Journal of Statistical Planning and Inference 149 (2014) 201–217

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Model-based clustering for multivariate partial ranking data

Julien Jacques [a,b,c,*], Christophe Biernacki [a,b,c]

[a] University Lille 1, France
[b] CNRS, France
[c] Inria, France

ARTICLE INFO

Article history:
Received 28 October 2013
Received in revised form
24 February 2014
Accepted 24 February 2014
Available online 4 March 2014

Keywords:
Multivariate ranking
Partial ranking
Mixture model
Insertion sort rank
SEM algorithm
Gibbs sampling

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 has the dual property to be a meaningful model through its location and scale parameters description and to be a kind of "physical" model through its derivation from the ranking generating process assumed to be an insertion sorting algorithm. The heterogeneity of the rank population is modeled by a mixture of ISR, whereas a 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, that allows us to simulate them during the estimation process. After having validated the estimation algorithm as well as the robustness of the model on simulated datasets, three real datasets were 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. The proposed model appears to be relevant in comparison with the most standard competitor ranking models (when available) and leads to significant interpretation for each application. In particular, regional alliances between European countries are exhibited in the Eurovision contest, which are often suspected but never proved.

1. Introduction

Ranking data occur when a number of subjects are asked to rank a list of objects according to a given order (personal preference or more objective criteria as for chronological rankings). These data are of great interest in human activities involving preferences, attitudes or choices in the context of Politics, Economics, Biology, Psychology, Marketing, etc. For instance, the *single transferable vote* voting system used in Ireland, Australia and New Zealand is based on preferential voting (Gormley and Murphy, 2008b). In Economics, it is sometimes more relevant to study the ranking of different economic actors according to some economical indicators rather than just the value of these indicators, since rank analysis focuses on comparisons between actors (Schwab, 2012). Around the mid-twentieth century, numerous probabilistic models for rank data were proposed, based on different assumptions about the origin of a rank datum. For a survey, refer to Marden (1995) for instance. Thurstone (1927) considers that a rank datum is the result of a ranking of latent continuous variables associated

* Corresponding author.

http://dx.doi.org/10.1016/j.jspi.2014.02.011
0378-3758

with each object to rank. Paired comparison models (Kendall and Smith, 1940; Mallows, 1957) assimilate a rank to the result of a paired comparison process. Parsimoniously modeling each paired comparison leads to the famous *Mallows Φ model* (Mallows, 1957) and its generalization to distance-based models (Fligner and Verducci, 1986). Multistage models (Fligner and Verducci, 1988; Plackett, 1975) assume that a rank is the result of an iterative process consisting in choosing the best object among the remaining ones at each step. Among this last class of models, the *Plackett–Luce model* (Luce, 1959; Plackett, 1975) is probably the most studied. More recently, Biernacki and Jacques (2013) propose the Insertion Sorting Rank (ISR) model as an effective and meaningful alternative for modeling ranking data with just a location and a scale parameter. The ISR model is set up by modeling the ranking generation 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, which limits their scope. Indeed, in a lot of applications, the study of rank data discloses heterogeneity, for instance due to different political meanings, different economical strategies, different human preferences, etc. Heterogeneous rankings have thus attracted a great deal of interest in the last decade: Murphy and Martin (2003) consider a mixture of distance-based models and apply it to the modeling of the American Psychological Association's (APA's) 1980 presidential election dataset (Diaconis, 1989), whereas Busse et al. (2007) adapt these models for tied and partial rankings. Very recently, Lee and Yu (2012) considered a weighted version of this mixture model family with applications in political studies. Mixtures of multistage models (Benter, 1994) and mixture of Plackett–Luce models have also been successfully applied to the clustering of Irish election data and college admission data by Gormley and Murphy (2006a,b, 2008a,b). 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 (Murphy and Martin, 2003).

On the other hand, multivariate ranking data have rarely been studied, despite a strong interest in satisfaction surveys or polls. Such ranking data occur when subjects are asked to rank several lists of objects. For instance, the real data application presented in Section 4.4 considers multivariate rankings of dimension four, each dimension corresponding to one of the four questions of a general knowledge test. To the best of our knowledge, the only work on this topic is the one of Bockenholt (1990), which extends the Thurstonian model to the multivariate case, but this extension suffers from numerical integration complexity.

Partial ranking, occurring when a subject does not rank all the objects, is probably more frequent than full ranking: refer for instance to the 2002 General Election for the Irish House of Parliament dataset, studied in Gormley and Murphy (2006a), in which 96% of the electors did not rank all the 14 candidates, or the APA's 1980 presidential election (studied in Section 4.3) which contains more than 60% of partial ranking. Murphy and Martin (2003)'s mixture model is extended to partial ranking by assuming a distribution on the missing entries according to a maximum entropy approach (Busse et al., 2007). Lebanon and Mao (2008) propose a non-parametric estimator based on kernel smoothing for the estimation of the distribution of partial ranking data, and a visualization technique based on multi-dimensional scaling in Kidwell et al. (2008).

As aforementioned, the experimental section of the paper (Section 4) will consider multivariate rankings (general knowledge test, Section 4.4) and partial rankings (APA election, Section 4.3). Additionally, multivariate partial rankings will be analyzed in Section 4.5, thanks to the Eurovision Song Contest dataset. This latter is composed of rankings of the ten preferred songs (among twenty) by a given number of European countries participating to this contest, and this for the last six contests (2007–2012). For each year, the rankings of the eight countries having participated to the six finals from 2007 to 2012 are considered. Thus, since the participating countries did not always rank these eight countries in their ten preferences, observed rankings are generally partial (in fact all observed rankings are partial). Moreover, since the six editions of the contests will be studied simultaneously, the rankings are multivariate (each dimension corresponding to one edition of the contest).

Our contribution consists in defining a clustering algorithm for multivariate partial ranking data based on an extension of the ISR model (Biernacki and Jacques, 2013), initially devoted to univariate full ranking. For this, a mixture model will be considered, with a conditional independence assumption on the multivariate ranking components, while preserving the location and scale parameters interpretation for each component and each variable. 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 datasets 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 such a wide application scope.

The paper is organized as follows: Section 2 briefly reviews the ISR model and extends this model for heterogeneous multivariate partial ranking data. The maximum likelihood estimation is considered in Section 3 by means 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 (preference order or other more objective criteria). The resulting ranking can be designated 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 positive integers.