



HAL
open science

Noise-free Latent Block Model for High Dimensional Data

Charlotte Laclau, Vincent Brault

► **To cite this version:**

Charlotte Laclau, Vincent Brault. Noise-free Latent Block Model for High Dimensional Data. Data Mining and Knowledge Discovery, In press, 10.1007/s10618-018-0597-3 . hal-01685777v1

HAL Id: hal-01685777

<https://hal.science/hal-01685777v1>

Submitted on 16 Jan 2018 (v1), last revised 29 Oct 2018 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Noise-free Latent Block Model for High Dimensional Data

Charlotte Laclau · Vincent Brault

Received: date / Accepted: date

Abstract Co-clustering is known to be a very powerful and efficient approach in unsupervised learning because of its ability to partition data based on both modes of a dataset. However, in high-dimensional context co-clustering methods may fail to provide a meaningful result due to the presence of noisy and/or irrelevant features. In this paper, we tackle this issue by proposing a novel co-clustering model which assumes the existence of a noise cluster, that contains all irrelevant features. An expectation-maximization (EM)-based algorithm is derived for this task, where the automatic variable selection as well as the joint clustering of objects and variables are achieved via a Bayesian framework. Experimental results on synthetic datasets show the efficiency of our model in the context of high-dimensional noisy data. Finally, we highlight the interest of the approach on two real datasets which goal is to study genetic diversity across the world.

1 Introduction

Clustering, which aims to partition data into groups (clusters) of similar objects, has a wide range of applications including information retrieval, bioinformatics, pattern recognition and image analysis. In many of these cases, and particularly in the case of high dimensional data, a significant proportion of the variables is not providing any relevant information. In this situation, attempting to learn while including this part of the data, which can be qualified as noises, strongly disrupts the clustering algorithms and can mask the existing structure. Despite the need for a theoretical and practical framework, where one can consider only the subset of relevant variables for partitioning the data, relatively little work has been proposed so far in this direction. Noise management remains a complex issue, which raises the question of its definition and, consequently, its modelling. Indeed, in unsupervised learning,

Charlotte Laclau
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG,
38000 Grenoble, France
E-mail: charlotte.laclau@univ-grenoble-alpes.fr

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

where there is no labels to guide this search, one can define the notion of noise in many different ways. For example, in genetic data analysis, the common approach to handle noise is to eliminate genes with a low variance, that is, with a homogeneous degree of expression across all individuals. This pre-processing step (or filtering) relies on the intrinsic properties of the variables to determine their relevance but completely ignores the possible interactions between the variables and the structure, and between the variables themselves. Other approaches have sought to weight the variables according to their discriminating power and to learn clusters simultaneously. These approaches are generally more efficient than the so-called filtering methods, but the weight calculation for each variable induces an algorithmic complexity which makes them impracticable in the context of high-dimensional data.

In the context of outlier detection, Dave (1991, 1993) and more recently Ben-David and Haghtalab (2014) proposed a novel formalism that allows to transform clustering algorithms, based on the notion of centroids (and hence distance), into robust algorithms for noisy objects. Their approach is based on the interesting concept of the existence of a potential noise cluster, i.e., a cluster that contains the set of noisy objects, without specifying or constraining them to be similar. Despite encouraging results, this type of approach has not been extended to the problem of noise variables and has remained limited to the framework of some metric approaches. To overcome these limitation, and to address the problem of noise cluster from the variable perspective, we propose to exploit the framework of co-clustering (or bi-clustering) (Hartigan, 1972; Mirkin, 1996), which aims to simultaneously cluster the sets of objects and variables into homogeneous blocks (or co-clusters). These blocks consist of subset of the data matrix composed of objects and variables strongly linked. In some sense, one can see co-clustering as a local variable selection approach which benefits from the knowledge of the object partition. Furthermore, in order to cover different aspects of the definition of noise, we propose a probabilistic approach, allowing a more flexible modelling of the noise. To this end, we assume that the data are generated according to a mixture of the product of two probability densities (Law et al, 2004): parameters of the first one are independent from the structure, while parameters of the second one are specific to each blocks.

As a result, our contributions can be summarized as follows:

- We design a novel probabilistic co-clustering model which relies on the assumption that there exists a variable cluster that contains only irrelevant features, referred to as the “noise” variable cluster, in the following. All the variables belonging to this cluster are assumed to be drawn from a probability distribution that does not depend on the structure of the data into groups. For the remaining data, we extract relevant partitions of both the features and the objects, which provide a clear interpretation of the data structure.
- The optimization of the model is carried out by a Variational EM-based (VEM) algorithm. In addition, we propose a Bayesian version and introduce Gibbs sampling on the different parameters to overcome the problem of vanishing clusters.
- In unsupervised learning, the estimation of the number of clusters is also a key point. To this end, we propose to adapt a model selection criteria, namely the Integrated Completed Likelihood (ICL) (Biernacki et al, 2000; Keribin et al, 2014).

For all the contributions mentioned above, we provide strong theoretical guarantees. Finally, we extensively validate our approach over synthetic datasets and study the relevance of the proposed model over a real dataset on genetic diversity.

The remainder of this paper is organized as follows. First, we give a brief overview of related works that tackle the problem of clustering and feature selection using model-based approaches in Section 2. We proceed by formally defining an appropriate latent block model

for simultaneous co-clustering and feature selection, named Noise-Free Latent Block model (NFLB) in Section 3. Section 4 describes three different optimization procedures, all derived from the Variational EM algorithm (VEM). In Section 5 we adapt the ICL criteria to the NFLB model and give an explicit formulation for its calculation. In Section 6, we provide a detailed theoretical analysis of the proposed model. Then, Section 7 illustrates the ability of our approach to identify irrelevant features on synthetic and one genetic dataset. We conclude this paper by summarizing the contributions and discussing possible perspectives of this work in Section 8.

2 Preliminary knowledge

This section gives an overview of clustering approaches which aim at performing either object or feature selection in the model-based framework. Also, we provide a description of the Latent Block Model (LBM), a probabilistic model for co-clustering which we use to develop our approach.

2.1 Clustering on noisy data

Noise-cluster and robust clustering. Several directions have been taken in developing robust clustering algorithms cluster (Dave, 1991, 1993; Cuesta-Albertos et al, 1997; Ester et al, 1996; Ben-David and Haghtalab, 2014). The concept of noise cluster was first introduced by Dave (1991, 1993) in a fuzzy centroid-based setting where assume the existence of a fictitious cluster which prototype (or center) is equidistant from all the objects of the dataset. Then objects which are further away from all other centers are assigned to this center and denoted as part of the noise cluster. Cuesta-Albertos et al (1997) proposed to use the concept of *trimming* in order to determine the subset of objects, of a predefined fixed size, whose removal leads to the maximum improvement of the objective function of k-means, and therefore of the clustering quality. They also extend the concept of trimming and provide strong theoretical guarantees for this family of approaches using influence functions (García-Escudero et al, 2008). Another family of approaches was explored by Ester et al (1996) to introduce the notion of noise cluster. They developed a density-based approach for clustering noisy objects where all objects in the dataset which belong to the sparse regions are assigned to the noise cluster. This concept of noise cluster was further extended and theoretically formalized by Ben-David and Haghtalab (2014), who proposed to generalize all these approaches to any prototype-based algorithms. For more references on robust clustering algorithms, we refer the reader to the survey of García-Escudero et al (2010).

The aforementioned approaches focus on the problem of noisy objects, or outliers, which generally represent a small proportion of the data. In this work, we aim to deal with high-dimensional data, therefore our goal is to target noisy features and to extend the concept of noisy cluster to the problem of feature selection. In addition, we propose to explore this concept in the probabilistic framework of mixture models, allowing more flexibility regarding the modelling of noise.

Model-based feature selection. Several work have been conducted in order to perform feature selection using the framework of mixture models. These approaches can be broadly divided into three categories The first proposes to cast the problem as a model selection problem, and includes the work of Raftery and Dean (2006), Maugis et al (2009) and Celeux et al

(2011). For instance, Raftery and Dean (2006) proposed to divide the initial set of variables into two groups: a set containing relevant features and a set containing irrelevant features, that are assumed to be dependent according to a linear relationship. Models in competition are compared based on their integrated log-likelihood and aim to maximize a two terms criterion, where the first term corresponds to the BIC resulting from the clustering on the set of relevant features while the second term relies on the linear regression of irrelevant variables on the set of relevant variables.

Another attempt to perform feature selection and clustering is to progressively add sparsity in the features by penalizing the log-likelihood function to optimize. Pan and Shen (2007) proposed a penalized clustering model for continuous data by considering a ℓ_1 penalty function focused on the mean of each cluster. The idea behind this approach is to define a feature as irrelevant if its mean is equal in all components. In the same line of reasoning several articles adapted or extended this concept of penalized model with variable selection (Wang and Zhu, 2008; Zhou et al, 2009).

A third way to perform feature selection in the model-based framework was introduced in (Law et al, 2004). The authors considered a Gaussian mixture model and decomposed the Normal distribution, commonly used in this case, into a product of two probability distributions. Parameters of the first distribution are specific to each cluster while parameters of the second distribution are independent from the clustering partition. This approach was further extended to other types of distributions (Wang and Kabán, 2005; Li and Zhang, 2008). The main drawbacks of these models is that even though they reduce the impact of the irrelevant features on the partition of the data, by assigning to them a low weight, they still consider them while computing the clustering partitions. For more details on model-based feature selection, we refer the reader to the survey of Bouveyron and Brunet-Saumard (2014).

Despite the good results for identifying irrelevant features in simple clustering, these methods present an increasing computational complexity with respect to the number of features, which make them impracticable in the context of high-dimensional data. In addition, they do not really perform feature selection, but rather feature weighting. In this work, we propose to extend the concept of noise cluster, introduced for outlier detection, to the problem of feature selection through a probabilistic co-clustering approach defined hereafter. Our goal is to exclude completely the irrelevant features from the clustering process without the need to choose any predefined threshold nor use any individual weighting schema.

2.2 Latent block models

Notation. We assume that the data are represented by a matrix $\mathbf{x} = \{x_{ij}, i \in I = \{1, \dots, n\}; j \in J = \{1, \dots, d\}\}$, where n and d denote the number of objects and variables, respectively. In this work we will limit ourselves to binary matrices, i.e. $x_{ij} \in \{0, 1\}$. A partition of I into g clusters is represented by the classification matrix $\mathbf{z} = (z_{ik}, i = 1, \dots, n, k = 1, \dots, g)$ where $z_{ik} = 1$ if element i belongs to cluster k and $z_{ik} = 0$ otherwise. In a similar way we define the partition of J into m clusters by $\mathbf{w} = (w_{j\ell}, j = 1, \dots, d, \ell = 1, \dots, m)$. Finally, sums and products related to rows, columns, row's cluster and column's cluster are subscripted by the letters i, j, k and ℓ without indicating the limits of variation which will be implicit. So, the sums \sum_i, \sum_j, \sum_k and \sum_ℓ stand for $\sum_{i=1}^n, \sum_{j=1}^d, \sum_{k=1}^g$ and $\sum_{\ell=1}^m$, respectively.

Latent Block Model. The co-clustering task can be embedded into a probabilistic framework, with the Latent Block Model (LBM), proposed by Govaert and Nadif (2003). Given a

matrix $\mathbf{x} \in \mathbb{R}^{n \times d}$, this model considered that the univariate random variables x_{ij} are conditionally independent knowing \mathbf{z} and \mathbf{w} , with parametrized probability density function (pdf) $f(x_{ij}; \alpha_{k\ell})$ if the row i belongs to the cluster k and the column j belongs to the cluster ℓ . The conditional pdf of \mathbf{x} knowing \mathbf{z} and \mathbf{w} can be expressed as

$$\prod_{i,j} f(x_{ij}; \alpha_{z_i w_j}) = \prod_{i,j,k,\ell} \{f(x_{ij}; \alpha_{k\ell})\}^{z_{ik} w_{j\ell}}.$$

In this case, the two sets I and J are assumed to be random samples so that the row and column labels become latent variables. This model is based on the following assumptions:

- Conditional independence defined before;
- Independent latent variables: the partitions $\mathbf{z}_1, \dots, \mathbf{z}_n$, $\mathbf{w}_1, \dots, \mathbf{w}_d$ are considered as latent variables and assumed to be independent:

$$p(\mathbf{z}, \mathbf{w}) = p(\mathbf{z})p(\mathbf{w}),$$

$$p(\mathbf{z}) = \prod_i p(z_i) \quad \text{and} \quad p(\mathbf{w}) = \prod_j p(w_j).$$

- For all i , the distribution of $p(z_i)$ is the multinomial distribution $\mathcal{M}(1; \pi_1, \dots, \pi_g)$ and does not depend on i . Similarly, for all j , the distribution of $p(w_j)$ is the multinomial distribution $\mathcal{M}(1; \tau_1, \dots, \tau_m)$ and does not depend on j .

The set parameter of the latent block model is denoted by $\theta = (\pi, \tau, \alpha)$, where $\pi = (\pi_1, \dots, \pi_g)$ and $\tau = (\tau_1, \dots, \tau_m)$ represent the mixing proportions and $\alpha_{k\ell}$ is the parameter of the distribution for the block (k, ℓ) .

The LBM can be easily adapted to data of different natures such as continuous, binary or discrete by considering appropriate distributions, like the Gaussian (Govaert and Nadif, 2013), Bernoulli and Multinomial (Keribin et al, 2014) ones. In all cases, the estimation of both partitions \mathbf{z} and \mathbf{w} and the set of parameters of the model is achieved by a Variational EM algorithm.

In real-world application, one can often encounter rows (i.e. objects) or columns (i.e. features) which are irrelevant, and by taking them into account, LBM usually creates a very high number of artificial blocks; for instance, by dividing relevant blocks to multiple ones based on noise. In this paper, we propose to overcome this *over-learning* problem by assuming, in the LBM, that there exists a variable noise cluster which contains all irrelevant features and that these features have no impact on the object clustering.

3 Noise-free Latent Block Model

In this section, we describe the modelling assumptions behind the proposed model. As an illustrative example, let us consider the United State Congressional Voting Records data ¹. This dataset contains votes of the members of the U.S. House of Representatives on various political issues. Each element is therefore coded by 1 for “yea” and 0 for “nay”. In analyzing these data, our goal is threefold.

1. We aim to determine the underlying political groups within the Congress, i.e., we assume g latent groups over the members.

¹ <https://archive.ics.uci.edu/ml/datasets.html>

2. We also aim to find the underlying political topics among the voted laws, i.e., we assume m latent groups over the political issues.
3. We aim to differentiate between relevant from non-informative political issues by assuming a set of laws for which the vote will not be influenced by the political group.

Consequently, we assume that the observed data matrix is generated according to a product of two mixtures of underlying probability density functions where the first mixture is associated to a relevant block structure while the second one contains only the irrelevant features and is therefore independent from the latent structure. Then, each member (resp. political issue) is associated with a random vector π_i (resp. τ_j), where π_{ik} (resp. $\tau_{j\ell}$) is the probability of a member i (resp. an issue j) belonging to group k (resp. ℓ). For each member and issue, the indicator vectors (z_1, \dots, z_n) where $z_i \in \{1, \dots, k, \dots, g\}$, and (w_1, \dots, w_d) where $w_j \in \{1, \dots, \ell, \dots, m\}$ represent the group membership of the i -th member and the j -th issue, respectively.

In addition, in order to deal with irrelevant features, our model relies on the idea that only a proportion ϕ of the laws are discriminant, and that the remaining proportion $(1 - \phi)$ of are just noise. To proceed we assume the existence of a noise cluster w_0 and introduce a novel parameter ϕ which role is to measure the proportion of relevant features. As a result, we now consider that \mathbf{w} is drawn according to $\prod_{j=1}^d \mathcal{M}(1; (1 - \phi), \phi\boldsymbol{\tau})$ in $\{0, \dots, m\}$.

On the one hand, features that belong to the noise cluster are assumed to be drawn according to a probability distribution, with parameters $\boldsymbol{\lambda} = (\lambda_j); j = 1, \dots, w_{+0}$, where w_{+0} is the number of features in the noise cluster. One can observe that $\boldsymbol{\lambda}$ does not depend on neither k or ℓ but is feature-specific, allowing to model different type of noise. For instance, a very small λ_j indicates that the noisy feature j mostly takes the value 0, while a high value represents a noisy feature with a majority of 1's. On the other hand, relevant features are assumed to be drawn from a probability distribution with parameters $\boldsymbol{\alpha} = (\alpha_{k\ell}); k = 1, \dots, g; \ell = 1, \dots, m$, which are block-specific.

In this work, we limit ourselves to the case of Bernoulli distribution, for the sake of clarity, but the extension to other density function (such as Gaussian or Multinomial), and therefore to other type of type (continuous or categorical) is straightforward.

Now, putting everything together we can define the Noise-free Latent Block model (NFLB) which postulates that a data matrix \mathbf{x} is drawn from the following generative procedure.

- Generating \mathbf{z} according to $\prod_{i=1}^n \mathcal{M}(1; \boldsymbol{\pi})$ in $\{1, \dots, g\}$.
- Generating \mathbf{w} according to $\prod_{j=1}^d \mathcal{M}(1; (1 - \phi), \phi\boldsymbol{\tau})$ in $\{0, \dots, m\}$.
- Generating \mathbf{x} with for each $j \in \{1, \dots, d\}$:
 - if $w_{j0} = 1$, $x_{.j} \sim \mathcal{B}(\lambda_j)^n$,
 - else $x_{.j} \sim \prod_{i=1}^n \mathcal{B}(\alpha_{z_i w_j})$.

This process is illustrated as a graphical model in Figure 1. The hyper-parameters related to a Bayesian framework are explained in the next Section.

The mixture density can be expressed as

$$\begin{aligned}
 p(\mathbf{x}; \theta) &= \sum_{\mathbf{z}, \mathbf{w}} p(\mathbf{x} | \mathbf{z}, \mathbf{w}; \theta) p(\mathbf{z}, \mathbf{w}; \theta) \\
 &= \sum_{\mathbf{z}, \mathbf{w}} \prod_{i,k} \pi_k^{z_{ik}} \times (1 - \phi)^{w_{+0}} \phi^{(d - w_{+0})} \prod_{j,\ell} \tau_\ell^{w_{j\ell}} \\
 &\quad \times \prod_{i,j} f(x_{ij}, \lambda_j)^{w_{j0}} \times \prod_{i,j,k,\ell} f(x_{ij}, \alpha_{k\ell})^{z_{ik} w_{j\ell}},
 \end{aligned}$$

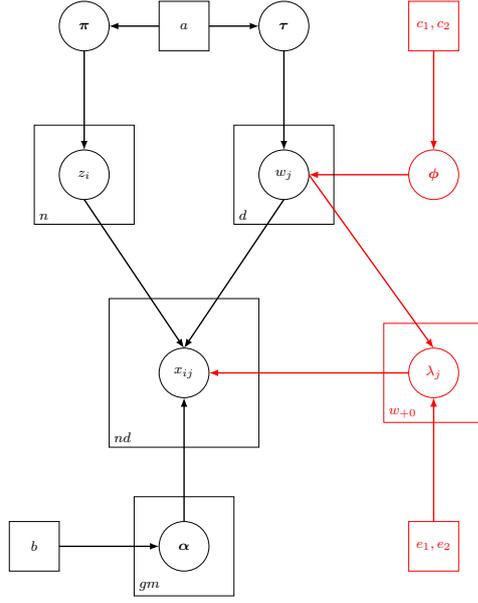


Fig. 1: Bayesian version of the Noise-free Latent Block model (NFLB). Elements in red represent the differences w.r.t. the classic LBM for Bernoulli distribution.

where $f(x_{ij}, \alpha_{k\ell}) = \alpha_{k\ell}^{x_{ij}} (1 - \alpha_{k\ell})^{(1-x_{ij})}$, $f(x_{ij}, \lambda_j) = \lambda_j^{x_{ij}} (1 - \lambda_j)^{(1-x_{ij})}$ and $\theta = (\pi, \phi, \tau, \lambda, \alpha)$ is the set of parameters of the model.

4 Parameter Estimation and Posterior Inference

Now, the goal is to compute the latent variables \mathbf{z} and \mathbf{w} and to estimate the set of parameters, θ . To proceed, one need to maximize the likelihood associated with our model, given by

$$\begin{aligned}
 \mathcal{L}(\theta) &= \sum_{(z,w) \in \mathcal{Z} \times \mathcal{W}} p(\mathbf{z}; \theta) p(\mathbf{w}; \theta) p(\mathbf{x} | \mathbf{z}, \mathbf{w}; \theta) \\
 &= \sum_{(z,w) \in \mathcal{Z} \times \mathcal{W}} \left(\underbrace{\prod_{i,k} \pi_k^{z_{ik}} \times (1-\phi)^{w_{+0}} \phi^{(d-w_{+0})} \prod_{j,\ell} \tau_\ell^{w_{j\ell}}}_A \right. \\
 &\quad \left. \times \underbrace{\prod_j \left(\lambda_j^{x_{+j}} (1-\lambda_j)^{(n-x_{+j})} \right)^{w_{j0}}}_B \times \underbrace{\prod_{i,j,k,\ell} \left(\alpha_{k\ell}^{x_{ij}} (1-\alpha_{k\ell})^{(1-x_{ij})} \right)^{z_{ik} w_{j\ell}}}_C \right), \quad (1)
 \end{aligned}$$

where $x_{+j} = \sum_{i=1}^n x_{ij}$.

The direct optimization of the likelihood for LBM's is a well-known issue: first it is intractable for large dataset, second deriving formula for the latent variables using a classic EM is challenging. To overcome this, Govaert and Nadif (2008) suggests to rather consider the maximization of the variational approximation of the likelihood, namely the Free Energy.

4.1 VE-step

The VE-step relies on the computation of the conditional expectation of the complete log-likelihood. First, we compute the probability for each object to belong to one of the row clusters $k = 1, \dots, g$:

$$s_{ik} = \frac{\pi_k \prod_{\ell=1}^m \prod_{j=1}^d [f(x_{ij}, \alpha_{k\ell})]^{w_{j\ell}}}{\sum_{k'} \pi_{k'} \prod_{\ell=1}^m \prod_{j=1}^d [f(x_{ij}, \alpha_{k'\ell})]^{w_{j\ell}}}, \quad (2)$$

where $s_{ik} = \mathbb{P}(z_{ik} = 1; x, \theta)$.

Then, we compute the probability for each feature to belong to one of the column clusters $\ell = 1, \dots, m$:

$$t_{j\ell} = \frac{\phi \tau_{\ell} \prod_{i,k} [f(x_{ij}, \alpha_{k\ell})]^{z_{ik}}}{\sum_{\ell'=1}^m \phi \tau_{\ell'} \prod_{i,k} [f(x_{ij}, \alpha_{k\ell'})]^{z_{ik}} + (1 - \phi) \prod_i [f(x_{ij}, \lambda_j)]}, \quad (3)$$

where $t_{j\ell} = \mathbb{P}(w_{j\ell} = 1; x, \theta)$. Finally, we distinguish one column cluster from the other ones which should contains only irrelevant features, i.e., non-discriminant ones:

$$t_{j0} = \frac{(1 - \phi) \prod_i [f(x_{ij}, \lambda_j)]}{\sum_{\ell'=1}^m \phi \tau_{\ell'} \prod_{i,k} [f(x_{ij}, \alpha_{k\ell'})]^{z_{ik}} + (1 - \phi) \prod_i [f(x_{ij}, \lambda_j)]} \quad (4)$$

where $t_{j0} = \mathbb{P}(w_{j0} = 1 | \mathbf{x}, \theta)$.

4.2 M-step

From B and C (equation 1), we obtain the estimators for the parameters of the Bernoulli densities given by

$$\hat{\lambda}_j = \frac{\sum_i x_{ij}}{n} = \frac{x_{+j}}{n}, \quad \hat{\alpha}_{k\ell} = \frac{\sum_{i,j} s_{ik} t_{j\ell} x_{ij}}{s_{+k} t_{+\ell}},$$

where $s_{+k} = \sum_i s_{ik}$ and $t_{+\ell} = \sum_j t_{j\ell}$.

From A (equation 1), and s.t. $\sum_k \pi_k = 1, \sum_{\ell=0}^m \tau_{\ell} = 1$, we obtain the proportion of row and column clusters,

$$\hat{\pi}_k = \frac{s_{+k}}{n}, \quad \hat{\tau}_{\ell} = \frac{t_{+\ell}}{d - t_{+0}} \quad \text{and} \quad \hat{\phi} = \frac{d - t_{+0}}{d},$$

where $t_{+0} = \sum_j t_{j0}$.

4.3 Bayesian Inference

While VEM is known to be an accurate way of estimating the parameters of the latent block models, one of the issue is its tendency to empty small clusters. To overcome this drawback, we also propose a Bayesian version of our framework.

A priori assumptions. In a Bayesian perspective, one can consider proper and independent non informative prior distributions for the mixing proportions π , τ , and for the parameters α and λ as a product of $g \times m$ and w_{+0} non informative priors on each Bernoulli parameter, respectively. Therefore, for the mixing proportions we have that $\pi, \tau \sim Dir(a, \dots, a)$ that is

$$p(\pi) \propto \prod_{k=1}^g \pi_k^{a-1} \quad \text{and} \quad p(\tau) \propto \prod_{\ell=1}^m \tau_\ell^{a-1},$$

and for the Bernoulli parameters, we have $\alpha \sim \mathcal{B}e(b, b)$, $\lambda \sim \mathcal{B}e(e_1, e_2)$, that is

$$p(\alpha) \propto \prod_{k,\ell} \alpha_{k\ell}^{b-1} (1 - \alpha_{k\ell})^{b-1}$$

and

$$p(\lambda) \propto \prod_{j:w_{j0}=1} \lambda_j^{e_1-1} (1 - \lambda_j)^{e_2-1}.$$

Finally, we also assume an a priori on the parameter ϕ

$$\phi \sim \mathcal{B}e(c_1, c_2) \quad \text{i.e.} \quad p(\phi) \propto \phi^{c_1-1} (1 - \phi)^{c_2-1}.$$

One can observe that we choose two different hyper-parameters for the Beta distributions associated to the λ 's and ϕ 's parameters, while we set them to be the same for the α 's. This choice and the values of these hyper-parameters will be discussed below. The graphical representation of the Bayesian version of the NFLB is presented in Figure 1.

In practice, one usually consider Jeffrey's or uniform's prior distributions on the parameters. The problem is that in the case of proportions parameters (i.e. π and τ) considering these types of distributions usually tends to empty some of the clusters (). To overcome this, Keribin et al (2014) have shown that taking $a = 4$ and $b = 1$ allows to prevent cluster from vanishing during the label assignment. For the parameters (e_1, e_2) and (c_1, c_2) , we choose to use uniform distribution so as not to put any type of a priori information. However, one should note that we can induce some specific behaviors depending on the availability of such a priori knowledge on the data:

- On the distribution of ϕ , we can introduce asymmetry if we have information about the proportion of irrelevant features. For instance, considering $c_2 \gg c_1$ implies that the proportion of noise is important compared to the proportion of relevant features.
- On the distribution of λ , we can also decide to model irrelevant features in different ways. On the one hand, we can consider that a feature is irrelevant if all the objects have a similar value for this latter, and in this case taking e_1 and e_2 smaller than 1 will encourage the λ 's to be close to the extremes 0 or 1. On the other hand, if we define noise as features with very mitigates values, taking $e_1 = e_2$ with high values tends to encourage λ 's close to 0.5.

Next, we present the new update formula for all parameters in this Bayesian framework.

New parameter estimation. Let us recall that we now consider the optimization of the Bayes formula given by

$$p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta)p(\theta).$$

From this formula, we can directly derive an EM algorithm for the computation of the MAP estimates.

- The E-Step remains the same than the one previously defined in the NFLB model and therefore the formula for computing the conditional probabilities of the labels are unchanged.
- The M-Step differs as the objective to maximize is augmented by the prior density. The Bayesian version combined with the variational approximation leads to

$$\hat{\pi}_k = \frac{s_{.k} + a - 1}{n + g(a - 1)}, \quad \hat{\tau}_\ell = \frac{t_{+\ell} + a - 1}{t_{+\ell} + m(a - 1)}, \quad (5)$$

where $s_{.k}$, $t_{+\ell}$ and $t_{+\ell}$ are the same than the ones defined in the variational approximation. For the parameters of the Bernoulli densities we have

$$\hat{\lambda}_j = \frac{\sum_i x_{ij} + e_1 - 1}{n + e_1 + e_2 - 2}, \quad \hat{\alpha}_{k\ell} = \frac{\sum_{i,j} s_{ik} t_{j\ell} x_{ij} + b - 1}{\sum_{i,j} s_{ik} t_{j\ell} + 2(b - 1)}. \quad (6)$$

Finally, the estimate of ϕ is given by

$$\hat{\phi} = \frac{t_{+\ell} + c_1 - 1}{d + c_1 + c_2 - 2}. \quad (7)$$

Now, going back to the U.S. votes data, we applied the aforementioned Bayesian version of the NFLB (for which the extension to the multinomial case is straightforward). The number of co-clusters is automatically determined by an appropriate model selection criteria defined in the next section. In this context, we obtained 5 clusters for the member of the congress and 6 clusters for the political issues (including the noise cluster). This structure is presented in Figure 2. From this figure, one can see a clear block structure for some of the political issues, where groups of congress members voted homogeneously 'yeah' or 'nay'. The noise cluster contains two of the laws (left to the red line).

In addition, we have an information regarding the official political group of each member; we know that among them 168 are labeled as Republicans and the remaining 267 as Democrats. By taking a closer look at the distribution of Republicans and Democrats among the clusters (see Table 1), one can see that cluster 2 is mainly composed of Republicans while clusters 4 and 5 are dominated by Democrats. Clusters 1 and 3 have a mixed representation of both, and therefore we can say that for these members, the vote was not clearly impacted by the political affiliation. Furthermore, our procedure separates the greatest non-voters, who chose to abstain from voting on more than 50% of the laws (cluster 1). Finally, in terms of variable cluster, the last political issue is not considered as being part of the noise cluster, and constitute a cluster by itself, because the non-voters all abstained on this one. Compared to the results reported by Keribin et al (2014) (see Table 1), our approach strengthens the politicization of each cluster. Also, we obtain a smaller number of cluster (both on the members and the laws) than Wyse and Friel (2012).

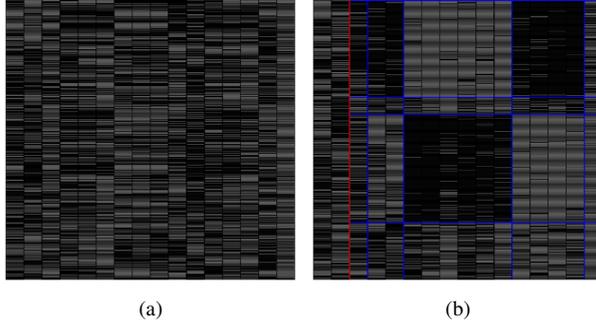


Fig. 2: Visualization of (a) the United State Congressional Voting Records data, where black cells indicate a “yea” and white cells represent a “nay”; (b) the same data matrix reorganized according to the partitions obtained with the NFLB model. Irrelevant features are isolated on the left side cluster delimited by the red line. Blue lines indicate the relevant block structure.

Table 1: Distribution of Democrats (Dem.) and Republicans (Rep.) among the row clusters obtained with NFLB and LBM (Keribin et al, 2014).

Cluster	NFLB		LBM	
	Rep.	Dem.	Rep.	Dem.
1	2	3	2	4
2	129	20	132	22
3	10	16	8	11
4	2	165	1	162
5	25	63	25	68

4.4 Gibbs sampling for the binary NFLB

As θ , the set of parameters, is assumed to be a random variable, Gibbs sampling aims to estimate the distribution $p(\mathbf{x}, \mathbf{z}, \mathbf{w}, \theta)$ with a Monte Carlo Markov Chain. Therefore, the idea is to simulate each of the parameter according to a conditional probability of the others:

1. Simulation of $\mathbf{z}^{(c+1)}$ according to $p(\mathbf{z} \mid \mathbf{x}, \mathbf{w}^{(c)}, \theta^{(c)})$ defined by:

$$p(\mathbf{z} \mid \mathbf{x}, \mathbf{w}^{(c)}, \theta^{(c)}) = \prod_{i=1}^n \mathcal{M}(1; s_{i1}, \dots, s_{ig})$$

where s_{ik} is defined by Equation (2).

2. Simulation of $\mathbf{w}^{(c+1)}$ according to $p(\mathbf{w} \mid \mathbf{x}, \mathbf{z}^{(c+1)}, \theta^{(c)})$ defined by:

$$p(\mathbf{w} \mid \mathbf{x}, \mathbf{z}^{(c)}, \theta^{(c)}) = \prod_{j=1}^d \mathcal{M}(1; t_{j0}, t_{j1}, \dots, t_{jm})$$

where $t_{j\ell}$ is defined by Equations (3) and (4).

3. Simulation of $\pi^{(c+1)}$ according to $\text{Dir}(a + z_{+1}^{(c+1)}, \dots, a + z_{+g}^{(c+1)})$.
4. Simulation of $\tau^{(c+1)}$ according to $\text{Dir}(a + w_{+1}^{(c+1)}, \dots, a + w_{+m}^{(c+1)})$.

5. Simulation of $\phi^{(c+1)}$ according to $\mathcal{B}e\left(c_1 + d - w_{+0}^{(c+1)}, c_2 + w_{+0}^{(c+1)}\right)$.
6. Simulation of $\lambda_j^{(c+1)}$ for any j such that $w_{j0} = 1$ according to $\mathcal{B}e\left(e_1 + x_{+j}, e_2 + n\right)$.
7. Simulation of $\alpha_{k\ell}^{(c+1)}$ according to

$$\prod_{k=1}^g \prod_{\ell=1}^m \mathcal{B}e\left(\sum_{i=1}^n \sum_{j=1}^d z_{ik}^{(c+1)} w_{j\ell}^{(c+1)} x_{ij}, z_{+k}^{(c+1)} w_{+\ell}^{(c+1)}\right).$$

The main goal of Gibbs sampling is to create a Markov chain that will give an overview of the posterior distribution of $(\mathbf{z}, \mathbf{w}, \boldsymbol{\theta})$. This overview relies on the idea that, as the number of iteration tends to infinity, the more likely a region is, the more parameters they will be in this specific region. One should also note, that the distribution has a great number of symmetries, due to the problem of *label switching*; for instance, if we switch two cluster numbers, let us say 1 and 2, the parameters will be different while they basically represent the same cluster. In order to solve this problem, we use the identifiability conditions given in Theorem 1 so as to order the clusters, then we average all simulated $\boldsymbol{\theta}$'s, and the assignment to a specific cluster is done according to a simple majority vote.

5 Model selection

Clustering, and by extension co-clustering, poses the question of defining an appropriate number of (co)-clusters. For mixture model, this problem is generally framed as a problem of model selection, and remains a difficult challenge, as choosing the best model according to the maximum likelihood usually leads to pick the most complex one. To overcome this issue, a popular approach is to penalize the likelihood of the model by, for instance, the number of parameters. In the framework of mixture models, two criteria are commonly used: the *Bayesian Information Criterion* (BIC) which is known to be consistent, and the *Integrated Completed Likelihood* (ICL), that aims to minimize the entropy. In the context of the LBM, empirical studies carried out by Keribin et al (2014) show that the two estimates provided by each of the criteria are asymptotically equal; this result is further supported by the recent work of Brault et al (2017). For the NFLB, we propose to use the ICL that, for each pair (g, m) , can be written as

$$\begin{aligned} \text{ICL}(g, m) &= p(\mathbf{x}, \mathbf{z}, \mathbf{w}; g, m) \\ &= \log \Gamma(ga) - g \log \Gamma(a) - \log \Gamma(n + ga) + \sum_{k=1}^g \log \Gamma(z_{+k} + a) \\ &\quad + \log \Gamma(c_1 + c_2) - \log \Gamma(c_1) - \log \Gamma(c_2) + \log \Gamma(ma) - m \log \Gamma(a) \\ &\quad + \log \Gamma(d - w_{+0} + c_1) + \log \Gamma(w_{+0} + c_2) - \log \Gamma(d + c_1 + c_2) \\ &\quad + \sum_{\ell=1}^m \log \Gamma(w_{+\ell} + a) - \log \Gamma(d - w_{+0} + ma) + w_{+0} [\log \Gamma(e_1 + e_2) \\ &\quad - \log \Gamma(e_1) - \log \Gamma(e_2) - \log \Gamma(n + e_1 + e_2)] \\ &\quad + \sum_{j=1}^d w_{j0} [\log \Gamma(x_{+j} + e_1) + \log \Gamma(n - x_{+j} + e_2)] \end{aligned}$$

$$\begin{aligned}
& +gm (\log \Gamma(2b) - 2 \log \Gamma(b)) + \sum_{k=1}^g \sum_{\ell=1}^m [\log \Gamma(N_{k\ell}^{\mathbf{z}\mathbf{w}} + b) \\
& + \log \Gamma(z_{+k}w_{+\ell} - N_{k\ell}^{\mathbf{z}\mathbf{w}} + b) - \log \Gamma(z_{+k}w_{+\ell} + 2b)] \tag{8}
\end{aligned}$$

where Γ is the Gamma function such that $\Gamma(n+1) = n!$, $N_{k\ell}^{\mathbf{z}\mathbf{w}} = \sum_{i=1}^n \sum_{j=1}^d z_{ik}w_{j\ell}x_{ij}$, and \mathbf{z} and \mathbf{w} are estimated by the algorithm. The calculations are derived from the appendix of Keribin et al (2014).

6 Theoretical Analysis

In this section, we propose to study the theoretical properties of the NFLB model. We start by showing that under reasonable assumptions, the model is identifiable, and proceed with results on its consistency.

Theorem 1 (Identifiability) *Consider the binary NFLB with $\boldsymbol{\pi}$ and $\boldsymbol{\tau}$ be the row and column mixing proportions and $\boldsymbol{\alpha} = (\alpha_{k,\ell})$ the $g \times m$ matrix of Bernoulli parameters. The binary NFLB is identifiable under one of these groups assumptions:*

- If the true partition $\mathbf{w}_{\cdot,0}^*$ of the null columns cluster known and :
 - For all $k \in 1, \dots, g$, $\pi_k > 0$ and the coordinates of the vector $\boldsymbol{\alpha}\boldsymbol{\tau}$ are distinct.
 - For all $\ell \in 1, \dots, m$, $\tau_\ell > 0$ and the coordinates of the vector $\boldsymbol{\pi}'\boldsymbol{\alpha}$ are distinct.
 - $n \geq 2m - 1$ and $d - w_{+0}^* \geq 2g - 1$.
- If every $\lambda_j = \lambda$ and
 - For all $k \in 1, \dots, g$, $\pi_k > 0$ and the coordinates of the vector $\phi\boldsymbol{\alpha}\boldsymbol{\tau}$ are distinct and distinct of $(1 - \phi)\lambda$.
 - For all $\ell \in 1, \dots, m$, $\tau_\ell > 0$ and the coordinates of the vector $\boldsymbol{\pi}'\boldsymbol{\alpha}$ are distinct and distinct of λ .
 - $n \geq 2m - 1$ and $d \geq 2g + 1$.

Proof For the first conditions, the key of the proof is in two steps:

- If the column j is in the cluster 0, the identifiability is guaranteed by the identifiability of the Bernoulli distribution.
- For the other columns, the identifiability is a corollary of the result of the latent block model (see Keribin et al (2014)).

For the second conditions, we just observe that we have a LBM where the first cluster has g times the same value, and, therefore we are in the conditions of Keribin et al (2014).

Now, we give result on the consistency of the estimators obtained in NFLB.

Theorem 2 (Consistency) *For the consistency, we need some assumptions on the parameters space Θ :*

- \mathcal{A}_1 *There exists a positive constant δ such that the space of parameters of $(\boldsymbol{\pi}^*, \boldsymbol{\phi}^*, \boldsymbol{\tau}^*)$ is included in $[\delta; 1 - \delta]^{g+1+m+g \times m}$.*
- \mathcal{A}_2 *The true set of parameter $\boldsymbol{\theta}^* = (\boldsymbol{\pi}^*, \boldsymbol{\phi}^*, \boldsymbol{\tau}^*)$ lies in the relative interior of the previous part of Θ .*
- \mathcal{A}_3 *All rows and columns of $\boldsymbol{\alpha}^*$ are unique.*
- $\mathcal{A}_{4.a}$ *These both assumptions are true:*
 - Every λ_j^* equal to the same $\lambda^* \in]\delta; 1 - \delta[$.

– No column of α^* is constant equal at λ .

$\mathcal{A}_{4.b}$ These both assumptions are true:

– The number w_{+0}^* of the noises columns is known.

– There exists $\gamma > 0$ such that for every λ_j and $(k, \ell) \in \{1, \dots, g\} \times \{1, \dots, m\}$, we have: $|\lambda_j - \alpha_{k\ell}| > \delta$.

Under the assumptions $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$ and ($\mathcal{A}_{4.a}$ or $\mathcal{A}_{4.b}$) the estimator of the maximum likelihood is consistent (up to one permutation of the cluster number).

Proof With the assumption $\mathcal{A}_{4.a}$, the model can be seen as a classic LBM under the assumption of Brault et al (2017). The sketch of the proof with $\mathcal{A}_{4.b}$ is that if the number of columns is fixed and the distance between the parameters is large enough, then the classification between noise or relevant is consistent and the rest is under the conditions of the estimation of the LBM.

7 Experimental Results

In order to give a comprehensive evaluation of the proposed models, we first conduct extensive experimentation on synthetic binary datasets, then demonstrate their efficiency on real DNA sequence datasets. In what follows, posterior probabilities returned by all approaches are converted into crisp partitions by using the Maximum A Posteriori (MAP) principle. Both the implementation of the model and the pre-processed real datasets can be found online².

7.1 Synthetic data

Setting. Monte-Carlo experiments are performed on $(n \times d)$, with $n \in \{100, 1000\}$ and $d \in \{60, 600\}$, synthetic binary datasets, arising from a $(g, m) = (5, 3)$ block structure with equal proportions in order to compare the behaviour of the proposed algorithm. The simulation process fits the generative process describe previously and we consider three scenarios differing in the percentage of irrelevant features (sampled from a $\mathcal{B}(\lambda)$, λ being randomly drawn) introduced: 10% and 50% and 80% successively. In addition, we consider easily (+), moderately (++) or hardly (+++) separated mixtures.

We compare B-NFLB, referred to as V-Bayes and its Gibbs version, referred to as Gibbs in the results in terms of two metrics: the co-clustering error (CE) (Patrikainen and Meila, 2006) and the co-clustering extension of the Adjusted Rand Index (CARI) (Robert and Vasseur, 2017). The CE is defined as follows

$$\text{CE}((\mathbf{z}, \mathbf{w}), (\hat{\mathbf{z}}, \hat{\mathbf{w}})) = e(\mathbf{z}, \hat{\mathbf{z}}) + e(\mathbf{w}, \hat{\mathbf{w}}) - e(\mathbf{z}, \hat{\mathbf{z}}) \times e(\mathbf{w}, \hat{\mathbf{w}}),$$

where $\hat{\mathbf{z}}$ and $\hat{\mathbf{w}}$ are the partitions of instances and variables estimated by the algorithm; \mathbf{z} and \mathbf{w} are the true partitions and $e(\mathbf{z}, \hat{\mathbf{z}})$ (resp. $e(\mathbf{w}, \hat{\mathbf{w}})$) denotes the error rate, i.e., the proportion of misclassified objects (resp. features). The CE is in $[0, 1]$, 0 being the case where the partitions are the same. CARI, is a symmetric index and takes the value 1 when the couples of true and estimated partitions agree perfectly, up to a permutation.

² The datasets can be found here: <https://github.com/laclauc/NFLB> and the code will be available upon publication.

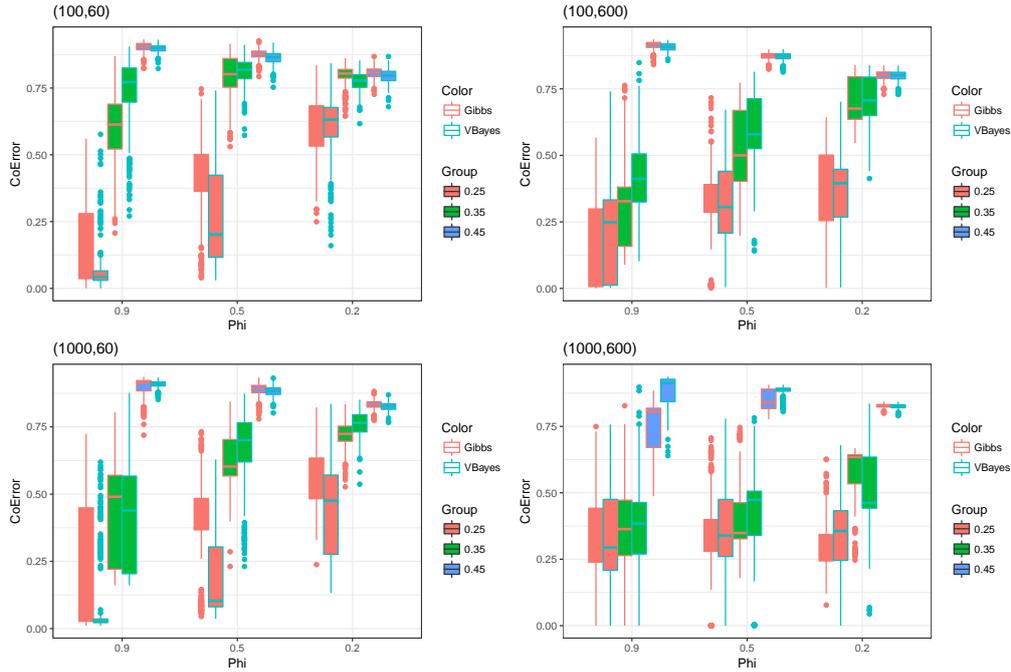


Fig. 3: CCE on data matrices as a function of the number of rows and columns, ϕ (x-axis on the plots), the degree of overlapping between co-clusters (colors of the boxplots) and the chosen algorithms (outline color of the boxplots).

Estimating the partitions knowing g and m . We compare the quality of the partitions z and w obtained with the V-Bayes algorithm and the V-Bayes with Gibbs sampling algorithm. For the first one, we randomly initialize the algorithm 10 times for each setting and keep the result that corresponds to the maximum likelihood. For the second one, we use 1000 iterations for the Gibbs sampling. In both cases, we set g and m to the values use during simulations.

In the following, we only present the results in terms of CCE for some of the settings as the results of CARI (and of other settings) were nearly the same. Figure 3 reports the results for different settings. One can observe that both algorithms give good performance in general and that none of them really seem to distinguish from the other. In addition, the quality of the estimation is inversely proportional to the difficulty of the data.

Estimating g and m . Next, we assess the ability of our approach to estimate the couple (g, m) on the generated data. For each simulated matrix, we vary the number of row clusters between 2 and 8 and the number of column clusters between 1 and 6. Table 2 reports the results obtained for $\phi = 0.5$ (i.e. 50% of the features relevant) and $[+]$ degree of separation. These results corresponds to the couple (g, m) that maximizes the ICL criterion given in Equation 8.

From this table, one can observe that in most cases, the ICL criterion correctly identifies both g and m . We can also see that it tends to underestimate the number of clusters when the number of rows and columns are too small, and that the estimation gets better as the

Table 2: Frequency of the models selected by the ICL criterion on 50 data matrices with well separated clusters, and 50% of irrelevant features.

		$d = 60$						$d = 600$									
		$m \backslash g$								$m \backslash g$							
		1	2	3	4	5	6	1	2	3	4	5	6				
$n = 100$	1																
	2	4															
	3		8	2													
	4			20	2												
	5			60	2					80	12						
	6			2						6							
	7									2							
	8																
$n = 1000$	1																
	2																
	3																
	4			2													
	5			80	14	4				95	3	3					
	6																
	7																
	8																

dimension of the data increases. In addition, one should note that when the number of rows is much greater than the number of columns ($n > d$), ICL may overestimate the number of classes in columns; in contrast, if the number of columns is greater than the number of rows ($d > n$), then g can be slightly overestimated.

In Table 3, we also report the results for the 600×1000 data with different levels of noise. One can see that the proposed approach is robust to the noise, as even with only 10% of relevant features ($\phi = 0.1$) the right couple (g, m) is correctly identified in more than 95% of the cases. However, when ϕ is low, we observe a behavior close to the one of LBM (Keribin et al, 2014), with a possible overestimation of the number of column clusters. This observation is coherent with the fact that when there is few noise, the NFLB approaches the classic LBM.

Table 3: Number of times (in %) that NFLB identifies the correct g and m for $(n, d) = (1000, 600)$ and $[+]$ datasets with different ϕ .

		$\phi = 0.9$						$\phi = 0.2$									
		$m \backslash g$								$m \backslash g$							
		1	2	3	4	5	6	1	2	3	4	5	6				
	1																
	2																
	3																
	4																
	5			89	11					96	2						
	6									2							
	7																
	8																

7.2 Genetic diversity through Microsatellites

A microsatellite is a DNA sequence formed by a continuous repetition of units usually composed of 1 to 4 nucleotides. The length of these sequences (i.e., the number of repeats) varies according to the species, but also from one individual to another and from one allele to the other. However, the location of these sequences in the genome is relatively similar between phylogenetically close species. In the following, we propose to apply the NFLB model to two microsatellite datasets, originally proposed to study the link between genetic diversity and geographical location of individuals.

Description and pre-processing. The first dataset, referred to as DIVERSITY³, was proposed by Rosenberg et al (2002), in order to investigate genetic diversity and population structure in the world using genotypes at 377 autosomal microsatellite loci in 1056 individuals from 52 populations. The second one, referred to as NATIVE⁴ is a subset of the data reported by Wang et al (2007), which is an extension of DIVERSITY. The original dataset consists of 678 microsatellite loci genotyped in 1484 individuals from 78 worldwide populations including 29 Native American populations. In this work, we propose to focus on the Native American populations from north, central and south America, arising from 27 different tribes.

Pre-processing to obtain binary data is done as in the original papers. In addition, for the DIVERSITY data we propose to assess the capacity of our model to exploit only partial information. The original data contains genotypes measured in base pairs, and we propose to extract three biased versions: the two first versions only contain 1 out of the 2 pair and are referred to as DIVERSITY₁ and DIVERSITY₂ in the following. The third one, referred to as DIVERSITY_{Comp} considers each element of the pair as separate individuals. Table 4 provides details on all datasets. Finally, for all datasets, we select the number of clusters based on the ICL criteria.

Table 4: Properties of the different pre-processed datasets in terms of size, sparsity (% of zeros).

Properties	DIVERSITY _{Comp}	DIVERSITY ₁	DIVERSITY ₂	NATIVE
n	2112	1056	1056	494
d	4689	3949	3867	5709
Sparsity	92.2%	90.8%	90.6%	88.1%

Results. Table 5 presents the estimated number of clusters as well as the proportion of variables denoted by $\hat{\phi}$. One can observe that when we only take into account partial information (DIVERSITY₁ and DIVERSITY₂), then only around one third on the features are considered as relevant, while for the full data, the model is more conservative and identify twice this proportion as relevant. For the NATIVE data, 85% of the features are clustered as irrelevant. This is in line with the fact that this dataset only studies individuals from the same continent, who have more characteristics in common (i.e. non discriminant). From this dataset, we also observe two interesting phenomenon: (1) our model separate all the tribes which

³ <https://rosenberglab.stanford.edu/data/rosenbergEtAl2002/diversitydata.stru>

⁴ <https://rosenberglab.stanford.edu/nativedata.html>

were added to the data by Wang et al (2007). This might be the results of a different coding of the newly gathered data. (2) The 19 clusters are usually associated to one or two tribes.

Table 5: Estimated parameters on all four datasets: number of row and column clusters (\hat{g} , \hat{m}), proportion of relevant features ($\hat{\phi}$).

Estimation	DIVERSITY _{Comp}	DIVERSITY ₁	DIVERSITY ₂	NATIVE
\hat{g}	16	4	6	19
\hat{m}	49	32	26	27
$\hat{\phi}$	66.9%	31.7%	37.7%	14.4%

Figure 4 crosses the clusters and the continent information. In both cases, we note that four of the clusters corresponds exactly to 4 of the continents. However, the three remaining ones are together in one cluster; for instance, Europe with Central, South Asia. We see two possible explanations: (1) there exist a well-known proximity between some countries from the different continents (e.g Russia is part of Europe but geographically close to central Asia); (2) we only use half of the genotypes to characterize the individuals.

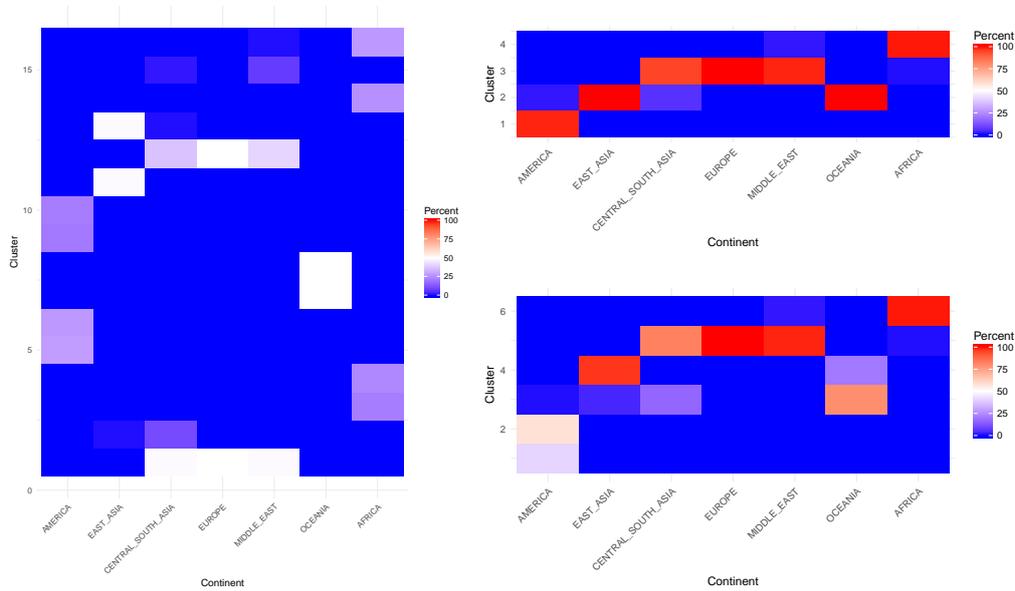


Fig. 4: Description of the clusters according to the Continent for the DIVERSITY datasets.

On DIVERSITY_{comp}, we observe that the clusters exclusively contain one or the other allele (explaining also the higher number of clusters). However, we also see that the clusters are slightly different from the agglomeration of partitions from DIVERSITY₁ and DIVERSITY₂. We have 16 clusters divided into 7 clusters for the first information and 9 clusters for the second one, against 4 and 6 for DIVERSITY₁ and DIVERSITY₂, respectively. Table 6 presents the number of individuals belonging to the same pair of clusters. One can see that to each cluster of the first allele we can associate one cluster build with the second allele.

Table 6: Number of individuals present in each cluster of allele 1 versus allele 2.

	2	4	6	7	10	11	12	14	15
1	45						464		34
3		48						4	
5			59		5				
8	2			39		1			
9			2		43				
13	5					235			
16		8						59	3

8 Conclusion

In this article, we proposed to study the problem of joint co-clustering and feature selection using the framework of mixture models. To this end, we propose a new framework that states the existence of a variable noise cluster, allowing a flexible definition of a noisy feature. From this framework, we derived two Bayesian models and adapted the ICL criteria for selecting an appropriate number of clusters. We first validated our approach on synthetic data and then apply it on a real-world application where the goal is to explore genetic diversity across the world. We were able to show the interest of feature selection in order to maintain good clustering results in presence of noise.

Although the results obtained with this new framework are promising, it admits some further improvements. First, we would like to extend our approach to stochastic variational inference (Hoffman et al, 2013), that is a scalable algorithm for approximating posterior distributions. On the other hand, it might be interesting to impose a penalty term on the number of variables contained in the noise cluster.

References

- Ben-David S, Haghtalab N (2014) Clustering in the presence of background noise. In: Proceedings of ICML, pp 280–288
- Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. *PAMI* 22(7):719–725
- Bouveyron C, Brunet-Saumard C (2014) Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis* 71:52–78
- Brault V, Keribin C, Mariadassou M (2017) Consistency and asymptotic normality of latent blocks model estimators. arXiv preprint arXiv:170406629
- Celeux G, Martin-Magniette ML, Maugis C, Raftery AE (2011) Letter to the editor: "A framework for feature selection in clustering". *Journal of the American Statistical Association* 106:383
- Cuesta-Albertos JA, Gordaliza A, Matrán C (1997) Trimmed k -means: an attempt to robustify quantizers. *The Annals of Statistics* 25(2):553–576
- Dave RN (1991) Characterization and detection of noise in clustering. *Pattern Recogn Lett* 12(11):657–664
- Dave RN (1993) Robust fuzzy clustering algorithms. In: [Proceedings 1993] Second IEEE International Conference on Fuzzy Systems, pp 1281–1286 vol.2
- Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of KDD, AAAI Press, pp 226–231

- García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A (2008) A general trimming approach to robust cluster analysis. *The Annals of Statistics* 36(3):1324–1345
- García-Escudero LA, Gordaliza A, Matrán C, Mayo-Iscar A (2010) A review of robust clustering methods. *Advances in Data Analysis and Classification* 4(2):89–109
- Govaert G, Nadif M (2003) Clustering with block mixture models. *Pattern Recognition* 36:463–473
- Govaert G, Nadif M (2008) Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis* 52(6):3233 – 3245
- Govaert G, Nadif M (2013) Co-clustering. Wiley Online Library
- Hartigan JA (1972) Direct Clustering of a Data Matrix. *Journal of the American Statistical Association* 67(337):123–129
- Hoffman MD, Blei DM, Wang C, Paisley J (2013) Stochastic variational inference. *J Mach Learn Res* 14(1):1303–1347
- Keribin C, Brault V, Celeux G, Govaert G (2014) Estimation and selection for the latent block model on categorical data. *Statistics and Computing* pp 1–16
- Law MHC, Figueiredo MAT, Jain AK (2004) Simultaneous feature selection and clustering using mixture models. *IEEE Trans Pattern Anal Mach Intell* 26:1154–1166
- Li M, Zhang L (2008) Multinomial mixture model with feature selection for text clustering. *Know-Based Syst* 21(7):704–708
- Maugis C, Celeux G, Martin-Magniette ML (2009) Variable selection for clustering with gaussian mixture models. *Biometrics* 65(3):701–709
- Mirkin BG (1996) *Mathematical classification and clustering. Nonconvex optimization and its applications*, Kluwer academic publ, Dordrecht, Boston, London
- Pan W, Shen X (2007) Penalized model-based clustering with application to variable selection. *J Mach Learn Res* 8:1145–1164
- Patrikainen A, Meila M (2006) Comparing subspace clusterings. *IEEE Transactions on Knowledge and Data Engineering* 18(7):902–916
- Raftery AE, Dean N (2006) Variable selection for model-based clustering. *Journal of the American Statistical Association* 101:168–178
- Robert V, Vasseur Y (2017) Comparing high dimensional partitions, with the coclustering adjusted rand indew. *CoRR* abs/1705.06760
- Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. *Science* 298(5602):2381–2385
- Wang S, Zhu J (2008) Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics* 64(2):440–448
- Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C, Mazzotti G, Poletti G, Hill K, Hurtado AM, Labuda D, Klitz W, Barrantes R, Bortolini MC, Salzano FM, Petzl-Erler ML, Tsuneto LT, Llop E, Rothhammer F, Excoffier L, Feldman MW, Rosenberg NA, Ruiz-Linares A (2007) Genetic variation and population structure in native americans. *PLoS Genetics* 3(11)
- Wang X, Kabán A (2005) Finding uninformative features in binary data
- Wyse J, Friel N (2012) Block clustering with collapsed latent block models. *Statistics and Computing* 22(2):415–428
- Zhou H, Pan W, Shen X (2009) Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics* 3:1473–1496