Computational Statistics and Data Analysis 115 (2017) 122–135

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A novel method for estimating the common signals for consensus across multiple ranked lists[☆]

Vendula Švendová, Michael G. Schimek [*]

Research Unit 'Statistical Bioinformatics', Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Auenbruggerplatz 2/5, 8036 Graz, Austria

ARTICLE INFO

Article history:
Received 10 May 2016
Received in revised form 29 April 2017
Accepted 23 May 2017
Available online 15 June 2017

Keywords:
Bootstrap
Distribution function
Indirect inference
Rank data
Signal estimation
Simulation

ABSTRACT

The ranking of objects, such as journals, institutions or biological entities, is broadly used to assess the relative quality or relevance of such objects. A multiple ranking is performed by a number of assessors (humans or machines) and inference about the nature of the observed rankings is desirable for evaluation, business or scientific purposes. The assessors' decisions are based on some inherent metric scale and depend on judgement and discriminatory ability, data to which we usually do not have access. An indirect inference approach is proposed that allows one to estimate those signal parameters that might be causal for the observed rankings obtained from several assessors, some of which may not necessarily provide the same decision quality. The order of the values represents a consensus ranking across the observed individual rankings. The standard errors of the estimated signal parameters are obtained through a non-parametric bootstrap. Hence, the signal variability can be evaluated object-wise for the purpose of quantifying the stability of the associated rank positions. As a result, such signal estimates can be used in the meta-analysis of conceptually similar evaluation exercises, studies or experiments, and in any data integration task where measurements on the metric scale are either unavailable, or not directly comparable. The suggested approach is validated on simulated rank data as well as on experimental rank data from current molecular medicine. The proposed algorithms were implemented and all calculations performed in the R environment. The source code is provided.

1. Introduction

Since the turn of the century an increasing interest in the statistical analysis of rank data has become evident. This interest is motivated, for instance, by the wish to compare rankings of educational institutions or multinational companies, by the demands of data mining and information retrieval tasks, and by the analytic requirements of high-throughput technologies, among many other applications. Research in recent years has focused on ways of combining related rank data from different sources or studies, often under the notion of *data integration* or *meta-analysis*. Rank-based methods have the advantage of being invariant to transformation and normalisation as long as the relative orderings are preserved. In addition, they are robust to outliers, although some information is inevitably lost compared to metric approaches. For studies that comprise different data types but have some commonality, rank-based methods offer the opportunity to integrate individual results

☆ https://github.com/svendula/MultiRankS.

* Corresponding author.
E-mail addresses: vendula.svendova@medunigraz.at (V. Švendová), michael.schimek@medunigraz.at (M.G. Schimek).

http://dx.doi.org/10.1016/j.csda.2017.05.010

in order to arrive at some consensus that is more conclusive than any of the individual studies (for an overview see Lin, 2010a). Most recently there has also been some interest in the identification of top-ranked objects, which are characterised by high concordance in their rank positions. This is motivated by the fact that long ranked lists are in most instances only informative towards the top. The practical issue with this is where should such a list be truncated in order to distinguish the informative upper section from the rest of the list. This problem has been solved by Hall and Schimek (2012), Schimek et al. (2012), and Sampath and Verducci (2013). Other current work concerns the stability of (top) ranked lists (Jurman et al., 2008; Hall and Miller, 2009, 2010).

A rather new field of application is official statistics. There it has become popular to rank regional units, such as the US Federal States, according to certain features such as the median household income. When these ranks are based on estimates derived from samples then they contain sampling errors as well as other errors, such as misreporting and nonresponse. In a cautionary note about rankings the United States Census Bureau writes "…the ranking of the estimates does not necessarily reflect the correct ranking of the unknown true values" (United States Census Bureau, 2012). Conclusions drawn exclusively from the rank order can be biased, as it is the size differences between the underlying estimates along with the size of their associated errors that provide indispensable additional information. So, for instance, the estimated rank order, with respect to some feature, of two states might be reversed when the true values are close to each other. The problem is that even for a full census – with no sampling errors involved – we do not know the true values.

The strategy in most other fields of data collection, such as in marketing, evaluation research or biotechnology, is to conduct multiple assessments or to replicate observations. The resulting rankings are then integrated in some way with the hope of achieving a single common representation of the individual ranked lists. Common practice is to apply rank aggregation techniques. These techniques have a long history starting with de Borda (1781). Recent methodological contributions, primarily Markov chain-based and cross entropy Monte Carlo-based methods, can be found in Dwork et al. (2001), DeConde et al. (2006), Sculley (2007), Lin and Ding (2009), Lin (2010b), and Kolde et al. (2012).

In this paper we do not suggest one more rank aggregation technique. We aim at the estimation of the unobserved true values and of their standard errors when only the rankings are available or can be observed. This means, for instance, that the concern of the United States Census Bureau can be handled when ranked objects are compared. We can provide additional relevant statistical information obtained solely from the rank data that can underpin the conclusions. However, beyond official statistics there are many more applications, for example in the molecular sciences, where a quantification of the phenomena that produced the observed rankings is useful, or where the combination of measurements on different scales or from different sources is needed.

We have developed an indirect inference method that provides the following results:

1. Estimates of the underlying true signals from multiple ranked lists, under the assumption that the involved objects are informative in the sense of high concordance in their rank positions.
2. Bootstrap standard errors of the signal estimates.
3. A consensus ranking derived from the signal estimates, i.e. without applying any data aggregation technique.
4. Stability assessment of the derived consensus ranking.

This article is organised in the following way: In Sections 2.1 and 2.2 we introduce our statistical model, its assumptions, and the distribution function approach allowing us to perform indirect inference. We then derive an algorithm for a specific objective function, using a Markov chain Monte Carlo optimisation (Sections 2.3 and 2.4). Section 3 is dedicated to the numerical evaluation of the model. Here, the simulation setting, the quality evaluation, and the error estimation are described. The simulation results, including a comparison with commonly used rank aggregation methods, are given in Section 4.2. The results are commented with respect to the four tasks mentioned above. Finally, an application from current molecular medicine is provided in Section 5.

2. The method

2.1. The statistical model

Let us consider a group of n assessors (humans or machines or studies) assessing p objects. We assume that the jth assessor either implicitly or explicitly observed random variables $X_{1j}, \ldots, X_{pj}$, which we shall call the *attributes*. In some but not all situations, these attributes can be observed or measured. In this article attributes are understood as a theoretical construct. The variable X_{ij} denotes the value of the attribute for the ith object as seen by the jth assessor. The order of these variables, say

$$X_{\sigma(1)j} > \cdots > X_{\sigma(p)j}, \tag{1}$$

defines the rankings that the jth assessor gave to the objects. Let $R_{1j}, \ldots, R_{pj}$, where $R_{ij} \in \mathbb{N}, R_{ij} \leq p$ denotes the rank of $X_{1j}, \ldots, X_{pj}$ according to (1). Then we say that $\{R_{ij}\}$ is the column *rank matrix* of the attributes $\{X_{ij}\}$. Ties can be handled by assigning random or average ranks, for example.

Furthermore, we suppose, that the values $X_{1j}, \ldots, X_{pj}$ follow the model

$$X_{ij} = \theta_i + Z_{ij}, \qquad i = 1, \ldots, p; j = 1, \ldots, n, \tag{2}$$