Canberra distance on ranked lists

Giuseppe Jurman, Samantha Riccadonna, Roberto Visintainer and Cesare Furlanello*
Fondazione Bruno Kessler
I-38123 Povo (Trento), Italy
{jurman, riccadonna, visintainer, furlan}@fbk.eu

Abstract

The Canberra distance is the sum of absolute values of the differences between ranks divided by their sum, thus it is a weighted version of the $L1$ distance. As a metric on permutation groups, the Canberra distance is a measure of disarray for ranked lists, where rank differences in top positions need to pay higher penalties than movements in the bottom part of the lists. Here we describe the distance by assessing its main statistical properties and we show extensions to partial ranked lists. We conclude providing two examples of use in functional genomics.

1 Introduction

The Canberra distance, introduced by Lance and Williams in the late sixties [1] as a software metric, is a weighted version of the classic $L1$ or Spearman's footrule which naturally extends to a metric on symmetric groups. This metric is particularly useful when comparing ranked lists in functional genomics. In fact, in the case of panels of biomarkers, we may require to differently penalize rank differences in the higher portion of the lists rather than those in the bottom section. The theory of metric methods on ranked data is relatively recent: an exhaustive reference is [2]. Pioneer work in this fiel has been carried on by Hoeffding [3] mixing statistics and permutation group theory and by Diaconis on distances for symmetric groups [4] *i.e.*, right-invariant distances. A few metrics on permutation groups are known, and for some of them a description in terms of their key properties (expected value, variance, distribution, extremal values) has been stated and proved [5]. We previously introduced the use of the Canberra distance in machine learning for computational biology and define an indicator of stability for ranked lists of biomarkers [6]. Although electively oriented towards computational biology, the potentiality of the Canberra distance can be of interest for the broader community dealing with rank comparison on general problems where the difference between relevant and negliglible objects (*i.e.*, with high and low rank) plays an important role.

Here we explicitly fin the approximated and exact (wherever feasible) expressions for the expectation value, the variance and the maximum value and argument, together with assessing its normality. All the aforementioned are expressed as functions of the harmonic numbers, the partial sums of the harmonic series summing the reciprocals of all natural numbers. Although a piece of classical number theory [7], computing with harmonic number is not trivial. Only recently papers have appeared providing closed[1] or at least recursive forms for more complex expressions involving sums and products of harmonic numbers [8].

After describing the key elements, here we discuss extensions of the Canberra distance to partial lists. In the last part, we show two applications for functional genomics. A molecular profilin and a study of variability of Canberra distances for differentially expressed biomarker lists are presented. All the algorithms described are implemented as functions of the `mlpy` Open Source Python library for machine learning, freely available at `https://mlpy.fbk.eu`.

*Corresponding author. Lab website: `http://mpba.fbk.eu`

[1] A closed form is an expression of a variable n that can be computed by applying a number of basic operations not depending on n; $\frac{n(n+1)}{2}$ is a closed form for the sum of the fi st n naturals, while $\sum_{i=1}^{n} i$ is not.

2 Properties of the Canberra distance

Preliminaries Given two real-valued vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, their Canberra distance and its natural extension to a distance on the permutation group on p objects S_p are define as follows:

$$\mathrm{Ca}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \frac{|\mathbf{x}_i - \mathbf{y}_i|}{|\mathbf{x}_i| + |\mathbf{y}_i|} \quad \text{and} \quad \mathrm{Ca}(\tau, \sigma) = \sum_{i=1}^{p} \frac{|\tau(i) - \sigma(i)|}{\tau(i) + \sigma(i)}, \text{ for } \tau, \sigma \in S_p.$$

Because of right-invariance, definin $\mathrm{Ca}_I(\sigma) = \mathrm{Ca}(\sigma, \mathrm{Id}_{S_p})$, the identity $\mathrm{Ca}(\sigma, \tau) = \mathrm{Ca}_I(\sigma\tau^{-1})$ holds and it will be repeatedly used hereafter.

For a natural $n \in \mathbb{N}$ and a positive exponent $a \in \mathbb{R}$, the generalized harmonic number $H_n^{(a)}$ is define as $H_n^{(a)} = \sum_{i=1}^{n} \frac{1}{i^a}$; when $a = 1$ the bracketed exponent is omitted and one speaks of harmonic number in short. Relations involving $H_n^{(a)}$ are dealt with by the identities proved in [8] expressing $H_n^{(a)}$ as functions of H_n. By Euler's formula, H_n can be expanded as

$$H_n = \log(n) + \gamma - \sum_{b=1}^{\infty} \frac{B_b}{n^b},$$

where $\gamma \approx 0.57721$ is the Euler-Mascheroni constant and B_b denotes the b-th Bernoulli number; truncating the formula at $b = 1$ the approximation reads as follows:

$$H_n = \log(n) + \gamma + \frac{1}{2n} + o(n^{-1}) \,. \tag{1}$$

In most of the proofs we follow the notations, the strategy and the results of Hoeffding's paper [3]: accordingly, we use the shorthands

$$c_p(i,j) = \frac{|i-j|}{i+j}, \quad f_p(i) = \frac{1}{p}\sum_{j=1}^{p} c_p(i,j) \quad \text{and}$$

$$d_p(i,j) = c_p(i,j) - \frac{1}{p}\sum_{g=1}^{p} c_p(i,g) - \frac{1}{p}\sum_{h=1}^{p} c_p(h,j) + \frac{1}{p^2}\sum_{h,g=1}^{p} c_p(h,g) \,.$$

Proofs are only sketched due to the length of the required computations: a few details are provided only for the case of the expectation value as an example of the following proofs.

The following analysis is to be considered as a combinatorial problem: as in [2, 5], the truly random permutations case is assumed. All permutations are supposed to be chosen independently and uniformly in S_p and the same applies to all the permutation pairs.

Results

R1. The expected value of the Canberra distance is

$$\begin{aligned} E\{\mathrm{Ca}(S_p)\} &= \frac{1}{|S_p|}\sum_{\sigma\in S_p} \mathrm{Ca}_I(\sigma) = \frac{1}{p!}(p-1)!\sum_{i,j=1}^{p} c_p(i,j) = \sum_{i=1}^{p} f_p(i) \\ &= \sum_{i=1}^{p} \frac{2i}{p}(2H_{2i} - H_i - H_{p+i} - 1) + 1 \\ &= \frac{2}{p}\left(\sum_{i=1}^{p} 2iH_{2i} - \sum_{i=1}^{p} iH_{n+i} - \sum_{i=1}^{p} iH_i - \sum_{i=1}^{p} i\right) + \sum_{i=1}^{p} 1 \\ &= \left(2p + 2 + \frac{1}{2p}\right) H_{2p} - \left(2p + 2 + \frac{1}{4p}\right) H_p - \left(p + \frac{3}{2}\right) \,. \end{aligned} \tag{2}$$

The proof is straightforward: the key step is the identity $f_p(i) = \frac{2i}{p}(2H_{2i} - H_i - H_{p+i} - 1) + 1$ together with the $\sum_{j=1}^{p} jH_{j+l} = \frac{(p-l)(p+l+1)}{2}H_{p+l+1} + \frac{l(l+1)}{2}H_{l+1} + \frac{p(2l-p-1)}{4}$.

R2. The approximation of the expected value by Euler's formula (1) **is**

$$\begin{aligned} E\{\mathrm{Ca}(S_p)\} &= \left(2p+2+\frac{1}{2p}\right)H_{2p} - \left(2p+2+\frac{1}{4p}\right)H_p - \left(p+\frac{3}{2}\right) \\ &= \left(2p+2+\frac{1}{2p}\right)\left(\log(2p)+\gamma+\frac{1}{4p}+o(p^{-1})\right) \\ &\quad - \left(2p+2+\frac{1}{4p}\right)\left(\log(p)+\gamma+\frac{1}{2p}+o(p^{-1})\right) - \left(p+\frac{3}{2}\right) \\ &= \frac{\log(p)}{4p} + \frac{\gamma}{4p} + \log(2)\left(2p+2+\frac{1}{2p}\right) - \left(\frac{1}{2}+\frac{1}{2p}\right) - \left(p+\frac{3}{2}\right) + o(1) \\ &= \log(2)(2p+2) - \frac{1}{2} - p - \frac{3}{2} + o(1) \\ &= \log(4)(p+1) - (p+2) + o(1)\ . \end{aligned} \tag{3}$$

The $o(1)$ term indicating the difference between the exact and the approximated values is about $4 \cdot 10^{-2}$ for $p = 20$, $2 \cdot 10^{-3}$ for $p = 10^3$ and it is less than 10^{-5} for $p > 10^5$.

R3. The variance can be computed starting from the identity

$$V\{\mathrm{Ca}(S_p)\} = \frac{1}{p-1}\sum_{i,j=1}^{p} d_p^2(i,j) = \frac{1}{p-1}\sum_{i,j=1}^{p} c_p^2(i,j) + E^2\{\mathrm{Ca}(S_p)\} - 2c_p(i,j)f_p(j)\ .$$

The expanded form involves tens of terms in H_p, H_p^2, H_{2p} and H_{2p}^2. In particular, for the two occurring sums $\sum_{i=1}^{p} i^2 H_{p+i}H_{2i}$ and $\sum_{i=1}^{p} i^2 H_{p+i}H_i$ no closed form is known.

R4. By integral approximation, the variance can be asymptotically estimated as

$$V\{\mathrm{Ca}(S_p)\} = \left(\frac{22}{3} + \frac{7}{3}\log^2(4) - \frac{4\pi^2}{9} - \frac{16}{3}\log(4)\right)p + o(p) \simeq 0.0375p + o(p)\ .$$

R5. Canberra distance is asymptotically normal. Since $d_p^2(i,j)$ reaches its maximum in $(1,1)$ and $d_p^2(1,1) = \frac{1}{p}\left[E\{\mathrm{Ca}(S_p)\} - 2(p+2-2H_{p+1})\right]$, the limit

$$\lim_p \frac{d_p^2(1,1)}{\frac{p-1}{p}V\{\mathrm{Ca}(S_p)\}} = 0$$

holds: then asymptotic normality follows from [3, Th. 1-4].

R6. Canberra distance has a maximal permutation. The two solutions $\rho_M,\ \rho_M^{-1}$ to the equation $\rho = \arg\max_{S_p}(\mathrm{Ca}_I)$ depend on the parity of p ($\rho_M = \rho_M^{-1}$ for even p):

$$\rho_M = \begin{pmatrix} 1 & 2 & \dots & \frac{p}{2} & \frac{p}{2}+1 & \frac{p}{2}+2 & \dots & p \\ \frac{p}{2}+1 & \frac{p}{2}+2 & \dots & p & 1 & 2 & \dots & \frac{p}{2} \end{pmatrix} \text{ or } \begin{pmatrix} 1 & 2 & \dots & \frac{p-1}{2} & \frac{p-1}{2}+1 & \frac{p-1}{2}+2 & \dots & p \\ \frac{p-1}{2}+1 & \frac{p-1}{2}+2 & \dots & p-1 & p & 1 & \dots & \frac{p-1}{2} \end{pmatrix}^{\pm 1}$$

respectively for even and odd p. The proof passes through restricting to smaller sets of permutations according to the following conditions:

1. if $\sigma \in S_p$ fi es an index $z \in \Omega_p = \{1,\dots,p\}$, then $\exists \tau \in S_n$ such that $\mathrm{Ca}_I(\tau) > \mathrm{Ca}_I(\sigma)$;
2. if $\exists\{i,\sigma(i)\} \subset \Omega_{\left[\frac{p}{2}\right]}$ then $\exists \tau \in S_p$ such that $\mathrm{Ca}_I(\tau) > \mathrm{Ca}_I(\sigma)$ and
   if $\exists\{i,\sigma(i)\} \subset \Omega_p \setminus \Omega_{\left[\frac{p}{2}\right]+1}$ then $\exists \tau \in S_p$ such that $\mathrm{Ca}_I(\tau) > \mathrm{Ca}_I(\sigma)$;
3. if $i \leq \left[\frac{p}{2}\right] - 1$ and $\sigma(i) > \sigma(i+1)$ then $\exists \tau \in S_p$ such that $\mathrm{Ca}_I(\tau) > \mathrm{Ca}_I(\sigma)$.

The maximum value is then $\mathrm{Ca}_I(\rho_M) =$

$$\begin{cases} 2r(H_{3r} - H_r) & \text{if } p = 4r \\ (2r+1)H_{6r} + rH_{3r+1} - (r+\frac{1}{2})H_{3r} - (2r+1)H_{2r+1} + \frac{1}{2}H_r & \text{if } p = 4r+1 \\ (2r+1)(2H_{6r+3} - H_{3r+1} - 2H_{2r+1} + H_r) & \text{if } p = 4r+2 \\ (2r+1)H_{6r+5} + \frac{1}{2}H_{3r+2} - (2r+1)H_{2r+1} - (r+1)H_{r+1} + (r+\frac{1}{2})H_r & \text{if } p = 4r+3\ , \end{cases}$$

which can be approximated by Euler's formula for any p by $\max_{S_p}(\mathrm{Ca}_I) = \frac{\log(3)}{2}p - \frac{2}{3} + o(1)$. The $o(1)$ term is $5.9 \cdot 10^{-4}$ for $p = 10^3$, $5.9 \cdot 10^{-5}$ for $p = 10^4$ and $6.0 \cdot 10^{-6}$ for $p = 10^5$.

Extensions to partial lists The Canberra distance can be naturally extended to partial lists in a few natural ways. If comparison among top-k lists is investigated for k fi ed, the Hausdorff version of the Canberra metric can be define [2]; the results in [9] show its equivalence with the Canberra distance with location parameter $l = k + 1$: $\mathrm{Ca}^{(k+1)}(\tau, \sigma) = \sum_{i=1}^{p} \frac{|\min\{\tau(i),k+1\} - \min\{\sigma(i),k+1\}|}{\min\{\tau(i),k+1\} + \min\{\sigma(i),k+1\}}$. The expected value of the Canberra distance with location parameter is

$$\begin{aligned} E\{\mathrm{Ca}^{(k+1)}(S_p)\} &= \frac{k}{p}\left(\left(2k+2+\frac{1}{2k}\right)H_{2k} - \left(2k+2+\frac{1}{4k}\right)H_k - \left(k+\frac{3}{2}\right)\right) \\ &\quad + \frac{2(p-k)}{p}(2(k+1)(H_{2k+1} - H_{k+1}) - k) \\ &= \frac{(k+1)(2p-k)}{p}\log(4) - \frac{2kp+3p-k-k^2}{p} + o(1)\ . \end{aligned}$$

Details of the top-k extension are described in [6], with application for profil ng in synthetic and oncological datasets and enrichment techniques.

If ranked lists of different lenghts $\sigma \in S_{l_1}$, $\tau \in S_{l_2}$ are to be compared by a distance d, the possible approach is to consider quotient groups of a larger group S_p, for $l_1, l_2 \leq p$: for f a suitable function, $d(\sigma, \tau) = f\left(\{d(\alpha, \beta) \colon \alpha \in S_{\tau_1}, \beta \in S_{\tau_2}\}\right)$, for $S_\xi = \{\rho \in S_p | \rho_{|\mathrm{supp}(\xi)} = \xi\}$. In [10] the case of d the Canberra distance and f the mean function is considered.

3 Applications on molecular signatures

The role of the Canberra distance as a stability indicator for ranked lists of molecular biomarkers was fi st described in [6], *e.g.*, as a measure of disarray among ranked lists produced by different classifie s or laboratories. On large scale projects such as the US-FDA led initiative MAQC-II [11], the stability indicator was applied coupled with accuracy metrics to evaluate potential sources of variability for microarray analysis, as the impact of batch preprocessing or normalization methods. Stability analysis was also applied in other ranking problems such as the analysis of gene enriched lists and the comparison of filte ing methods [6, 10]. In the f rst application in this section we present how to use the Canberra distance for model selection. In a second data intensive application we show how to quantify similarity among sets of ranked lists of differentially expressed probes under several conditions, available from the Broad Institute CMAP[2] repository of signatures.

List distances for model selection In a molecular profilin task, the degree of difference among the ranked lists produced during the feature selection process is as relevant as the classifie performance when selecting a model. We propose to couple the mean of the mutual Canberra distances among feature lists (as the disarray measure) with the classifie Area Under the ROC Curve (AUC) for model selection. As an example, we can choose a model with the optimal number of features in a binary classificatio task. The task aims at identifying positive TMPRSS2-ERG gene fusion cases from negative ones given two different cohorts of prostate cancer patients (Swedish WW, 103 positive and 352 negative and US PHS Cohort, 41 + 60). The Setlur dataset[3] consists of 6144 gene expression values from a custom Illumina DASL Assay originally described in [12]. We apply 10 times a 5-fold cross validation on the two cohorts separately, using Spectral Regression Discriminant Analysis [13] (SRDA) with $\alpha = 10^3$ as the classifie and the feature weighting algorithm, and Entropy based Recursive Feature Elimination (ERFE) as the ranking procedure [14]. In the diagnostic plot of Fig.1 we show the points corresponding to the different feature subsets for each model on an AUC versus normalized Canberra Distance space. The plot can be used as an effective tool for model selection purposes: although models have similar AUC, they show widely different levels of similarity of the corresponding ranked lists. For instance, on both cohorts, the model with 15 features has a better list similarity measure than the models with 50 and 100 features which reach the best AUC, without paying too much in terms of performance. From this perspective, the best possible compromise between similarity and performance is represented by the model with 25 features. Finally, note that the top-5 lists have small Canberra distance, but the selected features show relatively worse predictivity. On the other hand, due to the large number of irrelevant features

[2] Connectivity Map 02: `http://www.broadinstitute.org/cmap/`

[3] Available on GEO `http://www.ncbi.nih.gov/geo`, accession number GSE8402.

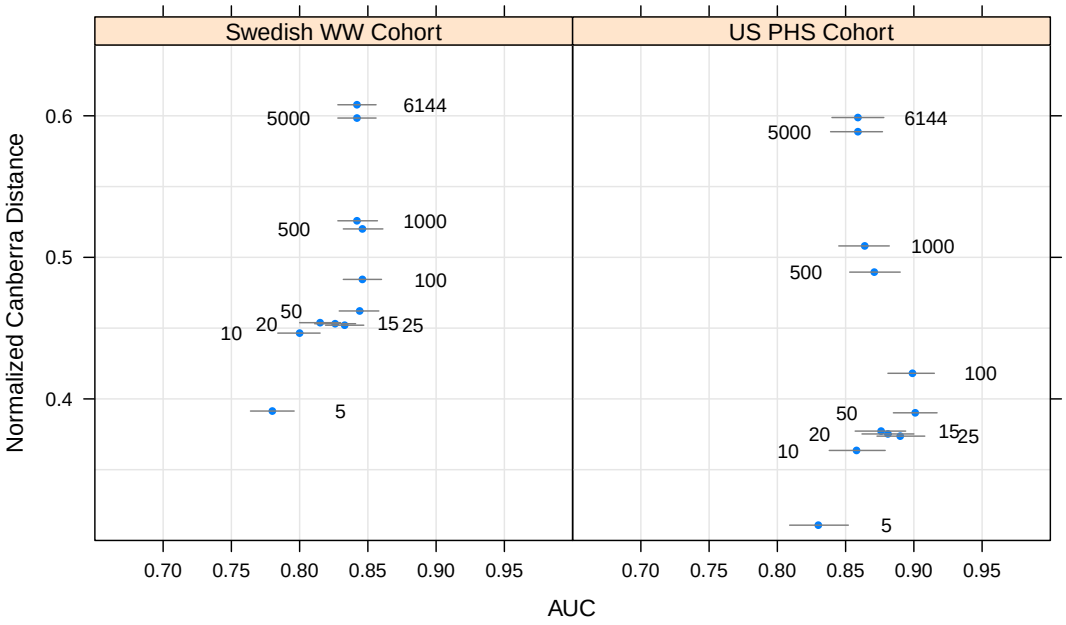

Figure 1: Accuracy-Similarity plots for the Setlur prostate cancer dataset, separately for the Swedish and US cohorts. Each point corresponds to a SRDA/ERFE model with different number of features, reported aside. Coordinates of each point are given by the AUC (bars: 95% bootstrap confidenc intervals) and the normalized Canberra distance, averaged on the $10 \times 5-$CV resampling.

(which are consistently low ranked with almost random weights), the models with more variables are characterized by an higher level of dissimilarity, which is however not heavily reflectin on the corresponding AUC values.

Similarity among CMAP signatures We consider the CMAP project that collects 6,100 lists of 22,283 gene expression probes from the HG-U133A platform and its variants [15]. The probes in the CMAP lists are decreasingly ranked according to their differential expressions between a treated and an untreated cell line by a specifi compound at different concentrations. In total, 1,309 compounds are tested at concentrations from 10^{-2} M down to 10^{-8} M on f ve different human tumoral cell lines in different batches. We computed the 18,601,950 values of the Canberra distances between all distinct pairs of lists by using the `mlpy` library on a high performance facility. Given the four compounds Haloperidol, LY-294002, Tanespimycin and Trichostatin A, the corresponding subset of z lists is extracted. Given the subset, potential sources of variability are different concentration, batch, platform and cell line. The distribution of the $z(z-1)/2$ Canberra distances normalized by the expected value (2) $E\{\mathrm{Ca}(S_{22283})\}$ is plotted in Fig.2(a). Distributions were computed using the R package `stats`. The mean distance is also explicitly drawn in the plots: a smaller mean (bottom panels) indicates an higher similarity among lists produced by the corresponding compounds, while mean values close to one (Haloperidol) indicate that the set of lists is not far from being randomly extracted within S_{22283}. The four compounds exhibit different histogram shapes and means, suggesting differences in variability. The Trichostatin A lists are the most uniform through the different experimental conditions. Trichostatin A and Tanespimycin have mean close to 0.8, while most of the distances of the other two compounds lie above this threshold. Moreover, the similarity among the histograms of Fig. 2(b) for the three cell lines HL60, MCF7 and PC3, accounting for the 99.5% of the data, shows that, on the CMAP data, cell line does not impact on variability.

Acknowledgments

The authors acknowledge funding by the European Union FP7 Project HiperDART. They also thank Andrea Gobbi for his help with the mathematical analysis, Davide Albanese for his precious work in developing and maintaining `mlpy`, and Silvano Paoli for his support with the HPC facility.

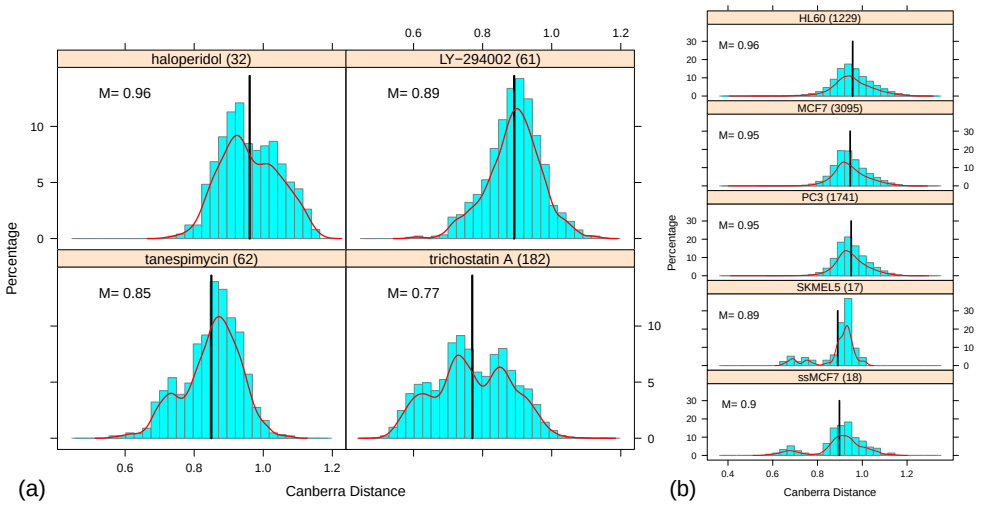

Figure 2: (a) Histogram and distribution density of the normalized Canberra distance among CMAP lists for the compounds Haloperidol, LY-294002, Tanespimycin and Trichostatin A (in parentheses: number of lists for the compound). The black vertical bar is the mean of all values for a compound. (b) Same for the f ve cell lines HL60, MCF7, PC3, SKMEL5 and ssMCF7.

References

[1] G.N. Lance and W.T. Williams. Mixed-Data Classificato y Programs I - Agglomerative Systems. *Aust. Comput. J.*, 1(1):15–20, 1967.

[2] D.E. Critchlow. *Metric methods for analyzing partially ranked data*. LNS 34. Springer, 1985.

[3] W. Hoeffding. A Combinatorial Central Limit Theorem. *Ann. Math. Stat.*, 22(4):558–566, 1951.

[4] P. Diaconis. *Group representations in probability and statistics*. Institute of Mathematical Statistics Lecture Notes – Monograph Series 11. IMS, 1988.

[5] P. Diaconis and R.L. Graham. Spearman's Footrule as a Measure of Disarray. *J. R. Stat. Soc. B*, 39:262–268, 1977.

[6] G. Jurman, S. Merler, A. Barla, S. Paoli, A. Galea, and C. Furlanello. Algebraic stability indicators for ranked lists in molecular profiling *Bioinformatics*, 24(2):258–264, 2008.

[7] R.L. Graham, D.E. Knuth, and O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison Wesley, Reading, Mass., 1989.

[8] J. Spieß. Some Identities Involving Harmonic Numbers. *Math. Comput.*, 55(192):839–863, 1990.

[9] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top-k lists. *SIAM J. Disc. Math.*, 17:134–160, 2003.

[10] G. Jurman, S. Riccadonna, R. Visintainer, and C. Furlanello. Algebraic Comparison of Partial Lists. Submitted, 2009.

[11] The MicroArray Quality Control (MAQC) Consortium. The MAQC-II Project: A comprehensive study of common practices for the development and validation of microarray-based predictive models. Submitted, 2009.

[12] S.R. Setlur, K.D. Mertz, Y. Hoshida, F. Demichelis, M. Lupien, S. Perner, A. Sboner, Y. Pawitan, O. Andrn, L.A. Johnson, J. Tang, H.O. Adami, S. Calza, A.M. Chinnaiyan, D. Rhodes, S. Tomlins, K. Fall, L.A. Mucci, P.W. Kantoff, M.J. Stampfer, S.O. Andersson, E. Varenhorst, J.E. Johansson, M. Brown, T.R. Golub, and M.A. Rubin. Estrogen-dependent signaling in a molecularly distinct subclass of aggressive prostate cancer. *J. Natl. Cancer Inst.*, 100(11):815–825, 2008.

[13] D. Cai, H. Xiaofei, and J. Han. Srda: An efficien algorithm for large-scale discriminant analysis. *IEEE Transactions on Knowledge and Data Engineering*, 20(1):1–12, 2008.

[14] C. Furlanello, M. Serafini S. Merler, and G. Jurman. Entropy-Based Gene Ranking without Selection Bias for the Predictive Classificatio of Microarray Data. *BMC Bioinformatics*, 4(1):54, 2003.

[15] J. Lamb, E.D. Crawford, D. Peck, J.W. Modell, I.C. Blat, M.J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K.N. Ross, M. Reich, H. Hieronymus, G. Wei, S.A. Armstrong, S.J. Haggarty, P.A. Clemons, R. Wei, S.A. Carr, E.S. Lander, and T.R. Golub. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*, 313(5795):1929–1935, 2006.