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

April 19, 2019

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. Analysing observations which are not quantitative variables but rankings has been often studied in social sciences. It has also become a popular problem in statistical learning thanks 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. More

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

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

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

Our aim is to predict outputs corresponding to permutations inputs. For instance, the permutation input can correspond to an ordering of tasks, in applications. In a workflow management system, there may be a large number of tasks that 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 [31], and logistics [7].

Another example of application is given by the maintenance of machines in a 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.

Another interesting case of output corresponding to a permutation input is of the form $\max_{x \in X} f(\sigma, x)$, where f is a function both acting on the permutation σ and some external variable x. This output corresponds to a worst case for the performance or the cost given by the permutation σ. Classical examples of this kind of output are the max distance criterion for Latin Hypercube Designs [26, 29] and the robust deviation for a tour in the robust traveling salesman problem [28]. In Section 3.3, we discuss and address the example of the max distance criterion.

In this paper, we will be in the framework of Gaussian processes indexed by S_N. 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 (symmetric positive kernel), 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 symmetric positive 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* [21], clustering rankings (see [8]) or kernels on rankings

for supervised learning. Constructing kernels over the set of permutations has been studied following several different ways. In [19], Kondor provides results about kernels in non-commutative finite groups and constructs *diffusion kernels* (which are positive definite) on S_N. These diffusion kernels are based on a discrete notion of neighborliness. Notice that the kernels considered therein are quite different from those considered in this paper. Furthermore, the diffusion kernels are not in general covariance functions because of their tricky dependency on permutations. Recently, [17] proves that the Kendall and Mallow's kernels are positive definite. Further, [24] extends this study characterizing both the feature spaces and the spectral properties associated with these two kernels. A real data set [9] on rankings is studied in [24]. The authors used a kernel regression to predict the age of a participant with his/her order of preference of six sources of news regarding scientific developments: TV, radio, newspapers and magazines, scientific magazines, the internet, school/university.

There are applications where not all of the items in (1) are ranked. Rather, a partial ranking is given (see for example the "sushi" dataset available at `http://www.kamishima.net` or movie datasets). The papers [20] and [17] provide kernels on partial rankings and deal with the complexity reduction of their computation.

The goal in this paper is threefold: first we define Gaussian processes indexed by S_N by providing a wide class of covariance kernels. We generalize previous results on the Mallow's kernel (see [17]). Second, we consider the Kriging models (see for instance [30]) that consist in inferring the values of a Gaussian random field given observations at a finite set of observation points. Here, the observations points are permutations. We study the asymptotic properties of the maximum likelihood estimator of the parameters of the covariance function. We also prove the asymptotic accuracy of the Kriging prediction under the estimated covariance parameters. Further, we provide simulations that illustrate the very good performances of the proposed kernels. Finally, we provide an application to Gaussian process based optimization of Latin Hypercube Designs. Last, we show that the Gaussian process framework may be adapted to the cases of learning with partially observed rankings. We define a class of covariance kernels on partial rankings, for which we show how to reduce the computation complexity. In simulations, we show that our suggested kernels yield more efficient Gaussian process predictions than the kernels given in [17].

The paper falls into the following parts. In Section 2, we recall some facts on S_N and provide some covariance kernels on this set. Asymptotic results on the estimation of the covariance function are presented in Section 3. Section 3 also contains an application to the optimization of Latin Hypercube Designs. Section 4 provides new covariance kernels for partial rankings with a comparison with the ones given in [17] in a numerical experiment. Section 5 concludes the paper. The proofs are all postponed to the appendix.

2 Covariance model for rankings

Recall that we define S_N as the set of all permutations on $E_N := \{1, \dots, N\}$. We aim at constructing kernels, or covariance functions, on S_N. We will base these kernels on the three following distances on S_N (see [13]). For any permutations π and σ of S_N let

- The Kendall's tau distance be defined by

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

  This distance counts the number of pairs on which the permutations disagree in ranking.

- The Hamming distance be defined by

$$d_H(\pi, \sigma) := \sum_{i=1}^{N} \mathbb{1}_{\pi(i)\neq\sigma(i)}. \quad (4)$$

- The Spearman's footrule distance be defined by

$$d_S(\pi, \sigma) := \sum_{i=1}^{N} |\pi(i) - \sigma(i)|. \quad (5)$$

These three 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 [13].

We aim at defining a Gaussian process indexed by permutations. Notice that, generally speaking, using the abstract Kolmogorov construction (see for example [12] 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

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

Of course, here the frame is much simpler as S_N is finite ($|S_N| = N!$), and the Gaussian distribution is obviously completely determined by its mean and covariance matrix. Hence, if we assume that the process is centered, we only have to build a covariance function on S_N. First, we recall the definition of a positive definite kernel on an abstract 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 say that K is 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.

These notions are 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

$$f = \sum_{\sigma \in S_N} a_\sigma K(., \sigma), \tag{6}$$

and K is of course an *universal kernel* (see [27]).

We now provide two different parametric families of covariance kernels. The members of these families have the general form

$$K_{\theta_1,\theta_2}(\sigma, \sigma') := \theta_2 \exp\left(-\theta_1 d(\sigma, \sigma')\right), \quad (\theta_1, \theta_2 > 0), \tag{7}$$

and

$$K_{\theta_1,\theta_2,\theta_3}(\sigma, \sigma') := \theta_2 \exp\left(-\theta_1 d(\sigma, \sigma')^{\theta_3}\right), \quad (\theta_1,\ \theta_2 > 0,\ \theta_3 \in [0, 1]). \tag{8}$$

Here, d is one of the three distances defined in (3), (4) and (5). More precisely, for the Kendall's (resp. Hamming's and Spearman's footrule) distance let $K^{\tau}_{\theta_1,\theta_2(,\theta_3)}$ (resp. $K^{H}_{\theta_1,\theta_2(,\theta_3)}$, $K^{S}_{\theta_1,\theta_2(,\theta_3)}$) be the corresponding covariance function. For concision, sometimes we will write $K_{\theta_1,\theta_2(,\theta_3)}$ (resp. d) for one of these three kernels (resp. distances).

We show in the next proposition that K_{θ_1,θ_2} is strictly positive definite.

Proposition 1. *For all $\theta_1 > 0$ and $\theta_2 > 0$, $K^{\tau}_{\theta_1,\theta_2}$, $K^{H}_{\theta_1,\theta_2}$, $K^{S}_{\theta_1,\theta_2}$ are strictly positive definite kernels on S_N.*

Remark 1. *In [24], the strictly positive definiteness of the Mallow's kernel, corresponding to $K^{\tau}_{\theta_1,\theta_2}$, is also shown. Our proof of Proposition 1 seems more direct than the one given in [24].*

We also have a similar result for $K_{\theta_1,\theta_2,\theta_3}$.

Proposition 2. *For all $\theta_1 > 0$, $\theta_2 \geq 0$ and $\theta_3 \in [0, 1]$, the maps $K^{\tau}_{\theta_1,\theta_2,\theta_3}$, $K^{H}_{\theta_1,\theta_2,\theta_3}$, $K^{S}_{\theta_1,\theta_2,\theta_3}$ are positive definite kernel on S_N.*

Propositions 1 and 2 enable to define Gaussian processes indexed by permutations.

3 Gaussian fields on the symmetric group

3.1 Asymptotic results

Let us consider a Gaussian process Y indexed by $\sigma \in S_N$, with zero mean and covariance function K_*. In a parametric setting, a classical assumption is that the

covariance function K_* belongs to some parametric set of the form

$$\{K_\theta; \theta \in \Theta\}, \tag{9}$$

where $\Theta \subset \mathbb{R}^p$ is given and for all $\theta \in \Theta$, K_θ is a covariance function. The hyperparameter 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 $(Y(\sigma_i) + \varepsilon_i, \sigma_i)$ for $i = 1, \ldots, n$, where $(\varepsilon_i)_i$ is an independent Gaussian white noise. Let us consider a sample of random permutations $\Sigma = (\sigma_1, \sigma_2, \cdots, \sigma_n) \in S_N$. Assume that we observe Σ and a random vector $y = (y_1, y_2, \cdots, y_n)^T$ defined by, for $i \leq N$,

$$y_i = Y(\sigma_i) + \varepsilon_i. \tag{10}$$

Here, Y is Gaussian process indexed by S_N and independent of Σ. We assume that Y is centered with covariance function $K_{\theta_1^*, \theta_2^*}$ (see (7) in Section 2) and that $(\varepsilon_i)_{i \leq n} \sim \mathcal{N}(0, \theta_3^* I_n)$. Y is the unknown process to predict and ε is an additive white noise. Notice that θ_3 denotes here the variance of the nugget effect while it is a power in Section 2. We keep the same name in order to use the compact notation θ for the hyperparameter of the model. The Gaussian process Y is stationary in the sense that for all $\sigma_1, \cdots, \sigma_n \in S_N$ and for all $\tau \in S_N$, the finite-dimensional distribution of Y 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 $\theta^* := (\theta_1^*, \theta_2^*, \theta_3^*)$. 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 estimator is defined as

$$\widehat{\theta}_{ML} = \widehat{\theta}_n \in \arg\min_{\theta \in \Theta} L_\theta \tag{11}$$

with

$$L_\theta := \frac{1}{n} \ln(\det R_\theta) + \frac{1}{n} y^t R_\theta^{-1} y, \tag{12}$$

where $R_\theta = [K_{\theta_1, \theta_2}(\sigma_i, \sigma_j) + \theta_3 \mathbb{1}_{i=j}]_{1 \leq i,j \leq n}$ is invertible since $\theta_3 > 0$. We assume that $\Theta \subset \prod_{i=1}^3 [\theta_{i,min}, \theta_{i,max}]$ for some given $0 < \theta_{i,min} \leq \theta_{i,max} < \infty$ $(i = 1, 2, 3)$.

When considering the asymptotic behaviour of the maximum likelihood estimator, two different frameworks can be studied: fixed domain and increasing domain asymptotic [30]. 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 $|i - j|$, (thus we need $N \to +\infty$). 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 asymptotically normal [25, 10, 11, 2]. Moreover, prediction performed using the estimated covariance parameter $\hat{\theta}_n$ 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.

We will consider hereafter the increasing-domain framework. We thus consider a sequence of Gaussian processes Y_n on S_{N_n}, with $N_n \underset{n\to+\infty}{\longrightarrow} +\infty$ and with $(\sigma_i^{(n)})_{i\leq n} \subset S_{N_n}$ are the observation points. However, for the sake of simplicity, we only write Y and $(\sigma_i)_{i\leq n}$ and the dependency on n is implicit. 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 $\forall i, j$, $d(\sigma_i, \sigma_j) \geq |i-j|^\beta$.
2. Condition 2: There exists $c > 0$ such that $\forall i,\ d(\sigma_i, \sigma_{i+1}) \leq c$.

Notice that β and c are assumed to be independent on n.

These conditions are natural under increasing-domain asymptotics. Indeed, Condition 1 provides asymptotic independence for pairs of observations with asymptotically distant indices. It allows to show 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 can be ensured with particular choices of sampling schemes for $(\sigma_1, \cdots, \sigma_n)$ (using the distances previously discussed).

As an example consider the following setting. We fix $k \in \mathbb{N}$. For $n \in \mathbb{N}, i \in [1:n]$, we choose $\sigma_i^{(n)} = \sigma_i = \tau_i c_i \in S_{k+n}$ (we have $N_n = k+n$) with $\tau_i \in S_k \times id_{[k+1:n+k]} := \{\sigma \in S_{n+k} |\ \sigma_{|[k+1:n+k]} = id\}$ a random permutation such that $(\tau_i)_i$ are independent (we do not make further assumptions on the law of τ_i). Let $c_i = (i+k \quad i+k-1 \quad \cdots \quad 1)$ the cycle defined by $c_i(1) = i+k$, $c_i(j) = j-1$ if $1 < j \leq i+k$ and $c_i(j) = j$ if $j > i+k$. Then, σ_i is a permutation such that $\sigma_i(1) = i+k$, $\sigma_i(j)$ is a random variable in $[2:k]$ if $1 < j \leq k+1$, $\sigma_i(j) = j-1$ if $k+1 < j \leq i+k$ and $\sigma_i(j) = j$ if $j > i+k$. 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. Indeed, the three distances in S_k are upper-bounded by $k(k-1)/2$, k and k^2 respectively.

The following theorems give both the consistency and the asymptotic normality of the estimator when the number of observations increases.

Theorem 1. *Let $\widehat{\theta}_{ML}$ be defined as in* (11)*. Assume that Conditions 1 and 2 hold. Then,*

$$\widehat{\theta}_{ML} \underset{n\to+\infty}{\xrightarrow{\mathbb{P}}} \theta^*. \tag{13}$$

Theorem 2. *Under the assumptions of Theorem 1, let M_{ML} be the 3×3 matrix defined by*

$$(M_{ML})_{i,j} = \frac{1}{2n} \operatorname{Tr}\left(R_{\theta^*}^{-1} \frac{\partial R_{\theta^*}}{\partial \theta_i} R_{\theta^*}^{-1} \frac{\partial R_{\theta^*}}{\partial \theta_j} \right). \tag{14}$$

Then

$$\sqrt{n} M_{ML}^{\frac{1}{2}} \left(\widehat{\theta}_{ML} - \theta^* \right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0, I_3). \tag{15}$$

Furthermore,

$$0 < \liminf_{n \to \infty} \lambda_{\min}(M_{ML}) \leq \limsup_{n \to \infty} \lambda_{\max}(M_{ML}) < +\infty, \tag{16}$$

where $\lambda_{\min}(M_{ML})$ (resp. $\lambda_{\max}(M_{ML})$) is the smallest (resp. largest) eigenvalue of M_{ML}.

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

$$\hat{Y}_{\widehat{\theta}_n}(\overline{\sigma}_n) = r_{\widehat{\theta}_n}^t(\overline{\sigma}_n) R_{\widehat{\theta}_n}^{-1} y \tag{17}$$

with

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

We point out that $\hat{Y}_{\widehat{\theta}_n}(\overline{\sigma}_n)$ is the conditional expectation of $Y(\overline{\sigma}_n)$ 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 behaves as if the true covariance parameter were known.

Theorem 3. *For any sequence $(\overline{\sigma}_n)_{n \in \mathbb{N}}$, with $\overline{\sigma}_n \in S_{N_n}$ for $n \in \mathbb{N}$, we have*

$$\left| \widehat{Y}_{\widehat{\theta}_{ML}}(\overline{\sigma}_n) - \widehat{Y}_{\theta^*}(\overline{\sigma}_n) \right| = o_{\mathbb{P}}(1). \tag{18}$$

The proofs of Theorems 1, 2 and 3 are given in the appendix. In [2] and [3], similar results for maximum likelihood are given for Gaussian fields indexed on $\mathbb{R}^d$ and on the set of all probability measures on $\mathbb{R}$ (see also [4]). In the appendix, we also discuss the similarities and differences between the proofs of Theorems 1, 2 and 3 and these given in [2] and [3].

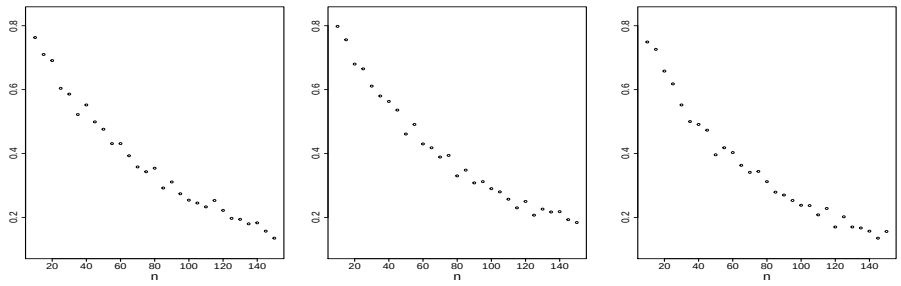

Figure 1: Monte Carlo estimates of $\mathbb{P}(\|\widehat{\theta}_n - \theta^*\| > 0.5)$ for different values of n, the number of observations, with $\theta^* = (0.1, 0.8, 0.3)$ and Kendall's tau distance, the Hamming distance and the Spearman's footrule distance from left to right.

3.2 Numerical experiments

As an illustration of Theorem 1, we provide a numerical illustration showing that the maximum likelihood is consistent. We generated the observations as discussed in Section 3 with $k = 3$. We recall that $N_n = k + n$ and $\sigma_i = \tau_i(i + k \; i + k - 1 \; \cdots 1) \in S_{k+n}$ where $\tau_i \in S_k \times id_{[k+1:k+n]}$ is 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 ranging from 20, 60 to 150. These densities have been estimated with a sample of 1000 values 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 $n = 150$. Further, we see that for $n = 150$, the Kendall's tau distance seems to give better estimates for 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, for a given $\overline{\sigma}_n$, we display estimates of the probability that the deviation between the prediction of $Y(\overline{\sigma}_n)$ given in (17) with the parameter $\hat{\theta}_n$ and the prediction of $Y(\overline{\sigma}_n)$ with the parameter θ^* exceeds 0.3. Indeed, Theorem 3 ensures us that this probability converges to 0 as $n \to +\infty$.

3.3 Application to the optimization of Latin Hypercube Designs

We consider here an application of Proposition 2 to find the best Latin Hypercube Design (LHD). A LHD is a design of experiments $(X_j)_{j \leq N} \in [0, 1]^d$ where, for each component $i \in [1 : d]$, the projections of $X_1, ..., X_N$ on the component i are

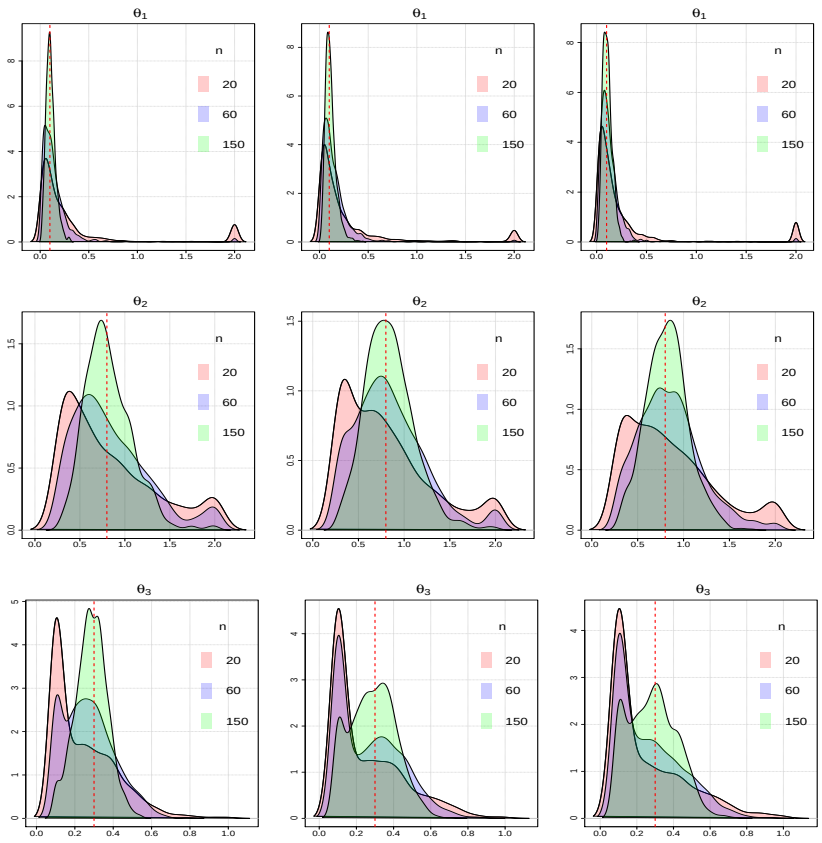

Figure 2: Density of the coordinates of $\widehat{\theta}_n$ for the number of observations $n = 20$ (in red), $n = 60$ (in blue), $n = 150$ (in green) with $\theta^* = (0.1, 0.8, 0.3)$ (represented by the red vertical line). We used the Kendall's tau distance, the Hamming distance and the Spearman's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

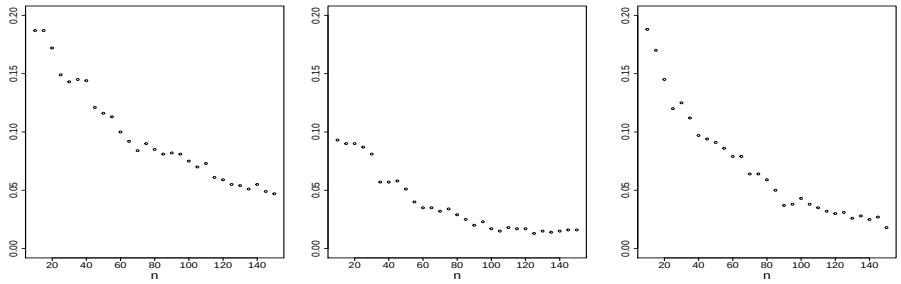

Figure 3: Monte Carlo estimates of $\mathbb{P}\left(\left|\widehat{Y}_{\widehat{\theta}_n}(\overline{\sigma}_n) - \widehat{Y}_{\theta^*}(\overline{\sigma}_n)\right| > 0.3\right)$ for different values of n, the number of observations, with $\theta^* = (0.1, 0.8, 0.3)$, $\overline{\sigma}_n = (1\ 4\ 6)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

equispaced in $[0,1]$ (see [26]). We will consider that each component of one X_j is equal to $k/(N-1)$ for some $k \in [0:N-1]$. So, for each LHD $(X_j)_{j\leq N}$, there exist $\sigma_2, ..., \sigma_d \in S_N$ such that for all $j \in [1:N]$, we have

$$X_j = \left(\frac{j-1}{N-1}, \frac{\sigma_2(j)-1}{N-1}, ..., \frac{\sigma_d(j)-1}{N-1} \right).$$

Hence, there is a bijection between the set of LHD with N points and the set S_N^{d-1}.

Now, if $(X_j)_{j\leq N}$ is a LHD, we can define its measure of space filling quality as

$$f((X_j)_{j\leq N}) = \sup_{x\in[0,1]^d} \min_{j\in[1:N]} \|x - X_j\|,$$

that is the largest distance of a point of $[0,1]^d$ to $(X_j)_{j\leq N}$. We remark that LHD minimizing f are called minimax [29]. Our aim is to find a minimax LHD $(X_j^*)_{j\leq N}$. However, given a LHD $(X_j)_{j\leq N}$, its quality $f((X_j)_{j\leq N})$ is not an obvious quantity and its computation is expensive.

To estimate this quantity, we suggest to generate N_{tot} random points $(x_l)_{l\leq N_{tot}}$ uniformly on $[0,1]^d$, to compute their distance to the LHD an to take the maximum value. This estimation is costly (because of the large number N_{tot}) and noisy (because of the randomness of the points $(x_l)_{l\leq N_{tot}}$). Thus, we suggest to model f by a Gaussian process realization and to apply the Expected Improvement (EI) strategy [18].

We thus assume that the unknown function f to minimize is a realization of a Gaussian process. We have to find a positive definite kernel on S_N^{d-1}. Thanks to Proposition 2, we have three positive definite kernels on S_N, thus on S_N^{d-1} (taking the tensor product of these kernels). Thus, we apply the EI strategy with these three kernels to find the best LHD with $N_{\max}$ calls to the function f. The $N_{\max}/2$ first LHDs are generated uniformly on S_N^{d-1} and the other ones are generated sequentially by following the EI strategy. The parameters of the covariance functions are estimated by maximum likelihood at each step. We refer to [18] for more details on EI.

In this experiment, to compare performances, we apply 5 methods:

- Random sampling, to generate $N_{\max}$ LHDs of the form $\{(X_j^{(i)})_{j\leq N};\ i \leq N_{\max}\}$ by generating $\sigma_2, ..., \sigma_d$ uniformly and independently;
- Simulated annealing, choosing that two LHDs $(\sigma_j)_{2\leq j\leq d}$ and $(\sigma'_j)_{2\leq j\leq d}$ are neighbours if there exist transpositions $\tau_2, ..., \tau_d$ such that for all $j \in [2:d]$, we have $\sigma'_j = \sigma_j\tau_j$;
- EI with Kendall distance;
- EI with Hamming distance;

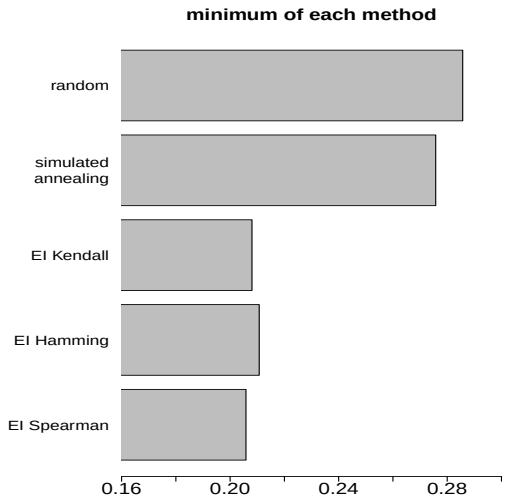

Figure 4: Minimal quality of LHD found by the five methods.

- EI with Spearman distance.

For each method, the performance indicator is $\min_{i=1,\dots,N_{\max}} f((X_j^{(i)})_{j\leq N})$. Here, we take $d = 3$, $N = 15$, $N_{\max} = 200$ and $N_{tot} = 27 \times 10^6$.

We can see in Figure 4 that the best LHDs are found by EI. The simulated annealing is slightly better than random sampling.

We display in Figure 5 the distributions of the qualities $\{f((X_j^{(i)})_{j\leq N}); i \leq N_{\max}\}$ for the five methods. We can notice that the simulated annealing does not explore the set of all the LHDs and does not find the best minimum. EI performs minimisation and exploration to find better minima. We can then provide the best LHD of EI. For example, the best LHD found by EI with the Kendall distance is the LHD given by the permutations

$$\begin{aligned} \sigma_2 &= (1, 4, 2, 14, 3, 13, 6, 7, 9, 12, 10, 5, 8, 11, 15), \\ \sigma_3 &= (4, 5, 13, 10, 11, 3, 6, 1, 8, 9, 7, 2, 14, 15, 12). \end{aligned}$$

To conclude, the kernels on permutations provided in Section 2 enable us to use EI that gives much better results than simulated annealing or random sampling to find the best LHD.

4 Covariance model for partial ranking

4.1 A new kernel on partial rankings

In application, it can happen that partial rankings rather than complete rankings are observed. A partial ranking aims at giving an order of preference between different

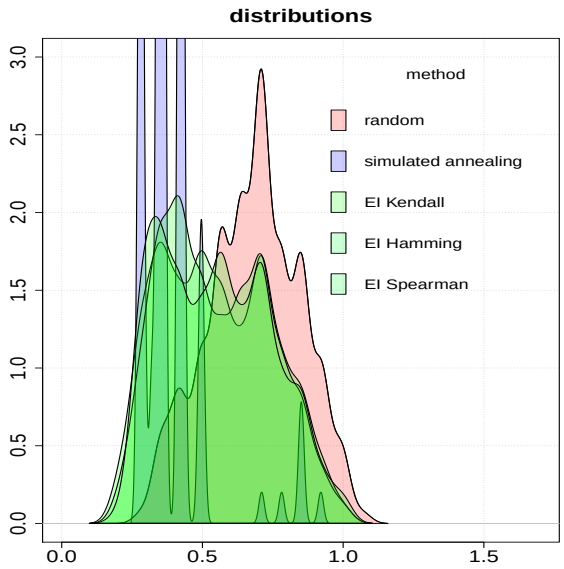

Figure 5: Distributions of the quality of LHDs for the five methods.

elements of X without comparing all the pairs in X. Hence, a partial ranking R is a statement of the form

$$X_1 \succ X_2 \succ \cdots \succ X_m, \tag{19}$$

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 is 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

$$\begin{aligned} E_R \;:=\; & \{\sigma \in S_N :\; \sigma(i_1) < \sigma(i_2) < \cdots < \sigma(i_m) \\ & \text{for any choice of } (x_{i_1}, \cdots, x_{i_m}) \in X_1 \times \cdots \times X_m \,\}. \end{aligned} \tag{20}$$

In the statistical literature, there is a natural way to extend a positive definite kernel K on S_N to the set of partial rankings (see [20], [17]). To do so, one considers for R and R' two partial rankings the following averaged kernel

$$\mathcal{K}(R, R') := \frac{1}{|E_R||E_{R'}|} \sum_{\sigma \in E_R} \sum_{\sigma' \in E_{R'}} K(\sigma, \sigma'). \tag{21}$$

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 [16]. Indeed, if $R_1, \cdots, R_n$ are partial rankings and if $(a_1, \cdots, a_n) \neq 0$, then

$$\sum_{i,j=1}^{n} a_i a_j \mathcal{K}(R_i, R_j) = \sum_{\sigma, \sigma' \in S_N} b_\sigma b_{\sigma'} K(\sigma, \sigma'), \tag{22}$$

where we set

$$b_\sigma := \sum_{i,\ \sigma \in R_i} \frac{a_i}{|E_{R_i}|}. \tag{23}$$

Observe that the computation of $\mathcal{K}$ is very costly. Indeed, we have to sum over $|E_R||E_{R'}|$ permutations. Several works aim to reduce the computation cost of this kernel (see [20, 22, 23]). However, its efficient computation remains an issue.

In the following, we provide another way to extend the kernels $K_{\theta_1,\theta_2,\theta_3}$ to partial rankings. We will provide computational simplifications for this extension. 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

$$d_{\text{avg}}(R, R') := \frac{1}{|E_R||E_{R'}|} \sum_{\sigma \in E_R} \sum_{\sigma \in E_{R'}} d(\sigma, \sigma'). \tag{24}$$

Since $d_{\text{avg}}(R, R) \neq 0$ in general, we need to define d_{partial} as follows

$$d_{\text{partial}}(R, R') := d_{\text{avg}}(R, R') - \frac{1}{2} d_{\text{avg}}(R, R) - \frac{1}{2} d_{\text{avg}}(R', R'). \tag{25}$$

Proposition 3. *$d_{partial}^{\frac{1}{2}}$ is a pseudometric on partial rankings (i.e. it satisfies the positivity, the symmetry, the triangular inequality and is equal to 0 on the diagonal $\{(R, R),\ R$ is a partial ranking$\}$).*

We further define

$$\mathcal{K}_{\theta_1,\theta_2,\theta_3}(R, R') := \theta_2 \exp(-\theta_1 d_{\text{partial}}(R, R')^{\theta_3}). \tag{26}$$

The next proposition warrants that this last function is in fact a covariance kernel, which will later enable to define Gaussian processes on partial rankings.

Proposition 4. *$\mathcal{K}_{\theta_1,\theta_2,\theta_3}$ is a positive definite kernel for the Kendall's tau distance, the Hamming distance and the Spearman's footrule distance.*

4.2 Kernel computation in partial ranking

At a first glance, the computation of the kernel $\mathcal{K}_{\theta_1,\theta_2,\theta_3}(R, R')$ on partial rankings may still appear very costly dur to the evaluation of d_{partial}. Indeed, we have to sum $|E_R||E_{R'}|$ elements for $d_{\text{avg}}(R, R')$, $|E_R|^2$ elements for $d_{\text{avg}}(R, R)$ and $|E_{R'}|^2$ elements for $d_{\text{avg}}(R', R')$. 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 $d_{S,\text{avg}}$) for the average distance in (24) 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

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_k} \succ X_{rest}, \tag{27}$$

where $X_{rest} := X \setminus \{x_{i_1}, \cdots, x_{i_k}\}$. It can be seen as the "highest rankings". In order to alleviate the notations, let just write $I = (i_1, \cdots, i_k)$ for this top-k partial ranking. The following proposition 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. $\tilde{I'}$).

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 $u'_1 = 4$.*

Proposition 5. *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 2. *Assume that $n = 7$, $I = (3, 2, 1, 4, 5)$ and $I' = (3, 5, 1, 6, 2)$. Proposition 5 leads to*

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

To compute the pseudometric d_{partial} defined in (25), we also need to compute $d_{\tau,\text{avg}}$ on the diagonal $\{(I,I)|\ I \text{ is a top-}k \text{ partial ranking}\}$. The following corollary gives these computations.

Corollary 1. *Let I be a top-k partial ranking. Then,*

$$\begin{aligned} d_{\tau,avg}(I,I) &= \frac{1}{2}\binom{N-k}{2}, \\ d_{H,avg}(I,I) &= N-k-1, \\ d_{S,avg}(I,I) &= (N-k)(N-k-1) + \frac{(N-k-1)(2N-2k-1)}{3}. \end{aligned}$$

In the case of the Hamming distance, we may step ahead and provide a simpler computational formula for the average distance between two partial rankings whenever their associated partitions share the same number of members (see Proposition 6 below). More precisely let R_1 and R_2 be two partial rankings such that

$$R_i = X_1^i \succ \cdots \succ X_k^i,\ i = 1,2, \tag{28}$$

assume also that for $j = 1,\cdots,k$, $|X_j^1| = |X_j^2|$ 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 lying in $\left[\sum_{l=1}^{j-1}\gamma_l + 1, \sum_{l=1}^{j}\gamma_l\right]$. Set further,

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

where S_{Γ_i} is the set of permutations on Γ_i. Notice that S_γ is nothing more than the subgroup of S_n letting invariant the sets 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 these extra notations and definitions, we are now able to compute $d_{H,\text{avg}}(R_1,R_2)$.

Proposition 6. *In the previous setting, we have*

$$d_{H,avg}(R_1,R_2) = |\{i,\ \Gamma(\pi_1(i)) \neq \Gamma(\pi_2(i))\}| + \sum_{j=1}^{k} \frac{\gamma_j}{N}(\gamma_j - 1), \tag{29}$$

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

Note that in (29), the term $|\{i,\ \Gamma(\pi_1(i)) \neq \Gamma(\pi_2(i))\}|$ counts the number of item $i \in [1:N]$ that are ranked differently in R_1 and R_2.

4.3 Numerical experiments

We have proposed in Section 4.1 a new kernel $\mathcal{K}_{\theta_1,\theta_2,\theta_3}$ defined by (26) on partial rankings. We show in Section 4.2 that in several cases (for example with top-k

kernel	$\mathcal{K}^{\tau}_{\theta_1,\theta_2,\theta_3}$	$\mathcal{K}^{H}_{\theta_1,\theta_2,\theta_3}$	$\mathcal{K}^{S}_{\theta_1,\theta_2,\theta_3}$	$\mathcal{K}_{\theta_1}$
rate	0.902	0.904	0.912	0.928
R^2	0.887	0.996	0.996	0.070

Table 1: Rate of test points that are in the 90% confidence interval and coefficient of determination for the four kernels.

partial rankings), we can reduce drastically the computation of this kernel. Another direction is given in [17] by considering the averaged Kendall kernel and reducing the computation of this kernel on top-k partial rankings. This kernel is available on the R package `kernrank`. We write $\mathcal{K}$ the averaged Kendall kernel, and we define $\mathcal{K}_{\theta_1} := \theta_1 \mathcal{K}$.

In this section, we compare our new kernel $\mathcal{K}_{\theta_1,\theta_2,\theta_3}$ with the averaged Kendall kernel $\mathcal{K}_{\theta_1}$ in a numerical experiment where an objective function indexed by top-k partial rankings is predicted, by Kriging. We take $N = 10$ and for simplicity, we take the same value $k = 4$ for all the top-k partial rankings. For a top-k partial ranking $I = (i_1, i_2, i_3, i_4)$, the objective function to predict is $f(I) := 2i_1 + i_2 - i_3 - 2i_4$. We make 500 noisy observations $(y_i)_{i\leq 500}$ with $y_i = f(I_i) + \varepsilon_i$, where $(I_i)_{i\leq 500}$ are i.i.d. uniformly distributed top-k partial rankings and $(\varepsilon_i)_{i\leq 500}$ are i.i.d. $\mathcal{N}(0, \lambda^2)$, with $\lambda = \frac{1}{2}$. As in Section 3, we estimate (θ, λ) by maximum likelihood. Then, we compute the predictions $(\widehat{y}'_i)_{i\leq 500}$ of $y' = (y'_i)_{i\leq 500}$, with y' the observations corresponding to 500 other test points $(I'_i)_{i\leq 500}$, that are i.i.d. uniform top-k partial rankings.

For the four kernels (our kernel $\mathcal{K}_{\theta_1,\theta_2,\theta_3}$ with the 3 distances and the averaged Kendall kernel $\mathcal{K}_{\theta_1}$), we provide the rate of test points that are in the 90% confidence interval together with the coefficient of determination R^2 of the predictions of the test points. Recall that

$$R^2 := 1 - \frac{\frac{1}{500}\sum_{i=1}^{500} (y'_i - \widehat{y}'_i)^2}{\frac{1}{500}\sum_{i=1}^{500} \left(y'_i - \overline{y'}\right)^2},$$

where $\overline{y'}$ is the average of y'. The results are provided in Table 1.

The rate of test points that are in the 90% confidence interval is close to 90% for the four kernels. We can deduce that the parameters (θ, λ) are well estimated by maximum likelihood, even for the averaged Kendall kernel $\mathcal{K}_{\theta_1}$.

However, we can see that the coefficient of determination of the averaged Kendall kernel $\mathcal{K}_{\theta_1}$ is close to 0. The predictions given by the averaged Kendall kernel $\mathcal{K}_{\theta_1}$ are nearly as bad as predicting with the empirical mean. In the opposite way the coefficient of determination of our kernels is larger than 0.9 for the Kendall distance, and larger than 0.99 for the Hamming distance and the Spearman distance. That means that the prediction given by our kernels are much better than the empirical mean.

To conclude, we provide a class of positive definite kernels $\mathcal{K}_{\theta_1,\theta_2,\theta_3}$ which

seems to be significantly more efficient than the averaged Kendall kernel K_{θ_1}, in the case of Gaussian process models on partial rankings.

5 Conclusion

In this paper, we provide a Gaussian process model for permutations. Following the recent works of [17] and [24], we propose kernels to model the covariance of such processes and show the relevance of such choices. Based on the three distances on the set of permutations, Kendall's tau, Hamming distance and Spearman's footrule distance, we obtain parametric families of relevant covariance models. To show the practical efficiency of these parametric families, we apply them to the optimization of Latin Hypercube Designs. In this framework, we prove under some assumptions on the set of observations, that the parameters of the model can be estimated and the process can be forecast using linear combinations of the observations, with asymptotic efficiency. 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 remark that our asymptotic setting corresponds to the increasing domain asymptotic framework for Gaussian processes on the Euclidean space. It would be interesting to extend our results to more general sets of permutations under designs that do not necessarily satisfy Conditions 1 and 2.

We also show that the Gaussian process framework can be extended to the case of partially observed ranks. This corresponds to many practical use cases. We provide new kernels on partial rankings, together with results that significantly simplify their computation. We show the efficiency of these kernels in simulations. We leave a specific asymptotic study of Gaussian processes indexed by partial rankings open for further research.

Acknowledgement

We are grateful to Jean-Marc Martinez for suggesting us the Latin Hypercube Design application.

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] F. Bachoc, A. Suvorikova, D. Ginsbourger, J.-M. Loubes, and V. Spokoiny. Gaussian processes with multidimensional distribution inputs via optimal transport and hilbertian embedding. *arxiv.org/abs/1805.00753v2*, 2019.

[5] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic Analysis on Semigroups*. Springer, Berlin, 1984.

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

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

[8] 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.

[9] E. Commission. Eurobarometer 55.2 (may-jun 2001), 2012.

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

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

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

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

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

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

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

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

[18] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

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

[20] R. Kondor and M. S. Barbosa. Ranking with kernels in Fourier space. In *COLT*, pages 451–463, 2010.

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

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

[23] M. Lomelí, M. Rowland, A. Gretton, and Z. Ghahramani. Antithetic and Monte Carlo kernel estimators for partial rankings. *Statistics and Computing*, Feb 2019.

[24] H. Mania, A. Ramdas, M. J. Wainwright, M. I. Jordan, and B. Recht. On kernel methods for covariates that are rankings. *Electron. J. Statist.*, 12(2):2537–2577, 2018.

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

[26] M. D. McKay, R. J. Beckman, and W. J. Conover. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):239–245, 1979.

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

[28] R. Montemanni, J. Barta, M. Mastrolilli, and L. M. Gambardella. The robust traveling salesman problem with interval data. *Transportation Science*, 41(3):366–381, 2007.

[29] T. J. Santner, B. J. Williams, W. Notz, and B. J. Williams. *The design and analysis of computer experiments*, volume 1. Springer, 2003.

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

[31] 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.

A Proofs for Sections 2 and 4

Proof of Proposition 1

Proof. We show that 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

$$K(\sigma, \sigma') := e^{-\nu d(\sigma,\sigma')} \tag{30}$$

is a strictly positive definite kernel.

Case of the Kendall's tau distance. It has been shown in Theorem 5 of [24] that K is a strictly positive definite kernel on S_n for the Kendall's tau distance. Nevertheless, we provide here an other shorter and easier proof. Let

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

Further, define

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

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

$$\begin{array}{rccc} N_k: & \{0,1\}^k & \longrightarrow & [1:2^k] \\ & (a_i)_{i\leq k} & \longmapsto & 1 + \sum_{i=1}^k a_i 2^{i-1}. \end{array}$$

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

$$M^{(k)}_{i,j} := \exp\left(-\nu \sum_{l=1}^k |N_k^{-1}(i)_l - N_k^{(-1)}(j)_l|\right), \quad (i,j \in [1:2^k]).$$

is the Kronecker product of k matrices A_ν defined by

$$A_\nu := \begin{pmatrix} 1 & e^{-\nu} \\ e^{-\nu} & 1 \end{pmatrix}, \quad (\nu > 0).$$

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

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

With the same computation, we have

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

We also have

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

and with the same computation,

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

So that we conclude the induction. Using this result with $k = \frac{N(N-1)}{2}$, we have that the matrix $\tilde{M}$ 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 and the Spearman's footrule 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 $(\exp(-\nu|i-j|))_{i,j\in[1:N]}$.

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

□

Lemma 1. *For all the three distances, there exist a constants $d_N \in \mathbb{N}^*$, $C_N \in \mathbb{R}$ and a function $\Phi : S_N \to \mathbb{R}^{d_N}$ such that $d(\sigma,\sigma') = C_N - \langle \Phi(\sigma), \Phi(\sigma')\rangle$. Here $\langle\cdot,\cdot\rangle$ denotes the standard scalar product on $\mathbb{R}^{d_N}$.*

Proof.

- $\frac{N(N-1)}{4} - d_\tau(\sigma,\sigma') = \frac{1}{2}\sum_{i<j} \mathbb{1}_{\sigma(i)<\sigma(j),\, \sigma'(i)<\sigma'(j)} + \mathbb{1}_{\sigma(i)>\sigma(j),\, \sigma'(i)>\sigma'(j)} - \frac{1}{2}\sum_{i<j} \mathbb{1}_{\sigma(i)<\sigma(j),\, \sigma'(i)>\sigma'(j)} + \mathbb{1}_{\sigma(i)>\sigma(j),\, \sigma'(i)<\sigma'(j)} = \langle \Phi(\sigma), \Phi(\sigma')\rangle$ 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

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

□

Proof of Proposition 2

Proof. Let us prove that d is a definite negative kernel, i.e. for all $c_1, ..., c_k \in \mathbb{R}$ such that $\sum_{i=1}^k c_i = 0$, we have $\sum_{i,j=1}^k c_i c_j d(\sigma_i,\sigma_j) \leq 0$. Let $c_1, ..., c_k \in \mathbb{R}$ such that $\sum_{i=1}^k c_i = 0$ and let $\sigma_1, ..., \sigma_k \in S_N$.

$$\sum_{i,j=1}^k c_i c_j d(\sigma_i,\sigma_j) \;=\; C_N \sum_{i,j=1}^k c_i c_j - \sum_{i,j=1}^k c_i c_j \langle \Phi(\sigma_i), \Phi(\sigma_j)\rangle \leq 0.$$

So, d is a negative definite kernel. Hence d^{θ_3} is a definite negative kernel for all $\theta_3 \in [0,1]$. The function $F : t \mapsto \theta_2 \exp(-\theta_1 t)$ is completely monotone, thus, using Schoenberg's theorem (see [5] for the definitions of these notions and Schoenberg's theorem), $K_{\theta_1,\theta_2,\theta_3}$ is a definite positive kernel. □

Proof of Proposition 3

Proof. Let us write

$$\Phi_{\text{avg}} : R \longmapsto \frac{1}{|E_R|} \sum_{\sigma \in E_R} \Phi(\sigma). \tag{31}$$

Then,

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

Thus,

$$\begin{aligned} d_{\text{partial}}(R, R') &= d_{\text{avg}}(R, R') - \frac{1}{2} d_{\text{avg}}(R, R) - \frac{1}{2} d_{\text{avg}}(R', R') \\ &= \frac{1}{2} \left[\left(C_N - d_{\text{avg}}(R, R) \right) + \left(C_N - d_{\text{avg}}(R', R') \right) - 2 \left(C_N - d_{\text{avg}}(R, R') \right) \right] \\ &= \frac{1}{2} \left(\|\Phi_{\text{avg}}(R)\|^2 + \|\Phi_{\text{avg}}(R')\|^2 - 2 \langle \Phi_{\text{avg}}(R), \Phi_{\text{avg}}(R') \rangle \right) \\ &= \|\Phi_{\text{avg}}(R) - \Phi_{\text{avg}}(R')\|^2. \end{aligned}$$

□

Proof of Proposition 4

Proof. Let us prove that d_{partial} is a definite negative kernel. We define

$$D_{\text{avg}}(R, R') := \Phi_{\text{avg}}(R)^T \Phi_{\text{avg}}(R'). \tag{32}$$

Let $(c_1, ..., c_k) \in \mathbb{R}^k$ such that $\sum_{i=1}^k c_i = 0$. We have

$$\begin{aligned}
\sum_{i,j=1}^{k} c_i c_j d_{\text{partial}}(R_i, R_j) &= \sum_{i,j=1}^{k} c_i c_j \left[d_{\text{avg}}(R_i, R_j) - \frac{1}{2} d_{\text{avg}}(R_i, R_i) - \frac{1}{2} d_{\text{avg}}(R_j, R_j) \right] \\
&= \sum_{i,j=1}^{k} c_i c_j d_{\text{avg}}(R_i, R_j) - \frac{1}{2} \sum_{i=1}^{k} c_i d_{\text{avg}}(R_i, R_i) \sum_{j=1}^{k} c_j \\
&\quad - \frac{1}{2} \sum_{j=1}^{k} c_j d_{\text{avg}}(R_j, R_j) \sum_{i=1}^{k} c_i \\
&= \sum_{i,j=1}^{k} c_i c_j d_{\text{avg}}(R_i, R_j) \\
&= \sum_{i,j=1}^{k} c_i c_j \left[C_N - D_{\text{avg}}(R_i, R_j) \right] \\
&= - \sum_{i,j=1}^{k} c_i c_j D_{\text{avg}}(R_i, R_j) \\
&\leq 0.
\end{aligned}$$

So, d_{partial} is a definite negative kernel, and we may conclude as in the proof of Proposition 2.

□

Proof of Proposition 5

Proof. Assume that σ (resp. σ') is a uniform random variable of 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 $\mathbb{E}(d_\tau(\sigma, \sigma'))$. Following the proof of Lemma 3.1 of [14], we have

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

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 now 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

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

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 $\#\{a\in J,\ \sigma(a)>u_l\}=k-u_l-r+l$. 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

$$\sum_{\substack{a\in J,\\ b\in\tilde{I}'}} K_{a,b}(\sigma,\sigma') = r\left(k+\frac{1-r}{2}\right) - \sum_{l=1}^{r} u'_l. \tag{33}$$

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 and does not change the other elements. 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, let compute $\mathbb{E}(d_S(\sigma,\sigma'))$. First, we define

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

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

It remains to compute all the expectations appearing here.

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,

$$\mathbb{E}(A_u(\sigma')) = r\frac{N+k+1}{2} - \sum_{j=1}^{r} u_j. \tag{34}$$

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 5. □

Proof of Proposition 6

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

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

□

B Proofs for Section 3

In the following, let us write $\|.\| = \|.\|_2$ for the operator norm (for a linear mapping of $\mathbb{R}^n$ with the Euclidean norm) of a squared matrix of size n, $\|.\|_F$ for 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]. We write all the proofs to be self-contained.

B.1 Lemmas

The following Lemmas are useful for the proofs of Theorems 1, 2 and 3.

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,*

$$\forall \alpha > 0,\ \liminf_{n\to+\infty} \inf_{\|\theta-\theta^*\|\geq\alpha} \frac{1}{n}\sum_{i,j=1}^{n} (R_{\theta,i,j} - R_{\theta^*,i,j})^2 > 0. \tag{35}$$

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

$$\liminf_{n\to+\infty} \frac{1}{n}\sum_{i,j=1}^{n} \left(\sum_{k=1}^{3} \lambda_i \frac{\partial}{\partial\theta_k} R_{\theta^*,i,j}\right)^2 > 0. \tag{36}$$

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

B.2 Proof of Theorem 1

Proof. Step 1: It suffices to prove that

$$\mathbb{P}\left(\sup_{\theta}|(L_\theta - L_{\theta^*}) - (\mathbb{E}(L_\theta|\Sigma) - \mathbb{E}(L_{\theta^*}|\Sigma))| \geq \epsilon \middle| \Sigma\right) \to_{n\to\infty} 0, \tag{37}$$

and for a fixed $a > 0$,

$$\mathbb{E}(L_\theta|\Sigma) - \mathbb{E}(L_{\theta^*}|\Sigma) \geq a\frac{1}{n}\sum_{i,j=1}^{n}(K_\theta(\sigma_i,\sigma_j) - K_{\theta^*}(\sigma_i,\sigma_j))^2. \tag{38}$$

Indeed, by contradiction, assume that we have (37), (38) but not the consistency of the maximum likelihood estimator. Then, writing the dependency of $\widehat{\theta}$ and $L(\theta)$ with n,

$$\exists\epsilon > 0,\ \exists\alpha > 0,\ \forall n \in \mathbb{N},\ \exists N_n \geq n,\ \mathbb{P}(|\widehat{\theta}_{N_n} - \theta^*| \geq \epsilon) \geq \alpha. \tag{39}$$

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:

$$\begin{aligned}
0 &\geq \inf_{\|\theta-\theta^*\|\geq\epsilon}(L_{N_n}(\theta) - L_{N_n}(\theta^*)) \\
&\geq \inf_{\|\theta-\theta^*\|\geq\epsilon}\mathbb{E}\left(L_{N_n}(\theta) - L_{N_n}(\theta^*)\right) - \sup_{\|\theta-\theta^*\|\geq\epsilon}|(L_\theta - L_{\theta^*}) - (\mathbb{E}(L_\theta) - \mathbb{E}(L_{\theta^*}))| \\
&\geq \inf_{\|\theta-\theta^*\|\geq\epsilon}\mathbb{E}\left(L_{N_n}(\theta) - L_{N_n}(\theta^*)\right) + o_{\mathbb{P}}(1) \\
&\geq a|R_\theta - R_{\theta^*}|^2 + o_{\mathbb{P}}(1) \quad \text{using (37),}
\end{aligned}$$

which is contradicted using (38). It remains to prove (37) and (38).

Step 2: We prove (37).
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 $z:=R_\theta^{-\frac{1}{2}}y$.

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

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

$$\sup_\theta\|L_\theta-\mathbb{E}(L_\theta)\|=o_{\mathbb{P}}(1)\tag{40}$$

Now, let us write $D_{\theta,\theta^*}:=\mathbb{E}(L_\theta|\Sigma)-\mathbb{E}(L_{\theta^*}|\Sigma)$. Thanks to (40),

$$\sup_\theta|L_\theta-L_{\theta^*}-D_\theta|\leq\sup_\theta|L_\theta-\mathbb{E}(L_\theta)|+|L_{\theta^*}-\mathbb{E}(L_{\theta^*})|=o_{\mathbb{P}}(1).\tag{41}$$

Step 3: We prove (38).

We have

$$\mathbb{E}(y^TR_\theta y|\Sigma)=\mathbb{E}(\mathrm{Tr}(y^TR_\theta y)|\Sigma)=\mathbb{E}(\mathrm{Tr}(R_\theta yy^T)|\Sigma))=\mathrm{Tr}(R_\theta\mathbb{E}(y^Ty)).$$

Thus

$$\mathbb{E}(L_\theta|\Sigma)=\frac{1}{n}\sum_{i=1}^n\ln(\det(R_\theta))+\frac{1}{n}\mathrm{Tr}(R_\theta^{-1}R_{\theta^*}),\tag{42}$$

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

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

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:

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

□

B.3 Proof of Theorem 2

Proof. First, we prove Equation (16). 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:

$$\liminf_{n \to \infty} \lambda_{\min}(M_{ML}) \geq C_{\min} > 0. \tag{43}$$

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 ([15]), we obtain

$$\limsup_{n \to \infty} \lambda_{\max}(M_{ML}) < +\infty, \tag{44}$$

that concludes the proof of Equation (16).

By contradiction, let us now assume that

$$\sqrt{n} M_{ML}^{\frac{1}{2}} \left(\widehat{\theta}_{ML} - \theta^* \right) \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}(0, I_3). \tag{45}$$

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:

$$\left| \mathbb{E}\left[g\left(\sqrt{n} M_{ML}^{\frac{1}{2}} (\widehat{\theta}_{ML} - \theta^*) \right) \right] - \mathbb{E}(g(U)) \right| \geq \xi, \tag{46}$$

with $U \sim \mathcal{N}(0, I_3)$. The rest of the proof consists in contradicting Equation (46).

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:

$$M_{ML} \underset{n\to\infty}{\longrightarrow} M_\infty, \tag{47}$$

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

We have:

$$\frac{\partial}{\partial \theta_i} L_\theta = \frac{1}{n}\left(\text{Tr}\left(R_\theta^{-1} \frac{\partial R_\theta}{\partial \theta_i} \right) - y^T R_\theta^{-1} \frac{\partial R_\theta}{\partial \theta_i} R_\theta^{-1} y \right). \tag{48}$$

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 $P^T P = I_n$ and D diagonal, $z_\sigma = P R_{\theta^*}^{-\frac{1}{2}} y$ (which is a vector of i.i.d. standard Gaussian variables, conditionally to $\Sigma = \sigma$), we have

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

Hence, we have

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

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, conditionally to $\Sigma = \sigma$, $\sqrt{n} \frac{\partial}{\partial \theta} L_{\theta^*}$ converges in distribution to $\mathcal{N}(0, 4 M_\infty)$.

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

$$\mathbb{E}\left(\exp\left(i\sum_{k=1}^{3}\lambda_k\sqrt{n}\frac{\partial}{\partial\theta_k}L_{\theta^*}\right)\right)\underset{n\to\infty}{\longrightarrow}\exp\left(-\frac{1}{2}\lambda^T(4M_\infty)\lambda\right). \tag{49}$$

Finally,

$$\sqrt{n}\frac{\partial}{\partial\theta}L_{\theta^*}\underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}}\mathcal{N}(0,4M_\infty). \tag{50}$$

Let us now compute

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

Thus, we have, a.s.

$$\mathbb{E}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}L_{\theta^*}\right)\underset{n\to+\infty}{\longrightarrow}(2M_\infty)_{i,j}, \tag{51}$$

and, using Lemmas 2 and 3,

$$V\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}L_{\theta^*}|\Sigma\right)\underset{n\to+\infty}{\longrightarrow}0. \tag{52}$$

Hence, a.s.

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}L_{\theta^*}\underset{n\to+\infty}{\overset{\mathbb{P}_{|\Sigma}}{\longrightarrow}}2M_\infty. \tag{53}$$

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

$$\frac{1}{n}\text{Tr}(A_\theta)+\frac{1}{n}y^TB_\theta y, \tag{54}$$

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

$$\sup_{\theta\in\Theta}\left\|\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}L_\theta\right\|=O_{\mathbb{P}_{|\Sigma}}(1). \tag{55}$$

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 (55), $r = o_{\mathbb{P}_{|\Sigma}}(|\widehat{\theta}_{ML} - \theta^*|)$. 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

$$\left(\widehat{\theta}_{ML} - \theta^*\right) = -\left[\left(\frac{\partial}{\partial\theta}\frac{\partial}{\partial\theta_k}L_{\theta^*}\right)^T + o_{\mathbb{P}_{|\Sigma}}(1)\right]^{-1}\frac{\partial}{\partial\theta_k}L_{\theta^*}. \tag{56}$$

Hence, using Slutsky lemma, Equation 53 and Equation 50, a.s.

$$\sqrt{n}\left(\widehat{\theta}_{ML} - \theta^*\right) \xrightarrow[n\to+\infty]{\mathcal{L}_{|\Sigma}} \mathcal{N}\left(0, (2M_\infty)^{-1}(4M_\infty)(2M_\infty)^{-1}\right) = \mathcal{N}\left(0, M_\infty^{-1}\right). \tag{57}$$

Moreover, using Equation (47), we have

$$\sqrt{n}M_{ML}^{\frac{1}{2}}\left(\widehat{\theta}_{ML} - \theta^*\right) \xrightarrow[n\to+\infty]{\mathcal{L}_{|\Sigma}} \mathcal{N}(0, I_3). \tag{58}$$

Hence, using dominated convergence theorem on Σ, we have

$$\sqrt{n}M_{ML}^{\frac{1}{2}}\left(\widehat{\theta}_{ML} - \theta^*\right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, I_3). \tag{59}$$

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

$$\left|\mathbb{E}\left[g\left(\sqrt{n}M_{ML}^{\frac{1}{2}}(\widehat{\theta}_{ML} - \theta^*\right)\right] - \mathbb{E}(g(U))\right| \geq \xi, \tag{60}$$

which is in contradiction with Equation (46). □

B.4 Proof of Theorem 3

Proof. Let $\overline{\sigma}_n \in S_{N_n}$. We have:

$$\left|\widehat{Y}_{\widehat{\theta}_{ML}}(\overline{\sigma}_n) - \widehat{Y}_{\theta^*}(\overline{\sigma}_n)\right| \leq \sup_{\theta\in\Theta}\left|\frac{\partial}{\partial\theta}\widehat{Y}_\theta(\overline{\sigma}_n)\right| \left|\widehat{\theta}_{ML} - \theta^*\right| \tag{61}$$

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

$$\left|\sup_{\theta\in\Theta}\frac{\partial}{\partial\theta_i}\hat{Y}_\theta(\overline{\sigma}_n)\right| = O_{\mathbb{P}}(1). \tag{62}$$

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

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

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

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

with respect to θ. Thus, we can write these fist and second derivatives as a sum of $w_\theta^T(\overline{\sigma}_n) W_\theta y$ where $w_\theta(\overline{\sigma}_n)$ is of the form $r_\theta(\overline{\sigma}_n)$ or $\frac{\partial}{\partial\theta_i} r_\theta(\overline{\sigma}_n)$ of $\frac{\partial^2}{\partial\theta_j\theta_i} r_\theta(\overline{\sigma}_n)$ 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

$$\int_\Theta \left| w_\theta^T(\overline{\sigma}_n) W_\theta y \right|^4 d_\theta = O_{\mathbb{P}}(1). \tag{63}$$

From Fubini-Tonelli Theorem (see [6]), we have

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

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

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

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

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

Thus

$$\mathbb{E}\left(\int_\Theta \left| w_\theta^T(\overline{\sigma}_n) W_\theta y \right|^4 d_\theta | \Sigma \right) \leq B^2 c \int_\Theta \| w_\theta^T(\overline{\sigma}_n) \|^2 d_\theta. \tag{64}$$

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

$$\sup_{\theta\in\Theta} \| w_\theta^T(\overline{\sigma}_n) \|^2 = \sup_\theta \sum_{i=1}^n \left(\frac{\partial^{|\alpha|}}{\partial\theta^\alpha} K_\theta(\overline{\sigma}_n, \sigma_i) \right)^2.$$

Thus, it suffices to bound this term. Using the proof of Lemma 3, there exists $A > 0, a > 0$ such that

$$\sup_\theta \left(\frac{\partial^{|\alpha|}}{\partial\theta^\alpha} K_\theta(\overline{\sigma}_n, \sigma_i) \right)^2 \leq A \exp\left(-a d(\overline{\sigma}_n, \sigma_i) \right).$$

Yet, choosing $i^* \in [1:n]$ such that $d(\overline{\sigma}_n, \sigma_{i^*}) \leq d(\overline{\sigma}_n, \sigma_i)$ for all $i \in [1:n]$, we have

$$d(\overline{\sigma}_n, \sigma_i) \geq \frac{1}{2} d(\sigma_i, \sigma_{i^*}).$$

Thus, we have

$$\begin{aligned}
\sup_\theta \sum_{i=1}^n \left(\frac{\partial^{|\alpha|}}{\partial \theta^\alpha} K_\theta(\overline{\sigma}_n, \sigma_i) \right)^2 &\leq A \sum_{i=1}^n \exp\left(-\frac{a}{2} d(\sigma_i, \sigma_{i^*}) \right) \\
&\leq A \sum_{i=1}^n \exp\left(-\frac{a}{2} |i - i^*|^\beta \right) \\
&\leq 2A \sum_{i=0}^{+\infty} \exp\left(-\frac{a}{2} i^\beta \right) \\
&\leq C.
\end{aligned}$$

That concludes the proof. □

B.5 Proofs of the 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 $\lambda_{\max}(R_\theta) \leq \lambda_{\max}((K_{\theta_1,\theta_2}(\sigma_i, \sigma_j))_{i,j}) + \theta_{3,\max}$ and we show that $\lambda_{\max}(K_{\theta_1,\theta_2}(\sigma_i, \sigma_j)_{i,j})$ is uniformly bounded using Gershgorin circle theorem ([15]).
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

$$\left| \frac{\partial^{|\alpha|} K_{\theta_1,\theta_2}(\sigma, \sigma')}{\partial \theta^\alpha} \right| \leq \max(1, \theta_{2,\max}) d(\sigma, \sigma')^{\alpha_1} e^{-\theta_{1,\min} d(\sigma,\sigma')}. \tag{65}$$

We conclude using Gershgorin circle theorem ([15]). □

Proof of Lemma 4

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

$$N(x) := \max(4c\theta_{2,\max}|x_1|, 2|x_2|, |x_3|), \tag{66}$$

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

$$\frac{1}{n}\sum_{i,j=1}^{n}(R_{\theta,i,j} - R_{\theta^*,i,j})^2. \tag{67}$$

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

$$k_\alpha^\beta \geq 4c\theta_{2,\max}\frac{2 + \ln(\theta_{2,\max}) - \ln(\theta_{2,\min})}{\alpha}. \tag{68}$$

   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 $n \geq k_\alpha$,

$$\begin{aligned}
& \frac{1}{n}\sum_{i,j=1}^{n}(R_{\theta,i,j} - R_{\theta^*,i,j})^2 \\
\geq\ & \frac{1}{n}\sum_{i=1}^{n-k_\alpha}(R_{\theta,i,i+k_\alpha}0 - R_{\theta^*,i,i+k_\alpha})^2 \\
\geq\ & \frac{1}{n}\sum_{i=1}^{n-k_\alpha} e^{-2\theta_{1,\max}ck_\alpha + 2\ln(\theta_{2,\min})} 4\sinh^2\left(\frac{(\theta_1^* - \theta_1)d(\sigma_i, \sigma_{i+k_\alpha}) + \ln(\theta_2) - \ln(\theta_2^*)}{2}\right) \\
\geq\ & C_{1,\alpha}\frac{n - k_\alpha}{n},
\end{aligned}$$

   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,

$$\left|\frac{(\theta_1^* - \theta_1)d(\sigma_i, \sigma_{i+1}) + \ln(\theta_2) - \ln(\theta_2^*)}{2}\right| \geq \frac{\alpha}{8\theta_{2,\max}}, \tag{69}$$

and we have

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

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,

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

Finally, if we write

$$C_\alpha := \min\left(C_{1,\alpha}, C_{2,\alpha}, \frac{\alpha^2}{2}\right), \tag{70}$$

we have

$$\inf_{N(\theta-\theta^*)\geq\alpha} \frac{1}{n}\sum_{i,j=1}^{n}(R_{\theta,i,j} - R_{\theta^*,i,j})^2 \geq \frac{n-k_\alpha}{n}C_\alpha. \tag{71}$$

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

$$\liminf_{n\to+\infty} \inf_{\|\theta-\theta^*\|\geq\alpha} \frac{1}{n}\sum_{i,j=1}^{n}(R_{theta,i,j} - R_{\theta^*,i,j})^2 \geq C_{\alpha/h} > 0. \tag{72}$$

□

Proof of Lemma 5

Proof. We have

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

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

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

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 $\left(\lambda_2-\lambda_1 d(\sigma_i,\sigma_j)\right)^2\geq\epsilon$ 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.

□