Novel methods for the statistical analysis of multiple and repeated rankings

Michael G. Schimek
Medical University of Graz, IMI-RU 'Statistical Bioinformatics', Graz, Austria -
michael.schimek@medunigraz.at

Vendula Švendová
Medical University of Graz, IMI-RU 'Statistical Bioinformatics', Graz, Austria -
vendula.svendova@medunigraz.at

Abstract

In recent years there has been an increasing interest in the statistics of ranked lists. This trend is motivated, for instance, by the wish to compare rankings of academic journals, by the need to process Web search engine results, and by the analytic requirements of high-throughput biotechnologies. Typically, such lists comprise between hundreds and tens of thousands of items (e.g. ordered by impact factors, URLs of profile pages, gene expression values). However, only a comparably small subset of k top-ranked items is informative and useful. Items listed in the top-range are typically characterized by a strong overlap of their rank positions when they are ranked by different instances of assessment. A central statistical task is the estimation of an overall k^* for a number of ranked lists comprising the same set of items, before one can fit a consolidated data model to the obtained sublists. We present recent methodological developments with a special focus on the case of correlated ranked lists. We introduce a novel approach for the analysis of repeated rankings, provide some simulation evidence, and apply it to omics time course data.

Keywords: Correlation, nonparametric inference; `R` package; top-k ranked lists.

1. Introduction

Typically, ranked lists from high-tech devices such as Web search engines or high-throughput biotechnologies comprise tens of thousands of items. As a direct result, statistical inference is rather demanding. However, only a comparably small subset of k top-ranked items is relevant in practice for further consideration. These items are characterized by a strong overlap of their rank positions when the lists are produced by independent assessors adopting identical selection rules (human-based or machine-based). On the other hand, when, for instance, rankings of academic journals are compared, the number of items is only in the hundreds, but usually we are then interested in these rankings across time and the analytic complexity goes up again. In summary, there are two scenarios: (i) very long ranked lists of independent assessments and (ii) relatively short ranked lists of dependent assessments (along the time axis). What both have in common is the fact that there are many more items to be ordered than assessors. In common voting problems (e.g. opinion polls) it is exactly the other way round. Occasionally, scenarios (i) and (ii) are merged, for example in time course microarray experiments. The inference task is always the same: identification of a stable top-ranking subset of items.
Hall and Schimek (2012) have provided us with a nonparametric inference method for paired ranked lists. Schimek, Myšičková and Budinská (2012) have proposed a strategy for inference in multiple ranked lists as long as these lists are independently generated. Moreover, they have contributed various simulation evidence for that situation. Schimek et al. (2015) describe the `TopKLists` `R` package in which, among other methods, the concerned inference procedures are implemented, and present an omics application. In this paper we focus on the practically most relevant case of dependencies across lists, with respect to simulated data as well as real data from a microarray time course experiment.

2. General assumptions

Let us have L input lists representing rank positions of the same set of N items. There are either several independently operating rank assignment mechanisms or just one rank assignment mechanism applied consecutively (i.e. along the time line). The rank position of an item might be the result of measuring the strength of evidence or of assessment based on expert knowledge or preference. The ranking of items is from 1 to N, for highest to lowest without ties, but missing assignments are allowed.
We assume a discrete space O that contains all N items, denoted by $o_i,\ i = 1, \ldots, N$. Since all items o can be associated with a unique label $i = 1, \ldots, N$, O can be viewed without loss of generality as a list $O = \{1, 2, \ldots, N\}$. Let us denote the rank of element o_i in O by $R(i)$ under a particular assignment. Then a permutation of $O, \tau(O) = \{1, 2, \ldots, N\}$, such that $R(i) \leq R(j)$ for any $i < j$ is a complete ranking of the items in O. We refer to $\tau(O)$ as a full ranked list, and to $R_\tau(i)$ as the rank of item o_i under the assignment mechanism (i.e. assessment) τ.
In most instances of integration of multiple and repeated rankings (N usually large or huge and L small), a common (i.e. consolidated) ranked list of full length is not desirable. Instead, one is only interested in a partial list (sub-space) $O' \subset O$ of length k which can be further analysed (e.g. a statistical model can be fitted) or immediately interpreted. Without loss of generality, we assume that the partial ranked list $\tau(O') = \{o'_1, o'_2, \cdots, o'_k\}$ is ordered according to their ranks such that $R(o'_i) < R(o'_j)$ for $i < j$. It is implicitly assumed that all the items that are in O but not in O' are ranked lower than k (i.e. have indices $k + 1, k + 2, \ldots, N$).

3. The inference method

In Hall and Schimek (2012) a nonparametric inference method for the truncation of paired ranked lists was developed. The implementation in `TopKLists` allows estimating the length, k, of a top-k list in the presence of irregular and missing assignments. 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 item 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. However, the above mentioned authors could prove that their theoretical results obtained under the assumption of complete independence also apply to the situation of m-dependence which is of special interest in our paper.
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 items, degenerates into noise (degradation of information). The estimation of $\hat{j}_0 - 1 = \hat{k}$ 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

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

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 the distance δ (for the choice of these tuning parameters see the vignette of `TopKLists` in Schimek et al., 2014).
The complex decision problem is solved via an iterative algorithm in `TopKLists`. The overall estimate $\hat{k}^*$ for the L rankings resulting from independent assessments τ is calculated in the following way: The inference procedure is executed for all possible pairs $M = (L^2 - L)/2$ of L lists, thus we obtain M values $\hat{k}_j$

$(j = 1, 2, \ldots, M)$. The overall top-k list length is then defined by $\hat{k}^* = \max_j(\hat{k}_j)$. Of course other criteria for top-$\hat{k}$-combination could be used as well, but maximum is most intuitive.

4. A construction concept for multiple and repeated rankings

Not only for studying the numerical behaviour of an inference method but also for practically motivated purposes, it is necessary to be able to simulate multiple and repeated ranked lists. How this can be achieved is the content of this section.
In such data the level of disagreement between the assessors' rankings need to be controlled or the data have to reflect a pre-specified dependence structure, for instance following a time course design. We will call *true measurement* the measurement (e.g. the *true* impact factor of an academic journal), that would be obtained under ideal conditions in absence of (random) errors (representing the various shortcomings of assessment). The true measurements form what we call the *true ranking* each assessor is trying to achieve.
We consider two scenarios, multiple ranked lists and repeated ranked lists. In the first scenario, several assessors or assignment mechanisms rank the items independently of each other. In the second scenario, only one assessor or assignment mechanism ranks the items at different time points, so there is an expected consecutive m-dependence because at each time point the assessor's decision process is reflecting the decisions at previous time points. Depending on the involved number of decisions in the past, we have 1-dependence, 2-dependence, and so on. Here, our focus is on 1-dependence, the most frequent case in practice.
Let us assume that there is a true measurement for each of the N items, say, vector $\Theta^{true} = (\Theta_1^{true}, \ldots, \Theta_N^{true})$. Each of the L assessments (whether multiple or repeated) results in a vector of values, say, $\Theta^\ell = (\Theta_1^\ell, \ldots, \Theta_N^\ell), \ell = 1, \ldots, L$, that differs from the Θ^{true} by some unknown random vector. We wish to simulate these Θ^ℓs and rank their values for both scenarios. Because concordance between the lists is considered informative and non-random only in a subset of k top-ranked items, we need to implement the dependency of the Θ^ℓs on Θ^{true} only for the first k index positions and let the remaining values follow a random distribution. Let us split the Θ^{true} and Θ^ℓ into the top-k values and the remaining $N - k$ values, denoting them as $\Theta^{true} = \begin{pmatrix} \theta^{true_k} \\ \theta^{true_{N-k}} \end{pmatrix}$ and $\Theta^\ell = \begin{pmatrix} \theta^{\ell_k} \\ \theta^{\ell_{N-k}} \end{pmatrix}$. When constructing multiple lists, there are two ways of introducing the random effect. In the first case, we assume a normal mean zero random variable with error variance σ^2, so that

$$\theta^{\ell_k} = \theta^{true_k} + N(0, \sigma^2) \qquad for\ \ell = 1, \ldots, L,$$

where σ^2 is small enough (otherwise the overlap of ranks in the top parts of the lists would be lost). In the second case we substitute the random variable by assuming there is a rank correlation ρ between θ^{true_k} and each of the θ^{ℓ_k}s, such that

$$r_s(\theta^{true_k}, \theta^{\ell_k}) = \rho \qquad for\ \ell = 1, \ldots, L,$$

where $r_s(x, y)$ denotes Spearman correlation between variables x and y. This can be achieved using the so-called Iman-Conover method (Iman & Conover, 1982) which yields approximate results. Therefore, there is an implicit small error (assumed as random) resulting form the Iman-Conover algorithm in each of the constructed data vectors. We can also use Iman-Conover for the simulation of 1-dependent (repeated) ranked lists. Their algorithm allows us to control the change in time by a pre-specified rank correlation between the consecutive pairs of lists, thus having

$$r_s(\theta^{1_k}, \theta^{2_k}) = \ldots = r_s(\theta^{L-1_k}, \theta^{L_k}) = \rho$$

for

$$\theta^{1_k} = \theta^{true_k} + N(0, \sigma^2).$$

The general procedure for introducing the required rank correlation is the same in both scenarios: We define $X_{k\times(L+1)}$, the input matrix, whose columns follow an arbitrary distribution. We generate the columns from the exponential distribution $\text{Exp}(\lambda)$. For practical reasons, we sort the matrix by its first column in ascending order and define the true measurement $\theta^{true_k} := X[, 1]$, the first column of the matrix X. This assures that the true ranking will be $1, 2, 3, \ldots, k$ and allows for easier visual inspection of the results. In case of multiple ranked lists we require that the (pre-specified) rank correlation is the same between θ^{true_k} and any of the θ^{ℓ_k},

while in the case of repeated ranked lists, we demand the same degree of correlation between each pair of the consecutive lists. To achieve these pairwise correlations, we create L sub-matrices, $X^i_{sub}, i = 2, 3, \ldots, L+1$. Each sub-matrix will consist of two columns

- $X^i_{sub} = (X[,1], X[,i]), i = 2, \ldots, L+1$ for multiple lists
- $X^i_{sub} = (X[,i], X[,i+1]), i = 2, \ldots, L+1$ for repeated lists,

where $X[,i]$ stands for the i-th column of the matrix X. Then we define the desired rank correlation matrix

$$C = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

between the two columns of X^i_{sub}. In the next step we have to generate the arbitrary set $\{a_j\}, j = 1, \ldots, k$ of k numbers (often called *scores*), and construct a matrix, say, $M_{k\times 2}$, whose columns represent two independent permutations of the set $\{a_j\}$. We fill the first column of M with ordered scores $\{a_j\}$ (again to preserve the *true* ranking) and fill the second column with the same scores randomly permuted. Such matrix M should have uncorrelated columns. Because C and the sample correlation matrix of M, say $T := corr(M)$, are symmetric and positive definite, we can find matrices P and Q, so that $C = P'P$ and $T = Q'Q$, using the Cholesky factorization scheme. Iman and Conover (1982) could prove that the rank correlation matrix of $S := MQ^{-1}P$ is almost exactly equal to the desired C. Now it is only left to reorganize the matrix X^i_{sub}, so it has the same rank order as the obtained S. The resulting two columns of the reordered matrix X^i_{sub} have approximately rank correlation ρ. Once we have generated all the $\theta^{1_k}, \theta^{2_k}, \ldots, \theta^{L_k}$ in the same manner, we create a matrix $Y_{k\times(L+1)}$ with columns $\theta^{true_k}, \theta^{1_k}, \theta^{2_k}, \ldots, \theta^{L_k}$. The columns of the matrix Y are the simulated measurements for each assessor (multiple lists), respectively for each time point (repeated lists). Ranking the values in each column will result in the desired L lists comprising the top-k ranks. For each of the Θ^ℓ, to obtain complete lists of N items, the remaining $N-k$ measurements are derived from $N(0, \sigma^2)$.

5. Some simulation evidence for the inference method

For each of the Spearman correlation-based scenarios we have generated two types of top-k lists: well and poorly separating from the (random) rest of the lists. For the case of well separating lists (we call them *cliff data*), we formed 5 ranked lists of 1000 items each, out of which 50 were preferentially ranked at the top. As described in the section above, we derived 50 values from the exponential distribution with $\lambda = 1.1$ and introduced a rank correlation of $\rho = 0.9$ between them. The remaining 950 values were generated as random following the normal distribution $N(0, 0.1^2)$. For the case of poorly separating lists (we call them *non-cliff data*), we proceeded as described above for the cliff data. However, in order to make it more difficult for the inference procedures to estimate k, some fuzziness was introduced between the index positions k and $k+10$ by generating 10 random values from the exponential distribution with $\lambda = 10$. In this setting rank correlation was introduced within the first 60 values. The remaining 940 values were again taken from $N(0, 0.1^2)$. For the situation of repeated rankings we assumed substantial 1-dependence by choosing ρ values of 0.9 and -0.9. All together we obtained 6 different types of datasets sharing most of their parameter settings for the purpose of comparison. In the simulations 1000 replicates were generated of each dataset and the median of the individual $\hat{k}$s was calculated. We expect the inference procedure to identify the top-50 most conforming items among the generated 1000 items. We applied the nonparametric inference method from the `TopKLists` package on these 6 types of datasets. In Table 1, for the multiple and repeated scenarios the calculated $Med(\hat{k})$ are displayed for pre-specified distances δ (appropriate δ values are usually not the same for multiple and repeated rankings). From Table 1 it is obvious that for multiple rankings, the k is confidently estimated near the true value of 50, even for the case of non-cliff data, for which it is generally harder to estimate the true k. In repeated rankings the overall concordance is influenced by the type of dependency, which is even more demanding for the inference procedure, that had been constructed for independent and low-dependent assessments. However, all obtained results are congruent with those seen for the simpler multiple scenario. We cannot observe any detrimental effect of substantial 1-dependence on the median of the resulting $\hat{k}$s.

	$\rho = 0.9$		$\rho = -0.9$
	Multiple	Repeated	Repeated
Cliff	$Med(\hat{k}) = 49$ ($\delta = 15$)	$Med(\hat{k}) = 49$ ($\delta = 20$)	$Med(\hat{k}) = 49$ ($\delta = 20$)
Non-cliff	$Med(\hat{k}) = 56$ ($\delta = 15$)	$Med(\hat{k}) = 53$ ($\delta = 20$)	$Med(\hat{k}) = 47$ ($\delta = 20$)

Table 1: Medians of the estimated $\hat{k}$ for the two different list types (cliff and non-cliff), in the multiple ($\rho = 0.9$) and in the repeated scenarios ($\rho = \pm 0.9$); the constant C is default and the pilot sample size $\nu = 10$ in all cases.

6. An omics example for the analysis of repeated rankings

As an example, we analysed with the described nonparametric inference method a time course experiment. The data set was retrieved from the NCBI GEO database (accession number GSE56899). It belongs to a study of Koh et al. (2014) using high-throughput methods of RNA analysis (i.e. microarrays and next-generation sequencing) to characterise the global landscape of circulating RNA in human subjects. Available for us were only the microarray gene expression values obtained from circulating cell-free RNA in the blood of 11 pregnant women and 4 controls. In that study blood samples were collected in each of the 3 trimesters, as well as after delivery (post-partum), resulting in 4 time-dependent complete samples from each woman (all of size 33297). As control, 4 samples were taken, 2 from non-pregnant women and 2 from men. All samples were analysed using Affymetrix Human Gene 1.0 ST Array. We have normalized the raw data applying the Bioconductor package `affy` (function `rma`). Differential expression analysis between the samples from pregnant and non-pregnant individuals was performed using the Bioconductor package `limma`, separately for each time point. Finally, for each time point (T1, T2, T3, and post) ranked lists were built for those annotated genes with values of $p < 0.1$, resulting in $N = 3386$ items (i.e. different gene symbols).
Our goal was twofold: (i) We wished to identify those top-k genes which are present in several ranked lists representing the 4 time points (not necessarily present in all of them). They are most likely maternal and not fetal because for the fetus we expect to observe some trend across the trimesters. The truncation point $\hat{k}$ can be estimated with the inference method. It should be noted that the procedure of Hall & Schimek (2012) is exploratory. Depending on the choice of the tuning parameters we are more or less strict with respect to overlap of items in the top positions of the lists. (ii) Our other goal was differentiating between top-ranked genes that present themselves highly conforming across all 4 lists and top-ranked genes that show some diversity in their rank positions across the lists. For that purpose the aggregation map introduced in Schimek et al. (2015) and implemented in the `TopKLists` package is quite useful. It provides a summary of the aggregate behaviour of all truncated ranked lists and will be described for our example data in the following. Figure 1 shows the aggregation map of the example data at the 4 time points.
The application of the inference method of `TopKLists` resulted in a $\hat{k}_{max} = 27$ of top-listed items. All the involved genes, as expected, are maternal. The aggregation map in Figure 1 is organised as follows: left the complete group of T1-post-T2-T3, in the middle the group of T2-T3-post, and right the group of T3-post. Note, this arrangement is driven by the data (the data structure enforced this ordering, not the time scale). The `aggmap` algorithm ordered the top-lists according to the degree of their pairwise overlap from left to right into the above groups. In the left group, for instance, T1, the reference list, and post have the strongest overlap, followed by T2 and T3. For each item there are triangles scaled from red (*very close*) to yellow (*far distant*) representing the distance d between the reference list (comprises the gene symbols) and that list where the triangle is printed (the number gives the actual distance between the concerned rank positions). An NA denotes a rank position beyond 3386 in this example. The complementary triangle is grey when an item is a member of the top-k list, and white otherwise. The rectangle of a gene symbol presents itself in grey when the percentage of $d \leq \delta$ across the columns of a group is above the threshold of 50%, and white otherwise. In Figure 1 there are various grey rectangles pointing at stable gene expression values along time. On the other hand, the white rectangles characterise values that vary with time, presumably in connection with maternal changes that go along with the development of the fetus.

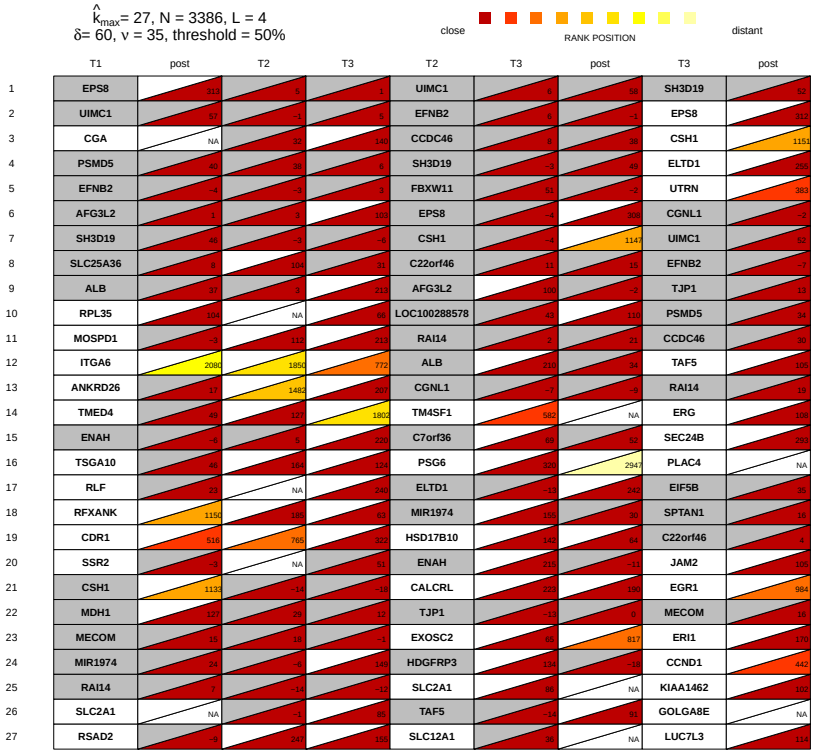

Figure 1: Aggregation map of time course microarray data: T1 - first trimester, T2 - second trimester, T3 - third trimester, post - after delivery.

References

Hall, P. & Schimek, M. G. (2012). Moderate deviation-based inference for random degeneration in paired rank lists. *Journal of the American Statistical Association, 107*, 661–672.

Iman, R. I. & Conover, W. J. (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics - Simulation and Computation, 11*, 311–334.

Koh, W. et al. (2014). Noninvasive in vivo monitoring of tissue-specific global gene expression in humans. *Proceedings of the National Academy of Sciences, 111*, 7361–7366.

Schimek, M. G, Myšičková, A. & Budinská, E. (2012). An inference and integration approach for the consolidation of ranked lists. *Communications in Statistics - Simulation and Computation, 41*, 1152–1166.

Schimek, M. G. et al. (2014a). TopKLists (`R` package version 1.0.3). `http://topklists.r-forge.r-project.org/`

Schimek, M. G. et al. (2014b). TopKLists: Analyzing multiple ranked lists (`R` Vignette). `http://cran.r-project.org/web/packages/TopKLists/vignettes/TopKLists.pdf`

Schimek, M. G. et al. (2015). TopKLists: a comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists. *Statistical Applications in Genetics and Molecular Biology*, to appear.