Free-knot spline smoothing for functional data

Daniel Gervini
University of Wisconsin – Milwaukee
Department of Mathematical Sciences
3200 N. Cramer St., Room E490
Milwaukee, WI 53211, USA
E-mail: gervini@uwm.edu

February 17, 2006

Abstract

This article introduces free-knot regression spline estimators for the mean and the variance components of a sample of curves. The asymptotic distribution of the mean estimator is derived, and asymptotic confidence bands are constructed. A comparative simulation study shows that free-knot splines estimate salient features of the functions (such as sharp peaks) more accurately than smoothing splines. This adaptive behavior is also illustrated by an analysis of weather data.

Key words and phrases: Functional data analysis; Karhunen–Loève decomposition; Longitudinal data analysis; Variance components.

1 Introduction

Many statistical problems today involve the analysis of samples of curves. Ideally, the statistician would observe a sample $X_1(t), \ldots, X_n(t)$ of independent and identically distributed realisations of a stochastic process $X(t)$, $t \in [a,b] \subset \mathbb{R}$. However, in practice only discrete and noisy measurements of those curves will be available. As an example, consider daily precipitation data (averaged over the years 1960 to 1994) for the Canadian cities of Calgary and Vancouver (Fig. 1a). Although the broad annual trends are easy to see, the finer details are obscured by random noise. Even the average over the whole sample of curves (consisting of 35 cities) is so variable that hardly any local features can be discerned, unless some kind of smoothing is applied (Fig. 9b).

This article proposes smooth estimators for the mean $\mu(t) = \mathrm{E}\{X(t)\}$ and the covariance function $\rho(s,t) = \mathrm{Cov}\{X(s), X(t)\}$. The existing approaches are based on individual smoothing of the sample curves, followed by cross-sectional averaging and covariance computation (Rice and Silverman (1991), Ramsay and Silverman (1997)). These methods, however, do not "borrow strength" from the combined dataset at the smoothing step, and are thus prone to oversmoothing (although Kneip (1994) offers some alternatives to ameliorate this problem).

In this article we introduce free-knot spline estimators of $\mu(t)$ and $\rho(s,t)$ that avoid individual smoothing. We show that this approach often produces better estimators than the smoothing splines of Ramsay and Silverman (1997), at the cost of a modest increase in computational complexity. In addition, free-knot spline estimation can be seen as a classical non-linear regression problem, so the asymptotic distribution of $\hat{\mu}(t)$ under the model is easy to derive and can be used for inference.

The idea of free-knot spline smoothing is not new, but due to increased computational power there has been renewed interest in this topic recently; see, for instance, Zhou, Shen and Wolfe (1998), Hansen and Kooperberg (2002) and Mao and Zhao (2003). Most of that work, however, deals with estimation of a single regression curve. Although the basic ideas can be extended to samples of curves, a number of issues arise. For instance, it is not realistic to assume that the data follow a mean-

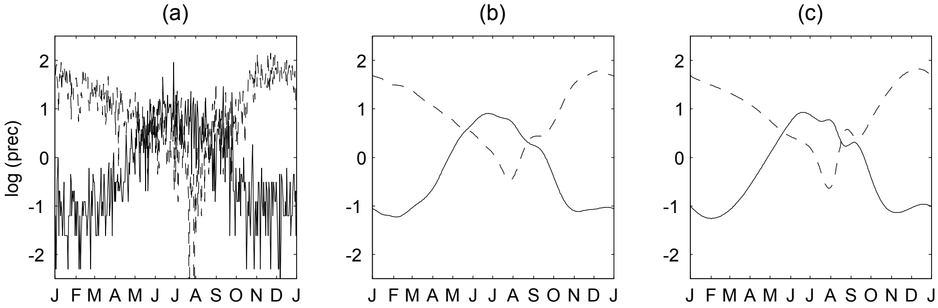

Figure 1: Logarithm of daily precipitation for Calgary (solid line) and Vancouver (dashed line); (a) raw data, (b) smoothing splines, and (c) free-knot regression splines.

plus-error model, as in Mao and Zhao (2003); the covariance function $\rho(s, t)$ induces within-subject correlations that affect the asymptotic distribution of $\hat{\mu}(t)$. Moreover, the asymptotic analysis appropriate in our context would assume that the number of curves n goes to infinity, rather than the number of observations per curve as in Mao and Zhao (2003). For these and other reasons, extending free-knot spline estimation and inference from the individual-curve setting to the functional data context is not trivial.

This article is organised as follows. Free-knot spline estimators of the mean and the covariance function are introduced in Section 2. Their asymptotic behavior is established in Section 3, and confidence intervals are derived. The finite-sample behavior of the estimators and the confidence intervals is studied by simulation in Section 4. A real-data example in Section 5 illustrates the comparative performance of the proposed estimators. Technical details are left to the Appendix.

2 Free-knot spline estimation

2.1 Estimation of the mean

As explained in the Introduction, we assume that the data are discrete and noisy realisations of sample curves, so $x_{ij} = X_i(t_j) + \varepsilon_{ij}$, where $a = t_1 < t_2 < \ldots < t_m = b$ and $\{\varepsilon_{ij}\}$ are i.i.d. random errors (throughout this paper we assume that the input grid $\{t_j\}$ is the same for all curves, but this assumption can be relaxed). Therefore, the m-dimensional vectors of observations follow the model

$$\begin{gathered}\mathbf{x}_i = \boldsymbol{\mu} + \boldsymbol{\eta}_i, \text{ where } \mu_j = \mu(t_j) \text{ for } j = 1, \ldots, m, \\ \text{and } \mathrm{E}(\boldsymbol{\eta}_i) = 0, \ \mathrm{V}(\boldsymbol{\eta}_i) < \infty \text{ for } i = 1, \ldots, n.\end{gathered} \tag{1}$$

A spline function of order r with given knots $\tau_1 < \cdots < \tau_p$ is a piecewise polynomial of degree $r-1$ with $r-2$ continuous derivatives in $[a,b]$. This family of functions, denoted by $\mathcal{S}_{r,p}(\boldsymbol{\tau})$, is a linear space of dimension $r+p$. The union of these spaces, $\mathcal{S}_{r,p} = \bigcup_{\boldsymbol{\tau}} \mathcal{S}_{r,p}(\boldsymbol{\tau})$, is usually known as the space of polynomial splines of order r with p free knots (Schumaker, 1980). We will construct an estimator of $\mu(t)$ in the space $\mathcal{S}_{r,p}$.

Given a vector of knots $\boldsymbol{\tau}$ in $[a,b]$, let $\boldsymbol{\kappa} = J(\boldsymbol{\tau})$ be the Jupp-transformed knots, defined as $\kappa_i = \log\{(\tau_{i+1} - \tau_i)/(\tau_i - \tau_{i-1})\}$ for $i = 1, \ldots, p$, where $\tau_0 = a$ and $\tau_{p+1} = b$. This one-to-one transformation maps constrained and increasing knot vectors $\boldsymbol{\tau}$ onto unconstrained and unordered vectors $\boldsymbol{\kappa}$, which has a number of practical and theoretical advantages (Jupp, 1978). Let $\boldsymbol{\beta}(t, \boldsymbol{\kappa}) \in \mathbb{R}^{r+p}$ be the vector of B-spline basis functions corresponding to $\boldsymbol{\kappa}$ (see e.g. Schumaker (1980) for definition and properties) and $B(\boldsymbol{\kappa})$ the $m \times (r+p)$ matrix whose jth row is $\boldsymbol{\beta}(t_j, \boldsymbol{\kappa})^\top$. We define $\hat{\boldsymbol{\mu}} = B(\hat{\boldsymbol{\kappa}})\hat{\mathbf{c}}$, where

$$(\hat{\mathbf{c}}, \hat{\boldsymbol{\kappa}}) = \underset{(\mathbf{c},\boldsymbol{\kappa}) \in \mathbb{R}^{r+p} \times \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \|\mathbf{x}_i - B(\boldsymbol{\kappa})\mathbf{c}\|^2. \tag{2}$$

The estimator of $\mu(t)$ for t outside the input grid will be $\hat{\mu}(t) = \boldsymbol{\beta}(t, \hat{\boldsymbol{\kappa}})^\top \hat{\mathbf{c}}$.

Note that for a given $\boldsymbol{\kappa}$, the optimal coefficient vector is $\hat{\mathbf{c}}(\boldsymbol{\kappa}) = \{B(\boldsymbol{\kappa})^\top B(\boldsymbol{\kappa})\}^{-1} B(\boldsymbol{\kappa})^\top \bar{\mathbf{x}}$, so

$$\hat{\boldsymbol{\kappa}} = \underset{\boldsymbol{\kappa} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \| \mathbf{x}_i - B(\boldsymbol{\kappa}) \{B(\boldsymbol{\kappa})^\top B(\boldsymbol{\kappa})\}^{-1} B(\boldsymbol{\kappa})^\top \bar{\mathbf{x}} \|^2. \tag{3}$$

This is a nonlinear minimisation problem. Finding the global minimiser (3) is not trivial, because the objective function may have several local minima (Jupp, 1978). Naive minimisation strategies, such as Newton–Raphson iterations starting from equispaced knots, will have trouble finding the global minimiser, or even a good local minimiser. More sophisticated procedures are discussed, for instance, in Hansen and Kooperberg (2002).

We propose the following stepwise knot-addition algorithm:

1. Initialization. Define a grid $\mathcal{T}_1 = \{s_1, \ldots, s_N\}$ in $[a, b]$, and let $J(\mathcal{T}_1)$ be the set of Jupp-transformed values of each $s \in \mathcal{T}_1$. Then:

   (a) Find the minimiser (3) among those $\kappa \in J(\mathcal{T}_1)$. Call it $\tilde{\kappa}_1$.

   (b) Use $\tilde{\kappa}_1$ as the starting point of a Gauss–Newton algorithm to find the (unrestricted) minimiser (3). Call it $\hat{\kappa}_1$ and define $\hat{\tau}_1 = J^{-1}(\hat{\kappa}_1)$.

2. Forward addition. Repeat for $k = 2, \ldots, p$:

   (a) Let $\mathcal{T}_k = \{(\hat{\boldsymbol{\tau}}_{k-1}; s) : s \in \mathcal{T}_1\} \subset \mathbb{R}^k$, where $(\hat{\boldsymbol{\tau}}_{k-1}; s)$ means that s is inserted among the elements of $\hat{\boldsymbol{\tau}}_{k-1}$ so as to form an increasing sequence. Let $J(\mathcal{T}_k)$ be the corresponding set of Jupp-transformed knots and let $\tilde{\boldsymbol{\kappa}}_k$ be the minimiser (3) restricted to $\boldsymbol{\kappa} \in J(\mathcal{T}_k)$.

   (b) Use $\tilde{\boldsymbol{\kappa}}_k$ as the starting point of a Gauss–Newton algorithm to find the unrestricted minimiser (3). Call it $\hat{\boldsymbol{\kappa}}_k$ and define $\hat{\boldsymbol{\tau}}_k = J^{-1}(\hat{\boldsymbol{\kappa}}_k)$.

This algorithm produces knot sequences $\{\hat{\boldsymbol{\tau}}_k\}$ and $\{\hat{\boldsymbol{\kappa}}_k\}$ of increasing dimensions, until $\hat{\boldsymbol{\kappa}}_p \in \mathbb{R}^p$ is reached. Therefore, the computing time is essentially equivalent to p runs of a Gauss–Newton algorithm. As $\mathcal{T}_1$ we take a grid of 20 equispaced points in $[a, b]$, so the computing time of steps 1a and 2a is negligible.

Although there is no guarantee that this (or any other) algorithm will find the global minimiser (3), we have found that it works well in practice. In our simulations and examples, the knots were added in the "right" order, in the sense that the knots associated with the most salient features of $\mu(t)$ are added first. This is important for model selection, since the optimal number of knots p is never known in practice and will be chosen on the basis of intermediate knot sequences.

For regression splines, p determines the bias/variance trade-off. Selection of p is a problem that has been extensively discussed in the literature (see e.g. Hastie, Tibshirani and Friedman (2001)). The most common approach is to find the p that minimises an estimate of the mean average squared error, $\mathrm{MASE}(\hat{\boldsymbol{\mu}}) = \mathrm{E}(\|\mathbf{x} - \hat{\boldsymbol{\mu}}\|^2)/m$. Usually $\mathrm{MASE}(\hat{\boldsymbol{\mu}})$ is estimated by cross validation, as in Rice and Silverman (1991) and Ramsay and Silverman (1997). However, cross-validating our estimator would be very time consuming, so we use instead the generalised cross-validation criterion of Craven and Wahba (1979). This criterion is defined as $\mathrm{GCV}(\hat{\boldsymbol{\mu}}) = \mathrm{ASE}(\hat{\boldsymbol{\mu}})/(1 - d/n)^2$, where $\mathrm{ASE}(\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} \|\mathbf{x}_i - \hat{\boldsymbol{\mu}}\|^2/mn$ and d is the effective number of parameters, $d = 2p + r$ in our case. Our proposal for estimating $\boldsymbol{\mu}$, then, is to use the above algorithm with a tentative (relatively large) p^*, computing the GCV criterion for all intermediate knot vectors $\{\hat{\boldsymbol{\kappa}}_k\}$ and choosing the $\hat{\boldsymbol{\kappa}}_p$ that minimises the GCV. This method worked very well in our simulations (see Section 4).

Although free-knot spline estimation is computationally more complex than smoothing spline estimation, many programming languages offer fast and efficient minimisation routines that can be used in steps 1b and 2b of the algorithm (we use the Matlab routine fminunc). This makes the whole estimation process, including selection of p, very fast. For instance, estimating the mean for each simulated sample in Section 4 took about 20 seconds on a Pentium 4 personal computer with a 2.2 GHz CPU.

2.2 Estimation of the covariance and its components

A stochastic process $X \in L^2[a,b]$ admits a essentially unique Karhunen–Loève decomposition

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} Z_k \psi_k(t), \tag{4}$$

where $\{Z_k\}$ are uncorrelated random variables with zero mean and variance $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$, and $\{\psi_k\}$ is an orthonormal basis of $L^2[a,b]$. Therefore the covariance function can be written as $\rho(s,t) = \sum_{k=1}^{\infty} \lambda_k \psi_k(s)\psi_k(t)$.

Usually the λ_ks decay to zero very rapidly, so only a few terms in (4) are significant. This motivates the following model for the terms $\boldsymbol{\eta}_i$ of model (1):

$$\boldsymbol{\eta}_i = \sum_{k=1}^{q} Z_{ik} \boldsymbol{\phi}_k + \boldsymbol{\varepsilon}_i, \tag{5}$$

$$\text{where } \{Z_{ik}\} \text{ are uncorrelated, } \mathrm{E}(Z_{ik}) = 0, \mathrm{V}(Z_{ik}) = \lambda_k,$$
$$\text{and } \{\varepsilon_{ij}\} \text{ are i.i.d., } \mathrm{E}(\varepsilon_{ij}) = 0, \mathrm{V}(\varepsilon_{ij}) = \sigma^2.$$

For identifiability (up to the sign of the $\boldsymbol{\phi}_k$s) we assume $\lambda_1 \geq \ldots \geq \lambda_q$ and $\boldsymbol{\phi}_k^\top \boldsymbol{\phi}_l / m = \delta_{kl}$. Model (5) implies that

$$\Sigma := \mathrm{E}(\boldsymbol{\eta}_i \boldsymbol{\eta}_i^\top) = \sum_{k=1}^{q} \lambda_k \boldsymbol{\phi}_k \boldsymbol{\phi}_k^\top + \sigma^2 I_m,$$

and then $\boldsymbol{\phi}_k / \sqrt{m}$ is an eigenvector of Σ with eigenvalue $\lambda_k m + \sigma^2$. In model (5) we are implicitly identifying ϕ_{kj} with $\psi_k(t_j)$, although this is only an approximation, since the m-dimensional discretisations of the ψ_ks are not exactly orthogonal in $\mathbb{R}^m$ (see Kneip (1994) for precise conditions under which $\sum_{j=1}^{m} \{\phi_{kj} - \psi_k(t_j)\}^2 / m$ converges to zero as m goes to infinity).

To estimate the variance components $\{\boldsymbol{\phi}_k\}$ we use free-knot splines again. Let $V_n = \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top / n$ be the model-free estimator of Σ, where $\hat{\boldsymbol{\mu}}$ is the estimator proposed in Section 2.1. Since $\boldsymbol{\phi}_1 / \sqrt{m}$ is the eigenvector of Σ associated with the largest eigenvalue, we define $\hat{\boldsymbol{\phi}}_1 = B(\hat{\boldsymbol{\kappa}})\hat{\mathbf{c}}$, where

$$(\hat{\mathbf{c}}, \hat{\boldsymbol{\kappa}}) = \underset{(\mathbf{c},\boldsymbol{\kappa})\in\mathbb{R}^{r+p}\times\mathbb{R}^p}{\operatorname{argmax}} \mathbf{c}^\top B(\boldsymbol{\kappa})^\top V_n B(\boldsymbol{\kappa})\mathbf{c}, \qquad (6)$$
$$\text{subject to } \frac{1}{m}\mathbf{c}^\top B(\boldsymbol{\kappa})^\top B(\boldsymbol{\kappa})\mathbf{c} = 1.$$

Again, it is convenient to solve (6) in two steps: first, find the optimal $\hat{\mathbf{c}}(\boldsymbol{\kappa})$ for a given knot sequence $\boldsymbol{\kappa}$, which is a common eigenvalue problem; then optimise $\boldsymbol{\kappa}$, which is a more difficult nonlinear problem. We use a stepwise knot-addition algorithm similar to the one proposed in Section 2.1; the necessary modifications are obvious, so we omit the details.

To select the optimal number of knots p for $\hat{\boldsymbol{\phi}}_1$ (which, in general, will not be the same p as for $\hat{\boldsymbol{\mu}}$) we use again a generalised cross-validation criterion. Let $s_{i1} = (\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\phi}}_1/m$. Since

$$\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i - \hat{\boldsymbol{\mu}} - s_{i1}\hat{\boldsymbol{\phi}}_1\|^2 = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - \hat{\boldsymbol{\mu}}\|^2 - \frac{1}{m}\hat{\mathbf{c}}^\top B(\hat{\boldsymbol{\kappa}})^\top V_n B(\hat{\boldsymbol{\kappa}})\hat{\mathbf{c}}, \qquad (7)$$

the maximiser (6) is the minimiser of the left-hand side of (7). Therefore, in analogy with $\mathrm{GCV}(\hat{\boldsymbol{\mu}})$, we define

$$\mathrm{GCV}(\hat{\boldsymbol{\phi}}_1) = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - \hat{\boldsymbol{\mu}} - s_{i1}\hat{\boldsymbol{\phi}}_1\|^2/(1-\frac{d}{n})^2$$

with $d = 2p + r$, and choose the p that minimises this criterion. We found in our simulations that $\mathrm{GCV}(\hat{\boldsymbol{\phi}}_1)$ selects a reasonable p in the great majority of cases.

The rest of the components of model (5) can be estimated sequentially. Given

$\hat{\boldsymbol{\phi}}_1, \ldots, \hat{\boldsymbol{\phi}}_{k-1}$, $k \geq 2$, we define $\hat{\boldsymbol{\phi}}_k = B(\hat{\boldsymbol{\kappa}})\hat{\mathbf{c}}$, where

$$\begin{gathered}(\hat{\mathbf{c}}, \hat{\boldsymbol{\kappa}}) = \underset{(\mathbf{c},\boldsymbol{\kappa}) \in \mathbb{R}^{r+p} \times \mathbb{R}^p}{\operatorname{argmax}} \mathbf{c}^\top B(\boldsymbol{\kappa})^\top V_n B(\boldsymbol{\kappa})\mathbf{c}, \\ \text{subject to } \frac{1}{m}\mathbf{c}^\top B(\boldsymbol{\kappa})^\top B(\boldsymbol{\kappa})\mathbf{c} = 1 \\ \text{and } \hat{\boldsymbol{\phi}}_j^\top B(\boldsymbol{\kappa})\mathbf{c} = 0 \text{ for all } j = 1, \ldots, k-1.\end{gathered} \tag{8}$$

The optimal number of knots p (which, in general, will be different for each component) can be chosen by generalised cross-validation as before.

Selection of the number of components q is also an important problem, but for reasons of space we cannot treat it in much detail here. We mention that q can be chosen either by some form of cross-validation or by sequential testing as in Kneip (1994). A less formal but more practical approach is to compute several components and keep those that explain a large proportion of the accumulated variance. This method works well when a few leading components explain most of the variance, which is often the case in practice.

To estimate the eigenvalues $\{\lambda_k\}$ and the error variance σ^2 we proceed as follows. Let $\hat{\xi}_k = \hat{\boldsymbol{\phi}}_k^\top V_n \hat{\boldsymbol{\phi}}_k / m$. Since $\boldsymbol{\phi}_k/\sqrt{m}$ is an eigenvector of Σ with eigenvalue $\lambda_k m + \sigma^2$, and $\mathrm{tr}(\Sigma) = (\sum_{k=1}^q \lambda_k + \sigma^2)m$, we find $\{\hat{\lambda}_k\}$ and $\hat{\sigma}^2$ by solving the equations

$$\begin{aligned}\hat{\xi}_k &= \hat{\lambda}_k m + \hat{\sigma}^2,\ k = 1, \ldots, q, \\ \mathrm{tr}(V_n) &= (\sum_{k=1}^q \hat{\lambda}_k + \hat{\sigma}^2)m.\end{aligned}$$

Explicitly,

$$\begin{aligned}\hat{\sigma}^2 &= \frac{m\ \mathrm{tr}(V_n) - \sum_{k=1}^q \hat{\xi}_k}{m - q}, \\ \hat{\lambda}_k &= \frac{\hat{\xi}_k}{m} - \hat{\sigma}^2,\ k = 1, \ldots, q.\end{aligned}$$

We then obtain a model-based estimator of Σ,

$$\hat{\Sigma} = \sum_{k=1}^{q} \hat{\lambda}_k \hat{\boldsymbol{\phi}}_k \hat{\boldsymbol{\phi}}_k^\top + \hat{\sigma}^2 I_m, \tag{9}$$

that will be used later to construct confidence intervals for $\mu(t)$.

In many applications it is also of interest to estimate the covariance function $\rho(s,t)$ and the individual curves $X_i(t)$. To estimate the covariance function we can use

$$\hat{\rho}(s,t) = \sum_{k=1}^{q} \hat{\lambda}_k \hat{\phi}_k(s) \hat{\phi}_k(t),$$

where $\hat{\phi}_k(t) = \boldsymbol{\beta}(t, \hat{\boldsymbol{\kappa}})^\intercal \hat{\mathbf{c}}$ and $(\hat{\mathbf{c}}, \hat{\boldsymbol{\kappa}})$ is given by (6) or (8). Individual smoothers of the sample curves are given by

$$\hat{X}_i(t) = \hat{\mu}(t) + \sum_{k=1}^{q} s_{ik} \hat{\phi}_k(t),$$

where $s_{ik} = (\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\phi}}_k / m$. These are the curves shown in Fig. 1b.

3 Asymptotics and inference

In this section we derive the asymptotic distribution of the free-knot spline estimator of $\mu(t)$ when the number of curves n goes to infinity, and construct confidence bands for μ. The number of knots p and the number of observations per curve m are assumed to be fixed.

For technical reasons, we restrict minimisation (3) to $\boldsymbol{\kappa}$ in a compact set $K \subset \mathbb{R}^p$. This is equivalent to requiring that, for some $\delta > 0$, $\boldsymbol{\tau}$ satisfies $\tau_i - \tau_{i-1} \geq \delta$ for all $i = 1, \ldots, p+1$. (From a practical point of view this restriction is not problematic, and in fact we suggest to implement it in the algorithm of Section 2.1 to avoid knot coalescence). We also assume $r \geq 3$, so B-splines are twice differentiable with respect to $\boldsymbol{\kappa}$. Finally, to simplify notation, we define $\boldsymbol{\theta} = (\mathbf{c}, \boldsymbol{\kappa})$.

Theorem 1 *Let* $P(\boldsymbol{\kappa}) = B(\boldsymbol{\kappa})\{B(\boldsymbol{\kappa})^\top B(\boldsymbol{\kappa})\}^{-1}B(\boldsymbol{\kappa})^\top$ *and* $\boldsymbol{\kappa}_0 = \operatorname{argmin}_{\boldsymbol{\kappa}\in K} ||\boldsymbol{\mu} - P(\boldsymbol{\kappa})\boldsymbol{\mu}||^2$. *If* $\boldsymbol{\kappa}_0$ *is unique, under the assumptions of the previous paragraph we have:*

1. $\hat{\boldsymbol{\kappa}} \xrightarrow{P} \boldsymbol{\kappa}_0$ *and* $\hat{\mathbf{c}} \xrightarrow{P} \mathbf{c}_0 := \{B(\boldsymbol{\kappa}_0)^\top B(\boldsymbol{\kappa}_0)\}^{-1}B(\boldsymbol{\kappa}_0)^\top\boldsymbol{\mu}$.

2. $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} \mathcal{N}(0, H_0^{-1}D_0H_0^{-1})$, *where* $D_0 = E\{\mathsf{D}f(\mathbf{x};\boldsymbol{\theta}_0)^\top\mathsf{D}f(\mathbf{x};\boldsymbol{\theta}_0)\}$, $H_0 = E\{\mathsf{D}^2f(\mathbf{x};\boldsymbol{\theta}_0)\}$, $f(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{x} - B(\boldsymbol{\kappa})\mathbf{c}\|^2$ *and* D *denotes the differential with respect to* $\boldsymbol{\theta}$.

3. Given $t \in [a,b]$, *let* $g_t(\boldsymbol{\theta}) = \boldsymbol{\beta}(t,\boldsymbol{\kappa})^\top\mathbf{c}$ *and* $\tilde{\mu}_0(t) = g_t(\boldsymbol{\theta}_0)$. *Then*

$$\sqrt{n}(\hat{\mu}(t) - \tilde{\mu}_0(t)) \xrightarrow{D} \mathcal{N}(0, \mathsf{D}g_t(\boldsymbol{\theta}_0)H_0^{-1}D_0H_0^{-1}\mathsf{D}g_t(\boldsymbol{\theta}_0)^\top).$$

The proof of Theorem 1, together with explicit expressions for D_0, H_0 and $\mathsf{D}g_t(\boldsymbol{\theta})$, are given in the Appendix.

Theorem 1 does not assume $\mu \in \mathcal{S}_{r,p}$ but requires uniqueness of the knots that provide the best spline approximation. Unfortunately, it is not easy to find a set of verifiable sufficient conditions for uniqueness of $\boldsymbol{\kappa}_0$. Nevertheless, the results in Theorem 1 are still valid if a consistent rule for selection among different global minimisers is implemented, such as choosing the $\hat{\boldsymbol{\kappa}}$ closest to zero (which would be equivalent to choosing the knot sequence $\hat{\boldsymbol{\tau}}$ closest to being equispaced).

When $\mu \in \mathcal{S}_{r,p}$, part 3 of Theorem 1 can be used to construct asymptotic confidence intervals for $\mu(t)$. It is shown in the Appendix that $D_0 = M_0^\top\Sigma M_0$ and $H_0 = M_0^\top M_0$, with $M_0 = M(\boldsymbol{\theta}_0)$ given by (11). Since $M(\boldsymbol{\theta})$ and $\mathsf{D}g_t(\boldsymbol{\theta})$ are continuous functions of $\boldsymbol{\theta}$, M_0 and $\mathsf{D}g_t(\boldsymbol{\theta}_0)$ are consistently estimated by $\hat{M}_0 = M(\hat{\boldsymbol{\theta}})$ and $\mathsf{D}g_t(\hat{\boldsymbol{\theta}})$, respectively. D_0 and H_0 are estimated by $\hat{D}_0 = \hat{M}_0^\top\hat{\Sigma}\hat{M}_0$ and $\hat{H}_0 = \hat{M}_0^\top\hat{M}_0$, respectively, with $\hat{\Sigma}$ given by (9). Theorem 2 gives conditions for consistency of $\hat{\Sigma}$.

Theorem 2 *If the components of model (5) satisfy* $\boldsymbol{\phi}_k = B(\boldsymbol{\kappa}_k)\mathbf{c}_k$ *for* $k = 1,\ldots,q$ *(where the* $\boldsymbol{\kappa}_k$*s do not have to be unique),* $\lambda_1 > \lambda_2 > \ldots > \lambda_q > 0$ *and* $\hat{\boldsymbol{\mu}} \xrightarrow{P} \boldsymbol{\mu}$, *then* $\|\hat{\boldsymbol{\phi}}_k\hat{\boldsymbol{\phi}}_k^\top - \boldsymbol{\phi}_k\boldsymbol{\phi}_k^\top\| \xrightarrow{P} 0$ *for* $k = 1,\ldots,q$ *(where* $\|\cdot\|$ *is the Frobenius norm). Moreover,* $\hat{\sigma}^2 \xrightarrow{P} \sigma^2$ *and* $\hat{\lambda}_k \xrightarrow{P} \lambda_k$ *for* $k = 1,\ldots,q$, *so* $\|\hat{\Sigma} - \Sigma\| \xrightarrow{P} 0$.

A confidence interval for $\mu(t)$ with asymptotic coverage $1-\alpha$ (for each t) is therefore given by

$$\hat{\mu}(t) \pm z_{\alpha/2}\{\mathsf{D}g_t(\hat{\boldsymbol{\theta}})\hat{H}_0^{-1}\hat{D}_0\hat{H}_0^{-1}\mathsf{D}g_t(\hat{\boldsymbol{\theta}})^\top\}^{\frac{1}{2}}/\sqrt{n}. \tag{10}$$

To construct a confidence band with simultaneous coverage probability $1-\alpha$ for all t in a given grid $\mathcal{G}$, we proceed as follows. Let $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, I_{2p+r})$, $W = \max_{t\in\mathcal{G}} |n^{-\frac{1}{2}}\mathsf{D}g_t(\hat{\boldsymbol{\theta}})\hat{H}_0^{-1}\hat{D}_0^{\frac{1}{2}}\mathbf{Z}|$ and w_α the $(1-\alpha)$-percentile of W (which can be easily estimated by simulation). Then the intervals $\hat{\mu}(t) \pm w_\alpha$ have simultaneous asymptotic coverage $1-\alpha$ for all $t \in \mathcal{G}$. Simultaneous confidence intervals are broader than the pointwise confidence intervals (10), but we have found in our simulations that for moderately large samples they can be narrow enough to be informative.

4 Simulations

The simulation study reported in this section had three goals: *(i)* to compare estimation errors of free-knot splines and smoothing splines, *(ii)* to assess the performance of generalised cross-validation as model-selection criterion, and *(iii)* to evaluate the finite sample coverage of the confidence intervals.

Since we mainly want to assess the adaptivity of free-knot splines, we chose three functions $\mu(t)$ with different degrees of smoothness (Fig. 2):

Model 1. A spline function of order four (cubic) in $[0,1]$ with knots $\boldsymbol{\tau} = [0.4, 0.6]$ and coefficients $\mathbf{c} = (0,1,0,1,0,0)$. This function is differentiable everywhere.

Model 2. A spline function of order four with knots $\boldsymbol{\tau} = [0.4, 0.6, 0.6, 0.6]$ and coefficients $\mathbf{c} = (0,1,0,0,1,0,0,0)$. This function is continuous but not differentiable at $t = 0.6$.

Model 3. The so-called "Doppler function" $\mu(t) = \{t(1-t)\}^{1/2}\sin\{2\pi(1+2^{(9-4k)/5})/(t+2^{(9-4k)/5})\}$, with $k = 5$. This function is differentiable everywhere but very variable and difficult to estimate, specially near the origin.

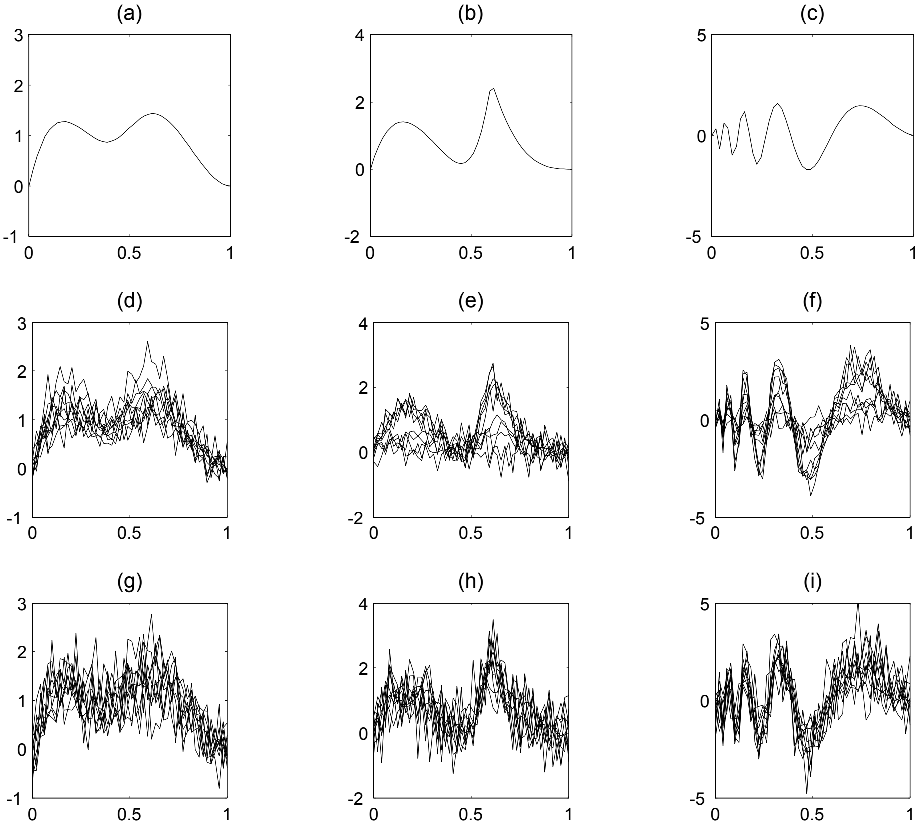

Figure 2: Model functions (a,b,c) and ten simulated curves for each model, with $\rho = 4$ (d,e,f) and $\rho = 1/4$ (g,h,i). Plots (a,d,g) correspond to Model 1, (b,e,h) to Model 2 and (c,f,i) to Model 3.

As input grid $\{t_j\}$ we took $m = 50$ equidistant points in $[0,1]$, and the functions were scaled so that $\|\boldsymbol{\mu}\|^2/m = 1$. As for the variance component model (5), we chose a simple but non-trivial model with only one component $\boldsymbol{\phi}_1 = \boldsymbol{\mu}$, with $Z_1 \sim \mathcal{N}(0, \lambda_1)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. We considered two variance ratios $\rho = \lambda_1/\sigma^2$, namely $\rho = 4$ and $\rho = 1/4$. The signal-to-noise ratio $m^{-1}\sum_{j=1}^{m}(\mu_j - \bar{\mu})^2/(\lambda_1 + \sigma^2)$ was set to 1 in both cases. The sample size was $n = 100$, and each sampling situation was replicated 500 times. Some of these curves are shown in Fig. 2.

The following estimators were computed for each sample:

- Raw mean and eigenvectors.

- Free-knot splines of order four, with p selected by generalised cross-validation among $p \leq p^*$, with $p^* = 10$, 15 and 20 for models 1, 2 and 3, respectively. The grid $\mathcal{T}_1$ for knot addition consisted of 20 equispaced points. For comparison, we also computed the "oracle" estimator corresponding to the p that minimises the root average squared error RASE $= \{\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2/m\}^{\frac{1}{2}}$ (this allows us to evaluate the GCV criterion, as no model-selection criterion can outperform the oracle estimator).

- Smoothing splines, as proposed by Rice and Silverman (1991) for the mean and by Silverman (1996) for the variance components (see also Ramsay and Silverman (1997)). We used cubic splines with knots $\{t_j\}$ and the squared norm of the second derivative as roughness penalty. The penalty parameter was chosen by generalised cross-validation among $\{.1^{3+.25(k-1)} : k = 1, \ldots, 29\}$ (we ran preliminary simulations to make sure this range was broad enough). We also computed the oracle estimator corresponding to the smoothing parameter that minimises the RASE.

Boxplots of simulated RASEs are shown in Fig. 3 (for the mean) and Fig. 4 (for the variance component). The indentations in the boxes are confidence intervals for the median RASE, what gives an idea of the Monte Carlo variability and the statistical significance of the results. We see that free-knot splines outperform smoothing splines

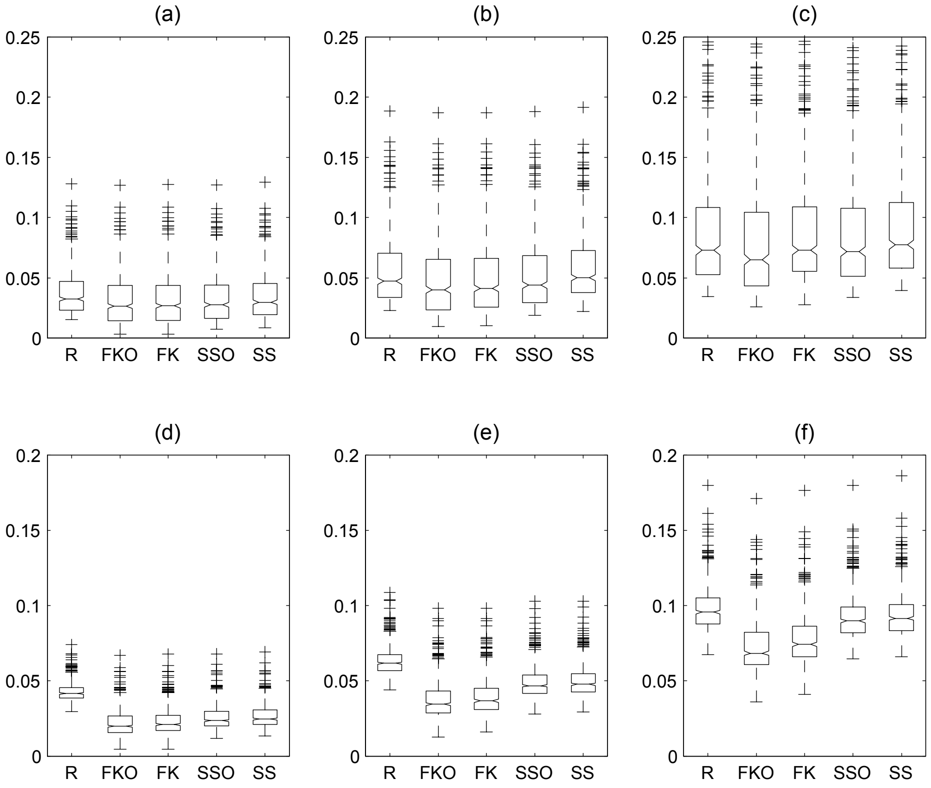

Figure 3: Simulated root average squared errors of estimators of $\boldsymbol{\mu}$ under Model 1 (a,d), Model 2 (b,e) and Model 3 (c,f), with $\rho = 4$ (a,b,c) and $\rho = 1/4$ (d,e,f). “R” is the raw mean, “FKO” the oracle free-knot spline, “FK” the GCV free-knot spline, “SSO” the oracle smoothing spline, and “SS” the GCV smoothing spline.

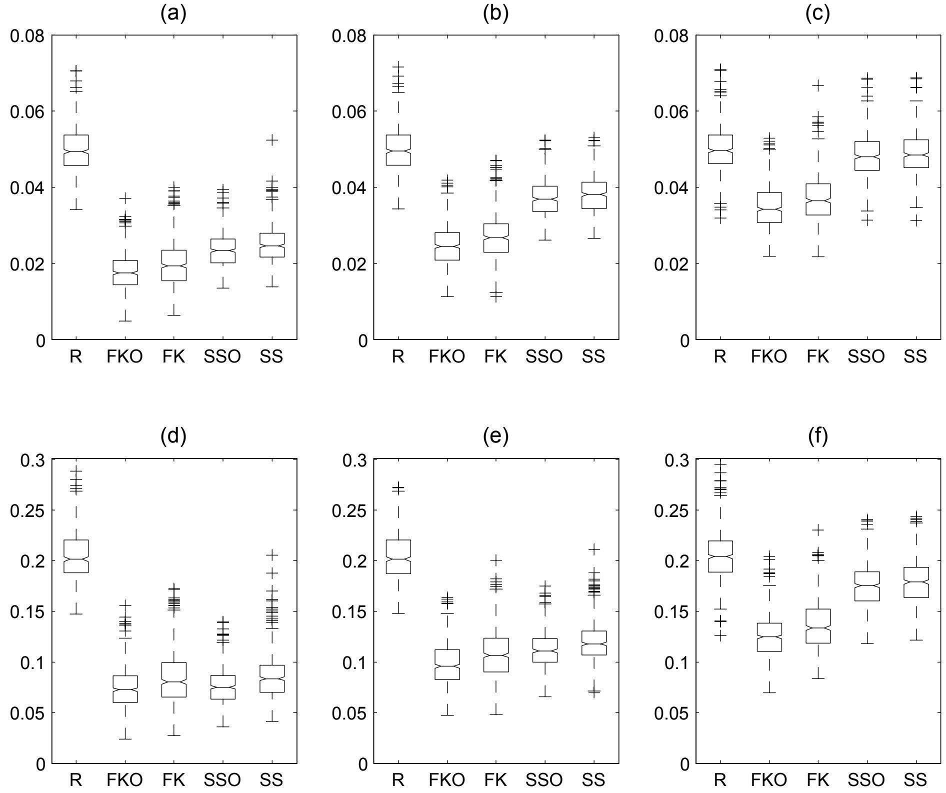

Figure 4: Simulated root average squared errors of estimators of ϕ_1. Subplots and labels are the same as in Fig. 3.

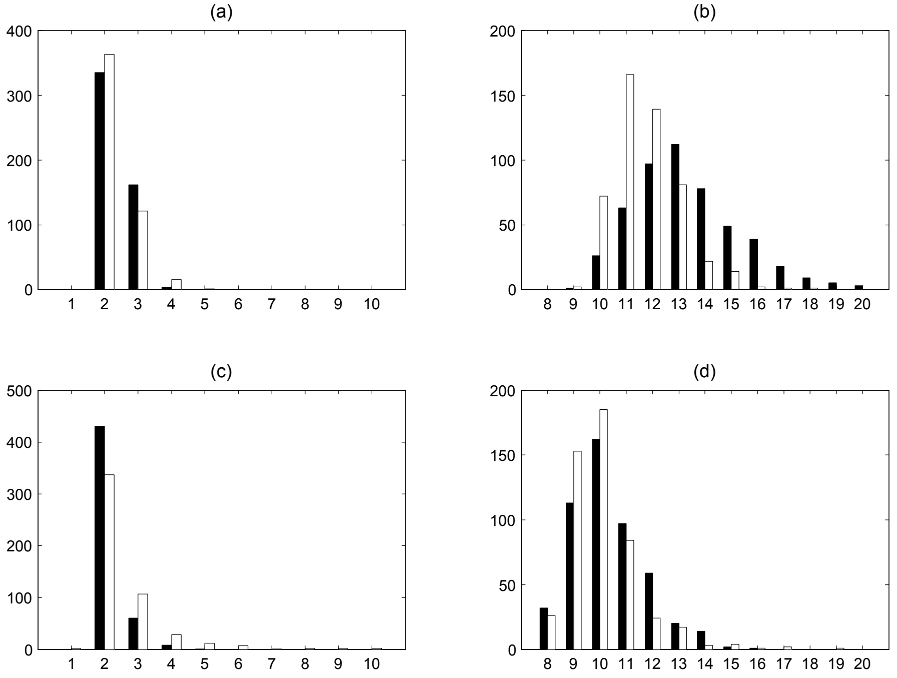

Figure 5: Frequency plots of the number of knots selected by the oracle estimator (black bars) and by GCV (white bars). Estimators of $\boldsymbol{\mu}$ (a,b) and $\boldsymbol{\phi}_1$ (c,d) for Model 1 (a,c) and Model 3 (b,d) with $\rho = 1/4$.

under all of the simulated models, the difference being most remarkable for Model 3 and $\rho = 1/4$.

The cross-validated estimators are very close to their "oracle" counterparts, so GCV is a good criterion for selecting the number of knots. For better insight, Fig. 5 shows bar plots of the number of knots p selected by the oracle estimator and by GCV for models 1 and 3 with $\rho = 1/4$. We see that GCV tends to be conservative, often selecting fewer knots than the oracle estimator, which is actually a good thing in this context.

The RASE is a global error measure across t. To assess the local adaptivity of free-knot splines, it is more instructive to plot bias, standard deviation and root mean

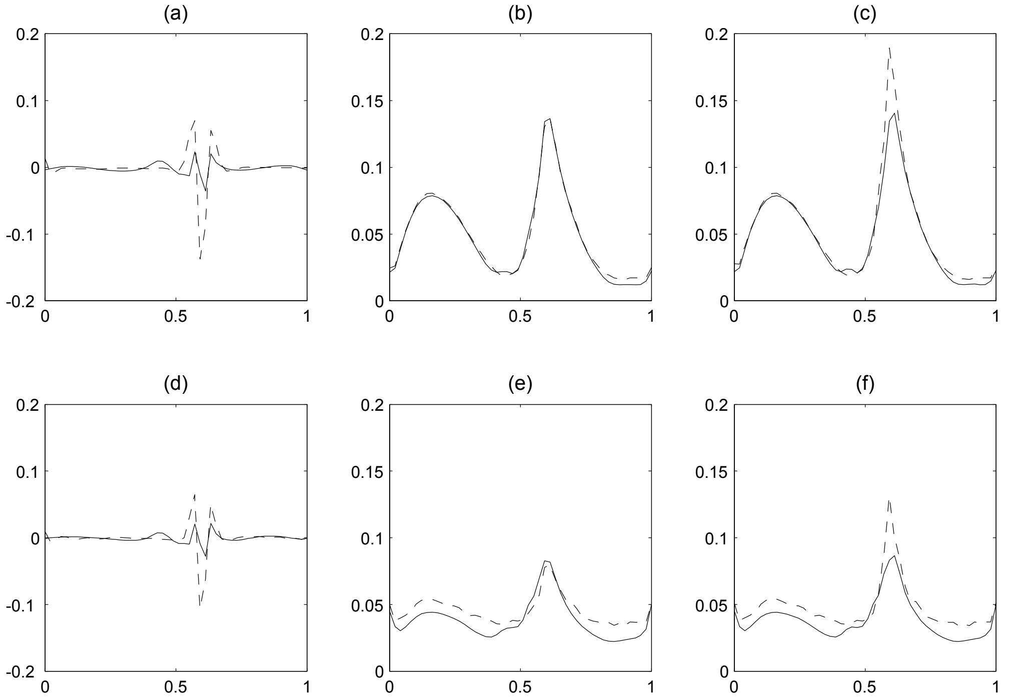

Figure 6: Simulated bias (a,d), standard deviation (b,e) and root mean squared error (c,f) of free-knot spline (solid line) and smoothing spline (dotted line) estimators of $\mu(t)$ under Model 2, with $\rho = 4$ (a,b,c) and $\rho = 1/4$ (d,e,f).

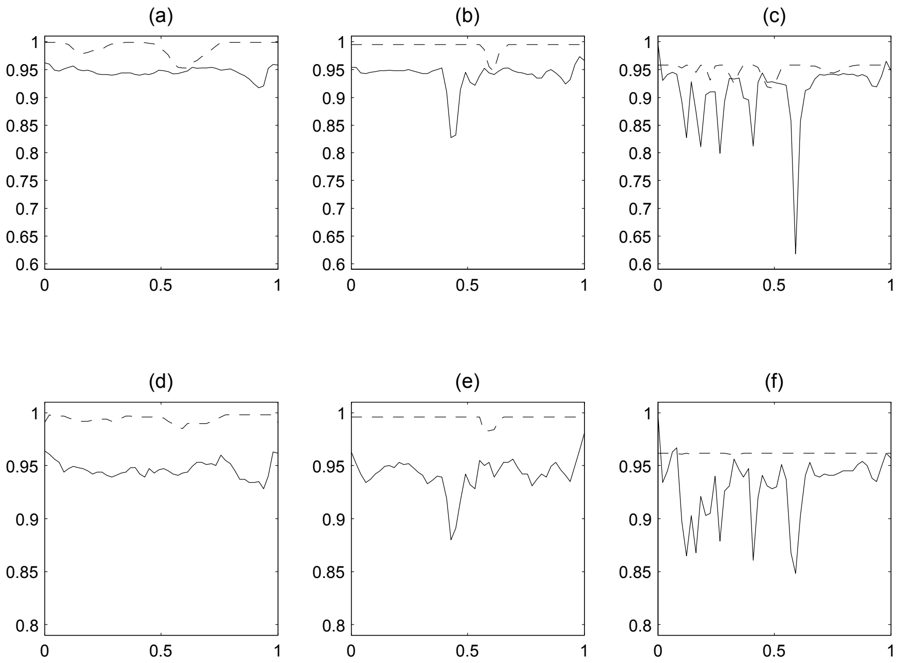

Figure 7: Simulated coverage probabilities of pointwise confidence bands (solid line) and simultaneous confidence bands (dashed line) for μ under Model 1 (a,d), Model 2 (b,e) and Model 3 (c,f), with $\rho = 4$ (a,b,c) and $\rho = 1/4$ (d,e,f).

squared error of the estimators as functions of t. For reasons of space we cannot do this for all of the simulated models, but as an example we show the results for Model 2 in Fig. 6. As expected, the lower RASE of the free-knot spline estimator is mainly due to the low bias at the peak, although the variance is smaller as well, especially for the high-noise case $\rho = 1/4$.

Now let us turn our attention to the confidence intervals. To assess their finite-sample coverage, we simulated 1000 samples from each model and computed 95% confidence bands. Fig. 7 shows empirical coverage probabilities as functions of t. Pointwise confidence intervals show coverage probabilities very close to the nominal ones for Models 1 and 2. The behavior deteriorates for Model 3, although the cov-

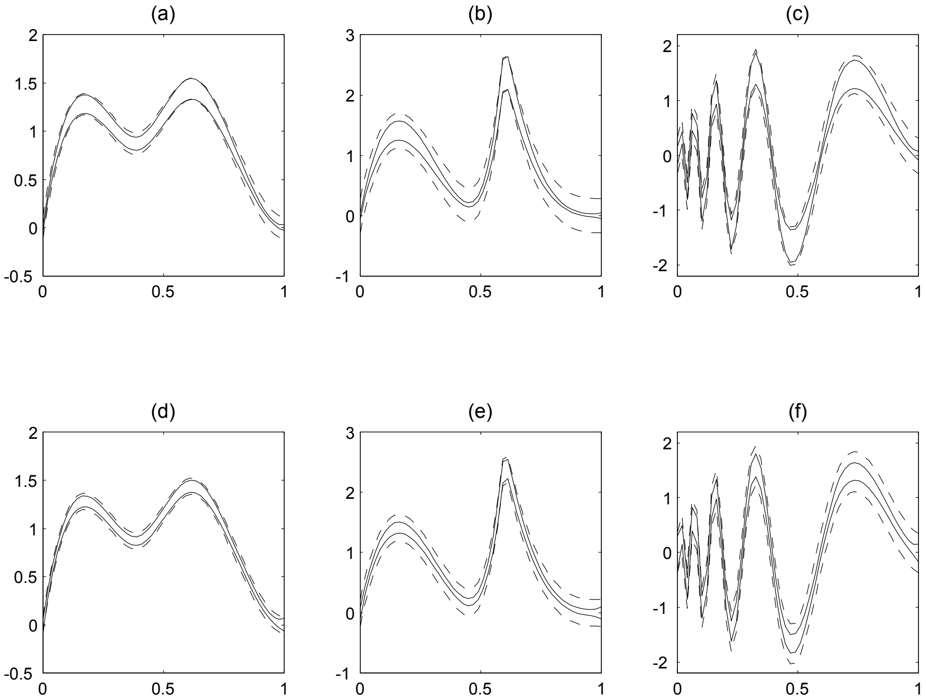

Figure 8: Median length pointwise confidence bands (solid line) and simultaneous confidence bands (dashed line) for μ. Subplots are as in Fig. 7.

erage probabilities remains over 85% for most values of t. Simultaneous confidence bands have coverage probabilities higher than 95% for most ts, as expected by construction, yet they are reasonably narrow. This is seen in Fig. 8, that shows median length intervals. Median length intervals are defined as $\tilde{\mu}(t) \pm 0.5l(t)$, where $\tilde{\mu}(t)$ is the median of the simulated $\hat{\mu}(t)$s and $l(t)$ is the median length of the simulated confidence intervals (these are not proper confidence intervals, but give a good idea of the average width of the actual intervals). Overall, we can say that both pointwise and simultaneous confidence intervals are accurate enough for statistical inference, at least for samples of this size.

5 Example: Daily precipitation in Canadian cities

Fig. 1a shows logarithms of daily precipitation (averaged over the years 1960 to 1994) for two Canadian cities, which are part of a larger dataset of 35 cities and is available for download on James Ramsay's website. This dataset has relatively few curves ($n = 35$) and many observations per curve ($m = 365$), and is extremely noisy, even after taking logarithms. This puts to test the ability of the estimators to detect important local features of the mean and variance components without undersmoothing.

We estimated the mean and variance components using free-knot splines and smoothing splines of order four (the details of implementation are the same as in Section 4). We added the restrictions $\mu(a) = \mu(b)$ and $\mu'(a) = \mu'(b)$ (and the same for the variance components), which are necessary for these cyclical data.

For the free-knot spline estimator of μ, generalised cross-validation chooses 5 knots. This represents a total of 14 parameters, although the two restrictions reduce the effective degrees of freedom to 12. For the smoothing spline estimator, GCV chooses a model with only 9.65 degrees of freedom. Fig. 9a shows both spline estimators and Fig. 9b the raw mean. The raw mean is so noisy that little information can be extracted from it, other than the fact that it rains more in summer than in winter; a slight bimodality is perceptible in the summer months, but given the

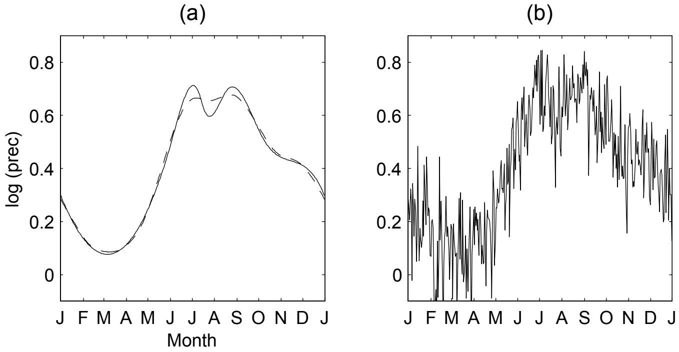

Figure 9: Mean daily precipitation of 35 Canadian cities, estimated by free-knot splines (a, solid), smoothing splines (a, dashed) and raw mean (b).

high variability, it would be hard to tell whether this is statistically significant or not. The free-knot spline estimator clearly shows that the average precipitation is bimodal during the summer months. The smoothing spline estimator, in contrast, misses this local feature.

The first four variance components were also estimated. The corresponding eigenvalues, for free-knot spline estimators, are $\hat{\lambda}_1 = .574$, $\hat{\lambda}_2 = .058$, $\hat{\lambda}_3 = .014$ and $\hat{\lambda}_4 = .006$. The contribution of $\hat{\boldsymbol{\phi}}_4$ is so small given the first four components (just 1%) that it does not seem necessary to compute higher order components. The first three components are shown in Fig. 10. Overall, both methods produce similar estimates (the average squared error is .2093 for free-knot splines and .2086 for smoothing splines), but qualitatively they are rather different. Free-knot splines exhibit more clear-cut peaks and troughs and are otherwise smooth, while smoothing splines show flatter local features but tend to be more wiggly over the whole range. This behavior is reflected in the individual curve estimators as well. We see in Fig. 1 that precipitation in Vancouver dips in the month of August, and this is more sharply estimated by free-knot splines than by smoothing splines.

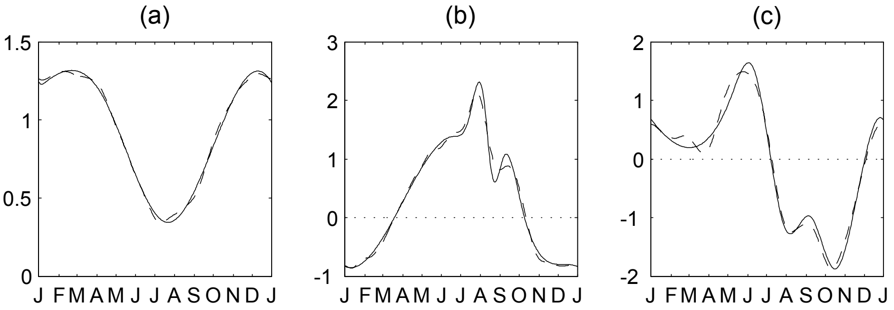

Figure 10: Estimated variance components for 35 Canadian cities, using free-knot splines (solid line) and smoothing splines (dashed line). First component (a), second component (b) and third component (c).

A Appendix: Asymptotics

Proof of Theorem 1: Given that $\boldsymbol{\kappa}$ is in a compact set K, $B(\boldsymbol{\kappa})$ and $\{B(\boldsymbol{\kappa})^\top B(\boldsymbol{\kappa})\}^{-1}$ are continuous functions of $\boldsymbol{\kappa}$, $\bar{\mathbf{x}}$ is bounded in probability, and $\hat{\mathbf{c}} = \{B(\hat{\boldsymbol{\kappa}})^\top B(\hat{\boldsymbol{\kappa}})\}^{-1} B(\hat{\boldsymbol{\kappa}})^\top \bar{\mathbf{x}}$, without loss of generality we can assume that $\mathbf{c}$ is in a compact set $C \subset \mathbb{R}^{p+r}$ and then

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in C \times K} \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}; \boldsymbol{\theta}).$$

Since the B-spline basis functions are differentiable with respect to $\boldsymbol{\kappa}$, there is an $L > 0$ such that

$$|f(\mathbf{x}; \boldsymbol{\theta}_1) - f(\mathbf{x}; \boldsymbol{\theta}_2)| \leq L \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \text{ for any } \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in C \times K.$$

It follows from Theorems 5.7 and 19.4 of van der Vaart (1998) that

$$\hat{\boldsymbol{\theta}} \xrightarrow{P} \operatorname*{argmin}_{\boldsymbol{\theta} \in C \times K} \mathrm{E}\{f(\mathbf{x}; \boldsymbol{\theta})\},$$

the minimiser being $\boldsymbol{\theta}_0 = (\mathbf{c}_0, \boldsymbol{\kappa}_0)$ with $\mathbf{c}_0$ and $\boldsymbol{\kappa}_0$ given in part 1. Part 2 follows from Theorem 5.23 of van der Vaart (1998), and part 3 is a consequence of the "delta method" (Theorem 3.1 of van der Vaart (1998)). ■

The partial derivatives of $f(\mathbf{x};\boldsymbol{\theta})$ are

$$\frac{\partial f}{\partial \kappa_k} = -(\mathbf{x} - B(\boldsymbol{\kappa})\mathbf{c})^\top \frac{\partial B(\boldsymbol{\kappa})}{\partial \kappa_k}\,\mathbf{c},$$

where matrix derivatives are understood componentwise, and

$$\mathsf{D}_{\mathbf{c}} f = -(\mathbf{x} - B(\boldsymbol{\kappa})\mathbf{c})^\top B(\boldsymbol{\kappa}).$$

Therefore $\mathsf{D}f(\mathbf{x};\boldsymbol{\theta}) = -(\mathbf{x} - B(\boldsymbol{\kappa})\mathbf{c})^\top M(\boldsymbol{\theta})$, where

$$M(\boldsymbol{\theta}) = \left[B(\boldsymbol{\kappa}), \frac{\partial B(\boldsymbol{\kappa})}{\partial \kappa_1}\,\mathbf{c}, \cdots, \frac{\partial B(\boldsymbol{\kappa})}{\partial \kappa_p}\,\mathbf{c}\right]. \tag{11}$$

Also note that $\mathsf{D}g_t(\boldsymbol{\theta}) = [\boldsymbol{\beta}(t,\boldsymbol{\kappa})^\top, \mathbf{c}^\top \mathsf{D}\boldsymbol{\beta}(t,\boldsymbol{\kappa})]^\top$.

Now let $M_0 = M(\boldsymbol{\theta}_0)$. Since $(\mathbf{c}_0, \boldsymbol{\kappa}_0)$ minimises $\|\boldsymbol{\mu} - B(\boldsymbol{\kappa})\mathbf{c}\|^2$ we have $(\boldsymbol{\mu} - B(\boldsymbol{\kappa}_0)\mathbf{c}_0)^\top M_0 = 0$, regardless of whether $\boldsymbol{\mu} = B(\boldsymbol{\kappa}_0)\mathbf{c}_0$ or not, and then $\mathrm{E}\{\mathsf{D}f(\mathbf{x};\boldsymbol{\theta}_0)\} = 0$ and $\mathrm{E}\{(\mathbf{x} - B(\boldsymbol{\kappa}_0)\mathbf{c}_0)(\mathbf{x} - B(\boldsymbol{\kappa}_0)\mathbf{c}_0)^\top\} = \Sigma$. Thus $\mathrm{E}\{\mathsf{D}f(\mathbf{x};\boldsymbol{\theta}_0)^\top \mathsf{D}f(\mathbf{x};\boldsymbol{\theta}_0)\} = M_0^\top \Sigma M_0$.

The second derivatives are

$$\begin{aligned}
\frac{\partial^2 f}{\partial \kappa_l \partial \kappa_k} &= \sum_{j=1}^m \left(\frac{\partial \boldsymbol{\beta}(t_j,\boldsymbol{\kappa})^\top}{\partial \kappa_l}\,\mathbf{c}\right)\left(\frac{\partial \boldsymbol{\beta}(t_j,\boldsymbol{\kappa})^\top}{\partial \kappa_k}\,\mathbf{c}\right) \\
&\quad - \sum_{j=1}^m (x_j - \boldsymbol{\beta}(t_j,\boldsymbol{\kappa})^\top \mathbf{c})\,\frac{\partial^2 \boldsymbol{\beta}(t_j,\boldsymbol{\kappa})^\top}{\partial \kappa_l \partial \kappa_k}\,\mathbf{c} \\
&= \mathbf{c}^\top \frac{\partial B(\boldsymbol{\kappa})^\top}{\partial \kappa_l}\frac{\partial B(\boldsymbol{\kappa})}{\partial \kappa_k}\,\mathbf{c} - (\mathbf{x} - B(\boldsymbol{\kappa})\mathbf{c})^\top \frac{\partial^2 B(\boldsymbol{\kappa})}{\partial \kappa_l \partial \kappa_k}\,\mathbf{c} \\
\mathsf{D}_{\mathbf{c}}(\mathsf{D}_{\mathbf{c}} f)^\top &= B(\boldsymbol{\kappa})^\top B(\boldsymbol{\kappa}) \\
\mathsf{D}_{\mathbf{c}}\Big(\frac{\partial f}{\partial \kappa_k}\Big) &= \mathbf{c}^\top \frac{\partial B(\boldsymbol{\kappa})^\top}{\partial \kappa_k} B(\boldsymbol{\kappa}) - (\mathbf{x} - B(\boldsymbol{\kappa})\mathbf{c})^\top \frac{\partial B(\boldsymbol{\kappa})}{\partial \kappa_k}.
\end{aligned}$$

If $\mu \in \mathcal{S}_{r,p}$ we have $\boldsymbol{\mu} = B(\boldsymbol{\kappa}_0)\mathbf{c}_0$ and then $\mathrm{E}\{\mathbf{x} - B(\boldsymbol{\kappa}_0)\mathbf{c}_0\} = 0$, which implies $\mathrm{E}\{\mathsf{D}^2 f(\mathbf{x};\boldsymbol{\theta}_0)\} = M_0^\top M_0$.

Note that $D_0 = \sum_{k=1}^q \lambda_k M_0^\top \boldsymbol{\phi}_k \boldsymbol{\phi}_k^\top M_0 + \sigma^2 H_0$ under model (5). Occasionally H_0 is nearly singular, so it is better to approximate H_0 and D_0 as follows. Let $\mathbf{m}(t_j;\boldsymbol{\theta})^\top = [\boldsymbol{\beta}(t,\boldsymbol{\kappa})^\top, \mathbf{c}^\top \mathsf{D}\boldsymbol{\beta}(t,\boldsymbol{\kappa})]^\top$, which is the jth row of $M(\boldsymbol{\theta})$. If $\max_j(t_j - t_{j-1}) = O(m^{-1})$,

$$\begin{aligned}\lim_{m\to\infty} M_0^\top M_0/m &= A := \int \mathbf{m}(t;\boldsymbol{\theta}_0)\mathbf{m}(t;\boldsymbol{\theta}_0)^\top\, dt,\\ \lim_{m\to\infty} M_0^\top \boldsymbol{\phi}_k/m &= \mathbf{b}_k := \int \mathbf{m}(t;\boldsymbol{\theta}_0)\boldsymbol{\phi}_k(t)\, dt.\end{aligned}$$

Then $H_0 \approx mA$ and $D_0 \approx \sum_{k=1}^q \lambda_k m^2 \mathbf{b}_k\mathbf{b}_k^\top + \sigma^2 mA$ for large m. We use these approximations instead of H_0 and D_0 when they are nearly singular. Of course, the integrals A and $\mathbf{b}_k$ are also approximated by averages, but here one can take a denser grid than the original $\{t_j\}$, which provides numerically more stable estimates of H_0 and D_0.

Proof of Theorem 2: We proceed by induction on k. Let

$$L_n(\boldsymbol{\phi}_k, \mathbf{l}, \boldsymbol{\zeta}) = \frac{1}{n}\sum_{i=1}^n \{\boldsymbol{\phi}_k^\top(\mathbf{x}_i - \boldsymbol{\mu})\}^2 + l_k(m - \|\boldsymbol{\phi}_k\|^2) - \sum_{j=1}^{k-1} l_j \boldsymbol{\phi}_j^\top \boldsymbol{\phi}_k,$$

where $\boldsymbol{\zeta} = (\boldsymbol{\mu}, \boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{k-1})$. Note that $L_n(\boldsymbol{\phi}_k, \mathbf{l}, \hat{\boldsymbol{\zeta}}_n)$ is the Lagrangian of the maximisation problem (8), and it is easy to verify that $\hat{l}_{n,k} = \hat{\boldsymbol{\phi}}_{n,k}^\top V_n \hat{\boldsymbol{\phi}}_{n,k}/m$ and $\hat{l}_{n,j} = (2/m)\hat{\boldsymbol{\phi}}_{n,j}^\top V_n \hat{\boldsymbol{\phi}}_{n,k}$ for $j = 1, \ldots, k-1$. Since $V_n \xrightarrow{P} \mathrm{V}(\boldsymbol{\eta})$, the eigenvalues of V_n are bounded in probability; therefore, $\{\hat{\mathbf{l}}_n\}$ is also bounded in probability and we can assume without loss of generality that $\mathbf{l}$ is in a compact set $C \in \mathbb{R}^k$. Then

$$(\hat{\boldsymbol{\phi}}_{n,k}, \hat{\mathbf{l}}_n) = \operatorname*{argmax}_{\Omega\times C} L_n(\boldsymbol{\phi}, \mathbf{l}, \hat{\boldsymbol{\zeta}}_n),$$

where $\Omega = \{\boldsymbol{\phi} \in \mathbb{R}^m : \boldsymbol{\phi} = B(\boldsymbol{\kappa})\mathbf{c},\ \|\boldsymbol{\phi}\|^2/m \le 1\}$.

Let $L_0(\boldsymbol{\phi}, \mathbf{l}, \boldsymbol{\zeta}) := \mathrm{E}\{L_n(\boldsymbol{\phi}, \mathbf{l}, \boldsymbol{\zeta})\}$, so

$$L_0(\boldsymbol{\phi}, \mathbf{l}, \boldsymbol{\zeta}) = \boldsymbol{\phi}_k^\top \Sigma \boldsymbol{\phi}_k + \{\boldsymbol{\phi}_k^\top(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\}^2 + l_k(m - \|\boldsymbol{\phi}_k\|^2) - \sum_{j=1}^{k-1} l_j \boldsymbol{\phi}_j^\top \boldsymbol{\phi}_k.$$

In particular,

$$L_0(\boldsymbol{\phi}, \mathbf{l}, \boldsymbol{\zeta}_0) = \boldsymbol{\phi}_k^\top \Sigma \boldsymbol{\phi}_k + l_k(m - \|\boldsymbol{\phi}_k\|^2) - \sum_{j=1}^{k-1} l_j \boldsymbol{\phi}_{0,j}^\top \boldsymbol{\phi}_k.$$

The maximiser of $L_0(\boldsymbol{\phi}, \mathbf{l}, \boldsymbol{\zeta}_0)$ is $(\boldsymbol{\phi}_{0,k}, \mathbf{l}_0)$, with $\mathbf{l}_0 := (0, \ldots, 0, \lambda_k m + \sigma^2)$. Therefore

$$\begin{array}{rcl} 0 & \leq & L_0(\boldsymbol{\phi}_{0,k}, \mathbf{l}_0, \boldsymbol{\zeta}_0) - L_0(\hat{\boldsymbol{\phi}}_{n,k}, \hat{\mathbf{l}}_n, \boldsymbol{\zeta}_0) \\ & \leq & a_n + b_n + c_n + d_n, \end{array}$$

where

$$\begin{array}{rcl} a_n & = & L_0(\boldsymbol{\phi}_{0,k}, \mathbf{l}_0, \boldsymbol{\zeta}_0) - L_0(\boldsymbol{\phi}_{0,k}, \mathbf{l}_0, \hat{\boldsymbol{\zeta}}_n), \\ b_n & = & L_0(\boldsymbol{\phi}_{0,k}, \mathbf{l}_0, \hat{\boldsymbol{\zeta}}_n) - L_n(\boldsymbol{\phi}_{0,k}, \mathbf{l}_0, \hat{\boldsymbol{\zeta}}_n), \\ c_n & = & L_n(\hat{\boldsymbol{\phi}}_{n,k}, \hat{\mathbf{l}}_n, \hat{\boldsymbol{\zeta}}_n) - L_0(\hat{\boldsymbol{\phi}}_{n,k}, \hat{\mathbf{l}}_n, \hat{\boldsymbol{\zeta}}_n), \\ d_n & = & L_0(\hat{\boldsymbol{\phi}}_{n,k}, \hat{\mathbf{l}}_n, \hat{\boldsymbol{\zeta}}_n) - L_0(\hat{\boldsymbol{\phi}}_{n,k}, \hat{\mathbf{l}}_n, \boldsymbol{\zeta}_0). \end{array}$$

(We used the fact that $L_n(\boldsymbol{\phi}_{0,k}, \mathbf{l}_0, \hat{\boldsymbol{\zeta}}_n) \leq L_n(\hat{\boldsymbol{\phi}}_{n,k}, \hat{\mathbf{l}}_n, \hat{\boldsymbol{\zeta}}_n)$, which follows from the definition of $(\hat{\boldsymbol{\phi}}_{n,k}, \hat{\mathbf{l}}_n)$ as maximiser). Since the function

$$M(\boldsymbol{\zeta}) := \sup_{\Omega \times C} |L_0(\cdot, \cdot, \boldsymbol{\zeta}) - L_0(\cdot, \cdot, \boldsymbol{\zeta}_0)|$$

is continuous and $\hat{\boldsymbol{\zeta}}_n \xrightarrow{P} \boldsymbol{\zeta}_0$ by inductive hypothesis, it is clear that a_n and d_n go to zero in probability. On the other hand,

$$\sup_{\Omega \times C} |L_n(\cdot, \cdot, \hat{\boldsymbol{\zeta}}_n) - L_0(\cdot, \cdot, \hat{\boldsymbol{\zeta}}_n)| \leq m \lambda_{\max}(V_n - \Sigma) + m \|\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}_0\|^2 \xrightarrow{P} 0,$$

so b_n and c_n also go to zero in probability. Therefore $L_0(\hat{\boldsymbol{\phi}}_{n,k}, \hat{\mathbf{l}}_n, \boldsymbol{\zeta}_0) \xrightarrow{P} L_0(\boldsymbol{\phi}_{0,k}, \mathbf{l}_0, \boldsymbol{\zeta}_0)$.

Now, since $\|\hat{\boldsymbol{\phi}}_{n,k} \hat{\boldsymbol{\phi}}_{n,k}^\top - \boldsymbol{\phi}_{0,k} \boldsymbol{\phi}_{0,k}^\top\| = \|\hat{\boldsymbol{\phi}}_{n,k} - \boldsymbol{\phi}_{0,k}\| \|\hat{\boldsymbol{\phi}}_{n,k} + \boldsymbol{\phi}_{0,k}\| / \sqrt{2}$ and $\|\hat{\boldsymbol{\phi}}_{n,k}\| =$

$\|\boldsymbol{\phi}_{0,k}\| = \sqrt{m}$, given $\varepsilon > 0$ we have

$$\begin{gathered}\{\|\hat{\boldsymbol{\phi}}_{n,k}\hat{\boldsymbol{\phi}}_{n,k}^\top - \boldsymbol{\phi}_{0,k}\boldsymbol{\phi}_{0,k}^\top\| > \varepsilon\} \cup \{\|\hat{\mathbf{l}}_n - \mathbf{l}_0\| > \varepsilon\} \subseteq \\ \{\|\hat{\boldsymbol{\phi}}_{n,k} - \boldsymbol{\phi}_{0,k}\| > \frac{\varepsilon}{\sqrt{2m}} \text{ and } \|\hat{\boldsymbol{\phi}}_{n,k} + \boldsymbol{\phi}_{0,k}\| > \frac{\varepsilon}{\sqrt{2m}}\} \cup \{\|\hat{\mathbf{l}}_n - \mathbf{l}_0\| > \varepsilon\} \subseteq \\ \{L_0(\boldsymbol{\phi}_{0,k}, \mathbf{l}_0, \boldsymbol{\zeta}_0) - L_0(\hat{\boldsymbol{\phi}}_{n,k}, \hat{\mathbf{l}}_n, \boldsymbol{\zeta}_0) > \delta\}\end{gathered}$$

for some $\delta > 0$, because $L_0(\boldsymbol{\phi}, \mathbf{l}, \boldsymbol{\zeta}_0)$ is continuous and has exactly two maximisers in the compact set $\Omega \times C$, namely $(\boldsymbol{\phi}_{0,k}, \mathbf{l}_0)$ and $(-\boldsymbol{\phi}_{0,k}, \mathbf{l}_0)$. Thus $\|\hat{\boldsymbol{\phi}}_{n,k}\hat{\boldsymbol{\phi}}_{n,k}^\top - \boldsymbol{\phi}_{0,k}\boldsymbol{\phi}_{0,k}^\top\| \xrightarrow{P} 0$ and $\hat{\mathbf{l}}_n \xrightarrow{P} \mathbf{l}_0$ as stated; the latter implies that $\hat{\boldsymbol{\phi}}_{n,k}^\top V_n \hat{\boldsymbol{\phi}}_{n,k}/m \xrightarrow{P} \lambda_k m + \sigma^2$, so $\{\hat{\lambda}_k\}$ and $\hat{\sigma}^2$ are also consistent.■

References

Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions. *Numer. Math.*, **31**, 377–403.

Hansen, M.K. and Kooperberg, C. (2002) Spline adaptation in extended linear models. *Stat. Sci.*, **17**, 2–51.

Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning. Data Mining, Inference and Prediction.* New York: Springer.

Jupp, D.L.B. (1978) Approximation to data by splines with free knots. *Siam J. Numer. Anal.*, **15**, 328–343.

Kneip, A. (1994) Nonparametric estimation of common regressors for similar curve data. *Ann. Statist.*, **22**, 1386–1427.

Mao, W. and Zhao, L.H. (2003) Free-knot polynomial splines with confidence intervals. *J. R. Statist. Soc.* B, **65**, 901–919.

Ramsay, J.O. and Silverman, B.W. (1997) *Functional Data Analysis.* New York: Springer.

Ramsay, J.O. and Silverman, B.W. (2002) *Applied Functional Data Analysis.* New York: Springer.

Rice, J.A. and Silverman, B.W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc.* B, **53**, 233–243.

Silverman, B.W. (1996) Smoothed functional principal components analysis by choice of norm. *Ann. Statist.*, **24**, 1–24.

Schumaker, L.L. (1981) *Spline Functions: Basic Theory.* New York: Wiley.

van der Vaart, A.W. (1998) *Asymptotic Statistics.* Cambridge: Cambridge University Press.

Zhou, S., Shen, X. and Wolfe, D. (1998) Local asymptotics for regression splines and confidence regions. *Ann. Statist.*, **26**, 1760–1782.