Gaussian Processes indexed on the symmetric group: prediction and learning

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

Gaussian field on the symmetric group: prediction and learning

F. Bachoc *[1], B. Broto†[2], F. Gamboa‡[1], and J-M. Loubes§[1]

[1]Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse Cedex 9
[2]CEA, LIST, Université Paris-Saclay, F-91120, Palaiseau, France

September 9, 2018

Abstract

In the framework of the supervised learning of a real function defined on an abstract 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 framework, we propose and study an harmonic analysis of the covariance operators that allows us to put into action the full machinery of Gaussian processes learning. We also consider our framework in the case of partial rankings.

Keywords
Learning, Gaussian processes, statistical ranking.

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 also become a very popular problem in statistical learning. It 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.

*francois.Bachoc@math.univ-toulouse.fr
†baptiste.broto@cea.fr
‡fabrice.gamboa@math.univ-toulouse.fr
§jean-michel.loubes@math.univ-toulouse.fr

More precisely, assume that we have a finite set $X = \{x_1, \cdots, x_n\}$ and we have to order the elements of X. A ranking on X is a statement of the form

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

Where all the $i_j, j = 1 \cdots, n$ are different. We can associate to this ranking the permutation σ defined by $\sigma(i_k) = k$. Reversely, to a permutation σ, we can associate the following ranking

$$x_{\sigma^{-1}(1)} \succ x_{\sigma^{-1}(2)} \succ \cdots \succ x_{\sigma^{-1}(n)}. \tag{2}$$

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 needs 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* ([18]) , clustering rankings (see [7]) or kernels on rankings for supervised learning. Constructing kernels over the set of permutations has been studied following several different ways. In [16], Kondor provides results about kernels 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. Notice 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 [17] deals with the complexity reduction of computing the kernel computation for partial ranking. Recently, [15] proved that the Kendall and Mallow's kernels are positive definite. Further, [20] 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. We generalize previous results on the Mallow's kernel (see [15]). Second, we consider the kriging models (see for instance [24]) which consist in inferring the values of a Gaussian random field given observations at a finite set of observation points. We study the asymptotic properties of the maximum likelihood estimator of the parameters of the covariance function. We also prove the asymptotic efficiency of the Kriging prediction under the estimated covariance parameters. We also provide simulations that illustrate the very good performances of the proposed kernels. Then, we show that this framework may be adapted to the cases of learning with partially observed rankings. Concerning partial ranking, we refer to [23] and [25] and references therein.

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

2 Covariance model for rankings

2.1 Positive kernels on ranking

We will use the following notations. Let S_n be the set of all permutations on $E_n = \{1, \dots, n\}$. In order to define a Gaussian process and to provide asymptotic results, we will embed S_n in an infinite set. For this, we will consider the space S_∞ of all permutations on $\mathbb{N}$ which coincides with identity when restricted to $[k+1 : +\infty[$ for some $k \in \mathbb{N}$.

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, consider the toy example called the *administrative nightmare* where 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 $p_{\sigma(1)}, p_{\sigma(2)}, \cdots, p_{\sigma(n)}, \cdots$ Since T is assumed to be the mean time to process the task, we can model the difference between this mean and the time needed to achieve the task for a particular choice of σ, $Y(\sigma) - T$, as 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. This toy example can serve as a model to understand complex decision procedures in a workflow management system where a large number of tasks may be done in different orders but are all necessary to achieve the goal. Workflow prediction or optimization problems currently occur in fields such as grid computing [26], and logistics [6].

Another example of application is given by maintenance of machines in a large supply line. Machines in a supply line need to be tuned or monitored in order to optimize the production of a good. The machines can be tuned in different orders, each corresponding to a permutation and these choices have an impact on the quality of the production of the goods, measured by a quantitative variable Y, for instance the amount of defects in the produced goods. Hence, the objective of the model will thus be to forecast the outcome of a specific order for the maintenance of the machines in order to optimize the production.

As we will consider increasing domains, if $\sigma \in S_n$ and if $n' > n$, we can also consider σ as a member of $S_{n'}$ by setting $\sigma(i) = i$ for all $n < i \leq n'$. With this convention, we can write $S_\infty = \bigcup_n S_n$. Furthermore, let $S_\mathbb{N}$ be the set of all permutations on the integers. This set is a non commutative group for the composition. Several distances can be considered on S_n. We will focus here on the three following distances (see [11]). For any permutations π and σ of S_n let

- The Kendall's tau distance defined by

$$d_\tau(\pi, \sigma) := \sum_{i<j} \left(\mathbb{1}_{\sigma(i)>\sigma(j),\ \pi(i)<\pi(j)} + \mathbb{1}_{\sigma(i)<\sigma(j),\ \pi(i)>\pi(j)} \right), \tag{3}$$

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

- The Hamming distance defined by

$$d_H(\pi, \sigma) := \sum_i \mathbb{1}_{\pi(i) \neq \sigma(i)}. \tag{4}$$

- The Spearman's footrule distance defined by

$$d_S(\pi, \sigma) := \sum_i |\pi(i) - \sigma(i)|. \tag{5}$$

- The Spearman's rank correlation distance defined by

$$d_{S2}(\pi, \sigma) := \sum_i |\pi(i) - \sigma(i)|^2. \tag{6}$$

These four distances are right-invariant. That is, for all $\pi,\ \sigma,\ \tau \in S_n$, $d(\pi, \sigma) = d(\pi\tau, \sigma\tau)$. Other right-invariant distances are discussed in [11]. We extend these three distances naturally on S_∞ and obtain a countably infinite discrete space. We then extend these distances on $S_\mathbb{N}$, considering infinite sums of non-negative numbers and assuming that the distances can be equal to $+\infty$. For example, the Kendall's tau distance is extended on $S_\mathbb{N}$ to

$$d_\tau(\pi, \sigma) = \sum_{\substack{i,j \in \mathbb{N}, \\ i<j}} \left(\mathbb{1}_{\sigma(i)>\sigma(j),\ \pi(i)<\pi(j)} + \mathbb{1}_{\sigma(i)<\sigma(j),\ \pi(i)>\pi(j)} \right). \tag{7}$$

We aim to define a Gaussian process indexed by permutations. Notice that generally speaking, using the abstract Kolmogorov construction (see for example

[10] Chapter 0), the law of a Gaussian random process $(Y_x)_{x\in E}$ indexed by an abstract set E is entirely characterized by its mean and covariance functions

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

and

$$K : (x, y) \mapsto \mathrm{Cov}(Y_x, Y_y).$$

Hence, if we assume that the process is centered, we have only to build a covariance function on $S_{\mathbb{N}}$. Indeed, we may put here in action the Kolmogorov machinery and a Gaussian process on $S_{\mathbb{N}}$ is well defined as soon as we know the mean and covariance functions. 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 $(x_1, \cdots, x_n) \in E^n$, 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 $(x_1, \cdots, x_n) \in E^n$ 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

$$f = \sum_{\sigma\in S_n} a_\sigma K(., \sigma), \tag{8}$$

and K is of course an *universal kernel* (see [22]). 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*

$$\mathit{Vect}\left\{\sum_{i=1}^{n} a_i K(., \sigma_i),\ n \in \mathbb{N},\ a_i \in \mathbb{R},\ \sigma_i \in S_\infty\right\} \tag{9}$$

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

Recall that given a positive kernel K on an abstract set E we may build an interesting Hilbert space of functions on E called *Reproducing Kernel Hilbert Space* (RKHS), associated to K. Roughly speaking, the computation of the scalar product on this Hilbert space relies on K. For more on RKHS, we refer to [4]. As the set defined in (9) is included in the RKHS of the kernel K, we have the following corollary.

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 parametric families of covariance kernels. The members of these families have the general form

$$K_{\theta_1,\theta_2}(\sigma,\sigma') := \theta_2 \exp\left(-\theta_1 d(\sigma,\sigma')\right), \quad (\theta_1,\theta_2 > 0). \tag{10}$$

Here, d is one of the three distances defined in (3), (4) and (5). More precisely, for the Kendall's (resp. Hamming's, Spearman's footrule and Spearman's rank correlation) distance let $K^{\tau}_{\theta_1,\theta_2}$ (resp. $K^{H}_{\theta_1,\theta_2}$, $K^{S}_{\theta_1,\theta_2}$, $K^{S2}_{\theta_1,\theta_2}$) be the corresponding covariance function. In short, sometimes we will write K_{θ_1,θ_2} (resp. d) for one of these three kernels (resp. distances). Notice that these covariance functions vanish when the distance between the permutation increases to infinity. Furthermore, the right-invariance of the distance d is inherited by the kernel K_{θ_1,θ_2}. We also introduce, for $\theta_1,\theta_2,\theta_3 > 0$ and $\theta := (\theta_1,\theta_2,\theta_3)^T$, the *nugget effect* modified kernel

$$K'_{\theta}(\sigma,\sigma') := K_{\theta_1,\theta_2}(\sigma,\sigma') + \theta_3 \mathbb{1}_{\sigma=\sigma'}. \tag{11}$$

We point out that the covariance matrices obtained from K'_{θ} are equal to those obtained from K_{θ_1,θ_2} plus $\theta_3 I_l$ (where l is the appropriate dimension). Hence, the kernel K'_{θ} is strictly positive as soon as K_{θ_1,θ_2} is a covariance function. These properties both hold and are formally stated in the following theorem and corollary.

Theorem 1. *For all $\theta_1 > 0$ and $\theta_2 > 0$, the maps $K^{\tau}_{\theta_1,\theta_2}$, $K^{H}_{\theta_1,\theta_2}$, $K^{S}_{\theta_1,\theta_2}$ and $K^{S2}_{\theta_1,\theta_2}$ are strictly positive definite kernel on S_n, S_∞ and $S_{\mathbb{N}}$.*

Remark 1. *In the proof of Theorem 1, we suggest an other proof of the strictly positive definiteness of the Mallow's kernel, which seems easier and shorter than in [20].*

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

2.2 Extension to partial ranking

2.2.1 A new kernel on partial rankings

In some statistical model, it could happen that we have to deal with partial ranking. A partial ranking aims at giving an order of preference between different elements of X without comparing all the pairs in X. A partial ranking R is a statement of the form

$$X_1 \succ X_2 \succ \cdots \succ X_m, \tag{12}$$

where $m < n$, and $X_1,\cdots,X_m$ are disjoint sets of $X = \{x_1,x_2,\cdots,x_n\}$. The partial ranking means that any element of X_j if preferred to any element of X_{j+1} but the elements of X_j cannot be ordered. Given a partial ranking R, we consider the following subset of S_n

$$\begin{aligned} E_R \;:=\; & \{\sigma \in S_n : \ \sigma(i_1) < \sigma(i_2) < \cdots < \sigma(i_m) \\ & \text{for any choice of } (x_{i_1},\cdots,x_{i_m}) \in X_1 \times \cdots \times X_m \}. \end{aligned} \tag{13}$$

There is a natural way to extend a positive definite kernel K on S_n to the set of partial rankings (see [17], [15]). To do so, one consider for R and R' two partial ranking the following averaged kernel

$$\mathcal{K}(R,R') := \frac{1}{|E_R||E_{R'}|}\sum_{\sigma\in E_R}\sum_{\sigma'\in E_{R'}} K(\sigma,\sigma'). \tag{14}$$

Here, $|E_R|$ denotes the cardinal of the set E_R. Notice that, if K is a positive definite kernel on permutations, then $\mathcal{K}$ is also a positive definite kernel ([14]). Indeed, if $R_1,\cdots,R_N$ are partial rankings and if $(a_1,\cdots,a_N)\neq 0$, then

$$\sum_{i,j=1}^{n} a_i a_j \mathcal{K}(R_i,R_j) = \sum_{\sigma,\sigma'\in S_n} b_\sigma b_{\sigma'} K(\sigma,\sigma'), \tag{15}$$

where we set

$$b_\sigma := \sum_{i,\ \sigma\in R_i} \frac{a_i}{|E_{R_i}|}. \tag{16}$$

The computation of $\mathcal{K}$ is very costly. Indeed, we have to sum over $|E_R||E_{R'}|$ permutations. Several works aim at reducing the computation cost of this kernel (see [17], [19]). However, its efficient computation remains an issue. We provide here a cheaper computational way to extend the kernels K_{θ_1,θ_2} and K'_θ to partial rankings. We first define the measure of dissimilarity d_{avg} on partial rankings as the mean of distances $d(\sigma,\sigma')$ $(\sigma\in E_R, \sigma'\in E_{R'})$. That is

$$d_{\text{avg}}(R,R') := \frac{1}{|E_R||E_{R'}|}\sum_{\sigma\in E_R}\sum_{\sigma\in E_{R'}} d(\sigma,\sigma'). \tag{17}$$

Then, we define

$$\mathcal{K}_{\theta_1,\theta_2}(R,R') := \theta_2\exp(-\theta_1 d_{\text{avg}}(R,R')), \tag{18}$$

and

$$\mathcal{K}'_\theta(R,R') := \theta_2\exp(-\theta_1 d_{\text{avg}}(R,R')) + \theta_3 \mathbb{1}_{R=R'}. \tag{19}$$

The next proposition warrants that these functions are in fact covariance kernels.

Proposition 2.

1. *$\mathcal{K}_{\theta_1,\theta_2}$ is a positive definite kernel for the Kendall's tau distance, the Hamming distance and the Spearman's footrule distance,*

2. *$\mathcal{K}'_\theta$ is a strictly positive definite kernel for the Kendall's tau distance, the Hamming distance and the Spearman's footrule distance.*

Remark 2. *In both cases (if we take the mean of the kernel or the mean of the distances), the values of $\mathcal{K}(R,R)$ depends on R and can be very close to 0. This means for the corresponding Gaussian process Y indexed by the partial rankings that the value $Y(R)$ is almost constant. To circumvent this problem, we may renormalize the kernel setting*

$$\mathcal{K}^{new}(R,R') := \frac{1}{\sqrt{\mathcal{K}(R,R)\mathcal{K}(R',R')}}\mathcal{K}(R,R'). \tag{20}$$

2.2.2 Kernel computation in partial ranking

At the first glance, the computation of the kernels $\mathcal{K}_{\theta_1,\theta_2}$ and $\mathcal{K}'_\theta$ on partial rankings may appears still very long as we have to sum $|E_R||E_{R'}|$ elements. However, this computation problem can be quite simplified. As we will show in this subsection, the mean of the distances is much easier to compute than the mean of exponential of distances. We write $d_{\tau,\text{avg}}$ (resp.$d_{H,\text{avg}}$ and $d_{S,\text{avg}}$), the distance mean in (17) when the distance on the permutations is d_τ (resp. d_H and d_S).
To begin with, let us consider the case of top-k partial rankings. A top-k partial ranking (or a top-k list) is a partial ranking of the form

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k} \succ X_{rest}, \tag{21}$$

where $X_{rest} := X \setminus \{x_{i_1}, \cdots, x_{i_k}\}$. In order to alleviate the notations, let just write $I = (i_1, \cdots, i_k)$ this top-k partial ranking. The proposition that follow shows that the computation cost to evaluate d_{avg} (and so the kernel values) might be reduced when the partial rankings are in fact top-k partial rankings. Before stating this proposition let us define some more mathematical objects. Let $I := (i_1, \cdots, i_k)$ and $I' := (i'_1, \cdots, i'_k)$ be two top-k partial ranking. Let

$$\{j_1, \cdots, j_p\} := \{i_1, \cdots, i_k\} \cap \{i'_1, \cdots, i'_k\}$$

where $j_1 < j_2 < \cdots < j_p$ and p is an integer not greater than k. Let, for $l = 1, \cdots p$, c_{j_l} (resp. c'_{j_l}) denotes the rank of j_l in I (resp. in I'). Further, let $r := k - p$ and define $\tilde{I}$ (resp. $\tilde{I'}$) as the complementary set of $\{j_1, \cdots, j_p\}$ in $\{i_1, \cdots, i_k\}$ (resp. in $\{i'_1, \cdots, i'_k\}$). Writing these two sets in ascending order, we may finally define for $j = 1, \cdots, r$, u_j (resp. u'_j) as the rank in I (resp I') of the j-th element of $\tilde{I}$ (resp. $\tilde{I'}$).

Proposition 3. *Let I and I' be two top k-partial rankings. Set $n' := n - k - 1$ and $m := n - |I \cup I'|$. Then,*

$$\begin{aligned}
d_{\tau,avg}(I, I') &= \sum_{1 \le l < l' \le p} \mathbb{1}_{(c_{j_l} < c_{j_{l'}}, c'_{j_l} > c'_{j_{l'}}) \text{ or } (c_{j_l} > c_{j_{l'}}, c'_{j_l} < c'_{j_{l'}})} + r(2k + 1 - r) \\
&\quad - \sum_{j=1}^{r} (u_j + u'_j) + r^2 + \binom{n-k}{2} - \frac{1}{2}\binom{m}{2}, \\
d_{H,avg}(I, I') &= \sum_{l=1}^{p} \mathbb{1}_{c_{j_l} \neq c'_{j_l}} + m\frac{n-k-1}{n-k} + 2r, \\
d_{S,avg}(I, I') &= \sum_{l=1}^{p} |c_{j_l} - c'_{j_l}| + r(n + k + 1) - \sum_{j=1}^{r} (u_j + u'_j) \\
&\quad + mn' - \frac{mn'(2n' + 1)}{3(n' + 1)}.
\end{aligned}$$

Notice that the sequences (c_{j_l}), (c'_{j_l}) and (u_j), (u'_j) are easily computable and so $d_{\text{avg}}(I, I')$ too. Let us discuss an easy example to handle the computation of the previous sequences.

Example 1. *Assume that $n = 7$, $I = (3, 2, 1, 4, 5)$ and $I' = (3, 5, 1, 6, 2)$. We have $(j_1, j_2, j_3, j_4) = (1, 2, 3, 5)$ (the items ranked by I and I', in inceasing order). Thus, $c_{j_1} = 3,\ c_{j_2} = 2,\ c_{j_3} = 1,\ c_{j_4} = 5$ and $c'_{j_1} = 3,\ c'_{j_2} = 5,\ c'_{j_3} = 1,\ c'_{j_4} = 2$. Further, $u_1 = 4$ and $u'_1 = 4$. So that, Proposition 3 leads to*

$$d_{\tau,avg}(I, I') = 6, \quad d_{S,avg}(I, I') = 4.5, \quad d_{S,avg}(I, I') = 11.5.$$

Combining the last proposition with (17) allows us to provide the following explicit evaluation of the kernels.

Corollary 3. *Let I be a k-top partial ranking. Then,*

$$\begin{aligned}
\mathcal{K}^{\tau}_{\theta_1,\theta_2}(I, I) &= \theta_2 \exp\left(-\frac{\theta_1}{2}\binom{n-k}{2}\right) \\
\mathcal{K}^{H}_{\theta_1,\theta_2}(I, I) &= \theta_2 \exp\left(-\theta_1(n-k-1)\right) \\
\mathcal{K}^{S}_{\theta_1,\theta_2}(I, I) &= \theta_2 \exp\left(-\theta_1\left[(n-k)(n-k-1) - \frac{(n-k-1)(2n-2k-1)}{3}\right]\right).
\end{aligned}$$

In the case of the Hamming distance, we may step ahead and provide a simpler computational formula for the distance between two partial ranking whenever their associated partitions share the same number of members (see Proposition 4 bellow). More precisely let R_1 and R_2 be two partial rankings such that

$$R_i = X^i_1 \succ \cdots \succ X^i_k,\ i = 1, 2, \tag{22}$$

assume also that for $j = 1, \cdots, k$, $|X^1_j| = |X^2_j|$ and let denote by γ_j this integer. Obviously, $n = \sum_{j=1}^{k} \gamma_j$ so that $\gamma := (\gamma_j)_j$ is an integer partition of n. Further, when $1 = \gamma_1 = \gamma_2 = \cdots = \gamma_{k-1}$ and $\gamma_k = n - k + 1$ one is in the top-$(k-1)$ partial ranking case. For $j = 1, \cdots, k$, let Γ_j be the set of all integers not lesser than $\sum_{l=1}^{j-1} \gamma_l + 1$ and not greater than $\sum_{l=1}^{j} \gamma_l$. Set further,

$$S_\gamma := S_{\Gamma_1} \times S_{\Gamma_2} \times \cdots \times S_{\Gamma_k}.$$

Notice that S_γ is nothing more than the subgroup of S_n letting invariant the sets S_{Γ_j} $(j = 1, \cdots, k)$. So that, for $i = 1, 2$, we can write E_{R_i} as a right coset $R_i = S_\gamma \pi_i$ for some $\pi_i \in E_{R_i}$. With this extra notations and definitions, we are now able to compute $d_{H,\text{avg}}(R_1, R_2)$.

Proposition 4. *In the previous frame, we have*

$$d_{H,avg}(R_1, R_2) = |\{i,\ \Gamma(\pi_1(i)) \neq \Gamma(\pi_2(i))\}| + \sum_{j=1}^{m} \frac{\gamma_j}{n}(\gamma_j - 1), \tag{23}$$

where, for $1 \leq l \leq n$, $\Gamma(l)$ is the integer j such that $l \in \Gamma_j$.

Another computation of d_{avg} is given in the proof of Theorem 2 of [15] for the Kendall's tau distance and for singleton partial rankings.

Remark 3. *The measure of dissimilarity between partial rankings d_{avg} depends on n, contrary to the distance between total rankings (or permutations). Indeed, let I and I' be two top-k partial rankings on an ambient space of cardinal $n_1 > 0$. If $n_2 > n_1$, I and I' are also top-k partial rankings of an ambient space of cardinal n_2. However, as shown in Proposition 3 the value of $d_{\text{avg}}(I, I')$ depends on the cardinal of the ambient space. Hence, partial rankings on S_∞ is not straightforward.*

3 Gaussian fields on the Symmetric group

Let us consider a Gaussian process Y indexed by $\sigma \in S_{\mathbb{N}}$ (or S_∞), 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

$$\{K_\theta; \theta \in \Theta\}, \tag{24}$$

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 $i = 1, \dots, n$. Let us consider an independent sample of random permutations $\Sigma = (\sigma_1, \sigma_2, \cdots, \sigma_n) \in S_{\mathbb{N}}$ (or S_∞). Assume that we observe Σ and a random vector $y = (y_1, y_2, \cdots, y_n)^T = (Y(\sigma_1), Y(\sigma_2), \cdots, Y(\sigma_n))^T$ defined by, for $\sigma \in S_{\mathbb{N}}$,

$$Y(\sigma) = Z(\sigma) + \varepsilon(\sigma). \tag{25}$$

Here, ε and Z are Gaussian processes indexed by $S_{\mathbb{N}}$ (or S_∞) and independent of Σ. We assume that Z is centered with covariance function $K_{\theta_1^*, \theta_2^*}$ (see (10) in Section 2) and that ε is centered with covariance function given by $\text{cov}(\varepsilon(\sigma), \varepsilon(\sigma')) = \mathbb{1}_{\sigma=\sigma'}$. Z is the unknown function to estimate and ε the noise. Thus, Y is a Gaussian process with zero mean and covariance function K'_{θ^*} defined by (11). The Gaussian process Y (resp. Z) is stationary in the sense that for all $\sigma_1, \cdots, \sigma_n \in S_{\mathbb{N}}$ and for all $\tau \in S_{\mathbb{N}}$, the finite-dimensional distribution of Y (resp. Z) at $\sigma_1, \cdots, \sigma_n$ is the same as the finite-dimensional distribution at $\sigma_1\tau, \cdots, \sigma_n\tau$.

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

$$\widehat{\theta}_{ML} = \widehat{\theta}_n \in \arg\min_{\theta \in \Theta} L_\theta \tag{26}$$

with

$$L_\theta := \frac{1}{n}\ln(\det R_\theta) + \frac{1}{n} y^t R_\theta^{-1} y, \tag{27}$$

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 ([24]). Under increasing-domain asymptotics, as $n \to \infty$, the observation points $\sigma_1, \cdots, \sigma_n$ 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 $(\sigma_1, \cdots, \sigma_n, \cdots)$ 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 ([21, 8, 9, 2]). Moreover, prediction performed using the estimated covariance parameter $\hat{\theta}$ is asymptotically as good as the one computed with θ^* as pointed out in [2]. 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 $\Sigma = (\sigma_1, \cdots, \sigma_n)$ 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 $\forall i, \ d(\sigma_i, \sigma_{i+1}) \le c$.

These conditions are natural under increasing-domain asymptotics. Indeed, Condition 1 provides asymptotic independence for pairs of observations with asymptotic distance locations and warrants that the variance of L_θ and of its gradient converges to 0. Condition 2 ensures the asymptotic discrimination of the covariance parameters (see Lemma 4 in the appendix). These conditions are ensured for particular choices of observations $(\sigma_1, \cdots, \sigma_n)$ for the three different distances previously considered. For example consider the following setting.

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 $c_n = (n + k \quad n + k - 1 \quad \cdots \quad 1)$ 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 $i > n + k$. Then, σ_n is a permutation such that $\sigma_n(1) = n + k$, $\sigma_n(i)$ is a random variable in $[2 : k]$ if $1 < i \le k + 1$, $\sigma_n(i) = i - 1$ if $k + 1 < i \le n + k$ and $\sigma_n(i) = i$ if $i > n + k$. A straightforward computation shows that the Conditions 1 and 2 are satisfied with $\beta = 1$ and $c = 1 + k(k - 1)/2$ for the Kendall's tau distance, $c = 2 + k$ for the Hamming distance, $c = 2 + 2k(k + 1)$ for the Spearman's footrule distance and $c = 2 + k^3$ for the Spearman's rank correlation distance. Indeed, the four distances in S_k are upper-bounded by $k(k - 1)/2$, k, k^2 and k^3 respectively.

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

Theorem 2. *Let $\widehat{\theta}_{ML}$ be defined as in* (26)*, then under Conditions 1 and 2, we get*

$$\widehat{\theta}_{ML} \xrightarrow[n\to+\infty]{\mathbb{P}} \theta^*. \tag{28}$$

The following theorem provides the asymptotic normality of the estimator.

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

$$(M_{ML})_{i,j} = \frac{1}{2n} \operatorname{Tr}\left(R_{\theta^*}^{-1} \frac{\partial R_{\theta^*}}{\partial \theta_i} R_{\theta^*}^{-1} \frac{\partial R_{\theta^*}}{\partial \theta_j} \right). \tag{29}$$

Then

$$\sqrt{n} M_{ML}^{\frac{1}{2}} \left(\widehat{\theta}_{ML} - \theta^* \right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, I_3). \tag{30}$$

Furthermore,

$$0 < \liminf_{n\to\infty} \lambda_{\min}(M_{ML}) \leq \limsup_{n\to\infty} \lambda_{\max}(M_{ML}) < +\infty. \tag{31}$$

Given the maximum likelihood estimator $\widehat{\theta}_n = \widehat{\theta}_{ML}$, 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

$$\hat{Y}_{\widehat{\theta}_n}(\sigma) = r_{\widehat{\theta}_n}^t(\sigma) R_{\widehat{\theta}_n}^{-1} y \tag{32}$$

with

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

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

The following theorem shows that the forecast with the estimated parameter is of the same order as if the true covariance parameter where known.

Theorem 4.

$$\forall \sigma \in S_{\mathbb{N}}, \quad \left| \widehat{Y}_{\widehat{\theta}_{ML}}(\sigma) - \widehat{Y}_{\theta^*}(\sigma) \right| = o_{\mathbb{P}}(1). \tag{33}$$

The proofs of Theorems 2, 3 and 4 are given in the appendix. In [2] and [3], similar results for maximum likelihood were given for Gaussian processes indexed by vectors and by one-dimension probability distributions. In the appendix, we also discuss the similarities and differences between the proofs of Theorems 2, 3 and 4 and these in [2] and [3].

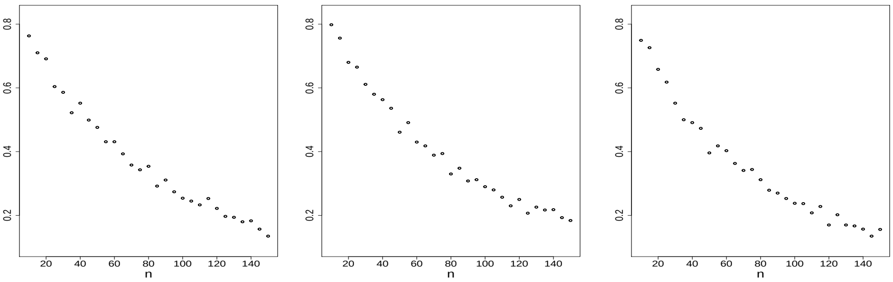

Figure 1: Estimates of $\mathbb{P}(\|\widehat{\theta}_n - \theta^*\| > 0.5)$ for different values of n, the number of observations, with $\theta^* = (0.1, 0.8, 0.3)$ and Kendall's tau distance, the Hamming distance and the Spearman's footrule distance from left to right.

4 Numerical illustrations

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

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 (32) with the parameter $\hat{\theta}_n$ 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 $n \to +\infty$.

5 Conclusion

In this paper, our aim is to provide a model of Gaussian Process indexed by permutations. For this, following the recent work of [15] or [20], we propose kernels

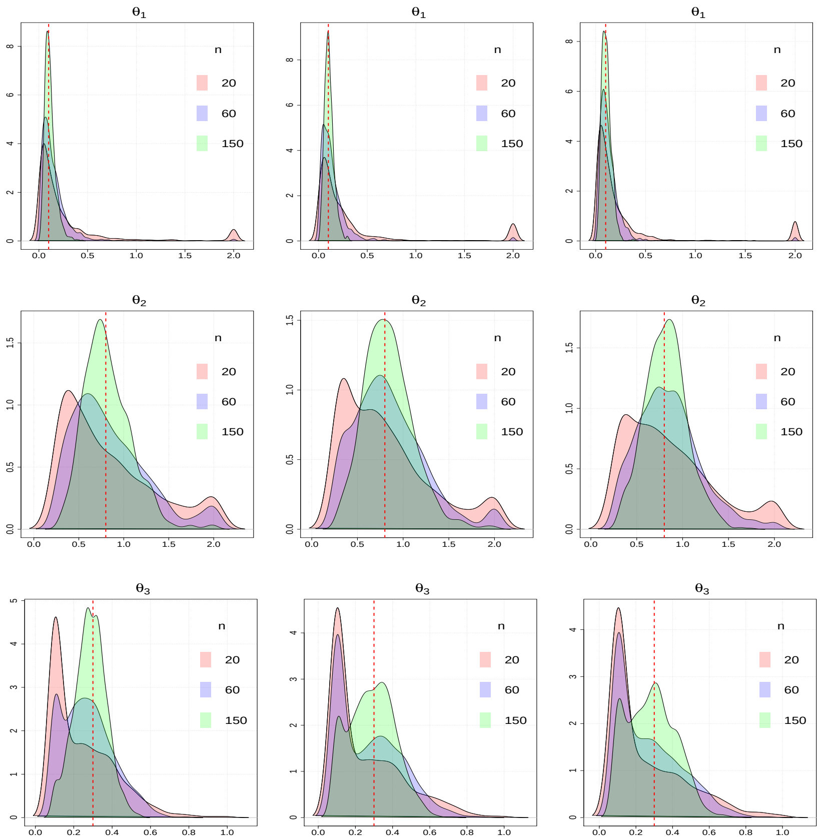

Figure 2: Density of the coordinates of $\widehat{\theta}_n$ for the number of observations $n = 20$ (in red), $n = 60$ (in blue), $n = 150$ (in green) with $\theta^* = (0.1, 0.8, 0.3)$ (represented by the red vertical line). We used the Kendall's tau distance, the Hamming distance and the Spearman's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

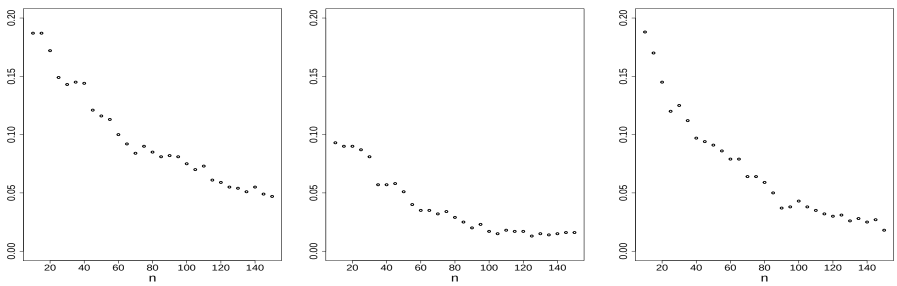

Figure 3: Estimates of $\mathbb{P}\left(\left|\widehat{Y}_{\widehat{\theta}_n}(\sigma) - \widehat{Y}_{\theta^*}(\sigma)\right| > 0.3\right)$ for different values of n, the number of observations, with $\theta^* = (0.1, 0.8, 0.3)$, $\sigma = (1\ 4\ 6)$ and the Kendall's tau distance, the Hamming distance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

to model the covariance of such processes and prove the validity of such choices. Based on the four distances on the set of permutations, Kendall's tau, Hamming distance, Spearman's footrule and Spearman's rank correlation, we obtain parametric families of valid covariance models. In this framework, we prove that under some assumptions on the set of observations, the parameters of the model can be estimated and that the process can also be forecast by linear interpolation of the observations. Such results enable to extend the well-known properties of Kriging methods to the case where the process is indexed by ranks and tackle a large variety of problems. We show that such framework can also be extended to the case of partially observed ranks, which corresponds to many practical use cases. We provide new kernels on partial rankings, together with results that significantly simplify their computation. Limitations of our methods come from the assumptions necessary to estimate the covariance parameters. Indeed, the observations are considered to satisfy the infinite domain assumption, which may prevent usual applications where the ranks apply to a fixed finite number of indices. This setting corresponds to the infill domain in the Gaussian Process literature, for which few results exist. This field of applications is currently under study.

6 Appendix

6.1 Proofs for Section 2

Proof of Proposition 1

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 $N_n \in \mathbb{N}$, $a_1^n, \cdots, a_{N_n}^n \in \mathbb{R}$ and $\sigma_1^n \cdots, \sigma_{N_n}^n \in S_n$ such that

$$f_n = \sum_{i=1}^{N_n} a_i^n K(., \sigma_i^n). \tag{34}$$

We conclude using that f is the pointwise limit of $(f_n)_n$. □

Proof of Theorem 1

Proof. Proof on S_n: 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

$$K(\sigma, \sigma') := e^{-\nu d(\sigma,\sigma')} \tag{35}$$

is a strictly positive definite kernel.

Case of the Kendall's tau distance. It is already shown in Theorem 5 of [20] that K is a strictly positive definite kernel on S_n for the Kendall's tau distance. Nevertheless, we suggest an other proof which seems easier and shorter. Let

$$\Phi : \begin{array}{ccc} S_n & \longrightarrow & \{0,1\}^{\frac{n(n-1)}{2}} \\ \sigma & \longmapsto & (\mathbb{1}_{\sigma(i)<\sigma(j)})_{1\leq i<j\leq n}. \end{array}$$

Let us write

$$M : \begin{array}{ccc} \{0,1\}^{\frac{n(n-1)}{2}} \times \{0,1\}^{\frac{n(n-1)}{2}} & \longrightarrow & \mathbb{R} \\ ((a_{i,j})_{i,j}, (b_{i,j})_{i,j}) & \longmapsto & \exp\left(-\nu \sum_{i<j} |a_{i,j} - b_{i,j}|\right). \end{array}$$

As Φ is an injective map, it suffices to show that M is a strictly positive definite kernel. For all $k \in \mathbb{N}^*$, we index the elements of $\{0,1\}^k$ using the following bijective map

$$N_k : \begin{array}{ccc} \{0,1\}^k & \longrightarrow & [1 : 2^k] \\ a & \longmapsto & 1 + \sum_{i=1}^k a_i 2^{i-1}. \end{array}$$

With this indexation, we let $\tilde{M}$ be the square matrix of size $2^{\frac{n(n-1)}{2}}$ defined by

$$\tilde{M}_{i,j} := M(N^{-1}_{\frac{n(n-1)}{2}}(i), N^{-1}_{\frac{n(n-1)}{2}}(j)).$$

By induction on k, we show that the $2^k \times 2^k$ matrix $M^{(k)}$ defined by

$$M_{i,j}^{(k)} := \exp\left(-\nu \sum_{l=1}^{k} |N_k^{-1}(i)_l - N_k^{(-1)}(j)_l|\right),$$

is the Kronecker product of k matrices A defined by

$$A := \begin{pmatrix} 1 & e^{-\nu} \\ e^{-\nu} & 1 \end{pmatrix}.$$

It is obvious for $k = 1$. Assume that it is true for some k. Thus, for all $i \leq 2^k$ and $j \leq 2^k$, we have

$$\begin{aligned}
(A \otimes M^{(k)})_{i,j} &= 1M_{i,j}^{(k)} \\
&= \exp\left(-\nu \sum_{l=1}^{k} |N_k^{-1}(i)_l - N_k^{(-1)}(j)_l|\right) \\
&= \exp\left(-\nu \sum_{l=1}^{k+1} |N_{k+1}^{-1}(i)_l - N_{k+1}^{(-1)}(j)_l|\right) \\
&= M_{i,j}^{(k+1)}.
\end{aligned}$$

With the same computation, we have

$$(A \otimes M^{(k)})_{i+2^k,j+2^k} = M_{i+2^k,j+2^k}^{(k+1)}.$$

We also have

$$\begin{aligned}
(A \otimes M^{(k)})_{i+2^k,j} &= e^{-\nu} M_{i,j}^{(k)} \\
&= \exp\left(-\nu \left[1 + \sum_{l=1}^{k} |N_k^{-1}(i)_l - N_k^{(-1)}(j)_l|\right]\right) \\
&= \exp\left(-\nu \sum_{l=1}^{k+1} |N_{k+1}^{-1}(i)_l - N_{k+1}^{(-1)}(j)_l|\right) \\
&= M_{i+2^k,j}^{(k+1)},
\end{aligned}$$

and with the same computation,

$$(A \otimes M^{(k)})_{i,j+2^k} = M_{i,j+2^k}^{(k+1)}.$$

That concludes the induction. Using this result with $k = \frac{n(n-1)}{2}$, we have that $\tilde{M}$ is the Kronecker product of positive definite matrices, thus is positive definite and so, M is a strictly positive definite kernel.

Case of the other distances. For the Hamming distance, the Spearman's footrule distance and the Spearman's rank correlation distance, we show that the kernel K is strictly positive definite on the set F of the functions from $[1:n]$ to $[1:n]$. We index these function using the following bijective map

$$J_n : \begin{array}{ccc} F & \longrightarrow & [1:n^n] \\ f & \longmapsto & 1 + \sum_{i=1}^n n^{f(i)-1}. \end{array}$$

Thus, it suffices to show that the $n^n \times n^n$ matrices $\tilde{M}$ defined by

$$\tilde{M}_{i,j} := K\left(J_n^{-1}(i), J_n^{-1}(j)\right),$$

are positive definite matrices for these three distances. Straightforward computations show that

- For the Hamming distance, $\tilde{M}$ is the Kronecker product of n matrices $(\exp(-\nu \mathbb{1}_{i\neq j}))_{i,j\in[1:n]}$.
- For the Spearman Footrule distance, $\tilde{M}$ is the Kronecker product of n matrices $(\exp(-\nu|i-j|))_{i,j\in[1:n]}$.
- For the Spearman's rank correlation distance, $\tilde{M}$ is the Kronecker product of n matrices $\left(\exp(-\nu|i-j|^2)\right)_{i,j\in[1:n]}$.

In all the cases, $\tilde{M}$ is the Kronecker product of positive definite matrices thus is a positive definite matrix.

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

Lemma 1. *Let $\sigma_1, \sigma_1 \in S_{\mathbb{N}}$. Then,*

$$d(\sigma_1, \sigma_2) < +\infty \iff \exists n,\ \sigma_2\sigma_1^{-1} \in S_n. \tag{36}$$

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

Proof of Proposition 2

Proof. It suffices to show that the kernel $\mathcal{K}_\nu(R, R') := \exp(-\nu d_{\text{avg}}(R, R'))$ is strictly positive definite. 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 $C_n - d(\sigma, \sigma') = \langle \Phi(\sigma), \Phi(\sigma') \rangle$. Indeed

- $\frac{n(n-1)}{4} - d_\tau(\sigma, \sigma') = \frac{1}{2} \sum_{i<j} \mathbb{1}_{\sigma(i)<\sigma(j),\, \sigma'(i)<\sigma'(j)} + \mathbb{1}_{\sigma(i)>\sigma(j),\, \sigma'(i)>\sigma'(j)} - \frac{1}{2} \sum_{i<j} \mathbb{1}_{\sigma(i)<\sigma(j),\, \sigma'(i)>\sigma'(j)} + \mathbb{1}_{\sigma(i)>\sigma(j),\, \sigma'(i)<\sigma'(j)} = \langle \Phi(\sigma), \Phi(\sigma') \rangle$ where $\Phi(\sigma) \in \mathbb{R}^{\frac{n(n-1)}{2}}$ is defined by $\Phi(\sigma)_{i,j} := \frac{1}{\sqrt{2}}(\mathbb{1}_{\sigma(i)>\sigma(j)} - \mathbb{1}_{\sigma(i)<\sigma(j)})$, for all $1 \leq i < j \leq n$.
- $n - d_H(\sigma, \sigma') = \sum_{i=1}^n \mathbb{1}_{\sigma(i)=\sigma(j)} = \langle \Phi(\sigma), \Phi(\sigma') \rangle$ where $\Phi(\sigma) \in \mathcal{M}_n(\mathbb{R})$ is defined by $\Phi(\sigma) := (\mathbb{1}_{\sigma(i)=j})_{i,j}$,
- $n^2 - d_S(\sigma, \sigma') = \sum_{i=1}^n \min(\sigma(i), \sigma'(i)) + n - \max(\sigma(i), \sigma'(i)) = \langle \Phi(\sigma), \Phi(\sigma') \rangle$ where $\Phi(\sigma) \in \mathcal{M}_n(\mathbb{R})^2$ is defined by

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

Let us write

$$\Phi_{\text{avg}} : R \longmapsto \frac{1}{|E_R|} \sum_{\sigma \in E_R} \Phi(\sigma). \tag{37}$$

Then,

$$\begin{aligned} C_n - d_{\text{avg}}(R, R') &= C_n - \frac{1}{|E_R||E_{R'}|} \sum_{\sigma \in E_R} \sum_{\sigma \in E_{R'}} d(\sigma, \sigma') \\ &= \frac{1}{|E_R||E_{R'}|} \sum_{\sigma \in E_R} \sum_{\sigma \in E_{R'}} C_n - d(\sigma, \sigma') \\ &= \frac{1}{|E_R||E_{R'}|} \sum_{\sigma \in E_R} \sum_{\sigma \in E_{R'}} \langle \Phi(\sigma), \Phi(\sigma') \rangle \\ &= \langle \Phi_{\text{avg}}(R), \Phi_{\text{avg}}(R') \rangle. \end{aligned}$$

We define

$$D_{\text{avg}}(R, R') := \langle \Phi_{\text{avg}}(R), \Phi_{\text{avg}}(R') \rangle. \tag{38}$$

Then, if we see the maps from $\{\text{partial rankings}\}^2$ to $\mathbb{R}$ as matrices indexed by partial rankings, 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

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

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

Proof of Proposition 3

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

First, we compute $\mathbb{E}(d_\tau(\sigma, \sigma'))$. We follow the proof of Lemma 3.1 of [12]. We have

$$\mathbb{E}(d_\tau(\sigma, \sigma')) = \sum_{i<j} \mathbb{E}(K_{a,b}(\sigma, \sigma')),$$

with

$$K_{a,b}(\sigma, \sigma') = \mathbb{1}_{(\sigma(a)<\sigma(b), \sigma'(a)>\sigma'(b)) \text{ or } (\sigma(a)>\sigma(b), \sigma'(a)<\sigma'(b))}.$$

We compute $\mathbb{E}(K_{a,b}(\sigma, \sigma'))$ for (a, b) in different cases. Let us write $J := \{j_1, \cdots, j_p\}$ and we keep the notation I (resp. I') for the set $\{i_1, ..., i_k\}$ (resp. $\{i'_1, ..., i'_k\}$). In this way, we have $I = J \sqcup \tilde{I}$ and $I' = J \sqcup \tilde{I}'$.

1. a and b are in J. There exists l and $l' \in [1 : p]$ such that $a = j_l$ and $b = j_{l'}$. Then

   $$K_{a,b}(\sigma, \sigma') = \mathbb{1}_{(c_{j_l} < c_{j_{l'}}, c'_{j_l} > c'_{j_{l'}}) \text{ or } (c_{j_l} > c_{j_{l'}}, c'_{j_l} < c'_{j_{l'}})}.$$

   Thus, the total contribution of the pairs in this case is

   $$\sum_{1 \le l < l' \le p} \mathbb{1}_{(c_{j_l} < c_{j_{l'}}, c'_{j_l} > c'_{j_{l'}}) \text{ or } (c_{j_l} > c_{j_{l'}}, c'_{j_l} < c'_{j_{l'}})}.$$

2. a and b both appear in one top-k partial ranking (say I) and exactly one of i or j, say i appear in the other top-k partial ranking. Let us call P_2 the set of (a, b) such that $a < b$ and (a, b) is in this case. We have

   $$\sum_{(a,b) \in P_2} K_{a,b}(\sigma, \sigma') = \sum_{\substack{a \in J, \\ b \in \tilde{I}}} K_{a,b}(\sigma, \sigma') + \sum_{\substack{a \in J, \\ b \in \tilde{I}'}} K_{a,b}(\sigma, \sigma')$$

Let us compute the first sum. Recall that $\tilde{I} = \{i_{u_1}, ..., i_{u_r}\}$.

$$\begin{aligned}\sum_{\substack{a\in J,\\ b\in\tilde{I}}} K_{a,b}(\sigma,\sigma') &= \sum_{b\in\tilde{I}}\sum_{a\in J} K_{a,b}(\sigma,\sigma') \\ &= \sum_{b\in\tilde{I}} \#\{a\in J,\ \sigma(a)>\sigma(b)\} \\ &= \sum_{l=1}^{r} \#\{a\in J,\ \sigma(a)>\sigma(i_{u_l})\}\end{aligned}$$

We order $u_1,\cdots,u_r$ such that $u_1<\cdots<u_r$. Let $l\in[1:r]$. Remark that $\sigma(i_{u_l})=u_l$. We have $\#\{a\in I,\ \sigma(a)>u_l\}=k-u_l$ and $\#\{a\in\tilde{I},\ \sigma(a)>u_l\}=r-l$, thus $\#\{a\in J,\ \sigma(i)>u_l\}=k-u_l-r+l$. Then,

$$\sum_{\substack{a\in J,\\ b\in\tilde{I}}} K_{a,b}(\sigma,\sigma') = r\left(k+\frac{1-r}{2}\right)-\sum_{l=1}^{r}u_l.$$

Likewise, we have

$$\sum_{\substack{a\in J,\\ b\in\tilde{I}'}} K_{a,b}(\sigma,\sigma') = r\left(k+\frac{1-r}{2}\right)-\sum_{l=1}^{r}u'_l. \tag{39}$$

Finally, the total contribution of the pairs in this case is

$$r(2k+1-r)-\sum_{j=1}^{r}(u_j+u'_j).$$

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

4. a and b do not appear in the same top-k partial ranking (say I). It is the only case where $K_{a,b}(\sigma,\sigma')$ is a non constant random variable. First, we show that in this case, $\mathbb{E}(K_{a,b}(\sigma,\sigma'))=1/2$. Assume for example that I does not contain a and b. Let $(a\ b)$ be the transposition which exchanges a and b. We have
$$\{\pi\in E_I,\pi(a)<\pi(b)\}=(a\ b)\{\pi\in E_I,\pi(a)>\pi(b)\}.$$
Thus, there are as many $\pi\in E_I$ such that $\pi(a)<\pi(b)$ as there are $\pi\in E_I$ such that $\pi(a)>\pi(b)$. That proves that $\mathbb{E}(K_{a,b}(\sigma,\sigma'))=1/2$.

Then, the total distribution of the pairs in this case is
$$\frac{1}{2}\left[\binom{|I^c|}{2}+\binom{|I'^c|}{2}-\binom{|I^c\cap I'^c|}{2}\right]=\binom{n-k}{2}-\frac{1}{2}\binom{m}{2}$$

That concludes the computation for the Kendall's tau distance.

To compute $\mathbb{E}(d_H(\sigma,\sigma'))$, it suffices to see that

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

Finally, we want to compute $\mathbb{E}(d_S(\sigma,\sigma'))$. Let us write

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

$$\mathbb{E}(d_S(\sigma,\sigma')) = \mathbb{E}(A_c) + \mathbb{E}(A_u(\sigma')) + \mathbb{E}(A_{u'}(\sigma)) + \mathbb{E}(R(\sigma,\sigma')).$$

We have to compute all the expected values.

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

$$\mathbb{E}(A_u(\sigma')) = r\frac{n+k+1}{2} - \sum_{j=1}^{r} u_j. \tag{40}$$

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

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

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

Then

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

That concludes the proof of Proposition 3. □

Proof of Proposition 4

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 $\sigma_j,\sigma_j'\sim\mathcal{U}(S_{\gamma_j})$.

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

□

6.2 Proofs for Section 3

In the following, let us write $\|.\| = \|.\|_2$ the operator norm (for an endomorphism of $\mathbb{R}^n$ with the Euclidean norm) of a squared matrix of size n, $\|.\|_F$ its Frobenius norm and if $M \in \mathcal{M}_n(\mathbb{R})$, let us define $|M|^2 := \frac{1}{n}\|M\|_F^2$.

The proofs of the three theorems of Section 3 are based on Lemmas 2 to 5. The proofs of these lemmas are new. Then, having at hand the lemmas, the proof of the theorems follows [3]. But, we write the proof to be self-contained.

6.2.1 Lemmas

The following Lemmas are useful for the proofs of Theorems 2, 3 and 4.

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,*

$$\forall \alpha > 0, \ \liminf_{n\to+\infty} \inf_{\|\theta-\theta^*\|\geq\alpha} \frac{1}{n}\sum_{i,j=1}^{n} (K'_\theta(\sigma_i, \sigma_j) - K'_{\theta^*}(\sigma_i, \sigma_j))^2 > 0. \tag{41}$$

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

$$\liminf_{n\to+\infty} \frac{1}{n}\sum_{i,j=1}^{n}\left(\sum_{k=1}^{3}\lambda_i \frac{\partial}{\partial\theta_k}K'_{\theta^*}(\sigma_i, \sigma_j)\right)^2 > 0. \tag{42}$$

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

6.2.2 Proof of Theorem 2

Proof. Step 1: It suffices to prove that

$$\mathbb{P}\left(\sup_\theta |(L_\theta - L_{\theta^*}) - (\mathbb{E}(L_\theta|\Sigma) - \mathbb{E}(L_{\theta^*}|\Sigma))| \geq \epsilon \middle| \Sigma\right) \to_{n\to\infty} 0, \tag{43}$$

and for a fixed $a > 0$,

$$\mathbb{E}(L_\theta|\Sigma) - \mathbb{E}(L_{\theta^*}|\Sigma) \geq a\frac{1}{n}\sum_{i,j=1}^{n}(K_\theta(\sigma_i, \sigma_j) - K_{\theta^*}(\sigma_i, \sigma_j))^2. \tag{44}$$

Indeed, by contradiction, assume that we have (43), (44) but not the consistency of the maximum likelihood estimator. Then, writing the dependency of $\widehat{\theta}$ and $L(\theta)$ with n,

$$\exists\epsilon > 0, \ \exists\alpha > 0, \ \forall n \in \mathbb{N}, \ \exists N_n \geq n, \ \mathbb{P}(|\widehat{\theta}_{N_n} - \theta^*| \geq \epsilon) \geq \alpha. \tag{45}$$

Thus, with probability at least α, we have, for all n:
$|\widehat{\theta}_{N_n} - \theta^*| \geq \epsilon$ thus $\inf_{|\theta-\theta^*|\geq\epsilon} L_{N_n}(\theta) \leq L_{N_n}(\widehat{\theta}_{N_n})$.
However, by definition of $\widehat{\theta}_{N_n}$, we have $L_{N_n}(\widehat{\theta}_{N_n}) \leq L_{N_n}(\theta^*)$.
Thus: $\inf_{|\theta-\theta^*|\geq\epsilon} L_{N_n}(\theta) \leq L_{N_n}(\theta^*)$.
Finally, with probability at least α:

$$\begin{aligned}
0 &\geq \inf_{\|\theta-\theta^*\|\geq\epsilon} \left(L_{N_n}(\theta) - L_{N_n}(\theta^*)\right) \\
&\geq \inf_{\|\theta-\theta^*\|\geq\epsilon} \mathbb{E}\left(L_{N_n}(\theta) - L_{N_n}(\theta^*)\right) - \sup_{\|\theta-\theta^*\|\geq\epsilon} |(L_\theta - L_{\theta^*}) - (\mathbb{E}(L_\theta) - \mathbb{E}(L_{\theta^*}))| \\
&\geq \inf_{\|\theta-\theta^*\|\geq\epsilon} \mathbb{E}\left(L_{N_n}(\theta) - L_{N_n}(\theta^*)\right) + o_{\mathbb{P}}(1) \\
&\geq a|R_\theta - R_{\theta^*}|^2 + o_{\mathbb{P}}(1) \quad \text{ using (43),}
\end{aligned}$$

which is contradicted using (44). It remains to prove (43) and (44).

Step 2: We prove (43).
For all $\sigma \in S_n$,

$$\begin{aligned}
\mathbb{V}(L_\theta|\Sigma=\sigma) &= \mathbb{V}\left(\frac{1}{n}\det(R_\theta) + \frac{1}{n}y^T R_\theta^{-1} y|\Sigma=\sigma\right) \\
&= \frac{2}{n^2}\mathrm{Tr}(R_{\theta^*}R_\theta^{-1}R_{\theta^*}R_\theta^{-1}) \\
&= \frac{2}{n^2}\|R_{\theta^*}R_\theta^{-1}\|_F^2 \\
&\leq \frac{2}{n}\|R_{\theta^*}\|_2\|R_\theta^{-1}\|_2 \\
&\leq \frac{C}{n},
\end{aligned}$$

Thus, for all σ,

$$\mathbb{V}(L_\theta|\Sigma=\sigma) = \mathbb{E}\left((L_\theta - \mathbb{E}(L_\theta|\Sigma=\sigma))^2|\Sigma=\sigma\right) \leq \frac{C}{n},$$

so

$$\mathbb{E}\left((L_\theta - \mathbb{E}(L_\theta|\Sigma=\sigma))^2\right) \leq \frac{C}{n},$$

thus $L_\theta - \mathbb{E}(L_\theta|\Sigma) = o_{\mathbb{P}}(1)$. Let us write $z := R_\theta^{-\frac{1}{2}} y$.

$$\begin{aligned}
\sup_\theta \left|\frac{\partial L_\theta}{\partial\theta}\right| &= \sup_\theta \frac{1}{n}\left(\mathrm{Tr}\left(R_\theta^{-1}\frac{\partial R_\theta}{\partial\theta}\right) + z^t R_{\theta^*}^{\frac{1}{2}} R_\theta^{-1}\frac{\partial R_\theta}{\partial\theta}R_\theta^{-1}R_{\theta^*}^{\frac{1}{2}} z\right) \\
&\leq \sup_\theta \left(\max\left(\|R_\theta^{-1}\|\left\|\frac{\partial R_\theta}{\partial\theta}\right\|, \|R_{\theta^*}\|\|R_\theta^{-2}\|\left\|\frac{\partial R_\theta}{\partial\theta}\right\|\right)\right)\left(1 + \frac{1}{n}|z|^2\right)
\end{aligned}$$

and so is bounded in probability conditionally to $\Sigma = \sigma$, uniformly in σ. Indeed $z \sim \mathcal{N}(0, I_n)$ thus $1/n\ \|z\|^2$ is bounded in probability.

Then $\sup_{k\in[1:p],\theta}\left|\frac{\partial L_\theta}{\partial \theta_k}\right|$ is bounded in probability.
Thanks to the pointwise convergence and the boundness of its derivatives, we have

$$\sup_{\theta}\|L_\theta - \mathbb{E}(L_\theta)\| = o_{\mathbb{P}}(1) \tag{46}$$

Now, let us write $D_{\theta,\theta^*} := \mathbb{E}(L_\theta|\Sigma) - \mathbb{E}(L_{\theta^*}|\Sigma)$. Thanks to (46),

$$\sup_{\theta}|L_\theta - L_{\theta^*} - D_\theta| \leq \sup_{\theta}|L_\theta - \mathbb{E}(L_\theta)| + |L_{\theta^*} - \mathbb{E}(L_{\theta^*})| = o_{\mathbb{P}}(1). \tag{47}$$

Step 3: We prove (44).
We have

$$\mathbb{E}(y^T R_\theta y|\Sigma) = \mathbb{E}(\text{Tr}(y^T R_\theta y)|\Sigma) = \mathbb{E}(\text{Tr}(R_\theta y y^T)|\Sigma)) = \text{Tr}(R_\theta \mathbb{E}(y^T y)).$$

Thus

$$\mathbb{E}(L_\theta|\Sigma) = \frac{1}{n}\sum_{i=1}^{n}\ln(\det(R_\theta)) + \frac{1}{n}\text{Tr}(R_\theta^{-1}R_{\theta^*}), \tag{48}$$

Let us write $\phi_1(M), \cdots, \phi_n(M)$ the eigenvalues of M. We have

$$\begin{aligned}
D_{\theta,\theta^*} &= \frac{1}{n}\ln(\det(R_\theta)) + \frac{1}{n}\text{Tr}(R_\theta^{-1}R_{\theta^*}) - \frac{1}{n}\ln(\det(R_{\theta^*})) - 1 \\
&= \frac{1}{n}\left(-\ln\left((\det(R_\theta^{-1})\det(R_{\theta^*})\right) + \text{Tr}(R_\theta^{-1}R_{\theta^*}) - 1\right) \\
&= \frac{1}{n}\left(-\ln\left((\det(R_{\theta^*}^{\frac{1}{2}}R_\theta^{-1}R_{\theta^*}^{\frac{1}{2}})\right) + \text{Tr}(R_{\theta^*}^{\frac{1}{2}}R_\theta^{-1}R_{\theta^*}^{\frac{1}{2}}) - 1\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(-\ln\left[\phi_i\left(R_{\theta^*}^{\frac{1}{2}}R_\theta^{-1}R_{\theta^*}^{\frac{1}{2}}\right)\right] + \phi_i\left(R_{\theta^*}^{\frac{1}{2}}R_\theta^{-1}R_{\theta^*}^{\frac{1}{2}}\right) - 1\right)
\end{aligned}$$

Thanks to Lemmas 3 and 4, the eigenvalues of R_θ and R_θ^{-1} are uniformly bounded in θ and Σ. Thus, there exist $a > 0$ and $b > 0$ such that for all σ, n and θ, we have

$$\forall i,\ a < \phi_i\left(R_{\theta^*}^{\frac{1}{2}}R_\theta R_{\theta^*}^{\frac{1}{2}}\right) < b.$$

Let us define $f(t) := -\ln(t) + t - 1$. f is minimal in 1 and $f'(1) = 0$ and $f''(1) = 1$. So there exists $A > 0$ such that for all $t \in [a, b]$, $f(t) \geq A(t-1)^2$.

Finally:

$$
\begin{aligned}
D_{\theta,\theta^*} &\geq \frac{A}{n}\sum_{i=1}^{n}\left(1-\phi_i(R_{\theta^*}^{\frac{1}{2}}R_\theta^{-1}R_{\theta^*}^{\frac{1}{2}})\right)^2 \\
&= \frac{A}{n}\mathrm{Tr}\left[\left(1-\phi_i(R_{\theta^*}^{\frac{1}{2}}R_\theta^{-1}R_{\theta^*}^{\frac{1}{2}})\right)^2\right] \\
&= \frac{A}{n}\mathrm{Tr}\left[\left(R_\theta^{-\frac{1}{2}}(R_\theta-R_{\theta^*})R_\theta^{-\frac{1}{2}}\right)^2\right] \\
&= \frac{A}{n}\left\|R_\theta^{-\frac{1}{2}}(R_\theta-R_{\theta^*})R_\theta^{-\frac{1}{2}}\right\|_F^2 \\
&\geq \frac{A}{n}\left\|R_\theta-R_{\theta^*}\right\|_F^2\left\|R_\theta^{\frac{1}{2}}\right\|_F^{-2}\left\|R_\theta^{\frac{1}{2}}\right\|_F^{-2} \\
&\geq a|R_\theta-R_{\theta^*}|^2.
\end{aligned}
$$

□

6.2.3 Proof of Theorem 3

Proof. First, we prove Equation (31). For all $(\lambda_1,\lambda_2,\lambda_3)\in\mathbb{R}^3$ such that $\|(\lambda_1,\lambda_2,\lambda_3)\|=1$, we have

$$
\begin{aligned}
\sum_{i,j=1}^{3}\lambda_i\lambda_j(M_{ML})_{i,j} &= \frac{1}{2n}\mathrm{Tr}\left(R_{\theta^*}^{-1}\left(\sum_{i=1}^{3}\lambda_i\frac{\partial R_{\theta^*}}{\partial\theta_i}\right)R_{\theta^*}^{-1}\left(\sum_{j=1}^{3}\lambda_j\frac{\partial R_{\theta^*}}{\partial\theta_j}\right)\right) \\
&= \frac{1}{2n}\mathrm{Tr}\left(R_{\theta^*}^{-\frac{1}{2}}\left(\sum_{i=1}^{3}\lambda_i\frac{\partial R_{\theta^*}}{\partial\theta_i}\right)R_{\theta^*}^{-\frac{1}{2}}R_{\theta^*}^{-\frac{1}{2}}\left(\sum_{j=1}^{3}\lambda_j\frac{\partial R_{\theta^*}}{\partial\theta_j}\right)R_{\theta^*}^{-\frac{1}{2}}\right) \\
&= \frac{1}{2n}\left\|R_{\theta^*}^{-\frac{1}{2}}\left(\sum_{i=1}^{3}\lambda_i\frac{\partial R_{\theta^*}}{\partial\theta_i}\right)R_{\theta^*}^{-\frac{1}{2}}\right\|_F^2 \\
&\geq \frac{1}{2n}\left\|R_{\theta^*}^{\frac{1}{2}}\right\|_F^{-2}\left\|\left(\sum_{i=1}^{3}\lambda_i\frac{\partial R_{\theta^*}}{\partial\theta_i}\right)\right\|_F^2\left\|R_{\theta^*}^{\frac{1}{2}}\right\|_F^{-2} \\
&\geq C\frac{1}{n}\left\|\frac{\partial R_{\theta^*}}{\partial\theta}\right\|_F^2 \\
&= C\left|\left(\sum_{i=1}^{3}\lambda_i\frac{\partial R_{\theta^*}}{\partial\theta_i}\right)\right|^2
\end{aligned}
$$

Hence, from Lemma 5, we obtain:

$$\liminf_{n\to\infty}\lambda_{\min}(M_{ML})\geq C_{\min}>0. \tag{49}$$

Moreover, we have

$$\begin{aligned}|(M_{ML})_{i,j}| &= \left|\frac{1}{2n}\mathrm{Tr}\left(R_{\theta^*}^{-1}\frac{\partial R_{\theta^*}}{\partial\theta_i}R_{\theta^*}^{-1}\frac{\partial R_{\theta^*}}{\partial\theta_j}\right)\right| \\ &\leq \frac{1}{2n}\left\|R_{\theta^*}^{-1}\frac{\partial R_{\theta^*}}{\partial\theta_i}\right\|_F\left\|R_{\theta^*}^{-1}\frac{\partial R_{\theta^*}}{\partial\theta_i}\right\|_F \\ &\leq \frac{1}{2}\left\|R_{\theta^*}^{-1}\frac{\partial R_{\theta^*}}{\partial\theta_i}\right\|_2\left\|R_{\theta^*}^{-1}\frac{\partial R_{\theta^*}}{\partial\theta_i}\right\|_2 \\ &\leq C_{max}.\end{aligned}$$

Using Gershgorin circle theorem ([13]), we obtain

$$\limsup_{n\to\infty}\lambda_{\max}(M_{ML}) < +\infty, \tag{50}$$

that concludes the proof of Equation (31).

By contradiction, let us now assume that

$$\sqrt{n}M_{ML}^{\frac{1}{2}}\left(\widehat{\theta}_{ML}-\theta^*\right) \underset{n\to+\infty}{\overset{\mathcal{L}}{\not\to}} \mathcal{N}(0,I_3). \tag{51}$$

Then, there exists a bounded measurable function $g:\mathbb{R}^3\to\mathbb{R}$, $\xi>0$ such that, up to extracting a subsequence, we have:

$$\left|\mathbb{E}\left[g\left(\sqrt{n}M_{ML}^{\frac{1}{2}}(\widehat{\theta}_{ML}-\theta^*\right)\right]-\mathbb{E}(g(U))\right|\geq\xi, \tag{52}$$

with $U\sim\mathcal{N}(0,I_3)$. The rest of the proof consists in contradicting Equation (52).

As $0 < C_{\min} \leq \lambda_{\min}(M_{ML}) \leq \lambda_{\max}(M_{ML}) \leq C_{\max}$, up to extraction another subsequence, we can assume that:

$$M_{ML} \underset{n\to\infty}{\longrightarrow} M_\infty, \tag{53}$$

with $\lambda_{\min}(M_\infty) > 0$.

We have:

$$\frac{\partial}{\partial\theta_i}L_\theta = \frac{1}{n}\left(\mathrm{Tr}\left(R_\theta^{-1}\frac{\partial R_\theta}{\partial\theta_i}\right) - y^T R_\theta^{-1}\frac{\partial R_\theta}{\partial\theta_i}R_\theta^{-1}y\right). \tag{54}$$

Let $\lambda=(\lambda_1,\lambda_2,\lambda_3)\in R^3$. For a fixed σ, denoting $\sum_{k=1}^3\lambda_k R_{\theta^*}^{-\frac{1}{2}}\frac{\partial R_{\theta^*}}{\partial\theta_k}R_{\theta^*}^{-\frac{1}{2}} = P^T D P$ with $P^TP=I_n$ and D diagonal, $z_\sigma = PR_{\theta^*}^{-\frac{1}{2}}y$ (which is a vector of i.i.d. standard Gaussian variables, conditionally to $\Sigma=\sigma$), we have

$$\begin{aligned}\sum_{k=1}^3\lambda_k\sqrt{n}\frac{\partial}{\partial\theta_k}L_{\theta^*} &= \frac{1}{\sqrt{n}}\left[\mathrm{Tr}\left(\sum_{k=1}^3\lambda_k R_\theta^{-1}\frac{\partial R_\theta}{\partial\theta_k}\right)-\sum_{i=1}^n\phi_i\left(\sum_{k=1}^3\lambda_k R_{\theta^*}^{-\frac{1}{2}}\frac{\partial R_{\theta^*}}{\partial\theta_k}R_{\theta^*}^{-\frac{1}{2}}\right)z_{\sigma,i}^2\right] \\ &= \frac{1}{\sqrt{n}}\left[\sum_{i=1}^n\phi_i\left(\sum_{k=1}^3\lambda_k R_{\theta^*}^{-\frac{1}{2}}\frac{\partial R_{\theta^*}}{\partial\theta_k}R_{\theta^*}^{-\frac{1}{2}}\right)(1-z_{\sigma,i}^2)\right]\end{aligned}$$

Hence, we have

$$\begin{aligned}
\mathbb{V}\left(\sum_{k=1}^{3}\lambda_k\sqrt{n}\frac{\partial}{\partial\theta_k}L_{\theta^*}|\Sigma\right) &= \frac{2}{n}\sum_{i=1}^{n}\phi_i^2\left(\sum_{k=1}^{3}\lambda_k R_{\theta^*}^{-\frac{1}{2}}\frac{\partial R_{\theta^*}}{\partial\theta_k}R_{\theta^*}^{-\frac{1}{2}}\right) \\
&= \frac{2}{n}\sum_{k,l=1}^{3}\lambda_k\lambda_l \mathrm{Tr}\left(\frac{\partial R_{\theta^*}}{\partial\theta_k}R_{\theta^*}^{-1}\frac{\partial R_{\theta^*}}{\partial\theta_l}R_{\theta^*}^{-1}\right) \\
&= \lambda^T(4M_{ML})\lambda \underset{n\to\infty}{\longrightarrow} \lambda^T(4M_\infty)\lambda.
\end{aligned}$$

Hence, for almost every σ, we can apply Lindeberg-Feller criterion to the variables $\frac{1}{\sqrt{n}}\phi_i\left(\sum_{k=1}^{3}\lambda_k R_{\theta^*}^{-\frac{1}{2}}\frac{\partial R_{\theta^*}}{\partial\theta_k}R_{\theta^*}^{-\frac{1}{2}}\right)(1-z_{\sigma,i}^2)$ to show that, conditionnally to $\Sigma=\sigma$, $\sqrt{n}\frac{\partial}{\partial\theta}L_{\theta^*}$ converges in distribution to $\mathcal{N}(0,4M\infty)$.

Then, using dominated convergence theorem on Σ, we show that:

$$\mathbb{E}\left(\exp\left(i\sum_{k=1}^{3}\lambda_k\sqrt{n}\frac{\partial}{\partial\theta_k}L_{\theta^*}\right)\right)\underset{n\to\infty}{\longrightarrow}\exp\left(-\frac{1}{2}\lambda^T(4M_\infty)\lambda\right). \tag{55}$$

Finally,

$$\sqrt{n}\frac{\partial}{\partial\theta}L_{\theta^*}\overset{\mathcal{L}}{\underset{n\to\infty}{\longrightarrow}}\mathcal{N}(0,4M_\infty). \tag{56}$$

Let us now compute

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_i\partial\theta_j}L_{\theta^*} &= \frac{1}{n}\mathrm{Tr}\left(-R_{\theta^*}^{-1}\frac{\partial R_{\theta^*}}{\partial\theta_i}R_{\theta^*}^{-1}\frac{\partial R_{\theta^*}}{\partial\theta_j}+R_{\theta^*}^{-1}\frac{\partial^2 R_{\theta^*}}{\partial\theta_i\partial\theta_j}\right) \\
&\quad +\frac{1}{n}y^T\left(2R_{\theta^*}^{-1}\frac{\partial R_{\theta^*}}{\partial\theta_i}R_{\theta^*}^{-1}\frac{\partial R_{\theta^*}}{\partial\theta_j}R_{\theta^*}^{-1}-R_{\theta^*}^{-1}\frac{\partial^2 R_{\theta^*}}{\partial\theta_i\partial\theta_j}R_{\theta^*}^{-1}\right)y.
\end{aligned}$$

Thus, we have, a.s.

$$\mathbb{E}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}L_{\theta^*}\right)\underset{n\to+\infty}{\longrightarrow}(2M_\infty)_{i,j}, \tag{57}$$

and, using Lemmas 2 and 3,

$$V\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}L_{\theta^*}|\Sigma\right)\underset{n\to+\infty}{\longrightarrow}0. \tag{58}$$

Hence, a.s.

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}L_{\theta^*}\overset{\mathbb{P}_{|\Sigma}}{\underset{n\to+\infty}{\longrightarrow}}2M_\infty. \tag{59}$$

Moreover, $\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}L_\theta$ can be written as

$$\frac{1}{n}\mathrm{Tr}(A_\theta)+\frac{1}{n}y^T B_\theta y, \tag{60}$$

where A_θ and B_θ are sums and products of the matrices R_θ^{-1} or $\frac{\partial^{|\beta|}}{\partial \theta^\beta}$ with $\beta \in [0 : 3]^3$. Hence, from Lemmas 2 and 3, we have

$$\sup_{\theta \in \Theta} \left\| \frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} L_\theta \right\| = O_{\mathbb{P}_{|\Sigma}}(1). \tag{61}$$

We know that, for $k \in \{1, 2, 3\}$

$$0 = \frac{\partial}{\partial \theta_i} L_{\widehat{\theta}_{ML}} = \frac{\partial}{\partial \theta_k} L_{\theta^*} + \left(\frac{\partial}{\partial \theta} \frac{\partial}{\partial \theta_k} L_{\theta^*} \right)^T (\widehat{\theta}_{ML} - \theta^*) + r$$

with some random r, such that

$$|r| \leq \sup_{\theta, i, j, k} \left| \frac{\partial^3}{\partial \theta_i \partial \theta_j \partial \theta_k} \right| |\widehat{\theta}_{ML} - \theta|^2.$$

Hence, from Equation (61), $r = o_{\mathbb{P}_{|\Sigma}}(|\widehat{\theta}_{ML} - \theta^*|)$. We then have

$$-\frac{\partial}{\partial \theta_k} L_{\theta^*} = \left[\left(\frac{\partial}{\partial \theta} \frac{\partial}{\partial \theta_k} L_{\theta^*} \right)^T + o_{\mathbb{P}_{|\Sigma}}(1) \right] \left(\widehat{\theta}_{ML} - \theta^* \right),$$

an so

$$\left(\widehat{\theta}_{ML} - \theta^* \right) = - \left[\left(\frac{\partial}{\partial \theta} \frac{\partial}{\partial \theta_k} L_{\theta^*} \right)^T + o_{\mathbb{P}_{|\Sigma}}(1) \right]^{-1} \frac{\partial}{\partial \theta_k} L_{\theta^*}. \tag{62}$$

Hence, using Slutsky lemma, Equation 59 and Equation 56, a.s.

$$\sqrt{n} \left(\widehat{\theta}_{ML} - \theta^* \right) \xrightarrow[n \to +\infty]{\mathcal{L}_{|\Sigma}} \mathcal{N}\left(0, (2M_\infty)^{-1}(4M_\infty)(2M_\infty)^{-1}\right) = \mathcal{N}\left(0, M_\infty^{-1}\right). \tag{63}$$

Moreover, using Equation (53), we have

$$\sqrt{n} M_{ML}^{\frac{1}{2}} \left(\widehat{\theta}_{ML} - \theta^* \right) \xrightarrow[n \to +\infty]{\mathcal{L}_{|\Sigma}} \mathcal{N}(0, I_3). \tag{64}$$

Hence, using dominated convergence theorem on Σ, we have

$$\sqrt{n} M_{ML}^{\frac{1}{2}} \left(\widehat{\theta}_{ML} - \theta^* \right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0, I_3). \tag{65}$$

To conclude, we have found a subsequence such that, after extracting,

$$\left| \mathbb{E}\left[g\left(\sqrt{n} M_{ML}^{\frac{1}{2}} (\widehat{\theta}_{ML} - \theta^* \right) \right] - \mathbb{E}(g(U)) \right| \geq \xi, \tag{66}$$

which is in contradiction with Equation (52). □

6.2.4 Proof of Theorem 4

Proof. Let $\sigma \in S_\infty$. We have:

$$\left|\widehat{Y}_{\widehat{\theta}_{ML}}(\sigma) - \widehat{Y}_{\theta^*}(\sigma)\right| \leq \sup_{\theta \in \Theta} \left|\frac{\partial}{\partial \theta} \widehat{Y}_\theta\right| \left|\widehat{\theta}_{ML} - \theta^*\right| \tag{67}$$

From Theorem 2, it is enough to show that, for $i \in \{1, 2, 3\}$

$$\left|\sup_{\theta \in \Theta} \frac{\partial}{\partial \theta_i} \hat{Y}_\theta(\sigma)\right| = O_{\mathbb{P}}(1). \tag{68}$$

From a version of Sobolev embedding theorem ($W^{1,4}(\Theta) \hookrightarrow L^\infty(\Theta)$, see Theorem 4.12, part I, case A in [1]), there exists a finite constant A_Θ depending only on Θ such that

$$\sup_{\theta \in \Theta} \left|\frac{\partial}{\partial \theta_i} \widehat{Y}_\theta(\sigma)\right| \quad \leq \quad A_\theta \int_\Theta \left|\frac{\partial}{\partial \theta_i} \widehat{Y}_\theta(\sigma)\right|^4 d\theta + A_\theta \sum_{j=1}^{3} \int_\Theta \left|\frac{\partial^2}{\partial \theta_j \theta_i} \widehat{Y}_\theta(\sigma)\right|^4 d\theta.$$

The rest of the proof consists in showing that these integrals are bounded in probability. We have to compute the derivatives of

$$\widehat{Y}_\theta(\sigma) = r_\theta^T(\sigma) R_\theta^{-1} y$$

with respect to θ. Thus, we can write these fist and second derivatives as a sum of $w_\theta^T(\sigma) W_\theta y$ where $w_\theta(\sigma)$ is of the form $r_\theta(\sigma)$ or $\frac{\partial}{\partial \theta_i} r_\theta(\sigma)$ of $\frac{\partial^2}{\partial \theta_j \theta_i} r_\theta(\sigma)$ and W_θ is product of the matrices R_θ^{-1}, $\frac{\partial}{\partial \theta_i} R_\theta$ and $\frac{\partial^2}{\partial \theta_j \theta_i} R_\theta$. It is sufficient to show that

$$\int_\Theta \left|w_\theta^T(\sigma) W_\theta y\right|^4 d_\theta = O_{\mathbb{P}}(1). \tag{69}$$

From Fubini-Tonelli Theorem (see [5]), we have

$$\mathbb{E}\left(\int_\Theta \left|w_\theta^T(\sigma) W_\theta y\right|^4 d_\theta\right) = \int_\Theta \mathbb{E}\left(\left|w_\theta^T(\sigma) W_\theta y\right|^4\right) d_\theta.$$

There exists a constant c so that for X a centred Gaussian random variable

$$\mathbb{E}\left(|X|^4\right) = c \mathbb{V}(X)^2,$$

hence

$$\begin{aligned}
\mathbb{E}\left(\int_\Theta \left|w_\theta^T(\sigma) W_\theta y\right|^4 d_\theta |\Sigma\right) &= C \int_\Theta \mathbb{V}\left(w_\theta^T(\sigma) W_\theta y |\Sigma\right)^2 d_\theta \\
&= c \int_\Theta \left(w_\theta^T(\sigma) W_\theta R_\theta^* W_\theta(\sigma) w_\theta(\sigma)\right)^2 d_\theta.
\end{aligned}$$

From Lemma 3, there exists $B < \infty$ such that, a.s.

$$\sup_\theta \|W_\theta R_{\theta^*} W_\theta\| < B.$$

Thus

$$E\left(\int_{\Theta}\left|w_{\theta}^{T}(\sigma)W_{\theta}y\right|^{4}d_{\theta}|\Sigma\right)\leq B^{2}c\int_{\Theta}\|w_{\theta}^{T}(\sigma)\|^{2}d_{\theta}. \tag{70}$$

Finally, for some $\alpha\in[0:2]^3$ such that $|\alpha|\leq 2$, we have

$$\begin{aligned}\sup_{\theta\in\Theta}\|w_{\theta}^{T}(\sigma)\|^{2} &= \sup_{\theta}\sum_{i=1}^{n}\left(\frac{\partial^{|\alpha|}}{\partial\theta^{\alpha}}K_{\theta}(\sigma,\sigma_i)\right)^{2}\\ &\leq C<+\infty,\end{aligned}$$

where C can be chosen independently of Σ, from condition 1 and using the proof of Lemma 3. We concludes the proof saying that the measure of Θ is finite. □

6.2.5 Proofs of lemmas

Proof of Lemma 2.

Proof. R_θ is the sum of a symmetric positive matrix and $\theta_3 I_n$. Thus, the eigenvalues are lower-bounded by $\theta_{3,\min}$. □

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 ([13]).
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

$$\left|\frac{\partial^{|\alpha|}K_{\theta_1,\theta_2}(\sigma,\sigma')}{\partial\theta^\alpha}\right|\leq\max(1,\theta_{2,\max})d(\sigma,\sigma')^{\alpha_1}e^{-\theta_{1,\min}d(\sigma,\sigma')}. \tag{71}$$

We conclude using Gershgorin circle theorem ([13]). □

Proof of Lemma 4

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

$$N(x) := \max(4c\theta_{2,\max}|x_1|, 2|x_2|, |x_3|), \tag{72}$$

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

$$\frac{1}{n}\sum_{i,j=1}^{n}(K'_\theta(\sigma_i,\sigma_j) - K'_{\theta^*}(\sigma_i,\sigma_j))^2. \tag{73}$$

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

$$k_\alpha^\beta \geq 4c\theta_{2,\max}\frac{2 + \ln(\theta_{2,\max}) - \ln(\theta_{2,\min})}{\alpha}. \tag{74}$$

   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 $n \geq k_\alpha$,

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

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

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,

$$\left|\frac{(\theta_1^* - \theta_1)d(\sigma_i, \sigma_{i+1}) + \ln(\theta_2) - \ln(\theta_2^*)}{2}\right| \geq \frac{\alpha}{8\theta_{2,\max}}, \tag{75}$$

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

$$C_\alpha := \min\left(C_{1,\alpha}, C_{2,\alpha}, \frac{\alpha^2}{2}\right), \tag{76}$$

we have

$$\inf_{N(\theta-\theta^*)\geq\alpha} \frac{1}{n}\sum_{i,j=1}^{n}(K'_\theta(\sigma_i,\sigma_j) - K'_{\theta^*}(\sigma_i,\sigma_j))^2 \geq \frac{n-k_\alpha}{n}C_\alpha. \tag{77}$$

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

$$\liminf_{n\to+\infty} \inf_{\|\theta-\theta^*\|\geq\alpha} \frac{1}{n}\sum_{i,j=1}^{n}(K'_\theta(\sigma_i,\sigma_j) - K'_{\theta^*}(\sigma_i,\sigma_j))^2 \geq C_{\alpha/h} > 0. \tag{78}$$

□

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.

□

References

[1] R. A. Adams and J. J. Fournier. *Sobolev spaces*, volume 140. Academic press, 2003.

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

[3] 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.

[4] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[5] P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

[6] M. Christopher. *Logistics & supply chain management*. Pearson UK, 2016.

[7] S. Clémençon, R. Gaudel, and J. 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, Sept. 2011.

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

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

[10] D. Dacunha-Castelle and M. Duflo. *Probability and statistics*, volume 2. Springer Science & Business Media, 2012.

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

[12] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. *SIAM Journal on discrete mathematics*, 17(1):134–160, 2003.

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

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

[15] Y. Jiao and J.-P. Vert. The Kendall and Mallows kernels for permutations. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

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

[17] R. Kondor and M. Barbosa. Ranking with kernels in Fourier space. *Preprint*, 2010.

[18] A. Korba, S. Clémençon, and E. Sibony. A Learning Theory of Ranking Aggregation. In *Artificial Intelligence and Statistics*, pages 1001–1010, 2017.

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

[20] H. Mania, A. Ramdas, M. J. Wainwright, M. I. Jordan, and B. Recht. On kernel methods for covariates that are rankings. *arXiv:1603.08035 [cs, stat]*, Mar. 2016. arXiv: 1603.08035.

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

[22] C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(Dec):2651–2667, 2006.

[23] E. Sibony, S. Clemençon, and J. Jakubowicz. Mra-based statistical learning from incomplete rankings. In *International Conference on Machine Learning*, pages 1432–1441, 2015.

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

[25] J.-P. Vert. Learning on the symmetric group. *Slides*, 2017.

[26] J. Yu, R. Buyya, and C. K. Tham. Cost-based scheduling of scientific workflow applications on utility grids. In *e-Science and Grid Computing, 2005. First International Conference on*, pages 8–pp. Ieee, 2005.