Model-Based Co-clustering for Ordinal Data

Julien Jacques, Christophe Biernacki

Model-Based Co-clustering for Ordinal Data

Julien Jacques[1,3]*, Christophe Biernacki[2,3]

[1] *Université de Lyon, Lyon 2, ERIC EA 3083, Lyon, France*

[2] *Laboratoire Paul Painlevé, UMR CNRS 8524, Université de Lille, Lille, France*

[3] *MODAL team, Inria Lille-Nord Europe*

Abstract

A model-based co-clustering algorithm for ordinal data is presented. This algorithm relies on the latent block model embedding a probability distribution specific to ordinal data (the so-called BOS or Binary Ordinal Search distribution). Model inference relies on a Stochastic EM algorithm coupled with a Gibbs sampler, and the ICL-BIC criterion is used for selecting the number of co-clusters (or blocks). The main advantage of this ordinal dedicated co-clustering model is its parsimony, the interpretability of the co-cluster parameters (mode, precision) and the possibility to take into account missing data. Numerical experiments on simulated data show the efficiency of the inference strategy, and real data analyses illustrate the interest of the proposed procedure.

Keywords: latent block model, EM algorithm, Gibbs sampler.

1. Introduction

Historically, clustering algorithms are used to explore data and to provide a simplified representation of data with a small number of homogeneous groups of individuals (i.e. clusters). With the big data phenomenon, the number of features becomes itself larger and larger, and traditional clustering methods are no more sufficient to explore such data sets. Indeed, the interpretation of a cluster of individuals using for instance a representative of this cluster (mean, mode, ...) is unfeasible since this representative is itself described by a very large number of features. Consequently, there is also a need to summarize the features by grouping them together into clusters.
Two approaches exist: bi-clustering and co-clustering. On the one hand, bi-clustering aims to identify blocks (or bi-clusters) defined as a subset of observations described by a subset of variables. These subsets can overlap. On the other hand, co-clustering aims to define both a partition of the observations and of the variables, and the blocks

*Corresponding author. Tel.: +33 478 772 609

Email addresses: julien.jacques@univ-lyon2.fr (Julien Jacques[1,3]), christophe.biernacki@univ-lille1.fr (Christophe Biernacki[2,3])

(or co-clusters) are obtained by crossing both partitions. The main differences is that blocks can overlap in bi-clustering and not in co-clustering, and moreover all features and observations have to belong to a block in co-clustering whereas not necessarily in bi-clustering. Figure 1 illustrates the differences between both approaches. This work focuses on the co-clustering problem as a natural extension of traditional partition clustering.

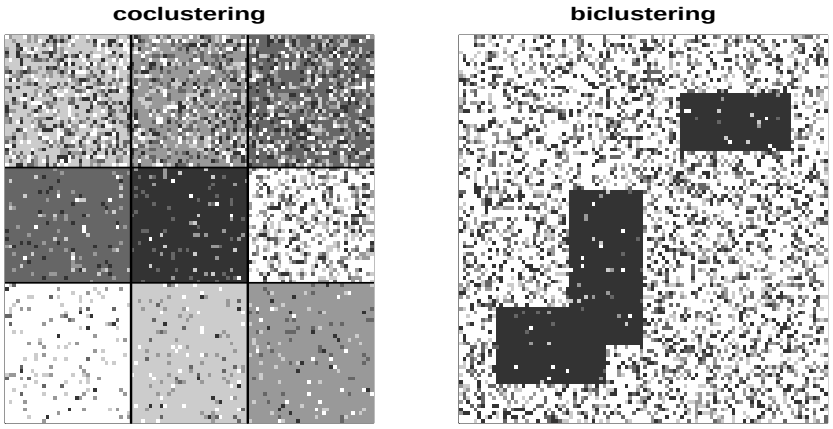

Figure 1: Co-clustering versus bi-clustering.

Co-clustering algorithms have been introduced to provide a solution by gathering into homogeneous groups both the observations and the features. Thus, the large data matrix can be summarized by a reduced number of blocks of data (or co-clusters). If the earliest (and most cited) methods are probably due to Hartigan (1972, 1975), the model-based approaches have recently proven their efficiency either for continuous, binary, count or contingency data (Govaert and Nadif, 2013; Pledger, 2014).

This work focuses on particular type of categorical data, ordinal data, occurring when the categories are ordered (Agresti, 2010). Ordinality is a characteristic of the meaning of measurements (Stevens, 1946), and distinct levels of an ordinal variable differ in degree of dissimilarity more than in quality (Agresti, 2010). Such data are very frequent in practice, as for instance in marketing studies where people are asked through questionnaires to evaluate some products or services on an ordinal scale (Dillon et al., 1994). Another examples can be found in medicine, when patients are asked to evaluate their quality of life on a Likert scale (see Cousson-Gélie (2000) for instance), or in vegetation sciences with the Braun-Blanquet scale (Podani, 2006). However, contrary to nominal categorical data (Celeux and Govaert, 2015), ordinal data have received less attention from a clustering point of view, and then, in face of such data, the practitioners often transform them into either quantitative data (associating an arbitrary number to each category, see Kaufman and Rousseeuw (1990) or Lewis et al. (2003) for instance) or into nominal data (ignoring the order information, see the Latent GOLD software

(Vermunt and Magidson, 2005)) in order to "recycle" easily related distributions. In order to avoid such extreme choices, some recent works have contributed to define clustering algorithms specific for ordinal data (Gouget, 2006; D'Elia and Piccolo, 2005; Podani, 2006; Giordan and Diana, 2011; Jollois and Nadif, 2011; Biernacki and Jacques, 2016; Ranalli and Rocci, 2016; Fernández et al., 2016). Nevertheless, when the number of features is large, the observations clustering can be not sufficient to summarize the data and a simultaneous clustering of the features could be meaningful.

In a co-clustering context Matechou et al. (2016) recently proposed an approach relying on the proportional odds model, itself assuming that the ordinal response has an underlying continuous latent variable. Unfortunately, the authors did not provide any code or package for their method and thus numerical comparisons are not possible. Let notice that the R package *biclust* (Kaiser et al., 2015) proposes several bi-clustering algorithms, whose bi-clustering goal is not the same than co-clustering.

In this work, we propose a model-based co-clustering algorithm relying on a recent distribution for ordinal data (BOS for *Binary Ordinal Search* model, Biernacki and Jacques (2016)), which has proven its efficiency for modeling and clustering ordinal data. One of the main advantage of the BOS model is its parsimony and the significance of its parameters. Indeed, in the present work, each co-cluster of data is summarized with only two parameters, one position parameter and one precision parameter. Another advantage of the co-clustering model we propose, is that it is able to take into account missing data by estimating them during the inference algorithm. Thus, the proposed co-clustering algorithm can be also used in a matrix completion task (see Candès and Recht (2009) for instance).

The paper is organized as follows. Section 2 proposes the co-clustering model whereas its inference and tools for selecting the number of co-clusters are presented in Section3. Numerical studies (Section 4) show the efficiency of the proposed approach, and two real data applications are presented in Section 5. A discussion concludes the paper in Section 6.

2. Latent block model for ordinal data

The data set is composed of a matrix of n observations (rows or individuals) of d ordinal variables (columns or features): $\mathbf{x} = (x_{ih})_{1\leq i\leq n, 1\leq h\leq d}$. For simplicity, the ordered levels of x_{ih} will be numbered $\{1, ..., m_h\}$, and all m_h's are assumed to be equal: $m_h = m$ $(1 \leq h \leq d)$. A natural approach for model-based co-clustering is to consider the latent block model (Govaert and Nadif (2013)), which itself relies on a probability distribution for the data. In the following, the BOS model for ordinal data is presented, then the latent block model and finally their combination for providing the proposed model.

2.1. The BOS model for ordinal data

The BOS model introduced in Biernacki and Jacques (2016) is a probability distribution for ordinal data parametrized by a precision parameter $\pi_{k\ell} \in [0, 1]$ and a position

parameter $\mu_{k\ell} \in \{1, ..., m\}$. This model has been built by their authors using the assumption that an ordinal variable is the result of a stochastic binary search algorithm within the ordered table $(1, ..., m)$. Advantage of such an algorithm is to use strictly (no more, no less) the order information conveyed by ordinal features.
Technically speaking, at the jth step of this binary search algorithm, e_j is the current interval in $\{1, \ldots, m\}$, and y_j the break point in this interval. The BOS distribution is then defined as follows:

$$\mathrm{p}(x_{ij}; \mu_{k\ell}, \pi_{k\ell}) = \sum_{e_{m-1},\ldots,e_1} \prod_{j=1}^{m-1} \mathrm{p}(e_{j+1}|e_j; \mu_{k\ell}, \pi_{k\ell})\mathrm{p}(e_1) \tag{1}$$

where

$$\begin{aligned}
\mathrm{p}(e_1) &= 1, \\
\mathrm{p}(e_{j+1}|e_j; \mu_{k\ell}, \pi_{k\ell}) &= \sum_{y_j \in e_j} \mathrm{p}(e_{j+1}|e_j, y_j; \mu, \pi)\mathrm{p}(y_j|e_j), \\
\mathrm{p}(y_j|e_j) &= \frac{1}{|e_j|}\mathbb{I}(y_j \in e_j), \\
\mathrm{p}(e_{j+1}|e_j, y_j; \mu_{k\ell}, \pi_{k\ell}) &= \pi_{k\ell}\mathrm{p}(e_{j+1}|y_j, e_j, z_j = 1; \mu_{k\ell}) + (1 - \pi_{k\ell})\mathrm{p}(e_{j+1}|y_j, e_j, z_j = 0), \\
\mathrm{p}(z_j|e_j; \pi_{k\ell}) &= \pi_{k\ell}\mathbb{I}(z_j = 1) + (1 - \pi_{k\ell})\mathbb{I}(z_j = 0), \\
\mathrm{p}(e_{j+1}|y_j, e_j, z_j = 0) &= \frac{|e_{j+1}|}{|e_j|}\mathbb{I}(e_{j+1} \in \{e_j^-, e_j^=, e_j^+\}), \\
\mathrm{p}(e_{j+1}|y_j, e_j, z_j = 1; \mu_{k\ell}) &= \mathbb{I}(e_{j+1} = \operatorname*{argmin}_{e \in \{e_j^-, e_j^=, e_j^+\}} \delta(e, \mu_{k\ell}))\mathbb{I}(e_{j+1} \in \{e_j^-, e_j^=, e_j^+\}),
\end{aligned}$$

with δ a "distance" between μ and an interval e (defined by $\delta(e, \mu_{k\ell}) = \min(|\mu_{k\ell} - b^-|, |\mu_{k\ell} - b^+|)$ if b^- and b^+ are the lower and upper limits of the interval e), with e_j^- the interval on the left of the break point y_j, $e_j = \{y_j\}$ the interval restricted to the break point y_j and e_j^+ the interval on the right of the break point.

It is shown in Biernacki and Jacques (2016) that the BOS distribution (1) is a polynomial function of $\pi_{k\ell}$ of degree $m-1$, in which the coefficients depend on the precision parameter $\mu_{k\ell}$. This distribution is especially flexible since it leads to a probability distribution evolving from a uniform distribution (when $\pi_{k\ell} = 0$) or to a distribution more and more peaked around the mode $\mu_{k\ell}$ (when $\pi_{k\ell}$ grows) until to a Dirac distribution at the mode $\mu_{k\ell}$ (when $\pi_{k\ell} = 1$). See Biernacki and Jacques (2016) for more detailed and illustration of this probability distribution. The shape of the BOS distribution for different values of μ and π is also displayed on Figure 2.

2.2. The latent block model

The latent block model assumes local independence, *i.e.* the $n \times d$ random variables $\mathbf{x}$ are assumed to be independent once the row partition $\mathbf{v} = (v_{ik})_{1 \le i \le n, 1 \le k \le K}$ and the

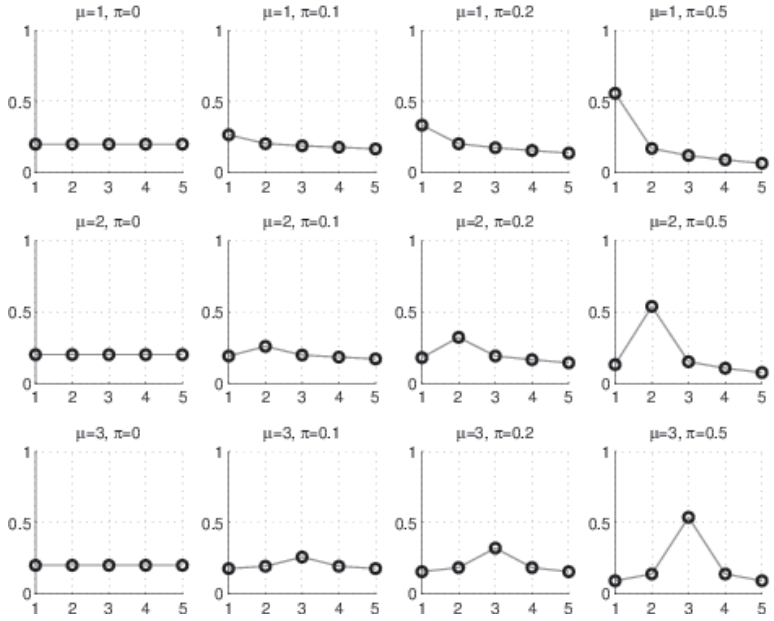

Figure 2: BOS distribution $\mathrm{p}(x;\mu,\pi)$: shape for $m = 5$ and for different values of μ and π.

column partition $\mathbf{w} = (w_{h\ell})_{1\leq h\leq d, 1\leq \ell\leq L}$ are fixed, where K and L are respectively the number of row and column clusters. Note that a standard binary partition is used for $\mathbf{v}$ ($v_{ik} = 1$ if row i belongs to cluster k and 0 otherwise) and $\mathbf{w}$ (similar than $\mathbf{v}$ but in column). Moreover, the row partition $\mathbf{v}$ and the column partition $\mathbf{w}$ are assumed to be independent. The latent block model can then be written as:

$$\mathrm{p}(\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{v}\in V}\sum_{\mathbf{w}\in W} \mathrm{p}(\mathbf{v};\boldsymbol{\theta})\mathrm{p}(\mathbf{w};\boldsymbol{\theta})\mathrm{p}(\mathbf{x}|\mathbf{v},\mathbf{w};\boldsymbol{\theta}) \tag{2}$$

where (below the straightforward range for i, h, k and ℓ are omitted):

- V is the set of all possible partitions of rows into K groups, W is the set of partitions of the columns into L groups,
- $\mathrm{p}(\mathbf{v};\boldsymbol{\theta}) = \prod_{ik} \alpha_k^{v_{ik}}$ and $\mathrm{p}(\mathbf{w};\boldsymbol{\theta}) = \prod_{h\ell} \beta_\ell^{w_{h\ell}}$ where α_k and β_ℓ are the row and column mixing proportions, belonging to $[0, 1]$ and adding to 1,
- $\mathrm{p}(\mathbf{x}|\mathbf{v},\mathbf{w};\boldsymbol{\theta}) = \prod_{ihk\ell} \mathrm{p}(x_{ih};\mu_{k\ell},\pi_{k\ell})^{v_{ik}w_{h\ell}}$ where $\mathrm{p}(x_{ih};\mu_{k\ell},\pi_{k\ell})$ is the probability of x_{ij} according to the BOS model (1) parametrized by $(\pi_{k\ell},\mu_{k\ell})$ with the so-called precision parameter $\pi_{k\ell} \in [0, 1]$ and position parameter $\mu_{k\ell} \in \{1, ..., m\}$, and
- $\boldsymbol{\theta} = (\pi_{k\ell}, \mu_{k\ell}, \alpha_k, \beta_\ell)$ is the whole mixture parameter.

2.3. The proposed model

The latent block model (2) for ordinal data proposed in this work can finally be written as:

$$\mathrm{p}(\mathbf{x};\boldsymbol{\theta}) \;=\; \sum_{\mathbf{v}\in V}\sum_{\mathbf{w}\in W}\prod_{ik}\alpha_k^{v_{ik}}\prod_{h\ell}\beta_\ell^{w_{h\ell}}\prod_{ihk\ell}\mathrm{p}(x_{ih};\mu_{k\ell},\pi_{k\ell})^{v_{ik}w_{h\ell}}. \tag{3}$$

The question to addressed is now its estimation.

3. Model inference

Missing data. In the present work, we consider the case in which the data $\mathbf{x}$ may be incomplete. We will notice $\check{\mathbf{x}}$ the set of observed data, $\hat{\mathbf{x}}$ the set of unobserved data and $\mathbf{x} = (\check{\mathbf{x}}, \hat{\mathbf{x}})$ the set of both observed and unobserved data. The inference algorithm which will now be described is able to take into account these missing data and to estimate them. We assume also that the whole missing process is Missing at Random (see Little and Rubin (2002)).

The inference of the model (3) consists in estimating $\boldsymbol{\theta}$ by maximizing the observed log-likelihood:

$$\ell(\boldsymbol{\theta};\check{\mathbf{x}}) = \sum_{\hat{\mathbf{x}}}\ln \mathrm{p}(\mathbf{x};\boldsymbol{\theta}). \tag{4}$$

The EM algorithm is computationally challenging in that co-clustering case (see Govaert and Nadif (2013)). Indeed, the E step of an EM algorithm requires the computation of the joint conditional distributions of the missing labels $\mathrm{p}(v_{ik}w_{h\ell} = 1|\check{\mathbf{x}};\boldsymbol{\theta}^{(q)})$ for $1 \leq i \leq n, 1 \leq k \leq K, 1 \leq h \leq d$ and $1 \leq \ell \leq L$, $\boldsymbol{\theta}^{(q)}$ being a current value of the parameter. Thus, the E step involves to compute too many terms that cannot be factorized as for a standard mixture due to the dependence conditionally on the observations of the row and column labels. Several alternatives to the EM algorithm are available, as the variational EM algorithm, the SEM-Gibbs algorithm or a Bayesian inference (Govaert and Nadif, 2013). In this paper we opt for the SEM-Gibbs, which is known to be efficient for avoiding spurious solutions, while being very simple to implement (Keribin et al., 2015).

3.1. SEM-Gibbs algorithm

The proposed SEM-Gibbs algorithm relies on an inner EM algorithm used in Biernacki and Jacques (2016) for the estimation of the BOS model. Starting from an initial value for the parameter ($\boldsymbol{\theta}^{(0)}$) and for the missing data ($\hat{\mathbf{x}}^{(0)}, \mathbf{w}^{(0)}$), the qth iteration of the SEM-Gibbs algorithm alternates the following SE and M steps ($q \geq 0$).

SE step. Execute a small number (at least 1) of successive iterations of the following three steps:

1. generate the row partition $v_{ik}^{(q+1)}|\hat{\mathbf{x}}^{(q)}, \check{\mathbf{x}}, \mathbf{w}^{(q)}$ for all $1 \leq i \leq n, 1 \leq k \leq K$:

$$\mathrm{p}(v_{ik} = 1|\hat{\mathbf{x}}^{(q)}, \check{\mathbf{x}}, \mathbf{w}^{(q)}; \boldsymbol{\theta}^{(q)}) = \frac{\alpha_k^{(q)} f_k(x_{i.}^{(q)}|\mathbf{w}^{(q)}; \boldsymbol{\theta}^{(q)})}{\sum_{k'} \alpha_{k'}^{(q)} f_{k'}(x_{i.}^{(q)}|\mathbf{w}^{(q)}; \boldsymbol{\theta}^{(q)})} \tag{5}$$

   where $x_{i.}^{(q)} = (x_{ih}^{(q)})_h$ and $f_k(x_{i.}^{(q)}|\mathbf{w}^{(q)}; \boldsymbol{\theta}^{(q)}) = \prod_{h\ell} \mathrm{p}(x_{ih}^{(q)}; \mu_{k\ell}^{(q)}, \pi_{k\ell}^{(q)})^{w_{h\ell}^{(q)}}$ and $x_{ih}^{(q)}$ being either $\check{x}_{ih}$ if it corresponds to an observed data or $\hat{x}_{ih}^{(q)}$ if not.
2. symmetrically, generate the column partition $w_{h\ell}^{(q+1)}|\hat{\mathbf{x}}^{(q)}, \check{\mathbf{x}}, \mathbf{v}^{(q+1)}$ for all $1 \leq h \leq d, 1 \leq \ell \leq L$:

$$\mathrm{p}(w_{h\ell} = 1|\hat{\mathbf{x}}^{(q)}, \check{\mathbf{x}}, \mathbf{v}^{(q+1)}; \boldsymbol{\theta}^{(q)}) = \frac{\beta_\ell^{(q)} g_\ell(x_{.h}^{(q)}|\mathbf{v}^{(q+1)}; \boldsymbol{\theta}^{(q)})}{\sum_{\ell'} \beta_{\ell'}^{(q)} g_{\ell'}(x_{.h}^{(q)}|\mathbf{v}^{(q+1)}; \boldsymbol{\theta}^{(q)})} \tag{6}$$

   where $x_{.h}^{(q)} = (x_{ih}^{(q)})_i$ and $g_\ell(x_{.h}^{(q)}|\mathbf{v}^{(q+1)}; \boldsymbol{\theta}^{(q)}) = \prod_{ik} \mathrm{p}(x_{ih}^{(q)}; \mu_{k\ell}^{(q)}, \pi_{k\ell}^{(q)})^{v_{ik}^{(q+1)}}$.
3. generate the missing data $\hat{x}_{ih}^{(q+1)}|\check{\mathbf{x}}, \mathbf{v}^{(q+1)}, \mathbf{w}^{(q+1)}$ following

$$\mathrm{p}(\hat{x}_{ih}|\check{\mathbf{x}}, \mathbf{v}^{(q+1)}, \mathbf{w}^{(q+1)}; \boldsymbol{\theta}^{(q)}) = \prod_{k\ell} \mathrm{p}(\hat{x}_{ih}; \mu_{k\ell}^{(q)}, \pi_{k\ell}^{(q)})^{v_{ik}^{(q+1)} w_{h\ell}^{(q+1)}}.$$

M step. Estimate $\boldsymbol{\theta}$, conditionally on $\hat{\mathbf{x}}^{(q+1)}, \mathbf{v}^{(q+1)}, \mathbf{w}^{(q+1)}$ obtained at the SE step (and also conditionally to $\check{\mathbf{x}}$), using the EM algorithm of Biernacki and Jacques (2016).

Choosing the parameter estimation. After a burn in period, the final estimation of the discrete parameter $\mu_{k\ell}$ is the mode of the sample distribution, and the final estimation of the continuous parameters $(\pi_{k\ell}, \alpha_k, \beta_\ell)$ is the mean of the sample distribution. It produces a final estimate $\hat{\boldsymbol{\theta}}$.

Estimating the partition and the missing data. After having chosen the parameter estimation $\hat{\boldsymbol{\theta}}$, a sample of $(\hat{\mathbf{x}}, \mathbf{v}, \mathbf{w})$ is generated with the Gibbs sampling described above in the SE step with $\boldsymbol{\theta}$ fixed to $\hat{\boldsymbol{\theta}}$. The final bi-partition $(\hat{\mathbf{v}}, \hat{\mathbf{w}})$ as well as the missing observation $\hat{\mathbf{x}}$ are estimated by the mode of their sample distributions.

3.2. Choice of the number of blocks

In order to select the numbers of blocks, K clusters in rows and L clusters in columns, some model selection criteria should be involved. The most classical ones, like BIC (Schwarz 1970), rely on penalizing the maximum log-likelihood value $\ell(\hat{\theta}; \check{\mathbf{x}})$. However, due to the dependency structure of the observed data $\check{\mathbf{x}}$, the value $\ell(\hat{\theta}; \check{\mathbf{x}})$ is not available (see Govaert and Nadif (2013); Keribin et al. (2015)). In addition and for the same reason, the penalization term of such standard criteria may not remain valid, what is the case for BIC for instance (see Keribin et al. (2015)).

Alternatively, an approximation of the ICL information criterion (Biernacki et al., 2001), called here ICL-BIC, can be invoked since allowing to overcome both previous problems due to the dependency structure in $\check{\mathbf{x}}$. The key point is that this latter vanishes since ICL relies on the completed latent block information $(\mathbf{v}, \mathbf{w})$, instead of integrating on it as it is the case in BIC. In particular, Keribin *et al.* (2014) detailed how to express ICL-BIC for the general case of categorical data. But, noticing that the BOS distribution can be simply viewed as a specific model for categorical data, it is possible to straightforwardly transpose the ICL-BIC expression given by these authors by following step by step their piece of work, with no new technical material. In addition, it is now proven that both BIC and ICL-BIC have the same behaviour for high values of the number of lines and/or columns, leading also to a consistent estimation of the number of blocks (see Keribin et al. (2015); Brault et al. (2017)). The resulting BOS-specific ICL-BIC is expressed by:

$$\text{ICL-BIC}(K, L) = \log \mathrm{p}(\check{\mathbf{x}}, \hat{\mathbf{v}}, \hat{\mathbf{w}}; \hat{\boldsymbol{\theta}}) - \frac{K-1}{2} \log n - \frac{L-1}{2} \log d - \frac{KL}{2} \log(nd) \qquad (7)$$

where $\hat{\mathbf{v}}, \hat{\mathbf{w}}$ and $\hat{\boldsymbol{\theta}}$ are the respective estimation of the row partition, column partition and model parameters obtained at the end of the estimation algorithm and where

$$\log \mathrm{p}(\check{\mathbf{x}}, \hat{\mathbf{v}}, \hat{\mathbf{w}}; \hat{\boldsymbol{\theta}}) = \sum_{ih : x_{ih} \in \check{\mathbf{x}}} \log \mathrm{p}(\check{x}_{ih}, \hat{v}_i, \hat{w}_h; \hat{\boldsymbol{\theta}}) + \sum_{ih : x_{ih} \in \hat{\mathbf{x}}} \log \mathrm{p}(\hat{v}_i, \hat{w}_h; \hat{\boldsymbol{\theta}})$$

with

$$\log \mathrm{p}(\check{x}_{ih}, \hat{v}_i, \hat{w}_h; \hat{\boldsymbol{\theta}}) = \sum_k \hat{v}_{ik} \log \hat{\alpha}_k + \sum_\ell \hat{w}_{h\ell} \log \hat{\beta}_\ell + \sum_{k\ell} \hat{v}_{ik} \hat{w}_{h\ell} \log \mathrm{p}(\check{x}_{ih}; \hat{\mu}_{k\ell}, \hat{\pi}_{k\ell})$$

and

$$\log \mathrm{p}(\hat{v}_i, \hat{w}_h; \hat{\boldsymbol{\theta}}) = \sum_k \hat{v}_{ik} \log \hat{\alpha}_k + \sum_\ell \hat{w}_{h\ell} \log \hat{\beta}_\ell.$$

The couple (K, L) leading to the maximum ICL-BIC value has then to be retained.

4. Numerical experiments on synthetic data sets

The convergence of the SEM-Gibbs algorithm and of the ICL-BIC criterion are theoretically known (see for instance Keribin et al. (2015)). The aim of this section is to investigate their behavior for finite sample size. Additionally, the influence of missing data on parameter estimation is investigated.

4.1. Algorithm and model-section criterion validation

Experimental setup. 50 data sets are simulated using the BOS distribution according to the following setup: $K = L = 3$ clusters in row and column, $d = 100$ ordinal variables with $m = 5$ levels and $n = 100$ observations. Two sets of values of $(\mu_{k\ell}, \pi_{k\ell})$ are chosen in order to build one simulation setting with well separated blocks (setting 1) and another one with more mixed blocks (setting 2). Values of model parameters are given in Table 1, and Figure 3 illustrates an example of original data and co-clustering result.

k/ℓ	1	2	3
1	(1,0.9)	(2,0.9)	(3,0.9)
2	(4,0.9)	(5,0.9)	(1,0.5)
3	(2,0.5)	(3,0.5)	(4,0.5)

k/ℓ	1	2	3
1	(1,0.2)	(2,0.2)	(3,0.2)
2	(4,0.2)	(5,0.2)	(1,0.1)
3	(2,0.1)	(3,0.1)	(4,0.1)

Table 1: Values of the BOS model parameters used for experiments, setting 1 (left) and setting 2 (right).

In order to select the number of iterations of the SEM-Gibbs algorithm to use, different numbers have been tested and the evolution of the model parameters and the partitions along with the iterations of the algorithm is plotted for each iteration number. Figure 4 plots this evolution for a SEM-Gibbs algorithm with 50 iterations and for setting 1. According to this representation, 50 iterations with a burn-in period of 20 iterations seem sufficient to obtain stability of the simulated chain. Moreover, in order to improve the initialization, the SEM-Gibbs algorithm is initialized with the marginal row and columns partitions obtained by *k-means*. The computing time with this setting is about one hour per simulation with an R code on an Intel Core i7 CPU 2.8GHz, 16Go RAM.

Empirical consistance of the SEM-Gibbs algorithm. Figure 5 and Table 2 illustrate the efficiency of the proposed estimation algorithm, by plotting the co-clustering results and the following indicators:

- `mu` (resp. `pi`): mean distance between the true $\boldsymbol{\mu}$ (resp. $\boldsymbol{\pi}$) and its estimated value $\hat{\boldsymbol{\mu}}$ (resp. $\hat{\boldsymbol{\pi}}$): $\Delta\boldsymbol{\mu} = \sum_{k=1}^{K}\sum_{\ell=1}^{L} |\mu_{k\ell} - \hat{\mu}_{k\ell}|/(KL)$ (resp. $\Delta\boldsymbol{\pi} = \sum_{k=1}^{K}\sum_{\ell=1}^{L} |\pi_{k\ell} - \hat{\pi}_{k\ell}|/(KL)$),
- `alpha` (resp. `beta`): mean distance between the true α (resp. β) and its estimated value $\hat{\alpha}$ (resp. $\hat{\beta}$): $\Delta\alpha = \sum_{k=1}^{K} |\alpha_k - \hat{\alpha}_k|/K$ (resp. $\Delta\beta = \sum_{\ell=1}^{L} |\beta_\ell - \hat{\beta}_\ell|/L$),
- `ARIr` (resp. `ARIc`): Adjusted Rand Index (ARI, Rand (1971)) for the row (resp. column) partition.

In order to evaluate the quality of the results, the ARI have been also computed using the true values of the model parameters (given in Table 1). The corresponding ARI

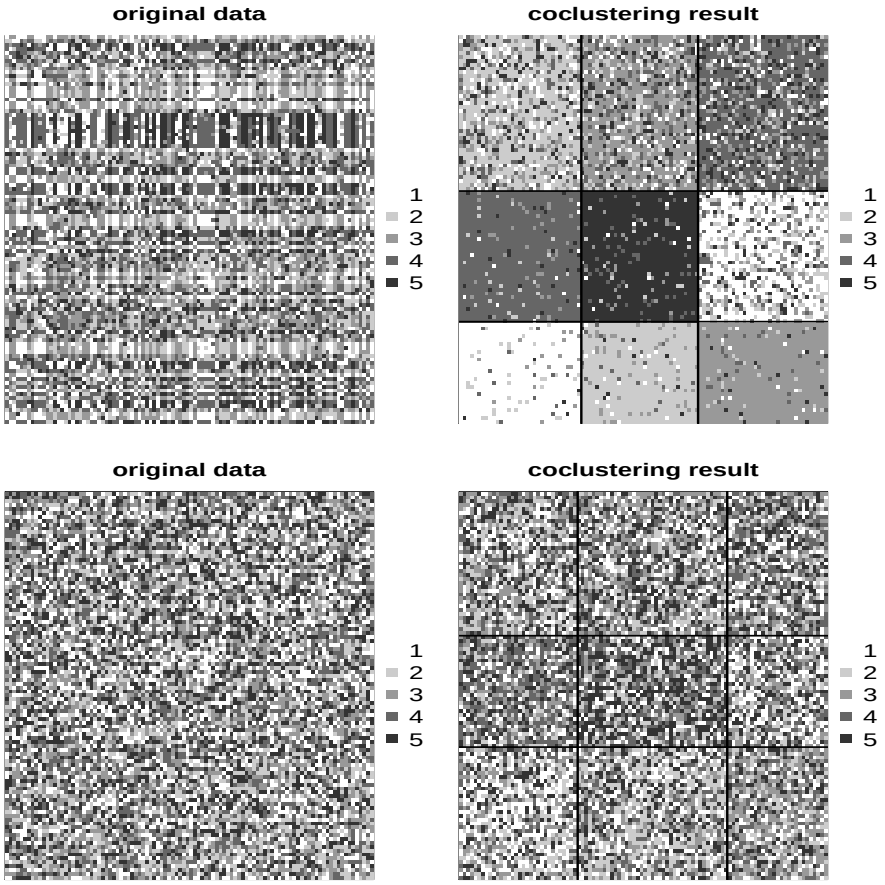

Figure 3: An example of data (left) and co-clustering results (right), for the experimental setting 1 (top) and setting 2 (bottom).

values are given in the two columns on the left of Table 2. Let first remark that the standard deviations for $\Delta\boldsymbol{\mu}$ are relatively large, what is due to the fact that μ belongs to a discrete space $\{1, \ldots, m\}$. The results for the setting 1 are excellent, what is not surprising since the blocks are well separated. For the setting 2, the estimations are as expected less accurate since the blocks are more mixed. But when comparing with the optimal ARI (computed using the true values of the model parameters), the results remain satisfying.

Efficiency of the ICL-BIC criterion to select the number of clusters. In this second experiment, the ability of ICL-BIC to retrieve the true number of clusters is tested. For this, data are simulated according to the previous experimental settings, and the ICL-BIC criterion is used to select the best number of clusters in row and in column among 2 to 4. Results presented in Table 3 show the ability of this criterion to retrieve the true number of clusters. The ICL-BIC criterion is very efficient in the first setting

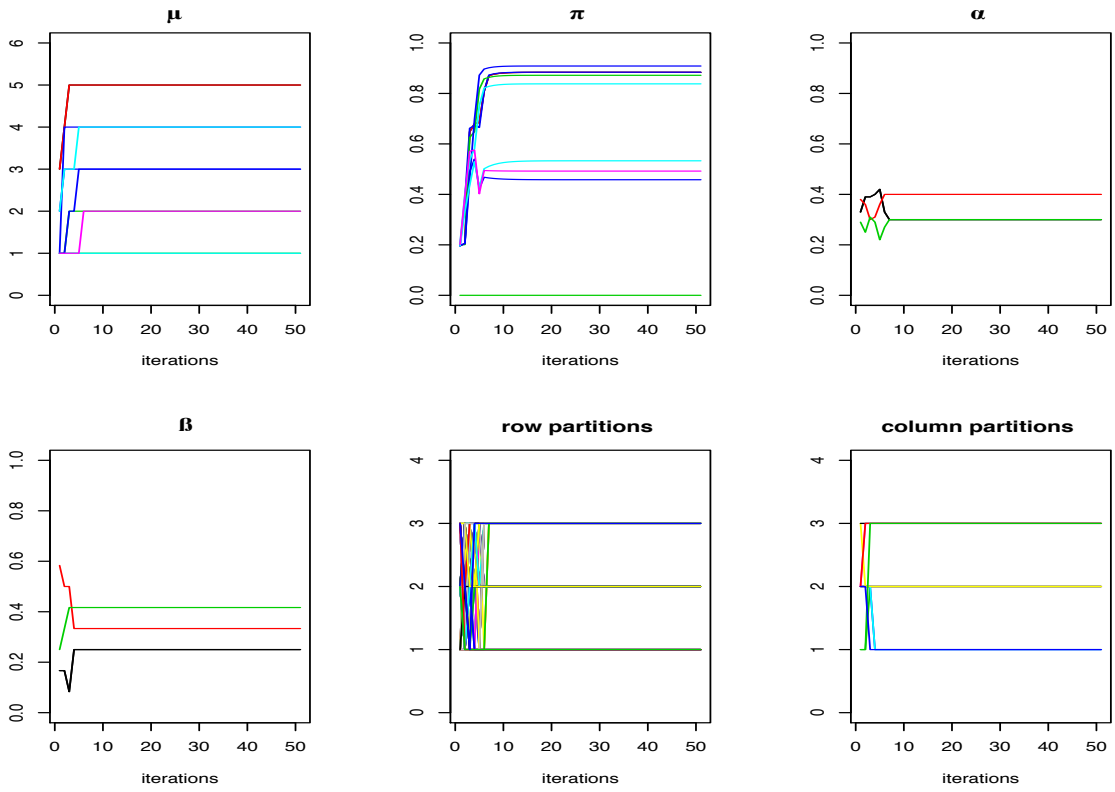

Figure 4: Evolution of the model parameters (one color per parameter $\mu_{k\ell}$, $\pi_{k\ell}$, α_k, β_ℓ) and the row/columns partitions (one color per v_{ik} and $w_{j\ell}$) during the SEM-Gibbs iterations

	$\Delta\boldsymbol{\mu}$	$\Delta\boldsymbol{\pi}$	$\Delta\alpha$	$\Delta\beta$	ARIr	ARIc
set. 1	0.16 (0.45)	0.03 (0.06)	0.05 (0.05)	0.05 (0.05)	0.97 (0.12)	0.96 (0.14)
				optimal ARI values	1 (0)	1 (0)
set. 2	0.68 (0.42)	0.06 (0.02)	0.06 (0.04)	0.07 (0.04)	0.58 (0.15)	0.59 (0.17)
				optimal ARI values	0.76 (0.09)	0.76 (0.09)

Table 2: Mean error of parameter estimation (and standard deviation) and mean ARI (s.d.) for the row and column partitions (`ARIr`,`ARIc`), for the experimental settings 1 and 2. Optimal ARI values have been obtained using the true model parameter values.

in which the clusters are well separated (the true numbers are selected in 92% of the 50 simulations), and, as expected, it is less efficient when clusters are more mixed (the true numbers are selected in 38% of the 50 simulations).

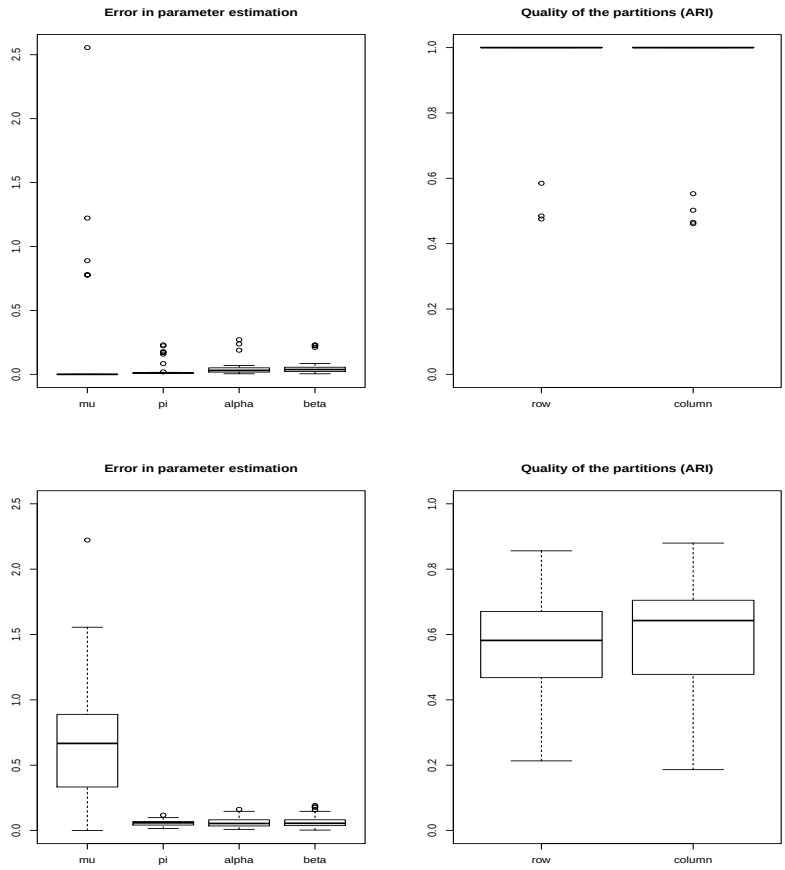

Figure 5: Error on parameter estimation (left) and ARI for the row and column partitions (right), for the experimental setting 1 (top) and setting 2 (bottom).

		L		
		2	3	4
K	2	0	0	0
	3	0	46	3
	4	0	1	0

		L		
		2	3	4
K	2	5	6	2
	3	1	19	10
	4	1	5	1

Table 3: Number of times the number of clusters K and L are selected (left: setting 1, right: setting 2).

4.2. Efficiency with missing data

In this section, a given percentage of missing data is introduced in the experimental setting 1 and 2 (from no missing data to 40%), and we study the impact of the presence of missing data onto the parameter estimation quality. Results are given in Figure 6. If

missing data has almost no impact on the *easy* experimental setting 1, they contribute to deteriorate the quality of the estimations in the experimental setting 2. So, if the clusters are well separated, what is expected if their number is selected by the ICL-BIC criterion, missing data has only a small impact on the co-clustering results. If the clusters are more mixed, the presence of missing data deteriorates the quality of estimation of the model parameter and of the partitions. In the real data application under study in the next section, the behavior of the proposed co-clustering algorithm in presence of (very) large proportion of missing data will be studied.

4.3. A more challenging experiment

The goal of this section is to investigate the behavior of the proposed approach with regards to a challenging dataset, having a large number of features, sample, blocks and having different block sizes.

Experimental setup. 20 data sets are simulated using the BOS distribution according to the following setup: $K = 12$ clusters in row and $L = 15$ clusters in column, $d = 10,000$ ordinal variables with $m = 6$ levels and $n = 1,000$ observations. All the blocks have parameters $\mu_{k\ell} = 1$ and $\mu_{k\ell} = 0.3$ except 15 blocks (among $K \times L = 180$ blocks) having $\mu_{k\ell} \in \{2, \ldots, 6\}$ and $\mu_{k\ell} = 0.9$. In order to defined blocks of heterogeneous sizes, the row proportions α_k (respectively column proportions β_ℓ) are sampled from a Dirichlet distribution of order 1 with parameter $\frac{1}{K}, \ldots, \frac{1}{K}$ (resp. $\frac{1}{L}, \ldots, \frac{1}{L}$). Figure 7 illustrates a sample of data (left), the true co-clustering (middle) and its estimation (right).

Results. The right plot of Figure 7 presents a co-clustering results (among the 20 data sets), which seems really satisfying since all the 9 blocks which are distinguishable on the true co-clustering plot are recovered (let remark that they are not at the same *places* in the data set since the cluster numbering is arbitrary). Moreover, Figure 8 displays the row and column ARIs distribution summary which are quite satisfying for 12 and 15 clusters in row and column, which are relatively large values with regards to many practical use cases.

5. Applications on real data

In this section the proposed co-clustering algorithm is used to analyse two real data sets. The first one is a survey on the quality of life of cancer patients whereas the second one is the Amazone Fine Food Review data.

5.1. Quality of life of cancer patients

The EORTC QLQ-C30 (Fayers et al., 2001) is a questionnaire developed to assess the quality of life of cancer patients. In this work the questionnaires filled in by 161 patients hospitalized for breast cancer are analyzed (see the Acknowledgment section for people and institutes who have contributed to collect the data). The EORTC QLQ-C30

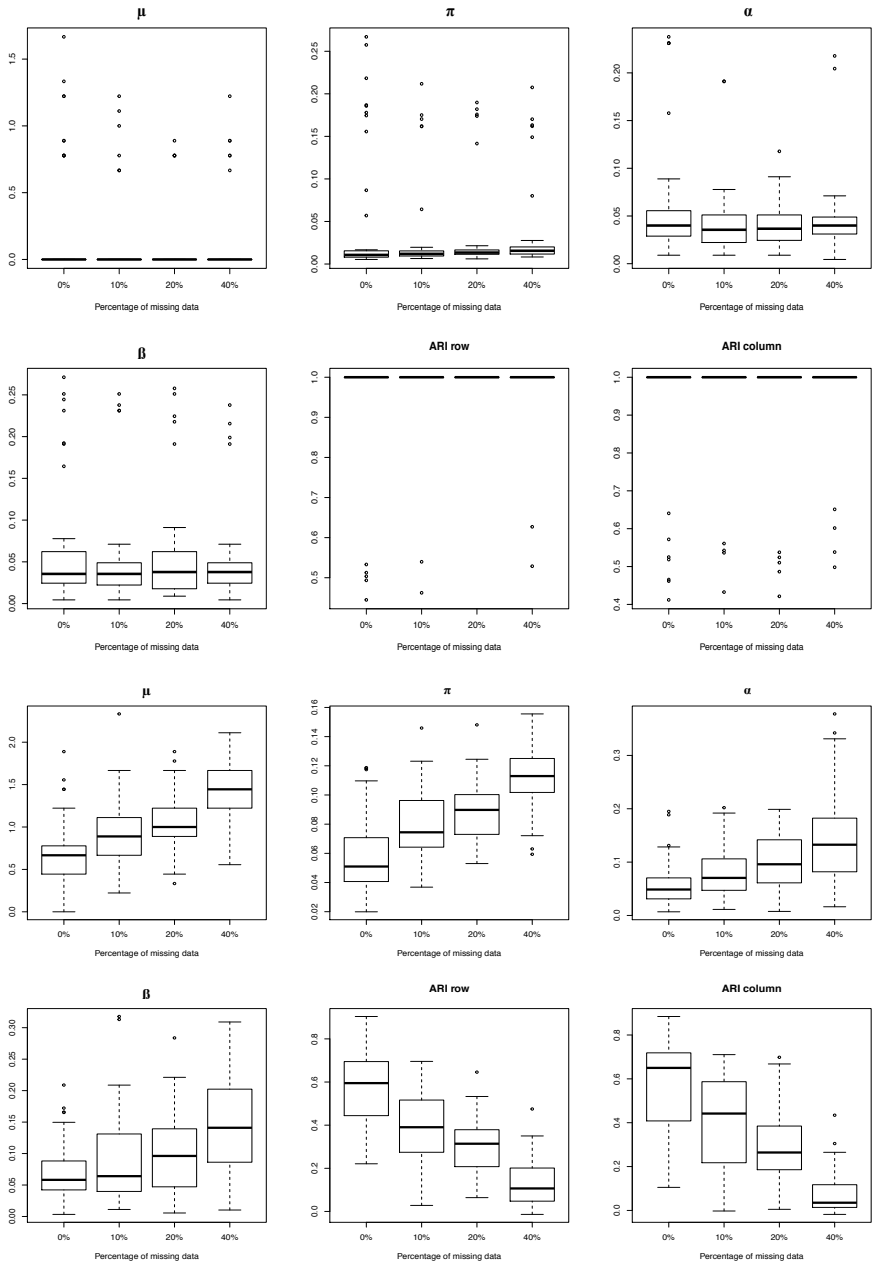

Figure 6: Error on parameter estimation and row and column ARI for different proportion of missing data, for the experimental setting 1 (two top lines) and 2 (two bottom lines).

questionnaire contains 30 questions for which the patients should answer with an ordinal scale. For the present co-clustering analysis only the first 28 (among 30) questions of the questionnaire are retained. For these questions the patients should answer on an

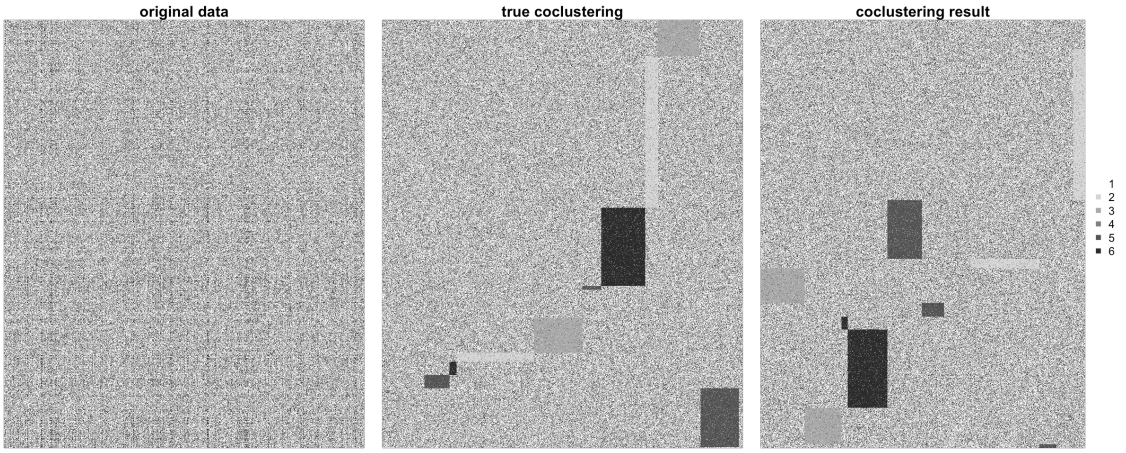

Figure 7: Data, true co-clustering and estimated co-clustering.

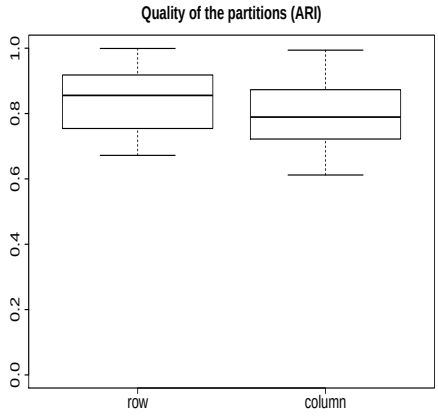

Figure 8: Row and column ARI.

ordinal scale with 4 categories ($m = 4$), from 1 (*not at all*) to 4 (*very much*). The two remaining questions, which are not taken into account in this analysis, are more general questions and should be answered on an ordinal scale with 7 categories. The data are plotted in the left panel of Figure 9.

Co-clustering is carried out for all row and column-clusters $(K, L) \in \{2, 3, 4\}^2$. The number of SEM-Gibbs iterations, tuned graphically in order to obtain stability of the simulated chain as described in Section 4.1, is fixed to 100 with a burn in period of 40 iterations. The ICL-BIC criterion selects 3 clusters in row and column (left panel of Table 4). The model parameters for $K = L = 3$ are given in Table 4 (right panel), and the co-clustering results are plotted in Figure 9 (right panel). On this figure, the numbering of the row-clusters is from the bottom to the top and the numbering of the

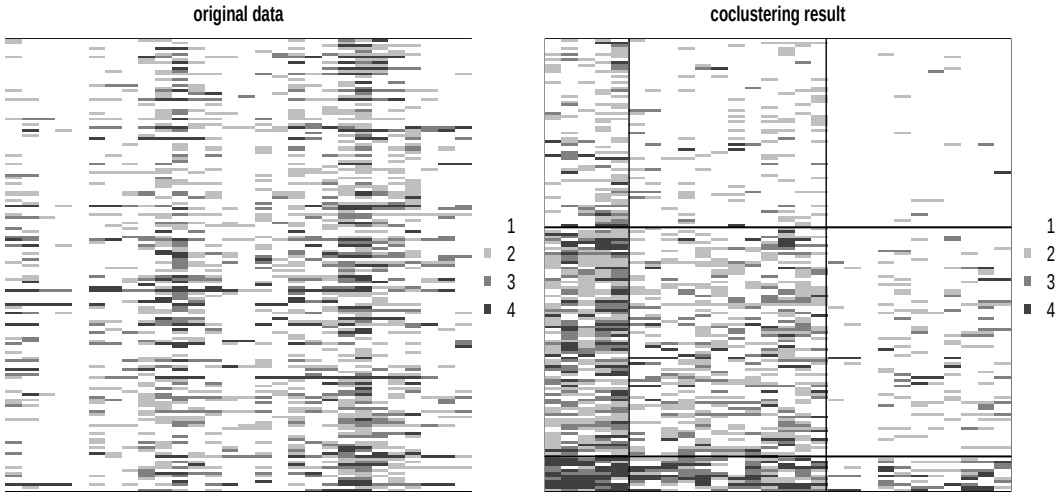

Figure 9: Original EORTC QLQ-C30 data (left) and co-clustering results into 3×3 blocks (right).

column-clusters is from the left to the right.
These results are particularly significant for the psychologists, as it is described below. The column-cluster 1 (left) can be interpreted by anxiety (for high scores) or quality of emotional life. The column-cluster 2 (middle) brings together the depressive symptoms items (loss of appetite, feeling weak, difficulty concentrating, irritable, depressed) and pain. The column-cluster 3 (right) is more difficult to interpret but with a common point which is the relationship to the other: there are physical quality of life items but that are associated with relationships with others. Since patients are hospitalized it seems logical that answers concerning the physical quality of life, symptoms and quality of social life are linked. For subjects, we would have in the first group (bottom) very few anxious patients, having an average quality of physical and social life and being rather depressed (12 patients). The second group (middle) concerns moderately anxious patients, but with poor or average quality of physical and social life, and feeling pretty moderately depressed (67 patients). This can be due to emotional suppression (false non-anxious) or they are really little depressed and anxious. The third group (top) corresponds to patients with rather high levels of depression, with very poor quality of physical and social life and feeling rather depressed (82 patients).

Comparison with competitors. Natural competitors of our BOS model for ordinal data are either continuous two-mode clustering methods or categorical two-mode clustering methods. For the former case, double k-means (Vichi, 2001) or two-mode Gaussian mixture analysis (Govaert and Nadif, 2013) could be used, as illustrated in Schepers et al. (2017). However, since double k-means seems to be not publicly available, we decide to restrict our attention to the latter which is available in the *BlockCluster* package (Bathia

		L		
		2	3	4
K	2	-3655	-3581	-3556
	3	-3642	**-3532**	-3548
	4	-3635	-3545	-3548

		ℓ		
		1	2	3
k	3	(1,0.60)	(1,0.84)	(1,0.98)
	2	(2,0.23)	(1,0.49)	(1,0.84)
	1	(4,0.59)	(1,$\simeq 0$)	(1,0.48)

Table 4: Value of the ICL-BIC criterion (left) for $(K, L) \in \{2, 3, 4\}^2$ for the BOS model and estimation of $(\mu_{k\ell}, \pi_{k\ell})$ (right) for the 9 co-clusters obtained on the EORTC QLQ-C30 data.

et al., 2016) for R. In addition, the BlockCluster package allows to perform categorical clustering methods, and thus we will use it also for its second functionality.
While performing the BlockCluster package for the continuous (Gaussian) case, only failed runs were observed for $(K, L) \in \{2, 3, 4\}^2$. This fact was already observed in Biernacki and Jacques (2016) since repeated measurements involved by ordinal data nearly systematically lead to degenerated Gaussian solutions (variance is zero). It indicates that continuous distributions (and in particular the Gaussian one) are not well suitable for such kind of data.

		L		
		2	3	4
K	2	-3652	-3573	-3569
	3	-3612	**-3535**	-3551
	4	-3606	**-3533**	-3559

Table 5: Value of the ICL criterion for $(K, L) \in \{2, 3, 4\}^2$ for BlockCluster categorical on the EORTC QLQ-C30 data.

We now compare our BOS model to the categorical co-clustering one as implemented in the BlockCluster package. For invoking this package, we have used hyperparameters for the Dirichlet distribution, associated to the mixing proportions (in row and in column) and to the level probabilities, both equal to 4 as recommended by Keribin et al. (2015). We can notice in Table 4 and Table 5 that both BOS and categorical models have quite similar ICL or ICL-BIC values (non asymptotic ICL values are available in BlockCluster for categorical data, thus avoiding the ICL-BIC approximation in this case). In particular they are in accordance for hesitating between the couple $(K, L) = (3, 3)$ and the couple $(K, L) = (4, 3)$, even if the former is more clearly selected by the BOS model. Advantage of the BOS model is twofold while leading to a similar number of blocks for $(K, L) = (3, 3)$ and a relatively close block structure (the row and column ARI between the results of both model are respectively 0.61 and 0.72). First, it is more parsimonious since BOS has only 13 continuous continuous and 9 discrete parameters for

BOS whereas the categorical model has 31 continuous parameters. Second, BOS is more easy to understand for the user, having a "Gaussian-like" meaning (mode, dispersion) as stated in the right of Table 4.

Quality of missing data imputation. Finally, in order to check on real data that the proposed methodology is efficient for imputing missing data, 10% of the EORTC QLQ-C30 data (451 observations over 28×161) have been totally randomly hidden (missing totally at random or MCAR mechanism (Little and Rubin, 2002)) and estimated by the proposed strategy. The experiment has been repeated 100 times, and Figure 10 displays the distribution of the estimation error $|x_{ij} - \hat{x}_{ij}|$ where x_{ij} is the hidden value and $\hat{x}_{ij}$ its estimation. Since the number of ordinal categories is equal to $m = 4$, this error belongs to $\{0, \ldots, 3\}$.
The quality of estimation of the missing data is very satisfying, with 60% of the missing observations perfectly estimated (null error) and more than 83% of them estimated with an error less than or equal to 1.

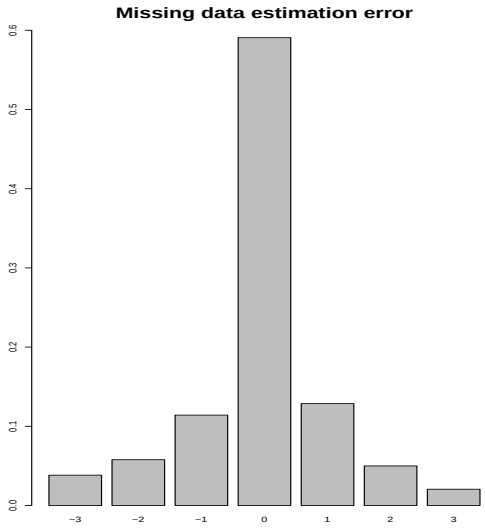

Figure 10: Relative frequency of estimation error when missing observations are artificially introduced in the EORTC QLQ-C30 data.

5.2. Amazon Fine Food Review data

The Amazon Fine Food Review data, available online on Kaggle website[1], corresponds to the ordinal assessment of products by customers. The assessment is done on an ordinal scale from 1 (lowest score) to 5 (highest score). The whole dataset is

[1]https://www.kaggle.com/snap/amazon-fine-food-reviews

composed of 256,059 customers and 74,258 products with about 500,000 products assessments. Thus, about 99.99737% of the data are missing. In order to illustrate our co-clustering method, we extract from this dataset the top 100 active customers and the top 100 evaluated products (Figure 11). In this sample of the whole dataset, *only* 86.44% of the data are missing. Given the large proportion of missing data the amount of available information in the data is relatively poor and since in this case also the proposed ICL-BIC criterion validity is weakened, we decide to fix the number of blocks to 4 (2 clusters in row and in column).

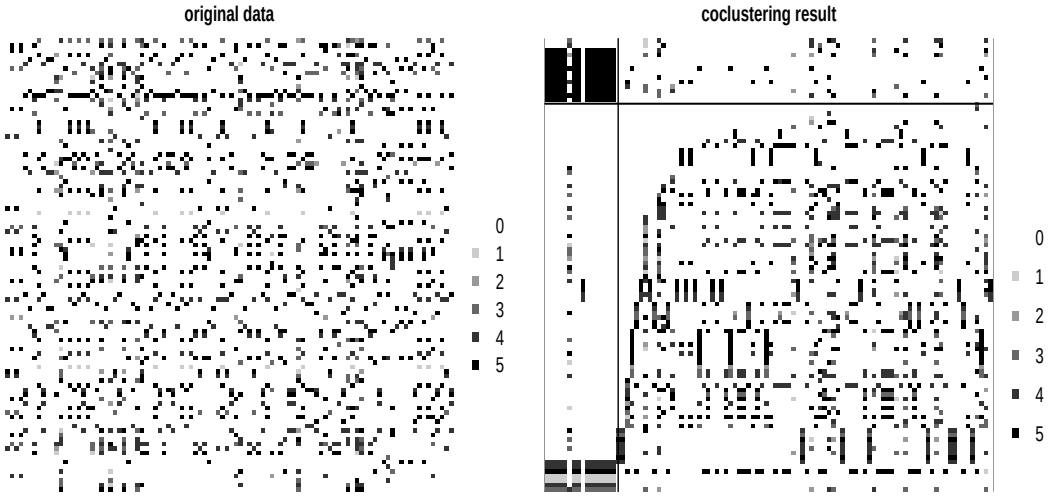

Figure 11: Top 100 Amazone Fine Food Review data (left) and co-clustering result (right).

The number of SEM-Gibbs iterations, tuned graphically in order to obtain stability of the simulated chain as described in Section 4.1, is fixed to 100 with a burn in period of 40 iterations. The corresponding co-clustering result is presented in the right panel of Figure 11, and parameter estimation for the six co-clusters are given in Table 6.

(μ, π)		ℓ	
		1	2
k	1	(5,0.98)	$\mathcal{U}$
	2	(5,0.45)	$\mathcal{U}$

Table 6: Value of (μ, π) for the 6 co-clusters obtained on the top 100 Amazone Fine Food Review data ($\mathcal{U}$: uniform distribution corresponding to $\pi_{12} \simeq 0$ and $\pi_{22} \simeq 0$).

Among the four co-clusters, two are essentially uniformly distributed ($\pi_{12} \simeq \pi_{22} \simeq 0$), and mainly group missing data (in white) together. Co-cluster (2,1) has a mode in 5 and is relatively dispersed ($\pi_{21} = 0.45$). Co-cluster (1,1) groups together people and

products with a distribution strangely very peaked in the highest scores ($\mu_{11} = 5$ and $\pi_{11} = 0.98$). In order to investigate this latter cluster, we look at the comments written by the customers about the products (these comments are available in the dataset), and we see that they all give exactly the same comment[2], what probably means that we have detected a group of false assessments.

In a second step, we increase the size of the data set by selecting the top $1,000$ customers and products, which contains 97.87% of missing data, and carry out a co-clustering with $K = L = 2$. Due to the very high proportion of missing data, all the obtained co-clusters are uniformly distributed ($\pi_{11} \simeq \pi_{12} \simeq \pi_{21} \simeq \pi_{22} \simeq 0$). Similar results would be obtained using the whole data set. As a conclusion, when the proportion of missing data is too high, the estimation of the missing data outweights the observed data and thus no relevant information can be obtained with our co-clustering strategy.

6. Discussion

In this paper a co-clustering algorithm for ordinal data is proposed. It relies on the latent block model using the parsimonious BOS distribution for ordinal data. Model inference is done through a SEM-Gibbs algorithm, which furthermore allows to tackle missing observations. The co-clustering results can be easily interpreted thanks to the meaningful parameters of the BOS distribution. Simulation study and real data analysis have contributed to show the efficiency and the practical interest of the proposed model. An R package is available upon request to the authors, and will be soon available on the CRAN.

If a practitioner is only interested in a clustering of individuals (rows), the proposed co-clustering algorithm provides a very parsimonious way to do this, by grouping all the features in a small number of groups and then modeling the features distributions with a very few number of parameters. Thus, it could be of practical use for high dimensional (row) clustering for ordinal data.

With the proposed approach, all the ordinal features must have the same number of categories. It could be interesting to extend this approach in order to be able to take into account features with different numbers of categories. The main gap is to be able to allow to features with different categories to be in same clusters. The latent block model does not allow this since it assumes that into a block the data share the same distribution, and so an alternative model have to be thought.

Finally, it could be interesting to take into account a temporal evolution in the data. For this, links between BOS co-clustering models at different time epoch has to be modeled, as for instance in a clustering context in Jacques and Biernacki (2010) or Hasnat et al. (2017).

[2] "*I'm addicted to salty and tangy flavors, so when I opened my first bag of Sea Salt & Vinegar Kettle Brand chips I knew I had a perfect complement to my vegetable trays of cucumber, carrot, celery and cherry tomatoes (...)*"

Acknowledgement

We thank Prof. Cousson-Gélie (Professor of Health Psychology, Laboratoire Epsylon Université Paul Valery Montpellier 3 & Université de Montpellier) for providing the EORTC QLQ-C30 data and for helpful discussion about the co-clustering results. We thank also INCa (Institut National du Cancer), Institut Lilly, Institut Bergonié, Centre Régional de Lutte Contre le Cancer de Bordeaux (C. Tunon de Lara, J. Delefortrie, A. Rousvoal, A. Avril, E. Bussières) and Laboratoire de Psychologie de l'Université de Bordeaux (C. Quintric and S. de Castro-Lévèque).
We also thank Margot Selosse for developing a C++ version of the R package (available soon on the CRAN).

References

Agresti, A., 2010. Analysis of ordinal categorical data. Wiley Series in Probability and Statistics. Wiley-Interscience, New York.

Bathia, P., Iovleff, S., Govaert, G., 2016. An r package and c++ library for latent block models: Theory, usage and applications. Journal of Statistical Software.

Biernacki, C., Celeux, G., Govaert, G., 2001. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (7), 719–725.

Biernacki, C., Jacques, J., 2016. Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. Statistics and Computing 26 (5), 929–943.

Brault, V., Keribin, C., Mariadassou, M., 2017. Consistency and asymptotic normality of latent blocks model estimators. Tech. rep., Version arXiv:1704.06629.

Candès, E. J., Recht, B., 2009. Exact matrix completion via convex optimization. Foundations of Computational Mathematics 9 (6), 717.

Celeux, G., Govaert, G., 2015. Latent Class Models for Categorical Data. In: Hennig, C., Meila, M., Murthag, F., Rocci, R. (Eds.), Handbook of Cluster Analysis. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Chapman & Hall/CRC, pp. 173–194.

Cousson-Gélie, F., 2000. Breast cancer, coping and quality of life: a semi-prospective study. European Review of Applied Psychology 3, 315–320.

D'Elia, A., Piccolo, D., 2005. A mixture model for preferences data analysis. Computational Statistics and Data Analysis 49 (3), 917–934.

Dillon, W. R., Madden, T. S., Firtle, N. H., 1994. Marketing Research in a Marketing Environment. Irwin.

Fayers, P., Aaronson, N., Bjordal, K., Groenvold, M., Curran, D., Bottomley, A., 2001. Eortc qlq-c30 scoring manual (3rd edition).

Fernández, D., Arnold, R., Pledger, S., 2016. Mixture-based clustering for the ordered stereotype model. Computational Statistics and Data Analysis 93, 46–75.

Giordan, M., Diana, G., 2011. A clustering method for categorical ordinal data. Communications in Statistics – Theory and Methods 40, 1315–1334.

Gouget, C., 2006. Utilisation des modèles de mélange pour la classification automatique de données ordinales. Ph.D. thesis, Université de Technologie de Compiègne.

Govaert, G., Nadif, M., 2013. Co-Clustering. Wiley-ISTE.

Hartigan, J., 1972. Direct clustering of a data matrix. Journal of the American Statistical Association 67 (337), 123–129.

Hartigan, J., 1975. Clustering algorithm. Wiley, New-York.

Hasnat, M. A., Velcin, J., Bonnevay, S., Jacques, J., 2017. Opinion mining from twitter data using evolutionary multinomial mixture models. Econometrics and Statistics , to appear.

Jacques, J., Biernacki, C., 2010. Extension of model-based classification for binary data when training and test populations differ. Journal of Applied Statistics 37 (5), 749–766.

Jollois, F.-X., Nadif, M., 2011. Classification de données ordinales : modèles et algorithmes. In: Proceedings of the 43th conference of the French Statistical Society. Bordeaux, France.

Kaiser, S., Santamaria, R., Khamiakova, T., Sill, M., Theron, R., Quintales, L., Leisch, F., De Troyer., E., 2015. biclust: BiCluster Algorithms. R package version 1.2.0. URL `https://CRAN.R-project.org/package=biclust`

Kaufman, L., Rousseeuw, P. J., 1990. Finding groups in data: An introduction to cluster analysis. Wiley.

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

Lewis, S. J. G., Foltynie, T., Blackwell, A. D., Robbins, T. W., Owen, A. M., Barker, R. A., 2003. Heterogeneity of parkinson's disease in the early clinical stages using a data driven approach. Journal of Neurology, Neurosurgery and Psychiatry 76, 343–348.

Little, R., Rubin, D., 2002. Statistical Analysis with missing data, 2nd Edition. Wiley.

Matechou, E., Liu, I., Fernandez, D., Farias, M., Gjelsvik, B., 2016. Biclustering models for two-mode ordinal data. Psychometrika 81 (3), 611–624.

Pledger, S.and Arnold, R., 2014. Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. Computational Statistics & Data Analysis 71, 241–261.

Podani, J., 2006. Braun-blanquet's legacy and data analysis in vegetation science. Journal of Vegetation Science 17, 113–117.

Ranalli, M., Rocci, R., 2016. Mixture models for ordinal data: A pairwise likelihood approach. Statistics and Computing 26 (1), 529–547.

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

Schepers, J., H-H., B., Mechelen, I., 2017. Maximal interaction two-mode clustering. Journal of Classification 75, 49–75.

Stevens, S., 1946. On the theory of scales of measurement. Science 103 (2684), 677–680.

Vermunt, J., Magidson, J., 2005. Technical Guide for Latent GOLD 4.0: Basic and Advanced. Statistical Innovations Inc., Belmont, Massachusetts.

Vichi, M., 2001. Double k-means Clustering for Simultaneous Classification of Objects and Variables. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 43–52.