Exploratory non-parametric and graphical procedures for top-k ranked lists

Michael G. Schimek
Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation
Auenbruggerplatz 2/V
8036 Graz, Austria
E-mail: michael.schimek@medunigraz.at

Introduction

Web search engines or microarray laboratory devices, among other new technologies, produce very long lists of distinct items or objects in rank order. The statistical task is to identify common top-ranking objects from two or more lists and to form sublists of consolidated items. In each list, the rank position might be due to a measure of strength of evidence, to a preference, or to an assessment either based on expert knowledge or a technical device. For each object, it is assumed that its rank assignment in one list is independent of its rank assignments in the other lists. The ranking is from 1 to N, without ties. Starting with the work of Mallows (1957), there is a substantial model-based literature on problems in combining rankings where the number of items N is relatively small, and significantly less than the number ℓ of assessment mechanisms. These parametric approaches cannot handle data of the type described above. Moreover, we are interested in problems where the reliability of rankings breaks down after the first (top) k objects due to error or lack of discriminatory information. Hence, we need distribution-free, and at the same time computationally highly efficient, approaches because list aggregation by means of brute force is limited to the situation where both N and ℓ are impractically small.

Here, we present a non-parametric inference procedure that allows us to test for random degeneration of paired rankings (Hall and Schimek, 2010) even under m-dependence of the assignments. The size of a reliable consensual sub-list obtained in this manner depends on various technical parameters to cope with irregular and incomplete rankings, typical for real data. This exploratory inference tool can provide the necessary input for rank aggregation procedures (Schimek, Mysickova, and Budinska, 2010). Here, our focus is on statistical graphics for data integration based on the estimated k's. The inference and the graphical procedures as well as others for rank aggregation were implemented in the `R` package `TopKLists` (for a description see Schimek et al., 2011). Two examples illustrate the capabilities of this new methodology.

Inference for top-k lists

Hall and Schimek (2010) have developed a computationally efficient moderate deviation-based inference procedure for random degeneration in paired rank lists. This non-parametric procedure gives an estimate of the point of degeneration j_0, where $j_0 - 1 = k$ is the length of the top list. It allows for various types of rank irregularities, missing rank assignments, and list lengths in the magnitude of thousands of objects. Overlap of rank positions in two input lists is represented by a sequence of indicators, where $I_j = 1$ if the ranking, given by the second assessor to the object ranked j by the first assessor, is not more than δ index positions distant from j, and otherwise $I_j = 0$. The variables I_j are assumed to follow a Bernoulli random distribution. This implies independence which is motivated by $k \ll N$ and a strong random contribution due to irregular assessments in real data. However, Hall and Schimek (2010) could show that their theoretical results also hold for m-dependence instead of complete independence.

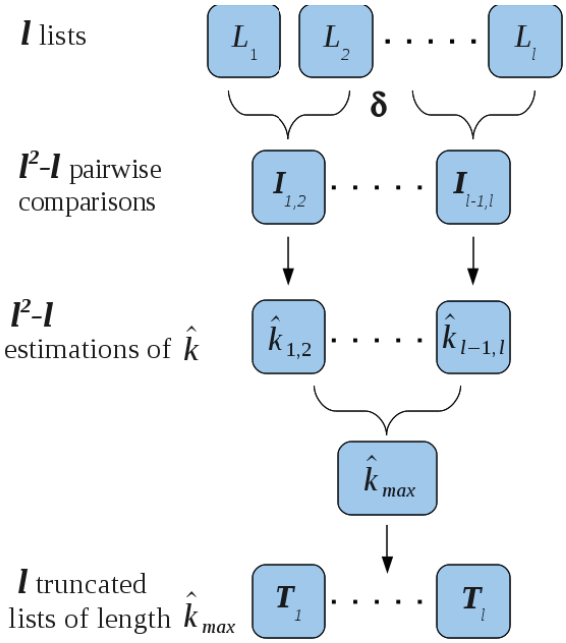

Figure 1: The inference concept to obtain ℓ truncated consensual lists from ℓ full ranked lists.

For the Bernoulli random variables $I_1, \ldots, I_N$, it is assumed that $p_j \geq \frac{1}{2}$ for each $j < j_0$, and $p_j = \frac{1}{2}$ for $j \geq j_0$, and in addition, a "general decrease" of p_j for increasing j that does not have to be monotone. The index j_0 is the rank position where the consensus information of the two lists, representing the same set of objects, degenerates into noise (degradation of information). The estimation of $\hat{j}_0$ is achieved via a *moderate deviation*-based approach. In theoretical analysis of the probability that an estimator, computed from a pilot sample size ν, exceeds a value z, the deviation above z is said to be a moderate deviation if its associated probability is polynomially small as a function of ν, and to be a large deviation if the probability is exponentially small in ν. In regular cases, the values of $z = z_\nu$ that are associated with moderate deviations are

$$z_\nu \equiv (C\,\nu^{-1}\,\log\nu)^{1/2},$$

where $C > \frac{1}{4}$. The null hypothesis H_0 that $p_k = \frac{1}{2}$ for ν consecutive values of k, versus the alternative H_1 that $p_k > \frac{1}{2}$ for at least one of the values of k, is rejected if and only if $\hat{p}_j^{\pm} - \frac{1}{2} > z_\nu$. The quantities $\hat{p}_j^{+}$ and $\hat{p}_j^{-}$ represent estimates of p_j computed from the ν data pairs I_ℓ for which ℓ lies immediately to the right of j, or immediately to the left of j, respectively. Under H_0, the variance of $\hat{p}_j^{\pm}$ equals $(4\nu)^{-1}$ hence, we can evaluate the above inference procedure in practice. However, apart from the pilot sample size ν and the constant C (the latter defaults to 0.251), statistical test results also depend on the distance δ (see next section).

The above described complex decision problem is solved via an iterative algorithm, adjustable for irregularity in the rankings. It is executed for all $(\ell^2 - \ell)/2$ pairs of input lists L_i, thus we obtain ℓ values $\hat{k}_j$ $(j = 1, 2, \ldots, \ell)$. In Figure 1, our strategy is outlined for the calculation of an overall index k^* (a function of the individual k's from the ℓ lists L_i, e.g. their maximum). Having obtained such an overall index, we arrive at truncated lists T_i, either aggregated by graphical means as described in

Table 1: Example of two rankings of $N = 15$ objects. The data streams and the sums of zeros for increasing δ values are displayed.

Obj.	L_1	L_2	$\delta=0$	$\delta=1$	$\delta=2$	$\delta=3$	$\delta=4$	$\delta=5$	$\delta=6$	$\delta=7$	$\delta=8$	$\cdots$	$\delta=14$
o_1	1	1	1	1	1	1	1	1	1	1	1	$\cdots$	1
o_2	2	8	0	0	0	0	0	0	1	1	1	$\cdots$	1
o_3	3	5	0	0	1	1	1	1	1	1	1	$\cdots$	1
o_4	4	3	0	1	1	1	1	1	1	1	1	$\cdots$	1
o_5	5	2	0	0	0	1	1	1	1	1	1	$\cdots$	1
o_6	6	4	0	0	1	1	1	1	1	1	1	$\cdots$	1
o_7	7	6	0	1	1	1	1	1	1	1	1	$\cdots$	1
o_8	8	7	0	1	1	1	1	1	1	1	1	$\cdots$	1
o_9	9	13	0	0	0	0	1	1	1	1	1	$\cdots$	1
o_{10}	10	11	0	1	1	1	1	1	1	1	1	$\cdots$	1
o_{11}	11	9	0	0	1	1	1	1	1	1	1	$\cdots$	1
o_{12}	12	12	1	1	1	1	1	1	1	1	1	$\cdots$	1
o_{13}	13	14	0	1	1	1	1	1	1	1	1	$\cdots$	1
o_{14}	14	10	0	0	0	0	1	1	1	1	1	$\cdots$	1
o_{15}	15	15	1	1	1	1	1	1	1	1	1	$\cdots$	1
#(0)			12	7	4	3	1	1	0	0	0	$\cdots$	0

the following or by stochastic rank aggregation (not considered here; for an overview see Lin, 2010).

Graphical δ selection

The input for the moderate deviation-based inference procedure is a sequence of I's, taking either zero or one, forming a data stream representing the concordance of the paired ranks of an object o. The data stream depends on some distance δ. The parameter δ is defined by the shift in index positions of a particular object o in one list, say L_i, with respect to the other list, say L_j. This means that we assume concordance (i.e. $I = 1$) for an arbitrary object characterized by rank positions in L_i versus L_j, maximal δ index values apart.

For the identification of an appropriate δ in real data analysis, we suggest the following strategy: Compute all data streams for $\delta \in [0, 1, 2, \ldots, N-1]$. Order the data stream vectors column-wise according to increasing δ values. In this way, we obtain a $N \times N$ matrix Δ. The ordered sequence of column sums (i.e. the #(0) for $\delta \in [0, 1, 2, \ldots, N-1]$) is the information we take advantage of in the so-called Δ-plot. It represents the reduction of discordance as a function of δ. When all column sums remain zero, complete concordance is attained. A reasonable choice of the distance parameter is associated with a distinct decline of the #(0)'s. Of course, prior information about the ranking mechanisms involved and the nature of the data is also relevant for the selection of δ. In Table 1, we display a toy example consisting of $N = 15$ objects in two rankings L_1 and L_2 (no missing assessments). As can be easily seen, $\delta = 7$ would be a good choice (a reduction of 5 from the previous count #(0)=12). For an example of a Δ-plot see Figure 2.

Graphical integration of partial lists

Our goal is to identify a subset of objects o_j that is characterized by high rank conformity across the lists. From the truncation procedure of Hall and Schimek (2010), we obtain $(\ell^2 - \ell)/2$ values $\hat{k}_j$ for a pre-specified distance parameter δ. For the integration of all ℓ truncated lists T_i of individual

lengths $\hat{k}_i$, we introduce a heatmap-like graph called *aggregation map*. A consolidated result based on irregular rankings can never be unique, hence we need such a tool. Let us have an index $p = 1, 2, \ldots$. We combine $\ell - 1$ aggregation levels (groupings of partial lists) in one display: For each group of $\ell - p$ truncated lists down to the smallest group consisting of just one pair of lists, we (i) select an arbitrary reference list L^0 under the condition that it comprises $\max_i(\hat{k}_i)$ objects among all pairwise comparisons in the group of rankings, (ii) print the names of its $\max_i(\hat{k}_i)$ objects vertically from the highest to the lowest rank position, and (iii) add the aggregation information for all remaining $\ell - p$ rankings (pairwise list combinations) in the group, ordered according to descending list length.

Table 2: Universum dataset of world's most attractive employers: rankings from the 2010 and 2009 global attraction index (sublists comprising 25 out of 50 published rank positions; NA's denote index values ranked lower than 25). In **boldface** are the two estimated top list lengths $\hat{k}$ of the employers' ranking.

Object	Rank 2010	Rank 2009
Google	1	1
KPMG	2	8
Ernst & Young	3	5
PricewaterhouseCoopers	4	2
Deloitte	5	10
Procter & Gamble	6	6
Microsoft	7	3
The Coca-Cola Company	8	13
J. P. Morgan	9	7
Goldman Sachs	10	4
L'Oréal	11	14
BMW	12	12
Sony	**13**	16
Johnson & Johnson	14	18
The Boston Consulting Group	15	11
McKinsey & Company	16	9
Morgan Stanley	**17**	15
Apple	18	NA
IBM	19	17
Deutsche Bank	20	19
Nestlé	21	24
Bank of America–Merrill Lynch	22	NA
IKEA	23	NA
adidas	24	NA
Accenture	25	23

The aggregation information per group and object consists of two measures represented by colored triangles outlined in an array, (1) **membership** in the top-k list, *yes* is denoted by the color 'grey' and *no* by the color 'white', (2) **distance** of the rank of an individual object $o \in L^0$ from its position in the other list, visualized by means of a color scale from 'red' *identical* to 'yellow' *far distant*. In addition, an integer value gives the numerical distance between the object's rank positions, a negative sign means ranked lower, and a positive sign means ranked higher in L with respect to L^0. For an example see Figure 3.

Two examples

Top ranking of world's most attractive employers: Universum, an employer branding firm, has conducted a survey for the second time to find the internationally most attractive employers based on about 130,000 career seekers (final year students of business and engineering from leading schools in the world). It calculates global talent attraction indices for business and for engineering graduates. These indices are transformed into ranked lists (only the first 50 companies are made publicly available; `www.universumglobal.com/top50`). Here, we focus an the business dataset and wish to identify a top-k list of brand names (objects) for the two consecutive years 2009 and 2010 (from a 2010 perspective, i.e. reference). The first 25 objects and their respective rankings are shown in Table 2.

It is reasonable to apply the Hall and Schimek (2010) algorithm because it can cope with the obvious correlation (m-dependence) of the rankings belonging to consecutive years. Before we can execute the inference procedure, we need to prespecify the distance parameter δ. Figure 2 depicts the Δ-plot. The first index where lack of concordance is starting to degrade takes a value of seven (see the subplot in the top right corner). Hence, $\delta = 7$ is an adequate choice to perform list truncation. The smallest appropriate pilot sample size is $\nu = 4$ (for smaller values the iterative procedure does not converge). We obtain a point of information degradation of $\hat{j}_0 = 14$. This means that the top ranked list consists of $\hat{k} = 13$ consolidated brand names. For a slightly larger $\nu = 10$ to account for additional irregularity in the data, we end up with an estimate of $\hat{j}_0 = 18$ ($\hat{k} = 17$). Instead of *Sony*, the top-k list ends with *Morgan Stanley*. This makes sense because *Apple* is a newcomer in the 2010 ranking and its new high-rank position drastically violates the rank conformity we evaluate here. Moreover, there are other companies, ranked a little bit lower than *Apple*, presenting themselves as movers from a distant position (see the NA's in Table 2).

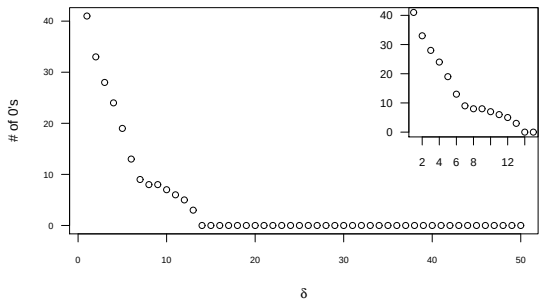

Figure 2: Δ-plot for the Universum data.

Integration of findings from microarray experiments: In Popovici et al. (2009), rankings of gene expression data from several large microarray studies were aggregated. Their goal was to identify a short list of control genes typical for cancer in general. Here we re-analyse three lists representing breast, prostate, and colon cancer, each of length $N = 10,000$. First, we execute the inference procedure for all pairwise combinations of the three input lists as described in Figure 1. As a consequence, lists are truncated at the overall index $k^* = 65$ for $\delta = 10$ and $\nu = 10$. Then, we integrate the obtained truncated lists by graphical means in an aggregation map (maximum of $(k^* + \delta) = 75$ objects displayed). The result is given in Figure 3. It becomes immediately clear that there is considerable overlap between the three truncated lists (indicated by the frequency of lower grey triangles), although stronger between prostate and breast. Gene RPL39 is an outlier with respect to colon cancer (denoted by a yellow upper triangle). The control genes selected in this way could all be biologically verified (for more details of our analysis see Schimek et al., 2011).

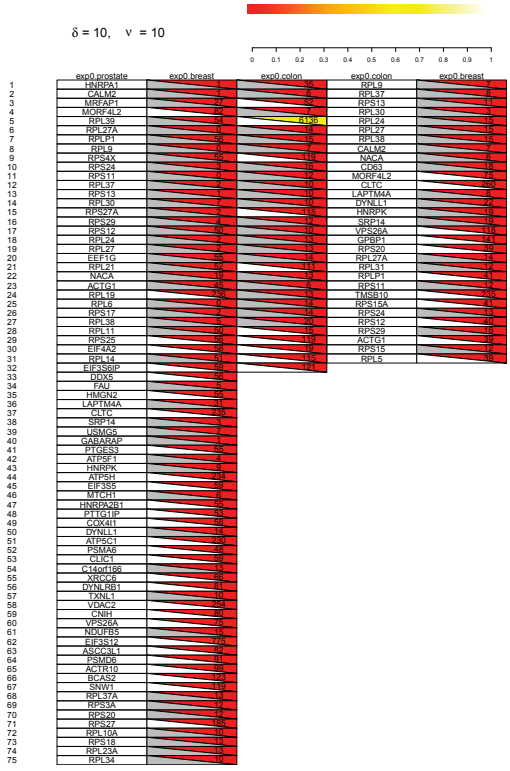

Figure 3: Microarray cancer data example: the aggregation map result of three truncated lists.

Conclusions

The described methodology for the calculation of top-k lists is exploratory, rather than a mechanism for putting a precise numerical value on a concise quantity, hence it cannot be compared to a formal test. Our goal is to offer applied statisticians a unified approach to the integration of ranked lists, and to prevent them from the common misconception that inference on and aggregation of several rankings is limited to a single unique solution. Conformity of two or more lists does not only depend on the assessment techniques but also on the structural relationship between the rankings such as distance and the stochastic nature of irregularities. As a direct consequence, several aggregation results can be obtained for one dataset. In most application fields, the consolidated outcome needs to be verified by human experts. This cannot be done without visualization tools. Such tools have been lacking so far, the aggregation map is a first attempt in this direction. To the benefit of the statistics community all the methods described here were implemented in the `R` package `TopKLists` on CRAN.

REFERENCES

Hall, P. and Schimek, M.G. (2010). Moderate deviation-based inference for random degeneration in paired rank lists. Under revision for *J. Amer. Statist. Assoc.*

Lin, S. (2010). Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**, 555-570.

Mallows, C.L. (1957). Non null ranking models I. *Biometrika*, **44**, 114-130.

Popovici, V. et al. (2009). Selecting control genes for RT-QPCR using public microarray data. *BMC Bioinformatics*, **10**, Article 42.

Schimek, M.G., Mysickova, A. and Budinska, E. (2010). An inference and integration approach for the consolidation of ranked lists. Accepted for publication in *J. Statist. Plan. Inference.*

Schimek, M.G. et al. (2011). Package `TopKLists` for rank-based genomic data integration. *IASTED CompBio 2011 Proceedings*, ACTA Press, Calgary, Canada.