The Latent Topic Block Model for the Co-Clustering of Textual Interaction Data

Laurent Bergé, Charles Bouveyron, Marco Corneli, Pierre Latouche

The Latent Topic Block Model for the Co-Clustering of Textual Interaction Data

Laurent R. Bergé[1,5]*, Charles Bouveyron[2,3], Marco Corneli[2] & Pierre Latouche[1,4]

[1] *Laboratoire MAP5, UMR CNRS 8145, Université Paris Descartes, Paris, France.*
[2] *Laboratoire J.A. Dieudonné, UMR CNRS 7351, Université Côte d'Azur, Nice, France.*
[3] *Epione, INRIA Sophia-Antipolis, Valbonne, France.*
[4] *Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne, Paris, France.*
[5] *Université du Luxembourg, 162a, Avenue de la Faïencerie, L-1511, Luxembourg.*

Abstract. In this paper, we consider textual interaction data involving two disjoint sets of individuals/objects. An example of such data is given by the reviews on web platforms (e.g. Amazon, TripAdvisor, etc.) where buyers comment on products/services they bought. We develop a new generative model, the latent topic block model (LTBM), along with an inference algorithm to simultaneously partition the elements of each set, accounting for the textual information. The estimation of the model parameters is performed via a variational version of the expectation maximization (EM) algorithm. A model selection criterion is formally obtained to estimate the number of partitions. Numerical experiments on simulated data are carried out to highlight the main features of the estimation procedure. Two real-world datasets are finally employed to show the usefulness of the proposed approach.

1 Introduction

In all aspects of everyday life, the recent digitalization of the systems has resulted in a massive generation of data, including text data. For instance, many e-commerce websites (such as Amazon or TripAdvisor) ask their clients to make comments about the products/services they bought. Similarly, major hospitals have fully numerical information systems which allow doctors to directly record surgery reports, biopsy reports or medical prescriptions about their patients. In these examples, the texts are the result of the interactions between individuals of type A (doctors, customers, etc.) and type B (patients, products, etc.). Since datasets have become larger and larger, clustering methods have been proposed as a tool to reduce the dimension and to provide a synthetic view of the available information. Taking the example of customers rating products, instead of focusing on the raw data, it is rather more relevant to look for coherent groups (also known as *clusters*) of both customers and goods. The task of simultaneously clustering the rows and the columns of such an array (here defined by the interactions between individuals and objects) is known as co-clustering. When the interactions are *textual* (e.g. a review), the text can provide significant information to perform a more realistic clustering. For instance, a group of users could review the same goods but with

*This work has been realized when L.R. Bergé was CNRS post-doctoral researcher at Laboratoire MAP5.

different arguments. Unfortunately, a large amount of the existing co-clustering methods do *not* account for the textual information. The aim of this paper is to tackle this issue by providing a new model-based method to co-cluster both individuals and objects while accounting for the textual content of their interactions.

1.1 Model-based co-clustering

Let us consider a binary table with M rows and P columns. A non-null entry at position (i, j) corresponds to an observed interaction between i (individual) and j (object). If no interaction occurred, the same entry is zero. Such table is called *incidence matrix* and can be used to summarize the observed interactions between two disjoint sets of M and P actors/objects, respectively. In the statistics and machine learning literature, methods for the co-clustering of rows and columns of an incidence matrix can be split into two main categories: deterministic approaches (see for instance George and Merugu, 2005; Banerjee et al., 2007; Wang and Huang, 2017) and model-based approaches. The selection of the number of row and column clusters (see Section 3.3) is one of the most important tasks in co-clustering analysis and model-based approaches provide a well defined framework for model selection. Moreover, model-based approaches are usually very flexible: accounting for groups of different sizes, the allow to manage different types of data. These are the main reasons motivating us to adopt the model-based point of view.

Several model-based methods for co-clustering are based on the the latent block model (LBM, Govaert and Nadif, 2003). In its original version, LBM assumes that the rows and the columns of the incidence matrix are clustered in hidden groups. The probability that the entry (i, j) of the matrix is 1 only depends on the row cluster of i and the column cluster of j. The model was extended later to deal with counting data (Govaert and Nadif, 2010), real data (Lomet, 2012), categorical data (Keribin et al., 2015) and ordinal data (Jacques and Biernacki, 2017). Several inference procedures have been proposed for LBM, including likelihood based methods (Govaert and Nadif, 2008), variational inference (Keribin et al., 2012), Bayesian inference (Keribin et al., 2012; Wyse and Friel, 2012) and greedy search approaches (Wyse et al., 2017). A recent algorithm (knows as *Largest Gaps*) of Brault and Channarond (2016) allows to perform clustering and model selection in LBM, only using the marginals (sum of the entries in rows and columns) of the incidence matrix, thus dramatically reducing the computational complexity of the estimation procedure. However, the algorithm is only consistent under certain assumptions concerning the degree distributions of rows and columns.

The contributions mentioned so far do not involve the analysis of textual data. For instance, if the individual i reviews the object j, the standard LBM would neglect the textual content of the review and just consider the entry (i, j) of the incidence matrix (equal to 1). However, the textual content of the review could be the key to perform a more realistic clustering. Before illustrating how we proceed to integrate the text analysis into a co-clustering model based method, we pass through some of the most well known statistical models for text analysis.

1.2 Statistical models for text corpora

One of the first contributions to the statistical modeling of texts is the work of Papadimitriou et al. (1998), based on latent semantic indexing (LSI) (Deerwester et al., 1990). LSI is known in particular for allowing the recovery of linguistic notions such as synonymy and polysemy from "term frequency - inverse document frequency" (tf-idf) data. Hofmann (1999) proposed

an alternative model for LSI, called probabilistic latent semantic analysis (pLSI), which models each word within a document using a mixture model. In pLSI, each mixture component is modeled by a multinomial random variable and the latent groups can be viewed as "topics". Thus, each word is generated from a single topic and different words in a document can be generated from different topics. However, pLSI has no model at the document level and may suffer from overfitting. Notice that pLSI can also be viewed has an extension of the mixture of unigrams, proposed by Nigam et al. (2000). The model which finally concentrates most desired features was proposed by Blei et al. (2003) and is called latent Dirichlet allocation (LDA). The LDA model has rapidly become a standard tool in statistical text analytics and is even used in different scientific fields such has image analysis (Lazebnik et al., 2006) or transportation research (Côme et al., 2014) for instance. The idea of LDA is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA is therefore similar to pLSI except that the topic distribution in LDA has a Dirichlet distribution. Several inference procedures have been proposed for LDA, including variational inference (Blei et al., 2003; Teh et al., 2006) and Bayesian inference, via Gibbs sampling (Phan et al., 2008). More recently Anandkumar et al. (2012); Podosinnikova et al. (2015) introduced new inference procedures based on moment matching and tensorial decomposition.

A limitation of LDA would be the inability to take into account possible topic correlations. This is due to the use of the Dirichlet distribution to model the variability among the topic proportions. To overcome this limitation, the correlated topic model (CTM) was also developed by Blei and Lafferty (2006). From another point of view, other works extend LDA to account for co-clustering of words and documents. In Shafiei and Milios (2006), given each document's specific parameter, words within a document are no longer i.i.d. They still follow a mixture distribution over latent topics depending on the segment (paragraph) of the document. In this way, each document is partitioned in segments that are topic homogeneous. Instead, Wang et al. (2009) use a Bayesian framework to simultaneously cluster documents and words of a document-term matrix, in such a way that topic proportions are no longer specific to each document but to each cluster of documents. This approach is partially related with the one we adopt as will we see in Section 2.3. More recently the work of Wang et al. (2009) was generalized by Kumar et al. (2016) who introduced statistical dependence between row and column clusters of a document-term matrix.

1.3 Contributions and organization of the paper

The present paper introduces a new model, the latent topic block model (LTBM), along with an inference algorithm, to simultaneously cluster the rows and columns of an incidence matrix accounting for the textual information associated with the non-null entries. In other words, a non-null entry (i, j) in the incidence matrix corresponds to a textual interaction between i and j, and the text is part of the generative model we introduce. Moreover, our approach aims at estimating the *topics* used for the textual interactions associated with the incidence matrix, thus allowing a deeper understanding of the co-clustering. In real data applications (Section 5), each topic (or argument) is identified by a list of most representative words. These words are used to decrypt the topic and therefore the clusters.

We stress that our aim is to cluster the rows/columns of an incidence matrix and not those of a document-term matrix. In this sense, (i) the approach introduced here can be seen as an extension of LBM and (ii) the co-clustering we perform is not the same as in Wang et al. (2009), who focus on the co-clustering of the rows and columns of a document-term matrix.

That said, our generative model is related to LDA, as discussed in Section 2.3. Finally, our model-based approach is related to the one introduced in Bouveyron et al. (2016) for network analysis. However, in that paper the scope is to cluster the nodes of a graph, whereas we aim here at co-clustering two disjoint sets of actors.

The present paper is organized as follows. Section 2 introduces the LTBM and details its relationship with other generative models. Section 3 describes the inference procedure adopted to estimate the model parameters. It also discusses further issues such as the initialization of the estimation algorithm and model selection. Section 4 focuses on experiments on synthetic data. The aim is to assess the capacity of the estimation algorithm to recover the true partitions (and the number of groups), when they are known. Finally, in Section 5, two real-world datasets are analysed in order to show the appeal of our methodology. The former dataset is collected from the Amazon e-commerce system, the latter is collected from the PubMed database.

2 The Latent Topic Block Model

In this section, we describe the latent topic block model that we introduce. The observed data is represented by an $M \times P$ incidence matrix A and a corpus of documents W. The observed connections (i.e. non null entries in A) are characterized by documents. If $A_{ij} = 1$, a set of D_{ij} documents W_{ij}^d, $d \in \{1, \ldots, D_{ij}\}$ is associated with the connection between i and j (e.g. the review of i of the good j). Hence $W_{ij} := \{W_{ij}^d\}_{d \leq D_{ij}}$ and $W = \{W_{ij}\}_{i \leq M, j \leq P}$.

2.1 Modeling of connections

Following LBM, we assume that the rows of A (the individuals) are grouped into Q latent *row* clusters. An hidden $M \times Q$ binary matrix Y is introduced such that the i-th row is denoted by Y_i and its entries are all zeros but one. In more details, $Y_{iq} = 1$ if and only if the i-th row of A belongs to the q-th row cluster, where $q \in \{1, \ldots, Q\}$. The rows $Y_1, \ldots, Y_M$ are assumed to be *independent* random vectors such that

$$\mathbb{P}(Y_{iq} = 1) = \rho_q,$$

$\forall i \in \{1, \ldots, M\}$, where $\rho_q > 0$ and $\sum_{q=1}^{Q} \rho_q = 1$. Similarly, the columns of A are grouped into L *column* clusters and an hidden $P \times L$ binary matrix is introduced such that the row X_j is an indicator of the cluster of the j-th column of A. For all $j \in \{1, \ldots, P\}$, we assume that

$$\mathbb{P}(X_{jl} = 1) = \delta_l,$$

where $\delta_l > 0$ and $\sum_{l=1}^{L} \delta_l = 1$. The label matrices Y and X are assumed to be independent. Hence

$$\begin{aligned} p\left(Y, X|\rho, \delta\right) &= p\left(Y|\rho\right) \times p\left(X|\delta\right) \\ &= \prod_{i=1}^{M} \prod_{q=1}^{Q} \rho_q^{Y_{iq}} \times \prod_{j=1}^{P} \prod_{l=1}^{L} \delta_l^{X_{jl}}. \end{aligned} \tag{1}$$

As in LBM, we assume that the probability of a connection between i and j only depends on their clusters. Thus, conditionally on Y and X, A_{ij} is assumed to be a random variable following a Bernoulli distribution

$$p(A_{ij}|X_{iq}Y_{jl} = 1) = \mathcal{B}(A_{ij}; \pi_{ql}) := \pi_{ql}^{A_{ij}}(1 - \pi_{ql})^{1 - A_{ij}}, \tag{2}$$

where $\pi_{ql} \in [0,1]$. We denote by π the $Q \times L$ matrix of the connections probabilities. Conditionally on Y and X, the entries of A are then all assumed to be independent

$$\begin{aligned} p\left(A|Y,X,\pi\right) &= \prod_{i=1}^{M}\prod_{j=1}^{P} p(A_{ij}|Y_i,X_j,\pi) \\ &= \prod_{i=1}^{M}\prod_{j=1}^{P}\left(\prod_{q=1}^{Q}\prod_{l=1}^{L}\left(\mathcal{B}\left(A_{ij},\pi_{ql}\right)\right)^{Y_{iq}X_{jl}}\right). \end{aligned} \tag{3}$$

Finally, the complete data likelihood of the model detailed is

$$p(A,Y,X|\pi,\rho,\delta) = p\left(A|Y,X,\pi\right)p(Y|\rho)p(X|\delta). \tag{4}$$

2.2 Modeling of documents

As pointed out previously, when $A_{ij} \neq 0$, a sequence of D_{ij} documents $W_{ij}^1,\ldots,W_{ij}^{D_{ij}}$ is associated with the interaction between the i-th individual and the j-th object. In LTBM, a document is defined as a set of words extracted from a common dictionary with V words. The number of words in W_{ij}^d (the d-th document sent from i to j) is N_{ij}^d and the n-th word in W_{ij}^d is denoted by W_{ij}^{dn}. As in the latent Dirichlet allocation (LDA, Blei et al., 2003) model, each word within a document follows a mixture distribution over a set of latent topics whose number K is unknown and must be estimated. However, while in LDA the topic proportions of words are specific to each document, in the model we propose, this proportions only depend on the row cluster of the i-th row of A and the column cluster of the j-th column of A.

Thus, we introduce a binary random vector Z_{ij}^{dn}, of length K, such that the k-th entry Z_{ij}^{dnk} is 1 if and only if W_{ij}^{dn} is sampled from the k-th topic, $k \in \{1,\ldots,K\}$, 0 otherwise. Then, conditional on Y_i and X_j, Z_{ij}^{dn} follows a multinomial distribution

$$Z_{ij}^{dn}|\{Y_{iq}X_{jl}A_{ij}=1\} \sim \mathcal{M}\left(1,\theta_{ql}=(\theta_{ql1},\ldots,\theta_{qlK})\right),$$

where $\theta_{qlk} \geq 0$ and $\sum_{k=1}^{K}\theta_{qlk}=1$. Moreover, conditional on Y_i, X_j and θ, the random vectors $Z_{ij}^{d1},\ldots,Z_{ij}^{N_{ij}^d}$ are all assumed to be independent. Therefore

$$\begin{aligned} p\left(Z|A,Y,X,\theta\right) &= \prod_{i=1}^{M}\prod_{j=1}^{P}\prod_{d=1}^{D_{ij}} p\left(Z_{ij}^d|Y_i,X_j,\theta\right)^{A_{ij}} \\ &= \prod_{i=1}^{M}\prod_{j=1}^{P}\prod_{d=1}^{D_{ij}}\left(\prod_{q=1}^{Q}\prod_{l=1}^{L} p\left(Z_{ij}^d|\theta_{ql}\right)^{Y_{iq}X_{jl}}\right)^{A_{ij}} \\ &= \prod_{i=1}^{M}\prod_{j=1}^{P}\prod_{d=1}^{D_{ij}}\left(\prod_{q=1}^{Q}\prod_{l=1}^{L}\left(\prod_{n=1}^{N_{ij}^d}\prod_{k=1}^{K}\theta_{qlk}^{Z_{ij}^{dnk}}\right)^{Y_{iq}X_{jl}}\right)^{A_{ij}}, \end{aligned} \tag{5}$$

where $Z_{ij}^d := (Z_{ij}^1,\ldots,Z_{ij}^{N_{ij}^d})$, $Z_{ij} := \{Z_{ij}^1,\ldots,Z_{ij}^{D_{ij}}\}$ and finally $Z := \{Z_{ij}\}_{i,j}$. For each pair (q,l) of row and column clusters, the model proportions θ_{ql} are themselves assumed to be independent random vectors, each one following a Dirichlet distribution

$$\theta_{ql} \sim \mathcal{D}(\alpha=(\alpha_1,\ldots,\alpha_K)). \tag{6}$$

	Indexes		Variables
i, j	Row/column indicators	Y_i	The *row* cluster indicator of the i-th row of A
q, l	Row/column clusters	X_j	The *column* cluster indicator of the j-th column of A
k	Topics	A_{ij}	The entry (i, j) of the $M \times P$ incidence matrix A
n	Word indicator	W_{ij}^{dn}	The n-th word in the d-th document W_{ij}^{d}
d	Document indicator	Z_{ij}^{dn}	The topic of W_{ij}^{dn}
v	Word v of the vocabulary	θ_{ql}	The topic proportions for the pair of groups (q, l)

	Numbers		Parameters
Q	Number of row clusters	ρ	Multinomial distribution parameters (Y)
L	Number of column clusters	δ	Multinomial distribution parameters (X)
K	Number of topics	π_{ql}	Connection probability between classes q and l
M	Number of individuals/rows	β_k	Multinomial distribution probabilities (k-th topic)
P	Number of objects/columns	α	Dirichlet distribution parameters (θ)
N_{ij}^{d}	Number of words in W_{ij}^{d}		
V	Number of vocables in the vocabulary		

Table 1: Notations in LTBM.

Given Z_{ij}^{dn}, the word W_{ij}^{dn} is finally assumed to be drawn from a multinomial distribution

$$W_{ij}^{dn}|Z_{ij}^{dnk} = 1 \sim \mathcal{M}(1, \beta_k = (\beta_{k1}, \ldots, \beta_{kV})), \tag{7}$$

where $\beta_{kv} > 0$ and $\sum_{v=1}^{V} \beta_{kv} = 1$, for all $k \in \{1, \ldots, K\}$. Henceforth, β denotes the $K \times V$ matrix whose k-th row is β_k. Notice that, unlike θ, the matrix β depends neither on the row clusters nor on the column clusters. Moreover, note that the number K of topics is the same for all cluster pairs. Thus, the following conditional distribution is obtained by independence

$$\begin{aligned} p(W|Z, A, \beta) &= \prod_{i=1}^{M}\prod_{j=1}^{P}\prod_{d=1}^{D_{ij}} p\left(W_{ij}^{d}|Z_{ij}^{d}, \beta\right)^{A_{ij}} \\ &= \prod_{i=1}^{M}\prod_{j=1}^{P}\prod_{d=1}^{D_{ij}} \left(\prod_{n=1}^{N_{ij}^{d}} p\left(W_{ij}^{dn}|Z_{ij}^{dn}, \beta\right)\right)^{A_{ij}} \\ &= \prod_{i=1}^{M}\prod_{j=1}^{P}\prod_{d=1}^{D_{ij}} \left(\prod_{n=1}^{N_{ij}^{d}}\prod_{k=1}^{K}\left(\prod_{v=1}^{V}(\beta_{kv})^{W_{ij}^{dnv}}\right)^{Z_{ij}^{dnk}}\right)^{A_{ij}}. \end{aligned} \tag{8}$$

Finally, the complete-data likelihood for the textual part of the model is obtained by conditioning

$$p(W, Z, \theta|A, Y, X, \beta, \alpha) = p(W|Z, A, \beta)p(Z|A, Y, X, \theta)p(\theta|\alpha). \tag{9}$$

All the notations used for the description of LTBM are given in Table 1. A graphical representation of the model can be seen in Figure 1.

2.3 Links with related models

As pointed out previously, the sampling scheme of the documents is similar to the one of LDA, but not identical. Indeed, assuming that Y and X are known, the joint distribution in (9) can

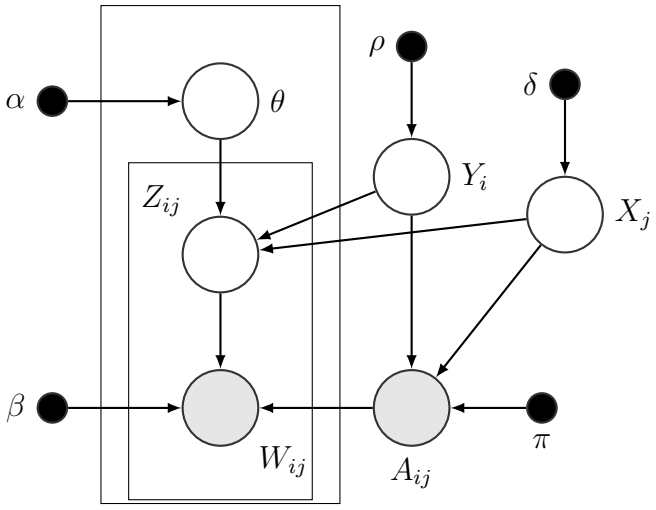

Figure 1: Graphical representation of LTBM.

be written as

$$
\begin{aligned}
p\left(W, Z, \theta | A, Y, X, \beta, \alpha\right) &= p\left(\theta|\alpha\right) \prod_{i=1}^{M} \prod_{j=1}^{P} p\left(W_{ij}, Z_{ij}|Y_i, X_j, \theta, \beta\right)^{A_{ij}} \\
&= p\left(\theta|\alpha\right) \prod_{i=1}^{M} \prod_{j=1}^{P} \left(p\left(W_{ij}|Z_{ij}, \beta\right) p\left(Z_{ij}|Y_i, X_j, \theta\right)\right)^{A_{ij}} \\
&= p\left(\theta|\alpha\right) \prod_{i=1}^{M} \prod_{j=1}^{P} \left(p\left(W_{ij}|Z_{ij}, \beta\right) \left(\prod_{q=1}^{Q} \prod_{l=1}^{L} p\left(Z_{ij}|\theta_{ql}\right)^{Y_{iq}X_{jl}}\right)\right)^{A_{ij}} \\
&= p(\theta|\alpha) \prod_{i=1}^{M} \prod_{j=1}^{P} \prod_{q=1}^{Q} \prod_{l=1}^{L} \left(p(W_{ij}|Z_{ij}, \beta) p(Z_{ij}|\theta_{ql})\right)^{Y_{iq}Y_{jl}A_{ij}} \\
&\equiv \prod_{q=1}^{Q} \prod_{l=1}^{L} p\left(W_{ql}|Z_{ql}, \beta\right) p\left(Z_{ql}|\theta_{ql}\right) p\left(\theta_{ql}|\alpha\right),
\end{aligned}
$$

where $W_{ql} =: \{W_{ij}|A_{ij}Y_{iq}X_{jl} = 1, \forall i, j\}$ is the meta document of all the exchanged documents involving the pair of clusters (q, l). Similarly $Z_{ql} =: \{Z_{ij}|A_{ij}Y_{iq}X_{jl} = 1, \forall i, j\}$ is the set of all the topic labels in the exchanged documents involving the pair of clusters (q, l). Therefore, the joint distribution of W and Z, can be seen from the LDA perspective: there are $Q \times L$ independent documents with their own topic proportions. However, notice that this analogy with LDA crucially hinges on the knowledge of Y and X. Since Y and X are unknown, LTBM is more general than LDA and the related inference more difficult.

When $K = 1$ and a single topic is associated with all the connections, the text analysis does not bring any further information and LTBM reduces to LBM. Finally, as mentioned in Section 1.3, LTBM is related to the stochastic topic block model (STBM, Bouveyron et al., 2016). However, while STBM characterizes textual interactions between actors/objects in the same set, LTBM models textual interactions between two disjoint sets of actors/objects.

3 Inference

The main goal of this section is to detail a variational expectation maximization (VEM, Dempster et al., 1977; Hathaway, 1986) algorithm to estimate the LTBM model parameters and to

provide estimates of Y, X as well as Z. After having introduced the optimization procedure, we focus our attention on the algorithm initialization (3.2) and on the model selection (3.3) task.

3.1 Variational inference

In this section, the values of Q (number of row clusters), L (number of column clusters) and K (number of topics) are assumed to be known. The choice of Q, L and K will be discussed in Section 3.3. In this context, we aim at estimating the hidden labels Y, X, the model parameters (π, ρ, δ) as well as β. In order to simplify both the exposition and the derivation of the results, the parameter α is considered fixed and it is not part of the inference procedure. In all the experiments in Section 4, we set $\alpha_1 = \cdots = \alpha_K = 1$ for all pairs (q, l), thus inducing a uniform distribution over the $(K-1)$-simplex of all θ_{ql}.

We focus on the following log-likelihood

$$\log p(W, A, Y, X|\pi, \rho, \delta, \beta) = \log p(W|A, Y, X, \beta) + \log p(A, Y, X|\pi, \rho, \delta), \tag{10}$$

where the second term on the right hand side of the equality is detailed in (4) and the first one is obtained from (9), by integrating out Z and θ. We aim at maximizing the above log-likelihood with respect to $(\beta, \rho, \delta, \pi)$ and (X, Y). Let us consider $\log p(W|A, Y, X, \beta)$ at first. Although this term is not explicitly tractable, we can take advantage of the following variational decomposition

$$\begin{aligned} \log p\left(W|A, Y, X, \beta\right) = & \int_\theta \sum_Z q\left(Z, \theta\right) \log \frac{p\left(W, Z, \theta|A, Y, X, \beta\right)}{q\left(Z, \theta\right)} \mathrm{d}\theta \\ & - \int_\theta \sum_Z q\left(Z, \theta\right) \log \frac{p\left(Z, \theta|W, A, Y, X, \beta\right)}{q\left(Z, \theta\right)} \mathrm{d}\theta, \end{aligned} \tag{11}$$

where $q(Z, \theta)$ is any distribution over the pair (Z, θ). The sum inside the integral is taken over the set of all the possible outcomes of Z. The last term on the right hand side of the above equation is the Kullback-Leibler (KL) divergence between the approximate and the true posterior distribution of the pair (Z, θ). The KL divergence is known to be positive and null if and only if $q(\cdot)$ is equal to $p(\cdot|W, A, Y, X, \beta)$. As a consequence, the first term on the right hand side of (11) is a lower bound for the integrated log-likelihood $\log p\left(W|A, Y, X, \beta\right)$. We denote it

$$\begin{aligned} \mathcal{L}(q(\cdot)|A, Y, X, \beta) : & = \int_\theta \sum_Z q\left(Z, \theta\right) \log \frac{p\left(W, Z, \theta|A, Y, X, \beta\right)}{q\left(Z, \theta\right)} \mathrm{d}\theta \\ & = \mathbb{E}_{Z,\theta}\left[\log \frac{p\left(W, Z, \theta|A, Y, X, \beta\right)}{q\left(Z, \theta\right)}\right], \end{aligned} \tag{12}$$

where the expectation is taken with respect to (Z, θ) following $q(\cdot)$. Since the posterior distribution of (Z, θ) is not tractable (due to similar arguments as in Blei et al., 2003), the following mean field variational approximation is adopted to specify $q(\cdot)$

$$q(Z, \theta) := q(\theta)q(Z) = q(\theta) \prod_{i=1}^{M} \prod_{j=1}^{P} \prod_{d=1}^{D_{ij}} \prod_{n=1}^{N_{ij}^d} q(Z_{ij}^{dn}), \tag{13}$$

corresponding to an independence assumption over the approximate posterior distribution. The basic idea is then to replace the log-likelihood in (11) by its lower bound

$$\log p\left(W|A, Y, X, \beta\right) \geq \mathcal{L}(q(\cdot)|A, Y, X, \beta)$$

and to use a VEM algorithm (Hathaway, 1986) to maximize the right hand side of the above inequality with respect to $q(\cdot)$ in (13) and β. We stress that this maximization corresponds to minimizing the KL divergence between the approximate and true posterior distribution of the pair (Z, θ). Thus, we can go back to (10) and notice that

$$\begin{aligned} \log p(W, A, Y, X|\pi, \rho, \delta, \beta) &= \log p(W|A, Y, X, \beta) + \log p(A, Y, X|\pi, \rho, \delta) \\ &\geq \mathcal{L}(q(\cdot)|A, Y, X, \beta) + \log p(A, Y, X|\pi, \rho, \delta), \end{aligned} \tag{14}$$

where the term $\log p(A, Y, X|\pi, \rho, \delta)$ is independent from the variational approximation adopted. The proposed estimation procedure is detailed in the following two steps:

1. Y and X being fixed, a VEM algorithm is applied to alternatively maximize the right hand side of (14) with respect to $q(\cdot)$ in (13) (E-step) and the model parameters (π, ρ, δ) and β (M-step), up to convergence.

2. The model parameters $(\pi, \rho, \delta, \beta)$ being fixed, a *greedy* search strategy is adopted to maximize the right hand side of (14) with respect to Y and X. Each row (respectively column) of A is switched from its current cluster to all the other row clusters (resp. column cluster) and the switch leading to the highest increase of the right hand side of (14) is finally retained.

The two steps above are iteratively repeated until convergence. The estimation algorithm detailed alternates a variational EM step with a classification step. Hence, it is referred to as the C-VEM algorithm. It was first used in Bouveyron et al. (2016) and it is built upon the Classification-EM (CEM) algorithm (Celeux and Govaert, 1991). Each step of the C-VEM algorithm is now analyzed in detail.

3.1.1 Variational EM step

In this section Y and X are fixed and we provide the updating formulas for $q(Z, \theta)$ and the model parameters.

Maximization of $\mathcal{L}$ with respect to $q(Z, \theta)$. The E step of the VEM algorithm is given by the following two propositions.

Proposition 1. *The VEM update step for distribution $q(Z_n^{ij})$ is given by*

$$q(Z_{ij}^{dn}) = \mathcal{M}(Z_{ij}^{dn}; 1, \phi_{ij}^{dn} = (\phi_{ij}^{dn1}, \ldots, \phi_{ij}^{dnK})),$$

where

$$\phi_{ij}^{dnk} \propto \left(\prod_{v=1}^{V} \beta_{kv}^{W_{ij}^{dnv}} \right) \prod_{q=1}^{Q} \prod_{l=1}^{L} \exp\left(\psi(\gamma_{qlk}) - \psi\left(\sum_{k'=1}^{K} \gamma_{qlk'} \right) \right)^{Y_{iq} X_{jl}}, \qquad \forall (n, k) \tag{15}$$

where ϕ_{ij}^{dnk} is the approximate posterior probability of word W_{ij}^{dn} being in topic k, γ_{qlk} is defined in the following proposition and $\psi(\cdot)$ denotes the digamma function.

Proof. In Appendix A.1. □

Proposition 2. *The VEM update step for distribution* $q(\theta)$ *is given by*

$$q(\theta) = \prod_{q=1}^{Q} \prod_{l=1}^{L} \mathcal{D}(\theta_{ql}; \gamma_{ql} = (\gamma_{ql1}, \ldots, \gamma_{qlK})),$$

where

$$\gamma_{qlk} = \alpha_k + \sum_{i=1}^{M} \sum_{j=1}^{P} \sum_{d=1}^{D_{ij}} \sum_{n=1}^{N_{ij}^d} A_{ij} Y_{iq} X_{jl} \phi_{ij}^{dnk}, \qquad \forall (q, l).$$

Proof. In Appendix A.2. □

Maximization of $\mathcal{L}$ with respect to the model parameters. The following proposition details the M step of the VEM algorithm providing the estimates of the model parameters $(\pi, \rho, \delta, \beta)$. These are obtained through the maximization of the lower bound in (14) .

Proposition 3. *The estimates of* (β, π, ρ) *and* δ *are given by*

$$\beta_{kv} \propto \sum_{i=1}^{M} \sum_{j=1}^{P} \sum_{d=1}^{D_{ij}} \sum_{n=1}^{N_{ij}^d} A_{ij} W_{ij}^{dnv} \phi_{ij}^{dnk}, \qquad \forall (k, v) \tag{16}$$

$$\pi_{ql} \propto \sum_{i=1}^{M} \sum_{j=1}^{P} Y_{iq} X_{jl} A_{ij}, \qquad \forall (q, l) \tag{17}$$

$$\rho_q \propto \sum_{i=1}^{M} Y_{iq}, \qquad \forall q, \tag{18}$$

$$\delta_l \propto \sum_{j=1}^{P} X_{jl}, \qquad \forall l. \tag{19}$$

Proof. In Appendix A.4. □

3.1.2 Maximization with respect to (Y, X)

The model parameters being estimated in the previous step, a greedy search strategy is now employed to maximize the lower bound on the right hand side of the inequality in (14) with respect to Y and X. Greedy search methods are generally used to solve combinatorial problems that are too computationally intensive to be treated ($Q^M L^P$ combinations to test in our case). However, these methods are not guaranteed to reach a global maximum (for non-convex objective functions). One way to solve this drawback is to run the greedy search several times, with different initializations, in order to choose the estimates leading to the highest value of the objective function.

Let us consider Y at first and assume that an initial clustering of the rows of A in Q cluster is provided (see Section 3.2 for more details). If the i-th row is in the q-th row cluster, the algorithm assesses the increase/decrease in the lower bound due to switching i to the cluster q', for each $q' \neq q$. The switch (if any) leading to the highest increase of the lower bound is actually performed and the entire routine is iteratively applied to all the rows of A until no further increase of the lower bound is possible. The maximization with respect to X is performed similarly.

Note that Wyse et al. (2017) used a greedy search approach similar to the one described above to perform inference in bipartite graphs using LBM. However, the greedy search algorithm they introduced is also used to perform model selection by maximizing the complete

data integrated log-likelihood of LBM. In our case, the number of clusters is assumed fixed during the maximization and a row (column) of A which is alone in its own cluster cannot be switched.

3.2 Initialization strategy

EM-like algorithms are known to be sensible to the initialization and are not guaranteed to converge toward a global maximum. A possible strategy consists in choosing several random initializations and finally retaining the estimate leading to the highest value of the lower bound of the observed data log-likelihood. Alternatively, the algorithm can be initialized relying on other (simpler) clustering algorithms (k-means, spectral clustering, etc.) hoping that the initialization provided is not so far from the global optimum. See Biernacki et al. (2003) for a further inspection of this point.

In our case, we need to get initial estimates of Y and X in order to perform the variational step detailed in Section 3.1.1. When testing the C-VEM algorithm on synthetic data, in Section 4, several initializations are compared. One of them proves to work very well and it is now described. Henceforth, this initialization is referred to as the *spectral* initialization strategy. Consider at first the label vector Y:

1. LDA is run on the whole set of documents W. We recall that each set of documents W_{ij}^d corresponds to a non null entry of A. Via LDA, we can estimate the main topic discussed in each document set $W_{ij} = \{W_{ij}^1, \dots, W_{ij}^{D_{ij}}\}$. Hence, an $M \times P$ matrix T is obtained such that $T_{ij} = k$ if $A_{ij} = 1$ *and* k is the main topic appearing in W_{ij}. If $A_{ij} = 0$, then $T_{ij} = 0$.

2. An $M \times M$ similarity matrix S is created such that its entry (i, i') is defined as

$$S_{i,i'} = \sum_{j=1}^{P} A_{ij} A_{i'j} \mathbf{1}_{\{T_{ij}=T_{i'j}\}},$$

   where $\mathbf{1}_{\mathcal{I}}$ denotes the indicator function on a set $\mathcal{I}$. The above equation states that if i and i' have a common connection j and they share the same main topic associated with this connection, then their similarity increases.

3. The spectral clustering algorithm can be used to produce an estimate of Y based on the graphs Laplacians associated with S (see von Luxburg, 2007, for a detailed review of graphs Laplacians and the spectral clustering algorithm).

An initial estimate of X can be obtained similarly and the VEM step (Section 3.1.1) can be implemented.

3.3 Model selection

The following proposition details a model selection criterion to estimate the number Q of row clusters, the number L of column clusters as well as the number K of topics, from the data.

Proposition 4. *A ICL criterion for LTBM can be obtained*

$$\begin{aligned} ICL_{LTBM} = &\max_{\beta} \mathcal{L}(q(\cdot)|A, Y, X, \beta) - K\frac{V-1}{2}\log(QL) \\ &+ \max_{\pi,\rho,\delta} \log p(A, Y, X|\pi, \rho, \delta, Q, L) - \frac{QL}{2}\log(MP) - \frac{Q-1}{2}\log M - \frac{L-1}{2}\log P. \end{aligned} \tag{20}$$

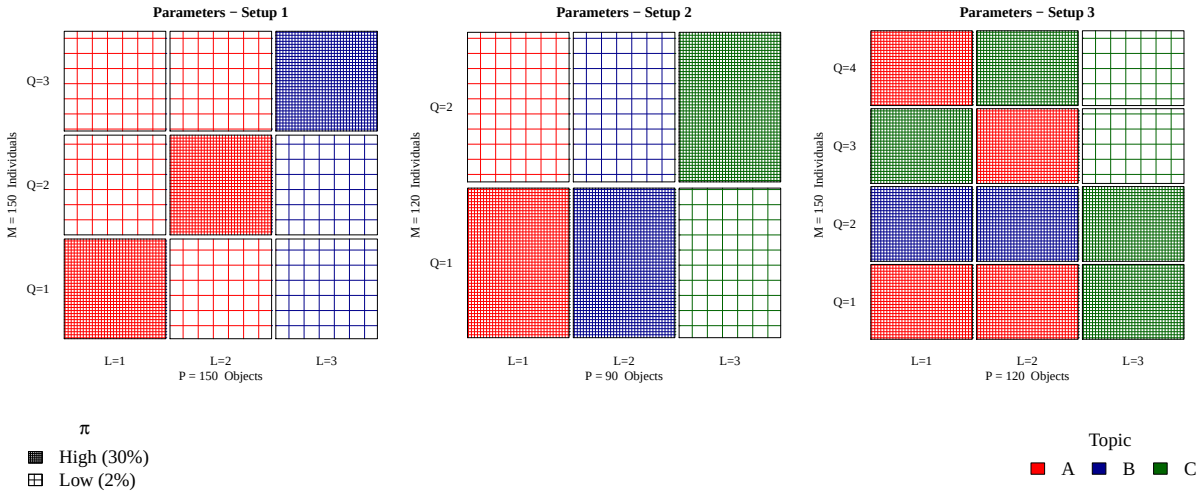

Figure 2: Schematic representation of the model parameters in each setup. Each image represents the parameters π and θ. The density of the coloured grids represents the probability of connections. All the individuals/objects are uniformly split into the row/column clusters. Different colors are associated with different topics.

Proof. In Appendix A.5. □

This result relies on two Laplace approximations, the Stirling formula and a variational estimation. In particular, the first term $\mathcal{L}(q(\cdot)|A, Y, X, \beta)$ on the right hand side of the equality is defined in (12). For a detailed description of the Laplace and Stirling approximations to obtain ICL we recommend the original paper of Biernacki et al. (2000).

4 Experiments on synthetic data

In this section, we carry out some experiments on simulated datasets to test the estimation strategy detailed in the previous section.

4.1 Synthetic datasets

Three different setups are considered. In each setup, an individual i connects with an object j either with a higher probability of 30% or with a lower probability of 2%. As detailed in Section 1.1, these probabilities only depend on the clusters of the pair (i, j). Once i and j are connected, a document is sampled and associated with the connection. The words in each document are extracted from three texts from BBC news. One text is about black holes in astrophysics (A), the second is about the birth of Princess Charlotte (B) and the third one focuses on UK politics (C). Notice that the three texts have been chosen different enough and specific, such that they can be considered as "pure" topics. The number of words in a document is drawn from a Poisson distribution of mean 100. The topic proportions in each text then only depend on the clusters of the corresponding pair (i, j).

The three setups are illustrated in more details in Figure 2. For instance, the first picture on the left hand side represents the connectivity matrix π in the first setup, where the number of row clusters is equal to the number of column clusters ($Q = L = 3$). Denser colored grids on the diagonal represent the higher connection probabilities of 30%. Two colors are

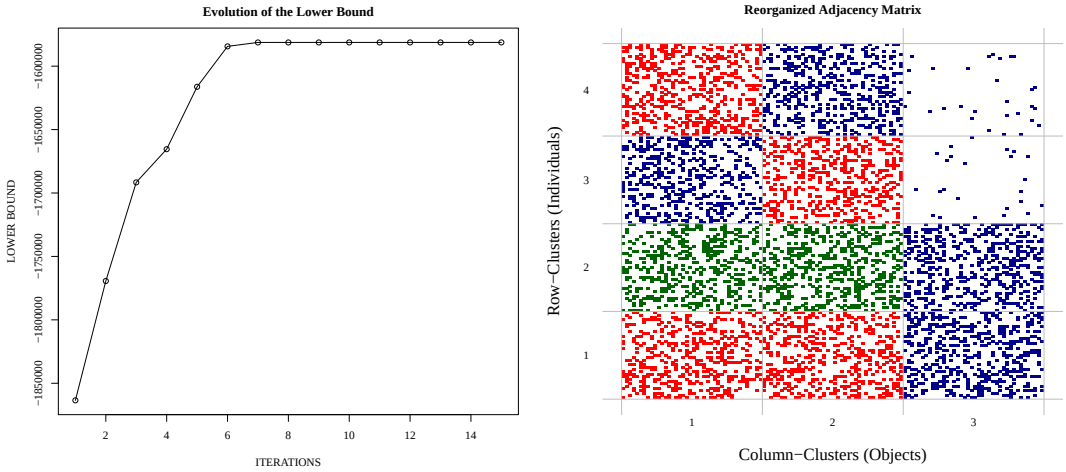

Figure 3: LTBM is fitted on a dataset simulated according to Setup 3. The evolution of the lower bound during the optimization process can be seen on the left. In the colored adjacency matrix on the right, rows and columns are reorganized according to the final clustering provided by the C-VEM algorithm.

associated with two different topics: (A) in red and (B) in blue. Notice that, relying solely on the textual analysis, it would be impossible to distinguish the three row clusters. Similarly, the first and the second column cluster would look as a single group.

An illustration. We now consider the third setup. An incidence matrix with $M = 150$ individuals and $P = 120$ objects is simulated according to the parametrization in Figure 2. The C-VEM algorithm is provided with the true values of K as well as Q and L, used to assign each row (respectively column) to its row (column) cluster. An illustration can be seen in Figure 3. On the left hand side, we can see the lower bound at each step of the C-VEM algorithm. The right pane of the same figure shows the reorganized adjacency matrix, in which rows and columns are permuted according to the clustering provided by the algorithm. As it can be seen, this matrix looks very similar (up to label switching) to the one on the right hand side of Figure 2.

4.2 Initialization performance

The aim of the present section is to compare the clustering results provided by the C-VEM algorithm when different initializations are adopted. The datasets we use are sampled according to the setups illustrated in the previous section. The C-VEM algorithm is always provided with the actual values of Q, L as well as K, and four initialization techniques are tested:

1. **Random.** Each row (resp. column) of the incidence matrix is randomly assigned to a row (column) cluster.

2. **k-means.** The k-means algorithm is applied to the rows (resp. columns) of the incidence matrix to estimate Y (X).

3. **Spectral.** The initialization strategy described in Section 3.2.

4. **LBM.** LBM is fitted on the incidence matrix to provide initial estimates of Y and X.

The quality of the estimates provided by the C-VEM algorithm is assessed via the adjusted Rand index (ARI, Rand, 1971). This index compares two partitions (for example the true Y and its estimate $\hat{Y}$ provided by C-VEM) and measures how close they are. More specifically, the ARI takes real values in $[0, 1]$, where a value of 1 means that the two classifications are identical (up to label switching).

Twenty datasets are simulated for each setup. The results are illustrated in Figure 4 which also reports statistics concerning the lower bound. The LBM initialization works well in situations where all the available information is encoded in π and the analysis of the textual content does not bring any further insight. In Setup 1, for example, the row and column clusters can be detected by solely looking at the interaction frequency and LBM is perfectly able to uncover the hidden partitions. In Setup 2, the column clusters 1 and 2 would be indistinguishable based on the interaction frequency. However, despite the LBM initialization, C-VEM does not remain trapped into local maxima and the ARI is still good in rows and columns. However, in Setup 3, the LBM initialization penalizes heavily the C-VEM algorithm.

Interestingly, the random initializations can lead to good results in all scenarios. However, the variance of the clustering results remain high. Finally, we clearly see that the spectral initialization outperforms its competitors: the ARI is 1 most of the time.

4.3 Model selection

So far, we have assumed that Q, L and K were known in advance. In this section, several datasets are sampled according to the different setups and the C-VEM algorithm is run on each dataset for different values of (Q, L, K). The model selection in (20) is then used to estimate the number of clusters/topics and the aim of this section is to assess how well it works.

For each setup, 50 independent datasets are generated. The C-VEM is run for all values of Q, L and K ranging from 1 to 10. Table 2 reports the results of the model selection. As it can be seen, the criterion selects the actual model most of the time and when it fails, the number of clusters/topics is misclassified by one unit (except for one case in Setup 1).

5 Real data

In the following, in order to illustrate the appeal of the methodology outlined in this paper, LTBM is fitted on two real datasets.

5.1 Amazon Fine Foods

This section focuses on a real dataset consisting of reviews of fine foods from Amazon. The dataset can be freely downloaded at the following address https://snap.stanford.edu/data/web-FineFoods.html. A time horizon of 10 years is considered, up to October 2012. The number of reported reviews is 568,464 and in the original dataset, each row corresponds to one review. Some additional information is reported for each review: the user/product numerical identifiers, a summary of the review and a rating attributed to the product by the user. The rating is expressed via an integer number spanning from 1 (very bad) to 5 (very good). The original dataset was preprocessed as follows. To focus on the most meaningful part of the data, we only considered the users reviewing more than 20 times and the products being reviewed more than 50 times. Each review was preprocessed in a classical way: very short words (less than three characters) and stop words were removed and punctuation and

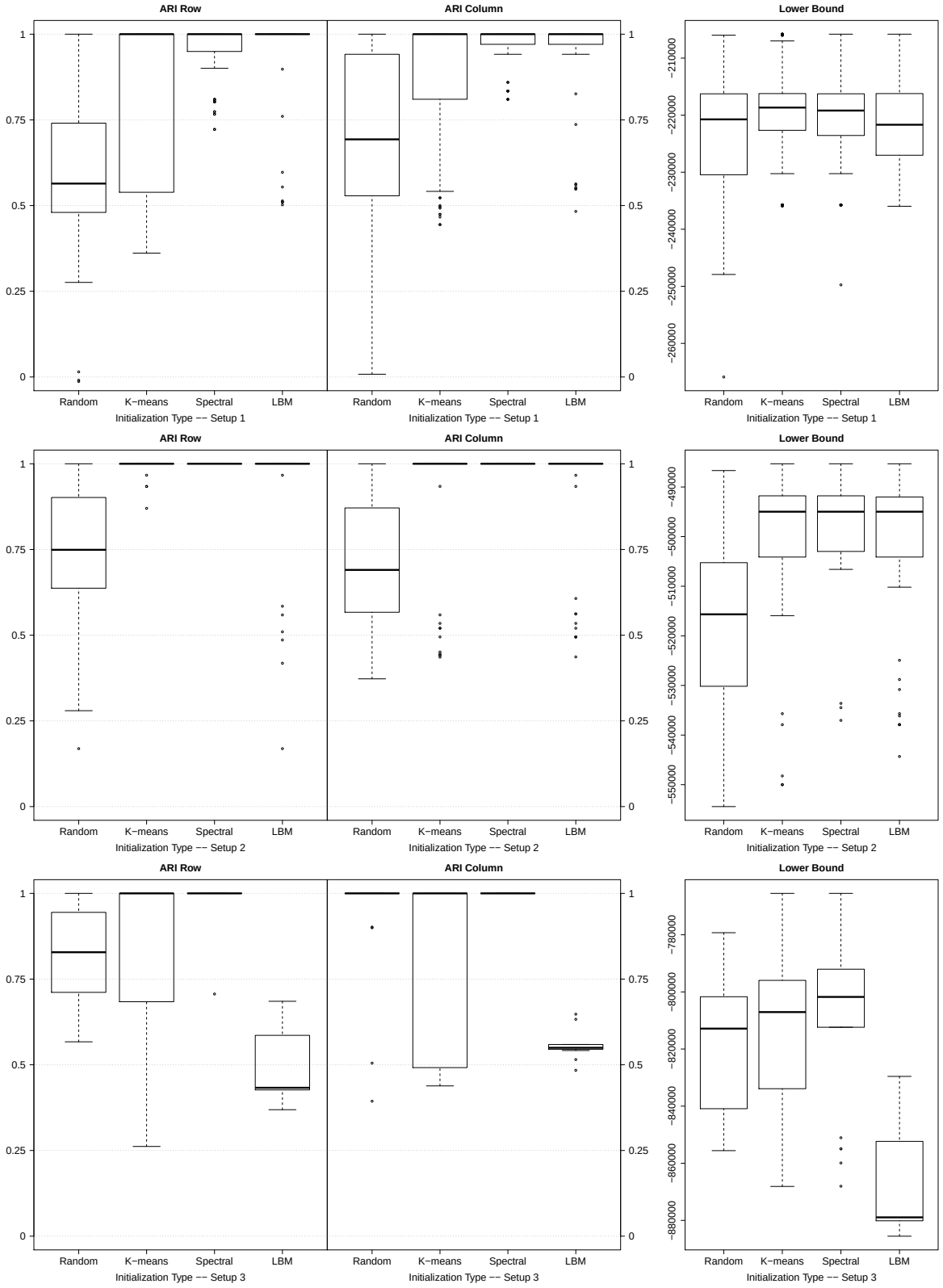

Figure 4: Four initialization techniques are compared on datasets sampled according to Setups 1,2 and 3. Twenty datasets are sampled for each setup.

Dataset Type: Setup 1

	K = **2**					K = 3			
$Q\backslash L$	2	**3**	4	5	$Q\backslash L$	2	3	4	5
2	3	0	2	0	2	1	0	0	0
3	0	43	2	0	3	2	0	0	0
4	0	2	0	0	4	0	0	0	0
5	0	0	0	0	5	0	0	0	0

Dataset Type: Setup 2

	K = **3**					K = 4			
$Q\backslash L$	2	**3**	4	5	$Q\backslash L$	2	3	4	5
2	0	47	0	0	2	0	2	0	0
3	0	1	0	0	3	0	0	0	0
4	0	0	0	0	4	0	0	0	0
5	0	0	0	0	5	0	0	0	0

Dataset Type: Setup 3

	K = **3**					K = 4			
$Q\backslash L$	2	**3**	4	5	$Q\backslash L$	2	3	4	5
2	0	0	0	0	2	0	0	0	0
3	0	0	0	0	3	0	0	0	0
4	0	47	0	0	4	0	2	0	0
5	0	0	1	0	5	0	0	0	0

Table 2: Model selection. The numbers in bold are the actual values of the parameters Q, L and K.

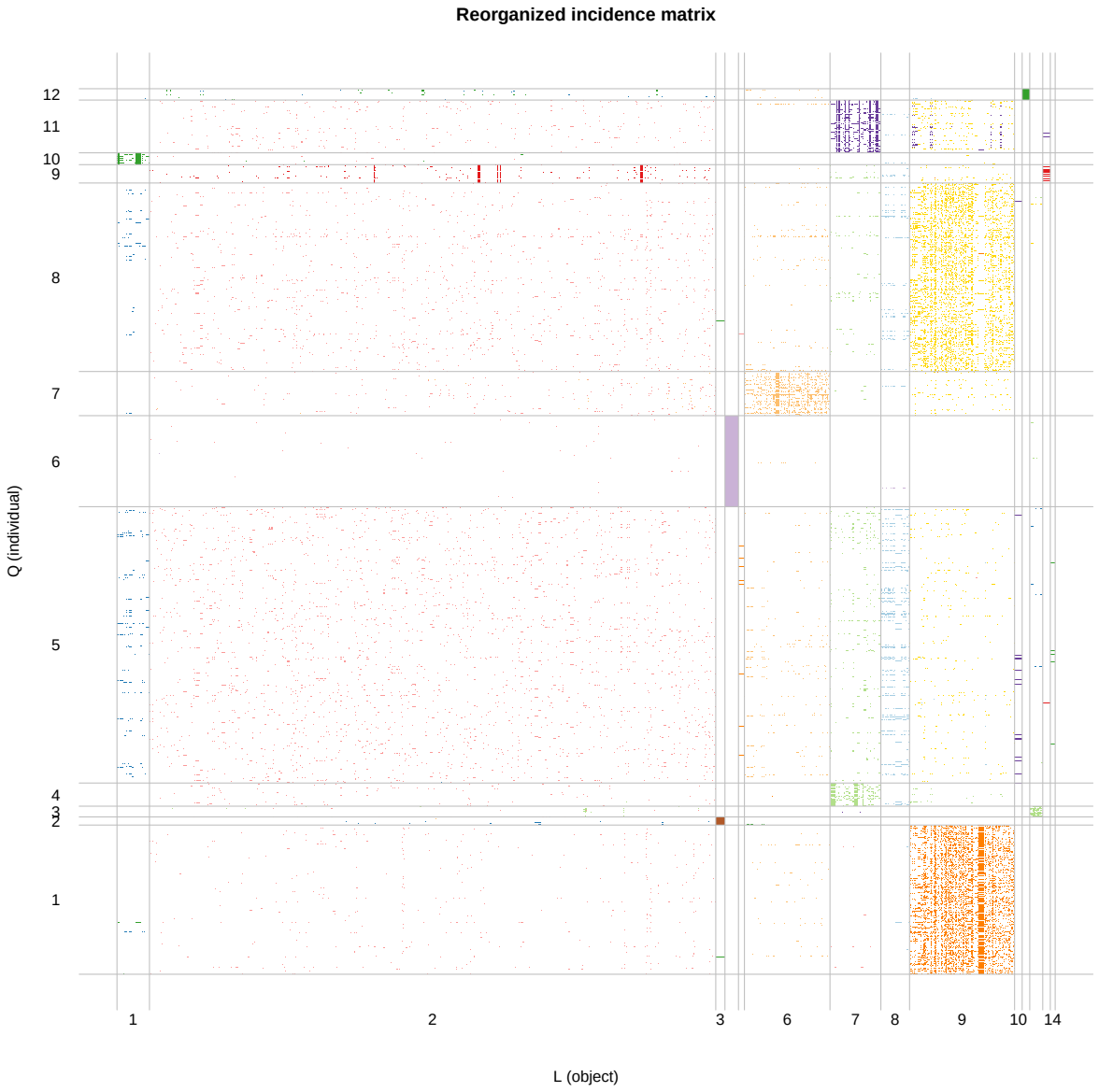

Figure 5: The Reorganized incidence matrix where nearby rows/columns belong to the same row/column cluster (delimited by grey lines). The colors of the cells mark the main topic used for the corresponding reviews. One colored dash marks an interaction/review between the corresponding pair.

numbers ignored. Moreover, the vocables appearing less than 10 times in the whole corpus were neglected. The main features of the resulting dataset are summarized in Table 3.

Number of users (M)	1644
Number of objects (P)	1733
Number of reviews	32836
Number of vocables (V)	11151

Table 3: Amazon fine food dataset statistics.

The C-VEM algorithm was run on the dataset for several initializations and we tested all values of Q, L and K in a range from 1 to 20. The model selection criterion in (20) finally

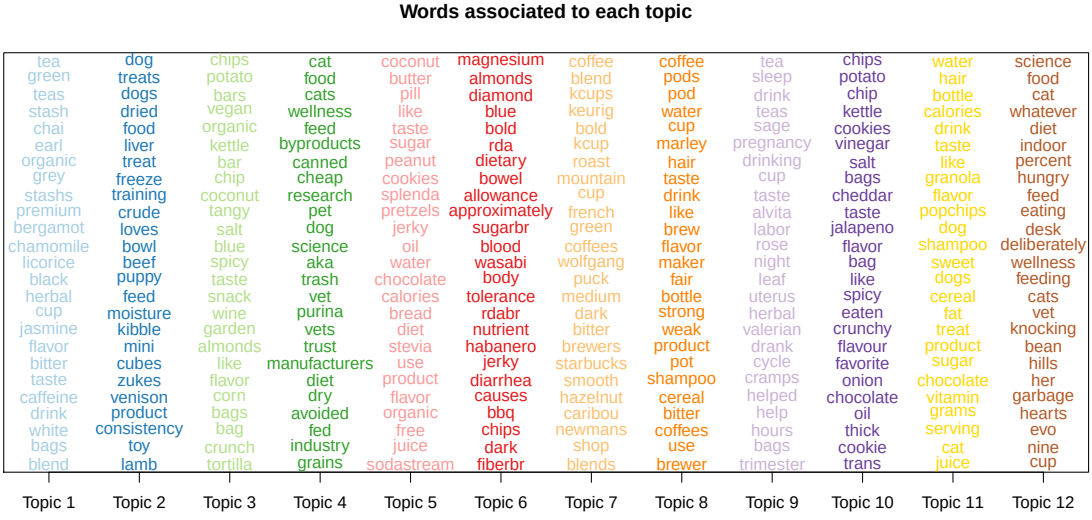

Figure 6: A list of the most representative words of each topic.

selected $Q = 12$ row clusters, $L = 14$ column clusters and $K = 12$ topics. The estimates of Y and X allowed us to permute the rows and the columns of the incidence matrix of the dataset, in such a way that nearby rows/columns belong to the same row/column cluster. The reorganized incidence matrix can be observed in Figure 5. The grey grid (horizontal and vertical lines) delimits the blocks estimated by the C-VEM algorithm. In the matrix, each colored dash corresponds to one user reviewing the corresponding object. The color of each block marks the main topic used for the reviews inside the block. A list of the most representative words of each topic can be seen in Figure 6. So, for instance the orange block at position $(1, 9)$ in the reorganized incidence matrix corresponds to the reviews that the users in the row cluster $q = 1$ made about the objects in the column cluster $l = 9$. The main topic used for such reviews is the orange one, namely Topic 8 whose key words are "coffee", "pods", "pod", etc. Figure 7 displays the estimated topic proportions θ_{ql} for each meta document associated with the pair (q, l). For instance, the entry $(12, 11)$ is blue and green, meaning that reviewers in row cluster $q = 12$ used Topics 2 and 4 (respectively) to review the items in column cluster $l = 11$. Topic 4 is the most used. In general, the denser the colored grids inside a cell, the higher the interaction frequency.

We now move to some (non-exhaustive) remarks about the results shown in the three figures.

1. The buyers in the row cluster $q = 2$ are mainly involved in reviewing a specific brand of cat food. Indeed, all the reviewed products in the column cluster $l = 3$ belong to the same brand. The mean rating for this pair is 2.51 but the mode is 1, corresponding to 289 very bad ratings. All the negative reviews are from different reviewers and contain the following sentence "Filler food is empty, leaves your cat always needing more". Notice that the most representative words of Topic 12 (brown) include "cat", "food" and "hungry". Since it is not credible that 289 different users adopt the very same sentence, we think that they are robots or anyway a fake.

2. In a similar fashion, the buyers in the row cluster $q = 6$ are mainly involved in reviewing a specific brand of tea, whose name "alvita" appears between the most representative

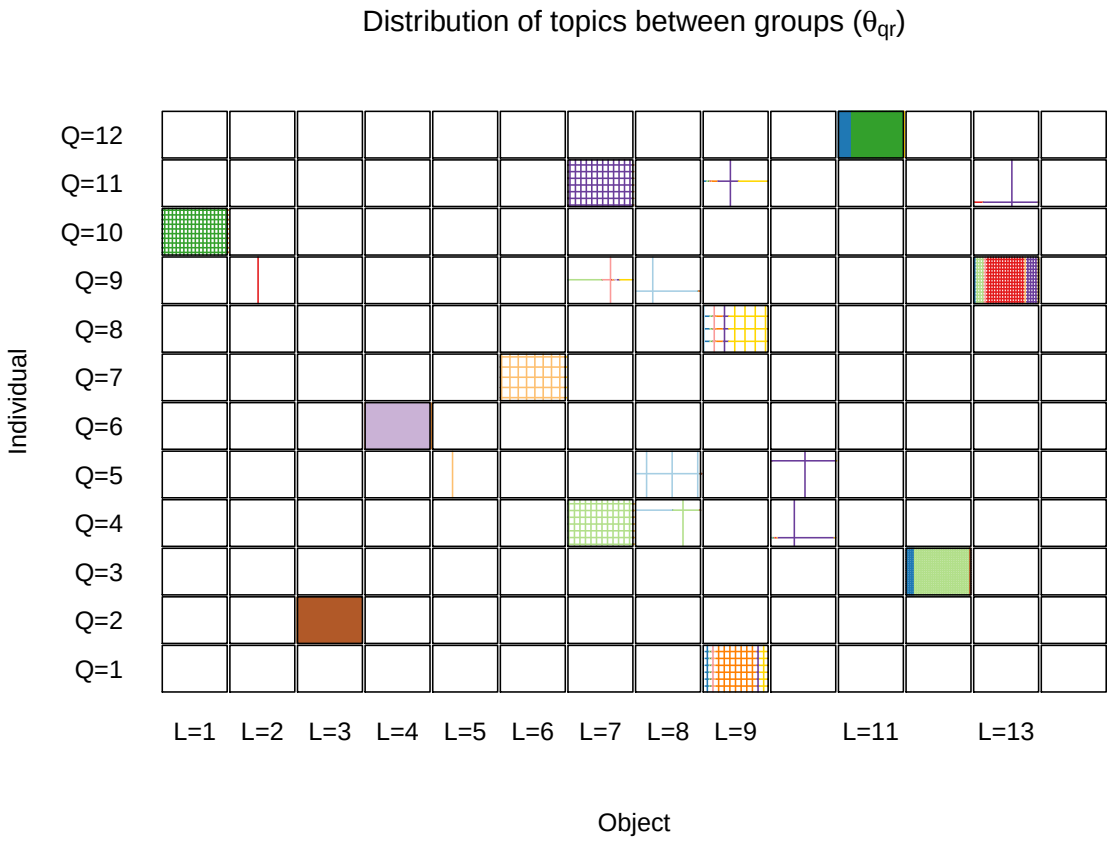

Figure 7: A graphical representation of the estimated topic proportions θ_{ql} for each meta-document. The denser the colored grids, the higher the frequency of the interactions/reviews.

words of Topic 9 (light violet). The products of this brand are grouped in the column cluster $l = 4$. The reviews are very positive and the most common rating is 5 (60 % of times). In this case, distinct users provide different reviews for each product of the brand, but strangely each reviewer gives the very same review for each product.

3. The column cluster $l = 11$ groups dog food items of the same brand "wellness". These products are *only* reviewed by the users in the row cluster $q = 12$, with very positive ratings. In this case too, the reviews for the same product are all different but the single review of each user is spread to all products of the brand. As it can be seen in Figure 7 there are two topics involved in these reviews: Topic 2 and Topic 4. Both of them contain the two key words "dog" and "food". Moreover, Topic 4 contains the key word "wellness" which is the brand of the dog food. Interestingly the words "cat" appears among the most representative words of Topic 4. This is due to the presence of some reviews in which reviewers discuss about differences between the dogs and the cats diet.

4. The first row cluster $(q = 1)$ consists of 277 buyers mainly reviewing the products in the ninth column cluster $(l = 9)$. The reviews associated with this pair of clusters are the 40% of the total number of reviews. The scores are globally positive with a mean value of 3.88. The reviewed products are foods and beverages. The most commented three items are an Italian style coffee drink, an energy drink and a Ginger lemon beverage drink. Topic 8 (orange) is the main topic associated with this pair of clusters and it

contains some key words like "coffee pods", "drink", "flavor", "taste", "like" and "fair". Interestingly, the items in the column cluster $l = 9$ are plenty reviewed by the users in row cluster $q = 8$. This cluster is made of 850 actors who globally behave as the users in cluster $q = 1$ except for the main topic used to review goods in columns cluster $l = 9$. Indeed, the main topic used for these reviews by the users in the row cluster $q = 1$ is Topic 8 whereas the main topic used by the reviewers in the row cluster $q = 8$ is Topic 11. These two topics are quite similar. However, Topic 11 seems to be more food oriented. For example the two key words "granola" and "popchips" are related with two brands producing chocolate and chips, respectively. These two words are not key words in Topic 8, which conversely has "marley" has a key word, which is a coffee brand. Notice that, if colors were removed the row clusters 1 and 8 would be indistinguishable and a LBM could fail to detect them both. This is an useful application of LTBM.

5. The row cluster $q = 9$ groups 34 users. This group is peculiar for two reasons

   i) It is the only group to massively use Topic 6 (red).

   ii) Its users are the main reviewers of the items in the column cluster $l = 13$.

   All the items in the column cluster $l = 13$ are foods of the same brand. More precisely, the items are packs of beef "jerky" (dried up) meat. There are 11 distinct reviews of any single product, but each review is repeated for all the products. The reviewers in group $q = 9$ also reviews the items in the column cluster $l = 2$. The main items that they review are packs of almonds of the same brand, whose name is "Blue Diamond". Notice that the three words "almond", "blue" and "diamond" appear as key words in Topic 6 (Figure 6). Globally, the reviews formulated by the users in the row cluster $q = 9$ are very positive.

6. The reviewers in the row cluster $q = 11$ are the main users of Topic 10. The first key words of this topic are "cheese", "potato", "kettle". Indeed, the most reviewed items in the column cluster $l = 7$ are the Kettle chips (several items of the same brand). Interestingly, the same objects are also reviewed by the users in the row cluster $q = 4$. This explains why the same key words appear on top of Topic 3 in Figure 6. However, the buyers in the row cluster $q = 11$ also review chocolate cookies (for instance), which is not the case for the row cluster $q = 4$. Indeed, "cookies" and "chocolate" are key words of Topic 10 but not of Topic 3.

5.2 PubMed data

This section focuses on a dataset extracted from the National Center for Biotechnology Information (NCBI) databases, via the free search engine PubMed (https:www.ncbi.nlm.nih.gov/pubmed/). The CRAN package RISmed (https://cran.r-project.org/package=RISmed) was used by the statistical software R to query PubMed about the keyword "colorectal", over a time horizon spanning from 2012 to 2016. As a result, we obtained a list of scientific articles, published during the selected time horizon, about colorectal pathologies. In order to capture the most significant information, the list of articles was pre-processed in order to remove the authors submitting less than 10 articles and the journals publishing less than 20 times over the considered time period. Moreover, only the abstracts written in English were taken into account. Finally, we obtained a table of 111 117 rows and 3 columns.

The data in the table previously described can be modelled via LTBM in the following way. Each author corresponds to an individual and one journal to an object. Overall, there are

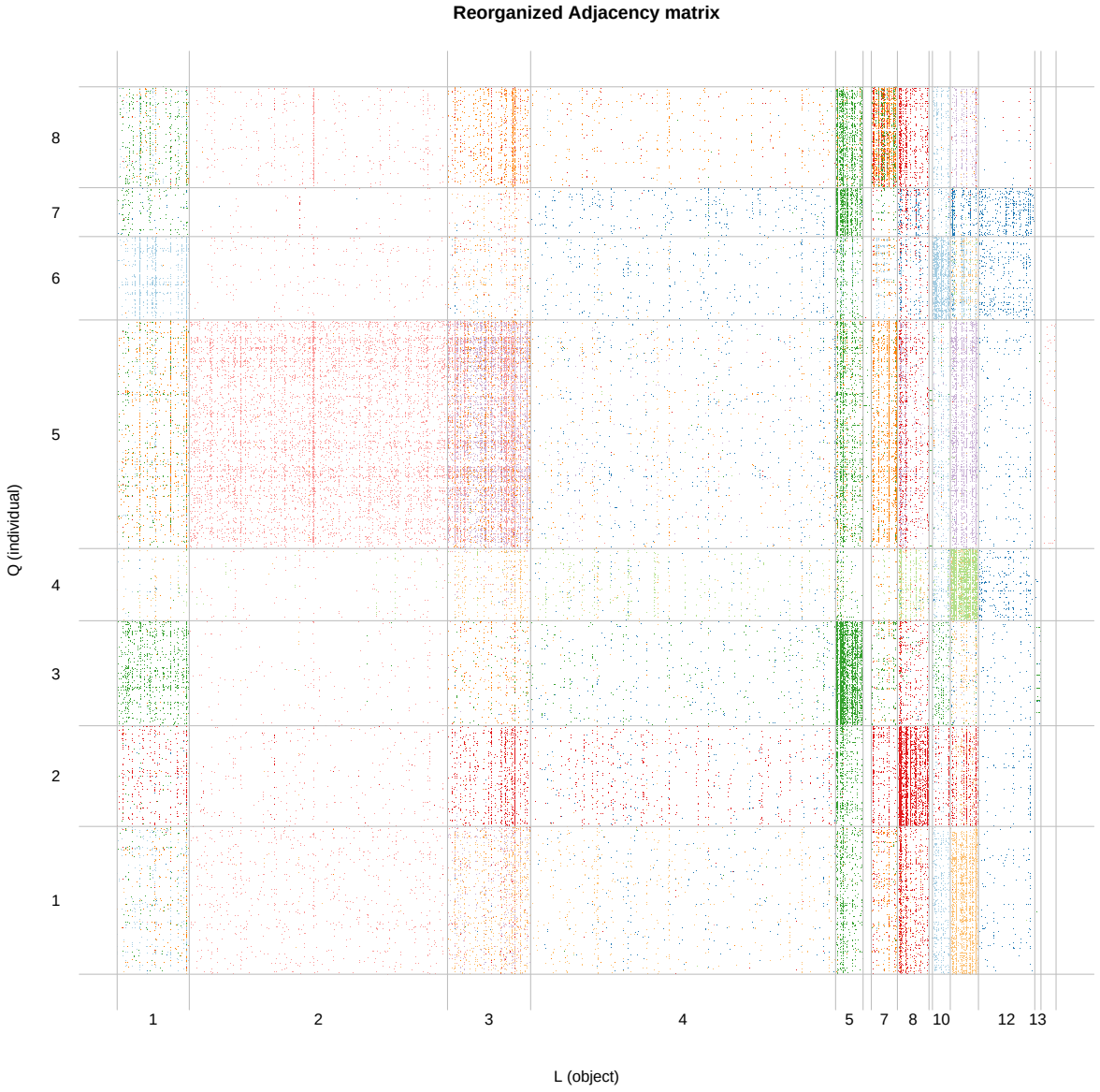

Figure 8: The Reorganized incidence matrix where nearby rows/columns belong to the same row/column cluster (delimited by grey lines). The colors of the cells mark the main topic used for the corresponding article. One coloured dash marks an interaction/article involving the corresponding pair.

$Q = 4918$ authors/individuals and $P = 1367$ objects/journals. If the i-th author published (one or more articles) on the j-th journal, the corresponding entry (i, j) of a $Q \times P$ incidence matrix was set to one, zero otherwise. The abstracts of the articles for the pair (i, j) were collected into W_{ij} and used for the textual analysis. Overall, 74 817 abstracts were analysed, corresponding to $V = 21\ 014$ different vocables. The C-VEM algorithm for LTBM was run on the data for several values of Q, L and K and several initialisations. Finally the algorithm selected $Q = 8$ row clusters, $L = 14$ column clusters and $K = 9$ topics. The estimates provided for Y and X were used to permute the rows and columns of the incidence matrix in order to obtain the reorganized incidence matrix in Figure 8, whose interpretation is the same as discussed in the previous section. Similarly, the most representative words for each topic can be seen in Figure 9. Some remarks about these figures can be done starting from the

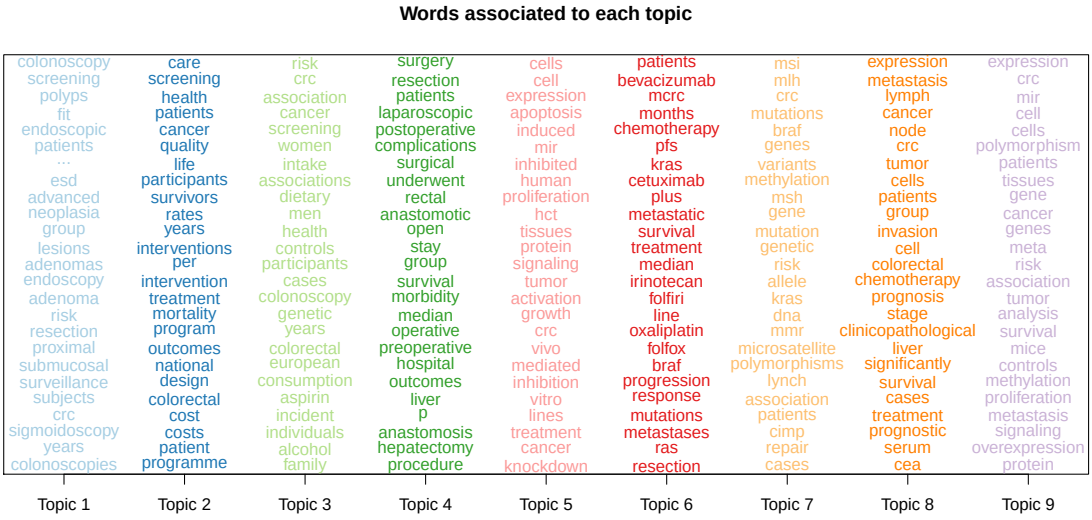

Figure 9: A list of the most representative words of each topic.

topic analysis.

Topic 1. Containing some key words like "colonoscopy", "screening" and "polyps" this topic is clearly related with prevention. Moreover it is widely correlated with the articles published in journals belonging to the column cluster $l = 10$. In particular it is the main topic related with the articles of the cluster pair $(q = 6, l = 10)$.

Topic 4. The key words are "surgery", "resection", "postoperative" and it is related with *surgical treatment*, *preparation* and *monitoring* for colorectal diseases. As in can be seen in Figure 8, Topic 4 is mainly associated with the column cluster $l = 5$ which contains journals like "Annals of surgical oncology", "Surgical endoscopy" and "Annals of surgery".

Topics 6 and 8. Topic 6 focuses on chemoterapy and other remedies for metastatic colorectal cancer (mCRC). In particular, "Bevacizumab", "Folfiri" and "Folfox" are names of specific drugs used to treat mCRC. Topic 8 also exhibits the words "chemoterapy" and "metastasis" and seems to be very closed to Topic 6 but no specific drug name is reported and this seems to be the main difference between the two topics. Moreover, Topic 6 is often the main topic in the journals in column cluster $l = 8$ (including for instance the "British journal of cancer", "Annals of oncology" and other American or European journals) whereas Topic 8 is central in column cluster $l = 7$ ("Oncology reports", "Anticancer research" and other Asiatic journals).

Topics 5 and 9. These two topics characterize articles on fundamental biology, with experiments "in vitro" on cells ("cell", "lines"). However, while Topic 5 seems to focus on signaling pathways in cells ("signal", "activation", "growth", "proliferation", "expression"), Topic 9 is more involved the role of mi-RNA ("mir") and some experiments are carried out on "mice". It is interesting to observe that these two topics are mainly used by authors in row cluster $q = 5$ but in different journal clusters: Topic 5 in column cluster $l = 2$ and Topic 9 in column cluster $l = 3$ (see Figures 8 and 10).

Topic 7. This topic is certainly related with *genetics*, which play a crucial role in colorectal diseases, especially the colorectal cancer (CRC). Not only the words "genes", "gene", "genetic" appear as key words, but also "msi" and "mlh" are technical abbreviations related with gene expressions or DNA repair. Topic 7 is central in the articles submitted by authors in row cluster $q = 1$ and published in journals in column cluster $l = 11$ (Figure 8). However, unlike Topic 4, is is not associated with one particular column cluster.

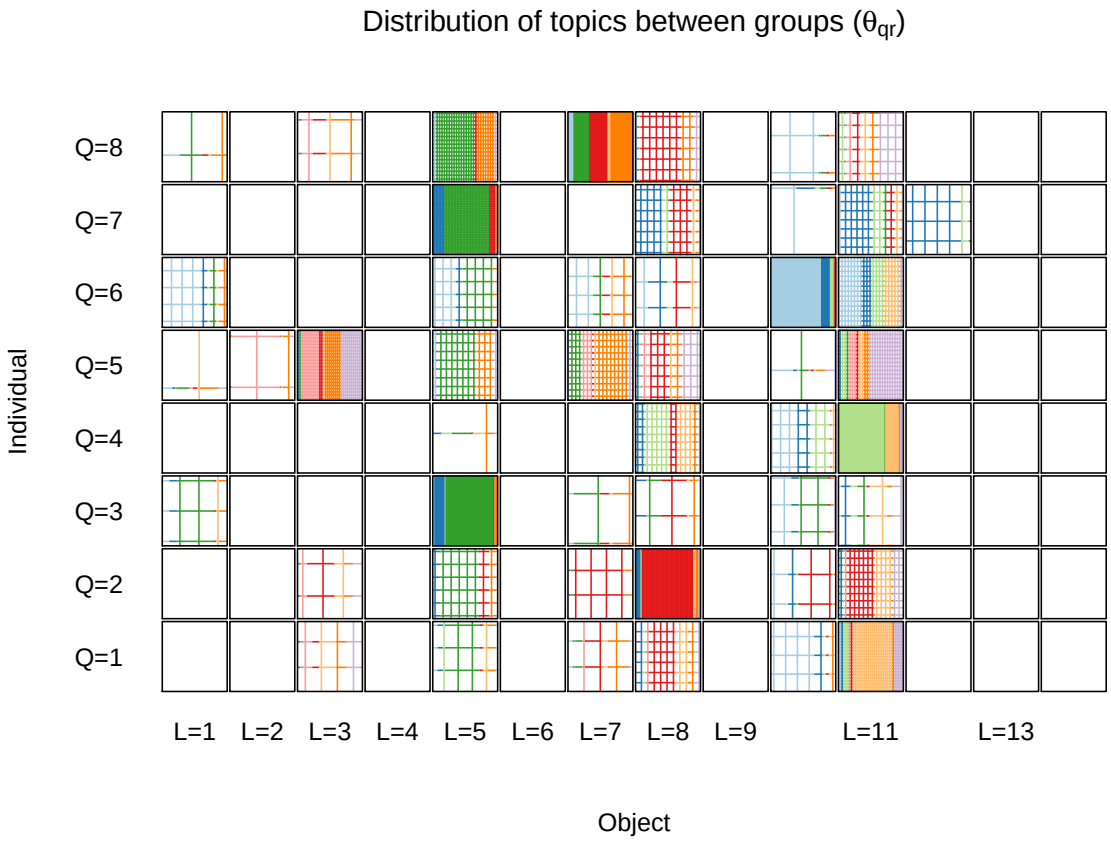

Figure 10: A graphical representation of the estimated topic proportions θ_{ql} for each meta-document. The denser the colored grids, the higher the frequency of the interactions/published articles.

6 Conclusion

This paper introduces a novel model-based co-clustering method to simultaneously cluster the rows and columns of an array of textual interaction data. The peculiarity of the model that we propose, called latent topic block model (LTBM) is the ability to account for both the presence/absence of interactions (as in a standard LBM) *and* the textual content associated with the interactions. We detailed an inference procedure to estimate the model parameters and obtained a model selection criterion to select the number of row clusters, columns clusters and discussed topics. Experiments on both simulated and real datasets were used to assess the appeal of the proposed methodology. In particular, we showed that LTBM allows a fine and improved understanding of two complex real-world datasets such as the Amazon reviews and the PubMed scientific publications.

A Appendix

A.1 Proof of Proposition 1

For all pairs (i, j) such that $A_{ij} = 1$, the update step for $q(Z_{ij}^{dn})$ (for all n) is given by

$$\begin{aligned}
\log q(Z_{ij}^{dn}) &= \mathbb{E}_{Z^{\setminus i,j,d,n},\theta}\left[\log p(W|A, Z, \beta) + \log p(Z|A, Y, X, \theta)\right] + \text{const} \\
&= \sum_{k=1}^{K} Z_{ij}^{dnk} \sum_{v=1}^{V} W_{ij}^{dnv} \log \beta_{kv} + \sum_{q=1}^{Q}\sum_{l=1}^{L} Y_{iq}X_{jl} \sum_{k=1}^{K} Z_{ij}^{dnk}\mathbb{E}_{\theta_{ql}}\left[\log \theta_{qlk}\right] + \text{const} \\
&= \sum_{k=1}^{K} Z_{ij}^{dnk} \left(\sum_{v=1}^{V} W_{ij}^{dnv} \log \beta_{kv} + \sum_{q=1}^{Q}\sum_{l=1}^{L} Y_{iq}X_{jl} \left(\psi(\gamma_{qlk}) - \psi\left(\sum_{k'=1}^{K} \gamma_{qlk'}\right)\right)\right) + \text{const},
\end{aligned}$$

where the constant term includes everything not depending on Z_{ij}^{dnk} and $\psi(\cdot)$ denotes the digamma function. The functional form of a multinomial distribution can be recognized in the above equation. Hence

$$q(Z_{ij}^{dn}) = \mathcal{M}(Z_{ij}^{dn}; 1, \phi_{ij}^{dn} = (\phi_{ij}^{dn1}, \ldots, \phi_{ij}^{dnK})),$$

where

$$\phi_{ij}^{dnk} \propto \left(\prod_{v=1}^{V} \beta_{kv}^{W_{ij}^{dnv}}\right) \prod_{q=1}^{Q}\prod_{l=1}^{L} \exp\left(\psi(\gamma_{qlk}) - \psi\left(\sum_{k'=1}^{K} \gamma_{qlk'}\right)\right)^{Y_{iq}X_{jl}}.$$

A.2 Proof of Propostion 2

The VEM update step for $q(\theta)$ can be obtained as follows

$$\begin{aligned}
\log q(\theta) &= \mathbb{E}_Z\left[\log p(Z|A, Y, X, \theta)\right] + p(\theta|\alpha) + \text{const} \\
&= \sum_{i=1}^{M}\sum_{j=1}^{P} A_{ij} \sum_{d=1}^{D_{ij}}\sum_{n=1}^{N_{ij}^d}\sum_{q=1}^{Q}\sum_{l=1}^{L} Y_{iq}X_{jl} \sum_{k=1}^{K} \mathbb{E}_Z[Z_{ij}^{dnk}] \log \theta_{qlk} + \sum_{q=1}^{Q}\sum_{l=1}^{L}\sum_{k=1}^{K}(\alpha_k - 1)\log \theta_{qlk} + \text{const} \\
&= \sum_{q=1}^{Q}\sum_{l=1}^{L}\sum_{k=1}^{K} \left(\alpha_k + \sum_{i=1}^{M}\sum_{j=1}^{P} A_{ij}Y_{iq}X_{jl} \sum_{d=1}^{D_{ij}}\sum_{n=1}^{N_{ij}^d} \phi_{ij}^{dnk} - 1\right) \log \theta_{qlk} + \text{const},
\end{aligned}$$

where the constant term includes everything not depending on θ. The functional form of factorizing Dirichlet distributions can be detected in the above equation. Therefore

$$q(\theta) = \prod_{q=1}^{Q}\prod_{l=1}^{L} \mathcal{D}(\theta_{ql}; \gamma_{ql} = (\gamma_{ql1}, \ldots, \gamma_{qlK})),$$

where

$$\gamma_{qlk} = \alpha_k + \sum_{i=1}^{M}\sum_{j=1}^{P}\sum_{d=1}^{D_{ij}}\sum_{n=1}^{N_{ij}^d} A_{ij}Y_{iq}X_{jl}\phi_{ij}^{dnk}.$$

A.3 Computing the lower bound

In this section, we compute the expectation in (12). In the following, the expectation is denoted by $\mathbb{E}_{Z,\theta}$ and it is taken with respect to the probability distribution defined in (13), conditionally on A, X, Y being known.

$$
\begin{aligned}
\mathcal{L}(q(\cdot)|A,Y,X,\beta) :&= \mathbb{E}_{Z,\theta}\left[\log \frac{p(W,Z,\theta|A,Y,X,\beta)}{q(Z,\theta)}\right] \\
&= \mathbb{E}_Z[\log p(W|A,Z,\beta)] + \mathbb{E}_{Z,\theta}[\log p(Z|A,Y,X,\theta)] + \mathbb{E}_\theta[\log p(\theta)] - \mathbb{E}_Z[\log q(Z)] - \mathbb{E}_\theta[\log q(\theta)] \\
&= \sum_{i=1}^{M}\sum_{j=1}^{P} A_{ij} \sum_{d=1}^{D_{ij}}\sum_{n=1}^{N_{ij}^d}\sum_{k=1}^{K} \phi_{ij}^{dnk} \sum_{v=1}^{V} W_{ij}^{dnv} \log \beta_{kv} \\
&+ \sum_{i=1}^{M}\sum_{j=1}^{P} A_{ij} \sum_{d=1}^{D_{ij}}\sum_{n=1}^{N_{ij}^d}\sum_{q=1}^{Q}\sum_{l=1}^{L} Y_{iq}X_{jl} \sum_{k=1}^{K} \phi_{ij}^{dnk} \left(\psi(\gamma_{qlk}) - \psi\left(\sum_{k'=1}^{K}\gamma_{qlk'}\right)\right) \\
&+ \sum_{q=1}^{Q}\sum_{l=1}^{L}\left(\log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) + \sum_{k=1}^{K}(\alpha_k-1)\left(\psi(\gamma_{qlk}) - \psi\left(\sum_{k'=1}^{K}\gamma_{qlk'}\right)\right)\right) \\
&- \sum_{i=1}^{M}\sum_{j=1}^{P} A_{ij} \sum_{d=1}^{D_{ij}}\sum_{n=1}^{N_{ij}^d}\sum_{k=1}^{K} \phi_{ij}^{dnk} \log \phi_{ij}^{dnk} \\
&- \sum_{q=1}^{Q}\sum_{l=1}^{L}\left(\log\Gamma\left(\sum_{k=1}^{K}\gamma_{qlk}\right) - \sum_{k=1}^{K}\log\Gamma(\gamma_{qlk}) + \sum_{k=1}^{K}(\gamma_{qlk}-1)\left(\psi(\gamma_{qlk}) - \psi\left(\sum_{k'=1}^{K}\gamma_{qlk'}\right)\right)\right).
\end{aligned}
\tag{21}
$$

A.4 Proof of Proposition 3

In order to obtain the stationary point in (16), we need to maximize the lower bound in (21) with respect to β_{kv}. Hence, all terms in (21) depending on β are grouped in the following objective function

$$
g(\beta,\lambda) := \sum_{i=1}^{M}\sum_{j=1}^{P} A_{ij} \sum_{d=1}^{D_{ij}}\sum_{n=1}^{N_{ij}^d}\sum_{k=1}^{K} \phi_{ij}^{dnk} \sum_{v=1}^{V} W_{ij}^{dnv} \log \beta_{kv} + \sum_{k=1}^{K}\lambda_k\left(\sum_{v=1}^{V}\beta_{kv} - 1\right),
$$

where K Lagrange multipliers are introduced to account for the constraint $\sum_{v=1}^{V}\beta_{kv} = 1$. Setting the derivative of $g(\cdot)$ with respect to β_{kv} equal to zero, we immediately obtain (16).

The optimal π_{ql} in (17) can be obtained by observing that the distribution $p(A|Y,X,\pi)$ in (4) is the only one involved with π. Hence, taking the log we get

$$
\log p(A|Y,X,\pi) = \sum_{i=1}^{N}\sum_{j=1}^{P}\sum_{q=1}^{Q}\sum_{l=1}^{L} Y_{iq}X_{jl}(A_{ij}\log\pi_{ql} + (1-A_{ij})\log(1-\pi_{ql}))
$$

and setting the derivative with respect to π_{ql} equal to zero, we obtain (17). Equations (19) and (18) are obtained in a very similar fashion.

A.5 Proof of Proposition 4

This section outlines the main steps needed to obtain (20). We preliminarily recall that the topic proportions θ_{ql} are assumed to be i.i.d. following a symmetric Dirichlet distribution with parameter $\alpha = 1$. First of all, we look for a BIC approximation of the following term

$$
\log p(W|A,Y,X) = \log\left(\int_\beta p(W|\beta,A,Y,X)p(\beta)d\beta\right),
$$

where $p(W|\beta,A,Y,X)$ appeares in (11), and $p(\beta)$ denotes a prior density function on β. Following the approach adopted in Bouveyron et al. (2016) and Than and Ho (2012), the number of i.i.d. observations is

equal to the number of meta documents (conditionally to $\{\beta, A, Y, X\}$). Therefore, we adopt the following BIC criterion to approximate $\log p(W|A, Y, X)$

$$\log p(W|A, Y, X) \approx \max_{\beta} p(W|\beta, A, Y, X) - K\frac{V-1}{2}\log(QL), \tag{22}$$

where the first term on the right hand side of the equality is replaced by the lower bound in (12) evaluated at $\hat{\beta}$ provided by (16) and $K(V-1)$ accounts for the number of free parameters in β. We now focus on the LBM part of the log-likelihood

$$\log p(A, Y, X|Q, L) = \log \int_{\pi,\rho,\delta} p(A, Y, X|\pi, \rho, \delta, Q, L)p(\pi, \rho, \delta)d\pi d\rho d\delta,$$

where $p(\pi, \rho, \delta)$ denotes any prior distribution over (π, ρ, δ). An ICL criterion is adopted to approximate the above log-likelihood:

$$\begin{aligned}\log p(A, Y, X|Q, L) \approx & \max_{\pi,\rho,\delta} \log p(A, Y, X|\pi, \rho, \delta, Q, L) \\ & - \frac{QL}{2}\log(MP) - \frac{Q-1}{2}\log M - \frac{L-1}{2}\log P.\end{aligned} \tag{23}$$

The same criterion was already introduced in the literature for model selection in binary LBM (see for instance Keribin et al., 2012). Equation (20) follows immediately from (22) and (23).

References

Anandkumar, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and Liu, Y.-K. (2012). A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 917–925.

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

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Machine Intel*, 7:719–725.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3):561–575.

Blei, D. and Lafferty, J. (2006). Correlated topic models. *Advances in neural information processing systems*, 18:147.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Bouveyron, C., Latouche, P., and Zreik, R. (2016). The stochastic topic block model for the clustering of vertices in networks with textual edges. *Statistics and Computing*.

Brault, V. and Channarond, A. (2016). Fast and consistent algorithm for the latent block model. *arXiv preprint arXiv:1610.09005*.

Celeux, G. and Govaert, G. (1991). A classification EM algorithm for clustering and two stochastic versions. Research Report RR-1364, INRIA. Projet CLOREC.

Côme, E., Randriamanamihaga, A., Oukhellou, L., and Aknin, P. (2014). Spatio-temporal analysis of dynamic origin-destination data using latent dirichlet allocation. application to the vélib? bike sharing system of paris. In *In Proceedings of 93rd Annual Meeting of the Transportation Research Board*.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

George, T. and Merugu, S. (2005). A scalable collaborative filtering framework based on co-clustering. In *Data Mining, Fifth IEEE international conference on*, pages 4–pp. IEEE.

Govaert, G. and Nadif, M. (2003). Clustering with block mixture models. *Pattern Recognition*, 36(2):463–473.

Govaert, G. and Nadif, M. (2008). Block clustering with bernoulli mixture models: Comparison of different approaches. *Computational Statistics & Data Analysis*, 52(6):3233–3245.

Govaert, G. and Nadif, M. (2010). Latent block model for contingency table. *Communications in Statistics?Theory and Methods*, 39(3):416–425.

Hathaway, R. J. (1986). Another interpretation of the em algorithm for mixture distributions. *Statistics & probability letters*, 4(2):53–56.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM.

Jacques, J. and Biernacki, C. (2017). Model-based co-clustering for ordinal data.

Keribin, C., Brault, V., Celeux, G., and Govaert, G. (2015). Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 25(6):1201–1216.

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

Kumar, S., Gao, X., and Welch, I. (2016). Co-clustering for dual topic models. In *Australasian Joint Conference on Artificial Intelligence*, pages 390–402. Springer.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE.

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

Nigam, K., McCallum, A., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134.

Papadimitriou, C., Raghavan, P., Tamaki, H., and Vempala, S. (1998). Latent semantic indexing: A probabilistic analysis. In *Proceedings of the tenth ACM PODS*, pages 159–168. ACM.

Phan, X.-H., Nguyen, L.-M., and Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, pages 91–100. ACM.

Podosinnikova, A., Bach, F., and Lacoste-Julien, S. (2015). Rethinking lda: moment matching for discrete ica. In *Advances in Neural Information Processing Systems*, pages 514–522.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Shafiei, M. M. and Milios, E. E. (2006). Latent dirichlet co-clustering. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 542–551. IEEE.

Teh, Y., Newman, D., and Welling, M. (2006). A collapsed variational bayesian inference algorithm for latent Dirichlet allocation. *Advances in neural information processing systems*, 18:1353–1360.

Than, K. and Ho, T. B. (2012). Fully sparse topic models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 490–505. Springer.

von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416.

Wang, P., Domeniconi, C., and Laskey, K. B. (2009). Latent dirichlet bayesian co-clustering. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 522–537. Springer.

Wang, S. and Huang, A. (2017). Penalized nonnegative matrix tri-factorization for co-clustering. *Expert Systems with Applications*, 78:64–73.

Wyse, J. and Friel, N. (2012). Block clustering with collapsed latent block models. *Statistics and Computing*, 22(2):415–428.

Wyse, J., Friel, N., and Latouche, P. (2017). Inferring structure in bipartite networks using the latent block-model and exact icl. *Network Science*, 5(1):45–69.