EVALUATING THE REPEATABILITY OF TWO STUDIES OF A LARGE NUMBER OF OBJECTS: MODIFIED KENDALL RANK-ORDER ASSOCIATION TEST

TIAN ZHENG, SHAW-HWA LO
DEPARTMENT OF STATISTICS, COLUMBIA UNIVERSITY

Abstract. In this paper, we propose a modified Kendall rank-order association test to address the difficulty of evaluating the repeatability of two studies on a large number of objects, most of which are undifferentiated. Nowadays statisticians face data of exploding dimensions from current scientific researches. The numbers of objects that need to be evaluated are frequently in the thousands or tens of thousands. Most of the time, however, the data are in the form of a enormous pool that contains a small set of “interesting” or “important” objects, which hold the answers to the scientific questions. For example, in microarray analysis, only a small fraction of genes are truly regulated due to the treatments under a single experiment, while the data contain expression levels from tens of thousands. The reliability of such studies is widely concerned, especially because of the fact that different studies on the same objects under the same condition usually result in nearly uncorrelated ranking of the objects. It is expected that the repeatability between two studies on the same objects should be high if the scientific signal is real. Since the repeatability between two datasets reflects in the association between the two sets of observed values, evaluating the extent and the significance of such association can be one way to measure the strength of the signals in the data. Due to the complex nature of the data, we consider the ranking association which is distribution-free. Conventional measures of association use all objects, which inevitably includes many uninformative ones, and thus do not have power when dealing the situation of only a few objects with true signals. In this paper, we propose a modified Kendall rank-order association test to address this difficulty. Through simulation results, we show that the proposed alteration to the classic Kendall rank-order correlation coefficient have desirable properties that will address many modern statistical needs.


Key words and phrases. Repeatability, Kendall rank-order, truncated rank, nonparametric association test.
Correspondences should be sent to Tian Zheng, Department of Statistics, Room 1005, Mail Code 4690, 1255 Amsterdam Avenue, Columbia University, New York, New York, 10027. Email correspondences should be addressed to tzheng@stat.columbia.edu.
Current technology developments in many scientific fields allow researchers to explore the research problems with more details, on a larger scale, involving many possible factors, and in huge dimensions simultaneously. Data collected in such ways, of unprecedentedly large sizes, provide new challenges to statisticians. In this paper, we will discuss one of them: evaluating the repeatability of two studies that rank a same large set of objects. In this paper, to avoid confusion with the statistical term “subject”, we name the units of evaluation objects.

It is believed that, despite the huge amount of data collected, only a small portion of the variables (factors) hold the keys to new scientific discovery or validation. Evaluating the importance of these variables (or factors) and ranking them accordingly has been a focus of many modern statistical research today. Such evaluation provides basis for dimension reduction, model selection, feature selection, machine learning, and etc. In this paper, for the convenience of discussion, we call the variables (or factors, or so-called dimensions in some context) objects of evaluation.

Nowadays, the number of objects that need to be evaluated are frequently in the thousand or tens of thousand, while only a few of them is relevant to the questions that are to be addressed by these data. Thus, the data are in the form of a enormous pool that contains a small set of “interesting” or important objects. Typical examples of such scenarios can be well found in microarray analysis. It is believed that only dozens or hundreds of genes are truly regulated due to the treatments under a single experiment, while the data contain expression levels from tens of thousands. Conventionally, genes of more interests are those that are highly expressed or those whose expression profiles correlate well with the treatments. The reliability of such studies is widely concerned, due to the fact that different studies on the same objects under the same condition usually result in different lists of highly expressed genes. It is hoped, however, that the repeatability between two studies should be high if the scientific signal is real. Based on such a belief that the repeatability between two datasets reflects in the association between the two sets of observed values, evaluating the extent and the significance of such association becomes a way to measure the strength of the signals in the data. Because of the complicated nature of such data, nonparametric measures of association are much desired since the distributions of the observed values are usually unknown. Due to this consideration, we study the repeatability between two studies through ranking associations.

Consider a simple example using a microarray dataset (van ’t Veer et al. 2002). In this study, gene expression levels of 24479 biological oligonucleotides in samples from 78 breast cancer patients were measured.
Figure 1. An example of two microarray experiment samples showing weak association.

(using 2-dye-hybridization experiments on DNA Microarrays). Among these 78 breast cancer patients, 44 remained disease-free for more than 5 years, while the other 34 patients developed metastases within 5 years (van ’t Veer et al. 2002). The goal of this study was to identify genes that were associated with the risk of developing metastases and use these genes for a better prediction of disease outcomes. To illustrate of the concepts and methods proposed in this paper, consider a small “experiment” using the data (downloaded from the paper web site). We randomly divide the data into two equal halves, each with 17 patients with poor prognosis (metastases within 5 years) and 22 patients with good prognosis (no metastases for more than 5 years). For the convenience of discussion, we call these two halves of the data sample 1 and sample 2. For each gene, in either sample, correlation between the log_{10} gene expression ratios (the real gene expression versus the background gene expression) and the prognosis label (1=poor, 0=good) is calculated. In such a
study, both positively and negatively correlated genes are regarded as important. Thus, we plot, in Figure 1, the absolute correlation values from sample 2 versus that of sample 1. One can expect that the truly important genes will be strongly (positively or negatively) correlated with the prognosis label, in both sample 1 and sample 2. It is also expected that some genes will display strong association by chance in one of the two samples. The noisy pattern in Figure 1 display a weak association between these two samples. Actually, the Kendall rank-order correlation coefficient of the 24479 genes expression levels is 0.00522 (p-value = 0.112). Another interesting pattern is that, of the 5000 top correlated genes selected based on the combined data of sample 1 and sample 2, only 17% demonstrated strong correlation with the prognosis label in both samples. It seems that the association between gene expressions and the cancer outcome are nearly random and uncorrelated between samples. It is noted, however, that the two samples have about 1131 top genes in common if one simply select the top 5000 ones with the highest absolute correlation values. Is it due to chance or true signal?

It is not a surprise that the association on all objects is low since the unimportant or “uninteresting” objects should be similar and undifferentiated in the analysis, except for random noises, and lead to random rankings, resulting in the overall nearly uncorrelated pattern as seen in Figure 1. However, it is of interests to examine the repeatability of the top genes since their repeated appearance in the top list may be due to true signals related to the mechanism of breast cancer. It is impossible to compute measures of association using only the top ranks since the objects do not overlap completely. On the other hand, measures of association using all objects inevitably include many uninformative ones, and thus do not have power when dealing the situation where only the top few matter. Simply examining the number of overlapping top objects will result in loss of power as we will show later. In this paper, we propose a modified Kendall rank-order association test to address this difficulty. Using rank-orders, this test is a non-parametric test of association, which does not rely on any assumption of the quantitative evaluation scores of the two studies. Through simulation results, we show the proposed alteration to the classic Kendall rank-order correlation coefficient have desirable properties that will address the needs of many modern statistical studies. For the example in Figure 1, the p-value derived for the modified Kendall rank-order test is $5.78 \times 10^{-5}$, indicating the association of the top genes are significant and supporting the existence of true biological signals.
In this section, we study the problem of repeatability through non-parametric tests of association between two sets of evaluations based on a same set of objects. To allow the tests to be distribution-free, the association is studied on the rankings not the actual observation or evaluation values.

2.1. Notations. Assume that \( n \) objects, \( S_1, S_2, \ldots, S_n \) are under evaluation. Let \( X_i, i = 1, \ldots, n \) and \( Y_i, i = 1, \ldots, n \) be two independent rankings (in decreasing order; In this article, we discuss rankings in the decreasing order unless otherwise noted.) received by these samples.

Denote \( \alpha_i \) as the true merit of object \( S_i \). The two sets of rankings are based on random observations and thus are random representations of the true ranking, \( \text{Rank}(\alpha_i) \). Here, the notation \( \text{Rank}(\cdot) \) is short for “the rank of ...”. For the convenience of the formulation, we assume that

\[
X_i = \text{Rank}(\alpha_i + \varepsilon_i),
\]
\[
Y_i = \text{Rank}(\alpha_i + \delta_i),
\]

where \( \varepsilon_i \)'s are independent random departures with distribution function \( F \), \( \delta_i \)'s are independent random departures with distribution function \( G \), and the \( \varepsilon \)'s are independent with the \( \delta \)'s. Here, the \( \alpha \)'s and the distribution functions \( F, G \) are introduced for the convenience of discussion and are neither assumed known nor used in the inference.

Without loss of generality, we assume that the objects \( S_1, \ldots, S_n \) are arranged in the order of their true merits, that is \( \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n \). Under the null hypothesis that there is no difference among the objects in terms of their true merits, i.e. \( \alpha_1 = \alpha_2 = \cdots = \alpha_n \), the ranking \( X \), now reduced to the \( \text{Rank}(\varepsilon) \), would be independent of the ranking \( Y, \text{Rank}(\delta) \). On the other hand, consider the extreme alternative where \( \alpha_1 > \alpha_2 > \cdots > \alpha_n \), the ranking \( X \) and the ranking \( Y \) will then be positively correlated, while the degree of correlation depends on the random variation of \( \varepsilon \)'s and \( \delta \)'s. Thus, the correlation between ranking \( X \) and ranking \( Y \) can be used to measure the variation of among the objects’ merit, relative to the random error variation.

2.2. Significance testing problem on ranking association. In practice, very frequently we need to separate out a small sample of objects with higher merit from the rest of the pool. If two rankings on the same objects are uncorrelated, the objects with high ranks may not be truly superior to the other objects. This is because it may be primarily due to luck that these objects are on the top of the list. On the other
hand, in studies that involve a large number of objects under evaluation (e.g., gene expression analysis), it is very common that the majority of the pool consists of objects that are quite similar if not identical in their true merits, except for several cases with higher ordered true merits. Under such circumstances, for a few objects among the top ranked, the rankings are correlated, while the rankings are uncorrelated for the rest of the observations with undifferentiated merits, even when the random error variations are low. Such a low correlation among the lower ranks dilutes the overall “coordination” between the ranking $X$ and the ranking $Y$. As a result, the overall correlation between $X$ and $Y$ would not be high. If, however, the objects are grouped into top ranks and low ranks (that is, higher rank value in a decreasing rank order), one would expect to observe a higher degree of correlation between the groupings according to the rankings $X$ and $Y$ than the correlation between the original rankings $X$ and $Y$, reflecting the important true merits of the few top objects.

Here we consider an inference question testing the null hypothesis $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_n \equiv \alpha$ versus a local alternative $H_a : \exists 1 \leq k_0 \ll n, \text{ s.t. } \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_{k_0} > \alpha_{k_0+1} = \alpha_{k_0+2} = \cdots = \alpha_{n-1} = \alpha_n$. We propose to study a modified Kendall rank-order test, which detects objects with true and high merits using the important association between the top ranks of two rankings and the association between the grouping of ranks, without effects from the noises in the lower ranks.

2.3. **Failure of the Kendall rank-order correlation coefficient.** We first examine the original Kendall rank-order correlation coefficient (Siegal and Castellan 1988, chapter 9). It uses the number of agreements and disagreements defined as follows: consider all possible pairs of ranks $(X_i, X_j)$ in which $X_i$ is lower than $X_j$, if

- the corresponding $Y_i$ is lower than $Y_j$, it is then an agreement;
- the corresponding $Y_i$ is higher than $Y_j$, it is then an disagreement.

Using these two counts, the Kendall rank-order correlation coefficient is formulated as

\[
T = \frac{\# \text{ agreements} - \# \text{ disagreements}}{\text{total number of pairs}}
\]  

(2.2)

Since the maximum possible values for the number of agreements and the number of disagreements are both the number of the total possible pairs, $T$ ranges between $-1$ and $1$, same as the conventional Pearson’s coefficient of correlation.
It is easy to show that

\begin{equation}
\# \text{agreements} = \sum_{i=1}^{n} \sum_{i \neq j} 1_{(X_i < X_j)} 1_{(Y_i < Y_j)},
\end{equation}

\begin{equation}
\# \text{disagreements} = \sum_{i=1}^{n} \sum_{i \neq j} 1_{(X_i < X_j)} 1_{(Y_i > Y_j)},
\end{equation}

and, assuming there are no tied observations,

\begin{equation}
\# \text{agreements} + \# \text{disagreements} = n(n-1)/2.
\end{equation}

Under the null hypothesis, \(E(\# \text{agreements}) = E(\# \text{disagreements}) = \frac{1}{4}n(n-1).\) And the variance of the number of agreements is

\[
\Var(\#\text{agreements}) = \Var\left(\sum_{i=1}^{n} \sum_{j \neq i} 1_{(X_i < X_j)} 1_{(Y_i < Y_j)}\right) \\
= \sum_{i=1}^{n} \sum_{j \neq i} \Var(1_{(X_i < X_j)} 1_{(Y_i < Y_j)}) \\
+ \sum_{i=1}^{n} \sum_{j \neq i} \sum_{(k,l) \neq (i,j)} \Cov(1_{(X_i < X_k)}, 1_{(Y_i < Y_l)}; 1_{(X_k < X_l)}, 1_{(Y_k < Y_l)}) \\
= \frac{3}{16}n(n-1) + 2\left(\frac{1}{9} + \frac{1}{36} - \frac{1}{8}\right)n(n-1)(n-2) - \frac{1}{16}n(n-1) \\
= \frac{1}{16}\left(\frac{4n}{9} + \frac{10}{9}\right)n(n-1).
\]

It follows that, under the null hypothesis,

\begin{equation}
E(T) = 0, \Var(T) = \Var\left(\frac{4(#\text{agreements}) - 2n(n-1)}{n(n-1)}\right) = \frac{2(2n+5)}{9n(n-1)}.
\end{equation}

Conventionally, for \(n > 10,\) the significance of \(T\) is evaluated using an approximating normal distribution (Siegel and Castellan 1988, chapter 9). From (2.2) and (2.7), one can see that such tests of association based on the Kendall rank-order correlation coefficient is equivalent to a z-test for the number of agreements. By constructing the test through the number of agreements, the computation is reduced by half.

The test is reasonably powerful to reject the null hypothesis when the majority of the observations are random representations of a sequence of completely ordered (differentiated) true merits. However, when only a few objects have a sequence of ordered \(\alpha\) values and the majority of the objects are similar in their \(\alpha\)’s, those undifferentiated objects will result in large numbers of random agreements and disagreements, and thus add substantial random noises to the true association on the few “important” cases. As a result, the
power of the test based on the Kendall rank-order correlation coefficients diminishes under local alternatives even the local association exists and is strong. Simulation results under two sets of local alternatives can be found in Figure 4. The powers of the Kendall rank-order correlation coefficients under different alternatives is the right ends of the curves. It is shown that the power is more affected when the number of important objects is smaller. More details and discussion on these examples are in section 3.2 and the simulation details are discussed in section 3.1.

2.4. Failure of the test based on the number of overlapping top objects. When the top objects are concerned, it is natural to consider a test based on the number of overlapping top objects as a solution to the problem in section 2.2.

Denote $k$ as the number of top objects that are considered. The number of overlapping top objects between ranking $X$ and ranking $Y$ can then be defined as

$$O = \sum_{i=1}^{n} \mathbf{1}_{(X_i \leq k)} \mathbf{1}_{(Y_i \leq k)}.$$  

It is easy to show that, under the null hypothesis,

$$E(O) = \frac{k^2}{n}, \quad \text{Var}(O) = \frac{k^2}{n} + \frac{k^2(k-1)^2}{n(n-1)} - \frac{k^4}{n^2}.$$  

Thus a simple test can be constructed using $\frac{O - E(O)}{\sqrt{\text{Var}(O)}}$, while the sampling distribution is approximated by the standard normal distribution. The exact significance can also be evaluated based the values of $k$ and $n$ under the null hypothesis through permutations.

The power of this test is satisfactory when the specification of $k$ is close to the true number of important objects, under the local alternatives we evaluate in this paper. Actually, under such ideal situations, the performance of this test is nearly comparable to that of the modified Kendall rank-order association test, which is to be discussed in the next section. However, the performance of the test based on the number of overlapping top objects deteriorates dramatically when the specified value of $k$ departs from the ideal specification, $k_0$ (the true number of objects with higher true merits). Such a trend, found in the simulation studies, is shown in Figure 5. Details of this example will be discussed in section 3.2.

2.5. Modified Kendall rank-order association test. In this subsection, we propose a solution to the problem raised in section 2.2 by evaluating the correlation between two rankings, while only a small number,
when $k \ll n$, of the top ranks are actually considered. The value of $k$ is pre-specified, which can be based on the knowledge of $k_0$ or inferred from previous results regarding the data.

The modified Kendall rank-order test statistic that we propose to use is based on the truncated ranks defined as $X^*_i = \min(X_i, k)$. Then the number of agreements as defined in the Kendall rank-order correlation coefficient can be calculated based on these truncated ranks. The test statistic is then the number of agreements standardized by the mean and standard deviation of its sampling distribution under the null hypothesis. Similar to the formulation of (2.3) and (2.4), we have,

\begin{equation}
\text{(2.10)} \quad \# \text{ agreements} = \sum_{i=1}^{n} \sum_{j \neq i} 1(X^*_i < X^*_j) 1(Y^*_i < Y^*_j) = \sum_{i=1}^{n} \sum_{j \neq i} 1(\min(X_i, k) < \min(X_j, k)) 1(\min(Y_i, k) < \min(Y_j, k))
\end{equation}

\begin{equation}
\text{(2.11)} \quad \# \text{ disagreements} = \sum_{i=1}^{n} \sum_{j \neq i} 1(X^*_i < X^*_j) 1(Y^*_i > Y^*_j) = \sum_{i=1}^{n} \sum_{j \neq i} 1(\min(X_i, k) < \min(X_j, k)) 1(\max(Y_i, k) > \max(Y_j, k))
\end{equation}

Under the null hypothesis, the ranking $X = \{X_1, \ldots, X_n\}$ is a random permutation of $1, \ldots, n$. Using (2.10) and (2.11), it is easy to derive that

\begin{equation}
E(\# \text{ agreements}) = E(\# \text{ disagreements}) = \frac{1}{4} n(n-1) \left(1 - \frac{(n-k+1)(n-k)}{n(n-1)}\right)^2.
\end{equation}

The variance of the number of the agreements under the null hypothesis can also be calculated as in (2.6).

\[
\text{Var}(\# \text{ agreements}) = \text{Var} \left( \sum_{i=1}^{n} \sum_{j \neq i} 1(X^*_i < X^*_j) 1(Y^*_i < Y^*_j) \right) = \sum_{i=1}^{n} \text{Var} \left( 1(X^*_i < X^*_j) 1(Y^*_i < Y^*_j) \right) + \sum_{i=1}^{n} \sum_{j \neq i} \sum_{(k,l) \neq (i,j)} \text{Cov} \left( 1(X^*_i < X^*_j) 1(Y^*_i < Y^*_j), 1(X^*_k < X^*_l) 1(Y^*_k < Y^*_l) \right) \\
= n(n-1) \left\{ \frac{1}{4} n(n-1) \left(1 - \frac{(n-k+1)(n-k)}{n(n-1)}\right)^2 \right. \\
+ (n-2)(n-3) \left( \frac{1}{4} \left( \frac{k-1}{\binom{n}{4}} \right) + \frac{1}{4} \left( \frac{n-k+1}{\binom{n}{4}} \right) + \frac{1}{6} \left( \frac{k-1}{\binom{n}{4}} \right)^2 \right) \\
+ (n-2)^2 \left( \frac{1}{9} \left( \frac{\binom{k}{3}}{\binom{n}{3}} \right)^2 \right) \\
+ (n-2)^2 \left( \frac{1}{9} \left( 1 - \frac{(n-k+1)^2}{\binom{n}{3}} \right)^2 \right) \\
\left. + \frac{1}{16} (n^2 - n) \left(1 - \frac{(n-k+1)(n-k)}{n(n-1)}\right)^4 \right\}.
\]
The modified Kendall rank-order test statistic is then defined as

\[ T^c = \frac{\# \text{ agreements} - E(\# \text{ agreements})}{\sqrt{\text{Var}(\# \text{ agreements})}}, \]

where the sampling distribution is approximated by the standard normal distribution. The exact significance can also be evaluated based the values of \( k \) and \( n \) under the null hypothesis through permutations.

**Figure 2.** Sampling distributions using truncated rankings versus full rankings under the null and alternative hypotheses. Each distribution is based on 5000 simulations on 1000 objects. The alternative hypothesis used is specified as in Figure 3 with \( \delta = 5 \) and \( k_0 = 50 \). The smooth curves in the plots represent the standard normal distribution used in the approximation of p-values.

The advantage of the modified statistic can be best illustrated as in Figure 2. The sampling distributions are simulated under the same null and alternative hypotheses for the test statistics of the original Kendall rank-order correlation coefficient and the modified Kendall rank-order association test statistic. For
the modified statistic, the sampling distribution under the alternative is well separated from the sampling distribution under the null hypothesis, while the distributions for the original statistic have a substantial overlap. It indicates that by focusing on the top ranks, as for the modified statistic, the signal becomes stronger because of the removal of a substantial amount of noise.

3. Simulations and Results

3.1. Simulation models. The local alternative can take a lot of forms of departures from the null hypothesis. For the evaluation of the proposed method, here we consider a class of alternatives of the form as shown in Figure 3. As mentioned in section 2.2, we assume the true merits $\alpha$’s are sorted. In each of the simulation model, a number, $k_0$, of $\alpha$’s are set to be higher than the rest of the objects that have identical merit values. For those that have higher merit values, we specify the values to have linear increments, while the highest $\alpha$ value, $\alpha_1$, is $\delta$ higher than the value of the undifferentiated objects. In other words, $\alpha_{i-1} - \alpha_i = \delta/k_0$, for $2 \leq i \leq k_0$ and $\alpha_{k_0} = 0$ for $k_0 + 1 \leq i \leq n$. For convenient simulations, $F$ and $G$ are assumed to be normal distributions with mean 0 and standard deviation $\sigma$.

![Simulation model diagram]

**Figure 3.** Alternative model used for simulations

3.2. Results. For the class of alternatives we are considering in the simulation studies, the strength of the signal from the objects with higher $\alpha$ values depends on the elevation of merit, $\delta$, and the noise standard deviation $\sigma$. It is expected that the higher the ratio between $\delta$ and $\sigma$, the more distinct the top objects are from the rest of the evaluation pool. Without loss of generality, we fix $\sigma$ to be 1 and only vary the value of $\delta$.

In Figure 4, power performance of the modified Kendall rank-order test on simulated data with 500 objects are plotted at each possible truncation values, $k$, under two specifications of $\delta$ and different values
of $k_0$. When $k = n$, the modified test becomes the original Kendall rank-order test. First, we observe that the power curves attain the peaks around the “right” specification of $k$, i.e., around the real value of $k_0$. Trimming too many objects (smaller values $k$) and too few (larger values of $k$) both result in loss of power. The most striking performance gain due to the advantage of the modified rank-order test is observed when the signal is weaker and the number of the objects with higher merits is smaller (say, $\delta = 3$, $k_0=10$). As shown in this example, the original Kendall rank-order statistic has little power when the proportion of true signal is less than 5%, while the modified test maintains a power of higher than 70% for a reasonable range of $k$ values around the true (unknown) value of $k_0$.

Figure 4. Power performance of the modified Kendall rank-order association test under two sets of alternatives. The power is estimated using 500 simulations. The models used for simulation are specified as in Figure 3 with $\delta = 3$ versus $\delta = 9$, while $k_0$ takes four different values.

Figure 5 shows the comparison of the power performance between the modified rank-order test and the test based on the number of overlapping top objects as described in section 2.4. The models used for this comparison have $\delta = 3$ and $\sigma = 1$, while $k_0$ varies. Considering the test based on the number of overlapping top objects, the clearest pattern in Figure 5 is that the power curves drop dramatically as the specified truncation number $k$ departs from the true value $k_0$. This method becomes powerless when the value $k$ get
closer to the total number of the objects, \( n \), by definition. On the other hand, the signal of the top objects is better preserved and reflected by the modified rank-order test, when \( k \) differs from \( k_0 \). This is due to the fact that the test using the overlap of top objects only reduces noises from the undifferentiated objects, while the modified rank-order test takes into account the informative order of the top ranks through the use of the truncated ranks and gain more power to detect the real signals.

![Diagram](image)

**Figure 5.** Power performance of the modified Kendall rank-order association test and the test based on the overlapping top objects. The power is estimated using 500 simulations. The model used for simulation is specified as in Figure 3 with \( \delta = 3 \) and \( \sigma = 1 \), while \( k_0 \) takes four different values.

4. Discussion and conclusion

4.1. **Departure from normality.** The sampling distributions of the modified Kendall rank-order test statistic can be well approximated by normal distributions when \( k \) is not very small. However, as based on the truncated ranks, when \( k \) and \( n \) are both small (say \( n \leq 30, k \leq 5 \)), the sampling distribution under the null hypothesis becomes more discrete. This is because, conditioning on the total number of agreements and disagreements combined, the distribution of the number of agreements is a binomial distribution with
probability 0.5. The combined total number of agreements and disagreements equals the number of informative pairs of objects (those that have at least one falling into the top \( k \) ranks in both \( X \) and \( Y \) rankings). This number becomes very small when \( n \) and \( k/n \) are both small. Actually, the expected total number of agreements and disagreements under the null hypothesis is

\[
E(\text{# agreements} + \text{# disagreements} \mid H_0) = \frac{n(n-1)}{2} \left( 1 - \left( \frac{n-k}{n-1} \right) \right)^2
\]

\[
= \frac{(2n-k-1)^2k^2}{2n(n-1)}.
\]

which is approximately \( 2k^2 \), for most cases we consider in this paper, where \( n \) is large and \( k \ll n \). Thus, the test derived based on the normal approximation should be acceptable when \( k \geq 10 \) and \( k \ll n \).

4.2. Disagreements versus agreements. The original Kendall rank-order correlation coefficient is calculated using both the number of agreements and disagreements, while adjusting for the numbers of ties in both the \( X \) and \( Y \) rankings. For the truncated ranks, the mean and variance of such a score becomes very complicated to evaluate. It is believed, though, that the number of disagreements should also contain useful information. This is an issue that remains to be addressed in the future.

4.3. More general alternatives. Other forms of departures from the null hypothesis can also take place in different contexts of research projects. For instance, in the microarray data example considered in the introduction, if we look at the coefficient of correlation, we should look at both ends of the fully-sorted “merit” list. On the positive value end will be the genes highly expressed in good prognosis patients but are suppressed in poor prognosis patients, while genes that expressed in good prognosis patients but not poor prognosis patients will have correlation values clustered around the lower end (large negative values). For a situation like this, a natural alteration of the proposed method will be changing the right-truncation of the rank used in this paper to center truncation, that is, modifying the rank by assigning a single rank value to the mid portion of the sorted observations.

4.4. Computational considerations. The computational complexity of the test statistic is the same as that of the original Kendall rank-order correlation coefficient, which is \( O(n^2) \). For a large value of \( n \), the sorting of the objects and the counting of the agreements can be time-consuming. However, the amount of computation needed can be handled by current computing power in a reasonably short time. For the example we have used in the introduction section, the calculation for 24479 genes took about 3 minutes to finish using R on a Pentium-IV PC, which should be acceptable by current computing standards and can be much improved using scientific computing languages such as C or Fortran. The mean and standard deviation
for the normal approximation only depend on the specification of $n$ and $k$, which is easy to calculate and can be prepared in advance.

4.5. **Conclusion.** In this paper, we proposed a modified Kendall rank-order association test, mainly for studying the repeatability of two studies on a large number of objects, most of which are undifferentiated. The method is to address new issues posed by the large noise-signal-ratio of current data sets. The test statistic is intuitional, based on the belief of the true signal, easy to implement and fast to calculate. The exact sampling distribution can be derived using permutation simulations or conveniently approximated by normal distributions in most practical situations. Simulations on a class of general alternatives show the substantial gain of power due to the proposed modification, comparing to the original Kendall rank-order coefficient of correlation. Through the use of rank-order based on truncated ranks, the test statistic still manages to capture the informative order of the objects with higher merits while removes the noises from the undifferentiated objects. We believe this new test can find important applications in current statistical studies that involves large number of objects of evaluation.

**References**
