

Article

On the Number of Independent Pieces of Information in a Functional Linear Model with a Scalar Response

Eduardo L. Montoya 

Department of Mathematics, California State University, Bakersfield, CA 93311, USA; emontoya2@csu.edu

Received: 18 September 2020; Accepted: 3 November 2020; Published: 5 November 2020



Abstract: In a functional linear model (FLM) with scalar response, the parameter curve quantifies the relationship between a functional explanatory variable and a scalar response. While these models can be ill-posed, a penalized regression spline approach may be used to obtain an estimate of the parameter curve. The penalized regression spline estimate will be dependent on the value of a smoothing parameter. However, the ability to obtain a reasonable parameter curve estimate is reliant on how much information is present in the covariate functions for estimating the parameter curve. We propose to quantify the information present in the covariate functions to estimate the parameter curve. In addition, we examine the influence of this information on the stability of the parameter curve estimator and on the performance of smoothing parameter selection methods in a FLM with a scalar response.

Keywords: eigenvalues; smoothing parameter; functional linear model

1. Introduction

Functional data analysis (FDA) continues to be an active and growing area of research as measurements from continuous processes are increasingly becoming prevalent in many fields. This type of data is functional data because they can be viewed as samples from curves. Recent key references in FDA include those of Hsing and Eubank [1] and Kokoszka and Reimherr [2]. Consider the following linear model where the response is a scalar, but the explanatory variable is function:

$$y_i = \alpha + \int_{\mathcal{T}} x_i(t)\beta(t)dt + \varepsilon_i \quad (1)$$

for $i = 1, 2, \dots, n$. The response y_i is a scalar value, $x_i(t)$ is a function, α is the intercept, $\beta(t)$ is a parameter curve, and ε_i is uncorrelated random noise with zero mean and constant variance σ^2 . Model (1) is also referred to as a scalar-on-function regression model. Here, we assume $\mathcal{T} = [0, 1]$, but note that any closed continuous domain \mathcal{T} in \mathbb{R} can be transformed to $[0, 1]$. Further, we assume that the covariate functions are known.

The objective for model (1) is to estimate the smooth parameter curve, $\beta(t)$. The parameter curve quantifies the relationship between the scalar response and a functional explanatory variable in the presence of uncertainty. A review of some common approaches for estimating $\beta(t)$ are provided in [3], and these approaches depend on some variant of a tuning parameter (smoothing parameter, number of knots, bandwidth, number of retained principal components, etc.), where the size of the tuning parameter controls the trade-off between the goodness-of-fit and the smoothness of the parameter curve estimate. Regarding inference for model (1), some recently proposed methodologies include goodness-of-fit test [4] and testing linearity in a FLM with a scalar response [5]. Tekbudak et al. [6] provided a comparison of these and other recent testing procedures for linearity in a FLM with a scalar response. In influence diagnostics, Cook's distance [7] and Peña's distance [8] are extended to a FLM with a scalar response [9].

In this paper, our aim is to quantify the amount of information present in the covariate functions for estimating the parameter curve when $\beta(t)$ is identifiable and assess the influence of the amount of this information on the numerical stability of the parameter curve estimator. In addition, we review the performance of different smoothing parameter selection methods based on the amount of information present in the covariate functions for estimating $\beta(t)$. To our knowledge, no study has explicitly focused on this aspect of model (1). Here, the concept of the numerical stability (referred to simply as stability hereafter) is tied to the idea that the parameter curve estimate will not substantially change when the set of covariate functions are slightly altered. In Section 2, we define a measure, denoted $\zeta(\mathbf{x}(t))$, to quantify the amount of information present in the covariate functions for estimating $\beta(t)$ when the parameter curve is identifiable. There are various computational methods to estimate the smooth parameter curve, $\beta(t)$, in model (1). Our approach is to estimate $\beta(t)$ using penalized regression spline estimation. Penalized regression spline estimation of $\beta(t)$ and common smoothing parameter selection methods are discussed in Section 3. Section 4 proposes measures to assess the performance of smoothing parameter selection methods and the stability of a parameter curve estimator. A simulation study that assesses the relationship between $\zeta(\mathbf{x}(t))$ and the stability of the parameter curve estimator, as well as examines the performance of different smoothing parameter selection methods under varying $\zeta(\mathbf{x}(t))$ s, is given in Section 5. Section 6 provides a real data application of $\zeta(\mathbf{x}(t))$. We conclude with discussion in Section 7.

2. Number of Independent Pieces of Information in a FLM

Model (1) can be ill-posed to varying degrees. An ill-posed problem refers to one for which no solution exists, the solution is not unique, or the solution is unstable [10]. Cardot et al. [11] provided theoretical conditions for the existence and uniqueness of a solution to (1), where the solution falls in the space spanned by the eigenfunctions of the functional covariate's covariance operator in which the model space is a separable Hilbert space of square integrable functions defined on $[0, 1]$. In practice, it is generally assumed that theoretical conditions for identifiability are satisfied when estimating $\beta(t)$ in model (1). In scalar-on-image regression models, Happ et al. [12] studied the impact of structural assumptions of the parameter image, such as smoothness and sparsity, on the model estimates, as well as measures to assess to what degree the assumptions are satisfied.

Our focus is assessing how much information is present in the covariate functions to estimate the parameter curve in model (1) when the parameter curve is identifiable. To our knowledge, no prior studies have given consideration to this aspect of model (1) and its influence on the stability of the parameter curve estimator. Given the different areas of application of model (1), assessing this relationship is essential as the stability of the parameter curve estimate would influence the reliability of model uncertainty estimates. We aim to study this aspect of model (1) by proposing to quantify the information present in the covariate functions to estimate the parameter curve. The idea underlying this work is motivated by the work of Wahba [13].

Wahba [13] proposed the idea of the number of independent pieces of information to gauge if one can obtain a reasonable solution in a type of general smoothing spline model (GSSM). In this context, the number of eigenvalues provided by the eigendecomposition of the inner-product of the representer of a bounded linear functional with itself when scaled by the reciprocal of the variance of the error component in the model that are greater than one are considered to be the number of independent pieces of information. If the number of independent pieces of information in a GSSM is large, then a solution to a GSSM is recoverable. However, no explicit criterion was given to quantify how many pieces are required. Motivated by this, we define a measure for the number of independent pieces of information in the covariate functions for estimating $\beta(t)$ in model (1). Let v_1, \dots, v_n be the eigenvalues

from the eigendecomposition of $\int_0^1 (\mathbf{x}(t)) (\mathbf{x}(t))^T dt$ where $\mathbf{x}(t) = (x_1(t), \dots, x_n(t))^T$. We define the number of independent pieces of information in the covariate functions for estimating $\beta(t)$ as

$$\zeta(\mathbf{x}(t)) = \sum_{i=1}^n \mathbb{I}_{(v_i/\sigma^2 > 1)}, \tag{2}$$

where σ^2 is the variance of the error component in model (1). Measure (2) may be estimated by plugging in an estimate of σ^2 . An estimate of σ^2 is provide in Section 3.

As an illustration of (2), Figure 1 contains four different sets of covariate functions that vary in their number of independent pieces of information. These covariate functions are used in a simulation study in Section 5. Figures in this study were produced using the R packages ggplot2 [14] and cowplot [15]. If there is less information present in the covariate functions to estimate $\beta(t)$ as quantified by (2), then one would anticipate that a reasonable or stable parameter curve estimator would be less feasible. To assess whether the size of $\zeta(\mathbf{x}(t))$ indicates the degree of stability of a solution to model (1), we propose a stability measure of a parameter curve estimator in Section 4. We address the relationship between $\zeta(\mathbf{x}(t))$ and the stability of the estimator in Section 5.

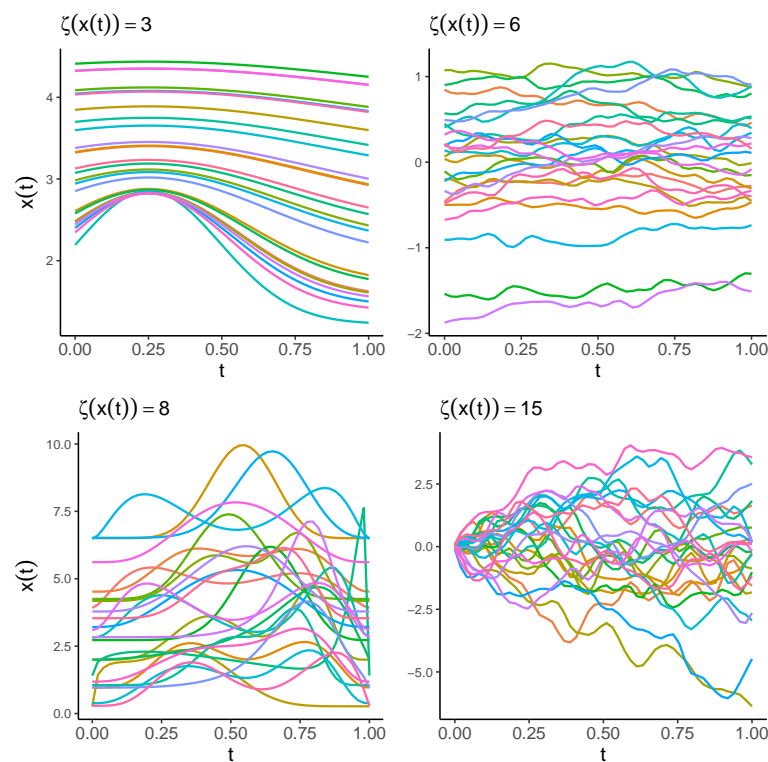


Figure 1. Four different sets of covariate functions of size $n = 25$ and their respective number of independent pieces of information when the signal-to-noise ratio is 0.10 in model (1).

3. Penalized Regression Spline Estimate of $\beta(t)$

Here, we review penalized regression spline estimation of the parameter curve $\beta(t)$. Assuming that the parameter curve is smooth in the sense that it lies in a Sobolev space of order 4, we may seek an estimate of $\beta(t)$ by minimizing

$$\sum_{i=1}^n \left(y_i - \alpha - \int_0^1 x_i(t)\beta(t)dt \right)^2 + \lambda \int_0^1 [\beta''(t)]^2 dt, \tag{3}$$

where the term $\int_0^1 (\beta''(t))^2 dt$ penalizes curvature in the estimate. To ease notation, from here forward, we assume the intercept $\alpha = 0$, but a non-zero α value is easily incorporated into the computational approaches discussed.

We use the method of regularized basis functions [16], with a B-spline basis of order 4 to minimize fitting criterion (3) with respect to the parameter curve $\beta(t)$. With this approach, each covariate function $x_i(t)$ and the parameter curve $\beta(t)$ are represented using a linear combination of B-spline basis functions. Let $x_i(t)$ be represented as $\sum_{j=1}^{J_x} k_{i,j} N_j(t)$ for $i = 1, \dots, n$, where $N_j(t)$ denotes the j th B-spline function of order 4. Similarly, we represent $\beta(t)$ as $\sum_{j=1}^{J_\beta} c_j N_j(t)$. Now, let $\sum_{j=1}^{J_x} k_{i,j} N_j(t) = \mathbf{K}_{[i]} \mathbf{N}_x(t)$ where $\mathbf{N}_x(t) = [N_1(t) \cdots N_{J_x}(t)]^T$ and $\mathbf{K}_{[i]}$ denotes the i th row of the matrix \mathbf{K} that has elements $\mathbf{K}_{[i,j]} = k_{i,j}$. Similarly, let $\sum_{j=1}^{J_\beta} c_j N_j(t) = \mathbf{N}_\beta(t)^T \mathbf{c}$ where $\mathbf{c} = (c_1, \dots, c_{J_\beta})^T$. In this basis representation framework, penalized least squares criterion (3) is re-expressed as finding \mathbf{c} to minimize

$$(\mathbf{y} - \mathbf{M}\mathbf{c})^T (\mathbf{y} - \mathbf{M}\mathbf{c}) + \lambda \mathbf{c}^T \mathbf{R}\mathbf{c}, \tag{4}$$

where $\mathbf{y} = (y_1, \dots, y_n)^T$, $\mathbf{M} = \mathbf{K} \int_0^1 \mathbf{N}_x(t) \mathbf{N}_\beta^T(t) dt$, and $\mathbf{R} = \int_0^1 [\mathbf{N}_\beta''(t)] [\mathbf{N}_\beta''(t)]^T dt$. For a given λ , an estimate of $\beta(t)$ is obtained by minimizing (4) with respect to \mathbf{c} via

$$\begin{aligned} \hat{\beta}_\lambda(t) &= \mathbf{N}_\beta(t)^T \hat{\mathbf{c}}_\lambda \\ &= \mathbf{N}_\beta(t)^T \left(\mathbf{M}^T \mathbf{M} + \lambda \mathbf{R} \right)^{-1} \mathbf{M}^T \mathbf{y}, \end{aligned}$$

where the subscript λ signifies the dependence of the solution on the value of the smoothing parameter.

Various data-driven approaches have been proposed to select the smoothing parameter λ in (3). These methods include Akaike’s information criterion, Akaike’s information criterion corrected, cross-validation, generalized cross-validation criterion, L-curve criterion, restricted maximum likelihood, Schwarz information criterion, etc. Since the size of λ controls the size of the penalty in (3), most of these data-driven methods consist of two parts: one that measures the goodness of fit of the model and another that quantifies the complexity of the parameter curve estimate. Thus, these methods attempt to achieve an optimal balance between how well the model fits the data and the smoothness of the parameter curve estimate (see [4,9,17–19], and others cited therein for examples of recent studies that have used one or more of these criteria to select the smoothing parameter in models of type (1).)

In our study, we restrict our discussion to the following commonly used data-driven criteria: Akaike’s information criterion corrected (*AICc*), cross-validation (*CV*), Schwarz information criterion (*SIC*), and the generalized cross-validation (*GCV*) criterion. The smoothing parameter selection methods used in our study are by no means exhaustive, nor are they meant to be. Rather, our intent is to explore if the amount of information present in the covariate functions for estimating $\beta(t)$ may effect the performance of given smoothing parameter selection method. For each criterion, the value of the smoothing parameter λ that minimizes the criterion is assumed to be a reasonable value for λ . Each criterion discussed here is dependent on the residuals sum of squares defined by

$$RSS_\lambda = \sum_{i=1}^n \left(y_i - \int_0^1 x_i(t) \hat{\beta}_\lambda(t) dt \right)^2$$

and the smoother matrix, $\mathbf{S}_\lambda = \mathbf{M}(\mathbf{M}^T \mathbf{M} + \lambda \mathbf{R})^{-1} \mathbf{M}^T$. With *CV*, for each λ , we obtain an estimate of $\beta(t)$ based on minimizing

$$\sum_{j=1, j \neq i}^N \left[y_j - \int_0^1 x_j(t) \beta(t) dt \right]^2 + \lambda \int_0^1 (\beta''(t))^2 dt.$$

Denote the minimizing solution by $\widehat{\beta}_\lambda^{(-i)}(t)$, where $(-i)$ symbolizes an estimate based on all observations except for the i th case. Since there are n cases that one can delete for a given λ , a cross-validation score is defined as

$$CV(\lambda) = N^{-1} \sum_{i=1}^N \left[y_i - \int_0^1 x_i(t) \widehat{\beta}_\lambda^{(-i)}(t) dt \right]^2.$$

A computationally friendly form of CV [16] is defined as

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \int_a^b x_i(t) \widehat{\beta}_\lambda(t) dt}{1 - S_{ii}} \right)^2,$$

where S_{ii} denotes the i th diagonal element of \mathbf{S}_λ .

The GCV criterion [20] replaces the diagonal elements of the smoother matrix in the CV formula by $tr(\mathbf{S}_\lambda)/n$ to obtain

$$GCV(\lambda) = \frac{RSS_\lambda}{n(1 - \gamma df_\lambda)^2}, \tag{5}$$

where $df_\lambda = tr(\mathbf{S}_\lambda)$. df_λ is referred to as the effective degrees of freedom [21]. RSS_λ and df_λ are commonly used to estimate σ^2 via $\widehat{\sigma}_\lambda^2 = RSS_\lambda / (n - df_\lambda)$. An estimate of σ^2 may then be used to estimate $\zeta(\mathbf{x}(t))$ via

$$\widehat{\zeta}(\mathbf{x}(t)) = \sum_{i=1}^n \mathbb{I}_{(v_i/\widehat{\sigma}^2 > 1)}. \tag{6}$$

The term γ in (5) represents an inflation of the effective degrees of freedom (EDF) for $\gamma > 1$. Inflation of the EDF is used as a measure to guard against GCV selecting a smoothing parameter that over-fits the data in non-parametric models [22]. Some simulation studies suggest 1.4 to be reasonable in non-parametric models [22,23]. The SIC criterion [24] may be expressed as

$$SIC(\lambda) = \ln \left(\frac{1}{n} RSS_\lambda \right) + df_\lambda \frac{\log(n)}{n}.$$

The AICc criterion proposed by Hurvich et al. [25] penalizes more complex estimates of $\beta(t)$ than does the SIC for smaller sample sizes, and it may be defined as

$$AICc(\lambda) = \ln \left(\frac{1}{n} RSS_\lambda \right) + \frac{2(df_\lambda + 1)}{n - df_\lambda - 2}.$$

4. Quantifying Stability of $\widehat{\beta}(t)$ and the Performance of Smoothing Parameter Selection Methods

An ideal value for the smoothing parameter, call it λ^* , may be considered one that minimizes the integrated squared error,

$$ISE(\beta, \widehat{\beta}_{\lambda^*}) = \int_0^1 \left(\beta(t) - \widehat{\beta}_{\lambda^*}(t) \right)^2 dt. \tag{7}$$

To assess the performance of the smoothing parameter selection methods discussed in Section 3, we use the median of the measure

$$K(\beta, \widehat{\beta}) = \frac{\int_0^1 \left(\beta(t) - \widehat{\beta}(t) \right)^2 dt}{\int_0^1 \left(\beta(t) - \widehat{\beta}_{\lambda^*}(t) \right)^2 dt}. \tag{8}$$

Note that any penalized regression spline estimate of the parameter curve will be dependent on the chosen value of the smoothing parameter, but the notation of this dependence is suppressed for better readability. Relative to $\beta(t)$, the estimator $\widehat{\beta}(t)$ is considered better the closer the median of (8)

is to 1. The further away the median of (8) is from 1, the poorer is the performance of the smoothing parameter selection method. Measure (8) was motivated by a measure proposed by Lee [26] to compare smoothing parameter selection methods in smoothing splines.

To assess the stability of a parameter curve estimator, we propose a leave-one-out measure motivated by the DFFITS [27] statistic. Specifically, we use the median of a leave-one-out integrated squared error measure,

$$ISE_{(n)}(\beta, \hat{\beta}_{\lambda^*}^{(-i)}) = \frac{1}{n} \sum_{i=1}^n \int_0^1 (\beta(t) - \hat{\beta}_{\lambda^*}^{(-i)}(t))^2 / \hat{\sigma}^{2(-i)} dt, \quad (9)$$

where $\hat{\beta}_{\lambda^*}^{(-i)}(t)$ and $\hat{\sigma}_{\lambda^*}^{2(-i)}$ represent estimators of $\beta(t)$ and σ^2 , respectively, based on all observations except for the i th case, and where λ^* minimizes criterion (7). A large value of the median of (9) would reflect a less stable estimator. If $\zeta(\mathbf{x}(t))$ is to be considered an appropriate measure of information in the covariate functions for estimating $\beta(t)$, then large values of $\zeta(\mathbf{x}(t))$ would be associated with small values of the median of (9) in the sense that slightly altering the set of covariate functions did not lead to large changes in the parameter curve estimate. Similarly, smaller values of $\zeta(\mathbf{x}(t))$ would be associated with larger values of the median of (9). This is evaluated with a simulation study in the next section.

5. A Simulation Study

Using a simulation study, we examine the stability of a parameter curve estimator and the performance of the smoothing parameter selection methods at varying levels of $\zeta(\mathbf{x}(t))$. The sampling distributions required to derive analytical formulas for the median of (8) and (9) are unattainable due to the dependence of the parameter curve estimator on the smoothing parameter. Therefore, a simulation estimate of the median of (8) is obtained by

$$SMK = \text{Median}(K_g(\beta, \hat{\beta})). \quad (10)$$

The subscript g denotes the g th simulated dataset. Similarly, we estimate the median of (9) via

$$SMISE_{(n)} = \text{Median}(ISE_g(\beta, \hat{\beta}_{\lambda^*})). \quad (11)$$

To obtain (10) and (11), simulated datasets are generated under various settings by the model,

$$y_i = \int_0^1 x_i(t)\beta(t)dt + \varepsilon_i,$$

for $i = 1, \dots, n$, where ε_i s are assumed to independent and identically normal random variables with mean 0 and standard deviation σ . Overall, under four different sets of covariate functions, $g = 2500$ simulated datasets were generated for each combination of $n \in \{25, 50, 100\}$, $x_i(t) \in \{x_{i1}(t), x_{i2}(t), x_{i3}(t), x_{i4}(t)\}$, $\beta \in \{\beta_1(t), \beta_2(t), \beta_3(t)\}$, and σ is chosen for each setting to ensure that $\kappa \in \{.10, .20\}$, where κ refers to the signal-to-noise ratio. As defined by Febrero-Bande et al. [9], $\kappa = \sigma/\psi$, where ψ denotes the standard deviation of $\int_0^1 x_i(t)\beta(t)dt$ for $i = 1, \dots, n$. The different sets of covariate functions used in this study were first produced on a discretized scale of 50 equally spaced values. To ensure sufficient flexibility in their functional representations, the parameter curve and the set covariate functions are represented as functions using the approach described in Section 3, with $J_\beta = 50$ and $J_x = 50$, respectively.

The simulations are performed using R [28], along with the extension and usage of code from the R package fda [29]. In addition, the R packages dplyr [30] and tidyr [31] were used for data management. The parameter curves in this study (shown in Figure 2) are defined as

$$\begin{aligned} \beta_1(t) &= \frac{1}{23}Beta_{[20,5]}(t) + \frac{1}{3}Beta_{[12,12]}(t) + \frac{1}{3}Beta_{[7,30]}(t), \\ \beta_2(t) &= \frac{7}{10}Beta_{[3,10]}(t) + \frac{3}{10}Beta_{[7,2]}(t), \\ \beta_3(t) &= \frac{1}{10}Beta_{[2,5]}(t) + \frac{9}{10}Beta_{[5,2]}(t), \end{aligned}$$

where

$$Beta_{[q,p]}(t) = \frac{\Gamma(q+p)}{\Gamma(q)\Gamma(p)}t^{q-1}(1-t)^{p-1} \text{ for } t \in [0, 1],$$

and they were chosen to assess the performance of (10) and (11) at varying levels of complexity of the parameter curve in terms of their approximate curvature.

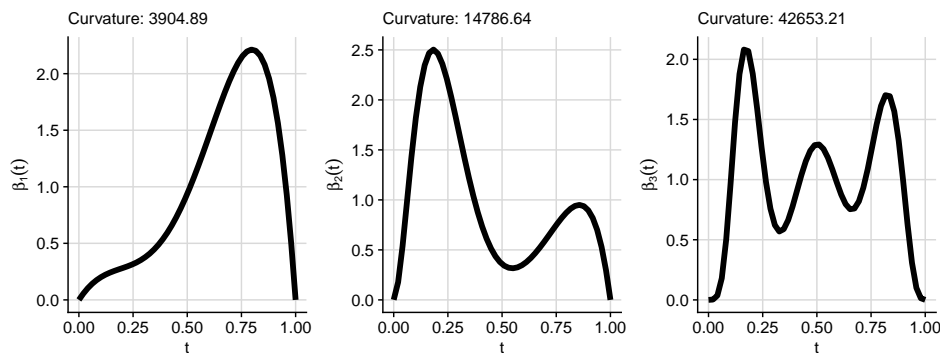


Figure 2. The parameter curves $\beta_1(t)$ (left), $\beta_2(t)$ (middle), and $\beta_3(t)$ (right) used in the simulation study.

The different sets of covariate functions were generated under different conditions to ensure varying levels of $\zeta(\mathbf{x}(t))$. The first set of covariate curves ($x_{i1}(t)$ s) are produced by generating realizations from a normal random variable with mean 0 and standard deviation η that are randomly shifted about the y-axis by η , where $\eta \sim Unif(1,30)$. We denote this first set of covariate curves as Covariate Set 1. Covariate Set 2 ($x_{i2}(t)$ s) is created by generating realizations from a Gaussian process having an exponential covariogram with variance parameter 10 and scale parameter 0.4. Covariate Set 3 ($x_{i3}(t)$ s) is produced by generating realizations from a mixture of beta random variables randomly shifted about the y-axis. Covariate Set 4 ($x_{i4}(t)$ s) is obtained by generating realizations from a Brownian motion process with variance parameter 2.7. As an illustration, the four different sets of covariate functions used in this study are shown in Figure 1 for a sample of size $n = 25$ at a specified signal-to-noise ratio.

Tables 1–3 provide the SMK when the parameter curves are $\beta_1(t)$, $\beta_2(t)$, and $\beta_3(t)$, respectively, under each simulation setting and smoothing parameter selection method. The results are presented on the log scale to better depict the differences. Note that a small value of the SMK does not reflect whether a reasonable estimator of $\beta(t)$ was obtained but rather reflects the performance of the given smoothing parameter selection method relative to criterion (7). Some of the differences between the smallest and the next smallest values of the SMK across the smoothing parameter selection methods for a given simulation scenario may not appear large. To better distinguish between the smallest and the next smallest values of the SMK, pairwise comparisons of the SMKs in a given simulation setting were performed using Mood’s median test with the Benjamini and Hochberg [32] correction as implemented in the R package RVAideMemoire [33] at significance level 0.05. Here, we consider the performance

of a smoothing parameter selection method to be favorable or best if it obtains the lowest *SMK* or an *SMK* that is not significantly different from the lowest *SMK*. We now summarize the favorable smoothing parameter selection method(s) that were most common across different simulation settings.

Under Covariate Set 1, *SIC* consistently obtained the lowest *SMK* across across all simulation settings. Under Covariate Set 2 at $n = 25$, $GCV_{\gamma=1.4}$, *AICc*, and *SIC* consistently preformed best under the different simulation settings. However, *SIC* tended to performed best across all simulation settings at $n = 50$. At $n = 100$ and the lower κ , *AICc* generally performed best across the three parameter curves. At the higher κ , *SIC* was the better smoothing parameter selection method. When the covariate functions assume the form of Covariate Set 3, *AICc* and $GCV_{\gamma=1.4}$ tended to perform just as well or better than the other smoothing parameter selection methods across all settings. Under Covariate Set 4 with $n = 25$, *AICc* was consistently among the better methods for all settings. For $n = 50$, *GCV* and *AICc* were favorable under parameter curves $\beta_1(t)$ and $\beta_2(t)$, whereas *SIC* was favorable under $\beta_3(t)$. When $n = 100$, *AICc* performed best or just as well as the other methods.

Table 1. The *SMK* under each smoothing parameter selection criterion for each simulation setting under parameter curve $\beta_1(t)$. A ⁺ next to a *SMK* for a given sample size indicates the lowest *SMK* across the different smoothing parameter selection methods. A * next to an *SMK* for a given sample size indicates that the *SMK* was not significantly different from lowest *SMK*.

		<i>n</i>	$\zeta(\mathbf{x}(t))$	<i>CV</i>	<i>GCV</i>	$GCV_{\gamma=1.4}$	<i>AICc</i>	<i>SIC</i>
Covariate Set 1	$\kappa = 0.10$	25	3	0.193	0.180	0.169 *	0.171 *	0.168 +
		50	3	0.193	0.182	0.177 *	0.180	0.173 +
		100	3	0.184 *	0.182 *	0.178 +	0.181 *	0.178 +
	$\kappa = 0.20$	25	2	0.190	0.183	0.139 *	0.151 *	0.138 +
		50	2	0.184	0.173	0.149 *	0.164	0.138 +
		100	3	0.174	0.174	0.168 *	0.173	0.158 +
Covariate Set 2	$\kappa = 0.10$	25	6	0.372	0.363	0.279 +	0.297 *	0.289 *
		50	9	0.478 *	0.478 *	0.450 +	0.459 *	0.454 *
		100	14	0.536 *	0.529 *	0.559 *	0.509 +	0.719
	$\kappa = 0.20$	25	3	0.165	0.146	0.086 +	0.099 *	0.088 *
		50	5	0.276	0.276	0.136	0.220	0.095 +
		100	8	0.449	0.453	0.373	0.445	0.261 +
Covariate Set 3	$\kappa = 0.10$	25	8	0.717	0.611 *	0.697	0.604 +	0.711
		50	10	0.412	0.400	0.357 *	0.352 +	0.412
		100	10	0.310	0.309	0.230 +	0.269	0.265
	$\kappa = 0.20$	25	7	0.671	0.665	0.597 *	0.580 +	0.628 *
		50	8	0.566 *	0.538 *	0.553 *	0.515 +	0.639
		100	9	0.446 *	0.446 *	0.409 +	0.422 *	0.617
Covariate Set 4	$\kappa = 0.10$	25	14	0.604 *	0.601 *	0.590 *	0.553 +	0.654
		50	17	0.483	0.455 *	0.460	0.411 +	0.599
		100	21	0.350	0.348	0.293 +	0.305 *	0.441
	$\kappa = 0.20$	25	10	0.539	0.521	0.417 *	0.410 +	0.455 *
		50	12	0.581 *	0.591 *	0.555 *	0.537 +	0.596 *
		100	16	0.550 *	0.549 *	0.538 *	0.523 +	0.739

Covariate Sets 1 and 2 tended to have a lower $\zeta(\mathbf{x}(t))$ in our study for a given sample size and κ , whereas Covariate Sets 3 and 4 had a higher $\zeta(\mathbf{x}(t))$. While a perfect one-to-one relationship does not appear evident between $\zeta(\mathbf{x}(t))$ and the performance of a smoothing parameter selection method, *SIC* tended to perform more favorably when $\zeta(\mathbf{x}(t))$ ranged between two and nine across the three different parameter curves. For the higher values of $\zeta(\mathbf{x}(t))$, *AICc* tended to perform just as well or better than the other methods more often than not. In addition, note that $GCV_{\gamma=1.4}$ obtained a lower *SMK* than *GCV* for almost all simulation settings.

Table 2. Analogous to Table 1 for parameter curve $\beta_2(t)$.

		n	$\zeta(x(t))$	CV	GCV	$GCV_{\gamma=1.4}$	AIC_C	SIC
Covariate Set 1	$\kappa = 0.10$	25	3	0.059	0.055	0.047 ⁺	0.049 [*]	0.047 ⁺
		50	3	0.279	0.260	0.079	0.191	0.065 ⁺
		100	3	0.621	0.619	0.494	0.604	0.289 ⁺
	$\kappa = 0.20$	25	2	0.052	0.050	0.040 [*]	0.043 [*]	0.039 ⁺
		50	3	0.070	0.063	0.055 [*]	0.059	0.051 ⁺
		100	3	0.298	0.281	0.080	0.248	0.063 ⁺
Covariate Set 2	$\kappa = 0.10$	25	6	0.289	0.307	0.215 ⁺	0.239 [*]	0.222 [*]
		50	9	0.397	0.398	0.332 [*]	0.367	0.298 ⁺
		100	14	0.567 [*]	0.560 [*]	0.633	0.542 ⁺	0.827
	$\kappa = 0.20$	25	3	0.133	0.115	0.048 ⁺	0.057 [*]	0.049 [*]
		50	5	0.188	0.176	0.078	0.146	0.051 ⁺
		100	8	0.387	0.382	0.297	0.363	0.222 ⁺
Covariate Set 3	$\kappa = 0.10$	25	8	0.488 [*]	0.495 [*]	0.498 [*]	0.464 ⁺	0.515 [*]
		50	10	0.416	0.427	0.383 [*]	0.363 ⁺	0.451
		100	10	0.413 [*]	0.417 [*]	0.392 [*]	0.378 ⁺	0.521
	$\kappa = 0.20$	25	7	0.407	0.452	0.360 ⁺	0.361 [*]	0.386 [*]
		50	8	0.446	0.418 [*]	0.411 [*]	0.382 ⁺	0.512
		100	9	0.482 [*]	0.487 [*]	0.429 ⁺	0.439 [*]	0.633
Covariate Set 4	$\kappa = 0.10$	25	16	0.425 ⁺	0.435 [*]	0.478	0.426 [*]	0.496
		50	20	0.349 [*]	0.344 [*]	0.390	0.324 ⁺	0.496
		100	23	0.394 [*]	0.407 [*]	0.472	0.377 ⁺	0.731
	$\kappa = 0.20$	25	12	0.477	0.481	0.449 [*]	0.426 ⁺	0.476
		50	16	0.440 [*]	0.425 [*]	0.513	0.413 ⁺	0.645
		100	17	0.481 [*]	0.471 [*]	0.519	0.429 ⁺	0.874

Table 3. Analogous to Table 1 for parameter curve $\beta_3(t)$.

		n	$\zeta(x(t))$	CV	GCV	$GCV_{\gamma=1.4}$	AIC_C	SIC
Covariate Set 1	$\kappa = 0.10$	25	3	0.095	0.091	0.081 ⁺	0.085 [*]	0.081 ⁺
		50	3	0.088	0.086	0.077 [*]	0.082	0.072 ⁺
		100	3	0.086	0.087	0.082 [*]	0.086	0.075 ⁺
	$\kappa = 0.20$	25	2	0.093	0.089	0.069 ⁺	0.075 [*]	0.069 ⁺
		50	3	0.094	0.092	0.078	0.087	0.068 ⁺
		100	3	0.093	0.092	0.084	0.091	0.075 ⁺
Covariate Set 2	$\kappa = 0.10$	25	6	0.077	0.072	0.060 ⁺	0.063 [*]	0.061 [*]
		50	9	0.091	0.088	0.071 [*]	0.077	0.068 ⁺
		100	14	0.316	0.321	0.273	0.302	0.245 ⁺
	$\kappa = 0.20$	25	3	0.075	0.069	0.046 ⁺	0.052 [*]	0.046 ⁺
		50	5	0.072	0.071	0.055 [*]	0.064	0.052 ⁺
		100	8	0.100	0.097	0.080 [*]	0.092	0.073 ⁺
Covariate Set 3	$\kappa = 0.10$	25	8	0.358	0.347	0.308 [*]	0.307 ⁺	0.323 [*]
		50	10	0.332	0.321 [*]	0.332	0.296 ⁺	0.386
		100	10	0.335 [*]	0.340 [*]	0.301 ⁺	0.311 [*]	0.409
	$\kappa = 0.20$	25	7	0.160	0.156	0.082 ⁺	0.092 [*]	0.095 [*]
		50	8	0.302	0.293	0.246 [*]	0.261	0.235 ⁺
		100	9	0.353	0.343 [*]	0.325 [*]	0.322 ⁺	0.376
Covariate Set 4	$\kappa = 0.10$	25	15	0.538 [*]	0.504 ⁺	0.538 [*]	0.520 [*]	0.550 [*]
		50	19	0.583 [*]	0.555 ⁺	0.772	0.612 [*]	0.869
		100	23	0.377 [*]	0.389 [*]	0.560	0.360 ⁺	1.158
	$\kappa = 0.20$	25	11	0.192	0.189	0.081 ⁺	0.088 [*]	0.098 [*]
		50	14	0.334	0.335	0.213	0.261	0.175 ⁺
		100	17	0.465 [*]	0.466 [*]	0.450 [*]	0.450 [*]	0.432 ⁺

Table 4 shows the $SMISE_{(n)}$ under each simulation setting. The results are presented on the log scale to better depict the differences. The results show similar patterns or trends under each κ . For a given parameter curve and sample size, the $SMISE_{(n)}$ decreases as $\zeta(\mathbf{x}(t))$ increases. Thus, the more information present in the covariate functions for estimating the parameter curve, the more stable the parameter curve estimator. We also note that, for a given covariate set, an increase in the sample size corresponds to a decrease in the $SMISE_{(n)}$ and to a non-decreasing $\zeta(\mathbf{x}(t))$. This reassures that $\zeta(\mathbf{x}(t))$ may be viewed as a measure of the amount of information present in the covariate functions in the sense that under a given covariate set, more observed data tends to increase $\zeta(\mathbf{x}(t))$ to varying degrees and provide a smaller $SMISE_{(n)}$. However, a large sample size does not imply a large $\zeta(\mathbf{x}(t))$, as illustrated by Covariate Set 1. Similarly, a small sample size does not imply a low $\zeta(\mathbf{x}(t))$, such as under Covariate Set 4. The differences in results under each κ are in part due to the scaling involved in (9) by $\hat{\sigma}^{2(-i)}$, where $\hat{\sigma}^{2(-i)}$ would tend to be higher under $\kappa = 0.20$.

Recall that for smoothing parameter selection under a covariate set with a low $\zeta(\mathbf{x}(t))$, SIC performed best or just as well as the other smoothing parameter selection methods. For covariate sets with a higher $\zeta(\mathbf{x}(t))$, $AICc$ was generally among the better smoothing parameter selection methods. To visualize the impact of $\zeta(\mathbf{x}(t))$ on the stability of the parameter curve estimate, Figure 3 shows the resulting approximate expected value of the parameter curve estimator (computed across all parameter curve estimates in the simulation study) plus or minus two times the approximate standard deviation of the parameter curve estimator when using the preferred smoothing parameter selection method suggested by Tables 1–3. For brevity, we only present the results for $n = 25$. The parameter curve estimator, under Covariate Set 1, showed much higher variability than under the other covariate sets. Further, the covariate set with a higher $\zeta(\mathbf{x}(t))$ (Covariate Set 4) reflected the lowest variability. Similar behavior was exhibited at $n = 50$ and $n = 100$. This behavior is consistent with the behavior of the $SMISE_{(n)}$ for a given covariate set and sample size. This reflects that a low $\zeta(\mathbf{x}(t))$ is associated with a less stable solution, which in turn may substantially increase the variability of a parameter curve estimator, where such variability would not be reflected in an observed confidence interval for the parameter curve.

In practice, $\zeta(\mathbf{x}(t))$ will need to be estimated due to its dependence on σ^2 . An estimate of this measure, $\hat{\zeta}(\mathbf{x}(t))$, is provided in (6) using the estimate of σ^2 provided in Section 3. The estimate of σ^2 will be dependent on a chosen value of the smoothing parameter. To better understand the impact of the chosen value of the smoothing parameter on (6), Tables 5–7 provide the average value of (6), computed as the average over all simulated datasets, when parameter curves are $\beta_1(t)$, $\beta_2(t)$, and $\beta_3(t)$, respectively, under each simulation setting and smoothing parameter selection method. On average, (6) provides a reasonable estimate of $\zeta(\mathbf{x}(t))$ in the simulation settings considered.

Table 4. The $SMISE_{(n)}$ under each simulation setting. The size of $\zeta(\mathbf{x}(t))$ is shown below each respective $SMISE_{(n)}$ in parentheses.

n	$\beta_1(t)$				$\beta_2(t)$				$\beta_3(t)$			
	$x_1(t)$	$x_2(t)$	$x_3(t)$	$x_4(t)$	$x_1(t)$	$x_2(t)$	$x_3(t)$	$x_4(t)$	$x_1(t)$	$x_2(t)$	$x_3(t)$	$x_4(t)$
	$\kappa = 0.10$				$\kappa = 0.10$				$\kappa = 0.10$			
25	10.44 (3)	5.13 (6)	4.83 (8)	2.49 (14)	11.12 (3)	5.76 (6)	5.56 (8)	3.51 (16)	11.32 (3)	5.56 (6)	5.91 (8)	3.62 (15)
50	9.66 (3)	4.41 (9)	3.07 (10)	1.28 (17)	10.27 (3)	5.02 (9)	3.94 (10)	2.21 (20)	10.61 (3)	5.01 (9)	4.67 (10)	2.54 (19)
100	9.31 (3)	3.94 (14)	2.67 (10)	0.61 (21)	9.18 (3)	4.44 (14)	3.29 (10)	1.51 (23)	10.46 (3)	4.85 (14)	4.47 (10)	1.85 (23)
	$\kappa = 0.20$				$\kappa = 0.20$				$\kappa = 0.20$			
25	8.95 (2)	4.01 (3)	3.86 (7)	1.74 (10)	9.96 (2)	4.69 (3)	4.48 (7)	3.02 (12)	9.97 (2)	4.34 (3)	4.28 (7)	2.79 (11)
50	8.36 (2)	3.49 (5)	2.65 (8)	0.67 (12)	9.28 (3)	4.00 (5)	3.29 (8)	1.71 (16)	9.11 (3)	3.88 (5)	3.57 (8)	1.82 (14)
100	8.06 (3)	3.17 (8)	1.99 (9)	0.18 (16)	8.61 (3)	3.74 (8)	2.79 (9)	1.07 (17)	8.93 (3)	3.74 (8)	3.38 (9)	1.27 (17)

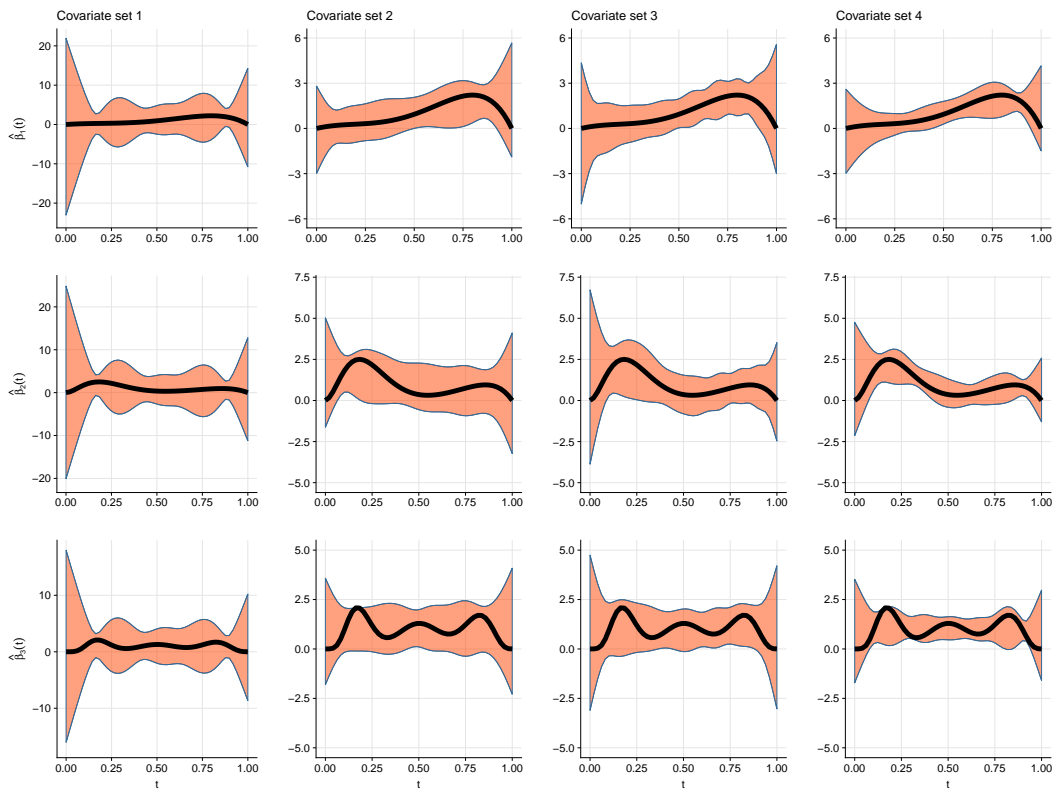


Figure 3. In each panel, the solid line represents the true parameter curve. The orange shaded region corresponds to the approximate expected value plus or minus two times the approximate standard deviation of the parameter curve estimator.

Table 5. The average value of (6) computed across all simulated datasets for each simulation setting under parameter curve $\beta_1(t)$.

		n	$\zeta(x(t))$	CV	GCV	$GCV_{\gamma=1.4}$	AIC_C	SIC
Covariate Set 1	$\kappa = 0.10$	25	3	2.71	2.71	2.71	2.71	2.71
		50	3	3.00	3.00	3.00	3.00	3.00
		100	3	3.00	3.00	3.00	3.00	3.00
	$\kappa = 0.20$	25	2	2.00	2.00	2.00	2.00	2.00
		50	2	2.49	2.50	2.49	2.49	2.49
		100	3	3.00	3.00	3.00	3.00	3.00
Covariate Set 2	$\kappa = 0.10$	25	6	6.16	6.19	6.05	6.08	6.08
		50	9	9.21	9.22	9.14	9.19	9.10
		100	14	14.18	14.18	14.14	14.17	14.09
	$\kappa = 0.20$	25	3	3.76	3.77	3.72	3.73	3.72
		50	5	5.33	5.33	5.30	5.32	5.29
		100	8	7.87	7.88	7.86	7.87	7.84
Covariate Set 3	$\kappa = 0.10$	25	8	7.92	7.99	7.82	7.85	7.90
		50	10	9.78	9.78	9.74	9.76	9.72
		100	10	10.32	10.32	10.30	10.31	10.27
	$\kappa = 0.20$	25	7	6.80	6.82	6.76	6.77	6.78
		50	8	7.74	7.75	7.69	7.72	7.65
		100	9	9.00	9.00	9.00	9.00	9.00
Covariate Set 4	$\kappa = 0.10$	25	14	14.84	14.87	14.58	14.61	14.76
		50	17	17.43	17.43	17.30	17.36	17.25
		100	21	21.19	21.19	21.14	21.17	21.11
	$\kappa = 0.20$	25	10	10.32	10.35	10.07	10.10	10.19
		50	12	12.99	13.00	12.87	12.93	12.80
		100	16	15.99	15.99	15.96	15.98	15.92

Table 6. Analogous to Table 5 for parameter curve $\beta_2(t)$.

		n	$\zeta(x(t))$	CV	GCV	GCV _{$\gamma=1.4$}	AIC _C	SIC
Covariate Set 1	$\kappa = 0.10$	25	3	2.96	2.96	2.96	2.96	2.96
		50	3	3.00	3.00	3.00	3.00	3.00
		100	3	3.00	3.00	3.00	3.00	3.00
	$\kappa = 0.20$	25	2	2.01	2.01	2.01	2.01	2.01
		50	3	2.99	2.99	2.99	2.99	2.99
		100	3	3.00	3.00	3.00	3.00	3.00
Covariate Set 2	$\kappa = 0.10$	25	6	6.18	6.17	6.06	6.08	6.08
		50	9	8.96	8.96	8.88	8.92	8.84
		100	14	14.01	14.01	13.96	14.00	13.88
	$\kappa = 0.20$	25	3	3.74	3.75	3.69	3.70	3.70
		50	5	5.17	5.16	5.12	5.14	5.10
		100	8	7.69	7.69	7.67	7.69	7.65
Covariate Set 3	$\kappa = 0.10$	25	8	8.08	8.16	8.01	8.03	8.09
		50	10	9.81	9.81	9.77	9.79	9.75
		100	10	10.34	10.34	10.32	10.33	10.29
	$\kappa = 0.20$	25	7	6.85	6.87	6.83	6.84	6.84
		50	8	7.81	7.82	7.75	7.79	7.69
		100	9	9.00	9.00	9.00	9.00	9.00
Covariate Set 4	$\kappa = 0.10$	25	16	16.62	16.65	16.28	16.31	16.64
		50	20	20.33	20.34	20.13	20.21	20.08
		100	23	23.35	23.35	23.30	23.32	23.25
	$\kappa = 0.20$	25	12	12.68	12.71	12.37	12.41	12.55
		50	16	15.48	15.49	15.32	15.40	15.24
		100	17	17.68	17.68	17.58	17.64	17.48

Table 7. Analogous to Table 5 for parameter curve $\beta_3(t)$.

		n	$\zeta(x(t))$	CV	GCV	GCV _{$\gamma=1.4$}	AIC _C	SIC
Covariate Set 1	$\kappa = 0.10$	25	3	2.93	2.93	2.93	2.93	2.93
		50	3	3.00	3.00	3.00	3.00	3.00
		100	3	3.00	3.00	3.00	3.00	3.00
	$\kappa = 0.20$	25	2	2.00	2.00	2.00	2.00	2.00
		50	3	2.90	2.90	2.90	2.90	2.90
		100	3	3.00	3.00	3.00	3.00	3.00
Covariate Set 2	$\kappa = 0.10$	25	6	6.11	6.12	6.02	6.04	6.04
		50	9	9.08	9.08	9.02	9.05	9.00
		100	14	14.05	14.05	14.00	14.04	13.97
	$\kappa = 0.20$	25	3	3.77	3.77	3.72	3.73	3.73
		50	5	5.22	5.22	5.18	5.20	5.17
		100	8	7.75	7.75	7.73	7.75	7.72
Covariate Set 3	$\kappa = 0.10$	25	8	7.87	7.94	7.77	7.79	7.84
		50	10	9.77	9.78	9.71	9.75	9.65
		100	10	10.32	10.32	10.29	10.31	10.25
	$\kappa = 0.20$	25	7	6.82	6.83	6.79	6.80	6.80
		50	8	7.77	7.78	7.70	7.74	7.66
		100	9	9.00	9.00	9.00	9.00	9.00
Covariate Set 4	$\kappa = 0.10$	25	15	15.71	15.75	15.38	15.41	15.59
		50	19	18.98	19.00	18.73	18.86	18.63
		100	23	22.59	22.60	22.47	22.56	22.26
	$\kappa = 0.20$	25	11	11.70	11.73	11.45	11.48	11.56
		50	14	14.36	14.36	14.26	14.30	14.22
		100	17	16.94	16.94	16.91	16.93	16.88

6. A Real Data Illustration

In this section, we illustrate the use of $\zeta(\mathbf{x}(t))$ with two real datasets: the gasoline dataset described by Kokoszka and Reimherr [2] and used by Reiss and Ogden [18] and the streamflow and precipitation data described by Masselot et al. [34]. Model (1) was applied to both datasets to model the relationship between a functional covariate and scalar response. The gasoline data consist of near-infrared reflectance spectra of 60 gasoline samples (measured in 2-nm intervals from 900 to 1700 nm), as well as the octane numbers for each gasoline sample. These data are available in the R package refund [35]. The aim of modeling these data using model (1) is to determine the association between the octane rating (response variable) and the near-infrared reflectance spectra curves (covariate curves). We represent the near-infrared reflectance spectra discretized measurements and the parameter curve as functions using the methods described in Section 3, with $J_x = 50$ and $J_\beta = 50$. The top left panel of Figure 4 contains the near-infrared reflectance spectra curves, $x_i(\omega)$ for $i = 1, \dots, 60$. The estimated number of independent pieces of information in these covariate functions for estimating parameter curve is 5. Since $\hat{\zeta}(\mathbf{x}(t)) = 5$, we use the *SIC* for smoothing parameter selection because it performed just as well or better in our simulation study when $\zeta(\mathbf{x}(t))$ was small. The upper right panel shows the parameter curve estimate along with 95% point-wise confidence intervals for $\beta(\omega)$ when using *SIC* for smoothing parameter selection. Note that the estimated parameter curve has a positive effect in the intervals (950, 1125) and (1325, 1475), implying that higher values of near-infrared reflectance spectra are associated with higher octane levels in these intervals. Lower values of near-infrared reflectance spectra are associated with lower octane levels in the intervals (1175, 1325) and (1525, 1650). For a given λ , approximate point-wise confidence intervals may be constructed using the variance of the parameter curve estimator, $\hat{\sigma}^2 \mathbf{N}_\beta^T(t) \mathbf{T} \mathbf{T}^T \mathbf{N}_\beta(t)$, where $\mathbf{T} = (\mathbf{M} \mathbf{M}^T + \lambda \mathbf{R})^{-1} \mathbf{M}^T$ [16]. Since $\zeta(\mathbf{x}(t))$ is low, our simulation study suggest that the variability of the parameter curve is greater than what is reflected by the confidence interval.

We briefly summarize the streamflow and precipitation data, referring to Masselot et al. [34] for further information on the study and the corresponding data. The data consist of yearly observations of the sum daily streamflow values from 1 July to 31 October, and yearly precipitation time series from 1 June to 31 October for years 1981–2012. These data were measured in areas of the Dartmouth River located in a region of the province of Quebec, Canada. In this study, investigators were interested in estimating and forecasting yearly total streamflow (scalar response) using the corresponding yearly precipitation profile (functional covariate). The precipitation time series and the parameter curve are represented as functions using the methods described in Section 3, with $J_x = 153$ and $J_\beta = 22$. However, since precipitation measurements are non-negative, we constrain the functional representation of the precipitation measurements to be non-negative by imposing non-negative constraints on the B-spline coefficients. The bottom left panel of Figure 4 contains the precipitation curves, $x_i(t)$ for $i = 1, \dots, 153$, covering the daily time domain from June to October in a given year. For these data, $\hat{\zeta}(\mathbf{x}(t)) = 29$. We use the *AICc* since it performed just as well or better than the other methods in our simulation study when $\zeta(\mathbf{x}(t))$ was large. The lower right panel shows the parameter curve estimate along with 95% point-wise confidence intervals for $\beta(t)$ when using *AICc* for smoothing parameter selection. The estimated parameter curve shows that the effect of precipitation on total streamflow is negative in June, as well as in October. Due to the size of the $\hat{\zeta}(\mathbf{x}(t))$, our simulation study suggests that this parameter curve estimate is more stable than the one estimated for the gasoline data.

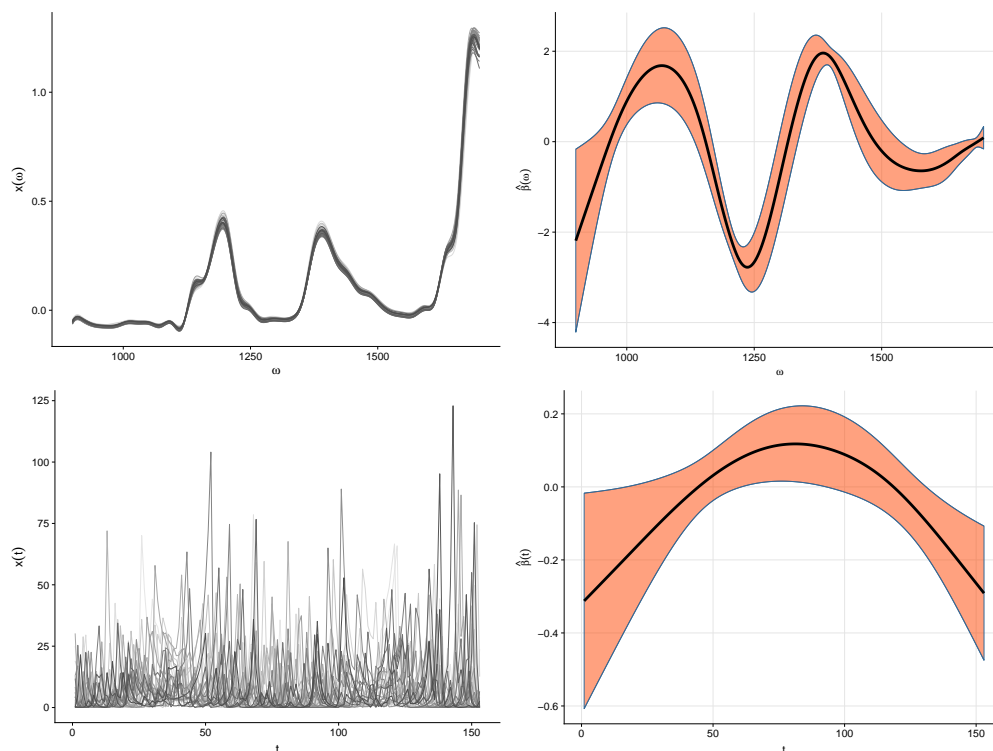


Figure 4. (top left) The reflectance spectra curves; (top right) the parameter curve estimate displaying the association between the octane rating and the near-infrared reflectance spectra curves is shown in the, together with a 95% point-wise confidence interval; and (bottom) analogous graphs for the streamflow and precipitation data.

7. Discussion

We present a measure, $\zeta(\mathbf{x}(t))$, for a FLM with a scalar response to determine how much information is present in the covariate curves for estimating the parameter curve $\beta(t)$ when the parameter curve is identifiable. To estimate the parameter curve in model (1), penalized regression spline estimation is used, and we summarize several commonly used methods for selecting the smoothing parameter. To assess the stability of the parameter curve estimator under varying levels of $\zeta(\mathbf{x}(t))$, we examine the $SMISE_{(n)}$ of a parameter curve estimator. The results show that the greater is $\zeta(\mathbf{x}(t))$, the more stable is the parameter curve estimator in that it produces a smaller $SMISE_{(n)}$ than when $\zeta(\mathbf{x}(t))$ is smaller. Further, we assess the impact of $\zeta(\mathbf{x}(t))$ on smoothing parameter selection, and, while a one-to-one relationship is not clear between $\zeta(\mathbf{x}(t))$ and the performance of a smoothing selection method, SIC tends to perform just as well as or better than other methods when $\zeta(\mathbf{x}(t))$ is small, whereas $AICc$ tends to perform just as well as or better than other methods when $\zeta(\mathbf{x}(t))$ is large.

Overall, our simulation study showed that the size of $\zeta(\mathbf{x}(t))$ impacts both the stability of a parameter curve estimator and the performance of the smoothing parameter selection methods. Future work will study if these results are consistent under alternative parameter curve estimation procedures. An interesting direction for future work is to investigate if shape constraints on the parameter curve could serve as a remedial measure for improving stability of the parameter curve estimator, particularly when $\zeta(\mathbf{x}(t))$ is low. Scenarios in which shape constraints are imposed on the parameter curve do arise in practice in a FLM with a scalar response (see [36,37] for some recent examples). Identifying problematic data in functional regression models remains critical and an on-going challenge. We hope this study encourages others to consider approximating $\zeta(\mathbf{x}(t))$ when applying model (1) so that the amount of information present in the covariate curves for estimating

parameter curve can be gauged. This, in turn, may provide guidance in choosing a smoothing parameter selection method, as well as considering the stability of the parameter curve estimate.

Funding: This research was supported in part by NSF grant HRD-1547784.

Acknowledgments: The reviewers of this manuscript are thanked for their helpful comments. The author expresses his gratitude to Pierre Masselot for sharing the streamflow and precipitation data.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Hsing, T.; Eubank, R. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*; John Wiley & Sons: Chichester, UK, 2015.
2. Kokoszka, P.; Reimherr, M. *Introduction to Functional Data Analysis*, 1st ed.; Chapman and Hall: Boca Raton, FL, USA, 2017.
3. Reiss, P.T.; Goldsmith, J.; Shang, H.L.; Ogden, R.T. Methods for Scalar-on-Function Regression. *Int. Stat. Rev.* **2017**, *85*, 228–249. [[CrossRef](#)] [[PubMed](#)]
4. García-Portugués, E.; González-Manteiga, W.; Febrero-Bande, M. A goodness-of-fit test for the functional linear model with scalar response. *J. Comput. Graph. Stat.* **2014**, *23*, 761–778. [[CrossRef](#)]
5. McLean, M.W.; Hooker, G.; Ruppert, D. Restricted likelihood ratio tests for linearity in scalar-on-function regression. *Stat. Comput.* **2015**, *25*, 997–1008. [[CrossRef](#)]
6. Tekbudak, M.Y.; Alfaro-Córdoba, M.; Maity, A.; Staicu, A.M. A comparison of testing methods in scalar-on-function regression. *arXiv* **2017**, arXiv:1710.05729.
7. Cook, R.D. Detection of influential observation in linear regression. *Technometrics* **1977**, *19*, 15–18.
8. Peña, D. A new statistic for influence in linear regression. *Technometrics* **2005**, *47*, 1–12. [[CrossRef](#)]
9. Febrero-Bande, M.; Galeano, P.; González-Manteiga, W. Measures of influence for the functional linear model with scalar response. *J. Multivar. Anal.* **2010**, *101*, 327–339. [[CrossRef](#)]
10. Groetsch, C.W. *Inverse Problems in the Mathematical Sciences*; Springer: Brunswick, GA, USA, 1993; Volume 52.
11. Cardot, H.; Ferraty, F.; Sarda, P. Spline estimators for the functional linear model. *Stat. Sin.* **2003**, *13*, 571–591.
12. Happ, C.; Greven, S.; Schmid, V.J. The impact of model assumptions in scalar-on-image regression. *Stat. Med.* **2018**, *37*, 4298–4317. [[CrossRef](#)]
13. Wahba, G. *Ill Posed Problems: Numerical and Statistical Methods for Mildly, Moderately and Severely Ill Posed Problems with Noisy Data*; Technical Report; Department of Statistics, University of Wisconsin: Madison, WI, USA, 1980.
14. Wickham, H. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*; R Package Version 3.3.1; 2019.
15. Wilke, C.O. *Cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'*; R Package Version 0.9.4; 2019.
16. Ramsay, J.O.; Silverman, B.W. *Functional Data Analysis*, 2nd ed.; Springer: New York, NY, USA, 2005.
17. Crambes, C.; Kneip, A.; Sarda, P. Smoothing splines estimators for functional linear regression. *Ann. Stat.* **2009**, *37*, 35–72. [[CrossRef](#)]
18. Reiss, P.T.; Ogden, R.T. Smoothing parameter selection for a class of semiparametric linear models. *J. R. Stat. Soc. Ser. B* **2009**, *71*, 505–523.
19. Shin, H.; Lee, S. An RKHS approach to robust functional linear regression. *Stat. Sin.* **2016**, *26*, 255–272. [[CrossRef](#)]
20. Craven, P.; Wahba, W. Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation. *Numer. Math.* **1979**, *31*, 377–403. [[CrossRef](#)]
21. Buja, A.; Hastie, T.; Tibshirani, R. Linear smoothers and additive models. *Ann. Stat.* **1989**, *17*, 453–510. [[CrossRef](#)]
22. Cummins, D.J.; Filloon, T.G.; Nychka, D. Confidence intervals for nonparametric curve estimates: Toward more uniform pointwise coverage. *J. Am. Stat. Assoc.* **2001**, *96*, 233–246. [[CrossRef](#)]
23. Kim, Y.J.; Gu, C. Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *J. R. Stat. Soc. Ser. B* **2004**, *66*, 337–356. [[CrossRef](#)]
24. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [[CrossRef](#)]

25. Hurvich, C.M.; Simonoff, J.S.; Tsai, C.L. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Stat. Soc. Ser. B* **1998**, *60*, 271–293. [[CrossRef](#)]
26. Lee, T.C. Smoothing parameter selection for smoothing splines: A simulation study. *Comput. Stat. Data Anal.* **2003**, *42*, 139–148. [[CrossRef](#)]
27. Belsley, D.A.; Kuh, E.; Welsch, R.E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; John Wiley & Sons: New York, NY, USA, 2005; Volume 571.
28. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018.
29. Ramsay, J.O.; Wickham, H.; Graves, S.; Hooker, G. *Fda: Functional Data Analysis*; R Package Version 2.4.7; 2017.
30. Wickham, H.; François, R.; Henry, L.; Müller, K. *Dplyr: A Grammar of Data Manipulation*; R Package Version 0.7.8; 2018.
31. Wickham, H.; Henry, L. *Tidyr: Easily Tidy Data with ‘Spread()’ and ‘Gather()’ Functions*; R Package Version 0.8.2; 2018.
32. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300. [[CrossRef](#)]
33. Hervé, M. *RVAideMemoire: Testing and Plotting Procedures for Biostatistics*; R Package Version 0.9-73; 2019.
34. Masselot, P.; Dabo-Niang, S.; Chebana, F.; Ouarda, T.B. Streamflow forecasting using functional regression. *J. Hydrol.* **2016**, *538*, 754–766.
35. Goldsmith, J.; Scheipl, F.; Huang, L.; Wrobel, J.; Gellar, J.; Harezlak, J.; McLean, M.W.; Swihart, B.; Xiao, L.; Crainiceanu, C.; et al. *Refund: Regression with Functional Data*; R Package Version 0.1-17; 2018.
36. Schipper, M.; Taylor, J.M.; Lin, X. Generalized monotonic functional mixed models with application to modelling normal tissue complications. *J. R. Stat. Soc. Ser. C* **2008**, *57*, 149–163. [[CrossRef](#)]
37. Montoya, E.L.; Meiring, W. On the relative efficiency of a monotone parameter curve estimator in a functional nonlinear model. *Stat. Comput.* **2013**, *23*, 425–436.

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).