Chilean Journal of Statistics
Vol. 2, No. 1, April 2011, 3–20

Asymptotic Statistics
Research Paper

Spearman's footrule: Asymptotics in applications

Pranab K. Sen[1,*], Ibrahim A. Salama[2] and Dana Quade[1]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, USA

[2]School of Business, North Carolina Central University, USA

(Received: 10 January 2011 · Accepted in final form: 25 March 2011)

Abstract

Spearman's footrule is a well known measure of disarray for ranked data that recently has found applications in many areas of research. Exact, computational, and asymptotic properties of the measure are discussed using a unified approach utilizing a Markovian property with an inherited martingale structure, extension to the partial rankings case is discussed. A method allowing the generation of weighted versions is introduced, and asymptotic normality is established.

Keywords: Asymptotic normality · Markov chain · Martingale · Partial rankings · Top list · Weighted measure of correlation.

Mathematics Subject Classification: Primary 60G42 · Secondary 62G99.

1. Introduction

Spearman's (1906) footrule is a well known measure of disarray for ranked data (see Diaconis and Graham, 1977) that recently has found applications in several areas of research, including aggregate rankings for search engines, bioinformatics, genomics, information science, litigation and management science; see Berman (1996), Kim et al. (2004), Fagin et al. (2006), Pihur et al. (2007, 2008), Bar-Ilan et al. (2007), Sufyan and Ahmed (2007), Jurman et al. (2008), Lin and Ding (2009), Lee and Yu (2010), and Powell and Reinhardt (2010). The measure has been extended to cases where we have incomplete, censored or multivariate data; see Critchlow (1985), Alvo and Charbonneau (1997), Sen et al. (2003), Salama and Quade (2004), Ubeda-Flores (2005), and Quade and Salama (2006). We also refer to the recent survey by Genest et al. (2010).

Beside theoretical interest, incomplete rankings (censored rankings) appear naturally in ranking the top k out of n objects. This is the case when a search engine responds to a search query. Out of a million items (related to the query), the search engine reports the top (say) 100 items. The search engine may report more, but (for all practical purposes) only the top 100 are of interest. The problem also appears in comparing the results of different search engines. Again, each search engine reports the top (say) 100 items, and we would like to find a measure that may be used to study if these results are "consistent" with each other (or, if we have an internal agreement among the several reported results).

[*]Corresponding author. Pranab K. Sen. Department of Biostatistics, University of North Carolina at Chapel Hill, 3101 McGavran-Greenberg Hall, CB # 7420, Chapel Hill, NC, 27599-7420, USA. Email: pksen@bios.unc.edu.

ISSN: 0718-7912 (print)/ISSN: 0718-7920 (online)

http://www.soche.cl/chjs

Two measures are widely used in this situation: Kendall's tau and Spearman's footrule. Most of the works (so far) have used these measures as "deterministic" measures, meaning comparison is done based on the actual value of the measure (not much about probability).

Another important problem where Spearman's footrule is used is in the determination of the top k genes (out of n genes, where k is relatively small and n is large). It is also used to study the reproducibility of microarray experiments (Kim et al., 2004), and to address stability issues (Boulesteix and Slawski, 2009), meaning if the produced "rankings" (of the objects under consideration) will remain approximately the same after small perturbations in the data. Another problem that uses Spearman's footrule is the aggregation of several ranks or partial ranks to produce a final list, in which Spearman's footrule and weighted versions of it are used; see Dwork et al. (2001), Fagin et al. (2003), Fagin et al. (2004), Pihur et al. (2009), and Lin (2010). Weighted versions for Spearman's rank correlation and Kendall's tau have been discussed in Salama and Quade (1982), Quade and Salama (1992), Mango (1997), Shieh (1998), Blest (2000), Costa et al. (2001), Costa and Soares (2005, 2007), Genst and Plante (2007), and Tarsitano (2009).

The exact distribution of Spearman's footrule is discrete in nature. This may be efficiently generated (under the assumptions of independence and uniformity) using the algorithm introduced in Salama and Quade (2002), which may be reasonable to use in the case of a moderate number of objects. But, in many applications, the number of objects being ranked (or considered for ranking) becomes increasingly large (tens of thousands), hence, there is a practical need for asymptotics. Diaconis and Graham (1977) provided a proof for the asymptotic normality using Hoeffding's (1951) combinatorial limit theorem. Another approach was considered by Sen and Salama (1983) and extended by Sen et al. (2003) to study the asymptotic normality for the partial rankings case, using a Markovian property, with an inherited martingale structure.

This paper is organized as follows. In Section 2, we discuss Spearman's footrule for full ranks (permutations), providing a representation as a linear combination of a Markov chain. This representation allows for an algorithm computing the exact distribution (of Spearman's footrule) in $O(n^4)$ time. This also conduces to a martingale structure leading to (a more general) limit theorem. In Section 3, we extend the results to the partial rankings case. In Section 4, we include weighted versions of the measure using the same framework considered in Section 3. We conclude by some remarks in Section 5.

2. The Full Ranking Case

In this section, we consider an equivalent representation of a metric known as Spearman's footrule with an inherited Markovian structure. This structure greatly facilitates the study of the exact and asymptotic properties of such a metric.

Let S_n be the set of all ($n!$) permutations of the first n integers $\{1, \ldots, n\}$. As in Diaconis and Graham (1977), we may define Spearman's footrule (D_n) on S_n by

$$D_n(\pi_n, \sigma_n) = \sum_{i=1}^{n} |\sigma_n(i) - \pi_n(i)|, \tag{1}$$

where $\sigma_n = (\sigma_n(1), \ldots, \sigma_n(n))$ and $\pi_n = (\pi_n(1), \ldots, \pi_n(n))$ are elements of S_n. The relationships of D_n with other commonly used non-parametric measures of association (such as Kendall tau and Spearman rho) and its asymptotic normality (under the assumption that σ_n and π_n are chosen independently and distributed uniformly in S_n) have been studied by Diaconis and Graham (1977).

2.1 Representation

First, we assume (without loss of generality) that $\pi_n = (1, \ldots, n)$, the identity permutation. Then, we may write

$$D_n(\sigma_n) = D_n(1_n, \sigma_n) = \sum_{i=1}^{n} |i - \sigma_n(i)|.$$

Second (as in Salama and Quade, 1982), for $\sigma_n \in S_n$, and for $j = 1, \ldots, n$, let

$$T_{n,i} = T_{n,i}(\sigma_n) = T_{n,i}(1_n, \sigma_n) = \sum_{j=1}^{i} I(\sigma_n(j) \leq i)$$

and

$$T_n = T_n(\sigma_n) = T_n(1_n, \sigma_n) = \sum_{i=1}^{n} T_{n,i}.$$

Then (see Sen and Salama, 1983) we have the following representation. For every $\sigma_n \in S_n$ and n (≥ 1), we have

$$T_n(\sigma_n) + \frac{1}{2} D_n(\sigma_n) = \frac{n(n+1)}{2}.$$

2.2 Markovian property

The importance of this representation lies in the following theorem.

Theorem 2.1 For every n (≥ 1), whenever σ_n is distributed uniformly on S_n and $\{T_{n,i}\colon i \leq n\}$ is a Markov chain, i.e., for every k $(\leq n-1)$ and $0 \leq r_1 \leq \cdots \leq r_k \leq r_{k+1}(\leq n)$,

$$\mathrm{P}(T_{n,k+1} = r_{k+1} | T_{n,j} = r_j, j \leq k) = \mathrm{P}(T_{n,k+1} = r_{k+1} | T_{n,k} = r_k).$$

Proof Let P be the set of all permutations (of $\{1, \ldots, n\}$) satisfying the condition $\{T_{n,1} = r_1, \ldots, T_{n,k} = r_k\}$. It is easy to see that for any $\sigma \in \mathrm{P}$, $T_{n,k+1}$ can only assume the values r_k, $r_k + 1$ and $r_k + 2$. If $(\sigma(1), \ldots, \sigma(n)) \in \mathrm{P}$, then among the set $\{\sigma(1), \ldots, \sigma(k)\}$, we have $k - r_k$ elements of the set $\{k+1, \ldots, n\}$. If we denote this set by A and its complement by A^c, then, we may have either of the following:

(i) $k + 1 \in A$. This happens with the (conditional) probability

$$\frac{\binom{n-k-1}{n-r_k-1}}{\binom{n-k}{k-r_k}} = \frac{(k - r_k)}{(n-k)}$$

and

(ii) $k + 1 \in A^c$. This happens with the (conditional) probability

$$1 - \frac{(k - r_k)}{(n-k)} = \frac{(n - 2k + r_k)}{(n-k)}.$$

In case (i), $T_{n,k+1}$ can assume only the values r_k+1 or r_k+2 with probabilities $[(n-k)-(k-r_k)]/(n-k)$ and $(k-r_k)/(n-k)$, respectively, while in case (ii), $T_{n,k+1}$ can assume only the values r_k and r_{k+1} with respective probabilities $[(n-k)-(k-r_k)-1]/(n-k)$ and $[(k-r_k)+1]/(n-k)$. Thus, the assumed values of $T_{n,k+1}$ (r_k,r_k+1 and r_k+2) and their respective (conditional) probabilities (given $T_{n,i}$, for $i \le k$) depend only on the value r_k assumed by $T_{n,k}$. ■

By a similar (elementary) argument to that of Theorem 2.1, we have the following lemma.

Lemma 2.2 Let σ_n to have a uniform distribution on S_n. Then, for every k and r

$$\mathrm{P}(T_{n,k}=r)=\frac{\binom{k}{r}\binom{n-k}{k-r}}{\binom{n}{k}}, \quad r=\max\{0,(2k-n)\},\ldots,k, \quad 1\le k\le n,$$

and, for every $k<q$ and $r\le s$,

$$\mathrm{P}(T_{n,k}=r,T_{n,q}=s)=\frac{\binom{n-q}{q-s}\sum_{u\ge r}\binom{k}{u}\binom{q-k}{s-u}\binom{u}{r}\binom{q-u}{k-r}}{n![k!(q-k)!(n-q)!]^{-1}}.$$

Based on results of Lemma 2.2, we have the following lemma.

Lemma 2.3

$$\mathrm{P}(T_{n,k+1}=s|T_{n,k}=r)=\begin{cases}\frac{(n-2k+r)(n-2k+r-1)}{(n-k)^2}, & s=r;\\ \frac{(n-2k+r)(2k-2r+1)}{(n-k)^2}, & s=r+1;\\ \frac{(k-r)^2}{(n-k)^2}, & s=r+2;\\ 0, & s\ge r+3 \text{ or } s<r.\end{cases}$$

Hence, for $k=0,\ldots,n-1$ (letting $T_{n,0}=0$),

$$\mathrm{E}(T_{n,k+1}|T_{n,k})=\frac{(n-k-1)^2}{(n-k)^2}T_{n,k}+\frac{2k+1}{n-k}-\frac{k}{(n-k)^2}.$$

From Lemma 2.3, we also

$$\mu_{n,k}=\mathrm{E}(T_{n,k})=\frac{k^2}{n}, \tag{2}$$

$$\gamma^2_{n,k}=\mathrm{Var}\,(T_{n,k})=\frac{k^2(n-k)^2}{n^2(n-1)}, \tag{3}$$

and

$$\gamma_{n,kq}=\mathrm{Cov}\,(T_{n,k},T_{n,q})=\frac{k^2(n-q)^2}{n^2(n-1)}, \quad k\le q. \tag{4}$$

Based on Equations (2)-(4), we have $\mu_n=\mathrm{E}(T_n)=n(2n+1)/6$ and $\gamma^2_n=\mathrm{Var}\,(T_n)=(n+1)(2n^2+7)/180$. We also write $\nu_{n,k}=\mathrm{E}(T^2_{n,k})=\gamma^2_{n,k}+\mu^2_{n,k}$, for $0\le k\le n$.

2.3 Exact distribution

The exact distribution of D_n given in Equation (1), –assuming σ and π are chosen independently at random from S_n– may be obtained by complete enumeration of all $n!$ elements of S_n. Ury and Kleinecke (1979) tabulated this distribution for $n = 2(1)10$, also providing a Monte Carlo approximation for $n = 11(1)15$. By modifying the complete enumeration approach, Franklin (1988) extended the exact tables to $n = 18$. An intrinsically different approach was used by Salama and Quade (1990) to further extend the tables to $n = 40$. Salama and Quade (2002) showed that the algorithm used can compute the exact distribution in a $O(n^4)$ polynomial time. We outline the algorithm as follows:

(i) For $1 \leq i \leq n$, let $L_i = i - T_i$, and reexpressed D_n as

$$\frac{D_n}{2} = \frac{n(n+1)}{2} - \sum_{i=1}^{n} T_i = \sum_{i=1}^{n} (i - T_i) = \sum_{i=1}^{n} L_i.$$

(ii) By direct substitution in first equation of Lemma 2.3, consider

$$\mathrm{P}(L_{i+1} = l_{i+1} | L_i = l_i) = \begin{cases} \frac{[(n-i)-l_i][(n-i)-l_i-1]}{(n-i)^2}, & l_{i+1} = l_i + 1; \\ \frac{[(n-i)-l_i](2l_i+1)}{(n-i)^2}, & l_{i+1} = l_i; \\ \frac{l_i^2}{(n-i)^2}, & l_{i+1} = l_i - 1. \end{cases}$$

Computations of the exact distribution of D_n is easier using the sequence $\{L_i\}$.

(iii) Let $\mathrm{P}_k^{(n)}$ be the matrix of dimension $(d_1(n,k), d_2(n,k))$ whose (X, Y) element is

$$\mathrm{P}_k^{(n)}[X][Y] = \mathrm{P}\left(\sum_{i=1}^{k} L_i = X, L_k = Y \right),$$

where $0 \leq k < n$. For every k and n, the minimum values of X and Y are both 0, assumed if (for example) $\sigma = (1, \ldots, n)$ and hence every $L_i = 0$. The maximum values occur if (for example) $\sigma = (n, n-1, \ldots, 1)$. In this case, if $k \leq n/2$, then $X = k(k+1)/2$ and $Y = k$. If $k > n/2$, then $Y = n - k$; but $X = (n^2/4) - (1/2)(n - k)(n - k - 1)$ if n is even and is smaller by 1/4 if n is odd. Thus for n even and odd, we have respectively

$$d_1(n,k) = \begin{cases} \frac{k(k+1)}{2} + 1, & \text{if } k \leq \frac{n}{2}; \\ \frac{n^2+4}{4} - \frac{1}{2}(n-k)(n-k-1), & \text{if } k > \frac{n}{2}, \end{cases} \quad \text{and}$$

$$d_1(n,k) = \begin{cases} \frac{k(k+1)}{2} + 1, & \text{if } k < \frac{n}{2}; \\ \frac{n^2+3}{4} - \frac{1}{2}(n-k)(n-k-1), & \text{if } k > \frac{n}{2}, \end{cases} \quad \text{while}$$

$$d_2(n,k) = \begin{cases} k + 1, & \text{if } k \leq \frac{n}{2}; \\ n - k + 1, & \text{if } k \geq \frac{n}{2}. \end{cases}$$

(iv) Obtain $\mathrm{P}_{k+1}^{(n)}$ from $\mathrm{P}_k^{(n)}$ using the double loop described in the Appendix and the algorithm for getting $\mathrm{P}_n^{(n)}$ is as follows:

(a) Let $\mathrm{P}_0^{(n)}$ be a 1×1 matrix with $\mathrm{P}_0^{(n)}[0][0] = 1$.

(b) For $k = 0, \ldots, n-1$, get $\mathrm{P}_{k+1}^{(n)}$ from $\mathrm{P}_k^{(n)}$.

2.4 Asymptotic normality

The advantage of the approach of Sen and Salama's (1983) using the Markovian property and the associated martingale structure is that it allows for a random sample of size n and it is easily extended to cover the censored ranks situation. We present a brief outline. We consider the standardized version $T_n^* = (T_n - \mathrm{E}(T_n))/\gamma_n$. For $1 \leq k \leq n-1$, we let $d_{nk} = [(n-k+1)(2k-1)-(k-1)]/[(n-k)(n-k+1)]^2$, and $d_{nk}^* = \sum_{j=1}^{k} d_{nk}$. If we write $Y_{nk} = T_{n,k}/(n-k)^2 - d_{nk}^*$, $B_{nk} = B(T_{nj}; j \leq k)$, then, for $k = 0, \ldots, n$, we have $\mathrm{E}(Y_{nk}|B_{n,k-1}) = Y_{n,k-1}$, for $1 \leq k \leq n-1$. Now, if we let $Z_{nk} = Y_{nk} - Y_{n,k-1}$, for $1 \leq k \leq n-1$, then the Z_{nk}'s are martingale differences, and we may write

$$T_n - \mathrm{E}(T_n) = \sum_{k=1}^{n-1} \left[\frac{1}{6}(n-k)(n-k+1)(2n-2k+1) \right] Z_{nk}.$$

If we let $c_{ni} = (n-i)(n-i+1)(2n-2i+1)/6[(n+1)(2n^2+7)/180]^{1/2}$ and $U_{ni} = c_{ni} Z_{ni}$ (setting $U_{n0} = 0$), for $1 \leq i \leq n-1$, then we have $T_n^* = \sum_{k=1}^{n-1} U_{nk}$, for $1 \leq k \leq n-1$, $\mathrm{E}(U_{nk}|B_{n,k-1}) = 0$, and $\sum_{i=1}^{n} \mathrm{E}(U_{ni}^2) = 1$, for $n \geq 1$.

Thus, T_n^* relates to a martingale array normalized by $\sum_{i=1}^{n} \mathrm{E}(U_{ni}^2) = 1$. To establish the asymptotic normality we need only to verify the following:

$$U_n^* = \sum_{i=1}^{n-1} U_{ni}^* = \sum_{i=1}^{n-1} \mathrm{E}(U_{ni}^2|B_{n,i-1}) \xrightarrow{\mathrm{p}} 1, \tag{5}$$

and, for every $\epsilon > 0$,

$$\sum_{i=1}^{n} \mathrm{E}[U_{ni}^2 I(|U_{ni}| > \epsilon)|B_{n,i-1}] \xrightarrow{\mathrm{p}} 0, \tag{6}$$

where $\xrightarrow{\mathrm{p}}$ denotes convergence in probability. Noting that $T_{n,k}$ are non-negative, it can be easily seen (for $1 \leq i \leq n-1$) that $|U_{ni}| \leq Cn^{-1/2}$ with probability 1, where C ($0 < C < +\infty$) does not depend on n. That ensures that Equation (6) holds for n adequately large. Further, since $\sum_{i=1}^{n} \mathrm{E}(U_{ni}^2) = 1$, to prove Equation (5), it suffices to show that $\mathrm{E}(U_n^* - 1)^2 \rightarrow 0$ as $n \rightarrow \infty$. Towards this, note that, for $1 \leq i \leq n-1$,

$$U_{ni}^* = \mathrm{E}(U_{ni}^2|B_{n,i-1}) = a_{ni}T_{n,i-1}^2 + b_{ni}T_{n,i-1} + g_{ni},$$

where $a_{ni} = O(n^{-3})$, $b_{ni} = O(n^{-2})$ and $g_{ni} = O(n^{-1})$, $1 \leq i \leq n-1$. Furthermore,

$$U_n^* - 1 = U_n^* - \mathrm{E}(U_n^*) = \sum_{i=1}^{n-1} [a_{in}(T_{n,i-1}^2 - \nu_{n,i-1}) + b_{ni}(T_{n,i-1} - \mu_{n,i-1})].$$

Noting that (for $1 \leq i \leq n-1$) $\mathrm{E}(T_{n,i-1}^2 - \nu_{n,i-1})^2 = O(n^3)$, $\mathrm{E}(T_{n,i-1} - \mu_{n,i-1})^2 = O(n)$, $\mathrm{E}(T_{n,i-1}^2 - \nu_{n,i-1})(T_{n,i-1} - \mu_{n,i-1}) = O(n^2)$ and $\mathrm{E}(T_{n,i-1} - \mu_{n,i-1})(T_{n,j-1} - \mu_{n,j-1}) = O(n)$, we obtain $\mathrm{E}(\bar{U}_n - 1)^2 = O(n^{-1})$, and Equation (5) holds.

3. Ranking the Top k Objects

In this section, we discuss an extension of Spearman's footrule to the partial rankings problem, i.e., ranking the top k items out of n objects.

Let σ be a permutation of S_n (the set of all $n!$ permutations of $\{1,\ldots,n\}$). Consider the right-censored case in which, for some $k = 1,\ldots,n$, we observe $\sigma_k = (\sigma_{1k},\ldots,\sigma_{nk})$, where

$$\sigma_{ik} = \sigma_i I(\sigma_i \leq k) + (k+\delta_k) I(\sigma_i > k),\ i = 1,\ldots,n,$$

with $I(B)$ being the indicator function od the set B, and the non-stochastic non-negative constant δ_k may be made to depend on k (which is determined as 1 in 3.2 below). Note that all censored ranks are greater than k. Hence, δ_k must be bigger or equal than 1. This situation corresponds to projecting S_n onto subspaces $S_n^{(k)}$ such that $S_n^{(0)} = 0 \subseteq S_n^{(1)} \subseteq \cdots \subseteq S_n^{(n)} = S_n$. For our subsequent analysis, we take π as the identity permutation $\pi = (1,\ldots,n)$, so that we may define $\pi_k = (\pi_{1k},\ldots,\pi_{nk})$ by

$$\pi_{ik} = iI(i \leq k) + (k+\delta_k) I(i > k),\ i = 1,\ldots,n.$$

With this notation, for any element σ_k of $S_n^{(k)}$, we define (without any loss of generality)

$$D_{nk}(\sigma_k) = D(\sigma_k,\pi_k) = \sum_{i=1}^{n} |\sigma_{ik} - \pi_{ik}|,\ k = 1,\ldots,n.$$

3.1 Representation

We have the following theorem due to Sen et al. (2003).

Theorem 3.1 For every $\sigma \in S_n$ and $k = 0,\ldots,n$,

$$D_{n,k}(\sigma) = 2\sum_{i=1}^{k-1} L_i(\sigma) + 2\delta_k L_k(\sigma). \tag{7}$$

3.2 Optimal choice of δ_k

Theorem 3.1 shows that $D_{n,k}(\sigma)$ is an increasing function of δ_k. In addition, we have already noted earlier that the δ_k cannot be smaller than 1. Thus, interpreting $D_{n,k}(\cdot)$ as a metric and minimizing it with respect to δ_k, it produces the normal choice $\delta_k = 1$, for all $k \leq n$; see Alvo and Cabilio (1995). We remark in passing that the choice $\delta_k = (n-k+1)/2$ made by Critchlow (1985), though it may seem natural in making the mean rank equal to $(n+1)/2$, does not correspond to the projection principle so that produces stochastically larger values of $D_{n,k}(\sigma)$. Therefore, in what follows, we take $\delta_k = 1$, hence

$$D_{n,k}(\sigma) = 2\sum_{i=1}^{k} L_i(\sigma),\ k = 0,\ldots,n. \tag{8}$$

3.3 Exact properties

Noting that $T_k(\sigma) = k - L_k(\sigma)$, for all $k \leq n$, we have $\mathrm{E}(L_k(\sigma)) = k(n-k)/n$, for $k = 1, \ldots, n$. Hence,

$$\mathrm{E}(D_{n,k}(\sigma)) = k(k+1)\left(1 - \frac{2k+1}{3n}\right), \quad k = 0, \ldots, n.$$

Also, we have

$$\mathrm{Cov}\,(L_k(\sigma), L_q(\sigma)) = \frac{k^2(n-q)^2}{n^2(n-1)}, \quad 1 \leq k \leq q \leq n. \tag{9}$$

Equation (9) leads to (for $1 \leq k \leq q \leq n$)

$$\begin{aligned} V_{nkq} &= \mathrm{Cov}\,(D_{n,k}(\sigma), D_{n,q}(\sigma)) \\ &= \frac{4}{n^2(n-1)}\left[\sum_{i=1}^{k}\sum_{j=1}^{i} j^2(n-i)^2 + \sum_{i=1}^{k}\sum_{j=i+1}^{n}(n-j)^2 i^2\right] \\ &= \frac{k(k+1)}{45n^2(n-1)}\big\{30[q(2k+1)+1-k^2]n^2 - 6(2k+1)\big[5q(q+1) \\ &\qquad - (k-1)(k+2)\big]n + 5(2k+1)q(q+1)(2q+1)\big\}. \end{aligned}$$

Letting $k = q$, we have

$$\begin{aligned} V_{nk} &= \mathrm{V}(D_{n,k}(\sigma)) \\ &= \frac{k(k+1)}{45n^2(n-1)}\big[30(k^2+k+1)n^2 - 12(2k+1)(2k^2+2k+1)n + 5k(k+1)(2k+1)^2\big]. \end{aligned}$$

3.4 Exact and asymptotic distributions

The exact distribution of $D_{n,k}(\cdot)$ can be easily obtained from the algorithm presented in Section 2. We construct the matrix $\mathrm{P}_k^{(n)}$ whose elements are $\mathrm{P}_k^{(n)}[X][Y] = \mathrm{P}(\sum_{i=1}^{k} L_i = X, L_k = Y)$. Now, $\mathrm{P}(\sum_{i=1}^{k} L_i = X) = \sum_y \mathrm{P}_k^{(n)}[X][Y]$. We write

$$\mathrm{Y}_i(\sigma) = \frac{T_i(\sigma)}{(n-i)^2} - \sum_{j=1}^{i} d_{ni}, \ i = 1, \ldots, n, \tag{10}$$

where

$$d_{nj} = \frac{(n-j+1)(2j-1)-(j-1)}{(n-j)^2(n-j+1)^2}, \ j = 1, \ldots, n. \tag{11}$$

Notice that $\sum_{j=1}^{i} d_{nj} = i^2/[n(n-i)^2]$ and that the partial sequence $\{Y_i(\sigma), 0 \leq i \leq n\}$ is a zero-mean martingale array. Now, expressing the $T_i(\cdot)$ in terms of the $L_i(\cdot)$, we have

$$L_i(\sigma) = \frac{i(n-i)}{n} - (n-i)^2 Y_i(\sigma), \ i = 1, \ldots, n. \tag{12}$$

Using Equations (12) and (8), we obtain

$$D_{n,k}(\sigma) = 2\sum_{i=1}^{k} \frac{i(n-i)}{n} - 2\sum_{i=1}^{k}(n-i)^2 Y_i(\sigma),\ k = 1, \ldots, n.$$

To capture the full implication of this representation along with the martingale characterization, we now formulate a permutational functional limit theorem which yields a stronger result than the asymptotic normality at fixed k. We consider an integer-valued sequence $\{k_n(t)\colon t \in [0,1]\}$ by letting

$$k_n(t) = \max_k \left\{ \frac{V_{nk}}{V_{nn}} \leq t \right\},\ t \in [0,1],$$

so that $k_n(t)$ is a non-decreasing jump function with jumps at each value of V_{nk}/V_{nn}, for $k \leq n$ (note V_{nk} is non-decreasing in k). Let us define a stochastic process $W_n = \{W_n(t),\, t \in [0,1]\}$ by letting $W_n(t) = V_{nn}^{-1/2}[D_{n,k_n(t)}(\sigma) - \mathrm{E}(D_{n,k_n(t)}(\sigma))]$, for $t \in [0,1]$. In this context, we may note that if n increases with $k/n \to t$ and $0 \leq t \leq 1$, then

$$\frac{\mathrm{E}(D_{n,k}(\sigma))}{n(n^2-1)} \to \frac{1}{3} t^2(3-2t).$$

Similarly, $V_{nk}/V_{nn} \to t^4(15 - 24t + 10t^3)$. Also, as n increases with $k/n \to t$, $q/n \to s$, and $0 \leq t < s \leq 1$, then

$$\frac{V_{nkq}}{V_{nn}} \to t^3(30s - 30s^2 + 10s^3 - 15t + 6t^2).$$

These expressions are useful in computing the covariance function of $W_n(t)$. Side by side, we consider a Gaussian function $W = \{W(t), t \in [0,1]\}$ on $[0,1]$ which has zero drift and covariance function

$$\gamma(t,s) = \mathrm{E}(W(t)W(s)) = t^3(30s - 30s^2 + 10s^3 - 15t + 6t^2),\ 0 \leq t \leq s \leq 1.$$

As such, using the functional central theorem for martingale arrays (viz. Theorem 2.4.2 of Sen, 1981), as adopted here under P_n (the uniform distribution over S_n), we arrive by some routine steps at the following result.

Theorem 3.2 Under P_n, as n increases, W_n weakly converges to W, in the Skorokhod-J_1 topology on $D[0,1]$.

As a direct consequence of this result, we claim that as n increases with $k/n \to t$, for some $t \in (0,1]$, the standardized version of $D_{n,k}(\sigma)$ is closely normally distributed. Further, $W_n(\cdot)$ is tight (as n increases), hence the uniform continuity in probability condition holds. This enables us to draw conclusions about the asymptotic distribution of $D_{n,k}(\sigma)$ when k itself is random.

4. Weighted Spearman's Footrule

In this, we include weighted versions of Spearman's footrule using the same framework considered in Section 3 where we proposed an extension of the Spearman's footrule to the partial rankings case.

Sometimes it is desirable to assign greater weight to the first ranked item, less to the second, and so on. For example, consider the following three permutations (rankings) in S_5: $\pi = (1,2,3,4,5)$, $\sigma_1 = (2,1,3,4,5)$, and $\sigma_2 = (1,2,3,5,4)$. Using Spearman's rank correlation, Spearman's footrule, or Kendall's tau, the correlation (distance) between π and σ_1 is the same as that between π and σ_2. But we notice that σ_1 disagrees with π on the first two positions, while σ_2 disagrees with π on the last two positions. Accordingly, we would like a new version (of these measures) to indicate that σ_2 is closer to π than σ_1 (or the correlation between π and σ_2 is higher than that between π and σ_1).

Regular (unweighted) measures of correlation assign the same weights to all observations. In contrast, weighted measures assign different weights. For example, higher weights to observations that agree on lower ranks. This notion is gaining more popularity in light of the increased interest in measures of correlations addressing the agreement between two rankings dealing with the top k out of n objects (as indicated in the web search situation, and the top number of genes situation). In some sense, the censored rankings measures are a form of weighted measure that (sort of) assigns a weight equal to one for observations that have a rank less than k, assigned by at least one "judge", and a weight of zero otherwise. Back to Spearman's footrule, Jurman et al. (2008) used the Canberra distance (see Lance and Williams, 1967) given by

$$\mathrm{Ca}(\pi,\sigma) = \sum_{i=1}^{n} \frac{|\pi(i) - \sigma(i)|}{\pi(i) + \sigma(i)}.$$

Replacing unranked items (when we have censored data) by $k+1$, Jurman et al. (2008) provided the following version of the Canberra distance under censoring

$$\mathrm{Ca}^{(k+1)}(\pi,\sigma) = \sum_{i=1}^{n} \frac{|\min(\pi(i), k+1) - \min(\sigma(i), k+1)|}{\min(\pi(i), k+1) + \min(\sigma(i), k+1)}.$$

They provided its expected value (up to $o(1)$) term by

$$\mathrm{E}(\mathrm{Ca}^{(k+1)}) = \frac{(k+1)(2n-k)}{n}\log(4) - \frac{2kn + 3n - k - k^2}{n}.$$

Another weighted version is related to the statistic T introduced in Salama and Quade (1982), where (assuming $\pi = (1,\ldots,n)$), $T(\sigma) = T(\pi,\sigma) = \sum_{i=1}^{n} T_i(\sigma)/i$. This was introduced as a weighted version of Spearman's footrule based on

$$D_n(\sigma) = D(\pi,\sigma) = \sum_{i=1}^{n} |i - \sigma(i)| = n(n+1) - 2\sum_{i=1}^{n} T_i(\sigma).$$

Note that $T(\sigma)$ may be recast as a metric on S_n as

$$d(\pi,\sigma) = n - \sum_{i=1}^{n} \frac{T_i(\pi,\sigma)}{i},$$

where (if we drop that $\pi = (1,\ldots,n)$), then, we may define $T_i(\pi,\sigma) = \#\{A_i(\pi) \cap A_i(\sigma)\}$, where $A_i(\pi)$ $(A_i(\sigma))$ is the set of objects that are ranked i or less by π (σ).

4.1 A weighted version for a given measure of correlation

In what follows, we discuss a method that provides a weighted version for a given measure of correlation (distance), both for full or partial (censored) ranking cases. For $\sigma_n = (\sigma_{1n}, \ldots, \sigma_{nn}) \in S_n$, and $k = 1, \ldots, n-1$, let $\sigma_n^{(k)} = (\sigma_{1n}^{(k)}, \ldots, \sigma_{nn}^{(k)})$, where

$$\sigma_{in}^{(k)} = \sigma_{in} I(\sigma_{in} \leq k) + (k+1) I(\sigma_{in} > k), \quad i = 1, \ldots, n.$$

Now, let $S_n^{(k)} = \{\sigma_n^{(k)} | \sigma_n \in S_n\}$, for $k = 1, \ldots, n-1$. Note that $S_n^{(k)}$ is the set of all possible rankings when we only rank the top k out of $n(\geq k)$ objects. This situation corresponds to projecting S_n onto subspaces $S_n^{(k)}$ such that $S_n^{(0)} = 0 \subseteq S_n^{(1)} \subseteq \cdots \subseteq S_n^{(n-1)} = S_n^{(n)} = S_n$. Let $D_{nk}(\cdot,\cdot)$ be an extension of $D_n(\cdot,\cdot)$ to $S_n^{(k)}$ (in the case of Spearman's rank correlation or Spearman's footrule, the extension to $S_n^{(k)}$ is clear. On the other hand, for Kendall's tau, we need to accommodate ties). Having defined a measure of correlation D_{nk} (over $S_n^{(k)}$) based on D_n (over S_n), we then define a weighted version $D_{nk}^{(W)}$ (over $S_n^{(k)}$) as follows

$$D_{nk}^{(W)}(\pi_n^{(k)}, \sigma_n^{(k)}) = \sum_{i=1}^{k} D_{ni}(\pi_n^{(i)}, \sigma_n^{(i)}).$$

In what follows, we provide a representation of a weighted version of Spearman's footrule based on the previous construction. To this end, we consider (as in Salama and Quade, 1982) the partial sequence $\{T_{nj}^{(k)}\colon 0 \leq j \leq k \leq n\}$, where for $k = 1, \ldots, n$, $T_{n0}^{(k)} = 0$, and

$$T_{nj}^{(k)} = \sum_{i=1}^{j} I(\sigma_{in}^{(k)} \leq j), \quad j = 1, \ldots, k \leq n.$$

At this juncture, we note that $T_{nj}^{(k)} = T_{nj}^{(n)}$, for $1 \leq j \leq k \leq n$. Accordingly, we use T_j to indicate $T_{nj}^{(n)}$, which is the same as $T_{nj}^{(k)}$, for $j = 1, \ldots, k \leq n$. Now, if we write Spearman's footrule as

$$D_n(\sigma) = \frac{1}{2} \sum_{i=1}^{n} |i - \sigma_{in}|,$$

then, D_{nk} admits the following representation (in terms of the T_j's)

$$D_{nk}(\sigma) = \frac{k(k+1)}{2} - \sum_{j=1}^{k} T_j,$$

with weighted version (based on the successive projection scenario) given by (as mentiond in Salama and Quade, 2004)

$$D_{nk}^{(W)}(\sigma) = \frac{k(k+1)(k+2)}{6} - \sum_{j=1}^{k} (k+1-j) T_j(\sigma).$$

4.2 Asymptotic normality

Setting $D^W_{n0} = 0$, our interest is to study the distributional behavior of the triangular array given by $\{D^W_{nk}\colon 0 \le k \le n; n \ge 1\}$ under the hypothesis of random ranking, i.e., under the permutational probability measure P_n associated with the discrete uniform distribution of σ on S_n. Noting that the permutational moments are given by

$$\mathrm{E}(T_j(\sigma)) = \frac{j^2}{n} = \mu_j, \quad 1 \le j \le k \le n,$$

and

$$\mathrm{E}((T_i(\sigma) - \mu_i)(T_j(\sigma) - \mu_j)) = \frac{i^2(n-j)^2}{n^2(n-1)}, \quad 1 \le i \le j \le k \le n,$$

we conclude that

$$\mathrm{E}(D^W_{n,k}(\sigma)) = \frac{k(k+1)(k+2)}{6}\left[1 - \frac{k+1}{2n}\right]$$

and

$$\begin{aligned} \mathrm{V}(D^W_{n,k}(\sigma)) &= \frac{k(k+1)(k+2)}{5040n^2(n-1)}[84n^2(k+1)(k^2+2k+2) \\ &\quad -8n(13k^4+52k^3+77k^2+50k+18)+35k(k+1)^3(k+2) \\ &= \frac{k^6}{5040n^2(n-1)}[84n^2 - 104nk + 35k^2] + o\left(\frac{k^6}{n}\right). \end{aligned}$$

Furthermore, for $1 \le j \le k \le n$, we have

$$\begin{aligned} V_{jkn} &= \mathrm{Cov}\,(D^W_{n,j}(\sigma), D^W_{n,k}(\sigma)) \\ &= \frac{k^4}{5040n^3}[n^2(210k^2 - 84jk - 42j^2) \\ &\quad -n(140k^3 + 84j^2k - 120j^3) + (35k^4 + 60j^3k - 60j^4)] + o\left(\frac{k^6}{n}\right). \end{aligned}$$

Sen and Salama (1983) studied the stochastic structure of $T_k(\sigma)$ incorporating a martingale approach that added convenience to the study of the asymptotic distribution theory. Based on second equation in Lemma 2.3 and Equations (10) and (11), we express $D_{nk}(\sigma)$ in terms of $Y_i(\sigma)$ as follows

$$D^W_{nk}(\sigma) = \mathrm{E}(D^W_{nk}(\sigma)) - 2\sum_{i=1}^{n}(n-i+1)(n-i)^2Y_i(\sigma).$$

To capture the full implication of the above representation along with the martingale characterization (of the partial sequence $\{Y_i(\sigma), 0 \le i \le n\}$ as a zero-mean martingale array), we proceed now to formulate a permutational functional central limit theorem. We consider an integer-valued sequence $\{h_n(t)\colon t \in [0,1]\}$ by letting

$$h_n(t) = \max_k \left\{\frac{V_{nk}}{V_{nn}} \le t\right\}, \quad t \in [0,1],$$

so that $h_n(t)$ is a non-decreasing jump function with jumps at each value of V_{nk}/V_{nn}, for $k \leq n$ (this is because V_{nk} is non-decreasing in k). Define a stochastic process $W_n = \{W_n(t)\colon t \in [0,1]\}$ by letting

$$W_n(t) = V_{nn}^{-\frac{1}{2}}\{D^W_{n,h_n(t)}(\sigma) - \mathrm{E}(D_{n,h_n(t)}(\sigma))\}, \quad t \in [0,1].$$

In this context, we may note that if n increases with $k/n \to t$, $0 \leq t \leq 1$, then

$$\frac{\mathrm{E}(D^W_{nk}(\sigma))}{n^3} \to \frac{t^3(2-t)}{12}.$$

Similarly,

$$\frac{V_{nkk}}{V_{nnn}} \to \frac{t^6(35t^2 - 104t + 84)}{15}.$$

Also, as n increases with $j/n \to s$, $k/n \to t$, and $0 \leq s < t \leq 1$, we have

$$\frac{V_{njk}}{V_{nnn}} \to \frac{t^4(35t^4 + 60s^3t - 60s^4 - 140t^3 - 84s^2t + 120s^3 + 210t^2 - 84st - 42s^2)}{15}.$$

These expressions are useful in computing the covariance function of $W_n(t)$. We consider a Gaussian function on the interval $[0,1]$, $W = \{W(t)\colon t \in [0,1]\}$, with zero drift and covariance function

$$\begin{aligned}\gamma(s,t) &= \mathrm{E}(W(s)W(t)) \\ &= \frac{1}{15}t^4(35t^4 + 60s^3t - 60s^4 - 140t^3 - 84s^2t + 120s^3 + 210t^2 - 84st - 42s^2).\end{aligned}$$

Using the functional limit theorem for a martingale array (Theorem 2.4.2 of Sen, 1981), as adapted here under P_n, we arrive by some routine steps at the following.

Theorem 4.1 Under random ranking, as n increases, W_n weakly converges to W in the Skorohod J_1-topology on $[0,1]$.

As a direct consequence of this weak convergence result, we claim that as n increases with $k/n \to t$, for some $t \in [0,1]$, the standardized version of $D^W_{nk}(\sigma)$ is closely normally distributed. Further, $W_n(\cdot)$ is tight (as n increases), and hence, the uniform continuity in probability condition holds. This enables us to derive the asymptotic distribution of $D^W_{nk}(\sigma)$ even when k is random.

5. Remarks

We conclude this work with the following remarks.

(i) The importance of asymptotic results in applications stems from the fact that, in many situations, the number of objects under consideration for ranking becomes increasingly large. This is clearly apparent in genomics, where the number of genes under consideration is in the thousands. We also see greater numbers dealing with the number of items we encounter in a web search problem. Let $O = \{O_1, \ldots, O_n\}$

be the set of objects under consideration for ranking (we may take $O = \{1, \dots, n\}$). Let R be the $(m \times n)$ matrix of ranks provided by m sources, where the ith row $R_i = (r_{i1}, \dots, r_{in})$, and r_{ij} is the rank assigned to object j by source i. We note that R_i can be a permutation (if all n objects are ranked), or a partial ranking (if only k out of n objects are ranked).

(ii) One of the problems in genomics is to identify the top k genes for further research. That is, given the matrix R, we need to identify the top k objects. In doing so, we encounter what is called the stability problem, meaning the effect of small perturbations of the data on the final result; see Jurman et al. (2008). The stability problem may be cast in terms of the variability among $R_1, \dots, R_m$, which, in turn, may be considered as the consistency or internal agreement among $R_1, \dots, R_m$. For example: let $A_1 = \{(1,2,3),(1,2,3),(1,2,3)\}$, and $A_2 = \{(1,2,3),(2,1,3),(3,2,1)\}$. Then, it is intuitively clear that set A_2 reflects more variability than set A_1. In turn, we can say that set A_1 reflects more consistency (or internal agreement) in ranking (the three objects) than set A_2. One way to measure the internal agreement is by means of a measure of correlation (or a metric) between the rows of R. If we let $d_n^k(\cdot,\cdot)$ be a metric (or its associated measure of correlation) on the set of objects represented by $R_1, \dots, R_m$, then as a measure of internal agreement we may use

$$d_{n,k}^*(R) = \sum_{i<j}^{m} d_n^k(R_i, R_j). \tag{13}$$

Spearman's footrule (along with Kendall's tau) is one of the measures used in such a situation (Quade and Salama, 2006). We can test for internal agreement, and here asymptotic results can be used to reach such conclusion. The conclusion that we have internal agreement greatly enhance our confidence in our choice of the top k objects.

(iii) If the number k is fixed in advance, then, we may test to see if we have internal agreement with respect to k. If the answer is yes, then we may proceed (using some score function) to produce the desired top k genes. If k is not fixed in advance, then we may proceed by searching for the "best" reasonable k that we have internal agreement upon. This search may be done using the sequence $d_{n,k}^*(R)$ (multiple inference problem).

(iv) In the web search situation, we need to rank the top k items. Results of different search engines may be reported as the matrix R. Now, we need to combine these rankings to produce one single ranking. One of the methods used in this process is to find a "ranking" R_0 such that the sum of the total distances from R_0 to $R_1, \dots, R_m$ is minimum. Spearman's footrule (and Kendall's tau) is used in this problem, along with a weighted version of it (see Pihur et al., 2009). It seems reasonable that before we try to search for one "ranking" that may be used to represent the entire set of ranks $\{R_1, \dots, R_m\}$, we test whether we have internal agreement among the elements of this set. If we can conclude this, then, we can start our search to produce such a representative. Again, due to the large number of objects under consideration (large n), the asymptotic theory becomes very relevant in the testing process.

(v) If we are only looking for the top k (disregarding the internal rankings within the top k), then, the ranking from each source is nothing but a binary classification of objects. If (for example) we denote items classified as top k by 1 and others by 2, then we may regard the final ranks as a sequence of on $\{1,2\}$. In this case, we may use the Kappa statistic (Cohen, 1960) as a measure for agreement between two sets of "ranks", which in turn is equivalent to using Spearman's footrule (the L_1 norm; and the L_2 in this case). Let $(r_{11}, \ldots, r_{1n})$ and $(r_{21}, \ldots, r_{2n})$ represent the classifications provided by the two sources (r_{ij} is 1 or 2). Then, $|r_{1j} - r_{2j}|$ is equal to 0 if the the two sources agree on classifying object j, and it is equal to 1 if they disagree. Accordingly, $\sum_{j=1}^{n} |r_{1j} - r_{2j}|$ is the number of objects on which the two sources disagree. The equivalence of Spearman's footrule and the Kappa statistic is based on the number of disagreements plus the number of agreements being equal to n. Using Spearman's footrule as the metric in Equation (13), we have what is equivalent to extending the Kappa statistic to testing the agreement among m sources.

(vi) Our discussion is based on m sources ranking (or partially ranking) n objects, which is similar to the m-ranking case in which the Friedman's (1937) statistic is used. If, for $j = 1, \ldots, n$, we let $r_{\cdot,j}$ be the mean ranks for object j, then the Friedman's statistic is based on $\sum_{j=1}^{n}(r_{\cdot,j} - (n+1)/2)^2$, which may be written as a function of $\sum_{i<j}^{n}(r_{\cdot,i} - r_{\cdot,j})^2$. Accordingly, the emphasis is on the average ranking of objects instead of inter-rater differences as reflected in the statistic based on Spearman's footrule (or other metrics).

Appendix

In this part, we present the algorithm that can be used to generate the matrix $\mathrm{P}_{k+1}^{(n)}$ from $\mathrm{P}_{k}^{(n)}$ discussed in Section 2.

```
For($j = 0; j \leq d_2(n,k); j = j + 1$)

    $$T[2] = \frac{[(n-k)-j][(n-k)-j-1]}{(n-k)^2};$$
    $$T[1] = \frac{[(n-k)-j](2j+1)}{(n-k)^2};$$
    $$T[0] = \frac{j^2}{(n-k)^2};$$

for($i = 0; i \leq d_1(n,k); i = i + 1$)

    $$\mathrm{P}_{k+1}^{(n)}[i+j+1][j+1] = \mathrm{P}_{k}^{(n)}[i][j] * T[2] + P_{k+1}^{(n)}[i+j+1][j+1];$$
    $$\mathrm{P}_{k+1}^{(n)}[i+j][j] = \mathrm{P}_{k}^{(n)}[i][j] * T[1] + P_{k+1}^{(n)}[i+j][j];$$

if($j > 0$)

    $$P_{k+1}^{(n)}[i+j-1][j-1] = P_{k}^{(n)}[i][j] * T[0] + P_{k+1}^{(n)}[i+j-1][j-1].$$

End of $i$ loop.
End of $j$ loop.
```

References

Alvo, M., Cabilio, P., 1995. Rank correlation methods for missing data. Canadian Journal of Statistics, 23, 345-358.

Alvo, M., Charbonneau, M., 1997. The use of Spearman's footrule in testing for trend when the data are incomplete. Communications in Statistics - Simulation and Computation, 26, 193-213.

Bar-Ilan, J., Levene, M., Lin, A., 2007. Some measures for computing citation databases. Journal of Informatics, 1, 26-34.

Berman, S.M., 1996. Rank inversions in scoring multipart examinations. The Annals of Applied Probability, 6, 992-1005.

Blest, D.C., 2000. Rank correlation – An alternative measure. Australian and New Zealand Journal of Statistics, 42, 101-111.

Boulesteix, A., Slawski, M., 2009. Stability and aggregation of ranked gene lists. Technical report 059. Department of Statistics, University of Munich, German.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Costa, J.P., Soares, C., Brazdil, P., 2001. Some improvements in the evaluation of methods to rank alternatives. Poster at Workshop on Non-linear Estimation and Classification. MSRI, Berkeley, California.

Costa, J.P., Soares, C., 2005. A weighted rank measure of correlation. Australian and New Zealand Journal of Statistics, 47, 515-529.

Costa J.P., Soares, C., 2007. Rejoinder to letter to the editor from C. Genest and J.-F. Plante concerning Pinto da Costa, J., Soares, C., 2005. A weighted rank measure of correlation. Australian and New Zealand Journal of Statistics, 49, 205-207.

Critchlow, D.E., 1985. Metric Methods for Analysing Partially Ranked Data. Springer-Verlag, Berlin and New York.

Diaconis, P., Graham, R.L., 1977. Spearman's footrule as a measure of disarray. Journal of The Royal Statistical Society Series B - Statistical Methodology, 39, 262-268.

Dwork, C., Kumar, R., Naor, M., Sivakumar, D., 2001. Rank aggregation methods for the web. In Proceedings of the 10th World Wide Web Conference, May 1-5, Hong Kong, pp. 613-622.

Fagin, R., Kumar, R., Sivakumar, D., 2003. Efficient similarity search and classification via rank aggregation. In Proceedings of ACM SIGMOD International Conference on Management of Data, June 9-12, 2003, San Diego, California.

Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., Vee, E., 2004. Comparing and aggregating rankings with ties. In Proceedings of ACM Symposium on Principles Of Database Systems (PODS), June 14-16, Paris, France, pp. 47-58.

Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., Vee, E., 2006. Comparing partial rankings. SIAM Journal Discrete Mathematics, 20, 628-648.

Franklin, L., 1988. Exact tables of Spearman's footrule for N=11(1)18 with estimate of convergence and errors for the normal approximation. Statistics and Probability Letters, 6, 399-406.

Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 32, 675-701.

Genest, C., Neslehova, J., Ben Ghorbal, N., 2010. Spearman's footrule and Gini's gamma: a review with complements. Journal of Nonparametric Statistics, 22, 937-954.

Genst, C., Plante, J.F., 2007. Re: Pinto da Costa, J., Soares C. (2005). A weighted rank correlation problem. Australian and New Zealand Journal of Statistics, 49, 203-204.

Hoeffding, W., 1951. A combinatorial limit theorem. The Annals of Mathematical Statistics, 22, 558-566.

Jurman, G., Merler, S., Barla, A., Poli, S., Galea, A., Furlanello, C., 2008. Algebraic stability indicators for ranked lists in molecular profiling. Bioinformatics, 24, 258-264.

Kim, B., Rha, S., Cho, G., Chung, H., 2004. Spearman's footrule as a measure of cDNA microarray reproducibility. Genomics, 84, 441-448.

Lance, G.N., Williams, W.T., 1967. Mixed-data classification program: I. Agglomerative systems. Australian Computer Journal, 1, 15-20.

Lee, P., Yu, P., 2010. Distance-based tree models for ranking data. Computational Statistics and Data Analysis, 54, 1672-1682.

Lin, S., 2010. Rank aggregation methods. Computational Statistics, 2, 555-570.

Lin, S., Ding, J., 2009. Integration of ranked lists via cross entropy monte carlo with application to mRNA and microRNA studies. Biometrics, 65, 9-18.

Mango, A., 1997. Rank correlation coefficient: A new approach. Computing Science and Statistics. Computational Statistics and Data Analysis on the Eve of the 21st Century. Proceedings of the Second World Congress of the IASC, 29, 471-476.

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

Pihur, V., Datta, S., Datta, S. 2008. Finding common genes in multiple cancer types through meta-analysis of microarray experiments: a rank aggregation approach. Genomics, 92, 400-403.

Pihur, V., Datta, S., Datta, S., 2009. RankAggreg, an R package for weighted rank aggregation. BMC Bioinformatics, 10, 62.

Powell, T., Reinhardt, I., 2010. Rank friction: an ordinal approach to persistent profitability. Strategic Management Journal, 31, 1244-1255.

Quade, D., Salama, I.A., 1992. A survey of weighted rank correlation. In Sen, P.K., Salama, I.A., (eds.). Order Statistics and Nonparametrics: Theory and Applications. Elsevier Science Publishers B.V., Amesterdam, pp. 213-225.

Quade, D., Salama, I.A., 2006. Concordance of complete or right-censored rankings based on Spearman's footrule. Communications in Statistics - Theory and Methods, 35, 1059-1069.

Salama, I.A., Quade, D., 1982. A nonparametric comparison of two multiple regressions by means of a weighted measure of correlation. Communications in Statistics - Theory and Methods, 11, 1185-1195.

Salama, I.A., Quade, D., 1990. A note on Spearman's footrule. Communications in Statistics - Simulation and Computation, 19, 591-601.

Salama, I.A., Quade, D., 2002. Computing the distribution of Spearman's footrule in $O(n^4)$ time. Journal of Statistical Computation and Simulation, 72, 895-898.

Salama, I.A., Quade, D., 2004. Agreement among censored rankings using Spearman's footrule. Communications in Statistics - Theory and Methods, 33, 1837-1850.

Sen, P.K., 1981. Sequential Nonparametrics: Invariance Principles and Statistical Inference. John & Wiley, New York.

Sen, P.K., Salama. I.A., 1983. The Spearman footrule and a Markov chain property. Statistics and Probability Letters, 1, 285-289.

Sen, P.K., Salama, I.A., Quade, D., 2003. Spearman's footrule under progressive censoring. Journal of Nonparametric Statistics, 15, 53-60.

Shieh, G.S., 1998. A weighted Kendall's tau statistic. Statistics and Probability Letters, 39, 17-24.

Spearman, C., 1906. Footrule for measuring correlation. The British Journal of Psychiatry, 2, 89-108.

Sufyan Beg, M., Ahmed, N., 2007. Web search enhancement by mining user action. Information Science, 177, 5203-5218.

Tarsitano, A., 2009. Comparing the effectiveness of rank correlation statistics. Working paper. Universit della Calabria, Dipartimento di Economia e Statistica, Italia.

Ubeda-Flores, M., 2005. Multivariate version of Blomqvist's Beta and Spearman's footrule. Annals of the Institute of Statistical Mathematics, 57, 781-788.

Ury, H.K., Kleinecke, D.C., 1979. Tables of the distribution of Spearman's footrule. Applied of Statistics, 28, 271-275.