Clustering of preference rankings: a non-parametric soft-clustering approach

Abstract  Typically, rank data consist of a set of individuals, or judges, who have ordered a set of items or objects according to their overall preference or some pre-specified criterion. When each judge has expressed his or her preferences according to his own best judgment, such data are characterized by systematic individual differences. In the literature several approaches have been proposed in order to decompose heterogeneous populations into a defined number of homogeneous groups. Often, these approaches work by assuming that the ranking process is governed by some distance-based models. We use the flexible class of methods proposed by Ben-Israel and Iyigun, which consists in a probabilistic-distance clustering approach, and define the disparity between a ranking and the center of a cluster as the Kemeny distance. This class of methods allows for probabilistic allocation of cases to classes, being a form of fuzzy clustering, rather than hard clustering, where the probability is unequivocally related to the chosen distance measure.

Antonio D’Ambrosio
University of Naples Federico II, Italy, e-mail: antdambr@unina.it

Willem J. Heiser
Leiden University, The Netherlands e-mail: heiser@fsw.leidenuniv.nl
1 Introduction

Preference rankings are data expressing individual’s preferences over a set of available alternatives. Statistical methods and models for the analysis of preference rankings can be distinguished in methods based on badness-of-fit functions, which aim to describe the structure of rank data, and methods based on probabilistic models, which aim to model either the ranking process or the population of judges (Marden, 1995). Within this latter category heterogeneity among the judges is assumed, and the goal is generally the identification of homogeneous sub-populations.

When, in addition to rank data, also some information about the judges are known, a variety of models and methods have been introduced, such as generalized linear-like models for rank data (Ditrich, Hatzinger and Katzenbeisser, 1998; Ditrich, Katzenbeisser and Hatzinger, 2000; Böckenholt, 2001; Gormley and Murphy 2008a) and recursive partitioning methods (D’Ambrosio, 2008; Lee and Yu, 2010; Strobl, Wickelmaier and Zeileis, 2011; D’Ambrosio and Heiser, 2016; Plaia and Sciandra, 2017).

Among clustering methods for rank data, mixtures of Bradley-Terry-Luce models and mixtures of distance-based models were proposed (Croon, 1989; Murphy and Martin, 2003; Gormley and Murphy, 2008b, Jacques and Biernacki, 2014).

We focus our attention to clustering of preference rankings. Most of the existing clustering methods for rank data are based on (mixtures of) distance-based models. Other approaches work under (Bayesian) Placket-Luce models (Mollica and Tardella, 2017) or under some recently introduced models for rank data, such as the Insertion Sorting Rank model (Jacques and Biernacki, 2014). In all these cases, either the choice of the distance measure or the assumption of the right model play a key role in obtaining the solution.

We propose a clustering approach for rank data that belong to the probabilistic distance clustering class of methods defined by Ben-Israel and Iyigun (2008), that allows a flexible way to find homogeneous sub-groups without any assumption on the model that in the population generates the preferences. In following this approach, according to which the probability of each judge to belong to a given cluster is unequivocally related to the distance between himself and the cluster center, we use the Kemeny distance (Kemeny and Snell, 1962). This choice is due to the consideration that the Kemeny distance is defined in the space of both full and tied rankings, and that it is the unique distance measure defined on the extended permutation polytope, which is generally accepted as the geometrical space of preference rankings (Heiser and D’Ambrosio, 2013; D’Ambrosio et al., 2017).
References