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

June 4, 2015

Abstract

We design a probability distribution for ordinal data by modeling the process generating data, which is assumed to rely only on order comparisons between categories. Contrariwise, most competitors often either forget the order information or add a nonexistent distance information. The data generating process 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 can be considered as missing values. Using then the classical latent class assumption, the previous univariate ordinal model is straightforwardly extended to model-based clustering for multivariate ordinal data. Parameters of this mixture model are estimated by an AECM 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, AECM algorithm.

1 Introduction

Clustering [39] is an important explanatory tool for practitioners to discover some hidden, but hopefully valuable, structures in data sets. Mixture models [38, 28, 8, 30] 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 [5], rankings [15], functional [17], categorical [11]...

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 services 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 [19] or [22] for instance) or into nominal data (ignoring the order information, see the Latent GOLD software [37]) in order to "recycle" easily related distributions.

We give now an overview of the main models and clustering algorithms used for ordinal data. First of all, it should be mentioned that ordinality is a characteristic of the meaning of measurements [35], and consequently it does not incur any restrictions for the shape of their distributions. Several approaches have been considered along years to define probability distributions for ordinal data: 1. modeling the cumulative probabilities rather than the individual probabilities, 2. constraining the multinomial model in order to take into account the ordinality of data, 3. assuming that ordinal data are the discretization of an underlying continuous latent variable, 4. constructing directly the distribution in order to obtain a set of desired properties for ordinal data. Note that approaches 2. and 4. can be sometimes connected. The first approach is inspired by the regression framework, modeling a function (logit, probit, log-log...) of the cumulative probabilities with a linear function of some covariables. For instance, the *cumulative logit model* states that:

$$\operatorname{logit} \mathrm{p}(X \leq x) = \beta_{0x} + \boldsymbol{\beta}^t \mathbf{t},$$

where X is the ordinal variable, x one of its categories, $\mathbf{t}$ the covariates and β_{0x} are increasing in x. In such models, the fact that $\boldsymbol{\beta}$ is common for all categories is justified by both parsimony and a latent variable interpretation of the model. Let us notice that, in the absence of covariates, the model is equivalent to a multinomial model. For a review of such models refer to [1] (or to [3] in the case of unobserved (latent) covariates). Such kind of models are used in a clustering context in the Latent GOLD software [37]. The second approach consists in introducing constraints on the multinomial model in order to take into account

the ordinality of the data. Earliest works are probably those of Vermunt [36], which define a general class of non-parametric models by imposing linear or log-linear inequality restrictions on probabilities. Following a similar idea, [12, 18] define probability distribution for ordinal data by imposing different models for the decrease of the probabilities on each side of the mode of the distribution. These models are used by their authors in a model-based clustering framework. In the third approach, the probability distribution for ordinal data is obtained by discretization of the probability distribution of an underlying continuous (Gaussian) variable. For instance, [12] builds a clustering model using this assumption and estimating the unknown discretization thresholds via an EM algorithm [7]. In [29], 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. In the fourth approach, the models are *artificially* constructed to obtain a probability distribution having some desired properties such as the presence of a unique mode, a decrease of the probabilities from each side of this mode, the possibility to achieve a uniform distribution... This is the case of the CUB model [6], which is defined as a mixture of two components, a binomial distribution (which allows to weight all the categories) and a uniform distribution:

$$\mathrm{p}(x;\xi,\alpha) = \alpha C_{m-1}^{x-1}\xi^{m-x}(1-\xi)^{x-1} + (1-\alpha)/m,$$

where m is the number of categories, $\xi \in [0,1]$ allows to control the mode position and $\alpha \in [0,1]$ manages a trade-off between a unimodal and a uniform distribution. The extended CUB model [14] considers additionally a third component: a Dirac distribution. The CUB model is also extended in order to get rid of the notion of distance between categories induced by the use of the binomial distribution (non linear CUB model (NLCUB), [24]). However, the CUB or NLCUB distribution can not be used easily in a clustering context since a mixture of CUB distributions is not identifiable. Indeed, in the mixture case, the uniform components of different CUB components are indistinguishable. Let us finally cite some other geometric clustering procedures for ordinal data. For instance, [32] consider hierarchical clustering with ad-hoc distance metric and similarity measures (Kendall's τ [20], Goodman-Kruskal's γ [10] or Somers' d [34]), and [9] define a non parametric clustering algorithm based on thresholding contingency tables.

In this work, an original approach is considered to define a new probability distribution for ordinal data. This approach consists in modeling the hypothetical data generating process. This general principle has already been successfully used in [4, 15] in the case of ranking data.

However, ranking and ordering data being totally different[1], the data generating process is also totally different: It was a *sorting algorithm* for ranking data and it is now a *search algorithm* for ordinal data as we describe now. A search algorithm is based only on comparisons between categories and thus the ordinal nature of data is wholly respected during the process and no link to any nominal or continuous distribution is necessary. The retained search algorithm is the binary search algorithm since it minimizes the averaged number of comparisons between objects [21]. 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 convenient 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 until eight categories, which is the case of most ordinal variables (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 [15, 16] for instance), but is not developed in this paper because of the relative rarity of such data. Finally, the model is extended to the clustering of multivariate ordinal data through a mixture model of the univariate and unimodal proposed distribution with a conditional independence assumption [11]. As a matter of fact, we do not claim that the process generating ordinal data has to follow a search algorithm and thus that any ordinal distribution should have to follow the proposed unimodal related distribution. Our contribution is just to exhibit, in an original but ordinal "compatible" way, a univariate, unimodal, meaningful (position and precision parameters) and parsimonious distribution which could be then used as a basic ingredient for multivariate model-based clustering. From a practical point of view also, this proposal seems to make sense in the real data sets studied at the end of the present paper according to the BIC model criterion [33].

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

[1]Ranking data occur when some judges are asking to sort m objects, quoted by $1, \ldots, m$, according to a preference order. Thus, a ranking data is a permutation of the whole data set $\{1, \ldots, m\}$. Ordinal data occur when some judges are asking to give a note among the set of ordered notes $1 < \ldots < m$. Thus, an ordinal data is a unique element of $\{1, \ldots, m\}$.

works.

2 Univariate model design

In this section, a new probability distribution for homogeneous ordinal data is designed. By homogeneous we mean that no cluster is present in the data. This distribution, quoted as BOS model in the sequel (for Binary Ordinal Search), will be parametrized by two meaningful parameters: a position parameter μ and a precision parameter π. The design of the model relies on the assumption that ordinal data is the result of a stochastic binary search algorithm. This choice will be motivated by two arguments: First, a search algorithm relies only on order comparisons (information present in ordinal data); Second, its binary version provides some optimality properties (see details below). The section is organized as follows: Section 2.1 depicts how ordinal data can be viewed as the result of a search algorithm, whereas in Section 2.2, stochastic events are introduced in this search algorithm in order to define a probability distribution on the whole set of categories. Section 2.3 presents the model properties and Section 2.4 describes maximum likelihood estimation of the model parameters.

2.1 Ordinal data as the outcome of a binary search algorithm

An *ordinal* variable μ, with m categories $\{l_1, \ldots, l_m\}$, is a categorical variable with categories whose order is significant. The total order relation between the categories is quoted in the sequel "$\prec$": $l_1 \prec \ldots \prec l_m$. Moreover, for simplicity, the categories will be numbered $\{1, ..., m\}$ according to their order. Following this notation, an ordinal variable μ is therefore an element of $\{1, \ldots, m\}$.

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 [21] 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 the combination of "$\succ$" 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 number of comparisons point of view [21], 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 the 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 with 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 previous three 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 (*i.e.* the number of elements) 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 $e = \{b^-, \ldots, b^+\}$:

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

  Notice that the arithmetic distance can be used here since the categories are numbered $\{1, ..., m\}$ according to their order (refer to the beginning of Section 2.1).

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)$)$[2] 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

Even though ordinal data are compatible with any distributional shape [35], some properties of the BOS model are very appealing from an interpretation point of view: unimodality, precision parameter governing the prominence of the mode, identifiability of the parameters.

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). Let us notice that the BOS model can consequently not model a distribution with two neighbouring modes.

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 $\mu \neq x$) points of view (see 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).

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

Identifiability The parameters (μ, π) of the distribution are identifiable if $\pi > 0$ (see Proposition A.7).

All propositions and related proofs are given in Appendix A. The shape of the BOS distribution for different values of μ and π is also displayed on Figure 1.

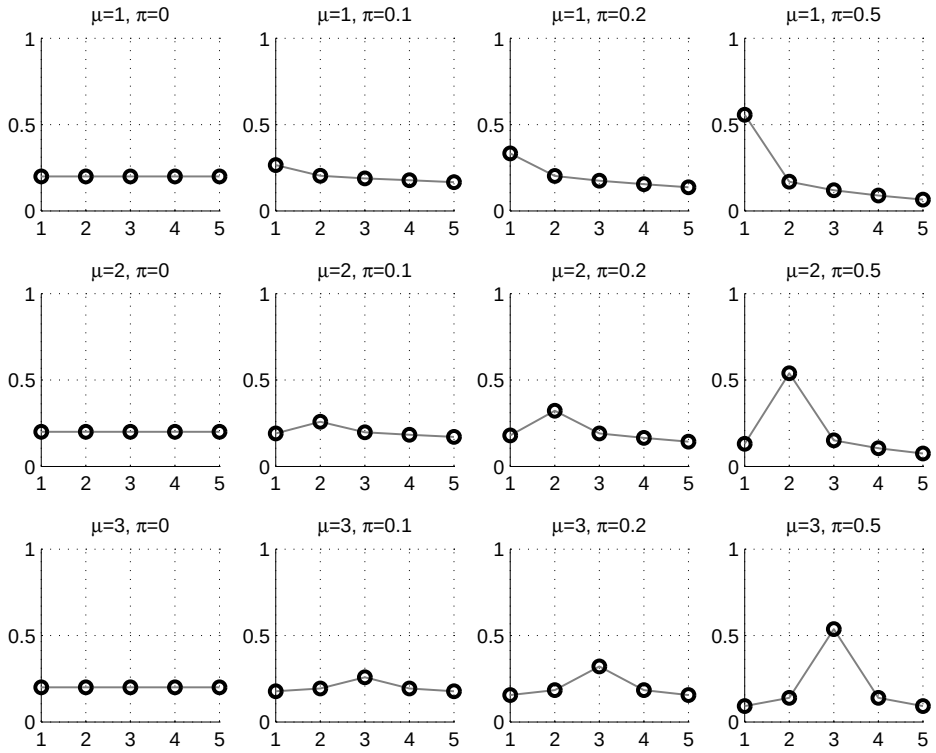

Figure 1: Distribution $\mathrm{p}(x; \mu, \pi)$: shape for $m = 5$ and for different values of μ and π.

2.4 Maximum likelihood estimation by EM algorithm

let us 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 [7] can be used since we have 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 $\mathbf{c} = (c_1, \ldots, c_n)$, $c_i = \{y_{ij}, z_{ij}, e_{ij}\}_{j=1,\ldots,m-1}$ 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

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

  Since each type of missing value evolves in a discrete space, this step requires to compute $\mathrm{p}(c_i, x_i; \mu, \pi)$ for all $c_i \in C_i$.

- **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. Consequently, this algorithm can be used for ordinal data with $m \leq 8$, which is the case for most ordinal variables. For instance, all the data sets studied in the reference book [1] concern ordinal data with fewer 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.

It's fair enough to say that this works for small scales and many scales are indeed small, but any wording that somehow implies that a bigger scale is an oddity should be avoided

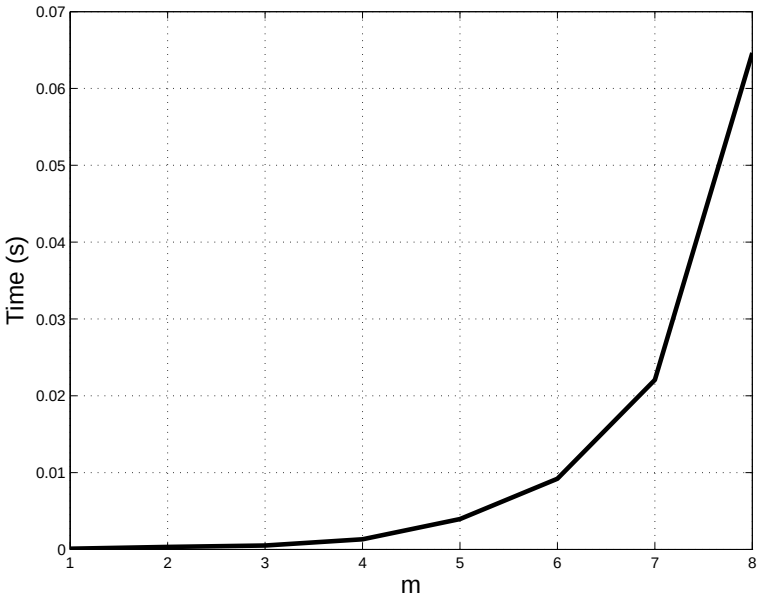

Figure 2: Time to compute the set $\mathcal{P}_i = \{\mathrm{p}(c_i, x_i; \mu, \pi), c_i \in C_i\}$ which is the key set to perform one iteration of the EM algorithm.

3 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.

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 commonly used for its simplicity in the clustering context (k-means [13], nominal data [11], ranking data [15]...), and experiments show that the clustering algorithm we proposed is relatively robust to this assumption (Section 4.1). 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 $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_g)$.

Remark The identifiability of this mixture model is not theoretically studied in this work, and could be the subject to a future work, inspired by [2] which gives conditions for the identifiability of a mixture of multivariate Bernoulli. Nevertheless, it should be noticed that no identifiability problems have been encountered in simulation studies and practical applications. See for instance experiments given at the end of Section 3.2 below.

3.2 Maximum likelihood estimation by AECM algorithm

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

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.

The proposed algorithm is known as Alternating Expectation-Conditional Maximization (AECM, [28, Section 8.7]). It has the same convergence properties than a generalized EM algorithm (GEM), increasing the (observed-data) likelihood at each step, even though the internal EM algorithm used in the M step is iterated only once.

AECM evaluation We provide now a set of simulated data to illustrate the behavior of our AECM estimation procedure. We simulate 20 data sets of size $n \in \{50, 500\}$ from a bivariate and a trivariate ($d \in \{2, 3\}$) mixture of two and three BOS components ($g \in \{2, 3\}$), each dimension having three categories ($m_h = 3$), with equal ($\alpha_k = 1/g$) or different mixing parameters ($\alpha_k = k/(\sum_{k'=1}^{g} k')$), other parameters being fixed to $\boldsymbol{\mu}_k^h = k$ and $\boldsymbol{\pi}_k^h = 0.5$, for $k = 1, \ldots, g$ and $h = 1, \ldots, d$. On each data set and with the true g value, the AECM algorithm is run 10 times with uniform random starting parameters, each run being stopped as soon as the following relative change of the log-likelihood is reached: $|\ell(\boldsymbol{\alpha}^{[q+1]}, \boldsymbol{\mu}^{[q+1]}, \boldsymbol{\pi}^{[q+1]}) - \ell(\boldsymbol{\alpha}^{[q]}\boldsymbol{\mu}^{[q]}, \boldsymbol{\pi}^{[q]})| < 10^{-3}$. Only the run providing the best log-likelihood is finally kept among the 10 ones. The mean of the Kullback-Leibler divergence of the estimated mixture parameter, the mean of the distance between each kind of parameter (mixing proportion $\boldsymbol{\alpha}$, mode $\boldsymbol{\mu}$, precision $\boldsymbol{\pi}$) and its estimate[3], and also all associated standard deviations are displayed in Table 1. The AECM procedure results improve a lot between $n = 50$ and $n = 500$, as would be required for consistency of parameter estimates.

On each data set, the four cluster number candidates $\{1, \ldots, 4\}$ are now considered with the previous tuning values for AECM. The estimated value $\hat{g}$ of g corresponds to this one leading to the largest BIC criterion value [33], BIC being defined by

$$\mathrm{BIC} = \hat{\ell} - 0.5\nu \ln(n), \tag{14}$$

[3]For avoiding the label cluster dependency, we keep the best distance over all label permutations.

	α_k equal								α_k different							
	$d=2$				$d=3$				$d=2$				$d=3$			
	$g=2$		$g=3$		$g=2$		$g=3$		$g=2$		$g=3$		$g=2$		$g=3$	
n	50	500	50	500	50	500	50	500	50	500	50	500	50	500	50	500
KL	.055	.005	.070	.007	.092	.007	.184	.011	.051	.005	.092	.006	.096	.007	.183	.011
	.044	*.003*	*.037*	*.002*	*.056*	*.005*	*.102*	*.004*	*.028*	*.005*	*.049*	*.003*	*.060*	*.004*	*.087*	*.004*
$\Delta\boldsymbol{\alpha}$	.129	.046	.126	.072	.111	.037	.124	.043	.185	.077	.132	.077	.145	.047	.126	.058
	.094	*.039*	*.068*	*.039*	*.084*	*.028*	*.051*	*.028*	*.116*	*.049*	*.089*	*.045*	*.072*	*.041*	*.085*	*.035*
$\Delta\boldsymbol{\mu}$	.012	.000	.091	.000	.025	.000	.083	.000	.037	.000	.116	.000	.025	.000	.250	.000
	.055	*.000*	*.114*	*.000*	*.081*	*.000*	*.129*	*.000*	*.122*	*.000*	*.144*	*.000*	*.061*	*.000*	*.186*	*.000*
$\Delta\boldsymbol{\pi}$	.173	.060	.269	.172	.178	.048	.252	.081	.221	.080	.268	.176	.188	.063	.258	.099
	.067	*.028*	*.058*	*.061*	*.054*	*.018*	*.051*	*.030*	*.092*	*.037*	*.060*	*.056*	*.051*	*.027*	*.043*	*.031*

Table 1: Mean of the Kullback-Leibler divergence (KL), of the mean distance between the true $\boldsymbol{\alpha}$ and its estimated value $\hat{\boldsymbol{\alpha}}$ ($\Delta\boldsymbol{\alpha} = \sum_{k=1}^{g} |\alpha_k - \hat{\alpha}_k|/g$), of the mean distance between the true $\boldsymbol{\mu}$ and its estimated value $\hat{\boldsymbol{\mu}}$ ($\Delta\boldsymbol{\mu} = \sum_{k=1}^{g}\sum_{h=1}^{d} |\mu_k^h - \hat{\mu}_k^h|/(gd)$), of the mean distance between the true $\boldsymbol{\pi}$ and its estimated value $\hat{\boldsymbol{\pi}}$ ($\Delta\boldsymbol{\pi} = \sum_{k=1}^{g}\sum_{h=1}^{d} |\pi_k^h - \hat{\pi}_k^h|/(gd)$), for the multivariate BOS mixture estimate provided by the AECM algorithm. The corresponding *standard deviation* of each is italic.

where $\hat{\ell}$ denotes the maximum log-likelihood value of the model and ν the corresponding number of continuous parameters. Table 2 displays the frequency of $\hat{g}$ retained by BIC. The results of the BIC criterion improve a lot between $n = 50$ and $n = 500$, as would be required for consistency of g value estimates.

	α_k equal								α_k different							
	$d=2$				$d=3$				$d=2$				$d=3$			
	$g=2$		$g=3$		$g=2$		$g=3$		$g=2$		$g=3$		$g=2$		$g=3$	
$\hat{g}\backslash n$	50	500	50	500	50	500	50	500	50	500	50	500	50	500	50	500
1	12	.	19	2	4	.	14	.	17	.	17	.	9	.	13	.
2	8	20	1	1	16	20	3	.	3	20	3	18	11	20	7	2
3	.	.	.	17	1	.	3	20	.	.	.	2	.	.	.	18
4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Table 2: Frequency (on 20 data sets) of each estimate $\hat{g}$ provided by the BIC criterion derived from the AECM algorithm.

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 [12, 29].

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 3). 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 [12]. In its Bayesian version [29], 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 $m = 8$ categories, which is the case for many ordinal scales.

model	continuous (ν)	*discrete*	$g = 2, d = 4, m^h = 4$
BOS mixture	$g - 1 + gd$	gd	9 (*+8 discrete*)
multinomial [11]	$g - 1 + g\sum_h (m^h - 1)$	.	25
Gaussian (diagonal) [5]	$g - 1 + 2gd$	.	17
Latent Gaussian [12]	$g - 1 + gd(d + 3)/2 + \sum_h (m^h - 1)$	.	41

Table 3: 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, through simulation studies, the robustness of the mixture of BOS models in a clustering context, and second to illustrate its usefulness on real data sets. In all the experiments reported in this section, the AECM algorithm is used with uniform random initializations for parameters and with a threshold for the relative change of the log-likelihood of 10^{-3}.

4.1 Simulated data

In these simulation studies, the robustness of the proposed clustering algorithm to *the quantization of categories* and to *the conditional dependency assumption* is evaluated.

Robustness to category quantization As it is argued throughout this paper, the BOS model building is well in accordance with the fundamental nature of ordinal data since it does not rely on any somewhat artificial continuous information. The goal of this first numerical study is to illustrate that the BOS mixture model is much more robust than the Gaussian

mixture model when they are both applied on any artificial continuous coding (quantization) of the categories.

To illustrate this important property for ordinal data, let us consider the following univariate distribution of ordinal data x (with $m = 4$ categories), based on a binned univariate Gaussian distribution ($h = 1, \ldots, 4$, $k = 1, 2$):

$$\mathrm{p}(x = h; a_k, b_1, \ldots, b_5) = \mathrm{p}(b_h \leq x^{gaus} < b_{h+1}) \quad \text{with} \quad x^{gaus} \sim \mathcal{N}(a_k, 1). \tag{15}$$

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

$$\exists! H \text{ s.t. } (1)\forall h \neq H, \mathrm{p}_{kH} \geq \mathrm{p}_{kh}, (2)\forall h < h' < H, \mathrm{p}_{kh} \leq \mathrm{p}_{kh'}, (3)\forall H < h < h', \mathrm{p}_{kh} \geq \mathrm{p}_{kh'}. \tag{16}$$

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 though it can be *very different from* a BOS distribution. Let us notice finally that this simulation scheme corresponds to a latent Gaussian model [12, 29].

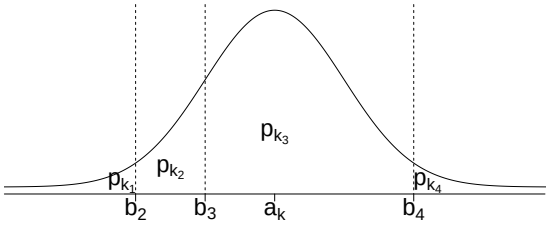

Figure 3: Illustration of the parameters a_k, b_2, b_3, b_4 and the associated probabilities p_{kh} ($1 \leq h \leq 4$).

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 rate $\mathcal{E}$ of the underlying Gaussian mixture. The Bayes error rate, which indicates the separation of the clusters, is defined

	$\mathcal{E} = 0.1$		$\mathcal{E} = 0.2$	
	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 4: Results of the robustness to category coding. 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 classification error rate, $\text{cor}(\cdot, \cdot)$ the empirical correlation, $\text{mean}(\cdot)$ the empirical mean, $\text{sd}(\cdot)$ the standard deviation.

for the present bi-component mixture as:

$$\mathcal{E} = 0.5 \sum_{h \in \Omega_1} \text{p}_{2h} + 0.5 \sum_{h \in \Omega_2} \text{p}_{1h}$$

where $\Omega_1 = \{1 \leq h \leq 4 : \text{p}_{1h} \geq \text{p}_{2h}\}$ (and symmetrically for Ω_2). We plan two different Bayes error rates $\mathcal{E} \in \{0.1, 0.2\}$.

- Conditionally to (16) (it means by throwing away realizations until (16) is fulfilled), the bounds b_h $(h = 1, \ldots, 5)$ are randomly chosen with the following 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[4] are estimated by their respective suitable EM and AECM algorithms. Table 4 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 classification error rates).

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

Robustness to conditional dependency The goal of this second simulation study is to show that the BOS mixture model is relatively robust to (*i.e.* little influenced by) the violation of the class independence assumption, in particular if clusters are well separated, situation of particular interest in clustering. For this, we define a specific mixture model in which the latent class assumption is violated, simulate data according to this model and compare our clustering results with the known partition according to the classification error rate.

For this, let us consider the following cumulative distribution function (cdf) of a *mixture of Gaussian copula*:

$$\mathrm{P}(\mathbf{x}; \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\pi}, \Gamma) = \sum_{k=1}^{g} \alpha_k \Phi_d(\Phi^{-1}(\mathrm{P}(x^1; \mu_k^1, \pi_k^1)), \ldots, \Phi^{-1}(\mathrm{P}(x^d; \mu_k^d, \pi_k^d)); \mathbf{0}, \Gamma), \qquad (17)$$

where $\mathrm{P}(x; \mu, \pi)$ 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 matrix. The Gaussian copula [31] 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 categories 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 rate $\mathcal{E}$. 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 (17) 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. Classification error rate $\hat{\mathcal{E}}$ is displayed in Table 5. 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 pro-

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 5: Results of the robustness to conditional dependency. Mean of the classification error rate $\hat{\mathcal{E}}$ is displayed (its standard deviation in parenthesis).

cedures. The evaluation is generally done on an ordinal scale, and the results are available on the agency website[5]. 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 of chosen studies (employability and further studies, EFS). The evaluations on a (classical) 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 model (which is a parsimonious version of the model used in the Latent GOLD software [37]) and with the Gaussian (diagonal) model, on the basis of the BIC criterion (see Equation (14)). Results in Table 6 show that the best model is a mixture of BOS with four components. Note that the

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

Gaussian model fails to estimate more than two components because of degeneracy problems appearing for larger number of clusters when considering ordinal variables as continuous ones, since many values are obviously exactly repeated.

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

Before analyzing the clustering into four groups with the BOS mixture model, let us graphically compare the clustering obtained by the three models for $g = 2$. Using the projection of data into the first and the third axes[6] 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.

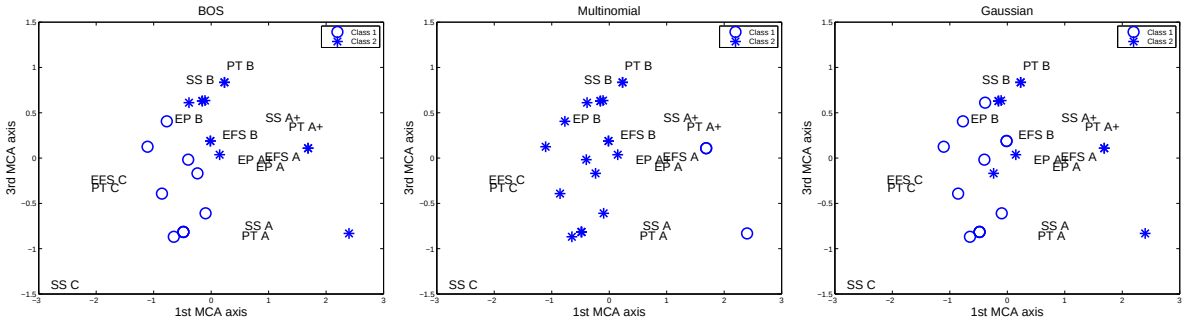

Figure 4: Clustering of the universities of the AERES data set into two clusters with the BOS mixture model and its competitors displayed in the first and the third axes of multiple correspondance analysis (MCA). Values of ordinal variables are also displayed in superposition ("PT B" for "pilot training with evaluation B", *etc.*).

let us now concentrate our analysis on the clustering into four groups with the BOS mixture selected by BIC (Figure 5 and Table 7). 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 the lowest scores at the four criteria.

[6]All partitions are better displayed in the first and the third axes than in the first and second axes.

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, due to the fact that splitting up the data set into clusters makes the clusters more homogeneous than the data set was before.

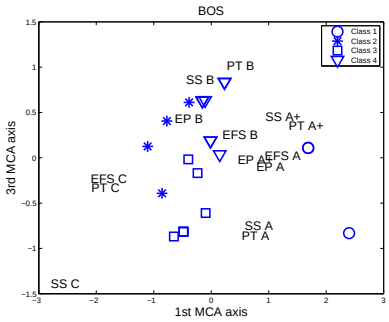

Figure 5: Clustering of the universitie of the AERES data set into four clusters with the BOS mixture model displayed in the first and the third axes of multiple correspondance analysis (MCA). Values of ordinal variables are also displayed in superposition ("PT B" for "pilot training with evaluation B", *etc.*).

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 7: 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 303 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 [23] and [1], has been considered in a clustering context using only the patients assessments after one and five months in [29]. In the present work, all assessments except the first one occurring before to take the treatment/placebo (baseline) are considered: one, three and five months. The first four patients description has been displayed in Table 8 for a comprehensive understanding.

			Arthritis Assessment		
subject	treatment	baseline	1 month	3 months	5 months
1	Auranofin	2	1	1	1
2	Auranofin	3	2	2	2
3	placebo	2	3	1	3
4	placebo	3	1	3	2

Table 8: Four observations from the Arthritis data set.

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. A mixture of multivariate BOS models is compared with the multinomial model (which is a parsimonious version of the model used in the Latent GOLD software [37]) and with the Gaussian (diagonal) model, on the basis of the BIC criterion. Results in Table 9 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.

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

Table 10 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) which increases month after month (π_k^hs globally increase) suggests that the presence or absence of effect of the treatment (or of the placebo) is more and more 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 10: Parameter estimation for the BOS mixture model with four clusters on the Arthritis data set.

5 Concluding remarks

We have designed a parsimonious distribution for ordinal variables, parametrized by two meaningful position and precision parameters. This model is easy to implement through an EM algorithm. Extension to multivariate model-based clustering has also been straightforwardly performed by making the assumption of conditional independence. Its use in real data sets showed its ability to discover clusters unrevealed 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 [27] proposed in [26]. 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 [25] 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[7]. 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 $\forall(x,\mu)\in\{1,\ldots,m\}^2$, $p(x;\mu,\pi)=\mathbb{I}(x=\mu)$.*

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*

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

[7]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).

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* $m \in \{1,\ldots,8\}$*, 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* $m \in \{1,\ldots,8\}$*. Conclusion follows.*

Proposition A.6 *(*Decreasing around μ if $0 < \pi < 1$*.)* $\forall (x,x') \in \{1,\ldots,m\}^2$ *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 $m \in \{1,\ldots,8\}$, 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 $m \in \{1,\ldots,8\}$. 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 $(\mu,\mu') \in \{1,\ldots,m\}^2$ with $\mu \neq \mu'$ such that $p(x;\mu,\pi) = p(x;\mu',\pi')$ for any $x \in \{1,\ldots,m\}$ 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 $x \in \{1,\ldots,m\}$ and any $\mu \in \{1,\ldots,m\}$ 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] E.S. Allman, C. Matias, and J.A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37(6A):3099–3132, 2009.

[3] D. J. Bartholomew, M. Knott, and I. Moustaki. *Latent Variable Models and Factor Analysis: A Unified Approach*. Wiley, 2011.

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

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

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

[7] 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.

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

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

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

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

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

[13] J.A. Hartigan and M.A. Wong. Algorithm as 1326 : A k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1978.

[14] M. Iannario and D. Piccolo. A program in r for cub models inference. Technical report, Universita di Napoli Federico II, Version 2.0, www.dipstat.unina.it, 2009.

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

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

[17] J. Jacques and C. Preda. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3):231–255, 2014.

[18] 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.

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

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

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

[22] 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.

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

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

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

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

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

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

[29] 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.

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

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

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

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

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

[35] S.S. Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946.

[36] J.K. Vermunt. A general class of nonparametric models for ordinal categorical data. *Sociological Methodology*, 29:187–223, 1999.

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

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

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