COMPARISON OF PARTIALLY RANKED LISTS

E. Stoimenova
Institute of Mathematics and Informatics
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
Sofia, BULGARIA
e-mail: jeni@math.bas.bg

Abstract

In this paper we introduce a measure of closeness of partial rankings based on a metric on permutations, and we analyze some of its properties.

1 Introduction

In many situations, there are different methods for analyzing the same data. For example, several methods exist for finding differentially expressed genes using RNA-seq data. They tend to produce similar, but not identical significant genes and rankings of the gene list. When comparing different methods applied to the same data, we are interested in how close are their outputs. The main idea is to define appropriate distance of the sample space. Further, the interpretation of the rough distance between two rankings should be made on the basis of its statistical significance. That means we need to know the distribution of the distance under some common hypotheses about a sample of rankings. In recent years, many new applications appear in different areas including bioinformatics pattern recognition, information retrivial [7], [6], [1], [4], [5], etc.

In this paper we define an appropriate mathematical framework that include special cases of partially ranked lists of genes. Any ranked list can be complete, which means all n genes are ranked, or incomplete, which means some genes are not ranked. The incomplete ranking include the case where the most significant k genes are ranked, with group $k+1$ consisting of the remaining genes. Any ranking of n items corresponds a permutation $\langle \alpha(1), \ldots, \alpha(n) \rangle$ from the set of all permutations S_n. We define appropriate distance measures on S_n in order to compare full or incomplete rankings or rankings of different types. The distance can be thought of as a measure of the similarity of the two rankings.

Let α and β be two permutations from S_n corresponding to two rankings and let d be a metric on the permutation group S_n. Then $d : S_n \times S_n \to [0, \infty)$ satisfies the usual axioms: $d(\alpha, \beta) \geq 0 \quad \forall \alpha, \beta \in S_n$, $d(\alpha, \beta) = 0 \Leftrightarrow \alpha = \beta$; $d(\alpha, \beta) = d(\beta, \alpha) \quad \forall \alpha, \beta \in S_n$; and the triangle inequality $d(\alpha, \beta) \leq d(\alpha, \gamma) + d(\gamma, \beta) \quad \forall \alpha, \beta, \gamma \in S_n$.

Invariance is natural in many problems. Right-invariance means that the distance does not depend on arbitrary labeling or reordering of the data:

$$d(\alpha, \beta) = d(\alpha\tau, \beta\tau).$$

Here $\alpha\tau$ is the product of two permutations α and τ and defined by $\alpha\tau(i) = \alpha(\tau(i))$. Right-invariant property allows to compute the distance between two permutations α and β through the the distance of $\alpha\beta^{-1}$ to the identity permutation.

Further in our analysis we are using a popular statistical measure of similarity on S_n called Spearman's ρ. For $\alpha, \beta \in S_n$ it is defined by

$$R^2(\alpha, \beta) = \sum_{i=1}^{n} (\alpha(i) - \beta(i))^2.$$

Strictly speaking, Spearman's ρ is not a metric in the above definition, however, its square root is the Euclidean metric on permutations. It is easy to see that Spearman's ρ is right-invariant. By right-invariance of a distance it is sufficient to study its statistical properties when one of the rankings is the identity permutation.

2 Complete or incomplete ranking

A ranking of n items is represented by an ordered n-tuple, which simply lists the items in their ranked order. The most preferred item is listed first, and the least preferred item appears in the n-th position. Any ranking corresponds to a permutation which is an element of the set S_n of permutations. Given a set of rankings, the problem of their comparison reduced to a problem of choosing appropriate measure of association on the set of all rankings. There are several usefull distance measures on S_n thoroughly discussed in statistical literature like Kendall's τ, Spearman's ρ, Spearman's footrule. Therefore, for two permutations $\alpha, \beta \in S_n$ the distance $d(\alpha, \beta)$ can be thought of as a measure of similarity of the two rankings. Excellent references on statistical analysis of rankings are the monographs by Diaconis [3], Critchlow [2], and Marden [8].

[**Classification into r ordered categories.**] Suppose the list of genes is splitted into several groups, so that there is a ranking between the groups and not necessarily within each group. It can be describe formally following Critchlow [2].

Let $n_1, \ldots, n_r$ be an ordered sequence of r strictly positive numbers summing to n. Such an ordered partition corresponds to a partial ranking with n_1 items in the first group, n_2 items in the second group and so on. No further information is conveyed about orderings within each group. The special case of ranking the top k items corresponds to $n_1 = \cdots = n_k = 1,\ n_{k+1} = k + 1$.

Formally, denote $N_1, \ldots, N_r$ are the following partition of $\{1, \ldots, n\}$:

$$\begin{aligned} N_1 &= \{1, \ldots, n_1\} \\ N_2 &= \{n_1 + 1, \ldots, n_1 + n_2\} \\ &\cdots \\ N_r &= \{n_1 + \cdots + n_{r-1} + 1, \ldots, n\}. \end{aligned} \tag{1}$$

Let S denote the subgroup of all rankings which permute the first n_1 items among the first n_1 ranks, and which permute the next n_2 items among the next n_2 ranks, and

so on. The equivalence class $[\alpha]$, that assigns the same set of ranks to the items from the each category as α, is the right coset $S\alpha$. There is a one-to-one correspondence between the partitioning "of type $n_1, \ldots, n_r$" and the right cosets of S.

2.1 Distances on partial rankings

In the above algebraic structure the problem of comparing of partial rankings is reduced to a problem of extending the metrics on the permutation group S_n to metrics on the corresponding coset space. We discuss an extension of the above metrics for the cases of partial rankings. One natural way of extending it is to construct the induced Hausdorff metrics. Its particular benefit is that it keeps the metric properties of the original distance.

Let the two partial rankings be of types $n_1, \ldots, n_r$. Denote n_{ij} the number of elements in the set $\{\alpha^{-1}(N_i) \cap \beta^{-1}(N_j)\}$. Then the function

$$R_{fv}(\alpha, \beta) = \sum_{i=1}^{r} \sum_{j=1}^{r} |c_i - c_j|^2 n_{ij}.$$

is a right-invariant metric on partial rankings induced by Spearman's ρ. Here $c_i = n_1 + \cdots + n_{i-1} + \frac{n_i+1}{2}$ is the average of the n_i numbers in the set N_i defined by (1).

The interpretation of this function is that it computes Spearman's ρ distance between the two rankings using the "pseudo-ranks" c_i and c_j instead the ordinary ranks to those items in $\{\alpha^{-1}(N_i) \cap \beta^{-1}(N_j)\}$. The function is called the "fixed vector" metric on S_n/S induced by Spearman's ρ. Its main advantage is that it preserves the distance properties and the right invariance as well [9]. Additionally, some useful statistical properties are known in the literature.

2.2 Comparing partial rankings of different types

We consider the most general case of comparing partial rankings of different types. Let the two partial rankings be of types $n_1, \ldots, n_r$ and $n'_1, \ldots, n'_{r'}$ respectively.

Let $N_1, \ldots, N_r$ be as defined in (1) and let $N'_1, \ldots, N'_r$ be a second partition of $\{1, \ldots, n\}$:

$$\begin{aligned} N'_1 &= \{1, \ldots, n'_1\} \\ N'_2 &= \{n'_1 + 1, \ldots, n'_1 + n'_2\} \\ &\cdot \quad \cdot \quad \cdot \\ N'_{r'} &= \{n'_1 + \cdots + n_{r'-1} + 1, \ldots, n\}. \end{aligned}$$

Let n_{ij} be the number of elements in the set $\{\alpha^{-1}(N_i) \cap \beta^{-1}(N'_j)\}$. Then

$$R_*(\alpha, \beta) = \sum_{i=1}^{r} \sum_{j=1}^{r'} |c_i - c'_j|^2 n_{ij}.$$

is a right-invariant metric on partial rankings. Here $c'_j = n'_1 + \cdots + n'_{j-1} + \frac{n'_j+1}{2}$ is the average of the n'_j numbers in the set N'_j defined by (2).

3 Large sample approximation of a distance distribution

Now, we estimate the mean and the variance of R^{2*} and find approximations of its distribution.

Definition 1. The metric d^* on S_n/S is asymptotically normally distributed if for partial rankings α^* and β^* the following limit distribution is valid

$$\lim_{n\to\infty} P\left(\frac{d^*(\alpha^*,\beta^*) - E\ d^*(\alpha^*,\beta^*)}{\sqrt{var(d^*(\alpha^*,\beta^*))}} \le x\right) = \Phi(x)$$

for all real numbers x, where Φ, is the standard normal cumulative distribution function.

The significance of the distance is useful to estimate the similarity between the two partial rankings. For this one needs to calculate the probability that d^* is less than or equal to the observed value $d^*(\alpha^*,\beta^*)$. This probability is the p-value for α^* and β^*. Smaller values of p indicate stronger evidence that α^* and β^* are "similar". To compute the p-value, Critchlow [2] finds the probability distribution of some popular metrics on permutations under the appropriate uniformity assumption.

The mean and the variance of R^{2*} are given by [2]:

$$E(R^{2*}) = \sum_{i=1}^{r}\sum_{j=1}^{r} |c_i - c_j|^2 \frac{n_i n_j}{n}$$

$$var(R^{2*}) = \frac{1}{n^2(n-1)} \sum_{i=1}^{r}\sum_{j=1}^{r}\sum_{k=1}^{r}\sum_{l=1}^{r} n_i n_j n_k n_l (|c_i - c_j|^2 + |c_k - c_l|^2 - 2|c_k - c_j|^2),$$

where $c_i = n_1 + \cdots + n_{i-1} + \frac{n_i+1}{2}$ is the average of the n_i numbers in the set $N_i = \{n_1 + \cdots + n_{i-1} + 1, \ldots, n_1 + \cdots + n_{i-1} + n_i\}$.

For equal partition sizes these reduce to

$$E(R^{2*}) = n^3 \frac{(r^2-1)}{6r^2}$$

$$var(R^{2*}) = \frac{n^6}{n-1} \frac{(r^2-1)^2}{6r^5}.$$

Normal approximation is valid for the distance distribution under the assumption that they were generated, independently, from a uniform distribution on all possible partial rankings. For equal partition sizes Gamma distribution with shape parameter $= \mu^2/\sigma^2 = n-1$ gives better approximation.

Example (Palejev & Stoimenova [10]). A simulation study is based on one million repetitions of gene sequences of size 13932. Each of them contains data for the significance of gene expression. Further, the genes are splitted into six groups by values

according the size of the p-values. Intervals reasonable for application are determined by $0, 10^{-4}, 10^{-3}, 10^{-2}, 0.05, 10^{-1}, 1$. For this unbalance case the distances between any two of the partial rankings are calculated. The distributions of the distances is shown on Figure 1. Gamma distribution approximation is also suitable for this case.

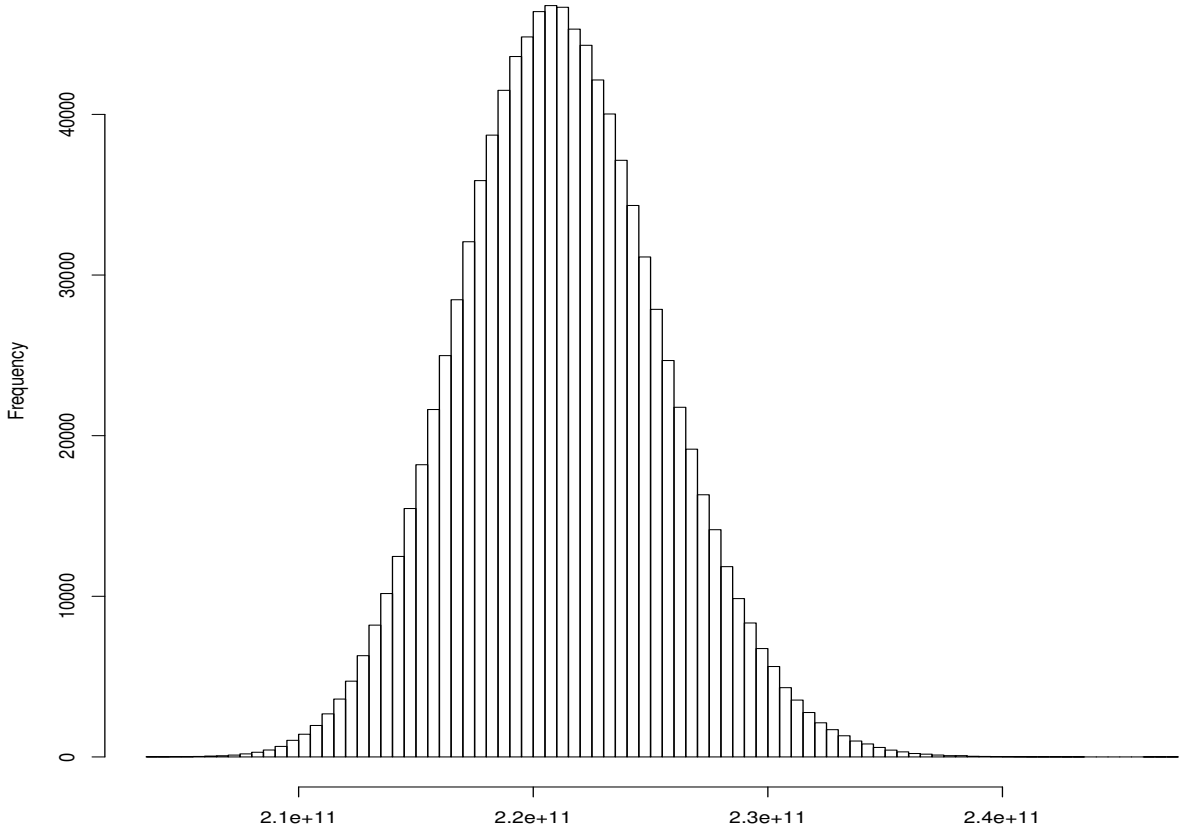

Figure 1: Distribution of distances between 2 random permutations

Acknowledgments. The author acknowledge funding by the Bulgarian fund for scientific investigations Project I02/19.

References

[1] Chan, C. H., Yan, F., Kittler, J., and Mikolajczyk, K. (2015). Full ranking as local descriptor for visual recognition: A comparison of distance metrics on S_n. *Pattern Recognition*, 48(4):134–160.

[2] Critchlow, D. E. (1985). *Metric methods for analyzing partially ranked data.* Lecture Notes in Statistics, 34. Berlin etc.: Springer-Verlag.

[3] Diaconis, P. (1988). *Group representations in probability and statistics.* IMS Lecture Notes-Monograph Series, 11. Hayward, CA: Institute of Mathematical Statistics.

[4] Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., and Vee, E. (2006). Comparing partial rankings. *SIAM J. Discrete Math.*, 20(3):628–648.

[5] Fagin, R., Kumar, R., and Sivakumar, D. (2003). Comparing top k lists. *SIAM J. Discrete Math.*, 17(1):134–160.

[6] Jurman G.,Riccadonna S. (2009). Canberra distance on ranked lists. In: *Proceedings of Advances in Ranking NIPS 09 Workshop*, pages 22–27.

[7] Jurman, G., Merler, S., Barla, A., Paoli, S., Galea, A., and Furlanello, C. (2007). Algebraic stability indicators for ranked lists in molecular profiling . *Bioinformatics*, 24(2):258–264.

[8] Marden, J. I. (1995). *Analyzing and modeling rank data.* Monographs on Statistics and Applied Probability. 64. London: Chapman.

[9] Rukhin, A. L. (1970). Certain statistical and probability problems on groups. 111:52–109.

[10] Palejev, D., Stoimenova, E. Comparison of incomplete ranked lists with application to RNA-seq differential expression methods. Working paper.