# Inference, aggregation and graphics for top-$k$ rank lists

**Michael G. Schimek**[1,2,*] **, Eva Budinska**[3]**,**
**Shili Lin**[4]**, Alena Mysickova**[5,2]

1. Medical University of Graz, Austria
2. Danube University Krems, Austria
3. Masaryk University Brno, Czech Republic
4. Ohio State University, USA
5. Humboldt University Berlin, Germany

* Contact author: michael.schimek@medunigraz.at

**Keywords:** Graphics Tool, Moderate Deviation, Ordered List, Rank Aggregation, `TopkLists` Package

Lists of common distinct objects in rank order are typical for various fields of application. The rank of an object belonging to the set of interest in a certain list indicates its respective position among all other objects. The rank position might be due to a measure of strength of evidence, to a consumer preference, or to an assessment either based on expert knowledge or a technical device. Let us assume that the rank assignment in each list is independent of the assignment in the other lists.

Let us have $\ell$ such lists $\tau_j$ $(j = 1, 2, \ldots, \ell)$ assigning rank positions to the same set of objects. The ranking is from 1 to $N$, without ties. Our goal is to identify a subset of objects that is characterized by high conformity across the lists. This implies that there is similarity between the rankings which can be evaluated by a distance measure $d$ (a permutation metric) such as Kendall's $\tau$ or Spearman's footrule. In practice, because of truncated rank lists and incomplete rankings of objects in some of the lists caused by missing assignments, we need to penalize these measures accordingly. Moreover, in most applications, especially for large or huge numbers $N$ of objects, it is not likely that consensus prevails, thus only the top-ranked elements are relevant. For the remainder objects their ordering is more or less at random. This is not only true for surveys of consumer preferences but also appropriate for search tasks in the Web and data integration in the field of biotechnology. In many instances we observe a general decrease, not necessarily monotone, of the probability for consensus rankings with increasing distance from the top rank position. Typically there is reasonable conformity in the rankings for the first, say $k$, elements of the lists, motivating the notion of *top-k rank lists*.

List aggregation by means of brute force is limited to the situation where both $N$ and $\ell$ are unrealistically small, and $k$ is known (e.g. 'ground truth' in Web search engines). Here our aim is to solve this computational problem for a realistic setting, firstly, via an algorithm for the selection of the $\hat{k}$'s for all $(\ell^2 - \ell)/2$ possible pairs of lists $\tau_j$, secondly, via a graphical tool monitoring the aggregation process of the thus obtained top-$k$ rank list information, and thirdly, via an algorithm for the calculation of a set of objects characterized by rankings of high conformity across the lists up to some global index $\bar{k}$. For the first task we take advantage of a moderate deviation-based inference procedure for random degeneration in paired rank lists (Hall and Schimek, 2009). The graphical tool is a newly developed type of heat map simultaneously displaying three-dimensional information representing the dynamics of the aggregation process based on the input from the inference procedure. For the last task an Order Explicit Algorithm (OEA) is combined with cross-entropy Monte Carlo (CEMC), as outlined in Lin and Ding (2009). For the same input lists the aggregation result does not only depend on the index $\bar{k}$ but also on the chosen distance measure and adopted concept for the handling of partial lists (incomplete sets of objects), apart from necessary tuning parameters. Therefore a graphical monitoring tool is much desirable.

Although the discussed methodology is quite general in terms of application, we take a special interest in the meta analysis of microarray experiments. Hence we apply the above algorithms, which we are implementing in the `R` package `TopkLists`, to both artificial and real gene expression data. `TopkLists` is based on the most recent algorithmic developments, allows for $N$ in the magnitude of thousands, and will serve as universal tool for the objective identification of informative objects (e.g. genes) conforming across rank lists.

## References

Hall P. and Schimek M.G. (2009). Moderate deviation-based inference for random degeneration in paired rank lists. Preprint.

Lin S. and Ding J. (2009). Integration of ranked lists via Cross Entropy Monte Carlo with applications to mRNA and microRNA studies. *Biometrics*, to appear.