Centre de recherche en économie et statistique

Série des Documents de Travail

n° 2016-08

Estimation of Low-Rank Covariance Function

V. Koltchinskii[1]
K.Lounici[2]
A.B.Tsybakov[3]

[1] Georgia Institute of Technology. E-mail: vlad@math.gatech.edu

[2] Georgia Institute of Technology. E-mail: klounici@math.gatech.edu

[3] CREST, ENSAE. E-mail: alexandre.tsybakov@ensae.fr

Estimation of Low-Rank Covariance Function

Koltchinskii, V.[1,2,a], Lounici, K.[3,4,a], Tsybakov, A.B.[5,6,b]

[a] *Georgia Institute of Technology, 686 Cherry St, Atlanta GA 30332, USA*
[b] *Laboratoire de Statistique, CREST-ENSAE, 3, av. P.Larousse, 92240 Malakoff, France.*

Abstract

We consider the problem of estimating a low rank covariance function $K(t,u)$ of a Gaussian process $S(t), t \in [0,1]$ based on n i.i.d. copies of S observed in a white noise. We suggest a new estimation procedure adapting simultaneously to the low rank structure and the smoothness of the covariance function. The new procedure is based on nuclear norm penalization and exhibits superior performances as compared to the sample covariance function by a polynomial factor in the sample size n. Other results include a minimax lower bound for estimation of low-rank covariance functions showing that our procedure is optimal as well as a scheme to estimate the unknown noise variance of the Gaussian process.

Key words: Gaussian process, Low rank Covariance Function, Nuclear norm, Empirical risk minimization, Minimax lower bounds, Adaptation

[1] vlad@math.gatech.edu

[2] Supported in part by NSF Grants DMS-1207808 and CCF-1415498

[3] klounici@math.gatech.edu

[4] Supported in part by NSF CAREER Grant DMS-1454515 and Simons Collaboration Grant 315477

[5] alexandre.tsybakov@ensae.fr

[6] Supported by GENES and by the French National Research Agency (ANR) under the grants IPANEMA (ANR-13-BSH1-0004-02) and Labex ECODEC (ANR - 11-LABEX-0047), and by "Chaire Economie et Gestion des Nouvelles Données", under the auspices of Institut Louis Bachelier, Havas-Media and Paris-Dauphine

1. Introduction

Let $X(t), t \in [0,1]$ be a Gaussian process satisfying the following stochastic differential equation:

$$dX(t) = S(t)dt + \sigma dW(t), \quad t \in [0,1], \tag{1}$$

where W is the standard Brownian motion, $\sigma > 0$ is the noise level, and

$$S(t) = \sum_{k=1}^{r} \sqrt{\lambda_k}\xi_k\varphi_k(t), \quad t \in [0,1].$$

Here ξ_k are i.i.d. standard Gaussian random variables independent of the Brownian motion W, $\{\varphi_k\}_{k=1}^r$ are unknown orthonormal functions in $L_2[0,1]$, possibly, with $r = \infty$, and the coefficients $\lambda_k > 0$ are unknown and such that $\sum_{k=1}^r \lambda_k < \infty$. The value of r is also unknown.

Assume that we observe n i.i.d. copies $X_1(t), \dots, X_n(t)$ of the process $X(t)$. In this paper, we study the problem of estimation of the covariance function of the stochastic process $S(\cdot)$,

$$K(t,u) = \mathbb{E}(S(t)S(u)) = \sum_{k=1}^{r} \lambda_k\varphi_k(t)\varphi_k(u), \quad t,u \in [0,1], \tag{2}$$

based on the observations $\{X_1(t), \dots, X_n(t), t \in [0,1]\}$. If $r = \infty$, the sum in (2) is understood in the sense of $L_2([0,1] \times [0,1])$-convergence. In short, (1) is a model of a "signal" (Gaussian stochastic process S) observed in a Gaussian white noise and the goal is to estimate the covariance of the signal based on a sample of such observations.

Statistical estimation of covariance functions has already received some attention in the literature. However, somewhat different setting was considered where the trajectories $X_i(\cdot)$ are observed at discrete time locations:

$$Y_{i,j} = S_i(T_{i,j}) + \sigma\eta_{i,j}, \quad 1 \le i \le n,\ 1 \le j \le m,$$

where S_i are i.i.d. copies of S, $\eta_{i,j}$ are i.i.d. $\mathcal{N}(0,1)$ and, for each i, the points $T_{i,j}$, $1 \le j \le m$, are equispaced in the interval $[0,1]$ or independent random variables with uniform distribution on $[0,1]$. In this setting, Yao et al. (2005) proposed a local smoothing estimation procedure assuming that the trajectories $X_i(\cdot)$ are well approximated by the projection on the linear span of

functions $\varphi_1, \ldots, \varphi_k$ for some known fixed k chosen by cross-validation. This procedure is computationally intensive as it requires to compute the eigenvalues and the inverse for n distinct $m \times m$ empirical covariance matrices of the trajectories X_i, $1 \leq i \leq n$, at each of the cross-validation steps. The results in Yao et al. (2005) provide theoretical guarantees for estimation of the covariance function and its eigenfunctions under the condition that the previous approximation is sufficiently precise. Hall et al. (2006) consider the same methodology and study the effect of the sampling rate on the estimation rate of the eigenfunctions. In a similar framework, Bunea and Xiao (2013) propose a simpler procedure to estimate the eigenfunctions and obtain theoretical guarantees on the estimation error. Their approach involves a dimension reduction step where the selection of the relevant eigenfunctions is performed by thresholding the eigenvalues of a correctly constructed empirical covariance matrix. In a similar setting, Bigot et al. (2010) consider the estimation of the covariance matrix of the process S at sample points rather than that of the covariance function. This problem can be reduced to multivariate regression and Bigot et al. (2010) develop a model selection approach to it resulting in some oracle inequalities.

Noteworthy, strong regularity conditions are usually imposed on the eigenfunctions φ_k in the existing literature. In Hall et al. (2006) the eigenfunctions are assumed to admit bounded derivatives of order at least two. In addition, the optimal bandwidth choice in the local smoothing approach used in Hall et al. (2006); Yao et al. (2005) requires the knowledge of smoothness degree of the eigenfunctions. In Bunea and Xiao (2013), the eigenfunctions are assumed to be continuously differentiable with bounded derivatives, the sequence of eigenvalues belongs to a Sobolev ball with regularity $\beta > 0$ and the optimal choice of the threshold in the dimension reduction step depends on β.

An interesting question is what are the optimal rates of estimation of the covariance function in a minimax sense. To our knowledge, it was not addressed in the literature.

In this paper, we assume that the trajectories $X_i(\cdot)$ are fully observed in time. Our aim is to understand the influence of the structure of the covariance function K on the estimation rate. The main contributions of this paper are as follows:

1. We propose a simple data-driven procedure to estimate the covariance function and prove oracle inequalities for it based on recent results on

high-dimensional matrix estimation.

2. We show that the proposed method is minimax optimal for estimation of K in the L_2-norm whereas the empirical covariance estimator is suboptimal.

2. Definitions and notations

Let $e_1(\cdot), e_2(\cdot), \ldots$ be an orthonormal basis of $L_2[0,1]$, which is assumed to be fixed throughout the paper. Denote by $\|\cdot\|_2$ the norms either of $L_2[0,1]$ or of $L_2([0,1]\times[0,1])$ (according to the context) and by $\langle\cdot,\cdot\rangle$ the corresponding inner products. For any integer $l \geq 1$, consider the orthogonal projection $S^{(l)} = \sum_{k=1}^{l}\langle e_k, S\rangle e_k$ of S onto the linear span of $\{e_1,\ldots,e_l\}$. Set

$$\dot{X}^{(l)} = \sum_{k=1}^{l}\int_0^1 e_k(t)dX(t)\ e_k, \quad \dot{W}^{(l)} = \sum_{k=1}^{l}\int_0^1 e_k(t)dW(t)\ e_k. \tag{3}$$

In view of (1), we have

$$\dot{X}^{(l)} = S^{(l)} + \sigma\dot{W}^{(l)}.$$

Similarly to (3), we define the processes

$$\dot{X}_i^{(l)} = \sum_{k=1}^{l}\int_0^1 e_k(t)dX_i(t)\ e_k, \quad i = 1,\ldots,n,$$

and consider the empirical covariance function

$$R_n^{(l)}(t,u) = \frac{1}{n}\sum_{i=1}^{n}\dot{X}_i^{(l)}(t)\dot{X}_i^{(l)}(u), \quad t,u\in[0,1].$$

Note that the expectation of $R_n^{(l)}(t,u)$ is

$$\begin{aligned}\mathbb{E}\left[R_n^{(l)}(t,u)\right] &= \mathbb{E}\left[S^{(l)}(t)S^{(l)}(u)\right] + \sigma^2 I^{(l)}(t,u)\\ &= K^{(l)}(t,u) + \sigma^2 I^{(l)}(t,u),\end{aligned}$$

with $I^{(l)}(t,u) = \sum_{k=1}^{l} e_k(t)e_k(u)$ and

$$K^{(l)}(t,u) = \mathbb{E}\left[S^{(l)}(t)S^{(l)}(u)\right] = \sum_{m=1}^{r}\lambda_m\varphi_m^{(l)}(t)\varphi_m^{(l)}(u)$$

where $\varphi_m^{(l)} = \sum_{k=1}^{l} \langle e_k, \varphi_m \rangle e_k$ is the orthogonal projection of φ_m onto the linear span of $\{e_1, \ldots, e_l\}$. In what follows, we will consider the set of functions

$$\mathcal{S}_l = \left\{ \sum_{j,k=1}^{l} s_{jk}(e_j \otimes e_k) : \ s_{jk} = s_{kj}, \ j,k = 1, \ldots, l \right\}$$

where $(e_j \otimes e_k)(t,s) = e_j(t)e_k(s)$. The set $\mathcal{S}_l$ consists of all symmetric kernels belonging to the linear span of $\{e_j \otimes e_k : j,k = 1, \ldots, l\}$. Note that K is not necessarily in $\mathcal{S}_l$ while $R_n^{(l)}, K^{(l)}, I^{(l)} \in \mathcal{S}_l$. It is easy to see that $K^{(l)}$ is the orthogonal projection of K onto $\mathcal{S}_l$.

If no ambiguity is caused, for any $A \in \mathcal{S}_l$, we will use the same symbol A to denote the corresponding symmetric $l \times l$ matrix. For any function $A \in \mathcal{S}_l$ or any $l \times l$ matrix A we denote by $\|A\|_1$ and $\|A\|_\infty$ its nuclear and spectral norms, respectively. The trace and the rank of matrix A are denoted by $\mathrm{tr}(A)$ and $\mathrm{rank}(A)$, and its Frobenius norm by $\|A\|_F$. Writing $A \geq 0$ for a matrix A means that A is non-negative definite.

3. Nuclear norm penalized estimator and its convergence rate

In this section, we assume that the noise level σ is known. For an integer $l \geq 1$, we define the estimator $\hat{A}^{(l)}$ of K as a solution of the following penalized minimization problem

$$\hat{A}^{(l)} \in \mathrm{argmin}_{A \in \mathcal{S}_l, A \geq 0} \left(\|R_n^{(l)} - A - \sigma^2 I^{(l)}\|_2^2 + \mu \|A\|_1 \right), \tag{4}$$

where $\mu > 0$ is a regularization parameter to be tuned. Note that here we have $\|A\|_1 = \mathrm{tr}(A)$. The solution of (4) is explicitly expressed via soft thesholding of the eigenvalues of the matrix $R_n^{(l)} - \sigma^2 I^{(l)}$ (cf. Koltchinskii et al. (2011)). The next theorem easily follows from the argument in the proof of Theorem 1 in Koltchinskii et al. (2011) (see also Lounici (2014)).

THEOREM 1. *Let $n, l \geq 1$ be integers and let $X_1(\cdot), \ldots, X_n(\cdot)$ be i.i.d. realizations of the process $X(\cdot)$ satisfying (1). If $\mu \geq 2\|R_n^{(l)} - K^{(l)} - \sigma^2 I^{(l)}\|_\infty$ then, for any K satisfying (2) with $\sum_{k=1}^{r} \lambda_k < \infty$ we have*

$$\|\hat{A}^{(l)} - K\|_2^2 \leq \inf_{A \in \mathcal{S}_l, A \geq 0} \left\{ \|A - K\|_2^2 + \min \left\{ 2\mu \|A\|_1, \frac{(1+\sqrt{2})^2}{8} \mu^2 \mathrm{rank}(A) \right\} \right\}.$$

This theorem is a deterministic fact as soon as we have a proper bound on a single random variable, namely, the spectral norm $\|R_n^{(l)} - K^{(l)} - \sigma^2 I^{(l)}\|_\infty$. In other words, all stochastic effects in our problem are localized in the behaviour of this random variable and the choice of μ is driven by it as well. The next lemma provides a probabilistic bound on this random variable.

LEMMA 2. *Let $n, l \geq 1$ be integers and let $X_1(\cdot), \ldots, X_n(\cdot)$ be i.i.d. realizations of the process $X(\cdot)$ satisfying (1). Set $\lambda_{\max} = \sup_{1 \leq j \leq r} \lambda_j$. For any $t > 0$ and $l \geq 1$, define*

$$\delta_n(l,t) = \max\left\{\sqrt{\frac{l+t}{n}}, \frac{l+t}{n}\right\}. \tag{5}$$

Then, with probability at least $1 - e^{-t}$, for any K satisfying (2) with $\sum_{k=1}^r \lambda_k < \infty$ we have

$$\|R_n^{(l)} - K^{(l)} - \sigma^2 I^{(l)}\|_\infty \leq C(\lambda_{\max} + \sigma^2)\delta_n(l,t),$$

for some absolute constant $C > 0$.

Proof. Set $\mathrm{x}_i(l) = (\int_0^1 e_1(t)dX_i(t), \ldots, \int_0^1 e_l(t)dX_i(t))^\top$ for any $1 \leq i \leq n$ and $\hat{B}_{n,l} = \frac{1}{n}\sum_{i=1}^n \mathrm{x}_i(l)\mathrm{x}_i(l)^\top$. Note that $\mathrm{x}_i(l)$ are i.i.d. normal random vectors with mean 0 and covariance matrix $B_l = K^{(l)} + \sigma^2 I^{(l)}$. Also $\|R_n^{(l)} - K^{(l)} - \sigma^2 I^{(l)}\|_\infty = \|\hat{B}_{n,l} - B_l\|_\infty$. Here, $I^{(l)}$ is the $l \times l$ identity matrix. Next,

$$\|\hat{B}_{n,l} - B_l\|_\infty \leq \|B_l\|_\infty \left\|\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top - I^{(l)}\right\|_\infty \leq (\lambda_{\max} + \sigma^2)\left\|\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top - I^{(l)}\right\|_\infty$$

where $Z_1, \ldots, Z_n$ are i.i.d. standard normal vectors in $\mathbb{R}^l$. Here we also used the fact that the following representation holds for random vectors $\mathrm{x}_i(l)$: $\mathrm{x}_i(l) = B_l^{1/2} Z_i$. Applying Theorem 5.39 in Vershynin (2012) to the random variable $\left\|\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top - I^{(l)}\right\|_\infty$ we get the result. □

Theorem 1 with Lemma 2 immediately imply the following result.

THEOREM 3. *Let $n, l \geq 1$ be integers and let $X_1(\cdot), \ldots, X_n(\cdot)$ be i.i.d. realizations of the process $X(\cdot)$ satisfying (1). Take*

$$\mu = c(\lambda_{\max} + \sigma^2)\delta_n(l,t),$$

for some sufficiently large absolute constant $c > 0$. Define

$$v_n(A, l, t) = \min\left\{(\lambda_{\max} + \sigma^2)\mathrm{tr}(A)\delta_n(l, t), (\lambda_{\max} + \sigma^2)^2\mathrm{rank}(A)\delta_n^2(l, t)\right\}.$$

Let $t > 0$. Then, with probability at least $1 - e^{-t}$, for any K satisfying (2) with $\sum_{k=1}^{r} \lambda_k < \infty$ we have

$$\|\hat{A}^{(l)} - K\|_2^2 \le \inf_{A \in \mathcal{S}_l, A \ge 0} \left\{\|A - K\|_2^2 + Cv_n(A, l, t)\right\} \tag{6}$$

with some absolute constant $C > 0$.

The bound (6) is the main oracle inequality that we will use now to obtain bounds on the risk of the estimator $\hat{A}^{(l)}$. It is easy to check that

$$v_n(A, l, t) \le (\lambda_{\max} + \sigma^2)^2\mathrm{rank}(A)\frac{l + t}{n}.$$

The above bound is trivial if $l + t \le n$. In the case $l + t > n$, it follows from the bound

$$(\lambda_{\max}+\sigma^2)\mathrm{tr}(A)\frac{l + t}{n} \le (\lambda_{\max}+\sigma^2)\lambda_{\max}\mathrm{rank}(A)\frac{l + t}{n} \le (\lambda_{\max}+\sigma^2)^2\mathrm{rank}(A)\frac{l + t}{n}.$$

Combining Theorem 3 with the fact that, for a random variable η, $\mathbb{E}[|\eta|] = \int_0^\infty \mathbb{P}(|\eta| \ge t)dt$ and taking $A = K^{(l)}$,

$$\mathbb{E}[\|\hat{A}^{(l)} - K\|_2^2] \le \|K^{(l)} - K\|_2^2 + C(\lambda_{\max} + \sigma^2)^2\frac{(r \wedge l)l}{n} \tag{7}$$

for some absolute constant $C > 0$, where we have used that $\mathrm{rank}(K^{(l)}) \le r \wedge l$. This inequality is valid for all K of the form (2), with finite or infinite r.

As a corollary, we get the following bound on the minimax risk over the class of covariance functions that admit a finite expansion with respect to the basis $\{e_k\}$. Denote by $\mathcal{K}_{r,l}(\lambda_{\max})$ the class of all covariance functions satisfying (2) such that $K \in \mathcal{S}_l$ and $\|K\|_\infty \le \lambda_{\max}$ where $\lambda_{\max}$ is a finite positive constant. Note that the system of functions $\{\varphi_k\}$ in this definition is not fixed and varies among all orthonormal systems in $L_2[0, 1]$.

Corollary 4. *Under the assumptions of Theorem 3, we have*

$$\sup_{K \in \mathcal{K}_{r,l}(\lambda_{\max})} \mathbb{E}[\|\hat{A}^{(l)} - K\|_2^2] \le C(\lambda_{\max} + \sigma^2)^2\frac{(r \wedge l)l}{n}$$

for some absolute constant $C > 0$.

It is interesting to compare the estimator $\hat{A}^{(l)}$ with the other natural estimator, which is the corrected empirical covariance function

$$\bar{A}^{(l)} \triangleq R_n^{(l)} - \sigma^2 I^{(l)}.$$

We have the following expression for the risk of $\bar{A}^{(l)}$.

PROPOSITION 5. *For any K satisfying (2) with $\sum_{k=1}^{r} \lambda_k < \infty$ we have*

$$\mathbb{E}[\|\bar{A}^{(l)} - K\|_2^2] = \|K^{(l)} - K\|_2^2 + \frac{\|B_l\|_2^2 + [\mathrm{tr}(B_l)]^2}{n}$$

where $B_l = K^{(l)} + \sigma^2 I^{(l)}$.

Proof. Set for brevity $B = B_l$, $\hat{B}_n = \hat{B}_{n,l}$, $\mathrm{x}_i = \mathrm{x}_i(l)$. Note that $\mathbb{E}(\bar{A}^{(l)}) = K^{(l)}$. The bias-variance decomposition of the risk of $\bar{A}^{(l)}$ yields

$$\mathbb{E}[\|\bar{A}^{(l)} - K\|_2^2] = \|K^{(l)} - K\|_2^2 + \mathbb{E}[\|R_n^{(l)} - \mathbb{E}(R_n^{(l)})\|_2^2].$$

Here, $\mathbb{E}[\|R_n^{(l)} - \mathbb{E}(R_n^{(l)})\|_2^2] = \mathbb{E}[\|\hat{B}_n - B\|_F^2] = \mathbb{E}\big[\big\|\frac{1}{n}\sum_{i=1}^n W_i\big\|_F^2\big]$ where $W_i = \mathrm{x}_i\mathrm{x}_i^\top - \mathbb{E}[\mathrm{x}_i\mathrm{x}_i^\top]$. Since the matrices W_i are i.i.d. we find $\mathbb{E}\big[\big\|\frac{1}{n}\sum_{i=1}^n W_i\big\|_F^2\big] = \mathbb{E}\,\mathrm{tr}\big(\frac{1}{n^2}\sum_{i,j=1}^n W_i^\top W_j\big) = \frac{1}{n}\mathrm{tr}\big(\mathbb{E}(W_1^\top W_1)\big) = \frac{1}{n}\big(\mathbb{E}(|\mathrm{x}_1|_2^4) - \mathrm{tr}(B^\top B)\big)$ where $|\cdot|_2$ denotes the Euclidean norm. Here, $\mathbb{E}(|\mathrm{x}_1|_2^4) - \mathrm{tr}(B^\top B) = \|B\|_2^2 + [\mathrm{tr}(B)]^2$ and the result follows. □

Since $\mathrm{tr}(B) \geq \sigma^2 l$, Proposition 5 implies

$$\mathbb{E}[\|\bar{A}^{(l)} - K\|_2^2] \geq \|K^{(l)} - K\|_2^2 + \frac{\sigma^4 l^2}{n}, \tag{8}$$

$$\inf_K \mathbb{E}[\|\bar{A}^{(l)} - K\|_2^2] \geq \frac{\sigma^4 l^2}{n} \tag{9}$$

where $\inf_K$ is the infimum over all K satisfying (2) with $\sum_{k=1}^{r} \lambda_k < \infty$. Comparing (9) with Corollary 4 we see that the risk of the empirical estimator $\bar{A}^{(l)}$ on the class $\mathcal{K}_{r,l}$ is of the order greater than the risk of our estimator $\hat{A}^{(l)}$ when r is smaller than l.

Our estimator also outperforms the estimator $\bar{A}^{(l)}$ for kernels K that do not admit a finite expansion with respect to the basis $\{e_k\}$, but satisfy some regularity conditions. To this end, we introduce a specific norm that can be naturally interpreted as a version of the Sobolev norm for covariance

functions. Fix the smoothness parameter $s > 0$. For any symmetric function $K : [0,1]^2 \to \mathbb{R}$, we define

$$\|K\|_{s,2} := \|\Delta^s K\|_2 = \left(\sum_{k,k' \geq 1} k^{2s} \langle K e_k, e_{k'} \rangle^2 \right)^{1/2},$$

where Δ is an operator admitting the matrix representation $\mathrm{diag}(1, 2, \cdots, k, \cdots)$ w.r.t the basis $(e_k)_{k \geq 1}$. Note that the norm $\|K\|_{s,2}$ depends on the basis $\{e_k\}$ but we do not indicate this dependence in the notation since $\{e_k\}$ is fixed. Note also that if K admits spectral representation (2), then

$$\|K\|_{s,2} = \left(\mathrm{tr}(\Delta^{2s} K^2) \right)^{1/2} = \left(\sum_{k=1}^{r} \lambda_k^2 \|\varphi_k\|_{s,2}^2 \right)^{1/2},$$

where we use the notation

$$\|\varphi\|_{s,2} = \|\Delta^s \varphi\| = \left(\sum_{k \geq 1} k^{2s} \langle \varphi, e_k \rangle^2 \right)^{1/2}$$

for a Sobolev type norm of a function $\varphi : [0,1] \to \mathbb{R}$.

Assumption 6. *Suppose the covariance function K has finite rank r and there exist constants $\lambda_{\max} > 0$, $s > 0$ and $\rho \geq 1$ such that $\|K\|_\infty \leq \lambda_{\max}$ and $\|K\|_{s,2} \leq \rho$.*

Denote by $\overline{\mathcal{K}}_r(s, \rho; \lambda_{\max})$ the class of all kernels K satisfying Assumption 6.

Theorem 7. *Given $r \geq 1, s > 0, \rho > 0$ and $\lambda_{\max} > 0$, set*

$$\ell := \max \left(\left\lceil \left(\frac{\rho^2}{(\lambda_{\max} + \sigma^2)^2} \frac{n}{r} \right)^{1/(2s+1)} \right\rceil, \left\lceil \left(\frac{\rho^2 n}{(\lambda_{\max} + \sigma^2)^2} \right)^{1/(2s+2)} \right\rceil \right).$$

Then, with some absolute constant $C > 0$,

$$\sup_{K \in \overline{\mathcal{K}}_r(s,\rho;\lambda_{\max})} \mathbb{E}[\|\hat{A}^{(\ell)} - K\|_2^2] \leq \tag{10}$$

$$C \min \left((\lambda_{\max} + \sigma^2)^{4s/(2s+1)} \rho^{2/(2s+1)} \left(\frac{r}{n} \right)^{2s/(2s+1)}, (\lambda_{\max} + \sigma^2)^{2s/(s+1)} \rho^{2/(s+1)} n^{-s/(s+1)} \right).$$

Proof. Since K satisfies Assumption 6, we have for any $l \geq 1$ that

$$\|K - K^{(l)}\|_2^2 = \sum_{k \geq l+1} \sum_{k'=1}^{\infty} \langle K e_k, e_{k'} \rangle^2 + \sum_{k' \geq l+1} \sum_{k=1}^{l} \langle K e_k, e_{k'} \rangle^2$$

$$\leq (l+1)^{-2s} \sum_{k \geq l+1} \sum_{k'=1}^{\infty} k^{2s} \langle K e_k, e_{k'} \rangle^2 + (l+1)^{-2s} \sum_{k' \geq l+1} \sum_{k=1}^{\infty} (k')^{2s} \langle K e_k, e_{k'} \rangle^2 \leq 2\rho^2 l^{-2s}.$$

Combining the previous display with (7), we find that, for any $l \geq 1$,

$$\mathbb{E}[\|\hat{A}^{(l)} - K\|_2^2] \leq 2\rho^2 l^{-2s} + C(\lambda_{\max} + \sigma^2)^2 \frac{(r \wedge l) l}{n}.$$

The minimum of the right-hand side of this inequality is achieved for l of the order of ℓ. By setting $l = \ell$, we obtain (10). □

Note that, if the rank r is small, the problem of estimation of covariance function K reduces to estimation of a small number r of eigenfunctions and eigenvalues of K. The rate in (10) is, in this case, of the order $O(n^{-2s/(2s+1)})$, which coincides with a standard minimax error rate of estimation of a function of one variable of smoothness s. On the other hand, when the rank r is large (say, $r = +\infty$), the estimation error rate becomes $O(n^{-s/(s+1)})$, which is the minimax rate of estimation of a function of two variables of smoothness s. Similar error rates where studied earlier in matrix completion problems for smooth kernels on graphs (see Koltchinskii and Rangel (2013)).

We consider now a class of kernels determined by the following assumption, which can be interpreted as a Sobolev type condition on the individual eigenfunctions φ_j.

ASSUMPTION 8. *The value r is finite and there exist constants $s > 0$, $c_* > 0$ such that, for any $1 \leq j \leq r$, $\|\varphi_j\|_{s,2} \leq c_*$.*

Denote by $\mathcal{K}_r(s, c_*; \lambda_{\max})$ the class of all kernels K defined by (2) with eigenfunctions φ_j satisfying Assumption 8 and such that $\|K\|_\infty < \lambda_{\max}$.

THEOREM 9. *Let $l_1 = \max\left(\lceil n^{\frac{1}{2s+1}} \rceil, \lceil (rn)^{\frac{1}{2(s+1)}} \rceil\right)$, $n \geq 1$, $1 \leq r < \infty$. For any $s > 0$, $c_* > 0$, $\lambda_{\max} > 0$ we have*

$$\sup_{K \in \mathcal{K}_r(s, c_*; \lambda_{\max})} \mathbb{E}[\|\hat{A}^{(l_1)} - K\|_2^2] \leq C \min\left(r n^{-\frac{2s}{2s+1}}, r^{\frac{1}{s+1}} n^{-\frac{s}{s+1}}\right) \qquad (11)$$

where $C > 0$ is a constant depending only on $\lambda_{\max}, \sigma$ and c_*.

Proof. It is enough to observe that, for all $K \in \mathcal{K}_r(s, c_*; \lambda_{\max})$,

$$\|K\|_{s,2}^2 = \sum_{k=1}^{r} \lambda_k^2 \|\varphi_k\|_{s,2}^2 \le c_*^2 \lambda_{\max}^2 r,$$

implying that $\mathcal{K}_r(s, c_*; \lambda_{\max}) \subset \overline{\mathcal{K}}_r(s, \rho; \lambda_{\max})$ with $\rho = c_* \lambda_{\max} \sqrt{r}$. Bound (11) now follows from (10). □

When r is a fixed constant and n is large, the rate in (11) is $O(n^{-\frac{2s}{2s+1}})$. The next theorem shows that this rate cannot be achieved by the corrected empirical covariance estimator $\bar{A}^{(l)}$ whatever is the choice of l.

THEOREM 10. *Let $n \ge 1$, $1 \le r < \infty$. There exists $c_* > 0$ such that for any $s > 0$, $\lambda_{\max} > 0$ we have*

$$\inf_{l \ge 1} \sup_{K \in \mathcal{K}_r(s, c_*; \lambda_{\max})} \mathbb{E}[\|\bar{A}^{(l)} - K\|_2^2] \ge C n^{-\frac{s}{s+1}} \tag{12}$$

where $C > 0$ is a constant that can depend only on $\lambda_{\max}, \sigma$, s and c_*.

Proof. Fix $l \ge 1$ and consider the function

$$\varphi_1(t) = C_1 \left(\sum_{k=1}^{l} \frac{e_k(t)}{k^{s+1}} + \sum_{k=l+1}^{2l} \frac{e_k(t)}{k^{s+1/2}} \right), \quad t \in [0, 1],$$

where C_1 is a normalizing constant, depending only on s, such that $\|\varphi_1\|_2 = 1$. By an easy computation, $\|\varphi_1\|_{s,2} \le c'$ for a constant c' depending only on s.

Set $\bar{K}(t, u) = \lambda_{\max} \varphi_1(t) \varphi_1(u)$. Then $\bar{K} \in \mathcal{K}_r(s, c_*; \lambda_{\max})$ with $c_* = c'$. Due to (8),

$$\begin{aligned} \sup_{K \in \mathcal{K}_r(s, c_*; \lambda_{\max})} \mathbb{E}[\|\bar{A}^{(l)} - K\|_2^2] &\ge \sup_{K \in \mathcal{K}_r(s, c_*; \lambda_{\max})} \|K^{(l)} - K\|_2^2 + \frac{\sigma^4 l^2}{n} \\ &\ge \|\bar{K}^{(l)} - \bar{K}\|_2^2 + \frac{\sigma^4 l^2}{n}. \end{aligned} \tag{13}$$

Observe that

$$\varphi_1 \otimes \varphi_1 = \varphi_1^{(l)} \otimes \varphi_1^{(l)} + (\varphi_1 - \varphi_1^{(l)}) \otimes \varphi_1^{(l)} + \varphi_1 \otimes (\varphi_1 - \varphi_j^{(l)}).$$

Therefore,

$$\|\varphi_1 \otimes \varphi_1 - \varphi_1^{(l)} \otimes \varphi_1^{(l)}\|_2^2 = \|(\varphi_1 - \varphi_1^{(l)}) \otimes \varphi_1^{(l)}\|_2^2 + \|\varphi_1 \otimes (\varphi_1 - \varphi_1^{(l)})\|_2^2$$
$$\geq \|\varphi_1\|_2^2 \|\varphi_1 - \varphi_1^{(l)}\|_2^2 = \|\varphi_1 - \varphi_1^{(l)}\|_2^2.$$

This implies that

$$\|\bar{K}^{(l)} - \bar{K}\|_2^2 \geq \lambda_{\max}^2 \|\varphi_1 - \varphi_1^{(l)}\|_2^2 \geq c\lambda_{\max}^2 l^{-2s}$$

for some constant $c > 0$ depending only on s. Using this inequality in (13) and taking the minimum over $l \geq 1$, we obtain the result. □

4. Adaptive Estimation

We observe that the optimal choice of the parameter l in theorems 7 and 9 depends on the unknown parameters ρ, s and r that quantify respectively the smoothness of the eigenfunctions of K and their number. In this section, we propose an adaptive estimator, which does not depend on s and r that attains the same rate as in Theorem 7 or in Theorem 9.

First, we describe a general method of aggregating estimators. Assume without loss of generality that the sample size n is even. We split the sample of n trajectories $\mathbb{X} = \{X_1, \ldots, X_n\}$ into two parts of equal size $n/2$, denoted $\mathbb{X}_1 = \{X_1, \ldots, X_{n/2}\}$ and $\mathbb{X}_2 = \{X_{n/2+1}, \ldots, X_n\}$. Fix an integer L. Using the sample $\mathbb{X}_1$, we construct a family of estimators $A^{(1)}, \ldots, A^{(L)}$ such that $A^{(l)} \in \mathcal{S}_l$, $1 \leq l \leq L$. These can be, for example, the estimators $\hat{A}^{(1)}, \ldots, \hat{A}^{(L)}$ defined in (4).

Consider the following adaptive selector of l:

$$\hat{l} = \arg\min_{1 \leq l \leq L} \{\|A^{(l)}\|_2^2 - 2\langle A^{(l)}, \tilde{R}_n^{(l)} - \sigma^2 I^{(l)} \rangle\}, \tag{14}$$

where $\tilde{R}_n^{(l)}(t,u) = \frac{2}{n} \sum_{i=n/2+1}^{n} \dot{X}_i^{(l)}(t) \dot{X}_i^{(l)}(u)$ is the projected empirical covariance function associated to the second subsample $\mathbb{X}_2$.

In the following theorem we assume that the first subsample is frozen, so we state the result for non-random functions $A^{(l)} \in \mathcal{S}_l$, $1 \leq l \leq L$.

THEOREM 11. *Let $A^{(l)}$, $1 \leq l \leq L$, be functions such that $A^{(l)} \in \mathcal{S}_l$. For any $t > 0$, with probability at least $1 - e^{-t}$ with respect to the subsample $\mathbb{X}_2$ we have*

$$\|A^{(\hat{l})} - K\|_2^2 \leq 2 \min_{1 \leq l \leq L} \|A^{(l)} - K\|_2^2 + C[\lambda_{\max} \vee \sigma^2]^2 \max\left\{ \frac{t + \log L}{n}, \left(\frac{t + \log L}{n} \right)^2 \right\}$$

for all K satisfying (2) with $\sum_{k=1}^{r}\lambda_k < \infty$. Here, $C > 0$ is an absolute constant.

Proof. Fix an arbitrary $\bar{l} \in \{1, \ldots, L\}$. Note that, by definition, $\{\mathcal{S}_l\}_{l\geq 1}$ is a nested sequence satisfying

$$\mathcal{S}_{l+1} = \mathcal{S}_l \oplus \text{l.s.}\left\{e_j \otimes e_{l+1} + e_{l+1} \otimes e_j,\, 1 \leq j \leq l\right\}.$$

Consequently, for any $1 \leq l, l' \leq L$, we have $\langle A^{(l)}, \tilde{R}_n^{(l)}\rangle = \langle A^{(l)}, \tilde{R}_n^{(l\vee l')}\rangle$. Similarly $\langle A^{(l)}, K\rangle = \langle A^{(l)}, K^{(l)}\rangle = \langle A^{(l)}, K^{(l\vee l')}\rangle$. Combining this observation with (14), we get

$$\begin{aligned}
&\|A^{(\hat{l})} - K\|_2^2 - \|A^{(\bar{l})} - K\|_2^2 \\
&\qquad = \|A^{(\hat{l})}\|_2^2 - 2\langle A^{(\hat{l})}, K^{(\hat{l})}\rangle - [\|A^{(\bar{l})}\|_2^2 - 2\langle A^{(\bar{l})}, K^{(\bar{l})}\rangle] \\
&\qquad \leq \|\hat{A}^{(\hat{l})}\|_2^2 - 2\langle A^{(\hat{l})}, \tilde{R}_n^{(\hat{l})} - \sigma^2 I^{(\hat{l})}\rangle - [\|A^{(\bar{l})}\|_2^2 - 2\langle A^{(\bar{l})}, \tilde{R}_n^{(\bar{l})} - \sigma^2 I^{(\bar{l})}\rangle] \\
&\qquad\qquad + 2\langle A^{(\hat{l})} - A^{(\bar{l})}, \tilde{R}_n^{(\hat{l}\vee\bar{l})} - K^{(\hat{l}\vee\bar{l})} - \sigma^2 I^{(\hat{l}\vee\bar{l})}\rangle \\
&\qquad \leq 2\langle A^{(\hat{l})} - A^{(\bar{l})}, \tilde{R}_n^{(\hat{l}\vee\bar{l})} - K^{(\hat{l}\vee\bar{l})} - \sigma^2 I^{(\hat{l}\vee\bar{l})}\rangle.
\end{aligned}$$

Here, $K^{(\hat{l}\vee\bar{l})} + \sigma^2 I^{(\hat{l}\vee\bar{l})} = \mathbb{E}[\tilde{R}_n^{(\hat{l}\vee\bar{l})}]$. Setting for brevity $m = \hat{l} \vee \bar{l}$ we deduce from the previous display that

$$\|A^{(\hat{l})} - K\|_2^2 - \|A^{(\bar{l})} - K\|_2^2 \leq 2U\|A^{(\hat{l})} - A^{(\bar{l})}\|_2 \leq \frac{1}{6}\|A^{(\hat{l})} - A^{(\bar{l})}\|_2^2 + 6U^2$$

where $U \triangleq \max_{l=1,\ldots,L}\langle U_l, \tilde{R}_n^{(m)} - \mathbb{E}[\tilde{R}_n^{(m)}]\rangle$ with $U_l = (A^{(\hat{l})} - A^{(\bar{l})})/\|A^{(\hat{l})} - A^{(\bar{l})}\|_2$ if $A^{(\hat{l})} \neq A^{(\bar{l})}$ and $U_l = 0$ otherwise. It follows from the last display and the bound

$$\frac{1}{6}\|A^{(\hat{l})} - A^{(\bar{l})}\|_2^2 \leq \frac{1}{3}\|A^{(\hat{l})} - K\|_2^2 + \frac{1}{3}\|A^{(\bar{l})} - K\|_2^2$$

that

$$\|A^{(\hat{l})} - K\|_2^2 \leq 2\|A^{(\bar{l})} - K\|_2^2 + 9U^2. \tag{15}$$

Since $\bar{l}$ is arbitrary, to complete the proof it suffices to bound the random variable U in probability. We first obtain a bound for each of the variables $\zeta_l = \langle U_l, \tilde{R}_n^{(m)} - \mathbb{E}[\tilde{R}_n^{(m)}]\rangle$. Note that associating U_l with the corresponding $m \times m$ matrices that we will also denote by U_l, we can write $\zeta_l = \langle U_l, \hat{B} - B\rangle$ where $\hat{B} = (2/n)\sum_{i=n/2+1}^{n} \mathrm{x}_i(m)\mathrm{x}_i(m)^\top$, $B = K^{(m)} + \sigma^2 I^{(m)} = \mathbb{E}[\mathrm{x}_i(m)\mathrm{x}_i(m)^\top]$,

and $\mathrm{x}_i(m)$ are i.i.d. normal vectors with mean 0 and covariance matrix B (cf. the proof of Lemma 2) and $\langle \cdot, \cdot \rangle$ is the inner product of matrices. It follows that

$$\begin{aligned}\zeta_l &= \langle B^{1/2} U_l B^{1/2}, \frac{2}{n} \sum_{i=n/2+1}^{n} Z_i Z_i^\top - I^{(m)} \rangle \\ &= \mathrm{tr}\Big(\frac{2}{n} \sum_{i=n/2+1}^{n} B^{1/2} U_l B^{1/2} Z_i Z_i^\top - B^{1/2} U_l B^{1/2} \Big) \\ &= \frac{2}{n} \sum_{i=n/2+1}^{n} Z_i^\top D Z_i - \mathrm{tr}(D)\end{aligned}$$

where $Z_1, \ldots, Z_n$ are i.i.d. standard normal vectors in $\mathbb{R}^m$ and $D = B^{1/2} U_l B^{1/2}$. By the Hanson-Wright inequality (see, e.g., Rudelson and Vershynin (2013)) we have that for any $t > 0$, with probability at least $1 - e^{-t}$,

$$\left| \frac{2}{n} \sum_{i=n/2+1}^{n} Z_i^\top D Z_i - \mathrm{tr}(D) \right| \leq C \left(\frac{\|D\|_\infty t}{n} + \|D\|_F \sqrt{\frac{t}{n}} \right) \tag{16}$$

where $C > 0$ is an absolute constant. Since $\|U_l\|_2 \leq 1$ when considering U_l as a function (which is equivalent to $\|U_l\|_F \leq 1$ when considering U_l as a matrix) and $\|B\|_\infty \leq \lambda_{\max} + \sigma^2$ we have $\|D\|_\infty \leq \|D\|_F \leq \lambda_{\max} + \sigma^2$. Thus, with probability at least $1 - e^{-t}$

$$|\zeta_l| \leq C(\lambda_{\max} \vee \sigma^2) \left(\sqrt{\frac{t}{n}} + \frac{t}{n} \right)$$

where $C > 0$ is an absolute constant. The union bound argument gives that, with probability at least $1 - e^{-t}$,

$$U^2 = \max_{l=1,\ldots,L} \zeta_l^2 \leq C(\lambda_{\max} \vee \sigma^2)^2 \left(\sqrt{\frac{t + \log L}{n}} + \frac{t + \log L}{n} \right)^2$$

where $C > 0$ is an absolute constant. Combining this with (15) proves the theorem. □

We now apply Theorem 11 to $A^{(l)} = \hat{A}^{(l)}$ where the estimators $\hat{A}^{(1)}, \ldots, \hat{A}^{(L)}$ are defined in (4). Combining Theorems 3, 11 and the fact that, for a random variable η, $\mathbb{E}[|\eta|] = \int_0^\infty \mathbb{P}(|\eta| \geq t)\, dt$ we get the following result.

THEOREM 12. *Let each of the estimators $\hat{A}^{(l)}$ satisfy the conditions of Theorem 3. Then*

$$\mathbb{E}\left[\|\hat{A}^{(\hat{l})} - K\|_2^2\right] \le C \min_{1\le l\le L} \inf_{A\in\mathcal{S}_l,\, A\ge 0} \left\{\|A-K\|_2^2 + v_n(A,l,l)\right\} + C[\lambda_{\max}\vee\sigma^2]^2 \frac{\log L}{n}$$

for all K satisfying (2) with $\sum_{k=1}^r \lambda_k < \infty$. Here, $C>0$ is an absolute constant.

We now fix $L = n$. Using Theorem 12, Theorem 7, Theorem 9 and Corollary 4, we obtain the following result.

THEOREM 13. *Let each of the estimators $A^{(l)} = \hat{A}^{(l)}$ satisfy the conditions of Theorem 3. Let $\hat{A}^{(\hat{l})}$ be the aggregated estimator with $\hat{l}$ defined in (14) with $L = n$.*

(i) For any $r \ge 1$, $c_* > 0$ and $s > 0$ such that $1 \le r \le n^{1+2s}$, we have

$$\sup_{K\in\mathcal{K}_r(s,c_*;\lambda_{\max})} \mathbb{E}\|\hat{A}^{(\hat{l})} - K\|_2^2 \le C \min\left(rn^{-\frac{2s}{2s+1}},\, r^{\frac{1}{s+1}} n^{-\frac{s}{s+1}}\right),$$

where $C>0$ is a constant that can depend only on $\lambda_{\max}, \sigma^2, c_*$, and s.

(ii) For any $r \ge 1$, $\rho \ge 1$, $s > 0$, $\lambda_{\max} > 0$, $\sigma^2 \ge 0$ such that $\rho^2 \le (\lambda_{\max}+\sigma^2)^2 \min\left(rn^{2s}, n^{1+2s}\right)$, we have

$$\sup_{K\in\overline{\mathcal{K}}_r(s,\rho;\lambda_{\max})} \mathbb{E}\|\hat{A}^{(\hat{l})} - K\|_2^2 \le C \min\left(\left(\frac{r}{n}\right)^{2s/(2s+1)},\, n^{-s/(s+1)}\right),$$

where $C>0$ is a constant that can depend only on $\lambda_{\max}, \sigma^2, \rho$, and s.

(iii) If $(r\wedge l)l \ge \log n$ and $l \le n$, then for any $\lambda_{\max} > 0$,

$$\sup_{K\in\mathcal{K}_{r,l}(\lambda_{\max})} \mathbb{E}[\|\hat{A}^{(\hat{l})} - K\|_2^2] \le C\frac{(r\wedge l)l}{n}$$

where $C>0$ is a constant that can depend only on $\lambda_{\max}$ and σ^2.

The conditions $r \le n^{1+2s}$ and $\rho^2 \le (\lambda_{\max}+\sigma^2)^2 \min\left(rn^{2s}, n^{1+2s}\right)$ are rather mild. Indeed, if r and ρ are fixed quantities, then these conditions are satisfied for n large enough. Theorem 13 shows that the estimator $\hat{A}^{(\hat{l})}$ is adaptive to the unknown parameters r and s on the scale of classes $\overline{\mathcal{K}}_r(s,\rho;\lambda_{\max})$ and $\mathcal{K}_r(s,c_*;\lambda_{\max})$ that no price is paid in the rate as compared to the non-adaptive estimators of Theorems 7 and 9. The same estimator is adaptive on the scale of classes $\mathcal{K}_{r,l}(\lambda_{\max})$, again with no price to be paid, for a wide range of values of l and r.

5. Estimation of σ^2

We now tackle the estimation of the unknown variance σ^2. We use the simple idea that $\langle e_l, S\rangle$ becomes negligible for large l when Assumption 8 is satisfied. Therefore, we propose the following (biased) estimator of σ^2 based on an independent copy X of the process (1):

$$\hat{\sigma}^2 = \frac{1}{M}\left\|\dot{X}^{(L+M)} - \dot{X}^{(L)}\right\|_2^2, \quad L = e^n, M \geq 1. \tag{17}$$

THEOREM 14. *Let $n, l \geq 1$ be integers and let $X_1(\cdot), \ldots, X_n(\cdot)$ be i.i.d. realizations of the process $X(\cdot)$ satisfying (1). Let Assumption 8 be satisfied. For any $t > 0$, we have with probability at least $1 - e^{-t}$*

$$|\hat{\sigma}^2 - \sigma^2| \lesssim \max\left\{c_*^2 r \lambda_{\max} L^{-2s}(1 \vee \sqrt{t} \vee t), \sigma^2\sqrt{\frac{t}{M}}, \frac{t}{M}\right\}.$$

Proof. We have, in view of Plancherel inequality, that

$$\begin{aligned}\hat{\sigma}^2 - \sigma^2 &= \frac{1}{M}\left\|S^{(L+M)} - S^{(L)}\right\|_2^2 + \frac{2}{M}\langle S^{(L+M)} - S^{(L)}, \dot{W}^{(L+M)} - \dot{W}^{(L)}\rangle \\ &\qquad + \frac{1}{M}\left\|\dot{W}^{(L+M)} - \dot{W}^{(L)}\right\|_2^2 - \sigma^2 \\ &= \frac{1}{M}\sum_{l=L}^{L+M}\langle S, e_l\rangle^2 + \frac{2}{M}\sum_{l=L}^{L+M}\langle S, e_l\rangle z_l + \frac{1}{M}\sum_{l=L}^{L+M} z_l^2 - \sigma^2 = I + II + III,\end{aligned} \tag{18}$$

where $z_L, \ldots, z_{L+M}$ are i.i.d. standard normal random variables also independent from S.

We now take the expectation

$$\mathbb{E}\left[\hat{\sigma}^2\right] - \sigma^2 = \frac{1}{M}\sum_{l=L}^{L+M}\sum_{j=1}^{r}\lambda_j\langle\varphi_j, e_l\rangle^2.$$

Note that $\langle\varphi_j, e_l\rangle^2 \leq \|\varphi_j - \varphi_j^{(l-1)}\|_2^2$. In view of Assumption 8, we get

$$\frac{1}{M}\sum_{l=L}^{L+M}\sum_{j=1}^{r}\lambda_j\langle\varphi_j, e_l\rangle^2 \leq r\lambda_{\max}\frac{c_*^2}{M}\sum_{l=L-1}^{L+M-1} l^{-2s} \lesssim c_*^2 r\lambda_{\max}L^{-2s}.$$

The bound in probability follows easily from the representation (18). Indeed, the second term can be treated using standard deviations bounds for Gaussian combined with a conditioning argument. The third term can be treated with a standard deviation inequality for chi-square distributions. The first term can be treated using (16) again. More specifically, set $\xi = (\xi_1, \dots, \xi_r)^\top$ and $A = (a_{j,j'})_{1 \le j,j' \le r}$ with

$$a_{j,j'} = \frac{\sqrt{\lambda_j \lambda_{j'}}}{M} \sum_{l=L}^{L+M} \langle \varphi_j, e_l \rangle \langle \varphi_{j'}, e_l \rangle.$$

Then, we have

$$\frac{1}{M} \sum_{l=L}^{L+M} \langle S, e_l \rangle^2 - \mathbb{E}\left[\frac{1}{M} \sum_{l=L}^{L+M} \langle S, e_l \rangle^2 \right] = \xi^\top A \xi - \mathbb{E}[\xi^\top A \xi],$$

with $\|A\|_F \lesssim c_*^2 r \lambda_{\max} L^{-2s}$ and $\|A\|_\infty \lesssim c_*^2 \sqrt{r} \lambda_{\max} L^{-2s}$.

An union bound argument gives the result. Details of the proof are omitted here.

□

6. Minimax lower bound

In this section, we show that the upper bound of Corollary 4 cannot be improved in a minimax sense.

THEOREM 15. *Let $1 \le r < \infty$ and let $\lambda_{\max} > 0$ be a given constant. Then there exist absolute constants $c_0 > 0$ and $0 < c_1 < 1$ such that, for any integers n and l satisfying $l \ge 2$, $n \ge l$, we have*

$$\inf_{\hat{K}_n} \sup_{K \in \mathcal{K}_{r,l}(\lambda_{\max})} \mathbb{P}\left(\|\hat{K}_n - K\|_2^2 \ge c_0 [\lambda_{\max} \wedge \sigma^2]^2 \frac{(r \wedge l) l}{n} \right) > c_1$$

where $\inf_{\hat{K}_n}$ denotes the infimum over all estimators of K.

Proof. Let first $r \le l/2$. Consider the vector-functions $e(t) = (e_1(t), \dots, e_l(t))$ and $\varphi(t) = (\varphi_1(t), \dots, \varphi_r(t))$ and a subset of $\mathcal{K}_{r,l}(\lambda_{\max})$ composed of kernels K satisfying (2) with $\lambda_j \equiv \gamma$ and

$$\varphi(t) = He(t)$$

for suitable $\gamma > 0$ and suitable $r \times l$ matrices H. Orthonormality of functions φ_j implies that H must satisfy $HH^\top = I_r$ where I_r is the $r \times r$ identity matrix, i.e., the rows of H should be orthonormal. To each such matrix H we associate a linear subspace U_H of $\mathbb{R}^l$, which is the linear span of the r rows of H. Clearly, $\dim(U_H) = r$ and $H^\top H$ is the orthogonal projector onto U_H in $\mathbb{R}^l$.

Note that the set of all such spaces U_H is the Grassmannian manifold $G_r(\mathbb{R}^l)$, i.e., the set of r-dimensional linear subspaces of $\mathbb{R}^l$. The Grassmannian manifold $G_r(\mathbb{R}^l)$ is a smooth manifold of dimension $d = r(l-r)$. A natural metric $d(\cdot,\cdot)$ on $G_r(\mathbb{R}^l)$ is defined as follows: for $U, \bar{U} \in G_r(\mathbb{R}^l)$,

$$d(U, \bar{U}) \triangleq \|P_U - P_{\bar{U}}\|_F = \|H^\top H - \bar{H}^\top \bar{H}\|_F$$

where P_U is the orthogonal projector onto U and H, $\bar{H}$ are the $r \times l$ matrices with orthonormal rows associated to U and $\bar{U}$ respectively. We refer to Mattila (1995) and Milnor and Stasheff (1974) for more details on the Grassmannian manifold.

From now on, we will identify $U \in G_r(\mathbb{R}^l)$ with the associated orthogonal projector $P_U = H^\top H$. The behavior of entropy numbers of the Grassmannian manifold is well studied (Szarek (1982), see also Proposition 8 in Pajor (1998)). In particular, for any $\epsilon \in (0,1)$ there exists a family of orthogonal projectors $\mathcal{U} \subset G_r(\mathbb{R}^l)$ such that

$$|\mathcal{U}| \geq \left\lfloor \frac{\bar{c}}{\epsilon} \right\rfloor^d \quad \text{and} \quad \bar{c}\epsilon\sqrt{r} \leq \|P - Q\|_F \leq \frac{1}{\bar{c}}\epsilon\sqrt{r},\ \forall P, Q \in \mathcal{U},\ P \neq Q, \quad (19)$$

for some small enough universal constant $\bar{c} > 0$. Here $|\mathcal{U}|$ denotes the cardinality of $\mathcal{U}$. We take in what follows $\epsilon = \bar{c}/2$. Set $N = |\mathcal{U}|$ and $\mathcal{U} = \{P_{(1)}, \ldots, P_{(N)}\}$. The associated H-matrices will be denoted by $H_1, \ldots, H_N$. Let K_j be a kernel of the form (2) with eigenvalues $\lambda_i \equiv \gamma, i = 1, \ldots, r$, and

$$\varphi(t) = H_j e(t), \quad j = 1, \ldots, N,$$

where $\gamma = a(\sigma^2 \wedge \lambda_{\max})\sqrt{\frac{l}{n}}$ and $a \in (0,1)$ is an absolute constant to be chosen later. Consider the set $\mathcal{K}' = \{K_1, \ldots, K_N\}$. Clearly, we have $\mathcal{K}' \subset \mathcal{K}_{r,l}(\lambda_{\max})$.

We now evaluate the Kullback-Leibler divergence between two probability measures induced by the observations $\{X_1(t), \ldots, X_n(t), t \in [0,1]\}$ corresponding to the kernels K_1 and K_j (with $j \neq 1$). Using the Girsanov formula and the fact that K_j is bilinear in $\{e_k\}$ it is easy to check that this divergence

is equal to the Kullback-Leibler divergence between the n-product distributions of the associated Gaussian vectors $\left(\int_0^1 e_1(t)dX(t), \ldots, \int_0^1 e_l(t)dX(t)\right)$. If $K = K_j$ this vector is distributed as $\mathcal{N}(0, \Sigma_j)$ with $\Sigma_j = \sigma^2 I_l + \gamma P_{(j)} = (\sigma^2+\gamma)P_{(j)} + \sigma^2 P_{(j)}^\perp$ and $P_{(j)}^\perp = I_l - P_{(j)}$. Denote the corresponding Gaussian measure by $\mathbb{P}_j$ and by $\mathbb{P}_j^{\otimes n}$ its n-product. Let $\mathrm{KL}(\mathbb{P}, \mathbb{Q})$ be the Kullback-Leibler divergence between two probability measures $\mathbb{P}$ and $\mathbb{Q}$.

It is easy to see that all matrices Σ_j have the same eigenvalues. Thus, for any $2 \le j \le N$ we have

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_1^{\otimes n}, \mathbb{P}_j^{\otimes n}) &= n\,\mathrm{KL}(\mathbb{P}_1, \mathbb{P}_j) \\
&= \frac{n}{2}\left[\mathrm{tr}(\Sigma_1^{-1}\Sigma_j) - l - \log\left(\det(\Sigma_1^{-1}\Sigma_j)\right)\right] \\
&= \frac{n}{2}\left[\mathrm{tr}(\Sigma_1^{-1}(\Sigma_j - \Sigma_1)\right].
\end{aligned}$$

Now, $\Sigma_1^{-1} = \frac{1}{\sigma^2+\gamma}P_{(1)} + \frac{1}{\sigma^2}P_{(1)}^\perp$, which yields

$$\begin{aligned}
\mathrm{tr}(\Sigma_1^{-1}(\Sigma_j - \Sigma_1)) &= \frac{\gamma}{\sigma^2+\gamma}\mathrm{tr}(P_{(1)}(P_{(j)} - P_{(1)})) + \frac{\gamma}{\sigma^2}\mathrm{tr}(P_{(1)}^\perp(P_{(j)} - P_{(1)})) \\
&= \left(\frac{\gamma}{\sigma^2+\gamma} - \frac{\gamma}{\sigma^2}\right)\left(\mathrm{tr}(P_{(1)}P_{(j)}) - r\right) \\
&= \frac{\gamma^2}{2(\sigma^2+\gamma)\sigma^2}\|P_{(1)} - P_{(j)}\|_F^2 \\
&\le \frac{r\gamma^2}{8(\sigma^2+\gamma)\sigma^2}
\end{aligned}$$

where we have used (19) with $\epsilon = \bar{c}/2$, and the fact that $\mathrm{tr}(P_{(1)}P_{(j)}) = r - \|P_{(1)} - P_{(j)}\|_F^2/2$. Combining the last two displays, we find

$$\mathrm{KL}(\mathbb{P}_1^{\otimes n}, \mathbb{P}_j^{\otimes n}) \le \frac{a^2(\lambda_{\max} \wedge \sigma^2)^2}{8(\sigma^2+\gamma)\sigma^2} rl \le \frac{a^2}{8} rl, \quad \forall\ 2 \le j \le N.$$

Recall that we assume $r \le l/2$, so that the dimension of the Grassmannian satisfies $d = r(l-r) \ge rl/2$. Consequently, in view of (19) with $\epsilon = \bar{c}/2$, we have $\log|\mathcal{U}| \ge rl(\log 2)/2$. Thus, choosing $a = (\log 2)^{1/2}/2$ we obtain

$$\mathrm{KL}(\mathbb{P}_1^{\otimes n}, \mathbb{P}_j^{\otimes n}) \le \frac{1}{16}\log|\mathcal{U}|, \quad \forall\ 2 \le j \le N.$$

Next, for any $1 \leq i, j \leq N$ with $i \neq j$,

$$\|K_i - K_j\|_2^2 = \gamma^2 \|H_i^\top H_i - H_j^\top H_j\|_F^2 = \gamma^2 \|P_{(i)} - P_{(j)}\|_F^2 \geq ca^2[\sigma^4 \wedge \lambda_{\max}^2]\frac{rl}{n},$$

where $c > 0$ is a absolute constant and the last inequality is due to (19). The result now follows from the last two displays by application of Theorem 2.5 in Tsybakov (2009).

Finally, consider the case $r > l/2$. Note that the classes $\mathcal{K}_{r,l}(\lambda_{\max})$ are nested in r. Assuming w.l.o.g. that l is even, we get that the minimax risk over $\mathcal{K}_{r,l}(\lambda_{\max})$ is bounded from below by the minimax risk on $\mathcal{K}_{l/2,l}(\lambda_{\max})$. But the minimax risk on $\mathcal{K}_{l/2,l}(\lambda_{\max})$ has been already treated above and we have proved that the lower rate is of the order l^2/n, which is the desired rate when $r > l/2$. □

REMARK 16. *It is possible to prove a minimax lower bound ensuring that the bound in Theorem 7 is optimal at least regarding the n dependence. Indeed, by a similar argument to that used in the proof of Theorem 15, we can prove the existence of an absolute constant* $0 < c_2 < 1$ *and a constant* $c_3 > 0$ *possibly depending on* $\sigma^2, \lambda_{\max}, \rho, r$ *such that, for any integer* $n \geq 1$ *we have*

$$\inf_{\hat{K}_n} \sup_{K \in \overline{\mathcal{K}}_r(s,\rho;\lambda_{\max})} \mathbb{P}\left(\|\hat{K}_n - K\|_2^2 \geq c_3 \min\left(n^{-\frac{2s}{2s+1}}, n^{-s/(s+1)}\right)\right) > c_2$$

where $\inf_{\hat{K}_n}$ *denotes the infimum over all estimators of* K*. Specifying the dependence of the minimax rate on parameters* $\sigma^2, \lambda_{\max}, \rho, r$ *remains an interesting open question. A similar argument should also provide a minimax lower bound for the class* $\mathcal{K}_r(s, c_*; \lambda_{\max})$ *matching the upper bound of Theorem 9 at least regarding the dependence in n.*

References

Bigot, J., Biscay, R., Loubes, J.-M., Muniz-Alvarez, L., 2010. Nonparametric estimation of covariance functions by model selection. Electron. J. Statist., 4, 822–855.

Bunea, F., Xiao, L., 2013. On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to FPCA.

Hall, P., Müller, H.-G., Wang, J.-L., 2006. Properties of principal component methods for functional and longitudinal data analysis. Ann. Statist. 34 (3), 1493–1517.

Koltchinskii, V., Lounici, K., Tsybakov, A. B., 2011. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. Ann. Statist. 39 (5), 2302–2329.

Koltchinskii, V. and Rangel, P., 2013. Low rank estimation of smooth kernels on graphs. Ann. Statist. 41 (2), 604–640.

Lounici, K., 2014. High-dimensional covariance matrix estimation with missing observations. Bernoulli 20 (3), 1029–1058.

Mattila, P., 1995. Geometry of sets and measures in Euclidean spaces. Vol. 44 of Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, fractals and rectifiability.

Milnor, J. W., Stasheff, J. D., 1974. Characteristic classes. Princeton University Press, Princeton, N. J.; University of Tokyo Press, Tokyo, Annals of Mathematics Studies, No. 76.

Pajor, A., 1998. Entropy of the Grassmann manifold. Convex Geometry Analysis, MSRI Publications 34, 181–188.

Rudelson, M., Vershynin, R., 2013. Hanson-Wright inequality and sub-Gaussian concentration. Electron. Commun. Probab. 18, no. 82, 9.

Szarek, S. J., 1982. Nets of Grassmann manifold and orthogonal group. In: Proceedings of research workshop on Banach space theory (Iowa City, Iowa, 1981). Univ. Iowa, Iowa City, IA, pp. 169–185.

Tsybakov, A. B., 2009. *Introduction to Nonparametric Estimation.* Springer, New York.

Vershynin, R., 2012. Introduction to the non-asymptotic analysis of random matrices. In: Compressed sensing. Cambridge Univ. Press, Cambridge, pp. 210–268.

Yao, F., Müller, H.-G., Wang, J.-L., 2005. Functional data analysis for sparse longitudinal data. J. Amer. Statist. Assoc. 100 (470), 577–590.