Rankclust: An R package for clustering multivariate partial rankings

Q. Grimonprez[a], **J. Jacques**[a,b] and **C. Biernacki**[a,b]

[a]Modal team
Inria Lille-Nord Europe
Villeneuve d'Ascq, France
quentin.grimonprez@inria.fr

[b]Laboratoire Paul Painlevé
Université Lille 1
Villeneuve d'Ascq
julien.jacques@polytech-lille.fr, christophe.biernacki@math.univ-lille1.fr

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 $\mathcal{O}_1, \ldots, \mathcal{O}_m$ according to their personal 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. 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.

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:

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

where $\mu \in \mathcal{P}_m$ is a *location parameter* and $\pi \in [\frac{1}{2}, 1]$ is a *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:

$$\mathrm{p}(x; \boldsymbol{\theta}) = \sum_{k=1}^{K} p_k \prod_{j=1}^{p} \mathrm{p}(x^j; \mu_k^j, \pi_k^j), \tag{2}$$

where the model parameter $\boldsymbol{\theta} = (\pi_k^j, \mu_k^j, p_k)_{k=1,\ldots,K\ ,j=1,\ldots,p}$ 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[1] and soon on the CRAN website[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

[1] C. Biernacki and J. Jacques. A generative model for rank data based on sorting algorithm. *Comput. Statist. Data Anal.*, 58:162–176, 2013.

[2] J. Jacques and C. Biernacki. Model-based clustering for multivariate partial ranking data. Technical Report 8113, Inria Research Report, 2012.

[1] `http://labomath.univ-lille1.fr/~jacques/`
[2] `http://cran.r-project.org/`