Nonparametric Covariance Function Estimation for Functional and Longitudinal Data

T. Tony Cai[1] and Ming Yuan[2]
University of Pennsylvania and Georgia Institute of Technology

(September 24, 2010)

Abstract

Covariance function plays a critical role in functional and longitudinal data analysis. In this paper, we consider nonparametric covariance function estimation using a reproducing kernel Hilbert space framework. A regularization method is introduced through a careful characterization of the function space in which a covariance function resides. It is shown that the procedure enjoys desirable theoretical and numerical properties. In particular, even though the covariance function is bivariate, the rates of convergence attained by the regularization method are very similar to those typically achieved for estimating univariate functions. Our results generalize and improve some of the known results in the literature both for estimating the covariance function and for estimating the functional principal components. The procedure is easy to implement and its numerical performance is investigated using both simulated and real data. In particular our method is illustrated in an analysis of a longitudinal CD4 count data from an HIV study.

Key words: Covariance function, convergence rate, functional data analysis, longitudinal data analysis, method of regularization, phase transition, reproducing kernel Hilbert space, Sobolev space, tensor product space.

[1]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104.
The research of Tony Cai was supported in part by NSF FRG Grant DMS-0854973.
[2]Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332. The research of Ming Yuan was supported in part by NSF Career Award DMS-0846234.

1 Introduction

Covariance function plays a critical role in functional and longitudinal data analysis. Important examples include functional principal component analysis, functional linear regression, and functional discriminant analysis. (see, e.g., Diggle et al., 2002; Ramsay and Silverman, 2002; 2005; Ferraty and Vieu, 2006). Covariance function summarizes the dependency of observations at different locations and characterizes many of the primary properties, such as smoothness and regularity, of the sample path (see, e.g., Stein, 1999). It is of significant interest to estimate the covariance function based on a random sample.

In the ideal setting where the whole random functions are observed in the continuum without measurement errors, it is known that under mild regularity conditions the sample covariance function converges to the population covariance function at the usual parametric rate, in terms of integrated squared error (see, e.g., Bosq, 2000). However, the sample covariance function is only of limited practical interest. In many applications such as longitudinal studies, it is only possible to observe each curve at discrete sampling locations and with measurement errors. In this paper, we study covariance function estimation in such a setting.

Let $X(\cdot)$ be a second-order stochastic process defined over a compact domain $\mathcal{T} \subset \mathbb{R}$, e.g., $\mathcal{T} = [0, 1]$. Its covariance function is defined as

$$C_0(s,t) = \mathbb{E}\left([X(s) - \mathbb{E}(X(s))][X(t) - \mathbb{E}(X(t))]\right), \qquad \forall s, t \in \mathcal{T}. \tag{1}$$

Let $\{X_1, X_2, \ldots, X_n\}$ be a collection of independent realizations of X. Suppose that we observe the random functions X_i at discrete points with measurement errors,

$$Y_{ij} = X_i(T_{ij}) + \varepsilon_{ij}, \qquad j = 1, \ldots, m; \quad i = 1, \ldots, n, \tag{2}$$

where the sampling locations T_{ij} are independently drawn from a common distribution on $\mathcal{T}$, and ε_{ij} are independent and identically distributed measurement errors with mean zero and finite variance σ_0^2. In addition, the random functions X, sampling locations T, and measurement errors ε are mutually independent. Such a model is commonly adopted and suitable for a large number of applications in functional and longitudinal data analysis (see, e.g., Shi, Weiss and Taylor, 1996; Staniswalis and Lee, 1998; James and Hastie, 2001; Diggle et al., 2002; Müller, 2005; Ramsay and Silverman, 2005). The requirement that the same number of observations are taken from different functions is not essential and only in place for ease of presentation and better illustration of the interplay of the number of curves and sampling frequency on estimating the covariance function. More general sampling schemes will be discussed in Section 6.

A number of nonparametric approaches have been introduced to estimate the covariance function based on model (2). See James, Hastie and Sugar (2000), Rice and Wu (2001), Yao, Müller and Wang (2005), Hall, Müller and Wang (2006), Paul and Peng (2009) among others. A two-step procedure is routinely taken in practice, where each curve is first estimated through smoothing and the covariance function C_0 is then estimated by the sample covariance function of the smoothed curves (see, e.g., Ramsay and Silverman, 2002). Despite its popularity, there is relatively little theoretical guidance as to how it performs. Hall, Müller and Wang (2006) considered covariance function estimation assuming that the sample path of X is twice differentiable and the curves are either densely sampled with $m \gg n^{1/4+\delta}$ for some $\delta > 0$ or sparsely sampled with m fixed. It was shown that when $m \gg n^{1/4+\delta}$, the two-step procedure can estimate C_0 as well as if the whole curves are observed, i.e., at the rate of $1/n$ in terms of integrated squared error. Alternatively, one could treat the problem of estimating C_0 as a bivariate smoothing task. In the case where m is held fixed as the number of curves n increases, Hall, Müller and Wang (2006) show that C_0 can be estimated by a local polynomial estimator at the rate of $n^{-2/3}$, the usual rate for bivariate smoothing. These results appear to suggest that the different techniques should be employed depending on the sampling frequency with the two-step procedure preferable in the densely sampled case, and the other approach more appropriate in the sparsely sampled case where m is finite. It also remains unclear what happens in more general setting where the random functions are of different smoothness with the number of sampling points on each curve possibly growing slowly with n.

The gap between the requirement on sampling frequencies leaves us in an uncomfortable situation to choose between these two methods in practice. It is therefore of great practical importance to address this dilemma and develop a more generally applicable and theoretically justifiable approach that can be used under different sampling frequencies. The main goal of this paper is to offer such a solution in a more general setting using a reproducing kernel Hilbert space (RKHS) framework. The proposed method is motivated by a careful examination of the function space in which the covariance function C_0 resides. By assuming that the sample path of X belongs to an RKHS, we show that C_0 necessarily comes from a tensor product space naturally connected with the differentiability of the sample path of X. This observation reveals the distinction between the covariance function estimation problem and the conventional bivariate smoothing problem as the parameter space is no longer the usual Sobolev type spaces. To take advantage of this special feature, a novel method of regularization for estimating C_0 is then introduced. The proposed method enjoys desirable theoretical properties. It is shown that the estimator converges to C_0 at the rate of $O_p((nm/\log n)^{-2\alpha/(2\alpha+1)} + n^{-1})$ under integrated squared error when assuming that the sample path of X is $\alpha(> 1/2)$ times differentiable. Our results characterize the joint effect of the number

of curves n and the sampling frequency m, and improves those available in the literature.

The rate exhibits an interesting phase transition effect. Namely, when the functions are sparsely sampled, i.e., $m \ll n^{1/2\alpha} \log n$, the convergence rates are determined by the total number of observations $N := nm$. On the other hand, when the functions are densely sampled, i.e., $m \gg n^{1/2\alpha} \log n$, the rates remain $1/n$ regardless of m. Perhaps more surprisingly, our results suggest that the proposed method is to a certain degree immune to the "curse of dimensionality". Since covariance functions are bivariate functions, one would naturally anticipate nonparametric estimation of C_0 to be more difficult than estimating a univariate function such as the mean function. It is, however, somewhat unexpected to see that the only price we pay here is at most a logarithmic factor.

Despite the generality of the method, we show that the estimators can be computed rather efficiently thanks to a representer theorem which makes our procedure easily implementable and enables us to take advantage of the existing techniques and algorithms for smoothing splines to compute the estimator. Moreover, we show that our estimate of the covariance function allows explicit calculation of the functional principal components and therefore avoids the usual numerical approximation (see, e.g., Ramsay and Silverman, 2005) where one has to trade numerical accuracy with computational efficiency. An R package implementing our method has been developed and is publicly available on the web. Numerical performance of the estimator is investigated using both simulated and real data. In particular our method is illustrated in an analysis of a longitudinal CD4 count data for a sample of AIDS patients.

The rest of the paper is organized as follows. Section 2 introduces the methodology for estimating the covariance function C_0. The implementation of the proposed method for covariance function estimation as well as computation of the functional principal components are discussed in detail in Section 3. The practical merits of the method are demonstrated by simulations and by an application to a longitudinal CD4 count dataset in Section 4. We study the rates of convergence of the proposed estimate in Section 5. Section 6 discusses possible extensions of our work and its connections with other related problems. The main results are proved in Section 7 and the proofs of some technical lemmas are given in the Appendix.

2 Methodology

In this section, we introduce a regularization method for estimating the covariance function based on discretely sampled data with noise. The sample path of X is assumed to be smooth in that it belongs to certain reproducing kernel Hilbert space (RKHS). Our method is motivated by a careful examination of the function space in which the covariance function C_0 resides. We first show that

C_0 necessarily comes from a tensor product space naturally connected with the differentiability of the sample path of X and then introduce our estimation procedure based on this important observation.

2.1 Covariance function and tensor product spaces

We begin by reviewing some basic facts of RKHS. Let $\mathcal{H}$ be a Hilbert space of functions on $\mathcal{T}$ associated with a inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. A symmetric bivariate function $K : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ is said to be its reproducing kernel if for every $t \in \mathcal{T}$, $K(\cdot, t) \in \mathcal{H}$ and $f(t) = \langle f, K(\cdot, t) \rangle_{\mathcal{H}}$ for every $f \in \mathcal{H}$. Since an RKHS $\mathcal{H}$ can be identified with its reproducing kernel K, we shall write $\mathcal{H}(K)$ in what follows to signify the correspondence. As a concrete example, Sobolev space $\mathcal{W}_2^2$ of periodic functions on $\mathcal{T} = [0, 1]$ that integrate to 0 and have square integrable second derivative forms an RKHS when endowed with squared norm $\int (f'')^2$ with reproducing kernel

$$K(s, t) = \frac{1}{4} B_2(s) B_2(t) - \frac{1}{24} B_4(|s - t|) \tag{3}$$

where $B_r(\cdot)$ is the rth Bernoulli polynomial. The readers are referred to Aronszajn (1950) for detailed discussions on RKHS.

Assuming that the reproducing kernel K is square integrable, Mercer's theorem states that K admits the following eigenvalue decomposition:

$$K(s, t) = \sum_{k \geq 1} \rho_k \varphi_k(s) \varphi_k(t) \tag{4}$$

where the nonnegative scalars $\rho_1 \geq \rho_2 \geq \ldots$ are the eigenvalues and the eigenfunctions $\{\varphi_k(\cdot) : k \geq 1\}$ are a set of orthonormal basis of $L_2(\mathcal{T})$, i.e.,

$$\langle \varphi_j, \varphi_k \rangle := \int_{\mathcal{T}} \varphi_j(t) \varphi_k(t) dt = \delta_{jk}, \tag{5}$$

where δ is Kronecker's delta.

In this paper we shall assume that the sample path of X belongs to an RKHS $\mathcal{H}(K)$. Then we can write

$$X(\cdot) = \sum_{k \geq 1} x_k \varphi_k(\cdot) \tag{6}$$

where $x_k = \langle X, \varphi_k \rangle_{L_2}$ is the Fourier coefficient of X with respect to $\{\varphi_k(\cdot) : k \geq 1\}$. It is worth pointing out that, despite the resemblance, the expression given by (6) differs from the usual Karhunen-Loève expansion. In (6), the basis functions $\{\varphi_k : k \geq 1\}$ are identified with the reproducing kernel K and the function space $\mathcal{H}(K)$ whereas the basis functions used in Karhunen-Loève expansion are the eigenfunctions of the covariance function C_0 and are not known a priori.

With the expansion (6), the first two moments of X can be given in terms of those for the Fourier coefficients:

$$\begin{aligned} \mu_0(t) &:= \mathbb{E}X(t) = \sum_{k\geq 1} \mathbb{E}(x_k)\varphi_k(t); \\ C_0(s,t) &= \sum_{j,k\geq 1} \mathrm{Cov}(x_k, x_j)\varphi_j(s)\varphi_k(t). \end{aligned}$$

The above expression naturally identifies C_0 with the tensor product space $\mathcal{H}(K \otimes K) := \mathcal{H}(K) \otimes \mathcal{H}(K)$, an RKHS associated with reproducing kernel

$$K \otimes K((s_1,t_1),(s_2,t_2)) = K(s_1,s_2)K(t_1,t_2). \tag{7}$$

With slight abuse of notation, let the operator $\otimes$ also denote the tensor product of real-valued functions on $\mathcal{T}$, i.e.,

$$f_1 \otimes f_2(s,t) = f_1(s)f_2(t). \tag{8}$$

Then the tensor product space $\mathcal{H}(K \otimes K)$ can be characterized as follows. If $f_1, f_2 \in \mathcal{H}(K)$, then $f_1 \otimes f_2 \in \mathcal{H}(K \otimes K)$ with

$$\|f_1 \otimes f_2\|_{\mathcal{H}(K\otimes K)} = \|f_1\|_{\mathcal{H}(K)}\|f_2\|_{\mathcal{H}(K)}. \tag{9}$$

Linear combinations of such tensor products are dense in $\mathcal{H}(K \otimes K)$. The norm of a linear combination is given by

$$\left\| \sum_{j=1}^{k} f_{j_1} \otimes f_{j_2} \right\|^2_{\mathcal{H}(K\otimes K)} = \sum_{i,j=1}^{k} \langle f_{i_1}, f_{j_1}\rangle_{\mathcal{H}(K)} \langle f_{i_2}, f_{j_2}\rangle_{\mathcal{H}(K)}. \tag{10}$$

Our first result, which is proved in Section 7, shows that the covariance function C_0 is a member of $\mathcal{H}(K \otimes K)$.

Theorem 1 *If the sample path of X belongs to $\mathcal{H}(K)$ almost surely such that $\mathbb{E}\|X\|^2_{\mathcal{H}(K)} < \infty$. Then C_0 belongs to the tensor product space $\mathcal{H}(K \otimes K)$.*

Since $\{\varphi_k : k \geq 1\}$ is an orthonormal basis of $L_2(\mathcal{T})$, it follows that $\{\varphi_j(\cdot)\varphi_k(\cdot) : j,k \geq 1\}$ forms an orthonormal basis of $L_2(\mathcal{T} \times \mathcal{T})$. For a function $f \in L_2(\mathcal{T} \times \mathcal{T})$, write

$$f(s,t) = \sum_{j,k\geq 1} f_{jk}\varphi_j(s)\varphi_k(t). \tag{11}$$

Recall that $\{\rho_k : k \geq 1\}$ is the set of eigenvalues of K. It is clear (see, e.g., Wahba, 1990) that the tensor product space $\mathcal{H}(K \otimes K)$ contains all functions f such that

$$\|f\|^2_{\mathcal{H}(K\otimes K)} := \sum_{j,k\geq 1} \rho_j^{-1}\rho_k^{-1} f_{jk}^2 < \infty. \tag{12}$$

As an example, we revisit Sobolev space $\mathcal{W}_2^2$ defined on $\mathcal{T} = [0,1]$. In this case, $\mathcal{H}(K \otimes K)$ consists of periodic bivariate functions such that all its derivatives $\partial^{k+l} f(s,t)/\partial s^k \partial t^l$ are square integrable for $0 \le k, l \le 2$. It is noteworthy that this space differs from the second order Sobolev space defined on a two dimensional torus.

2.2 Covariance function estimate

Theorem 1 shows that the covariance function C_0 belongs to the tensor product space $\mathcal{H}(K \otimes K)$. We are now in position to introduce the regularization estimate of C_0 based on this important fact.

It is easy to see that the random vector $Y_{i\cdot} := (Y_{i1}, \ldots, Y_{im})'$ has mean $\mu_{i\cdot} := (\mu_0(T_{i1}), \ldots, \mu_0(T_{im}))$ and covariance matrix $\Sigma_i := [C_0(T_{ij_1}, T_{ij_2})]_{1 \le j_1, j_2 \le m} + \sigma_0^2 I$. If the mean function μ_0 is known, one can regress $[Y_{ij_1} - \mu_0(T_{ij_1})][Y_{ij_2} - \mu_0(T_{ij_2})]$ on (T_{ij_1}, T_{ij_2}) to estimate C_0. In the light of Theorem 1, we consider the following method of regularization:

$$\hat{C}_\lambda = \underset{C \in \mathcal{H}(K \otimes K)}{\operatorname{argmin}} \left\{ \ell_n(C) + \lambda \|C\|^2_{\mathcal{H}(K \otimes K)} \right\}, \tag{13}$$

where

$$\ell_n(C) = \frac{1}{nm(m-1)} \sum_{i=1}^{n} \sum_{1 \le j_1 \ne j_2 \le m} \left([Y_{ij_1} - \mu_0(T_{ij_1})][Y_{ij_2} - \mu_0(T_{ij_2})] - C(T_{ij_1}, T_{ij_2}) \right)^2 \tag{14}$$

and $\lambda \ge 0$ is a tuning parameter that balances the fidelity to the data measured by ℓ_n and smoothness of the estimate measured by the squared RKHS norm.

Of course, the mean function μ_0 is typically unknown in practice. In this case, we will replace μ_0 by its estimate in defining ℓ_n:

$$\ell_n(C) = \sum_{i=1}^{n} \sum_{1 \le j_1 \ne j_2 \le m} \left([Y_{ij_1} - \hat{\mu}(T_{ij_1})][Y_{ij_2} - \hat{\mu}(T_{ij_2})] - C(T_{ij_1}, T_{ij_2}) \right)^2, \tag{15}$$

where $\hat{\mu}$ is an estimate of μ_0. The problem of estimating the mean function μ_0 has been extensively studied. In particular, Cai and Yuan (2010) investigated the optimal rates for estimating μ_0 under various settings. It was shown that with appropriate tuning parameters, the smoothing splines estimate obtained by regressing Y_{ij} on T_{ij} altogether achieves the optimal rate of convergence. We shall therefore adopt the proposed spline estimate. Interested readers are referred to Cai and Yuan (2010) for more detailed discussions on optimal estimation of the mean function.

Remark 1 Other loss function may also be used in place of ℓ_n to measure the goodness of the fit to the data in (13). One particular example occurs when the random function X follows a Gaussian process and the measurement error ϵ follows a centered normal distribution. In this case, it is not

hard to see that the random vector $Y_{i\cdot}$ follows a multivariate normal distribution. It is then natural to use the negative log-likelihood instead of ℓ_n in (13). Here we opt for ℓ_n because of its general applicability.

3 Computation

We now turn to the implementation of the procedure as well as computation of the functional principal components. The estimator $\hat{C}_\lambda$ defined in the previous section is particularly easy to calculate. Although the minimization in (13) is taken over infinite dimensional spaces, the solutions can actually be found in finite dimensional spaces thanks to the so-called representer lemma.

Lemma 2 *There exist $m \times m$ symmetric matrices $A_1, \ldots, A_n$ whose diagonal entries are zeros such that*

$$\hat{C}_\lambda(s,t) = \sum_{i=1}^{n} \left\{ [K(s, T_{ij})]_{1\le j\le m} A_i [K(t, T_{ij})]'_{1\le j\le m} \right\}. \tag{16}$$

Lemma 2 can be proved by a similar argument as that of Theorem 1.3.1 in Wahba (1990) and we omit the proof here for brevity. Write

$$G_{i_1 i_2} = [K(T_{i_1 j_1}, T_{i_2 j_2})]_{1\le j_1, j_2\le m}, \qquad 1 \le i_1, i_2 \le n. \tag{17}$$

Observe that for any function $\hat{C}_\lambda$ that admits representation (16), its squared RKHS norm $\|\hat{C}_\lambda\|^2_{\mathcal{H}(K\otimes K)}$ can be written as

$$\|\hat{C}_\lambda\|^2_{\mathcal{H}(K\otimes K)} = \sum_{i_1, i_2=1}^{n} \mathrm{trace}(G_{i_1 i_2} A_{i_2} G_{i_2 i_1} A_{i_1}), \tag{18}$$

which is a convex quadratic function of the entries of A_i's. Because ℓ_n is also convex and quadratic, computing $\hat{C}_\lambda$, or equivalently solving for the A_i's becomes a convex quadratic programming problem and can be solved fairly easily. The readers are also referred to Wahba (1990) and Gu (2002) for further discussions on algorithms and strategies to solve this type of problems efficiently.

Although our main focus is on the covariance function estimation, representation (16) also gives rise to easily computable functional principal component estimates which can be useful in many applications. By Mercer's theorem, C_0 admits the following spectral decomposition:

$$C_0(s,t) = \sum_{k\ge 1} \theta_k \psi_k(s) \psi_k(t) \tag{19}$$

where $\{\psi_k : k \ge 1\}$ is a set of orthonormal basis on L_2 and $\theta_1 \ge \theta_2 \ge \ldots$ are the associated eigenvalues. Note that $\{(\theta_k, \psi_k) : k \ge 1\}$ generally differs from the eigen system $\{(\rho_k, \varphi_k) : k \ge 1\}$ of the reproducing kernel K. Estimating the functional principal components is one of the most

fundamental tasks in functional data analysis and has attracted a lot of attention in the literature (see, e.g., Ramsay and Silverman, 2005). In particular, ψ_ks are commonly estimated by the eigenfunctions of a covariance function estimation. Computing the eigenfunctions of a symmetric bivariate function is generally non-trivial. Typically this is done by discretizing the covariance function estimation and approximate its eigenfunctions by the respective eigenvectors (see, e.g., Rice and Silverman, 1991; Capra and Müller, 1997). One therefore trades the computational complexity with approximation accuracy. Fortunately in our case, the eigenfunctions of $\hat{C}_\lambda$ can actually be computed explicitly without resorting to such numerical techniques thanks to the representation (16).

Let

$$A = \begin{bmatrix} A_1 & 0 & 0 & \dots & 0 \\ 0 & A_2 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & A_n \end{bmatrix}, \tag{20}$$

and

$$Q = \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} & \dots & Q_{1n} \\ Q_{21} & Q_{22} & Q_{23} & \dots & Q_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ Q_{n1} & Q_{n2} & Q_{n3} & \dots & Q_{nn} \end{bmatrix} \tag{21}$$

where

$$Q_{i_1 i_2} = \left[\int_{\mathcal{T}} K(s, T_{i_1 j_1}) K(s, T_{i_2 j_2}) ds \right]_{1 \le j_1, j_2 \le m}, \qquad 1 \le i_1, i_2 \le n. \tag{22}$$

Lemma 3 *The eigenfunctions of $\hat{C}_\lambda$ defined by (13) can be expressed as*

$$\hat{\psi}_k(\cdot) = U_k' \mathbf{g}(\cdot), \qquad k = 1, \dots, n, \tag{23}$$

where U_k is the k-th column of $U = Q^{-1/2} V$ and V is the eigenvectors of $Q^{1/2} A Q^{1/2}$, and

$$\mathbf{g}(\cdot) = (K(\cdot, T_{11}), \dots, K(\cdot, T_{1m}), K(\cdot, T_{21}), \dots, K(\cdot, T_{nm}))'. \tag{24}$$

4 Numerical Experiments

This section considers the finite sample performance of our estimator. We shall first investigate the numerical performance of the estimator through simulation studies and then apply our method to the analysis of a longitudinal CD4 dataset. The theoretical properties of the estimator will be investigated in Section 5.

4.1 Simulation studies

As mentioned in Section 3, the estimator $\hat{C}_\lambda$ of the covariance function C_0 can be calculated efficiently by convex quadratic programming. To demonstrate the merits of the proposed estimator in finite sample settings, we carried out a set of simulation studies with different combinations of sampling frequency, smoothness, and sample size. A collection of random functions X_is were generated independently as follows:

$$X(t) = \sum_{k=1}^{50} \zeta_k Z_k \cos(k\pi t), \qquad t \in [0,1], \tag{25}$$

where Z_ks were independently sampled from the uniform distribution on $[-\sqrt{3}, \sqrt{3}]$ and $\zeta_k = (-1)^{k+1} k^{-\alpha}$. It is not hard to see that the covariance function of X is

$$C_0(s,t) = \sum_{k=1}^{50} k^{-2\alpha} \cos(k\pi s) \cos(k\pi t). \tag{26}$$

Fifty random functions were simulated from this model with $\alpha = 2$. From each function, m random locations were uniformly generated from $[0,1]$ and noisy observations of the function is obtained following model (2). Two values of m are considered, $m = 5$ and $m = 10$. The error standard deviation σ_0 is set to be 0.368 to yield an average signal to noise ratio of $2:1$. We apply the proposed method to the simulated datasets to obtain covariance function estimates as well as estimates of the functional principal components. As is common in most smoothing methods, the choice of the tuning parameter λ plays an important role in the determining the performance of $\hat{C}_\lambda$. Data-driven optimal choice of the tuning parameter is generally difficult. Here we apply the commonly used practical strategy of empirically choosing the value of λ through five fold cross validation.

Figures 1 and 2 provide a visual inspection of a typically simulated data along with the proposed covariance function estimation as well as the functional principal components. The fifty curves were given in the top left panel of Figure 1. The top right panel of Figure 1 shows the observed data for $m = 5$ where observations from the same curve are connected together. Also given in the lower panels of Figure 1 are the first two estimated functional principal components based on $m = 5$ or $m = 10$ observations on each curve together with the those obtained from the sample covariance function as well as the truth.

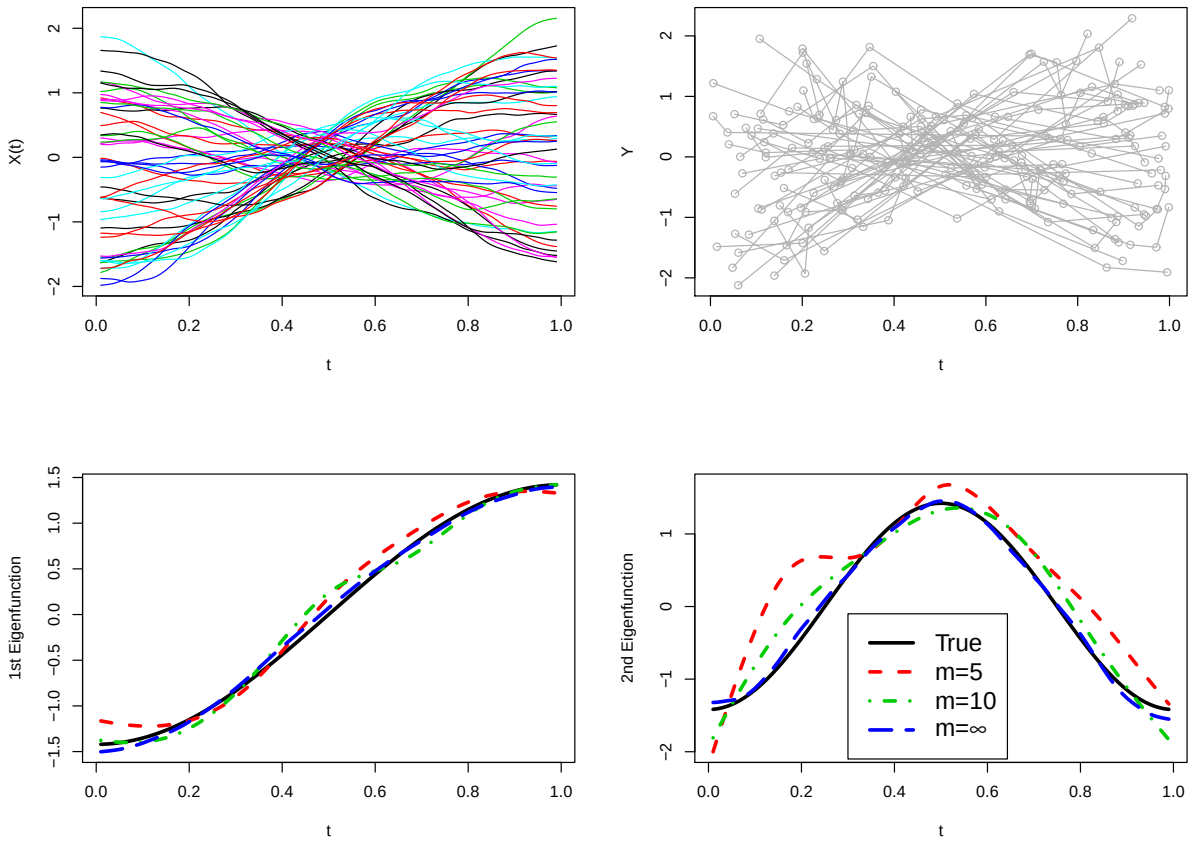

Figure 1: A typical simulated dataset: The top left panel shows 50 simulated functions. Noisy observations at $m = 5$ random locations on each curve are given in the top right panel. The bottom panels give the first two estimated functional principal components. The solid black lines correspond to the truth. The red dashed and green dot-dashed represent those estimated with $m = 5$ and 10 observations on each curve respectively. The blue long dashed lines are estimates obtained from the sample covariance function, which is only computable when each curve is observed completely without noise.

Covariance function estimates obtained with $m = 5$ and 10 respectively are plotted along with C_0 in Figure 2. For comparison, we also included the sample covariance function based on observing each curve completely and without noise. In a certain sense, it reflects how well an estimate can perform when $m = \infty$ observations on each curve. It is evident that our estimate captures the main characteristics of C_0 with as few as five observations on each curve. When the sampling frequency increases, the quality of the estimate improves. The difference between our estimate with $m = 10$ and the sample covariance function or C_0 is negligible, which is also supported by our theoretical development given in Section 5.

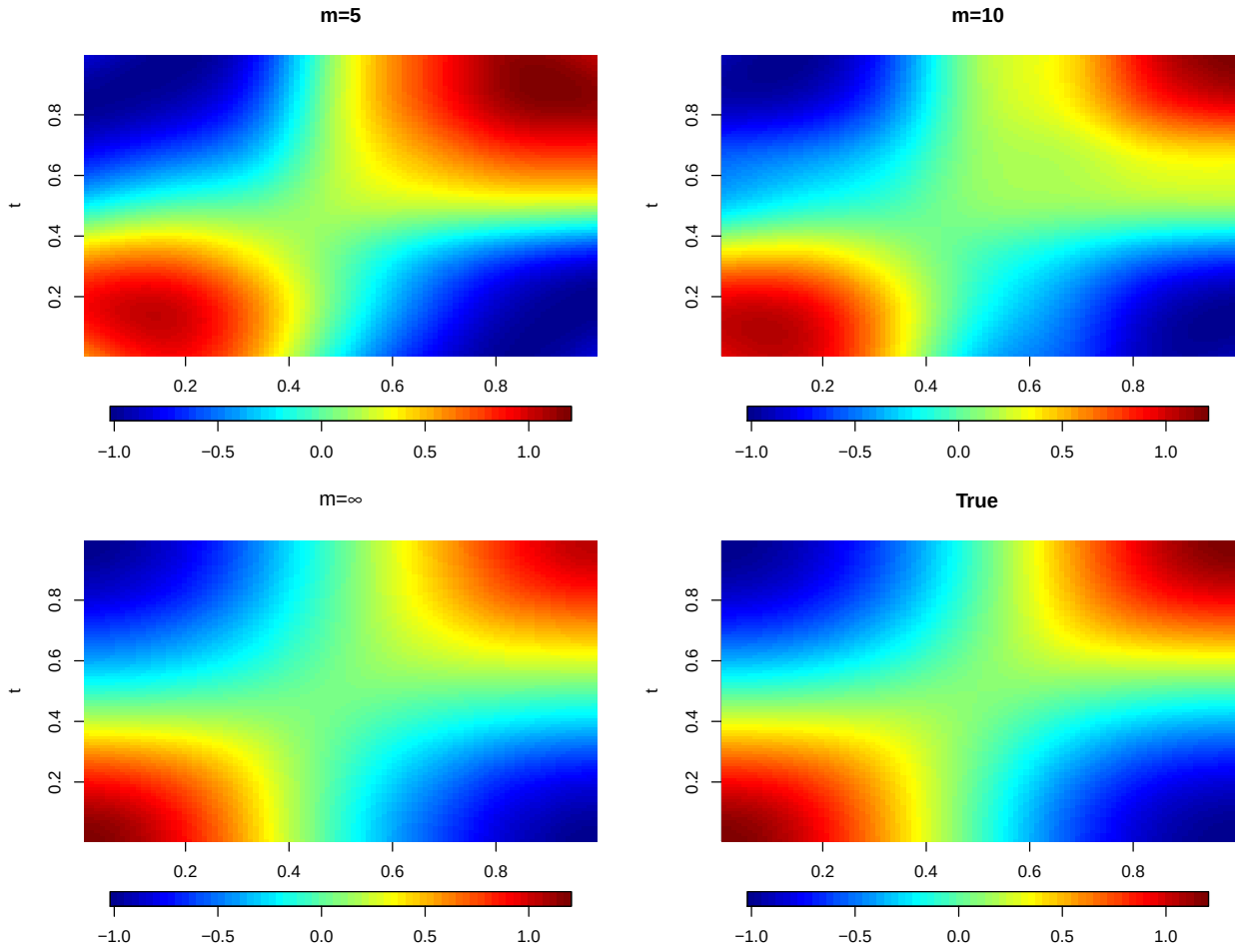

Figure 2: Covariance function estimation with different sampling frequency: Fifty Xs were first simulated. Noisy observations are obtained at $m = 5$ or 10 random locations. The corresponding covariance function estimate is given for each sampling frequency. Also given is the true covariance function and the sample covariance function computed with each function observed completely.

Clearly the value of α in our simulation model determines the smoothness of the simulated curves and subsequently the difficulty of estimation. For comparison purpose, we now consider $\alpha = 1$. Figure 3 shows a typical set of fifty simulated curves as well as covariance function estimates with different sampling frequencies. It is obvious from the figure that the true covariance function is not as smooth as the case when $\alpha = 2$. As a result, higher sampling frequency is required to yield estimates of similar qualities.

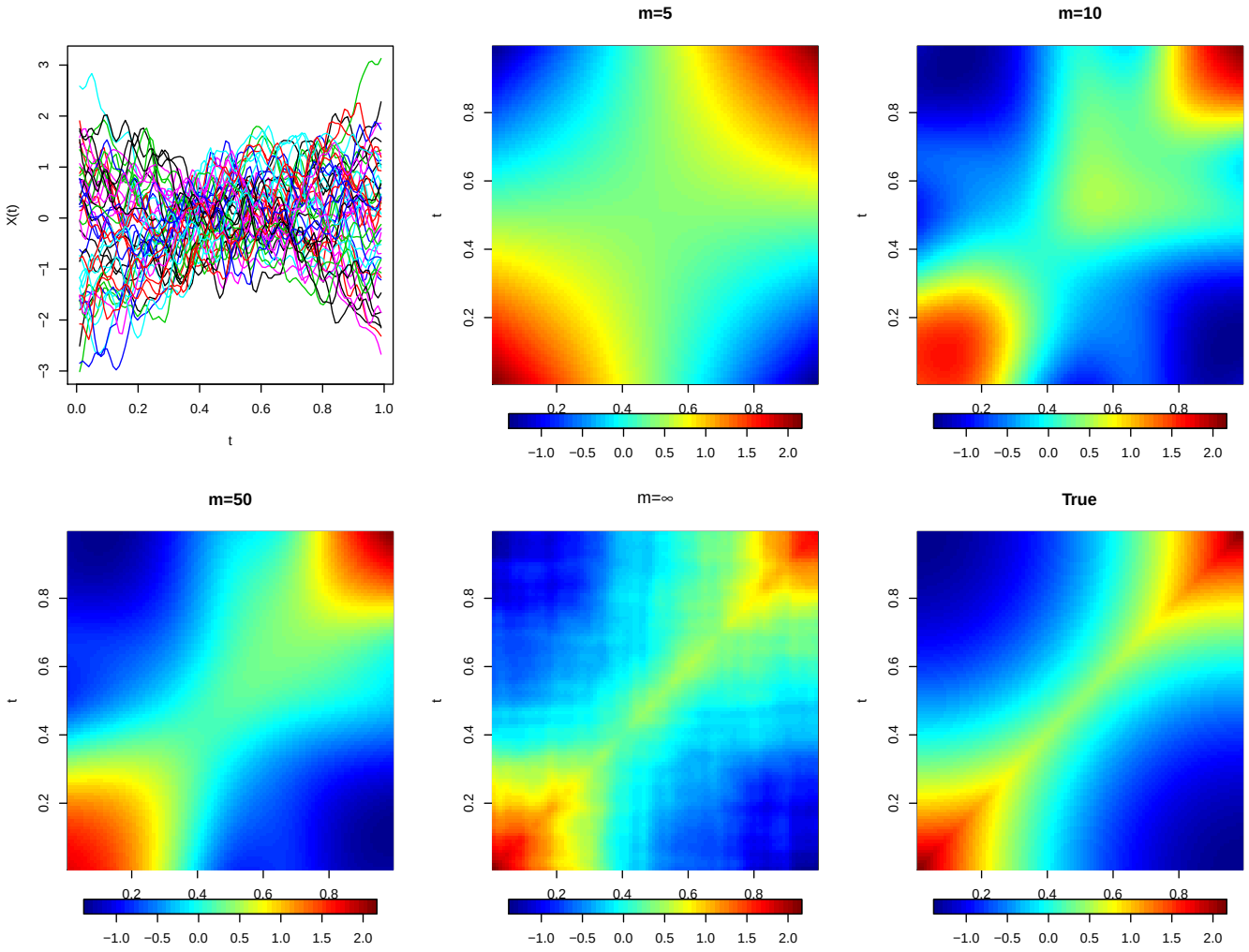

Figure 3: A typical simulated dataset: The top left panel shows the fifty simulated functions, followed by covariance function estimates obtained with $m = 5, 10$ and 50 randomly sampled observations from each curve. The bottom middle panel gives the sample covariance function based on the fifty curves. The bottom right panel is the true covariance function.

To obtain further insight, we repeat the experiment for 200 times with both $n = 50$ and $n = 100$ curves. To fix ideas, we focus on the case when $\alpha = 2$, which corresponds to the situation when the sample path of X is twice differentiable. The estimation error measured by integrated squared error for $m = 5$, 10 or ∞ are reported in Table 1. It is evident from Table 1 that the proposed method performs very well. It can also be observed that the performance improves as the sampling frequency m or sample size n increases.

	$m = 5$		$m = 10$		$m = \infty$	
$n = 50$	0.0229	(0.0011)	0.0142	(0.0005)	0.0046	(0.0004)
$n = 100$	0.0142	(0.0005)	0.0073	(0.0003)	0.0025	(0.0002)

Table 1: Average integrated squared error $\|\hat{C} - C_0\|^2_{L_2}$: averaged over two hundred runs, the numbers in parentheses are the standard errors.

4.2 Longitudinal CD4 count data analysis

To further illustrate the usefulness of the proposed covariance function estimator, we now apply it to a longitudinal CD4 count dataset. The data, reported by Kaslow et al. (1987), recorded CD4+ cell counts for a total of 369 infected men enrolled in the Multicenter AIDS Cohort Study. The human immune deficiency virus (HIV) causes AIDS by attacking CD4+ cells and reducing an individual's ability to fight infections. A healthy person has around 1100 CD4+ cells per milliliter of blood. CD4+ cells decrease in number from infection. Therefore the cell count constitutes a critical assessment of the health of the immune system and progression of the disease. In this particular study, the patients were scheduled to have their cell counts measured twice a year. Because of missing appointments among other factors, the actual times of measurement is random and relatively sparse. The number of observations per subject ranges from 1 to 12 yielding a total of 2376 records. This dataset is a classical example of longitudinal data analysis and further details can be found in Diggle et al. (2002).

One of the main objectives of the analysis of the data is to characterize the time course of the cell counts. This can be accomplished by examining the covariance function. We applied the proposed method to this dataset. The covariance function estimate along with the first three functional principal components are given in the left panels of Figure 4. From a statistical modeling perspective, it is of particular interest to check if the time course can be modeled by a stationary process. From the covariance function estimate, however, this does not appear to be the case.

Along with the CD4+ counts, several other variables are also recorded in this dataset. In

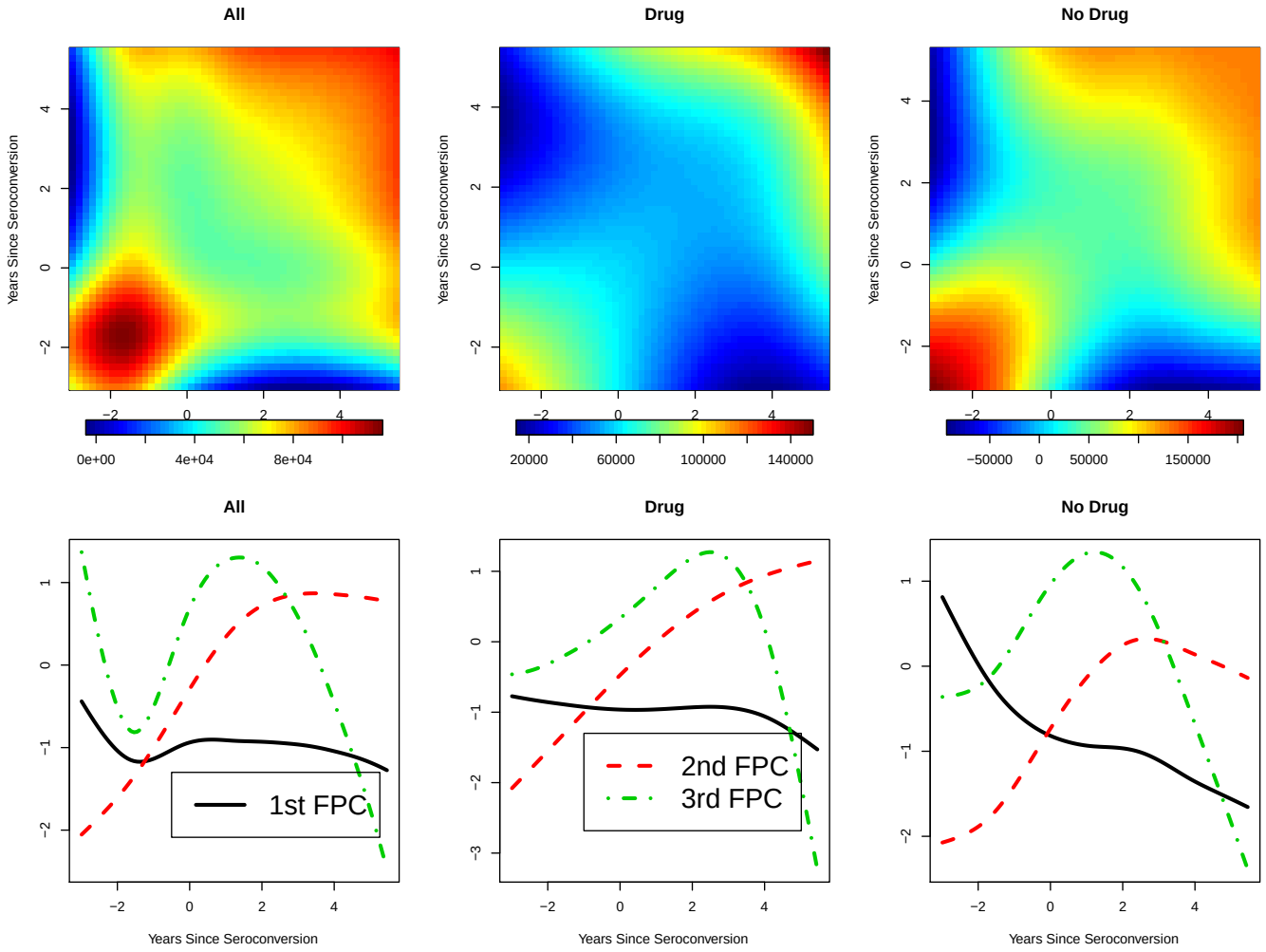

Figure 4: Covariance function and functional principal component estimation for the CD4 data. The left panels correspond to the estimates obtained with the full data including a total of 369 men. The middle panels are estimates obtained with only record from those who used recreational drug throughout the study. The right panels are estimates obtained with only record from those who did not use recreational drug at all throughout the study. Note that the top panels are in different color scales in order to better demonstrate how the function value changes within each panel.

particular, the participants were asked whether or not they were using recreational drug between visits. Among the 369 men, 209 used recreational drug throughout the study, 36 stayed away from it altogether. Of interest here is whether or not the two more homogeneous sub-populations of those who used recreational drug on a regular basis or those who did not use it at all may display different behaviors and therefore require separate modeling. To this end, we estimate the covariance function as well as the functional principal components for both subsets and the results are given in the middle and left panels of Figure 4 for comparison purpose. We note that the covariance function estimates for the full data and the subsets are plotted with different color scale for better visualization within each plot. It is interesting to note that many of the main characteristics of the covariance function are shared in all three estimates. The estimate obtained from the full data, however, displays strong covariation in the early phase of the study, which is absent from the estimate based on only those who use drug regularly. Such discrepancy can be further witnessed from the first three principal components. Although they are quite similar for time greater than 0, significant and consistent differences can be observed at earlier times.

5 Rates of Convergence

In this section we investigate the theoretical properties of the covariance function estimator $\hat{C}_\lambda$ and establish the rate of convergence. Let $\mathcal{P}(\alpha; M_0, c_0)$ be the collection of probability measures of X such that

(a) the sample path of X belongs to $\mathcal{H}(K)$ almost surely and $\mathbb{E}\|X\|^2_{\mathcal{H}(K)} < M_0$;

(b) K is a Mercer kernel with eigenvalues satisfying $\rho_k \sim k^{-2\alpha}$;

(c) there exists a numerical constant $c_0 > 0$ such that $\mathbb{E}X^4(t) \leq c_0[\mathbb{E}X^2(t)]^2$ for any $t \in \mathcal{T}$, and

$$\mathbb{E}\left(\int_{\mathcal{T}} X(t)f(t)dt\right)^4 \leq c_0\left(\mathbb{E}\left(\int_{\mathcal{T}} X(t)f(t)dt\right)^2\right)^2, \tag{27}$$

for any $f \in L_2(\mathcal{T})$.

The first two conditions essential specify the smoothness of X, which is more specifically related to the decay rate of ρ_k. For example, when Sobolev space $\mathcal{W}^2_\alpha([0,1])$ is considered, it is well known that $\rho_k \sim k^{-2\alpha}$. The last condition concerns the fourth moment of X and is satisfied with $c_0 = 3$ when X follows a Gaussian process.

The next theorem establishes the rate of convergence of $\hat{C}_\lambda$ in terms of integrated squared error.

Theorem 4 *Assume that $\mathbb{E}\varepsilon^4 < \infty$, T_{ij} are independent and identically distributed with a density bounded away from zero on $\mathcal{T}$ and the mean function estimate $\hat{\mu}$ satisfies*

$$\|\hat{\mu} - \mu_0\|_{L_2}^2 = O_p\left(\left(\frac{\log n}{mn}\right)^{\frac{2\alpha}{2\alpha+1}} + \frac{1}{n}\right). \tag{28}$$

Then

$$\lim_{D\to\infty} \limsup_{n\to\infty} \sup_{\mathcal{L}(X)\in\mathcal{P}(\alpha;M_0,c_0)} P\left(\|\hat{C}_\lambda - C_0\|_{L_2}^2 > D\left(\left(\frac{\log n}{mn}\right)^{\frac{2\alpha}{2\alpha+1}} + \frac{1}{n}\right)\right) = 0, \tag{29}$$

if $\lambda \asymp (\frac{\log n}{mn})^{\frac{2\alpha}{2\alpha+1}}$.

The condition on the mean function estimation (28) is satisfied by the spline estimate suggested by Cai and Yuan (2010). In particular, it was shown that the mean function can be estimated at the rate of $O_p\left((mn)^{-\frac{2\alpha}{2\alpha+1}} + n^{-1}\right)$ in terms of integrated squared error. It is quite surprising to note that even though the covariance function is of higher dimension than the mean function, the two rates at most differ by a factor of $(\log n)^{\frac{2\alpha}{2\alpha+1}}$. This is in stark contrast with the common wisdom. Consider, for example, the case when the sample path of X belongs to Sobolev space $\mathcal{W}_2^2([0,1])$, i.e., twice differentiable functions on $[0,1]$. It is natural to think of C_0 as a member of the second order Sobolev space on the unit square $\mathcal{W}_2^2([0,1]^2)$, i.e., twice differentiable functions on $[0,1]^2$. As shown by Stone (1982), the optimal rate for estimating functions from $\mathcal{W}_2^2([0,1])$ or $\mathcal{W}_2^2([0,1]^2)$ is $M^{-4/5}$ and $M^{-2/3}$ respectively if independent observations are available at M different locations. For estimating the covariance function, a total of nm^2 observations $\{(Y_{ij_1} - \mu_0(T_{ij_1}))(Y_{ij_2} - \mu_0(T_{ij_2})) : 1 \le i \le n, 1 \le j_1 \ne j_2 \le m\}$ are obtained. But there are significant redundancy among these observations as we are observing nm Y_{ij}s after all. Also the data are correlated due to the functional nature of X_is.

To appreciate the importance of the association between C_0 and the tensor product space, we first consider the case when m is finite. The number of observations in this case is of the order $O(n)$ for estimating both the mean and covariance functions. Cai and Yuan (2010) showed that the mean function can be estimated at the optimal rate of $n^{-4/5}$. Similarly, the rate of $n^{-2/3}$ can be achieved, in particular, by the local polynomial estimate of Hall, Müller and Wang (2006) using the fact that C_0 is twice differentiable. But as shown by Theorem 1, the space from which C_0 comes is the tensor product space $\mathcal{W}_2^2([0,1]) \otimes \mathcal{W}_2^2([0,1])$, which is much smaller than $\mathcal{W}_2^2([0,1]^2)$. As a consequence, the rate attained by our method, $(n/\log n)^{-4/5}$, is much better than the best possible rate of $n^{-2/3}$ for estimating functions from second order Sobolev space on $[0,1]^2$.

It is also interesting to compare our results with those obtained by Paul and Peng (2009) who consider a particular setting where C_0 can be approximated by a function with a fixed number of

non-vanishing eigenvalues and eigenfunctions from $\mathcal{W}_2^4$. Under various regularity conditions, they show that when m is bounded, the restricted MLE that based on this particular structure can achieve the convergence rate of $O_p((n/\log n)^{-8/9})$. See, e.g., Theorem 2.1 and Corollary 2.1 of Paul and Peng (2009). The rate matches that from Theorem 4 because in this case it is well known that $\alpha = 4$. Despite the similarity, however, we note that the two approaches and therefore their implications are very different. The particular covariance function model studied by Paul and Peng (2009) is very specialized. The restricted MLE discussed in their paper uses the knowledge of this particular model whereas our method does not require such information and achieves the same rate for a much more general class of covariance functions. Moreover, we do not require many of the technical conditions imposed by Paul and Peng (2009).

The situation when m is not finite is more complex because it is then necessary to address how much data redundancy and correlation may affect our ability to estimate the covariance function. Theorem 4 reveals an curious phase transition behavior of the convergence rate. When the functions are densely sampled, i.e., $m \gg n^{\frac{1}{2\alpha}} \log n$, the sampling frequency does not matter and the rate is determined solely by the number of curves:

$$\|\hat{C} - C_0\|_{L_2}^2 = O_p\left(n^{-1}\right). \tag{30}$$

This suggests that when sampled frequently enough, we can estimate the covariance function as well as if the entire functions are observable. But when the functions are sparsely sampled, i.e., $m \ll n^{\frac{1}{2\alpha}} \log n$, the rate is jointly determined by the number of curves and the number of observations on each curve through the total number of observations $N := nm$:

$$\|\hat{C} - C_0\|_{L_2}^2 = O_p\left((N/\log N)^{-\frac{2\alpha}{2\alpha+1}}\right). \tag{31}$$

Little is known about how other methods may behave when m is not finite. One exception is Hall, Müller and Wang (2006) who showed that, when assuming that the sample path of X is twice differentiable, the two step procedure where one estimates each curve by smoothing and then estimates C_0 by the sample covariance function of those smoothed curves achieves the convergence rate of $1/n$ if $m \gg n^{1/4+\delta}$ for some $\delta > 0$. Our result is comparable and slightly better in that we only require $m \gg n^{\frac{1}{2\alpha}} \log n$ to achieve $1/n$ convergence rate for general $\alpha > 1/2$.

Although not our main focus, we note that convergence rates of the estimated functional principal components can be easily derived from Theorem 4. To this end, we consider the theoretical properties of $\hat{\psi}_k$ for $k \leq K_0$ where K_0 is fixed. Without loss of generality, we assume that $\langle \psi_k, \hat{\psi}_k \rangle \geq 0$. Then we have the following result.

Corollary 5 *Under the conditions of Theorem 4, if θ_k is of multiplicity one, then*

$$\lim_{D\to\infty}\limsup_{n\to\infty}\sup_{\mathcal{L}(X)\in\mathcal{P}(\alpha;M_0,c_0)} P\left(\|\hat{\psi}_k-\psi_k\|_{L_2}^2 > D\left((mn/\log n)^{-\frac{2\alpha}{2\alpha+1}}+n^{-1}\right)\right)=0, \qquad (32)$$

if $\lambda\sim(mn/\log n)^{-\frac{2\alpha}{2\alpha+1}}$.

The results for estimating the functional principal components again improves the one given in Paul and Peng (2009) who obtained the rate of $O_p((n/\log n)^{-8/9})$ in the case of $\alpha=4$ under a much more restrictive setting. Theoretical analysis of functional principal component analysis can also be found in Yao, Müller and Wang (2005) and Hall, Müller and Wang (2006) among others. In particular, Hall, Müller and Wang (2006) considered the case when m is bounded and showed that, when assuming that the sample path of X has bounded second derivative, the optimal rate for estimating the functional principal components is $O_p(n^{-2\alpha/(2\alpha+1)})$ and can be achieved by local polynomial regression procedures. They also show that when $m\gg n^{1/4+\delta}$ for some $\delta>0$, the optimal rate is $O_p(1/n)$ and can be achieved by the sample covariance function of pre-smoothed X_is. Since the estimating strategy differs between the two situations, it is not clear how to deal with the intermediate cases. In contrast, the functional principal components computed from our covariance function estimation in a vanilla fashion are applicable to all cases. Under much weaker conditions, it achieves the optimal rate of $O_p(1/n)$ when $m\gg n^{1/4}\log n$. The threshold is slightly better than that ($n^{1/4+\delta}$) identified by Hall, Müller and Wang (2006). Furthermore, it attains a rate within a factor $(\log n)^{4/5}$ of the optimal when m is bounded.

Finally, we note that the rate achieved by the estimator $\hat{C}_\lambda$ can not be much improved. As stated in the next theorem, even if C_0 is known a priori to be of rank one, i.e., X has a single principal component, the best rate attainable is $O((mn)^{-\frac{2\alpha}{2\alpha+1}}+n^{-1})$, which differs from that of $\hat{C}$ by at most a factor of $(\log n)^{2\alpha/(2\alpha+1)}$. To this end, let $\mathcal{P}'(\alpha;M_0)$ be the collection of probability measures of X such that $X(\cdot)=x\psi(\cdot)$ where x is a centered random variable with bounded second moment and $\|\psi\|_{\mathcal{H}(K)}^2\le M_0$. It is clear that $\mathcal{P}'(\alpha;M_0)\subset\mathcal{P}(\alpha;M_0,c_0)$.

Theorem 6 *There exists a constant $d>0$ depending only on M_0 and σ_0^2 such that for any estimate $\tilde{C}$ base on observations $\{(T_{ij},Y_{ij}):1\le i\le n,1\le j\le m\}$,*

$$\limsup_{n\to\infty}\sup_{\mathcal{L}(X)\in\mathcal{P}'(\alpha;M_0)} P\left(\|\tilde{C}-C_0\|_{L_2}^2 > d\left((nm)^{-\frac{2\alpha}{2\alpha+1}}+n^{-1}\right)\right)>0. \qquad (33)$$

6 Discussions

We have developed a regularization method for covariance function estimation based on a random sample of discrete observations with measurement errors. Despite its similarity to the standard

bivariate smoothing problem, we show that covariance function estimation can be fundamentally different from the usual Sobolev space based bivariate smoothing due to the fact that it lies in a particular type of tensor product space. It is shown that the proposed method enjoy superior theoretical properties to some of the known results in the literature both for estimating the covariance functions and for estimating the functional principal components. The procedure is also easily implementable.

In the present paper, for ease of presentation we have assumed that there are equal number of sampling points on each curve to better demonstrate the joint effects of the number of curves n and the sampling frequency m on estimating the covariance function. It should be noted that this assumption is not essential and can be easily relaxed. In a more general setting, one may have different number of sampling points on different curves. Let m_i be the number of sampling points on the ith curve. Following the same argument, it can be shown that the proposed method enjoys the same properties if the numbers of sampling points on individual curves are of the same order of magnitude. That is, there exist constants $c_1 \geq c_2 > 0$ such that

$$c_2 m \leq \min_{1 \leq i \leq n} m_i \leq \max_{1 \leq i \leq n} m_i \leq c_1 m.$$

More generally, the number of sampling points m_i may be itself random. In such a case, our results continue to hold for $m := \mathbb{E}(m_i)$ if there exists a constant $0 < \sigma_*^2 < \infty$ such that $var(m_i) \leq \sigma_*^2$.

We also note that although we have focused on random curves defined on a compact subset $\mathcal{T}$ of the real line, the proposed method can be readily applied to functional data defined on more general compact domains and the convergence rates established in Section 5 continue to hold with minor modifications. Such extension could be useful, for instance, when modeling images. For example, if the sample path of X belongs to the α-order Sobolev space on $[0,1]^d$, the rate of convergence for the proposed estimator given in Theorem 4 is changed to $(nm/\log n)^{2\alpha/(2\alpha+d)} + n^{-1}$.

The study of the covariance function estimation given in this paper is expected to have implications in a number of related problems such as functional linear regression, classification or clustering where the estimate of the covariance kernel plays a prominent role. We leave these topics for future research.

7 Proofs

We prove the main results in this section. Throughout this section we use C to denote a covariance function and the lower case c stands for a positive constant which may vary from place to place. Two technical lemmas used in the proofs of the main results are proved in the Appendix.

7.1 Proof of Theorem 1

Note that $C_0(s,t) = g_0(s,t) - \mu_0(s)\mu_0(t)$ where $g_0(s,t) = \mathbb{E}X(s)X(t)$. By Jensen's inequality,

$$\|\mu_0\|^2_{\mathcal{H}(K)} = \|\mathbb{E}X\|^2_{\mathcal{H}(K)} \leq \mathbb{E}\|X\|^2_{\mathcal{H}(K)} < \infty, \tag{34}$$

which implies that $\mu_0 \in \mathcal{H}(K)$. Therefore, $\mu_0 \otimes \mu_0 \in \mathcal{H}(K \otimes K)$. It remains to show that $g_0 \in \mathcal{H}(K \otimes K)$.

It suffices to verify that

$$\|g_0\|^2_{\mathcal{H}(K\otimes K)} = \sum_{j,k\geq 1} \rho_j^{-1}\rho_k^{-1}\mathbb{E}(x_j x_k)^2 < \infty. \tag{35}$$

Observe that $\mathbb{E}(x_j x_k)^2 \leq \mathbb{E}(x_j^2)\mathbb{E}(x_k^2)$. Therefore,

$$\begin{aligned}
\|g_0\|^2_{\mathcal{H}(K\otimes K)} &\leq \sum_{j,k\geq 1} \rho_j^{-1}\rho_k^{-1}\mathbb{E}(x_j^2)\mathbb{E}(x_k^2) \\
&= \left(\sum_{k\geq 1} \rho_k^{-1}\mathbb{E}(x_k^2)\right)^2 \\
&= \left(\mathbb{E}\|X\|^2_{\mathcal{H}(K)}\right)^2 < \infty.
\end{aligned}$$

This completes the proof of Theorem 1. ∎

7.2 Proof of Lemma 3

We first show that $\hat{\psi}_k$s are orthonormal. Observe that

$$\int_{\mathcal{T}} \mathbf{g}(s)\mathbf{g}'(s)ds = Q. \tag{36}$$

Therefore,

$$\int_{\mathcal{T}} \hat{\psi}_{k_1}(s)\hat{\psi}_{k_2}(s)ds = U'_{k_1}QU_{k_2} = V'_{k_1}Q^{-1/2}QQ^{-1/2}V_{k_2} = \delta_{k_1k_2}, \tag{37}$$

where δ is the Kronecker's delta.

Denote by $\hat{\theta}_k$ the eigenvalues of $Q^{1/2}AQ^{1/2}$, then

$$\begin{aligned}
\sum_k \hat{\theta}_k\hat{\psi}_k(s)\hat{\psi}_k(t) &= \mathbf{g}'(s)\left(\sum_k \hat{\theta}_k U_k U'_k\right)\mathbf{g}(t) \\
&= \mathbf{g}'(s)\left(Q^{-1/2}V\,\mathrm{diag}(\hat{\theta}_1,\ldots)V'Q^{-1/2}\right)\mathbf{g}(t) \\
&= \mathbf{g}'(s)A\mathbf{g}(t) \\
&= \hat{C}_\lambda(s,t),
\end{aligned}$$

which completes the proof. ∎

7.3 Proof of Theorem 4

For brevity, we shall consider here only the case when $\mu_0(\cdot)$ is known and (14) is employed in defining $\hat{C}_\lambda$. The effect of this assumption in establishing the convergence rate of $\hat{C}_\lambda$ is negligible because of the condition (28), which essentially controls the error term that may come with using an estimated mean function instead of the true mean function. The proof follows from the same line of argument when using $\hat{\mu}$ but gets considerably more tedious.

For technical purpose, we first introduce a class of norms on $\mathcal{H}(K \otimes K)$. To this end, recall that any $f \in \mathcal{H}(K \otimes K)$ can be written as

$$f(s,t) = \sum_{j,k \geq 1} f_{jk} \varphi_j(s) \varphi_k(t) \tag{38}$$

where $f_{jk} = \langle f, \varphi_j(\cdot)\varphi_k(\cdot) \rangle_{L_2}$. By the construction of $\{\varphi_k : k \geq 1\}$, we have

$$\|f\|_{L_2}^2 = \sum_{j,k \geq 1} f_{jk}^2, \qquad \text{and} \qquad \|f\|_{\mathcal{H}(K \times K)}^2 = \sum_{j,k \geq 1} \gamma_{jk}^{-1} f_{jk}^2 \tag{39}$$

where $\gamma_{jk} = \rho_j \rho_k$. Define, for $0 \leq a \leq 1$,

$$\|f\|_a^2 = \sum_{j,k \geq 1} \gamma_{jk}^{-a} f_{jk}^2. \tag{40}$$

It is clear that $\|\cdot\|_0 = \|\cdot\|_{L_2}$ and $\|\cdot\|_1 = \|\cdot\|_{\mathcal{H}(K \otimes K)}$.

For brevity, we shall assume that $\mu_0(\cdot) = 0$ and T follows a uniform distribution without loss of generality. Recall that

$$\hat{C}_\lambda = \underset{C \in \mathcal{H}(K \otimes K)}{\operatorname{argmin}} \left\{ \ell_{mn}(C) + \lambda \|C\|_{\mathcal{H}(K \otimes K)}^2 \right\} \tag{41}$$

where

$$\ell_{mn}(C) = \frac{1}{nm(m-1)} \sum_{i=1}^{n} \sum_{1 \leq j \neq k \leq m} \left(Y_{ij} Y_{ik} - C(T_{ij}, T_{ik}) \right)^2. \tag{42}$$

Observe that

$$\begin{aligned} \ell_\infty(C): &= \mathbb{E} \ell_{mn}(C) \\ &= \frac{1}{m(m-1)} \mathbb{E} \left(\sum_{1 \leq j \neq k \leq m} \left[Y_{1j} Y_{1k} - C(T_{1j}, T_{1k}) \right]^2 \right) \\ &= \operatorname{Var}\left(Y_{11} Y_{12} \right) + \mathbb{E} \left(\left[C(T_{11}, T_{12}) - C_0(T_{11}, T_{12}) \right]^2 \right). \end{aligned}$$

Write

$$\bar{C}_\lambda = \underset{C \in \mathcal{H}(K \otimes K)}{\operatorname{argmin}} \left\{ \ell_\infty(C) + \lambda \|C\|_{\mathcal{H}(K \otimes K)}^2 \right\}. \tag{43}$$

Then

$$\hat{C}_\lambda - C_0 = (\bar{C}_\lambda - C_0) + (\hat{C}_\lambda - \bar{C}_\lambda). \tag{44}$$

The two terms can be regarded as deterministic and stochastic error respectively. Write

$$\ell_{\infty,\lambda}(C) = \ell_\infty(C) + \lambda\|C\|^2_{\mathcal{H}(K\otimes K)};$$
$$\ell_{mn,\lambda}(C) = \ell_{mn}(C) + \lambda\|C\|^2_{\mathcal{H}(K\otimes K)}.$$

Denote $G_\lambda = D^2\ell_{\infty,\lambda}(\bar{C}_\lambda)$ and

$$\tilde{C}_\lambda = \bar{C}_\lambda - G_\lambda^{-1} D\ell_{mn,\lambda}(\bar{C}_\lambda), \tag{45}$$

where D stands for the Fréchet derivatives. It is clear that the stochastic error can be further decomposed as

$$\hat{C}_\lambda - \bar{C}_\lambda = \left(\hat{C}_\lambda - \tilde{C}_\lambda\right) + \left(\tilde{C}_\lambda - \bar{C}_\lambda\right). \tag{46}$$

We now bound $\bar{C}_\lambda - C_0$, $\hat{C}_\lambda - \tilde{C}_\lambda$ and $\tilde{C}_\lambda - \bar{C}_\lambda$ separately.

We first consider the deterministic error $\bar{C}_\lambda - C_0$.

Lemma 7 *There exists a constant $c > 0$ such that*

$$\left\|\bar{C}_\lambda - C_0\right\|_a^2 \le c\lambda^{1-a}\|C_0\|^2_{\mathcal{H}(K\otimes K)}.$$

Proof. Write

$$C_0(s,t) = \sum_{k_1,k_2=1}^{\infty} a_{k_1k_2}\varphi_{k_1}(s)\varphi_{k_2}(t), \qquad \bar{C}_\lambda(s,t) = \sum_{k_1,k_2=1}^{\infty} \bar{b}_{k_1k_2}\varphi_{k_1}(s)\varphi_{k_2}(t).$$

Then

$$\bar{b}_{k_1,k_2} = (1+\lambda\gamma_{k_1k_2}^{-1})^{-1} a_{k_1k_2}. \tag{47}$$

It follows that

$$\begin{aligned}
\left\|\bar{C}_\lambda - C_0\right\|_a^2 &= \sum_{k_1,k_2=1}^{\infty} (1+\gamma_{k_1k_2}^{-1})^a (\bar{b}_{k_1k_2} - a_{k_1k_2})^2 \\
&= \sum_{k_1,k_2=1}^{\infty} (1+\gamma_{k_1k_2}^{-1})^a \left(\frac{\lambda\gamma_{k_1k_2}^{-1}}{1+\lambda\gamma_{k_1k_2}^{-1}}\right)^2 a_{k_1k_2}^2 \\
&\le c\lambda^2 \sup_{k_1,k_2} \frac{\gamma_{k_1k_2}^{-(1+a)}}{\left(1+\lambda\gamma_{k_1k_2}^{-1}\right)^2} \sum_{k=1}^{\infty} \gamma_{k_1k_2}^{-1} a_{k_1k_2}^2 \\
&\le c\lambda^{1-a}\|C_0\|^2_{\mathcal{H}(K\otimes K)}.
\end{aligned}$$

Hereafter, we use c to denote a generic positive constant which may take different values at each appearance. ∎

Next, we consider $\tilde{C}_\lambda - \bar{C}_\lambda$.

Lemma 8 *There exists a constant $c > 0$ such that*

$$\mathbb{E}\left\|\tilde{C}_\lambda - \bar{C}_\lambda\right\|_a \leq c\left(n^{-1} + n^{-1}m^{-1}\lambda^{-(a+(1/2\alpha))}\log(1/\lambda) + n^{-1}\lambda^{1-(a+(1/2\alpha))}\log(1/\lambda)\right). \tag{48}$$

Proof. By definition

$$D\ell_{mn,\lambda}(\bar{C}_\lambda) + D^2\ell_{\infty,\lambda}(\bar{C}_\lambda)(\tilde{C}_\lambda - \bar{C}_\lambda) = 0. \tag{49}$$

Notice that

$$D\ell_{mn,\lambda}(\bar{C}_\lambda) = D\ell_{mn,\lambda}(\bar{C}_\lambda) - D\ell_{\infty,\lambda}(\bar{C}_\lambda) = D\ell_{mn}(\bar{C}_\lambda) - D\ell_\infty(\bar{C}_\lambda). \tag{50}$$

Therefore

$$\begin{aligned}
\mathbb{E}\left[D\ell_{mn,\lambda}(\bar{C}_\lambda)f\right]^2 &= \mathbb{E}\left[D\ell_{mn}(\bar{C}_\lambda)f - D\ell_\infty(\bar{C}_\lambda)f\right]^2 \\
&= \frac{1}{n^2m^2(m-1)^2}\mathrm{Var}\left[\sum_{i=1}^n \sum_{1\leq j<k\leq m}\left(\left[Y_{ij}Y_{ik} - \bar{C}_\lambda(T_{ij},T_{ik})\right]f(T_{ij},T_{ik})\right)\right] \\
&= \frac{1}{nm^2(m-1)^2}\mathrm{Var}\left[\sum_{1\leq j\neq k\leq m}\left(\left[Y_{1j}Y_{1k} - \bar{C}_\lambda(T_{1j},T_{1k})\right]f(T_{1j},T_{1k})\right)\right] \\
&= \frac{1}{nm^2(m-1)^2}\mathrm{Var}\left[\sum_{1\leq j\neq k\leq m}\left(\left[C_0(T_{1j},T_{1k}) - \bar{C}_\lambda(T_{1j},T_{1k})\right]f(T_{1j},T_{1k})\right)\right] \\
&\quad + \frac{1}{nm^2(m-1)^2}\mathbb{E}\left[\mathrm{Var}\left(\sum_{1\leq j\neq k\leq m} Y_{1j}Y_{1k}f(T_{1j},T_{1k})\middle| T\right)\right]
\end{aligned}$$

We now bound the two terms separately. The second term can be bounded using the following technical lemma which is proved in the Appendix.

Lemma 9 *Let $Y_j = X(T_j) + \epsilon_j$ where X is a mean zero second order stochastic process whose probability law belongs to $\mathcal{P}(\alpha; M_0, c_0)$, T_j are independent uniform random variables defined on $\mathcal{T}$, and ϵ_j are independent measurement error with mean zero and variance σ_0^2. Denote, for $f, g \in \mathcal{L}_2(\mathcal{T})$,*

$$U = \sum_{1\leq j\neq k\leq m} Y_jY_kf(T_j)g(T_k). \tag{51}$$

Then

$$\mathbb{E}\left[\mathrm{Var}(U|T)\right] \leq 4c_0 m^4 \mathbb{E}\left(\int X(s)f(s)ds\right)^2 \mathbb{E}\left(\int X(s)g(s)ds\right)^2 + O(m^3). \tag{52}$$

In particular, if $f = \varphi_{k_1}$ *and* $g = \varphi_{k_2}$*, then*

$$\mathbb{E}\left[\mathrm{Var}(U|T)\right] \leq 4c_0 m^4 a_{k_1k_1} a_{k_2k_2} + O(m^3), \tag{53}$$

where $\{a_{jk} : j, k \geq 1\}$ *is the Fourier coefficients of* C_0 *with respect to orthonormal basis* $\{\varphi_j \otimes \varphi_k : j, k \geq 1\}$.

The second term can now be bounded using Lemma 9. By taking $f = \varphi_{k_1} \otimes \varphi_{k_2}$ in Lemma 9, we have

$$\mathbb{E}\left[\mathrm{Var}\left(\sum_{1\leq j<k\leq m} Y_{1j}Y_{1k}f(T_{1j}, T_{1k})\Big|T\right)\right] \leq cm^4 a_{k_1k_1}a_{k_2k_2} + O(m^3). \tag{54}$$

To bound the first term, denote by

$$V_m = \sum_{1\leq j<k\leq m} \left(\left[C_0(T_{1j}, T_{1k}) - \bar{C}_\lambda(T_{1j}, T_{1k})\right] f(T_{1j}, T_{1k})\right). \tag{55}$$

Observe that

$$\begin{aligned}
\mathrm{Var}(V_m) &\leq \mathbb{E}(V_m^2) \\
&= \sum_{\substack{1\leq j<k\neq m \\ 1\leq j'\neq k'\leq m}} \mathbb{E}\big(\left[C_0(T_{1j}, T_{1k}) - \bar{C}_\lambda(T_{1j}, T_{1k})\right] f(T_{1j}, T_{1k}) \\
&\qquad \times \left[C_0(T_{1j'}, T_{1k'}) - \bar{C}_\lambda(T_{1j'}, T_{1k'})\right] f(T_{1j'}, T_{1k'})\big) \\
&\leq \frac{m!}{(m-4)!}\left(\int_{\mathcal{T}\times\mathcal{T}} \left[C_0(s,t) - \bar{C}_\lambda(s,t)\right] f(s,t)dsdt\right)^2 + cm^3 \\
&\leq m^4 \int_{\mathcal{T}\times\mathcal{T}} \left[C_0(s,t) - \bar{C}_\lambda(s,t)\right]^2 dsdt \int_{\mathcal{T}\times\mathcal{T}} f^2(s,t)dsdt + cm^3 \\
&\leq c\left(m^4\lambda \int_{\mathcal{T}\times\mathcal{T}} f^2(s,t)dsdt + m^3\right).
\end{aligned}$$

Take $f = \varphi_{k_1} \otimes \varphi_{k_2}$, i.e., $f(s,t) = \varphi_{k_1}(s)\varphi_{k_2}(t)$. Then the first term can be bounded by $c_0 n^{-1}(\lambda + m^{-1})$.

In summary, we have

$$\mathbb{E}\left[D\ell_{mn,\lambda}(\bar{C}_\lambda)\varphi_{k_1} \otimes \varphi_{k_2}\right]^2 \leq cn^{-1}\left(a_{k_1k_1}a_{k_2k_2} + \lambda + m^{-1}\right). \tag{56}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left\|\tilde{C}_\lambda - \bar{C}_\lambda\right\|_a^2 &= \mathbb{E}\left\|G_\lambda^{-1} D\ell_{n,\lambda}(\bar{C})\right\|_a^2 \\
&= \mathbb{E}\left[\sum_{k_1,k_2\geq 1}(1+\gamma_{k_1k_2}^{-1})^a(1+\lambda\gamma_{k_1k_2}^{-1})^{-2}\left(D\ell_{n,\lambda}(\bar{b})\varphi_{k_1}\otimes\varphi_{k_2}\right)^2\right] \\
&\leq cn^{-1}\left[\sum_{k_1,k_2\geq 1}(1+\gamma_{k_1k_2}^{-1})^a(1+\lambda\gamma_{k_1k_2}^{-1})^{-2}a_{k_1k_1}a_{k_2k_2}\right] \\
&\quad + cn^{-1}(\lambda+m^{-1})\sum_{k_1,k_2\geq 1}(1+\gamma_{k_1k_2}^{-1})^a(1+\lambda\gamma_{k_1k_2}^{-1})^{-2}
\end{aligned}$$

Observe that

$$\begin{aligned}
&\left[\sum_{k_1,k_2\geq 1}(1+\gamma_{k_1k_2}^{-1})^a(1+\lambda\gamma_{k_1k_2}^{-1})^{-2}a_{k_1k_1}a_{k_2k_2}\right] \\
\leq\ &\sum_{k_1,k_2\geq 1}(1+\gamma_{k_1}^{-1})^a(1+\gamma_{k_2}^{-1})^a a_{k_1k_1}a_{k_2k_2} \\
=\ &\left(\sum_{k_1\geq 1}(1+\gamma_{k_1}^{-1})^a a_{k_1k_1}\right)^2 \\
\leq\ &\left(\sum_{k_1\geq 1}(1+\gamma_{k_1}^{-1})a_{k_1k_1}\right)^2.
\end{aligned}$$

Recall that

$$\begin{aligned}
a_{k_1k_1} &= \int_{\mathcal{T}\times\mathcal{T}}\varphi_{k_1}(s)\varphi_{k_1}(t)C_0(s,t)dsdt \\
&= \mathbb{E}\left(\int_{\mathcal{T}}X(t)\varphi_{k_1}(t)dt\right)^2.
\end{aligned}$$

We get

$$\sum_{k_1\geq 1}(1+\gamma_{k_1}^{-1})a_{k_1k_1} = \mathbb{E}\sum_{k_1=1}^{\infty}(1+\gamma_{k_1}^{-1})\left(\int_{\mathcal{T}}X(t)\varphi_{k_1}(t)dt\right)^2 = \mathbb{E}\left\|X\right\|_K^2. \tag{57}$$

Hence the first term can be bounded by cn^{-1}. Together with the fact that (see, e.g., Lin, 2000)

$$\sum_{k_1,k_2=1}^{\infty}(1+\gamma_{k_1k_2}^{-1})^a(1+\lambda\gamma_{k_1k_2}^{-1})^{-2} \sim \lambda^{-(a+(1/2\alpha))}\log(1/\lambda), \tag{58}$$

The proof can now be completed by collecting all these inequalities. ∎

It now remains to bound $\hat{C}_\lambda - \tilde{C}_\lambda$.

Lemma 10 *If there exists some* $1/(2\alpha) < b \leq 1$ *such that*

$$\frac{\log(1/\lambda)}{nm\lambda^{2b+1/(2\alpha)}} \to 0, \tag{59}$$

then for any $0 \leq a \leq b$,

$$\left\|\hat{C}_\lambda - \tilde{C}_\lambda\right\|_a^2 = o_p\left(n^{-1} + n^{-1}\lambda^{1-a-\frac{1}{2\alpha}}\log(1/\lambda) + (nm)^{-1}\lambda^{-a-\frac{1}{2\alpha}}\log(1/\lambda)\right). \tag{60}$$

Before proving Lemma 10, we state the following technical lemma which is proved in the Appendix.

Lemma 11 *Under the conditions of Theorem 4, if* $1/2\alpha < b \leq 1$, *then for any* $0 \leq a \leq b$

$$\left\|\hat{C}_\lambda - \tilde{C}_\lambda\right\|_a^2 = O_p\left(\frac{\log(1/\lambda)}{nm\lambda^{a+b+1/(2\alpha)}}\right)\left\|\hat{C}_\lambda - \bar{C}_\lambda\right\|_b^2. \tag{61}$$

Proof of Lemma 10: By Lemma 11,

$$\left\|\hat{C}_\lambda - \tilde{C}_\lambda\right\|_a^2 = O_p\left(\frac{\log(1/\lambda)}{nm\lambda^{a+b+1/(2\alpha)}}\right)\left\|\hat{C}_\lambda - \bar{C}_\lambda\right\|_b^2. \tag{62}$$

.

If $a > 1/2\alpha$, then taking $b = a$ yields

$$\left\|\hat{C}_\lambda - \tilde{C}_\lambda\right\|_a^2 = O_p\left(\frac{\log(1/\lambda)}{nm\lambda^{2a+1/(2\alpha)}}\right)\left\|\hat{C}_\lambda - \bar{C}_\lambda\right\|_a^2. \tag{63}$$

Assuming that

$$\frac{\log(1/\lambda)}{nm\lambda^{2a+1/(2\alpha)}} \to 0, \tag{64}$$

then

$$\left\|\hat{C}_\lambda - \tilde{C}_\lambda\right\|_a = o_p\left(\left\|\hat{C}_\lambda - \bar{C}_\lambda\right\|_a\right). \tag{65}$$

Together with the triangular inequality

$$\left\|\tilde{C}_\lambda - \bar{C}_\lambda\right\|_a \geq \left\|\hat{C}_\lambda - \bar{C}_\lambda\right\|_a - \left\|\hat{C}_\lambda - \tilde{C}_\lambda\right\|_a = (1 - o_p(1))\left\|\hat{C}_\lambda - \bar{C}_\lambda\right\|_a. \tag{66}$$

Therefore,

$$\begin{aligned}\left\|\hat{C}_\lambda - \bar{C}_\lambda\right\|_a^2 &= O_p\left(\left\|\tilde{C}_\lambda - \bar{C}_\lambda\right\|_a^2\right) \\ &= O_p\left(n^{-1} + n^{-1}\lambda^{1-a-\frac{1}{2\alpha}}\log(1/\lambda) + (nm)^{-1}\lambda^{-a-\frac{1}{2\alpha}}\log(1/\lambda)\right).\end{aligned}$$

Now consider the case when $a \leq 1/2\alpha$. From the previous discussion, we know that for any $b > 1/(2\alpha)$ such that

$$\frac{\log(1/\lambda)}{nm\lambda^{2b+1/(2\alpha)}} \to 0, \tag{67}$$

we have

$$\left\|\hat{C}_\lambda - \bar{C}_\lambda\right\|_b^2 = O_p\left(n^{-1} + n^{-1}\lambda^{1-b-\frac{1}{2\alpha}}\log(1/\lambda) + (nm)^{-1}\lambda^{-b-\frac{1}{2\alpha}}\log(1/\lambda)\right) \tag{68}$$

Hence,

$$\begin{aligned}\left\|\hat{C}_\lambda - \tilde{C}_\lambda\right\|_a^2 &= O_p\left(\frac{\log(1/\lambda)}{nm\lambda^{a+b+1/(2\alpha)}}\right) \\ &\quad \times O_p\left(n^{-1} + n^{-1}\lambda^{1-b-\frac{1}{2\alpha}}\log(1/\lambda) + (nm)^{-1}\lambda^{-b-\frac{1}{2\alpha}}\log(1/\lambda)\right) \\ &= o_p\left(n^{-1}\lambda^{b-a} + n^{-1}\lambda^{1-a-\frac{1}{2\alpha}}\log(1/\lambda) + (nm)^{-1}\lambda^{-a-\frac{1}{2\alpha}}\log(1/\lambda)\right).\end{aligned}$$

To sum up, if there exists some $1/(2\alpha) < b \leq 1$ such that

$$\frac{\log(1/\lambda)}{nm\lambda^{2b+1/(2\alpha)}} \to 0, \tag{69}$$

then for any $0 \leq a \leq b$,

$$\left\|\hat{C}_\lambda - \tilde{C}_\lambda\right\|_a^2 = o_p\left(n^{-1} + n^{-1}\lambda^{1-a-\frac{1}{2\alpha}}\log(1/\lambda) + (nm)^{-1}\lambda^{-a-\frac{1}{2\alpha}}\log(1/\lambda)\right). \blacksquare \tag{70}$$

Combining the bounds on $\bar{C}_\lambda - C_0$, $\tilde{C}_\lambda - \bar{C}_\lambda$ and $\hat{C}_\lambda - \tilde{C}_\lambda$, we have

$$\|\hat{C}_\lambda - C_0\|_{L_2}^2 = O_p\left(\lambda + n^{-1} + n^{-1}\lambda^{1-\frac{1}{2\alpha}}\log(1/\lambda) + (nm)^{-1}\lambda^{-\frac{1}{2\alpha}}\log(1/\lambda)\right). \tag{71}$$

Taking

$$\lambda \sim (nm/\log n)^{-\frac{2\alpha}{2\alpha+1}} \tag{72}$$

yields that

$$\|\hat{C}_\lambda - C_0\|_{L_2}^2 = O_p\left((nm/\log n)^{-\frac{2\alpha}{2\alpha+1}} + n^{-1}\right). \blacksquare \tag{73}$$

7.4 Proof of Corollary 5

As shown by Bhatia, Davis and McIntosh (1983),

$$\sup_{k\geq 1}\delta_k\|\hat{\psi}_k - \psi_k\| \leq 8^{1/2}\|\hat{C} - C\|_{L_2}, \tag{74}$$

where $\delta_k = \min_{1\leq j\leq k}(\theta_j - \theta_{j+1})$. The proof can then be completed by Theorem 4 and the fact that θ_k is of multiplicity one. $\blacksquare$

7.5 Proof of Theorem 6

First note that, by taking $d > 0$ small enough,

$$\limsup_{n\to\infty} \sup_{L(X)\in\mathcal{P}'(\alpha;M_0)} P\left(\|\tilde{C} - C_0\|_{L_2}^2 > dn^{-1}\right) > 0, \tag{75}$$

by taking $\psi(\cdot)$ to be a constant function. It suffices to show that for some $d > 0$,

$$\limsup_{n\to\infty} \sup_{L(X)\in\mathcal{P}'(\alpha;M_0)} P\left(\|\tilde{C} - C_0\|_{L_2}^2 > d(nm)^{-2\alpha/(2\alpha+1)}\right) > 0. \tag{76}$$

In what follows, we shall assume that T follows a uniform distribution on $\mathcal{T}$ for brevity. We shall also assume that $M_0 = 1$ and $\rho_1 = 1$ without loss of generality. Let $X(\cdot) = \beta\psi(\cdot)$ where $P(\beta = 1) = P(\beta = -1) = 1/2$ and $\psi \in \mathcal{H}(K)$. Clearly $L(X) \in \mathcal{P}'(\alpha; 1)$ if and only $\|\psi\|_{\mathcal{H}(K)}^2 \le 1$. Let $M = c_M(nm)^{1/(2\alpha+1)}$. For a binary vector $b \in \{\pm 1\}^M$, denote

$$\psi_b(\cdot) = M^{-1/2} \sum_{k=M+1}^{2M} b_{k-M}\rho_k^{1/2}\varphi_k(\cdot). \tag{77}$$

It is not hard to see that

$$\|\psi_b\|_{\mathcal{H}(K)}^2 = M^{-1} \sum_{k=M+1}^{2M} b_{k-M}^2 = 1. \tag{78}$$

Furthermore,

$$\|\psi_b - \psi_{b'}\|_{L_2}^2 \ge cM^{-1} \sum_{k=M+1}^{2M} k^{-2\alpha}(b_{k-M} - b'_{k-M})^2 \ge cM^{-(2\alpha+1)}H(b, b') \tag{79}$$

where $H(\cdot,\cdot)$ represents the Hamming distance.

By Varshamov-Gilbert bound, there exists a collection of binary sequences $\{b^{(1)}, \ldots, b^{(N)}\} \subset \{\pm 1\}^m$ such that $N \ge 2^{M/8}$, and

$$H(b^{(j)}, b^{(k)}) \ge M/8, \qquad \forall 1 \le j < k \le N. \tag{80}$$

Therefore,

$$\|\psi_{b^{(j)}} - \psi_{b^{(k)}}\|_{L_2} \ge c(nm)^{-\alpha/(2\alpha+1)}. \tag{81}$$

Let Π_k be the probability measure of (X, T, ε) such that $\psi = \psi_k := (1 + \psi_{b^{(k)}})/2$ and Π_0 be such that $\psi = \psi_0 := 1/2$. It is easy to derive that

$$\|C_0(\psi_k) - C_0(\psi_{k'})\|_{L_2}^2 = \|\psi_k \otimes \psi_k - \psi_{k'} \otimes \psi_{k'}\|_{L_2}^2 \ge c(nm)^{-2\alpha/(2\alpha+1)}, \tag{82}$$

where $C_0(\psi_k)$ is the covariance function of X when $\psi = \psi_k$. On the other hand, the Kullback-Leibler distance from probability measure Π_k to Π_0 can be bounded by

$$KL(\Pi_k|\Pi_0) \le cnm\|\psi_k - \psi_0\|_{L_2}^2 \le c(nm)^{1/(2\alpha+1)} \le c\log N. \tag{83}$$

The proof can now be completed using Fano's lemma by taking c_M large enough. ∎

References

[1] Aronszajn, N. (1950), Theory of reproducing kernels, *Transactions of the American Mathematical Society*, **68**, 337-404.

[2] Bhatia, R., Davis, C. and Mcintosh, A. (1983), Perturbation of spectral subspaces and solution of linear operator equations, *Linear Algebra and Its Applications*, **52/53**, 45-67.

[3] Bosq, D. (2000), *Linear Processes in Function Spaces: Theory and Applications*, New York: Springer.

[4] Cai, T.T. and Yuan, M. (2010), Optimal estimation of the mean function based on discretely sampled functional data: Phase transition, *Manuscript*.

[5] Capra, W.B., and Müller, H.G. (1997), An accelerated-time model for response curves, *Journal of the American Statistical Association*, **92**, 72-83.

[6] Diggle, P., Heagerty, P., Liang, K. and Zeger, S. (2002), *Analysis of Longitudinal Analysis*, 2nd ed. Oxford University Press.

[7] Ferraty, F. and Vieu, P. (2006), *Nonparametric Functional Data Analysis: Methods, Theory, Applications and Implementations*, Springer, New York.

[8] Gu, C. (2002), *Smoothing Spline ANOVA Models*, Springer, New York.

[9] Hall, P., Müller, H. and Wang, J. (2006), Properties of principal component methods for functional and longitudinal data analysis, *Annals of Statistics*, **34**, 1493-1517.

[10] Lin, Y. (2000), Tensor product space ANOVA models, *Annals of Statistics*, **28**, 734-755.

[11] Müller, H. (2005), Functional modeling and classication of longitudinal data (with discussion), *Scandinavia Journal of Statistics*, **32**, 223-246.

[12] James, G. M. and Hastie, T. J. (2001), Functional linear discriminant analysis for irregularly sampled curves, *Journal of the Royal Statistical Society (Series B)*, **63**, 533-550.

[13] James, G. M., Hastie, T. J. and Sugar, C. (2000), Principal component models for sparse functional data, *Biometrika*, **87**, 587-602.

[14] Kaslow, R., Ostrow, D., Detels, R., Phair, J., Polk, B. and Rinaldo, C. (1987), The multicenter AIDS cohort study: Rationale, organization and selected characteristics of the participants, *American Journal of Epidemiology*, **126**, 310-318.

[15] Paul, D. and Peng, J. (2009), Consistency of restricted maximum likelihood estimators of principal components, *Annals of Statistics*, **37**, 1229-1271.

[16] Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies.* Springer, New York.

[17] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis,* 2nd Edition. Springer, New York.

[18] Rice, J., and Silverman, B. (1991), Estimating the mean and covariance structure nonparametrically when the data are curves, *Journal of the Royal Statistical Society, Ser. B*, **53**, 233-243.

[19] Rice, J. A. and Wu, C. O. (2001), Nonparametric mixed effects models for unequally sampled noisy curves, *Biometrics*, **57**, 253-259.

[20] Shi , M., Weiss , R. and Taylor, J. (1996), An analysis of paediatric CD4 counts for acquired immune deciency syndrome using exible random curves, *Applied Statistics*, **45**, 151-163.

[21] Staniswalis, J. and Lee, J. (1998), Nonparametric regression analysis of longitudinal data, *Journal of the American Statistical Association*, **93**, 1403-1418.

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

[23] Stone, C. (1982), Optimal global rates of convergence for nonparametric regression, *Annals of Statistics*, **10**, 1040-1053.

[24] Wahba, G. (1990), *Spline Models for Observational Data*, SIAM, Philadelphia.

[25] Yao, F., Müller, H., and Wang, J. (2005), Functional data analysis for sparse longitudinal data, *Journal of the American Statistical Association*, **100**, 577-590.

Appendix

In this appendix we prove Lemmas 9 and 11 which are used in the proof of the main results given in Section 7.

A.1 Proof of Lemma 9

Observe that

$$\begin{aligned} U &= \sum_{1\le j\neq k\le m} X(T_j)X(T_k)f(T_j)g(T_k) + \sum_{1\le j\neq k\le m} \epsilon_j X(T_k)f(T_j)g(T_k) \\ &\quad + \sum_{1\le j\neq k\le m} \epsilon_k X(T_j)f(T_j)g(T_k) + \sum_{1\le j\neq k\le m} \epsilon_j\epsilon_k f(T_j)g(T_k) \\ &=: U_1 + U_2 + U_3 + U_4. \end{aligned}$$

By Cauchy-Schwartz inequality,

$$\mathbb{E}\left[\mathrm{Var}(U|T)\right] \le \mathbb{E}[\mathbb{E}(U^2|T)] \le 4\left(\mathbb{E}(U_1^2) + \mathbb{E}(U_2^2) + \mathbb{E}(U_3^2) + \mathbb{E}(U_4^2)\right). \tag{84}$$

We now bound the four terms on the rightmost hand side separately.

We begin with $\mathbb{E}(U_1^2)$.

$$\begin{aligned} \mathbb{E}(U_1^2) &= \mathbb{E}\left(\sum_{1\le j_1\neq j_2\neq j_3\neq j_4\le m} X(T_{j_1})X(T_{j_2})X(T_{j_3})X(T_{j_4})f(T_{j_1})g(T_{j_2})f(T_{j_3})g(T_{j_4})\right) \\ &\quad +\mathbb{E}\left(\sum_{1\le j_1\neq j_2\neq j_3\le m} X(T_{j_1})X^2(T_{j_2})X(T_{j_3})f(T_{j_1})g(T_{j_2})f(T_{j_2})g(T_{j_3})\right) \\ &\quad +\cdots\cdots+ \\ &\quad +\mathbb{E}\left(\sum_{1\le j_1\neq j_2\le m} X^2(T_{j_1})X^2(T_{j_2})X(T_{j_3})f^2(T_{j_1})g^2(T_{j_2})\right) \\ &= m(m-1)(m-2)(m-3)\mathbb{E}\left[\left(\int X(s)f(s)ds\right)^2\left(\int X(s)g(s)ds\right)^2\right] \\ &\quad +m(m-1)(m-2)\mathbb{E}\left[\left(\int X(s)f(s)ds\right)\left(\int X(s)g(s)ds\right)\left(\int X^2(s)f(s)g(s)ds\right)\right] \\ &\quad +m(m-1)(m-2)\mathbb{E}\left[\left(\int X(s)f(s)ds\right)^2\left(\int X^2(s)g^2(s)ds\right)\right] \\ &\quad +m(m-1)(m-2)\mathbb{E}\left[\left(\int X(s)g(s)ds\right)^2\left(\int X^2(s)f^2(s)ds\right)\right] \\ &\quad +m(m-1)\mathbb{E}\left[\left(\int X^2(s)f^2(s)ds\right)\left(\int X^2(s)g^2(s)ds\right)\right] \end{aligned}$$

By Cauchy-Schwartz inequality,

$$\begin{aligned}
&\mathbb{E}\left[\left(\int X(s)f(s)ds\right)^2\left(\int X(s)g(s)ds\right)^2\right] \\
\leq\ &\left[\mathbb{E}\left(\int X(s)f(s)ds\right)^4\right]^{1/2}\left[\mathbb{E}\left(\int X(s)g(s)ds\right)^4\right] \\
\leq\ &c_0\mathbb{E}\left(\int X(s)f(s)ds\right)^2\mathbb{E}\left(\int X(s)g(s)ds\right)^2,
\end{aligned}$$

where the last inequality holds because of (27). Next, note that

$$\begin{aligned}
&\mathbb{E}\left[\left(\int X(s)f(s)ds\right)\left(\int X(s)g(s)ds\right)\left(\int X^2(s)f(s)g(s)ds\right)\right] \\
\leq\ &\left(\mathbb{E}\left[\left(\int X(s)f(s)ds\right)^2\left(\int X(s)g(s)ds\right)^2\right]\right)^{1/2}\left(\mathbb{E}\left[\left(\int X^2(s)f(s)g(s)ds\right)^2\right]\right)^{1/2}
\end{aligned}$$

Observe that

$$\begin{aligned}
\mathbb{E}\left[\left(\int X^2(s)f(s)g(s)ds\right)^2\right] &= \int_{\mathcal{T}^2}\mathbb{E}[X^2(s)X^2(t)]f(s)g(s)f(t)g(t)dsdt \\
&\leq \int_{\mathcal{T}^2}\left(\mathbb{E}[X^4(s)]\mathbb{E}[X^4(t)]\right)^{1/2}f(s)g(s)f(t)g(t)dsdt \\
&\leq c_0\left(\int_{\mathcal{T}}C(s,s)f(s)g(s)ds\right)^2.
\end{aligned}$$

We have

$$\begin{aligned}
&\mathbb{E}\left[\left(\int X(s)f(s)ds\right)\left(\int X(s)g(s)ds\right)\left(\int X^2(s)f(s)g(s)ds\right)\right] \\
\leq\ &c_0\left[\mathbb{E}\left(\int X(s)f(s)ds\right)^2\mathbb{E}\left(\int X(s)g(s)ds\right)^2\right]^{1/2}\left(\int_{\mathcal{T}}C(s,s)f(s)g(s)ds\right).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}\left[\left(\int X(s)f(s)ds\right)^2\left(\int X^2(s)g^2(s)ds\right)\right] &\leq c_0\mathbb{E}\left(\int X(s)f(s)ds\right)^2\left(\int_{\mathcal{T}}C(s,s)g^2(s)ds\right), \\
\mathbb{E}\left[\left(\int X(s)g(s)ds\right)^2\left(\int X^2(s)f^2(s)ds\right)\right] &\leq c_0\mathbb{E}\left(\int X(s)g(s)ds\right)^2\left(\int_{\mathcal{T}}C(s,s)f^2(s)ds\right), \\
\mathbb{E}\left[\left(\int X^2(s)f^2(s)ds\right)\left(\int X^2(s)g^2(s)ds\right)\right] &\leq c_0\left(\int_{\mathcal{T}}C(s,s)f^2(s)ds\right)\left(\int_{\mathcal{T}}C(s,s)g^2(s)ds\right).
\end{aligned}$$

Collecting all these inequalities, we conclude that

$$\mathbb{E}(U_1^2) \leq c_0m^4\mathbb{E}\left(\int X(s)f(s)ds\right)^2\mathbb{E}\left(\int X(s)g(s)ds\right)^2 + O(m^3). \tag{85}$$

Now consider $\mathbb{E}(U_2^2) = \mathbb{E}(U_3^2)$.

$$\begin{aligned}
\mathbb{E}(U_2^2) &= \mathbb{E}\left(\sum_{1\le j_1\neq j_2\neq j_3\le m} \epsilon_{j_1}^2 X(T_{j_2})X(T_{j_3})f^2(T_{j_1})g(T_{j_2})g(T_{j_3})\right) \\
&\qquad +\mathbb{E}\left(\sum_{1\le j_1\neq j_2\le m} \epsilon_{j_1}^2 X^2(T_{j_2})f^2(T_{j_1})g^2(T_{j_2})\right) \\
&= m(m-1)(m-2)\sigma_0^2\left(\int_{\mathcal{T}} f^2(s)ds\right)\left(\int_{\mathcal{T}^2} C(s,t)g(s)g(t)dsdt\right) \\
&\qquad +m(m-1)\sigma_0^2\left(\int_{\mathcal{T}} f^2(s)ds\right)\left(\int_{\mathcal{T}} C(s,s)g^2(s)ds\right).
\end{aligned}$$

At last, we look at $\mathbb{E}(U_4^2)$.

$$\begin{aligned}
\mathbb{E}(U_4^2) &= \mathbb{E}\left(\sum_{1\le j_1\neq j_2\le m} \epsilon_{j_1}^2\epsilon_{j_2}^2 f^2(T_{j_1})g^2(T_{j_2})\right) \\
&= m(m-1)\sigma_0^4\left(\int_{\mathcal{T}} f^2(s)ds\right)\left(\int_{\mathcal{T}} g^2(s)ds\right).
\end{aligned}$$

In summary, we have

$$\mathbb{E}(U^2) \le 4c_0 m^4 \mathbb{E}\left(\int X(s)f(s)ds\right)^2 \mathbb{E}\left(\int X(s)g(s)ds\right)^2 + O(m^3). \tag{86}$$

The second statement of the lemma follows immediately from the fact that

$$a_{kk} = \int_{\mathcal{T}\times\mathcal{T}} \varphi_k(s)\varphi_k(t)C_0(s,t)dsdt = \mathbb{E}\left(\int_{\mathcal{T}} X(t)\varphi_k(t)dt\right)^2. \blacksquare \tag{87}$$

A.2 Proof of Lemma 11

By definition

$$G_\lambda(\hat{C}_\lambda - \tilde{C}_\lambda) = D^2\ell_{\infty,\lambda}(\bar{C}_\lambda)(\hat{C}_\lambda - \tilde{C}_\lambda). \tag{88}$$

First order condition implies that

$$D\ell_{mn,\lambda}(\hat{C}_\lambda) = D\ell_{mn,\lambda}(\bar{C}_\lambda) + D^2\ell_{mn,\lambda}(\bar{C}_\lambda)(\hat{C}_\lambda - \bar{C}_\lambda) = 0, \tag{89}$$

where we used the fact that $\ell_{mn,\lambda}$ is quadratic. Together with (49), we have

$$\begin{aligned}
D^2\ell_{\infty,\lambda}(\bar{C}_\lambda)(\hat{C}_\lambda - \tilde{C}_\lambda) &= D^2\ell_{\infty,\lambda}(\bar{C}_\lambda)(\hat{C}_\lambda - \bar{C}_\lambda) + D^2\ell_{\infty,\lambda}(\bar{C}_\lambda)(\bar{C}_\lambda - \tilde{C}_\lambda) \\
&= D^2\ell_{\infty,\lambda}(\bar{C}_\lambda)(\hat{C}_\lambda - \bar{C}_\lambda) - D^2\ell_{mn,\lambda}(\bar{C}_\lambda)(\hat{C}_\lambda - \bar{C}_\lambda) \\
&= D^2\ell_{\infty}(\bar{C}_\lambda)(\hat{C}_\lambda - \bar{C}_\lambda) - D^2\ell_{mn}(\bar{C}_\lambda)(\hat{C}_\lambda - \bar{C}_\lambda).
\end{aligned}$$

Therefore,

$$\hat{C}_\lambda - \tilde{C}_\lambda = G_\lambda^{-1}\left[D^2\ell_\infty(\bar{C}_\lambda)(\hat{C}_\lambda - \bar{C}_\lambda) - D^2\ell_{mn}(\bar{C}_\lambda)(\hat{C}_\lambda - \bar{C}_\lambda)\right]. \tag{90}$$

Then

$$\begin{aligned}\left\|\hat{C}_\lambda - \tilde{C}_\lambda\right\|_a^2 &= \sum_{k_1,k_2=1}^{\infty}(1+\gamma_{k_1k_2}^{-1})^a(1+\lambda\gamma_{k_1k_2}^{-1})^{-2} \\ &\quad\times\left[\frac{1}{nm(m-1)}\sum_{i=1}^{n}\sum_{1\le j\ne k\le m}h(T_{ij},T_{ik}) - \int_{\mathcal{T}\times\mathcal{T}}h(s,t)dsdt\right]^2,\end{aligned}$$

where

$$h(s,t) = (\hat{C}_\lambda(s,t) - \bar{C}_\lambda(s,t))\varphi_{k_1}(s)\varphi_{k_2}(t) \in \mathcal{H}(K\otimes K). \tag{91}$$

Write

$$h = \sum_{j_1,j_2\ge 1} h_{j_1j_2}\varphi_{j_1}\otimes\varphi_{j_2}. \tag{92}$$

Then

$$\begin{aligned}&\left[\frac{1}{nm(m-1)}\sum_{i=1}^{n}\sum_{1\le j\ne k\le m}h(T_{ij},T_{ik}) - \int_{\mathcal{T}\times\mathcal{T}}h(s,t)dsdt\right]^2 \\ =\ &\left[\sum_{j_1,j_2\ge1}h_{j_1j_2}\left(\frac{1}{nm(m-1)}\sum_{i=1}^{n}\sum_{1\le j\ne k\le m}\varphi_{j_1}(T_{ij})\varphi_{j_2}(T_{ik}) - \int_{\mathcal{T}\times\mathcal{T}}\varphi_{j_1}(s)\varphi_{j_2}(t)dsdt\right)\right]^2 \\ \le\ &\sum_{j_1,j_2\ge1}\gamma_{j_1j_2}^{-b}h_{j_1j_2}^2\sum_{j_1,j_2\ge1}\gamma_{j_1j_2}^{b}\left(\frac{1}{nm(m-1)}\sum_{i=1}^{n}\sum_{1\le j\ne k\le m}\varphi_{j_1}(T_{ij})\varphi_{j_2}(T_{ik}) - \int_{\mathcal{T}\times\mathcal{T}}\varphi_{j_1}(s)\varphi_{j_2}(t)dsdt\right)^2.\end{aligned}$$

Observe that

$$\begin{aligned}&\mathbb{E}\left(\frac{1}{nm(m-1)}\sum_{i=1}^{n}\sum_{1\le j\ne k\le m}\varphi_{j_1}(T_{ij})\varphi_{j_2}(T_{ik}) - \int_{\mathcal{T}\times\mathcal{T}}\varphi_{j_1}(s)\varphi_{j_2}(t)dsdt\right)^2 \\ =\ &\frac{1}{n}\mathbb{E}\left(\frac{1}{m(m-1)}\sum_{1\le j\ne k\le m}\varphi_{j_1}(T_{ij})\varphi_{j_2}(T_{ik}) - \int_{\mathcal{T}\times\mathcal{T}}\varphi_{j_1}(s)\varphi_{j_2}(t)dsdt\right)^2 \\ \le\ &\frac{1}{nm^2(m-1)^2}\mathbb{E}\left(\sum_{1\le j\ne k\le m}\varphi_{j_1}(T_{ij})\varphi_{j_2}(T_{ik})\right)^2 \\ =\ &\frac{1}{nm^2(m-1)^2}\sum_{\substack{1\le j\ne k\le m\\ 1\le j'\ne k'\le m}}\mathbb{E}\left[\varphi_{j_1}(T_{ij})\varphi_{j_2}(T_{ik})\varphi_{j_1}(T_{ij'})\varphi_{j_2}(T_{ik'})\right] \\ \le\ &\frac{1}{n}\left[\left(\frac{1}{m^2(m-1)^2}\frac{m!}{(m-4)!}-1\right)\left(\int_{\mathcal{T}\times\mathcal{T}}\varphi_{j_1}(s)\varphi_{j_2}(t)\right)^2 + cm^{-1}\right] \\ \le\ &\frac{c}{nm}\end{aligned}$$

where we used the fact that

$$\left(\int_{\mathcal{T}\times\mathcal{T}} \varphi_{j_1}(s)\varphi_{j_2}(t)dsdt\right)^2 \leq \int_{\mathcal{T}\times\mathcal{T}} \varphi_{j_1}^2(s)\varphi_{j_2}^2(t)dsdt = 1. \tag{93}$$

Therefore, whenever $b > 1/2\alpha$,

$$\begin{aligned}
\left\|\hat{C}_\lambda - \tilde{C}_\lambda\right\|_a^2 &\leq O_p\left(\frac{1}{nm}\right) \sum_{k_1,k_2\geq 1} (1+\gamma_{k_1k_2}^{-1})^a (1+\lambda\gamma_{k_1k_2}^{-1})^{-2} \|h\|_b^2 \\
&\leq O_p\left(\frac{1}{nm}\right) \sum_{k_1,k_2\geq 1} (1+\gamma_{k_1k_2}^{-1})^a (1+\lambda\gamma_{k_1k_2}^{-1})^{-2} \left\|\hat{C}_\lambda - \bar{C}_\lambda\right\|_b^2 \|\varphi_{k_1}\otimes\varphi_{k_2}\|_b^2 \\
&= O_p\left(\frac{1}{nm}\right) \left\|\hat{C}_\lambda - \bar{C}_\lambda\right\|_b^2 \sum_{k_1,k_2\geq 1} (1+\lambda\gamma_{k_1k_2}^{-1})^{-2} (1+\gamma_{k_1k_2}^{-1})^{a+b} \\
&= O_p\left(\frac{\log(1/\lambda)}{nm\lambda^{a+b+1/(2\alpha)}}\right) \left\|\hat{C}_\lambda - \bar{C}_\lambda\right\|_b^2 . \blacksquare
\end{aligned}$$