On Rank Correlation and the Distance Between Rankings

Ben Carterette
Department of Computer & Information Sciences
University of Delaware
Newark, DE, USA 19716
carteret@cis.udel.edu

ABSTRACT

Rank correlation statistics are useful for determining whether a there is a correspondence between two measurements, particularly when the measures themselves are of less interest than their relative ordering. Kendall's τ in particular has found use in Information Retrieval as a "meta-evaluation" measure: it has been used to compare evaluation measures, evaluate system rankings, and evaluate predicted performance. In the meta-evaluation domain, however, correlations between systems confound relationships between measurements, practically guaranteeing a positive and significant estimate of τ regardless of any actual correlation between the measurements. We introduce an alternative measure of distance between rankings that corrects this by explicitly accounting for correlations between systems over a sample of topics, and moreover has a probabilistic interpretation for use in a test of statistical significance. We validate our measure with theory, simulated data, and experiment.

Categories and Subject Descriptors: H.3 [**Information Storage and Retrieval**]; H.3.4 [**Systems and Software**]: Performance Evaluation

General Terms: Experimentation, Measurement

Keywords: information retrieval, evaluation, rank correlation, distance measure

1. INTRODUCTION

Ranking is a ubiquitous problem in Information Retrieval. Retrieval systems rank documents by estimated relevance; evaluations like the Text Retrieval Conference (TREC) rank systems by evaluation measures; systems rank queries by predicted difficulty; users rank systems by preference. Evaluating a ranking requires some measure of comparison between two rankings. In the case of ranking documents, there are a wide variety of evaluation measures that implicitly compare a ranking of documents to a perfect ranking; for the other tasks listed above, rank correlation measures, in particular Kendall's τ, have become *de facto* standards.

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.

Kendall's τ is appealingly intuitive: given two different rankings of the same m items, count the number of pairs that are concordant—in the same order in both rankings—and discordant—in reverse order. If P is the concordance count and Q the discordance count,

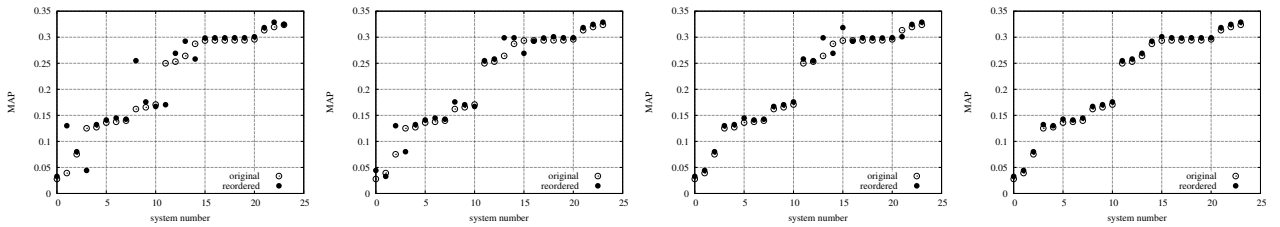
$$\tau = \frac{P-Q}{P+Q}.$$

τ ranges from -1 to 1, with 1 meaning the two rankings are identical and -1 meaning one is in reverse of the other. A τ of 0 means that 50% of the pairs are concordant and 50% discordant. Every value of τ maps directly to a percentage of concordant pairs (assuming no ties).

Maurice Kendall introduced τ in 1938 and subsequently developed much of its theory in his 1948 monograph *Rank Correlation Methods* [7]. As Kendall explains, rank correlation methods are ideally suited to situations where measurements are subjective or difficult in practice. However, correlation measures are meaningful when samples are drawn independently and with identical sampling distributions (i.i.d.). When samples are not i.i.d., the interpretation becomes unclear: correlation between items in a sample can confound correlation in the measurements. Suppose, for example, we want to measure the rank correlation between two evaluation measures. If we calculate τ between rankings of a system that is nearly perfect, a middle-of-the-road system, and a system with a fatal bug, it does not matter how similar the measures are. If they capture anything at all about performance, they are guaranteed to be highly correlated.

Some recent work has revealed strange behavior by Kendall's τ when comparing rankings of systems [2, 8, 11, 13]. At a high level, much of this can be explained by correlations between systems: when they are ranking the same documents for the same topics, they will be so highly correlated that it is unclear whether τ has any meaning whatsoever. In Section 2 we explain the high-level problem in more detail, along with other hurdles to interpreting a τ correlation.

If our samples are not independent and identically distributed, a measure of distance between rankings should take into account the likelihood of the particular rankings. In Section 3 we develop such a measure, along with a significance test for the hypothesis that two rankings are the same. In Section 4 we present theoretical results about our measure and compare our rank distance measure to Kendall's τ over random rerankings as well as some classic experiments from IR evaluation studies.



(a) $\tau = 0.91; d_{rank} = 14.74$. (b) $\tau = 0.90; d_{rank} = 7.90$. (c) $\tau = 0.89; d_{rank} = 3.52$. (d) $\tau = 0.90; d_{rank} = 0.60$.

Figure 1: Four different reorderings of 24 systems over 149 topics. Open circles are "true" *MAP*; filled circles are some estimate. Kendall's τ is 0.9 ± 0.01 for all four; our measure d_{rank} decreases from left to right.

2. THE TROUBLE WITH KENDALL'S TAU

Kendall's τ and other rank correlation statistics make particular assumptions that result in unclear interpretation in some of the situations they are used in information retrieval. In particular, its use in meta-evaluation studies carries some significant problems. Consider Figure 1, which shows 24 retrieval systems submitted to the TREC 2007 Million Query Track [1] sorted by increasing *MAP* over 149 topics (TREC topics 701-850). Note that there are "bins" of roughly equal-quality systems. Four reorderings are shown. Visual inspection suggests they are getting better going from left to right: Figure 1(a) shows systems swapped between bins (in particular the eighth-ranked system has moved up to rank 12); Figure 1(b) shows smaller between-bin swaps; Figure 1(c) shows no between-bin swaps but some big within-bin swaps; and Figure 1(d) shows only small (visually almost imperceptible) within-bin swaps. Kendall's τ cannot pick up on any difference between these: each of the four rankings has a τ of 0.9 ± 0.01 with the original ranking. This highlights the first, and most obvious, problem with Kendall's τ and other rank correlation statistics: they treat all pairwise swaps equally, even when the probability of a swap is low.

A related but more subtle problem is that Kendall's τ treats all swaps as statistically *independent*: it assumes that if there is no correlation, then knowing that systems i and k swapped says nothing about whether j and k swapped. This is certainly not true of retrieval systems; for example, if systems i and j are very similar to one another, a swap between j and k is much more likely given a swap between i and k. Intuitively, very similar systems will tend to "stick together". The result of this is that τ has a very high baseline even when there is no correlation, much higher than the $\tau = 0$ that it is supposed to have for uncorrelated variables. This is nicely illustrated by Soboroff et al., who show that correlations around 0.4–0.5 are achieved even when relevance is assigned to retrieved documents randomly [13].

The variance in τ over a sample of systems is seldom discussed. A sample of m systems with correlation τ between two measurements has 95% confidence interval [7]:

$$\left(\tau \pm 1.96\sqrt{\tfrac{2}{m}}\sqrt{1 + \tfrac{2 \cdot 1.96^2}{m} - \tau^2}\right) / \left(1 + \tfrac{2 \cdot 1.96^2}{m}\right) \qquad (1)$$

This means that with 25 systems that have a τ of 0.9, the true correlation is somewhere between 0.389 and 0.987 with 95% confidence. This is an unacceptably large range for the way τ is used in IR. In general the variance is of order $\sqrt{2/m}$, and it is not easy to reduce it by sampling additional systems. Contrast this to measures like mean average precision calculated over a set of n topics: the variance in the mean is of order $\sqrt{1/n}$, and in principle (if not in practice) we can reduce variance by using more topics.

We "know" the variance is actually much less than this because we consistently see similar τ correlations in meta-evaluation studies and between different evaluation measures. These observations are another consequence of the systems not being independent. We argue that the sample space we are actually interested in is the topic space: given that two rankings of systems are correlated over a particular set of topics, would they still be correlated, and would the correlation be as high, if run over a different set?

To summarize, the problems with Kendall's τ (or any rank correlation statistic) in meta-evaluation studies are:

1. all pairs are treated equally;
2. pairs are assumed statistically independent;
3. sample space orthogonal to the space of interest;
4. high variance over the system sample space.

Taken together, these make τ difficult to use for anything but the most high-level approximation of rank similarity.

2.1 Proposed Solutions

The first problem we discuss above—the fact that τ treats all pairs equally—has probably been noticed by anyone who has used τ to evaluate a ranking of systems. Nearly all the proposed alternatives to τ and other rank correlation measures address this problem.

Cormack and Lynam proposed estimating power and bias of pairwise system comparisons by effectively calculating τ over only pairs with significant differences [5] (Sakai independently did something similar, naming it "discriminative power" [10]). Yilmaz et al. proposed a "top-heavy" variant that uses average precision-like weights to make higher-ranked systems more important [17]. Melucci proposed that τ be replaced by a different rank correlation statistic, Kolmogorov-Smirnov's D [8]. None of these address the problems of independence or of the sample space.

Discriminative power is the only one to take into account variance over the topic space (via its determination of statistical significance). It does not, however, eliminate the system space as a source of variance. Moreover, because the pairs that are significantly different are the ones most likely to be agreed upon, and because there is even greater correlation among significant pairs than randomly-chosen pairs, discriminative power will be very high by default. Taking a very high baseline together with wide confidence intervals results in a statistic for which the upper bound (1.0) is almost always within the standard error.

We consider discriminative power to be a step in the right direction: it takes the topic sample into account, and makes use of differences between systems in the form of paired significance tests. It is simply not a big enough step. Instead of comparing each pair of systems independently, we must compare all the systems simultaneously. We must eliminate any variance due to systems, since it is likely impossible to have anything resembling a random sample of systems. Finally, we must be able to test hypotheses about rankings over a sample of topics.

Our goal, therefore, is to develop a measure of rank similarity with the following properties:

1. penalizes swaps between items that are very different more than swaps between items that are similar;
2. penalizes swaps between items conditional on swaps among the other pairs;
3. assumes a fixed population of systems with a random sample of topics.

Number 1 ensures that pairs are not all treated equally. Number 2 explicitly incorporates correlations among pairs. Number 3 corrects the sample space. The next section describes our measure.

3. RANK DISTANCE

One way to accomplish all three criteria above is to estimate the *probability* of observing a particular alternative ranking of systems given a baseline ranking based on measurements of system results over a sample of topics. This is what our rank distance measure does. We will present the measure first, then justify it with examples and analogies.

Suppose we have an $n \times m$ matrix $\mathbf{X}$ of average precision values for m systems over n topics.[1] Let $\boldsymbol{\mu}$ be a vector of m mean APs over n topics. Let $\mathbf{y}$ be a vector of m means of some other evaluation measure that provides an alternative ranking of systems. Reorder $\boldsymbol{\mu}$ and the columns of $\mathbf{X}$ in increasing order of $\mathbf{y}$. The rank distance from $\boldsymbol{\mu}$ to $\mathbf{y}$ is:

$$d^2_{rank}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{X}) = \min_{\boldsymbol{\theta}} n(\boldsymbol{\theta} - \boldsymbol{\mu}_\Delta)' {\boldsymbol{\Sigma}_\Delta}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_\Delta) \quad (2)$$
$$\text{s.t. } \theta_i > 0 \quad \forall i$$

where $\boldsymbol{\mu}_\Delta$ is the vector of differences between adjacent values in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_\Delta$ is the covariance matrix for differences in AP between adjacent columns in $\mathbf{X}$. Essentially, we find a point $\boldsymbol{\theta}$ that has maximum probability under a distribution defined by $\mathbf{X}$ subject to the constraint that it preserves the same ranking as $\mathbf{y}$. Algorithm 1 shows pseudocode for setting up and solving the minimization problem; more detail is provided below.

3.1 Derivation (by Example)

Consider the following example: average precision and precision@10 computed for four topics run by three systems.

$$AP = \begin{bmatrix} 0.283 & 0.481 & 0.516 \\ 0.017 & 0.399 & 0.544 \\ 0.075 & 0.300 & 0.277 \\ 0.183 & 0.662 & 0.616 \end{bmatrix} \quad p@10 = \begin{bmatrix} 0.8 & 0.8 & 0.8 \\ 0.2 & 0.7 & 0.5 \\ 0.3 & 0.5 & 0.5 \\ 0.7 & 1.0 & 1.0 \end{bmatrix}$$

where rows correspond to topics and columns to systems. The system means are:

$$MAP = \begin{bmatrix} 0.140 & 0.461 & 0.488 \end{bmatrix}' \quad p@10 = \begin{bmatrix} 0.5 & 0.75 & 0.7 \end{bmatrix}'$$

[1]We use bold letters to denote vectors and matrices.

Let us refer to the systems as A, B, and C by column. Clearly system A is much worse than B or C by *MAP*; in fact, with only four topics, the differences are significant by a paired t-test ($p < 0.05$). Systems B and C are similar and not significantly different. System A performs worse by precision@10 as well, though not significantly so. Systems B and C have swapped by mean precision@10.

To make our requirements concrete, we desire a distance measure that penalizes swaps between A, B or A, C much more harshly than swaps between B, C; penalizes separation of B and C, e.g. placing A in between them; and has variance over the topic sample rather than the system sample.

Consider the Mahalanobis distance [16]: if $\mathbf{x}$ and $\mathbf{y}$ are vectors drawn from a m-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then the Mahalanobis distance between $\mathbf{x}$ and $\mathbf{y}$ is defined to be:

$$d^2(\mathbf{x}, \mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{x} - \mathbf{y})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mathbf{y}).$$

This is not a distance between rankings, as $\mathbf{x}$ and $\mathbf{y}$ are not lists of ranks.[2] It is, however, inversely proportional to the probability of observing the vector $\mathbf{x}$ assuming the true mean is $\mathbf{y}$ [6]. Note that with a one-dimensional normal distribution with mean μ and variance σ^2, Mahalanobis distance reduces to square of the common Z-statistic $(x - y)^2/\sigma^2$.

If $\mathbf{x}$ and $\mathbf{y}$ are vectors of means (i.e. each value in $\mathbf{x}$ is the average of measurements to a random sample of n items such as topics), there is an analogue to the Mahalanobis distance in the same way the t-statistic is an analogue to the Z-statistic. It is called *Hotelling's* T^2:

$$T^2 = nd^2 = n(\mathbf{x} - \mathbf{y})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})$$

T^2 is inversely proportional to the probability of observing mean vector $\mathbf{x}$ assuming mean vector $\mathbf{y}$ is the true mean [6]. In the one-dimensional case, this becomes $n(x - y)^2/\sigma^2 = (x - y)^2/(\sigma^2/n)$, which should be recognizable as the square of the t-statistic.

We are interested in probabilities of rankings, so d^2 and T^2 are not directly applicable. But a ranking can be defined in terms of the signs of differences between values. Sort the columns of both matrices in increasing order of p@10 and take the difference between adjacent columns:

$$AP_\Delta = \begin{bmatrix} 0.233 & -0.035 \\ 0.527 & -0.145 \\ 0.202 & 0.023 \\ 0.433 & 0.046 \end{bmatrix} \quad p@10_\Delta = \begin{bmatrix} 0.0 & 0.0 \\ 0.5 & 0.2 \\ 0.2 & 0.0 \\ 0.3 & 0.0 \end{bmatrix}$$

The system mean differences are:

$$MAP_\Delta = \begin{bmatrix} 0.349 & -0.028 \end{bmatrix}' \quad p@10_\Delta = \begin{bmatrix} 0.2 & 0.05 \end{bmatrix}'$$

from which we can tell that $B > C > A$ by p@10. The reason for reordering the columns is that it ensures all differences in mean p@10 will be positive, and therefore that the vector has no ambiguity about the ranking.

This effectively reduces a ranking to a point in $(m - 1)$-dimensional space. Covariance matrices determine which points in that space—that is, which rankings—might have

[2]If we map values in $\mathbf{x}$ and $\mathbf{y}$ to their respective ranks and set $\boldsymbol{\Sigma}$ to the $m \times m$ identity matrix, the resulting Mahalanobis distance is proportional to Spearman's ρ. In other words, ignoring correlations between items produces a common rank correlation statistic. This shows explicitly how a traditional rank correlation statistic ignores dependence.

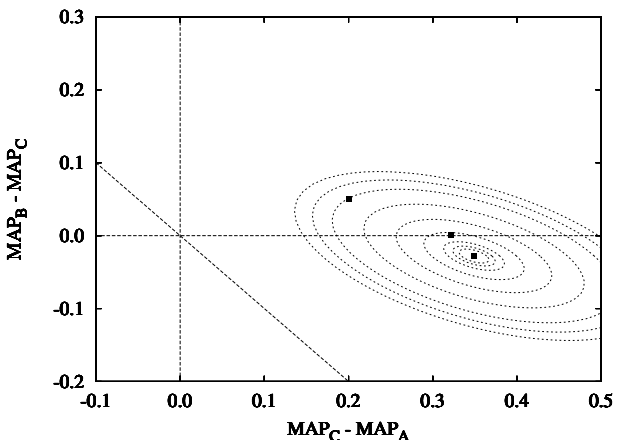


Figure 2: Contours of the bivariate normal distribution $\boldsymbol{\mu}_{\boldsymbol{\Delta}} = [0.349\ -0.028]'$, $n\boldsymbol{\Sigma}_{\boldsymbol{\Delta}} = \begin{bmatrix} 0.025 & -0.007 \\ -0.007 & 0.007 \end{bmatrix}$. Points are the mean, the precision values $(0.20, 0.05)$, and the minimum-distance rank-preserving point $(0.322, \epsilon)$.

high or low probability. The covariances are:

$$\Sigma_{AP_\Delta} = \begin{bmatrix} 0.0246 & -0.0071 \\ -0.0071 & 0.0073 \end{bmatrix} \quad \Sigma_{p@10_\Delta} = \begin{bmatrix} 0.02 & 0.01 \\ 0.01 & 0.01 \end{bmatrix}$$

Figure 2 shows a contour plot of the joint distribution of $MAP_C - MAP_A$ and $MAP_B - MAP_C$. The point in the center is the mean, $(.349, -.028)$. Precision@10 is at $(.2, 0.05)$. The innermost contour line is the 2.5% confidence interval where T^2 values are lowest; the outermost is the 97.5% interval (i.e. 97.5% of the total probability is inside the outermost contour). T^2 values get progressively higher for points radiating out from the mean.

Note that there are regions in which every point defines the same relative ordering of systems. In Quadrant I (upper right), we have $MAP_C > MAP_A$ and $MAP_B > MAP_C$, and therefore the ranking is $B > C > A$. In Quadrant IV (lower right), we have $MAP_C > MAP_A$ and $MAP_B < MAP_C$. The ranking could be $B > C > A$ or $B > A > C$. The diagonal line separates the two possibilities; note that most of the probability mass is in the $C > B > A$ region, representing the "true" ranking by *MAP*. Quadrants II and III (upper left and lower left, respectively) are where $MAP_A > MAP_C$.

The regions representing rankings in which pairs with significant differences have been swapped have negligible probability. All of the probability mass is split between two regions in which the ranking is correct: either $C > B > A$ or $B > C > A$. We want to leverage this property, which holds in general, for our distance measure.

Even though every point in Quadrant I represents an acceptable ranking, and every point in the upper region of Quadrant IV represents the "true" ranking, many of these points have very large T^2 from the mean and thus effectively zero probability. The point $(0.2, -0.1)$, for instance, represents a correct ranking but is so far outside the contours that it has negligible probability. The point $(0.20, 0.05)$, which is the mean of precision@10 in our example, is in an outer contour and thus has relatively low probability despite being an acceptable ranking. In this respect these points are indistinguishable from points that represent entirely wrong rankings. This is a problem that needs to be resolved.

Algorithm 1 Calculate d_{rank} from a vector of means $\mathbf{y}$ to a vector of means $\boldsymbol{\mu}$.

Require: vector $\mathbf{y}$ of means, $n \times m$ matrix $\mathbf{X}$ of values, vector $\boldsymbol{\mu}$ of column means of $\mathbf{X}$.

1: $i_k \leftarrow$ the index of the kth-greatest value of $\mathbf{y}$. (e.g. i_1 is the index of the maximum value, i_2 the index of the second-greatest value, etc.).
2: $\mathbf{X}_{\boldsymbol{\Delta}} \leftarrow \begin{bmatrix} \mathbf{X}_{\cdot i_1} - \mathbf{X}_{\cdot i_2} & \mathbf{X}_{\cdot i_2} - \mathbf{X}_{\cdot i_3} & \dots & \mathbf{X}_{\cdot i_{m-1}} - \mathbf{X}_{\cdot i_m} \end{bmatrix}$, where $\mathbf{X}_{\cdot j}$ is the jth column of $\mathbf{X}$.
3: $\boldsymbol{\mu}_\Delta \leftarrow \begin{bmatrix} \mu_{i_1} - \mu_{i_2} & \mu_{i_2} - \mu_{i_3} & \dots & \mu_{i_{m-1}} - \mu_{i_m} \end{bmatrix}'$.
4: $\boldsymbol{\Sigma}_{\boldsymbol{\Delta}} \leftarrow (\mathbf{X}_{\boldsymbol{\Delta}} - \boldsymbol{\mu}_\Delta)(\mathbf{X}_{\boldsymbol{\Delta}} - \boldsymbol{\mu}_\Delta)'/(n-1) + \lambda\mathbf{I}$ (i.e. $\mathrm{Cov}(\mathbf{X}_{\boldsymbol{\Delta}})$ regularized by λ; $\lambda > 0$ iff $m \geq n$).
5: $\boldsymbol{\theta} \leftarrow \mathrm{QPSOLVE}(n{\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}}^{-1}, -\boldsymbol{\mu}_\Delta)$.
6: $d^2_{rank} \leftarrow n(\boldsymbol{\theta} - \boldsymbol{\mu}_\Delta){\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_\Delta)$.
7: **return** $|\sqrt{d^2_{rank}}|$.

Instead of measuring the distance from a particular point to the mean, we will find a point in the same region—i.e. another point that preserves the ranking—that has *minimum distance* to the mean. This point is unique in each region. In our example, the minimum-distance rank-preserving point for the differences in mean precision@10 is at $(0.322, \epsilon)$, where ϵ is an infinitesimal number greater than zero. Being in the inner contours, that point has fairly high probability; its T^2 distance from the mean is 0.65, which is quite low. Based on that, we can conclude that the ranking by precision at 10 is not very different from a ranking by *MAP* for these three systems.

Thus we have our measure in Eq. 2. Given a vector of means $\mathbf{y}$ and a matrix of AP or other values $\mathbf{X}$ with means $\boldsymbol{\mu}$, the rank distance from $\mathbf{y}$ to $\boldsymbol{\mu}$ is:

$$d^2_{rank}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{X}) = \min_{\boldsymbol{\theta}} n(\boldsymbol{\theta} - \boldsymbol{\mu}_\Delta)'{\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_\Delta)$$
$$\text{s.t. } \theta_i > 0 \quad \forall i$$

where $\boldsymbol{\theta}$ is the minimum-distance rank-preserving point. The constraints $\theta_i > 0$ ensure that whatever vector of differences minimizes the distance preserves the ordering of values in $\mathbf{y}$.

Algorithm 1 shows how to set up the problem. At line 2, $\mathbf{X}_{\cdot j}$ is the jth column of $\mathbf{X}$, i.e. the values of AP or another measure by system j for n topics. Subtracting columns finds the differences in AP between systems that are adjacent in the ranking by $\mathbf{y}$, exactly as $(m-1)$ paired t-tests would. The covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}$ ensures that the t-tests are not performed independently. When $m \geq n$, the covariance matrix is singular and noninvertible; we work around that case by regularizing $\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}$ (line 4). A simple regularization procedure adds a constant λ to the diagonal [15]; we used $\lambda \approx 10^{-5}$. Since covariance matrices are always positive semidefinite, the formula is convex and solvable in polynomial time using quadratic programming methods [3]. The algorithm assumes a quadratic program solver $\mathrm{QPSOLVE}(\mathbf{A}, \mathbf{b_0})$ that minimizes $\mathbf{b}'\mathbf{A}\mathbf{b}$ w.r.t. $\mathbf{b}$ s.t. $\mathbf{b} > \mathbf{b_0}$. Our unoptimized implementation in `R` ran in 0.003 seconds on average, compared to 0.0004 seconds for `R`'s built-in implementation of Kendall's τ.

To conclude this section, consider the following table.

ranking	ABC	ACB	BAC	CAB	BCA	CBA
τ	-1	-1/3	-1/3	1/3	1/3	1
dsc pow	0	0	1/2	1/2	1	1
d_{rank}	4.88	4.88	4.88	4.88	0.65	0.00

While τ and discriminative power see differences among the first four rankings, d_{rank} effectively considers them all so bad as to be beyond repair. The first two have swapped both significant pairs by putting A first. The second two have only swapped one significant pair, but have also violated the rule that very similar items should not be separated. Discriminative power cannot capture this case and thus views these two as better than the former two. The last two, which only swap B and C, are the best, but d_{rank} can make a subtle distinction between them while discriminative power cannot. Kendall's τ makes an extreme distinction between them, saying that swapping B and C is as bad as swapping A and B.

Our measure has an interpretation problem: in principle it is unbounded, with larger values meaning greater distances. What is a "good" rank distance? Is 0.65 good? In fact, the magnitude of rank distance is highly dependent on the number of systems being ranked, the number of topics they are evaluated over, and the overall interdependence of the systems by the baseline measure. Fortunately, we can avoid worrying about a direct interpretation by introducing a statistical hypothesis test for rank distance with an easy-to-interpret p value.

3.2 Rank Distance Hypothesis Test

Statistical hypothesis tests are used to estimate the probability that two measurements are different, or that a measurement is different from some value, given a sample of items. In our case, we wish to estimate the probability that d_{rank} is zero given a sample of topics. If it is zero, the rankings are identical. If it is far from zero, the rankings are significantly different. When it is slightly greater than zero due to a few pairwise swaps, the question is whether the swaps are "acceptable" given the variance in the topic space: if we used a different sample of topics, would we see a similar result? If we could evaluate over the entire population of topics, would the rankings be identical?

Many significance tests are based on knowing the distribution of the test statistic under the null hypothesis. For the t-test, the t-statistic has a t distribution. Hotelling's T^2 has an F distribution. Given these distributions, we can determine the probability (p-value) of a particular value of t or T^2 under the null hypothesis; if it is low, we reject the null hypothesis that two values are the same. Our rank distance measure has a distribution defined in terms of the cumulative density function of the multivariate normal distribution: the probability of a particular d_{rank} is the total probability in the region corresponding to the ranking induced by $\mathbf{y}$. For example, the p-value of $d_{rank} = 0.65$ in our example above is the integral of the normal density function over all points in Quadrant I. In practice this becomes difficult to compute as m grows, but there is a simple *bootstrap sampling* procedure we can use to estimate it.

Our approach is similar to that of Sakai [9] and Smucker et al. [12] for pairwise tests. We sample topics with replacement, calculate all m system means over those topics, and then calculate d_{rank} from the resampled means to the original means. Over B trials we gain an empirical distribution of d_{rank}. A bootstrap p-value estimate is simply the proportion of the d_{rank}s in our empirical distribution that are no less than a given value d_r. If the p-value is sufficiently small, we reject the hypothesis that the rankings are the same. Algorithm 2 shows the details.

Algorithm 2 Estimate a p-value for a given rank distance d_r using bootstrap sampling.

Require: Rank distance d_r with baseline matrix $\mathbf{X}$ from Alg. 1; number of bootstrap samples B; $S \leftarrow 0$.

1: **for** $i \leftarrow 1$ to B **do**
2: $\quad\mathbf{X}^* \leftarrow$ an $n \times m$ matrix constructed by sampling rows in $\mathbf{X}$ with replacement.
3: $\quad\boldsymbol{\mu}^* \leftarrow$ column means of $\mathbf{X}^*$.
4: $\quad$**if** $d_{rank}(\boldsymbol{\mu}^*|\boldsymbol{\mu}, \mathbf{X}) \geq d_r$ **then** $S \leftarrow S + 1$.
5: **end for**
6: $p \leftarrow S/B$.

In our example above, $d_{rank} = 0.65$ with probability 0.21 and $d_{rank} = 0$ with probability 0.79 in the bootstrap distribution (with $B = 10000$ trials). The two rankings are statistically indistinguishable. The other four rankings are so rare as to have empirical probability zero; they are clearly significantly different from the baseline.

3.3 Software Download and Use

We have implemented our measure in C++ and `R`. The software is able to read in outputs from `trec_eval -q` to construct the matrix $\mathbf{X}$; a separate input file provides the ranking $\mathbf{y}$ to evaluate. Both implementations report the rank distance and a bootstrapped p-value suitable for dissemination. The bootstrap distribution can be saved to disk so that new p-values can be computed quickly in a series of batch experiments. We have also published complete distribution tables for common TREC datasets and measures; the software and distribution tables can be found at `http://ir.cis.udel.edu/~carteret/dRank/`.

4. ANALYSIS AND EVALUATION

Because our rank distance measure is entirely new, evaluation is somewhat difficult. We will first provide some simple theoretical results. We then compare d_{rank} to τ over random rerankings of TREC systems to show how they are different, and re-evaluate past work in meta-evaluation.

4.1 Theory

We can prove some basic things about our measure that may help understand or accept it. The first result is that the rank distance is zero if and only if two rankings are identical.

THEOREM 1. *Let $\mathbf{y}$ be a real-valued $1 \times m$ vector and $\mathbf{X}$ a real-valued $n \times m$ matrix with column means $\boldsymbol{\mu}$. Then $d_{rank}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{X}) = 0$ if and only if values of $\mathbf{y}$ have the same ranking as values in $\boldsymbol{\mu}$.*

PROOF. Let $\boldsymbol{\mu}_{\boldsymbol{\Delta}}, \boldsymbol{\Sigma}_{\boldsymbol{\Delta}}$ be defined as in Alg. 1. Let $\boldsymbol{\theta}$ be the vector that minimizes $d_{rank}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{X})$.

$\Rightarrow$: Suppose the rankings are the same. Since values of $\boldsymbol{\mu}$ are in the same order as values of $\mathbf{y}$, line 3 of Alg. 1 results in a vector $\boldsymbol{\mu}_{\boldsymbol{\Delta}}$ that is positive in every value. All values of $\boldsymbol{\theta}$ are positive by design. It is then simple to show that because both vectors are all positive, the minimum of $(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\Delta}})\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}{}^{-1}(\boldsymbol{\theta} - \mu_{\boldsymbol{\Delta}})$ is zero for any $\boldsymbol{\Sigma}_{\boldsymbol{\Delta}}$.

$\Leftarrow$: Suppose $d_{rank}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{X}) = 0$. Since $\boldsymbol{\theta}$ must be positive, this can only happen if $\boldsymbol{\mu}_{\boldsymbol{\Delta}} > 0$ as well. And if $\boldsymbol{\mu}_{\boldsymbol{\Delta}} > 0$, the ranking is the same. □

The following results shows how our measure goes above and beyond Kendall's τ by penalizing a ranking more when

pairs that are "more different" are swapped (Thm. 2) and by penalizing a ranking that "breaks up" a pair of similar systems (Thm. 3). Both theorems use the same setup. Let $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ induce identical rankings of m systems based on matrices $\mathbf{X}$ and $\mathbf{Y}$, respectively. Let $\mathbf{y}$ induce an alternative ranking in which system i and $i+1$ are swapped, but $i+1$ and $i+2$ are in the correct order (i.e. $y_{i+1} > y_i > y_{i+2}$ while $\mu_i > \mu_{i+1} > \mu_{i+2}$ and $\nu_i > \nu_{i+1} > \nu_{i+2}$).

THEOREM 2. *Let $t(\mathbf{x}-\mathbf{y})$ be Student's paired t-statistic. All else being equal, if $t(\mathbf{X}_{\cdot i} - \mathbf{X}_{\cdot(i+1)}) > t(\mathbf{Y}_{\cdot i} - \mathbf{Y}_{\cdot(i+1)})$, then $d_{rank}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{X}) > d_{rank}(\mathbf{y}|\boldsymbol{\nu}, \mathbf{Y})$.*

The proof follows from the fact that when all else is equal, the difference in the two rank distances reduces to a difference in t-statistics for systems $i, (i+1)$ by measure X and $i, (i+1)$ by measure Y. If we accept Student's t as a measure of distance (i.e. greater t means less similar), the result follows easily. Note in particular that the smaller the t-statistic, the less different the systems are, and therefore the lower d_{rank} is. The implication is that d_{rank} captures everything about a ranking that discriminative power does.

THEOREM 3. *Let $\rho(\mathbf{x}, \mathbf{y})$ be the Pearson correlation between vectors $\mathbf{x}$ and $\mathbf{y}$. All else being equal, if $\rho(\mathbf{X}_{\cdot i} - \mathbf{X}_{\cdot(i+1)}, \mathbf{X}_{\cdot i} - \mathbf{X}_{\cdot(i+2)}) > \rho(\mathbf{Y}_{\cdot i} - \mathbf{Y}_{\cdot(i+1)}, \mathbf{Y}_{\cdot i} - \mathbf{Y}_{\cdot(i+2)})$, then $d_{rank}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{X}) > d_{rank}(\mathbf{y}|\boldsymbol{\nu}, \mathbf{Y})$.*

The proof is algebraic, simple to work out though not easy to transcribe; with limited space, we will omit it. The consequence of this theorem is that the more similar $i+1$ and $i+2$ are, the greater the penalty from swapping i to be in between them.

4.2 Experimental Data

The standard datasets used in evaluation studies are the complete retrieval results of systems submitted to TREC tracks over the years. For most tracks, submitted systems retrieve up to 1,000 documents for each of a set of topics; the data consists of the full ranked list for every topic along with the relevance judgments (*qrels*) for those topics.

We present results with the 2007 Million Query Track runs (TB topic subset; 24 runs over 149 topics) and the 2004 Robust Track runs (110 runs over 249 topics).

4.3 Comparing Tau and Rank Distance

Figure 3 shows the ranks that the 24 Million Query Track runs are "allowed" to appear at given a 0.9 threshold for Kendall's τ or the significance threshold for the rank distance (for the MQ runs, the bootstrapped critical value for $\alpha = 0.05$ is 2.62). In these plots, each (x = system, y = rank) point indicates the probability of system x appearing at rank y for all rankings that have $\tau \geq 0.9$ or $d_{rank} \leq 2.62$. Lighter values mean higher probability; points are colored on a logarithmic scale. Note that τ "allows" each system a distribution over about 5 ranks, centered on its actual rank. Rank distance, on the other hand, constrains some systems to staying where they are (in particular, systems that are significantly different from those they are adjacent to), but others to move considerably. The group of 7 systems numbered 15–21 in particular have nearly free rein (with some constraints) by virtue of there being no significant differences between them.

The fact that statistical significance is not transitive produces some interesting effects. The fourth system is allowed by rank distance to appear at ranks 4 through 8, while the fifth is only allowed to appear at ranks 4 through 6. This is because there is a significant differences between systems 5 and 7, but not between 4 and 7.

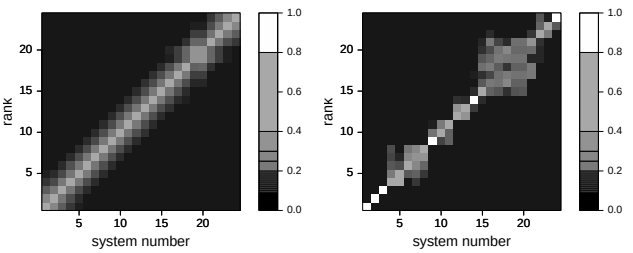


Figure 3: Kendall's τ and d_{rank} constrain the possible rankings of systems in very different ways. On the left, the distribution of possible ranks when $\tau \geq 0.9$ is identical for each system. On the right, the distribution of possible ranks when $d_{rank} \leq 2.62$ ($p > 0.05$) does not allow significantly different pairs to swap.

Kendall's τ and d_{rank} disagree quite a bit on whether a ranking is "good". Over two very large samples of rerankings (one sampled using d_{rank}'s sampling distribution, one sampled using τ's), over 25% of rankings with $\tau > 0.9$ are significantly different by d_{rank} from the true ranking. Conversely, rankings with τs as low as 0.8 can still be statistically indistinguishable from the true ranking.

Incidentally, the plot for discriminative power with threshold 1, which is not shown, has very nearly the same contours as the plot for rank distance, but the blocks within those contours are solidly colored. Discriminative power cannot capture the finer constraints imposed by dependencies among systems and allows for instance systems 19 and 20 (which are equal in *MAP* to four decimal places) to appear at any rank from 15 to 21 with up to 5 systems separating them. Rank distance insists that these two systems always be ranked next to each other, though they may be in any order relative to one another.

4.4 Comparing Evaluation Measures

It is well-known that different evaluation measures correlate highly. Voorhees and Harman give τ correlations between rankings by different evaluation measures for the TREC-7 data [14]. What we would really like to know is whether different measures actually measure different system attributes *despite* being highly correlated. Since they are highly correlated by default, rank correlations cannot tell us much about differences in system rankings between measures. Our measure can.

Table 1(a) shows τ correlations between rankings of 110 Robust systems over 249 topics by traditional TREC evaluation measures along with the 95% confidence interval calculated by Kendall's formula (1). There is quite a bit of overlap in the confidence intervals.

Table 1(b) shows the rank distances between the same rankings. As explained above, rank distance is not symmetric: it depends on which measure is chosen as the "baseline". This table presents distance from the measure on the row (the baseline) to the measure on the column and the bootstrap p-value of the hypothesis test. Note that these distances are substantially higher than the MQ critical value

	P10	P30	R-prec	*MAP*	MRR	bpref
P10	1	0.84 (0.64, 0.93)	0.81 (0.60, 0.92)	0.80 (0.59, 0.91)	0.73 (0.50, 0.86)	0.80 (0.59, 0.91)
P30		1	0.90 (0.72, 0.97)	0.89 (0.70, 0.96)	0.63 (0.39, 0.79)	0.89 (0.70, 0.96)
R-prec			1	0.92 (0.74, 0.98)	0.61 (0.36, 0.78)	0.93 (0.76, 0.98)
MAP				1	0.58 (0.33, 0.75)	0.96 (0.80, 0.99)
MRR					1	0.60 (0.35, 0.77)
bpref						1

(a) Kendall's τ (and 95% c.i.) between rankings by different evaluation measures averaged over 249 topics.

	P10	P30	R-prec	*MAP*	MRR	bpref
P10	0	10.11, p=0.00	13.06, p=0.00	14.11, p=0.00	15.08, p=0.00	13.49, p=0.00
P30	9.92, p=0.00	0	9.24, p=0.00	7.86, p=0.00	20.05, p=0.00	8.44, p=0.00
R-prec	15.29, p=0.00	8.16, p=0.00	0	**5.83**, p=0.37	22.63, p=0.00	**4.73**, p=0.91
MAP	25.44, p=0.00	14.42, p=0.00	9.51, p=0.00	0	29.18, p=0.00	**3.52**, p=1.00
MRR	13.27, p=0.00	16.45, p=0.00	18.89, p=0.00	18.95, p=0.00	0	19.03, p=0.00
bpref	18.15, p=0.00	9.72, p=0.00	**6.38**, p=0.14	**3.15**, p=1.00	23.95, p=0.00	0

(b) Rank distance from the ranking by the measure on the row to the ranking by the measure on the column, with bootstrap p-values. Nonsignificant differences are bolded.

Table 1: Comparisons between ranking by common evaluation measures over the 110 Robust systems.

of 2.62 and the maximum distance of 4.88 in our example in Section 3.1. The values of d_{rank} depend heavily on the number of systems and topics as well as the degree of interdependence between systems. Thus the bootstrap p-value is often better-suited for reporting; this is shown as well. Rankings that are *not* significantly different are bolded.

The table shows that rankings by bpref and *MAP* and rankings by bpref and R-prec are indistinguishable from each other; bpref does a very good job of approximating both *MAP* and R-prec. A ranking by *MAP* is not significantly different from a baseline ranking by R-prec, but the converse is not true, probably because *MAP* is more tightly constrained than R-prec. Overall, because nearly all the rank distances are significant, this table shows that apart from bpref & *MAP* and bpref & R-prec, different measures really do measure different system attributes, and rankings by different measures can capture systems that are designed to optimize one attribute over another. This is much more enlightening than the τ table, in which all the values are positive and significant, and whose confidence intervals overlap with each other substantially.

4.5 Incomplete Relevance Judgments

Shallow Pools. It is commonly understood that evaluation over shallow judgment pools correlates highly to evaluation over a deep pool. Again, this should not be surprising given what we now know about rank correlation methods.

Table 2 shows τ correlation and rank distance between a baseline ranking of Robust systems by official *MAP* and the ranking by *MAP* calculated over a shallow pool. While τ reaches nearly 0.9 with a depth 10 pool, the rank distance test rejects the hypothesis that the rankings are the same at that point. It takes a deeper pool before the ranking are statistically indistinguishable.

Incomplete Judgments. Buckley and Voorhees introduced the *bpref* measure for evaluation with incomplete test collections [4]. To test it, they compared bpref, *MAP*, R-prec, and P10 over increasingly incomplete sets of relevance judgments. We duplicated their experiment on the Robust collection, calculating d_{rank} and the rank test p-value in addition to τ between the official ranking of systems by each measure and the ranking of systems by the same measure

depth	judgments	τ (95% c.i.)	d_{rank}, p-value
1	4,789	0.67 (0.43, 0.82)	24.31, p=0.00
5	16,478	0.83 (0.62, 0.93)	14.81, p=0.00
10	28,569	0.89 (0.70, 0.96)	11.57, p=0.00
25	60,099	0.96 (0.80, 0.99)	**5.65**, p=0.49
50	101,626	0.98 (0.83, 1.00)	**1.76**, p=1.00

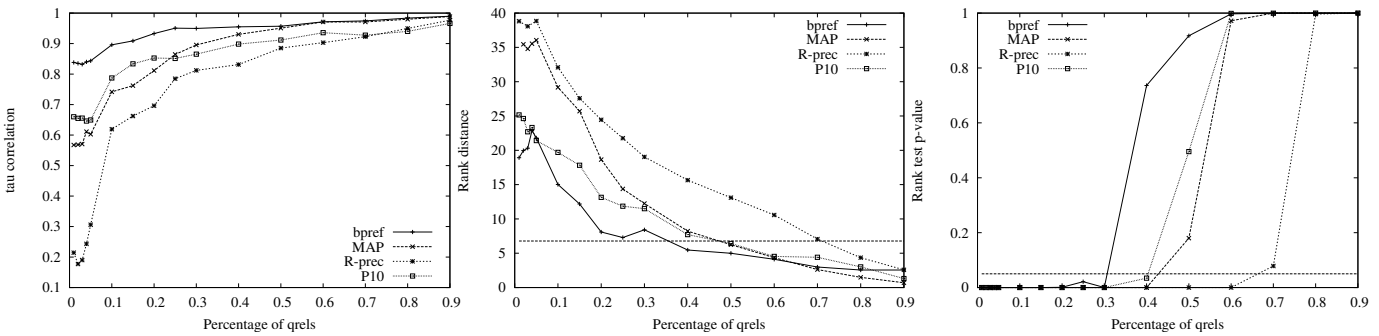
Table 2: Kendall's τ and rank distance to official ranking when evaluating over shallow pools. Bolded rank distances indicate rankings that are not significantly different.

calculated over the reduced set. (For full details of the methodology we refer the reader to the original paper.)

Figure 4(a) shows τ correlation between a measure calculated with a reduced qrels set and the same measure over all the qrels. bpref seems to be more robust to missing judgments, having a higher τ with very incomplete sets; R-precision seems to be least robust to missing judgments. This agrees with the results of Buckley and Voorhees.

Figure 4(b) tells a slightly different story. This shows rank distance decreasing as the qrels becomes more complete. While the relative performance of the four measures is roughly the same, bpref's strong performance with τ turns out to be somewhat misleading: it requires nearly twice as many judgments to achieve a good d_{rank} (one that is not significantly different from 0) than a "good" τ of 0.9. However, bpref is still more robust to missing judgments than any of the other measures.

Finally, Figure 4(c) shows the p-value of the rank hypothesis test increasing as qrels completeness increases. Recall that a low p-value means that the ranking is significantly different; higher p-value means more similar to the "true" ranking. Here we see that bpref very quickly "loses" significance at about 30% qrels. The other three measures lose significance in the interval $0.4-0.7$, meaning that more than half the full collection was needed before the rankings were indistinguishable. All four measures very quickly transition from "significant" to "not significant"; this suggests there is a sort of "phase transition" due to the number of judgments, at which point the variance constrains the ranking enough that it is unlikely to change drastically.



(a) Kendall's τ. (b) Rank distance. (c) Rank test p-value.

Figure 4: Reduced qrels. Horizontal lines indicate significance; values above the line in (b) and below the line in (c) are significant at $\alpha = 0.05$. Every value in (a) is significant.

No Relevance Judgments. Soboroff et al. examined ranking systems by, in effect, random judgments [13]. They randomly sampled subsets of documents to be labeled relevant. They then used those pseudo-relevance judgments ("pseudo-rels") to evaluate and rank systems. They found a positive and significant τ correlation to the true ranking.

We duplicated the experiment on the Robust systems. Over 50 sets of pseudo-rels that were assembled (sampling from the pool without duplicates), the average rank distance was 26.99 with a standard deviation of 1.58, on the same order as the distance between MRR and other measures. The average τ correlation was 0.656 with a standard deviation of 0.021. The hypothesis that the ranking by pseudo-rels was equivalent to the ranking by the qrels was rejected ($p \approx 0$). By contrast, the τ correlation indicated that the ranking was significantly correlated—a noninformative result.

5. CONCLUSION

We have introduced a new measure of the distance between rankings that solves the problems with using Kendall's τ over a sample of correlated items. We have demonstrated its power by analogy to powerful statistical methods like the t-test and Hotelling's T^2, through theoretical results about relative increases, through direct comparison to Kendall's τ over simulated rerankings, and through examples well-known in the IR literature.

We grant that our measure is not intuitive. However, the problems with using Kendall's τ in meta-evaluation are so severe that the intuition it offers is a false hope. *At best*, τ can only provide the loosest guide to interpreting ranking results; at worst it is entirely misleading in both directions: a high τ does not necessarily indicate a good ranking, and a "low" τ is far from a guarantee of a bad ranking.

Furthermore, the assumptions required by our method are much weaker than those required by Kendall's τ. On the surface our analogies to Mahalanobis distance may give the appearance that we are assuming measures like average precision are normally distributed over topics. The actual assumption is substantially weaker: only that the vector of means is normally distributed over samples of n topics. This is a consequence of multivariate generalizations to the Central Limit Theorem, which say that the distribution a sum of random variables converges to normal (regardless of whether the random variables are scalars or vectors). Kendall's τ, on the other hand, requires the assumption that systems are independently sampled, which is most certainly not true even in approximation. Furthermore, the failure of τ's assumption imposes much greater error on the statistic than the failure of any assumption made by d_{rank}.

We believe our method can be adapted to comparing rankings of documents as well, and possibly used as an objective function in learning to rank. We will be investigating this further in the future.

6. REFERENCES

[1] J. Allan, B. Carterette, J. A. Aslam, V. Pavlu, B. Dachev, and E. Kanoulas. Overview of the TREC 2007 Million Query Track. In *Proceedings of TREC*, 2007.

[2] J. Aslam and R. Savell. On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In *Proceedings of SIGIR*, pages 361–362, 2003.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of SIGIR*, pages 25–32, 2004.

[5] G. V. Cormack and T. R. Lynam. Power and bias of subset pooling strategies. In *Proceedings of SIGIR*, pages 837–838, 2007.

[6] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall, 1982.

[7] M. Kendall. *Rank Correlation Methods*. Griffin, London, UK, fourth edition, 1970.

[8] M. Melucci. On rank correlation in information retrieval evaluation. *SIGIR Forum*, 41(1):18–33, 2007.

[9] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of SIGIR*, pages 525–532, 2006.

[10] T. Sakai. Alternatives to bpref. In *Proceedings of SIGIR*, pages 71–78, 2007.

[11] M. Sanderson and I. Soboroff. Problems with kendall's tau. In *Proceedings of SIGIR*, pages 839–841, 2007.

[12] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of CIKM*, pages 623–632, 2007.

[13] I. Soboroff, C. Nicholas, and P. Cahan. Ranking Retrieval Systems without Relevance Judgments. In *Proceedings of SIGIR*, pages 66–73, 2001.

[14] E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[15] D. I. Warton. Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *Journal of the American Statistical Association*, 103(481):340–349, March 2008.

[16] L. Wasserman. *All of Statistics*. Springer, 2006.

[17] E. Yilmaz, J. A. Aslam, and S. Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of SIGIR*, pages 587–594, 2008.