The Kendall and Mallows Kernels for Permutations

Yunlong Jiao YUNLONG.JIAO@MINES-PARISTECH.FR
Jean-Philippe Vert JEAN-PHILIPPE.VERT@MINES-PARISTECH.FR
MINES ParisTech – CBIO, PSL Research University, Institut Curie, INSERM U900, Paris, France

Abstract

We show that the widely used Kendall tau correlation coefficient, and the related Mallows kernel, are positive definite kernels for permutations. They offer computationally attractive alternatives to more complex kernels on the symmetric group to learn from rankings, or to learn to rank. We show how to extend the Kendall kernel to partial rankings or rankings with uncertainty, and demonstrate promising results on high-dimensional classification problems in biomedical applications.

1. Introduction

Kernel-based algorithms have been proved successful in numerous applications and enjoy great popularity in the machine learning community (Cortes & Vapnik, 1995; Vapnik, 1998; Schölkopf & Smola, 2002; Shawe-Taylor & Cristianini, 2004). The essential idea behind these methods is to define a positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ over an input space $\mathcal{X}$, which can often be thought of as a measure of similarity, and which implicitly defines an embedding $\Phi : \mathcal{X} \to \mathcal{F}$ of the input space $\mathcal{X}$ to a Hilbert space $\mathcal{F}$ in which the kernel becomes an inner product:

$$\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \quad K(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle_{\mathcal{F}}.$$

Kernel methods operate implicitly in the Hilbert space $\mathcal{F}$, which can be high-dimensional, by only manipulating the kernel function between data. This *kernel trick* is particularly interesting when $K(\mathbf{x}, \mathbf{x}')$ is inexpensive to evaluate, compared to $\Phi(\mathbf{x})$ and $\Phi(\mathbf{x}')$. In particular, kernel methods have found many applications where the input data are discrete or structured, such as strings or graphs, thanks to the development of numerous kernels for these data (Haussler, 1999; Kashima et al., 2003; Gärtner et al., 2004; Shawe-Taylor & Cristianini, 2004; Schölkopf et al., 2004; Vishwanathan et al., 2009).

In this paper, we are interested in developing and studying positive definite kernels for a particular type of discrete data, namely, *permutations*. A permutation is a 1-to-1 mapping from a finite set into itself. Permutations are ubiquitous in many applications involving rankings or partial rankings, such as analyzing data describing the preferences or votes of a population (Diaconis, 1988), learning or tracking correspondances between sets of objects (Huang et al., 2009), or estimating a single ranking that best represents a collection of individual rankings (Ailon et al., 2008). Another potentially rich source of ranking data comes from real-valued vectors in which the relative order of the values of multiple features is more important than their absolute magnitude. For example, in the case of high-dimensional gene expression data, Geman et al. (2004) showed that simple classifiers based on binary comparisons between the expression of different genes in a sample show competitive prediction accuracy with much more complex classifiers built on the quantitative gene expression levels, a line of thoughts that have been further investigated by Tan et al. (2005); Xu et al. (2005); Lin et al. (2009). In these approaches, a n-dimensional vector is thus first transformed into a permutation by sorting its entries, and a classifier is trained on the resulting permutations.

Working with permutations is, however, computationally challenging. There are $n!$ permutations of n items, suggesting that various simplifications or approximations are necessary in pursuit of efficient algorithms to analyze or learn permutations. Such simplifications include for example, reducing ranks to a series of binary decisions (Ailon et al., 2008; Balcan et al., 2008), or estimating a parametric distribution over permutations (Lebanon & Mao, 2008; Helmbold & Warmuth, 2009; Huang et al., 2009).

In this context, it is surprising that relatively little attention has been paid to the problem of defining positive definite kernels between permutations, which could pave the way to the use of computationally efficient kernel methods in problems involving permutations. A notable exception is the work of Kondor (2008); Kondor & Barbosa (2010), who exploit the fact that the set of permutations endowed with the composition operation forms a group,

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

called the *symmetric group* (Diaconis, 1988; Huang et al., 2009), on which right-invariant positive definite kernels are fully characterized by Bochner's theorem (Kondor, 2008; Fukumizu et al., 2008). They derive interesting kernels, such as a diffusion kernel for rankings or partial rankings, which however remains prohibitive to compute when the number of ranked items is large.

In this paper we take one step further towards the development of computationally attractive kernels for permutations, by noticing that two widely-used and computationally efficient measures of similarity between permutations, the Kendall tau correlation coefficient and the Mallows kernel, are positive definite. These kernels compare two permutations of n items in terms of $\binom{n}{2}$ pairwise comparisons, but can be computed in $O(n \log n)$, paving the way to the use of kernel methods for problems involving rankings or permutations of a large number of items. In particular, the feature space of the Kendall and Mallows kernels is precisely the space of binary pairwise comparisons defined by Geman et al. (2004), and we show that instead of selecting a few features in this space as the Top Scoring Pairs (TSP)-family classifiers do (Geman et al., 2004; Tan et al., 2005; Xu et al., 2005; Lin et al., 2009), one can simply work with *all* pairs with the kernel trick. We further study how the new kernels can be extended to partial rankings, and to uncertain rankings which is particularly relevant when the rankings are obtained by sorting a real-valued vector where ties or near-ties occur. Finally, we demonstrate promising results of the underlying kernels on a large benchmark of high-dimensional biomedical data classification problems.

2. The Kendall and Mallows Kernel for Total Rankings

Let us first fix some notations. Given a list of n items $\{x_1, x_2, \dots, x_n\}$, a *total ranking* is a strict ordering of the n items of the form

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_n} \,, \tag{1}$$

where $\{i_1, \dots, i_n\}$ are distinct indices in $\{1, 2, \dots, n\} =: [1, n]$. Each total ranking can equivalently be represented by a *permutation* $\sigma : [1, n] \to [1, n]$ such that $\sigma(i) \neq \sigma(j)$ for $i \neq j$, and $\sigma(i) = j$ indicates that a ranker assigns rank j to item i. For example, the ranking $x_2 \succ x_4 \succ x_3 \succ x_1$ is associated to the permutation $\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 2 & 3 \end{pmatrix}$, meaning $\sigma(1) = 1$, $\sigma(2) = 4$, etc.. There are $n!$ different total rankings, and we denote by $\mathbb{S}_n$ the set of all permutations over n items. Endowed with the composition operation $\sigma_1\sigma_2(i) = \sigma_1(\sigma_2(i))$, $\mathbb{S}_n$ is a group called the *symmetric group*.

Given two permutations $\sigma, \sigma' \in \mathbb{S}_n$, the number of concordant and discordant pairs between σ and σ' are respectively

$$\begin{aligned} n_c(\sigma, \sigma') = \sum_{i<j} \big[& \mathbb{1}_{\{\sigma(i)<\sigma(j)\}} \mathbb{1}_{\{\sigma'(i)<\sigma'(j)\}} \\ & + \mathbb{1}_{\{\sigma(i)>\sigma(j)\}} \mathbb{1}_{\{\sigma'(i)>\sigma'(j)\}} \big] \,, \\ n_d(\sigma, \sigma') = \sum_{i<j} \big[& \mathbb{1}_{\{\sigma(i)<\sigma(j)\}} \mathbb{1}_{\{\sigma'(i)>\sigma'(j)\}} \\ & + \mathbb{1}_{\{\sigma(i)>\sigma(j)\}} \mathbb{1}_{\{\sigma'(i)<\sigma'(j)\}} \big] \,. \end{aligned}$$

As their names suggest, $n_c(\sigma, \sigma')$ and $n_d(\sigma, \sigma')$ count how many pairs of items are respectively in the same or opposite order in the two rankings σ and σ'. n_d is frequently used as a distance between permutations, often under the name *Kendall tau distance*, and underlies two popular similarity measures between permutations:

- The *Mallows kernel* defined for any $\lambda \geq 0$ by
$$K_M^\lambda(\sigma, \sigma') = e^{-\lambda n_d(\sigma, \sigma')} \,, \tag{2}$$
- The *Kendall kernel* defined as
$$K_\tau(\sigma, \sigma') = \frac{n_c(\sigma, \sigma') - n_d(\sigma, \sigma')}{\binom{n}{2}} \,. \tag{3}$$

The Mallows kernel plays a role on the symmetric group similar to the Gaussian kernel on Euclidean space, for example for statistical modeling of permutations (Mallows, 1957; Critchlow, 1985; Fligner & Verducci, 1986; Meilă et al., 2007) or nonparametric smoothing (Lebanon & Mao, 2008), and the Kendall kernel (Kendall, 1938; 1948) is probably the most widely used measure of rank correlation coefficient. In spite of their pervasiveness, to the best of our knowledge the following property has been overlooked:

Theorem 1. *The Mallows kernel K_M^λ, for any $\lambda \geq 0$, and the Kendall kernel K_τ are positive definite.*

Proof. Consider the mapping $\Phi : \mathbb{S}_n \to \mathbb{R}^{\binom{n}{2}}$ defined by

$$\Phi(\sigma) = \left(\frac{1}{\sqrt{\binom{n}{2}}} (\mathbb{1}_{\{\sigma(i)>\sigma(j)\}} - \mathbb{1}_{\{\sigma(i)<\sigma(j)\}}) \right)_{1 \leq i < j \leq n} .$$

Then one immediately sees that, for any $\sigma, \sigma' \in \mathbb{S}_n$,

$$K_\tau(\sigma, \sigma') = \Phi(\sigma)^\top \Phi(\sigma') \,,$$

showing that K_τ is positive definite, and that

$$\begin{aligned} \|\Phi(\sigma) - \Phi(\sigma')\|^2 &= K_\tau(\sigma, \sigma) + K_\tau(\sigma', \sigma') - 2K_\tau(\sigma, \sigma') \\ &= 1 + 1 - 2\Big(\frac{n_c(\sigma, \sigma') - n_d(\sigma, \sigma')}{\binom{n}{2}} \Big) \\ &= \frac{4}{\binom{n}{2}} n_d(\sigma, \sigma') \,, \end{aligned}$$

showing that n_d is conditionally positive definite (Schoenberg, 1938) and therefore that K_M^λ is positive definite for all $\lambda \geq 0$. □

Although the Kendall and Mallows kernels correspond respectively to a linear and Gaussian kernel on a $\binom{n}{2}$-dimensional embedding of $\mathbb{S}_n$ such that they can in particular be computed in $O(n^2)$ time by a naive implementation of pair-by-pair comparison, it is interesting to notice that more efficient algorithms based on divide-and-conquer strategy can significantly speed up the computation, up to $O(n \log n)$ using a technique based on Merge Sort algorithm (Knight, 1966). Computing in $O(n \log n)$ a kernel corresponding to a $O(n^2)$-dimensional embedding of $\mathbb{S}_n$ is a typical example of the *kernel trick*, which allows to scale kernel methods to larger values of n than what would be possible for methods working with the explicit embedding.

3. Extension to Partial Rankings

In this section we show how the Kendall kernel K_τ can efficiently be adapted to *partial rankings*, a situation frequently encountered in practice. For example, in a movie recommendation system, each user only grades a subset of movies that he has watched according to personal preference. As another example, in a chess tournament, each match results in an relative ordering between two contestants, and one would like to find globally a single ranking that best represents the large collection of binary outcomes.

As opposed to a total ranking (1), a partial ranking is in general of the form $X_1 \succ X_2 \succ \cdots \succ X_k$, where $X_1, \ldots, X_k$ are k disjoint subsets of n items $\{x_1, \ldots, x_n\}$. For example, $\{x_2, x_4\} \succ x_6 \succ \{x_3, x_8\}$ in a social survey could represent the fact that items 2 and 4 are ranked higher by an interviewee than item 6, which itself is ranked higher than items 3 and 8. Note that it is uninformative of the relative order of items 2 and 4, nor of how item 1 is rated.

To extend any kernel K over $\mathbb{S}_n$ to a kernel over the set of partial rankings, we represent a partial ranking by the set $R \subset \mathbb{S}_n$ of permutations which are compatible with all partial orders described by the partial ranking, and adopt the *convolution kernel* between two partial rankings R and R' as

$$K(R, R') = \frac{1}{|R||R'|} \sum_{\sigma \in R} \sum_{\sigma' \in R'} K(\sigma, \sigma'). \tag{4}$$

As a convolution kernel, it is positive definite as long as K is positive definite (Haussler, 1999). However, a naive implementation to compute (4) typically requires $O((n-k)!(n-k')!)$ operations when the number of observed items in R, R' is respectively k and k', which can quickly become prohibitive.

We now show that we can circumvent the computational burden of naively implementing (4) with the Kendall kernel on at least two particular cases of partial rankings:

a) An *interleaving partial ranking* is of the form

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k}, \quad k \leq n,$$

where we have a total ranking for k out of n items. This type of partial ranking is frequently encountered in real life, for example if each person is able to vote for only a few candidates in an election example, or in case there exist interleaved inaccessible values. The interleaving partial ranking corresponds to the set of permutations compatible with it:

$$A_{i_1,\ldots,i_k} = \{\sigma \in \mathbb{S}_n | \sigma(i_a) > \sigma(i_b) \text{ if } a < b, a, b \in [1, k]\}. \tag{5}$$

b) A *top-k partial ranking* is of the form

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k} \succ X_{\text{rest}}, \quad k \leq n,$$

where we have a total ranking for k out of n items and also know that these k items are ranked higher than all the other items. For example, the top k hits returned by a search engine leads to a top k partial ranking, or so does a survey on top k favorite flavors of ice cream. The top-k partial ranking corresponds to the set of compatible permutations

$$B_{i_1,\ldots,i_k} = \{\sigma \in \mathbb{S}_n | \sigma(i_a) = n + 1 - a, a \in [1, k]\}. \tag{6}$$

Indeed, the following holds:

Theorem 2. *The Kendall kernel between two interleaving partial rankings of respectively k and m observed items, or between a top-k partial ranking and a top-m partial ranking, of form (4) can be computed in $O(k \log k + m \log m)$ operations.*

Proof. We prove this theorem by showing explicitly how to compute the Kendall kernel between two partial rankings of type (5) or (6) in supplementary material. Checking the correctness of the algorithms is tedious but easy once we notice that most of the additive terms in the Kendall kernel for partial rankings cancel each other out because of the symmetry of the compatible set of full permutations. This also means that such a fast algorithm is not likely to exist for the Mallows kernel over partial rankings taking form (4). Note that in both algorithms, the first step is the computationally most expensive one, where we need to identify the total ranking restricted to the observed items in partial rankings. This can be achieved by any sorting algorithm, leading the algorithms to time complexity $O(k \log k + m \log m)$ overall. □

4. The Kendall Kernel for Quantitative Vectors

When data to analyze are n-dimensional real-valued quantitative vectors, converting them to permutations in $\mathbb{S}_n$ by ranking their entries can be beneficial in cases where we trust more the relative ordering of the values than their absolute magnitudes. For example, an interesting line of work in the analysis of gene expression data promotes the development of classifiers built upon relative reversals of pairwise feature comparison, based on the observations that gene expression measurements are subject to various measurement errors such as technological biases and normalization issues, while assessing whether a gene is more expressed than another gene is generally a more robust task (Geman et al., 2004; Tan et al., 2005; Xu et al., 2005; Lin et al., 2009). This suggests that the Kendall kernel can be relevant for analyzing quantitative vectors as well. It now takes the form of dot product as

$$K_\tau(\mathbf{x}, \mathbf{x}') = \Phi(\mathbf{x})^\top \Phi(\mathbf{x}')\,, \tag{7}$$

where $\Phi : \mathbb{R}^n \to \mathbb{R}^{\binom{n}{2}}$ is defined for $\mathbf{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$ by

$$\Phi(\mathbf{x}) = \left(\frac{1}{\sqrt{\binom{n}{2}}} \left(\mathbb{1}_{\{x_i > x_j\}} - \mathbb{1}_{\{x_i < x_j\}} \right) \right)_{1 \le i < j \le n} . \tag{8}$$

In this case, the interpretation of the Kendall kernel in terms of concordant and discordant pairs (3) is obviously still valid, with the caveats that in the presence of ties between entries of $\mathbf{x}$, say two coordinates i and j such that $x_i = x_j$, the tied pair $\{x_i, x_j\}$ will be neither concordant nor discordant. This implies in particular that if $\mathbf{x}$ has ties or so does $\mathbf{x}'$, then $|K_\tau(\mathbf{x}, \mathbf{x}')| < 1$ strictly. As for permutations, the fast implementation of Kendall kernel also applies to quantitative vectors in $O(n \log n)$ time, even in the presence of ties (Knight, 1966).

Feature mapping (8) is by construction very sensitive to the presence of entry pairs that are "almost ties" but not "sheer ties". In fact, each entry of $\Phi(\mathbf{x})$ is, up to a normalization constant, the Heaviside step function which takes discrete values in $\{-1, 0, +1\}$, and thus can change abruptly even when $\mathbf{x}$ changes slightly but reverses the order of two entries whose values are close. We propose to make the mapping more robust by assuming a random noise ϵ added to $\mathbf{x}$ and checking where $\Phi(\mathbf{x} + \epsilon)$ is, on average (similarly to, e.g., Muandet et al., 2012). In other words, we consider a smoother mapping $\Psi : \mathbb{R}^n \to \mathbb{R}^{\binom{n}{2}}$ defined by

$$\Psi(\mathbf{x}) = \mathbb{E}\Phi(\mathbf{x} + \epsilon), \tag{9}$$

where ϵ is a n-dimensional random vector, and the corresponding kernel

$$G\left(\mathbf{x}, \mathbf{x}'\right) = \Psi(\mathbf{x})^\top \Psi(\mathbf{x}')\,. \tag{10}$$

Denoting $\tilde{\mathbf{x}} := \mathbf{x} + \epsilon$ the randomly jittered vector, we deduce from (8) that Ψ is equivalently written as

$$\Psi(\mathbf{x}) = \left(\frac{1}{\sqrt{\binom{n}{2}}} \left(\mathbb{P}\left(\tilde{x}_i > \tilde{x}_j\right) - \mathbb{P}\left(\tilde{x}_i < \tilde{x}_j\right) \right) \right)_{1 \le i < j \le n} .$$

Depending on the noise distribution, various kernels are obtained. In particular, assuming that $\epsilon \sim (\mathcal{U}[-\frac{a}{2}, \frac{a}{2}])^n$ the n-dimensional uniform noise of window size a centered at 0, the (i,j)-th entry of $\Psi(\mathbf{x})$ for all $i < j$ becomes

$$\Psi_{ij}(\mathbf{x}) = \frac{1}{\sqrt{\binom{n}{2}}} g_a(x_i - x_j)\,, \tag{11}$$

where

$$g_a(t) := \begin{cases} 1 & t \ge a \\ 2(\frac{t}{a}) - (\frac{t}{a})^2 & 0 \le t \le a \\ 2(\frac{t}{a}) + (\frac{t}{a})^2 & -a \le t \le 0 \\ -1 & t \le -a \end{cases} . \tag{12}$$

g_a is odd, continuous, piecewise quadratic between $[-a, a]$ and constant elsewhere at ± 1, and thus can be viewed as smoothed version of the Heaviside step function to compare any two entries x_i and x_j from their difference $x_i - x_j$.

Although the kernel (10) can be an interesting alternative to the Kendall kernel (7), we unfortunately lose for G the computational trick that allows to compute K_τ in $O(n \log n)$. Specifically, we have two ways to compute G:

- *Exact evaluation.* The first alternative is to compute explicitly the $\binom{n}{2}$-vector representation $\Psi(\mathbf{x})$ in the feature space by (11) and (12), and then take the dot product to obtain G. The computational cost is therefore linear with the dimension of the feature space, i.e. $O(n^2)$.

- *Monte Carlo approximation.* The second alternative requires the observation that $G(\mathbf{x}, \mathbf{x}') = \mathbb{E}\Phi(\tilde{\mathbf{x}})^\top \mathbb{E}\Phi(\tilde{\mathbf{x}}') = \mathbb{E}K_\tau(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$, where $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}'$ are independently noise-perturbed versions of $\mathbf{x}$ and $\mathbf{x}'$, and we can thus approximate G by a D^2-sample mean:

  $$G_D(\mathbf{x}, \mathbf{x}') = \frac{1}{D^2} \sum_{i=1}^{D} \sum_{j=1}^{D} K_\tau\left(\tilde{\mathbf{x}}^i, \tilde{\mathbf{x}}'^j\right)\,, \tag{13}$$

  where $\tilde{\mathbf{x}}^1, \ldots, \tilde{\mathbf{x}}^D$ are i.i.d. noisy versions of $\mathbf{x}$, and the same for $\mathbf{x}'$. Since computing K_τ has complexity $O(n \log n)$, the computational cost of G_D is $O(D^2 n \log n)$

We note that the second alternative is faster to compute than the first one as long as, up to constants, $D^2 < n / \log n$, and small values of D are thus favored. In this case, however, the approximation performance can be unappealing. To better understand the trade-off between the two alternatives, there is therefore a pressing need to understand how large D should be to ensure that the approximation error is not detrimental to learning with the approximate kernel G_D instead of G.

For this purpose, let us consider for example the case where the smoothed kernel G is used to train a Support Vector Machine (SVM) from a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \subset (\mathbb{R}^n \times \{-1,+1\})^m$, specifically to estimate a function $h(\mathbf{x}) = \mathbf{w}^\top \Psi(\mathbf{x})$ by solving

$$\min_{\mathbf{w}} F(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \widehat{R}(\mathbf{w}), \tag{14}$$

where $\widehat{R}(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^m \ell(y_i \mathbf{w}^\top \Psi(\mathbf{x}_i))$ is the empirical loss, with $\ell(y_i \mathbf{w}^\top \Psi(\mathbf{x}_i)) = \max(0, 1 - y_i \mathbf{w}^\top \Psi(\mathbf{x}_i))$ the hinge loss associated to the i-th point, λ the regularization parameter. Now suppose that instead of training the SVM with smoothed feature mapping on the original points $\{\Psi(\mathbf{x}_i)\}_{i=1,\dots,m}$, we first randomly jitter $\{\mathbf{x}_i\}_{i=1,\dots,m}$ at each point D times, resulting in $\{\tilde{\mathbf{x}}_i^j\}_{i=1,\dots,m; j=1,\dots,D}$, and then replace each $\Psi(\mathbf{x}_i)$ by the D-sample empirical average of jittered points mapped by Φ into the feature space, that is

$$\Psi_D(\mathbf{x}_i) := \frac{1}{D}\sum_{j=1}^{D} \Phi(\tilde{\mathbf{x}}_i^j)\,.$$

Note that $\Psi_D(\mathbf{x}_i)^\top \Psi_D(\mathbf{x}_j) = G_D(\mathbf{x}_i, \mathbf{x}_j)$, hence training a SVM with the Monte Carlo approximate G_D instead of exact version G is equivalent to solving (14) with $\{\Psi_D(\mathbf{x}_i)\}_{i=1,\dots,m}$ in the hinge loss instead of $\{\Psi(\mathbf{x}_i)\}_{i=1,\dots,m}$. The following theorem quantifies the approximation performance in terms of objective function F.

Theorem 3. *For any $0 \le \delta \le 1$, the solution $\widehat{\mathbf{w}}_D$ of the SVM trained with the Monte Carlo approximation (13) with D random jittered samples for each training point satisfies, with probability greater than $1 - \delta$,*

$$F(\widehat{\mathbf{w}}_D) \le \min_{\mathbf{w}} F(\mathbf{w}) + \sqrt{\frac{8}{\lambda D}}\left(2 + \sqrt{8 \log \frac{m}{\delta}}\right)\,.$$

The proof is left in supplementary material. It is known that compared to the exact solution of (14), an $O(m^{-1/2})$-approximate solution is sufficient to reach the optimal statistical accuracy (Bottou & Bousquet, 2008). This accuracy can be attained in our analysis when $D = O(m/\lambda)$, and since typically $\lambda \sim m^{-1/2}$ (Steinwart, 2005), this suggests that it is sufficient to take D of order $m^{3/2}$. Going back to the comparison strategy of the two alternatives G and G_D, we see that the computational cost of computing the full $m \times m$ Gram matrix with the exact evaluation is $O(m^2 n^2)$, while the cost of computing the approximate Gram matrix with $D = O(m^{3/2})$ random samples is $O(m^2 D^2 n \log n) = O(m^5 n \log n)$. This shows that, up to constants and logarithmic terms, the Monte Carlo approach is interesting when $m = o(n^{1/3})$, otherwise the exact evaluation using explicit computation in the feature space is preferable.

5. Relationship to the Diffusion Kernel on $\mathbb{S}_n$

It is interesting to relate the Mallows kernel (2) to the diffusion kernel on symmetric group proposed by Kondor & Barbosa (2010), which is the diffusion kernel (Kondor & Lafferty, 2002) on the Cayley graph of $\mathbb{S}_n$ generated by adjacent transpositions with left-multiplication. This graph, illustrated for a specific case $n = 4$ in Figure 1, is defined by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \mathbb{S}_n$ as vertices, and undirected edge set $\mathcal{E} = \{\{\sigma, \pi\sigma\} : \sigma \in \mathbb{S}_n, \pi \in Q\}$, where $Q = \{(i, i+1) | i = 1, \dots, n-1\}$ the set of all adjacent transpositions. Note Q is symmetric in the sense that $\pi \in Q \Leftrightarrow \pi^{-1} \in Q$, and the graph adjacency relation is a right-invariant relation, that is $\sigma \sim \sigma' \Leftrightarrow \sigma'\sigma^{-1} \in Q$. The corresponding graph Laplacian is the matrix Δ with

$$\Delta_{\sigma,\sigma'} = \begin{cases} 1 & \text{if } \sigma \sim \sigma' \\ -(n-1) & \text{if } \sigma = \sigma' \\ 0 & \text{otherwise} \end{cases},$$

where $n - 1$ is the degree of vertex σ (number of edges connected with vertex σ), and the diffusion kernel on $\mathbb{S}_n$ is finally defined as

$$K_{\text{dif}}^{\beta}(\sigma, \sigma') = [e^{\beta\Delta}]_{\sigma,\sigma'} \tag{15}$$

for some diffusion parameter $\beta \in \mathbb{R}$, where $e^{\beta\Delta}$ is the matrix exponential. K_{dif}^{β} is a right-invariant kernel on the symmetric group (Kondor & Barbosa, 2010, Proposition 2), and we denote by $\kappa_{\text{dif}}^{\beta}$ the positive definite function induced by K_{dif}^{β} such that $K_{\text{dif}}^{\beta}(\sigma, \sigma') = \kappa_{\text{dif}}^{\beta}(\sigma'\sigma^{-1})$. Since it is straightforward that the Mallows kernel K_M^{λ} is also right-invariant, we denote by κ_M^{λ} the positive definite function induced by the Mallows kernel K_M^{λ} such that $K_M^{\lambda}(\sigma, \sigma') = \kappa_M^{\lambda}(\sigma'\sigma^{-1})$.

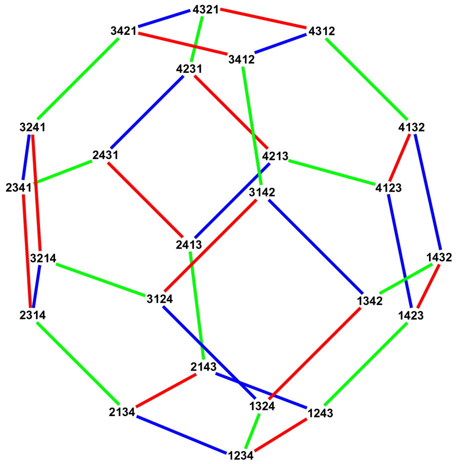

Figure 1. Cayley graph of $\mathbb{S}_4$, generated by the transpositions (1 2) in blue, (2 3) in green, and (3 4) in red.

Interestingly, the Mallows kernel has a similar interpretation. Indeed, it is well-known that the Kendall tau distance

$n_d(\sigma, \sigma')$ is the minimum number of adjacent swaps required to bring σ to σ', i.e. $n_d(\sigma, \sigma')$ equals to the shortest path distance on the Cayley graph, or simply written

$$n_d(\sigma, \sigma') = d_{\mathcal{G}}(\sigma, \sigma'). \tag{16}$$

Different from the diffusion kernel for which communication between permutations is a diffusion process over the graph, the Mallows kernel $K_M^\lambda = e^{-\lambda n_d} = e^{-\lambda d_{\mathcal{G}}}$ considers exclusively the shortest path over the graph when expressing the similarity between permutations.

A notable advantage of the Mallows kernel over the diffusion kernel is that the Mallows kernel enjoys faster evaluation. On one hand if data examples are total rankings, i.e. $\sigma, \sigma' \in \mathbb{S}_n$, evaluating $K_{\text{dif}}^\beta(\sigma, \sigma')$ would require exponentiating a $n!$-dimensional Laplacian matrix by naive implementation, and can reduce to exponentiating matrices of smaller sizes by careful analysis in the Fourier space, which still remains problematic if working dimension n is large (Kondor & Barbosa, 2010). However, evaluating $K_M^\lambda(\sigma, \sigma')$ only takes $O(n \log n)$ time. On the other hand if data examples are partial ranking of size $k \ll n$, i.e. $R, R' \subset \mathbb{S}_n$, and we take convolution kernel (4), the analysis of exploring the sparsity of the Fourier coefficients of the group algebra of partial rankings R, R' of size k reduces the evaluation of both the diffusion kernel and the Mallows kernel to $O((2k)^{2k+3})$ time, provided that the exponential kernel Fourier matrices $[\hat{\kappa}(\mu)]_{\geq [\dots]_{n-k}}$ are precomputed before any kernel evaluations take place (Kondor & Barbosa, 2010, Theorem 13). To avoid notation overflow, we simply point out that the complexity bound should be further refined if we additionally consider the sparsity of the Fourier coefficients $\hat{\kappa}_M(\mu)$ for the Mallows kernel. In fact, since $\kappa_M(\sigma)$ depends only on the destination of the ordered item pairs $\{(i,j)\}_{i<j}$ sent by permutation σ, the Fourier coefficient $\hat{\kappa}_M(\mu)$ is zero whenever $\mu \lhd (n-2, 1, 1)$ with respect to dominance order indexed by integer partition (Huang et al., 2009), regardless of k, which renders a huge interest in terms of computational issue.

6. Experimental Results

Datasets. We investigate the performance of classifying high-dimensional biomedical data, motivated by previous work demonstrating the relevance of replacing numerical features by pairwise comparisons in this context (Geman et al., 2004; Tan et al., 2005; Xu et al., 2005; Lin et al., 2009). For that purpose, we collected 10 datasets related to human cancer research publicly available online (Li et al., 2003; Schroeder et al., 2011; Shi et al., 2011), as summarized in Table 1. The features are proteomic spectra relative intensities for the *Ovarian Cancer* dataset and gene expression levels for all the others. The contrasting classes are typically "Non-relapse v.s. Relapse" in terms of cancer prognosis, or "Normal v.s. Tumor" in terms of cancer identification. The datasets have no missing values, except the *Breast Cancer 1* dataset for which we performed additional preprocessing to remove missing values as follows: first we removed two samples (both labeled "relapse") from the training set that have around 10% and 45% of missing gene values; next we discarded any gene whose value was missing in at least one sample, amounting to a total of 3.5% of all genes.

Methods. We compare the Kendall kernel to other standard kernels (linear, homogeneous 2nd-order polynomial and Gaussian RBF with bandwidth set with "median trick"), using SVM (with regularization parameter C) and Kernel Fisher Discriminant (KFD, without tuning parameter) as classifiers. In addition, we include in the benchmark classifiers based on Top Scoring Pairs (TSP) (Geman et al., 2004), namely (1-)TSP, k-TSP (Tan et al., 2005)[1] and APMV (all-pairs majority votes, i.e. $\binom{n}{2}$-TSP). Finally we also test SVM with various kernels using as input only top features selected by TSP (Shi et al., 2011).

In all experiments, each kernel is centered (on the training set) and scaled to unit norm in the feature space. For KFD-based models, we add 10^{-3} on the diagonal of the centered and scaled kernel matrix, as suggested by (Mika et al., 1999). The Kendall kernel we use in practice is a soft version to (7) in the sense that the extremes ± 1 can still be attained in the presence of ties, specifically we use

$$K_\tau(\mathbf{x}, \mathbf{x}') = \frac{n_c(\mathbf{x}, \mathbf{x}') - n_d(\mathbf{x}, \mathbf{x}')}{\sqrt{(n_0 - n_1)(n_0 - n_2)}},$$

where $n_0 = \binom{n}{2}$ and n_1, n_2 are the number of tied pairs in $\mathbf{x}, \mathbf{x}'$ respectively.

Except for three datasets that are split into training and test sets, in which case we report the performance on the test set, we perform a 5-fold cross-validation repeated 10 times and report the mean performance over the $5 \times 10 = 50$ splits to evaluate the performance of the different methods. In addition, on each training set, an internal 5-fold cross-validation is performed to tune parameters, namely the C parameter of SVM-based models optimized over a grid ranging from 10^{-2} to 10^3 in log scale, and the number k of TSP in case of feature selection (ranging from 1 to 5000 in log scale).

Results. Table 2 and Figure 2 (Left) summarize the performance of each model across the datasets. A SVM with the Kendall kernel achieves the highest average prediction accuracy overall (79.39%), followed by a linear SVM

[1] While the original k-TSP algorithm selects only top k *disjoint* pairs with the constraint that k is *less than* 10, we do not restrict ourselves to any of these two conditions since we consider k-TSP in this study essentially a feature pair scoring algorithm.

Table 1. Information of biomedial datasets.

Dataset	No. of features	No. of samples (training/test)		Reference
		C_1	C_2	
Breast Cancer 1	23624	44/7 (Non-relapse)	32/12 (Relapse)	(van 't Veer et al., 2002)
Breast Cancer 2	22283	142 (Non-relapse)	56 (Relapse)	(Desmedt et al., 2007)
Breast Cancer 3	22283	71 (Poor Prognosis)	138 (Good Prognosis)	(Wang et al., 2005)
Colon Tumor	2000	40 (Tumor)	22 (Normal)	(Alon et al., 1999)
Lung Adenocarcinoma 1	7129	24 (Poor Prognosis)	62 (Good Prognosis)	(Beer et al., 2002)
Lung Cancer 2	12533	16/134 (ADCA)	16/15 (MPM)	(Gordon et al., 2002)
Medulloblastoma	7129	39 (Failure)	21 (Survivor)	(Pomeroy et al., 2002)
Ovarian Cancer	15154	162 (Cancer)	91 (Normal)	(Petricoin et al., 2002)
Prostate Cancer 1	12600	50/9 (Normal)	52/25 (Tumor)	(Singh et al., 2002)
Prostate Cancer 2	12600	13 (Non-relapse)	8 (Relapse)	(Singh et al., 2002)

Table 2. Prediction accuracy (%) of different models across datasets.

	Average	BC1	BC2	BC3	CT	LA1	LC2	MB	OC	PC1	PC2
SVMkdtALL	79.39	78.95	71.31	67.34	85.78	70.98	97.99	63.67	99.48	100	58.4
SVMlinearTOP	77.16	84.21	69.29	67.11	84.19	63.92	97.32	65.17	99.41	85.29	55.7
SVMlinearALL	76.09	78.95	71.67	64.27	86.73	70.23	97.99	62.67	99.64	73.53	55.17
SVMkdtTOP	75.5	52.63	70.61	65.81	85.46	67.7	97.99	58.33	99.92	97.06	59.47
SVMpolyALL	74.54	68.42	71.62	63.66	78.43	70.53	98.66	61.17	99.28	79.41	54.23
KFDkdtALL	74.33	63.16	59.41	67.22	85.46	59.08	99.33	59.33	98.73	97.06	54.57
kTSP	74.03	57.89	58.22	64.47	87.23	61.7	97.99	56	99.92	100	56.83
SVMpolyTOP	73.99	63.16	69.44	66.26	79.14	65.98	99.33	60	99.21	88.24	49.1
KFDlinearALL	71.81	63.16	60.43	67.52	77.26	57.24	97.99	59.5	100	73.53	61.43
KFDpolyALL	71.39	63.16	60.48	67.38	75.1	58.52	97.99	60.33	100	73.53	57.43
TSP	69.71	68.42	49.58	57.8	85.61	58.96	95.97	52.67	99.8	76.47	51.83
SVMrbfALL	69.31	63.16	71.41	65.87	81.18	70.84	93.96	63.83	98.85	26.47	57.5
KFDrbfALL	66.39	63.16	60.48	66.03	83.71	58.73	97.32	59.67	98.46	26.47	49.87
APMV	61.91	84.21	65.98	33.96	64.49	33.6	89.93	42.17	85.19	73.53	46

trained on a subset of features selected from the top scoring pairs (77.16%) and a standard linear SVM (76.09%). The SVM with Kendall kernel outperforms all the other methods at a P-value of 0.07 according to a Wilcoxon rank test. We note that even though models based on KFD generally are less accurate than those based on SVM, the relative order of the different kernels is consistent between KFD and SVM, adding evidence that the Kendall kernel provides an interesting alternative to other kernels in this context. The performance of TSP and k-TSP, based on majority vote rules, are comparatively worse than those of SVM using the same features, as already observed by Shi et al. (2011).

Figure 2 further shows how the performance of different kernels depends on the choice of the C parameter or the SVM (Middle), and on the number of features used (Right), on some representative datasets. We observe that compared to other kernels, a SVM with the Kendall kernel is relatively insensitive to hyper-parameter C especially when C is large, which corresponds to a hard-margin SVM. This may explain in part the success of SVM in this setting, since the risk of choosing a bad C during training is reduced. Regarding the number of features used in case of feature selection, we notice that it does not seem to be beneficial to perform feature selection in this problem, explaining why the Kendall kernel which uses all pairwise comparisons between features outperforms other kernels restricted to a subset of these pairs.

Finally, as a proof of concept we empirically compare on one dataset the smooth alternative (10) and its Monte Carlo approximate (13) with the original Kendall kernel. Figure 3 shows how the performance varies with the amount of noise added to the samples (Left), and how the performance varies with the number of samples in the Monte Carlo scheme for a given amount of noise (Right). It confirms that the smooth alternative (10) can improve the performance of the Kendall kernel, and that the amount of noise (window size) should be considered as a parameter of the kernel to be optimized. Although the D^2-sample Monte Carlo approximate kernel (13) mainly serves as a fast estimate to the exact evaluation of (10), it shows that the idea of jittered input with specific noise can also bring a tempting benefit for data analysis with Kendall kernel, even when D is small. This also justifies the motivation of our proposed smooth alternative (10). Last but not least, despite the fact that the convergence rate of D^2-sample Monte Carlo approximate to the exact kernel evaluation is guaranteed by Theorem 3, experiments show that the convergence in practice is typically faster than the theoretical bound, and even faster in

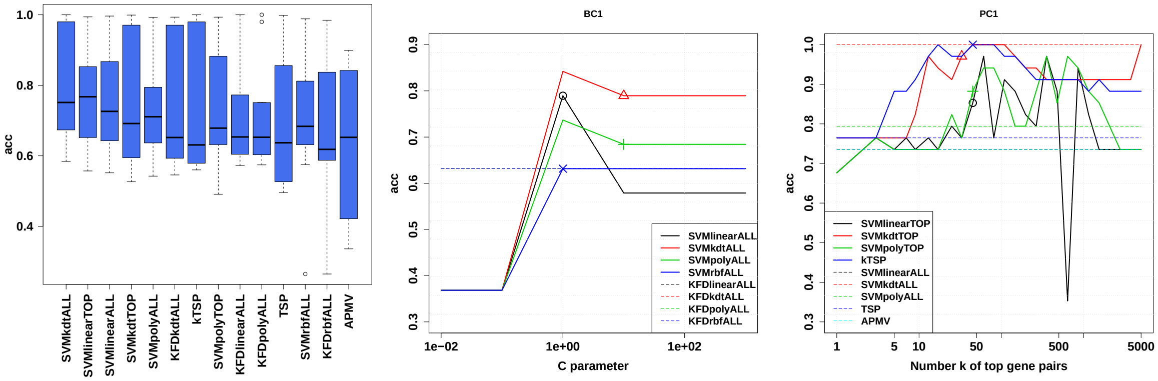

Figure 2. **Left**: Model performance comparison (ordered by decreasing average accuracy across datasets). **Middle**: Sensitivity of kernel SVMs to C parameter on the *Breast Cancer 1* dataset. **Right**: Impact of TSP feature selection on the *Prostate Cancer 1* dataset. (Special marks are returned by cross-validation.)

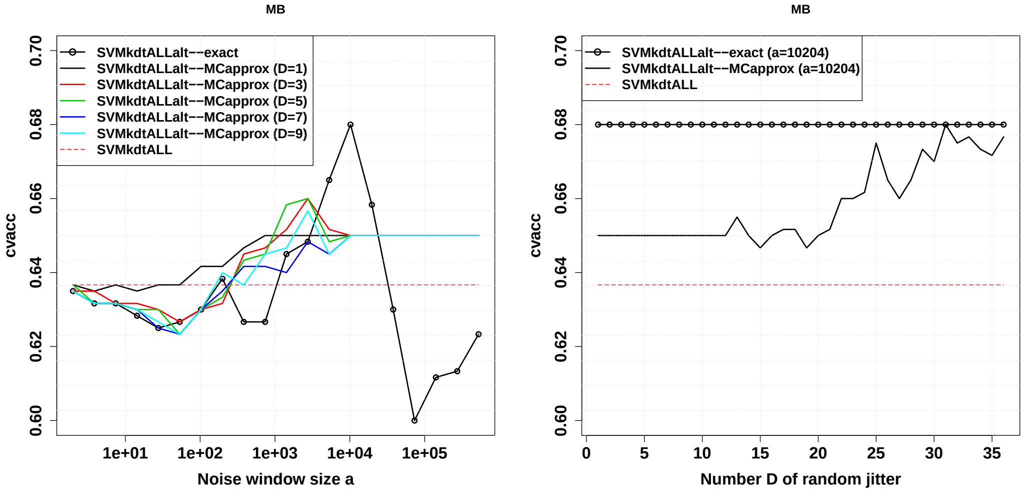

Figure 3. **Left**: Empirical performance of smoothed alternative to Kendall kernel on the *Medulloblastoma* dataset. **Right**: Empirical convergence of Monte Carlo approximate at the fixed window size attaining maximum underlying accuracy from the left plot.

case that the window size a is small. This is due to the fact that the convergence rate is also dependent of the observed data distribution in the input space, for which we have not made any specific assumption in our analysis.

7. Conclusion

Based on the observation that the classical Kendall tau correlation between total rankings is a positive definite kernel, we presented some extensions and applications pertaining to learning with the Kendall kernel and the related Mallows kernel. We showed that both kernels can be evaluated efficiently in $O(n \log n)$ time, and that the Kendall kernel can be extended to partial rankings containing k items out of n in $O(k \log k)$ time. When permutations are obtained by sorting real-valued vectors, we proposed an extension of the Kendall kernel based on random perturbations of the input vector to increase its robustness to small variations, and discussed two possible algorithms to compute it. We further highlighted a connection between the fast Mallow kernel and the diffusion kernel of Kondor & Barbosa (2010). We also reported promising experimental results on biomedical data demonstrating that for highly noisy data, the Kendall kernel is competitive or even outperforms other state-of-the-art kernels. We leave for future work further applications of kernel methods to permutations with these kernels, such as clustering of rankings with kernel k-means as an alternative to existing techniques based on mixtures of Mallows models.

Acknowledgments

This work was supported by the European Union 7th Framework Program through the Marie Curie ITN MLPM grant No 316861, and by the European Research Council grant ERC-SMAC-280032.

References

Ailon, N., Charikar, M., and Newman, A. Aggregating inconsistent information: ranking and clustering. *J. ACM*, 55(5):23:1–23:27, 2008.

Alon, U., Barkai, N., Notterman, D. A., et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U. S. A.*, 96(12): 6745–6750, 1999.

Balcan, M.-F., Bansal, N., Beygelzimer, A., et al. Robust reductions from ranking to classification. *Mach. Learn.*, 72(1–2):139–153, 2008.

Beer, D. G., Kardia, S. L. R., Huang, C.-C., et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, 8(8):816–824, Aug 2002.

Bottou, L. and Bousquet, O. The tradeoffs of large scale learning. In Platt, J.C., Koller, D., Singer, Y., and Roweis, S. (eds.), *Adv. Neural. Inform. Process Syst.*, volume 20, pp. 161–168. Curran Associates, Inc., 2008.

Cortes, C. and Vapnik, V. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995.

Critchlow, Douglas E. *Metric methods for analyzing partially ranked data*, volume 34 of *Lecture Notes in Statistics*. Springer New York, 1985.

Desmedt, C., Piette, F., Loi, S., et al. Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin. Cancer Res.*, 13(11):3207–3214, 2007.

Diaconis, P. *Group representations in probability and Statistics*, volume 11 of *Lecture Notes–Monograph Series*. Institut of Mathematical Statistics, Hayward, CA, 1988.

Fligner, M. A. and Verducci, J. S. Distance based ranking models. *J. R. Stat. Soc. Ser. B*, 48(3):359–369, 1986.

Fukumizu, K., Gretton, A., Schölkopf, B., and Sriperumbudur, B. K. Characteristic kernels on groups and semigroups. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds.), *Adv. Neural. Inform. Process Syst.*, volume 21, pp. 473–480. 2008.

Gärtner, T., Lloyd, J.W., and Flach, P.A. Kernels and distances for structured data. *Mach. Learn.*, 57(3):205–232, 2004.

Geman, D., d'Avignon, C., Naiman, D. Q., and Winslow, R. L. Classifying gene expression profiles from pairwise mRNA comparisons. *Stat. Appl. Genet. Mol. Biol.*, 3(1): Article19, 2004.

Gordon, G. J., Jensen, R. V., Hsiao, L.-L., et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res.*, 62(17):4963–4967, 2002.

Haussler, D. Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, UC Santa Cruz, 1999.

Helmbold, D. P. and Warmuth, M. K. Learning permutations with exponential weights. *J. Mach. Learn. Res.*, 10: 1705–1736, 2009.

Huang, J., Guestrin, C., and Guibas, L. Fourier theoretic probabilistic inference over permutations. *J. Mach. Learn. Res.*, 10:997–1070, 2009.

Kashima, H., Tsuda, K., and Inokuchi, A. Marginalized Kernels between Labeled Graphs. In Faucett, T. and Mishra, N. (eds.), *Proceedings of the Twentieth International Conference on Machine Learning*, pp. 321–328, New York, NY, USA, 2003. AAAI Press.

Kendall, M. G. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

Kendall, M. G. *Rank correlation methods*. Griffin, 1948.

Knight, W. R. A computer method for calculating Kendall's tau with ungrouped data. *J. Am. Stat. Assoc.*, 61(314):436–439, 1966.

Kondor, I. R. *Group theoretical methods in machine learning*. PhD thesis, Columbia University, 2008.

Kondor, I. R. and Lafferty, J. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the Nineteenth International Conference on Machine Learning*, volume 2, pp. 315–322, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

Kondor, R. I. and Barbosa, M. S. Ranking with kernels in fourier space. In Kalai, A. T. and Mohri, M. (eds.), *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pp. 451–463. Omnipress, 2010.

Lebanon, G. and Mao, Y. Non-parametric modeling of partially ranked data. *J. Mach. Learn. Res.*, 9:2401–2429, 2008.

Li, J., Liu, H., and Wong, L. Mean-entropy discretized features are effective for classifying high-dimensional biomedical data. In Zaki, M. J., Wang, J. T.-L., and Toivonen, H. (eds.), *Proceedings of the 3nd ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD 2003), August 27th, 2003, Washington, DC, USA*, pp. 17–24, 2003.

Lin, X., Afsari, B., Marchionni, L., et al. The ordering of expression among a few genes can provide simple cancer biomarkers and signal BRCA1 mutations. *BMC bioinformatics*, 10:256, 2009.

Mallows, C. L. Non-null ranking models. i. *Biometrika*, 44 (1/2):114–130, 1957.

Meilă, M., Phadnis, K., Patterson, A., and Bilmes, J. Consensus ranking under the exponential model. In *Proceedings of the Twenty-Third Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-07)*, pp. 285–294, Corvallis, Oregon, 2007. AUAI Press.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., and Müller, K.R. Fisher discriminant analysis with kernels. In Hu, Y.-H., Larsen, J., Wilson, E., and Douglas, S. (eds.), *Neural Networks for Signal Processing IX*, pp. 41–48. IEEE, 1999.

Muandet, K., Fukumizu, K., Dinuzzo, F., and Schölkopf, B. Learning from distributions via support measure machines. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Adv. Neural. Inform. Process Syst.*, volume 25, pp. 10–18. Curran Associates, Inc., 2012.

Petricoin, E. F., Ardekani, A. M., Hitt, B. A., et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359(9306):572–577, 2002.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, 2002.

Schoenberg, I. J. Metric spaces and positive definite functions. *Trans. Am. Math. Soc.*, 44(3):522–536, 1938.

Schölkopf, B. and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2002.

Schölkopf, B., Tsuda, K., and Vert, J.-P. *Kernel Methods in Computational Biology*. MIT Press, The MIT Press, Cambridge, Massachussetts, 2004.

Schroeder, M., Haibe-Kains, B., Culhane, A., et al. *breastCancerTRANSBIG: Gene expression dataset published by Desmedt et al. [2007] (TRANSBIG).*, 2011. URL http://compbio.dfci.harvard.edu/. R package version 1.2.0.

Shawe-Taylor, J. and Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004.

Shi, P., Ray, S., Zhu, Q., and Kon, M. A. Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction. *BMC Bioinformatics*, 12:375, 2011.

Singh, D., Febbo, P. G., Ross, K., et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209, 2002.

Steinwart, I. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Trans. Inform. Theory*, 51(1):128–142, 2005.

Tan, A. C., Naiman, D. Q., Xu, L., Winslow, R. L., and Geman, D. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21(20):3896–3904, 2005.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, 2002.

Vapnik, V. N. *Statistical Learning Theory*. Wiley, New-York, 1998.

Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. Graph kernels. *J. Mach. Learn. Res.*, 10:1–41, 2009.

Wang, Y., Makedon, F. S., Ford, J. C., and Pearlman, J. HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. *Bioinformatics*, 21(8):1530–1537, 2005.

Xu, L., Tan, A. C., Naiman, D. Q., Geman, D., and Winslow, R. L. Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics*, 21(20):3905–3911, 2005.