Bernoulli **26**(4), 2020, 2847–2875
https://doi.org/10.3150/20-BEJ1209

Asymptotic properties of penalized splines for functional data

LUO XIAO

Department of Statistics, North Carolina State University, Raleigh, USA. E-mail: lxiao5@ncsu.edu

Penalized spline methods are popular for functional data analysis but their asymptotic properties have not been established. We present a theoretic study of the L_2 and uniform convergence of penalized splines for estimating the mean and covariance functions of functional data under general settings. The established convergence rates for the mean function estimation are mini-max rate optimal and the rates for the covariance function estimation are comparable to those using other smoothing methods.

Keywords: L_2 convergence; functional data analysis; nonparametric regression; penalized splines; uniform convergence

1. Introduction

Functional data analysis (FDA) has been popular in the last two decades; for textbooks on FDA, see [9,16] and [10] and for recent reviews, see [13] and [20]. Functional data methods are nonparametric and classical smoothing methods such as local polynomial methods, splines and wavelets have all played important roles in the development of functional data methods. Among them, local polynomial smoothing has been more often employed because of its ease of theoretic study; see [31] for a summary of theoretical results on FDA and the references therein. The paper [2] established mini-max optimal rates of estimation of the mean function for functional data.

The present paper is concerned about the asymptotic properties of penalized spline estimators for FDA. Penalized splines have become widely used for methods development and applications in recent years because of its computational simplicity and its connection to mixed effects models [18]. The methods have also found lots of successes in FDA, see, e.g., the R package *refund* [11] and the references therein. However, it seems a consensus that the asymptotic theory of penalized splines is difficult to derive and many theoretic gaps haven not been filled even for the standard nonparametric regression setting. To our best knowledge, there are very few theoretic studies of penalized splines for FDA. When the functional data are observed on a fixed common and dense grid, [3] studied the L_2 convergence rate of penalized spline estimators. When functional data are observed densely for each subject under a deterministic design, [28] established the uniform convergence (no rate was given) of mean and covariance estimators based on a mix of penalized splines and local polynomials. When the covariance of functional data is modeled parametrically, [4] studied the L_2 convergence rate of penalized splines for estimating the mean function. Despite those works, asymptotic theory of penalized splines for FDA in general has not yet been developed.

The contribution of the paper is a theoretic study of the L_2 and uniform convergence of penalized spline estimators for estimating the mean and covariance functions under general functional

1350-7265

data settings. First, following [2], both fixed common design and random independent design will be considered. For the former, the results in [3] are special cases of our theoretic results for mean function estimation. As far as we know, our work is the first one to study the theoretical properties of penalized splines for functional data under the random independent design. Moreover, the uniform convergence of penalize splines for nonparametric regression has been elusive until recent works [22,27] and the present paper extends those works for functional data.

Second, under either design, our theoretic study is unified as it includes the three different types of functional data: "sparse", "dense" and "ultra-dense". Generally speaking, these types can be differentiated by their different rates of the average number of observations per subject compared to the number of subjects and have different asymptotic properties, thus causing "phase transition" [2]. See [31] for a first definition of the three types of functional data for random design and see also the discussion for Table 1. We shall establish the rate-optimality of penalized splines for estimating the mean function for each type of data under either design and also derive the corresponding rates for the number of basis functions and the smoothing parameter.

Third, the convergence rates derived in the paper include as special cases both the regression spline type asymptotic rates and the smoothing spline type asymptotic rates, thereby extending the two-type asymptotics in the standard nonparametric regression [5,22]. An interesting finding is that, for ultra-dense functional data, to achieve the mini-max optimal rates of mean function estimation for functional data, the rates of number of basis functions and the smoothing parameter can be more flexible than in nonparametric regression.

Fourth, similar to [31], the study considers a general weighting scheme in aggregating the multiple observations from each subject, which includes as special cases the methods of standard equal weight per observation [30] and the equal weight per subject [2,12]. As a result, a comparison of the two types of weights can be made for penalized splines, similar to that in [31] for local polynomials.

Last but not the least, the theoretic study of covariance function estimation takes into account the mean function estimation (see Section 4.1), while existing theoretic works either assumed that the mean function is known (e.g., [31]) or used a non-conventional approach (e.g., [12]).

The theoretic study is based on recent advances in the asymptotic theory of penalized splines [22,23,27] and earlier theoretic works, for example, [5]. However, in order to establish the theoretic properties of penalized splines for FDA, a few new techniques are employed. For example, the technique for studying the variability of penalized splines for functional data can be useful for the theoretic derivation of penalized splines for general correlated data; see, for example, the derivations between (A.7) and (A.10) in the proof of Theorem 3.1. Another example is, for the random independent design, new techniques are used to derive the uniform convergence of penalized splines; see the proof of Theorem 3.4 on how to deal with the dependence of H_n on the design points and the derivation of (A.14). These new techniques may be useful for theoretic studies of penalized splines under other contexts.

1.1. Functional data model

Suppose that the observed data takes the form $\{(T_{ij}, Y_{ij}) \in \mathcal{T} \times \mathbb{R}, 1 \le j \le N_i, 1 \le i \le n\}$, where Y_{ij} is subject i's jth observation at time $T_{ij} \in \mathcal{T}$ and $\mathcal{T}$ is a compact time interval. Without loss of generality, let $\mathcal{T} = [0, 1]$. The data has n subjects and subject i has N_i observations. We shall consider both fixed design and random design for T_{ij}, however, N_is are assumed non-random. Consider the functional data model

$$Y_{ij} = \mu(T_{ij}) + X_i(T_{ij}) + \epsilon_{ij}, \tag{1.1}$$

where $\mu(t)$ is the fixed mean function, X_i is a zero-mean random function that models subject i's smooth deviation from the mean, and ϵ_{ij}s are random errors. Define $\sigma(s, t) = \mathrm{Cov}\{X_i(s), X_i(t)\}$, the covariance function of the random functions, and $\sigma_\epsilon^2 = \mathbb{E}\epsilon_{ij}^2$, the error variance. We shall use penalized splines to estimate the mean and covariance functions and the focus is on the asymptotic theory of penalized spline estimators for functional data.

1.2. Notation

We use $'$ to denote the transpose of a matrix. The Euclidean norm is $\|\cdot\|_2$, Frobenius norm is $\|\cdot\|_F$ and the operator norm is $\|\cdot\|_{\mathrm{op}}$. For a matrix $A = (a_{ij})$, $\|A\|_{\max} = \max_{i,j} |a_{ij}|$ and $\|A\|_\infty = \max_i \sum_j |a_{ij}|$.

For two scalars, let $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. The notation $a \preceq b$ denotes that there exists an absolute constant $c > 0$ such that $a \leq cb$ for sufficiently large n. And $a \simeq b$ means that $a \preceq b$ and $b \preceq a$, that is, a and b are rate-wise equivalent. Also $a \ll b$ means $a = o(b)$ and $a \gg b$ means $b = o(a)$. For two square matrices A and B, $A \leq B$ denotes that $B - A$ is positive semidefinite, $A \preceq B$ means that there exists an absolute constant $c > 0$ such that $A \leq cB$ for sufficiently large n, and $A \simeq B$ means that $A \preceq B$ and also $B \preceq A$.

For a univariate continuous function g over $\mathcal{T}$, $\|g\|$ denotes its supreme norm. Similarly, if g is a d-variate continuous function over $\mathcal{T}^d$, $\|g\|$ also denotes it supreme norm. We also use $\|g\|_{L_2}$ to denote the L_2 norm of g over $\mathcal{T}^d$. For a positive integer p, denote by $\mathcal{C}^p(\mathcal{T})$ the class of functions with continuous pth derivatives over $\mathcal{T}$. Similarly, denote by $\mathcal{C}^p(\mathcal{T}^2)$ the class of bivariate functions such that if $\sigma \in \mathcal{C}^p(\mathcal{T}^2)$, then $\partial^{2p}\sigma(s,t)/\partial s^p \partial t^p$ is continuous in $\mathcal{T}^2$. Alternatively, we may also define $\mathcal{C}^p(\mathcal{T}^2)$ to be functions such that for any $0 \leq i \leq p$, $\partial^p \sigma(s,t)/\partial s^i \partial t^{p-i}$ is continuous in $\mathcal{T}^2$. For the latter, the theoretical results on covariance function estimation require $p \geq 2$ as the proofs require $\partial^2\sigma(s,t)/\partial s \partial t$ to be continuous in $\mathcal{T}^2$.

Let $\underline{T}_i = (T_{i1}, T_{i2}, \ldots, T_{iN_i})'$ and $\underline{T} = (\underline{T}_1', \underline{T}_2', \ldots, \underline{T}_n')'$. We use the following notation in [31]. Let $\bar{N} = n^{-1}\sum_{i=1}^n N_i$, $\bar{N}_{S2} = n^{-1}\sum_{i=1}^n N_i^2$ and $\bar{N}_H = (n^{-1}\sum_{i=1}^n N_i^{-1})^{-1}$. Note that $\bar{N}_{S2} \geq (\bar{N})^2$ and $\bar{N} \geq \bar{N}_H$.

1.3. Organization of the paper

The rest of the paper is organized as follows. Section 2 introduces the penalized spline estimator for estimating the mean function and the penalized spline estimator for estimating the covariance function, respectively. Section 3 gives the L_2 and uniform convergence rates of penalized splines for mean function estimation. Section 4 gives the L_2 and uniform convergence rates of penalized splines for covariance function estimation. The Appendix gives the proofs of theorems for mean function estimation as well as some technical lemmas.

The supplement [24] contains additional proofs for mean function estimation and all proofs of theorems for covariance function estimation.

2. Penalized spline estimators

2.1. Mean function estimation

Let $B(t) = \{B_1(t), \ldots, B_K(t)\}'$ be the collection of B-spline basis functions of order m (degree $m-1$) and constructed from equally-spaced knots. Here K equals the number of interior knots plus the order of B-splines. We approximate the mean function $\mu(t)$ by the spline function $B'(t)\boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \mathbb{R}^K$ is a coefficient vector to be determined. Penalized splines estimate $\boldsymbol{\theta}$ via the minimization

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \sum_{j=1}^{N_i} w_i \left\{ Y_{ij} - B'(T_{ij})\boldsymbol{\theta} \right\}^2 + \lambda \boldsymbol{\theta}' P \boldsymbol{\theta} \right],$$

where $w_i > 0$ are fixed weights to be specified and satisfy $\sum_{i=1}^n N_i w_i = 1$, the penalty matrix $P \in \mathbb{R}^{K\times K}$ is positive semidefinite and to be specified, and λ is a smoothing parameter that balances data fit and smoothness of the fit. We shall focus on P-splines in [6] so that $\boldsymbol{\theta}' P \boldsymbol{\theta}$ equals the squared sum of the qth order consecutive differences of the coefficient vector $\boldsymbol{\theta}$. However, the theoretical results to be established can be easily adapted to other types of penalized splines, for example, penalized splines in [14] and [17]. Please refer to [22] for a detailed theoretic comparison of the different types of penalized splines. One common choice of the weights is $w_i = N^{-1}$ where $N = \sum_{i=1}^n N_i$, that is, each observation has equal weight (denoted OBS). Alternatively as in [12], $w_i = (nN_i)^{-1}$, which implies that each subject has the same weight (denoted SUBJ). For the theoretic study, we shall employ a general set of weights as in [31], which includes the above choices as special cases.

Let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iN_i})' \in \mathbb{R}^{N_i}$, $B_i = [B(T_{i1}), B(T_{i2}), \ldots, B(T_{iN_i})]' \in \mathbb{R}^{N_i \times K}$ and $W_i = w_i I_{N_i}$. Furthermore, let $Y = (Y_1', Y_2', \ldots, Y_n')'$, $B = [B_1', B_2', \ldots, B_n']'$ and $W = \text{blockdiag}(W_1, W_2, \ldots, W_n)$. Also let $G_n = \sum_{i=1}^n B_i' W_i B_i = B'WB$ and $H_n = G_n + \lambda P$. Then, $\widehat{\boldsymbol{\theta}} = H_n^{-1}(B'WY)$ and the penalized spline estimate of the mean function is $\widehat{\mu}(t) = B'(t)\widehat{\boldsymbol{\theta}}$. The smoothing parameter λ can be selected by cross-validation.

2.2. Covariance function estimation

To estimate the covariance function $\sigma(s,t)$, a standard procedure consists of two steps [29]. In the first step, an empirical estimate of the covariance function is constructed. Let $\widehat{\mu}$ be any estimate of the mean function and $\widetilde{e}_{ij} = Y_{ij} - \widehat{\mu}(T_{ij})$ be the residuals and $\widetilde{\sigma}_{ij_1j_2} = \widetilde{e}_{ij_1}\widetilde{e}_{ij_2}$ be the auxiliary variables. Then $\{\widetilde{\sigma}_{ij_1j_2} : 1 \le j_1 \ne j_2 \le N_i, i = 1, \ldots, n\}$ is a collection of empirical estimates of the covariance function. In the second step, the empirical estimates are smoothed using a bivariate smoother. Our interest is penalized spline smoothing and we shall use a modified version of the sandwich smoother [26], which is based on tensor-product splines.

Let $\bar{B}(t) = \{\bar{B}_1(t), \ldots, \bar{B}_L(t)\}'$ be the collection of B-splines of order m and constructed from equally-spaced knots. To simplify the theoretic study, we use the same order of B-splines as in the univariate smoothing although the number of basis functions is L, which can be different from K. Throughout the paper, we use the bar symbol to denote that the dimensions are related to L instead of K. The covariance function $\sigma(s,t)$ is modeled as a tensor-product spline $\mathcal{H}(s,t) = \sum_{1\le k,\ell\le L} \theta_{k\ell}\bar{B}_k(s)\bar{B}_\ell(t)$, where $\Theta = (\theta_{k\ell}) \in \mathbb{R}^{L\times L}$ is a coefficient matrix to be determined. Let $\mathrm{vec}(\cdot)$ be an operator that stacks the columns of a matrix into a vector and $\boldsymbol{\theta}_\sigma = \mathrm{vec}(\Theta)$. Consider the penalized least squares

$$\mathrm{PLS} = \sum_{i=1}^{n}\Bigg[v_i \sum_{1\le j_1\ne j_2\le N_i}\big\{\widetilde{\sigma}_{ij_1j_2} - \mathcal{H}(T_{ij_1}, T_{ij_2})\big\}^2\Bigg] + \boldsymbol{\theta}_\sigma' P_\sigma \boldsymbol{\theta}_\sigma,$$

where the penalty matrix P_σ is to be specified and $v_i > 0$ are weights such that $\sum_{i=1}^n v_i N_i(N_i - 1) = 1$. Let $\bar{B}_i = [\bar{B}(T_{i1}), \ldots, \bar{B}(T_{iN_i})]' \in \mathbb{R}^{N_i\times L}$ and $\bar{G}_n = (n\bar{N})^{-1}\sum_{i=1}^n \bar{B}_i'\bar{B}_i$. We use the penalty matrix

$$P_\sigma = \lambda_\sigma \bar{G}_n \otimes \bar{P} + \lambda_\sigma \bar{P} \otimes \bar{G}_n + \lambda_\sigma^2 \bar{P}\otimes\bar{P},$$

where λ_σ is a smoothing parameter and $\bar{P}$ is similar to P but has dimension $L\times L$. The above penalty is similar to the penalty matrix used by the sandwich smoother [26]. Alternatively, one may use the penalty in [7] as in [8] and [25] or the one in [21]. With those penalties, similar L_2 convergence rates can be established with some modification of the proofs; however, it remains unclear how to derive the corresponding uniform convergence rates.

By minimizing PLS with respect to Θ, we obtain the estimate

$$\widehat{\Theta} = (\widehat{\theta}_{k\ell}) = \arg\min_{\Theta} \mathrm{PLS}$$

and the estimate of the covariance function $\sigma(s,t)$ is $\widehat{\sigma}(s,t) = \bar{B}'(s)\widehat{\Theta}\bar{B}(t)$. Let $\mathrm{tr}(\cdot)$ be the trace operator. It is easy to derive that

$$\boldsymbol{\theta}_\sigma' P_\sigma\boldsymbol{\theta}_\sigma = \lambda_\sigma\,\mathrm{tr}(\bar{P}\Theta\bar{G}_n) + \lambda_\sigma\,\mathrm{tr}\big(\bar{P}\Theta'\bar{G}_n\big) + \lambda_\sigma^2\,\mathrm{tr}(\bar{P}\Theta\bar{P}).$$

Note that the right-hand side of the above equality is the same if Θ is replaced by Θ'. In addition, the auxiliary variables are also symmetric, that is, $\widetilde{\sigma}_{ij_1j_2} = \widetilde{\sigma}_{ij_2j_1}$. Thus, we can derive that $\widehat{\Theta} = \widehat{\Theta}'$ and hence $\widehat{\sigma}(s,t) = \widehat{\sigma}(t,s)$, that is, the estimate is also a symmetric function.

Denote by $\otimes$ the Kronecker product operator. Then $\mathcal{H}(s,t) = \bar{B}'(s,t)\boldsymbol{\theta}_\sigma$, where $\bar{B}(s,t) = \bar{B}(t)\otimes\bar{B}(s)$. Let $\widetilde{\Sigma}_i = (\widetilde{\sigma}_{ij_1j_2}) \in \mathbb{R}^{N_i\times N_i}$ and $\widetilde{\boldsymbol{\sigma}}_i^* = \mathrm{vec}^*(\widetilde{\Sigma}_i)$, where $\mathrm{vec}^*(\cdot)$ is an operator that is the same as $\mathrm{vec}(\cdot)$ except that it excludes the diagonal elements of a square matrix. Let A_i be the sub-matrix of $\bar{B}_i\otimes\bar{B}_i$ that excludes the rows corresponding to the same T_{ij}. Finally, let $V_i = v_i I_{N_i(N_i-1)}$. Then

$$\mathrm{PLS} = \sum_{i=1}^{n}\big(\widetilde{\boldsymbol{\sigma}}_i^* - A_i\boldsymbol{\theta}_\sigma\big)'V_i\big(\widetilde{\boldsymbol{\sigma}}_i^* - A_i\boldsymbol{\theta}_\sigma\big) + \boldsymbol{\theta}_\sigma' P_\sigma\boldsymbol{\theta}_\sigma.$$

Let $G_{\sigma,n} = \sum_{i=1}^n A_i' V_i A_i$ and $H_{\sigma,n} = G_{\sigma,n} + P_\sigma$. Then

$$\widehat{\boldsymbol{\theta}}_\sigma = \mathrm{vec}(\widehat{\Theta}) = H_{\sigma,n}^{-1}\left(\sum_{i=1}^n A_i' V_i \widetilde{\boldsymbol{\sigma}}_i^*\right).$$

It follows that $\widehat{\sigma}(s,t) = \bar{B}'(s,t)\widehat{\boldsymbol{\theta}}_\sigma$. The smoothing parameter λ_σ may be selected by cross-validation.

3. Asymptotic properties of mean function estimator

In this section, we establish the asymptotic properties of penalized splines introduced in Section 2.1 for estimating the mean function. We first make the following two assumptions, required for both the fixed common design and random independent design.

Assumption 1. (a) The random functions X_i are independent and identically distributed with zero-mean function and covariance function $\sigma(s,t)$; (b) The random errors ϵ_{ij} are independent from the random functions X_i and are independent and identically distributed with mean zero and variance $\sigma_\epsilon^2 < \infty$; (c) $\|\sigma\| < \infty$.

Assumption 2. (a) The number of basis functions K satisfies $K \geq n^{\delta_1}$ for some constant $\delta_1 > 0$ and $K = o(n)$; (b) The smoothing parameter λ satisfies $\lambda = o(n^{-\delta_2})$ for some constant $\delta_2 > 0$.

In Section 3.1, we consider the fixed common design, where all functions are discretely observed at a fixed and equally-spaced set of time points. Then in Section 3.2, we consider the random independent design, where the observed time points are independent and identically distributed. We now introduce some notation. Let $h = K^{-1}$ and $h_e = h \vee \lambda^{1/(2q)}$. Due to its similar role as the bandwidth parameter for kernel methods, the latter notation h_e may be called the "effective bandwidth" for penalized splines. The notation h shall be frequently used throughout the proofs, in accordance with existing theoretic works on spline estimators. To quantify the variance of penalized splines for functional data, we define $\tau_1 = \sum_{i=1}^n N_i w_i^2$ and $\tau_2 = \sum_{i=1}^n N_i(N_i - 1)w_i^2$. Recall that n is the number of subjects, N_i is the number of observations for the ith subject and w_is are fixed weights and satisfy $\sum_{i=1}^n N_i w_i = 1$.

3.1. Fixed common design

Assumption 3 (Fixed common design). (a) $N_i = \bar{N}$ for all i and $T_{ij} = (j - 1/2)/\bar{N}$. (b) $\bar{N} \geq n^{\delta_3}$ for some constant $\delta_3 > 0$; (c) There exists a sufficiently small constant $c_0 > 0$ such that $K \leq c_0 \bar{N}$.

Remark. For covariance function estimation, we shall assume that the number of marginal basis functions L satisfies $L \leq c_0 \bar{N}$ in (c).

Recall from Section 2.1 that we estimate the mean function using penalized splines that are constructed from order m B-spline basis functions and an order q difference penalty. In addition, the number of B-splines is K and the smoothing parameter for the penalty is λ.

Theorem 3.1 (Mean function: L_2 convergence under fixed common design). *Suppose that Assumptions 1–3 hold. If $\mu \in \mathcal{C}^p(\mathcal{T})$ with $q \le p \wedge m$, then*

$$\mathbb{E}\big(\|\widehat{\mu} - \mu\|_{L_2}^2\big) = O\big(K^{-2m}\big) + o\big(K^{-2p}\big) + O\big(\lambda^2 h_e^{-2q}\big) + O\big(\tau_1 h_e^{-1} + \tau_2\big).$$

Similar to [5] and [22] for standard nonparametric regression, the term $O(K^{-2m}) + o(K^{-2p})$ is the order of the integrated and squared approximation bias of spline functions, $O(\lambda^2 h_e^{-2q})$ is the order of the integrated and squared shrinkage bias from the smoothness penalty, and $O(\tau_1 h_e^{-1} + \tau_2)$ is the order of the integrated variability of penalized splines. In particular, $O(\tau_2)$ is due to the correlation between the observations within each subject. So if the observations within each subject are uncorrelated, i.e., the covariance function is always 0, then the variability term would be $O(\tau_1 h_e^{-1})$.

It is worth mentioning that the proof of Theorem 3.1 actually gives a slightly stronger result:

$$\sup_{t \in \mathcal{T}} \mathbb{E}\big[\big\{\widehat{\mu}(t) - \mu(t)\big\}^2\big] = O\big(K^{-2m}\big) + o\big(K^{-2p}\big) + O\big(\lambda^2 h_e^{-2q}\big) + O\big(\tau_1 h_e^{-1} + \tau_2\big).$$

The above result is stronger because the time interval $\mathcal{T}$ is compact. A similar observation can also be made for Theorem 3.3 and will be omitted.

Theorem 3.1 holds for general weights and in the rest of the subsection, we focus on the weights $w_i = (n\bar{N})^{-1}$. Recall that $\bar{N}$ is the average number of observations per subject. The above weights correspond to both the OBS and SUBJ weights for the fixed common design. In this case $\tau_1 = (n\bar{N})^{-1}$ and $\tau_2 = n^{-1} - (n\bar{N})^{-1}$. Because $h_e^{-1} = O(K)$ and $K = O(\bar{N})$ by Assumption 3, $\tau_1 h_e^{-1} = O(n^{-1})$ and $O(\tau_1 h_e^{-1} + \tau_2) = O(n^{-1})$, that is, the variance of penalized splines is $O(n^{-1})$.

Following [5] and [22], we now discuss the two-type asymptotic properties of penalized splines for functional data. For the regression spline type asymptotics, we assume $\lambda = o(K^{-m-q})$, which leads to $h_e \sim h \sim K^{-1}$.

Corollary 3.1 (L_2 convergence: regressione spline type asymptotics under fixed common design). *Suppose that Assumptions 1–3 hold. If $\mu \in \mathcal{C}^p(\mathcal{T})$ with $q \le p \wedge m$ and $\lambda = o(K^{-m-q})$, then*

$$\mathbb{E}\big(\|\widehat{\mu} - \mu\|_{L_2}^2\big) = O\big(K^{-2m}\big) + o\big(K^{-2p}\big) + O\left(\frac{1}{n}\right). \tag{3.1}$$

The L_2 rate in (3.1) is the same as that for regression splines. For regression splines, to achieve optimal rate of convergence, the order of B-splines needs to match the degree of smoothness of the mean function. So we now let $m = p$. Then the rate in (3.1) becomes $O(K^{-2p} + n^{-1})$. It is sharp because if $c_0 c_1(\bar{N} \wedge n^{1/(2p)}) \le K \le c_0 \bar{N}$ for c_0 as in Assumption 3 and some constant

Table 1. Optimal L_2 rates for mean function estimation and corresponding rates for $\bar{N}$, average number of observations per subject, and K, the number of basis functions, for regression spline type asymptotics under the fixed common design and the smoothing parameter λ satisfies $\lambda = o(K^{-m-q})$

Case	$\bar{N}$	K	$\mathbb{E}(\\|\widehat{\mu} - \mu\\|_{L_2}^2)$
Sparse	$o(n^{\frac{1}{2p}})$	$\bar{N}$	$O(\bar{N}^{-2p})$
Dense	$n^{\frac{1}{2p}}$	$n^{\frac{1}{2p}}$	$O(n^{-1})$
Ultra-dense	$\gg n^{\frac{1}{2p}}$	$n^{\frac{1}{2p}} \ll K \preceq \bar{N}$	$O(n^{-1})$

$0 < c_1 < 1$, then

$$\mathbb{E}\big(\|\widehat{\mu} - \mu\|_{L_2}^2\big) = O\left(\bar{N}^{-2p} + \frac{1}{n}\right),$$

which is the optimal L_2 rate [2]. [3] obtained a slightly loose L_2 rate that leads to optimal rate under more stringent conditions, for example, $\bar{N}$ needs to be much larger. For the regression spline type asymptotics, the condition $\lambda K^{-m-q} = o(1)$ means that the smoothing parameter λ does not matter as long as it is sufficiently small.

For the random independent design, [31] identified "sparse", "dense", and "ultra-dense" cases for functional data. In particular, the ultra-dense case means that not only a parametric rate is obtained but also the bias of the estimator is rate-wise smaller than the parametric rate. These cases can be differentiated by the rate of $\bar{N}$, the average number of observations per subject. For regression spline type asymptotics under the fixed common design, it turns out $n^{1/(2p)}$ is the transition rate between the cases for $\bar{N}$, and for those three cases the optimal L_2 rates as well as the corresponding required rates for the number of basis functions K are given in Table 1.

It is interesting that the optimal L_2 rate depends on the number of knots in a somewhat flexible way for the ultra-dense case in Table 1. For the ultra-dense case where $\bar{N} \gg n^{1/(2p)}$, the optimal and parametric rate can be attained whenever $c_0 c_1 n^{1/(2p)} \le K \le c_0 \bar{N}$. This is different from the theoretical results in standard nonparametric regression function estimation using penalized splines, where the number of knots for the optimal L_2 error rate is $n^{1/(2p+1)}$; see, for example, [5] and [22]. Here the required number of knots has a higher rate ($n^{1/(2p)}$) so as to achieve the optimal L_2 rate (Note that the optimal L_2 rate for standard nonparametric function estimation is $n^{-2p/(2p+1)}$.), and its rate can be anywhere between $n^{1/(2p)}$ and $\bar{N}$, showing flexibility of the number of knots for optimal estimation of the mean function for functional data.

Next, we consider the smoothing spline type asymptotics of penalized splines and we assume λh_e^{-2q} is sufficiently large, or equivalently, λK^{2q} is sufficiently large.

Corollary 3.2 (**L_2 convergence: smoothing spline type asymptotics under fixed common design).** *Suppose that Assumptions 1–3 hold. If $\mu \in \mathcal{C}^p(\mathcal{T})$ with $q \le p \wedge m$ and $\lambda K^{2q} \ge C$ for some sufficiently large constant C, then*

$$\mathbb{E}\big(\|\widehat{\mu} - \mu\|_{L_2}^2\big) = O\left(\lambda + \frac{1}{n}\right). \tag{3.2}$$

Table 2. Optimal L_2 rates for mean function estimation and corresponding rates for $\bar{N}$, average number of observations per subject, and λ, the smoothing parameter, for smoothing spline type asymptotics under the fixed common design and the number of basis functions K satisfies λK^{2p} is sufficiently large

Case	$\bar{N}$	λ	$\mathbb{E}(\|\widehat{\mu}-\mu\|_{L_2}^2)$
Sparse	$o(n^{\frac{1}{2p}})$	$\bar{N}^{-2q}$	$O(\bar{N}^{-2p})$
Dense	$n^{\frac{1}{2p}}$	n^{-1}	$O(n^{-1})$
Ultra-dense	$\gg n^{\frac{1}{2p}}$	$\bar{N}^{-2p} \preceq \lambda \ll n^{-1}$	$O(n^{-1})$

The L_2 rate in (3.2) is the same as that for smoothing splines for functional data and shows that the smoothing parameter λ determines the rate of convergence. For fixed common design, $K \leq \bar{N}$ as in Assumption 3. Thus, the condition $\lambda K^{2q} \geq C$ in Corollary 3.2 implies that $\lambda \geq C\bar{N}^{-2q}$. Therefore, if $\bar{N}^{-2q} \preceq \lambda \preceq \bar{N}^{-2q} \vee n^{-1}$, then

$$\mathbb{E}\big(\|\widehat{\mu}-\mu\|_{L_2}^2\big) = O\left(\bar{N}^{-2q} + \frac{1}{n}\right),$$

which is the rate for smoothing splines and will be again rate-optimal [2] if we let $q = p$ and $m \geq p$, that is, the order of penalty is the same as the degree of smoothness of the mean function and the order of B-splines is no smaller than the order of penalty.

Table 2 gives the three cases of functional data with corresponding optimal rates. As discussed in [2] and other works, if $\bar{N} \geq n^{1/(2p)}$, the functional data is "dense" and a parametric L_2 rate is obtained. [31] further identified an "ultra-dense" case when the order of the bias of the local linear smoother decays faster than the parametric rate. The ultra-dense case could also happen for penalized splines and occurs if $\bar{N}^{-2p} = o(n^{-1})$ and λ satisfies $\lambda \geq \bar{N}^{-2p}$ and $\lambda = o(n^{-1})$.

For ultra-dense functional data as in Table 2, that is, $\bar{N} \gg n^{1/(2p)}$, the optimal L_2 rate also depends on the smoothing parameter λ in a flexible way, similar to the role of the number of knots for regression spline type asmptotics; see the discussion after Corollary 3.1. Indeed, the optimal rate can be attained when $\bar{N}^{-2p} \preceq \lambda \preceq n^{-1}$. Again, this is different from the theoretical results in standard nonparametric regression function estimation using penalized splines, where for the optimal L_2 rate, $\lambda \sim n^{-2p/(2p+1)}$; see, for example, [5] and [22]. Here λ has a lower rate (n^{-1}) so as to achieve the parametric L_2 rate, and its rate can be anywhere between $\bar{N}^{-2p}$ and n^{-1}, showing that, for optimal estimation of the mean function for functional data, λ can be rate-wise smaller as long as it is not too small.

We conclude the discussion on the L_2 rates for fixed common design by comparing the rates of required numbers of knots as well as the smoothing parameter for optimal estimation under both types of asymptotics. First, note that the transition rate for $\bar{N}$ is the same for both types of asymptotics. Next, recall that for regression spline type asymptotics (with $m = p$ and $q \leq p$), the conditions are $\lambda = o(K^{-p-q})$ and $n^{1/(2p)} \wedge \bar{N} \preceq K \preceq \bar{N}$, and for smoothing spline type asymptotics (with $q = p$ and $m \geq p$), the conditions are $\bar{N}^{-2q} \preceq \lambda \preceq \bar{N}^{-2q} \vee n^{-1}$ and $1 \preceq \lambda K^{2p}$. For the sparse and dense cases with $\bar{N} \preceq n^{1/(2p)}$, regression spline type asymptotics give $K \sim \bar{N}$ and $\lambda = o(\bar{N}^{-p-q})$ while smoothing spline type asymptotics give $K \sim \bar{N}$ and $\lambda \sim \bar{N}^{-2p}$. Thus, for the sparse and dense cases, the required number of knots are rate-wise the same for optimal estimation under the two-type asymptotics. For the ultra dense case with $\bar{N} \gg n^{1/(2p)}$, regression spline type asymptotics give $n^{1/(2p)} \preceq K \preceq \bar{N}$ and $\lambda = o(K^{-p-q})$ while smoothing spline type asymptotics give $n^{1/(2p)} \preceq K \preceq \bar{N}$, $\bar{N}^{-2q} \preceq \lambda \preceq n^{-1}$ and $1 \preceq \lambda K^{2p}$. So for the ultra-dense case, the required numbers of knots are somehow equally flexible for optimal estimation under the two types of asymptotics.

Now we establish the uniform convergence of penalized splines for estimating the mean function. Compared to the L_2 convergence, additional assumptions, that is, Assumptions 4 and 5 below, are needed. The additional assumptions are standard for establishing uniform convergence.

Assumption 4 (Uniform convergence). There exists a constant $\tau > 2$ such that

$$\mathbb{E}\|X_i\|^\tau < \infty, \qquad \mathbb{E}|\epsilon_{ij}|^\tau < \infty.$$

Assumption 5 (Uniform convergence under fixed common design). (a)

$$\frac{1}{(\tau_1 + \tau_2) h_e^2 \log n} \left(\frac{n}{\log n} \right)^{2/\tau - 2} = O(1),$$

where $\tau > 2$ is in Assumption 4; (b) $(n\bar{N}) \max_i w_i < \infty$.

Theorem 3.2 (Mean function: Uniform convergence under fixed common design). *Suppose that Assumptions 1–5 hold. If $\mu \in \mathcal{C}^p(\mathcal{T})$ with $q \le p \wedge m$, then*

$$\|\widehat{\mu} - \mu\| = O\big(K^{-m}\big) + o\big(K^{-p}\big) + O\big(\lambda h_e^{-q}\big) + O\big\{\big(\tau_1 h_e^{-1} + \tau_2\big)^{1/2} (\log n)^{1/2}\big\} \quad a.s.$$

For the uniform convergence rate, $O(K^{-m}) + o(K^{-p})$ is the order of approximation bias of spline functions, $O(\lambda h_e^{-q})$ is the order of the shrinkage bias from the smoothness penalty, and $O\{(\tau_1 h_e^{-1} + \tau_2)^{1/2} (\log n)^{1/2}\}$ is the order of the variability of penalized splines. Compared to the L_2 rate, the difference is the log term, which is common for uniform convergence.

The derived uniform convergence rates agree with those obtained under standard nonparametric regression setting on the bias of penalized splines; see [22,27]. The order of the variability term, that is, $O\{(\tau_1 h_e^{-1} + \tau_2)^{1/2} (\log n)^{1/2}\}$, under functional data setting can be rate-wise bigger because of correlation between observations within each subject. Indeed, under standard nonparametric regression setting with the same weights, the corresponding variability order would be $O\{(\tau_1 h_e^{-1} \log n)^{1/2}\}$.

Similar to the L_2 rates, we may consider the two-type asymptotics of penalized splines and create tables similar to Tables 1 and 2 for the best uniform convergence rates. Due to the log term in the order of the variability term, the best uniform rate will be slower than the optimal L_2 rate and the transition rates in terms of the number of observations per subject $\bar{N}$ for the three different types of functional data are slightly smaller. Specifically, the best uniform convergence rate is $O(\bar{N}^{-2p} + \log n/n)$ and the transition rate for $\bar{N}$ now becomes $(n/\log n)^{1/(2p)}$. Consequently, the required rates for the number of basis functions and the smoothing parameter will change accordingly and the details are omitted.

3.2. Random independent design

For functional data with random independent design, we impose Assumptions 6 and 7.

Assumption 6 (Random independent design). The design points T_{ij} are independent and identically distributed with the cumulative distribution function Q, where Q has a positive and continuously differentiable density function ρ over $\mathcal{T}$.

Assumption 7 (Random independent design). (a)

$$\tau_1 = o\big(h^2/\log n\big), \qquad \max_i N_i w_i = o(h/\log n);$$

(b)

$$\tau_1^{-2}\left(\sum_{i=1}^n N_i w_i^4\right) = o\big(h^2/\log n\big), \qquad \tau_1^{-1}\max_i N_i w_i^2 = o(h/\log n);$$

(c)

$$\tau_2^{-2}\left\{\sum_{i=1}^n N_i(N_i-1)(4N_i-6)w_i^4\right\} = o\big(h^4/\log n\big),$$
$$\tau_2^{-1}\max_i N_i(N_i-1)w_i^2 = o\big(h^2/\log n\big).$$

The conditions in Assumption 7 are required to ensure that relevant empirical cumulative distribution functions converge at desired orders; see Lemmas A.7, A.8 and A.9 for details. They generally hold for both the OBS and SUBJ weights. Recall that for the former each observation has the same weight in forming the least squares and for the latter each subject has the same weight.

Theorem 3.3 (Mean function: L_2 convergence under random independent design). *Suppose that Assumptions 1–2, 6–7 hold. If $\mu \in \mathcal{C}^p(\mathcal{T})$ with $q \le p \wedge m$, then*

$$\mathbb{E}\big(\|\widehat{\mu}-\mu\|_{L_2}^2|\underline{T}\big) = O\big(K^{-2m}\big) + o\big(K^{-2p}\big) + O\big(\lambda^2 h_e^{-2q}\big) + O\big(\tau_1 h_e^{-1} + \tau_2\big) \quad a.s.$$

Remark. The L_2 rates are the same under both designs; see Theorem 3.1. In addition, the results hold almost surely and hence are stronger than results in [31], which hold in probability.

Similar to the discussions after Theorem 3.1 for the fixed common design, we shall now consider separately two types of asymptotics: regression spline type and smoothing spline type.

Corollary 3.3 (L_2 convergence: regression spline type asymptotics under random independent design). *Suppose that Assumptions 1–2, 6–7 hold. If $\mu \in \mathcal{C}^p(\mathcal{T})$ with $q \le p \wedge m$ and*

$\lambda = o(K^{-m-q})$, *then*

$$\mathbb{E}\big(\|\widehat{\mu} - \mu\|_{L_2}^2 | \underline{T}\big) = O\big(K^{-2m}\big) + o\big(K^{-2p}\big) + O(\tau_1 K + \tau_2) \quad a.s.$$

Moreover, if $m = p$ *and*

$$\tau_1^{-1/(2p+1)} \wedge \tau_2^{-1/(2p)} \preceq K \preceq \tau_1^{-1/(2p+1)} \vee (\tau_2/\tau_1), \tag{3.3}$$

then

$$\mathbb{E}\big(\|\widehat{\mu} - \mu\|_{L_2}^2 | T\big) = O\big(\tau_1^{2p/(2p+1)} + \tau_2\big) \quad a.s. \tag{3.4}$$

We now apply the corollary to the OBS weight, where $w_i = (n\bar{N})^{-1}$, and the SUBJ weight, where $w_i = (nN_i)^{-1}$. For OBS weight, $\tau_1 = (n\bar{N})^{-1}$ and $\tau_2 = (\bar{N}_{S2} - \bar{N})/\{n(\bar{N})^2\}$, where $\bar{N}_{S2} = n^{-1}\sum_{i=1}^n N_i^2$. Hence,

$$\mathbb{E}\big(\|\widehat{\mu} - \mu\|_{L_2}^2 | \underline{T}\big) = O\left\{(n\bar{N})^{-2p/(2p+1)} + \frac{1}{n}\frac{\bar{N}_{S2}}{(\bar{N})^2}\right\} \quad \text{a.s.}$$

For SUBJ weight, $\tau_1 = (n\bar{N}_H)^{-1}$ and $\tau_2 = n^{-1}(1 - \bar{N}_H^{-1})$, where $\bar{N}_H = (n^{-1}\sum_{i=1}^n N_i^{-1})^{-1}$ is the geometric mean of the numbers of observations from the subjects. Hence,

$$\mathbb{E}\big(\|\widehat{\mu} - \mu\|_{L_2}^2 | \underline{T}\big) = O\left\{(n\bar{N}_H)^{-2p/(2p+1)} + \frac{1}{n}\right\} \quad \text{a.s.}$$

Therefore, the penalized spline smoother with SUBJ weight achieves the optimal L_2 rate under the independent design [2]. The same holds for the OBS weight if $\bar{N}_{S2}/(\bar{N})^2$ is bounded. Because $\bar{N} \geq \bar{N}_H$ and $\bar{N}_{S2} \geq (\bar{N})^2$, similar to local linear smoothers [31], the penalized spline smoother with OBS weight may have a rate-wise smaller asymptotic bias while the penalized spline smoother with SUBJ weight could have a rate-wise smaller asymptotic variance.

If $\tau_1^{2p/(2p+1)} = o(\tau_2)$, then the rate of the number of basis functions K to achieve (3.4) has some flexibility and can be any value between $\tau_2^{-1/(2p)}$ and τ_2/τ_1, similar to the observation before for the common design. For SUBJ weight, this happens if $\bar{N}_H n^{-1/(2p)} \to \infty$. As for OBS weight, if we assume $\bar{N}_{S2}/(\bar{N})^2$ is bounded, this happens if $\bar{N}n^{-1/(2p)} \to \infty$.

For SUBJ weight, Table 3 gives a summary of different cases of functional data. Note that condition (3.3) now becomes

$$(n\bar{N}_H)^{1/(2p+1)} \wedge n^{1/(2p)} \preceq K \preceq (n\bar{N}_H)^{1/(2p+1)} \vee \bar{N}_H.$$

A similar table for the OBS weight can be made under the assumption that $\bar{N}_{S2}/(\bar{N})^2$ is bounded. For random design, $\bar{N}_H$ differentiates the three types of functional data and the transition rate for $\bar{N}_H$ is $n^{1/(2p)}$, the same as that for fixed common design.

We now discuss the sparse case in Table 3. To simplify the discussion, further assume that $N_i = \bar{N}$ for all i. Then, $\bar{N}_H = \bar{N}$ as well. We see that the L_2 rate is $O\{(n\bar{N})^{-(2p)/(2p+1)}\}$ and the required rate for K is $(n\bar{N})^{1/(2p+1)}$, which are the corresponding optimal rates for standard

Table 3. Optimal L_2 rates for mean function estimation and corresponding rates for $\bar{N}_H = (n^{-1}\sum_{i=1}^n N_i^{-1})^{-1}$ and the number of basis functions K for regression spline type asymptotics under the random independent design with SUBJ weight and the smoothing parameter λ satisfies $\lambda = O(K^{-p-q})$

Case	$\bar{N}_H$	K	$\mathbb{E}(\Vert\widehat{\mu}-\mu\Vert_{L_2}^2\vert\underline{T})$
Sparse	$o(n^{\frac{1}{2p}})$	$(n\bar{N}_H)^{\frac{1}{2p+1}}$	$O\{(n\bar{N}_H)^{-\frac{2p}{2p+1}}\}$ a.s.
Dense	$n^{\frac{1}{2p}}$	$n^{\frac{1}{2p}}$	$O(n^{-1})$ a.s.
Ultra-dense	$\gg n^{\frac{1}{2p}}$	$n^{\frac{1}{2p}} \ll K \preceq \bar{N}_H$	$O(n^{-1})$ a.s.

nonparametric regression with $n\bar{N}$ independent observations. This means when $\bar{N} = o(n^{1/(2p)})$, that is, sparse functional data, we may ignore the correlation between functional data observations and simply treat them as independent observations. Similar discussion can be made for the sparse case for the large number of knots scenario as well and will be omitted.

Corollary 3.4 (L_2 convergence: smoothing spline type asymptotics under random independent design). *Suppose that Assumptions 1–2, 6–7 hold. If $\mu \in \mathcal{C}^p(\mathcal{T})$ with $q \le p \wedge m$ and $\lambda K^{2q} \ge C$ for some sufficiently large constant C, then*

$$\mathbb{E}\big(\Vert\widehat{\mu}-\mu\Vert_{L_2}^2\vert\underline{T}\big) = O\big(\lambda + \tau_1\lambda^{-1/(2q)} + \tau_2\big) \quad a.s.$$

Moreover, if $q = p$ and $\tau_1^{2q/(2q+1)} \wedge (\tau_2/\tau_1)^{-2q} \le \lambda \le \tau_2 \vee \tau_1^{2q/(2q+1)}$, then

$$\mathbb{E}\big(\Vert\widehat{\mu}-\mu\Vert_{L_2}^2\vert T\big) = O\big(\tau_1^{2p/(2p+1)} + \tau_2\big) \quad a.s. \tag{3.5}$$

Again we apply the corollary to the OBS weight and SUBJ weight. For OBS weight, $\tau_1 = (n\bar{N})^{-1}$ and $\tau_2 = (\bar{N}_{S2} - \bar{N})/\{n(\bar{N})^2\}$, hence

$$\mathbb{E}\big(\Vert\widehat{\mu}-\mu\Vert_{L_2}^2\vert\underline{T}\big) = O\left\{(n\bar{N})^{-2p/(2p+1)} + \frac{1}{n}\frac{\bar{N}_{S2}}{\bar{N}^2}\right\} \quad \text{a.s.}$$

For SUBJ weight, $\tau_1 = (n\bar{N}_H)^{-1}$ and $\tau_2 = n^{-1}(1-\bar{N}_H^{-1})$, hence

$$\mathbb{E}\big(\Vert\widehat{\mu}-\mu\Vert_{L_2}^2\vert\underline{T}\big) = O\left\{(n\bar{N}_H)^{-2p/(2p+1)} + \frac{1}{n}\right\} \quad \text{a.s.}$$

For SUBJ weight, Table 4 gives a summary of different cases of functional data.

If $\tau_1^{2p/(2p+1)} = o(\tau_2)$, then the rate of λ to achieve (3.5) has some flexibility and can be any value between $(\tau_2/\tau_1)^{-2p}$ and τ_2, similar to the observation before for the common design. For SUBJ weight, this happens if $\bar{N}_H n^{-1/(2p)} \to \infty$. As for OBS weight, if we assume $\bar{N}_{S2}/\bar{N}^2$ is bounded, this happens if $\bar{N}n^{-1/(2p)} \to \infty$.

We now consider the uniform convergence of $\widehat{\mu}$ under the random independent design. Assumption 8 is needed for establishing the uniform convergence under the random independent design.

Table 4. Optimal L_2 rates for mean function estimation and corresponding rates for $\bar{N}_H = (n^{-1}\sum_{i=1}^n N_i^{-1})^{-1}$ and the smoothing parameter λ for smoothing spline type asymptotics under the random independent design with SUBJ weight and the number of basis functions K satisfies λK^{2p} is sufficiently large

Case	$\bar{N}_H$	λ	$\mathbb{E}(\Vert\widehat{\mu}-\mu\Vert_{L_2}^2 \mid \underline{T})$
Sparse	$o(n^{\frac{1}{2p}})$	$(n\bar{N}_H)^{-\frac{2p}{2p+1}}$	$O\{(n\bar{N}_H)^{-\frac{2p}{2p+1}}\}$ a.s.
Dense	$n^{\frac{1}{2p}}$	n^{-1}	$O(n^{-1})$ a.s.
Ultra-dense	$\gg n^{\frac{1}{2p}}$	$\bar{N}_H^{-2p} \preceq \lambda \ll n^{-1}$	$O(n^{-1})$ a.s.

Assumption 8 (Uniform convergence under random independent design). (a)

$$\frac{1}{(\tau_1 h_e + \tau_2 h_e^2)\log n}\left(\frac{n}{\log n}\right)^{2/\tau-2} = O(1),$$

where $\tau > 2$ is in Assumption 4; (b)

$$n \max_i N_i w_i < \infty, \qquad (n\bar{N}) \max_i w_i < \infty.$$

Theorem 3.4 (Mean function: Uniform convergence under random independent design). *Suppose that Assumptions 1–2, 4, 6–8 hold. If $\mu \in \mathcal{C}^p(\mathcal{T})$ with $q \le p \wedge m$, then*

$$\Vert\widehat{\mu}-\mu\Vert = O\big(K^{-m}\big) + o\big(K^{-p}\big) + O\big(\lambda h_e^{-q}\big) + O\Big\{\big(\tau_1 h_e^{-1} + \tau_2\big)^{1/2}(\log n)^{1/2}\Big\} \quad a.s.$$

The difference between the uniform rate and L_2 rate is the log term in the last term, in order to have a uniform control over the variability of penalized splines. Similar discussions can be made as those after Theorem 3.2 in terms of best rate estimation and hence are omitted.

4. Asymptotic properties of covariance function estimator

In this section, we establish the asymptotic properties of penalized splines introduced in Section 2.2 for estimating the covariance function. We first make the following two assumptions, required for both the fixed common design and random independent design. Recall that L is the number of marginal basis functions for bivariate penalized splines and λ_σ is the associated smoothing parameter.

Assumption 9.

$$\sup_{t\in\mathcal{T}} \mathbb{E}X_i(t)^4 < \infty, \qquad \mathbb{E}\epsilon_{ij}^4 < \infty.$$

Assumption 10. (a) $L \ge n^{\delta_4}$ for some constant $\delta_4 > 0$ and $L = o(n)$; (b) $\lambda_\sigma = o(n^{-\delta_5})$ for some constant $\delta_5 > 0$.

4.1. Assumptions on mean function estimator

Since the covariance function estimation involves the mean function estimator, to control the effect of the latter, we may need the following assumptions. Let $U_1 = o(1)$ be a non-random value such that

$$\Big(\sup_{t\in\mathcal{T}} \mathbb{E}\big[\big\{\widehat{\mu}(t) - \mu(t)\big\}^4 |\underline{T}\big]\Big)^{1/4} = O(U_1) \tag{4.1}$$

for the fixed common design and almost surely for the random independent design. The above assumption involves fourth moment of the mean function estimator, thus we establish the following lemmas for the penalized spline mean function estimator.

Lemma 4.1 (Mean function: 4th moment convergence under fixed common design). *Suppose that Assumptions in Theorem 3.1 hold. In addition, suppose that Assumption 9 holds. Let*

$$U_1 = \big\{K^{-2m} + o\big(K^{-2p}\big) + \lambda^2 h_e^{-2q} + \tau_1 h_e^{-1} + \tau_2\big\}^{1/2}. \tag{4.2}$$

If $\mu \in \mathcal{C}^p(\mathcal{T})$ with $q \le p \wedge m$, then the penalized spline estimator $\widehat{\mu}$ satisfies

$$\sup_{t\in\mathcal{T}} \mathbb{E}\big[\big\{\widehat{\mu}(t) - \mu(t)\big\}^4\big] = O\big(U_1^4\big).$$

Lemma 4.2 (Mean function: 4th moment convergence under random independent design). *Suppose that Assumptions in Theorem 3.3 hold. In addition, suppose that Assumption 9 holds. Let U_1 be as in (4.2). If $\mu \in \mathcal{C}^p(\mathcal{T})$ with $q \le p \wedge m$, then the penalized spline estimator $\widehat{\mu}$ satisfies*

$$\sup_{t\in\mathcal{T}} \mathbb{E}\big[\big\{\widehat{\mu}(t) - \mu(t)\big\}^4\big] = O\big(U_1^4\big) \quad a.s.$$

Similarly, let $U_2 = o(1)$ be a non-random value such that

$$\|\widehat{\mu} - \mu\| = O(U_2) \quad \text{a.s.} \tag{4.3}$$

See Theorems 3.2 and 3.4 for the form of U_2 for the penalized spline estimator for the fixed common design and random independent random design, respectively.

One note is that if U_1 and U_2 correspond to the optimal rates of convergence for mean function estimation and the mean function is sufficiently smooth, then the convergence rates of covariance function estimation are not affected by the mean function estimation.

We now introduce some notation. Let $h_\sigma = L^{-1}$ and $h_{\sigma,e} = h_\sigma \vee \lambda_\sigma^{1/(2q)}$. To quantify the variability of penalized splines for estimating the covariance function, we define $\widetilde{\tau}_1 = \sum_{i=1}^n v_i^2 N_i(N_i - 1)$, $\widetilde{\tau}_2 = \sum_{i=1}^n v_i^2 N_i(N_i - 1)(N_i - 2)$, and $\widetilde{\tau}_3 = \sum_{i=1}^n v_i^2 N_i(N_i - 1)(N_i - 2)(N_i - 3)$. The terms $\widetilde{\tau}_2$ and $\widetilde{\tau}_3$ are due to correlation between functional data observations.

Table 5. Best L_2 rates for covariance function estimation and corresponding rates for the number of observations per subject $\bar{N}$ and the number of marginal basis functions L for regression spline type asymptotics under the fixed common design and the smoothing parameter λ_σ satisfies $\lambda_\sigma = O(L^{-p-q})$

Case	$\bar{N}$	L	$\mathbb{E}(\\|\widehat{\sigma} - \sigma\\|_{L_2}^2)$
Sparse	$o(n^{\frac{1}{2p}})$	$\bar{N}$	$O(\bar{N}^{-2p})$
Dense	$n^{\frac{1}{2p}}$	$n^{\frac{1}{2p}}$	$O(n^{-1})$
Ultra-dense	$\gg n^{\frac{1}{2p}}$	$n^{\frac{1}{2p}} \ll L \preceq \bar{N}$	$O(n^{-1})$

4.2. Fixed common design

Theorem 4.1 (Covariance function: L_2 convergence under fixed common design). *Suppose that Assumptions 1, 3, 9 and 10 hold and (4.1) holds. If $\sigma \in \mathcal{C}^p(\mathcal{T}^2)$ with $q \le p \wedge m$, then*

$$\mathbb{E}\big(\|\widehat{\sigma} - \sigma\|_{L_2}^2\big) = O\big(U_1^2\big) + O\big(L^{-2m}\big) + o\big(L^{-2p}\big) + O\big(\lambda_\sigma^2 h_{\sigma,e}^{-2q}\big) + O\big(\widetilde{\tau}_1 h_{\sigma,e}^{-2} + \widetilde{\tau}_2 h_{\sigma,e}^{-1} + \widetilde{\tau}_3\big).$$

Similar to the discussions for the mean function estimation with penalized splines, for covariance function estimation, we could derive asymptotics of penalized splines similar to regression splines and smoothing splines and identify the corresponding best convergence rate. Moreover, we could determine the transition rate for the sparse, dense and ultra-dense cases for covariance function estimation.

To simplify the discussion, in the rest of the subsection, we consider the OBS weight which means that $v_i = \{n\bar{N}(\bar{N}-1)\}^{-1}$. Note that for the fixed common design, the OBS weight and the SUBJ weight are the same. It follows that $\widetilde{\tau}_1 = \{n\bar{N}(\bar{N}-1)\}^{-1}$, $\widetilde{\tau}_2 = (\bar{N}-2)/\{n\bar{N}(\bar{N}-1)\}$, and $\widetilde{\tau}_3 = (\bar{N}-2)(\bar{N}-3)/\{n\bar{N}(\bar{N}-1)\}$. Because $h_{\sigma,e}^{-1} = O(\bar{N})$, $O(\widetilde{\tau}_1 h_{\sigma,e}^{-2} + \widetilde{\tau}_2 h_{\sigma,e}^{-1} + \widetilde{\tau}_3) = O(n^{-1})$. Furthermore, we shall assume that the term $O(U_1^2)$ can be removed. It is easy to show that the above term does not affect the rate of covariance function estimation when it corresponds to the optimal rate for mean function estimation and the mean function is sufficiently smooth.

Corollary 4.1 (Covariance function: L_2 convergence for regression spline type asymptotics under fixed common design). *Suppose that Assumptions 1, 3, 9 and 10 hold and (4.1) holds. If $\sigma \in \mathcal{C}^p(\mathcal{T}^2)$ with $q \le p \wedge m$ and $\lambda_\sigma = o(L^{-m-q})$, then*

$$\mathbb{E}\big(\|\widehat{\sigma} - \sigma\|_{L_2}^2\big) = O\big(L^{-2m}\big) + o\big(L^{-2p}\big) + O\bigg(\frac{1}{n}\bigg).$$

For regression spline type asymptotics under the fixed common design, $n^{1/(2p)}$ is the transition rate for $\bar{N}$ between the sparse, dense and ultra-dense cases. In order to obtain optimal rate, we now let $m = p$. Then for those three cases the optimal L_2 rates as well as the corresponding required rates for L are given in Table 5. Here we use the word "best" instead of "optimal" because the mini-max rate for covariance function estimation has not been established. It is our belief that those best rates are actually optimal.

Table 6. Best L_2 rates for covariance function estimation and corresponding rates for $\bar{N}$ and λ_σ for smoothing spline asymptotics under the fixed common design and $\lambda_\sigma L^{2p}$ is sufficiently large

Case	$\bar{N}$	λ_σ	$\mathbb{E}(\\|\widehat{\sigma}-\sigma\\|_{L_2}^2)$
Sparse	$o(n^{\frac{1}{2p}})$	$\bar{N}^{-2p}$	$O(\bar{N}^{-2p})$
Dense	$n^{\frac{1}{2p}}$	n^{-1}	$O(n^{-1})$
Ultra-dense	$\gg n^{\frac{1}{2p}}$	$\bar{N}^{-2p} \preceq \lambda_\sigma \ll n^{-1}$	$O(n^{-1})$

Corollary 4.2 (Covariance function: L_2 convergence for smoothing spline type asymptotics under fixed common design). *Suppose that Assumptions 1, 3, 9 and 10 hold and (4.1) holds. If $\sigma \in \mathcal{C}^p(\mathcal{T}^2)$ with $q \le p \wedge m$ and $\lambda_\sigma L^{2q} \ge C$ for a sufficiently large constant C, then*

$$\mathbb{E}\big(\|\widehat{\sigma}-\sigma\|_{L_2}^2\big) = O(\lambda_\sigma) + O\left(\frac{1}{n}\right).$$

For regression spline type asymptotics under the fixed common design, again $n^{1/(2p)}$ is the transition rate for $\bar{N}$ between the sparse, dense and ultra-dense cases. In order to obtain optimal rate, we now let $q = p$. Then for those three cases the best L_2 rates as well as the corresponding required rates for L are given in Table 6.

Now we establish the uniform convergence of penalized splines for estimating the covariance function. Assumption 11 is standard for establishing the uniform convergence of covariance estimation, while Assumption 12 is needed to establish the uniform convergence of covariance estimation under the fixed common design.

Assumption 11 (Uniform convergence). There exists a constant $\tau_\sigma > 2$ such that

$$\mathbb{E}\|X_i\|^{2\tau_\sigma} < \infty, \qquad \mathbb{E}|\epsilon_{ij}|^{2\tau_\sigma} < \infty.$$

Assumption 12 (Uniform convergence under fixed common design). (a)

$$\frac{1}{(\widetilde{\tau}_1+\widetilde{\tau}_2+\widetilde{\tau}_3)h_{\sigma,e}^4 \log n}\left(\frac{n}{\log n}\right)^{2/\tau_\sigma - 2} = O(1),$$

where $\tau_\sigma > 2$ is in Assumption 11; (b) $(n\bar{N}^2)\max_i v_i < \infty$.

Theorem 4.2 (Covariance function: Uniform convergence under fixed common design). *Suppose that Assumptions 1, 3, 10–12 hold and (4.3) holds. If $\sigma \in \mathcal{C}^p(\mathcal{T}^2)$ with $q \le p \wedge m$, then*

$$\begin{aligned}\|\widehat{\sigma}-\sigma\| = {} & O(U_2) + O\big(L^{-m}\big) + o\big(L^{-p}\big) + O\big(\lambda_\sigma h_{\sigma,e}^{-q}\big) \\ & + O\Big[\big\{\big(\widetilde{\tau}_1 h_{\sigma,e}^{-2} + \widetilde{\tau}_2 h_{\sigma,e}^{-1} + \widetilde{\tau}_3\big)\log n\big\}^{1/2}\Big] \quad a.s.\end{aligned}$$

4.3. Random independent design

Assumption 13 ensures the almost sure convergence of of an empirical cumulative distribution function. Compared with the conditions in Assumption 7 for mean function estimation, there are fewer conditions here. However, the established results are in probability and hence are weaker than those for mean function estimation, which hold almost surely.

Assumption 13 (Random independent design).

$$\widetilde{\tau}_1 = o\big(h_\sigma^4/\log n\big), \qquad \max_i N_i(N_i-1)v_i = o\big(h_\sigma^2/\log n\big).$$

Theorem 4.3 (Covariance function: L_2 convergence under random independent design). *Suppose that $N_i \geq 4$ for all i. Suppose that Assumptions 1, 6, 9–10 and 13 hold and (4.1) holds almost surely. If $\sigma \in \mathcal{C}^p(\mathcal{T}^2)$ with $q \leq p \wedge m$, then*

$$\mathbb{E}\big(\|\widehat{\sigma}-\sigma\|_{L_2}^2|\underline{T}\big) = O_{\mathbb{P}}\big(U_1^2\big) + O_{\mathbb{P}}\big(L^{-2m}\big) + o_{\mathbb{P}}\big(L^{-2p}\big) + O_{\mathbb{P}}\big(\lambda_\sigma^2 h_{\sigma,e}^{-2q}\big) + O_{\mathbb{P}}\big(\widetilde{\tau}_1 h_{\sigma,e}^{-2} + \widetilde{\tau}_2 h_{\sigma,e}^{-1} + \widetilde{\tau}_3\big).$$

Remark. The above derived result holds in probability, rather than almost surely as in Theorem 3.3 for $\mathbb{E}(\|\widehat{\mu}-\mu\|_{L_2}^2|\underline{T})$.

To simplify the discussion, assume that $N_i = \bar{N}$ for all i. Then the SUBJ weights are the same as the OBJ weights and $v_i = \{n\bar{N}(\bar{N}-1)\}^{-1}$. It follows that

$$\widetilde{\tau}_1 h_{\sigma,e}^{-2} + \widetilde{\tau}_2 h_{\sigma,e}^{-1} + \widetilde{\tau}_3 = O\big\{\big(n\bar{N}^2 h_{\sigma,e}^2\big)^{-1} + (n\bar{N}h_{\sigma,e})^{-1} + n^{-1}\big\} = O\big\{\big(n\bar{N}^2 h_{\sigma,e}^2\big)^{-1} + n^{-1}\big\}.$$

Assume now U_1^2 is negligible.

Corollary 4.3 (Covariance function: L_2 convergence for regression spline type asymptotics under random independent design). *Suppose that $N_i \geq 4$ for all i. Suppose that Assumptions 1, 6, 9–10 and 13 hold and (4.1) holds almost surely. If $\sigma \in \mathcal{C}^p(\mathcal{T}^2)$ with $q \leq p \wedge m$ and $\lambda_\sigma = o(L^{-m-q})$, then*

$$\mathbb{E}\big(\|\widehat{\sigma}-\sigma\|_{L_2}^2|\underline{T}\big) = O_{\mathbb{P}}\big(L^{-2m}\big) + o_{\mathbb{P}}\big(L^{-2p}\big) + O_{\mathbb{P}}\big\{\big(n\bar{N}^2 h_\sigma^2\big)^{-1} + n^{-1}\big\}.$$

Moreover, if $m = p$ and $(n\bar{N}^2)^{1/(2p+2)} \wedge n^{1/(2p)} \leq L \leq (n\bar{N}^2)^{1/(2p+2)} \vee \bar{N}$, we obtain the best rate

$$\mathbb{E}\big(\|\widehat{\sigma}-\sigma\|_{L_2}^2|\underline{T}\big) = O_{\mathbb{P}}\big\{\big(n\bar{N}^2\big)^{-2p/(2p+2)} + n^{-1}\big\}.$$

The quantity $n^{1/(2p)}$ is the transition rate for $\bar{N}$ between the different cases for functional data; see Table 7. In particular, for the sparse case, the best rate is the optimal rate for estimating bivariate smooth functions with $n\bar{N}$ independent data points [19]. However, it is worth noting that under further assumptions on the covariance functions, that is, the eigenvalues of the covariance functions decay at certain rates, a better rate might be achievable [1,15].

Table 7. Best L_2 rates for covariance function estimation and corresponding rates for $\bar{N}$, average number of observations per subject, and L, the number of marginal basis functions, for regression spline asymptotics under the random independent design and the smoothing parameter λ_σ satisfies $\lambda_\sigma = O(L^{-p-q})$

Case	$\bar{N}$	L	$\mathbb{E}(\\|\widehat{\sigma} - \sigma\\|_{L_2}^2 \| \underline{T})$
Sparse	$o(n^{\frac{1}{2p}})$	$(n\bar{N}^2)^{\frac{1}{2p+2}}$	$O_{\mathbb{P}}\{(n\bar{N}^2)^{-\frac{2p}{2p+2}}\}$
Dense	$n^{\frac{1}{2p}}$	$n^{\frac{1}{2p}}$	$O_{\mathbb{P}}(n^{-1})$
Ultra-dense	$\gg n^{\frac{1}{2p}}$	$n^{\frac{1}{2p}} \ll K \preceq \bar{N}$	$O_{\mathbb{P}}(n^{-1})$

Corollary 4.4 (Covariance function: L_2 convergence for smoothing spline asymptotics under random independent design). *Suppose that $N_i \geq 4$ for all i. Suppose that Assumptions 1, 6, 9–10 and 13 hold and (4.1) holds almost surely. If $\sigma \in \mathcal{C}^p(\mathcal{T}^2)$ with $q \leq p \wedge m$ and $\lambda_\sigma L^{2q} \geq C$ for some sufficiently large C, then*

$$\mathbb{E}\big(\|\widehat{\sigma} - \sigma\|_{L_2}^2 | \underline{T}\big) = O_{\mathbb{P}}(\lambda_\sigma) + O_{\mathbb{P}}\big\{\big(n\bar{N}^2\lambda_\sigma^{1/q}\big)^{-1} + n^{-1}\big\}.$$

Moreover, if $q = p$ and $(n\bar{N}^2)^{-2p/(2p+2)} \wedge \bar{N}^{-2p} \leq \lambda_\sigma \leq (n\bar{N}^2)^{-2p/(2p+2)} \vee n^{-1}$, we obtain the best rate

$$\mathbb{E}\big(\|\widehat{\sigma} - \sigma\|_{L_2}^2 | \underline{T}\big) = O_{\mathbb{P}}\big\{\big(n\bar{N}^2\big)^{-2p/(2p+2)} + n^{-1}\big\}.$$

For the large number of knots scenario under the random independent design, $n^{1/(2q)}$ is the transition rate for $\bar{N}$ between the sparse, dense and ultra-dense cases and for those three cases the optimal L_2 rates as well as the corresponding required rates for λ_σ are given in Table 8.

Finally, we establish the uniform convergence of covariance function estimation under random independent design and Assumption 14 is needed.

Assumption 14 (Uniform convergence under random independent design). (a)

$$\frac{1}{(\widetilde{\tau}_1 h_{\sigma,e}^2 + \widetilde{\tau}_2 h_{\sigma,e}^3 + \widetilde{\tau}_3 h_{\sigma,e}^4)\log n}\left(\frac{n}{\log n}\right)^{2/\tau_\sigma - 2} = O(1),$$

Table 8. Best L_2 rates for covariance function estimation and corresponding rates for $\bar{N}$, average number of observations per subject, and L, the number of marginal basis functions, for the large number of knots scenario under the random independent design and the smoothing parameter λ_σ satisfies $\lambda_\sigma L^{2p} \geq C$ for some sufficiently large constant C

Case	$\bar{N}$	λ_σ	$\mathbb{E}(\\|\widehat{\sigma} - \sigma\\|_{L_2}^2 \| \underline{T})$
Sparse	$o(n^{\frac{1}{2p}})$	$(n\bar{N}^2)^{-\frac{2p}{2p+2}}$	$O_{\mathbb{P}}\{(n\bar{N}^2)^{-\frac{2p}{2p+2}}\}$
Dense	$n^{\frac{1}{2p}}$	n^{-1}	$O_{\mathbb{P}}(n^{-1})$
Ultra-dense	$\gg n^{\frac{1}{2p}}$	$\bar{N}^{-2p} \preceq \lambda_\sigma \ll n^{-1}$	$O_{\mathbb{P}}(n^{-1})$

where $\tau_\sigma > 2$ is in Assumption 11; (b)

$$n \max_i N_i(N_i - 1)v_i < \infty, \qquad (n\bar{N}) \max_i (N_i - 1)v_i < \infty.$$

Theorem 4.4 (Covariance function: Uniform convergence under random independent design). *Suppose that $N_i \geq 4$ for all i. Suppose that Assumptions 1, 6, 10–11 and 13–14 hold and (4.3) holds. If $\sigma \in \mathcal{C}^p(\mathcal{T}^2)$ with $q \leq p \wedge m$, then*

$$\|\widehat{\sigma} - \sigma\| = O(U_2) + O\big(L^{-m}\big) + o\big(L^{-p}\big) + O\big(\lambda_\sigma h_{\sigma,e}^{-q}\big) \\ + O\big[\big\{\big(\widetilde{\tau}_1 h_{\sigma,e}^{-2} + \widetilde{\tau}_2 h_{\sigma,e}^{-1} + \widetilde{\tau}_3\big)\log n\big\}^{1/2}\big] \quad a.s.$$

Appendix: Technical proofs

A.1. Proofs of theorems for mean function estimation

Notation. For the design points, define $Q_{ni}(t) = N_i^{-1}\sum_{j=1}^{N_i} 1_{\{T_{ij}\leq t\}}$, where $1_{\{\cdot\}}$ is an indicator function that equals 1 if the statement inside the bracket is true and 0 otherwise. Define $Q_n(t) = \sum_{i=1}^n w_i N_i Q_{ni}(t)$ and $\widetilde{Q}_n(t) = \tau_1^{-1}\sum_{i=1}^n w_i^2 N_i Q_{ni}(t)$. The functions $Q_n(t)$ and $\widetilde{Q}(t)$ are empirical cumulative distributions functions and they shall be shown or assumed to converge to a cumulative distribution function $Q(t)$ with a density function $\rho(t)$. For the fixed common design with Assumption 3, $Q_n(t) = \widetilde{Q}_n(t) = Q_{n1}(t)$ and that $\|Q_n - Q\| = O(\bar{N}^{-1})$. Here $Q(t) = t$ and $\rho(t) = 1$ over $\mathcal{T}$. Also define

$$R_n(s,t) = \tau_2^{-1}\sum_{i=1}^n N_i(N_i - 1)w_i^2 1_{\{T_{ij}\leq s\}}1_{\{T_{ij}\leq t\}}.$$

Finally, for a matrix $A = (a_{ij})$, we shall use A_+ to denote that $A_+ = (|a_{ij}|)$. Note that $\|A_+\|_{\max} = \|A\|_{\max}$ and $\|A_+\|_\infty = \|A\|_\infty$.

Proof of Theorems 3.1 and 3.3. The proof is similar for the fixed common design and random independent design. We first consider the common design and then adapt the proof to the independent design.

First, since $\mathcal{T}$ is the unit interval,

$$\mathbb{E}\|\widehat{\mu} - \mu\|_{L_2}^2 \leq \big\|\mathbb{E}\big\{(\widehat{\mu} - \mu)^2\big\}\big\| \leq \|\mathbb{E}\widehat{\mu} - \mu\|^2 + \big\|\mathrm{var}\big\{\widehat{\mu}(\cdot)\big\}\big\|.$$

We shall consider the bias $\|\mathbb{E}\widehat{\mu} - \mu\|^2$ and the variance $\|\mathrm{var}\{\widehat{\mu}(\cdot)\}\|$ separately.

Let η_μ be the spline function in Lemma A.1 such that

$$\|\eta_\mu - \mu\| = O\big(h^m\big) + o\big(h^p\big). \tag{A.1}$$

Let $\boldsymbol{\mu}_i = \{\mu(T_{i1}), \ldots, \mu(T_{iN_i})\}'$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \ldots, \boldsymbol{\mu}_n')'$. Similarly, let $\boldsymbol{\eta}_{\mu,i} = \{\eta_\mu(T_{i1}), \ldots, \eta_\mu(T_{iN_i})\}'$ and $\boldsymbol{\eta}_\mu = (\boldsymbol{\eta}_{\mu,1}', \ldots, \boldsymbol{\eta}_{\mu,n}')'$. Recall that $\widehat{\mu}(t) = B'(t)H_n^{-1}(B'WY)$, where $G_n =$

$B'WB$ and $H_n = G_n + \lambda P$. Then

$$
\begin{aligned}
&\mathbb{E}\{\widehat{\mu}(t)\} \\
&\quad = B'(t)H_n^{-1}(B'W\boldsymbol{\mu}) \\
&\quad = B'(t)G_n^{-1}(B'W\boldsymbol{\mu}) - B'(t)H_n^{-1}(\lambda P)G_n^{-1}(B'W\boldsymbol{\mu}) \\
&\quad = B'(t)G_n^{-1}(B'W\boldsymbol{\eta}_\mu) + B'(t)G_n^{-1}\{B'W(\boldsymbol{\mu} - \boldsymbol{\eta}_\mu)\} - B'(t)H_n^{-1}(\lambda P)G_n^{-1}(B'W\boldsymbol{\mu}).
\end{aligned}
$$

Since η_μ is a spline function, $\eta_\mu(t) = B'(t)\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^K$. Thus, $\boldsymbol{\eta}_\mu = B\boldsymbol{\beta}$ and

$$
B'(t)G_n^{-1}(B'W\boldsymbol{\eta}_\mu) = B'(t)\boldsymbol{\beta} = \eta_\mu(t).
$$

Let $\boldsymbol{\alpha} = B'W(\boldsymbol{\mu} - \boldsymbol{\eta}_\mu) \in \mathbb{R}^K$ and $\boldsymbol{\gamma} = G_n^{-1}(B'W\boldsymbol{\mu}) \in \mathbb{R}^K$. Then,

$$
\mathbb{E}\{\widehat{\mu}(t)\} - \mu(t) = (\eta_\mu - \mu)(t) + B'(t)G_n^{-1}\boldsymbol{\alpha} - B'(t)H_n^{-1}(\lambda P)\boldsymbol{\gamma}.
$$

Because of the non-negativity and unity of B-spline functions, that is, $B_k(t) \geq 0$ and $\sum_k B_k(t) = 1$ for any $t \in \mathcal{T}$, we derive that

$$
\|\mathbb{E}\widehat{\mu} - \mu\| \leq \|\eta_\mu - \mu\| + \|G_n^{-1}\boldsymbol{\alpha}\|_{\max} + \|H_n^{-1}(\lambda P)\boldsymbol{\gamma}\|_{\max}. \tag{A.2}
$$

We first consider $\|G_n^{-1}\boldsymbol{\alpha}\|_{\max}$. Let α_k be the kth element of $\boldsymbol{\alpha}$. Then

$$
\alpha_k = \sum_{i=1}^{n} w_i \sum_{j=1}^{N_i} B_k(T_{ij})\{\mu(T_{ij}) - \eta_\mu(T_{ij})\} = \int B_k(s)\{\mu(s) - \eta_\mu(s)\}\, dQ_n(s),
$$

where $Q_n(s) = \sum_{i=1}^{n} w_i \sum_{j=1}^{N_i} 1_{\{T_{ij} \leq s\}}$ is an empirical cumulative distribution function. By Assumption 3 for the common design and Lemma A.2,

$$
\|\boldsymbol{\alpha}\|_{\max} = o(h^{p+1}). \tag{A.3}
$$

Since $\|G_n^{-1}\boldsymbol{\alpha}\|_{\max} \leq \|G_n^{-1}\|_\infty \|\boldsymbol{\alpha}\|_{\max}$ and $\|G_n^{-1}\|_\infty = O(h^{-1})$ by Lemma A.4,

$$
\|G_n^{-1}\boldsymbol{\alpha}\|_{\max} = o(h^p). \tag{A.4}
$$

Next, by equality (A.19) in Lemma A.6,

$$
\|H_n^{-1}(\lambda P)\boldsymbol{\gamma}\|_{\max} = O(\lambda h_e^{-q}). \tag{A.5}
$$

Combining (A.1), (A.2), (A.4) and (A.5),

$$
\|\mathbb{E}\widehat{\mu} - \mu\| = O(h^m) + o(h^p) + O(\lambda h_e^{-q}). \tag{A.6}
$$

Next, we consider the variance $\|\operatorname{var}\{\widehat{\mu}(\cdot)\}\|$. First,

$$
\operatorname{var}\{\widehat{\mu}(t)\} = B'(t)H_n^{-1}B'W\operatorname{var}(Y)WBH_n^{-1}B(t).
$$

Let

$$\widetilde{\Gamma} = (\widetilde{\gamma}_{k\ell}) = B'W\,\mathrm{var}(Y)WB = \sum_{i=1}^{n} B_i' W_i\,\mathrm{var}(Y_i) W_i B_i. \tag{A.7}$$

Then,

$$\mathrm{var}\{\widehat{\mu}(t)\} = B'(t)H_n^{-1}\widetilde{\Gamma}H_n^{-1}B(t). \tag{A.8}$$

Note that $\mathrm{var}(Y_i) = \{\sigma(T_{ij_1}, T_{ij_2})\}_{1\le j_1, j_2 \le N_i} + \sigma_\epsilon^2 I_{N_i}$. By (A.7),

$$\widetilde{\gamma}_{k\ell} = \sum_{i=1}^{n} w_i^2 \left\{ \sum_{1\le j\le N_i} B_k(T_{ij})B_\ell(T_{ij}) + \sum_{1\le j_1, j_2\le N_i} B_k(T_{ij_1})B_\ell(T_{ij_2})\sigma(T_{ij_1}, T_{ij_2}) \right\}.$$

Define $\Gamma_1 = (\gamma_{1k\ell})$ and $\Gamma_2 = (\gamma_{2k\ell})$ with

$$\gamma_{1k\ell} = \sum_{i=1}^{n} w_i^2 \left\{ \sigma_\epsilon^2 \sum_{1\le j\le N_i} B_k(T_{ij})B_\ell(T_{ij}) \right\}$$

and

$$\gamma_{2k\ell} = \sum_{i=1}^{n} w_i^2 \left\{ \sum_{1\le j_1, j_2\le N_i} B_k(T_{ij_1})B_\ell(T_{ij_2}) \right\}. \tag{A.9}$$

Note that $|\widetilde{\gamma}_{k\ell}| \le \sigma_\epsilon^2 \gamma_{1k\ell} + \|\sigma\|\gamma_{2k\ell}$. Let $\Gamma = (\gamma_{k\ell}) = \sigma_\epsilon^2\Gamma_1 + \|\sigma\|\Gamma_2$. For a matrix $A = (a_{ij})$, we shall use A_+ to denote that $A_+ = (|a_{ij}|)$. By the linearity of terms,

$$\begin{aligned} B'(t)H_n^{-1}\widetilde{\Gamma}H_n^{-1}B(t) &\le B'(t)\big(H_n^{-1}\big)_+(\widetilde{\Gamma})_+\big(H_n^{-1}\big)_+B(t) \\ &\le B'(t)\big(H_n^{-1}\big)_+\Gamma\big(H_n^{-1}\big)_+B(t). \end{aligned}$$

Therefore, by (A.8),

$$\mathrm{var}\{\widehat{\mu}(t)\} \le B'(t)\big(H_n^{-1}\big)_+\Gamma\big(H_n^{-1}\big)_+B(t).$$

By the unity and non-negativity of B-spline bases, that is, $\sum_k B_k(t) = 1$ and $B_k(t) \ge 0$ for any $t \in \mathcal{T}$,

$$\big\|\mathrm{var}\{\widehat{\mu}(\cdot)\}\big\| \le \big\|\big(H_n^{-1}\big)_+\Gamma\big(H_n^{-1}\big)_+\big\|_{\max} \tag{A.10}$$

For the fixed common design, $N_i = \bar{N}$ and $T_{ij} = (j - 1/2)/\bar{N}$ for all j; see Assumption 3(a). Thus, there are $O(\bar{N}h)$ js such that $B_k(\frac{j-1/2}{\bar{N}}) \neq 0$ and the O notation is uniform with respect to k. Hence, $\gamma_{2k\ell} = O\{(\tau_1+\tau_2)h^2\}$ uniformly for k and ℓ, that is, $\|\Gamma_2\|_{\max} = O\{(\tau_1+\tau_2)h^2\}$. It is also easy to show that $\|\Gamma_1\|_{\mathrm{op}} = O(\tau_1 h)$. By Lemma A.6, $\|H_n^{-1}\|_{\max} = O(h_2^{-1})$ and $\|H_n^{-1}\|_\infty = O(h^{-1})$. Thus,

$$\big\|\big(H_n^{-1}\big)_+\Gamma\big(H_n^{-1}\big)_+\big\|_{\max} \le \sigma_\epsilon^2 \big\|H_n^{-1}\big\|_\infty \big\|H_n^{-1}\big\|_{\max}\|\Gamma_1\|_{\mathrm{op}} + \|\sigma\|\big\|H_n^{-1}\big\|_\infty^2\|\Gamma_2\|_{\max}$$

$$= O\big(\tau_1 h_e^{-1}\big) + O(\tau_1 + \tau_2) = O\big(\tau_1 h_e^{-1} + \tau_2\big).$$

By (A.10),

$$\big\|\mathrm{var}\big\{\widehat{\mu}(\cdot)\big\}\big\| = O\big(\tau_1 h_e^{-1} + \tau_2\big).$$

By combining with (A.6), the proof is complete for the fixed common design.

Now we consider the independent design. The expectation (variance) in the above proof becomes now conditional expectation (conditional variance) given the design points $\underline{T}$. For the bias, (A.1) and (A.2) always hold. Equality (A.4) holds almost surely if so does (A.3). We write α_k as $\alpha_{k1} + \alpha_{k2}$, where

$$\alpha_{k1} = \int B_k(s)\big\{\mu(s) - \eta_\mu(s)\big\}\, dQ(s),$$

$$\alpha_{k2} = \int B_k(s)\big\{\mu(s) - \eta_\mu(s)\big\}\, d(Q_n - Q)(s).$$

It is easy to show that $\max_k |\alpha_{k1}| = o(h^{p+1})$ by Lemma A.2. By integration by parts, α_{k2} can be rewritten as

$$\alpha_{k2} = \int g_k(s)(Q_n - Q)(s)\, ds,$$

where $g_k(s) = \partial[B_k(s)\{\mu(s) - \eta_\mu(s)\}]/\partial s$. By Lemma A.7, $\|Q_n - Q\| = o(h)$ almost surely. Using Lemma A.1, we can show that $\max_k |\alpha_{k2}| = o(h^{p+1})$ almost surely. Next, (A.5) also holds almost surely. It follows that (A.6) holds almost surely.

For the variance, the derivation remains valid with slight changes. Now we let $\Gamma_1 = (\gamma_{1k\ell})$ and $\Gamma_2 = (\gamma_{2k\ell})$ with

$$\gamma_{1k\ell} = \int B_k(s) B_\ell(s)\, d\widetilde{Q}_n(s),$$

$$\gamma_{2k\ell} = \iint B_k(s) B_\ell(t)\, d_s\, d_t R_n(s,t).$$

The proof is complete if $\|\Gamma_1\|_{\mathrm{op}} = O(h)$ and $\|\Gamma_2\|_{\max} = O(h^2)$ almost surely.

By Lemma A.8 and a proof similar to that of G_n in Lemma A.3, $\|\Gamma_1\|_{\mathrm{op}} = O(h)$ almost surely. We now consider Γ_2. Define $F = (f_{k\ell}) \in \mathbb{R}^{K\times K}$ with

$$f_{k\ell} = \iint B_k(s) B_\ell(t) \rho(s) \rho(t)\, ds\, dt.$$

We first derive that

$$\gamma_{2k\ell} - f_{k\ell} = \iint B_k(s) B_\ell(t)\, d_s\, d_t \big\{R_n(s,t) - Q(s)Q(t)\big\}.$$

Since $\|F\|_{\max} = O(h^2)$, it is sufficient to show that the term above is uniformly $O(h^2)$ almost surely. Denote $R_n(s,t) - Q(s)Q(t)$ by $Z(s,t)$. Also denote $B_k(s)B_\ell(t)$ by $g_{k\ell}(s,t)$. By integration by parts, $\gamma_{2k\ell} - f_{k\ell}$ can be written as

$$-\int_s \frac{\partial g_{k\ell}(s,1)}{\partial s} Z(s,1)\,ds - \int_t \frac{\partial g_{k\ell}(1,t)}{\partial t} Z(1,t)\,ds + \iint \frac{\partial^2 g_{k\ell}(s,t)}{\partial s\,\partial t} Z(s,t)\,ds\,dt.$$

By Lemma A.9, each of the above three terms is uniformly $O(h^2)$ almost surely and we have now proved $\|\Gamma_2\|_{\max} = O(h^2)$ almost surely. And the proof is now complete for the independent design as well. $\square$

Proof of Theorem 3.2. First,

$$\|\widehat{\mu} - \mu\| \le \|\mathbb{E}\widehat{\mu} - \mu\| + \|\widehat{\mu} - \mathbb{E}\widehat{\mu}\|.$$

By equality (A.6) for the fixed common design, it suffices to show that

$$\|\widehat{\mu} - \mathbb{E}\widehat{\mu}\| = O\bigl[\bigl\{\bigl(\tau_1 h_e^{-1} + \tau_2\bigr)\log n\bigr\}^{1/2}\bigr] \quad \text{a.s.}$$

and the proof is provided in the supplement [24]. $\square$

Proof of Theorem 3.4. First,

$$\|\widehat{\mu} - \mu\| \le \bigl\|\mathbb{E}(\widehat{\mu}|\underline{T}) - \mu\bigr\| + \bigl\|\widehat{\mu} - \mathbb{E}(\widehat{\mu}|\underline{T})\bigr\|.$$

As equality (A.6) (with the expectation replaced by conditional expectation) holds almost surely for the random independent design, it suffices to show that

$$\bigl\|\widehat{\mu} - \mathbb{E}(\widehat{\mu}|\underline{T})\bigr\| = O\bigl[\bigl\{\bigl(\tau_1 h_e^{-1} + \tau_2\bigr)\log n\bigr\}^{1/2}\bigr] \quad \text{a.s.} \tag{A.11}$$

Let $e_{ij} = X_i(T_{ij}) + \epsilon_{ij}$, $\boldsymbol{e}_i = (e_{i1}, \dots, e_{iN_i})'$ and $\boldsymbol{e} = (\boldsymbol{e}_1', \dots, \boldsymbol{e}_n')'$. Define

$$\widetilde{u}(t) = \widehat{\mu}(t) - \mathbb{E}\bigl\{\widehat{\mu}(t)|\underline{T}\bigr\} = B'(t)H_n^{-1}\bigl(B'W\boldsymbol{e}\bigr)$$

and

$$u(t) = B'(t)H^{-1}\bigl(B'W\boldsymbol{e}\bigr) = \sum_{i=1}^{n}\sum_{j=1}^{N_i} a_{ij}(t)w_i e_{ij},$$

where $H = G + \lambda P$ and

$$a_{ij}(t) = B'(t)H^{-1}B(T_{ij}).$$

Note that H does not depend on the design points. Then by proof similar to that of $\|u\|$ in the proof of Theorem 3.2,

$$\|u\| = O\bigl[\bigl\{\bigl(\tau_1 h_e^{-1} + \tau_2\bigr)\log n\bigr\}^{1/2}\bigr] \quad \text{a.s.} \tag{A.12}$$

The main change to the proof is that we need to establish

$$\left\| \sum_i \mathbb{E}\left[\left\{ \sum_j a_{ij}(\cdot) w_i e_{ij} 1_{\{|e_{ij}| \le \widetilde{L}_n\}} \right\}^2 \right] \right\| = O\big(\tau_1 h_e^{-1} + \tau_2\big). \tag{A.13}$$

Using the same notation as in the proof of Theorem 3.2, we first have

$$\sum_i \mathbb{E}\left[\left\{ \sum_j a_{ij}(t) w_i e_{ij} 1_{\{|e_{ij}| \le \widetilde{L}_n\}} \right\}^2 \middle| \underline{T} \right] = B'(t) H^{-1} \widetilde{\Gamma} H^{-1} B(t). \tag{A.14}$$

Using the same technique in the proof of Theorem 3.3 for the random independent design,

$$B'(t) H^{-1} \widetilde{\Gamma} H^{-1} B(t) = B'(t) \big(H^{-1}\big)_+ \Gamma \big(H^{-1}\big)_+ B(t),$$

where the big O notation is uniform with respect to $t \in \mathcal{T}$ and $\Gamma = (\gamma_{k\ell})$ is defined as in the proof of Theorem 3.3. Thus,

$$\sum_i \mathbb{E}\left[\left\{ \sum_j a_{ij}(t) w_i e_{ij} 1_{\{|e_{ij}| \le \widetilde{L}_n\}} \right\}^2 \right] = B'(t) \big(H^{-1}\big)_+ \mathbb{E}(\Gamma) \big(H^{-1}\big)_+ B(t),$$

and furthermore,

$$\left\| \sum_i \mathbb{E}\left[\left\{ \sum_j a_{ij}(\cdot) w_i e_{ij} 1_{\{|e_{ij}| \le \widetilde{L}_n\}} \right\}^2 \right] \right\| \le \big\| \big(H^{-1}\big)_+ \mathbb{E}(\Gamma) \big(H^{-1}\big)_+ \big\|_{\max}. \tag{A.15}$$

With slight abuse of notation, let $\Gamma_1 = (\gamma_{1k\ell})$ and $\Gamma_2 = (\gamma_{2k\ell})$ with

$$\gamma_{1k\ell} = \int B_k(s) B_\ell(s) \rho(s)\, ds,$$

$$\gamma_{2k\ell} = \iint B_k(s) B_\ell(t) \rho(s) \rho(t)\, dt.$$

Then by a proof similar to that of Theorem 3.3 for the random design and (A.15), we could establish (A.13). Then, (A.12) holds.

Because of (A.12), the proof is complete if

$$\|\widetilde{u} - u\| = o\big(\|u\|\big) \quad \text{a.s.} \tag{A.16}$$

We derive that

$$\widetilde{u}(t) - u(t) = -B'(t) H_n^{-1} (G_n - G) H^{-1} \big(B' W \boldsymbol{e}\big).$$

Thus,

$$\|\widetilde{u} - u\| \le \big\| H_n^{-1} (G_n - G) H^{-1} \big(B' W \boldsymbol{e}\big) \big\|_{\max}$$

$$
\begin{aligned}
&\le \|H_n^{-1}\|_\infty \|G_n - G\|_\infty \|G^{-1}\|_\infty \|GH^{-1}(B'W\boldsymbol{e})\|_{\max} \\
&= o(h^{-1}) \|GH^{-1}(B'W\boldsymbol{e})\|_{\max} \quad \text{a.s.}
\end{aligned}
$$

Since

$$
\|GH^{-1}(B'W\boldsymbol{e})\|_{\max} = \left\| \int B(s)\rho(s)u(s)\,ds \right\|_{\max} = O(h\|u\|),
$$

equality (A.16) holds. And the proof is complete. □

A.2. Technical lemmas for mean function estimation

The following lemma is adapted from Lemma 3.1 and Remark 3.1 in [22].

Lemma A.1. *Suppose that Assumption 2(a) holds. If $\mu \in \mathcal{C}^p(\mathcal{T})$, then there exists a spline function $\eta_\mu(t) = B'(t)\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^K$ such that*

$$
\|\mu^{(i)} - \eta_\mu^{(i)}\| = O(h^{m-i}) + o(h^{p-i})
$$

for $i = 0$ and 1.

Lemma A.2. *Suppose that Assumption 2(a) holds and $p \ge 1$. Let η_μ be the spline function in Lemma A.1 and $F(\cdot)$ be any cumulative distribution function in $\mathcal{T}$ that depends on $\underline{T}$. Then for $i = 0$ and 1,*

$$
\max_k \left| \int B_k(t)\{\mu^{(i)}(t) - \eta_\mu^{(i)}(t)\}\,dF(t) \right| = o(h^{p+1-i}) + o(h^{p-i}\|F - Q\|).
$$

Proof. The proof is similar to that of Lemma 3.1 in [22] and hence is omitted. □

For the fixed common design, suppose that Assumptions 2 and 3 always hold. Then Lemmas A.3, A.4, and A.5 below can be established by proofs similar to those of Lemmas 6.2, 6.3 and 6.4 in [32]. Thus, the proofs are omitted. For the independent design, suppose that Assumptions 2, 6 and 7(a) always hold. Then these lemmas still hold almost surely because of Lemma A.7. Let $G = \int B(s)B'(s)\rho(s)\,ds$.

Lemma A.3.

$$
G_n \simeq G \simeq hI.
$$

Lemma A.4. *Denote the (i, j)th element of G_n^{-1} by α_{ij}. There exists a constant $c > 0$ and $0 < \gamma < 1$ such that, for large n,*

$$
|\alpha_{ij}| \le ch^{-1}\gamma^{|i-j|}.
$$

In addition,

$$\|G_n^{-1}\|_\infty = O(h^{-1}).$$

Remark A.1. The same inequalities can be established for G^{-1} with a similar proof.

Lemma A.5.

$$\|G_n - G\|_{\max} = O(\|Q_n - Q\|_{\max}),$$
$$\|G_n^{-1} - G^{-1}\|_{\max} = O(h^{-2}\|Q_n - Q\|_{\max}),$$
$$\|G_n^{-1} - G^{-1}\|_\infty = O(h^{-2}\|Q_n - Q\|_{\max}).$$

Because of Lemmas A.3, A.4 and A.5, the following lemma [27] holds for the fixed common design. For the independent design, if Assumption 7(a) holds, then the lemma also holds almost surely.

Lemma A.6. *The following equalities hold*:

$$\|H_n^{-1}\|_{\max} = O(h_e^{-1}), \tag{A.17}$$
$$\|H_n^{-1}\|_\infty = O(h^{-1}), \tag{A.18}$$
$$\|H_n^{-1}P\boldsymbol{\gamma}\|_{\max} = O(h_e^{-q}). \tag{A.19}$$

The following lemmas establish the convergence of several empirical cumulative distributions under the random independent design.

Lemma A.7. *Suppose that Assumptions 2 and 6 hold. If Assumption 7(a) holds, then $\|Q_n - Q\| = o(h)$ almost surely.*

Proof. The proof is given in the supplement [24]. □

By proofs similar to that for Lemma A.7, we can establish the following lemmas.

Lemma A.8. *Suppose that Assumptions 2 and 6 hold. If Assumption 7(b) holds, then $\|\widetilde{Q}_n - Q\| = o(h)$ almost surely.*

Lemma A.9. *Let $Z(s,t) = R_n(s,t) - Q(s)Q(t)$. Suppose that Assumptions 2 and 6 hold. If Assumption 7(c) holds, then $\|Z\| = o(h^2)$ almost surely.*

Acknowledgements

Luo Xiao was partially supported by Grant Number R01NS091307 from National Institute of Neurological Disorders and Stroke (NINDS) and Grant Number R56AG064803 from National Institute on Aging (NIA).

Supplementary Material

Supplement to "Asymptotic properties of penalized splines for functional data" (DOI: 10.3150/20-BEJ1209SUPP; .pdf). We provide additional proofs for mean and covariance functions estimation.

References

[1] Cai, T. and Yuan, M. (2012). Nonparametric covariance function estimation for functional and longitudinal data. Technical report, Univ. Pennsylvania, Philadelphia, PA.

[2] Cai, T.T. and Yuan, M. (2011). Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *Ann. Statist.* **39** 2330–2355. MR2906870 https://doi.org/10.1214/11-AOS898

[3] Cardot, H. (2000). Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *J. Nonparametr. Stat.* **12** 503–538. MR1785396 https://doi.org/10.1080/10485250008832820

[4] Chen, H., Wang, Y., Paik, M.C. and Choi, H.A. (2013). A marginal approach to reduced-rank penalized spline smoothing with application to multilevel functional data. *J. Amer. Statist. Assoc.* **108** 1216–1229. MR3174699 https://doi.org/10.1080/01621459.2013.826134

[5] Claeskens, G., Krivobokova, T. and Opsomer, J.D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika* **96** 529–544. MR2538755 https://doi.org/10.1093/biomet/asp035

[6] Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.* **11** 89–121. MR1435485 https://doi.org/10.1214/ss/1038425655

[7] Eilers, P.H.C. and Marx, B.D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemom. Intell. Lab. Syst.* **66** 159–174.

[8] Goldsmith, J., Bobb, J., Crainiceanu, C.M., Caffo, B. and Reich, D. (2011). Penalized functional regression. *J. Comput. Graph. Statist.* **20** 830–851. MR2878950 https://doi.org/10.1198/jcgs.2010.10007

[9] Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications. Springer Series in Statistics*. New York: Springer. MR2920735 https://doi.org/10.1007/978-1-4614-3655-3

[10] Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Wiley Series in Probability and Statistics*. Chichester: Wiley. MR3379106 https://doi.org/10.1002/9781118762547

[11] Huang, L., Scheipl, F., Goldsmith, J., Gellar, J., Harezlak, J., Mclean, M., Swihart, B., Xiao, L., Crainiceanu, C.M., Reiss, P., Chen, Y., Greven, S., Huo, L., Kundu, M.G. and Wrobel, J. (2015). R package *mgcv*: Methodology for regression with functional data (version 0.1-13). Available at https://cran.r-project.org/web/packages/refund/index.html.

[12] Li, Y. and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Statist.* **38** 3321–3351. MR2766854 https://doi.org/10.1214/10-AOS813

[13] Morris, J.S. (2015). Functional regression. *Annu. Rev. Stat. Appl.* **2** 321–359.

[14] O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statist. Sci.* **1** 502–527. MR0874480

[15] Paul, D. and Peng, J. (2009). Consistency of restricted maximum likelihood estimators of principal components. *Ann. Statist.* **37** 1229–1271. MR2509073 https://doi.org/10.1214/08-AOS608

[16] Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR2168993

[17] Ruppert, D. and Carroll, R.J. (2000). Theory & methods: Spatially-adaptive penalties for spline fitting. *Aust. N. Z. J. Stat.* **42** 205–223.

[18] Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics* **12**. Cambridge: Cambridge Univ. Press. MR1998720 https://doi.org/10.1017/CBO9780511755453

[19] Stone, C.J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642

[20] Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016). Functional data analysis. *Annu. Rev. Stat. Appl.* **3** 257–295.

[21] Wood, S.N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* **62** 1025–1036. MR2297673 https://doi.org/10.1111/j.1541-0420.2006.00574.x

[22] Xiao, L. (2019). Asymptotic theory of penalized splines. *Electron. J. Stat.* **13** 747–794. MR3925516 https://doi.org/10.1214/19-ejs1541

[23] Xiao, L. (2019). Asymptotics of bivariate penalised splines. *J. Nonparametr. Stat.* **31** 289–314. MR3941214 https://doi.org/10.1080/10485252.2018.1563295

[24] Xiao, L. (2020). Supplement to "Asymptotic properties of penalized splines for functional data." https://doi.org/10.3150/20-BEJ1209SUPP

[25] Xiao, L., Li, C., Checkley, W. and Crainiceanu, C. (2018). Fast covariance estimation for sparse functional data. *Stat. Comput.* **28** 511–522. MR3761337 https://doi.org/10.1007/s11222-017-9744-8

[26] Xiao, L., Li, Y. and Ruppert, D. (2013). Fast bivariate P-splines: The sandwich smoother. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 577–599. MR3065480 https://doi.org/10.1111/rssb.12007

[27] Xiao, L. and Nan, Z. (2019). Uniform convergence of penalized splines. Unpublished manuscript.

[28] Yao, F. and Lee, T.C.M. (2006). Penalized spline models for functional principal component analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 3–25. MR2212572 https://doi.org/10.1111/j.1467-9868.2005.00530.x

[29] Yao, F., Müller, H.-G., Clifford, A.J., Dueker, S.R., Follett, J., Lin, Y., Buchholz, B.A. and Vogel, J.S. (2003). Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics* **59** 676–685. MR2004273 https://doi.org/10.1111/1541-0420.00078

[30] Yao, F., Müller, H.-G. and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. MR2160561 https://doi.org/10.1198/016214504000001745

[31] Zhang, X. and Wang, J.-L. (2016). From sparse to dense functional data and beyond. *Ann. Statist.* **44** 2281–2321. MR3546451 https://doi.org/10.1214/16-AOS1446

[32] Zhou, S., Shen, X. and Wolfe, D.A. (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist.* **26** 1760–1782. MR1673277 https://doi.org/10.1214/aos/1024691356

Received July 2019 and revised January 2020