Communications in Statistics - Simulation and Computation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lssp20

An Inference and Integration Approach for the Consolidation of Ranked Lists

Michael G. Schimek [a] , Alena Myšičková [b] & Eva Budinská [c]

[a] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria

[b] Max Planck Institute for Molecular Genetics, Berlin, Germany

[c] Swiss Institute of Bioinformatics, Lausanne, Switzerlandand Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic

Version of record first published: 02 Apr 2012.

To cite this article: Michael G. Schimek , Alena Myšičková & Eva Budinská (2012): An Inference and Integration Approach for the Consolidation of Ranked Lists, Communications in Statistics - Simulation and Computation, 41:7, 1152-1166

To link to this article: http://dx.doi.org/10.1080/03610918.2012.625843

Communications in Statistics—Simulation and Computation®, 41: 1152–1166, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610918.2012.625843

An Inference and Integration Approach for the Consolidation of Ranked Lists

MICHAEL G. SCHIMEK[1], ALENA MYŠIČKOVÁ[2], AND EVA BUDINSKÁ[3]

[1]Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria
[2]Max Planck Institute for Molecular Genetics, Berlin, Germany
[3]Swiss Institute of Bioinformatics, Lausanne, Switzerland and Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic

In this article, we describe a new approach that combines the estimation of the lengths of highly conforming sublists with their stochastic aggregation, to deal with two or more rankings of the same set of objects. The goal is to obtain a much smaller set of informative common objects in a new rank order. The input lists can be of large or huge size, their rankings irregular and incomplete due to random and missing assignments. A moderate deviation-based inference procedure and a cross-entropy Monte Carlo technique are used to handle the combinatorial complexity of the task. Two alternative distance measures are considered that can accommodate truncated list information. Finally, the outlined approach is applied to simulated data that was motivated by microarray meta-analysis, an important field of application.

Keywords Cross-entropy Monte Carlo; Kendall's τ; Moderate deviation; Partial list; Random degeneration; Rank aggregation; Spearman's footrule; Top-k ranked list.

Mathematics Subject Classification 62G86.

1. Introduction

In various fields of application, such as consumer preference, election of political representatives or social choice, we have to consolidate lists of rankings for the same set of objects or subjects. The ranking is usually performed by selected persons (assessors or voters) and recently also by technical devices (e.g., omics platforms). When the number of items is small and the number of assessors is large,

Received August 20, 2010; Accepted October 15, 2010

Address correspondence to Michael G. Schimek, Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Auenbruggerplatz 2/V, Graz 8036, Austria; E-mail: michael.schimek@medunigraz.at sometimes called the majority vote problem, statistical methods are available (see, e.g., Mallows, 1957; Fligner and Verducci, 1986; Cohen et al., 1999). However, we are interested in the case where the number of items is large or even huge and the number of assessors (ranking mechanisms) is small. This is typical for tasks in genomics when information needs to be combined across multiple biological experiments and laboratory platforms. For instance, in microarray studies, the use of different technologies means that not all studies yield comparable gene expression levels. However, we might want to derive a global list of genes which are differentially expressed between two conditions such as *pathogenic* vs. *normal* based on the complete experimental evidence. This task requires the meta-analysis of the available experimental outcomes.

Regrettably, combinatorial solutions are NP-hard (e.g., see Fagin et al., 2003) and standard statistical techniques no longer apply. Alternative, computationally feasible techniques are in demand. For example, similarities of ordered gene lists for the meta-analysis of microarray experiments are studied in Yang et al. (2006). Most recently, a data integration approach has been proposed (Lin and Ding, 2009), which enables dealing with a few lists of up to a few hundred items. Both concepts are interesting but inappropriate for the consolidation of the very long lists (usually with rank information degradation) that we address here. An additional complication is dealing with lists which do not necessarily comprise the same set of objects. In response to these demands, we suggest a new statistical concept which: (i) allows us to truncate lists in a pairwise manner when the degree of overlap of rank positions becomes erratic; and (ii) helps to aggregate the truncated lists thereby obtained. We finally end up with a consensus-based subset of objects in a new rank order. Our approach avoids computational complexity to such an extent that data sets of reasonable size can be handled. For instance, high-throughput biotechnologies are very demanding because they typically produce tens of thousands of measurements in each experiment.

For the rest of this article let us assume that the rank assignment in each list does not depend on the assignment in the other lists, i.e., ranking persons or devices act stochastically independent of each other. Further more, let us have ℓ such input lists representing rank positions of the same set of N objects. The ranking of objects is from 1 to N, without ties, where 1 denotes the highest and N the lowest ranking.

We assume a discrete space O that contains all N objects, denoted by o_i, $i = 1, \ldots, N$. Since all objects 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. Under the assignment mechanism τ, we refer to $\tau(O)$ as a full ranked list, and to $R_\tau(i)$ as the rank of object o_i.

In the particular applications we have in mind, a full ranked list is neither desirable nor available. Instead, one is only interested in a partial list (sub-space) $O' \subset O$ of length k. 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 $k+1, k+2, \ldots, N$).

In this article, we will have ℓ such lists of length k_l ($l = 1, 2, \ldots, \ell$). Note that neither the lengths of the lists nor the underlying spaces $O_1, O_2, \ldots, O_\ell$ necessarily need to be the same.

Our goal is to identify a subset of objects that is characterized by high conformity across the ℓ lists, more specifically, to arrive at a new ranking of the items in $O' = \bigcup_{l=1}^{\ell} O'_l$, or a top-$k$ list of O', that integrates the information contained in the partial (truncated) 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 τ (not to be confused with the already introduced τ denoting a ranked list) or Spearman's footrule. The problem with such a measure is: (i) that we need to have a complete ranking of the objects in all lists (no missing assignments); and (ii) that the amount of consensus (the probability for overlap in rank position for each object across the lists) is assumed to remain the same from the highest to the lowest rank. However, in most applications, especially for large or huge numbers N of objects, it is unlikely that consensus prevails. This is true for consumer preferences of products as well as for many applications in science and technology. Typically, we can observe a general decrease, not necessarily monotone, of the probability for consensus rankings with increasing distance from the top rank position. Moreover, it is often the case that there is reasonable conformity in the rankings for the first k items of the lists, motivating the notion of *top-k ranked lists*. Hence, for our task, we are required to aggregate ℓ lists O'_l of various lengths k_l under the condition of missing assignments (incomplete rankings) because a specific object might not be a member of each of the partial lists. To deal with this condition the distance measure d needs to be adapted to subspaces.

The idea of distance measure modification will be taken up following the introduction of Kendall's τ or Spearman's footrule, respectively. Then our stochastic approach will be discussed with respect to inference (i.e., estimation of the k_l's) and integration (i.e., aggregation of objects belonging to the subspace O' characterized by high conformity of items across lists). Finally, various simulation evidence motivated by gene expression (microarray) data is provided. However, the scope of our method goes far beyond microarray data integration as will be pointed out in the conclusions.

2. Measures of Distance and Their Adaptation to Partial Lists

A measure of distance is essential when objects are aggregated across ranked lists. Apart from this special application, measures of distance or metrics are relevant for non-parametric rank-based statistical methods in general. Such measures have been treated extensively, for instance in Marden (1995).

Here, we discuss two measures, Kendall's τ distance (Kendall, 1938) and Spearman's footrule distance (Spearman, 1906), which can both be modified to deal with truncated lists with presumably different underlying spaces. For a thorough discussion of space aspects in ranking procedures, see Lin (2010).

2.1. *Kendall's τ*

Kendall's τ is equal to the number of adjacent pairwise exchanges required to convert one ranking, or permutation, to another. Essentially, this means counting the number of pairwise discordances between the two lists. Let us have two full ranked lists τ_1 and τ_2 on space O, and pairwise discordances between the two lists

$$d_k(i, j) = I\left[(R_{\tau_1}(i) - R_{\tau_1}(j))(R_{\tau_2}(i) - R_{\tau_2}(j)) < 0\right]$$

for $i, j = 1, \ldots, N$. $I(\cdot)$ is an indicator function that takes the value of zero or one depending on whether the condition within the brackets is satisfied or not. Then, Kendall's τ distance is given by

$$K(\tau_1, \tau_2) = \sum_{\{i,j\} \in O} d_k(i, j).$$

Its maximum is $N(N-1)/2$ where N is the list length.

As pointed out already, a complication is handling incomplete rankings obtained from truncated lists. Conventional distance measures cannot be applied directly since they deal only with comparing one permutation against another over the same set of objects (requires complete rankings throughout). But Kendall's τ can be modified to account for incomplete rankings as shown below.

Let us have two truncated lists, τ'_1 and τ'_2 of length k_1 resp. k_2. Let τ_1 and τ_2 be the underlying associated ranking mechanisms over the space O. Assume that the sub-space O' contains all the objects that are present in either O'_1 or O'_2 ($O' = O'_1 \cup O'_2$). Let O'^c_m denote the complement of O'_m for $m = 1, 2$. For each $i \in O'$, if $i \in O'_l$, then the rank $R_{\tau_m}(i)$ is defined as in the original list; if $i \in O_m \cap O'^c_m$, then define $R_{\tau_m}(i) = k_m + 1$, otherwise the rank is left undefined. For each pair of items $i, j \in O'$, let D be the collection of pairs of items that are in both lists. Furthermore, let B denote the collection of pairs with the following property: both items of each pair belong to either one of the lists but not to both. We define

$$d'_k(i, j) = \begin{cases} I\left[(R_{\tau_1}(i) - R_{\tau_1}(j))(R_{\tau_2}(i) - R_{\tau_2}(j)) < 0\right] & \text{if } (i, j) \in D \cap B^c \\ p & \text{otherwise,} \end{cases} \tag{1}$$

where $I(\cdot)$ is an indicator function and p a penalty parameter taking values between zero and one, usually set to $\frac{1}{2}$. This definition due to Lin (2010) differs slightly from the one given in DeConde et al. (2006). The Kendall's τ distance for the discordances given in (1) amounts to

$$K(\tau'_1, \tau'_2) = \sum_{\{i,j\} \in O'} d'_k(i, j).$$

Hence, the modified measure can be evaluated in a similar fashion to the original one.

2.2. *Spearman's Footrule*

An alternative measure of distance is Spearman's footrule. Let us assume again two permutations τ_1 and τ_2 of a set O of objects. Spearman's footrule distance is the sum of the absolute differences between the ranks of the two lists over all items in O

$$S(\tau_1, \tau_2) = \sum_{i \in O} |R_{\tau_1}(i) - R_{\tau_2}(i)|,$$

where $R_{\tau_m}(i)$ is the rank of object i in list τ_m ($m = 1, 2$). As can be seen from the formulae, Spearman's footrule takes the actual rankings of the items into consideration, whereas in Kendall's τ, only relative rankings matter. The maximum Spearman's distance is $N^2/2$ for N even, and $(N+1)(N-1)/2$ for N odd, which corresponds to the situation in which the two lists are exactly the reverse of each other.

Spearman's footrule can also be modified to allow for partial lists. Let us compare two truncated lists corresponding to the subsets O'_1 and O'_2. Again, we consider $O' = O'_1 \cup O'_2$. For each $i \in O'$, if $i \in O_m \cap O'^c_m$ (O'^c_m is the complement of O'_m) we define $R_{\tau_m}(i) = k_m + 1$. Because in Spearman's footrule not just the relative orderings are taken into consideration, one cannot simply leave the rankings of some of the objects undefined. Instead, we also let $R_{\tau_m}(i) = k_m + 1$ if $i \notin O_m$. Applying this modification, we can compute the Spearman's footrule distance for incomplete rankings in the same manner as for complete rankings.

Note that not all distance measures are also metrics (especially when adapted to partial lists). For the mathematical theory behind distance measures we refer to Fagin et al. (2003).

3. A Combined Inference and Integration Approach

A methodology is needed that allows us to consolidate ℓ lists, i.e., to calculate a new top-k list τ^* (an ordered set of objects) that is characterized by rankings of high conformity across the assessments up to position k^*.

Our data-driven approach consists of two algorithmic steps. In step one, we estimate the lengths k_l of the truncated lists O'_l from all pairwise combinations of the ℓ full lists O_l, exploiting the complete information contained in these lists. All items ranked equal or higher than some overall index k^* (a function of the individual k's), are kept for step two, the aggregation of objects under a stochastic optimization criterion. Rank aggregation is computationally extremely expensive, thus truncated lists as input are essential to reduce the computational burden (usually $k_l \ll N$ in empirical data).

3.1. *Degeneration of Pairwise Rank Information*

Hall and Schimek (2012) developed a moderate deviation-based inference procedure for random degeneration in paired ranked lists. The concept of moderate deviations was originally introduced in the context of wavelets by Donoho and Johnstone (1994).

In practice, the degree of correspondence, i.e., overlap in rank positions for an arbitrary object, between ranked lists (full or partial) is not high because of irregular and incomplete rankings due to random and missing assignments. However, the procedure proposed by Hall and Schimek (2012) is specifically designed to deal with such complications. For each combination of two full lists, an estimate $\hat{k}$ for the length of the partial (top-k) list can be obtained via a moderate deviation argument. The probability that an estimator, computed from a pilot sample size v, 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 v, and to be a large deviation if the probability is exponentially small in v.

Let us have 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 $I_j = 0$. Further, let us assume: (i) independent or modestly correlated (m-dependent) Bernoulli random variables $I_1, \ldots, I_N$, with $p_j \geq \frac{1}{2}$ for $1 \leq j \leq j_0 - 1$, $p_{j_0-1} > \frac{1}{2}$, and $p_j = \frac{1}{2}$ for $j \geq j_0$; (ii) a "general decrease" of p_j for increasing j that does not need to be monotone. The index j_0 is the point of degeneration into noise and needs to be estimated ($\hat{j}_0 - 1 = \hat{k}$).

For a pilot sample size v, the values of $z = z_v$ that are associated with moderate deviations are

$$z_v \equiv \left(Cv^{-1} \log v\right)^{1/2}, \tag{2}$$

where $C > 0$ is a constant. The quantities

$$\hat{p}_j^+ = \frac{1}{v} \sum_{n=j}^{j+v-1} I_n \quad \text{and} \quad \hat{p}_j^- = \frac{1}{v} \sum_{n=j-v+1}^{j} I_n \tag{3}$$

represent estimates of p_j computed from the v data pairs I_n for which n lies immediately to the right of j, or immediately to the left of j, respectively.

The constant C is chosen so that z_v in (2) is a moderate-deviation bound for testing the null hypothesis H_0 that $p_k = \frac{1}{2}$ for v consecutive values of k, vs. the alternative H_1 that $p_k > \frac{1}{2}$ for at least one of the values of k. In particular, assuming that H_0 applies to the v consecutive values of k in the respective series at (3), we reject H_0 if and only if $\hat{p}_j^{\pm} - \frac{1}{2} > z_v$. Under H_0, the variance of $\hat{p}_j^{\pm}$ equals $(4v)^{-1}$. This implies a value for the constant $C > \frac{1}{4}$.

The complex inference problem is solved via an iterative algorithm, adjustable for irregularity in the rankings. The overall estimate $\hat{k}^*$ for the ℓ lists τ_l is calculated in the following way. The inference procedure is executed for all possible pairs $L = (\ell^2 - \ell)$ of lists τ_l, thus we obtain L values $\hat{k}_j$ ($j = 1, 2, \ldots, L$). The overall top-k list length is then defined by $\hat{k}^* = \max_j(\hat{k}_j)$, a conservative assumption chosen to avoid loosing any information from the pairwise comparisons.

3.2. *Rank Aggregation of Several Lists*

When we wish to aggregate rank orders in a stochastic manner we need to have an optimization criterion, which itself is specific to the choice of a distance measure. Measures conforming with the so-called generalized Kemeny guidelines are most appropriate for this task (see Dwork et al., 2001). Kendall's τ distance is among them.

Let us have rankings $\tau_1, \tau_2, \ldots, \tau_\ell$ (these are usually from truncated lists) as input. Let $O' = \bigcup_{l=1}^{\ell} O'_l$ and τ be the consensus ranking with respect to O', assuming that the k_l's are fixed. Then our goal is to find an estimate of τ (i.e., an ordered subset of O') that minimizes the sum of weighted distances between τ and each of the lists τ_l. We seek τ^* such that

$$\tau^* = \arg\min_{\tau} \left\{ \sum_{l=1}^{\ell} w_l d(\tau, \tau_l), \tau \subset O' \right\}, \tag{4}$$

where $w = (w_1, w_2, w_3 \ldots, w_\ell)$ is a weight vector that can be used to specify prior information on the relative importance of the input lists, and d is a distance measure. Note again, the ranked lists can be of different lengths and from different spaces.

When $w_l = \frac{1}{\ell}$ for all l and $d(\tau, \tau_l) = K(\tau, \tau_l)$, denoting Kendall's distance measure, the estimate τ^* in Eq. (4) reduces to the Kemeny optimal aggregation (for details see Schimek et al., in press).

The actual computation of the optimal aggregation of full lists of size N, or even partial lists when k is large, constitutes a severe combinatorial problem, as already mentioned. To overcome this obstacle, Markov chain (MC) approaches have been devised (e.g., see DeConde et al., 2006). Consensus rankings (majority preferences) between pairs of items across lists are formed. The assumption that assessors continuously compare pairs of alternatives during their decision process leads naturally to a MC representation. A decision matrix characterizes the potential transitions between alternative decisions. The limiting equilibrium distribution represents the global assessment of all objects. An advantage of MC approaches is that they do not require all the lists to comprise the same objects. A drawback is the associated computational effort.

A more recent approach to solve (4) is cross-entropy Monte Carlo (CEMC) introduced by Rubinstein (1997) for estimating probabilities of rare events in complex stochastic networks and then followed up with complicated combinatorial optimization problems. Several authors have taken advantage of this basic principle for efficient rank aggregation. We prefer the CEMC approach of Lin and Ding (2009) to the one of Pihur et al. (2007) because the former authors have introduced a new Order Explicit Algorithm (OEA) which will be described below. It is motivated by the fact that the orders of the objects in the optimal list are explicitly given in the probability matrix $\boldsymbol{v}$. As a result, Lin and Ding's algorithm is computationally much more efficient, and equally important, it permits modified distance measures as we have considered in this article.

Let us assume a random matrix $\mathbf{X} = (X_{jr})_{N\times k}$ with each component variable X taking the values 0 or 1, and with the constraints of its columns summing up to 1 and its rows summing up to at most 1. This implies that each realization x of $\mathbf{X}$ uniquely determines an ordered list of length k by the position of 1's in each column from left to right. The length k of the aggregated top-k list can be any number not exceeding the size of the union of the full lists, but usually much smaller than N. Let $\boldsymbol{v} = (p_{jr})_{N\times k}$ denote the corresponding probability matrix (each column sums to 1). For each column variable, $\mathbf{X}_r = (X_{1r}, X_{2r}, \ldots, X_{Nr})$, a multinomial distribution with sample size 1 and probability vector $\boldsymbol{v}_r = (p_{1r}, p_{2r}, \ldots, p_{Nr})$ under the constraints of the joint column variables is assumed. Any realization x of $\mathbf{X}$ uniquely determines the corresponding top-k candidate list without reference to the probability matrix $\boldsymbol{v}$. That is, $A = f(\boldsymbol{x}) = (x_{jr} \mid x_{jr} = 1, j = 1, 2, \ldots, N, r = 1, 2, \ldots, k)$. The 1's in each of the k columns make up the top-k list, in that order. Given the 1-to-1-correspondence between A and $\boldsymbol{x}$, finding A^* is equivalent to finding $\boldsymbol{x}^*$ that minimizes $\Phi\{f(\boldsymbol{x})\}$.

Using CEMC, $\boldsymbol{x}^*$ can be obtained by iteratively updating the parameter matrix $\boldsymbol{v}$ such that, iteration by iteration, $P_{\boldsymbol{v}}(\boldsymbol{x})$ will place more and more of its probability mass on the $\boldsymbol{x}$'s that are in the "neighborhood" of $\boldsymbol{x}^*$. Loosely speaking, $\boldsymbol{x}$ is called a neighbor of $\boldsymbol{x}^*$ if the corresponding value of the objective function, $y = \Phi\{f(\boldsymbol{x}; \boldsymbol{v})\}$, is close to the minimum y^*. Let $\boldsymbol{v}$ be the current estimate of the parameter matrix. The next parameter update $\boldsymbol{v}'$ is chosen to minimize the cross entropy $CE(Q^*, P_{\boldsymbol{v}'})$ between the distributions $P_{\boldsymbol{v}'} = P_{\boldsymbol{v}'}(\boldsymbol{x})$ and Q^*, where Q^* is the ideal but unobtainable importance sampling distribution for estimating the rare probability $b = P_{\boldsymbol{v}}[\Phi\{f(\boldsymbol{x};\, \boldsymbol{v})\} \leq y]$,

$$Q^*(\boldsymbol{x}) = \frac{I[\Phi\{f(\boldsymbol{x};\, \boldsymbol{v})\} \leq y] P_{\boldsymbol{v}}(\boldsymbol{x})}{b}.$$

Minimizing $CE(Q^*, P_{\boldsymbol{v}'})$ is equivalent to maximizing

$$\sum_{\boldsymbol{x}} \{I[\Phi\{f(\boldsymbol{x};\, \boldsymbol{v})\} \leq y] \log P_{\boldsymbol{v}'}(\boldsymbol{x})\} P_{\boldsymbol{v}}(\boldsymbol{x}) = E_{\boldsymbol{v}}\left[I[\Phi\{f(\boldsymbol{x};\, \boldsymbol{v})\} \leq y] \log P_{\boldsymbol{v}'}(\boldsymbol{x})\right],$$

which is now free from the probability b to be estimated.

Suppose $\boldsymbol{x}_i = (x_{ijr})_{N\times k}$, $i = 1, 2, \ldots, m$, is a sample drawn from $P_{\boldsymbol{v}}(\boldsymbol{x})$ with the current parameter specification $\boldsymbol{v}$ and the corresponding candidate top-k lists denoted as $\tau_i = f(\boldsymbol{x}_i)$, $i = 1, 2, \ldots, m$. Then,

$$\boldsymbol{v}_{\text{new}} = \arg\max_{\boldsymbol{v}'} \left\{ \frac{1}{m} \sum_{i=1}^{m} I[\Phi\{f(\boldsymbol{x}_i;\, \boldsymbol{v})\} \leq y] \log P_{\boldsymbol{v}'}(\boldsymbol{x}_i) \right\} \tag{5}$$

$$= \left[\frac{\sum_{i=1}^{m} I\{\Phi(\tau_i) \leq y\} x_{ijr}}{\sum_{i=1}^{m} I\{\Phi(\tau_i) \leq y\}} \right]_{j=1,\ldots,N;\, r=1,\ldots,k,} \tag{6}$$

can be used in the update for the next parameter matrix $\boldsymbol{v}'$. In addition, the threshold value y can also be updated iteratively. Equations (5) and (6), respectively, lead to the construction of a sequence, $y_0, y_1, \ldots$, which converges to a value y_∞ close to y^* (Margolin, 2005). Similarly, $\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots$, converges to $\boldsymbol{v}_\infty$, with the corresponding $P_{\boldsymbol{v}_\infty}(\boldsymbol{x})$ placing most of its probability mass on the $\boldsymbol{x}$'s that satisfy $\Phi\{f(\boldsymbol{x};\, \boldsymbol{v})\} \leq y_\infty$ (Lin and Ding, 2009).

A final remark, when aggregating the top-k lists, the distances in the optimization criterion to be computed are those between the candidate aggregate list and each of the input lists τ_l. Suppose the lengths of the input lists k_l are not all the same, then Kendall's τ as well as Spearman's footrule distance need to be scaled (to be independent of the list length). For instance, the maximum feasible distance of Kendall's τ would be a reasonable scaling factor. The same applies to Spearman's footrule.

4. Simulation Evidence

In our approach, the algorithm for inference in step one, as well as the algorithm for integration in step two, require a number of technical and tuning parameters to be set. However, most crucial as input to step two is the data-driven choice of the top-k list lengths of the truncated lists, which also depend on such parameters. The ad-hoc choice of the k's, which has prevailed so far, conflicts with our goal of obtaining an informative ordered subset of objects that is characterized by a high consensus in the rankings across all ℓ lists.

In various simulation experiments, we studied the impact of the parameter choice on the aggregation results. Due to page limitations, we will focus on selected findings.

4.1. *Outline of Simulation Study*

We generated random samples of microarray data based on a 10-vs.-10 experiment (number of replicates in each experiment) of N genes, where the first k genes in 10 replicates are differentially expressed. For an introduction to the statistical analysis of such experiments, see, e.g., McLachlan et al. (2004).

We simulated the microarray data as follows: a data matrix of N rows and 2×10 columns was randomly generated, each row representing one gene and each column representing one subject. All 2×10 subjects are random samples from a Normal distribution with mean $\mu_0 = 0$ and standard deviation $\sigma_0 = 1$ (following Peng et al., 2003). Additionally, to the first $\frac{k}{2}$ positions of the last 10 subjects, Normal random errors with different mean values $\mu_1 = 0.5, 0.75, 1, 1.25, 1.5$ and standard deviation $\sigma_1 = 1$, and to the positions $(\frac{k}{2} + 1), \ldots, k$, Normal random errors with mean values $\mu_2 = -0.5, -0.75, -1, -1.25, -1.5$ and standard deviation $\sigma_2 = 1$, were added. The purpose of this was to simulate upregulated and downregulated genes in a microarray experiment. The combination of those, in absolute value identical, μ_1 and μ_2, generated five settings of increasing effect size.

The SAM package ("Significance analysis of microarrays") was adopted to the simulated microarray data to obtain ranked lists of expressed genes based on their p-values. SAM is available in the R library samr (for details see, Tusher et al., 2001). It was applied 5 times on 5×20 different subjects to obtain $\ell = 5$ different ranked lists of N genes.

Finally, for the generation of a complete artificial data set per effect size, the whole simulation procedure was executed 100 times resulting in 100 simulations of 5 ranked lists of length N.

All tuning parameters, apart from the pilot sample size v, in the inference and integration procedures were fixed in compliance with Hall and Schimek (2012) and Lin and Ding (2009) (default settings). The pilot sample size v plays the role of a smoothing parameter in moderate deviation-based testing and is therefore critical with respect to the calculation of $\hat{j}_0$. We considered values $v \in [2, 6, 10, 14, \ldots, 98]$. The simulations were performed for distance values $\delta \in [0, 4, 8, 12, \ldots, 40]$, where $\delta = 20$ is most adequate for these data, and both distance measures, Kendall's τ and Spearman's footrule. The practice of δ-choice is beyond the scope of this article (for details, see, Schimek and Budinská, 2010; Hall and Schimek, 2012).

4.2. *Simulation Study Results*

The following results of our simulation study are given for $N = 100$ and $k = 10$. We display two types of plots, one for the estimation of the overall top-k list length (defined as the maximum from all $L = 10$ pairwise comparisons of rankings), and the other for the aggregation of the five lists applying Kendall's τ and Spearman's footrule.

Figures 1 and 2 show series of boxplots of estimated $\hat{k}$'s, plotted for various values of the pilot sample size v and one selected distance $\delta = 20$. Figure 1 represents the cases of small effect size. Its top graph corresponds to the estimates for means $\mu_1 = 0.5$ and $\mu_2 = -0.5$, the middle one to the estimates for means $\mu_1 = 0.75$ and $\mu_2 = -0.75$, and the bottom one to the estimates for means $\mu_1 = 1$ and $\mu_2 = -1$. The dashed horizontal line represents the true value of k. The obtained estimates for moderate effect size are summarized in Fig. 2 (has the same outline as Fig. 1).

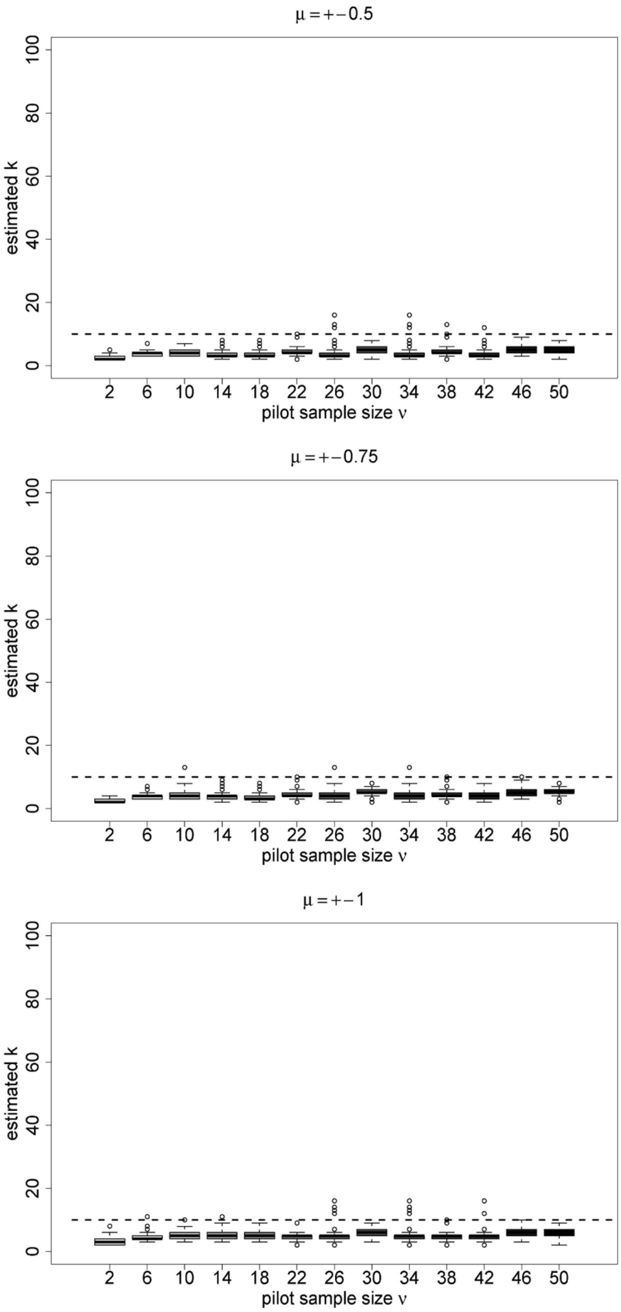

Figure 1. Boxplots of estimates $\hat{k}$ for the simulated data ($N = 100$ objects, true $k = 10$) under $\delta = 20$ and $\nu = 2, 6, 10, \ldots, 50$; top: $\mu_1 = 0.5$, $\mu_2 = -0.5$, middle: $\mu_1 = 0.75$, $\mu_2 = -0.75$, and bottom: $\mu_1 = 1$, $\mu_2 = -1$ (small effect sizes).

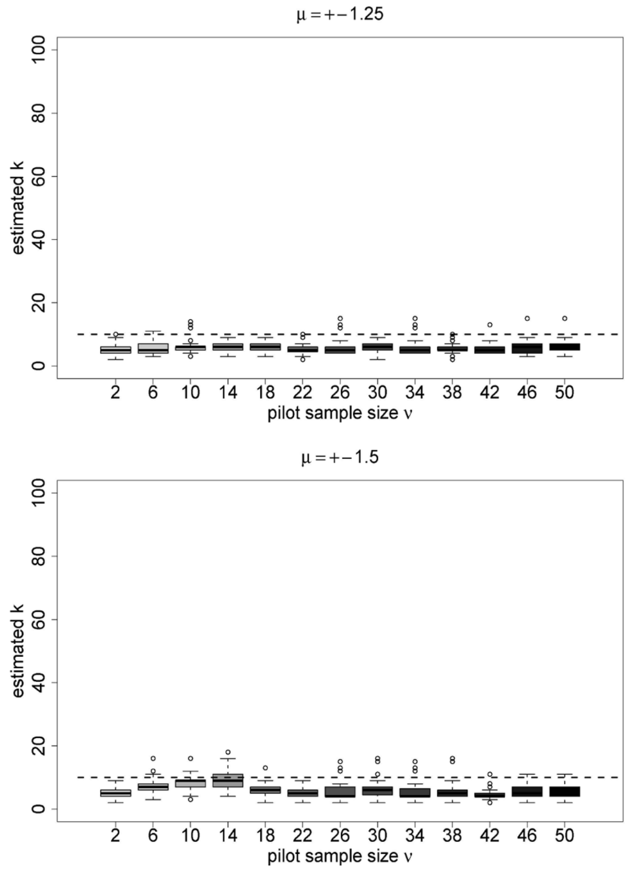

Figure 2. Boxplots of estimates $\hat{k}$ for the simulated data ($N = 100$ objects, true $k = 10$) under $\delta = 20$ and $v = 2, 6, 10, \ldots, 50$; top: $\mu_1 = 1.25$, $\mu_2 = -1.25$, and bottom: $\mu_1 = 1.5$, $\mu_2 = -1.5$ (moderate effect sizes).

The top graph corresponds to $\mu_1 = 1.25$ and $\mu_2 = -1.25$, and the bottom graph to $\mu_1 = 1.5$ and $\mu_2 = -1.5$.

We can observe that a small effect size in the simulated data goes hand in hand with a slight underestimation of the true value of k, hardly influenced by the choice of the pilot sample size v unless chosen far too large. This underestimation owes to the fact that some of the differentially expressed genes are not present in the top section of the ranked list due to randomness. However, genes sampled from the standard Normal can be positioned among the top rankings. When considering the estimated $\hat{k}^*$ from the simulated data with larger effect size, the precision is much higher. The results are stable for a wide range of pilot sample sizes v. For the strongest effect ($\mu_1 = 1.5$ and $\mu_2 = -1.5$), the parameter $v = 14$ seems to be the best overall choice. In this optimal situation, 54% of all estimated $\hat{k}^*$ lie in the interval $[8, 11]$. In general, we can claim that for short lists and a small true value of k, the selection of a rather small value of v is appropriate. When random samples of artificial microarray data are constructed as in this simulation study, irregularities occur in the rankings, and a perfect reconstruction of the true top-k list length is not feasible. However, for an adequate value of δ and a parameter v not too far from its optimum, the inference procedure can accomplish estimates close to the true k.

Figures 3 and 4 concern the aggregation results for the five truncated input lists when Kendall's τ and Spearman's footrule are applied. From both figures, one can conclude that the first 10 (truly differentially expressed) genes are those that

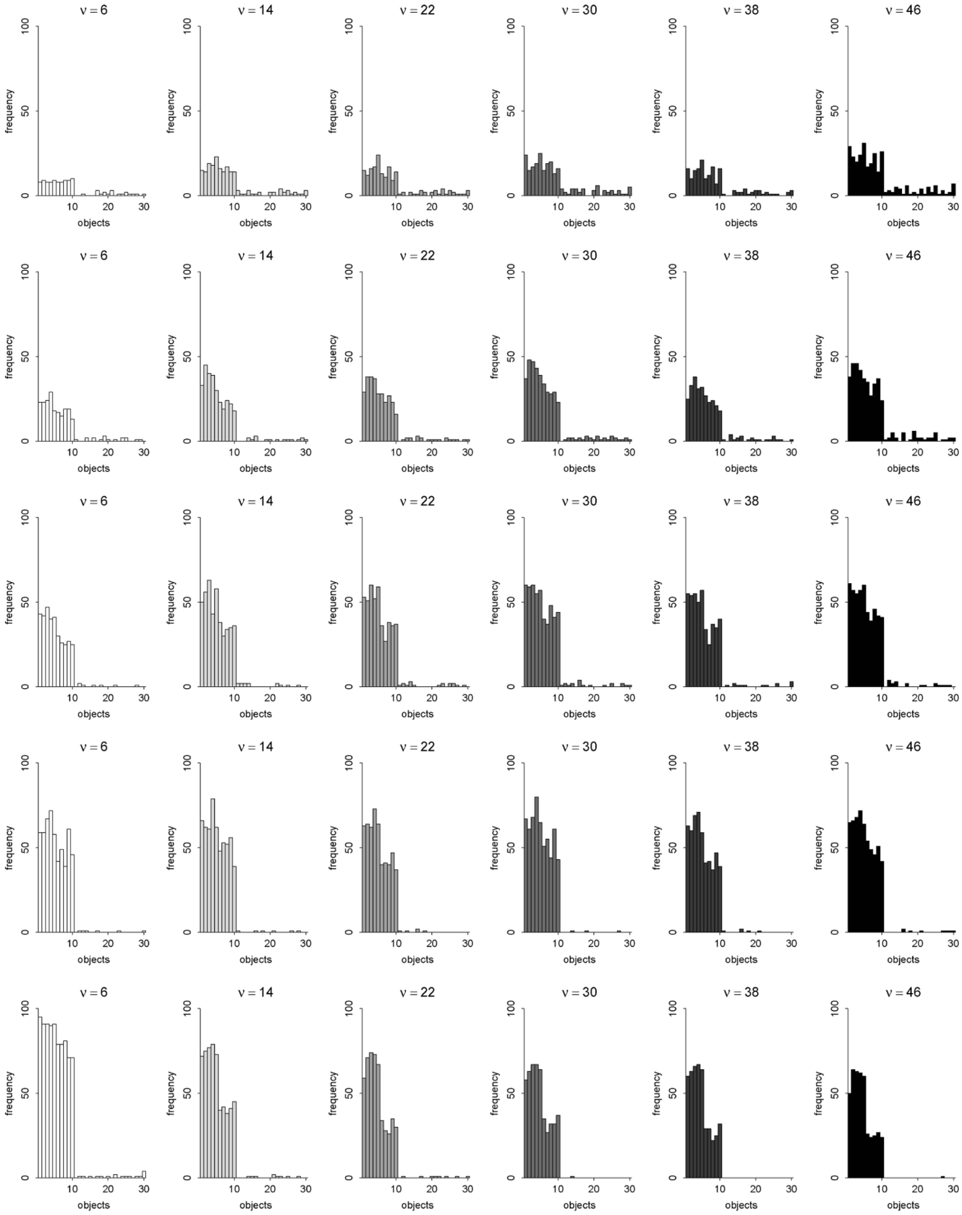

Figure 3. Aggregation results for Kendall's τ under increasing effect size from top ($\mu_1 = 0.5$, $\mu_2 = -0.5$) to bottom ($\mu_1 = 1.5$, $\mu_2 = -1.5$) – frequency of appearance of the first 30 objects (of 100) in the simulated data for $\delta = 20$ and $v = 6, 14, 22, 30, 38, 46$.

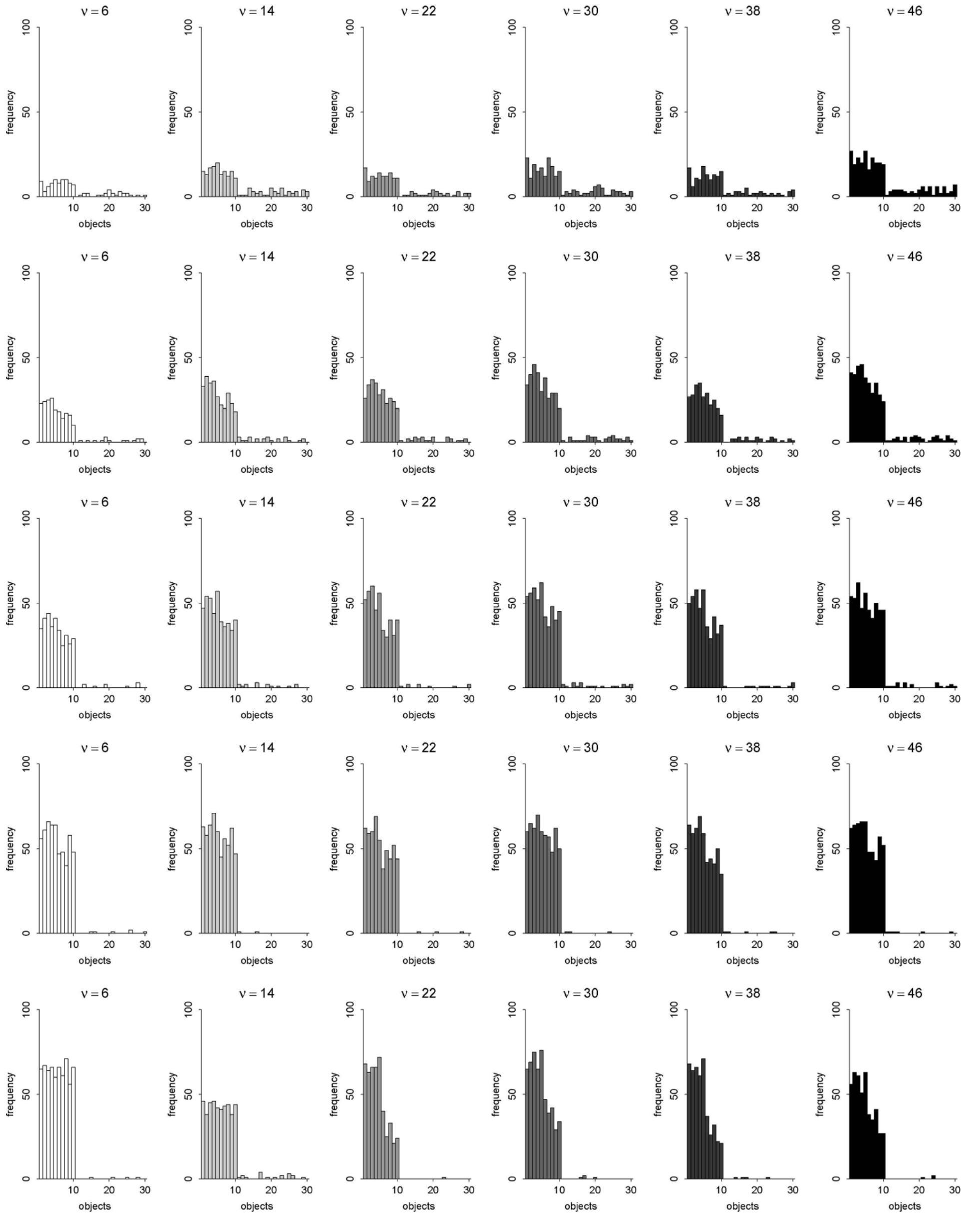

Figure 4. Aggregation results for Spearman's footrule under increasing effect size from top ($\mu_1 = 0.5$, $\mu_2 = -0.5$) to bottom ($\mu_1 = 1.5$, $\mu_2 = -1.5$) – frequency of appearance of the first 30 objects (of 100) in the simulated data for $\delta = 20$ and $\nu = 6, 14, 22, 30, 38, 46$.

are most frequently identified as part of the top-k list. Even for the smallest effect size (first row of histograms in Figs. 3 and 4), the frequency to appear in the top-k list is significantly higher for the first 10 genes than for the remaining ones. With increasing effect size, this frequency goes up compared to the rest of the gene set (only the first 30 genes are displayed in Figs. 3 and 4). The obtained results are consistent for a wide range of pilot sample sizes v. For cases of small effect size, larger values of v perform better, whereas for larger effect sizes, the whole interval $v \in [6, 46]$ is adequate. Let us finally look into potential differences between the aggregation results due to the adoption of Kendall's τ versus Spearman's footrule. The direct comparison of the corresponding histograms in Figs. 3 and 4 makes it clear: there are only minor differences with respect to the applied distance measure. Kendall's τ tends to produce higher frequencies for the reconstruction of the top 10 objects compared to Spearman's footrule, the former occasionally favoring the first 3–5 of the top list. In summary, all obtained results (histograms) demonstrate a perfect separation between the 10 true top-ranked objects and the remainder for a wide range of pilot sample sizes v, independently of the applied distance measure.

5. Conclusion

In various applications, for example in molecular biology, the consolidation of full ranked lists is neither desirable with respect to the research goal, nor feasible because of the high dimensionality of the data. Until now there has been a lack of statistical procedures to deal with several lists of a dimension of hundreds or even thousands of ranked objects. In this article, we have introduced an approach that is powerful enough to handle multiple ranked lists of arbitrary length, while avoiding combinatorial complexity. Firstly, it allows us to truncate the full lists in a data-driven manner, yielding partial lists, and secondly to stochastically integrate them into one aggregated list of objects in a new consolidated rank order.

The involved inference procedure is based on all possible pairwise assessments and is designed to deal with irregular and incomplete rankings. The integration procedure takes advantage of a modification of Kendall's τ or of Spearman's footrule distance measure, allowing for missing rank information resulting from the truncation of the full lists.

We have illustrated the features and advantages of the new approach on simulated data in the context of current microarray-based research. However, the described methodology could be applied in many other areas such as consumer preference research or in the consolidation of Web search engine results. Last, but not least, it should be mentioned that all calculations were carried out with the β-version of the `R` package `TopKLists` developed by the first author and collaborators.

Acknowledgments

We are grateful to Shili Lin (Ohio State University, USA) for the `R`-version of her package `TopKCEMC`, and to Gig Searle (Medical University of Graz, Austria) for her critical reading of our manuscript. We acknowledge the feedback of three anonymous referees which lead to a substantial improvement of the article.

References

Cohen, W. C., Schapire, R. S., Singer, Y. (1999). Learning to order things. *Journal of Artificial Intelligence Research* 10:243–270.

Deconde, R. P., Hawley, S., Falcon, S., Clegg, N., Knudsen, B., Etzioni, R. (2006). Combined results of microarray experiments: a rank aggregation approach. *Statistical Applications Genetics and Molecular Biology* 5(1):Article 15.

Donoho, D. L., Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81:425–455.

Dwork, C., Kumar, R., Naor, M., Sivakumar, D. (2001). Rank aggregation methods for the web. http://www10.org/ cdrom/papers/577/.

Fagin, R., Kumar, R., Sivakumar, D. (2003). Comparing top-*k* lists. *SIAM Journal of Discrete Mathematics* 17:134–160.

Fligner, M. A., Verducci, J. S. (1986). Distance based ranking models. *Journal of the Royal Statistical Society B* 48:359–369.

Hall, P., Schimek, M. G. (2012). Moderate deviation-based inference for random degeneration in paired rank lists. To appear in *Journal of American Statistical Association*.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika* 30:91–93.

Lin, S. (2010). Space oriented rank-based integration. *Statistical Applications in and Genetics and Molecular Biology* 9(1):Article 20.

Lin, S., Ding, J. (2009). Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA studies. *Biometrics* 65:9–18.

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

Marden, J. I. (1995). *Analyzing and Modeling Rank Data*. London: Chapman & Hall.

Margolin, L. (2005). On the convergence of the cross-entropy method. *Annal. Operations Res.* 134:201–214.

McLachlan, G. J., Do, K.-A., Ambroise, C. (2004). *Analyzing Microarray Gene Expression Data*. Hobokin, NJ: Wiley.

Peng, X., Wood, C. L., Blalock, E. M., Chen, K. C., Landfield, P. W., Stronberg, A. J. (2003). Statistical implications of pooling RNA samples for microarray experiments. *BMC Bioinformatics* 4(26):1471–2105.

Pihur, V., Datta, S., Datta, S. (2007). Weighted rank aggregation of cluster validation measures: A Monte Carlo cross-entropy approach. *Bioinformatics* 23:1607–1615.

R Development Core Team. (2009). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Rubinstein, R. Y. (1997). Optimization of computer simulation models with rare events. *Europian Journal of Operational Research* 99:89–112.

Schimek, M. G., Budinská, E. (2010). Visualization techniques for the integration of rank data. *COMPSTAT 2010. Proceedings in Computational Statistics*. Heidelberg: Physica, pp. 1637–1644.

Schimek, M. G., Lin, S., Wang, N. (in press). *Statistical Integration of Omics Data*. New York: Springer.

Spearman, C. (1906). A footrule for measuring correlation. *British Journal of Psychology* 2:89–108.

Tusher, V., Tibshirani, R., Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. of the National Academy of Sciety USA* 98:5116–5121.

Yang, X., Bentink, S., Scheid, S., Spong, R. (2006). Similarities of ordered gene lists. *Journal of Bioinformatics and Computational Biology* 4:693–708.