Inference for the Top-k Rank List Problem

Peter Hall\textsuperscript{1} and Michael G. Schimek\textsuperscript{2}

\textsuperscript{1} Department of Mathematics and Statistics, The University of Melbourne
Parkville, VIC, 3010, Australia, halpstat@ms.unimelb.edu.au
\textsuperscript{2} Institute for Medical Informatics, Statistics and Documentation, Medical
University of Graz
Auenbruggerpl. 2, 8036 Graz, Austria, michael.schimek@meduni-graz.at

Abstract. Consider a problem where \(N\) items (objects or individuals) are judged by assessors using their perceptions of a set of performance criteria, or alternatively by technical devices. In particular, two assessors might rank the items between 1 and \(N\) on the basis of relative performance, independently of each other. We aggregate the rank lists in that we assign \textit{one} if the two assessors agree, and \textit{zero} otherwise. How far can we continue into this sequence of 0’s and 1’s before randomness takes over? In this paper we suggest methods and algorithms for addressing this problem.

Keywords: ordered list, moderate deviation bound, nonparametric inference, rank aggregation, random degeneration, top-k rank list

1 Introduction

In various fields of application we are confronted with lists of distinct objects in rank order. The ordering might be due to a measure of strength of evidence or to an assessment based on expert knowledge or a technical device. The ranking might also represent some measurement taken on the objects which might not be comparable across the lists, for instance, because of different assessment technologies or levels of measurement error.

In this paper our interest is to consolidate such lists of common objects, under the assumption of a general decrease of the probability for consensus of rankings with increasing distance from the top rank position. This assumption is reasonable for the rank aggregation applications we have in mind, and is equivalent to the notion of random degeneration of paired rank information (when our input consists of two lists). Applications include the combined analysis of gene expression measurements across experiments and array platforms, data integration of results from internet search engines, or the determination of consensus rankings in customer surveys. The first two examples involve an additional aspect we wish to consider, namely lists of extreme length, say, ten thousand and more (usually resulting from high-dimensional analysis problems). The longer the lists, the more likely we are to observe non-overlapping ranks.

Rank order problems are not new. They have been intensively studied in psychometrics in the nineteen twenties (e.g. Thurstone, 1931), and later on
in biometric experimental design problems (e.g. Bradley and Terry, 1952). A more recent account is given in the book by Kendall and Gibson (1990), however, the classical statistical methodology cannot cope with very long lists.

The last few years have seen an increasing research interest in rank aggregation, focusing on computational approaches that allow calculation of a consolidated list that satisfies a suitable minimum total distance criterion with respect to two or more input lists. The goal of combining information across multiple large data sources, studies or experiments is highly challenging indeed. The naive approach would be combinatorial, however, even for moderately sized data sets it would be NP hard (see e.g. Fagin et al., 2003). Current attempts to overcome this difficulty follow two different strategies. Dwork et al. (2001), and Fagin, Kumar and Sivakumar (2003), developed Markov chain meta-search algorithms for the internet, summarizing majority preferences between pairs of objects across lists. DeConde et al. (2006) applied the Markov process framework to microarray findings obtained across different array platforms. Lin and Ding (2008) derived a cross-entropy Monte Carlo method for the integration of rank lists in genomic studies. What these new algorithms have in common is the fact that they are still computationally highly demanding. Our experience for the simple two-list integration case, with the cross-entropy Monte Carlo method, is that the number of rankings needs to be limited to about two hundred. Because of that, partial instead of full lists are analyzed all the time, and the list length $k$ of a so-called top-$k$ list is chosen arbitrarily. Such an ad-hoc approach is dissatisfying. This motivates us to propose a moderate deviation-based inference concept for identifying the $k$ as the rank position where the consensus information of the two lists, representing the same objects, degenerates into noise. To cope with the above-mentioned demands, we need to provide a mathematical concept, as well as an algorithmic solution, that can cope with very long lists of the order of tens of thousands. In the problems that motivate this work, the total number of objects is much larger than the number of comparable rankings before noise prevails. In particular, assessors or technical devices often agree in the case of many of the first 100 or so objects, but can give extremely noisy rankings to the remaining objects, out of perhaps 10,000.

In Section 2 a sequence that represents the paired rank information of two lists is introduced. Then in Section 3, for the problem outlined in Section 1, a simplified mathematical model is proposed, and a suitable algorithm is derived. In Section 4 we study the numerical properties of the proposed algorithm via simulations, and give recommendations for the choice of technical and tuning parameters. The theoretical properties of our methodology are derived in Section 5, summarized in three theorems 1. Finally we illustrate our top-$k$ list inference procedure on real data from a study in molecular medicine.

---

1 A journal paper also providing the proofs is in preparation