Model-Based Clustering of Multivariate Ordinal Data Relying on a Stochastic Binary Search Algorithm

Christophe Biernacki, Julien Jacques

Model-Based Clustering of Multivariate Ordinal Data Relying on a Stochastic Binary Search Algorithm

Christophe Biernacki Julien Jacques

July 26, 2014

Abstract

We design the first univariate probability distribution for ordinal data which strictly respects the ordinal nature of data. More precisely, it relies only on order comparisons between modalities. Contrariwise, most competitors either forget the order information or add a nonexistent distance information. The proposed distribution is obtained by modeling the data generating process which is assumed, from optimality arguments, to be a stochastic binary search algorithm in a sorted table. The resulting distribution is natively governed by two meaningful parameters (position and precision) and has very appealing properties: decrease around the mode, shape tuning from uniformity to a Dirac, identifiability. Moreover, it is easily estimated by an EM algorithm since the path in the stochastic binary search algorithm is missing. Using then the classical latent class assumption, the previous univariate ordinal model is straightforwardly extended to model-based clustering for multivariate ordinal data. Again, parameters of this mixture model are estimated by an EM algorithm. Both simulated and real data sets illustrate the great potential of this model by its ability to parsimoniously identify particularly relevant clusters which were unsuspected by some traditional competitors.

Keywords. Ordinal data, binary search algorithm, latent variables, EM algorithm.

1 Introduction

Clustering [34] is an important explanatory tool for practitioners to discover some hidden, but hopefully valuable, structures in data sets. Mixture models [33, 25, 7, 27] have now become a standard approach thanks to their ability to dive clustering into a well-posed mathematical context for parameter estimation and model selection (in particular for the selection of the

number of clusters). Additionally, mixture models are able to retrieve and generalize several classical geometric methods and have successful use in very numerous practical situations. In this model-based clustering context, classifying data relies thus on the availability of a suitable distribution probability for the kinds of data at hand which can be numerical [3], rankings [12], functional [14], categorical [10],...

One particular type of categorical data is ordinal data, occurring when the categories are ordered. Such data are very frequent in practice, as for instance in marketing studies where people are asked through questionnaires to evaluate some products or service on an ordinal scale. However, contrary to nominal categorical data, ordinal data have received less attention from a model-based 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 [16] or [19] for instance) or into nominal data (ignoring the order information, see the Latent GOLD software [32]) in order to "recycle" easily related distributions. Indeed, the main works really specific to ordinal data are essentially concentrated on regression with ordinal responses [1]. This paper intends to fill this gap by providing a more valuable ordinal data probability distribution than existing ones.

We give now an overview of the main models and clustering algorithms used for ordinal data. A first approach to ordinal data clustering is a hierarchical clustering with ad-hoc distance metric and similarity measures: Kendall's τ [17], Goodman-Kruskal's γ [9] or Somers' d [31] (see [29] for example). The CUB model [4] is an alternative probability distribution for ordinal data built as a mixture of three components: a binomial distribution which allows to weight all the categories, a Dirac distribution and a uniform distribution. This model is extended in order to get rid of the notion of distance between modalities induced by the use of the binomial distribution (non linear CUB model (NLCUB), [21]). However, the CUB or NLCUB distribution can not be used easily in a clustering context since a mixture of CUB is not identifiable. Several other approaches to ordinal data clustering consist of constraining the multinomial model in order to take into account the order information. This is done by assuming different models for the decrease of the probabilities on each side of the mode of the distribution [11, 15]. Another approach, probably the most successful, assumes that ordinal data are discretizations of latent Gaussian variables, following the earliest model of Everitt [6]. For instance, [11] builds a clustering model using this assumption and estimating the unknown discretization thresholds via an EM algorithm [5]. In [26], a similar approach is extended to multivariate ordinal data by considering a conditional independence assumption and by using a Bayesian paradigm for the model estimation. Let finally cite [8] which proposes a non parametric clustering algorithm based on thresholding contingency tables.

Recently, new probability distributions [2, 12] for ranking data (another kind of categorical data) have been successfully introduced by modeling the data generating process. In the present work a probability distribution for ordinal data is built on a quite similar scheme, assuming that ordinal data results from a search algorithm in a table of ordered categories. By considering a search algorithm based only on comparisons between categories, the ordinal nature of data is wholly respected and none link to any nominal or continuous distribution is necessary. The retained search algorithm is the binary search algorithm since it is optimal in averaged number of comparisons between objects [18]. The parametrization of the model with a position parameter (the modal category) and a precision parameter follows naturally from the model construction, as well as expected properties like the existence of a unique mode, the decrease of the probability distribution on each side of the mode, the possibility to have a uniform or a Dirac distribution. Maximum likelihood estimation of model parameters can be simply performed using an EM algorithm, since the path in the binary search algorithm can be viewed as a particular latent variable. However, this estimation is of combinatorial complexity (as for models based on latent Gaussian variables) but is easily tractable for ordinal data with eight categories, which is the case of main ordinal data (for instance in the reference book [1] all ordinal data have less than eight categories). Extension to more than eight categories could be considered by using a SEM-Gibbs-like estimation algorithm (see [12, 13] for instance), but is not developed in this paper because of the rarity of such data. Finally, the model is extended to the clustering of multivariate ordinal data through a mixture model with a conditional independence assumption [10].

The paper is organized as follows. Section 2 presents the new probability distribution for ordinal data, its properties and the EM parameter estimation algorithm. Section 3 presents the clustering algorithm for multivariate ordinal data, whereas Section 4 is devoted to numerical applications on simulated and real data. Section 5 concludes and draws possible future works.

2 Univariate model design

2.1 Ordinal data as the outcome of a binary search algorithm

A univariate *ordinal* variable μ with m categories $\{l_1, \ldots, l_m\}$ is a *nominal* variable where the m categories are full ordered according to the given order relation "$\prec$": $l_1 \prec \ldots \prec l_m$. In the following, only the indexes of the categories will be used for describing an ordinal variable, thus $\mu \in \{1, \ldots, m\}$.

The major assumption of this work is to consider μ as the outcome of a *search process* within the ordered table $(1, \ldots, m)$ (see [18] for a description of main search algorithms). More precisely, the exact value of μ is unknown at the beginning of the process but basic order comparisons "$\succ$" and "$\prec$" between the unknown μ and any element of $\{1, \ldots, m\}$ are available (note that equality "$=$" is deduced from "$\succ$" and "$\prec$"). Then, the exact value of μ is gradually revealed after using successively these basic order comparisons through the table. The great advantage of this approach is to strictly respect the ordinal nature of μ relying only on order relations between possible elements.

We make also the additional assumption that some *erroneous comparisons* could arise through the search process, so, a possibly erroneous outcome $x \in \{1, \ldots, m\}$ (it means possibly different from μ) can be obtained. Obviously, the value of x depends both on the involved search process and also on the way the wrong comparisons occur. In order to minimize the number of potentially wrong comparisons, it is necessary to minimize the number of comparisons performed during the search process. Since the binary search algorithm is optimal from this point of view [18], we retain it as the best candidate for generating ordinal data.

We recall now the general principle of a binary algorithm in which we added possibility to perform wrong comparisons. It iterates in $m-1$ iterations and the outcome is a data $x \in \{1, \ldots, m\}$. Starting from an interval $e_j = (b_j^-, \ldots, b_j^+) \subset (1, \ldots, m)$, the jth iteration $(1 \leq j \leq m-1)$ is divided into three steps:

Step 1 (break point y_j): choose an element $y_j \in e_j$ to break the interval e_j;

Step 2 (accuracy z_j): choose the accuracy $z_j \in \{0, 1\}$ of the order comparison between y_j and μ where $z_j = 0$ when the comparison is blind (what means not using μ to perform comparison) and where $z_j = 1$ when the comparison is perfect (what means using μ to perform comparison). Thus, wrong comparisons (according to μ) can only arise from blind comparisons.

Step 3 (subinterval e_{j+1}): choose a new search interval $e_{j+1} \in \{e_j^-, e_j^=, e_j^+\}$ depending on y_j and z_j where $e_j^- = (b_j^-, \ldots, y_j - 1)$ is the interval on the left of the break, $e_j^= = \{y_j\}$ is restricted to the break point y_j, $e_j^+ = (y_j + 1, \ldots, b_j^+)$ is the interval on the right of the break.

The sequence of e_j, y_j, z_j resulting from the binary search algorithm is the following:

$$e_1 = \{1, \ldots, m\} \to y_1 \to z_1 \to e_2 \to \ldots \to y_{m-1} \to z_{m-1} \to e_m = \{x\} \to y_m = x. \quad (1)$$

A standard binary algorithm usually would select the break point y_j at the middle of the

interval e_j. Moreover no errors would be considered with such a standard deterministic algorithm, so $z_j = 1$ for all j and, consequently, the interval e_{j+1} would be necessarily the one containing the researched value μ. However, in order to obtain a probability distribution on the whole set of categories $\{1, \ldots, m\}$, we assume that the values of y_j, z_j and e_j are unknown (missing values) and thus we propose to model our lack of knowledge by particular random variables. It leads to a stochastic binary algorithm, defining itself a probabilistic model on the ordinal variable $x \in \{1, \ldots, m\}$, that we describe both now.

2.2 Stochastic binary algorithm and related probabilistic model

A probability distribution is now associated to each decision taken at the three steps of the previous binary search algorithm. It will provide at the end a probabilistic model on $x \in \{1, \ldots, m\}$.

Starting with $\mathrm{p}(e_1) = 1$, each of the three previous steps is defined as follows at iteration j:

Step 1 (break point y_j): y_j is uniform in e_j, so, $\mathbb{I}(.)$ being the indicator function and $|e_j| = b_j^+ - b_j^-$,

$$\mathrm{p}(y_j|e_j) = \frac{1}{|e_j|}\mathbb{I}(y_j \in e_j); \tag{2}$$

Step 2 (accuracy z_j): $z_j \in \{0, 1\}$ is drawn from a Bernoulli of parameters $\pi \in [0, 1]$:

$$\mathrm{p}(z_j|e_j; \pi) = \pi\mathbb{I}(z_j = 1) + (1 - \pi)\mathbb{I}(z_j = 0); \tag{3}$$

Step 3 (subinterval e_{j+1}): the distribution of $e_{j+1} \in \{e_j^-, e_j^=, e_j^+\}$ depends on y_j and z_j:

- **If the comparison is blind** ($z_j = 0$), it is independent on μ and e_{j+1} is chosen with a probability proportional to the length of the three allowed intervals e_j^-, $e_j^=$ and e_j^+:

$$\mathrm{p}(e_{j+1}|y_j, e_j, z_j = 0) = \frac{|e_{j+1}|}{|e_j|}\mathbb{I}(e_{j+1} \in \{e_j^-, e_j^=, e_j^+\}); \tag{4}$$

- **If the comparison is perfect** ($z_j = 1$), the interval containing μ (or the closest interval to μ if μ is no more in the allowed intervals at this step) is retained almost surely:

$$\mathrm{p}(e_{j+1}|y_j, e_j, z_j = 1; \mu) = \mathbb{I}(e_{j+1} = \underset{e \in \{e_j^-, e_j^=, e_j^+\}}{\operatorname{argmin}} \delta(e, \mu))\mathbb{I}(e_{j+1} \in \{e_j^-, e_j^=, e_j^+\}), \tag{5}$$

where δ measures a "distance" between μ and an interval $e = (b^-, \ldots, b^+)$:

$$\delta(e, \mu) = \min(|\mu - b^-|, |\mu - b^+|). \tag{6}$$

From this algorithm, it is easy to express the distribution of x. First, we marginalize on z_j by using (3), (4) and (5)

$$\mathrm{p}(e_{j+1}|e_j, y_j; \mu, \pi) = \pi \mathrm{p}(e_{j+1}|y_j, e_j, z_j = 1; \mu) + (1 - \pi)\mathrm{p}(e_{j+1}|y_j, e_j, z_j = 0). \tag{7}$$

Then we marginalize on y_j by combining with (2)

$$\mathrm{p}(e_{j+1}|e_j; \mu, \pi) = \sum_{y_j \in e_j} \mathrm{p}(e_{j+1}|e_j, y_j; \mu, \pi)\mathrm{p}(y_j|e_j). \tag{8}$$

The last step consists of expressing $\mathrm{p}(x; \mu, \pi)$ $(= \mathrm{p}(e_m; \mu, \pi)$ from $(1))$[1] thanks to the Markovian properties of the e_j's and by using (8)

$$\begin{aligned} \mathrm{p}(x; \mu, \pi) &= \sum_{e_{m-1}, \ldots, e_1} \mathrm{p}(e_m, e_{m-1}, \ldots, e_1; \mu, \pi) && (9) \\ &= \sum_{e_{m-1}, \ldots, e_1} \prod_{j=1}^{m-1} \mathrm{p}(e_{j+1}|e_j; \mu, \pi)\mathrm{p}(e_1). && (10) \end{aligned}$$

In the following we will note this model BOS for *Binary Ordinal Search*.

2.3 Properties of the BOS model

The distribution $\mathrm{p}(x; \mu, \pi)$ benefits from very appealing properties and its parameters have a nice meaning. All propositions and related proofs are given in Appendix A. Its shape for different values of μ and π is also displayed on Figure 1.

Mode μ is a *position* parameter indicating the mode of the distribution (see Proposition A.3). In addition, this mode is unique and probabilities monotically decrease around it (see Proposition A.6).

Precision π is a *precision* parameter since the mode μ is uniformly more pronounced when π grows both from absolute ($\mathrm{p}(\mu; \mu, \pi)$ is an increasing function of π) and relative ($\mathrm{p}(\mu; \mu, \pi) - \mathrm{p}(x; \mu, \pi)$ is an increasing function of π for all $\mu \neq x$) point of views (see

[1] We will do the slight abuse of notation $\mathrm{p}(x; \mu, \pi) = \mathrm{p}(\{x\}; \mu, \pi)$.

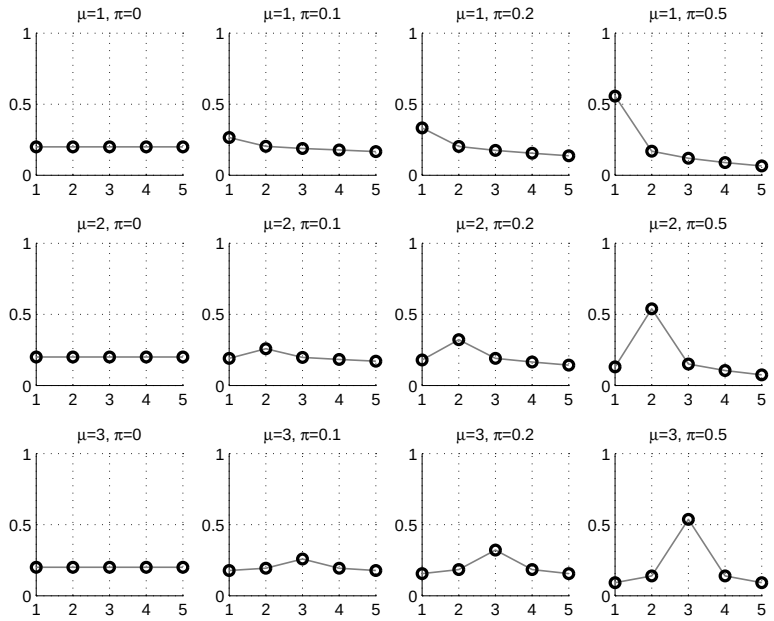

Figure 1: Distribution $\mathrm{p}(x;\mu,\pi)$: shape for $m = 5$ and for different values of μ and π.

Propositions A.4 and A.5 respectively). At the limit, the distribution is a Dirac in μ when $\pi = 1$ (see Proposition A.2). On the contrary, the distribution is uniform on $\{1,\ldots,m\}$ when $\pi = 0$ (see Proposition A.1).

Identifiability The distribution is identifiable on its parameters (μ,π) if $\pi > 0$ (see Proposition A.7).

2.4 Maximum likelihood estimation by EM algorithm

Let consider a n-sample $(x_1,\ldots,x_n)$ independent and identically distributed from $\mathrm{p}(x;\mu,\pi)$. The aim is to estimate the parameters (μ,π) by maximizing the log-likelihood $\ell(\mu,\pi) = \sum_{i=1}^{n} \ln \mathrm{p}(x_i;\mu,\pi)$. The EM algorithm [5] can be used since we are faced to a model with missing values: the break points y_{ij}, the accuracies z_{ij} and the subintervals e_{ij} ($i = 1,\ldots,n$ and $j = 1,\ldots,d$). These missing values are in the following quoted by $\mathrm{c} = (c_1,\ldots,c_n)$, $c_i = (\{y_{ij}, z_{ij}, e_{ij}\}_{j=1,\ldots,m-1})$ being the missing values associated to individual x_i.

Starting from initial parameters $(\mu,\pi)^{[0]}$ and noting C_i the space where c_i stands, the qth iteration ($q \geq 1$) of the algorithm is the following:

- **E Step**: for all $c_i \in C_i$ ($i = 1,\ldots,n$), compute the conditional probabilities of the

missing data

$$\mathrm{p}(c_i|x_i;\mu^{[q]},\pi^{[q]}) = \frac{\mathrm{p}(c_i,x_i;\mu^{[q]},\pi^{[q]})}{\mathrm{p}(x_i;\mu^{[q]},\pi^{[q]})}. \tag{11}$$

- **M Step**: maximize over $\mu^{[q+1]} \in \{1,\ldots,m\}$ the expected conditional completed log-likelihood

$$\bar{\ell}_c(\mu^{[q+1]}) = \sum_{i=1}^{n} \sum_{c_i \in C_i} \mathrm{p}(c_i|x_i;\mu^{[q]},\pi^{[q]}) \ln \mathrm{p}(x_i,c_i;\mu^{[q+1]},\pi^{[q+1]}) \tag{12}$$

where

$$\pi^{[q+1]} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m-1} \mathrm{p}(z_{ij}=1|x_i;\mu^{[q]},\pi^{[q]})}{n(m-1)}. \tag{13}$$

The algorithm is stopped typically when a predefined threshold $\epsilon > 0$ is reached in the relative change of the log-likelihood: $|\ell(\mu^{[q+1]},\pi^{[q+1]}) - \ell(\mu^{[q]},\pi^{[q]})| < \epsilon$.

In practice, it is easier to run the whole EM algorithm for each possible value of $\mu \in \{1,\ldots,m\}$ and finally to retain the run leading to the largest log-likelihood value.

Computational cost Computing the log-likelihood $\ell(\mu,\pi)$ and all the terms in Equations (11), (12) and (13) present in both the E Step and the M Step relies on computing the key set $\mathcal{P}_i = \{\mathrm{p}(c_i,x_i;\mu,\pi), c_i \in C_i\}$. Because of the combinatorial process generating the data, the cardinal of $\mathcal{P}_i$ is of exponential order and then its computational cost is of exponential order also. However, ordinal data are expected to usually have a low number of categories, typically $m \leq 8$. For instance, in the reference book [1], all ordinal data have less than 8 categories. In this case, the computational time of $\mathcal{P}_i$ remains very low as it is illustrated in Figure 2 (we used a MATLAB code on a Intel Core i7-3537U CPU 2.00GHz, 8Go RAM). Note also that computing $\mathcal{P}_i$ for a n-sample is then no really more complex since only m different modality values are present in this whole sample.

3 Latent class extension for clustering

In this section a clustering algorithm for multivariate ordinal data is proposed, based upon a mixture of multivariate BOS models built with a conditional independence assumption.

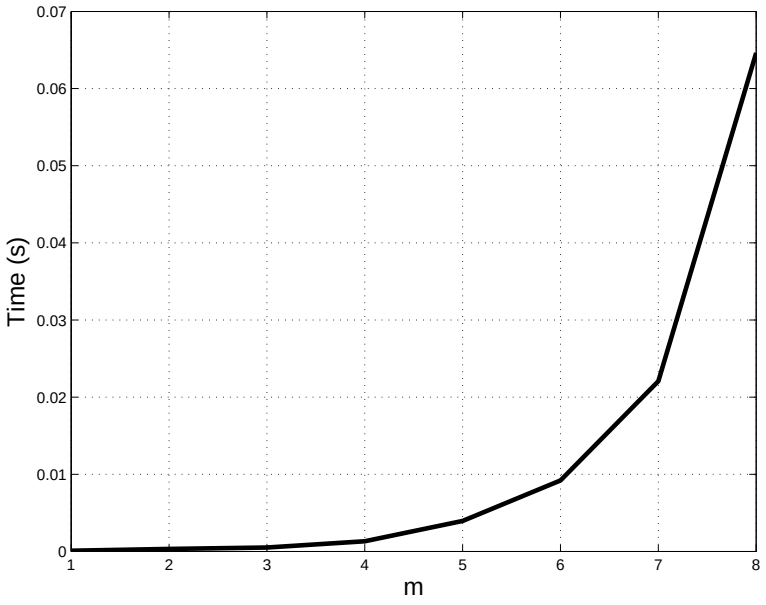

Figure 2: Time to compute the set $\mathcal{P}_i = \{\mathrm{p}(c_i, x_i; \mu, \pi), c_i \in C_i\}$ which is the key set to perform one iteration of the EM algorithm.

3.1 Multivariate probabilistic model for clustering

Let $\mathbf{x} = (x_h)_{1 \leq h \leq d}$ be a multivariate ordinal variable, where the hth component x_h is an ordinal variable of m_h categories. Let $\boldsymbol{w} = (w_1, \ldots, w_g) \in \{0, 1\}^g$ be a group indicator variable, such that $w_k = 1$ if the observation belongs to cluster k and $w_k = 0$ otherwise. w is assumed to follow a one order multinomial distribution:

$$w \sim \mathcal{M}(1, \alpha_1, \ldots, \alpha_g)$$

where α_k is the mixing proportion of cluster k, $\alpha_k > 0$ and $\sum_{k=1}^{g} \alpha_k = 1$. Conditionally on cluster k, the distribution of $\mathbf{x}$ is assumed to be:

$$\mathrm{p}(\mathbf{x}|w_k = 1; \boldsymbol{\mu}_k, \boldsymbol{\pi}_k) = \prod_{h=1}^{d} \mathrm{p}(x_h; \mu_k^h, \pi_k^h)$$

where $\boldsymbol{\mu}_k = (\mu_k^1, \ldots, \mu_k^d)$ and $\boldsymbol{\pi}_k = (\pi_k^1, \ldots, \pi_k^d)$. This conditional independence assumption states that, conditionally on the belonging to cluster k, the d ordinal responses of an individual are independently drawn from d univariate BOS models of parameters μ_k^h and π_k^h ($h = 1, \ldots, d$). Conditional independence is a standard assumption for categorical data [10].

The marginal distribution of $\mathbf{x}$ is then

$$\mathrm{p}(\mathbf{x};\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\pi}) = \sum_{k=1}^{g} \alpha_k \mathrm{p}(\mathbf{x}|w_k = 1;\boldsymbol{\mu}_k,\boldsymbol{\pi}_k)$$

with $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_g)$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_g)$ and $\boldsymbol{\pi} = (\boldsymbol{\pi}_1,\ldots,\boldsymbol{\pi}_g)$.

3.2 Maximum likelihood estimation by EM algorithm

Let consider a n-sample $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$ of multivariate ordinal data, independent and identically distributed from $\mathrm{p}(\mathbf{x};\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\pi})$. As for the parameter estimation in the homogeneous univariate case (Section 2.4), the model parameters $(\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\pi})$ can be estimated by maximizing the log-likelihood $\ell(\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\pi}) = \sum_{i=1}^{n} \mathrm{p}(\mathbf{x}_i;\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\pi})$ using an EM algorithm, with one additional latent variable: the group indicators $(\boldsymbol{w}_1,\ldots,\boldsymbol{w}_n)$ of the observed data. Thus, the missing values associated to the individual $\mathbf{x}_i$ are now $c_i = \left\{\boldsymbol{w}_i, \{y_{ijk}^h, z_{ijk}^h, e_{ijk}^h\}_{1\leq j\leq m^h-1,\, 1\leq k\leq g,\, 1\leq h\leq d}\right\}$ with straightforward notations. Starting from initial parameters $(\boldsymbol{\alpha},\boldsymbol{\mu},\boldsymbol{\pi})^{[0]}$, the qth iteration $(q \geq 1)$ of the algorithm is the following:

- **E Step**: compute the conditional probabilities, for all $1 \leq i \leq n$ and $1 \leq k \leq g$

$$\mathrm{p}(w_{ik} = 1|\mathbf{x}_i;\boldsymbol{\alpha}^{[q]},\boldsymbol{\mu}^{[q]},\boldsymbol{\pi}^{[q]}) = \frac{\alpha_k^{[q]}\mathrm{p}(\mathbf{x}_i|w_{ik}=1;\boldsymbol{\mu}_k^{[q]},\boldsymbol{\pi}_k^{[q]})}{\sum_{k'=1}^{g}\alpha_{k'}^{[q]}\mathrm{p}(\mathbf{x}_i|w_{ik'}=1;\boldsymbol{\mu}_{k'}^{[q]},\boldsymbol{\pi}_{k'}^{[q]})}.$$

- **M Step**: update the estimation of the model parameters as follows:
  - for the mixing proportion $(1 \leq k \leq g)$,

$$\alpha_k^{[q+1]} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{p}(w_{ik} = 1|\mathbf{x}_i;\boldsymbol{\alpha}^{[q]},\boldsymbol{\mu}^{[q]},\boldsymbol{\pi}^{[q]}),$$

  - for the mode and precision parameters, for all $(k,h) \in \{1,\ldots,g\}\times\{1,\ldots,d\}$, $(\mu_k^h,\pi_k^h)^{[q+1]}$ are estimated by an internal EM algorithm (as described in Section 2.4), by weighting each observation with $\mathrm{p}(w_{ik} = 1|\mathbf{x}_i;\boldsymbol{\alpha}^{[q]},\boldsymbol{\mu}^{[q]},\boldsymbol{\pi}^{[q]})$.

As for the EM algorithm described in Section 2.4, the algorithm is stopped when a predefined threshold is reached in the relative change of the log-likelihood.

3.3 Comparison with state of the art methods

Competitors to the proposed model-based algorithm (multivariate BOS mixture) are mainly the multinomial model (which ignores the order information), the Gaussian model (which assigns an arbitrary numerical value to each category) and the models based on the discretization of latent Gaussian variables [11, 26].

From a model complexity point of view (*i.e.* the number of continuous parameters), the multivariate BOS mixture model is the most parsimonious model (see Table 1). From a probabilistic point of view, the multinomial model, which is an efficient competitor for large data sets thanks to its high flexibility, has the drawback to be totally non identifiable for univariate ordinal data and thus to be unavailable in this case. From a computational point of view, the model based on the discretization of latent Gaussian variables suffers from numerical integration difficulties in its frequentist version [11]. In its Bayesian version [26], the main problem is the large amount of hyper parameters to fix, whose choice should significantly influence the resulting clustering. The less computational demanding estimation methods are the multinomial and the Gaussian models. However, as we noted at the end of Section 2.4, the BOS is particularly fast to estimate until $m = 8$ modalities, which corresponds to a frequent limit for ordinal data.

model	continuous	*discrete*	$g = 2, d = 4, m^h = 4$
BOS mixture	$g - 1 + gd$	gd	9 (*+8 discrete*)
multinomial	$g - 1 + g\sum_h (m^h - 1)$	-	25
Gaussian (diagonal)	$g - 1 + 2gd$	-	17
Latent Gaussian [11]	$g - 1 + gd(d + 3)/2 + \sum_h (m^h - 1)$	-	41

Table 1: Number of continuous and discrete parameters for the BOS mixture and its competitors.

4 Numerical illustration

The aim of this section is first to evaluate the robustness of the BOS model (through simulation studies) and second to illustrate its usefulness on real data sets.

4.1 Simulated data

In these simulation studies, the robustness of the BOS model to *data scale* and to *conditional dependency* is evaluated.

Robustness to data scale As it is argued throughout this paper, the BOS model is totally in accordance with the fundamental nature of ordinal data which consists in ordered modalities without notion of distances. To illustrate this essential property in a model-based clustering context, let consider the following univariate distribution of ordinal data x (with $m = 4$ modalities), based on a binned univariate Gaussian distribution ($h = 1, \ldots, 4$, $k = 1, 2$):

$$\mathrm{p}(x = h; a_k, b_1, \ldots, b_5) = \mathrm{p}(b_h \leq x^{gaus} < b_{h+1}) \quad \text{with} \quad x^{gaus} \sim \mathcal{N}(a_k, 1). \tag{14}$$

$\mathcal{N}(a, 1)$ designates the Gaussian distribution of center a and unit variance, the bounds $b_1, \ldots, b_5$ satisfying the following constraint ($\{h, h', H\} \in \{1, \ldots, 4\}^3$)

$$\exists! H \text{ s.t. } (1)\forall h \neq H, \mathrm{p}_{kH} \geq \mathrm{p}_{kh}, (2)\forall h < h' < H, \mathrm{p}_{kh} \leq \mathrm{p}_{kh'}, (3)\forall H < h < h', \mathrm{p}_{kh} \geq \mathrm{p}_{kh'}. \tag{15}$$

The goal of such constraints is to obtain decreasing probabilities $\mathrm{p}_{kh} = \mathrm{p}(x = h; a_k, b_1, \ldots, b_5)$ around a unique modal interval. In this manner, each distribution $\mathrm{p}_k = \{\mathrm{p}_{kh}\}_{1 \leq h \leq 4}$ respects the property of monotonic decrease around the mode held by the BOS distribution, even if it can be *very different from* a BOS distribution.

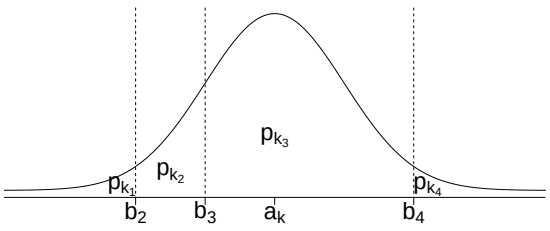

Figure 3: Illustration of the parameters a_k, b_2, b_3, b_4 and the associated probabilities p_{kh} ($1 \leq h \leq 4$).

Then, 100 samples of size $n = 100$ are drawn from the bi-component mixture $0.5\mathrm{p}_1 + 0.5\mathrm{p}_2$ where parameters a_1, a_2, b_1, b_2, b_3, b_4 and b_5 are fixed as follows for each n-sample. Figure 3 illustrates the parameters a_k, b_2, b_3, b_4 and the associated probabilities p_{kh} ($1 \leq h \leq 4$).

- $a_1 = 0$ and a_2 is deduced from a given Bayes error value $\mathcal{E}$ of the underlying Gaussian mixture. We plan two different Bayes error values $\mathcal{E} \in \{0.1, 0.2\}$.
- Conditionally to 15, the bounds b_h ($h = 1, \ldots, 5$) are randomly chosen with the following

	$\mathcal{E} = 0.1$		$\mathcal{E} = 0.2$	
	BOS	Gaussian	BOS	Gaussian
$\text{cor}(\Delta, \hat{\mathcal{E}})$	0.0436	0.5144	0.0128	0.5173
$\text{mean}(\hat{\mathcal{E}})$	0.2676	0.3332	0.3601	0.3796
$\text{sd}(\hat{\mathcal{E}})$	0.1605	0.1611	0.1144	0.1192

Table 2: Results of the robustness to data scale. We have noted $\Delta = |(b_4 - b_3) - (b_3 - b_2)|$ the difference between the ranges of the second and the third intervals, $\hat{\mathcal{E}}$ the empirical error, $\text{cor}(\cdot, \cdot)$ the empirical correlation, $\text{mean}(\cdot)$ the empirical mean, $\text{sd}(\cdot)$ the standard deviation.

design

$$b_1 = -\infty, \quad b_2 \sim \mathcal{N}(0, 1), \quad b_3|b_2 \sim b_2 + u_1, \quad b_4|b_3 \sim b_3 + 8u_2, \quad b_5 = +\infty,$$

where u_1 and u_2 designate two independent random variables from the uniform distribution on $[0, 1]$. In this way, the second and the third intervals can possibly have very different ranges $b_3 - b_2$ and $b_4 - b_3$, the difference of these ranges being noted $\Delta = |(b_4 - b_3) - (b_3 - b_2)|$. It is an important matter for illustrating robustness of the BOS model to scale data.

On each n-sample, a two component BOS model and a two component Gaussian model[2] are estimated by their respective suitable EM algorithm. Table 2 displays results which clearly indicate that the BOS model is drastically less sensitive to the choice of bounds b_h than the Gaussian model (see the correlation values) and also that the BOS model provides more accurate partitioning (see the mean and the standard deviation of the empirical error values).

Robustness to conditional dependency In order to evaluate the robustness of the BOS mixture model to class dependency, we define a specific mixture model in which the latent class assumption is violated. For this, let consider the following cumulative distribution function (cdf) of a *mixture of Gaussian copula*:

$$\text{P}(\mathbf{x}; \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\pi}, \Gamma) = \sum_{k=1}^{g} \alpha_k \Phi_d(\Phi^{-1}(\text{P}(x^1; \mu_k^1, \pi_k^1)), \ldots, \Phi^{-1}(\text{P}(x^d; \mu_k^d, \pi_k^d)); \mathbf{0}, \Gamma), \tag{16}$$

where $\text{P}(x; \mu, \pi)$ is the univariate BOS cdf, $\Phi(\cdot)$ is the cdf of $\mathcal{N}(0, 1)$ and $\Phi_d(\cdot; \mathbf{0}, \Gamma)$ is the cdf of the d-variate normal distribution with center $\mathbf{0}$ and with *correlation matrix* Γ as covariance

[2]Note that a two component multinomial model is not performed in this univariate situation since it is not identifiable.

ρ	$\mathcal{E} = 0.1$	$\mathcal{E} = 0.2$	$\mathcal{E} = 0.3$
0.0	0.1083 (0.0088)	0.2269 (0.0692)	0.3407 (0.0537)
0.4	0.1125 (0.0846)	0.2469 (0.0908)	0.3893 (0.0790)
0.8	0.1391 (0.0798)	0.2613 (0.0759)	0.4151 (0.0800)

Table 3: Results of the robustness to conditional dependency. Mean of the empirical Bayes error $\hat{\mathcal{E}}$ is displayed (its standard deviation in parenthesis).

matrix. The Gaussian copula [28] has the nice property to easily define a kind of multivariate BOS model for which marginal distributions are exactly univariate BOS and for which pairwise component correlations are totally described by the correlation matrix Γ.

In our numerical experiments, we consider bivariate data ($d = 2$) with four modalities each ($m_1 = m_2 = 4$), a bi-component mixture ($g = 2$) with BOS parameters $\boldsymbol{\mu}_1 = (2, 3)$, $\boldsymbol{\mu}_2 = (3, 2)$, $\boldsymbol{\pi}_1 = \boldsymbol{\pi}_2 = (\pi, \pi)$, with equal mixture coefficients ($\alpha_1 = \alpha_2 = 0.5$), and with a correlation matrix Γ where the unique correlation coefficient is noted ρ. The value of π is fixed from a given value of the Bayes error $\mathcal{E}$. Note that in the case $\rho = 0$, the latent class assumption (made by the BOS model) is satisfied.

Twenty samples of size $n = 1000$ are drawn from (16) with crossed designs $\mathcal{E} \in \{0.1, 0.2, 0.3\}$ and $\rho \in \{0.0, 0.4, 0.8\}$ and a bi-component BOS mixture model is estimated on each sample with the EM algorithm. Estimated Bayes error $\hat{\mathcal{E}}$ is displayed in Table 3. It appears that $\hat{\mathcal{E}}$ are poorly influenced by the class conditional dependence, especially when components are well-separated, situation of particular interest in the clustering framework.

4.2 Application to AERES evaluation

Data Created by the French programme law on research of 2006 and up and running since March 2007, the AERES (Evaluation Agency for Research and Higher Education) is tasked with evaluating research and higher education institutions, research organisations, research units, higher education programs and degrees and with approving their staff evaluation procedures. The evaluation is generally done on an ordinal scale, and the results are available on the agency website[3]. In the present study, the Bachelor's degrees evaluation of March 2011 for the 23 universities of the French academies of Bordeaux, Toulouse, Lyon, Montpellier, Grenoble are analyzed. The universities have been evaluated through four criteria: programme leadership (pilot training, PT), educational project (EP), schemes for helping students to succeed (support success, SS), integration of graduates into the job market and continuation

[3]http://www.aeres-evaluation.com/

of chosen studies (employability and further studies, EFS). The evaluations on a ordinal scale {A+, A, B, C} are available in Appendix B.

Global analysis First of all, a global analysis is performed by estimating the BOS model ($g = 1$) on the whole data set by the one-cluster EM algorithm, and the corresponding parameter estimation is obtained:

$$\hat{\boldsymbol{\mu}} = (\text{B,A,B,B}) \quad \text{and} \quad \hat{\boldsymbol{\pi}} = (0.37, 0.39, 0.27, 0.59).$$

This preliminary analysis exhibits that these French universities have relatively good educational projects, but are less efficient in term of program leadership, schemes for helping students to succeed, and integration of graduates into the job market and continuation of chosen studies. Moreover, the evaluations concerning the support success are more heterogeneous (low value of $\hat{\pi}_3$) than the employability and further studies (higher value for $\hat{\pi}_4$). A clustering analysis may allow to identify more homogeneous groups of universities.

Cluster analysis In a second step, a clustering is performed with one to six clusters. Mixture of multivariate BOS models are compared with the multinomial and with the Gaussian (diagonal) models, on the basis of the BIC criterion [30] defined by

$$\text{BIC} = \hat{\ell} - 0.5\nu \ln(n), \tag{17}$$

where $\hat{\ell}$ denotes the maximum log-likelihood value of the model at hand and ν the corresponding number of continuous parameters. The model having the largest BIC is selected. Results in Table 4 show that the best model is a mixture of BOS with four components. Note that the Gaussian model fails in estimating more than two components because of degeneracy problems appearing for larger number of clusters when considering ordinal variables as continuous ones.

Model	$g = 1$	$g = 2$	$g = 3$	$g = 4$	$g = 5$	$g = 6$
BOS	-111.90	-109.14	-107.80	**-104.25**	-108.49	-114.28
Multinomial	**-108.37**	-111.50	-120.08	-135.01	-151.84	-169.75
Gaussian (diagonal)	**-105.76**	-109.31	NaN	NaN	NaN	NaN

Table 4: Values of the BIC criterion for the BOS mixture model and its competitors for 1 to 6 clusters on the AERES data set.

Before to analyze the clustering into four groups with the BOS mixture model, let graph-

ically compare the clustering obtained by the three models for $g = 2$. Using the projection of data into the first principal plan of multiple correspondance analysis (Figure 4), we note that the partition obtained by the proposed algorithm is closer to the Gaussian partition than to the multinomial one.

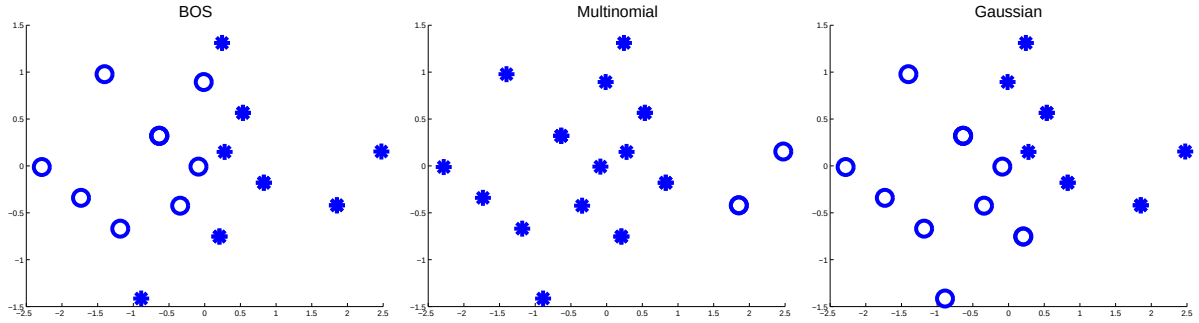

Figure 4: Clustering of the AERES data set into two clusters (circles and stars) with the BOS mixture model and its competitors displayed in the first principal plan of multiple correspondance analysis.

Let now concentrate our analysis on the clustering into four groups with the BOS mixture selected by BIC (Figure 5 and Table 5). The first cluster is composed by universities with homogeneous high scores for all criteria, whereas the third cluster consists also in homogeneous but lower scores. The second cluster is composed by universities having heterogeneous scores for the different criteria: good and very good for the educational project and the support success, but poor for the employability and further studies of the students. Finally, the fourth cluster is composed by the universities having almost the lower scores at the four criteria. Notice finally that all groups and all evaluation criteria have much more precision (higher value of π_k^h) than in the preliminary global analysis.

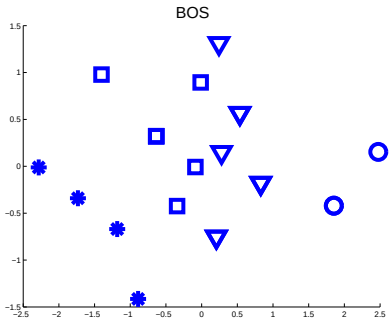

Figure 5: Clustering of the AERES data set into four clusters with the BOS mixture model displayed in the first principal plan of multiple correspondance analysis.

cluster	proportion α_k	mode $\boldsymbol{\mu}_k$	precision $\boldsymbol{\pi}_k$
1	0.30	(A, A, A, B)	(0.89, 0.62, 0.83, 0.83)
2	0.18	(B, A, A+, C)	(0.36, 0.73, 0.99, 0.99)
3	0.38	(B, B, B, B)	(0.86, 0.48, 0.69, 0.99)
4	0.13	(C, B, B, C)	(0.99, 0.99, 0.62, 0.99)

Table 5: Parameter estimation for the BOS mixture model with four clusters on the AERES data set.

4.3 Application to Arthritis data

Data The data come from a clinical trial in which patients suffering from rheumatoid arthritis are randomly assigned to a treatment group and to a placebo group. At the start of the study and after one month, three months and five months of treatment, patients assessed their arthritis on the ordinal scale: good (quoted by 3), fair (2), poor (1). This data, early studied in [20] and [1], has been considered in a clustering context using only the patients assessments after one and five months in [26]. In the present work, all assessments (one, three and five months) are considered.

Global analysis We begin by checking the estimated BOS parameters for the one cluster case ($g = 1$) and we obtain:

$$\hat{\boldsymbol{\mu}} = (2, 3, 3) \quad \text{and} \quad \hat{\boldsymbol{\pi}} = (0.15, 0.10, 0.22).$$

It appears that arthritis seems to evolve favorably over the months but precision of the patient perception is very heterogeneous (low π^h values).

Cluster analysis In a second step, a clustering is performed with one to six clusters. Mixture of multivariate BOS models are compared with the multinomial and with the Gaussian (diagonal) models, on the basis of the BIC criterion. Results in Table 6 show that the best model of all is a mixture of BOS with three components. Note that the Gaussian model fails again in estimating more than three components.

Table 7 displays estimated parameters of the BOS model with three components. Each cluster corresponds to a stable perception of the disease evolution (good, fair and poor) over the months but with important differences concerning precision. Precision of clusters 1 (poor arthritis) and 3 (good) increases months after months (π_k^hs globally increase) thus suggests that the effect or the no effect of the treatment (or of the placebo) is more and more

Model	$g = 1$	$g = 2$	$g = 3$	$g = 4$	$g = 5$	$g = 6$
BOS	-941.65	-873.75	**-853.58**	-861.78	-868.03	-879.50
Multinomial	-936.18	**-861.53**	-862.37	-879.73	-897.82	-917.56
Gaussian (diagonal)	**-1008.65**	-1028.48	-1048.32	NaN	NaN	NaN

Table 6: Values of the BIC criterion for the BOS mixture model and its competitors for 1 to 6 clusters on the Arthritis data set.

significant in opposite directions. Precision of cluster 2 (more than the half of the patients) is quite intermediate and without any important evolution over the time, grouping patients with quite heterogeneous level of Arthritis.

cluster	proportion α_k	mode $\boldsymbol{\mu}_k$	precision $\boldsymbol{\pi}_k$
1	0.11	(1, 1, 1)	(0.67, 0.88, 0.97)
2	0.55	(2, 2, 2)	(0.47, 0.42, 0.44)
3	0.34	(3, 3, 3)	(0.74, 0.72, 0.93)

Table 7: Parameter estimation for the BOS mixture model with four clusters on the Arthritis data set.

5 Concluding remarks

We have designed a natural, parsimonious and meaningful distribution for ordinal variables which has the great advantage to totally respect the true nature of such data: ordered categories without any underlying numerical distance. This model is easy to implement through an EM algorithm. The resulting distribution could thus be seen as an efficient tradeoff between multinomial and Gaussian ones.

Extension to multivariate model-based clustering has also been straightforwardly performed by making the assumption of conditional independence, which is classical for categorical data. Its use in real data sets showed its ability to discover unrevealed clusters by traditional competitor methods.

In our future works, we plan to reuse our ordinal model in a biclustering context where, similarly to the one-way clustering developed in the present paper, there is again a lack of really specific distributions for ordinal data. This new biclustering approach will have to be compared to this one based on the proportional odds model [24] proposed in [23]. Another pursuit of the work is to break down the latent class assumption by introducing dependence

between ordinal variables. One way to do this could be to use copulas as in [22] in the case of mixed data.

Software A R package is under construction, and a Matlab code is available on request.

APPENDIX A. The BOS properties: statements and proofs

From Equations (7), (8) and (10) in Section 2.2, it is easily seen that the BOS model has a polynomial expression related to π of degree less than or equal to $m-1$:

$$\mathrm{p}(x;\mu,\pi) = \sum_{j=0}^{m-1} a_j(m,\mu,x)\pi^j. \tag{18}$$

From this statement, we have developed a *formal* MATLAB code computing the *exact* expression of each coefficient $a_j(m,\mu,x)$ for any x, μ and m values. For instance, for $m=5$, $\mu=2$ and $x=4$, we obtain the following exact expression:

$$\begin{aligned}\mathrm{p}(4;2,\pi) &= a_0(5,2,4) + a_1(5,2,4)\pi + a_2(5,2,4)\pi^2 + a_3(5,2,4)\pi^3 + a_4(5,2,4)\pi^4 \\ &= \frac{1}{5} - \frac{33}{200}\pi - \frac{457}{7200}\pi^2 + \frac{2}{75}\pi^3 + \frac{13}{7200}\pi^4.\end{aligned}$$

This formal calculus tool is used in the proofs of all the following properties. Its advantage is to provide very straightforward proofs. Its drawback is to provide proofs for only some selected values of m. However, we have chosen $m \in \{1,\ldots,8\}$ since it answers to most practical situations[4]. For instance in the reference book [1] all ordinal data have less than 8 categories.

Proposition A.1 *($\pi = 0$: Uniformity.) If $\pi = 0$, then $\forall(\mu,x) \in \{1,\ldots,m\}^2$, $p(x;\mu,\pi) = m^{-1}$.*

Proof *The formal calculus leads to $a_0(m,\mu,x) = m^{-1}$ for all $m \in \{1,\ldots,8\}$. Conclusion follows by setting $\pi = 0$ in (18).*

Proposition A.2 *($\pi = 1$: Dirac in μ.) If $\pi = 1$, then $\forall(x,\mu) \in \{1,\ldots,m\}^2$, $p(x;\mu,\pi) = \mathbb{I}(x=\mu)$.*

[4]We conjecture that all propositions hold for any m value. In particular, we have proved the simplest of them (Propositions (A.1), (A.2), (A.4)) in this general situation (these general proofs are not given in this paper).

Proof *The formal calculus leads to* $\sum_{j=0}^{m-1} a_j(m,\mu,\mu) = 1$ *for all* $m \in \{1,\ldots,8\}$. *Conclusion follows by setting* $\pi = 1$ *in (18).*

Proposition A.3 *(*μ: mode if $\pi > 0$.*) If* $\pi > 0$, *then* $\forall(\mu,x) \in \{1,\ldots,m\}^2$ *such that* $\mu \neq x$, $p(\mu;\mu,\pi) > p(x;\mu,\pi)$.

Proof *For* $\pi = 1$, *Proposition A.2 already established that* μ *is the mode. We consider now* $0 < \pi < 1$. *From (18),* $p(\mu;\mu,\pi) - p(x;\mu,\pi)$ *is a polynomial of degree less than or equal to* $m-1$. *Moreover, from Proposition A.1,* $\pi = 0$ *is a root of* $p(\mu;\mu,\pi) - p(x;\mu,\pi)$, *thus we can factorize it by* π

$$p(\mu;\mu,\pi) - p(x;\mu,\pi) = \pi \left(\sum_{j=0}^{m-2} b_j(m,\mu,x)\pi^j \right). \tag{19}$$

Since $0 < \pi < 1$ *implies that* $0 < \pi^j < 1$ *for any* $j > 1$, *the following lower bound is directly obtained*

$$p(\mu;\mu,\pi) - p(x;\mu,\pi) > \pi \left(b_0(m,\mu,x) + \sum_{j=1}^{m-2} b_j^-(m,\mu,x) \right) \tag{20}$$

where $b_j^-(m,\mu,x) = \min(0, b_j(m,\mu,x))$. *All* $b_j(m,\mu,x)$ *are obtained from a formal polynomial Euclidian division for all* $m \in \{1,\ldots,8\}$, *allowing to show by formal calculus that* $b_0(m,\mu,x) + \sum_{j=1}^{m-2} b_j^-(m,\mu,x) > 0$ *for all* $m \in \{1,\ldots,8\}$. *Conclusion follows.*

Proposition A.4 *(*μ: absolute growing of its probability with π.*)* $\forall \mu \in \{1,\ldots,m\}$, $p(\mu;\mu,\pi)$ *is an increasing function of* π.

Proof *The formal calculus leads to* $a_j(m,\mu,\mu) \geq 0$ *for all* $m \in \{1,\ldots,8\}$ *and for all* $j \in \{1,\ldots,m-1\}$. *Conclusion follows directly from (18).*

Proposition A.5 *(*μ: relative growing of its probability with π.*)* $\forall(\mu,x) \in \{1,\ldots,m\}^2$ *such that* $\mu \neq x$, $p(\mu;\mu,\pi) - p(x;\mu,\pi)$ *is an increasing function of* π.

Proof *The key is to prove the differential (in* π*) expression* $p'(\mu;\mu,\pi) - p'(x;\mu,\pi) > 0$. *From (18),* $p'(\mu;\mu,\pi) - p'(x;\mu,\pi)$ *is a polynomial of degree less than or equal to* $m-2$

$$p'(\mu;\mu,\pi) - p'(x;\mu,\pi) = \sum_{j=0}^{m-2} b_j(m,\mu,x)\pi^j. \tag{21}$$

Since $0 < \pi < 1$ *implies that* $0 < \pi^j < 1$ *for any* $j > 1$, *the following lower bound is directly obtained*

$$p'(\mu;\mu,\pi) - p'(x;\mu,\pi) > b_0(m,\mu,x) + \sum_{j=1}^{m-2} b_j^-(m,\mu,x) \tag{22}$$

where $b_j^-(m,\mu,x) = \min(0, b_j(m,\mu,x))$. *All* $b_j(m,\mu,x)$ *are obtained from a formal derivation of a polynomial for all* $m \in \{1,\dots,8\}$, *allowing to show by formal calculus again that* $b_0(m,\mu,x) + \sum_{j=1}^{m-2} b_j^-(m,\mu,x) > 0$ *for all* $m \in \{1,\dots,8\}$. *Conclusion follows.*

Proposition A.6 *(*Decreasing around μ if $0 < \pi < 1$*.)* $\forall (x,x') \in \{1,\dots,m\}^2$ *such that* $x' < x < \mu$ *or* $\mu > x > x'$, $p(x;\mu,\pi) > p(x';\mu,\pi)$.

Proof *From (18),* $p(x;\mu,\pi) - p(x';\mu,\pi)$ *is a polynomial of degree less than or equal to* $m-1$. *Moreover, from Propositions A.1 and A.2 respectively,* $\pi = 0$ *and* $\pi = 1$ *are roots of* $p(x;\mu,\pi) - p(x';\mu,\pi)$, *thus we can factorize it by* $\pi(1-\pi)$

$$p(x;\mu,\pi) - p(x';\mu,\pi) = \pi(1-\pi)\left(\sum_{j=0}^{m-3} b_j(m,\mu,x,x')\pi^j\right). \tag{23}$$

Since $0 < \pi < 1$ *implies that* $0 < \pi^j < 1$ *for any* $j > 1$, *the following lower bound is directly obtained*

$$p(x;\mu,\pi) - p(x';\mu,\pi) > \pi(1-\pi)\left(b_0(m,\mu,x,x') + \sum_{j=1}^{m-3} b_j^-(m,\mu,x,x')\right) \tag{24}$$

where $b_j^-(m,\mu,x,x') = \min(0, b_j(m,\mu,x,x'))$. *All* $b_j(m,\mu,x,x')$ *are obtained from a formal polynomial Euclidian division for all* $m \in \{1,\dots,8\}$, *allowing to show by formal calculus again that* $b_0(m,\mu,x,x') + \sum_{j=1}^{m-2} b_j^-(m,\mu,x,x') > 0$ *for all* $m \in \{1,\dots,8\}$. *Conclusion follows.*

Proposition A.7 *(*Identifiability*)* *If* $0 < \pi \leq 1$, *identifiability holds.*

Proof *The identifiability problem could concern* μ *and/or* π.

- *First, there exists no couple* $(\mu,\mu') \in \{1,\dots,m\}^2$ *with* $\mu \neq \mu'$ *such that* $p(x;\mu,\pi) = p(x;\mu',\pi')$ *for any* $x \in \{1,\dots,m\}$ *and any* $(\pi,\pi') \in (0,1]^2$. *Indeed, otherwise both* μ *and* μ' *should be modes of both distributions, which is impossible by unicity of the mode (Propositions A.3 and A.6).*

- *Second, there exists no couple* $(\pi,\pi') \in (0,1]^2$ *with* $\pi \neq \pi'$ *such that* $p(x;\mu,\pi) = p(x;\mu,\pi')$ *for any* $x \in \{1,\dots,m\}$ *and any* $\mu \in \{1,\dots,m\}$ *since* $p(x;\mu,\pi)$ *is a polynomial of order at least one. This last statement comes from the fact that* $p(x;\mu,\pi)$ *varies from* $1/m$ *to 0 or 1 when* π *varies (respectively when* $x \neq \mu$ *and* $x = \mu$*) from Propositions A.1 and A.2, thus depends on the value of* π.

APPENDIX B. AERES data

University	PT	EP	SS	EFS
Bordeaux 1	A	A	A	B
Bordeaux 2	A+	A	A+	A
Bordeaux 3	B	A	B	B
Bordeaux 4	B	A	A+	A
Pau	C	B	B	C
Toulouse 1	B	B	B	B
Toulouse 2	B	B	A	B
Toulouse 3	A	A	A+	A
Champollion	A	B	B	B
Lyon 1	A	A+	A	A
Lyon 2	B	A	B	B
Lyon 3	B	A+	B	B
St Etienne	A	B	A	B
Montpellier 1	B	A	A	B
Montpellier 2	A	A	A	B
Montpellier 3	B	B	A	B
Nîmes	C	B	C	C
Perpignan	B	B	B	B
Grenoble 1	B	B	A+	A
Grenoble 2	A	A	B	B
Grenoble 3	C	B	B	C
Savoie	A	A	A	B

References

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

[2] C. Biernacki and J. Jacques. A generative model for rank data based on sorting algorithm. *Computational Statistics and Data Analysis*, 58:162–176, 2013.

[3] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *The Journal of the Pattern Recognition Society*, 28:781–793, 1995.

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

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977. With discussion.

[6] B.S. Everitt. A finite mixture model for the clustering of mixed-mode data. *Statistics & Probability Letters*, 6(305–309), 1988.

[7] C. Fraley and A.E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Society*, 97:611—612, 2002.

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

[9] L. A. Goodman and W. H. Kruskal. Measures of association for cross classifications. *Journal of the American Statistical Association*, 49:732–764, 1954.

[10] L.A Goodman. Explanatory latent structure models using both identifiable and unidentifiable models. *Biometrika*, 61:215–231, 1974.

[11] C. Gouget. *Utilisation des modèles de mélange pour la classification automatique de données ordinales*. PhD thesis, Université de Technologie de Compiègne, 2006.

[12] J. Jacques and C. Biernacki. Model-based clustering for multivariate partial ranking data. *Journal of Statistical Planning and Inference*, 149:201–217, 2014.

[13] J. Jacques, Q. Grimonprez, and C. Biernacki. Rankcluster: An r package for clustering multivariate partial rankings. *The R Journal*, in press, 2014.

[14] J. Jacques and C. Preda. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, in press:DOI 10.1007/s11634–013–0158–y, 2014.

[15] F-X. Jollois and M. Nadif. Classification de données ordinales : modèles et algorithmes. In *Proceedings of the 41th conference of the French Statistical Society*, Bordeaux, France, 2011.

[16] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: An introduction to cluster analysis*. Wiley, 1990.

[17] M. G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33:239–251, 1945.

[18] D.E. Knuth. *Sorting and Searching: Volume3. The art of Computer Programming*. Addison-Wesley Professional, second edition, 1998.

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

[20] S.R. Lipsitz, K. Kim, and L. Zhao. Analysis of repeated categorical data using generalized estimating equations. *Statistics in Medecine*, 13:1149–1163, 1994.

[21] Marica Manisera and Paola Zuccolotto. Modeling rating data with nonlinear {CUB} models. *Computational Statistics and Data Analysis*, 78(0):100 – 118, 2014.

[22] M. Marbac, C. Biernacki, and V. Vandewalle. Model-based clustering of Gaussian copulas for mixed data. *ArXiv e-prints*, May 2014.

[23] E. Matechou, I. Liu, S. Pledger, and R. Arnold. Biclustering models for ordinal data. In *NZSA 2011 Conference*, Auckland, New-Zealand, 2011.

[24] P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42:109–142, 1980.

[25] G. McLachlan and D. Peel. *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York, 2000.

[26] D. McParland and C. Gormley. *Algorithms from and for Nature and Life: Studies in Classification, Data Analysis, and Knowledge Organization*, chapter Clustering Ordinal Data via Latent Variable Models, pages 127–135. Springer, Switzerland, 2013.

[27] V. Melnykov and R. Maitra. Finite mixture models and model-based clustering. *Statistics Surveys*, 4:80–116, 2010.

[28] R.B. Nelsen. *An introduction to copulas*. Springer, 1999.

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

[30] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[31] R. H. Somers. A new assymmetric measure of association for ordinal variables. *American Sociological Review*, 27:799–811, 1962.

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

[33] J.H. Wolfe. Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5:329–350, 1970.

[34] R. Xu and D.C. Wunsch. *Clustering*. John Wiley and Sons, 2009.