Search for top-k consensus objects in multiple ranked lists: TopKInference versus other recent procedures

Vendula Švendová[1], Michael G. Schimek[1,2]
[1]Medical University of Graz, Graz, Austria
[2]Corresponding author's email: michael.schimek@medunigraz.at

Abstract

In recent years an increasing interest in the statistics of ranked lists can be seen, primarily stimulated by new biotechnologies as well as novel Web tools. Typically, such lists comprise tens of thousands of objects (e.g. genes in highthroughput analysis or URLs in Web search engines). However, only a comparably small subset of k top-ranked objects is informative. These objects are characterized by a strong overlap of their rank positions when they are assessed several times. The statistical task is to identify these top-k elements. Until recently there has not been a formal approach to this challenging task. Because of the high dimensionality of the data sets of interest and the associated severe multiplicity problem when classical inference procedures are adopted, it is hard to derive a methodologically convincing solution. Hall and Schimek [2] have suggested a moderate deviation-based inference procedure implemented in the `R` procedure `TopKInference`. Because of the high practical relevance of the top-k ranked lists problem, recently also other approaches have been suggested, usually in the domain of genomics. We compare [2] in detail with the most comprehensive method of Plaisier et al. [4].

Keywords: Heatmap, hypergeometric distribution, inference, moderate deviation, rank order, top-k ranked list.

1 Introduction

Ranking a set of objects or items is a widely used practice in many fields of application. Judgements of items can be performed by humans or machines. Judges or assessing devices are assumed to rank all objects from 1 to N (without ties), independently of each other, and their rankings might differ substantially. The result is always the same: L ranked lists when there are L assessments. What these lists have in common is the prespecified set of objects. A fundamental statistical task is the specification of a top-ranked subset on which the assessors agree apart from irregularities caused by the complexity and the size of the decision processes involved. Apart from the resulting random perturbations of index values, the obtained subsets are most likely to differ in the composition of the top-ranked items. As a direct consequence, these truncated ranked lists do not share all of their objects (problem of missing values). Note, when N is in the thousands and tens of thousands, the computational burden of this task is enormous.

Hall and Schimek (2012) in [2] have attempted to tackle all the above mentioned problems. Their procedure allows to estimate a consolidated subset of length k under irregular and missing assignments, even when the transition between the top part of the compared lists and the rest is fuzzy (lack of "cliff" event). Moreover, under the assumption of a distance parameter $\delta = 0$ (for details see later), symmetry of the ranked lists prevails with the consequence that no reference list L^0 (ground truth) needs to be specified. The approach of [2] is neutral with regard to origin and type of observations forming the basis for the rankings. Most of the mentioned features are not shared by other approaches aiming at the goal of top-k list identification. The rank-rank hypergeometric overlap (`RRHO`) approach in [4], for instance, is limited to gene expression

measurements with no missing observations. Genomics is currently an important application field of ranking procedures. The proposal in [6], as another example, requires the specification of a reference list, does not allow for fuzzyness, and is limited to a rather small N. In the following, we shortly describe the approaches in [2] and [4], propose a quite general data model, and compare the concerned algorithms for simulated data. Finally, results are given and conclusions are drawn.

2 Methods

Let us first describe the nonparametric inference approach for the truncation of paired ranked lists in [2]. Its iterative algorithm implemented in `TopKInference`, a module of the `R` package `TopKLists` allows estimating the length, k, of a top-k list. 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 assignments in real data. A unique feature of [2] is that the assumption of complete list independence can be relaxed (proofs also hold for m-dependence).

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 need not 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 top-k list length is obtained via $\hat{k} = \hat{j} - 1$ from $\hat{j}$, which itself is calculated by 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 $z_\nu \equiv \left(C\,\nu^{-1}\,\log\nu\right)^{1/2}$, 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_m for which m 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, inference results also depend on δ which characterizes the typical shift of items between two empirical rankings.

The second approach considered here is the rank-rank hypergeometric overlap (`RRHO`) method due to Plaisier et al. (2010) in [4]. It aims primarily at the graphical characterization of overlap in two lists and as a by-product estimates the top lists length k. In distinction to [2] this method was developed against the background of differential microarray expression analysis with the goal of making shifts of expression levels explicit, given a common set of genes (hence no missing observations allowed). These shifts are associated with binary experimental conditions. Elevated expression values are directional (plus for overexpressed genes, minus for underexpressed genes). Therefore not only the absolute value (as in [2]) but also the sign is relevant for the rankings with unchanged genes positioned in the middle. However, the statistical approach of [4] is much more general than one would expect from the type of journal it was published in. Because of this fact, it is highly interesting to compare the `TopKInference` procedure with the `RRHO` procedure.

The core part of the `RRHO` approach are multiplicity-corrected (because of the high-dimensionality of the inference problem) t-tests. More precisely, $-\log_{10}$-transformed p-values are calculated. `RRHO` iterates entirely through both ranked lists which comprise exactly the same items (i.e. genes in [4]). It is calculated if the amount of items that are above the current thresholds in both sublists (i.e. overlapping items) is significantly more or less than would be expected by random chance according to the hypergeometric distribution. This procedure results in a $N \times N$ matrix of hypergeometric p-values, where N is the total list length. The hypergeometric distribution is widely used in microarray analysis to determine the degree of enrichment or overlap of particular subsets of genes. A p-value is obtained from the cumulative distribution function by the summation of the probability distribution from one extremity of the distribution to the value of k, the number of overlapping objects,

$$H(k; s, M, N) = \begin{cases} \sum_{j=0}^{k} h(j; s, M, N) & \text{for} \quad k \leq \overline{k} \\ \sum_{j=k}^{s} h(j; s, M, N) & \text{for} \quad k > \overline{k}, \end{cases}$$

where $\overline{k} = s(M/N)$ is the expected value for the hypergeometric distribution (s and M denote the rank thresholds of the two lists). For the full range of s's and M's, the iterations slide through the rank thresholds in the two ranked lists, and the hypergeometric p-values of observed overlap are calculated. The authors of [4] take advantage of the symmetry of the involved cumulative distribution function upon exchange of s and M, thus limiting the necessary calculations to one direction. Finally, they suggest an analytical correction for multiple hypothesis testing [1] based on the false discovery rate (FDR). For a list of hypergeometric p-values in increasing order,

$$\widetilde{p}_j = \min_{k=j,\ldots,N} \left(\frac{N}{k} H_N p_k \right),$$

where $H_N = \sum_{k=1}^{N} k^{-1}$ is a harmonic number and j is the position of the p-value in the ordered list of all p-values. Once the ranks corresponding to the most statistically significant overlap between the two lists are determined, the FDR of the observed subset of k overlapping items can be determined by calculating

$$\text{FDR}_{\text{sublist}} = \frac{k_{\text{expected}}}{k_{\text{observed}}},$$

where $k_{\text{expected}} = (sM)/N$.

Both `TopKInference` and `RRHO` work on pairs of ranked lists. Strategies have already been developed to aggregate the results from all possible pairs of L lists to obtain an overall top-k list length k^* (for details see [5]). These are also applied here for the comparison of the two methods based on 5 simulated ranked lists.

3 Data Simulation

To test the performance of the described approaches, we used simulated data. We generated 5 lists of 1000 objects each, out of which 50 were preferentially ranked at the top. This was achieved by generating 50 values from the exponential distribution with $\lambda = 1.1$ and 950 random values from the normal distribution $N(0, 0.1^2)$. The λ-parameter controls the distribution of the preselected 50 objects across the whole index field of each ranked list. This means that a small number of these 50 objects might take arbitrary values that do not qualify them for the top range of the list due to randomness. The complete set of 1000 values was assigned to the object names 'obj1',...,'obj1000' and subsequently these object names were ranked according to their actual values. The

design of these 5 lists follows the usual assumption of independent assessments. However, there are practical situations, e.g. in Web queries when more than one provider is boosting the ranking of the same commercial portal, where dependence between ranked lists is induced. Hence, we decided to also study the behavior of the two methods of interest under the violation of the independence assumption of ranked lists. In the comparative analysis we allowed for independence and three different degrees of rank correlation, $\rho = \{0.2, 0.4, 0.8\}$. For the construction of these dependent lists we adopted the Iman-Conover method of [3]. Our simulation procedure is depicted in Figure 1.

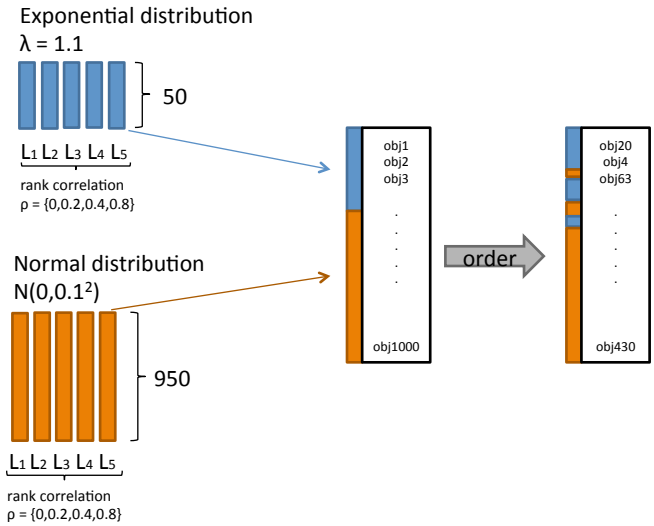


Figure 1: Scheme of data simulation with Iman-Conover method

4 Comparison and Results

For the comparison we analyzed $L = 5$ simulated lists, as described in the section 3. The `TopKInference` method of [2] is implemented in the `R` package `TopKLists`. The `RRHO` method of [4] is implemented in a Web application at http://systems.crump.ucla.edu/rankrank/. Due to the nature of the simulated data, the correctly estimated k should lie in the interval between 45 and 50 and certainly not exceed the number of planted elements, i.e. 50. The estimation of k's using both methods under independence and different degrees of correlation are listed in Table 1. Under modest correlation, k is correctly estimated by both methods. However, this does not hold for $\rho = 0.8$: `RRHO` fails to recognize the planted elements, while `TopKInference` still yields a correct estimate of k.

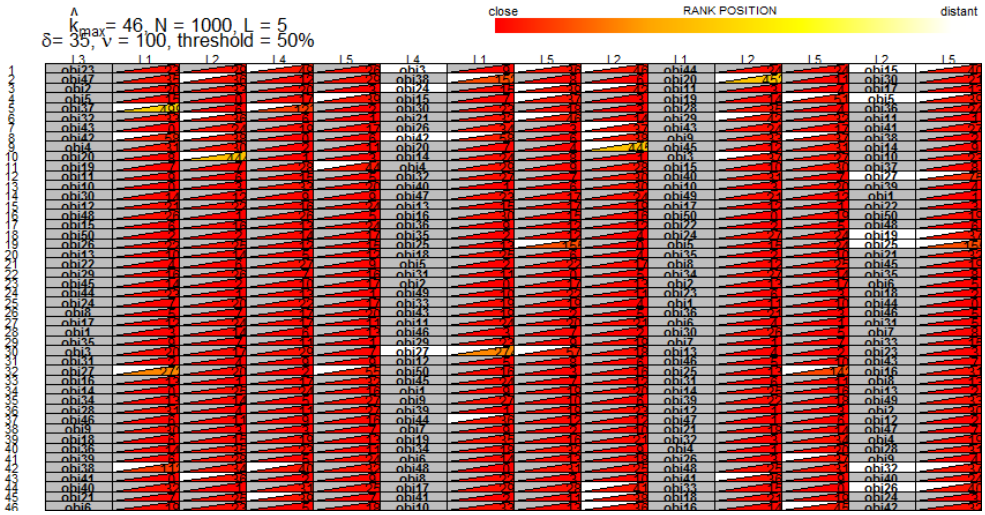
The graphical results of both methods are shown in Figures 2 and 3. The aggregation map of `TopKInference` shows the top-k truncated lists. There are $L - 1$ aggregation groups combined in one display, each of which has another reference list. Each object's membership in the top-k list is denoted by its colour: gray objects are members while white objects are not (threshold of 50%). The distance (number in the triangle) of the rank of an individual item in the reference list from its position in the other list, is denoted by a triangle colour scaled from red (identical) to yellow (far distant). Figure 2 shows the aggregation maps for independence ($\rho = 0$) and substantial correlation ($\rho = 0.8$). The individual tuning parameters are given in the figures. To exemplify the use of the `RRHO` heatmaps (Figure 3), we selected the first two input lists. The heatmaps represent the $-\log_{10}$ transformed p-values. The start position for both rankings is located in the bottom left corner of each heatmap, while the end position is located in the

top right corner. Red colour signifies the highest observed overlap. For a reliable top-k estimation the red pixels need to be concentrated in the bottom left corner. It can be easily seen that for the extreme correlation of $\rho = 0.8$ those top-50 planted elements are not identified.

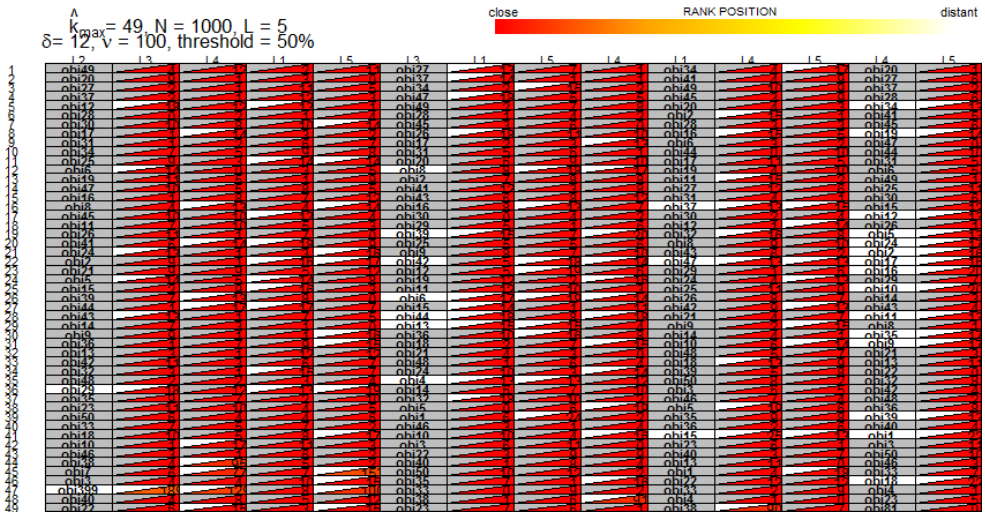
In summary both described methods, `TopKInference` and `RRHO`, perform equally well when estimating top-k for independent or slightly correlated ranked lists. For extreme correlations, which might appear in Web queries, the `RRHO` fails to detect the true k, while the `TopKInference` estimates k correctly.

Table 1: Estimated overall top list length k^* obtained from 5 input lists

	`TopKInference`	`RRHO`
$\rho = 0$	46	46
$\rho = 0.2$	47	47
$\rho = 0.4$	47	47
$\rho = 0.8$	49	416

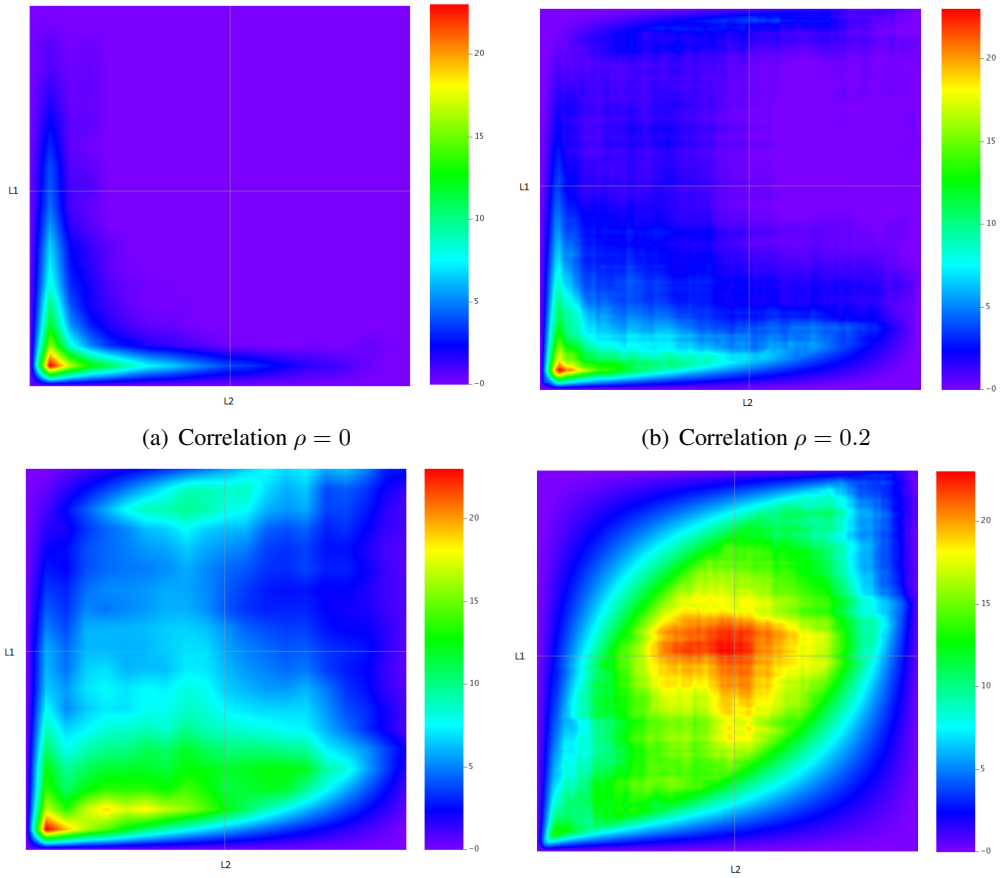


(a) Correlation $\rho = 0$



(b) Correlation $\rho = 0.8$

Figure 2: `TopKInference` aggregation maps



(a) Correlation $\rho = 0$ (b) Correlation $\rho = 0.2$

(c) Correlation $\rho = 0.4$ (d) Correlation $\rho = 0.8$

Figure 3: `RRHO` heatmaps after multiplicity correction

References

[1] Benjamini, Y. and Yekutieli, D. (2001) "The control of the false discovery rate in multiple testing under dependency." *Ann. Statist.*, 29, 1165-1188.

[2] Hall, P. and Schimek, M. G. (2012) "Moderate-deviation-based inference for random degeneration in paired rank lists", *J. Amer. Statist. Assoc.*, 107, 661-672.

[3] Iman, R. L. and Conover, W. J.(1982) "A distribution-free approach to including rank correlation among input variables", *Commun. Statist. - Simula. and Computa.*, 11(3), 311-334.

[4] Plaisier, S. B. et al. (2010) "Rank-rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures", *Nucl. Acids Res.*, 38, 17, e169.

[5] Schimek, M. G. et al. (2011) "Package "TopKLists" for rank-based genomic data integration." Proceedings of IASTED CompBio Conference, 434-440, DOI: 10.2316/P.2011.742-032.

[6] Sampath, S. and Verducci, J. (2012) "Is there a partial consensus ordering between rankings?". Unpublished manuscript of JSM 2012 presentation.