Rankclust: An R package for clustering multivariate partial rankings

Q. Grimonprez\textsuperscript{a}, J. Jacques\textsuperscript{a,b} and C. Biernacki\textsuperscript{a,b}

\textsuperscript{a}Modal team
Inria Lille-Nord Europe
Villeneuve d’Ascq, France
quentin.grimonprez@inria.fr

\textsuperscript{b}Laboratoire Paul Painlevé
Université Lille 1
Villeneuve d’Ascq
julien.jacques@polytech-lille.fr, christophe.biernacki@math.univ-lille1.fr

\textbf{Keywords} : model-based clustering, multivariate ranking, partial ranking.

1 Introduction

Ranking data occur when a number of subjects are asked to rank a list of objects $O_1, \ldots, O_m$ according to their personal order of preference. The resulting ranking can be designed by its \textit{ordering} representation $x = (x_1, \ldots, x_m) \in \mathcal{P}_m$ which signifies that Object $O_{x_h}$ is the \textit{hth} ($h = 1, \ldots, m$), where $\mathcal{P}_m$ is the set of the permutations of the first $m$ integers. These data are of great interest in human activities involving preferences, attitudes or choices like Politics, Economics, Biology, Psychology, Marketing, \textit{etc}. For instance, the voting system \textit{single transferable vote} occurring in Ireland, Australia and New Zealand, is based on preferential voting.

2 Mixture of multivariate ISR model

Starting from 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, [1] state the following ISR model:

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

where $\mu \in \mathcal{P}_m$ is a \textit{location parameter} and $\pi \in [\frac{1}{2}, 1]$ is a \textit{scale parameter}. The numbers $G(x,y,\mu)$ and $A(x,y)$ are respectively the number of good paired comparisons and the total number of paired comparisons of objects during the sorting process (see [1] for more details). Recently, [2] propose a model-based clustering algorithm for multivariate rankings, i.e. when a datum is composed of several rankings, potentially partial (when some objects have not been ranked). For this, they extend the ISR model by assuming that, given a group $k$, the components of a multivariate ranking are independent:

$$p(x; \theta) = \sum_{k=1}^{K} p_k \prod_{j=1}^{p} p(x_j; \mu_j^k, \pi_j^k),$$  \hspace{1cm} (2)
where the model parameter $\theta = (\pi_k^j, \mu_k^j, \rho_k)_{k=1,...,K, j=1,...,p}$ are estimated by the mean of a SEM-Gibbs algorithm. The resulting algorithm is able to cluster ranking data sets with full and/or partial rankings, univariate or multivariate. To the best of our knowledge, this is the only clustering algorithm for ranking data with a so wide application scope.

3 The Rankclust package

This algorithm has been implemented in C++ and is available through the Rankclust package for R, available on the author webpage\textsuperscript{1} and soon on the CRAN website\textsuperscript{2}.

The main function rankclust() performs cluster analysis for multivariate rankings and is able to take into account partial ranking.

This function has only one mandatory arguments: data, which is a matrix composed of the observed ranks in their ordering representation. The user can specify the number of clusters (1 by default) he wants to estimate or provide a list of clusters numbers. In that case, the user can choose either the BIC or ICL criterion to select the best number of clusters among his list.

The outputs of rankclust() are of different nature:

- the estimation of the model parameters as well as the ’distances’ between the final estimation and the current value at each iteration of the SEM-Gibbs algorithm. These distances can be used as indicators of the estimation variability.

- the estimated partition. Additionally, for each cluster, the probability and the entropy for all the cluster’s members are given. This information helps the user in its interpretation of the clusters.

- for each partial ranking, an estimation of the missing positions.

4 Application

The use of the Rankclust package will be illustrated by the analysis of the European countries votes at the Eurovision song contest from 2007 to 2012.

References


\textsuperscript{1}http://labomath.univ-lille1.fr/~jacques/

\textsuperscript{2}http://cran.r-project.org/