Model-based clustering for multivariate partial ranking data

Julien Jacques, Christophe Biernacki

Model-based clustering for multivariate partial ranking data

Julien JACQUES, Christophe BIERNACKI

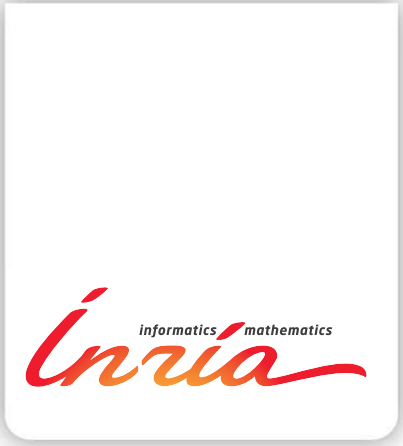
RESEARCH REPORT

N° 8113

October 2012

Project-Teams MΘDAL

ISSN 0249-6399 ISRN INRIA/RR--8113--FR+ENG



Model-based clustering for multivariate partial ranking data

Julien JACQUES*, Christophe BIERNACKI†

Project-Teams MΘDAL

Research Report n° 8113 — October 2012 — 22 pages

Abstract: This paper proposes the first model-based clustering algorithm dedicated to multivariate partial ranking data. This is an extension of the Insertion Sorting Rank (ISR) model for ranking data, which is a meaningful and effective model obtained by modelling the ranking generating process assumed to be a sorting algorithm. The heterogeneity of the rank population is modelled by a mixture of ISR, whereas conditional independence assumption allows the extension to multivariate ranking. Maximum likelihood estimation is performed through a SEM-Gibbs algorithm, and partial rankings are considered as missing data, what allows to simulate them during the estimation process. After having validated the estimation algorithm on simulations, three real datasets are studied: the 1980 American Psychological Association (APA) presidential election votes, the results of French students to a general knowledge test and the votes of the European countries to the Eurovision song contest. For each application, the proposed model shows relevant adequacy and leads to significant interpretation. In particular, regional alliances between European countries are exhibited in the Eurovision contest, which are often suspected but never proved.

Key-words: Multivariate ranking, partial ranking, mixture model, Insertion Sort Rank, SEM algorithm, Gibbs sampling

* julien.jacques@polytech-lille.fr
† christophe.biernacki@math.univ-lille1.fr

RESEARCH CENTRE
LILLE – NORD EUROPE

Parc scientifique de la Haute-Borne
40 avenue Halley - Bât A - Park Plaza
59650 Villeneuve d'Ascq

Classification automatique de données de rang multivariées incomplètes

Résumé : Nous proposons le premier modèle de classification automatique pour données de rang multivariées potentiellement incomplètes. Ce modèle est une extension du modèle *Insertion Sorting Rank* (ISR) pour données de rang, qui est un modèle efficace et signifiant obtenu en modélisant le processus de génération des données. L'hétérogénéité des données est traitée à l'aide d'un modèle de mélange, tandis qu'une hypothèse classique d'indépendance conditionnelle permet de prendre en compte les rangs multivariés. L'estimation des paramètres du modèle est réalisée par maximum de vraisemblance à l'aide d'un algorithme SEM-Gibbs. Les données incomplètes sont considérées comme des données manquantes, ce qui permet de les simuler durant le processus d'estimation. Après avoir validé la stratégie d'estimation sur données simulées, trois jeux de données ont été étudiés : les votes lors de l'élection du président de l'American Psychological Association de 1980, les résultats d'étudiants français lors d'un test de culture générale, et les votes des pays lors du concours de l'Eurovision. Pour chaque application, le modèle proposé a montré une très bonne qualité d'ajustement et à conduit à des interprétations intéressantes. Notamment, pour le concours de l'Eurovision, nous avons mis à jour des alliances géographiques entre pays voisins, ce qui a souvent été suspecté pour ce concours mais jamais prouvé.

Mots-clés : Données de rang multivariées, rangs partiels, modèle de mélange, tri par insertion, algorithme SEM, échantillonneur de Gibbs

1 Introduction

Ranking data occur when a number of subjects are asked to rank a list of objects according to their personal order of preference. These data are of great interest in human activities involving preferences, attitudes or choices like Politics, Economics, Biology, Psychology, Marketing, *etc.* For instance, the voting system *single transferable vote* occurring in Ireland, Australia and New Zeeland, is based on preferential voting [18]. In Economics, it is sometimes more relevant to study the ranking of different economic actors according to some economical indicators rather than the only value of these indicators, since rank analysis focuses on comparisons between actors [27]. Around the mid twentieth century, numerous probabilistic models for rank data have been proposed, based on different assumptions about the origin of a rank datum. For a survey, refer for instance to [24]. Thurstone considers that a rank datum is the result of a ranking of latent continuous variables associated with each object to rank [30]. Paired comparison models [19, 23] liken a rank to the result of a paired comparison process. Modelling parsimoniously each paired comparison leads to the famous *Mallows* Φ *model* [23] and its generalization to distance-based models [11]. Multistage models [12, 26] assume that a rank is the result of an iterative process consisting in choosing at each step the best object among the remaining ones. Among this last class of models, the *Plackett-Luce model* [22, 26] is probably the most studied. More recently, [2] propose the Insertion Sorting Rank (ISR) model as an effective and meaningful alternative for modelling ranking data. The ISR model is set up by modelling the ranking generating process, assumed to be a sorting algorithm in which a stochastic event has been introduced at each comparison between two objects.

All these models consider homogeneous, full and univariate ranking data, what limits their scope. Indeed, in a lot of applications, the study of rank data discloses heterogeneity, due for instance to different political meanings, different economical strategies, different human preferences, *etc.* Heterogeneous rankings have thus attracted a great interest in the last decade: [25] consider mixture of distance-based models and apply it to the modelling of the American Psychological Association's (APA) 1980 presidential election dataset [9], while [4] adapt these models for tied and partial rankings. Mixture of multistage models [1] and Plackett-Luce models have also been successfully applied to the clustering of Irish election data and college admission data by [15, 16, 17, 18]. If mixture of multistage models leads to interesting adequacy power, mixture of distance-based models have more meaningful parameters (and in a lower number), and moreover are simple to implement [25]. On the other side, multivariate ranking data have been rarely studied, despite a strong interest in satisfaction surveys or polls. [3] extends Thurstonian model to the multivariate case, but this extension suffers from numerical integration complexity.

Partial ranking is probably most frequent than full ranking: refer for instance to the 2002 General Election for the Irish House of Parliament dataset, studied in [15], in which 96% of the electors did not rank all the 14 candidates, or the APA's 1980 presidential election which contains more than 60% of partial ranking. [25]'s mixture model is extended to partial ranking by assuming a distribution on the missing entries according to a maximum entropy approach [4]. [21] propose a non-parametric estimator based on kernel smoothing for the estimation of the distribution of partial ranking data, and a visualisation technique based on multi-dimensional scaling in [20].

Our contribution consists in defining a clustering algorithm for multivariate partial ranking data based on an extension of the ISR model [2], initially devoted to univariate full ranking. For this, a mixture model will be considered, with a conditional independence assumption on the multivariate ranking components. The missing entries in the partial ranking will be considered as missing data and inferred in the estimation procedure. Thus, the proposed algorithm will be able to cluster ranking data sets with full and/or partial ranking, univariate or multivariate. To

the best of our knowledge, this is the only clustering algorithm for ranking data with a so wide application scope.

The paper is organised as follows: Section 2 briefly reviews the ISR model and extends this model for heterogeneous multivariate partial ranking data. Maximum likelihood estimation is considered in Section 3 by the mean of a SEM-Gibbs algorithm. Section 4 illustrates the relevance of the mixture of multivariate ISR through simulation study and three real applications, and finally Section 5 concludes the paper.

2 The isr model for heterogeneous multivariate partial ranks

2.1 The univariate isr model

Rank data arise when judges or subjects are asked to rank several objects $\mathcal{O}_1, \ldots, \mathcal{O}_m$ according to a given order of preference. The resulting ranking can be designed by its *ordering* representation $x = (x^1, \ldots, x^m) \in \mathcal{P}_m$ which signifies that Object $\mathcal{O}_{x^h}$ is the hth ($h = 1, \ldots, m$), where $\mathcal{P}_m$ is the set of the permutations of the first m integers. Based on the assumption that a rank datum is the result of a sorting algorithm based on paired comparisons, and that the judge who ranks the objects uses the insertion sort because of its optimality properties, [2] state the following ISR model (see Appendix A.1 for details):

$$\mathrm{p}(x; \mu, \pi) = \frac{1}{m!} \sum_{y \in \mathcal{P}_m} \mathrm{p}(x|y; \mu, \pi) = \frac{1}{m!} \sum_{y \in \mathcal{P}_m} \pi^{G(x,y,\mu)} (1-\pi)^{A(x,y)-G(x,y,\mu)}, \tag{2.1}$$

where $\mu \in \mathcal{P}_m$ is the modal ranking, also named the *reference* or *central* ranking, $\pi \in [\frac{1}{2}, 1]$ is the probability of good paired comparison (in the sort algorithm) according to μ, and the sum over $y \in \mathcal{P}_m$ corresponds to all the possible initial presentation orders of the objects to rank (with identical prior probabilities equal to $1/m!$). The term $G(x, y, \mu)$ is equal to the number of good paired comparisons during the sorting process leading to return x when the presentation order is y, whereas $A(x, y)$ corresponds to the total number of paired comparisons (good or wrong). Their precise definitions are given in Appendix A.2 and proofs can be found in [2].
This model has several interesting properties: the distribution is uniform when $\pi = \frac{1}{2}$; μ is the mode of the distribution whereas its *opposite* ranking $\bar{\mu}$ ($\bar{\mu} = \mu \circ \bar{e}$ with $\bar{e} = (m, \ldots, 1)$ the permutation of total inversion) is the rank of smallest probability; the mode of the distribution is uniformly more pronounced when π grows; identifiability of the parameters occurs once $\pi > \frac{1}{2}$.

2.2 Mixture of multivariate isr

Let now redefine $x = (x^1, \ldots, x^p)$ as a *multivariate* rank, in which $x^j = (x^{j1}, \ldots, x^{jm_j})$, for $1 \leq j \leq p$, is a rank of m_j objects.

The population of multivariate ranks is assumed to be composed of K groups in proportions p_k ($p_k \in [0, 1]$ and $\sum_{k=1}^{K} p_k = 1$). Given a group k, the components $x^1, \ldots, x^p$ of the multivariate rank datum x are assumed to be sampled from independent ISR distributions with reference ranking $\mu_k^1, \ldots, \mu_k^p \in \mathcal{P}_{m_j}$ and good paired comparison probability $\pi_k^1, \ldots, \pi_k^p \in [\frac{1}{2}, 1]$. This conditional independence assumption, classical for categorical data and called latent class model [10, 6], is considered since ranking can be viewed as specific categorical data and also because it is straightforward to implement. In addition, we will see in the study of the Eurovision contest (Section 4.4) that the model seems relatively robust to this assumption.

The unconditional probability of a rank x is then

$$\mathrm{p}(x;\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k \prod_{j=1}^{p} \frac{1}{m_j!} \sum_{y \in \mathcal{P}_{m_j}} \mathrm{p}(x^j|y;\mu_k^j,\pi_k^j), \tag{2.2}$$

where $\boldsymbol{\theta} = (\pi_k^j, \mu_k^j, p_k)_{k=1,\ldots,K\ ,j=1,\ldots,p}$ and $\mathrm{p}(x^j|y;\mu_k^j,\pi_k^j)$ is defined in (2.1).

2.3 Partial ranking

Each component x^j of x can be full or partial. Let $\check{x}^j$ be the rank x^j in which the *unobserved* positions are replaced by 0, and $\hat{x}^j$ be x^j with 0 in the place of the *observed* positions. With these notations we have $x^j = \hat{x}^j + \check{x}^j$. Let $\check{I}^j \subset \{1,\ldots,m_j\}$ be the set of indices corresponding to observed ranking positions in x^j, and $\hat{I}^j \subset \{1,\ldots,m_j\}$ the set of unobserved positions indices ($\check{I}^j \cup \hat{I}^j = \{1,\ldots,m_j\}$). Let also define $\hat{x} = (\hat{x}^1,\ldots,\hat{x}^p)$ and $\check{x} = (\check{x}^1,\ldots,\check{x}^p)$. In order to illustrate these notations, we consider the following example in dimension $p = 1$ with $m_1 = 5$: let assume that the judge does not rank objects in third and forth positions and returns $\check{x} = (2,5,0,0,3)$. The objects $\mathcal{O}_1$ and $\mathcal{O}_4$ have then not been ranked. Possible $\hat{x}$ can be either $(0,0,1,4,0)$ or $(0,0,4,1,0)$, and then x can be either $(2,5,1,4,3)$ or $(2,5,4,1,3)$. Frequently, the objects in the top positions will be ranked and the missing ones will be at the end of the ranking, but our model does not impose such situation and is able to work with partial ranking whatever are the positions of the missing data.

3 Estimation

3.1 Likelihood expression

Let introduce the latent variable z which records the group membership of the observations. The latent variable $z = (z^1,\ldots,z^K)$ is defined such that $z^k = 1$ if the observation belongs to group k and $z^k = 0$ otherwise. Let $\mathbf{x} = \{x_1,\ldots,x_n\}$ be a sample of n multivariate rankings and $\mathbf{z} = \{z_1,\ldots,z_n\}$ the corresponding latent variables. Let $\check{I}_i^j$ and $\hat{I}_i^j$ be respectively the sets of indices of observed and unobserved positions in the jth component x_i^j of the ith observation x_i. Similarly, let $\hat{x}_i^j$ and $\check{x}_i^j$ correspond to the previous notations for the jth component of the ith observation, $\check{x}_i = \{\check{x}_i^1,\ldots,\check{x}_i^p\}$ and $\hat{x}_i = \{\hat{x}_i^1,\ldots,\hat{x}_i^p\}$. Let also define $\check{\mathbf{x}} = \{\check{x}_i; i = 1,\ldots,n\}$ and $\hat{\mathbf{x}} = \{\hat{x}_i; i = 1,\ldots,n\}$. Let $y_i = (y_i^1,\ldots,y_i^p) \in \mathcal{P}_{m_1} \times \ldots \times \mathcal{P}_{m_p}$ be the presentation orders of the objects for the ith observation and $\mathbf{y} = \{y_1,\ldots,y_n\}$.
Assuming that triplets (x_i, y_i, z_i) arise independently ($i = 1,\ldots,n$), the observed-data log-likelihood of model (2.2) is:

$$l(\boldsymbol{\theta};\check{\mathbf{x}}) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} p_k \prod_{j=1}^{p} \frac{1}{m_j!} \sum_{y \in \mathcal{P}_{m_j}} \sum_{x \in \mathcal{X}_i^j} \mathrm{p}(x|y;\mu_k^j,\pi_k^j) \right),$$

where $\mathcal{X}_i^j = \{x \in \mathcal{P}_{m_j} : x^h = \check{x}_i^{jh}, \forall h \in \check{I}_i^j\}$ is the set of all the rankings compatible with the observed part $\check{x}_i^j$ of x_i^j.

The maximization of this likelihood is not straightforward since several missing data occurs: the group membership z_i, the presentation order y_i and the unobserved ranking position $\hat{x}_i$. In such a situation, a convenient way to maximize the likelihood is to consider an EM algorithm

[8], which relies on the following completed-data log-likelihood:

$$l_c(\boldsymbol{\theta};\mathbf{x},\mathbf{y},\mathbf{z}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_i^k \sum_{j=1}^{p} \ln\left(\frac{p_k}{m_j!}\mathrm{p}(x_i^j|y_i^j;\mu_k^j,\pi_k^j)\right).$$

Since this completed log-likelihood is not linear for all three type of missing data, the E step of the EM algorithm is intractable. In this work a SEM-Gibbs approach is proposed to overcome this difficulty.

3.2 SEM-Gibbs algorithm

The fundamental idea of this algorithm is to reduce the computational complexity that is present in both E and M steps of EM by removing all explicit and extensive use of the conditional expectations of any product of missing data. It relies on the SEM algorithm [14, 5] which generates the latent variables y_i, z_i and $\hat{x}_i$ at a so-called stochastic step (S step) from the conditional probabilities computed at the E step. Then these latent variables are directly used in the M step. However, the advantage with SEM-Gibbs algorithm relies on the fact that the latent variables are generated without calculating conditional probabilities at the E step, thanks to a Gibbs algorithm. The proposed SEM-Gibbs algorithm proceeds in the following two steps (SE-Gibbs step and M step), after starting from initial values $\hat{\mathbf{x}}^{\{0\}} = (\hat{x}_i^{j\{0\}})_{1\leq i\leq n, 1\leq j\leq p}$, $\mathbf{y}^{\{0\}} = (y_i^{j\{0\}})_{1\leq i\leq n, 1\leq j\leq p}$, $\mathbf{z}^{\{0\}} = (z_i^{k\{0\}})_{1\leq i\leq n, 1\leq k\leq K}$ and $\boldsymbol{\theta}^{\{0\}}$, and for Q_{SEM} iterations.

The SE-Gibbs step It consists of three sub-steps that are now described at the qth iteration:

- Generate $y_i^{j\{q\}}|\{z_i^{\{q-1\}}, x_i^{j\{q-1\}}, (\mu_k,\pi_k)^{\{q-1\}}\}$, where $x_i^{j\{q-1\}} = \{x_i^{jh\{q-1\}}; 1\leq h\leq m_j\}$ with $x_i^{jh\{q-1\}} = \check{x}_i^{jh}$ if $h\in \check{I}_i^j$ and $x_i^{jh\{q-1\}} = \hat{x}_i^{jh\{q-1\}}$ otherwise. For this, we consider a Gibbs sampler generating a chain $y_i^{j\{q,0\}},\ldots,y_i^{j\{q,R_j\}}$, in which the last value $y_i^{j\{q,R_j\}}$ (for R_j large enough, see Section 3.4) is retained for $y_i^{j\{q\}}$.
  Starting from $y_i^{j\{q,0\}} = y_i^{j\{q-1\}}$, the Gibbs sampler generates $r\in\{1,\ldots,R_j\}$ sequences $y_i^{j\{q,r\}}$ where $(y_i^{jh\{q,r\}},\cdot)$ is generated according to the following distribution:

  $$p\left(y_i^{jh}, y_i^{jh+1}|y_i^{j1\to h-1\{q,r\}}, y_i^{jh+2\to m_j\{q,r-1\}}, x_i^{j\{q-1\}}; (\mu_k,\pi_k)^{\{q-1\}}\right)$$

  for $h\in\{1,\ldots,m_j-2\}$, and $(y_i^{jm_j-1\{q,r\}}, y_i^{jm_j\{q,r\}})$ according to

  $$p\left(y_i^{jm_j-1}, y_i^{jm_j}|y_i^{j1\to m_j-2\{q,r\}}, x_i^{j\{q-1\}}; (\mu_k,\pi_k)^{\{q-1\}}\right),$$

  where k is such that $z_i^{k\{q-1\}} = 1$. The notation $y_i^{ja\to b}$ means $y_i^{ja},\ldots,y_i^{jb}$.

- Generate $z_i^{\{q\}}|\{y_i^{\{q\}}, x_i^{\{q-1\}}; \boldsymbol{\theta}^{\{q-1\}}\} \sim \mathcal{M}(t_i^{1\{q\}},\ldots,t_i^{K\{q\}})$ with, for $k=1,\ldots,K$:

  $$t_i^{k\{q\}} \propto p_k^{\{q-1\}} \prod_{j=1}^{p} \mathrm{p}(x_i^{j\{q-1\}}|y_i^{j\{q\}}; \mu_k^{j\{q-1\}}, \pi_k^{j\{q-1\}})$$

  where $x_i^{\{q-1\}} = \{x_i^{j\{q-1\}}; 1\leq j\leq p\}$.

- Generate $\hat{x}_i^{j\{q\}}|\{z_i^{\{q\}}, y_i^{j\{q\}}, \check{x}_i^j; \boldsymbol{\theta}^{\{q-1\}}\}$ following a similar sequential scheme as for $y_i^{j\{q\}}$. Let $\hat{I}_i^j = (\hat{I}_i^j(h))_{h=1,\ldots,\hat{m}_i^j}$ where $\hat{m}_i^j = |\hat{I}_i^j|$ is the number of objects with unobserved position in x_i^j. The chain $\hat{x}_i^{j\{q,0\}}, \ldots, \hat{x}_i^{j\{q,R_j\}}$ is generated by the Gibbs sampler as follows: starting from an arbitrary $\hat{x}_i^{j\{q,0\}}$, draw R_j sequences $\hat{x}_i^{j\{q,r\}}$ ($r \in \{1, \ldots, R_j\}$) where $(\hat{x}_i^{j\hat{I}_i^j(h)\{q,r\}}, \cdot)$ is generated according to

$$p\left(\hat{x}_i^{j\hat{I}_i^j(h)}, \hat{x}_i^{j\hat{I}_i^j(h+1)} | \hat{x}_i^{j\hat{I}_i^j(1\to h-1)\{q,r\}}, \hat{x}_i^{j\hat{I}_i^j(h+2\to\hat{m}_i^j)\{q,r-1\}}, \check{x}_i^j, y_i^{j\{q\}}; (\mu_k, \pi_k)^{\{q-1\}}\right)$$

for $h \in \{1, \ldots, \hat{m}_i^j - 2\}$ and $(\hat{x}_i^{j\hat{I}_i^j(\hat{m}_i^j-1)\{q,r\}}, \hat{x}_i^{j\hat{I}_i^j(\hat{m}_i^j)\{q,r\}})$ according to

$$p\left(\hat{x}_i^{j\hat{I}_i^j(\hat{m}_i^j-1)}, \hat{x}_i^{j\hat{I}_i^j(\hat{m}_i^j)} | \hat{x}_i^{j\hat{I}_i^j(1\to\hat{m}_i^j-2)\{q,r\}}, \check{x}_i^j, y_i^{j\{q\}}; (\mu_k, \pi_k)^{\{q-1\}}\right).$$

where k is such that $z_i^{k\{q\}} = 1$. The last value $\hat{x}_i^{j\{q,R_j\}}$ of the Gibbs chain is retained for $\hat{x}_i^{j\{q+1\}}$.

The M step The M step consists in computing the parameter value $\boldsymbol{\theta}^{\{q\}}$ which maximizes the completed log-likelihood computed at the previous step:

$$\boldsymbol{\theta}^{\{q\}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\, l_c(\boldsymbol{\theta}; \{\check{\mathbf{x}}, \hat{\mathbf{x}}^{\{q\}}\}, \mathbf{y}^{\{q\}}, \mathbf{z}^{\{q\}}),$$

where $\Theta = [\frac{1}{2}, 1]^p \otimes_{j=1}^p \mathcal{P}_{m_j} \times \Delta^K$ with $\Delta^K = \{p_k : 1 \leq k \leq K; 0 \leq p_k \leq 1, \sum_{k=1}^K p_k = 1\}$ is the unit K-simplex, and where $\hat{\mathbf{x}}^{\{q\}}$, $\mathbf{y}^{\{q\}}$ and $\mathbf{z}^{\{q\}}$ are simulated in the E step. The maximum for the mixing proportions are:

$$p_k^{\{q\}} = \frac{1}{n}\sum_{i=1}^n z_i^{k\{q\}}.$$

Given that the whole exploration of $\mathcal{P}_m$ is intractable to estimate each modal rank $\mu_k^{j\{q+1\}}$, a Gibbs sampling is used. The justification of the use of such an algorithm can be found in a Bayesian setting, in which the maximum *a posteriori* of the distribution $\mathrm{p}(\mu_k^j|\check{\mathbf{x}}, \hat{\mathbf{x}}^{\{q\}}, \mathbf{y}^{\{q\}}, \mathbf{z}^{\{q\}}; \pi_k^j)$ is equivalent to the argument maximizing the completed log-likelihood when a uniform prior on μ_k^j is considered, since:

$$\mathrm{p}(\mu|\mathbf{x}, \mathbf{y}, \mathbf{z}; \pi, p) \propto \exp(l_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{y}, \mathbf{z})$$

where $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{p})$ with $\boldsymbol{\mu} = \{\mu_k^j\}_{k=1,\ldots,K, j=1,\ldots,p}$, $\boldsymbol{\pi} = \{\pi_k^j\}_{k=1,\ldots,K, j=1,\ldots,p}$ and $\boldsymbol{p} = \{p_k\}_{k=1,\ldots,K}$. The Gibbs sampling proceeds as follows. Starting from initial values $\pi_k^{j\{q,0\}}$ and $\mu_k^{j\{q,0\}}$, the probabilities $\pi_k^{j\{q\}}$ and the modal ranks $\mu_k^{j\{q\}}$ are estimated according to a Gibbs sampling alternating on R_j iterations the two following steps:

(i) $\mu_k^{j\{q,r+1\}}|\{\check{\mathbf{x}}, \hat{\mathbf{x}}^{\{q\}}, \mathbf{y}^{\{q\}}, \mathbf{z}^{\{q\}}; \pi_k^{j\{q,r\}}\}$ is simulated following a similar scheme as for $y_i^{j\{q\}}$: $(\mu_k^{jh\{q,r+1\}}, \cdot)$ is simulated according to

$$p\left(\mu_k^{jh}, \mu_k^{jh+1} | (\mu_k^{j1\to h-1\{q,r+1\}}, \mu_k^{jh+2\to m_j\{q,r\}}, x_i^{j\{q\}}, y_i^{j\{q\}}; \pi_k^{j\{q,r\}}\right)$$

for $h \in \{1, \ldots, m_j - 2\}$ and $(\mu_k^{jm_j-1\{q,r+1\}}, \mu_k^{jm_j\{q,r+1\}})$ according to

$$p\left(\mu_k^{jm_j-1}, \mu_k^{jm_j} | \mu_k^{j1\to m_j-2\{q,r+1\}}, x_i^{j\{q\}}, y_i^{j\{q\}}; \pi_k^{j\{q,r\}}\right),$$

(ii) $\pi_k^{j\{q,r+1\}} = \frac{\sum_{i=1}^n z_i^k G(\check{x}_i^j, \hat{x}_i^{j\{q\}}, y_i^{j\{q\}}, \mu_k^{j\{q,r+1\}})}{\sum_{i=1}^n z_i^k A(\check{x}_i^j, \hat{x}_i^{j\{q\}}, y_i^{j\{q\}})}$.

The retained value for $\pi_k^{j\{q+1\}}$ and $\mu_k^{j\{q+1\}}$ is the couple in the sequence of size R_j maximizing the completed log-likelihood.

Choice of $\boldsymbol{\theta}^{\{q+1\}}$ The SEM-Gibbs algorithm is stopped after a given number Q_{SEM} of iterations. After removing a burn-in period of B_{SEM} iterations, the estimation $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is obtained as follows: for each distinct value of $\{\mu_k^j; 1 \leq j \leq p, 1 \leq k \leq K\}$ in the sequence $\mu_k^{j\{q\}}$ ($B_{SEM} \leq q \leq Q_{SEM}$), take the mean $\bar{\pi}_k^j$ of the $\pi_k^{j\{q\}}$ and $\bar{p}_k$ of the $p_k^{\{q\}}$, and retain the parameters $(\bar{p}_k, \mu_k^j, \bar{\pi}_k^j)_{1\leq j\leq p, 1\leq k\leq K}$ leading to the highest log-likelihood. Since the log-likelihood is computationally intractable, it has to be approximated (Section 3.3). In such a situation, label switching can traditionally occur and could need to use some procedures as described in [7] and [29]. However, in our context, we can expect for no label switching when the clusters are well separated, while the case of non-well separated clusters will be avoided because not retained by the model selection criterion defined in Section 3.5.

3.3 Likelihood approximation

Since the observed-data log-likelihood is not tractable, it is approximated by:

$$l(\boldsymbol{\theta}; \check{\mathbf{x}}) \approx -\sum_{i=1}^{n} \ln \left(\frac{1}{Q_l} \sum_{q=B_l}^{Q_l} \frac{1}{p(x_i|(\hat{x}_i, y_i, z_i)^{\{q\}}; \hat{\boldsymbol{\theta}})} \right) \tag{3.1}$$

where $(\hat{x}_i, y_i, z_i)^{\{q\}}$ arise independently from $p(\hat{x}_i, y_i, z_i|\check{x}_i; \hat{\boldsymbol{\theta}})$. The simulation of these triplets is carried out sequentially as in the SE-Gibbs step with Q_L iterations.

3.4 Choice of R_j, Q_{SEM}, B_{SEM}, Q_l, B_l

The size R_j of the Gibbs sampling, used in both SE-Gibbs and M steps, will be chosen greater than $\frac{m_j(m_j-1)}{2}$ which is the maximum Kendall distance between two ranks of size m_j, so that any rank of $\mathcal{P}_{m_j}$ can be reached with non-null probability from any arbitrary initialisation $y_i^{j\{q,0\}}$ for instance.
The number Q_{SEM} of iterations of the SEM-Gibbs algorithm and the size B_{SEM} of the burn-in period, as the sizes Q_l and B_l for the likelihood approximation, will be tuned empirically thanks to simulation study. We will see in Section 4.1 than these numbers have not to be very high to produce good estimations.

3.5 Model selection

In order to select the number K of components in the mixture, the BIC criterion [28] is used:

$$\text{BIC} = -2l(\hat{\boldsymbol{\theta}}; \check{\mathbf{x}}) + (Kp + K - 1)\log(n),$$

where $l(\hat{\boldsymbol{\theta}}; \check{\mathbf{x}})$ is the maximum log-likelihood (in practice it will be approximated by (3.1)) and $Kp + K - 1$ the number of continuous parameters (proportions p_k and probabilities of good paired comparisons π_k^j).

3.6 Estimation of the missing ranking positions

For each partial ranking, *i.e.* when $\breve{x}_i$ $(1 \leq i \leq n)$ is not full, the corresponding full ranking x_i can be estimated using the Gibbs chain simulated in Section 3.3. The estimation of x_i can be unconditional, using the mode of the empirical distribution generated by the chain $(\hat{x}_i^{\{q\}})_{q \in \{B_L, \ldots, Q_l\}}$, or conditional to the cluster k when considering only the $\hat{x}_i^{\{q\}}$'s of this chain such that $z_i^{k\{q\}} = 1$.

4 Numerical experiments

In this section, after having demonstrated the efficiency of the SEM-Gibbs estimation algorithm through a simulation study, the interest of the proposed mixture of multivariate ISR is illustrated on three real datasets: the APA election dataset which contains full and partial rankings; the results of students to a general knowledge test, which consists of multivariate full rankings; the Eurovision dataset containing multivariate partial rankings.
In each application, the SEM-Gibbs algorithm is used with the following iteration numbers: $R_j = 10$ (number of iterations of the Gibbs sampling, $1 \leq j \leq p$), $Q_{SEM} = 100$ (number of SEM iterations), $B_{SEM} = 10$ (size of the burn-in period in the SEM-Gibbs algorithm) and $Q_L = 100$ (number of iterations used for the likelihood approximation) and $B_l = 10$ (size of the burn-in period for the likelihood approximation). For each estimation, the SEM-Gibbs algorithm is launched 20 times, and the solution corresponding to the best approximated likelihood is retained.

4.1 Evaluation of the SEM-Gibbs algorithm on simulation

The goal of this experiment is to validate the proposed estimation algorithm by a simulation study, in particular in presence of partial rankings. For this, ranking data are simulated according to a mixture of two bivariate ISR distributions (with equal proportions $p_1 = p_2 = 0.5$). The first component of the mixture is parametrized by $\mu_1^1 = (1, 2, 3, 4, 5)$, $\mu_1^2 = (3, 4, 1, 5, 2)$, $\pi_1^1 = 0.8$ and $\pi_1^2 = 0.9$, and the second by $\mu_2^1 = (5, 4, 3, 2, 1)$, $\mu_2^2 = (2, 5, 4, 1, 3)$, $\pi_2^1 = 0.7$ and $\pi_2^2 = 0.95$. Note that the components of the mixture are relatively well separated, and good estimation even for small sample size can be expected. In order to evaluate the proposed estimation strategy, two samples of size $n = 200$ and $n = 4\,000$ are simulated and a mixture of two bivariate ISR distributions is estimated with the SEM-Gibbs strategy described in Section 3. As to evaluate the estimation strategy in presence of partial ranking, some partial rankings are introduced as follows: for each dimension, in $d\%$ of the rankings two values are deleted (for instance a simulated rank $(3, 4, 1, 5, 2)$ is replaced by $(3, 0, 0, 5, 2)$), in $\frac{d}{2}\%$ of the rankings three values are deleted and in $\frac{d}{4}\%$ of it four values are deleted.

The efficiency of the estimation of the modal rankings μ_k^j is illustrated by computing the averaged, minimum and maximum Kendall distance between the true modal rank and the estimated one over 20 simulations. Similarly, the efficiency of the estimation of the probabilities π_k^j of good paired comparison is illustrated by the averaged, minimum and maximum absolute value of the difference between the true probability and its estimation. Results are given in Table 1, for four proportions of partial rankings $(d = 0, 5, 10, 20)$.

The modal ranking estimation is perfect with both sample size until $d = 5$ (8.75% of missing values in the sampled ranks). For higher proportions of partial rankings, the estimations are still very satisfying. Similarly, estimation of the probabilities of good paired comparisons is very satisfying when $d \leq 5$, and relatively correct for $d \geq 10$. Comparing the two sample sizes, we note that the larger sample size leads to better estimations only in the case of full rankings. Indeed,

			$K(\mu, \hat{\mu}_{\text{SEM-gibbs}})$						$\lvert\pi - \hat{\pi}_{\text{SEM-gibbs}}\rvert$					
		n→	200	4 000	200	4 000	200	4 000	200	4 000	200	4 000	200	4 000
d	j	k	mean		best		worst		mean		best		worst	
0	1	1	0	0	0	0	0	0	0.025	0.005	0.023	0.004	0.027	0.006
0	1	2	0	0	0	0	0	0	0.013	0.006	0.011	0.005	0.017	0.007
0	2	1	0	0	0	0	0	0	0.002	0.002	0.001	0.001	0.004	0.002
0	2	2	0	0	0	0	0	0	0.022	0.002	0.021	00.02	0.023	0.003
5	1	1	0	0	0	0	0	0	0.014	0.071	0.011	0.070	0.016	0.072
5	1	2	0	0	0	0	0	0	0.044	0.070	0.038	0.070	0.048	0.071
5	2	1	0	0	0	0	0	0	0.017	0.041	0.015	0.041	0.020	0.042
5	2	2	0	0	0	0	0	0	0.030	0.022	0.028	0.021	0.032	0.023
10	1	1	0	0.025	0	0	0	0.5	0.021	0.101	0.014	0.089	0.039	0.299
10	1	2	0.025	0	0	0	0.1	0	0.088	0.084	0.083	0.057	0.109	0.087
10	2	1	0	0	0	0	0	0	0.028	0.065	0.023	0.046	0.049	0.039
10	2	2	0	0	0	0	0	0	0.051	0.030	0.041	0.013	0.066	0.031
20	1	1	0	0	0	0	0	0	0.110	0.128	0.090	0.067	0.130	0.136
20	1	2	0.41	0.05	0.1	0	0.6	0.5	0.164	0.140	0.144	0.133	0.181	0.197
20	2	1	0	0	0	0	0	0	0.050	0.066	0.034	0.004	0.066	0.074
20	2	2	0	0.045	0	0	0	0.5	0.067	0.099	0.049	0.061	0.085	0.431

Table 1: Results of simulation: distance between the true parameter and the estimated one for different proportion d of partial ranking.

for a given proportion of partial ranking, both sample sizes lead to similar results. This can be explained by the fact that the proportion of partial ranking has been fixed and then growing the sample size leads to grows equivalently the number of incomplete rankings. We conclude from these experiments that the SEM-Gibbs algorithm is an efficient algorithm to estimate a mixture of multivariate ISR, even in presence of a moderate proportion of partial rankings, and we can use it in the following real applications.

4.2 The APA election

This dataset is famous in the ranking data literature [9, 25, 4]. It consists of votes for the 1980 American Psychological Association presidential election. Five candidates were present, and the votes consist of ranking this five candidates in order of preference. Among these candidates, candidates noted 1 and 2 are research psychologists, 4 and 5 are clinical psychologists whereas candidate 3 is a community psychologist. A total of 15 449 votes were cast in the election among which 5 738 ranked all five candidates.

In this application, we firstly compare the mixture of Mallows models using both Cayley and Kendall distances, already successfully applied on this dataset in [25], with our mixture of ISR on the 5 738 full rankings. Mixture of ISR is estimated by the SEM-Gibbs algorithm. In order to compare ISR and Mallows mixtures, the BIC criterion is used: Figure 1 (left) shows its values for 1 to 10 groups. According to the BIC criterion, the mixture of ISR distributions provides a better modelling of the APA dataset than the mixture of Mallows Φ model.

Using now the BIC criterion to select the number of groups, we decide to select 4 groups, although BIC can be slightly lower for 10 groups but a plateau appears after 4 groups. In addition, it allows to simplify the interpretation of the parameter estimation (given in Table 2). Examining these estimated parameters, we remark that the two first groups, representing about 45% of the population of voters, ranks first the community psychologist, then the research

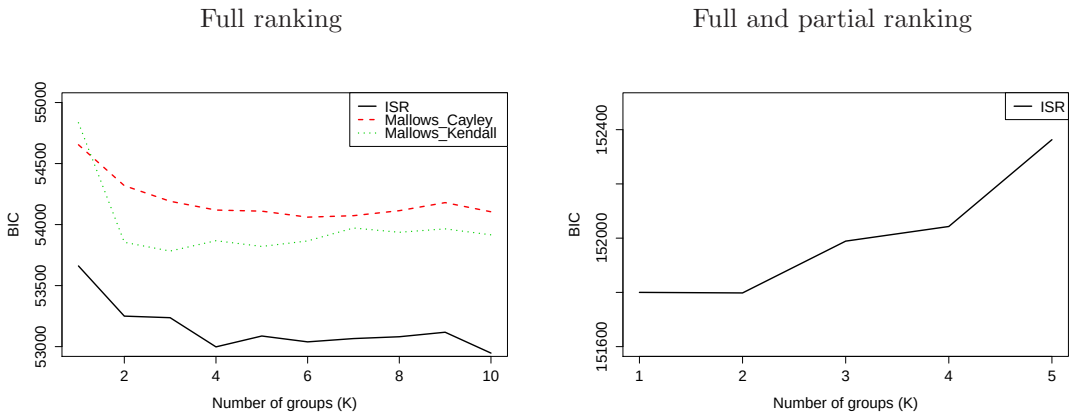


Figure 1: Value of the BIC criterion on the APA dataset. On the left, for 1 to 10 groups, with mixture of Mallows models (using Kendall and Cayley distance) and mixture of ISR. On the right, for 1 to 5 groups, with mixture of ISR taking into account partial ranking. .

psychologists and finally the clinical psychologists. The difference between the first and second groups are the order of the two research psychologists and of the two clinical psychologists. The two others groups, representing about 55% of the voters, do opposite ranking, but with more dispersion since the probability of good paired comparison are slightly lower than for the two first groups. They rank first the clinical psychologists, then the research psychologists and finally the community psychologist. As previously, the difference between these two groups consists in the different rankings within each sub-category of psychologists.

μ_k	π_k	p_k
(3,1,2,5,4)	0.738	0.343
(3,2,1,4,5)	0.712	0.113
(4,5,1,2,3)	0.644	0.399
(5,4,2,1,3)	0.716	0.144

Table 2: Results of estimation of a mixture of ISR with 4 groups on the APA full ranking dataset.

Our SEM-Gibbs algorithm allows moreover to take into account the partial rankings present in these election, which represent about 63% of the total votes. Using these partial rankings, the BIC criterion (Figure 1, right) leads to an hesitation between an homogeneous ISR model, with parameters $\mu = (3, 1, 5, 4, 2)$ and $\pi = 0.527$, and a mixture with two components, in which a majority component ($p_1 = 0.991$) has the same parameters as the homogeneous ISR, and a minority component ($p_1 = 0.009$) is parametrized by $\mu_2 = (5, 1, 3, 2, 4)$ and $\pi_2 = 0.601$.

We conclude this experiment with some words about the results of this election [9]. The elected candidate, using the system of proportional voting system, was candidate 1, which appears second in the modal ranking(s) of the ISR model using partial and full rankings (with 1 or 2 groups). Examining the 4-components mixture estimated on only the full ranking, we see that the elected candidate has always been ranked in intermediary positions, never in the top or the last position of any cluster modal ranking. On the other side, the community candidate (3), which appears on the top position of two groups (and also when using partial rankings) has certainly not been elected since he was also ranked in the last position by a lot of people, as

indicate the modal ranks of the third and fourth groups. Contrary to other voting system, in which we vote only for one candidate and then contribute only to its election, this election system based on proportional voting allows also to contribute to the non-election of other candidates.

The fact that the cluster structure is less evident when taking into account both partial and full ranking than only full ranking can be explained as follows: voters who have strong convictions on which candidate they want to elect and which one they want not, rank all the five candidates with their favourite candidate in the top position and the candidate they do not like in the last position. The resulting cluster structure is then very clear. On the other side, partial rankings are due to voters who rank only their favourite candidate(s) and then, the difference between voters and consequently the cluster structure appears to be less evident. Nevertheless, the elected candidate has always a stable position (second) when considering one or two clusters.

4.3 General knowledge test

In this application, the results of a general knowledge test given to two groups of students are studied. These students were in 2010 in third year (40 students) and fourth year (30 students) of the Statistics and Computer Science department of Polytech'Lille Engineering School (France). This test contains four questions about literature, sports, basic mathematics and cinema. Additionally to the students answers, their gender (girl or boy) and their year of study are registered. The four questions of the test were the following:

- *Literature*. Rank these four French writers according to their date of birth: $\mathcal{O}_1$ = Hugo , $\mathcal{O}_2$ = Molière, $\mathcal{O}_3$ = Camus, $\mathcal{O}_4$ = Rousseau. The correct answer is $\mu^* = (2, 4, 1, 3)$.
- *Sport*. Rank these four national football teams according to increasing number of wins in the football World Cup: $\mathcal{O}_1$ = France, $\mathcal{O}_2$ = Germany, $\mathcal{O}_3$ = Brasil, $\mathcal{O}_4$ = Italy. The correct answer is $\mu^* = (1, 2, 4, 3)$.
- *Mathematics*. Rank these four number in increasing order: $\mathcal{O}_1 = \pi/3$, $\mathcal{O}_2 = \ln 1$, $\mathcal{O}_3 = e^2$, $\mathcal{O}_4 = (1 + \sqrt{5})/2$. The correct answer is $\mu^* = (2, 1, 4, 3)$.
- *Cinema*. Rank chronologically the following Quentin Tarantino movies: $\mathcal{O}_1$ = Inglourious Basterds, $\mathcal{O}_2$ = Pulp Fiction, $\mathcal{O}_3$ = Reservoir Dogs, $\mathcal{O}_4$ = Jackie Brown. The correct answer is $\mu^* = (3, 2, 4, 1)$.

Figure 2 plots the test results for each question on polytopes. The size of the points is proportional to the number of observations.

A mixture of multivariate ISR distributions is estimated on these data, with a number of groups from 1 to 4. Using the BIC criterion (Figure 3), three groups are selected. The corresponding parameters are given by Table 3.

		Literature		Sport		Mathematics		Cinema	
k	p_k	μ_k^1	π_k^1	μ_k^2	π_k^2	μ_k^3	π_k^3	μ_k^4	π_k^4
1	0.4	(2,4,1,3)	0.839	(1,2,4,3)	1	(2,1,4,3)	0.932	(3,4,2,1)	0.765
2	0.271	(2,4,1,3)	0.849	(1,4,2,3)	1	(2,1,4,3)	0.952	(4,3,2,1)	0.795
3	0.329	(2,4,1,3)	0.710	(3,2,4,1)	0.657	(2,1,4,3)	0.896	(2,3,4,1)	0.648

Table 3: Parameters of a mixture of multivariate ISR distribution with 3 groups on the general knowledge test data.

Examining the estimated parameters leads to say that the Literature and Mathematics questions are easy for all students, since the modal rankings of the three groups correspond to the right

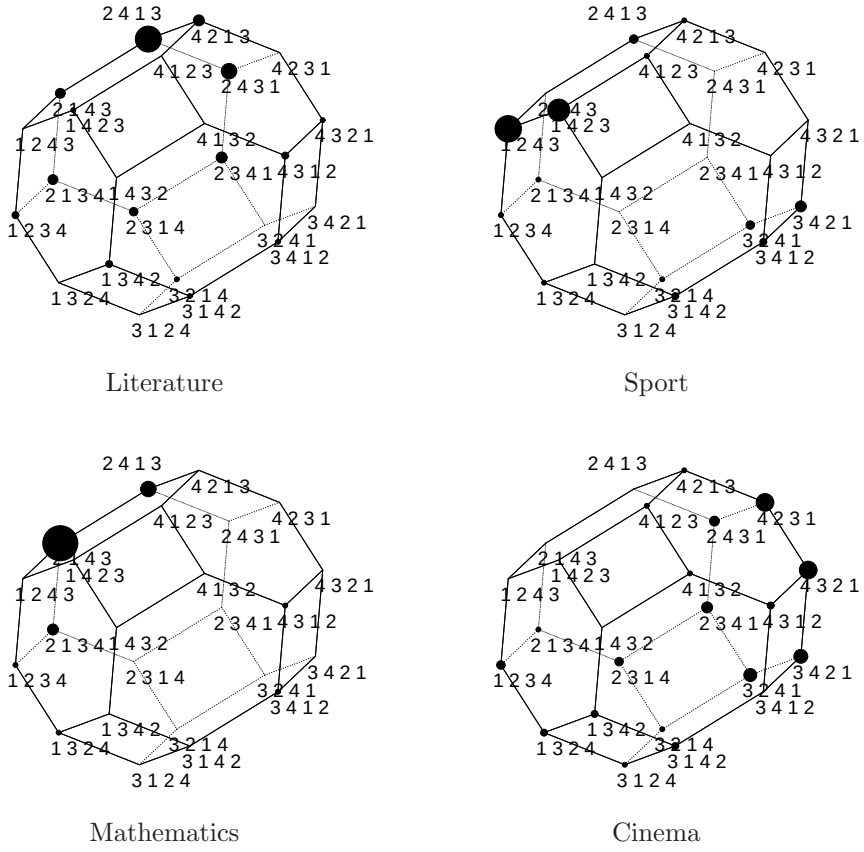


Figure 2: Empirical distribution of the 70 student's answers to the general knowledge test.

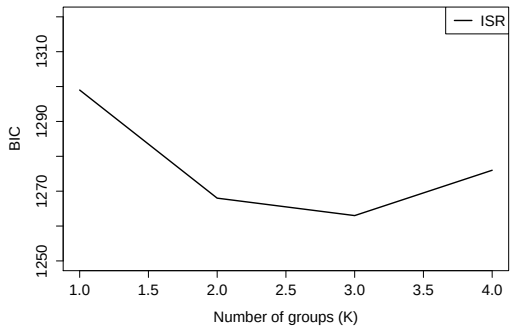


Figure 3: Value of the BIC criterion with mixture of ISR for the general knowledge test.

answers. The sport question is maybe more difficult since only the first group, representing 40% of the students, has a modal rank corresponding to the true ranking. The Cinema question is the more difficult since none group has the right answer as modal rank.

The first cluster, composed of the majority of students (40%), represents students knowing often the right answers (for 3 questions over 4) and relatively self-confident. The second cluster, of smaller size (27%), contains students with lower knowledge in Sport, but equally self-confident. Finally, the last cluster, of intermediary size (33%), corresponds to students rather low in almost all areas (except Mathematics) and with little confidence in them. We note here that the confidence is quite correlated with the knowledge of the true answer. Students are somewhere aware of their potential weakness.

Table 4 gives the confusion matrices between the estimated partition and the girl/boy and 3rd/4th year repartitions. This table exhibits that girls are more present in the third group than in the other groups. Since girls are generally known for being less interested in football than boys, this can explain why this third group have a modal ranking for the sport question so far from the true ranking.

	sexe		
class	girl	boys	
1	5.7	35.7	41.4
2	8.6	18.6	27.2
3	21.4	10	31.4
	35.7	64.3	100

	year of study		
class	3rd	4th	
1	28.6	12.9	41.5
2	17.1	10	27.1
3	11.4	20	31.4
	57.1	42.9	100

Table 4: Repartition (in %) of girls/boys and 3rd/4th years student in the three class estimated by the mixture of ISR.

4.4 The Eurovision Song Contest

The Eurovision Song Contest is an annual competition held among active member countries of the European Broadcasting Union. Each member country submits a song to be performed on live television and then casts votes for the other countries' songs to determine the most popular song in the competition. The vote consists in ranking ten preferred song in order of preference. We consider in these experiments the votes of the $n = 34$ countries who participate to the competitions from 2007 to 2012. During this six years, only 8 countries have participated to the six finals of the competition: 1: France, 2: Germany, 3: Greece, 4: Romania, 5: Russia, 6: Spain, 7: Ukraine and 8: United Kingdom. The studied dataset is then composed of multivariate ranking ($p = 6$ corresponding to the six contests between 2007 and 2012), each rank being of size $m = 8$ (only the votes for the 8 countries which participated to each final are considered) and all ranking being partial. The absence of full ranking signifies that none country participating to the votes has ranked all of the 8 considered countries in its 10 preferences. This dataset is challenging since the number of observations ($n = 34$) is small compared to the size of the rank ($m = 8$) and the presence of partial rankings (precisely, 57.7% of the rankings elements are missing).

A mixture of multivariate ISR distributions is estimated on this dataset, with a number of groups from 1 to 6. The BIC criterion (Figure 4) leads to select 5 groups, whose parameters are given by Table 5. Interesting analysis of these data can be deduced from these parameters, especially using the modal rankings. Indeed, in 2007, all the groups seem to have globally voted in the same way. Indeed, among the 8 considered countries the best were Ukraine (7), Russia (5) and Greece (3) who finished respectively second, third and seventh. These three countries

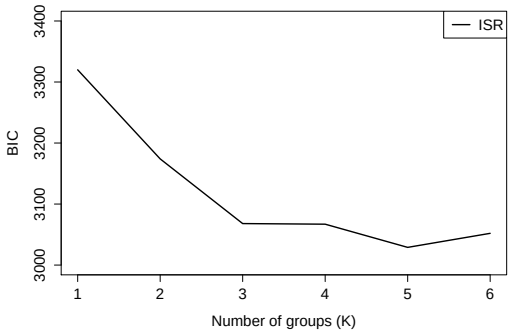


Figure 4: Value of the BIC criterion with mixture of ISR for the Eurovision datatest.

k	p_k
1	0.353
2	0.088
3	0.235
4	0.176
5	0.147

	2007		2008		2009	
k	μ_k^1	π_k^1	μ_k^2	π_k^2	μ_k^3	π_k^3
1	(3,7,5,2,4,6,8,1)	0.831	(3,5,7,6,2,4,8,1)	0.874	(3,1,8,2,4,7,6,5)	0.845
2	(5,7,3,2,1,8,4,6)	0.915	(5,1,7,3,2,4,6,8)	0.889	(1,5,3,2,6,7,4,8)	0.886
3	(5,7,3,4,6,2,8,1)	0.888	(7,5,3,6,4,8,1,2)	0.886	(5,7,8,1,4,3,2,6)	0.747
4	(7,5,3,6,4,2,8,1)	0.921	(5,7,1,3,4,6,8,2)	0.852	(8,1,4,2,6,3,5,7)	0.892
5	(7,5,4,6,3,2,8,1)	0.911	(5,1,7,4,3,2,8,6)	0.921	(5,1,8,3,7,6,2,4)	0.949
	2010		2011		2012	
k	μ_k^4	π_k^4	μ_k^5	π_k^5	μ_k^6	π_k^6
1	(3,7,2,1,6,4,5,8)	0.838	(3,6,7,1,2,4,8,5)	0.763	(6,5,2,4,3,8,7,1)	0.863
2	(2,5,4,3,7,1,8,6)	0.875	(2,8,5,3,6,7,4,1)	0.967	(2,5,8,6,7,1,4,3)	0.881
3	(4,3,2,1,5,7,6,8)	0.855	(7,8,1,2,5,4,3,6)	0.789	(5,4,7,2,6,8,3,1)	0.825
4	(2,4,1,8,5,7,6,3)	0.972	(2,8,4,1,7,6,3,5)	0.889	(5,2,4,7,3,1,6,8)	0.909
5	(2,7,5,6,4,1,3,8)	0.869	(5,7,3,8,2,4,6,1)	0.803	(5,7,3,1,4,8,2,6)	0.703

Table 5: Parameters of a mixture of multivariate ISR distribution with 5 groups on the Eurovision dataset.

appear in the first three positions of the modal rank of each group, except for the group 5, in which Romania if preferred to Greece. Another remark concerns the votes of the group 5. In each six years, Ukraine and Russia (7 and 5) have been well sorted by this group of countries, since they always appear in a good position in the modal rankings of this group. This suggests the existence of geographical voting alliances often suspected in this contest. In order to verify this assumption, we plot on Figure 5 the estimated countries clustering in 5 groups (group 1: red, group 2: blue, group 3: yellow, group 4: green and group 5: gray). This clustering has been obtained by simulating the latent variable z. For this, the SE-Gibbs step has been repeated 1 000 times, using the estimated parameters given in Table 5.

This map confirms the existence of geographical alliances. Indeed, group 1 (red) contains essentially West European countries, group 2 (blue, which is the smallest group) contains some Northern countries, group 3 (yellow) contains Mediterranean countries, group 4 (green) is maybe more dispersed and finally group 5 (gray) contains essentially East European countries. Note that this last group is the one we previously detect to be used to rank Ukraine and Russia in the first positions. This clustering in 5 groups means that countries of a same cluster tend to have similar votes. The regional proximity of the country within each cluster confirms the assumption

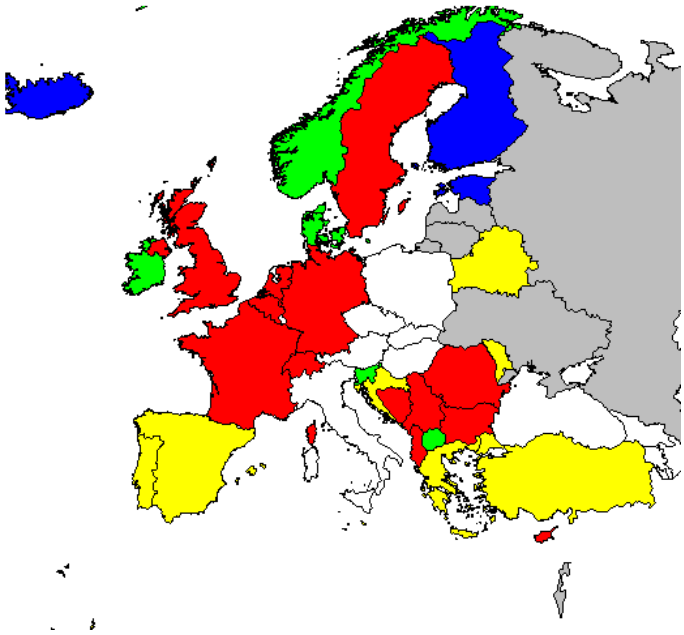


Figure 5: Classification of European countries according to their vote to the Eurovision competition.

of geographical alliances.

A last remark on this dataset concerns the independence assumption we made in order to work with multidimensional rankings. The geographical alliances we just exhibited tend to suggest that this assumption is not satisfied. Nevertheless, the interesting obtained results from an interpretation point of view allow to think that the proposed model is relatively robust to this assumption.

5 Conclusion

Mixture of ISR models provides an efficient tool to model heterogeneous population of ranks. The estimation strategy, based on a SEM-Gibbs algorithm, is moreover an elegant way to take into account partial ranking data, which are very frequent in real applications. Using this model on the challenging Eurovision dataset has allowed to exhibit and confirm regional alliances between countries participating this contest, which has been often suspected and criticized. In [13], a study of the Eurovision contest votes from 1975 to 2000, based on Monte Carlo simulations, has also concluded to regional alliances. A more precise examination of the structure of the geographical blocs of alliances allowed them to exhibit that centrally placed countries within these blocs have a higher probability of being future winners.

In this work, the conditional independence assumption has been used to take into account multivariate ranking. This is a first approach that has the advantage of simplicity. The introduction of correlation between rankings is a great challenge which has to be raised in order to define even more relevant models for multivariate ranking.

Computer code

An **R** package for clustering of multivariate partial ranking data using our methodology is available on request from the authors and will be soon available on the CRAN website[1].

Acknowledgement

We thank Quentin Grimonprez for developing a C++ implementation of the SEM-Gibbs algorithm, and Pr. Brendan Murphy for providing the idea to work on the Eurovision contest.

[1] http://cran.r-project.org/

A Definitions and notations for isr

A.1 Definitions and notations

The sorting algorithm modelled by ISR operates as follows. First, the current object in the presentation order y is placed before (on the left of) the already sorted objects, and is compared to the first object after. If the relative position of both objects in this pair is correct (according to μ), this pair order is unchanged with probability π and this process is restarted with the next object in y. Otherwise, the pair order is reversed (with probability π) and a new pair comparison is performed with the next object after (if it exists). And so forth, until obtaining the final ranking x. Table 6 illustrates this algorithm on an example.

step j	unsorted	sorted
start	$y = \boxed{1}\boxed{3}\boxed{2}$	-
1	$\boxed{3}\boxed{2}$	$x^{(1)} = \boxed{1}$
2	$\boxed{2}$	$\boxed{3} \overset{?}{\leftrightarrow} \boxed{1}$ $x^{(2)} = \boxed{3}\boxed{1}$
3	-	$\boxed{2} \overset{?}{\leftrightarrow} \boxed{3}\boxed{1}$ $\boxed{3}\boxed{2} \overset{?}{\leftrightarrow} \boxed{1}$ $x = \boxed{3}\boxed{1}\boxed{2}$

Table 6: An example to illustrate the stochastic insertion sort algorithm considered by the ISR model, with $\mu = (1, 2, 3)$, $y = (1, 3, 2)$ and $x = (3, 1, 2)$. The notation $x^{(j)}$ means the ranking of the j first objects in y in the order imposed by x

We now define the notations used in Section 2.1.

A.2 Notations

Let $x^{-1} = (x_1^{-1}, \dots, x_m^{-1}) \in \mathcal{P}_m$ be the ranking representation of a rank data, which contains the ranks given to the objects and means that $\mathcal{O}_i$ is in the x_i^{-1}th position.

- $A(x, y) = \sum_{j=1}^{m} A_j^-(x, y) + A_j^+(x, y)$ is the total number of *all* paired comparisons for the whole sorting process, where
  - $A_j^-(x, y)$, the number of *all* comparisons of the current object with the objects already ranked (according to x) on its left (if they exist), is the cardinal of $\mathcal{A}_j^-(x, y) = \{i : x_{y_i}^{-1} < x_{y_j}^{-1}, 1 \leq i < j\}$, the set of the indices of the presentation order y for which the already sorted objects $\mathcal{O}_{y_1}, \dots, \mathcal{O}_{y_{j-1}}$ are ranked in x before the current object $\mathcal{O}_{y_j}$, and consequently *on its left*.
  - $A_j^+(x, y)$, which indicates if the current object $\mathcal{O}_{y_j}$ is compared, at the j step of the sorting, with the object ranked in x just *on its right*, is the cardinal of $\mathcal{A}_j^+(x, y) = \{i : i = \arg\min_{1 \leq i' < j}\{i' : x_{y_{i'}}^{-1} > x_{y_j}^{-1}\}\}$ containing the index of the rank y designating the object sorted in x just after (so *on the right* of) $\mathcal{O}_{y_j}$ among the already sorted objects $\mathcal{O}_{y_1}, \dots, \mathcal{O}_{y_{j-1}}$, if it exists.

- $G(x,y,\mu) = \sum_{j=1}^{m} \left(\sum_{i \in \mathcal{A}_j^-(x,y)} \delta_{y_i y_j}(\mu) + \sum_{i \in \mathcal{A}_j^+(x,y)} \delta_{y_j y_i}(\mu) \right)$ is the total number of *good* paired comparisons for the whole sorting process, with $\delta_{ii'}(\mu) = \mathbf{1}\{\mu_i^{-1} < \mu_{i'}^{-1}\}$ is equal to 1 if $\mathcal{O}_i$ is correctly ranked before $\mathcal{O}_{i'}$ (according to μ), 0 otherwise $(i, i' = 1, \ldots, m, i \neq i')$.

References

[1] W. Benter. Computer-based horse race handicapping and wagering systems: A report. In W.T. Ziemba, V.S. Lo, and D.B. Haush, editors, *Efficiency of racetrack betting markets*. London: Academic Press, 1994.

[2] C. Biernacki and J. Jacques. A generative model for rank data based on sorting algorithm. *Comput. Statist. Data Anal.*, pages in press, DOI 10.1016/j.csda.2012.08.008, 2012.

[3] U. Bockenholt. Multivariate thurstonian models. *Psychometrika*, 55(2):391–403, 1990.

[4] L.M. Busse, P. Orbanz, and J.M. Buhmann. "cluster analysis of heterogeneous rank data". In *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.

[5] G. Celeux and J. Diebolt. The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2(1):73–82, 1985.

[6] G. Celeux and G. Govaert. Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8:157–176, 1991.

[7] G. Celeux, M. Hurn, and C. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95:957–970, 2000.

[8] 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.

[9] P. Diaconis. A generalization of spectral analysis with applications to ranked data. *Annals of Statistics*, 17:949–979, 1989.

[10] B. S. Everitt. *An introduction to latent variable models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1984.

[11] M.A. Fligner and J.S. Verducci. Distance based ranking models. *J. Roy. Statist. Soc. Ser. B*, 48(3):359–369, 1986.

[12] M.A. Fligner and J.S. Verducci. Multistage ranking models. *J. Amer. Statist. Assoc.*, 83(403):892–901, 1988.

[13] Derek Gatherer. Comparison of eurovision song contest simulation with actual results reveals shifting patterns of collusive voting alliances. *Journal of Artificial Societies and Social Simulation*, 9(2):1, 2006.

[14] A. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Matching Intelligence*, 6:721–741, 1984.

[15] I.C. Gormley and T.B. Murphy. Analysis of Irish third-level college applications data. *J. Roy. Statist. Soc. Ser. A*, 169(2):361–379, 2006.

[16] I.C. Gormley and T.B. Murphy. A latent space model for rank data. In *"Proceedings of the 23th International Conference on Machine Learning"*, Pittsburgh, PA, 2006.

[17] I.C. Gormley and T.B. Murphy. Exploring voting blocs within the irish electorate: A mixture modeling approach. *J. Amer. Statist. Assoc.*, 103(483):1014–1027, 2008.

[18] I.C. Gormley and T.B. Murphy. A mixture of experts model for rank data with applications in election studies. *Annals of Applied Statistics*, 2(4):1452–1477, 2008.

[19] M.G. Kendall and B.B. Smith. On the method of paired comparisons. *Biometrika*, 31:324–345, 1940.

[20] P. Kidwell, G. Lebanon, and W.S. Cleveland. Visualizing incomplete and partially ranked data. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1356–1363, 2008.

[21] G. Lebanon and Y. Mao. Non-parametric modeling of partially ranked data. *J. Mach. Learn. Res.*, 9:2401–2429, 2008.

[22] R.D. Luce. *Individual choice behavior: A theoretical analysis*. John Wiley & Sons Inc., New York, 1959.

[23] C.L. Mallows. Non-null ranking models. I. *Biometrika*, 44:114–130, 1957.

[24] J.I. Marden. *Analyzing and modeling rank data*, volume 64 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1995.

[25] T.B. Murphy and D. Martin. Mixtures of distance-based models for ranking data. *Comput. Statist. Data Anal.*, 41(3-4):645–655, 2003.

[26] R.L. Plackett. The analysis of permutations. *J. Roy. Statist. Soc. Ser. C Appl. Statist.*, 24(2):193–202, 1975.

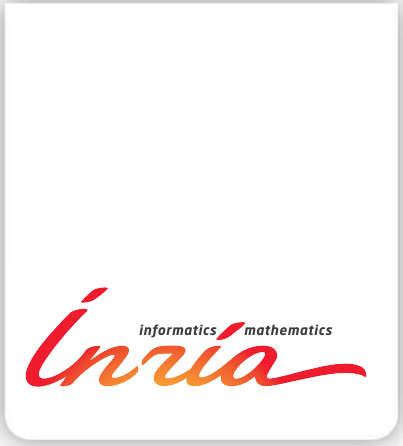
[27] K. Schwab. The global competitiveness report 2012 - 2013. Technical report, World Economic Forum, Geneva Switzerland, 2012.

[28] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[29] M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society. Serie B*, 62(4):795–809, 2000.

[30] L.L. Thurstone. A law of comparative judgment. *Psychological Review*, 79:281–299, 1927.

Contents



RESEARCH CENTRE
LILLE – NORD EUROPE

Parc scientifique de la Haute-Borne
40 avenue Halley - Bât A - Park Plaza
59650 Villeneuve d'Ascq

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399