Electronic Journal of Statistics
Vol. 12 (2018) 2537–2577
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1437

On kernel methods for covariates that are rankings

Horia Mania, Aaditya Ramdas, Martin J. Wainwright,[*] Michael I. Jordan[†] and Benjamin Recht[‡]

Departments of Statistics and EECS
University of California, Berkeley, CA, 94720
e-mail: hmania@berkeley.edu; aramdas@berkeley.edu; wainwrig@berkeley.edu; jordan@berkeley.edu; brecht@berkeley.edu

Abstract: Permutation-valued features arise in a variety of applications, either in a direct way when preferences are elicited over a collection of items, or an indirect way when numerical ratings are converted to a ranking. To date, there has been relatively limited study of regression, classification, and testing problems based on permutation-valued features, as opposed to permutation-valued responses. This paper studies the use of reproducing kernel Hilbert space methods for learning from permutation-valued features. These methods embed the rankings into an implicitly defined function space, and allow for efficient estimation of regression and test functions in this richer space. We characterize both the feature spaces and spectral properties associated with two kernels for rankings, the Kendall and Mallows kernels. Using tools from representation theory, we explain the limited expressive power of the Kendall kernel by characterizing its degenerate spectrum, and in sharp contrast, we prove that the Mallows kernel is universal and characteristic. We also introduce families of polynomial kernels that interpolate between the Kendall (degree one) and Mallows (infinite degree) kernels. We show the practical effectiveness of our methods via applications to Eurobarometer survey data as well as a Movielens ratings dataset.

Keywords and phrases: Mallows kernel, Kendall kernel, polynomial kernel, representation theory, Fourier analysis, symmetric group.

Received September 2017.

1. Introduction

Ranked data arises naturally in any context in which preferences are expressed over a collection of alternatives. Familiar examples include election data, ratings of consumer items, or choice of schools. Preferences can be expressed directly via relative comparisons of alternatives, or indirectly via scores assigned to the different alternatives. Preferences are also often expressed implicitly; e.g., through

[*]Martin J. Wainwright was partially supported by Air Force Office of Scientific Research AFOSR-FA9550-14-1-0016, Office of Naval Research DOD ONR-N00014, and NSF grant CIF-31712-23800.

[†]Michael I. Jordan was partially supported by the Mathematical Data Science program of the Office of Naval Research.

[‡]Benjamin Recht was partially supported by ONR awards N00014-15-1-2620, N00014-13-1-0129, and N00014-14-1-0024 and NSF awards CCF-1148243 and CCF-1217058.

click activity on the web. In this paper, we consider datasets in which each covariate corresponds to a complete ranking over a set of d alternatives (that is, a permutation belonging to the symmetric group), and we study regression, classification and testing problems with such data.

As a running example to which we return in Section 6.2, we consider the Eurobarometer 55.2 survey conducted in several European countries in 2001, recently published by the European Opinion Research Group [1]. Each respondent was asked to indicate their preferences over sources of information about scientific developments; their options were: TV, radio, newspapers/magazines, scientific magazines, the internet, and school/university. Therefore, each observation in the survey contained a ranking of $d = 6$ objects, along with other covariates such as the participant's age, gender, etc; a snippet is shown in Table 1. Many natural questions arise from this dataset. Can we predict a person's age/gender from their ranking? Do men and women (or old and young) have the same distribution over sources of information? The primary goal of this paper is to develop and analyze some principled methods for answering such questions.

TABLE 1
Snippet of the Eurobarometer 55.2 *survey data.*

Respondent	Gender	Age	Ranking of news sources
1	F	32	TV > Radio > School/University > Newspapers/Mags. > Web > Sci. Mags.
2	F	84	TV > Radio > Newspapers/Mags. > School/University > Sci. Mags. > Web
3	F	65	TV > Newspapers/Mags. > Sci. Mags. > Radio > School/University > Web
4	M	29	Web > Radio > Newspapers/Mags. > TV > Sci. Mags. > School/University

There is a large existing literature on the use of rank statistics for testing and inference; see, for instance, the book by Lehmann and D'Abrera [2] and references therein. However, this body of work does not address problems in which the ranking themselves act as covariates. Thus, inferential problems in which the rankings are naturally viewed as covariates are generally simplified in various ways. For example, in the original report on the Eurobarometer survey data by the European Opinion Research Group [1], the authors measured only the frequency with which each of the six sources of information was ranked in the first or second position. Their analysis did not distinguish between respondents' first and the second preferences and disregarded the information encoded in their bottom four preferences. When covariates have been included, the analysis is generally strongly parametric; for example, Francis et al. [3] analyze the same dataset by extending the classical Bradley-Terry model (1952) to incorporate covariates such as sex and age.

Our focus in the current paper is on nonparametric models in which the covariates are rankings. We build on work of Jiao and Vert [5], who discuss the use of Mercer kernels for ranking data. Kernels on the symmetric group induce an inner-product structure on permutations by implicitly embedding them into a suitable Reproducing Kernel Hilbert Space (RKHS). This space is defined by a bivariate kernel function, and the representer theorem [6] allows problems of regression and testing to be reduced to the computation of the kernel values $k(\sigma, \pi)$ for pairs of permutations (σ, π). We view kernel-based methodology

as particularly appropriate for ranking problems: in particular, it allows us to transition from the cumbersome setting of the non-Abelian symmetric group of permutations to the familiar setting of Hilbert spaces. This methodology does not require us to make generative or probabilistic assumptions, and is practically viable as long as kernel evaluations are computationally efficient.

1.1. Kernels on the symmetric group

There is a rich theoretical understanding of kernels on Euclidean spaces, including the linear, polynomial, Matern, Laplace and Gaussian kernels [7]. The latter three are especially popular because they are *translation-invariant*—meaning that $k(x, y) = k(x + z, y + z)$ for all $x, y, z \in \mathbb{R}^d$—*characteristic*—meaning that the maximum mean discrepancy over the unit ball of the RKHS defines a metric on the space of distributions—and *universal*—meaning that the RKHS is dense in the space of square-integrable functions. The latter property ensures that any square-integrable decision boundary or regression function can be approximated arbitrarily well by a sequence of elements from the RKHS. While kernels on Euclidean spaces are well understood, there is a a growing interest in understanding properties of kernels on non-standard groups. For example, Fukumizu et al. [8] provide necessary and sufficient conditions for kernels on groups and semi-groups to be characteristic.

Permutations lie within the symmetric group, and various kernels for this non-Abelian group have been proposed [5, 9, 10]. Many of these kernels, including the Kendall and Mallows kernels considered in this paper, are *right-invariant*, meaning that they are invariant to a re-indexing of the underlying objects. This property is desirable for our applications: otherwise, the kernel similarity between a pair of permutations would depend on how the items were indexed. Much of the focus in the theoretical literature on kernel methods has focused on the *bi-invariant* class of kernels, which are both right- and left-invariant. A prominent example of a bi-invariant kernel is the diffusion kernel, which is quite well understood [see, e.g., 11]. Unfortunately, such kernels are not suitable for our applications, since for any bi-invariant kernel, the value between a pair of rankings that rank a specific item in positions one and two respectively would be the same as if they ranked it in positions, say, one and twenty. We thus focus on right-invariant kernels, such as the Kendall and Mallows kernels, and aim to bring the understanding of these kernels to the level of the bi-invariant kernels. In particular, we analyze the feature maps and spectral properties of the Kendall and Mallows kernels, as well as a new class of polynomial kernels.

There is also a mathematical literature on metrics on the symmetric group, including Cayley's metric, Ulam's metric, and Spearman's footrule. However, with the important exception of nearest-neighbor methods, most statistical analysis methods are more compatible with inner-product representations (kernels, similarities) than with metrics (distances, dissimilarities). For example, support vector machines, logistic regression, ridge regression, PCA, and many other methods can all be "kernelized."

1.2. Contributions

After a presentation of basic background on kernel methods on the symmetric group in Section 2, we begin our development by presenting an analysis of the Kendall and Mallows kernels from a primal point of view in Section 3. In particular, in Proposition 1, we prove that the Gram matrix associated with the Kendall kernel always has rank $\binom{d}{2}$, and we discuss the statistical implications of this result. Then, in Proposition 2, we present a novel finite-dimensional feature map for the Mallows kernel. At first glance, this result may seem surprising because the Mallows kernel appears analogous to the Gaussian kernel on the reals, which does *not* have a finite-dimensional feature map. This surprise is alleviated, however, by noting that all kernels on finite domains must admit some finite-dimensional feature map. We will show, moreover, that our feature maps are not just finite dimensional, but also interpretable and easy to describe.

In Euclidean spaces, there exists a large body of work on the spectral properties of kernels, meaning the decay rates of their eigenvalues. This decay informs the statistical analysis of kernel methods, providing leverage on the ability of kernels to discriminate between distributions, or estimate decision boundaries and regression functions. Motivated by this past work, in Section 4 we study the spectra of the Kendall and Mallows kernels, proceeding via a non-Abelian variant of Bochner's theorem [8, 11]. This analysis requires a foray into representation theory [12, 13]. We provide as much background on representation theory as is necessary to understand our theorem statements, leaving the proofs and fuller development of representation theory for the Appendix. Theorem 3 fully characterizes the Fourier spectrum of the Kendall kernel. In particular, we show that it has only two nonzero irreducible representations, both of which turn out to be rank-one matrices; this degeneracy suggests the kernel is useful only for a limited range of problems. Theorem 5 provides a first-principles proof of the fact that the Mallows kernel is universal and characteristic; i.e, every irreducible representation is a strictly positive-definite matrix.

In Section 5, we propose and analyze natural families of polynomial kernels of degree p that interpolate between the Kendall and Mallows kernels (corresponding to $p = 1$ and $p = \infty$ respectively). We study their (primal) feature maps and (dual) spectra and, in Theorem 6, we prove that $p = d - 1$ suffices for the kernel to be universal and characteristic.

In addition to these theoretical insights, we also present the results of various experiments with our kernel representations. In our first set of experiments, we apply kernel methods to a simulated data set in order to illustrate our predicted differences in the empirical power of two-sample hypothesis tests using different kernels for rankings, and discuss on which instances we expect the Kendall or Mallows kernels to have higher power. We then apply these kernel-based tests to the Eurobarometer survey data, and we also fit kernel SVM and kernel regression models to this data in order to showcase the usefulness of kernel methods to leverage ranking data. Our two-sample tests find that men and women do have significantly different preferences—the classifiers have a test error of 34% for predicting if the respondent was old or young. Moreover, the regression from

rankings to age has a test prediction error of about 11 years. We also studied a data set consisting of ratings for movies in which we transformed the users' ratings across movie genres into rankings. We find significant evidence for males and females having different preferences over movie genres, a simple illustration of the possible utility of converting absolute ratings into relative rankings.

2. An overview of kernel methods for rankings

In order to understand the use of kernels for permutation-valued features, we first need to introduce some standard terminology.

Symmetric group $\mathbb{S}_d$. There is a natural one-to-one correspondence between permutations and rankings. Let the set $[d] := \{1, 2, \ldots, d\}$ represent labels of a collection of d objects. Any permutation $\sigma : [d] \to [d]$ defines a ranking, in which $\sigma(i)$ is the rank of object i. The set of all permutations forms a group with the standard function composition $\sigma \circ \sigma'$; that is, $\pi = \sigma \circ \sigma' \iff \pi(i) = \sigma(\sigma'(i))$. This group is known as the symmetric group and it is denoted by $\mathbb{S}_d$.

Universal RKHS. A kernel is a bivariate function, $k \colon \mathbb{S}_d \times \mathbb{S}_d \to \mathbb{R}$, such that for any collection of rankings the associated Gram matrix is positive semidefinite. We let $\mathcal{F}_k$ denote the reproducing kernel Hilbert space (RKHS) induced by the kernel k; $\mathcal{F}_k$ is a set of functions defined by the closure of the span of $\{k(\sigma, \cdot)\}_{\sigma \in \mathbb{S}_d}$. The inner product between two functions $f = \sum_{j=1}^{d!} a_j k(\sigma_j, \cdot)$ and $g = \sum_{j=1}^{d!} b_j k(\sigma_j, \cdot)$ is defined to be $\langle f, g \rangle_{\mathcal{F}_k} = \sum_k \sum_j a_k b_j k(\sigma_k, \sigma_j)$. This inner product induces the RKHS norm $\|f\|_{\mathcal{F}_k} = \sqrt{\langle f, f \rangle_{\mathcal{F}_k}}$. If k is a kernel on a space $\mathcal{X}$ (say $\mathbb{S}_d$) and ℓ is a kernel on $\mathcal{Y}$ (say $\mathbb{R}^p$), then $m := k \times \ell$ is a kernel on the space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$; that is, for $z = (x, y), z' = (x', y')$, we have $m(z, z') = k(x, x')\ell(y, y')$. Naturally, we can recurse this process to define kernels on domains involving a variety of data types, showcasing their generality. For compact metric spaces, a continuous kernel k is called *universal* if the RKHS $\mathcal{F}_k$ defined by it is dense, in L_∞ norm, in the space of continuous functions [14]. In our setting, a kernel k is universal if and only if any real-valued function f on $\mathbb{S}_d$ can be written as a linear combination of functions $k(\pi, \cdot)$, with $\pi \in \mathbb{S}_d$; that is, $\mathcal{F}_k$ contains all possible functions.

Feature maps. Mercer's theorem [7, Proposition 2.11], when applied to a finite domain such as $\mathbb{S}_d$, guarantees that any kernel $k \colon \mathbb{S}_d \times \mathbb{S}_d \to \mathbb{R}$ admits a finite-dimensional feature map $\Phi \colon \mathcal{X} \to \mathbb{R}^m$ such that

$$k(\sigma, \sigma') = \langle \Phi(\sigma), \Phi(\sigma') \rangle_{\mathbb{R}^m}, \quad \text{for all } \sigma, \sigma' \in \mathbb{S}_d.$$

For example, such a feature map can be defined by the columns of the square root of the $d! \times d!$ kernel matrix. In light of this characterization, kernels correspond to inner products in appropriate feature spaces, and can be thought of as a measure of similarity between rankings. Feature maps are not unique, and many

different feature maps may give rise to the same kernel. Nonetheless, the feature maps considered in this work are easy to define in closed form and offer valuable insights into the properties of the Kendall and Mallows kernels.

Right-invariance. A function $F\colon \mathbb{S}_d \times \mathbb{S}_d \to \mathbb{R}$ is called *right-invariant* if $F(\sigma, \sigma') = F(\sigma \circ \pi, \sigma' \circ \pi)$ for all permutations $\sigma, \sigma', \pi \in \mathbb{S}_d$. By setting $\pi = \sigma^{-1}$, we see that this property holds if and only if $F(\sigma, \sigma') = f(\sigma' \circ \sigma^{-1})$ for some function $f : \mathbb{S}_d \to \mathbb{R}$. For kernels, we overload notation by using k to refer to both F and f by k; usage will be clear from the context. Right-invariance of kernels is desirable for applications involving rankings since it ensures that the kernel values remain unchanged by a relabeling of the objects being ranked. Furthermore, as we discuss later, right-invariance enables us to use Fourier analysis to study the kernels.

The Kendall and Mallows kernels. All the kernels that we study in this paper measure the similarity between two rankings through the number of pairs of objects that they order in the same way or in opposite ways. More precisely, letting $n_d(\sigma, \sigma')$ and $n_c(\sigma, \sigma')$ denote (respectively) the number of *discordant* and *concordant* pairs between permutations σ and σ', we have the relations

$$n_d(\sigma, \sigma') := \sum_{i<j} \left[\mathbb{1}_{\{\sigma(i)<\sigma(j)\}} \mathbb{1}_{\{\sigma'(i)>\sigma'(j)\}} + \mathbb{1}_{\{\sigma(i)>\sigma(j)\}} \mathbb{1}_{\{\sigma'(i)<\sigma'(j)\}} \right], \quad \text{and} \tag{1a}$$

$$n_c(\sigma, \sigma') = \binom{d}{2} - n_d(\sigma, \sigma'), \tag{1b}$$

where equality (1b) follows because any pair of indices is either concordant or discordant. Of particular interest are the *Kendall kernel*, denoted by k_τ, and the *Mallows kernel*, denoted by k_m^ν, where ν is a user-chosen bandwidth parameter. They each depend only on the number of discordant/concordant pairs, and are defined by

$$k_\tau(\sigma, \sigma') := \frac{n_c(\sigma, \sigma') - n_d(\sigma, \sigma')}{\binom{d}{2}}, \quad \text{and} \tag{2a}$$

$$k_m^\nu(\sigma, \sigma') := \exp\left(-\nu \cdot n_d(\sigma, \sigma')\right). \tag{2b}$$

Jiao and Vert [5] show that k_τ and k_m^ν are indeed positive semidefinite and that they can be computed in $\mathcal{O}(d \log d)$ time. Furthermore, it is not hard to check that the number of discordant pairs between two permutations is right-invariant, and in fact $n_d(\sigma, \sigma') = \mathrm{i}(\sigma' \circ \sigma^{-1})$, where $\mathrm{i}(\pi)$ denotes the number of inversions of the permutation π. See Appendix C.2 for a short proof of this fact.

Therefore, kernels that depend only on the number of discordant pairs are right-invariant, which is one of the reasons behind our particular interest in the Kendall and Mallows kernels. Another reason is that the Kendall kernel corresponds almost directly to the classical Kendall-τ metric on $\mathbb{S}_d$, and the Mallows kernel is reminiscent of the popular Mallows distribution over $\mathbb{S}_d$. Later, we introduce a family of polynomial kernels that interpolate between these two kernels. While these are not the only kernels of interest, they are natural starting points.

Kernel regression on $\mathbb{S}_d$. Consider the problem of *kernel ridge regression* [6], where we fit a linear model in the feature space $\mathcal{H}$, which generally induces a nonlinear model in the original space. Given a set of n observations $\{(\pi_i, y_i)\}_{i=1}^n$, kernel ridge regression fits a function $f\colon \mathbb{S}_d \to \mathbb{R}$ to the data by solving the optimization problem

$$f^* \;:=\; \arg\min_{f\in\mathcal{F}_k} \left\{ \sum_{i=1}^n (y_i - f(\pi_i))^2 + \lambda \|f\|_{\mathcal{F}_k}^2 \right\}, \tag{3}$$

where $\lambda \in \mathbb{R}^+$ is a regularization parameter. If k is universal, then the estimate f^* can approximate any function $f : \mathbb{S}_d \to \mathbb{R}$ arbitrarily well. Conversely, if $\mathcal{F}_k$ is not universal, then we may suffer from an approximation error even in the limit of infinite data.

Representer theorem. Note that kernel ridge regression is never directly performed as written above—indeed, the representer theorem [6] implies that f^* lies in the span of $\{k(\pi_i, \cdot)\}_{i=1}^n$, meaning that the optimum can be written as $f^* = \sum_{i=1}^n w_i^* k(\sigma_i, \cdot)$ for some vector w^*. This allows us to rewrite the optimization problem (3) as

$$w^* := \arg\min_{w\in\mathbb{R}^n} \left\{ \|y - M_k w\|_2^2 + \lambda \|w\|_2^2 \right\},$$

where M_k is the $n \times n$ Gram matrix whose entries are $M_{k,ij} = k(\nu_i, \nu_j)$. This quadratic program has the explicit solution $w^* = (M_k + \lambda I_n)^{-1} y$.

Characteristic kernels. Any kernel k on a domain $\mathcal{X}$ induces a pseudo-metric on the set of probability distributions on $\mathcal{X}$, known as the *maximum mean discrepancy* [15, 16, 17], which in our setting of $\mathcal{X} = \mathbb{S}_d$ is given by

$$\mathrm{MMD}_k(P, Q) = \sup_{\|f\|_{\mathcal{F}_k} \le 1} \left[\mathbb{E}_{\sigma\sim P}[f(\sigma)] - \mathbb{E}_{\pi\sim Q}[f(\pi)] \right]. \tag{4}$$

Given a feature map $\Phi\colon \mathbb{S}_d \to \mathcal{H}$ induced by k, we define a mean embedding of P by $\mu_{k,P} = \mathbb{E}_{\sigma\sim P}[\Phi(\sigma)]$. Elementary computations [17] show that

$$\mathrm{MMD}_k(P, Q) = \|\mu_{k,P} - \mu_{k,Q}\|_{\mathcal{H}}. \tag{5}$$

The kernel is said to be *characteristic* if MMD_k actually defines a metric on the set of probability distributions—that is, if $\mathrm{MMD}_k(P, Q) = 0$ if and only if $P = Q$.

Two-sample testing on $\mathbb{S}_d$. Let P and Q be probability distributions over $\mathbb{S}_d$, and consider testing the null hypothesis $H_0 : P = Q$ against the alternative $H_1 : P \neq Q$, using samples $\{\alpha_i\}_{i=1}^{n_1} \overset{i.i.d.}{\sim} P$ and $\{\beta_1\}_{i=1}^{n_2} \overset{i.i.d.}{\sim} Q$. One approach to this testing problem is to estimate a pseudo-metric between P and Q, and reject H_0 if the estimate is large. For example, Gretton et al. [17] define the statistic

$$T_k(\alpha, \beta) = \frac{1}{n_1(n_1-1)} \sum_{i\neq j} k(\alpha_i, \alpha_j) + \frac{1}{n_2(n_2-1)} \sum_{i\neq j} k(\beta_i, \beta_j) - \frac{2}{n^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} k(\alpha_i, \beta_j),$$

which is an unbiased estimator of MMD_k^2, and consider the test $T_k(\alpha, \beta) > t^*$ for some threshold t^* (which can be determined, for example, by bootstrapping or permutation testing). We use this nonparametric framework for two-sample testing as a jumping-off point in our investigation of the statistical properties of kernels on permutations. Specifically, we investigate the interpretability of this class of tests. Given a kernel k, to what kind of differences between P and Q is the test sensitive? If the null hypothesis is *not* rejected, does that mean the two probability distributions are equal or that they simply have the same low-order moments (for some appropriate notion of moment)?

One may understand the test and the pseudo-metric MMD_k by studying the kernel k. For example, since MMD_k is not always a metric, this test would have trivial power against alternatives $P \neq Q$ whenever $\mathrm{MMD}_k(P, Q) = 0$. Hence, it is useful to understand when the MMD_k could be equal to zero, even though $P \neq Q$. The results presented in the next section offer answers to these questions. For example, Proposition 1 shows that the MMD induced by the Kendall kernel is *not* a metric, and in fact it is far from being a metric. In sharp contrast, Theorem 5 guarantees that MMD induced by the Mallows kernel is a metric; i.e., $\mathrm{MMD}_{k_m^\nu}(P, Q) = 0$ only when $P = Q$.

3. Feature spaces of the Kendall and Mallows kernels

In this paper, we make extensive use of sets $\{i, j\}$, and as a matter of convention, we always write any such set with the smaller element appearing first and the larger one appearing second, meaning that implicit in the notation $\{i, j\}$ is the fact that $i < j$. For example, the objects 1 and 2 always appear as $\{1, 2\}$ and never $\{2, 1\}$.

Jiao and Vert [5] constructed a feature map $\Phi_\tau \colon \mathbb{S}_d \to \mathbb{R}^{\binom{d}{2}}$ for the Kendall kernel defined by

$$\Phi_\tau(\sigma)_{\{i,j\}} = \sqrt{\binom{d}{2}^{-1}} \left(2\mathbb{1}(\sigma(i) < \sigma(j)) - 1\right), \tag{6}$$

which is easily seen to satisfy $k_\tau(\sigma, \sigma') = \Phi_\tau(\sigma)^\top \Phi_\tau(\sigma')$. Using this map we can give an interpretation of the MMD operator of Eq. (5). We first fix an ordering $\sigma_1, \sigma_2, \ldots, \sigma_{d!}$ of the elements of $\mathbb{S}_d$. We also fix an ordering $t_1, t_2, \ldots, t_{\binom{d}{2}}$ of the sets $\{a, b\}$ with $a, b \in [d]$, denoting this set by $\mathcal{T}^*$. We use M_τ to denote the $\mathbb{R}^{\binom{d}{2} \times d!}$ matrix whose columns are indexed by the rankings σ_j, whose rows are indexed by the tuples t_i, and whose j-th column is the vector $\Phi_\tau(\sigma_j)$. With this notation, if we view the distributions P and Q as probability vectors in $[0, 1]^{d!}$, the MMD in Eq. (5) is equal to $\|M_\tau(P - Q)\|_2$.

We also define the matrix $A_\tau \in \{0, 1\}^{\binom{d}{2} \times d!}$ with columns and rows indexed similarly, and entries

$$(A_\tau)_{\{a,b\},\sigma} = \begin{cases} 1 & \text{if } \sigma(a) < \sigma(b) \\ 0 & \text{if } \sigma(a) > \sigma(b). \end{cases}$$

With this notation in place, we can now state the following result.

Proposition 1. *The maximum mean discrepancy* MMD_{k_τ} *between two probability distributions* P, Q *on* $\mathbb{S}_d$ *is zero if and only if* $A_\tau(P-Q)=0$. *Moreover, the matrix* A_τ *has rank* $\binom{d}{2}$.

Straightforward algebra shows that $\frac{1}{2}\sqrt{\binom{d}{2}}M_\tau(P-Q)=A_\tau(P-Q)$, proving the first part of the proposition; the second part is proved in Appendix A.1. We remark that $(A_\tau P)_{\{a,b\}}=P(\sigma(a)<\sigma(b))$, and hence the MMD_{k_τ} corresponds to the Euclidean distance between the vectors of probabilities of the events $\{\sigma:\sigma(a)<\sigma(b)\}$ under the distributions P and Q. As a parallel to the linear kernel in $\mathbb{R}^m$, the Kendall kernel detects a difference between two probability distributions only if they differ in mean, where we define the mean as the vectors of probabilities of events $\{\sigma\colon\sigma(a)<\sigma(b)\}$.

How many probability distributions have the same mean embedding as P under the Kendall kernel? A probability distribution Q over $\mathbb{S}_d$ is a vector in $\mathbb{R}^{d!}$ that is contained in the unit simplex, a subset of a hyperplane of dimension $d!-1$. Proposition 1 shows that for each P in the interior of the unit simplex of $\mathbb{R}^{d!}$ there is a subspace $V\subset\mathbb{R}^{d!}$ of dimension $d!-\binom{d}{2}-1$ such that for each $\gamma\in V$ there exists $\epsilon>0$ such that $P+\epsilon\gamma$ is a probability distribution over $\mathbb{S}_d$ and $\mathrm{MMD}_{k_\tau}(P+\epsilon\gamma,P)=0$. In other words, as d increases, the fraction of the directions that the Kendall kernel cannot distinguish goes to one. This observation shows that the Kendall kernel is far from being a metric on the probability simplex in $\mathbb{R}^{d!}$. We offer a Fourier transform perspective on this fact in Theorem 3, showing in particular that the Kendall kernel can detect only low-frequency differences between two probability distributions.

We next describe a finite-dimensional feature map for the Mallows kernel.

Proposition 2. *Let* $\mathcal{P}(\mathcal{T}^*)$ *denote the power set of* $\mathcal{T}^*$. *Then, a feature map of the Mallows kernel* k_m^ν *is given by a map* $\Phi_m\colon\mathbb{S}_d\to\mathcal{P}(\mathcal{T}^*)$. *In particular, if* $s_1,s_2,\ldots,s_r$ *are distinct elements of* $\mathcal{T}^*$, *we set*

$$\Phi_m(\sigma)_\emptyset = \left(\frac{1+\exp(-\nu)}{2}\right)^{\frac{1}{2}\binom{d}{2}} \quad \text{for all } \sigma\in\mathbb{S}_d, \text{ and} \tag{7a}$$

$$\Phi_m(\sigma)_{s_1s_2\ldots s_r} = \left(\frac{1+\exp(-\nu)}{2}\right)^{\frac{1}{2}\binom{d}{2}}\left(\frac{1-\exp(-\nu)}{1+\exp(-\nu)}\right)^{\frac{r}{2}}\prod_{i=1}^{r}\overline{\Phi}(\sigma)_{s_i}, \tag{7b}$$

where $\overline{\Phi}(\sigma)_{s_i}=2\mathbb{1}_{\{\sigma(a_i)<\sigma(b_i)\}}-1$ *when* $s_i=\{a_i,b_i\}$.

Since the Kendall feature map Φ_τ has components $2\mathbb{1}(\sigma(a)<\sigma(b))-1$, the components of the mean embedding $\mathbb{E}_{\sigma\sim P}\Phi_\tau(\sigma)$ capture the probabilities of pairwise comparisons: $2P(\sigma(a)<\sigma(b))-1$. However, Eq. (7) shows that the features of the Mallows kernel are of the form $\prod_{i=1}^{r}(2\mathbb{1}(\sigma(a_i)<\sigma(b_i))-1)$, up to a scalar factor. By expanding the product, we see that the expected value of the feature map (7) has components equal to

$$\mathbb{E}_{\sigma\sim P}\Phi_m(\sigma)_{s_1 s_2 \ldots s_r} = C\sum_{j=0}^{r}(-1)^j 2^{r-j} \sum_{\substack{A\subset\{s_1,\ldots,s_r\}\\ |A|=r-j}} P\left(\sigma(a)<\sigma(b),\, \forall\{a,b\}\in A\right),$$

where the scale factor C depends only on dimension d, bandwidth ν, and r. Therefore, the mean embedding with respect to the Mallows kernel captures the probabilities that different rankings of objects, expressed as concurrent pairwise orderings, occur. Since any ranking of d objects can be fully characterized by $d-1$ pairwise comparisons, the above feature map suggests that the Mallows kernel is characteristic, a fact that is established later in Theorem 5.

4. Fourier analysis of the Kendall and Mallows kernels

We start by defining basic concepts that allow us to state our results about the Fourier transforms of the Kendall and Mallows kernels. Elementary treatments of Fourier analysis on groups are provided by Kondor [11] and Huang et al. [18], with a concise summary given by Kondor and Barbosa [10]. In addition to the basic concepts discussed in those treatments, our proofs will also require some more advanced machinery, as found, for example, in Diaconis [12], Sagan [19], or Fulton and Harris [13]. We introduce these more advanced concepts in Appendix B.

The Fourier transform of a function $f\colon \mathbb{S}_d \to \mathbb{C}$ takes the form

$$\widehat{f}(\rho_\lambda) := \sum_{\sigma\in\mathbb{S}_d} f(\sigma)\rho_\lambda(\sigma), \tag{8}$$

where ρ_λ is a matrix-valued function to be defined shortly. As a contrast with the Fourier transform for functions defined over $\mathbb{R}$, instead of being indexed by a frequency ξ, the Fourier transform is indexed by λ, which is a *partition* of d—a non-increasing sequence of integers that sum to d. Furthermore, instead of the standard exponential basis functions $\exp(i\xi x)$, the terms ρ_λ are functions from $\mathbb{S}_d$ to $\mathbb{C}^{d_\lambda\times d_\lambda}$ and are called representations.

More precisely, a *representation* of the symmetric group is a matrix-valued function $\rho\colon \mathbb{S}_d \to \mathbb{C}^{d_\rho\times d_\rho}$ such that $\rho(\sigma)$ is invertible and $\rho(\sigma\circ\sigma') = \rho(\sigma)\rho(\sigma')$ for all permutations $\sigma,\sigma'\in\mathbb{S}_d$. The integer d_ρ is called the dimension of the representation. As an immediate consequence of the definitions, it follows that

$$\rho(e) = I_{d_\rho} \quad \text{and} \quad \rho(\sigma)^{-1} = \rho(\sigma^{-1}) \qquad \text{for all } \sigma\in\mathbb{S}_d.$$

A representation ρ is *reducible* if it is equivalent to the direct sum of two representations. Explicitly, a representation ρ is reducible if there exist two representations ρ_1 and ρ_2 and an invertible matrix $C\in\mathbb{C}^{d_\rho\times d_\rho}$ such that

$$\rho(\sigma) = C^{-1}\left[\rho_1(\sigma)\oplus\rho_2(\sigma)\right]C = C^{-1}\begin{pmatrix}\rho_1(\sigma) & \mathbf{0}\\ \mathbf{0} & \rho_2(\sigma)\end{pmatrix}C \quad \text{for all } \sigma\in\mathbb{S}_d.$$

A representation that is not reducible is called *irreducible*. For brevity, we refer to irreducible representations as *irreps*. The symmetric group has a finite number

of distinct irreps (an explanation of the meaning of "distinct" is provided in Appendix B), which have a standard indexing by finite sequences of positive integers $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_r)$ such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$ and $\sum_{i=1}^{r} \lambda_i = d$. Such sequences are called *partitions* of d and $\lambda \vdash d$ means that λ is a partition of d.

Returning to Eq. (8) and using the terminology just introduced, the Fourier transform of a function on the symmetric group can be described as a mapping from the irreps ρ_λ to matrices in $\mathbb{C}^{d_\lambda \times d_\lambda}$. This version of the Fourier transform shares many similar properties with its counterpart over real numbers, including the Fourier inversion formula and the Plancherel formula. In particular, we note that in this context Bochner's theorem states that a right-invariant kernel $k\colon \mathbb{S}_d \times \mathbb{S}_d \to \mathbb{C}$ is positive definite if and only if the matrix $\widehat{k}(\rho_\lambda)$ is positive semi-definite for all partitions $\lambda \vdash d$ [8, 11]. For a precise statement of Bochner's theorem and related results, we refer the reader to Appendix B.

We introduce some notation for the standard partial ordering of the partitions of d. Given any two partitions $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_r)$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_l)$, we say that $\lambda \trianglerighteq \mu$ if $\sum_{i=1}^{j} \lambda_i \geq \sum_{i=1}^{j} \mu_i$ for all $j \leq \min\{l, r\}$. We say $\lambda \triangleleft \mu$ whenever it is not true that $\lambda \trianglerighteq \mu$. The irreps of the symmetric group inherit the same partial ordering.

Equipped with this background, we now turn to the statements of our results on the spectral properties of the Kendall and Mallows kernels, as well as a discussion of some of their consequences. We begin with a theorem that characterizes the spectrum of the Kendall kernel.

Theorem 3. *The Kendall kernel has the following properties:*

(a) When $d = 2$, the Fourier transform of the Kendall kernel is equal to 0 at $\rho_{(2)}$ and equal to 2 at $\rho_{(1,1)}$.

(b) When $d \geq 3$, the Fourier transform $\widehat{k_\tau}$ of the Kendall kernel is zero at all irreducible representations except for $\rho_{(d-1,1)}$ and $\rho_{(d-2,1,1)}$. Furthermore, at both of the latter two representations, the Fourier transform $\widehat{k_\tau}$ has rank one.

While Proposition 1 implies that the Fourier spectrum of the Kendall kernel does not have full support, Theorem 3 provides a more refined characterization of the spectrum. Understanding the decay of eigenvalues of kernel operators has been a key step in characterizing the statistical rates of kernel-based estimators in other settings [20]. Theorem 3 is thus a first step towards understanding the statistical properties of methods that use the Kendall kernel. In particular, the MMD for a kernel k defined on the symmetric group can be expressed in the Fourier domain as

$$\mathrm{MMD}_k^2(P, Q) = \frac{1}{d!} \sum_{\lambda \vdash d} d_\lambda \operatorname{tr}\left[\left(\widehat{P}(\rho_\lambda) - \widehat{Q}(\rho_\lambda) \right)^\top \widehat{k}(\rho_\lambda) \left(\widehat{P}(\rho_\lambda) - \widehat{Q}(\rho_\lambda) \right) \right]. \quad (9)$$

This result follows from the Fourier inversion formula, and we prove it in Appendix C.3 for completeness. From Eq. (9), it immediately follows that a kernel k

is characteristic if and only if $\widehat{k}(\rho_\lambda)$ is positive definite for all $\lambda \vdash d$. The following corollary follows immediately from Theorem 3 and Eq. (9), and it characterizes the discriminative properties of the Kendall kernel in the frequency domain.

Corollary 4. *When $d \geq 3$, the MMD pseudo-metric for the Kendall kernel is given by*

$$\mathrm{MMD}_\tau(P,Q)^2 = \frac{1}{d!} \sum_{\lambda \in \left\{ \substack{(d-1,1),\\(d-2,1,1)} \right\}} d_\lambda \operatorname{tr}\left[\left(\widehat{P}(\rho_\lambda) - \widehat{Q}(\rho_\lambda) \right)^\top \widehat{k_\tau}(\rho_\lambda) \left(\widehat{P}(\rho_\lambda) - \widehat{Q}(\rho_\lambda) \right) \right]. \tag{10}$$

This result follows by combining the Fourier-analytic characterization of Theorem 3 with expression (9). Corollary 10 shows that most differences between $\widehat{P}$ and $\widehat{Q}$ do not contribute to $\mathrm{MMD}_k(P,Q)$. The only differences that contribute to MMD_k are the $(d-1) \times (d-1)$ matrix $\widehat{P}(\rho_{(d-1,1)}) - \widehat{Q}(\rho_{(d-1,1)})$ and the $\binom{d-1}{2} \times \binom{d-1}{2}$ matrix $\widehat{P}(\rho_{(d-2,1,1)}) - \widehat{Q}(\rho_{(d-2,1,1)})$. To be more precise, the Kendall kernel can differentiate between P and Q if and only if their Fourier transforms at $\rho_{(d-1,1)}$ or $\rho_{(d-2,1,1)}$ differ along a single direction aligning with the only eigenvector with a non-zero eigenvalue of $\widehat{k_\tau}(\rho_{(d-1,1)})$ or $\widehat{k_\tau}(\rho_{(d-2,1,1)})$.

We now turn to Fourier analysis of the Mallows kernel (2b). Despite its superficial similarity to the Kendall kernel, it has very different properties.

Theorem 5. *The Fourier transform $\widehat{k_m^\nu}$ of the Mallows kernel is strictly positive definite at all irreducible representations ρ_λ.*

Note that Theorem 5 corrects an assertion in the paper of Jiao and Vert [5]; the authors of that work suggested that since the Mallows kernel depends only on the relative rankings of pairs of objects, the Fourier transform $\widehat{k_m^\nu}$ should be expected to be zero at all irreps $\lambda \lhd (d-2,1,1)$. Theorem 5 shows that this natural intuition does not actually hold.

Theorem 5 also has implications for the universality of the Mallows kernel. Gretton et al. [17] show that a universal and continuous kernel on a compact metric space is characteristic—hence, a kernel on $\mathbb{S}_d$ is universal if and only if it is characteristic. As with Theorem 3, Theorem 5 has implications for the kernel MMD induced by the Mallows kernel. In particular, it shows that the Mallows kernel is both characteristic and universal, and hence $\mathrm{MMD}_{k_m^\nu}$ is a metric on probability distributions over $\mathbb{S}_d$.

5. A family of polynomial-type kernels

Based on our results thus far, it is natural to suspect that there exists a family of kernels interpolating between the relative simplicity of the Kendall kernel, which is analogous to a linear kernel on $\mathbb{R}^d$, and the richness of the Mallows kernel, which is analogous to a Gaussian kernel on $\mathbb{R}^d$. This intuition motivates us to introduce three families of polynomial-type kernels on the symmetric group,

defined as follows:

$$k^p(\sigma,\sigma') := (1 + k_\tau(\sigma,\sigma'))^p \tag{11a}$$

$$\overline{k}^p(\sigma,\sigma') := \left(1 + \frac{k_\tau(\sigma,\sigma')}{p}\right)^p, \quad \text{and} \tag{11b}$$

$$\overline{k}^{p,\nu}(\sigma,\sigma') := \exp\left(-\frac{\nu}{2}\binom{d}{2}\right)\left(1 + \nu\binom{d}{2}\frac{k_\tau(\sigma,\sigma')}{2p}\right)^p. \tag{11c}$$

We refer to these three kernels as the *polynomial kernel*, the *normalized polynomial kernel*, and the *ν-normalized polynomial kernel* of degree k, respectively. Since each kernel depends only on the number of discordant pairs, they are all right-invariant. Moreover, each kernel is positive semidefinite, since they can each be written as a polynomial function of the Kendall kernel with non-negative coefficients.

Theorem 6. *The Fourier transforms of the three polynomial kernels k^p, $\overline{k}^p$, $\overline{k}^{p,\nu}$ are zero at all irreducible representations ρ_λ with $\lambda \lhd (\max\{d-2p,1\},1,\ldots,1)$. Furthermore, when $p \geq d-1$, the Fourier transform of the three polynomial kernels is strictly positive definite at all irreducible representations.*

The first part of the theorem shows that the polynomial kernels of degree p do not detect differences between distributions at irreps ρ_λ with λ not higher in the partial ordering than the partition $(\max\{d-2p,1\},1,\ldots,1)$. Intuitively, as the degree of the polynomial kernels increases they are able to detect more differences between probability distributions. The second part of the theorem shows that the polynomial kernels of degree at least $d-1$ detect all differences between probability distributions.

The appeal of defining the second and third kernels, $\overline{k}^p$ and $\overline{k}^{p,\nu}$, in addition to the first one, is two-fold. On the one hand, in practice, the kernel k^p becomes difficult to evaluate when p is large because $k^p(\sigma,\sigma) = 2^p$. On the other hand, the two normalized kernels satisfy the relations

$$\lim_{p\to\infty} \overline{k}^p(\sigma,\sigma') = \exp(k_\tau(\sigma,\sigma')) \quad \text{and} \quad \lim_{p\to\infty} \overline{k}^{p,\nu}(\sigma,\sigma') = \exp(-\nu n_d(\sigma,\sigma')). \tag{12}$$

The first limit is a constant times the Mallows kernel k_m^ν with bandwidth $\nu = 2\binom{d}{2}^{-1}$, while the second limit is precisely the Mallows kernel k_m^ν. This observation suggests we can infer properties about the Mallows kernel via the ν-normalized polynomial kernel. Indeed, our proof of Theorem 5 makes use of this fact.

5.1. Feature maps of the polynomial kernels

We now consider the feature spaces associated with the polynomial kernels. We show here how the dimensions of the feature spaces increase as the degree of the kernels increases, eventually leading to the feature space of the Mallows

kernel (up to constants). We give a recursive construction of the feature maps $\Phi_p \colon \mathbb{S}_d \to \mathbb{R}^{\binom{1+\binom{d}{2}}{p}}$ that satisfy the relation $k^p(\sigma, \sigma') = \Phi_p(\sigma)^\top \Phi_p(\sigma')$. First, we use the feature map of the Kendall kernel to construct Φ_1; in particular, the map $\Phi_1 \colon \mathbb{S}_d \to \mathbb{R}^{1+\binom{d}{2}}$ is defined by

$$\Phi_1(\sigma)_{t_0} := 1 \quad \text{and} \quad \Phi_1(\sigma)_{t_r} := \sqrt{\binom{d}{2}^{-1}} \left(2\mathbb{1}_{\{\sigma(i_r)<\sigma(j_r)\}} - 1\right),$$

where the coordinates are indexed by the unordered pair $t_0 = \{-1, 0\}$ and the $\binom{d}{2}$ unordered pairs $t_r = \{i_r, j_r\}$ with $i_r, j_r \in [d]$ and $i_r < j_r$. We denote the set of these unordered pairs by

$$\mathcal{T} \coloneqq \left\{t_0, t_1, \ldots, t_{\binom{d}{2}}\right\}. \tag{13}$$

The feature map Φ_1 clearly satisfies $k^1(\sigma, \sigma') = \Phi_1(\sigma)^\top \Phi_1(\sigma')$. Now we use the map Φ_{p-1} to construct a feature map Φ_p for $p \geq 1$. By definition, we have

$$\begin{aligned} k^p(\sigma, \sigma') &= (1 + k_\tau(\sigma, \sigma'))\,(1 + k_\tau(\sigma, \sigma'))^{p-1} = \Phi_1(\sigma)^\top \Phi_1(\sigma') \Phi_{p-1}(\sigma')^\top \Phi_{p-1}(\sigma) \\ &= \operatorname{tr}\left(\Phi_1(\sigma)^\top \Phi_1(\sigma') \Phi_{p-1}(\sigma')^\top \Phi_{p-1}(\sigma)\right) \\ &= \operatorname{tr}\left(\left(\Phi_1(\sigma)\Phi_{p-1}(\sigma)^\top\right)^\top \Phi_1(\sigma')\Phi_{p-1}(\sigma')^\top\right). \end{aligned}$$

Therefore, the polynomial kernel of degree p between σ and σ' is equal to the inner product of the matrices $\Phi_1(\sigma)\Phi_{p-1}(\sigma)^\top$ and $\Phi_1(\sigma')\Phi_{p-1}(\sigma')^\top$. By induction, we see that Φ_p can be obtained from Φ_1 by taking the outer product with itself p times, meaning that the embedding $\Phi_p : \mathbb{S}_d \to \mathbb{R}^{\binom{1+\binom{d}{2}}{p}}$ can be expressed in terms of a sequence $s_1, s_2, \ldots, s_p$ of elements of $\mathcal{T}$ as

$$\Phi_p(\sigma)_{s_1 s_2 \ldots s_p} = \prod_{i=1}^{p} \Phi_1(\sigma)_{s_i}. \tag{14}$$

It is clear that as the degree of the polynomial kernels increases, the kernels capture more information about the probability distribution of the data. Proposition 2 together with Eq. (14) show that when the degree of the polynomial kernels is at least $\binom{d}{2}$, their feature sets contain all the features of the Mallows kernel (up to constants). This offers another perspective on how the polynomial kernels interpolate between the Kendall kernel and the Mallows kernel.

6. Empirical results

We now present an empirical exploration of our kernel-based methodology. We present results for simulated data and for two real-world datasets—the European Union survey Eurobarometer data and the large-scale MovieLens dataset.

6.1. Experiments with simulated data

Setup. We evaluate the empirical power of two-sample hypothesis tests based on the Kendall and Mallows kernel U-statistics. In order to do so, we chose pairs of probability distributions P and Q over $\mathbb{S}_d$ and then sampled i.i.d. rankings $\alpha_1, \alpha_2, \ldots, a_n$ from P and $\beta_1, \beta_2, \ldots, \beta_n$ from Q. The size of the rankings was fixed to $d = 5$. The hypothesis tests considered here reject the null when $T_k(\alpha, \beta) > t^*$, where the threshold t^* is chosen by permutation testing so as to ensure the probability of a false positive is at most 0.05.

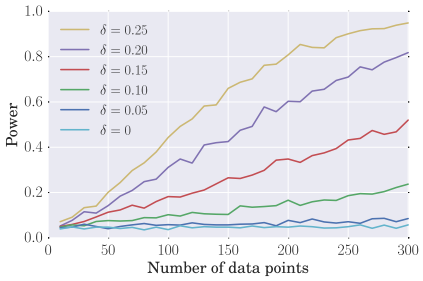
We fixed P to be the uniform distribution over $\mathbb{S}_d$ and chose a distribution Q such that $\|A_\tau(P - Q)\|_2 = \delta$ for different values of δ, where A_τ is the matrix discussed in Proposition 1. For each value of δ there are many distributions Q that are at the prescribed distance from P. When $\delta > 0$, we first sampled uniformly a direction from the complement of the null space of A_τ and then chose the distribution Q at the prescribed distance away from P in that direction. When $\delta = 0$, we sampled uniformly a direction from the null space of A_τ and then chose the distribution Q which is the farthest away from P in that direction.

Once P and Q were fixed, we sampled i.i.d. sets of rankings $\{\alpha_i\}_{i=1}^n$ and $\{\beta_i\}_{i=1}^n$ from the distributions P and Q respectively. We varied the sample size n from 10 to 300 in increments of 10. For each pair of sample sets $\{\alpha_i\}_{i=1}^n$ and $\{\beta_i\}_{i=1}^n$ we used 200 permutations of these $2n$ data points to estimate the rejection threshold t^*. To get estimates of the power of the kernel tests, for each value of n, we sampled 1000 data sets from the fixed distributions P and Q and ran the tests on them, measuring the frequency with which the tests rejected the null hypothesis.

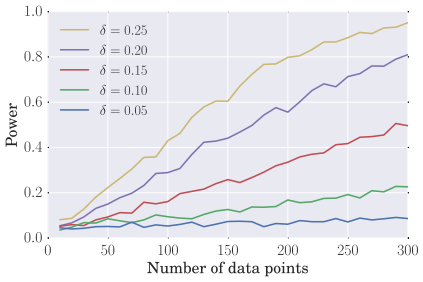
Discussion of results. Recall that $\mu_{k,P} = \mathbb{E}_{\sigma\sim P}\Phi(\sigma)$ denotes the mean embedding of the probability distribution P with respect to the feature map $\Phi\colon \mathbb{S}_d \to \mathbb{R}^m$ of the kernel k. Similarly, we define the covariance matrix of P as $\Sigma_{k,P} = \mathbb{E}_{\sigma\sim P}\Phi(\sigma)\Phi(\sigma)^\top - \mu_{k,P}\mu_{k,P}^\top$. Elementary computations [21] show that

$$\begin{aligned}
\mathbb{E}T_k(\alpha, \beta) &= \|\mu_{k,P} - \mu_{k,Q}\|_2^2 \\
\operatorname{Var} T_k(\alpha, \beta) &= \frac{2}{n(n-1)} \operatorname{tr}\left(\Sigma_{k,P}^2\right) + \frac{2}{n(n-1)} \operatorname{tr}\left(\Sigma_{k,Q}^2\right) + \frac{4}{n^2} \operatorname{tr}\left(\Sigma_{k,P}\Sigma_{k,Q}\right) \\
&\quad + \frac{4}{n}(\mu_{k,P} - \mu_{k,Q})^\top \Sigma_{k,P}(\mu_{k,P} - \mu_{k,Q}) \\
&\quad + \frac{4}{n}(\mu_{k,P} - \mu_{k,Q})^\top \Sigma_{k,Q}(\mu_{k,P} - \mu_{k,Q}).
\end{aligned}$$

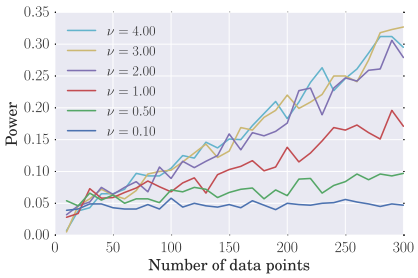
Ramdas et al. [22] showed that for real-valued data, when d and n are sufficiently large, the power of kernel U-statistic tests scales roughly like $\Psi(n\delta^2/V)$ for sufficiently small δ, where Ψ is the Gaussian CDF, and V is a term independent of n which depends on the variance. The kernels over the symmetric



(a) Power of the Kendall kernel MMD.



(b) Power of the Mallows kernel MMD ($\nu = 0.22$).



(c) Power of the Mallows kernel MMD ($\delta = 0$, but $P \neq Q$).

FIG 1. *The empirical power of the MMD two-sample test with the Kendall kernel is shown in plot (a) and with the Mallows kernel in plots (b,c) as a function of the number of data points n. Note that the power of both the Kendall and Mallows kernels increases as δ and n increase. Moreover, the Kendall kernel has trivial power* 0.05 *when $\delta = 0$ (but $P \neq Q$), while the Mallows kernel achieves achieves non-trivial power in this setting as long as the bandwidth is chosen large enough. Finally, note that as the bandwidth ν of the Mallows kernel is increased the variance of the corresponding U-statistic increases as well, as is indicated by the jagged lines shown in plot (c).*

group do not satisfy the necessary assumptions to apply the results of that work, but we observe a similar behavior in our simulations. For instance, Figure 1a shows the empirical power of the Kendall kernel test as function of n for different values of δ. As expected, the power of the test increases as δ increases. More interestingly, observe that for certain values of n a doubling of δ translates into roughly four times more power. Finally, note that when $\delta = 0$ the Kendall kernel test has trivial power 0.05. This behavior meets our expectations based on Proposition 1 and the results of Ramdas et al. [22].

Proposition 2 shows that as the bandwidth ν decreases, the weight of the features $\mathbb{1}_{\{\sigma\colon \sigma(a)<\sigma(b)\}}$ increases relative to higher order features. Therefore, when $\delta > 0$, we expect that the Mallows kernel with a small bandwidth will match the performance of the Kendall kernel. Figure 1b corroborates this intuition—it shows the power of the Mallows kernel with bandwidth $\nu = 0.22$. In order

to choose the bandwidth 0.22, we used a heuristic based on Proposition 2. We chose the largest bandwidth ν such that the weight put by the Mallows kernel on the pairwise comparison features is twice the weight put on all higher-order features. More explicitly, based on Eq. (7), we chose ν to be the maximum value such that

$$\left(\frac{1-\exp(-\nu)}{1+\exp(-\nu)}\right)^{\frac{1}{2}} \geq 2\sum_{r\geq 2}^{\infty}\left(\frac{1-\exp(-\nu)}{1+\exp(-\nu)}\right)^{\frac{r}{2}}. \tag{15}$$

The features constructed in Proposition 2 are interpretable and the precise weighting derived in Eq. (7) can be used to derive heuristics like (15) for choosing the bandwidth ν of the Mallows kernel.

When $\delta = 0$, the low-order features do not capture the difference between P and Q, and therefore a higher bandwidth should yield higher power. In Figure 1c we see that the Mallows kernel has power against the null hypothesis $P = Q$ even when $\delta = 0$. These results agree with the fact that the Mallows kernel is characteristic (Theorem 5). As expected, when $\delta = 0$ a higher bandwidth yields more power, but at the cost of a higher variance of the statistic, as indicated by the jagged curves shown in Figure 1c.

6.2. Survey data

Dataset and methods. In this section, we showcase the use of kernels for hypothesis testing, classification, and regression on a real rankings dataset: the European Union survey Eurobarometer 55.2 [1]. As part of this survey, collected in 2001 across all countries belonging to the European Union, participants expressed their views on topics ranging from the single currency, agriculture, to science and technology. Participants were selected via a multi-stage stratified random sampling method, and there were 16130 respondents in total. As part of the survey, participants were asked to rank in the order of preference six sources of news regarding scientific developments: TV, radio, newspapers and magazines, scientific magazines, the internet, school/university. The data set also includes demographic information such as gender and age; a snippet of the dataset is shown in Table 1.

We removed all respondents who did not provide a complete ranking over the six sources of news, leaving 12216 participants. Then, we split the data set in two distinct ways: across gender, and across age groups (40 or younger and over 40). Out of the 12216 participants, 5915 were men, 6301 were women, 5985 were 40 or younger, 6231 were over 40.

We ran two-sample hypothesis tests across these groups with both the Kendall and the Mallows kernels. Furthermore, we fitted a kernel SVM with the Mallows kernel to predict the age group of participants. Finally, we fitted a kernel ridge regression model with the Mallows kernel to predict the age of participants. For both the classification and regression tasks, we used the Scikit-Learn Python package [23] to fit the models. The bandwidth of the Mallows kernel and the regularization parameter were chosen by cross-validation.

Results and discussion. For the hypothesis tests across gender, we sub-sampled 300 participants from each of the two groups and ran a permutation test with 400 permutations, using the Kendall and the Mallows ($\nu = 1$) kernel U-statistics. We obtained p-values equal to 0.075 and 0.412 respectively. After increasing the number of samples from each group to 600, we obtained p-values equal to 0.002 and 0.002 respectively.

For the hypothesis tests across age groups we sub-sampled 30 participants from each of the two group and ran a permutation test with 400 permutations, using the Kendall and the Mallows ($\nu = 1$) kernel U-statistics. We obtained p-values equal to 0.007 and 0.477 respectively. After increasing the number of samples from each group to 50, we obtained p-values equal to 0.002 and 0.005 respectively. We note that fewer samples than for the tests across gender were required to reject the null hypothesis. For the type of rankings considered here, we did expect a large discrepancy across age groups. In general, young participants are more likely to attend schools or universities, making them more likely to rank highly these institutions as preferred source of information. Moreover, in 2001, it was to be expected that younger participants were more accustomed to the internet than older participants.

For the classification task across age groups, we fit a kernel SVM model using the Mallows kernel. We split the 12216 participants randomly into a training set of 10000 participants, and a test set of 2216 participants. The bandwidth for the Mallows kernel was chosen to be 0.1 through cross-validation. We obtained an error rate of 34%, which is better than chance.

For the regression task to predict age, we fit a kernel ridge regression model using the Mallows kernel. We split the 12216 participants randomly into a training set of 10000 participants, and a test set of 2216 participants. The bandwidth for the Mallows kernel was chosen to be 0.1 through cross-validation. The model predicted the age of the respondents in the test set with an average ℓ_1-error of 11 years.

6.3. Movie ratings

Dataset and methods. Not all rankings come in the form of explicit orderings of alternatives. The MovieLens 1M dataset contains about one million ratings of movies provided by 6000 users of the website movielens.org. For each user in the dataset we are given their gender, age, and occupation, and for each movie we are given its classifications into genres. Each movie can belong to multiple genres such as action, drama, thriller and comedy. The movies contained in this dataset are split into a total of 18 genres. The ratings are measured on a 5-star scale.

For each movie genre we counted the number of movies belonging to that genre, and then kept only the ratings to movies belonging to at least one of the ten most popular genres. Then, for each user we computed the average of the ratings across the ten movie genres. Finally, we removed all users that did not record at least one rating for each of the ten movie genres. The total number of users remaining in the dataset after these procedures was 4428.

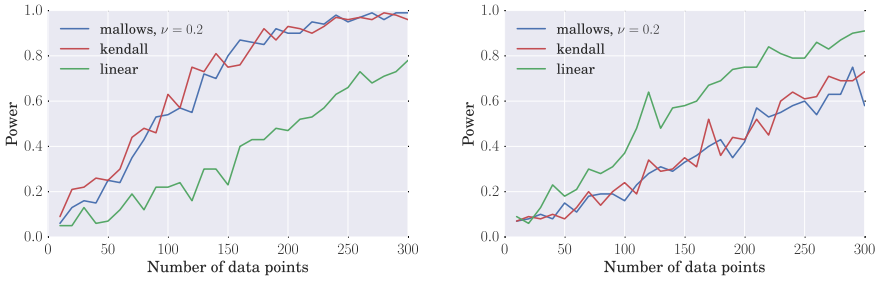
We split this data in two distinct ways: across gender, and across age groups (younger than 35 and 35 or older). Given this data we ran two-sample hypothesis tests across the groups by using the standard linear kernel U-statistic for data in $\mathbb{R}^{10}$. Furthermore, for each user we transformed the average ratings into rankings in the obvious way (the highest average rating takes rank one and so on), breaking ties randomly. Given the data in this new format, we ran two-sample hypothesis tests across the groups by using the Mallows ($\nu = 0.2$) and the Kendall kernel U-statistics. For all three U-statistics, we sub-sampled n samples from each group of users and used 200 permutations to determine the rejection threshold. For each sample size n, we ran 100 trials to estimate the empirical power of the hypothesis tests.

Results and discussion. Our findings are summarized in Figure 2. All three tests reject their respective null hypotheses with power going to one as the number of data points used increases. It is interesting to note that depending on the split of the data, either the Kendall and Mallows tests have more power than the linear kernel, or the other way around. Of course, the linear kernel tests the equality of the probabilities of the average ratings in $\mathbb{R}^{10}$, whereas the Kendall and the Mallows kernel are testing for differences in their respective feature spaces. In particular, the linear kernel would reject the null even when the distribution of preference over movies is the same across groups, but the distribution of scores is not the same (e.g., when one group gives higher scores on average than the other group). Therefore, for certain applications, it is more natural to study the users' distributions of preferences between movie genres by working with rankings rather than with the scores directly.

7. Conclusions

In this paper, we provided feature map and Fourier-analytic characterizations for various right-invariant kernels: the Kendall and Mallows kernels, and a novel family of polynomial kernels. We showed that the Kendall kernel is nearly degenerate in two ways: its Gram matrix has rank $\binom{d}{2}$, and it has only two nonzero Fourier matrices, both of which have rank one. We constructed a $2^{\binom{d}{2}}$ feature map for the Mallows kernel and showed that its Gram matrix has full rank $d!$. This shows that the Mallows kernel is both universal and characteristic, and, in Fourier space, this means that the Mallows kernel has a strictly positive definite Fourier transform at all the irreducible representations. Moreover, the Mallows kernel allows the feature map to be obtained in closed form, which informs the choice of bandwidth, as we showed for the Mallows kernel in the two-sample testing experiments.

We thus see that the Kendall and Mallows kernels are quite different, even though both of them depend only on counting discordant pairs between rankings. There is a natural analogy between these kernels in the space of permutations to the linear and Gaussian kernels in Euclidean space. Building on this analogy, we proposed a new class of polynomial kernels that smoothly interpo-



(a) Hypothesis tests across gender. (b) Hypothesis tests across age groups.

FIG 2. *The empirical power of the Mallows, Kendall, and linear kernel tests applied to average scores of movie genres data. This data was obtained by averaging users' movie scores included in the MovieLens 1M dataset and then splitting the users in two ways: by gender, and by age. For both settings, all tests achieve a non-trivial power. However, we emphasize that the linear kernel, the Kendall, and Mallows kernels test different null hypotheses.*

late between the Kendall and Mallows extremes, yielding a hierarchy of kernels that are sensitive to differences between distributions at an increasingly dense set of frequencies.

Many properties of the Fourier transform of the Mallows and polynomial kernels are still not understood. For example, unlike the case of the Kendall kernel, we do *not* have closed-form descriptions of the Fourier matrices for these kernels. Such concise expressions would not only be of mathematical interest, but could also be useful for computing in the spectral domain. It would also be interesting to understand the properties of these kernels when applied to partial rankings (top-k or random-k), which is even harder because partial rankings do not jointly form a group. We view the current results on kernels for full rankings as an important step towards developing and rigorously analyzing kernel methods for partial rankings.

In Section 6, we studied the empirical power of kernel U-statistic two-sample tests with the Kendall and Mallows kernels under different sets of alternatives. The scaling of the power with the distance between the alternatives is similar to that of the linear and Gaussian kernel over real data. It would be interesting to characterize the power of two-sample tests using the Kendall or Mallows kernel as a function of the number of samples, the number of objects and an appropriate signal-to-noise ratio. Our final set of experiments involved data transformed from raw numerical scores (ratings of movies) into rankings, a transformation also explored by Jiao and Vert [5]. This type of reduction to rank statistics, while well studied in the context of classical rank-based methods for testing [2], merits further study in the context of permutation-based covariates. It offers invariance to arbitrary monotone transformations of the covariates, and hence a way of protecting against model mismatch and/or covariate biases.

Outline of the Appendix

The appendix includes all the proofs of the results presented in the main text. Sections A.1 and A.2 contain the proofs of the results presented in Section 3 of the main text. In Sections A.3, A.4, and A.5 we prove our results concerning the Fourier spectra of the Kendall, polynomial, and Mallows kernels. Section B contains further background material on representation theory and Fourier analysis on the symmetric group needed in the proofs. Finally, Section C contains some proofs of miscellaneous claims that are used throughout.

Appendix A: Proofs of main results

A.1. Proof of Proposition 1

We prove that all the basis vectors of $\mathbb{R}^{\binom{d}{2}}$ are in the span of the matrix A_τ. To achieve this we order the tuples $t_i = (a_i, b_i)$, with $a_i < b_i$, as follows. The tuples t_i and t_j are ordered $t_i < t_j$ if and only if $a_i < a_j$ or $a_i = a_j$ and $b_i < b_j$. With total ordering fixed the i-th coordinate of $\mathbb{R}^{\binom{d}{2}}$ corresponds to the tuple t_i.

Then it is enough to prove that for any $1 \leq j \leq \binom{d}{2}$ the vector $v_j = \sum_{i=1}^{j} e_i$ is equal to the column $(A_\tau)_\sigma$ for some appropriate $\sigma \in \mathbb{S}_d$, where e_i is the i-th standard basis vectors of $\mathbb{R}^{\binom{d}{2}}$.

We will construct inductively the permutations π_j such that $(A_\tau)_{\pi_j} = v_j$. Observe that the identity permutation $\pi_{\binom{d}{2}}(i) = i$ satisfies $(A_\tau)_{\pi_{\binom{d}{2}}} = v_{\binom{d}{2}}$. Also, note that if we swap the ranks of d and $d-1$ in the permutation $\pi_{\binom{d}{2}}$, we obtain a permutation $\pi_{\binom{d}{2}-1}$ such that $(A_\tau)_{\pi_{\binom{d}{2}-1}} = v_{\binom{d}{2}-1}$.

Assume we have constructed π_{j+1} such that $(A_\tau)_{\pi_{j+1}} = v_{j+1}$. We construct π_j such that $(A_\tau)_{\pi_j} = v_j$. Let $t_{j+1} = (a, b)$ be the tuple corresponding to the $j+1$-st coordinate. Since $(A_\tau)_{\pi_{j+1}} = v_{j+1}$, we have $\pi_{j+1}(a) < \pi(r)$ for all $a < r \leq b$, and $\pi_{j+1}(a) > \pi_{j+1}(r)$ for all $r > b$. Moreover $\pi_{j+1}(r) > \pi_{j+1}(r+1)$ for all $a < r < b$. Therefore, if we choose $\pi_j(r) = \pi_{j+1}(r)$ for all r distinct from a and b and $\pi_j(a) = \pi_{j+1}(b)$, $\pi_j(b) = \pi_{j+1}(a)$, we find that $(A_\tau)_{\pi_j} = v_j$. The conclusion follows.

A.2. Proof of Proposition 2

We begin with the observation that the ν-normalized polynomial kernel converges to the Mallows kernel as its degree increases:

$$\overline{k}^{p,\nu}(\sigma, \sigma') = e^{-\frac{\nu}{2}\binom{d}{2}} \left(1 + \nu \binom{d}{2} \frac{k_\tau(\sigma, \sigma')}{2p}\right)^p \xrightarrow{p\to\infty} e^{-\nu n_d(\sigma,\sigma')} = k_m^\nu(\sigma, \sigma').$$

We construct a feature map for the Mallows kernel by exploiting this observation. Specifically, we derive a feature map of the ν-normalized polynomial kernel

and compute its limit as the degree p of the kernel goes to infinity. Similar to Section 3, we define the feature map $\overline{\Phi} : \mathbb{S}_d \to \mathbb{R}^{\binom{d}{2}+1}$:

$$\overline{\Phi}(\sigma)_{t_r} := 2\mathbb{1}_{\{\sigma(a_r)<\sigma(b_r)\}} - 1,$$

where the coordinates are indexed by the ordered pairs $t_r = \{a_r, b_r\}$ with $a_r, b_r \in [d]$ and $a_r < b_r$. Let $\mathcal{T}^*$ denote the set of $\binom{d}{2}$ such tuples. Then, a binomial expansion yields

$$\begin{aligned}\overline{k}^{p,\nu}(\sigma,\sigma') &= e^{-\frac{\nu}{2}\binom{d}{2}} \left(1 + \frac{\nu}{2p}\sum_{i=1}^{\binom{d}{2}} \overline{\Phi}(\sigma)_{t_i}\overline{\Phi}(\sigma')_{t_i}\right)^p \\ &= e^{-\frac{\nu}{2}\binom{d}{2}} \sum_{c_0+\ldots+c_{\binom{d}{2}}=p} \frac{p!}{c_0!c_1!\ldots c_{\binom{d}{2}}!}\left(\frac{\nu}{2p}\right)^{p-c_0} \prod_{i=1}^{\binom{d}{2}} \overline{\Phi}(\sigma)_{t_i}^{c_i} \prod_{j=1}^{\binom{d}{2}} \overline{\Phi}(\sigma')_{t_j}^{c_j}.\end{aligned}$$

Note that $(\overline{\Phi}(\sigma))_{t_i}^2 = 1$ for any $t_i \in \mathcal{T}^*$ and any $\sigma \in \mathbb{S}_d$. For any $A \subset \mathcal{T}^*$ we denote $\overline{\Phi}(\sigma)_A := \prod_{t_i \in A} \overline{\Phi}(\sigma)_{t_i}$, and $\overline{\Phi}(\sigma)_\emptyset = 1$. Hence, we can simplify the above expression to

$$\overline{k}^{p,\nu}(\sigma,\sigma') = e^{-\frac{\nu}{2}\binom{d}{2}} \sum_{A\subset\mathcal{T}^*} \overline{\Phi}(\sigma)_A\overline{\Phi}(\sigma')_A \sum_{\substack{c_0+\ldots+c_{\binom{d}{2}}=p \\ c_i \text{ odd when } t_i\in A \\ c_i \text{ even when } t_i\notin A}} \frac{p!}{c_0!c_1!\ldots c_{\binom{d}{2}}!}\left(\frac{\nu}{2p}\right)^{p-c_0}.$$

By symmetry the second sum on the right-hand side depends only on the power p and the size of the set A. Therefore, if we define

$$\delta(p,r) = \sum_{\substack{c_0+\ldots+c_{\binom{d}{2}}=p \\ c_i \text{ odd when } 1\le i\le r \\ c_i \text{ even when } r<i}} \frac{p!}{c_0!c_1!\ldots c_{\binom{d}{2}}!}\left(\frac{\nu}{2p}\right)^{p-c_0},$$

we find that

$$\overline{k}^{p,\nu}(\sigma,\sigma') = e^{-\frac{\nu}{2}\binom{d}{2}} \sum_{A\subset\mathcal{T}^*} \overline{\Phi}(\sigma)_A\overline{\Phi}(\sigma')_A\delta(p,|A|).$$

We are left to compute the limit of $\delta(p,|A|)$ as $p \to \infty$, and to this end we construct a generating function for the sequence $\delta(p,r)$ by defining

$$F(z) := p!\left(\frac{\nu}{2p}\right)^p e^{\frac{2p}{\nu}z}\left(\frac{e^z - e^{-z}}{2}\right)^r \left(\frac{e^z + e^{-z}}{2}\right)^{\binom{d}{2}-r}.$$

By Taylor expanding each term e^z individually we see that the function $F(z)$ is the generating function of $d(p,r)$ for $0 \le r \le \binom{d}{2}$. More precisely, $\delta(p,r)$ is the

p-th coefficient of the generating function $F(z)$. By expanding $F(z)$ into a linear combination of exponentials we can compute $\delta(p,r)$ and study its asymptotic behavior:

$$\begin{aligned}
F(z) &= p!\left(\frac{\nu}{2p}\right)^p \frac{e^{\frac{2p}{\nu}z-\binom{d}{2}z}}{2^{\binom{d}{2}}}\left(e^{2z}-1\right)^r\left(e^{2z}+1\right)^{\binom{d}{2}-r} \\
&= \frac{p!}{2^{\binom{d}{2}}}\left(\frac{\nu}{2p}\right)^p e^{\frac{2p}{\nu}z-\binom{d}{2}z}\sum_{i=0}^{r}\sum_{j=0}^{\binom{d}{2}-r} e^{2(i+j)z}(-1)^{r-i}\binom{r}{i}\binom{\binom{d}{2}-r}{j} \\
&= \frac{p!}{2^{\binom{d}{2}}}\left(\frac{\nu}{2p}\right)^p \sum_{i=0}^{r}\sum_{j=0}^{\binom{d}{2}-r} e^{\left(\frac{2p}{\nu}-\binom{d}{2}+2(i+j)\right)z}(-1)^{r-i}\binom{r}{i}\binom{\binom{d}{2}-r}{j}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\delta(p,r) &= \frac{1}{2^{\binom{d}{2}}}\left(\frac{\nu}{2p}\right)^p \sum_{i=0}^{r}\sum_{j=0}^{\binom{d}{2}-r}\left(\frac{2p}{\nu}-\binom{d}{2}+2(i+j)\right)^p (-1)^{r-i}\binom{r}{i}\binom{\binom{d}{2}-r}{j} \\
&= \frac{1}{2^{\binom{d}{2}}}\sum_{i=0}^{r}\sum_{j=0}^{\binom{d}{2}-r}\left(1-\frac{\frac{\nu}{2}\left(\binom{d}{2}-2(i+j)\right)}{p}\right)^p (-1)^{r-i}\binom{r}{i}\binom{\binom{d}{2}-r}{j} \\
&\xrightarrow{p\to\infty} \frac{1}{2^{\binom{d}{2}}}\sum_{i=0}^{r}\sum_{j=0}^{\binom{d}{2}-r} e^{-\frac{\nu}{2}\left(\binom{d}{2}-2(i+j)\right)}(-1)^{r-i}\binom{r}{i}\binom{\binom{d}{2}-r}{j} \\
&= \frac{e^{-\frac{\nu}{2}\binom{d}{2}}}{2^{\binom{d}{2}}}\sum_{i=0}^{r}\sum_{j=0}^{\binom{d}{2}-r} e^{\nu i}e^{\nu j}(-1)^{r-i}\binom{r}{i}\binom{\binom{d}{2}-r}{j} \\
&= \frac{e^{-\frac{\nu}{2}\binom{d}{2}}}{2^{\binom{d}{2}}}\left(e^{\nu}-1\right)^r\left(e^{\nu}+1\right)^{\binom{d}{2}-r}.
\end{aligned}$$

The conclusion follows.

A.3. Proof of Theorem 3

For $d=2$ and $d=3$, the irreps ρ_λ are easy to describe in closed form [12]; in particular, we have

$$\widehat{k_\tau}(\rho_{(2)})=0, \quad \widehat{k_\tau}(\rho_{(1,1)})=2$$
$$\widehat{k_\tau}(\rho_{(3)})=0, \quad \widehat{k_\tau}(\rho_{(2,1)})=\begin{pmatrix}\frac{2}{3} & \frac{2}{\sqrt{3}}\\ \frac{2}{\sqrt{3}} & 2\end{pmatrix}, \quad \widehat{k_\tau}(\rho_{(1,1,1)})=\frac{2}{3}.$$

Accordingly, it remains to prove Theorem 3 when $d\geq 4$. Each representation $\rho\colon \mathbb{S}_d\to\mathbb{C}^{d_\rho\times d_\rho}$ defines a collection of d_ρ^2 functions $\sigma\mapsto\rho(\sigma)_{ij}$ on the symmetric

group. An important result in the representation theory states that the functions defined by the irreps ρ_λ form a basis for the space of functions over the symmetric group. To exploit this fact, we express the Kendall kernel as a linear combination of the functions defined by the overcomplete representation $\tau_{(d-2,1,1)}$ defined in Section B, Eq. (23).

Lemma 7. *The Kendall function $\sigma \mapsto k_\tau(\sigma)$ is a linear combination of the functions defined by the representation $\tau_{(d-2,1,1))}$.*

Proof. A rough sketch of the argument is as follows. The Kendall function is a linear combination of indicator functions $\mathbb{1}_{\{\sigma(i)>\sigma(j)\}}$ (plus a constant). The result follows because each of these functions is a linear combination of the indicator functions $\mathbb{1}_{\{\sigma(i)=l,\sigma(j)=r\}}$, which are exactly the functions defined by $\tau_{(d-2,1,1)}$.

Formally, to prove the claim it suffices to express the function k_τ as a linear combination of the functions defined by $\rho_{(d)}$, $\rho_{(d-1,1)}$, $\rho_{(d-2,2)}$, and $\rho_{(d-2,1,1)}$. James' submodule theorem states that

$$\tau_{(d-2,1,1)} \equiv \rho_{(d)} \oplus \rho_{(d-1,1)} \oplus \rho_{(d-1,1)} \oplus \rho_{(d-2,2)} \oplus \rho_{(d-2,1,1)}.$$

Therefore, we just have to show that the Kendall function is a linear combination of the functions defined by $\tau_{(d-2,1,1)}$. We have

$$\begin{aligned} k_\tau(\sigma) = 1 - 2\frac{\mathrm{i}(\sigma)}{\binom{d}{2}} &= 1 - 2\binom{d}{2}^{-1} \sum_{i<j} \mathbb{1}_{\{\sigma(i)>\sigma(j)\}} \\ &= 1 - 2\binom{d}{2}^{-1} \sum_{i<j} \sum_{l<r} \mathbb{1}_{\{\sigma(i)=r,\sigma(j)=l\}}. \end{aligned}$$

The functions $\mathbb{1}_{\{\sigma(i)=r,\sigma(j)=l\}}$ are defined by $\tau_{(d-2,1,1)}$ by construction, completing the proof. □

Our next step in proving Theorem 3 is to compute the Fourier transform of the Kendall transform at the representations $\tau_{(d)}$, $\tau_{(d-1,1)}$, $\tau_{(d-2,2)}$, and $\tau_{(d-2,1,1)}$. Our next lemma summarizes the results of these computations, which though technical are conceptually straightforward. First define

- vector $u \in \mathbb{R}^d$ with entries $u_i = d - 2i + 1$ for $i = 1, \ldots, d$
- vector $w \in \mathbb{R}^{\binom{d}{2}}$ with entries $w_{\{i,j\}} = 2d - 2(i + j) + 2$ for $1 \le i < j \le d$.
- vectors v_1, v_2, and v_3 in $\mathbb{R}^{d(d-1)}$ with entries

$$\begin{aligned} [v_1]_{(i,j)} &= 1 - 2\mathbb{1}_{\{i>j\}}, \\ [v_2]_{(i,j)} &= d - 2i + 2 - 2\mathbb{1}_{\{i<j\}}, \\ [v_3]_{(i,j)} &= d - 2j + 2 - 2\mathbb{1}_{\{i>j\}}. \end{aligned}$$

With this notation, we have the following:

Lemma 8. *The Fourier transform of the Kendall kernel satisfies the identities:*

$$\widehat{k_\tau}(\tau_{(d)}) = 0, \quad \widehat{k_\tau}(\tau_{(d-1,1)}) = \frac{(d-2)!}{\binom{d}{2}} uu^\top, \tag{16a}$$

$$\widehat{k_\tau}(\tau_{(d-2,2)}) = \frac{(d-3)!}{\binom{d}{2}} ww^\top, \qquad \textit{and} \tag{16b}$$

$$\widehat{k_\tau}(\tau_{(d-2,1,1)}) = \frac{(d-2)!}{\binom{d}{2}} v_1 v_1^\top + \frac{(d-3)!}{\binom{d}{2}} v_2 v_2^\top + \frac{(d-3)!}{\binom{d}{2}} v_3 v_3^\top. \tag{16c}$$

Proof. Lemma 8 states in closed form the values of $\widehat{k}_\tau$ evaluated at the four representations $\tau_{(d)}$, $\tau_{(d-1,1)}$, $\tau_{(d-2,2)}$, and $\tau_{(d-2,1,1)}$. We compute these values one at a time.

Computing $\widehat{k}_\tau(\tau_{(d)})$. We first show that $\widehat{k}_\tau(\tau_{(d)}) = 0$. Recall that $\tau_{(d)}$ is the trivial representation, equal to 1 at all permutations. Therefore, we need to check that $\sum_{\sigma\in\mathbb{S}_d} 1 - 2\binom{d}{2}^{-1} \mathrm{i}(\sigma) = 0$. Note that we have

$$\sum_{\sigma\in\mathbb{S}_d} \mathrm{i}(\sigma) = \sum_{\sigma\in\mathbb{S}_d} \sum_{i<j} \mathbb{1}_{\{\sigma(i)>\sigma(j)\}} = \sum_{i<j} \sum_{\sigma\in\mathbb{S}_d} \mathbb{1}_{\{\sigma(i)>\sigma(j)\}} = \sum_{i<j} \frac{d!}{2} = \frac{d!\binom{d}{2}}{2},$$

so that the conclusion follows.

Computing $\widehat{k}_\tau(\tau_{(d-1,1)})$. In this case, we show that $\widehat{k}_\tau(\tau_{(d-1,1)}) = \frac{(d-2)!}{\binom{d}{2}} vv^\top$, where the vector $v \in \mathbb{R}^d$ has components $v_r = d - 2r + 1$. Consider the functions g_{ij} on $\mathbb{S}_d$ defined by $g_{ij}(\sigma) = 1 - 2\mathbb{1}_{\{\sigma(i)>\sigma(j)\}}$, for all $i < j$. Then

$$k_\tau(\sigma) = \frac{1}{\binom{d}{2}} \sum_{i<j} g_{ij}(\sigma) \;\text{ and hence }\; \widehat{k}(\rho) = \frac{1}{\binom{d}{2}} \sum_{i<j} \widehat{g}_{ij}(\rho),$$

for any representation ρ.

We compute $\widehat{g}_{ij}(\tau_{(d-2,2)})$ for each tuple $i < j$ and then sum up the results. The rows of $\tau_{(d-1,1)}$ are indexed by tabloids of shape $(d-1,1)$. Each of these tabloids is fully specified by the index contained in the second row. We identify the tabloids of shape $(d-1,1)$ with those indices. Let t_1 and t_2 be two indices in $[d]$. Then

$$\widehat{g}_{ij}(\tau_{(d-1,1)})_{t_1 t_2} = \sum_{\sigma\in\mathbb{S}_d} \left(1 - 2\mathbb{1}_{\{\sigma(i)>\sigma(j)\}}\right) \mathbb{1}_{\{\sigma(t_1)=t_2\}}.$$

There are three cases to consider. First, suppose that t_1 is distinct from both i and j. There are $(d-1)!$ permutations σ that satisfy $\sigma(t_1) = t_2$, out of which exactly half satisfy $g_{ij}(\sigma) = 1$ and the other half satisfy $g_{ij}(\sigma) = -1$. Therefore, we are guaranteed that $\widehat{g}_{ij}(\sigma)_{t_1 t_2} = 0$ when $t_1 \notin \{i, j\}$.

In the second case we may assume that $t_1 = i$. Then, out of the $(d-1)!$ permutations that satisfy $\sigma(i) = t_2$ there are $(t_2 - 1)(d-2)!$ permutations that

satisfy $\sigma(i) > \sigma(j)$ and $(d-t_2)(d-2)!$ that satisfy the opposite inequality. Hence, when $t_1 = i$, we have $\widehat{g}_{ij}(\tau_{(d-1,1)}) = (d-2t_2+1)(d-2)!$.

The remaining (third) case is when $t_1 = j$. Then, out of the $(d-1)!$ permutations with $\sigma(j) = t_2$ there are $(d-t_2)(d-2)!$ with $\sigma(i) > \sigma(j)$ and $(t_{12}-1)(d-2)!$ with $\sigma(i) < \sigma(j)$. Therefore $\widehat{g}_{ij}(\tau_{(d-1,1)}) = -(d-2\tau_2+1)(d-2)!$ when $t_1 = j$. To summarize, we have

$$\widehat{g}_{ij}(\tau_{(d-1,1)})_{t_1 t_2} = \begin{cases} 0 & \text{if } t_1 \notin \{i,j\} \\ (d-2t_2+1)(d-2)! & \text{if } t_1 = i \\ (2t_2-d-1)(d-2)! & \text{if } t_1 = j. \end{cases}$$

Now we need to sum the Fourier transforms of the functions g_{ij} to obtain the Fourier transform of k_τ. We have

$$\begin{aligned} \widehat{k}_\tau(\tau_{(d-1,1)})_{t_1 t_2} &= \binom{d}{2}^{-1} \sum_{i<j} \widehat{g}_{ij}(\tau_{(d-1,1)}) \\ &= \binom{d}{2}^{-1} \sum_{t_1 = i < j} (d-2t_2+1)(d-2)! \\ &\quad + \binom{d}{2}^{-1} \sum_{i<j=t_1} (2t_2-d-1)(d-2)! \\ &= \binom{d}{2}^{-1} (d-t_1)(d-2t_2+1)(d-2)! \\ &\quad + \binom{d}{2}^{-1} (t_1-1)(2t_2-d-1)(d-2)! \\ &= \binom{d}{2}^{-1} (d-2t_1+1)(d-2t_2+1)(d-2)!, \end{aligned}$$

as claimed.

Computing $\widehat{k}_\tau(\tau_{(d-2,2)})$. Now, we show that $\widehat{k}_\tau(\tau_{(d-2,2)}) = \frac{(d-3)!}{\binom{d}{2}} w w^\top$, where the vector $w \in \mathbb{R}^{\binom{d}{2}}$ has entries $w_{\{r_1,r_2\}} = 2d - 2(r_1+r_2) + 2$ for all r_1 and r_2 such that $1 \le r_1 < r_2 \le d$.

The entries of $\widehat{k}_\tau(\tau_{(d-2,2)})$ are indexed by tabloids of shape $(d-2,2)$ which can be identified with the set of two indices contained in the second row. Therefore we can identify the tabloids of shape $(d-2,2)$ with sets of two indices. Fix two such sets $t_1 = \{t_{11}, t_{12}\}$ and $t_2 = \{t_{21}, t_{22}\}$. As before, $g_{ij}(\sigma) := 1 - 2\mathbb{1}_{\{\sigma(i)>\sigma(j)\}}$. For these functions, we have

$$\begin{aligned} \widehat{g}_{ij}(\tau_{(d-2,2)})_{t_1 t_2} &= \sum_{\sigma \in \mathbb{S}_d} g_{ij}(\sigma) \mathbb{1}_{\{\sigma(\{t_{11},t_{12}\}) = \{t_{21},t_{22}\}\}} \\ &= \sum_{\sigma \in \mathbb{S}_d} \left(1 - 2\mathbb{1}_{\{\sigma(i)>\sigma(j)\}}\right) \mathbb{1}_{\{\sigma(t_{11})=t_{21}, \sigma(t_{12})=t_{22}\}} \\ &\quad + \sum_{\sigma \in \mathbb{S}_d} \left(1 - 2\mathbb{1}_{\{\sigma(i)>\sigma(j)\}}\right) \mathbb{1}_{\{\sigma(t_{11})=t_{22}, \sigma(t_{12})=t_{21}\}}. \end{aligned}$$

By breaking into four cases, similar to the proof the computation of $\widehat{k}_\tau(\tau_{(d-2,2)})$, we obtain

$$\widehat{g}_{ij}(\tau_{(d-2,2)})_{t_1t_2} = \begin{cases} 0 & \text{if } \{t_{11},t_{12}\}\cap\{i,j\}=\emptyset \\ 0 & \text{if } \{t_{11},t_{12}\}=\{i,j\} \\ (2d-2(t_{21}+t_{22})+2)(d-3)! & \text{if } \{t_{11},t_{12}\}\cap\{i\}=\{i\} \\ (2(t_{21}+t_{22})-2d-2)(d-3)! & \text{if } \{t_{11},t_{12}\}\cap\{i\}=\{j\}. \end{cases}$$

Summing the terms $\widehat{g}_{ij}(\tau_{(d-2,2)})$ over pairs $i<j$ yields the result.

Computing $\widehat{k}_\tau(\tau_{(d-2,1,1)})$. We show that

$$\widehat{k}_\tau(\tau_{(d-2,1,1)}) = \frac{(d-2)!}{\binom{d}{2}} v_1 v_1^\top + \frac{(d-3)!}{\binom{d}{2}} v_2 v_2^\top + \frac{(d-3)!}{\binom{d}{2}} v_3 v_3^\top,$$

where v_1, v_2, and v_3, are the vectors in $\mathbb{R}^{d(d-1)}$ defined by

$$\begin{aligned} [v_1]_{(r_1,r_2)} &= 1 - 2\mathbb{1}_{\{r_1>r_2\}}, \\ [v_2]_{(r_1,r_2)} &= d - 2r_1 + 2 - 2\mathbb{1}_{\{r_1<r_2\}}, \\ [v_3]_{(r_1,r_2)} &= d - 2r_2 + 2 - 2\mathbb{1}_{\{r_1>r_2\}}. \end{aligned}$$

The same ideas used in the computation of $\widehat{k}_\tau(\tau_{(d-2,2)})$ apply here as well, but the analysis is a bit more detailed because there are more cases to consider. The entries of $\widehat{k}_\tau(\tau_{(d-2,1,1)})$ are indexed by tabloids of shape $(d-2,1,1)$. These tabloids are completely specified by the entries contained in the second and third rows. Hence, we can identify them with ordered tuples in $[d]^2$. Fixing two such tuples $t_1=(t_{11},t_{12})$ and $t_2=(t_{21},t_{22})$, with $t_{11}\neq t_{12}$ and $t_{21}\neq t_{22}$, we then have

$$\begin{aligned} \widehat{g}_{ij}(\tau_{(d-2,1,1)})_{t_1t_2} &= \sum_{\sigma\in\mathbb{S}_d} g_{ij}(\sigma)\mathbb{1}_{\{\sigma(t_{11})=t_{21},\sigma(t_{12})=t_{22}\}} \\ &= \sum_{\sigma\in\mathbb{S}_d} \left(1-2\mathbb{1}_{\{\sigma(i)>\sigma(j)\}}\right)\mathbb{1}_{\{\sigma(t_{11})=t_{21},\sigma(t_{12})=t_{22}\}}. \end{aligned}$$

Arguments similar to the ones used in the computation of $\widehat{g}_{ij}(\tau_{(d-1,1)})$ enable us to compute $\widehat{g}_{ij}(\tau_{(d-2,1,1)})$ as well. In order to make the result more readable, let us split them into to cases: $t_{21}<t_{22}$ and $t_{21}>t_{22}$. Then we obtain

$$\widehat{g}_{ij}(\tau_{(d-2,1,1)})_{t_1t_2} = \begin{cases} 0 & \text{if } \{t_{11},t_{12}\}\cap\{i,j\}=\emptyset \\ (d-2)! & \text{if } t_{11}=i,\ t_{12}=j,\ t_{21}<t_{22} \\ -(d-2)! & \text{if } t_{11}=j,\ t_{12}=i,\ t_{21}<t_{22} \\ (d-2t_{21})(d-3)! & \text{if } t_{11}=i,\ t_{12}\neq j,\ t_{21}<t_{22} \\ (d-2t_{22}+2)(d-3)! & \text{if } t_{11}\neq j,\ t_{12}=i,\ t_{21}<t_{22} \\ (2t_{21}-d)(d-3)! & \text{if } t_{11}=j,\ t_{12}\neq i,\ t_{21}<t_{22} \\ (2t_{22}-d-2)(d-3)! & \text{if } t_{11}\neq i,\ t_{12}=j,\ t_{21}<t_{22}, \end{cases} \tag{17a}$$

$$\widehat{g}_{ij}(\tau_{(d-2,1,1)})_{t_1t_2} = \begin{cases} 0 & \text{if } \{t_{11},t_{12}\}\cap\{i,j\}=\emptyset \\ -(d-2)! & \text{if } t_{11}=i,\ t_{12}=j,\ t_{21}>t_{22} \\ (d-2)! & \text{if } t_{11}=j,\ t_{12}=i,\ t_{21}>t_{22} \\ (d-2t_{21}+2)(d-3)! & \text{if } t_{11}=i,\ t_{12}\neq j,\ t_{21}>t_{22} \\ (d-2t_{22})(d-3)! & \text{if } t_{11}\neq j,\ t_{12}=i,\ t_{21}>t_{22} \\ (2t_{21}-d-2)(d-3)! & \text{if } t_{11}=j,\ t_{12}\neq i,\ t_{21}>t_{22} \\ (2t_{22}-d)(d-3)! & \text{if } t_{11}\neq i,\ t_{12}=j,\ t_{21}>t_{22}. \end{cases} \tag{17b}$$

The conclusion then follows by computing the sum $\sum_{i<j} g_{ij}(\tau_{(d-2,1,1)})_{t_1t_2}$ in the four possible cases obtained from the orderings of t_{11} and t_{12}, and of t_{21} and t_{22}. □

At this point, Theorem 3 follows from Lemma 7 and Lemma 8. Lemma 7 together with decomposition (28) imply that $\widehat{k_\tau}(\rho_\lambda)=0$ for all $\lambda \lhd (d-2,1,1)$ because the functions defined by irreps are a basis for the space of functions on $\mathbb{S}_d$. Next, observe that $\widehat{k_\tau}(\tau_{(d)})=0$ is equivalent to $\widehat{k_\tau}(\rho_{(d)})=0$. Then, since the matrix $\widehat{k_\tau}(\tau_{(d-1,1)})$ has rank one, the decomposition (26) of the representation $\tau_{(d-1,1)}$ implies that $\widehat{k_\tau}(\rho_{(d-1,1)})$ has rank one as well. Also, since both matrices $\widehat{k_\tau}(\tau_{(d-1,1)})$ and $\widehat{k_\tau}(\tau_{(d-2,2)})$ have rank one, from decomposition (27) of the representation $\tau_{(d-2,2)}$ we obtain $\widehat{k_\tau}(\rho_{(d-2,2)})=0$. Finally, since the matrix $\widehat{k_\tau}(\tau_{(d-2,1,1)})$ has rank three, from decomposition (28) of the representation $\tau_{(d-2,1,1)}$ we know that $\widehat{k_\tau}(\rho_{(d-2,1,1)})$ has rank one, which completes the proof of Theorem 3.

A.4. Proof of Theorem 6

We use an approach similar to the proof of Theorem 3.

Lemma 9. *The kernels k^p, $\overline{k}^p$, and $\overline{k}^{p,\nu}$ are linear combinations of the functions defined by the representation $\tau_{(\max\{d-2p,1\},1,\dots,1)}$.*

Proof. We express the function $\sigma \mapsto k^p(\sigma)$ as a linear combination of the functions defined by the representation $\tau_{(\max\{d-2p,1\},1,\dots,1)}$. The same property can be proved for $\overline{k}^p$ and $\overline{k}^{p,\nu}$ analogously.

We first analyze the case $2p<d$. By definition, we have

$$\begin{aligned} k^p(\sigma) = (1+k_\tau(\sigma))^p &= \left(2-\frac{2}{\binom{d}{2}}\,\mathrm{i}(\sigma)\right)^p = 2^p\sum_{r=1}^{p}(-1)^r\binom{p}{r}\,\mathrm{i}(\sigma)^r \\ &= 2^p\sum_{r=1}^{p}(-1)^r\binom{p}{r}\left(\sum_{i<j}\mathbb{1}_{\{\sigma(i)>\sigma(j)\}}\right)^r, \end{aligned}$$

showing that the polynomial kernel k^p is a linear combination of products of functions $\mathbb{1}_{\{\sigma(i)>\sigma(j)\}}$. These products contain at most p terms, which means there are at most $2p$ values $\sigma(i_1)$, $\sigma(i_2)$, ..., $\sigma(i_{2p})$ on which the products depend. The indicator functions of events of the form $\{\sigma(i_1) = j_1, \ldots, \sigma(i_{2p}) = j_{2p}\}$ form a basis for all the functions that depend only on the values $\sigma(i_1)$, $\sigma(i_2)$, ..., $\sigma(i_{2p})$. The conclusion follows for the case $2p < d$ because these indicator functions are exactly the functions defined by the representation $\tau_{(\max\{d-2p,1\},1,\ldots,1)}$.

The case $2p \geq d$ follows analogously once we note that any product of $2p$ indicator functions $\mathbb{1}_{\{\sigma(i)>\sigma(j)\}}$ is determined by $d-1$ values $\{\sigma(i_1), \ldots, \sigma(i_{d-1})\}$. (To be clear, this is because given $d-1$ such values, the d^{th} value is fixed). □

Then, by the James submodule theorem and the linear independence of the functions defined by irreps, we find that the Fourier transforms of the three polynomial kernels are zero at all irreps ρ_λ with $\lambda \lhd (\max\{d-2p, 1\}, 1, \ldots, 1)$. The first part of Theorem 6 is now proved.

To prove the second part of Theorem 6 we make use of feature maps of the three polynomial kernels. Up to constants, the feature maps for the three kernels k^p, $\overline{k}^p$, and $\overline{k}^{p,\nu}$ are the same. For simplicity, we work with the kernel k^p. All the arguments presented here extend to the other two polynomial kernels as well.

We now give a recursive construction of the feature maps $\Phi_p \colon \mathbb{S}_d \to \mathbb{R}^{\left(1+\binom{d}{2}\right)^p}$ that satisfy the relation $k^p(\sigma, \sigma') = \Phi_p(\sigma)^\top \Phi_p(\sigma')$. First, we use the feature map of the Kendall kernel to construct Φ_1; in particular, the map $\Phi_1 \colon \mathbb{S}_d \to \mathbb{R}^{1+\binom{d}{2}}$ is defined by

$$\Phi_1(\sigma)_{t_0} := 1 \quad \text{and} \quad \Phi_1(\sigma)_{t_r} := \sqrt{\binom{d}{2}^{-1}} \left(2\mathbb{1}_{\{\sigma(i_r)<\sigma(j_r)\}} - 1\right),$$

where the coordinates are indexed by the unordered pair $t_0 = \{-1, 0\}$ and the $\binom{d}{2}$ unordered pairs $t_r = \{i_r, j_r\}$ with $i_r, j_r \in [d]$ and $i_r < j_r$. We denote the set of these unordered pairs by

$$\mathcal{T} := \left\{t_0, t_1, \ldots, t_{\binom{d}{2}}\right\}. \tag{18}$$

The feature map Φ_1 clearly satisfies $k^1(\sigma, \sigma') = \Phi_1(\sigma)^\top \Phi_1(\sigma')$. Now we use the map Φ_{p-1} to construct a feature map Φ_p for $p \geq 1$. By definition, we have

$$\begin{aligned} k^p(\sigma, \sigma') &= (1 + k_\tau(\sigma, \sigma'))\,(1 + k_\tau(\sigma, \sigma'))^{p-1} = \Phi_1(\sigma)^\top \Phi_1(\sigma') \Phi_{p-1}(\sigma')^\top \Phi_{p-1}(\sigma) \\ &= \operatorname{tr}\left(\Phi_1(\sigma)^\top \Phi_1(\sigma') \Phi_{p-1}(\sigma')^\top \Phi_{p-1}(\sigma)\right) \\ &= \operatorname{tr}\left(\left(\Phi_1(\sigma)\Phi_{p-1}(\sigma)^\top\right)^\top \Phi_1(\sigma')\Phi_{p-1}(\sigma')^\top\right). \end{aligned}$$

Therefore, the polynomial kernel of degree p between σ and σ' is equal to the inner product of the matrices $\Phi_1(\sigma)\Phi_{p-1}(\sigma)^\top$ and $\Phi_1(\sigma')\Phi_{p-1}(\sigma')^\top$. By induction, we see that Φ_p can be obtained from Φ_1 by taking the outer product with itself p times, meaning that the embedding $\Phi_p : \mathbb{S}_d \to \mathbb{R}^{\left(1+\binom{d}{2}\right)^p}$ can be expressed as

$$\Phi_p(\sigma)_{s_1 s_2 \ldots s_p} = \prod_{i=1}^{p} \Phi_1(\sigma)_{s_i}, \tag{19}$$

where $s_1, s_2, \ldots, s_p$ is a sequence of elements of $\mathcal{T}$.

The following lemma is the key result that allows us to show that the three polynomial kernels of degree greater or equal than $d-1$ are characteristic.

Lemma 10. *The vectors* $\{\Phi_{d-1}(\sigma) \mid \sigma \in \mathbb{S}_d\}$ *are linearly independent.*

Since it is more involved, we defer this proof to Section C.1; here we provide some intuition for the argument. By construction, each entry of Φ_{d-1} is equal to a product of up to $d-1$ terms $2\mathbb{1}_{\{\sigma(i)<\sigma(j)\}} - 1$ times a constant. The key property that makes the result true is that the indicator functions $\mathbb{1}_{\{\sigma=\sigma_r\}}$ can be expressed as a product of $d-1$ indicator functions $\mathbb{1}_{\{\sigma(i)<\sigma(j)\}}$. For example, when $d=3$, the product $\mathbb{1}_{\{\sigma(1)<\sigma(3)\}}\mathbb{1}_{\{\sigma(3)<\sigma(2)\}}$ is equal to the indicator function of the permutation $[1,3,2]$. Moreover, the degree $d-1$ is the smallest with this property.

As mentioned previously, a universal kernel on $\mathbb{S}_d$ is also characteristic. Hence, it suffices to show that the polynomial kernel k^{d-1} is universal. Therefore, it is enough to check that the Gram matrix $M_\tau = [k^{d-1}(\sigma_i, \sigma_j)]$ is invertible, where $\sigma_1, \sigma_2, \ldots, \sigma_{d!}$ enumerate all the elements of $\mathbb{S}_d$. The Gram matrix can be written as

$$M_\tau = \begin{bmatrix} \Phi_{d-1}(\sigma_1)^\top \\ \vdots \\ \Phi_{d-1}(\sigma_{d!})^\top \end{bmatrix} \begin{bmatrix} \Phi_{d-1}(\sigma_1) & \cdots & \Phi_{d-1}(\sigma_{d!}) \end{bmatrix}$$

because $k^{d-1}(\sigma_i, \sigma_j) = \Phi_{d-1}^\top(\sigma_i)\Phi_{d-1}(\sigma_j)$. From Lemma 10, we know that the vectors Φ_{d-1} are independent, and hence the Gram matrix M_τ is full rank, which completes the proof of Theorem 6.

A.5. Proof of Theorem 5

By Bochner's theorem and the Fourier inversion theorem it suffices to show that the Mallows kernel is characteristic or universal.

We first give a direct proof that the Mallows kernel is universal. Theorem 6 shows that the ν-normalized polynomial kernel $\overline{k}^{p,\nu}$ defined by

$$\overline{k}^{p,\nu}(\sigma, \sigma') = e^{-\frac{\nu}{2}\binom{d}{2}} \left(1 + \nu \binom{d}{2} \frac{k_\tau(\sigma, \sigma')}{2p}\right)^p$$

is characteristic and universal when the degree p is greater or equal than $d-1$. Moreover, we saw that as the degree p increases to infinity, the kernel $\overline{k}^{p,\nu}$ converges to the Mallows kernel k_m^ν. Therefore, it is not surprising that the Mallows kernel is universal since it is the limit of universal kernels.

Let us now make this rough argument precise. We need to show that the Gram matrix $M_m = [k_m^\nu(\sigma_i, \sigma_j)]$ is strictly positive definite; here the permutations $\sigma_1, \sigma_2, \ldots, \sigma_{d!}$ enumerate the elements of $\mathbb{S}_d$.

Recall that the Hadamard product between two matrices A and B of the same dimensions, denoted by $A \circ B$, is formed by taking elementwise-product of the entries; we use $A^{\circ p}$ to denote the Hadamard product of the matrix A with itself p times. By Schur's theorem, the Hadamard product $A \circ B$ of any two PSD matrices is also PSD. Let $M_\tau = \frac{\nu}{2}\binom{d}{2}[k_\tau(\sigma_i, \sigma_j)]$. Performing a Taylor series expansion of the exponential function yields

$$e^{\frac{\nu}{2}\binom{d}{2}} M_m = e^{\frac{\nu}{2}\binom{d}{2}}[k_m^\nu(\sigma_i, \sigma_j)] = e^{\frac{\nu}{2}\binom{d}{2}}[e^{-\nu n_d(\sigma_i, \sigma_j)}] = \sum_{i=0}^{\infty} \frac{1}{i!} M_\tau^{\circ i},$$

where the series on the right-hand side is entry-wise absolutely convergent. For some $0 \leq \alpha_i \leq 1$, re-arranging terms yields

$$\begin{aligned} e^{\frac{\nu}{2}\binom{d}{2}} M_m &= \sum_{i=0}^{d-1} \binom{d-1}{i} \frac{1}{(d-1)^i} M_\tau^{\circ i} + \sum_{i=0}^{\infty} \alpha_i \frac{1}{i!} M_\tau^{\circ i} \\ &= \left(1 + \frac{M_\tau}{d-1}\right)^{\circ(d-1)} + \sum_{i=0}^{\infty} \alpha_i \frac{1}{i!} M_\tau^{\circ i}. \end{aligned} \tag{20}$$

The first term on the right-hand side of (20) is the Gram matrix of the ν-normalized polynomial kernel of degree $d-1$, and thus it is a strictly positive definite matrix. The second term is a positive semi-definite matrix because of Schur's theorem. Hence M_m is strictly positive definite and Theorem 5 is now proved.

For completeness, we show that the Mallows kernel is characteristic in two other ways. First of all, because of the feature embedding of the Kendall kernel, it can be viewed as the standard Gaussian kernel on $\mathbb{R}^{\binom{d}{2}}$ restricted to $2^{\binom{d}{2}}$. Then, since the Gaussian kernel is characteristic, the Mallows kernel has to be characteristic.

As yet another proof, we note that the result of Theorem 5 can be obtained via a more abstract argument, using the results of Christmann and Steinwart [24]. Given a compact metric space X and a separable Hilbert space $\mathcal{H}$, let $\Psi\colon X \to \mathcal{H}$ a continuous and injective map. The authors show that the kernel k on $X \times X$ given by

$$k(x, y) = e^{-\nu \|\Psi(x) - \Psi(y)\|_{\mathcal{H}}^2} \tag{21}$$

is universal. The symmetric group is a compact metric space and we can choose $\Psi = \Phi$, the feature map of the Kendall kernel. We can thus conclude that the kernel defined in Eq. (21) is universal and characteristic; since it equals the Mallows kernel up to constants, the claim of Theorem 5 follows.

Appendix B: Background in representation theory

In this section, we present further notions and results about the representation theory for the symmetric group. Our exposition is brief and covers only the

essential results needed in our work. For a more detailed introduction good resources include the thesis of Kondor [11] and the appendices by Huang et al. [18], with a concise summary also given by Kondor and Barbosa [10]. More detailed presentations can be found in Diaconis [12], Sagan [19], or Fulton and Harris [13], ordered according to increasing levels of abstraction.

Groups

A group $(G, \cdot)$ is a set G endowed with a multiplicative operation $\cdot\colon G \times G \to G$ such that

(a) there exists an element $e \in G$ called the identity element such that $e \cdot g = g \cdot e = g$ for all $g \in G$.
(b) $g_1 \cdot (g_2 \cdot g_3) = (g_1 \cdot g_2) \cdot g_3$ for all $g_1, g_2, g_3 \in G$.
(c) for any element $g \in G$, there exists $g^{-1} \in G$ such that $g \cdot g^{-1} = g^{-1} \cdot g = e$.

It is easy to check that $(\mathbb{R}, +)$ or $(\mathbb{R}, \cdot)$ are examples of groups. It is also straightforward to check that the set of permutations together with the operation of composition form a group, called the symmetric group. Notice that we do not require $g_1 \cdot g_2 = g_2 \cdot g_1$. A group with this property is called *commutative* or *abelian*. Abelian groups are easier to study than non-abelian ones. Unfortunately, the symmetric group is not abelian.

Equivalent representations

Two representations ρ_1 and ρ_2 are **equivalent** if they have the same dimension and if there exists an invertible matrix C such that $\rho_1(\sigma) = C^{-1}\rho_2(\sigma)C$ for all $\sigma \in \mathbb{S}_d$. In other words, two representations are equivalent if there exists a change of basis that makes one of them equal to the other. We use $\rho_1 \equiv \rho_2$ to denote the equivalence of the representations ρ_1 and ρ_2.

For any representation ρ_1, there exists an equivalent representation ρ_2 such that each matrix $\rho_2(\sigma)$ is unitary (i.e., $\rho_2(\sigma)^* \coloneqq \overline{\rho_2(\sigma)}^\top = \rho_2(\sigma)^{-1} = \rho_2(\sigma^{-1})$). Therefore, we can always assume that the representations we are working with are unitary.

Furthermore, in the case of the irreps of the symmetric group, there exist bases such that each representation ρ_λ is real, and hence orthogonal. The irreps in these bases are known as Young's orthogonal representations, and throughout this paper we work with these forms of ρ_λ.

Irreps

We already said that an irreducible representation is a representation that is not equivalent to a direct sum of representations. The symmetric group, in fact any finite group, has a finite number of pairwise inequivalent irreps. Let us consider a maximal set of pairwise inequivalent irreps. There can be multiple such sets, but they are the same up to equivalence. To be more precise, between

two maximal sets of irreps there exists a bijection such that an irrep in the first set is mapped to an equivalent irrep in the other set.

A fundamental result in representation theory states than any representation is equivalent to a direct sum of irreps. That is, each representation ρ can be decomposed into the direct of sum of some irreducible representations ρ_1, ρ_2, $\ldots$, ρ_k with some multiplicities m_1, m_2, $\ldots$, m_k:

$$\rho \equiv \bigoplus_{i=1}^{k} \bigoplus_{j=1}^{m_i} \rho_i.$$

Let us recall that the entries of each representation $\rho\colon \mathbb{S}_d \to \mathbb{C}^{d_\rho \times d_\rho}$ define d_ρ^2 functions $\sigma \mapsto \rho(\sigma)_{ij}$ on the symmetric group. The functions defined by Young's orthogonal representations form a basis for the space of functions $f\colon \mathbb{S}_d \to \mathbb{C}$. This result is important and this work exploits it extensively.

The Fourier transform

We saw that the **Fourier transform** of a function $f\colon \mathbb{S}_d \to \mathbb{C}$ is a map from representations to matrices, and it is given by

$$\widehat{f}(\rho) = \sum_{\sigma \in \mathbb{S}_d} f(\sigma)\rho(\sigma),$$

where ρ is a representation of the symmetric group.

This Fourier transform has properties similar to those of its counterpart over the real numbers. First of all, there exists a **Fourier inversion formula** and it takes the form

$$f(\sigma) = \frac{1}{d!} \sum_{\lambda \vdash d} d_\lambda \operatorname{tr}\left(\rho_\lambda(\sigma^{-1})\widehat{f}(\rho_\lambda)\right).$$

The Fourier transform on the symmetric also satisfies the **Plancherel formula**:

$$\sum_{\sigma \in \mathbb{S}_d} f(\sigma^{-1})g(\sigma) = \frac{1}{d!} \sum_{\lambda \vdash d} d_\lambda \operatorname{tr}\left(\hat{f}(\rho_\lambda)\hat{g}(\rho_\lambda)\right).$$

A third familiar property is that the Fourier transform of the convolution of two functions is the product of the Fourier transforms of the individual functions. The **convolution** of two functions f and g on $\mathbb{S}_d$ is defined by

$$f * g(\pi) = \sum_{\sigma \in \mathbb{S}_d} f(\pi\sigma^{-1})g(\sigma).$$

Bochner's Theorem

Bochner's Theorem for locally compact abelian groups states that any positive definite function on such a group can be written as the Fourier transform of

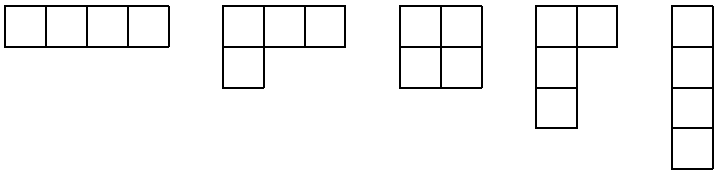
a unique measure over the dual group. In the case of the non-abelian finite symmetric group, such a result follows directly from the Fourier transform and the Fourier inversion formula. The goal of Bochner's theorem in this setting is to offer a characterization of positive definite functions and of the functions that lead to characteristic kernels.

Theorem 11 ([8, 11]). *A right-invariant function $k\colon \mathbb{S}_d \times \mathbb{S}_d \to \mathbb{R}$ defines positive definite kernel if and only if the Fourier transform $\widehat{k}$ of the associated univariate function is positive semi-definite at all irreps (i.e., $\widehat{k}(\rho_\lambda) \succeq 0$ for all $\lambda \vdash d$). Moreover, the kernel k is characteristic if and only if $\widehat{k}(\rho_\lambda)$ is positive definite for all $\lambda \vdash d$.*

Ferrer diagrams, Young tableaux, and Young tabloids

As mentioned in Section 4, it is natural to index the irreps of $\mathbb{S}_d$ by partitions λ of d. The exact correspondence is not easy to describe, but it is useful to understand how to visualize the partitions λ and the corresponding irrep ρ_λ.

The partitions $\lambda \vdash d$ are represented graphically in the form of **Ferrer's diagrams**. The diagram of a partition $\lambda = (\lambda_1, \ldots, \lambda_r)$ is formed by boxes placed in rows such that row i contains λ_i boxes. For example, the partitions of 4 are (4), $(3,1)$, $(2,2)$, $(2,1,1)$, and $(1,1,1,1)$, represented as:



In this graphical representation, a wider partition is higher in the partial ordering, while a taller partition is lower in the partial ordering.

A Ferrer diagram with the elements of the set $\{1, 2, \ldots, d\}$ in its boxes is called a **Young tableau**. Young tableaux in which the rows are viewed as sets are called **Young tabloids**. To emphasize that the rows of a Young tabloid are not ordered we drop the vertical lines in the graphical representation. For example, the Young tabloids of the partition $(2,1)$ are

$$\begin{array}{|c|c|}\hline 1 & 2 \\ \hline 3 \\ \cline{1-1}\end{array} \qquad \begin{array}{|c|c|}\hline 1 & 3 \\ \hline 2 \\ \cline{1-1}\end{array} \qquad \begin{array}{|c|c|}\hline 2 & 3 \\ \hline 1 \\ \cline{1-1}\end{array}$$

In what follows, we adopt the shorthand notation $\sigma(\{1,3\}) := \{\sigma(1), \sigma(3)\}$. When we are interested in the subset of permutations $\sigma \sim P$ that satisfy $\sigma(\{1,3\}) = \{2,5\}$, $\sigma(\{2,4\}) = \{1,4\}$ and $\sigma(\{5\}) = \{3\}$, we express this as the permutations that satisfy

$$\sigma\left(\begin{array}{|c|c|}\hline 1 & 3 \\ \hline 2 & 4 \\ \hline 5 \\ \cline{1-1}\end{array}\right) = \begin{array}{|c|c|}\hline 2 & 5 \\ \hline 1 & 4 \\ \hline 3 \\ \cline{1-1}\end{array} \tag{22}$$

Overcomplete representations and James' submodule theorem

In studying irreps or the Fourier transforms of functions it is often useful to consider reducible representations that have an easy to understand interpretation and contain copies of the irreps. We have seen in Section A.3 that the representations τ_λ play such a role. We now define these representations for a general partition λ.

Let $\{t_1\}$, $\{t_2\}$, ..., $\{t_l\}$ be an enumeration of all Young tabloids[1] of some partition $\lambda \vdash d$. The representation τ_λ takes values in $\mathbb{R}^{l \times l}$ and is defined by

$$[\tau_\lambda(\sigma)]_{ij} = \begin{cases} 1 & \text{if } \sigma(\{t_i\}) = \{t_j\} \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

We note that the Fourier transform of a probability measure P at the representation τ_λ encodes marginal probabilities:

$$[\widehat{P}(\tau_\lambda)]_{ij} = \sum_{\sigma \in \mathbb{S}_d} P(\sigma)[\tau_\lambda(\sigma)]_{ij} = P(\sigma(\{t_i\}) = \{t_j\}).$$

Therefore the Fourier transform at this representation has a concrete interpretation in the "time domain." Nonetheless, because of the Fourier inversion formula we want to understand the properties of the kernel functions at irreps. James' Submodule Theorem give a decomposition of τ_λ into irreps. We state the form of the theorem presented by Huang et al. [18].

Theorem. [James' Submodule Theorem] *There exist orthogonal matrices C_λ and integers $K_{\lambda\mu} \geq 0$ so that*

$$C_\lambda^\top \tau_\lambda(\sigma) C_\lambda = \bigoplus_{\mu \trianglerighteq \lambda} \bigoplus_{l=1}^{K_{\lambda\mu}} \rho_\mu(\sigma), \text{ for all } \sigma \in \mathbb{S}_d. \tag{24}$$

Furthermore, $K_{\lambda\lambda} = 1$ for all $\lambda \vdash d$.

The integers $K_{\lambda,\mu}$ are known as Kostka's numbers and there are methods to compute them. For example, we have already mentioned in Section A.3 that

$$\tau_{(n)} \equiv \rho_{(n)} \tag{25}$$
$$\tau_{(n-1,1)} \equiv \rho_{(n)} \oplus \rho_{(n-1,1)} \tag{26}$$
$$\tau_{(n-2,2)} \equiv \rho_{(n)} \oplus \rho_{(n-1,1)} \oplus \rho_{(n-2,2)} \tag{27}$$
$$\tau_{(n-2,1,1)} \equiv \rho_{(n)} \oplus \rho_{(n-1,1)} \oplus \rho_{(n-1,1)} \oplus \rho_{(n-2,2)} \oplus \rho_{(n-2,1,1)}. \tag{28}$$

Appendix C: Miscellaneous proofs

In this appendix, we collect the proofs of various other results.

[1] It is standard to use $\{t\}$ to denote a Young tabloids and t to denote a Young tableaux because the former are equivalence classes of the latter.

C.1. Proof of Lemma 10

Recall from Eq. (18) that for $i_r, j_r \in [d]$ with $i_r < j_r$, we use $t_r = \{i_r, j_r\}$ to denote unordered pairs with an additional $t_0 = \{-1, 0\}$ for convenience, and $\mathcal{T}$ to denote the set of all such $\binom{d}{2} + 1$ unordered pairs. In Eq. (19), the definition of the feature map $\Phi_{d-1} : \mathbb{S}_d \to \mathbb{R}^{\left(\binom{d}{2}+1\right)^{d-1}}$ implies that

$$\forall s_1, s_2, \dots s_{d-1} \in \mathcal{T}, \quad \Phi_{d-1}(\sigma)_{s_1 s_2 \dots s_{d-1}} = C_{s_1 s_2 \dots s_{d-1}} \prod_{s_r \neq t_0} (2\mathbb{1}_{\{\sigma(i_r) < \sigma(j_r)\}} - 1)$$

where $C_{s_1 s_2 \dots s_{d-1}} = \binom{d}{2}^{|\{r : s_r \neq t_0\}|/2}$ is a positive constant independent of σ, and the product is only over $s_r \neq t_0$ since $\Phi_1(\sigma)_{t_0} = 1$. We use the convention that an empty product evaluates to 1.

Now define a new feature map $\overline{\Phi}_{d-1} : \mathbb{S}_d \to \mathbb{R}^{\left(\binom{d}{2}+1\right)^{d-1}}$ as

$$\forall \sigma \in \mathbb{S}_d, \ \forall s_1, s_2, \dots s_{d-1} \in \mathcal{T}, \quad \overline{\Phi}_{d-1}(\sigma)_{s_1 s_2 \dots s_{d-1}} = \prod_{s_r \neq t_0} (2\mathbb{1}_{\{\sigma(i_r) < \sigma(j_r)\}} - 1)$$

Let C represent an invertible diagonal matrix of the constants $C_{s_1 s_2 \dots s_{d-1}}$. Then note that

$$\forall \sigma \in \mathbb{S}_d, \ \Phi_{d-1}(\sigma) = C \overline{\Phi}_{d-1}(\sigma).$$

Hence, the vectors $\{\Phi_{d-1}(\sigma)\}_{\sigma \in \mathbb{S}_d}$ are linearly independent if and only if the vectors $\{\overline{\Phi}_{d-1}(\sigma)\}_{\sigma \in \mathbb{S}_d}$ are linearly independent. We work with $\{\overline{\Phi}_{d-1}(\sigma)\}_{\sigma \in \mathbb{S}_d}$ because its entries are always ± 1.

Claim. *If $\{\alpha(\sigma)\}_{\sigma \in \mathbb{S}_d}$ are $d!$ real coefficients such that*

$$\sum_{\sigma \in \mathbb{S}_d} \alpha(\sigma) \overline{\Phi}_{d-1}(\sigma) = \mathbf{0} \ \in \ \mathbb{R}^{\left(\binom{d}{2}+1\right)^{d-1}}, \tag{29}$$

then each coefficient $\alpha(\sigma)$ is equal to zero.

In what follows, we drop repeated occurrences of t_0 when indexing the coordinates of $\overline{\Phi}_{d-1}$ without risking confusion. For example, $\overline{\Phi}_{d-1}(\sigma)_{t_2}$ means $\overline{\Phi}_{d-1}(\sigma)_{t_2 t_0 \dots t_0}$, where t_0 is repeated $d-2$ times. Now observe that $\overline{\Phi}_{d-1}(\sigma)_{t_0} = 1$ for all σ, implying that

$$\sum_{\sigma} \alpha(\sigma) = 0. \tag{30a}$$

By construction, we have $\overline{\Phi}_{d-1}(\sigma)_{t_r} = 2\mathbb{1}_{\{\sigma(i_r) < \sigma(j_r)\}} - 1$, and hence

$$\sum_{\{\sigma(i) < \sigma(j)\}} \alpha(\sigma) - \sum_{\{\sigma(i) > \sigma(j)\}} \alpha(\sigma) = 0 \qquad \text{for all sets } s = \{i, j\} \in \mathcal{T}. \tag{30b}$$

Eqs. (30a) and (30b) imply that

$$\sum_{\{\sigma(i)<\sigma(j)\}} \alpha(\sigma) = 0 \quad \text{and} \quad \sum_{\{\sigma(i)>\sigma(j)\}} \alpha(\sigma) = 0. \tag{31}$$

From now on, for a given unordered pair $s = \{i, j\} \in \mathcal{T}$, we introduce the shorthand notation $+s := \{\sigma \colon \sigma(i) < \sigma(j)\}$, with $-s$ denoting its complement. Moreover, we write several such signed unordered pairs next to each other we mean the intersection of the two sets. For example $+s_1 - s_2$ means $s_1 \cap s_2^c$.

We use induction to show that each $\alpha(\sigma)$ is zero. Assume that for some fixed integer p and for all choices of p unordered pairs $s_1, \ldots, s_p \in \mathcal{T}$ and for all possible binary signs $\epsilon_1, \ldots, \epsilon_p \in \{+1, -1\}$ the following holds:

$$\sum_{\epsilon_1 s_1 \ldots \epsilon_p s_p} \alpha(\sigma) = 0.$$

We show that this property holds for all choices of $p+1$ unordered pairs and binary signs. The base case $p = 1$ has been shown in Eq. (31).

Fix a sequence of $p+1$ distinct pairs s_1, s_2, …, and s_p, all distinct from t_0. Then, each sequence $\epsilon_1 s_1$, $\epsilon_2 s_2$, …, $\epsilon_p s_p$ can be encoded with a vector in $\{-1, +1\}^{p+1}$. For a given sign vector ϵ in $\{-1, +1\}^{p+1}$ let $\operatorname{sign}(\epsilon)$ be equal to the product of the entries of ϵ. Therefore, $\operatorname{sign}(\epsilon)$ is $+1$ if the vector ϵ contains an even number of -1 entries, and is -1 otherwise. Then, we have

$$\sum_{\sigma \in \mathbb{S}_d} \alpha(\sigma) \overline{\Phi}_{d-1}(\sigma_i)_{s_1 \ldots s_{p+1}} = 0 \implies \sum_{\epsilon \in \{-1,+1\}^{p+1}} \operatorname{sign}(\epsilon) \sum_{\epsilon_1 s_1 \ldots \epsilon_{p+1} s_{p+1}} \alpha(\sigma) = 0. \tag{32}$$

The signed pairs $-s_1$ and $+s_1$ are complements of each other. Therefore, we have

$$\sum_{-s_1 \epsilon_2 s_2 \ldots \epsilon_{p+1} s_{p+1}} \alpha(\sigma) + \sum_{+s_1 \epsilon_2 s_2 \ldots \epsilon_{p+1} s_{p+1}} \alpha(\sigma) = \sum_{\epsilon_2 s_2 \ldots \epsilon_{p+1} s_{p+1}} \alpha(\sigma).$$

This property holds for all pairs s_j, not just for s_1. Furthermore, by the induction step we know that the right hand side of the above equation equals zero. More generally, if ϵ and ξ are two sign vectors in $\{-1, +1\}^{p+1}$ that differ only in a coordinate, we have

$$\underbrace{\sum_{\epsilon_1 s_1 \epsilon_2 s_2 \ldots \epsilon_{p+1} s_{p+1}} \alpha(\sigma)}_{f(\epsilon)} + \underbrace{\sum_{\xi_1 s_1 \xi_2 s_2 \ldots \xi_{p+1} s_{p+1}} \alpha(\sigma)}_{f(\xi)} = 0. \tag{33}$$

Let G be the standard graph on the hypercube $\{-1, +1\}^{p+1}$, i.e. the graph with node set equal to $\{-1, +1\}^{p+1}$ that connects to nodes by an edge only if they differ in a single coordinate. Then, observation (33) immediately implies that for any two sign vectors ϵ and ξ at distance two of each other in the graph G,

we have $f(\epsilon) = f(\xi)$. In fact, it is straightforward that any pair of nodes ϵ and ξ that are at an even distance apart satisfy $f(\epsilon) = f(\xi)$.

It is easily checked that two nodes ϵ and ξ are at an even distance away only if $\operatorname{sign}(\epsilon) = \operatorname{sign}(\xi)$. Therefore, if $\operatorname{sign}(\epsilon) = \operatorname{sign}(\xi)$, then $f(\epsilon) = f(\xi)$. Moreover, Eq. (32) implies that

$$\sum_{\epsilon\colon\ \operatorname{sign}(\epsilon)=1} f(\epsilon) - \sum_{\epsilon\colon\ \operatorname{sign}(\epsilon)=-1} f(\epsilon) = 0.$$

We also know that $\sum_{\epsilon\in\{-1,+1\}^{p+1}} f(\epsilon) = 0$ because $\sum_{\sigma\in\mathbb{S}_d} \alpha(\sigma) = 0$. Therefore, $\sum_{\operatorname{sign}(\epsilon)=1} f(\epsilon) = 0$ and $\sum_{\operatorname{sign}(\epsilon)=-1} f(\epsilon) = 0$. But the terms inside each of these sums are equal to each other, hence $f(\epsilon) = 0$ for all $\epsilon \in \{-1,+1\}^{p+1}$. This completes the induction step.

Finally, because for any permutation σ there exists a sequence of $d-1$ unordered pairs $\mathbb{1}_{\{\sigma(i)<\sigma(j)\}}$ that uniquely determine it, for each permutation σ we can choose sets $\epsilon_1 s_1, \ldots, \epsilon_{d-1} s_{d-1}$ such that σ is the only permutation that is contained in all of them. Then, by what have proven so far, we find $\alpha(\sigma) = 0$ for all $\sigma \in \mathbb{S}_d$ and the conclusion follows.

C.2. Proving that $n_d(\sigma, \sigma')$ is right-invariant

We need to check that $n_d(\sigma, \sigma') = n_d(\sigma \circ \pi, \sigma' \circ \pi)$ for all $\pi \in \mathbb{S}_d$. By definition, we have

$$\begin{aligned}
&\sum_{i<j} \left[\mathbb{1}_{\{\sigma(i)<\sigma(j)\}}\mathbb{1}_{\{\sigma'(i)>\sigma'(j)\}} + \mathbb{1}_{\{\sigma(i)>\sigma(j)\}}\mathbb{1}_{\{\sigma'(i)<\sigma'(j)\}}\right] = \\
&= \sum_{i<j} \mathbb{1}_{\{\sigma(\pi(i))<\sigma(\pi(j))\}}\mathbb{1}_{\{\sigma'(\pi(i)))>\sigma'(\pi(j))\}} \\
&\quad + \sum_{i<j} \mathbb{1}_{\{\sigma(\pi(i))>\sigma(\pi(j))\}}\mathbb{1}_{\{\sigma'(\pi(i))<\sigma'(\pi(j))\}}
\end{aligned}$$

The permutation π just maps the sets $\{i, j\}$ bijectively to the sets $\{\nu(i), \nu(j)\}$. Since we are summing over all the pairs, it means that the two sums must be equal. By choosing $\pi = \sigma^{-1}$ we get that $n_d(\sigma, \sigma') = n_d(e, \sigma' \circ \sigma^{-1})$, where e is the identity permutation. By definition $n_d(e, \sigma' \circ \sigma^{-1}) = \mathrm{i}(\sigma' \circ \sigma^{-1})$.

C.3. MMD_k in Fourier domain

We show that for any kernel k on $\mathbb{S}_d$ the maximum mean discrepancy can satisfies the identity:

$$\mathrm{MMD}_k^2(P, Q) = \frac{1}{d!} \sum_{\lambda\vdash d} d_\lambda \operatorname{tr}\left[\left(\widehat{P}(\rho_\lambda) - \widehat{Q}(\rho_\lambda)\right)^\top \widehat{k}(\rho_\lambda)\left(\widehat{P}(\rho_\lambda) - \widehat{Q}(\rho_\lambda)\right)\right] \tag{34}$$

Let α_1, α_2 be two independent random permutations sampled according to the probability distribution P. Similarly β_1 and β_2 are independent and sampled according to Q. The Fourier inversion formula ensures that

$$k(\alpha_1, \alpha_2) = k(\alpha_1 \alpha_2^{-1}) = \frac{1}{d!} \sum_{\lambda \vdash d} d_\lambda \operatorname{tr}\left[\widehat{k}(\lambda) \rho_\lambda(\alpha_2 \alpha_1^{-1})\right] \tag{35}$$

$$= \frac{1}{d!} \sum_{\lambda \vdash d} d_\lambda \operatorname{tr}\left[\rho_\lambda(\alpha_1)^\top \widehat{k}(\lambda) \rho_\lambda(\alpha_2)\right], \tag{36}$$

where the last equality follows because the irrep ρ_λ is one of Young's orthogonal representations.

Taking expectation with respect to α_1 and α_2 yields

$$\mathbb{E}k(\alpha_1, \alpha_2) = \frac{1}{d!} \sum_{\lambda \vdash d} d_\lambda \operatorname{tr}\left[\widehat{P}(\lambda)^\top \widehat{k}(\lambda) \widehat{P}(\lambda)\right]. \tag{37}$$

In an analogous manner, we have

$$\mathbb{E}k(\beta_1, \beta_2) = \frac{1}{d!} \sum_{\lambda \vdash d} d_\lambda \operatorname{tr}\left[\widehat{Q}(\lambda)^\top \widehat{k}(\lambda) \widehat{Q}(\lambda)\right] \quad \text{and}$$

$$\mathbb{E}k(\alpha_1, \beta_1) = \frac{1}{d!} \sum_{\lambda \vdash d} d_\lambda \operatorname{tr}\left[\widehat{Q}(\lambda)^\top \widehat{k}(\lambda) \widehat{P}(\lambda)\right].$$

Given these pieces, the conclusion follows because

$$\mathrm{MMD}_k^2(P, Q) = \mathbb{E}k(\alpha_1, \alpha_2) + k(\beta_1, \beta_2) - k(\alpha_1, \beta_2) - k(\alpha_2, \beta_1).$$

In particular, see the paper Gretton et al. [17] for a proof of this last identity.

We note that the Fourier expansion (34) of the MMD_k^2 shows that the kernel k is characteristic if and only if $\hat{k}$ is strictly positive definite at all irreps.

References

[1] Brussels European Opinion Research Group. Eurobarometer 55.2 (May–June 2001), 2012.

[2] Erich L. Lehmann and Howard J.M. D'Abrera. *Nonparametrics: Statistical Methods Based on Ranks*. Springer New York, 2006. MR2279708

[3] Brian Francis, Regina Dittrich, and Reinhold Hatzinger. Modeling heterogeneity in ranked responses by nonparametric maximum likelihood: How do Europeans get their scientific knowledge? *The Annals of Applied Statistics*, 4:2181–2202, 2010. MR2829952

[4] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs the method of paired comparisons. *Biometrika*, 39:324–345, 1952. MR0070925

[5] Yunlong Jiao and Jean-Philippe Vert. The Kendall and Mallows kernels for permutations. *Proceedings of the International Conference on Machine Learning*, 32:1935–1944, 2015.

[6] George Kimeldorf and Grace Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971. MR0290013

[7] Bernhard Schölkopf and Alex J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[8] Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Bharath K Sriperumbudur. Characteristic kernels on groups and semigroups. In *Advances in Neural Information Processing Systems*, pages 473–480, 2009.

[9] Risi Imre Kondor and John Lafferty. Diffusion kernels on graphs and other discrete structures. *Proceedings of the International Conference on Machine Learning*, 19:315–322, 2002.

[10] Risi Imre Kondor and Marconi S. Barbosa. Ranking with kernels in Fourier space. *Conference on Learning Theory*, 23:451–463, 2010.

[11] Risi Imre Kondor. Group theoretical methods in machine learning. *unpublished Ph.D. dissertation, Columbia University*, 2008.

[12] Persi Diaconis. Group representations in probability and statistics. *IMS Lecture Notes-Monograph Series*, 11:i–192, 1988. MR0964069

[13] William Fulton and Joe Harris. *Representation Theory*, volume 129. Springer Science & Business Media, 1991. MR1153249

[14] Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *The Journal of Machine Learning Research*, 2:67–93, 2002. MR1883281

[15] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. MR1450938

[16] Svetlozar T. Rachev, Lev Klebanov, Stoyan V. Stoyanov, and Frank Fabozzi. *The Methods of Distances in the Theory of Probability and Statistics*. Springer Science & Business Media, 2013. MR3024835

[17] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012. MR2913716

[18] Jonathan Huang, Carlos Guestrin, and Leonidas Guibas. Fourier theoretic probabilistic inference over permutations. *Journal of Machine Learning Research*, 10:997–1070, 2009. MR2520800

[19] Bruce Sagan. *The Symmetric Group: Representations, Combinatorial Algorithms, and Symmetric Functions*, volume 203. Springer Science & Business Media, 2013. MR1824028

[20] Martin J. Wainwright. *High-dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2017.

[21] Song Xi Chen and Ying-Li Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics*, 38:808–835, 2010. MR2604697

[22] Aaditya Ramdas, Sashank J. Reddi, Barnabas Poczos, Aarti Singh, and Larry Wasserman. Adaptivity and computation-statistics tradeoffs for kernel and distance based high dimensional two sample testing. *arXiv preprint arXiv:1508.00655*, 2015.

[23] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Brucher Matthieu, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. MR2854348

[24] Andreas Christmann and Ingo Steinwart. Universal kernels on non-standard input spaces. *Advances in Neural Information Processing Systems*, pages 406–414, 2010.