Gaussian Processes indexed on the symmetric group: prediction and learning

François Bachoc, Baptiste Broto, Fabrice Gamboa, Jean-Michel Loubes

Gaussian Processes indexed on the symmetric group: prediction and learning

François Bachoc[b], Baptiste Broto[a], Fabrice Gamboa[b] and Jean-Michel Loubes[b]

[a]*CEA* and [b]*Institut de Mathématiques de Toulouse*

March 9, 2018

Abstract

In the framework of the supervised learning of a real function defined on a space $\mathcal{X}$, the so called Kriging method stands on a real Gaussian field defined on $\mathcal{X}$. The Euclidean case is well known and has been widely studied. In this paper, we explore the less classical case where $\mathcal{X}$ is the non commutative finite group of permutations. In this setting, we propose and study an harmonic analysis of the covariance operators that enables to consider Gaussian processes models and forecasting issues. Our theory is motivated by statistical ranking problems.

keywords : Gaussian processes, Ranks, Kernel Methods.

1 Introduction

The problem of ranking a set of items is a fundamental task in today's data driven world. Analyzing observations which are not quantitative variables but rankings has been often studied in social sciences. Nowadays, it has become very popular in statistical learning. This is mainly due to the generalization of the use of automatic recommendation systems. Rankings are labels that model an order over a finite set $E_n := \{1, \dots, n\}$. Hence, an observation is a set of preferences between these n points. It is thus a one to one relation σ acting from E_n onto E_n. In other words, σ lies in the finite symmetric group S_n of all permutations of E_n.

In this paper, our aim is to predict a function defined on the permutation group and for this we will use the framework of Gaussian processes indexed on this set. Actually, Gaussian process models rely on the definition of a covariance function that characterizes the correlations between values of the process at different observation points. As the notion of similarity between data points is crucial, *i.e.* close location inputs are likely to have similar target values, covariance functions are the key ingredient in using Gaussian processes for prediction. Indeed, the covariance operator contains nearness or similarity informations. In order to obtain a satisfying model one need to choose a covariance function (*i.e.* a positive definite kernel) that respects the structure of the index space of the dataset.

A large number of applications gave rise to recent researches on ranking including "ranking aggregation" ([KCS17]) , clustering rankings (see [CGJ11]) or kernels on rankings for supervised learning. Constructing kernels over the set of permutations has been tackled in several manners. In [Kon08], Kondor provides results about kernel in non-commutative finite groups and constructs "diffusion kernels" (which are positive definite) on the permutation group S_n. These diffusion kernels are based on a discrete notion of neighborliness. We remark that the kernels considered therein are very different from those considered here. Furthermore, the diffusion kernels are not in general covariance functions because

of their tricky dependency on permutations. The paper [KB10] deals with the complexity reduction of computing the kernel computation for partial ranking. Recently, [JV17] proved that the Kendall and Mallow's kernels are positive definite. [MRW$^+$16] extended this study characterizing both the feature spaces and the spectral properties associated with these two kernels.

The goal in this paper is twofold : first we define Gaussian processes indexed by permutations by providing a class of covariance kernels. They generalize previous results on the Mallow's kernel (see [JV17]). Second, we study the asymptotic properties of the maximum likelihood estimator of the parameters of the covariance function and the properties of the prediction of the Gaussian Process associated. We prove the asymptotic accuracy of the Kriging prediction under the estimated covariance parameters. We also provide simulations that illustrate the performances of the studied kernels.

The paper falls into the following parts. In Section 2 we recall generalities on the set of permutations and provide some covariance kernels. Asymptotic results on the estimation of the covariance function are presented in Section 3. Section 4 is devoted to numerical illustration. Finally, Section 5 deals with the special case of partial rankings. The proofs are postponed to the appendix.

2 Covariance model for rankings

We will use the following notations. Let S_n be the set of permutations on $E_n = \{1, \dots, n\}$. In order to define a Gaussian process and in order to provide asymptotic results, we require the process to be defined over an infinite set. For this, we will consider the space $S_\infty := \bigcup_n S_n \otimes Id_{\mathbb{N}^* \setminus E_n}$, where $Id_{\mathbb{N}^* \setminus E_n}$ is the identity operator on $\mathbb{N}^* \setminus E_n$. This corresponds to a set of observations where rankings are given on the first $\{1, \dots, n\}$ elements while leaving the other invariant.

This framework can be seen as a model to simulate long processes where it is possible to change the order of the tasks, leading to several outcomes. For example in process mining, consider that we have to collect firms from a large number of people (we assume that there is a countably infinite number of them) to process out an administrative document. There is a predefined sequential order for the document signatures, resulting in an overall time of treatment T. We call p_i the i-th person who signs the document according to this predefined order. Let us now call $Y(\sigma)$ the processing time required when the order of signatures is given by $p_{\sigma(1)}, p_{\sigma(2)}, ..., p_{\sigma(n)}, ...$ Assume that $Y - T$ is a realization of a Gaussian process with zero-mean and covariance function K_*. Our aim is to predict the time $Y(\sigma)$, for new permutations σ, for instance in the aim of finding the order resulting in the shortest processing time for the document. Another example is given by a collection of machines in a supply line that need to be tuned in order to optimize the production of a good. The machines can be tuned in different orders, each corresponding to a permutation. The objective of the model will thus be to forecast the outcome of a specific order for the machines.

Recall that we set $S_\infty := \bigcup_n S_n \otimes Id_{\mathbb{N}^* \setminus E_n}$. Furthermore, let $S_\mathbb{N}$ be the set of permutations on the integers. As we will consider increasing domains, if $\sigma \in S_n$ and if $n' > n$, we can consider σ to be in $S_{n'}$ with $\sigma(i) = i$ for all $n < i \leq n'$. With this simplification, we can write $S_\infty = \bigcup_n S_n$. Several distances can be considered on S_n. We will focus here on the three following distances (see [Dia88]). For any permutations π and σ of S_n let

- The Kendall's tau distance be defined by

$$d_\tau(\pi, \sigma) := \sum_{i<j} \mathbb{1}_{\sigma(i)>\sigma(j),\, \pi(i)<\pi(j)} + \mathbb{1}_{\sigma(i)<\sigma(j),\, \pi(i)>\pi(j)}, \tag{2.1}$$

  that is, it counts the number of pairs on which the permutations disagree in ranking.

- The Hamming distance be defined by

$$d_H(\pi,\sigma) := \sum_i \mathbb{1}_{\tau(i)\neq\sigma(i)}. \tag{2.2}$$

- The Spearman's footrule distance be defined by

$$d_S(\pi,\sigma) := \sum_i |\tau(i)-\sigma(i)|. \tag{2.3}$$

These three distance are right-invariant. That is, $\forall \pi,\ \sigma,\ \tau \in S_n$, $d(\pi,\sigma) = d(\pi\tau,\sigma\tau)$. Other right-invariant distances are discussed in [Dia88]. We extend the last distances naturally on S_∞ and obtain a countably infinite discrete space. We then extend these distances on $S_\mathbb{N}$, taking infinite sums of positive numbers and assuming that the distances can be equal to $+\infty$. For example, the Kendall's tau distance is extended to

$$d_\tau(\pi,\sigma) = \sum_{\substack{i,j\in\mathbb{N},\\ i<j}} \mathbb{1}_{\sigma(i)>\sigma(j),\ \pi(i)<\pi(j)} + \mathbb{1}_{\sigma(i)<\sigma(j),\ \pi(i)>\pi(j)}. \tag{2.4}$$

We observe that $S_\mathbb{N}$ is still a group with the composition.

We aim to define a Gaussian process indexed by permutations. Let us recall that the law of a Gaussian random process $(Y_x)_{x\in E}$ indexed by a set E is entirely characterized by its mean and covariance functions

$$M : x \mapsto \mathbb{E}(X_x)$$

and

$$K : (x,y) \mapsto \mathrm{Cov}(X_x, X_y).$$

Hence we have only to build a covariance function on $S_\mathbb{N}$.
We recall the definition of a positive definite kernel on a space E. A symmetric map $k : E\times E \to \mathbb{R}$ is called a "positive definite kernel" if for all $n\in\mathbb{N}$ and for all $(x_1,...,x_n)\in E^n$, the matrix $(K(x_i,x_j))_{i,j}$ is positive semi-definite.
In this paper, we call k a "strictly positive definite kernel" if k is symmetric and for all $n\in\mathbb{N}$ and for all $(x_1,...,x_n)\in E^n$ such that $x_i\neq x_j$ if $i\neq j$, the matrix $(K(x_i,x_j))_{i,j}$ is positive definite.

This notion is particularly interesting for S_n (and any finite set). Indeed, if k is a strictly positive definite kernel, then for any function $f : S_n \to \mathbb{R}$, there exists $(a_\sigma)_{\sigma\in S_n}$ such that:

$$f = \sum_{\sigma\in S_n} a_\sigma k(.,\sigma), \tag{2.5}$$

and k is of course an "universal kernel" (see [SFL11]). The last decomposition is no longer true neither in S_∞ nor in $S_\mathbb{N}$, but we have a result a little bit weaker than the universality of the kernel in S_∞.

Proposition 1. *If k is a strictly positive definite kernel on S_∞, then*

$$Vect\left\{\sum_{i=1}^n a_i k(.,\sigma_i),\ n\in\mathbb{N},\ a_i\in\mathbb{R},\ \sigma_i\in S_\infty\right\} \tag{2.6}$$

is dense for the pointwise convergence topology in the space of all the functions on S_∞.

Proof. Let $f : S_\infty \to \mathbb{R}$ and let f_n be the restriction of f on S_n. The kernel k is strictly definite positive on S_n so there exists $N_n\in\mathbb{N}$, $a_1^n,...,a_{N_n}^n\in\mathbb{R}$ and $\sigma_1^n...,\sigma_{N_n}^n\in S_n$ such that

$$f_n = \sum_{i=1}^{N_n} a_i^n k(.,\sigma_i^n). \tag{2.7}$$

Hence f is the pointwise limit of $(f_n)_n$. □

Corollary 1. *Let k be a strictly positive definite kernel on S_∞ and let $\mathcal{F}$ be its RKHS. Then, $\mathcal{F}$ is dense, in the pointwise convergence topology, in the space of all the functions on S_∞.*

We provide now three different covariance kernels. They share the following type

$$K_{\theta_1^*,\theta_2^*}(\sigma,\sigma') := \theta_2^* \exp\left(-\theta_1^* d(\sigma,\sigma')\right), \tag{2.8}$$

where d is one of the three distances discussed previously. More precisely, for the Kendall's tau distance, let $K^\tau_{\theta_1,\theta_2}$ be the corresponding covariance function; for the Hamming distance, let $K^H_{\theta_1,\theta_2}$ be the corresponding covariance function; and for the Spearman's footrule distance, let $K^S_{\theta_1,\theta_2}$ be the corresponding covariance function. We will write K_{θ_1,θ_2} (resp. d) for all three kernels (resp. distances). Note that when $d(\sigma,\sigma') = +\infty$, we have $K_{\theta_1,\theta_2}(\sigma,\sigma') = 0$. Note further that the right-invariance of the distances is inherited by the kernel K_{θ_1,θ_2}.

Finally, let $\theta \in (\mathbb{R}_+^*)^3$ and let us write K'_θ for the following covariance kernel

$$K'_\theta(\sigma,\sigma') := K_{\theta_1,\theta_2}(\sigma,\sigma') + \theta_3 \mathbb{1}_{\sigma=\sigma'}. \tag{2.9}$$

In our case, we have assumed that $K_{\theta_1^*,\theta_2^*}$ is a covariance function, so that K'_θ is a strictly positive definite kernel. The following theorem proves this assumption.

Theorem 1. *For all $\theta_1 > 0$ and $\theta_2 > 0$, the maps $K^\tau_{\theta_1,\theta_2}$, $K^H_{\theta_1,\theta_2}$ and $K^S_{\theta_1,\theta_2}$ are strictly positive definite kernel on S_n, on S_∞ and on $S_\mathbb{N}$.*

Corollary 2. *The kernel K'_θ is strictly positive definite on S_n, on S_∞ and on $S_\mathbb{N}$.*

3 Gaussian fields on the Symmetric group

Let us consider a Gaussian process Y indexed by $\sigma \in S_\mathbb{N}$, with zero mean and unknown covariance function K_*. A classical assumption is that the covariance function K_* belongs to a parametric set of the form

$$\{K_\theta; \theta \in \Theta\}, \tag{3.1}$$

with $\Theta \subset \mathbb{R}^p$ and where for all $\theta \in \Theta$, K_θ is a covariance function. The quantity θ is generally called the covariance parameter. In this framework, $K_* = K_{\theta^*}$ for some parameter $\theta^* \in \Theta$.

The parameter θ^* is estimated from noisy observations of the values of the Gaussian process on several inputs. Namely $(y_i = Y(\sigma_i), \sigma_i)$ for $i = 1, \ldots, n$. Actually, let us consider an independent sample of random permutations $\Sigma = (\sigma_1, \sigma_2, ..., \sigma_n) \in S_\mathbb{N}$. Assume that we observe Σ and a realization $y = (y_1, y_2, ..., y_n)^T \in \mathbb{R}^n$ of the random vector $Y = (Y(\sigma_1), Y(\sigma_2), ..., Y(\sigma_n))^T$ defined by

$$Y(\sigma_k) = Z(\sigma_k) + \varepsilon_k. \tag{3.2}$$

Here, $\varepsilon \sim \mathcal{N}(0, \theta_3^* I_n)$ is independent of Σ, and Z is a Gaussian process indexed by $S_\mathbb{N}$ independent of Σ and ε. We assume that Z is centered with covariance function $K_{\theta_1^*,\theta_2^*}$ (see (2.8) in Section 2). Thus, Y is a Gaussian process with zero mean and covariance function K'_{θ^*} defined by (2.9). The Gaussian process Y (resp. Z) is stationary in the sense that for all $\sigma_1, ..., \sigma_n \in S_\mathbb{N}$ and for all $\tau \in S_\mathbb{N}$, the finite-dimensional distribution of Y (resp. Z) at $\sigma_1, ..., \sigma_n$ is the same as the finite-dimensional distribution at $\sigma_1\tau, ..., \sigma_n\tau$.

Several techniques have been proposed for constructing an estimator $\hat{\theta} = \hat{\theta}(\sigma_1, y_1, ..., \sigma_n, y_n)$ of θ^*. Here, we shall focus on the maximum likelihood one. It is widely used in practice and has received a lot of theoretical attention. The maximum likelihood estimate is defined as

$$\widehat{\theta}_{ML} = \widehat{\theta}_n \in \arg\min_{\{\theta \in \Theta\}} L_\theta \tag{3.3}$$

with

$$L_\theta := \frac{1}{n}\ln(\det R_\theta) + \frac{1}{n} y^t R_\theta^{-1} y, \tag{3.4}$$

where $R_\theta = [K'_\theta(\sigma_i, \sigma_j)]_{1\le i,j\le n}$. We consider that $\Theta \subset \prod_{i=1}^3 [\theta_{i,min}, \theta_{i,max}]$ for some given $0 < \theta_{i,min} \le \theta_{i,max} < \infty$ $(i = 1, 2, 3)$.

When considering the asymptotic behaviour of the Maximum Likelihood Estimate, two different frameworks can be studied: fixed domain and increasing domain asymptotics ([Ste99]). Under increasing-domain asymptotics, as $n \to \infty$, the observation points $\sigma_1, ..., \sigma_n$ are such that $\min_{i\neq j} d(\sigma_i, \sigma_j)$ is lower bounded and $d(\sigma_i, \sigma_j)$ becomes large with $|i - j|$. Under fixed-domain asymptotics, the sequence (or triangular array) of observation points $(\sigma_1, ..., \sigma_n, ...)$ is dense in a fixed bounded subset. For a Gaussian field on $\mathbb{R}^d$, under increasing-domain asymptotics, the true covariance parameter θ^* can be estimated consistently by maximum likelihood. Furthermore, the maximum likelihood estimator is asymptoticly normal ([MM84, CL93, CL96, Bac14]). Moreover, prediction performed using the estimated covariance parameter $\hat{\theta}$ is asymptotically as good as the one computed with θ^* as pointed out in [Bac14]. Finally, note that in the Symmetric group, the fixed-domain framework can not be considered (contrary to the input space $\mathbb{R}^d$) since S_n is a finite space and S_∞ is a discrete space.

We will consider hereafter the increasing-domain framework. Hence, we observe values of the Gaussian process on the permutations $\Sigma = (\sigma_1, ..., \sigma_n)$ that are assumed to fulfill the following assumptions

1. Condition 1: There exists $\beta > 0$ such that $\forall i, j,\ d(\sigma_i, \sigma_j) \ge |i - j|^\beta$.
2. Condition 2: There exists $c > 0$ such that $\forall i,\ d(\sigma_i, \sigma_{i+1}) \le c$.

Such conditions are ensured for particular choices of observations $(\sigma_1, ..., \sigma_n)$ for the three different distances previously considered. For example consider the following setting.

Lemma 1. *We fix $k \in \mathbb{N}$ and we choose $\sigma_n = \tau_n c_n \in S_{k+n}$ with $\tau_n \in S_k$ a random permutation such that $(\tau_n)_n$ are independent (we do not make further assumptions on the law of τ_n). Let $c_n = (n+k \quad n+k-1 \quad ... \quad 1)$ the cycle defined by $c_n(1) = n + k$, $c_n(i) = i - 1$ if $1 < i \le n + k$ and $c_n(i) = i$ if $i > n + k$. Finally, σ_n is a permutation such that $\sigma_n(1) = n + k$, $\sigma_n(i)$ is a random variable in $[2 : k]$ if $1 < i \le k + 1$ or if $i = n + k$, $\sigma_n(i) = i - 1$ if $k + 1 < i < n + k$ and $\sigma_n(i) = i$ if $i > n + k$. The conditions are satisfied with $\beta = 1$ and $c = 1 + k(k-1)/2$ for the Kendall's tau distance, $c = 1 + k$ for the Hamming distance and $c = 2 + k$ for the Spearman's footrule distance.*

Remark 1. *If there is a $\sigma_1 \in S_N$ for $N \in \mathbb{N}$, Condition 2 ensures that all the observations $(\sigma_n)_n$ belong to S_∞. More generally, using the stationarity of the Gaussian process Y and writing σ_n instead of $\sigma_n \sigma_1^{-1}$ we can assume that all the observations belong to S_∞.*

The following theorem ensures the consistency of the estimator when the number of observations increases.

Theorem 2. *Let $\widehat{\theta}_{ML}$ be defined as in* (3.3)*, then under Conditions 1 and 2, we get*

$$\widehat{\theta}_{ML} \xrightarrow[n\to+\infty]{\mathbb{P}} \theta^*. \tag{3.5}$$

The following Lemmas are useful for the proof of Theorem 2 (and of Theorems 3 and 4 below). Their proofs are postponed to the appendix.

Lemma 2. *The eigenvalues of R_θ are lower-bounded by $\theta_{3,\min} > 0$ uniformly in n, θ and Σ.*

Lemma 3. *For all $\alpha = (\alpha_1, \alpha_2, \alpha_3) \in \mathbb{N}^3$, with $|\alpha| = \alpha_1 + \alpha_2 + \alpha_3$ and with $\partial\theta^\alpha = \partial\theta_1^{\alpha_1}\partial\theta_2^{\alpha_2}\partial\theta_3^{\alpha_3}$, the eigenvalues of $\frac{\partial^{|\alpha|} R_\theta}{\partial\theta^\alpha}$ are upper-bounded uniformly in n, θ and Σ.*

Lemma 4. *Uniformly in* Σ*,*

$$\forall \alpha > 0, \ \liminf_{n\to+\infty} \inf_{\|\theta-\theta^*\|\geq\alpha} \frac{1}{n}\sum_{i,j=1}^{n} (K'_\theta(\sigma_i,\sigma_j) - K'_{\theta^*}(\sigma_i,\sigma_j))^2 > 0. \tag{3.6}$$

Lemma 5. $\forall (\lambda_1,\lambda_2,\lambda_3) \neq (0,0,0)$*, uniformly in* Σ*,*

$$\liminf_{n\to+\infty} \frac{1}{n}\sum_{i,j=1}^{n} \left(\sum_{k=1}^{3} \lambda_i \frac{\partial}{\partial\theta_k} K'_{\theta^*}(\sigma_i,\sigma_j)\right)^2 > 0. \tag{3.7}$$

With these lemmata we are ready to prove the main asymptotic results.

Proof. of Theorem 2. We follow the proof of Theorem V.9 of [BGLV17]. We first show that for all $\epsilon > 0$, almost surely,

$$\mathbb{P}\left(\sup_\theta |(L_\theta - L_{\theta^*}) - (\mathbb{E}(L_\theta|\Sigma) - \mathbb{E}(L_{\theta^*}|\Sigma))| \geq \epsilon \middle| \Sigma\right) \to_{n\to\infty} 0. \tag{3.8}$$

We then prove that, for a fixed $a > 0$,

$$\mathbb{E}(L_\theta|\Sigma) - \mathbb{E}(L_{\theta^*}|\Sigma) \geq a\frac{1}{n}\sum_{i,j=1}^{n} (K_\theta(\sigma_i,\sigma_j) - K_{\theta^*}(\sigma_i,\sigma_j))^2. \tag{3.9}$$

We conclude since (3.8), (3.9) and Lemma 4 imply consistency. □

The following theorem provides the asymptotic normality of the estimator.

Theorem 3. *Let* M_{ML} *be the* 3×3 *matrix defined by*

$$(M_{ML})_{i,j} = \frac{1}{2n}\operatorname{Tr}\left(R_{\theta^*}^{-1}\frac{\partial R_{\theta^*}}{\partial\theta_i}R_{\theta^*}^{-1}\frac{\partial R_{\theta^*}}{\partial\theta_j}\right). \tag{3.10}$$

Then

$$\sqrt{n}M_{ML}^{\frac{1}{2}}\left(\widehat{\theta}_{ML} - \theta^*\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, I_3). \tag{3.11}$$

Furthermore,

$$0 < \liminf_{n\to\infty} \lambda_{\min}(M_{ML}) \leq \limsup_{n\to\infty} \lambda_{\max}(M_{ML}) < +\infty. \tag{3.12}$$

Proof. We proceed as in the proof of Theorem V.10 in [BGLV17]. First, we prove (3.12). We then use a proof by contradiction: we assume that (3.11) is not true. So, there exists a bounded measurable function $g:\mathbb{R}^3\to\mathbb{R}$ and $\xi > 0$ so that, up to extracting a subsequence

$$\left|\mathbb{E}\left[g\left(\sqrt{n}M_{ML}^{\frac{1}{2}}(\widehat{\theta}_{ML} - \theta^*\right)\right] - \mathbb{E}(g(U))\right| \geq \xi, \tag{3.13}$$

with $U\sim\mathcal{N}(0,I_3)$. As in [BGLV17], we prove that, extracting another subsequence, we have:

$$\sqrt{n}M_{ML}^{\frac{1}{2}}\left(\widehat{\theta}_{ML} - \theta^*\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, I_3). \tag{3.14}$$

which is in contradiction with (3.13). □

Given the maximum likelihood estimator $\widehat{\theta}_{ML}$, the value $Y(\sigma)$, for any input $\sigma \in S_{\mathbb{N}}$, can be predicted by plugging the estimated parameter in the conditional expectation (or posterior mean) expression for Gaussian processes. Hence $Y(\sigma)$ is predicted by

$$\hat{Y}_{\widehat{\theta}}(\sigma) = r^t_{\widehat{\theta}}(\sigma) R^{-1}_{\widehat{\theta}} y \tag{3.15}$$

with

$$r_{\widehat{\theta}}(\sigma) = \begin{bmatrix} K'_{\widehat{\theta}}(\sigma, \sigma_1) \\ \vdots \\ K'_{\widehat{\theta}}(\sigma, \sigma_n) \end{bmatrix}.$$

We point out that $\hat{Y}_{\widehat{\theta}}(\sigma)$ is the conditional expectation of $Y(\sigma)$ given $y_1, ..., y_n$, when assuming that Y is a centered Gaussian process with covariance function $K_{\widehat{\theta}}$.

Theorem 4.

$$\forall \sigma \in S_{\mathbb{N}}, \quad \left| \widehat{Y}_{\widehat{\theta}_{ML}}(\sigma) - \widehat{Y}_{\theta^*}(\sigma) \right| = o_{\mathbb{P}}(1). \tag{3.16}$$

Proof. We follow the same guidelines as in Theorem V.11 in [BGLV17], showing that, for $k \in \{1, 2, 3\}$

$$\left| \sup_{\theta \in \Theta} \frac{\partial}{\partial \theta_k} \hat{Y}_{\theta}(\sigma) \right| = O_{\mathbb{P}}(1) \tag{3.17}$$

□

4 Numerical illustrations

To illustrate Theorem 2, we suggest a numerical application to show that the maximum likelihood is consistent. We generated the observations suggested in Section 3 with $k = 3$. We recall that $\sigma_n = \tau_n(n + k\ n + k - 1\ ...1) \in S_{k+n}$ with $\tau_n \in S_k$ a random permutation.

Remark 2. *This choice of observations σ_n can model real cases. Recall the example given in Section 2, where $Y(\sigma)$ is the time for a document to be signed in the order $p_{\sigma(1)}, p_{\sigma(2)}, ..., p_{\sigma(n)}, ...$ To estimate θ^*, we have to observe a realization of the time Y at $\sigma_1, \sigma_2, ...$ with $\sigma_n = \tau_n(n + k\ n + k - 1\ ...1) \in S_{k+n}$ and $\tau_n \in S_k$ is a random permutation. Assume that the k first persons $p_1, ..., p_k$ are in the same office We begin to give the document to the person p_{n+k}. p_{n+k} signs the document, then, observing that he/she is the first one to sign, gives the document to one of the k first persons, who then sign in a random order.*

To highlight the dependency with n, we write $\widehat{\theta}_n := \widehat{\theta}_{ML}$ the maximum likelihood estimator for n observations. For each value of n, we estimate the probability $\mathbb{P}(\|\widehat{\theta}_n - \theta^*\| > \varepsilon)$ using a Monte-Carlo method and a sample of 1000 values of $\mathbb{1}_{\|\widehat{\theta}_n - \theta^*\| > \varepsilon}$. Figure 1 depicts these estimates for $\varepsilon = 0.5$, $\theta^* = (0.1, 0.8, 0.3)$ and $\Theta = [0.02, 2] \times [0.3, 2] \times [0.1, 1]$.

In Figure 2, we display the density of the coordinates of the maximum likelihood estimator for different values of n (20, 60 and 150). These densities have been estimated with a 1000 sample of the maximum likelihood estimator. We observe that the densities can be far from the true parameter for $n = 20$ or $n = 60$ but are quite close to it for $n = 150$. We can see that for $n = 150$, the Kendall's tau distance seems to give better estimates of θ^*_3. However, the computation time of the distance matrix is much longer with the Kendall's tau distance than with the other distances.

In Figure 3, we display estimates of the probability that the absolute value of the prediction of $Y(\sigma)$ given in (3.15) with the parameter $\hat{\theta}_{ML}$ minus the prediction of $Y(\sigma)$ with the parameter θ^* is greater than 0.3. Theorem 4 ensures us that this probability converges to 0 when $n \to +\infty$.

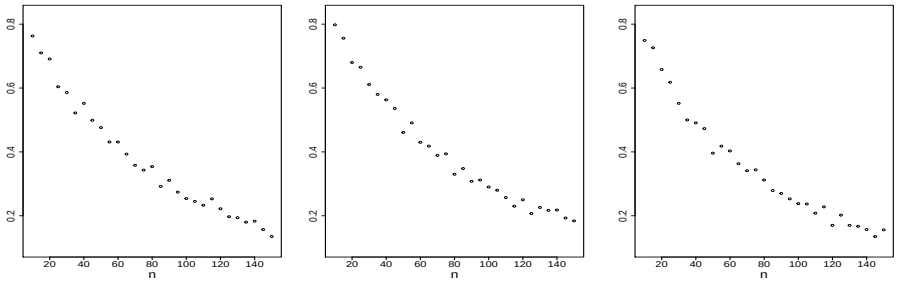

Figure 1: Estimates of $\mathbb{P}(\|\widehat{\theta}_n - \theta^*\| > 0.5)$ for different values of n, the number of observations, with $\theta^* = (0.1, 0.8, 0.3)$ and Kendall's tau distance, the Hamming distance and the Spearman's footrule distance from left to right.

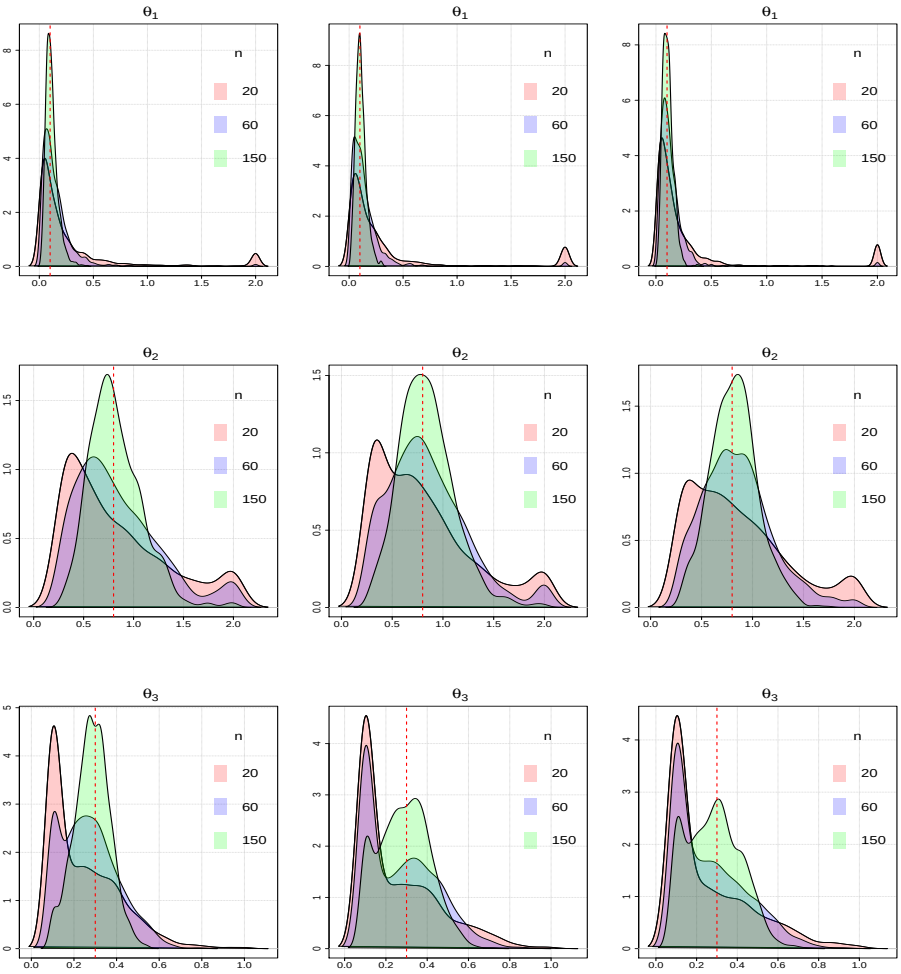

Figure 2: Density of the coordinates of $\widehat{\theta}_n$ for the number of observations $n = 20$ (in red), $n = 60$ (in blue), $n = 150$ (in green) with $\theta^* = (0.1, 0.8, 0.3)$ (represented by the red vertical line). We used the Kendall's tau distance, the Hamming distance and the Spearman's footrule distance from left to right.

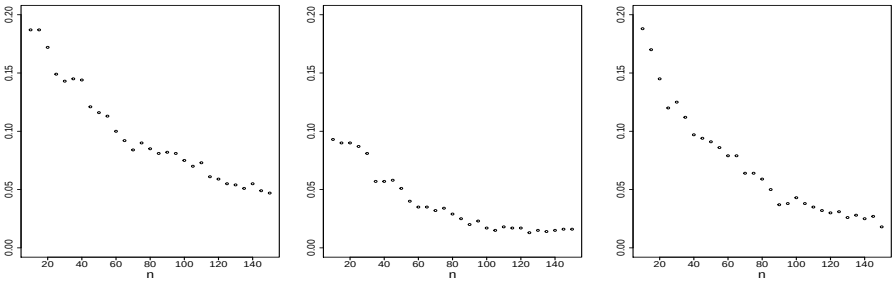

Figure 3: Estimates of $\mathbb{P}\left(\left|\widehat{Y}_{\widehat{\theta}_{ML}}(\sigma) - \widehat{Y}_{\theta^*}(\sigma)\right| > 0.3\right)$ for different values of n, the number of observations, with $\theta^* = (0.1, 0.8, 0.3)$, $\sigma = (1\ 4\ 6)$ and the Kendall's tau distance, the Hamming distance and the Spearman's footrule distance from left to right.

5 Partial Rankings

5.1 Introduction

In many situations, when n is large, preferences are not given for all points $i \in \{1, \dots, n\}$ but only for a small number k of points $\{i_1, \dots, i_k\}$. This situation occurs often in social science. When considering statistical models which analyzes the behavior of human decision in a consumer behaviour, partial rankings are often considered. Actually, given k objects governed by a large number n of variables, how can we model the decision to choose one object rather than another ? Objects are described by a set of quantitative variables $X = (X^1, \dots, X^n)' \in \mathbb{R}^n$. These variables are representative of a specific property of each object. For instance when buying a bike one may be interested in the weight, the price, the number of velocities, the height, or any other quantitative or qualitative descriptors. Each consumer when confronted to the choice of a product, chooses to give more importance to certain variables while discarding others. The consumer selects a small number of variables (features) that are essential in his choice, ranks these variables according to its preferences while the others play little importance.

In the general framework, we have a finite set $X = \{x_1, ..., x_n\}$. A partial ranking aims at giving an order of preference between different elements of X. A partial ranking R is a statement of the form

$$X_1 \succ X_2 \succ ... \succ X_m, \tag{5.1}$$

where $X_1, ..., X_m$ are disjoint set of $X = \{x_1, x_2, ..., x_n\}$. This partial ranking means that any element of X_j if preferred to any element of X_{i+1}. We can associate to the partial ranking R the set E_R of S_n defined by

$$E_R := \{\sigma \in S_n,\ \forall (x_{i_1}, ..., x_{i_k}) \in X_1 \times ... \times X_m,\ \sigma(i_1) < \sigma(i_2) < ... < \sigma(i_m)\}. \tag{5.2}$$

Remark 3. *In [KB10] and [JV17], the set E_R is defined by the set of the permutations σ such that $\forall (x_{i_1}, ..., x_{i_k}) \in X_1 \times ... \times X_m,\ \sigma(i_1) > \sigma(i_2) > ... > \sigma(i_k)$. They chose this definition to simplify their computations but in this way the ranking mapped to $\{\sigma\}$ is*

$$x_{\sigma^{-1}(i_n)} \succ ... \succ x_{\sigma^{-1}(i_1)}.$$

The definition (5.2) *seems to be more natural because we map $\{\sigma\}$ to the ranking*

$$x_{\sigma^{-1}(i_1)} \succ ... \succ x_{\sigma^{-1}(i_n)}.$$

The first natural way to extend a positive definite kernel $K : S_n \times S_n \to \mathbb{R}$ on the partial rankings (see [KB10], [JV17],...) is letting

$$K(R, R') := \frac{1}{|E_R||E_{R'}|} \sum_{\sigma \in E_R} \sum_{\sigma' \in E_{R'}} K(\sigma, \sigma'). \tag{5.3}$$

If K is a positive definite kernel on permutations, then K defined by (5.3) is a positive definite kernel on partial ranking ([Hau99]). We also can see this saying that if $R_1, ..., R_N$ are partial rankings and if $(a_1, ..., a_N) \neq 0$, then

$$\sum_{i,j=1}^{n} a_i a_j K(R_i, R_j) = \sum_{\sigma, \sigma' \in S_n} b_\sigma b_{\sigma'} K(\sigma, \sigma'), \tag{5.4}$$

letting

$$b_\sigma := \sum_{i,\ \sigma \in R_i} \frac{a_i}{|E_{R_i}|}. \tag{5.5}$$

Remark 4. *The values of $K(R, R)$ depends on R. It can be very closed to 0, that means for a Gaussian process Y indexed by the partial rankings that the value $Y(R)$ is almost constant. To circumvent this problem, we can define a new kernel*

$$K^{new}(R, R') := \frac{1}{\sqrt{K(R, R)K(R', R')}} K(R, R'). \tag{5.6}$$

The computation of this kernel seems to be very long because we have to sum over $|E_R||E_{R'}|$ permutations. In the following, we aim to reduce this computation. We focus especially on the following kernel on S_n:

$$K(\sigma, \sigma') := e^{-\nu d(\sigma, \sigma')}, \tag{5.7}$$

where d if the Kendall's tau distance, the Hamming distance or the Spearman's footrule distance. These kernels are interesting for two reasons: they are strictly positive definite and they are easy to interpret (more than a kernel defined by a matrix exponential).

5.2 Direct computations

The first idea is to simplify the expression of (5.3). However, this does not seem to be a simple task, that is why we take a particular framework. In this section, we assume that all the items are ranked, i.e. $(X_j)_{1 \leq j \leq m}$ is a partition of $[1 : n]$. Let $\gamma_j := |X_j|$ and $\gamma = (\gamma_j)_{1 \leq j \leq m}$ is a partition of n. This computation has always been done in [LM08] for the Kendall's tau distance. Let us sum up the result that interests us in the following proposition.

Proposition 2. *[LM08]*
Let γ be a partition of n. For all $j \in [1 : m]$, let $g_j := \sum_{l=1}^{j-1} \gamma_l$ and let

$$\begin{aligned} a_j^\gamma &:= |\{(s,t),\ s < t,\ g_j + 1 \leq \tau(t) < \tau(s) \leq g_{j+1}\}| \\ b_{jl}^\gamma(\tau) &:= |\{(s,t),\ s < t,\ g_j + 1 \leq \tau(t) \leq g_{j+1} < g_l + 1 \leq \tau(s) \leq g_{l+1}\}| \end{aligned}$$

Then, if $E_{R_i} = S_{\gamma^i} \pi_i$ for $i = 1, 2$,

$$K(R_1, R_2) = \frac{1}{|S_{\gamma^1}||S_{\gamma^2}|} \left(\sum_{\sigma \in \pi_1 \pi_2^{-1} S_{\gamma^2}} e^{-\nu \sum_{1=j<l}^{m} b_{jl}^{\gamma^1}(\sigma)} \right) \left(\prod_{s=1}^{m} \prod_{r=1}^{\gamma_s^1} \sum_{h=0}^{r} e^{-\nu h} \right) \tag{5.8}$$

Now, we do the same work with the Hamming distance. Before, we need to introduce a new notation.

Definition 1. *We define*

$$c_{n,d} := |\{\sigma \in S_n,\ d_H(\sigma, id) = d\}| = \frac{n!}{(n-d)!} \sum_{k=0}^{d} \frac{(-1)^k}{k!}, \tag{5.9}$$

the number of permutations of n elements which move exactly d elements.

Now we give a proposition similar than Proposition 2 with the Hamming distance.

Proposition 3. *Let us define*

$$\begin{aligned} a_j^\gamma &:= |\{i \in [1:n],\ i \neq \tau(i),\ (i, \tau(i)) \in [g_j + 1 : g_{j+1}]\}| \\ b_{jl}^\gamma(\tau) &:= |\{i \in [1:n],\ l \neq j,\ i \in [g_j + 1 : g_{j+1}],\ \tau(i) \in [g_l + 1 : g_{l+1}]\}| \end{aligned}$$

Then,

$$K(R_1, R_2) = \frac{1}{|S_{\gamma^1}||S_{\gamma^2}|} \left(\sum_{\sigma \in \pi_1 \pi_2^{-1} S_{\gamma^2}} e^{-\nu \sum_{j,l=1}^m b_{jl}^{\gamma^1}(\sigma)} \right) \left(\prod_{s=1}^{m} \sum_{h=0}^{\gamma_s^1} c_{\gamma^1, h} e^{-\nu h} \right). \tag{5.10}$$

5.3 Fourier Transform of the kernel on partial ranking

5.3.1 Notations

In this section, we use the usual kernel on partial ranking defined by

$$K(R, R') := \frac{1}{|E_R||E_{R'}|} \sum_{\sigma \in E_R} \sum_{\sigma' \in E_{R'}} K(\sigma, \sigma'). \tag{5.11}$$

We assume that the kernel K on the set of permutation is right-invariant and we write $k(\sigma) := K(\sigma, id)$. We extend the work of [KB10]. We compute the Fourier transform for general partial rankings, i.e. statement R of the form

$$X_1 \succ X_2 \succ ... \succ X_m \tag{5.12}$$

Let $\gamma = (\gamma_1, ..., \gamma_m)$ be the size of the $(X_j)_{1 \leq j \leq m}$, let k be the sum of the γ_j and let $\tilde{\gamma}$ be the partition of n defined by $\tilde{\gamma} := (\gamma_1, \gamma_2, ..., \gamma_m, n-k)$. Let Π_k^n be the set of interleaving of $[1:k]$ with $[k+1:n]$

$$\begin{aligned} \sigma(i) \leq \sigma(j) \text{ si } i < j \leq k \\ \sigma(i) \leq \sigma(j) \text{ si } k+1 \leq i < j \end{aligned}$$

Then, writing $\Pi_n^n = \{id\}$, we have (as in [KB10] but generalized for all partial ranking)

$$E_R := \Pi_k^n S_{\tilde{\gamma}} \pi_R, \tag{5.13}$$

where $\pi_R \in S_n$ is such that $\pi(X_j) = \left[1 + \sum_{l=1}^{j-1} \gamma_l : \sum_{l=1}^{j} \gamma_l\right]$. Finaly, let us write $E_\gamma := \Pi_k^n S_{\tilde{\gamma}}$.

5.3.2 Reduction of number of terms

We just generalize the works of [KB10] for general partial rankings. Let R_i ($i = 1, 2$) be the partial rankings defined by

$$E_{R_i} = E_{\gamma^i} \pi_{R_i}. \tag{5.14}$$

As in [KB10], we identify a set A of permutation with the function of S_n which associates to σ the number 1 if $\sigma \in A$ and 0 otherwise. Proposition 6 of [KB10] gives

$$K(R_1, R_2) = \frac{1}{n! |E_{R_1}||E_{R_2}|} \sum_{\lambda \vdash n} d_\lambda Tr\left(\widehat{E_{R_1}}(\rho_\lambda)^* \, \widehat{k}(\rho_\lambda) \widehat{E_{R_2}}(\rho_\lambda) \right). \tag{5.15}$$

The next proposition (which generalizes Proposition 8 of [KB10]) show how the sum in the previous equation can be reduced to a lower number of terms.

Proposition 4. *Let $c := \max(\gamma_1^1, ..., \gamma_{m_1}^1, n - k_1, \gamma_1^2, ..., \gamma_{m_2}^2, n - k_2)$ and let Λ_c^n be the set of Young's diagrams of n boxes with at least c boxes in their first row. Then*

$$K(R_1, R_2) = \frac{1}{n! |E_{R_1}| |E_{R_2}|} \sum_{\lambda \in \Lambda_c^n} d_\lambda Tr\left(\widehat{E_{R_1}}(\rho_\lambda)^* \, \widehat{k}(\rho_\lambda) \widehat{E_{R_2}}(\rho_\lambda) \right). \tag{5.16}$$

5.3.3 Reduction of each remaining term

Here, we assume that the partial rankings for which we want to compute the kernel $K(R_1, R_2)$ have always the same forms γ. For example, assume that all these partial rankings are top-k lists and partial rankings of the form $x_{i_1} \succ x_{i_2} \succ ... \succ x_{i_k}$. In this case, we just have $\gamma = (1, 1, ..., 1, n - k)$ and $\gamma = (1, 1, ..., 1)$. The next proposition show that the computation of (5.16) can still be reduced.

Proposition 5. *Let R_i ($i = 1, 2$) be the partial rankings defined by*

$$E_{R_i} = E_{\gamma^i} \pi_i. \tag{5.17}$$

Then

$$K(R_1, R_2) = \frac{1}{n!} \sum_{\lambda \in \Lambda_c^n} d_\lambda Tr\left(\rho_\lambda(\pi_2 \pi_1^{-1}) q_\lambda(E_{\gamma^1}) \widehat{k}(\rho_\lambda) p_\lambda(E_{\gamma^2}) \right), \tag{5.18}$$

where

$$\begin{aligned} q_\lambda(E_\gamma) &:= \frac{1}{|E_\gamma|} \sum_{\sigma \in E_\gamma} \rho_\lambda(\sigma^{-1}), \\ p_\lambda(E_\gamma) &:= \frac{1}{|E_\gamma|} \sum_{\sigma \in E_\gamma} \rho_\lambda(\sigma). \end{aligned}$$

Corollary 3. *If all the items are ranked, i.e. $k = n$, then we have*

$$K(R_1, R_2) = \frac{1}{n!} \sum_{\lambda \in \Lambda_c^n} d_\lambda Tr\left(\rho_\lambda(\pi_2 \pi_1^{-1})^T p_\lambda(S_{\tilde{\gamma}^1}) \widehat{k}(\rho_\lambda) p_\lambda(S_{\tilde{\gamma}^2}) \right), \tag{5.19}$$

where $p_\lambda(S_{\tilde{\gamma}})$ is a projector on $W_\lambda^\gamma = \{x \in \mathbb{C}^{d_\lambda}, \ \forall \sigma \in S_{\tilde{\gamma}}, \ \rho_\lambda(\sigma) x = x\}$.

5.4 A new kernel on partial ranking

However, $|E_R|$ and $|E_{R'}|$ are often large numbers. The computation of the kernel on partial rankings can be very long.
That is why we suggest another manner to extend the kernels $K\theta_1, \theta_2$ and K'_θ on partial rankings. If d is the Kendall's tau distance, the Hamming distance or the Spearman's footrule distance, we define the measure of dissimilarity d_{avg} as the mean of distances $d(\sigma, \sigma')$ for $\sigma \in E$ and $\sigma' \in E'$:

$$d_{avg}(R, R') := \frac{1}{|E_R||E_{R'}|} \sum_{\sigma \in E_R} \sum_{\sigma \in E_{R'}} d(\sigma, \sigma'). \tag{5.20}$$

Then, we define

$$K_{\theta_1,\theta_2}(R, R') := \theta_2 \exp(-\theta_1 d_{avg}(R, R')), \tag{5.21}$$

and

$$K'_\theta(R, R') := \theta_2 \exp(-\theta_1 d_{avg}(R, R')) + \theta_3 \mathbb{1}_{R=R'}. \tag{5.22}$$

Proposition 6. *K_{θ_1,θ_2} is a positive definite kernel on the set of partial rankings of S_n, and K'_θ is a strictly positive definite kernel.*

The computation of this new kernel on partial rankings is still very long: we have to sum $|E_R||E_{R'}|$ elements. However, this form can be easily reduced. Let us take the example of top-k partial rankings. A top-k partial ranking (or a top-k list) is a partial ranking of the form

$$x_{i_1} \succ x_{i_2} \succ ... \succ x_{i_k} \succ X_{rest}. \tag{5.23}$$

Let us write $I = (i_1, ..., i_k)$ this top-k partial ranking. The following proposition show how to reduce the computation of d_{avg} (and so of the kernel of partial rankings) for the fixed domain framework.

Proposition 7. *Let I and I' be partial rankings. Let the set $[1:k] = \{c_1, ..., c_p\} \sqcup \{u_1, ..., u_r\}$ such that $i_{c_j} \in I \setminus I'$ and $i_{u_j} \in I \cap I'$. Similarly, let $[1:k] = \{c'_1, ..., c'_p\} \sqcup \{u'_1, ..., u'_r\}$ such that $i'_{c'_j} \in I' \setminus I$ and $i'_{u'_j} \in I' \cap I$. We order the integers c_j and c'_j so that $i_{c_j} < i_{c_{j+1}}$ and $i'_{c'_j} < i'_{c'_{j+1}}$. Remark that $i_{c_j} = i'_{c'_j}$. Let us write $n' := n - k - 1$ and $m := n - |I \cup I'|$. Then*

$$d_{\tau,avg}(I, I') = \sum_{\{j,j'\} \subset [1:p]} \mathbb{1}_{(c_j < c_{j'}, c'_j > c'_{j'}) \text{ or } (c_j > c_{j'}, c'_j < c'_{j'})} + r(2k+1-r) - \sum_{j=1}^{r} (u_j + u'_j) + r^2 + \binom{n-k}{2}, \tag{5.24}$$

$$d_{H,avg}(I, I') = \sum_{j=1}^{p} \mathbb{1}_{c_j \neq c'_j} + m\frac{n-k-1}{n-k} + 2r, \tag{5.25}$$

and

$$d_{S,avg}(I, I') = \sum_{j=1}^{p} |c_j - c'_j| + r(n+k+1) - \sum_{j=1}^{r} (u_j + u'_j) + \frac{2m}{n'} \left(\frac{n'+1}{2} - \frac{1}{3}n'^2 - \frac{1}{2}n' - \frac{1}{6} \right). \tag{5.26}$$

Remark 5. *These computations show that $K(I, I)$ can be easily computed. We have*

$$K_\tau(I, I) = \theta_2 \exp\left(-\theta_1 \binom{n-k}{2} \right) \tag{5.27}$$

$$K_H(I, I) = \theta_2 \exp\left(-\theta_1 (n-k-1) \right) \tag{5.28}$$

$$K_H(I, I) = \theta_2 \exp\left(-2\theta_1 \frac{n-k}{n-k-1} \left(\frac{n-k}{2} - \frac{(n-k-1)^2}{3} - \frac{n-k-1}{2} - \frac{1}{6} \right) \right). \tag{5.29}$$

The computation of K^{new} defined by (5.6) *become easy for the top k lists.*

For the Hamming distance, we have a more general result for all partial ranking such that all items are ranked, with same sizes (i.e. $\gamma^1 = \gamma^2$). We take the previous notations.

Proposition 8.

$$d_{H,avg}(S_\gamma \pi_1, S_\gamma \pi_2) = |\{i,\ \Gamma(\pi_1(i)) \neq \Gamma(\pi_2(i))\}| + \sum_{j=1}^{m} \frac{\gamma_j}{n}(\gamma_j - 1) \tag{5.30}$$

We can also find in the proof of Theorem 2 of [JV17] how to compute $d_{\tau,avg}(R, R')$ for R and R' partial rankings of the form

$$x_{i_1} \succ x_{i_2} \succ ... \succ x_{i_k}. \tag{5.31}$$

Remark 6. *The measure of dissimilarity between partial rankings d_{avg} depends on n, contrary to the distance between total ranking (or permutations). For example, let I and I' be two tok-k lists of $[1 : n_1]$. If $n_2 > n_2$, we can assume that I and I' are top-k lists of $[1 : n_2]$. However, we can see in Proposition 7 that $d_{avg,n_1}(I, I') \neq d_{avg,n_2}(I, I')$.*

6 Appendix

6.1 Proofs of main Results

Proof of Theorem 1

Proof. First part: we show that the map K_{θ_1,θ_2} is a strictly positive definite kernel on S_n. It suffices to prove that, if $\nu > 0$, the map K defined by

$$K(\sigma, \sigma') := e^{-\nu d(\sigma,\sigma')} \tag{6.1}$$

is a strictly positive definite kernel. It is already shown in Theorem 5 of [MRW$^+$16] for the Kendall's tau distance. Let us prove it for the Hamming distance. We follow the proof of Theorem 5 in [MRW$^+$16]. For simplicity, we keep the same notations for a map $S_n \times S_n \to \mathbb{R}$ and a matrix indexed by the permutations of S_n. Let

$$D(\sigma, \sigma') := n - d_H(\sigma, \sigma') = \sum_{i=1}^{n} \mathbb{1}_{\sigma(i)=\sigma(j)}. \tag{6.2}$$

We see that if we define the vector $\Phi \in \mathbb{R}^{n^2}$ by

$$\Phi(\sigma) := (\mathbb{1}_{\sigma(i)=j})_{i,j}, \tag{6.3}$$

we have

$$D(\sigma, \sigma') = \Phi(\sigma)^T \Phi(\sigma'). \tag{6.4}$$

Hence D is a Grammian matrix and thus a positive semi-definite matrix.
Let us write $M := e^{\nu n} K$ seen as a matrix indexed by the elements of S_n. We use the Hadamard product $A \circ B$ of the matrix A and B, defined by the element-wise product. As in [MRW$^+$16], we have

$$\begin{aligned} M &= \sum_{i=0}^{n} \frac{\nu^i}{i!} D^{\circ i} + \sum_{i \geq n+1} \frac{\nu^i}{i!} D^{\circ i} \\ &= \left(1 + \frac{\nu}{n} D\right)^{\circ n} + \sum_{i \geq 0} \alpha_i \nu^i D^{\circ i} \end{aligned}$$

with

$$\alpha_i := \begin{cases} \frac{(d-1)^i-(d-1)(d-2)...(d-i)}{(d-1)^i} > 0 & \text{if } i \leq n, \\ \frac{1}{i!} > 0 & \text{if } i \geq n+1. \end{cases}$$

As in [MRW$^+$16], the second term is a positive semi-definite matrix because of Schur's theorem. We have to prove that the first term is a positive definite matrix. For this, we will show that the first term is a Gramian matrix of independent vectors.

Let us write $\Phi_1(\sigma) \in \mathbb{R}^{1+n^2}$ the vector defined by

$$\Phi_1(\sigma)_{c_0} = 1 \quad \Phi_1(\sigma)_{c_r} = \sqrt{\tfrac{\nu}{n}} \mathbb{1}_{\sigma(i_r)=j_r}, \tag{6.5}$$

where the coordinates are indexed by the ordered pair $c_0 = (-1, 0)$ and the n^2 ordered pairs $c_r = (i_r, j_r) \in [1:n]^2$. We see that $(1 + \frac{\nu}{n} D)$ is the Grammian matrix of $\{\Phi_1(\sigma),\ \sigma \in S_n\}$. As in proof of Theorem 6 in [MRW$^+$16], we can show by induction on p that if we define $\Phi_p : S_n \to \mathbb{R}^{(1+n^2)^p}$ by

$$\Phi_p(\sigma)_{c_1 c_2 ... c_p} = \prod_{i=1}^{p} \Phi_1(\sigma)_{c_i}, \tag{6.6}$$

then $(1 + \frac{\nu}{n} D)^{\circ p}$ is the Grammian matrix of $\{\Phi_p(\sigma),\ \sigma \in S_n\}$. We use this results for $p = n$. Finally we have to show that the vectors $\{\Phi_n(\sigma), \sigma \in S_n\}$ are linearly independent. We then will have that $(1 + \frac{\nu}{n} D)^{\circ n}$ is the Grammian matrix of independent vectors so it is a definite positive matrix and we may conclude that the claim holds on S_n and so on S_∞.

Assume that their exists $(a_\sigma)_{\sigma \in S_n}$ such that

$$\sum_{\sigma \in S_n} \alpha_\sigma \Phi_n(\sigma) = 0 \quad \in \mathbb{R}^{\left(1+n^2\right)^n}. \tag{6.7}$$

Let $\sigma_0 \in S_n$. Let us write $j_i := \sigma_0(i)$ and c_i the ordered pair (i, j_i). We look at the coordinate $c_1, ..., c_n$ of the equation (6.7), we have

$$\begin{aligned} 0 &= \sum_{\sigma \in S_n} \alpha_\sigma \Phi_{n-1}(\sigma)_{c_1,...,c_n} \\ &= \sum_{\sigma \in S_n} \alpha_\sigma \left(\frac{\nu}{n}\right)^{\frac{n}{2}} \mathbb{1}_{\sigma(1)=i_1,...,\sigma(n)=i_n} \\ &= \alpha_{\sigma_0} \left(\frac{\nu}{n}\right)^{\frac{n}{2}}. \end{aligned}$$

Thus, $\alpha_{\sigma_0} = 0$, and that is true for all permutation σ_0. We have shown that $K^H_{\theta_1,\theta_2}$ is a strictly positive definite kernel on S_n.

Let us prove it now for $K^S_{\theta_1,\theta_2}$. We follow the same idea. Let

$$D(\sigma, \sigma') := n^2 - d_S = \sum_{i=1}^{n} \min(\sigma(i), \sigma'(i)) + n - \max(\sigma(i), \sigma'(i)),$$

$$\Phi(\sigma)_{i,j,1} := \begin{cases} 1 & \text{if } j \leq \sigma(i) \\ 0 & \text{otherwise,} \end{cases} \qquad \Phi(\sigma)_{i,j,2} := \begin{cases} 0 & \text{if } j < \sigma(i) \\ 1 & \text{otherwise.} \end{cases}$$

and $M := e^{\nu n^2} K$. We have

$$M = \left(1 + \frac{\nu}{n} D\right)^{\circ n} + \sum_{i \geq 0} \alpha_i \nu^i D^{\circ i} \tag{6.8}$$

and it remains to show that the first matrix is positive definite. Let $\Phi_1(\sigma)_c := 1$ if $c = (0,0,0)$ and $\Phi_1(\sigma)_c := \sqrt{\frac{\nu}{n}}\Phi(\sigma)_c$ if $c \in [1:n] \times [1:n] \times \{1,2\}$. As for the Hamming distance, we define Φ_n by induction and we want to prove that the vectors $\{\Phi_n(\sigma), \sigma \in S_n\}$ are linearly independent. Assume that we have (6.7). Let $\sigma_0 \in S_n$ and choose $c_i := (i, \sigma(i), 1)$. Then $\alpha_{\sigma_0} = 0$.

Second part: let us prove now that the claim is also true on $S_{\mathbb{N}}$. We will use the following lemma.

Lemma 6. *Let $\sigma_1, \sigma_1 \in S_{\mathbb{N}}$. Then,*

$$d(\sigma_1, \sigma_2) < +\infty \iff \exists n,\ \sigma_2\sigma_1^{-1} \in S_n. \tag{6.9}$$

Proof. It is obvious for the Hamming distance and the Spearman's footrule distance. Let us prove it for the Kendall's tau distance. Assume that $d_\tau(id, \sigma) < +\infty$. Let us write $N := \max\{j,\ \exists i < j,\ \sigma(i) > \sigma(j)\}$.

Let us prove that σ is the identity on $[N+1 : +\infty[$. By contradiction, assume that $\exists n_1 \leq N,\ \sigma(n_1) \geq N+1$. Then there exists $n_2 \geq N+1,\ \sigma(n_1) \leq N$. Thus $n_1 < n_2$, but $\sigma(n_1) > \sigma(n_2)$, that contradicts the maximality of N.

Thus, σ is an increasing permutation on $[N+1 : +\infty[$, so it is the identity on this set. □

In order to prove that K_{θ_1,θ_2} is strictly positive definite on $S_{\mathbb{N}}$, the idea is to boil down to S_n using the previous lemma and using the positivity on S_n.

Let $\sim$ be the equivalence relation defined by: $i \sim j \ \Leftrightarrow\ d(i,j) < +\infty$.

Let $(\sigma_1, ..., \sigma_n) \in \mathcal{S}_{\mathbb{N}}$ and let $(a_1, ..., a_n) \neq 0$. Let $C_1, ..., C_K$ be the equivalence classes formed by $\{\sigma_1, ..., \sigma_n\}$. Then

$$\begin{aligned}\sum_{i,j} a_i a_j K_{\theta_1,\theta_2}(\sigma_i, \sigma_j) &= \sum_{k=1}^{K} \sum_{(i,j)\in C_k^2} a_i a_j K_{\theta_1,\theta_2}(\sigma_i, \sigma_j) \\ &= \sum_{k=1}^{K} \sum_{(i,j)\in C_k^2} a_i a_j K_{\theta_1,\theta_2}(\sigma_i\tau_k, \sigma_j\tau_k)\end{aligned}$$

where $\tau_k \in S_{\mathbb{N}}$ such that $\exists n_k,\ \forall i \in C_k,\ \tau_k\sigma_i \in S_{n_k}$ (we can choose for example $\tau_k = \sigma_{i_0}^{-1}$ with i_0 any element of C_k). We know that the kernel K_{θ_1,θ_2} is strictly positive definite on S_{n_k}, so all the terms of the previous sum over k are non-negative and at least one is positive. □

6.2 Proofs of Lemmata

Proof of Lemma 3.

Proof. It is easy to prove when $\alpha_1 = \alpha_2 = 0$. Indeed:

1. If $\alpha_3 = 0$, then $\lambda_{\max}\left(K'_\theta(\sigma_i, \sigma_j)_{i,j}\right) \leq \lambda_{\max}\left(K_{\theta_1,\theta_2}(\sigma_i, \sigma_j)_{i,j}\right) + \theta_{3,\max}$ and we show that $\lambda_{\max}\left(K_{\theta_1,\theta_2}(\sigma_i, \sigma_j)_{i,j}\right)$ is uniformly bounded using Gershgorin circle theorem ([Ger31]).
2. If $\alpha_3 = 1$, then $\frac{\partial^{|\alpha|} R_\theta}{\partial\theta^\alpha} = I_n$.
3. If $\alpha_3 > 1$, then $\frac{\partial^{|\alpha|} R_\theta}{\partial\theta^\alpha} = 0$.

Then, we suppose that $(\alpha_1, \alpha_2) \neq (0,0)$. Thus,

$$\frac{\partial^{|\alpha|} R_\theta}{\partial \theta^\alpha} = \frac{\partial^{|\alpha|} \left(K_{\theta_1,\theta_2}(\sigma_i,\sigma_j)_{i,j}\right)}{\partial \theta^\alpha}.$$

It does not depend on α_3 so we can assume that $\alpha \in \mathbb{N}^2$. We have

$$\left| \frac{\partial^{|\alpha|} K_{\theta_1,\theta_2}(\sigma,\sigma')}{\partial \theta^\alpha} \right| \leq \max(1, \theta_{2,\max}) d(\sigma,\sigma')^{\alpha_1} e^{-\theta_{1,\min} d(\sigma,\sigma')}. \tag{6.10}$$

We conclude using Gershgorin circle theorem ([Ger31]). □

Proof of Lemma 4

Proof. Let N be the norm on $\mathbb{R}^3$ defined by

$$N(x) := \max(4c\theta_{2,\max}|x_1|, 2|x_2|, |x_3|), \tag{6.11}$$

with c as in Condition 2. Let $\alpha > 0$. We want to find a positive lower-bound over $\theta \in \Theta \setminus B_N(\theta^*, \alpha)$ of

$$\frac{1}{n} \sum_{i,j=1}^{n} (K'_\theta(\sigma_i,\sigma_j) - K'_{\theta^*}(\sigma_i,\sigma_j))^2. \tag{6.12}$$

Let $\theta \in \Theta \setminus B_N(\theta^*, \alpha)$.

1. If $|\theta_1 - \theta_1^*| \geq \alpha/(4c\theta_{2,\max})$. Let $k_\alpha \in \mathbb{N}$ be the first integer such that

$$k_\alpha^\beta \geq 4c\theta_{2,\max} \frac{2 + \ln(\theta_{2,\max}) - \ln(\theta_{2,\min})}{\alpha}. \tag{6.13}$$

   Then, for all $i \in \mathbb{N}^*$,

$$\left| \frac{(\theta_1^* - \theta_1) d(\sigma_i, \sigma_{i+k_\alpha}) + \ln(\theta_2) - \ln(\theta_2^*)}{2} \right| \geq 1.$$

   For all $n \geq k_\alpha$,

$$\begin{aligned}
& \frac{1}{n} \sum_{i,j=1}^{n} (K'_\theta(\sigma_i,\sigma_j) - K'_{\theta^*}(\sigma_i,\sigma_j))^2 \\
\geq\ & \frac{1}{n} \sum_{i=1}^{n-k_\alpha} (K'_\theta(\sigma_i,\sigma_{i+k_\alpha}) - K'_{\theta^*}(\sigma_i,\sigma_{i+k_\alpha}))^2 \\
\geq\ & \frac{1}{n} \sum_{i=1}^{n-k_\alpha} e^{-2\theta_{1,\max} c k_\alpha + 2\ln(\theta_{2,\min})} 4\sinh^2\left(\frac{(\theta_1^* - \theta_1) d(\sigma_i, \sigma_{i+k_\alpha}) + \ln(\theta_2) - \ln(\theta_2^*)}{2} \right) \\
\geq\ & C_{1,\alpha} \frac{n - k_\alpha}{n},
\end{aligned}$$

   where we write $C_{1,\alpha} = e^{-2\theta_{1,\max} c k_\alpha + 2\ln(\theta_{2,\min})} 4\sinh^2(1)$.

2. If $|\theta_1 - \theta_1^*| \leq \alpha/(4c\theta_{2,\max})$.

(a) If $|\theta_2 - \theta_2^*| \geq \alpha/2$. We have

$$\begin{aligned}\frac{|\theta_1 - \theta_1^*|}{2} d(\sigma_i, \sigma_{i+1}) &< \frac{\alpha}{8\theta_{2,\max}} \\ &= \frac{\alpha}{4\theta_{2,\max}} - \frac{\alpha}{8\theta_{2,\max}} \\ &\leq \frac{|\ln(\theta_2^*) - \ln(\theta_2)|}{2} - \frac{\alpha}{8\theta_{2,\max}}.\end{aligned}$$

Thus,

$$\left| \frac{(\theta_1^* - \theta_1) d(\sigma_i, \sigma_{i+1}) + \ln(\theta_2) - \ln(\theta_2^*)}{2} \right| \geq \frac{\alpha}{8\theta_{2,\max}}, \tag{6.14}$$

and we have

$$\begin{aligned}&\frac{1}{n} \sum_{i,j=1}^{n} (K'_\theta(\sigma_i, \sigma_j) - K'_{\theta^*}(\sigma_i, \sigma_j))^2 \\ \geq\ & \frac{1}{n} \sum_{i=1}^{n-1} (K'_\theta(\sigma_i, \sigma_{i+1}) - K'_{\theta^*}(\sigma_i, \sigma_{i+1}))^2 \\ \geq\ & \frac{1}{n} \sum_{i=1}^{n-1} e^{-2\theta_{1,\max} c + 2\ln(\theta_{2,\min})} 4 \sinh^2 \left(\frac{\alpha}{8\theta_{2,\max}} \right) \\ =\ & C_{2,\alpha} \frac{n-1}{n},\end{aligned}$$

where we write $C_{2,\alpha} := e^{-2\theta_{1,\max} c + 2\ln(\theta_{2,\min})} 4 \sinh^2 \left(\frac{\alpha}{8\theta_{2,\max}} \right)$.

(b) If $|\theta_2 - \theta_2^*| < \alpha/2$, we have $|\theta_3 - \theta_3^*| \geq \alpha$. Thus,

$$\begin{aligned}&\frac{1}{n} \sum_{i,j=1}^{n} (K'_\theta(\sigma_i, \sigma_j) - K'_{\theta^*}(\sigma_i, \sigma_j))^2 \\ \geq\ & \frac{1}{n} \sum_{i=1}^{n} (K'_\theta(\sigma_i, \sigma_i) - K'_{\theta^*}(\sigma_i, \sigma_i))^2 \\ =\ & \frac{1}{n} \sum_{i=1}^{n} (\theta_2 + \theta_3 - \theta_2^* - \theta_3^*)^2 \\ \geq\ & \frac{\alpha^2}{4}.\end{aligned}$$

Finally, if we write

$$C_\alpha := \min \left(C_{1,\alpha}, C_{2,\alpha}, \frac{\alpha^2}{2} \right), \tag{6.15}$$

we have

$$\inf_{N(\theta - \theta^*) \geq \alpha} \frac{1}{n} \sum_{i,j=1}^{n} (K'_\theta(\sigma_i, \sigma_j) - K'_{\theta^*}(\sigma_i, \sigma_j))^2 \geq \frac{n - k_\alpha}{n} C_\alpha. \tag{6.16}$$

To conclude, there exists $h > 0$ such that $\|.\|_2 \leq h N(.)$ thus

$$\liminf_{n \to +\infty} \inf_{\|\theta - \theta^*\| \geq \alpha} \frac{1}{n} \sum_{i,j=1}^{n} (K'_\theta(\sigma_i, \sigma_j) - K'_{\theta^*}(\sigma_i, \sigma_j))^2 \geq C_{\alpha/h} > 0. \tag{6.17}$$

□

Proof of Lemma 5

Proof. We have

$$\begin{aligned}
\frac{\partial}{\partial\theta_1}K'_{\theta^*}(\sigma_i,\sigma_j) &= -d(\sigma_i,\sigma_j)e^{-\theta_1^* d(\sigma_i,\sigma_j)},\\
\frac{\partial}{\partial\theta_2}K'_{\theta^*}(\sigma_i,\sigma_j) &= e^{-\theta_1^* d(\sigma_i,\sigma_j)},\\
\frac{\partial}{\partial\theta_3}K'_{\theta^*}(\sigma_i,\sigma_j) &= \mathbb{1}_{i=j}
\end{aligned}$$

Let $(\lambda_1,\lambda_2,\lambda_3)\neq(0,0,0)$. We have

$$\begin{aligned}
&\frac{1}{n}\sum_{i,j=1}^{n}\left(\sum_{k=1}^{3}\lambda_k\frac{\partial}{\partial\theta_k}K'_{\theta^*}(\sigma_i,\sigma_j)\right)^2\\
=\;&\frac{1}{n}\sum_{i\neq j=1}^{n}\left(\sum_{k=1}^{2}\lambda_k\frac{\partial}{\partial\theta_k}K'_{\theta^*}(\sigma_i,\sigma_j)\right)^2+(\lambda_2+\lambda_3)^2\\
=\;&\frac{1}{n}\sum_{i\neq j=1}^{n}e^{-2\theta_1^* d(\sigma_i,\sigma_j)}\left(\lambda_2-\lambda_1 d(\sigma_i,\sigma_j)\right)^2+(\lambda_2+\lambda_3)^2.
\end{aligned}$$

If $\lambda_1\neq 0$, then for conditions 1 and 2, we can find $\epsilon>0,\tau>0,k\in\mathbb{Z}$ so that for $|i-j|=k$, we have $(\lambda_2-\lambda_1 d(\sigma_i,\sigma_j))^2\geq\epsilon$ and $e^{-2\theta_1^* d(\sigma_i,\sigma_j)}\geq\tau$. This concludes the proof in the case $\lambda_1\neq 0$. The proof in the case $\lambda_1=0$ can then be obtained by considering the pairs $(j,j+1)$ in the above display.

□

6.3 Proofs of Partial rankings

Proof of Proposition 3

Proof. We follow the different steps of [LM08].

Lemma 7. *For $\pi\in S_n$, $q>0$, and a composition $\gamma=(\gamma_1,...,\gamma_m)$ we have*

$$\sum_{\sigma\in S_\gamma\pi}q^{d_H(\sigma,id)}=q^{\sum_{j,l=1}^{m}b_{jl}^{\gamma}(\pi)}\prod_{j=1}^{m}\sum_{h=0}^{\gamma_j}c_{\gamma_j,h}q^h. \tag{6.18}$$

Proof.

$$\begin{aligned}
\sum_{\sigma\in S_\gamma\pi}q^{d_H(\sigma,id)} &= \sum_{\sigma\in S_\gamma\pi}q^{\sum_{j=1}^{m}a_j^{\gamma}(\sigma)+\sum_{j,l=1}^{r}b_{jl}^{\gamma}(\tau)}\\
&= q^{\sum_{j,l=1}^{r}b_{jl}^{\gamma}(\pi)}\sum_{\sigma\in S_\gamma\pi}q^{\sum_{j=1}^{m}a_j^{\gamma}(\sigma)}\\
&= q^{\sum_{j,l=1}^{r}b_{jl}^{\gamma}(\pi)}\sum_{\sigma\in S_\gamma\pi}q^{\sum_{j=1}^{m}a_j^{\gamma}(\sigma)}\\
&= q^{\sum_{j,l=1}^{r}b_{jl}^{\gamma}(\pi)}\prod_{j=1}^{m}\sum_{\sigma_j\in S_{\gamma_j}}q^{d_H(\sigma_j,id)}\\
&= q^{\sum_{j,l=1}^{m}b_{jl}^{\gamma}(\pi)}\prod_{j=1}^{m}\sum_{h=0}^{\gamma_j}c_{\gamma_j,h}q^h.
\end{aligned}$$

□

Corollary 4. *For $\pi, \tau \in S_n$, $q > 0$, and a partition $\gamma = (\gamma_1, ..., \gamma_m)$ we have*

$$\sum_{\sigma \in S_\gamma \pi} q^{d_H(\sigma\tau, id)} = q^{\sum_{j,l=1}^r b_{jl}^\gamma(\pi\tau)} \prod_{j=1}^m \sum_{h=0}^{\gamma_j} c_{\gamma_j,h} q^h. \tag{6.19}$$

Proof. $(S_\gamma \pi)\tau = S_\gamma(\pi\tau)$ so

$$\sum_{\tau \in S_\gamma \pi} q^{d_H(\sigma\tau, id)} = \sum_{\sigma' \in S_\gamma(\pi\tau)} q^{d(\sigma')}$$

□

We use Equation (6.19) to conclude the proof of Proposition 3.

$$\begin{aligned}
&\sum_{\sigma \in S_{\gamma^1}\pi_1} \sum_{\sigma' \in S_{\gamma^2}\pi_2} e^{-\nu d_H(\sigma,\sigma')} \\
= &\sum_{\sigma \in S_{\gamma^1}} \sum_{\sigma' \in S_{\gamma^2}} e^{-\nu d_H(\sigma\pi_1,\sigma'\pi_2)} \\
= &\sum_{\sigma \in S_{\gamma^1}} \sum_{\sigma' \in S_{\gamma^2}} e^{-\nu d_H(\sigma\pi_1\pi_2^{-1}\sigma'^{-1},id)} \\
= &\sum_{\sigma' \in S_{\gamma^2}} \sum_{\sigma \in S_{\gamma^1}} e^{-\nu d_H(\sigma\pi_1\pi_2^{-1}\sigma',id)} \\
= &\sum_{\sigma' \in S_{\gamma^2}} e^{-\nu \sum_{j,l=1}^r b_{jl}^\lambda(\pi_1\pi_2^{-1}\sigma')} \prod_{j=1}^m \sum_{h=0}^{\gamma_j} c_{\gamma_j^1,h} e^{-\nu h}.
\end{aligned}$$

□

Proof of Proposition 4

Proof. It suffices to prove that if $E_R = E_\gamma \pi_R$, then $\widehat{E_R}(\rho_\lambda) = 0$ for all $\lambda \notin \Lambda_h^n$ where $h := \max(\tilde{\gamma}_k)_k$ (with $\tilde{\gamma}_{m+1} = n - k$). For all $j \in [1 : m+1]$, let $g_j := \sum_{l=1}^{j-1} \tilde{\gamma}_l$ and let S_{Γ_j} be the subgroup of S_n which permutes only the integers between $g_j + 1$ and g_{j+1}. Then, we have

$$S_{\tilde{\gamma}} = \prod_{j=1}^{m=1} S_{\Gamma_j} \tag{6.20}$$

Now, let $\tau := (1\ 2\ ...\ n)$. Then, for all $j \in [1 : m+1]$

$$S_{\Gamma_j} = \tau^{g_j} S_{\gamma_j}, \tag{6.21}$$

where S_{γ_j} is the subgroup of S_n which permutes only the integers between 1 and γ_j. Proposition 7 of [KB10] says that the subgroup S_{γ_j} of S_n is bandlimited on $\Lambda_{\gamma_j}^n$ i.e. for all $\Lambda_{\gamma_j}^n$, we have $\widehat{S_j}(\rho_\lambda) = 0$. Thus, using Equation (6.21) and the fact that $\widehat{uv}(\rho_\lambda)\widehat{u}(\rho_\lambda)\widehat{v}(\rho_\lambda)$, we have that S_{Γ_j} is also bandlimited on $\Lambda_{\gamma_j}^n$. Then, Equation (6.20) shows that $S_{\tilde{\gamma}}$ is bandlimited on Λ_h^n. We concludes the proof using that $E_R = \Pi_k^n S_{\tilde{\gamma}} \pi_R$. □

Proof of Proposition 5

Proof.

$$\begin{aligned} & K(R_1, R_2) \\ = & \frac{1}{n!|E_{R_1}|E_{R_2}|} \sum_{\lambda \in \Lambda_c^n} d_\lambda Tr\left(\widehat{E_{R_1}}(\rho_\lambda)^* \, \widehat{k}(\rho_\lambda) \widehat{E_{R_2}}(\rho_\lambda)\right) \\ = & \frac{1}{n!} \sum_{\lambda \in \Lambda_c^n} d_\lambda Tr\left[\left(\frac{1}{|E_{R_1}|} \sum_{\sigma \in E_{R_1}} \rho_\lambda(\sigma)^*\right) \widehat{k}(\rho_\lambda) \left(\frac{1}{|E_{R_2}|} \sum_{\sigma \in E_{R_2}} \rho_\lambda(\sigma)\right)\right] \\ = & \frac{1}{n!} \sum_{\lambda \in \Lambda_c^n} d_\lambda Tr\left(\rho_\lambda(\pi_1^{-1}) q_\lambda(E_{\gamma^1}) \widehat{k}(\rho_\lambda) p_\lambda(E_{\gamma^2}) \rho_\lambda(\pi_2)\right) \end{aligned}$$

□

Proof of Proposition 6

Proof. It suffices to show that the kernel $K_\nu(R, R') := \exp(-\nu d_{avg}(R, R'))$ is strictly positive definite. We have seen that for all the three distances, there exist constants C_n and d_n and a function $\Phi : S_n \to R^{d_n}$ such that $C_n - d(\sigma, \sigma') = \Phi(\sigma)^T \Phi(\sigma')$.

Let us write

$$\Phi_{avg} : R \longmapsto \frac{1}{|E_R|} \sum_{\sigma \in E_R} \Phi(\sigma). \tag{6.22}$$

Then,

$$\begin{aligned} C_n - d_{avg}(R, R') &= C_n - \frac{1}{|E||E'|} \sum_{\sigma \in E_R} \sum_{\sigma \in E_{R'}} d(\sigma, \sigma') \\ &= \frac{1}{|E||E'|} \sum_{\sigma \in E_R} \sum_{\sigma \in E_{R'}} C_n - d(\sigma, \sigma') \\ &= \frac{1}{|E||E'|} \sum_{\sigma \in E_R} \sum_{\sigma \in E_{R'}} \Phi(\sigma)^T \Phi(\sigma') \\ &= \Phi_{avg}(R)^T \Phi_{avg}(R'). \end{aligned}$$

$$D_{avg}(R, R') := \Phi_{avg}(R)^T \Phi_{avg}(R'). \tag{6.23}$$

D_{avg} is a Gramian matrix thus a semi-definite positive matrix.
We use the Hadamard product $A \circ B$ of the matrix A and B, defined by the element-wise product. We have

$$e^{\nu C_n} K_\nu \sum_{i=0}^{+\infty} \frac{\nu^i}{i!} D_{avg}^{\circ i}.$$

Then, $e^{\nu C_n} K_\nu$ is a positive semi-definite matrix because of Schur's theorem. □

Proof of Proposition 7

Proof. Assume that σ (resp. σ') is an uniform random variable on E_I (resp. $E_{I'}$). We have to compute $\mathbb{E}(d_{avg}(\sigma, \sigma'))$ for the three distances: Kendall's tau, Hamming and Spearman's footrule.

First, we prove (5.24). We follow the proof of Lemma 3.1 of [FKS03]. We have

$$\mathbb{E}(d_{\tau,avg}(\sigma,\sigma')) = \sum_{i<j} \mathbb{E}(K_{i,j}(\sigma,\sigma')), \tag{6.24}$$

with

$$K_{i,j}(\sigma,\sigma') = \mathbb{1}_{(\sigma(i)<\sigma(j),\sigma'(i)>\sigma'(j)) \text{ or } (\sigma(i)>\sigma(j),\sigma'(i)<\sigma'(j))}. \tag{6.25}$$

We compute $\mathbb{E}(K_{i,j}(\sigma,\sigma'))$ for (i,j) in different cases.

1. i and j are in $I \cap I'$. Let us call them a and b. Their exists j and $j' \in [1:p]$ such that $a = i_{c_j}$ and $b = i_{c_{j'}}$. Then
$$K_{a,b}(\sigma,\sigma') = \mathbb{1}_{(c_j<c_{j'},c'_j>c'_{j'}) \text{ or } (c_j>c_{j'},c'_j<c'_{j'})}. \tag{6.26}$$
Thus, the total contribution of the pairs in this case is
$$\sum_{\{j,j'\}\subset[1:p]} \mathbb{1}_{(c_j<c_{j'},c'_j>c'_{j'}) \text{ or } (c_j>c_{j'},c'_j<c'_{j'})}. \tag{6.27}$$

2. i and j both appear in one top-k list (say I) and exactly one of i or j, say i appear in the other top-k list. Let us call P_2 the set of (i,j) such that $i<j$ and (i,j) is in this case. We have
$$\begin{aligned}\sum_{(i,j)\in P_2} K_{i,j}(\sigma,\sigma') &= \sum_{\substack{i\in I\cap I',\\ j\in(I\cup I')\setminus(I\cap I')}} K_{i,j}(\sigma,\sigma') \\ &= \sum_{\substack{i\in I\cap I',\\ j\in I\setminus I'}} K_{i,j}(\sigma,\sigma') + \sum_{\substack{i\in I\cap I',\\ j\in I'\setminus I}} K_{i,j}(\sigma,\sigma')\end{aligned}$$
Let us compute the first sum. We order $u_1,...,u_r$ such that $u_1 < ... < u_r$. Let $l \in [1:r]$. We have $\#\{i \in I,\ \sigma(i) > \sigma(i_{u_l}) = u_l\} = k - u_l$ and $\#\{i \in I \setminus I',\ \sigma(i) > \sigma(i_{u_l})\} = r - l$ thus $\#\{i \in I \cap I',\ \sigma(i) > \sigma(i_{u_l})\} = k - u_l - r + l$.
$$\begin{aligned}\sum_{\substack{i\in I\cap I',\\ j\in I\setminus I'}} K_{i,j}(\sigma,\sigma') &= \sum_{l=1}^{r} \#\{i \in I\cap I',\ \sigma(i) > \sigma(i_{u_l})\} \\ &= r\left(k + \frac{1-r}{2}\right) - \sum_{l=1}^{r} u_l.\end{aligned}$$
Likewise, we have
$$\sum_{\substack{i\in I\cap I',\\ j\in I'\setminus I}} K_{i,j}(\sigma,\sigma') = r\left(k + \frac{1-r}{2}\right) - \sum_{l=1}^{r} u'_l. \tag{6.28}$$
Finally, the total contribution of the pairs in this case is
$$r(2k+1-r) - \sum_{j=1}^{r}(u_j + u'_j). \tag{6.29}$$

3. i, but not j, appears in one top k list (say I), and j, but not i, appears in the other top k list (I'). Then $K_{i,j}(\sigma,\sigma') = 1$ and the total contribution of these pairs is r^2.

4. i and j do not appear in the same top k list (say I). It is the only case where $K_{i,j}(\sigma, \sigma')$ is a non constant random variable. First, we show that in this case, $\mathbb{E}(K_{i,j}(\sigma, \sigma')) = 1/2$. Assume for example that I do not contain i and j. Fix $\sigma' \in E_{I'}$. Let $(i\ j)$ be the transposition which exchange i and j. We have

$$\{\sigma \in E_I, \sigma(i) < \sigma(j)\} = (i\ j)\{\sigma \in E_I, \sigma(i) > \sigma(j)\}. \tag{6.30}$$

Thus, there are as many $\sigma in E_I$ such that $\sigma(i) < \sigma(j)$ as there are $\sigma \in E_I$ such that $\sigma(i) > \sigma(j)$. That proves that $\mathbb{E}(K_{i,j}(\sigma, \sigma')) = 1/2$.

Then, the total distribution of the pairs in thus case is

$$\frac{1}{2}\binom{|I^c|}{2} + \frac{1}{2}\binom{|I'^c|}{2} = \binom{n-k}{2} = \frac{1}{2}(n-k)(n-k-1). \tag{6.31}$$

Finally, we have proved (5.24).

To prove Equation (5.25), it suffices to see that

$$\begin{aligned} d_{H,avg}(I, I') &= \mathbb{E}\left(\sum_{i=1}^{n} \mathbb{1}_{\sigma(i)\neq\sigma'(i)}\right) \\ &= \sum_{j=1}^{p} \mathbb{1}_{c_j \neq c'_j} + \mathbb{E}\left(\sum_{i\neq I\cup I'} \mathbb{1}_{\sigma(i)\neq\sigma'(i)}\right) \\ &\quad + \mathbb{E}\left(\sum_{j=1}^{r} \mathbb{1}_{u_j \neq \sigma'(i_{u_j})}\right) + \mathbb{E}\left(\sum_{j=1}^{r} \mathbb{1}_{\sigma(i_{u'_j})\neq u'_j}\right) \\ &= \sum_{j=1}^{p} \mathbb{1}_{c_j \neq c'_j} + m\frac{n-k-1}{n-k} + 2r. \end{aligned}$$

We want to prove (5.26). Let us write

- $A_c := \sum_{j=1}^{p} |c_j - c'_j|$
- $A_u(\sigma') := \sum_{j=1}^{r} |u_j - \sigma'(i_{u_j})|$
- $A_{u'}(\sigma) := \sum_{j=1}^{r} |\sigma(i'_{u'_j}) - u'_j|$
- $R(\sigma, \sigma') := \sum_{i\neq I\cup I'} |\sigma(i) - \sigma'(i)|$.

$$d_{S,avg}(I, I') = \mathbb{E}(A_c) + \mathbb{E}(A_u(\sigma')) + \mathbb{E}(A_{u'}(\sigma)) + \mathbb{E}(R(\sigma, \sigma')).$$

We have to compute all the expected values.

1. $\mathbb{E}(A_c) = A_c$.
2. $E(A_u(\sigma')) = \sum_{j=1}^{r} \mathbb{E}(|u_j - \sigma'(i_{u_j})|)$. If $\sigma' \in B_{I'}$, then $\sigma'(i_{u_j}) \in [k+1 : n]$ so:

$$\mathbb{E}(|u_j - \sigma'(i_{u_j})|) = \mathbb{E}(\sigma'(i_{u_j} - u_j) = \frac{n+k+1}{2} - u_j.$$

Finally,

$$E(A_u(\sigma')) = r\frac{n+k+1}{2} - \sum_{j=1}^{r} u_j. \tag{6.32}$$

3. $E(A_{u'}(\sigma)) = r\frac{n+k+1}{2} - \sum_{j=1}^{r} u'_j$.

4. $E(R(\sigma,\sigma')) = \sum_{i\neq I\cup I'} \mathbb{E}(|\sigma(i)-\sigma'(i)|)$. $\sigma(i)$ and $\sigma'(i)$ are independent uniform random variables on $[k+1:n]$.

$$\begin{aligned}\mathbb{E}(|\sigma(i)-\sigma'(i)|) &= \sum_{j=1}^{n-k-1} j\mathbb{P}(|\sigma(i)-\sigma'(i)|=j)\\ &= \sum_{j=1}^{n-k-1} j2\frac{n-k-1-j}{(n-k-1)^2}.\end{aligned}$$

Then

$$\mathbb{E}(R(\sigma,\sigma')) = \frac{2m}{n'^2}\sum_{j=1}^{n'} j(n'-j) = \frac{2m}{n'^2}\left(\frac{n'(n'+1)}{2} - \frac{1}{3}n'^3 - \frac{1}{2}n'^2 - \frac{1}{6}n'\right).$$

So,

$$\mathbb{E}(R(\sigma,\sigma')) = \frac{2m}{n'}\left(\frac{n'+1}{2} - \frac{1}{3}n'^2 - \frac{1}{2}n' - \frac{1}{6}\right). \tag{6.33}$$

That concludes Equation (5.26). □

Proof of Proposition 8

Proof. We define

$$\begin{aligned}a_j^{\gamma}(\sigma,\sigma') &:= |\{i\in[1:n],\ \sigma(i)\in\Gamma_j, \sigma'(i)\in\Gamma_j,\ \sigma(i)\neq\sigma'(i)\}|\\ b_{j,l}^{\gamma}(\sigma,\sigma') &:= |\{i\in[1:n],\ \sigma(i)\in\Gamma_j, \sigma'(i)\in\Gamma_l,\ j\neq l\}|\end{aligned}$$

Now, assume that $\sigma,\sigma'\sim\mathcal{U}(S_\gamma)$ and $\sigma_j,\sigma'_j\sim\mathcal{U}(S_{\gamma_j})$.

$$\begin{aligned}\mathbb{E}\left(d_H(\sigma,\sigma')\right) &= \mathbb{E}\left(\sum_{j,l=1}^{m} b_{j,l}^{\gamma}(\sigma\pi_1,\sigma'\pi_2) + \sum_{j=1}^{m} a_j^{\gamma}(\sigma\pi_1,\sigma'\pi_2)\right)\\ &= \sum_{j,l=1}^{m} b_{j,l}^{\gamma}(\pi_1,\pi_2) + \sum_{j=1}^{m} |\{i,\ \pi_1(i),\pi_2(i)\in\Gamma_j\}|\frac{\gamma_j-1}{\gamma_j}\\ &= |\{i,\ \Gamma(\pi_1(i))\neq\Gamma(\pi_2(i))\}| + \sum_{j=1}^{m}\frac{\gamma_j}{n}(\gamma_j-1)\end{aligned}$$

□

References

[Bac14] François Bachoc. Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes. *Journal of Multivariate Analysis*, 125:1–35, 2014.

[BGLV17] F. Bachoc, F. Gamboa, J. M. Loubes, and N. Venet. A Gaussian process regression model for distribution inputs. *IEEE Transactions on Information Theory*, PP(99):1–1, 2017.

[CGJ11] Stéphan Clémençon, Romaric Gaudel, and Jérémie Jakubowicz. Clustering Rankings in the Fourier Domain. In *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 343–358. Springer, Berlin, Heidelberg, September 2011.

[CL93] N. Cressie and S.N Lahiri. The asymptotic distribution of REML estimators. *Journal of Multivariate Analysis*, 45:217–233, 1993.

[CL96] N. Cressie and S.N Lahiri. Asymptotics for REML estimation of spatial covariance parameters. *Journal of Statistical Planning and Inference*, 50:327–341, 1996.

[Dia88] Persi Diaconis. Group representations in probability and statistics. *Lecture Notes-Monograph Series*, 11:i–192, 1988.

[FKS03] Ronald Fagin, Ravi Kumar, and Dakshinamurthi Sivakumar. Comparing top k lists. *SIAM Journal on discrete mathematics*, 17(1):134–160, 2003.

[Ger31] S. Gerschgorin. Uber die abgrenzung der eigenwerte einer matrix. *Izvestija Akademii Nauk SSSR, Serija Matematika*, 7(3):749–754, 1931.

[Hau99] David Haussler. Convolution kernels on discrete structures. Technical report, Technical report, Department of Computer Science, University of California at Santa Cruz, 1999.

[JV17] Yunlong Jiao and Jean-Philippe Vert. The Kendall and Mallows kernels for permutations. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[KB10] Risi Kondor and Marconi Barbosa. Ranking with kernels in Fourier space. 2010.

[KCS17] Anna Korba, Stephan ClÃ©menÃ§on, and Eric Sibony. A Learning Theory of Ranking Aggregation. In *Artificial Intelligence and Statistics*, pages 1001–1010, 2017.

[Kon08] Imre Risi Kondor. *Group Theoretical Methods in Machine Learning*. PhD Thesis, Columbia University, New York, NY, USA, 2008.

[LM08] Guy Lebanon and Yi Mao. Non parametric modeling of partially ranked data. *Journal of Machine Learning Research*, 9(Oct):2401–2429, 2008.

[MM84] K. V. Mardia and R. J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71:135–146, 1984.

[MRW+16] Horia Mania, Aaditya Ramdas, Martin J. Wainwright, Michael I. Jordan, and Benjamin Recht. On kernel methods for covariates that are rankings. *arXiv:1603.08035 [cs, stat]*, March 2016. arXiv: 1603.08035.

[SFL11] Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

[Ste99] M.L Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York, 1999.