbol{\pi}^\star} = 0.25 - 0.5\varepsilon.$$

- Arithmetic proportions:

$$\boldsymbol{\pi}^\star = \begin{pmatrix} 0.1 & 0.15 & 0.2 & 0.25 & 0.3 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\rho}^\star = \begin{pmatrix} 0.1 & 0.2 & 0.3 & 0.4 \end{pmatrix}$$

  with the following parameters

$$\pi^\star_{\min} = 0.1 \text{ and } \delta_{\boldsymbol{\pi}^\star} = 0.1 - 0.2\varepsilon.$$

The number of rows n and the number of columns d vary between 20 and 4000 by step 20 and for each configuration, 100 matrices were simulated. For the choice of the thresholds S_g, we studied four cases:

1. We first propose a constant oracle threshold to illustrate the Theorem 1, which suggests that any constant threshold strictly between 0 and $\delta_{\boldsymbol{\pi}^\star}$ gives the consistency:

$$S_1 = \delta_{\boldsymbol{\pi}^\star}/2.$$

[2] For the sake of repeatability, all codes are available on the following Gitlab: https://gricad-gitlab.univ-grenoble-alpes.fr/braultv/largest-gaps

2. In practice, there is mostly no reason why the parameter $\delta_{\boldsymbol{\pi}^\star}$ could be known. In this case, Theorem 2 claims that we need to use a varying threshold $S_g^{n,d}$ instead, such that $S_g^{n,d}$ tends to 0 but not too fast (slower than $\sqrt{2\log n/d}$). If the threshold decreases too slowly, it may be larger than $\delta_{\boldsymbol{\pi}^\star}$ and the smallest gaps could be undetected. On the opposite, if the threshold decreases too fast, we may detect too many gaps. In the simulation, we studied three possibilities:
   (a) Faster threshold: $S_2^{n,d} = \sqrt{2\log(n)/d}\left(1+10^{-10}\right)$.
   (b) Middle threshold: $S_3^{n,d} = 2\sqrt{2\log(n)/d}$.
   (c) Slower threshold: $S_4^{n,d} = (\log(n)/d)^{1/4}$.

Figures 2 and 3 show the proportions of bad estimations of $g^\star$ following the parameters ε (in rows) as function of the number of rows n, the numbers of columns d and the thresholds used (columns). For each figure, the number of columns (d) increases on the x-axis and the number of rows (n) increases on the y-axis. The red color corresponds to a bad estimate while the blue color corresponds to a good estimate (the more blue it is, the better).

As expected, it appears that the best threshold is the oracle $S_1 = \delta_{\boldsymbol{\pi}^\star}/2$ but this threshold cannot be used in practice because $\delta_{\boldsymbol{\pi}^\star}$ is unknown. For the scaled thresholds, $S_2^{n,d} = \sqrt{2\log n/d}\left(1+10^{-10}\right)$ is the best.

We can see that the larger the number of rows n is, the worse the estimation is and the larger the number of columns d is, the better the estimation is. In the case of $n = d$, the quality of the estimation increases with n. In particular, the model selection can be generalized for the case of [8] and the results would be similar. $\pi^\star_{\min}$ has a weak effect because it is rare to have an empty class but the effect of $\delta_{\boldsymbol{\pi}^\star}$ seems to be greater.

6.2 Estimation of the parameters

To illustrate the convergence of the the estimation $\widehat{\boldsymbol{\theta}}$ to the true parameter $\boldsymbol{\theta}^\star$, the same experimental design is chosen with the number of rows n and the number of columns d vary between 40 and 4000 by step 40 and, for each configuration, 100 matrices were simulated. To estimate the quality of the estimation, the distance d^∞ introduced in Section 5.1 is calculated. As the distance equals $+\infty$ when the number of clusters of $\widehat{\boldsymbol{\theta}}$ is different of its of $\boldsymbol{\theta}^\star$, we chose to represent the mean of the finite values with the size of the point corresponding at the number of finite values (see Figure 4 for the balanced case and Figure 7 for the arithmetic case).

We observe that the error decreases with the number of observations and the results are identical when the numbers of clusters are correctly estimated. The optimization of the LG algorithm thus depends on the choice of the threshold. In particular, we find that the oracle threshold (S_1) finds the right number of clusters faster than the evolutionary thresholds. In particular,the threshold $S_4^{n,d}$ does not appear on the graphic because it underestimates the number of blocks.

6.3 Comparison of computing times

To conclude the part of the simulations, the computation times are estimated and compared with a classical procedure. For that, the plan presented previously is

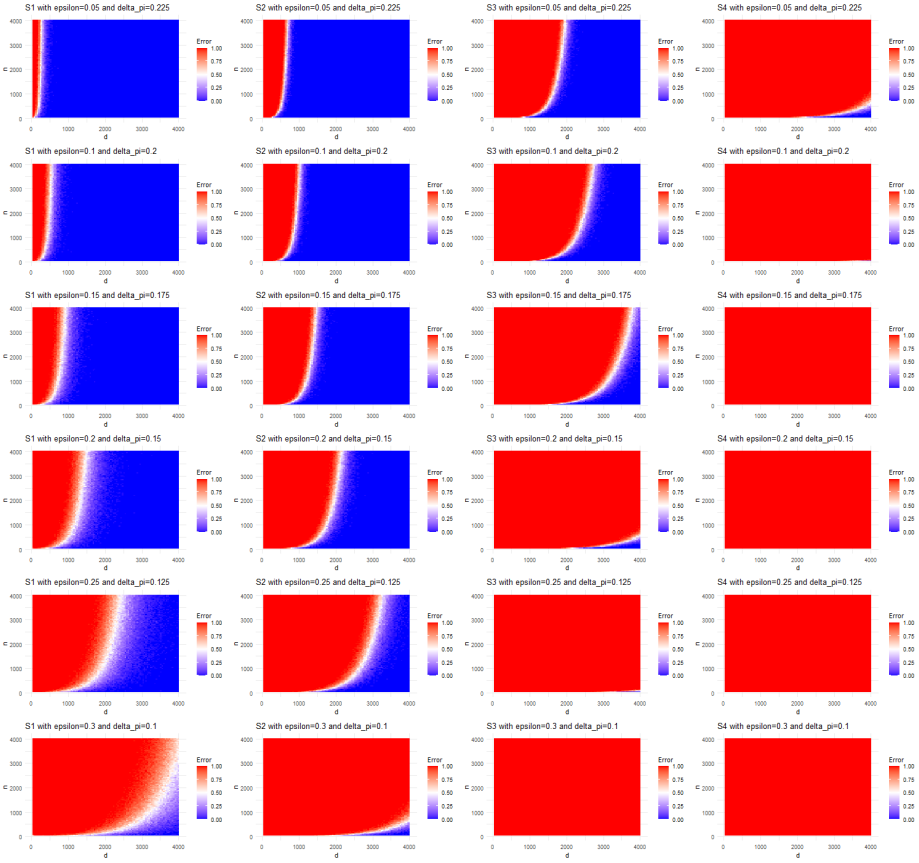

Fig. 2 Proportions of bad estimations of $g^\star$ for the parameter $\varepsilon \in \{0.05, \ldots, 0.3\}$ (in rows) following the thresholds (in columns) used for the balanced case: for each graphic, the number of rows n (y-axis) and the number of columns d (x-axis) varies between 20 and 4000.

taken again by taking only the square matrices (n=d) with $n \in \{100, 200, \ldots, 1\,400\}$ for the balanced case and $n \in \{100, 200, \ldots, 1\,300\}$ for the arithmetic. Six procedures are studied:

- the algorithm *LG* with the four previous threshold: S_1, $S_2^{n,d}$, $S_3^{n,d}$ and $S_4^{n,d}$;
- the algorithm *variational Bayes* with the hyperparameters $(4, 4)$ as proposed by [17] knowing the true number of parameters $g^\star$ and $m^\star$ (named *Simple VBayes* for the next);
- as the number of blocks is usually unknown, the combination of the algorithm *variational Bayes* and the *Integrated Complete-data Likelihood* (see [3, 17]) is studied with the hyperparameters $(4, 4)$ on a 2×7 square grid (named *VBayes+ICL* for the next).

For the implementation, the `R` package `blockcluster` (version 4.5.1; see [23]) with the function `coclusterBinary` is used. For each configuration, 20 matrices are simulated and procedures are evaluated on two criteria:

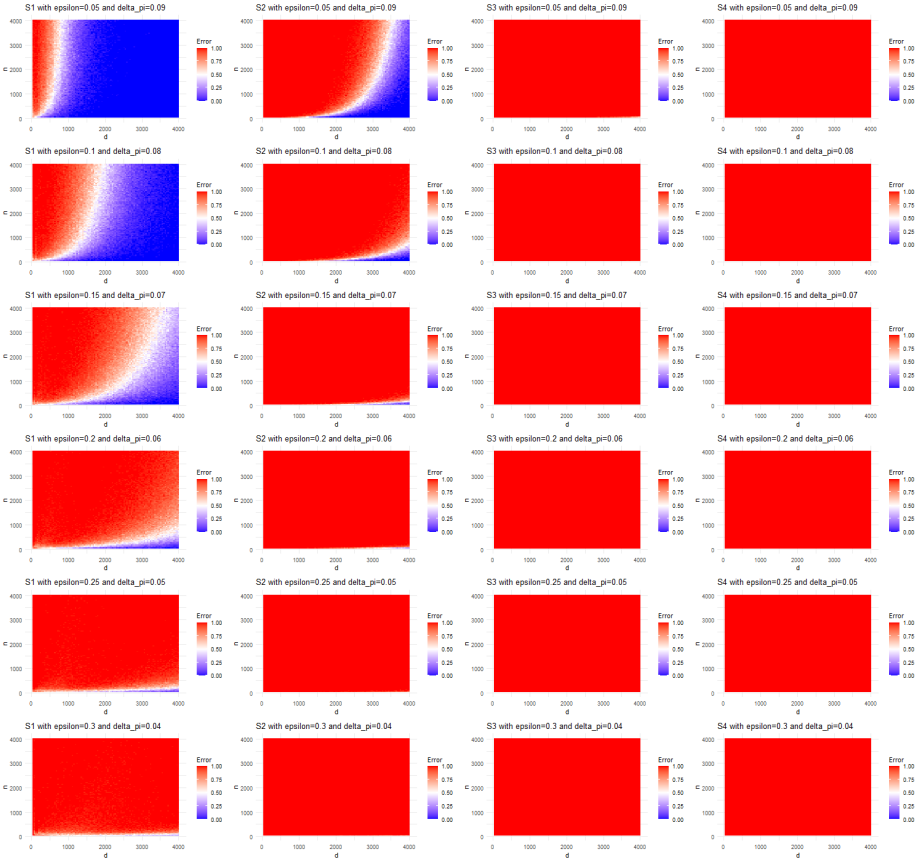

Fig. 3 Proportions of bad estimations of $g^\star$ for the parameter $\varepsilon \in \{0.05, \ldots, 0.3\}$ (in rows) following the thresholds (in columns) used for the arithmetic case: for each graphic, the number of rows n (y-axis) and the number of columns d (x-axis) varies between 20 and 4000.

– the mean computation times with 10 runs; for this, the package `microbenchmark` (version 1.4.9; see [21]) with the function `microbenchmark` is used and plans are launched on a cluster[3].
– the estimation quality of the number of clusters per row after an estimation of each matrix (except for *Simple VBayes* as the number is assumed be known); results are averaged over the 20 matrices.

The results are displayed in the Figures 5 (balanced case) and 6 (arithmetic case) where the computation times (in seconds with logarithmic scale; on the top) and the quality of the estimations (averaged for each ε on over the 20 matrices; on the bottom) have been grouped by matrix size (regardless of the values of ε). A detailed version for each ε is available on the Figures 8, 9, 10 and 11 in supple-

[3] Cluster `Luke44`, 28 cores, 128Go RAM, GPU 2xK40m, 2xIntel Xeon E5-2680 2.40 GHz; more informations on the url https://scalde.gricad-pages.univ-grenoble-alpes.fr/web/pages/presentation.html

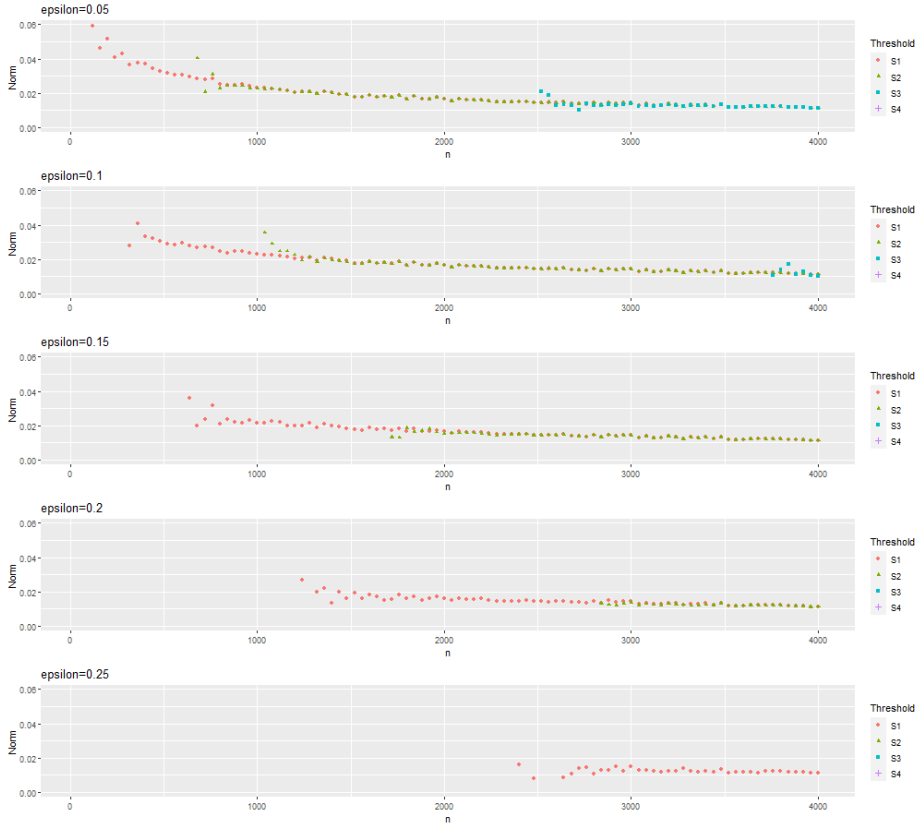

Fig. 4 Average estimate of the distance $d^{\infty}\left(\boldsymbol{\theta}^{\star}, \widehat{\boldsymbol{\theta}}\right)$ between the true parameters and the estimated parameters following ε (rows) in the balanced case: for each graphic, the threshold is represented by different symbols (● for S_1, ▲ for $S_2^{n,d}$, ■ for $S_3^{n,d}$ and + for $S_3^{n,d}$) and the size for the number of finite values used; the number of rows n varies between 40 and 4000 and d is supposed equal n.

mentary material. Moreover, the means and standard deviations (in milliseconds) are represented in the Tables 1 and 2.

On the quality, the procedure *VBayes+ICL* estimates better than the other the number of clusters in rows and the differences between the four threshold are the same than the section 6.1.

For the computing times, the order of *Largest Gaps* is that of the millisecond while that of *VBayes+ICL* is of the second (and the minute after $n = 300$). Moreover, the computing time for the *Largest Gaps* seems to increase linearly with n (the side length of the square matrix) as stated in the section 3.2 while *VBayes+ICL* increases much faster than linearly. Moreover, *Largest Gaps* does not seem to be impacted by the configuration while the convergence time of *VBayes+ICL* (and even *Simple VBayes* who knows the right number of clusters) is larger for the case of arithmetic proportions. And finally, the computation time of *VBayes+ICL* depends on the maximum choice of the number of clusters (fixed

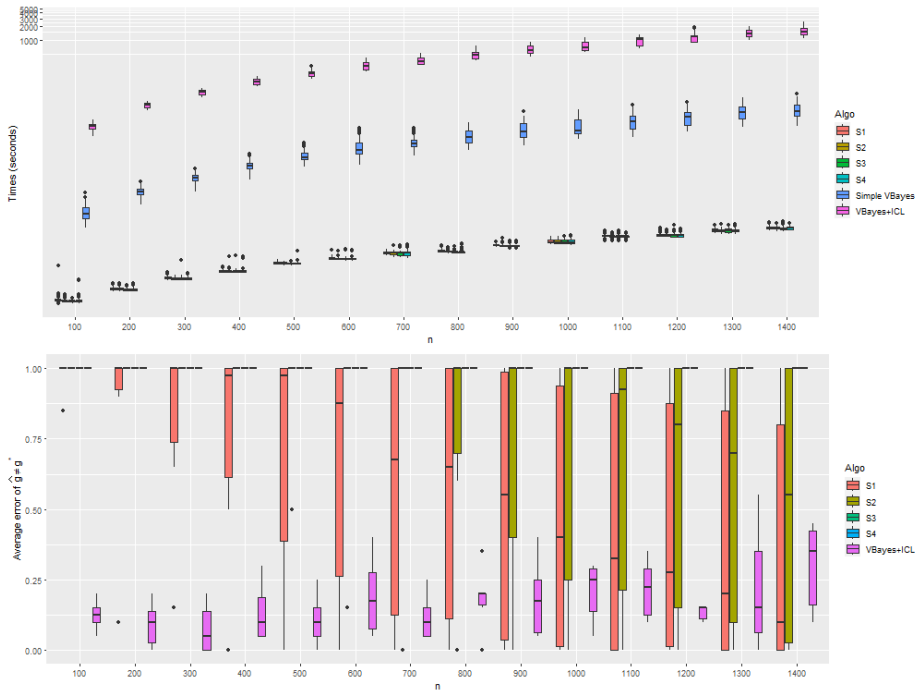

Fig. 5 On the top, boxplots of the computing times (in seconds; logarithmic scale) for each procedure (colour) in function of the numbers of rows and columns ($n = d$) over the 1200 simulations (200 per ε) for the balanced case. On the bottom, boxplots of quality of estimations averaged over the 20 matrices for each ε for each procedure (colour) in function of the numbers of rows and columns ($n = d$) for the balanced case.

here quite close to the true result) but unknown in practice while *Largest Gaps* does not need this kind of parameter and the time is almost independent of the number of blocks.

7 Conclusion and discussion

First of all, *Largest Gaps* is a co-clustering algorithm, which has nice theoretical properties: its computational cost is much lower than most known algorithms. And it provides a consistent procedure under the Latent Block Model, for all inference tasks: model selection, classification and estimation of the model parameters. Since the algorithm is simple, the consistency is rather easy to obtain. Note that in this article, only binary matrices have been studied, but the model as well as the method and the proofs can be directly extended for distributions which have the same concentration properties, for example compactly-supported distributions, where the support is known.

As a consistent algorithm, the advantage of the LG algorithm is the simplicity: it is a simple and original way to exploit the concentration of the marginal distributions of the matrix $\mathbf{x}$ under LBM. But it lacks robustness: a single outlier in

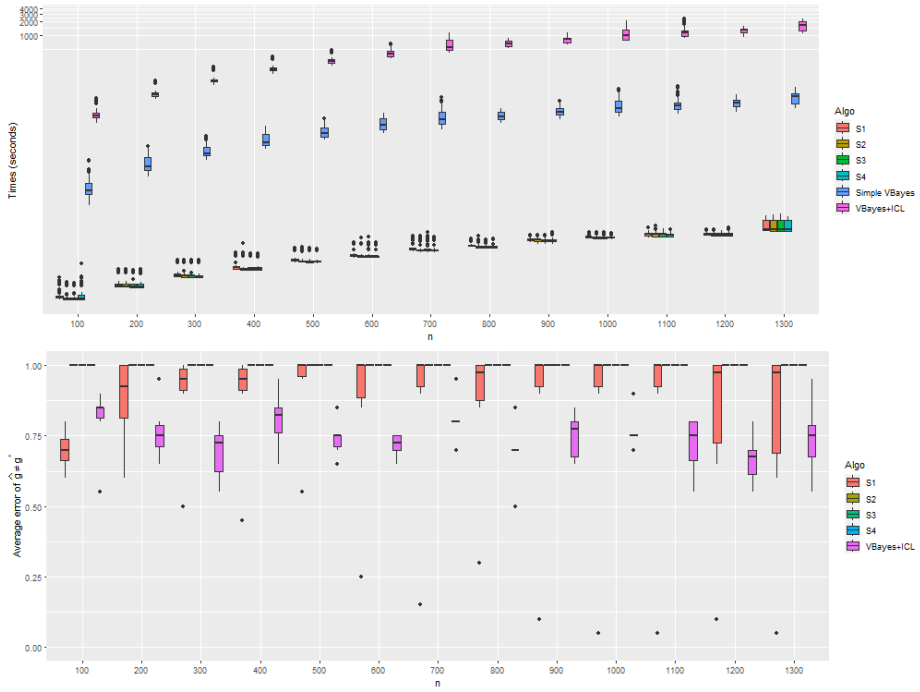

Fig. 6 On the top, boxplots of the computing times (in seconds; logarithmic scale) for each procedure (colour) in function of the numbers of rows and columns ($n = d$) over the 1200 simulations (200 per ε) for the arithmetic case. On the bottom, boxplots of quality of estimations averaged over the 20 matrices for each ε for each procedure (colour) in function of the numbers of rows and columns ($n = d$) for the balanced case.

the marginal distribution can affect the whole classification, which is not desirable by the use on real-world datasets. Moreover, the procedure is essentially based on the choice of threshold and simulations have shown that there is still room for improvement in this choice. Many other algorithms, also based on this nice feature, could be used instead and might be better in practice, but harder to analyze from a theoretical point of view. For example, fitting a binomial mixture model on the variables $\overline{X_{(1)\bullet}}, \dots \overline{X_{(n)\bullet}}$ with an EM-algorithm, could be fruitful as well.

The contribution of this article actually goes beyond the LG algorithm. It shows special features of the latent structure of the LBM, and their consequences. In particular, when the asymptotic regime of the model is reached, the latent structure is almost obvious, and moreover it can be pick either one but not the two from a summary of the data. Indeed even basic algorithms like LG can retrieve the latent clusters from variables $\overline{X_{(1)\bullet}}, \dots \overline{X_{(n)\bullet}}$ (sums of the rows and of the columns of the matrix $\mathbf{x}$), whereas most known classification algorithms are unusable, because of their complexity.

A consequence of this remark is that the LBM should be used sparingly on very large real data sets: if the number of clusters requested is too small and if the blocks are sufficiently contrasted then the marginals should highlight the clusters.

Table 1 Computing times for the different procedures in columns (LG with the four threshold, $VBayes$ with the true number of clusters and $VBayes$ coupled with the ICL criterion on a 2×7 square grid) following the number of rows and columns (in rows) for the balanced case. Each cell represents the average (in milliseconds) and the standard deviation (in parenthesis) over the 1200 simulations (200 per ε).

	S_1	$S_2^{n,d}$	$S_3^{n,d}$	$S_4^{n,d}$	Simple VBayes	VBayes+ICL
n=100	2 (0.9)	2 (0.1)	2 (0.1)	2 (0.8)	171 (71.5)	12765 (2180.1)
n=200	3 (0.4)	3 (0.3)	3 (0.3)	3 (0.3)	468 (102)	37193 (4385.4)
n=300	6 (0.5)	6 (0.4)	6 (0.9)	6 (0.4)	930 (204.1)	71594 (8711.8)
n=400	8 (0.6)	8 (1)	8 (1)	8 (1.4)	1789 (461.6)	123632 (18941.2)
n=500	13 (0.8)	12 (0.6)	12 (0.6)	12 (1.2)	2934 (846.2)	187319 (29995)
n=600	16 (1.4)	16 (1)	16 (1.8)	16 (2.1)	4597 (2044)	280348 (57861.3)
n=700	21 (1.5)	21 (2)	21 (2)	21 (2.6)	5724 (1718.6)	356932 (61458.8)
n=800	23 (1.6)	22 (1.3)	22 (1.4)	22 (1.9)	8050 (2900.3)	490034 (107033.6)
n=900	31 (1.9)	30 (2.3)	30 (1.9)	30 (1.8)	11057 (4469.8)	629238 (136921.4)
n=1000	37 (3.6)	37 (4)	37 (3.9)	37 (4)	13357 (5909.4)	761629 (189071.3)
n=1100	49 (5.3)	48 (4.6)	49 (5)	48 (4.6)	17068 (6931.2)	977295 (212403.4)
n=1200	52 (4.8)	52 (4.9)	52 (5.4)	52 (5.3)	21133 (7826.5)	1206144 (300568.7)
n=1300	66 (6.5)	65 (6.3)	65 (7.2)	65 (6.6)	27259 (10467)	1427121 (278186.1)
n=1400	75 (6)	75 (6.2)	73 (6.5)	74 (6.4)	30831 (12370.8)	1601915 (346575)

If this is not the case, the use of the LBM should be questioned. The bound of the theorem 1 would give an estimate of the quality of the use of this model.

Finally, it appears in the simulations that the estimate of the number of clusters is underestimated. Moreover, the distribution of marginals from real-world data rarely has such obvious deviations as assumed by the model asymptotic. To overcome this problem, it would be interesting to estimate the row clusters with a mixture model on the variables $(\overline{X_{(1)\bullet}}, \dots, \overline{X_{(n)\bullet}})$; this will be the subject of future work.

Table 2 Computing times for the different procedures in columns (LG with the four threshold, $VBayes$ with the true number of clusters and $VBayes$ coupled with the ICL criterion on a 2×7 square grid) following the number of rows and columns (in rows) for the arithmetic case. Each cell represents the average (in milliseconds) and the standard deviation (in parenthesis) over the 1200 simulations (200 per ε).

	S_1	$S_2^{n,d}$	$S_3^{n,d}$	$S_4^{n,d}$	Simple VBayes	VBayes+ICL
n=100	2 (0.9)	2 (0.7)	2 (0.7)	2 (1.2)	530 (302.3)	20819 (7768.3)
n=200	4 (1.4)	4 (1.4)	4 (1.4)	4 (1.4)	1677 (693.7)	58494 (18294.9)
n=300	7 (2.4)	7 (2.4)	7 (2.4)	7 (2.4)	3183 (1145.8)	117450 (34865.2)
n=400	10 (3.2)	10 (3.7)	10 (3.2)	9 (3.3)	5637 (2119)	203479 (55356.2)
n=500	14 (3.9)	14 (4)	13 (3.9)	13 (3.9)	8503 (2862.3)	304668 (76486.8)
n=600	18 (5.2)	18 (5.5)	18 (5.9)	18 (5.3)	12423 (3889.1)	435626 (100281.1)
n=700	25 (6.5)	24 (6.6)	24 (7)	24 (6.8)	17794 (7596.6)	622631 (180216.9)
n=800	27 (2.1)	26 (2.9)	25 (2.6)	25 (2.5)	18475 (3680)	684012 (94177.3)
n=900	36 (3.2)	35 (3.3)	35 (3.2)	35 (3.8)	22635 (4697.3)	828716 (106707.4)
n=1000	41 (2.8)	41 (3)	41 (3.4)	40 (2.2)	31799 (13840.7)	1130163 (371089.7)
n=1100	47 (4.4)	47 (5.5)	47 (5.7)	46 (4.6)	32203 (11270.2)	1226172 (290324.2)
n=1200	47 (3.3)	46 (3.6)	47 (3.6)	47 (4.2)	34399 (6683.5)	1295896 (167982.4)
n=1300	72 (21.2)	72 (21.3)	72 (21.8)	71 (21)	45497 (11230.9)	1673698 (353671.7)

A Proof of Theorem 1

In this appendix, we present demonstrations of the concentration inequality. We first estimate the probability of having the right number of clusters and then the right classification knowing that we have the right number of clusters. Finally, we evaluate the quality of the parameter estimates.

First of all, note that $\{\widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^\star\} \subset \{\widehat{g} = g^\star\}$ and $\{\widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^\star\} \subset \{\widehat{m} = m^\star\}$, hence :

$$
\begin{aligned}
&\mathbb{P}\left(\widehat{g} \neq g^\star \text{ or } \widehat{m} \neq m^\star \text{ or } \widehat{\mathbf{Z}} \not\equiv_{\mathcal{Z}} \mathbf{z}^\star \text{ or } \widehat{\mathbf{W}} \not\equiv_{\mathcal{W}} \mathbf{w}^\star \text{ or } d^\infty\left(\widehat{\boldsymbol{\theta}}^{s_{\mathcal{Z}},t_{\mathcal{W}}}, \boldsymbol{\theta}^\star\right) > t\right) \\
&= \mathbb{P}\left(\widehat{\mathbf{Z}} \not\equiv_{\mathcal{Z}} \mathbf{z}^\star \text{ or } \widehat{\mathbf{W}} \not\equiv_{\mathcal{W}} \mathbf{w}^\star \text{ or } d^\infty\left(\widehat{\boldsymbol{\theta}}^{s_{\mathcal{Z}},t_{\mathcal{W}}}, \boldsymbol{\theta}^\star\right) > t\right) \\
&= \mathbb{P}\left(\widehat{\mathbf{Z}} \not\equiv_{\mathcal{Z}} \mathbf{z}^\star \text{ or } \widehat{\mathbf{W}} \not\equiv_{\mathcal{W}} \mathbf{w}^\star\right) + \mathbb{P}\left(\left\{d^\infty\left(\widehat{\boldsymbol{\theta}}^{s_{\mathcal{Z}},t_{\mathcal{W}}}, \boldsymbol{\theta}^\star\right) > t\right\} \setminus \left\{\widehat{\mathbf{Z}} \not\equiv_{\mathcal{Z}} \mathbf{z}^\star \text{ or } \widehat{\mathbf{W}} \not\equiv_{\mathcal{W}} \mathbf{w}^\star\right\}\right) \\
&= \mathbb{P}\left(\widehat{\mathbf{Z}} \not\equiv_{\mathcal{Z}} \mathbf{z}^\star \text{ or } \widehat{\mathbf{W}} \not\equiv_{\mathcal{W}} \mathbf{w}^\star\right) + \mathbb{P}\left(d^\infty\left(\widehat{\boldsymbol{\theta}}^{s_{\mathcal{Z}},t_{\mathcal{W}}}, \boldsymbol{\theta}^\star\right) > t, \widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^\star, \widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^\star\right) \\
&\leq \mathbb{P}\left(\widehat{\mathbf{Z}} \not\equiv_{\mathcal{Z}} \mathbf{z}^\star\right) + \mathbb{P}\left(\widehat{\mathbf{W}} \not\equiv_{\mathcal{W}} \mathbf{w}^\star\right) \\
&\quad + \mathbb{P}\left(d^\infty\left(\widehat{\boldsymbol{\theta}}^{s_{\mathcal{Z}},t_{\mathcal{W}}}, \boldsymbol{\theta}^\star\right) > t \middle| \widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^\star, \widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^\star\right) \mathbb{P}\left(\widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^\star, \widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^\star\right) \\
&\leq \mathbb{P}\left(\widehat{\mathbf{Z}} \not\equiv_{\mathcal{Z}} \mathbf{z}^\star\right) + \mathbb{P}\left(\widehat{\mathbf{W}} \not\equiv_{\mathcal{W}} \mathbf{w}^\star\right) + \mathbb{P}\left(d^\infty\left(\widehat{\boldsymbol{\theta}}^{s_{\mathcal{Z}},t_{\mathcal{W}}}, \boldsymbol{\theta}^\star\right) > t \middle| \widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^\star, \widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^\star\right)
\end{aligned}
$$

To complete the proof, we then need to bound from above the terms of this inequality. The two first terms are bounded using Proposition 1 in Appendix A.1, and the last term is bounded with Proposition 2 in Appendix A.2.

A.1 Concentration inequality on $\widehat{\mathbf{Z}}$

Let us first define the following events.

- There is at least one individual in each row class, denoted by

$$A_{g^\star} = \bigcap_{k=1}^{g^\star} \left\{ Z^\star_{+k} \neq 0 \right\}.$$

- Denoting D the random variable equal to the maximal distance between $\overline{X_{i\bullet}}$ and the center of the class of row i:

$$D = \max_{1 \leq k \leq g^\star} \sup_{\substack{1 \leq i \leq n \\ \text{with } z^\star_{i,k}=1}} \left| \overline{X_{i\bullet}} - \tau_k \right|,$$

  we also define:

$$A_{S_g} = \{2D < S_g < \delta_{\boldsymbol{\pi}^\star} - 2D\} \text{ and } A_{id} = A_{g^\star} \cap A_{S_g}.$$

Lemma 1 (Interesting event)

$$A_{id} \subset \{\widehat{g} = g^\star\} \cap \{\widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^\star\}$$

Proof On the event A_{S_g}, for any two rows $i \neq i' \in \{1, \ldots, n\}$, we have two possibilities:

- Either the rows i and i' are in the same class k, and then on A_{S_g}, we have:

$$\left| \overline{X_{i\bullet}} - \overline{X_{i'\bullet}} \right| \leq \left| \overline{X_{i\bullet}} - \tau_k \right| + \left| \overline{X_{i'\bullet}} - \tau_k \right| \leq 2D < S_g.$$

- Or row i is in the class k and row i' in the class $k' \neq k$, and on the event A_{S_g}, we have:

$$\begin{aligned}
\left| \overline{X_{i\bullet}} - \overline{X_{i'\bullet}} \right| &= \left| \overline{X_{i\bullet}} - \tau_{k'} - \left(\overline{X_{i'\bullet}} - \tau_{k'} \right) \right| \\
&\geq \left| \overline{X_{i\bullet}} - \tau_{k'} \right| - \left| \overline{X_{i'\bullet}} - \tau_{k'} \right| \\
&\geq \left| \overline{X_{i\bullet}} - \tau_{k'} \right| - D \\
&\geq \left| \tau_k - \tau_{k'} \right| - \left| \overline{X_{i\bullet}} - \tau_k \right| - D \\
&\geq \delta_{\boldsymbol{\pi}^\star} - 2D \\
&> S_g.
\end{aligned}$$

Therefore, $G_i = \overline{X_{(i)\bullet}} - \overline{X_{(i-1)\bullet}}$ is less than S_g if and only if both rows $(i-1)$ and (i) are in the same class. On A_{S_g}, the algorithm hence finds the true classification. Moreover, on $A_{g^\star}$, there is at least one row in each class, then the algorithm finds the true number of classes. As a conclusion, on A_{id}, both $\widehat{g} = g^\star$ and $\widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^\star$ are satisfied.

Lemma 2 (Concentration inequality for A_{id}) *Under Assumption (I) and if $S_g \in]0, \delta_{\boldsymbol{\pi}^\star}[$ and $S_m \in]0, \delta_{\boldsymbol{\rho}^\star}[$ for n, d large enough:*

$$\mathbb{P}\left(\overline{A_{id}}\right) \leq 2n \exp\left(-\frac{d}{2} \min(\delta_{\boldsymbol{\pi}^\star} - S_g, S_g)^2 \right) + g^\star \left(1 - \pi^\star_{\min}\right)^n$$

where $\overline{A_{id}}$ is the complementary of the event A_{id}.

Proof Using an union bound, we first obtain:

$$\mathbb{P}\left(\overline{A_{id}}\right) \leq \mathbb{P}\left(\overline{A_{g^\star}}\right) + \mathbb{P}\left(\overline{A_{S_g}}\right)$$

Now we bound from above each of these terms. Again with an union bound:

$$\begin{aligned}
\mathbb{P}\left(\overline{A_{g^\star}}\right) &= \mathbb{P}\left(\bigcup_{k=1}^{g^\star} \overline{\left\{Z^\star_{+k} \neq 0\right\}}\right) \\
&\leq \sum_{k=1}^{g^\star} \mathbb{P}\left(\overline{\left\{Z^\star_{+k} \neq 0\right\}}\right) \\
&\leq \sum_{k=1}^{g^\star} \mathbb{P}\left(Z^\star_{+k} = 0\right) \\
&\leq \sum_{k=1}^{g^\star} \prod_{i=1}^{n} \mathbb{P}\left(Z^\star_{i,k} = 0\right) \\
&\leq \sum_{k=1}^{g^\star} \prod_{i=1}^{n} (1 - \pi^\star_k) \\
&\leq \sum_{k=1}^{g^\star} \prod_{i=1}^{n} (1 - \pi^\star_{\min}) \\
&\leq g^\star \left(1 - \pi^\star_{\min}\right)^n ,
\end{aligned}$$

which gives the upper bound of the first term. Secondly:

$$\begin{aligned}
A_{S_g} &= \{2D < S_g < \delta_{\boldsymbol{\pi}^\star} - 2D\} \\
&= \{2D < S_g, 2D < \delta_{\boldsymbol{\pi}^\star} - S_g\} \\
&= \left\{D < \frac{1}{2}\min(\delta_{\boldsymbol{\pi}^\star} - S_g, S_g)\right\}.
\end{aligned}$$

Denoting $t = \min(\delta_{\boldsymbol{\pi}^\star} - S_g, S_g)$,

$$\begin{aligned}
\mathbb{P}\left(\overline{A_{S_g}}\right) &= \mathbb{P}\left(D \geq \frac{t}{2}\right) \\
&= \mathbb{P}\left(\bigcup_{k=1}^{g^\star} \bigcup_{i|z^\star_{i,k}=1} \left\{|\overline{X_{i\bullet}} - \tau_k| \geq \frac{t}{2}\right\}\right) \\
&\leq \sum_{k=1}^{g^\star} \sum_{i|z^\star_{i,k}=1} \mathbb{P}\left(|\overline{X_{i\bullet}} - \tau_k| \geq \frac{t}{2}\right).
\end{aligned}$$

Moreover for all $i \in \{1, \ldots, n\}$, given $z^\star_{i,k} = 1$, X_{i+} has a binomial distribution $\mathcal{B}in\left(d, \tau_k\right)$. The concentration properties of this distribution are then exploited through the Hoeffding inequality:

$$\mathbb{P}\left(|\overline{X_{i\bullet}} - \tau_k| \geq \frac{t}{2}\right) = \mathbb{P}\left(|X_{i+} - d\tau_k| \geq \frac{dt}{2}\right) \leq 2e^{-\frac{1}{2}dt^2}.$$

And as a conclusion, the bound of the second term is:

$$\mathbb{P}\left(\overline{A_{S_g}}\right) \leq \sum_{k=1}^{g^\star} \sum_{i|z^\star_{i,k}=1} 2e^{-\frac{1}{2}dt^2} = 2ne^{-\frac{1}{2}dt^2}.$$

With these two lemmas, the following proposition is obtained

Proposition 1 *Under Assumption (I) and if* $S_g \in]0, \delta_{\boldsymbol{\pi}^\star}[$ *and* $S_m \in]0, \delta_{\boldsymbol{\rho}^\star}[$ *for* n, d *large enough:*

$$\mathbb{P}\left(\widehat{\mathbf{Z}} \not\equiv_{\mathcal{Z}} \mathbf{z}^\star\right) \leq 2n \exp\left(-\frac{d}{2}\min(\delta_{\boldsymbol{\pi}^\star} - S_g, S_g)^2\right) + g^\star\left(1 - \pi^\star_{\min}\right)^n.$$

$$\mathbb{P}\left(\widehat{\mathbf{W}} \not\equiv_{\mathcal{W}} \mathbf{w}^\star\right) \leq 2d \exp\left(-\frac{n}{2}\min(\delta_{\boldsymbol{\rho}^\star} - S_m, S_m)^2\right) + m^\star\left(1 - \rho^\star_{\min}\right)^d.$$

Proof Lemma 1 tells that whenever the event A_{id} is satisfied, then both true number of row classes and their true classification are obtained. Lemma 2 provides an upper bound of $\mathbb{P}\left(\overline{A_{id}}\right)$. From these lemmas, it is directly deduced that:

$$\begin{aligned}\mathbb{P}\left(\widehat{\mathbf{Z}} \not\equiv_{\mathcal{Z}} \mathbf{z}^\star\right) &\leq \mathbb{P}\left(\overline{A_{id}}\right)\\ &\leq 2n \exp\left(-\frac{d}{2}\min(\delta_{\boldsymbol{\pi}^\star} - S_g, S_g)^2\right) + g^\star\left(1 - \pi^\star_{\min}\right)^n,\end{aligned}$$

which is Proposition 1.

A.2 Concentration inequality on $d^\infty\left(\widehat{\boldsymbol{\theta}}^{s_{\mathcal{Z}}, t_{\mathcal{W}}}, \boldsymbol{\theta}^\star\right) > t$

In this part, the proof of the inequality on $d^\infty\left(\widehat{\boldsymbol{\theta}}^{s_{\mathcal{Z}}, t_{\mathcal{W}}}, \boldsymbol{\theta}^\star\right) > t$ is detailled.

Proposition 2 *For all* $t > 0$*, we have:*

$$\begin{aligned}&\mathbb{P}\left(d^\infty\left(\widehat{\boldsymbol{\theta}}^{s_{\mathcal{Z}}, t_{\mathcal{W}}}, \boldsymbol{\theta}^\star\right) > t \middle| \widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^\star, \widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^\star\right)\\ &\leq 2g^\star m^\star\left[1 - \pi^\star_{\min}\rho^\star_{\min}\left(1 - e^{-2t^2}\right)\right]^{nd} + 2g^\star e^{-2nt^2} + 2m^\star e^{-2dt^2}\end{aligned}$$

The proof consists in obtaining three bounds: one for each parameter. The inequalities on $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ are an application of the Hoeffding inequality and are similar to [8] for the row class proportions. To obtain the inequality for $\boldsymbol{\alpha}$, it is necessary to study the conditional probability, given the true partition $(\mathbf{z}^\star, \mathbf{w}^\star)$. Apart from the problem of two asymptotic behaviors, the proof is similar to [8].

In the sequel, and for ease of reading, we remove the superscripts $s_{\mathcal{Z}}$ and $t_{\mathcal{W}}$. Therefore, for all $t > 0$:

$$\begin{aligned}&\mathbb{P}\left(d^\infty\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}^\star\right) > t \middle| \widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^\star, \widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^\star\right)\\ &= \mathbb{P}\left(\max\left(\left\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^\star\right\|_\infty, \left\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star\right\|_\infty, \left\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\star\right\|_\infty\right) > t \middle| \widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^\star, \widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^\star\right)\\ &\leq \mathbb{P}\left(\left\|\widehat{\boldsymbol{\pi}} - \boldsymbol{\pi}^\star\right\|_\infty > t \middle| \widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^\star, \widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^\star\right) + \mathbb{P}\left(\left\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}^\star\right\|_\infty > t \middle| \widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^\star, \widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^\star\right)\\ &\quad + \mathbb{P}\left(\left\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^\star\right\|_\infty > t \middle| \widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^\star, \widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^\star\right)\\ &\leq \sum_{k=1}^{g^\star} \mathbb{P}\left(\left|\widehat{\pi_k} - \pi^\star_k\right| > t \middle| \widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^\star, \widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^\star\right) + \sum_{\ell=1}^{m^\star} \mathbb{P}\left(\left|\widehat{\rho_\ell} - \rho^\star_\ell\right| > t \middle| \widehat{\mathbf{Z}} \equiv_{\mathcal{Z}} \mathbf{z}^\star, \widehat{\mathbf{W}} \equiv_{\mathcal{W}} \mathbf{w}^\star\right)\end{aligned}$$

$$+\sum_{k=1}^{g^\star}\sum_{\ell=1}^{m^\star}\mathbb{P}\left(|\widehat{\alpha_{k\ell}}-\alpha_{k\ell}^\star|>t\left|\widehat{\mathbf{Z}}\equiv_{\mathcal{Z}}\mathbf{z}^\star,\widehat{\mathbf{W}}\equiv_{\mathcal{W}}\mathbf{w}^\star\right.\right).$$

The upper bounds of the first and second terms are the same as [8]; only the last term is different. For $\widehat{\alpha_{k\ell}}$, first note that when $\widehat{\mathbf{Z}}\equiv_{\mathcal{Z}}\mathbf{z}^\star$ and $\widehat{\mathbf{W}}\equiv_{\mathcal{W}}\mathbf{w}^\star$

$$\widehat{\alpha_{k\ell}}=\widetilde{\alpha_{k\ell}}=\frac{1}{z_{+k}^\star w_{+\ell}^\star}\sum_{(i,j)\left|z_{i,k}^\star w_{j,\ell}^\star=1\right.}X_{ij}$$

and given $(\mathbf{z}^\star,\mathbf{w}^\star)$, the Hoeffding inequality gives for all $t>0$:

$$\begin{aligned}\mathbb{P}\left(|\widehat{\alpha_{k\ell}}-\alpha_{k\ell}^\star|>t\left|\widehat{\mathbf{Z}}\equiv_{\mathcal{Z}}\mathbf{z}^\star,\widehat{\mathbf{W}}\equiv_{\mathcal{W}}\mathbf{w}^\star\right.\right)&=\mathbb{P}\left(|\widehat{\alpha_{k\ell}}-\alpha_{k\ell}^\star|>t\left|\widehat{\mathbf{Z}}\equiv_{\mathcal{Z}}\mathbf{z}^\star,\widehat{\mathbf{W}}\equiv_{\mathcal{W}}\mathbf{w}^\star\right.\right)\\&\leq\mathbb{P}\left(|\widehat{\alpha_{k\ell}}-\alpha_{k\ell}^\star|>t\right)\\&\leq\mathbb{E}_{\mathbf{Z}^\star,\mathbf{W}^\star}\left[\mathbb{P}\left(|\widehat{\alpha_{k\ell}}-\alpha_{k\ell}^\star|>t|\,\mathbf{Z}^\star,\mathbf{W}^\star\right)\right]\\&\leq\mathbb{E}_{\mathbf{Z}^\star,\mathbf{W}^\star}\left[2e^{-2Z_{+k}^\star W_{+\ell}^\star t^2}\right].\end{aligned}$$

But, as $Z_{+k}^\star=\sum_{i=1}^n Z_{i,k}^\star$ and $W_{+\ell}^\star=\sum_{j=1}^d W_{j\ell}^\star$ and the variables are independents, the expectation is:

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}^\star,\mathbf{W}^\star}\left[2e^{-2Z_{+k}^\star W_{+\ell}^\star t^2}\right]&=2\mathbb{E}_{\mathbf{Z}^\star,\mathbf{W}^\star}\left[\exp\left(-2\sum_{i=1}^n Z_{i,k}^\star\sum_{j=1}^d W_{j\ell}^\star t^2\right)\right]\\&=2\mathbb{E}_{\mathbf{Z}^\star,\mathbf{W}^\star}\left[\prod_{i=1}^n\exp\left(-2Z_{i,k}^\star\sum_{j=1}^d W_{j\ell}^\star t^2\right)\right]\\&=2\prod_{i=1}^n\mathbb{E}_{\mathbf{Z}^\star,\mathbf{W}^\star}\left[\exp\left(-2Z_{i,k}^\star\sum_{j=1}^d W_{j\ell}^\star t^2\right)\right]\\&=2\prod_{i=1}^n\prod_{j=1}^d\mathbb{E}_{\mathbf{Z}^\star,\mathbf{W}^\star}\left[\exp\left(-2Z_{i,k}^\star W_{j\ell}^\star t^2\right)\right]\end{aligned}$$

Since, for all $i\in\{1,\dots,n\}$, $j\in\{1,\dots,d\}$, $k\in\{1,\dots,g^\star\}$ and $\ell\in\{1,\dots,m^\star\}$, the variable $Z_{i,k}^\star W_{j\ell}^\star$ is a Bernoulli of parameter $\pi_k^\star\rho_\ell^\star$, then:

$$\begin{aligned}\mathbb{E}_{\mathbf{Z}^\star,\mathbf{W}^\star}\left[\exp\left(-2z_{i,k}^\star w_{j,\ell}^\star t^2\right)\right]&=\pi_k^\star\rho_\ell^\star\exp\left(-2\times1\times t^2\right)+\left(1-\pi_k^\star\rho_\ell^\star\right)\exp\left(-2\times0\times t^2\right)\\&=\pi_k^\star\rho_\ell^\star\exp\left(-2t^2\right)+1-\pi_k^\star\rho_\ell^\star\\&=1-\pi_k^\star\rho_\ell^\star\left(1-e^{-2t^2}\right).\end{aligned}$$

Finally, the inequality is:

$$\begin{aligned}\mathbb{P}\left(|\widehat{\alpha_{k\ell}}-\alpha_{k\ell}^\star|>t\left|\widehat{\mathbf{Z}}\equiv_{\mathcal{Z}}\mathbf{z}^\star,\widehat{\mathbf{W}}\equiv_{\mathcal{W}}\mathbf{w}^\star\right.\right)&\leq\left[1-\pi_k^\star\rho_\ell^\star\left(1-e^{-2t^2}\right)\right]^{nd}\\&\leq\left[1-\pi_{\min}^\star\rho_{\min}^\star\left(1-e^{-2t^2}\right)\right]^{nd}\end{aligned}$$

and the result is obtained by the sum on each cluster.

B Proof of Theorem 2: consistency

The proof is based on Theorem 1, as $n \to +\infty$ and $d \to +\infty$ and by the assumption (I), the following limits are obtained for every $t > 0$:

$$g^\star \left(1 - \pi^\star_{\min}\right)^n + m^\star \left(1 - \rho^\star_{\min}\right)^d \underset{n,d\to+\infty}{\longrightarrow} 0$$
$$\text{and} \quad g^\star e^{-2nt^2} + m^\star e^{-2dt^2} \underset{n,d\to+\infty}{\longrightarrow} 0.$$

For the same reasons, as soon as t is positive, $0 < \pi^\star_{\min}\rho^\star_{\min}\left(1 - e^{-2t^2}\right) < 1$ and

$$g^\star m^\star \left[1 - \pi^\star_{\min}\rho^\star_{\min}\left(1 - e^{-2t^2}\right)\right]^{nd} \underset{n,d\to+\infty}{\longrightarrow} 0.$$

For the last terms, the assumption (TA) gives that

$$\frac{n}{\log(d)} \underset{n,d\to+\infty}{\longrightarrow} +\infty \quad \text{and} \quad \frac{d}{\log(n)} \underset{n,d\to+\infty}{\longrightarrow} +\infty$$

which results in

$$ne^{-\frac{1}{2}d\delta^2_{\boldsymbol{\pi}^\star}} + de^{-\frac{1}{2}n\delta^2_{\boldsymbol{\rho}^\star}} \underset{n,d\to+\infty}{\longrightarrow} 0$$

and, as $S_g^{n,d}$ tends to zero,

$$ne^{-d\frac{\left(\delta_{\boldsymbol{\pi}^\star} - S_g^{n,d}\right)^2}{2}} \underset{n,d\to+\infty}{\longrightarrow} 0.$$

Thanks the same Assumption (TA), there exists a positive constant $C > \sqrt{2}$ such that for n and d large enough

$$S_g^{n,d}\sqrt{\frac{d}{\log n}} > C \Longrightarrow \frac{S_g^{n,d}}{\sqrt{2}}\sqrt{\frac{d}{\log n}} > \frac{C}{\sqrt{2}} > 1$$

$$\begin{aligned} ne^{-d\frac{S_g^{n,d\,2}}{2}} &= \exp\left[\log n - d\frac{S_g^{n,d\,2}}{2}\right] \\ &= \exp\left[\log n\left(1 - \left(\sqrt{\frac{d}{\log n}}\frac{S_g^{n,d}}{\sqrt{2}}\right)^2\right)\right] \\ &\leq \exp\left[\log n \underbrace{\left(1 - \frac{C}{\sqrt{2}}\right)}_{<0}\right] \\ &\underset{n,d\to+\infty}{\longrightarrow} 0. \end{aligned}$$

C Proof of Theorem 3: consistency in the sparse case

First, if $\pi^{\star\,n,d}_{\min}$ tends to zero (and as $g^{\star n,d} \leq 1/\pi^{\star\,n,d}_{\min}$), the series expansion of $t \mapsto \log(1-t)$ is used

$$g^{\star n,d}\left(1 - \pi^{\star\,n,d}_{\min}\right)^n = \exp\left[\log g^{\star n,d} + n\log\left(1 - \pi^{\star\,n,d}_{\min}\right)\right]$$

$$
\begin{aligned}
&\leq \exp\left\{n\left[\frac{1}{n}\log\left(\frac{1}{\pi_{\min}^{\star\, n,d}}\right)+\log\left(1-\pi_{\min}^{\star\, n,d}\right)\right]\right\}\\
&\leq \exp\left\{n\left[-\frac{1}{n}\log\left(\pi_{\min}^{\star\, n,d}\right)-\pi_{\min}^{\star\, n,d}+o\left(\pi_{\min}^{\star\, n,d}\right)\right]\right\}\\
&\leq \exp\left\{n\left[-\pi_{\min}^{\star\, n,d}+o\left(\pi_{\min}^{\star\, n,d}\right)\right]\right\}\\
&\leq \exp\left[-n\pi_{\min}^{\star\, n,d}+o\left(n\pi_{\min}^{\star\, n,d}\right)\right]
\end{aligned}
$$

and, by the assumption (MA.2), $n\pi_{\min}^{\star\, n,d}$ tends to infinity and

$$
g^{\star n,d}\left(1-\pi_{\min}^{\star\, n,d}\right)^{n} \underset{n,d\to+\infty}{\longrightarrow} 0.
$$

For the same reasons, for all $t>0$

$$
\begin{aligned}
m^{\star n,d}\left(1-\rho_{\min}^{\star\, n,d}\right)^{n} &\underset{n,d\to+\infty}{\longrightarrow} 0\\
\text{and } g^{\star}m^{\star}\left[1-\pi_{\min}^{\star}\rho_{\min}^{\star}\left(1-e^{-2t^2}\right)\right]^{nd} &\underset{n,d\to+\infty}{\longrightarrow} 0.
\end{aligned}
$$

Moreover, as

$$
\begin{aligned}
g^{\star n,d}e^{-2nt^2} &= \exp\left(\log g^{\star n,d}-2nt^2\right)\\
&\leq \exp\left\{-n\left[\frac{1}{n}\log\left(\pi_{\min}^{\star\, n,d}\right)+2t^2\right]\right\}\\
&\leq \exp\left[-n2t^2+o\left(n\right)\right].
\end{aligned}
$$

then

$$
g^{\star n,d}e^{-2nt^2} \underset{n,d\to+\infty}{\longrightarrow} 0 \quad \text{and} \quad m^{\star n,d}e^{-2dt^2} \underset{n,d\to+\infty}{\longrightarrow} 0.
$$

Finally, the assumption (MA.1) implies that there exists a positive constant $C>2$ such that for n and d large enough

$$
\begin{aligned}
\frac{\delta_{\boldsymbol{\pi}^\star}^{n,d}}{S_g^{n,d}}>C &\Leftrightarrow \delta_{\boldsymbol{\pi}^\star}^{n,d}>S_g^{n,d}C\\
&\Leftrightarrow \delta_{\boldsymbol{\pi}^\star}^{n,d}>S_g^{n,d}(C-1)+S_g^{n,d}\\
&\Leftrightarrow \delta_{\boldsymbol{\pi}^\star}^{n,d}-S_g^{n,d}>S_g^{n,d}\underbrace{(C-1)}_{>1}>S_g^{n,d}
\end{aligned}
$$

and, for n and d large enough, $\min\left(\delta_{\boldsymbol{\pi}^\star}^{n,d}-S_g^{n,d},S_g^{n,d}\right)$ is $S_g^{n,d}$ and the assumptions (I) and (TA) allow to conclude.

D Supplementary Material

In this supplementary material, additional figures from the experiments in the section 6 are presented.

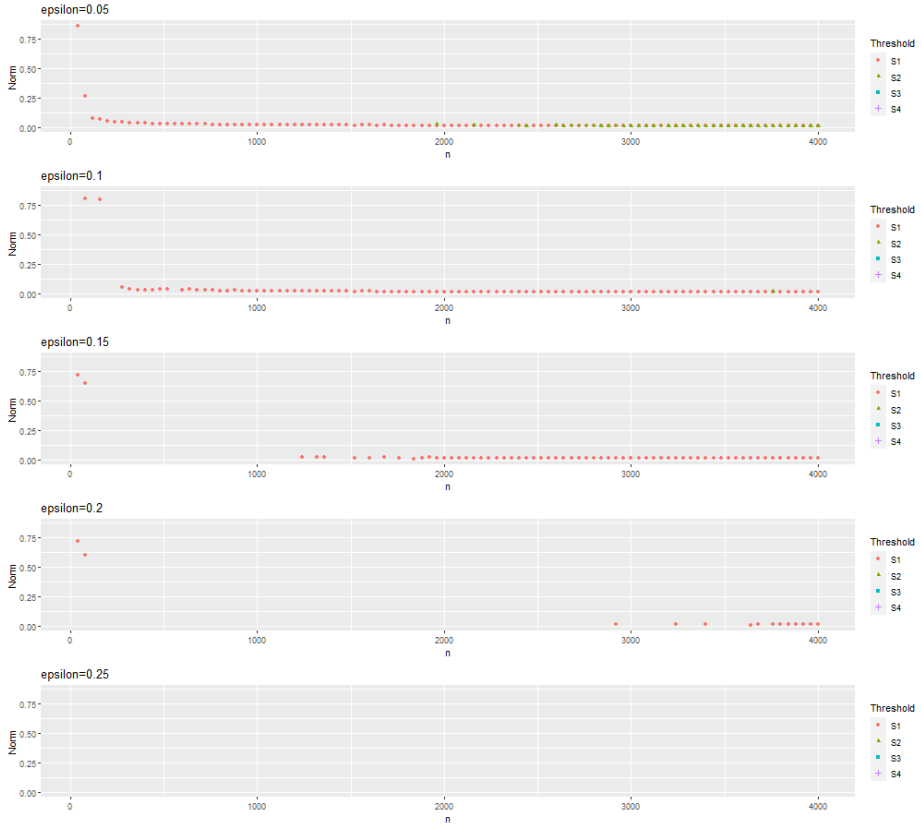

Fig. 7 Average estimate of the distance $d^{\infty}\left(\boldsymbol{\theta}^{\star}, \widehat{\boldsymbol{\theta}}\right)$ between the true parameters and the estimated parameters following ε (rows) in the arithmetic case: for each graphic, the threshold is represented by different symbols (● for S_1, ▲ for $S_2^{n,d}$, ■ for $S_3^{n,d}$ and + for $S_3^{n,d}$) and the size for the number of finite values used; the number of rows n varies between 40 and 4000 and d is supposed equal n.

References

1. Barbillon, P., Donnet, S., Lazega, E., Bar-Hen, A.: Stochastic block models for multiplex networks: an application to a multilevel network of researchers. Journal of the Royal Statistical Society: Series A (Statistics in Society) **180**(1), 295–314 (2017)
2. Bennett, J., Lanning, S.: The netflix prize. In: Proceedings of KDD cup and workshop, vol. 2007, p. 35 (2007)
3. Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE transactions on pattern analysis and machine intelligence **22**(7), 719–725 (2000)
4. Bouveyron, C., Latouche, P., Zreik, R.: The stochastic topic block model for the clustering of vertices in networks with textual edges. Statistics and Computing **28**(1), 11–31 (2018)
5. Brault, V., Keribin, C., Mariadassou, M., et al.: Consistency and asymptotic normality of latent block model estimators. Electronic journal of statistics **14**(1), 1234–1268 (2020)
6. Celeux, G., Chauveau, D., Diebolt, J.: On Stochastic Versions of the EM Algorithm. Rapport de recherche RR-2514, INRIA (1995). URL http://hal.inria.fr/inria-00074164

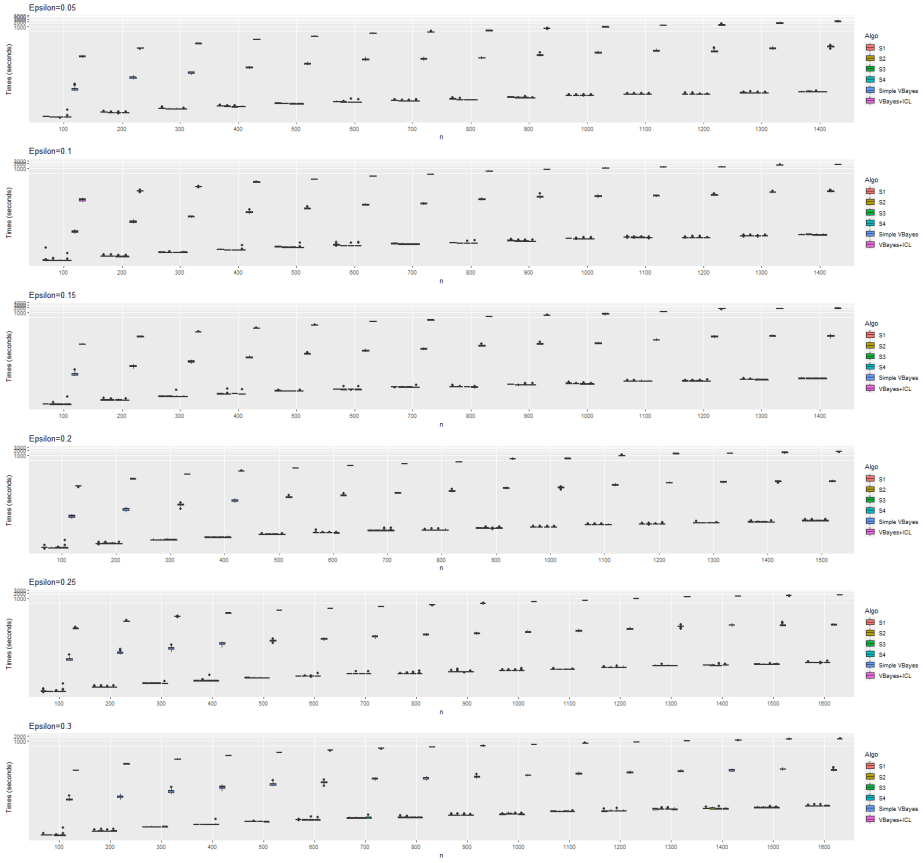

Fig. 8 Boxplots of the computing times (in seconds; logarithmic scale) for each procedure (colour) in function of the numbers of rows and columns ($n = d$) for each ε (in rows) over the 200 simulations for the balanced case.

7. Celisse, A., Daudin, J.J., Pierre, L., et al.: Consistency of maximum-likelihood and variational estimators in the stochastic block model. Electronic Journal of Statistics **6**, 1847–1899 (2012)
8. Channarond, A., Daudin, J.J., Robin, S., et al.: Classification and estimation in the stochastic blockmodel based on the empirical degrees. Electronic Journal of Statistics **6**, 2574–2601 (2012)
9. Govaert, G.: Classification croisée. Thèse d'état, Université Pierre et Marie Curie (1983)
10. Govaert, G., Nadif, M.: Clustering with block mixture models. Pattern Recognition **36**, 463–473 (2003)
11. Hartigan, J.A.: Clustering Algorithms, 99th edn. John Wiley & Sons, Inc., New York, NY, USA (1975)
12. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bitter, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M.: Gene-expression profiles in hereditary breast cancer. New Eng. J. Med. **344**, 539–548 (2001)
13. Iannario, M.: On the identifiability of a mixture model for ordinal data. Metron **68**(1), 87–94 (2010)
14. Jagalur, M., Pal, C., Learned-Miller, E., Zoeller, R.T., Kulp, D.: Analyzing in situ gene expression in the mouse brain with image registration, feature extraction and block clustering. BMC Bioinformatics **8**(Suppl 10), S5 (2007)

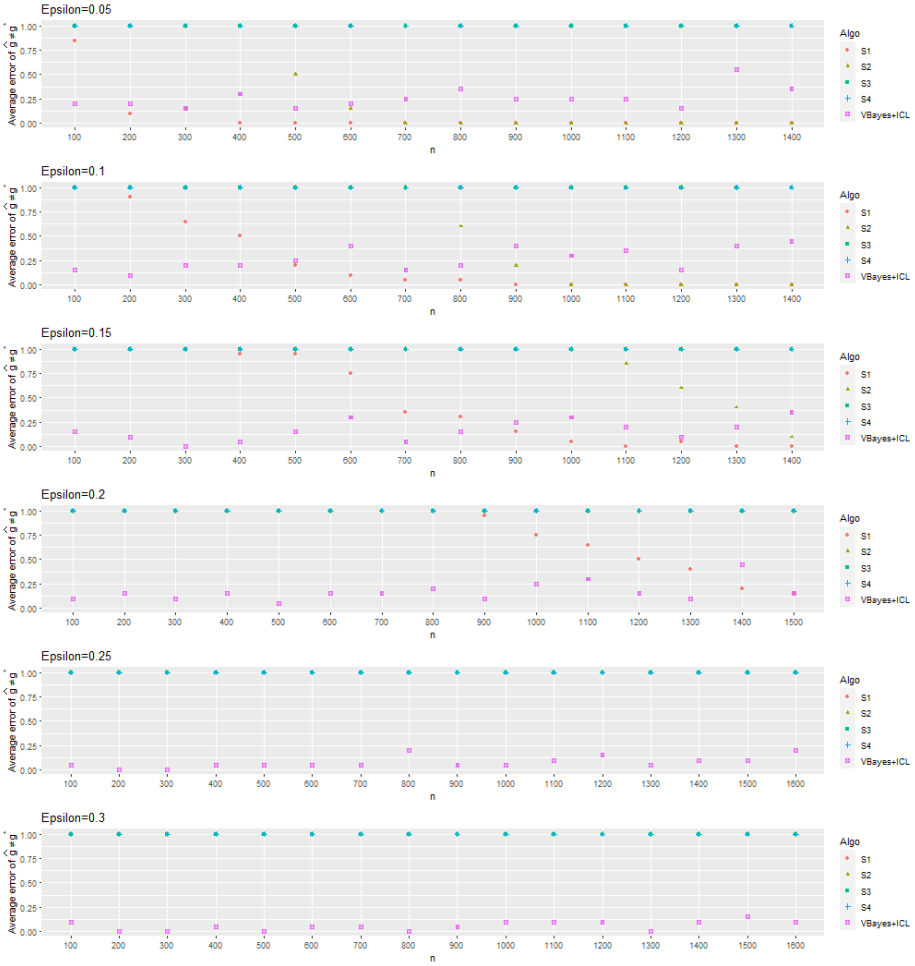

Fig. 9 Quality of estimations averaged over the 20 matrices for each ε (in rows) for each procedure (colour) in function of the numbers of rows and columns ($n = d$) for the balanced case.

15. Keribin, C.: Cluster or co-cluster the nodes of oriented graphs? Journal de la société française de statistique **162**(1), 46–69 (2021)
16. Keribin, C., Brault, V., Celeux, G., Govaert, G.: Model selection for the binary latent block model. In: 20th International Conference on Computational Statistics. Limassol, Chypre (2012). URL http://hal.inria.fr/hal-00778145
17. Keribin, C., Brault, V., Celeux, G., Govaert, G.: Estimation and selection for the latent block model on categorical data. Statistics and Computing pp. 1–16 (2014). DOI 10.1007/s11222-014-9472-2. URL http://dx.doi.org/10.1007/s11222-014-9472-2
18. Laclau, C., Redko, I., Matei, B., Bennani, Y., Brault, V.: Co-clustering through optimal transport. In: International Conference on Machine Learning, pp. 1955–1964. PMLR (2017)
19. Mariadassou, M., Matias, C., et al.: Convergence of the groups posterior distribution in latent or stochastic block models. Bernoulli **21**(1), 537–573 (2015)

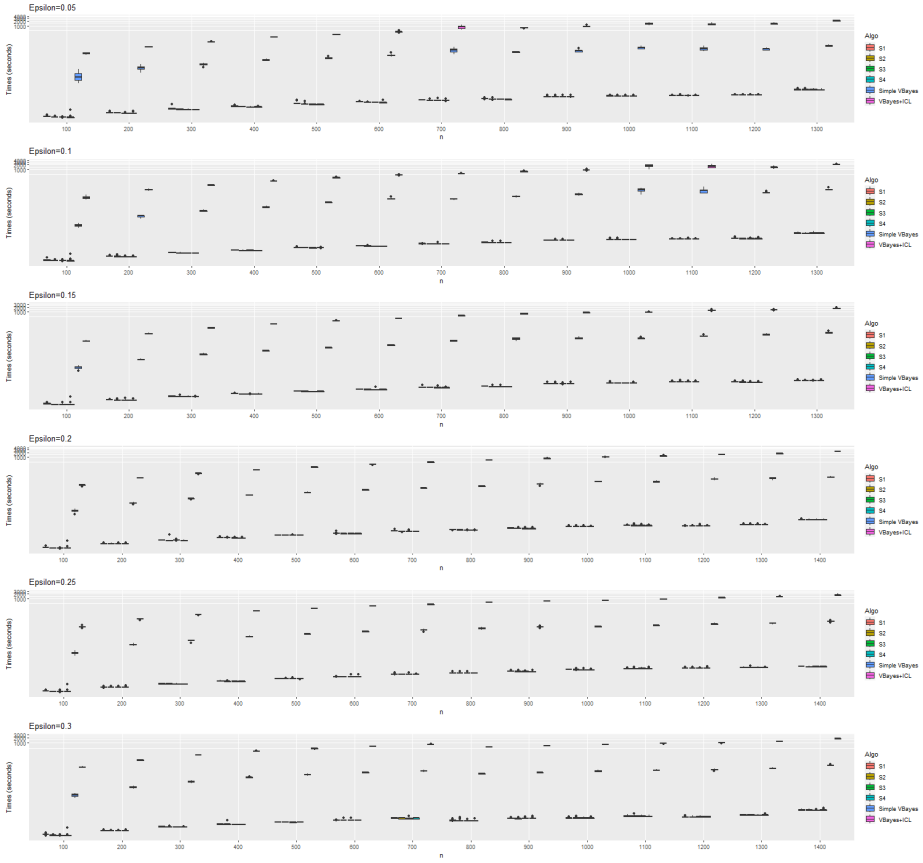

Fig. 10 Boxplots of the computing times (in seconds; logarithmic scale) for each procedure (colour) in function of the numbers of rows and columns ($n = d$) for each ε (in rows) over the 200 simulations for the arithmetic case.

20. Maugis, C., Michel, B.: A non asymptotic penalized criterion for gaussian mixture model selection. ESAIM: Probability and Statistics **15**, 41–68 (2011)
21. Mersmann, O.: microbenchmark: Accurate Timing Functions (2021). URL https://CRAN.R-project.org/package=microbenchmark. R package version 1.4.9
22. Shan, H., Banerjee, A.: Bayesian co-clustering. In: Eighth IEEE International Conference on Data Mining, 2008. ICDM'08, pp. 530–539 (2008)
23. Singh Bhatia, P., Iovleff, S., Govaert, G.: blockcluster: An R package for model-based co-clustering. Journal of Statistical Software **76**(9), 1–24 (2017). DOI 10.18637/jss.v076.i09
24. Tabouy, T., Barbillon, P., Chiquet, J.: Variational inference for stochastic block models from sampled data. Journal of the American Statistical Association **115**(529), 455–466 (2020)
25. Wyse, J., Friel, N.: Block clustering with collapsed latent block models. Statistics and Computing pp. 1–14 (2010)

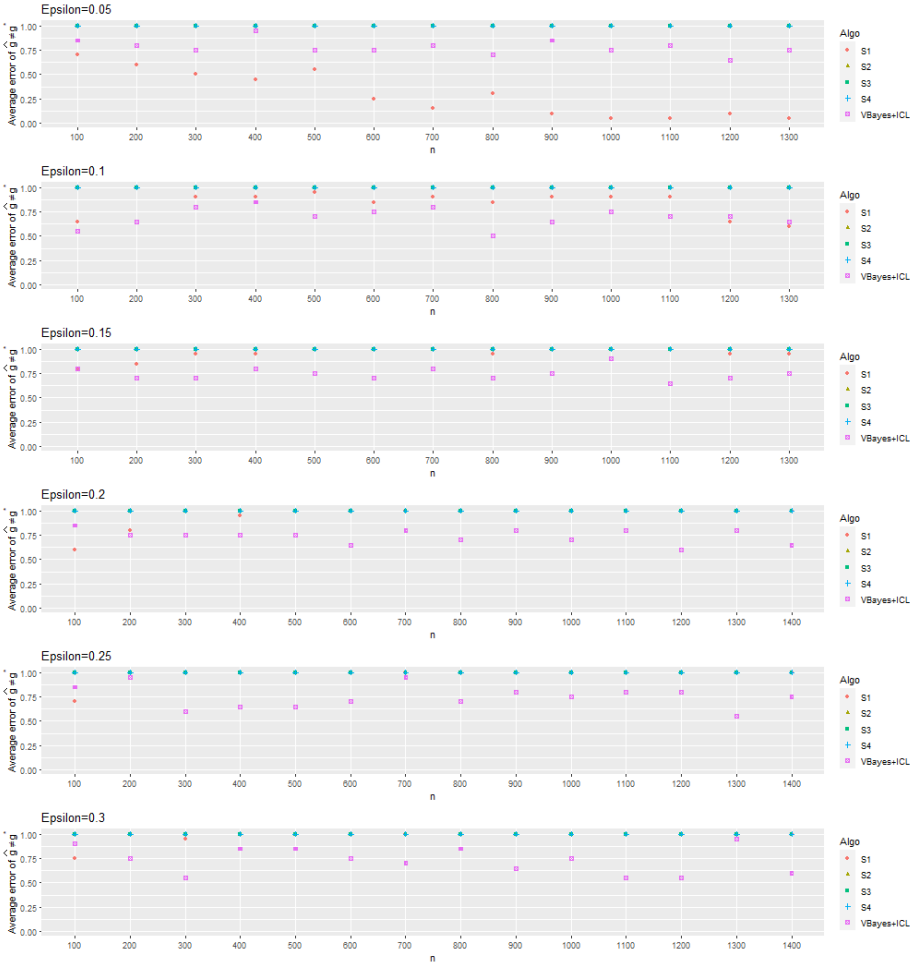

Fig. 11 Quality of estimations averaged over the 20 matrices for each ε (in rows) for each procedure (colour) in function of the numbers of rows and columns ($n = d$) for the arithmetic case.