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, Université 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: co-clustering, ordinal data, SEM-Gibbs algorithm.

1. Introduction

Historically, clustering algorithms are used to explore data and to provide a simplified representation of them 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. 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 of features.
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

*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])

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 or contingency data (Govaert and Nadif, 2013).

This work focuses on particular type of categorical data, ordinal data, occurring when the categories are ordered (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 example can be in medicine, when patients are asked to evaluate their quality of life on a Likert scale (see Cousson-Gélie (2000) for instance). However, contrary to nominal categorical data, 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; Jollois and Nadif, 2011; D'Elia and Piccolo, 2005; Podani, 2006; Giordan and Diana, 2011; Biernacki and Jacques, 2016). In a co-clustering context, Matechou et al. (2016) recently proposed an approach relying on the proportional odds model (PO), 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.

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 is presented below.

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 $\mathbf{v} = (v_{ik})_{1 \leq i \leq n, 1 \leq k \leq K}$ and the 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 $\mathbf{w}$. The latent block model can be written:

$$\mathrm{p}(\mathbf{x}; \theta) = \sum_{\mathbf{v} \in V} \sum_{\mathbf{w} \in W} \mathrm{p}(\mathbf{v}; \theta) \mathrm{p}(\mathbf{w}; \theta) \mathrm{p}(\mathbf{x} | \mathbf{v}, \mathbf{w}; \theta) \tag{1}$$

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

- V the set of all possible partitions of rows into K groups, W the set of partitions of the columns into L groups,
- $\mathrm{p}(\mathbf{v}; \theta) = \prod_{ik} \alpha_k^{v_{ik}}$ and $\mathrm{p}(\mathbf{w}; \theta) = \prod_{h\ell} \beta_\ell^{w_{h\ell}}$ where α_k and β_ℓ are the row and column mixing proportions, belonging to $[0, 1]$ and summing to 1,
- $\mathrm{p}(\mathbf{x} | \mathbf{v}, \mathbf{w}; \theta) = \prod_{ihk\ell} \mathrm{p}(x_{ih}; \mu_{k\ell}, \pi_{k\ell})^{v_{ik} w_{h\ell}}$ where $\mathrm{p}(x_{ih}; \mu_{k\ell}, \pi_{k\ell})$ is the probability of x_{ij} according to the BOS model (Biernacki and Jacques, 2016) parametrized by $(\pi_{k\ell}, \mu_{k\ell})$ with the so-called precision parameter $\pi_{k\ell} \in [0, 1]$ and position parameter $\mu_{k\ell} \in \{1, ..., m\}$ (detail of $\mathrm{p}(x_{ih}; \mu_{k\ell}, \pi_{k\ell})$ is given below),
- $\theta = (\pi_{k\ell}, \mu_{k\ell}, \alpha_k, \beta_\ell)$ is the whole mixture parameter.

This latent block model relies on the BOS distribution for ordinal data which is now presented.

Ordinal model. 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 in which e_j is the current interval in $\{1, \ldots, n\}$, and y_j the break point in this interval. The BOS distribution is defined as follows:

$$\mathrm{p}(x_{ij}; \mu_{k\ell}, \pi_{k\ell}) = \sum_{e_{m-1}, ..., e_1} \prod_{j=1}^{m-1} \mathrm{p}(e_{j+1} | e_j; \mu_{k\ell}, \pi_{k\ell}) \mathrm{p}(e_1) \tag{2}$$

where

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

with δ a "distance" between μ and an interval $e=\{b^-,\dots,b^+\}$ defined by $\delta(e,\mu)=\min(|\mu-b^-|,|\mu-b^+|)$ and with

$$\mathrm{p}(z_j|e_j;\pi_{k\ell})=\pi\mathbb{I}(z_j=1)+(1-\pi_{k\ell})\mathbb{I}(z_j=0)\quad\text{and}\quad\mathrm{p}(y_j|e_j)=\frac{1}{|e_j|}\mathbb{I}(y_j\in e_j).$$

It is shown in Biernacki and Jacques (2016) that the BOS distribution (2) 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 probability distribution evolving from the uniform distribution (when $\pi_{k\ell}=0$) to a distribution more and more picked 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 1.

The latent block model (1) for ordinal data can finally be written:

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

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)).

3. Model inference

The aim is to estimate θ by maximizing the observed log-likelihood

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

For computational reasons, EM algorithm is not feasible in that co-clustering case (see Govaert and Nadif (2013)), thus we opt for one of its stochastic version denoted by SEM-Gibbs (Keribin et al., 2010).

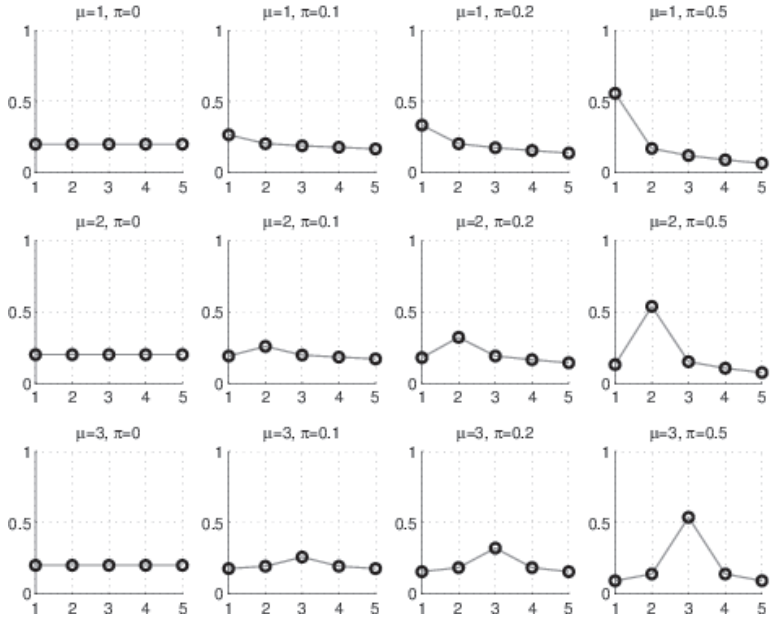

Figure 1: Distribution $\mathrm{p}(x;\mu,\pi)$: shape for $m=5$ and for different values of μ and π.

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 ($\theta^{(0)}$) 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 $1 \leq i \leq n, 1 \leq k \leq K$:

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

   where $f_k(x_{i.}^{(q)}|\mathbf{w}^{(q)}; \theta^{(q)}) = \prod_{h\ell} \mathrm{p}(x_{ih}^{(q)}; \mu_{k\ell}^{(q)}, \pi_{k\ell}^{(q)})^{w_{h\ell}^{(q)}}$ 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 $1 \leq h \leq d, 1 \leq \ell \leq L$:

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

   where $g_\ell(x_{.h}^{(q)}|\mathbf{v}^{(q+1)}; \theta^{(q)}) = \prod_{ik} \mathrm{p}(x_{ih}^{(q)}; \mu_{k\ell}^{(q)}, \pi_{k\ell}^{(q)})^{v_{ik}^{(q+1)}}$.

3. generate the missing data $\hat{x}_{ih}^{(q+1)}|\check{\mathbf{x}}, \mathbf{v}^{(q+1)}, \mathbf{w}^{(q+1)}$ following

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

M step. Estimate θ, 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 $\hat{\theta}$.

Estimating the partition and the missing data. After having chosen the parameter estimation $\hat{\theta}$, a sample of $(\hat{\mathbf{x}}, \mathbf{v}, \mathbf{w})$ is generated with the Gibbs sampling described above with θ fixed to $\hat{\theta}$. 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, we propose to adapt to our situation the ICL-BIC criterion developed in Keribin et al. (2014) in the case of co-clustering of categorical data:

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

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

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

with

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

and

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

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

4. Numerical experiments on synthetic data sets

This section aims to show the efficiency of the SEM-Gibbs algorithm for model parameter estimation as well as the efficiency of the ICL-BIC criterion to choose the number of co-clusters. 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 2 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 are 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 3 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 a Intel Core i7 CPU 2.8GHz, 16Go RAM.

Empirical consistance of the SEM-Gibbs algorithm. Figure 4 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. $\Delta\beta = \sum_{\ell=1}^{L} |\beta_\ell - \hat{\beta}_\ell|/L$),

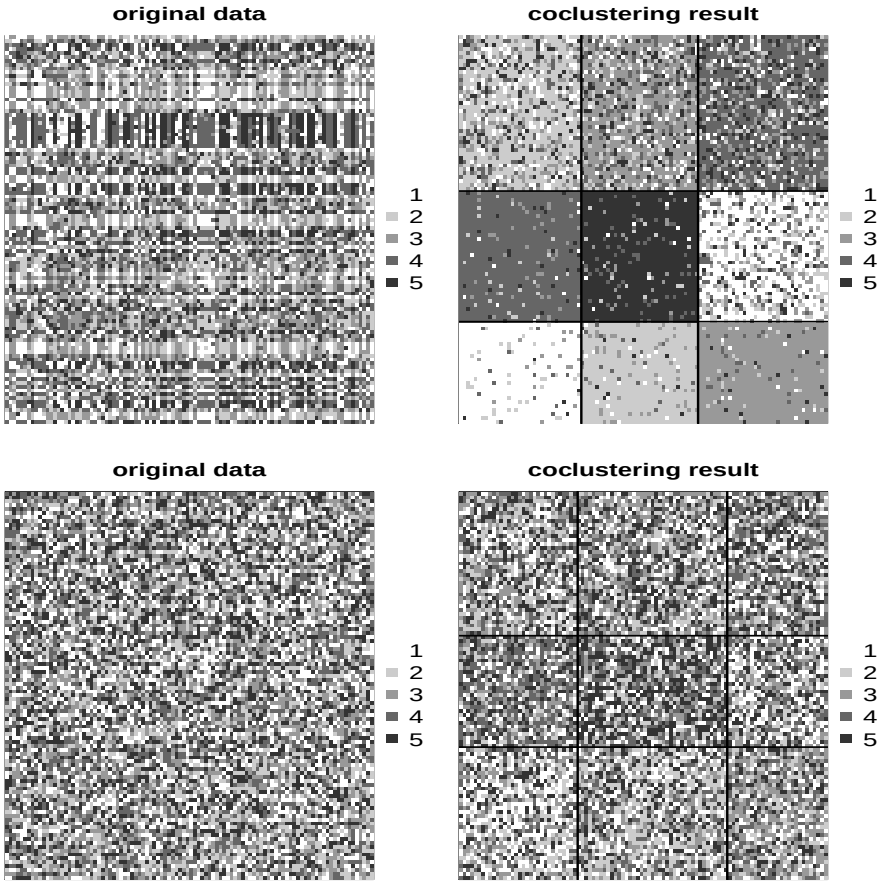

Figure 2: An example of data (left) and co-clustering results (right), for the experimental setting 1 (top) and setting 2 (bottom).

- `ARIr` (resp. `ARIc`): Adjusted Rand Index (ARI) for the row (resp. column) partition.

As it can be seen on Figure 4 and Table 2, the proposed algorithm achieves to obtain very satisfying estimations for the model parameter as well as for the row and column partitions.

	$\Delta\boldsymbol{\mu}$	$\Delta\boldsymbol{\pi}$	$\Delta\alpha$	$\Delta\beta$	`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)
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)

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.

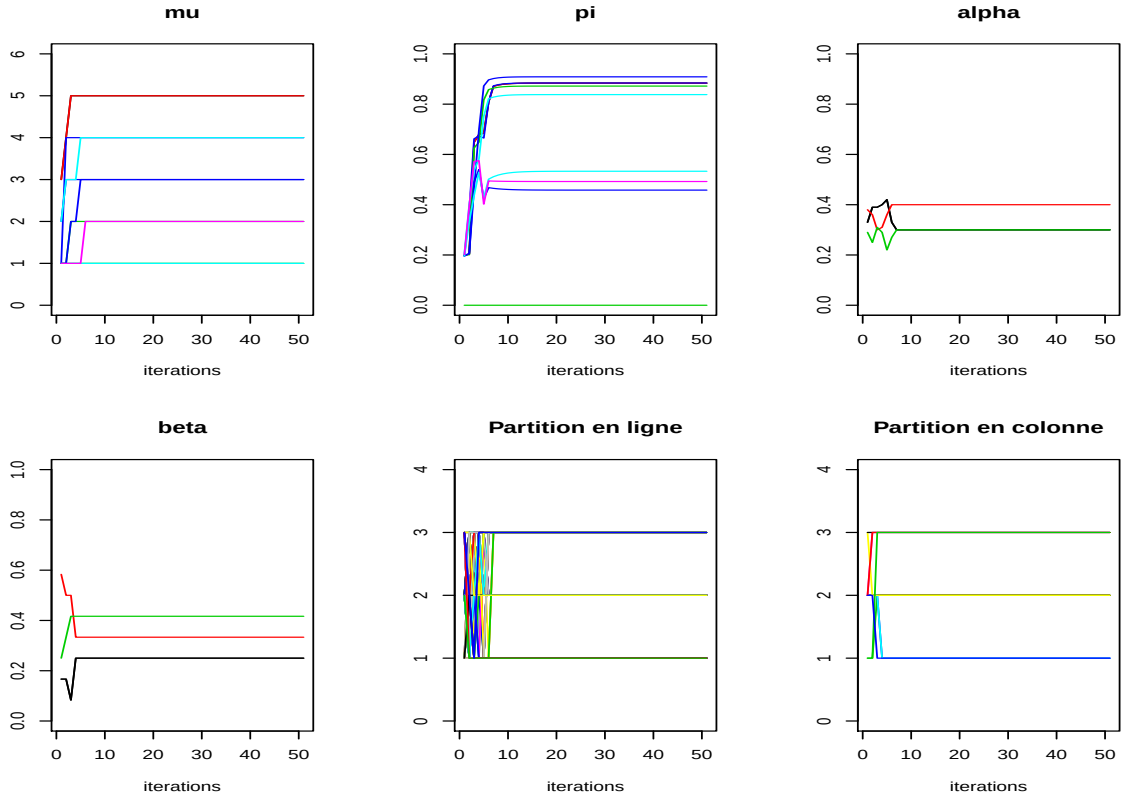

Figure 3: Evolution of the model parameters (one color per parameter $\mu_{k\ell}$, $\pi_{k\ell}$, α_k, β_ℓ) and the row/columns partitions (one color per v_{ik} and $w_{j\ell}$) during the SEM-Gibbs iterations

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 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).

4.2. Efficiency with missing data

In this section, we introduce a given percentage of missing data 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 5. If missing data has almost no impact on the *easy* experimental setting 1, they contribute

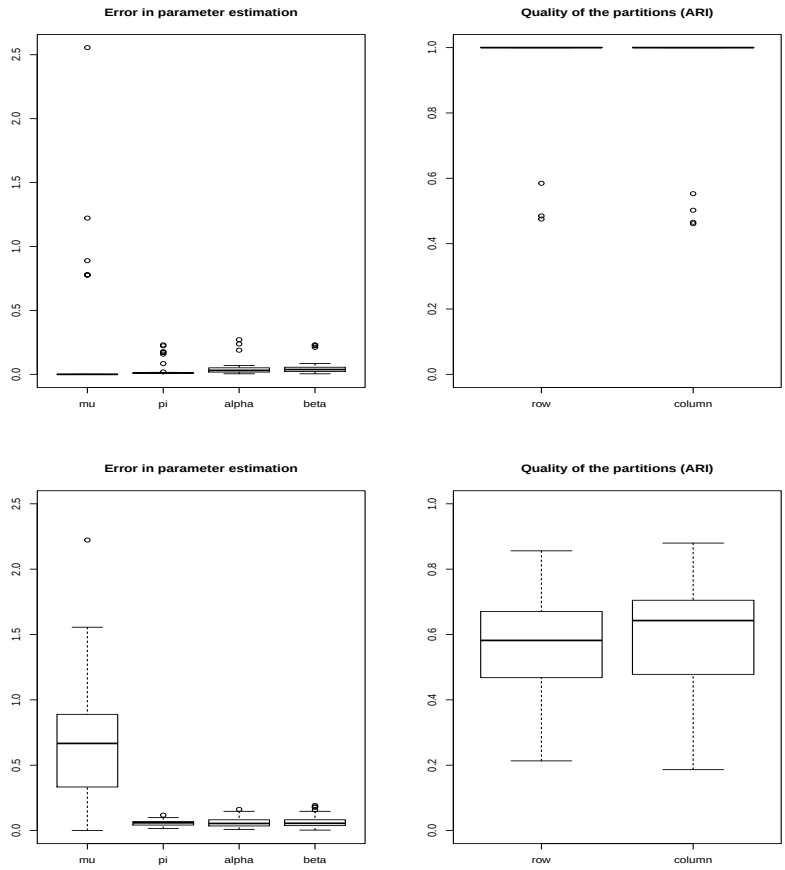

Figure 4: Error on parameter estimation (left) and ARI for the row and column partitions (right), for the experimental setting 1 (top) and setting 2 (bottom).

		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).

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.

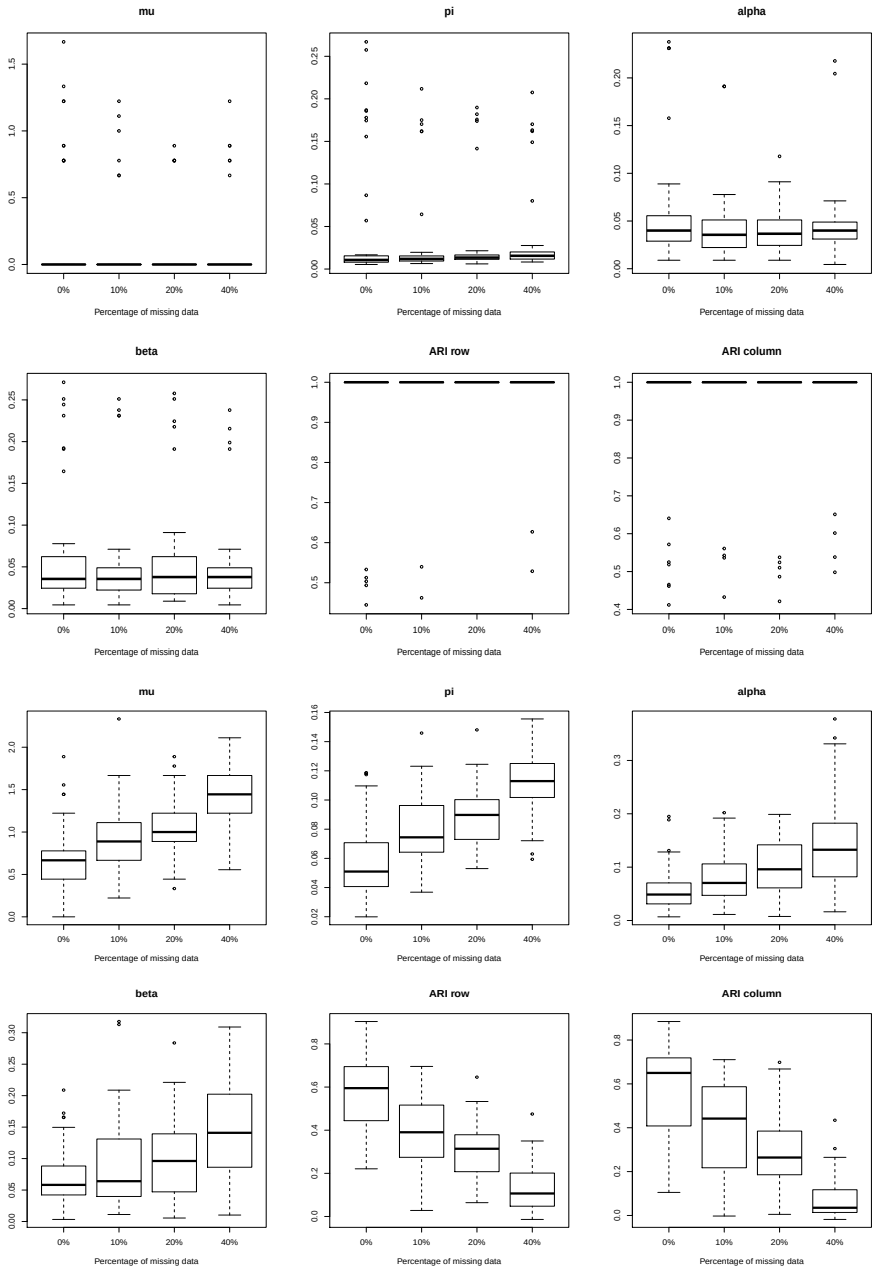

Figure 5: Error on parameter estimation and row and column ARI for different proportion of missing data, for the experimental setting 1 (two top lines) and 2 (two bottom lines).

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 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 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 6.

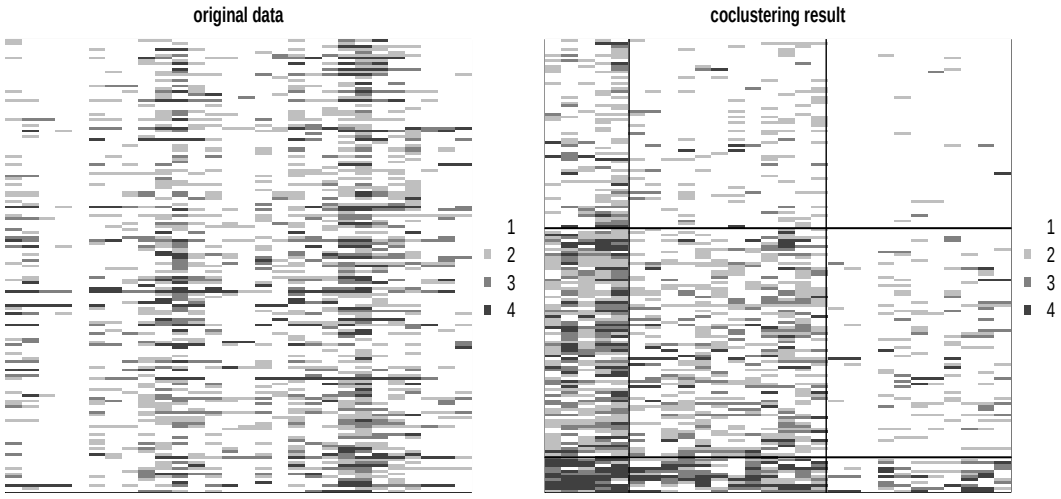

Figure 6: Original EORTC QLQ-C30 data (left) and co-clustering results into 3×3 blocks (right).

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 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 6 (right panel). On this figure, the numbering of the row-clusters is from the bottom to the top

and the numbering of the 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).

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

		ℓ		
		1	2	3
	3	(1,0.60)	(1,0.84)	(1,0.98)
k	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 $(K, L) \in \{2, 3, 4\}^2$ and estimation of $(\mu_{k\ell}, \pi_{k\ell})$ (right) for the 9 co-clusters obtained on the EORTC QLQ-C30 data.

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) and estimated by the proposed strategy. The experiment has been repeated 100 times, and Figure 7 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.

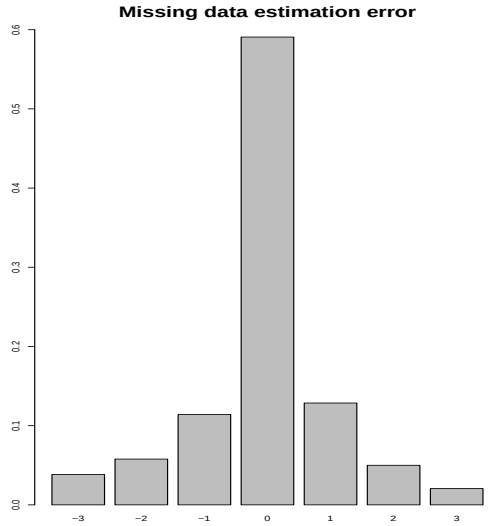

Figure 7: Relative frequency of estimation error when missing observations are artificially introduced in the EORTC QLQ-C30 data.

5.2. Amazone Fine Food Review data

The Amazone 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 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 8). In this sample of the whole dataset, *only* 86.44% of the data are missing. Since with a so 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).

The number of SEM-Gibbs iterations, tuned graphically 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 8, and parameter estimation for the six co-clusters are given in Table 5.

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

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

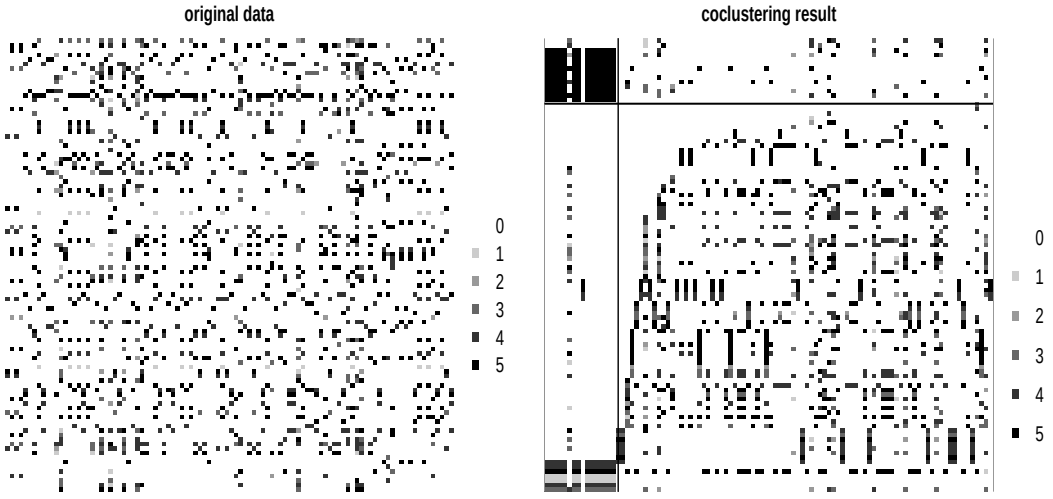

Figure 8: Top 100 Amazone Fine Food Review data (left) and co-clustering result (right).

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

Table 5: 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 $\pi_{22} \simeq 0$).

$\pi_{11} = 0.98$). 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.

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

[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 (...)*"

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 the implementation of a faster version including C++ programming is under study.

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 thinked.

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).

References

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

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.

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

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).

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.

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.

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

Keribin, C., Brault, V., Celeux, G., Govaert, G., 2014. Estimation and selection for the latent block model on categorical data. Statistics and Computing 25 (6), 1201–1216.

Keribin, C., Govaert, G., Celeux, G., 2010. Estimation d'un modèle à blocs latents par l'algorithme SEM. In: Proceedings of the 42th conference of the French Statistical Society. Marseille, France.

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.

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

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