Report on the paper 'Weighted Distance-Based Models for Ranking Data Using the R Package rankdist' proposed for publication in Journal of Statistical Software

Overview

The paper presents the R package rankdist which implements clustering algorithms for ranking data based of several distance-based models. The authors also propose a new probability distribution for ranking data. The architecture of the package is presented and an application on a famous ranking data set is used to illustrate the use of the package. The paper is well written, easy to read and to understand. Nevertheless, several gaps should be addressed before the paper could be accepted for publication.

Main remark

1. In Section 5.3, all the optimizations for the dispersion parameters are done using numerical optimization tools. I think that is because no closed forms are available for the maximization. The author should study the behavior of these optimizations tools for the estimation of λ with a complete simulation study. Similarly, in Section 5.4, an heuristic is proposed for estimating π_0: does-it work ? As for λ a complete simulation study should be carry out in order to show that this heuristic does the job. Moreover, the choice of the initial ranking π_{00} in not explicitly stated: which π_{00} should be used ?

2. The proposed probability distribution is compared only with other distance-based model, whereas, in particular for the APA dataset studied in this paper, Jacques & Biernacki (2014) show that their model overperforms distance based models from a BIC point of view. The comparison of the proposed probability distribution should be done at least with other models for which R package are available, and the authors should at least show that the model they proposed overperforms competitors on some datasets.

3. Can the proposed package consider tied or missing rankings in other positions than in the last ones ? Can multivariate ranking be taken into account ?

4. The authors should explain and illustrate which functionalities their package offers more compared to the Rankcluster package, which propose clustering algorithm for multivariate, partial and tied rankings.

Minor remarks

1. What about the computing time of the proposed algorithm ? In the mixture case, only rankings with 5 objects are considered. Is the approach computationally feasible for a clustering with 15 objects (as the data sets studied in Section 7.2, but in the univariate case). Computing time should be mentioned in the simulation study required by the main remarks.

2. Update Jacques & Biernacki technical paper (2012) into JSPI 2014.

3. Section 4.2, line 9: I don't understand what do you mean by $\frac{1}{(t-q)!}$ *observation*: did you introduce in the dataset all the possible rankings with weight $\frac{1}{(t-q)!}$?

4. For the APA study, can you plot the same figure than Figure 3 for all rankings in order to facilitate what is the impact of taking into account the partial rankings.