The Annals of Statistics
2016, Vol. 44, No. 1, 1–30
DOI: 10.1214/15-AOS1323
© Institute of Mathematical Statistics, 2016

FUNCTIONAL LINEAR REGRESSION WITH POINTS OF IMPACT

BY ALOIS KNEIP[1], DOMINIK POSS[2] AND PASCAL SARDA

Universität Bonn and Institut de Mathématiques de Toulouse

The paper considers functional linear regression, where scalar responses $Y_1, \ldots, Y_n$ are modeled in dependence of i.i.d. random functions $X_1, \ldots, X_n$. We study a generalization of the classical functional linear regression model. It is assumed that there exists an unknown number of "points of impact," that is, discrete observation times where the corresponding functional values possess significant influences on the response variable. In addition to estimating a functional slope parameter, the problem then is to determine the number and locations of points of impact as well as corresponding regression coefficients. Identifiability of the generalized model is considered in detail. It is shown that points of impact are identifiable if the underlying process generating $X_1, \ldots, X_n$ possesses "specific local variation." Examples are well-known processes like the Brownian motion, fractional Brownian motion or the Ornstein–Uhlenbeck process. The paper then proposes an easily implementable method for estimating the number and locations of points of impact. It is shown that this number can be estimated consistently. Furthermore, rates of convergence for location estimates, regression coefficients and the slope parameter are derived. Finally, some simulation results as well as a real data application are presented.

1. Introduction. We consider linear regression involving a scalar response variable Y and a functional predictor variable $X \in L^2([a,b])$, where $[a,b]$ is a bounded interval of $\mathbb{R}$. It is assumed that data consist of an i.i.d. sample (X_i, Y_i), $i = 1, \ldots, n$, from (X, Y). The functional variable X is such that $\mathbb{E}(\int_a^b X^2(t)\,dt) < +\infty$ and for simplicity the variables are supposed to be centered in the following: $\mathbb{E}(Y) = 0$ and $\mathbb{E}(X(t)) = 0$ for $t \in [a,b]$ a.e.

In this paper, we study the following *functional linear regression model with points of impact*

$$Y_i = \int_a^b \beta(t) X_i(t)\,dt + \sum_{r=1}^{S} \beta_r X_i(\tau_r) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where ε_i, $i = 1, \ldots, n$ are i.i.d. centered real random variables with $\mathbb{E}(\varepsilon_i^2) = \sigma^2 < \infty$, which are independent of $X_i(t)$ for all t, $\beta \in L^2([a,b])$ is an unknown,

Received July 2014; revised February 2015.

[1]Supported by the DFG through GRK 1707 and the Hausdorff Center for Mathematics.

[2]Supported by the DFG through GRK 1707.

MSC2010 subject classifications. Primary 62G08, 62M99; secondary 62J05.

Key words and phrases. Functional linear regression, model selection, stochastic processes, nonstandard asymptotics.

bounded slope function and $\int_a^b \beta(t)X_i(t)\,dt$ describes a common effect of the whole trajectory $X_i(\cdot)$ on Y_i. In addition, the model incorporates an unknown number $S \in \mathbb{N}$ of "points of impact," that is, *specific* time points $\tau_1, \ldots, \tau_S$ with the property that the corresponding functional values $X_i(\tau_1), \ldots, X_i(\tau_S)$ possess some significant influence on the response variable Y_i. The function $\beta(t)$, the number $S \geq 0$, as well as τ_r and β_r, $r = 1, \ldots, S$, are unknown and have to be estimated from the data. Throughout the paper, we will assume that all points of impact are in the interior of the interval, $\tau_r \in (a, b)$, $r = 1, \ldots, S$. Standard functional linear regression with $S = 0$ as well as the point impact model of McKeague and Sen (2010), which assumes $\beta(t) \equiv 0$ and $S = 1$, are special cases of the above model.

If $S = 0$, then (1.1) reduces to $Y_i = \int_a^b \beta(t)X_i(t)\,dt + \varepsilon_i$. This model has been studied in depth in theoretical and applied statistical literature. The most frequently used approach for estimating $\beta(t)$ then is based on functional principal components regression [see, e.g., Frank and Friedman (1993), Bosq (2000), Cardot, Ferraty and Sarda (1999), Cardot, Mas and Sarda (2007) or Müller and Stadtmüller (2005) in the context of generalized linear models]. Rates of convergence of the estimates are derived in Hall and Horowitz (2007) and Cai and Hall (2006). Alternative approaches and further theoretical results can, for example, be found in Crambes, Kneip and Sarda (2009), Cardot and Johannes (2010), Comte and Johannes (2012) or Delaigle and Hall (2012).

There are many successful applications of the standard linear functional regression model. At the same time, results are often difficult to analyze from the points of view of model building and substantial interpretation. The underlying problem is that $\int_a^b \beta(t)X_i(t)\,dt$ is a weighted average of the whole trajectory $X_i(\cdot)$ which makes it difficult to assess specific effects of local characteristics of the process. This lead James, Wang and Zhu (2009) to consider "interpretable functional regression" by assuming that $\beta(t) = 0$ for most points $t \in [a, b]$ and identifying subintervals of $[a, b]$ with nonzero $\beta(t)$.

A different approach based on impact points is proposed by Ferraty, Hall and Vieu (2010). For a pre-specified $q \in \mathbb{N}$, they aim to identify a function g as well as those design points $\tau_1, \ldots, \tau_q$ which are "most influential" in the sense that $g(X_i(\tau_1), \ldots, X_i(\tau_q))$ provides a best possible prediction of Y_i. Nonparametric smoothing methods are used to estimate g, while $\tau_1, \ldots, \tau_q$ are selected by a cross-validation procedure. The method is applied to data from spectroscopy, where it is of practical interest to know which values $X_i(t)$ have greatest influence on Y_i.

To our knowledge, McKeague and Sen (2010) are the first to explicitly study identifiability and estimation of a point of impact in a functional regression model. For centered variables, their model takes the form $Y_i = \beta X_i(\tau) + \varepsilon_i$ with a single point of impact $\tau \in [a, b]$. The underlying process X is assumed to be a fractional Brownian motion with Hurst parameter H. The approach is motivated by the analysis of gene expression data, where a key problem is to identify individual genes associated with the clinical outcome. McKeague and Sen (2010) show that consistent estimators are obtained by least squares, and that the estimator of τ has the

rate of convergence $n^{-1/(2H)}$. The coefficient β can be estimated with a parametric rate of convergence $n^{-1/2}$.

There also exists a link between our approach and the work of Hsing and Ren (2009) who for a given grid $t_1, \ldots, t_p$ of observation points propose a procedure for estimating linear combinations $m(X_i) = \sum_{j=1}^{p} c_j X_i(t_j)$ influencing Y_i. Their approach is based on an RKHS formulation of the inverse regression dimension-reduction problem which for any $k = 1, 2, 3, \ldots$ allows to determine a suitable element $(\hat{c}_1, \ldots, \hat{c}_p)^T$ of the eigenspace spanned by the eigenvectors of the k leading eigenvalues of the empirical covariance matrix of $(X_i(t_1), \ldots, X_i(t_p))^T$. They then show consistency of the resulting estimators $\hat{m}(X_i)$ as $n, p \to \infty$ and then $k \to \infty$. Note that (1.1) necessarily implies that $Y_i = m(X_i) + \varepsilon_i$, where as $p \to \infty$ $m(X_i)$ may be written as a linear combination as considered by Hsing and Ren (2009). Their method therefore offers a way to determine consistent estimators $\hat{m}(X_i)$ of $m(X_i)$, although the structure of the estimator will not allow a straightforward identification of model components.

Assuming a linear relationship between Y and X, (1.1) constitutes a unified approach which incorporates the standard linear regression model as well as specific effects of possible point of impacts. The latter may be of substantial interest in many applications.

Although in this paper we concentrate on the case of unknown points of impact, we want to emphasize that in practice also models with pre-specified points of impact may be of potential importance. This in particular applies to situations with a functional response variable $\mathcal{Y}_i(t)$, defined over the same time period $t \in [a, b]$ as X_i. For a specified time point $\tau \in [a, b]$, the standard approach [see, e.g., He, Müller and Wang (2000)] will then assume that $Y_i := \mathcal{Y}_i(\tau) = \int_a^b \beta_\tau(t) X_i(t)\, dt + \varepsilon_i$, where $\beta_\tau \in L^2([a, b])$ may vary with τ. But the value $X_i(\tau)$ of X_i at the point τ of interest may have a specific influence, and the alternative model $Y_i := \mathcal{Y}_i(\tau) = \int_a^b \beta_\tau(t) X_i(t)\, dt + \beta_1 X_i(\tau) + \varepsilon_i$ with $S = 1$ and a fixed point of impact may be seen as a promising alternative. The estimation procedure proposed in Section 5 can also be applied in this situation, and theoretical results imply that under mild conditions β_1 as well as $\beta_\tau(t)$ can be consistently estimated with nonparametric rates of convergence. A similar modification may be applied in the related context of functional autoregression, where $X_1, \ldots, X_n$ denote a stationary time series of random function, and $\mathcal{Y}(\tau) \equiv X_i(\tau)$ is to be predicted from X_{i-1} [see, e.g., Bosq (2000)].

The focus of our work lies on developing conditions ensuring identifiability of the components of model (1.1) as well as on determining procedures for estimating number and locations of points of impact, regression coefficients and slope parameter.

The problem of identifiability is studied in detail in Section 2. The key assumption is that the process possesses "specific local variation." Intuitively, this means that at least some part of the local variation of $X(t)$ in a small neighborhood

$[\tau-\epsilon,\tau+\epsilon]$ of a point $\tau\in[a,b]$ is essentially uncorrelated with the remainder of the trajectories outside the interval $[\tau-\epsilon,\tau+\epsilon]$. Model (1.1) is uniquely identified for all processes exhibiting specific local variation. It is also shown that the condition of specific local variation is surprisingly weak and only requires some suitable approximation properties of the corresponding Karhunen–Loève basis.

Identifiability of (1.1) does not impose any restriction on the degree of smoothness of the random functions X_i or of the underlying covariance function. The same is true for the theoretical results of Section 5 which yield rates of convergence of coefficient estimates, provided that points of impact are known or that locations can be estimated with sufficient accuracy.

But nonsmooth trajectories are advantageous when trying to identify points of impact. In order to define a procedure for estimating number and locations of points of impact, we therefore restrict attention to processes whose covariance function is nonsmooth at the diagonal. It is proved in Section 3 that any such process has specific local variation. Prominent examples are the fractional Brownian motion or the Ornstein–Uhlenbeck process. From a practical point of view, the setting of processes with nonsmooth trajectories covers a wide range of applications. Examples are given in Section 7 and in the supplementary material [Kneip, Poss and Sarda (2015)], where the methodology is applied to temperature curves and near infrared data.

An easily implementable and computationally efficient algorithm for estimating number and locations of points of impact is presented in Section 4. The basic idea is to perform a decorrelation. Instead of regressing on $X_i(t)$, we analyze the empirical correlation between Y_i and a process $Z_{\delta,i}(t):=X_i(t)-\frac{1}{2}(X_i(t-\delta)+X_i(t+\delta))$ for some $\delta>0$. For the class of processes defined in Section 3, $Z_{\delta,i}(t)$ is highly correlated with $X_i(t)$ but only possesses extremely weak correlations with $X_i(s)$ if $|t-s|$ is large. This implies that under model (1.1) local maxima $\widehat{\tau}_r$ of the empirical correlation between Y_i and $Z_{\delta,i}(t)$ should be found at locations close to existing points of impact. The number S is then estimated by a cut-off criterion. It is proved that the resulting estimator $\widehat{S}$ of S is consistent, and we derive rates of convergence for the estimators $\widehat{\tau}_r$. In the special case of a fractional Brownian motion and $S=1$, we retrieve the basic results of McKeague and Sen (2010).

In Section 5, we introduce least squares estimates of $\beta(t)$ and β_r, $r=1,\dots,S$, based on a Karhunen–Loève decomposition. Rates of convergence for these estimates are then derived. A simulation study is performed in Section 6, while applications to a dataset is presented in Section 7. The Appendix is devoted to the proofs of some of the main results. The remaining proofs as well as the application of our method to a second dataset are gathered in the supplementary material.

2. Identifiability. Our setup implies that $X_1,\dots,X_n$ are i.i.d. random functions with the same distribution as a generic $X\in L^2([a,b])$. In the following, we will additionally assume that X possesses a continuous covariance function $\sigma(t,s)$, $t,s\in[a,b]$.

In a natural way, the components of model (1.1) possess different interpretations. The linear functional $\int_a^b \beta(t)X_i(t)\,dt$ describes a *common effect* of the whole trajectory $X_i(\cdot)$ on Y_i. The additional terms $\sum_{r=1}^S \beta_r X_i(\tau_r)$ quantify *specific effects* of the functional values $X_i(\tau_1), \ldots, X_i(\tau_S)$ at the points of impact $\tau_1, \ldots, \tau_S$. Identifiability of an impact point τ_r quite obviously requires that at least some part of the local variation of $X_i(t)$ in small neighborhoods of τ_r, is uncorrelated with the remainder of the trajectories. This idea is formalized by introducing the concept of "specific local variation."

DEFINITION 1. A process $X \in L^2([a,b])$ with continuous covariance function $\sigma(\cdot,\cdot)$ possesses *specific local variation* if for any $t \in (a,b)$ and all sufficiently small $\epsilon > 0$ there exists a real random variable $\zeta_{\epsilon,t}(X)$ such that with $f_{\epsilon,t}(s) := \frac{\operatorname{cov}(X(s),\zeta_{\epsilon,t}(X))}{\operatorname{var}(\zeta_{\epsilon,t}(X))}$ the following conditions are satisfied:

(i) $0 < \operatorname{var}(\zeta_{\epsilon,t}(X)) < \infty$,
(ii) $f_{\epsilon,t}(t) > 0$,
(iii) $|f_{\epsilon,t}(s)| \le (1+\epsilon) f_{\epsilon,t}(t)$ for all $s \in [a,b]$,
(iv) $|f_{\epsilon,t}(s)| \le \epsilon \cdot f_{\epsilon,t}(t)$ for all $s \in [a,b]$ with $s \notin [t-\epsilon, t+\epsilon]$.

The definition of course implies that for given $t \in (a,b)$ and small $\epsilon > 0$ any process X with specific local variation can be decomposed into

$$X(s) = X_{\epsilon,t}(s) + \zeta_{\epsilon,t}(X) f_{\epsilon,t}(s), \qquad s \in [a,b], \tag{2.1}$$

where $X_{\epsilon,t}(s) = X(s) - \zeta_{\epsilon,t}(X) f_{\epsilon,t}(s)$ is a process which is uncorrelated with $\zeta_{\epsilon,t}(X)$. If $\sigma_{\epsilon,t}(\cdot,\cdot)$ denotes the covariance function of $X_{\epsilon,t}(s)$, then obviously

$$\sigma(s,u) = \sigma_{\epsilon,t}(s,u) + \operatorname{var}\big(\zeta_{\epsilon,t}(X)\big) f_{\epsilon,t}(s) f_{\epsilon,t}(u), \qquad s,u \in [a,b]. \tag{2.2}$$

By condition (iv), we can infer that for small $\epsilon > 0$ the component $\zeta_{\epsilon,t}(X) f_{\epsilon,t}(s)$ essentially quantifies local variation in a small interval around the given point t, since $\frac{f_{\epsilon,t}(s)^2}{f_{\epsilon,t}(t)^2} \le \epsilon^2$ for all $s \notin [t-\epsilon, t+\epsilon]$. When X is a standard Brownian motion it is easily verified that conditions (i)–(iv) are satisfied for $\zeta_{\epsilon,t}(X) = X(t) - \frac{1}{2}(X(t-\epsilon) + X(t+\epsilon))$. Then $f_{\epsilon,t}(s) := \frac{\operatorname{cov}(X(s),\zeta_{\epsilon,t}(X))}{\operatorname{var}(\zeta_{\epsilon,t}(X))} = 1$ for $t = s$, while $f_{\epsilon,t}(s) = 0$ for all $s \in [a,b]$ with $|t-s| \ge \epsilon$. Figure 1 illustrates the decomposition of $X(s)$ in $X_{\epsilon,t}(s)$ and $\zeta_{\epsilon,t}(X) f_{\epsilon,t}(s)$ for a trajectory of a Brownian motion.

The following theorem shows that under our setup all impact points in model (1.1) are uniquely identified for any process possessing specific local variation. Recall that (1.1) implies that

$$m(X) := \mathbb{E}(Y|X) = \int_a^b \beta(t)X(t)\,dt + \sum_{r=1}^S \beta_r X(\tau_r).$$

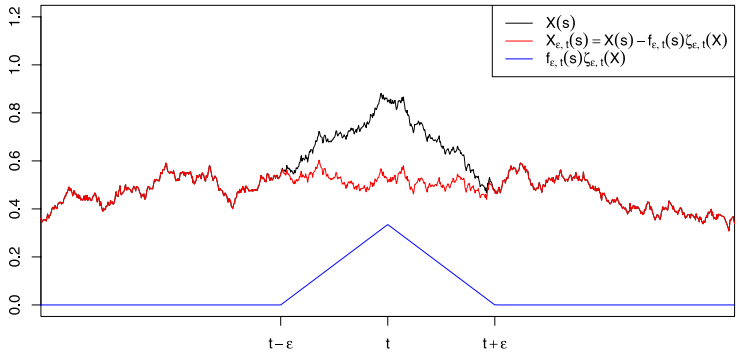

FIG. 1. *The figure illustrates the decomposition of a trajectory from a Brownian motion X (black line) in $X_{\epsilon,t}$ (red line) and $\zeta_{\epsilon,t}(X)\, f_{\epsilon,t}$ (blue line). The component $\zeta_{\epsilon,t}(X)\, f_{\epsilon,t}$ can be seen to quantify the local variation of X in an interval around t.*

THEOREM 1. *Under our setup, assume that X possesses specific local variation. Then, for any bounded function $\beta^* \in L^2([a,b])$, all $S^* \geq S$, all $\beta_1^*, \ldots, \beta_{S^*}^* \in \mathbb{R}$, and all $\tau_1, \ldots, \tau_{S^*} \in (a,b)$ with $\tau_k \notin \{\tau_1, \ldots, \tau_S\}$, $k = S+1, \ldots, S^*$, we obtain*

$$\mathbb{E}\left(\left(m(X) - \int_a^b \beta^*(t)X(t)\,dt - \sum_{r=1}^{S^*} \beta_r^* X_i(\tau_r)\right)^2\right) > 0, \tag{2.3}$$

whenever $\mathbb{E}((\int_a^b (\beta(t) - \beta^*(t))X(t)\,dt)^2) > 0$, *or* $\sup_{r=1,\ldots,S} |\beta_r - \beta_r^*| > 0$, *or* $\sup_{r=S+1,\ldots,S^*} |\beta_r^*| > 0$.

The question arises whether it is possible to find general conditions which ensure that a process possesses specific variation. From a theoretical point of view, the Karhunen–Loève decomposition provides a tool for analyzing this problem.

For $f, g \in L^2([a,b])$ let $\langle f, g\rangle = \int_a^b f(t)g(t)\,dt$ and $\|f\|$ the associated norm. We will use $\lambda_1 \geq \lambda_2 \geq \cdots$ to denote the nonzero eigenvalues of the covariance operator Γ of X, while $\psi_1, \psi_2, \ldots$ denote a corresponding system of orthonormal eigenfunctions. It is then well known that X can be decomposed in the form

$$X(t) = \sum_{r=1}^{\infty} \langle X, \psi_r\rangle \psi_r(t), \tag{2.4}$$

where $\mathbb{E}(\langle X, \psi_r\rangle^2) = \lambda_r$, and $\langle X, \psi_r\rangle$ is uncorrelated with $\langle X, \psi_l\rangle$ for $l \neq r$.

The existence of specific local variation requires that the structure of the process is not too simple in the sense that the realizations X_i a.s. lie in a finite dimensional subspace of $L^2([a,b])$. Indeed, if Γ only possesses a finite number $K < \infty$ of nonzero eigenvalues, then model (1.1) is not identifiable. This is easily verified: $X(t) = \sum_{r=1}^{K} \langle X, \psi_r\rangle \psi_r(t)$ implies that $\int_a^b \beta(t)X(t)\,dt = \sum_{r=1}^{K} \alpha_r \langle X, \psi_r\rangle$ with

$\alpha_r = \langle \psi_r, \beta \rangle$. Hence, there are infinitely many different collections of K points $\tau_1, \ldots, \tau_K$ and corresponding coefficients $\beta_1, \ldots, \beta_K$ such that

$$\int_a^b \beta(t)X(t)\,dt = \sum_{s=1}^{K} \alpha_s \langle X, \psi_s \rangle = \sum_{s=1}^{K} \langle X, \psi_s \rangle \sum_{r=1}^{K} \beta_r \psi_s(\tau_r) = \sum_{r=1}^{K} \beta_r X(\tau_r).$$

Most work in functional data analysis, however, relies on the assumption that Γ possesses infinitely many nonzero eigenvalues. In theoretically oriented papers, it is often assumed that $\psi_1, \psi_2, \ldots$ form a complete orthonormal system of $L^2([a,b])$ such that $\| \sum_{r=1}^{\infty} \langle f, \psi_r \rangle \psi_r - f \| = 0$ for any function $f \in L^2([a,b])$.

The following theorem shows that X possesses specific local variation if for a suitable class of functions L^2-convergence generalizes to L^∞-convergence.

For $t \in (a,b)$ and $\epsilon > 0$, let $\mathcal{C}(t, \epsilon, [a,b])$ denote the space of all continuous functions $f \in L^2([a,b])$ with the properties that $f(t) = \sup_{s \in [a,b]} f(s) = 1$ and $f(s) = 0$ for $s \notin [t-\epsilon, t+\epsilon]$.

THEOREM 2. *Let $\psi_1, \psi_2, \ldots$ be a system of orthonormal eigenfunctions corresponding to the nonzero eigenvalues of the covariance operator Γ of X. If for all $t \in (a,b)$ there exists an $\epsilon_t > 0$ such that*

$$\lim_{k \to \infty} \inf_{f \in \mathcal{C}(t,\epsilon,[a,b])} \sup_{s \in [a,b]} \left| f(s) - \sum_{r=1}^{k} \langle f, \psi_r \rangle \psi_r(s) \right| = 0 \tag{2.5}$$

for every $0 < \epsilon < \epsilon_t$,

then the process X possesses specific local variation.

The message of the theorem is that existence of specific local variation only requires that the underlying basis $\psi_1, \psi_2, \ldots$ possesses suitable approximation properties. Somewhat surprisingly, the degree of smoothness of the realized trajectories does not play any role.

As an example consider a standard Brownian motion defined on $[a,b] = [0,1]$. The corresponding Karhunen–Loève decomposition possesses eigenvalues $\lambda_r = \frac{1}{(r-0.5)^2\pi^2}$ and eigenfunctions $\psi_r(t) = \sqrt{2}\sin((r-1/2)\pi t)$, $r = 1, 2, \ldots$. In the Supplementary Appendix B [Kneip, Poss and Sarda (2015)], it is verified that this system of orthonormal eigenfunctions satisfies (2.5). Although all eigenfunctions are smooth, it is well known that realized trajectories of a Brownian motion are a.s. not differentiable. This can be seen as a consequence of the fact that the eigenvalues $\lambda_r \sim \frac{1}{r^2}$ decrease fairly slowly, and, therefore, the sequence $\mathbb{E}((\sum_{r=1}^{k} \langle X, \psi_r \rangle \psi_r'(t))^2) = \sum_{r=1}^{k} \lambda_r (\psi_r'(t))^2$ diverges as $k \to \infty$. At the same time, another process with the same system of eigenfunctions but exponentially decreasing eigenvalues $\lambda_r^* \sim \exp(-r)$ will a.s. show sample paths possessing an

infinite number of derivatives. Theorem 2 states that any process of this type still has specific local variation.

3. Covariance functions which are nonsmooth at the diagonal. In the following, we will concentrate on developing a theoretical framework which allows to define an efficient procedure for estimating number and locations of points of impact.

Although specific local variation may well be present for processes possessing very smooth sample paths, it is clear that detection of points of impact will profit from a high local variability which goes along with nonsmoothness. As pointed out in the Introduction, we also believe that assuming nonsmooth trajectories reflect the situation encountered in a number of important applications. McKeague and Sen (2010) convincingly demonstrate that genomics data lead to sample paths with fractal behavior. All important processes analyzed in economics exhibit strong random fluctuations. Observed temperatures or precipitation rates show wiggly trajectories over time, as can be seen in our application in Section 7. Furthermore, any growth process will to some extent be influenced by random changes in environmental conditions. In functional data analysis, it is common practice to smooth observed (discrete) sample paths and to interpret nonsmooth components as "errors." We want to emphasize that, unless observations are inaccurate and there exists some important measurement error, such components are an intrinsic part of the process. For many purposes as, for example, functional principal component analysis, smoothing makes a lot of sense since local variation has to be seen as nuisance. But in the present context local variation actually is a key property for identifying impact points.

Therefore, further development will focus on processes with nonsmooth sample paths which will be expressed in terms of a nonsmooth diagonal of the corresponding covariance function $\sigma(t,s)$. It will be assumed that $\sigma(t,s)$ possesses nonsmooth trajectories when passing from $\sigma(t,t-\Delta)$ to $\sigma(t,t+\Delta)$, but is twice continuously differentiable for all (t,s), $t \neq s$. An example is the standard Brownian motion whose covariance function $\sigma(t,s) = \min(t,s)$ has a kink at the diagonal. Indeed, in view of decomposition (2.2) a nonsmooth transition at diagonal may be seen as a natural consequence of pronounced specific local variation.

For a precise analysis, it will be useful to reparametrize the covariance function. Obviously, the symmetry of $\sigma(t,s)$ implies that

$$\begin{aligned}\sigma(t,s) &= \sigma\big(\tfrac{1}{2}\big(t+s+|t-s|\big), \tfrac{1}{2}\big(t+s-|t-s|\big)\big)\\ &=: \omega^*\big(t+s, |t-s|\big) \qquad \text{for all } t,s \in [a,b].\end{aligned}$$

Instead of $\sigma(t,s)$, we may thus equivalently consider the function $\omega^*(x,y)$ with $x = t+s$ and $y = |t-s|$. When passing from $s = t-\Delta$ to $s = t+\Delta$, the degree of smoothness of $\sigma(t,s)$ at $s=t$ is reflected by the behavior of $\omega^*(2t,y)$ as $y \to 0$.

First, consider the case that σ is twice continuously differentiable and for fixed x and $y > 0$ let $\frac{\partial}{\partial y_+}\omega^*(x, y)|_{y=0}$ denote the right (partial) derivative of $\omega^*(x, y)$ as $y \to 0$. It is easy to check that in this case for all $t \in (a, b)$ we obtain

$$\begin{aligned}\frac{\partial}{\partial y_+}\omega^*(2t, y)\Big|_{y=0} &= \frac{\partial}{\partial y}\sigma\Big(t + \frac{y}{2}, t - \frac{y}{2}\Big)\Big|_{y=0} \\ &= \frac{1}{2}\Big(\frac{\partial}{\partial s}\sigma(s, t)\Big|_{s=t} - \frac{\partial}{\partial s}\sigma(t, s)\Big|_{s=t}\Big) = 0.\end{aligned} \tag{3.1}$$

In contrast, any process with $\frac{\partial}{\partial y_+}\omega^*(x, y)|_{y=0} \neq 0$ is nonsmooth at the diagonal. If this function is smooth for all other points (x, y), $y > 0$, then the process, similar to the Brownian motion, possesses a kink at the diagonal. Now note that, for any process with $\sigma(t, s) = \omega^*(t + s, |t - s|)$ continuously differentiable for $t \neq s$ but $\frac{\partial}{\partial y_+}\omega^*(x, y)|_{y=0} < 0$, it is possible to find a twice continuously differentiable function $\omega(x, y, z)$ with $\sigma(t, s) = \omega(t, s, |t - s|)$ such that $\frac{\partial}{\partial y_+}\omega^*(t + t, y)|_{y=0} = \frac{\partial}{\partial y}\omega(t, t, y)|_{y=0}$.

In a still more general setup, the above ideas are formalized by Assumption 1 below which, as will be shown in Theorem 3, provides sufficient conditions in order to guarantee that the underlying process X possesses specific variation. We will also allow for unbounded derivatives as $|t - s| \to 0$.

ASSUMPTION 1. For some open subset $\Omega \subset \mathbb{R}^3$ with $[a, b]^2 \times [0, b - a] \subset \Omega$, there exists a twice continuously differentiable function $\omega : \Omega \to \mathbb{R}$ as well as some $0 < \kappa < 2$ such that for all $t, s \in [a, b]$

$$\sigma(t, s) = \omega\big(t, s, |t - s|^\kappa\big). \tag{3.2}$$

Moreover,

$$0 < \inf_{t\in[a,b]} c(t) \qquad \text{where } c(t) := -\frac{\partial}{\partial z}\omega(t, t, z)\Big|_{z=0}. \tag{3.3}$$

One can infer from (3.1) that for every twice continuously differentiable covariance function σ there exists some function ω such that (3.2) holds with $\kappa = 2$. But note that formally introducing $|t - s|^\kappa$ as an extra argument establishes an easy way of capturing nonsmooth behavior as $|t - s| \to 0$, since σ is not twice differentiable at the diagonal if $\kappa < 2$. In Assumption 1, the value of $\kappa < 2$ thus quantifies the degree of smoothness of σ at the diagonal. A very small κ will reflect pronounced local variability and extremely nonsmooth sample paths. There are many well-known processes satisfying this assumption.

Fractional Brownian motion with Hurst coefficient $0 < H < 1$ on an interval $[a, b]$, $a > 0$: The covariance function is then given by

$$\sigma(t, s) = \tfrac{1}{2}\big(t^{2H} + s^{2H} - |t - s|^{2H}\big).$$

In this case, Assumption 1 is satisfied with $\kappa = 2H$, $\omega(t,s,z) = \frac{1}{2}(t^{2H} + s^{2H} - z)$ and $c(t) = 1/2$.

Ornstein–Uhlenbeck process with parameters $\sigma_u^2, \theta > 0$: The covariance function is then defined by

$$\sigma(t,s) = \frac{\sigma_u^2}{2\theta}\big(\exp\big(-\theta|t-s|\big) - \exp\big(-\theta(t+s)\big)\big).$$

Then Assumption 1 is satisfied with $\kappa = 1$, $\omega(t,s,z) = \frac{\sigma_u^2}{2\theta}(\exp(-\theta z) - \exp(-\theta(t+s)))$ and $c(t) = \sigma_u^2/2$.

Theorem 3 below now states that any process respecting Assumption 1 possesses specific local variation. In Section 2, we already discussed the structure of an appropriate r.v. $\zeta_{\epsilon,t}(X)$ for the special case of a standard Brownian motion. The same type of functional may now be used in a more general setting.

For $\delta > 0$ and $[t-\delta, t+\delta] \subset [a,b]$, define

$$Z_\delta(X,t) = X(t) - \tfrac{1}{2}\big(X(t-\delta) + X(t+\delta)\big). \tag{3.4}$$

THEOREM 3. *Under our setup, assume that the covariance function σ of X satisfies Assumption 1. Then X possesses specific local variation, and for any $\epsilon > 0$ there exists a $\delta > 0$ such that conditions* (i)–(iv) *of Definition 1 are satisfied for $\zeta_{\epsilon,t}(X) = Z_\delta(X,t)$, where $Z_\delta(X,t)$ is defined by* (3.4).

4. Estimating points of impact. When analyzing model (1.1), a central problem is to estimate number and locations of points of impact. Recall that we assume an i.i.d. sample (X_i, Y_i), $i = 1,\dots,n$, where X_i possesses the same distribution as a generic X. Furthermore, we consider the case that each X_i is evaluated at p equidistant points $t_j = a + \frac{j-1}{p-1}(b-a)$, $j = 1,\dots,p$.

REMARK. Note that all variables have been assumed to have means equal to zero. Any practical application of the methodology introduced below, however, should rely on centered data to be obtained from the original data by subtracting sample means. Obviously, the theoretical results developed in this section remain unchanged for this situation with however substantially longer proofs.

Determining $\tau_1,\dots,\tau_S$ of course constitutes a model selection problem. Since in practice the random functions X_i are observed on a discretized grid of p points, one may tend to use multivariate model selection procedures like Lasso or related methods. But these procedures are multivariate in nature and are not well adapted to a functional context. An obvious difficulty is the linear functional $\int_a^b \beta(t)X_i(t)\,dt \approx \frac{1}{p}\sum_{j=1}^p \beta(t_j)X_i(t_j)$ which contradicts the usual sparseness assumption by introducing some common effects of all variables. But even if $\int_a^b \beta(t)X_i(t)\,dt \equiv 0$, results may heavily depend on the number p of observations

per function. Note that in our functional setup for any fixed $m \in \mathbb{N}$ we necessarily have $\operatorname{Var}(X_i(t_j) - X_i(t_{j-m})) \to 0$ as $p \to \infty$. Lasso theory, however, is based on the assumption that variables are not too heavily correlated. For example, the results of Bickel, Ritov and Tsybakov (2009) indicate that convergence of parameter estimates *at least* requires that $\sqrt{n/\log p}(\operatorname{Var}(X_i(t_j) - X_i(t_{j-1}))) \to \infty$ as $n \to \infty$. This follows from the distribution version of the restricted eigenvalue assumption and Theorem 5.2 of Bickel, Ritov and Tsybakov (2009) [see also Zhou, van de Geer and Bühlmann (2009) for a discussion on correlation assumptions for selection models]. As a consequence, standard multivariate model selection procedures cannot work unless the number p of grid points is sufficiently small compared to n.

In this paper, we propose a very simple approach which is based on the concepts developed in the preceding sections. The idea is to identify points of impact by determining the grid points t_j, where $Z_{\delta,i}(t_j) := Z_\delta(X_i, t_j)$ possesses a particularly high correlation with Y_i.

The motivation of this approach is easily seen when considering our regression model (1.1) more closely. Note that $Z_{\delta,i}(t)$ is strongly correlated with $X_i(t)$, but it is "almost" uncorrelated with $X_i(s)$ for $|t - s| \gg \delta$. This in turn implies that the correlation between Y_i and $Z_{\delta,i}(t)$ will be comparably high if and only if a particular point t is close to a point of impact. More precisely, Lemmas 3 and 4 in the Supplementary Appendix C [Kneip, Poss and Sarda (2015)] show that as $\delta \to 0$ and $\min_{r \neq s} |\tau_s - \tau_r| \gg \delta$

$$\mathbb{E}\big(Z_{\delta,i}(t_j)Y_i\big) = \beta_r c(\tau_r)\delta^\kappa + O\big(\max\{\delta^{\kappa+1}, \delta^2\}\big) \qquad \text{if } |t_j - \tau_r| \approx 0,$$

$$\mathbb{E}\big(Z_{\delta,i}(t_j)Y_i\big) = O\big(\max\{\delta^{\kappa+1}, \delta^2\}\big) \qquad \text{if } \min_{r=1,\dots,S} |t_j - \tau_r| \gg \delta.$$

Moreover, assuming that the process X possesses a Gaussian distribution, then since $\operatorname{Var}(Z_{\delta,i}(t_j)) = O(\delta^\kappa)$ [see (A.3) in the proof of Theorem 3], the Cauchy–Schwarz inequality lead to $\operatorname{Var}(Z_{\delta,i}(t_j)Y_i) = O(\delta^\kappa)$, and hence

$$\left| \frac{1}{n} \sum_{i=1}^{n} Z_{\delta,i}(t_j)Y_i - \mathbb{E}\big(Z_{\delta,i}(t_j)Y_i\big) \right| = O_P\left(\sqrt{\frac{\delta^\kappa}{n}}\right).$$

These arguments indicate that points of impact may be estimated by using the locations of sufficiently large local maxima of $|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(t_j)Y_i|$. A sensible identification will require a suitable choice of $\delta > 0$ in dependence of the sample size n. If δ is too large, it will not be possible to distinguish between the influence of points of impact which are close to each other. On the other hand, if δ is too small compared to n (as, e.g., $\delta^k \sim n^{-1}$), then "true" maxima may perish in a flood of random peaks.

The situation is illustrated in Figure 2. It shows a simulated example of the regression model (1.1) with $n = 5000$, $\beta(t) \equiv 0$, and $S = 5$ points of impact. The error term is standard normal, while X_i are independent realizations of an

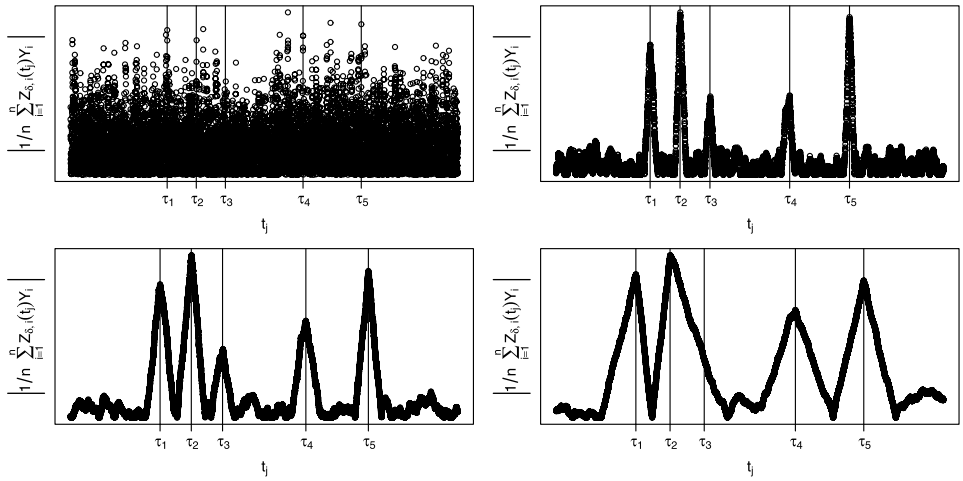

FIG. 2. *The figure shows* $|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t_j)Y_i|$ *for different choices of* δ *in a point of impact model with* 5 *points of impact whose locations are indicated by vertical lines. The upper left panel corresponds to a very small* δ*, where the noise level overlays the signal. By increasing* δ *the location of the points of impact becomes more and more visible. By choosing* δ *too large, as in the lower right panel, we are not able to distinguish between the influence of points of impact in close vicinity anymore.*

Ornstein–Uhlenbeck process with $\theta = 5$ and $\sigma_u = 3.5$, evaluated over $p = 10{,}001$ equidistant grid points in the interval $[0, 1]$. The figure shows the behavior of $|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t_j)Y_i|$ for different choices $\delta = 10/10{,}001 \approx 5/n$, $\delta = 142/10{,}001 \approx 1/\sqrt{n}$, $\delta = 350/10{,}001 \approx 2.47/\sqrt{n}$, and $\delta = 750/10{,}001 \approx 5.3/\sqrt{n}$.

In order to consistently estimate S, our estimation procedure requires to exclude all points t in an interval of size $\sqrt{\delta}$ around the local maxima of $|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t_j)Y_i|$ from further considerations. The vertical lines in Figure 2 indicate the true location of the points of impact, whereas the tick marks on the horizontal axis represent our possible candidates for τ when applying the following estimation procedure.

Estimation procedure: Choose some $\delta > 0$ such that there exists some $k_\delta \in \mathbb{N}$ with $1 \le k_\delta < \frac{p-1}{2}$ and $\delta = k_\delta(b-a)/(p-1)$. In a first step, determine for all $j \in \mathcal{J}_{0,\delta} := \{k_\delta + 1, \dots, p - k_\delta\}$

$$Z_{\delta,i}(t_j) := X_i(t_j) - \tfrac{1}{2}\big(X_i(t_j - \delta) + X_i(t_j + \delta)\big).$$

Iterate for $l = 1, 2, 3, \dots$:

- Determine
$$j_l = \arg\max_{j \in \mathcal{J}_{l-1,\delta}} \left| \frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t_j)Y_i \right|$$
and set $\widehat{\tau}_l := t_{j_l}$.
- Set $\mathcal{J}_{l,\delta} := \{j \in \mathcal{J}_{l-1,\delta} \,|\, |t_j - \widehat{\tau}_l| \ge \sqrt{\delta}/2\}$, that is, eliminate all points in an interval of size $\sqrt{\delta}$ around $\widehat{\tau}_l$. Stop iteration if $\mathcal{J}_{l,\delta} = \varnothing$.

Choose a suitable cut-off parameter $\lambda > 0$.

- Estimate S by

$$\widehat{S} = \arg \min_{l=0,1,2,\ldots} \left| \frac{(1/n)\sum_{i=1}^{n} Z_{\delta,i}(\widehat{\tau}_{l+1})Y_i}{((1/n)\sum_{i=1}^{n} Z_{\delta,i}(\widehat{\tau}_{l+1})^2)^{1/2}} \right| < \lambda.$$

- $\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{S}}$ then are the final estimates of the points of impact.

A theoretical justification for this estimation procedure is given by Theorem 4. Its proof along with the proofs of Propositions 1 and 2 below can be found in the Supplementary Appendix C. Theory relies on an asymptotics $n \to \infty$ with $p \equiv p_n \geq Ln^{1/\kappa}$ for some constant $0 < L < \infty$. It is based on the following additional assumption on the structure of X and Y.

ASSUMPTION 2. (a) $X_1, \ldots, X_n$ are i.i.d. random functions distributed according to X. The process X is *Gaussian* with covariance function $\sigma(t, s)$.

(b) The error terms $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $N(0, \sigma^2)$ r.v. which are independent of X_i.

THEOREM 4. *Under our setup and Assumptions 1 as well as 2 let $\delta \equiv \delta_n \to 0$ as $n \to \infty$ such that $\frac{n\delta^\kappa}{|\log \delta|} \to \infty$ as well as $\frac{\delta^\kappa}{n^{-\kappa+1}} \to 0$. As $n \to \infty$ we then obtain*

$$\max_{r=1,\ldots,\widehat{S}} \min_{s=1,\ldots,S} |\widehat{\tau}_r - \tau_s| = O_P\big(n^{-1/k}\big). \tag{4.1}$$

Additionally, assume that $\delta^2 = O(n^{-1})$ and that the algorithm is applied with cut-off parameter

$$\lambda \equiv \lambda_n = A\sqrt{\frac{\operatorname{Var}(Y_i)}{n} \log\left(\frac{b-a}{\delta}\right)} \qquad \text{where } A > \sqrt{2}.$$

Then

$$P(\widehat{S} = S) \to 1 \qquad \text{as } n \to \infty. \tag{4.2}$$

The theorem of course implies that the rates of convergence of the estimated points of impact depend on κ. If $\kappa = 1$ as, for example, for the Brownian motion or the Ornstein–Uhlenbeck process, then $\max_{r=1,\ldots,\widehat{S}} \min_{s=1,\ldots,S} |\widehat{\tau}_r - \tau_s| = O_P(n^{-1})$. Arbitrarily fast rates of convergence can be achieved for very nonsmooth processes with $\kappa \ll 1$.

A suitable choice of δ satisfying the requirements of the theorem for all possible $\kappa < 2$ is $\delta = Cn^{-1/2}$ for some constant C.

Recall that for $l > 1$, our algorithm requires that $\widehat{\tau}_l$ is determined only from those points t_j which are not in $\sqrt{\delta}/2$-neighborhoods of any previously selected $\widehat{\tau}_1, \ldots, \widehat{\tau}_{l-1}$. This implies that for any δ the number M_δ of iteration steps is finite,

and $M_\delta = O(\frac{b-a}{\sqrt{\delta}/2})$ is the maximal possible number of "candidate" impact points which can be detected for a fixed n and $\delta \equiv \delta_n$. The size of these intervals is due to the use of the cut-off criterion for estimating S. It can easily be seen from the proof of the theorem that in order to establish (4.1) it suffices to eliminate all points in $\delta|\log\delta|$ neighborhoods of $\widehat{\tau}_1, \ldots, \widehat{\tau}_{l-1}$ which is a much weaker restriction.

We also want to emphasize that the cut-off value provided by the theorem heavily relies on the Gaussian assumption. A different approach that may work under more general conditions is to consider all selected local maxima $\widehat{\tau}_1, \ldots, \widehat{\tau}_{M_\delta}$ and to estimate S by usual model selection criteria like BIC.

This is quite easily done if it can additionally be assumed that, in model (1.1), $\beta(t) = 0$ for all $t \in [a, b]$. One may then apply a best subset selection by regressing Y_i on all possible subsets of $X_i(\widehat{\tau}_1), \ldots, X_i(\widehat{\tau}_{M_\delta})$, and by calculating the residual sum of squares RSS_s for each subset of size s. An estimate $\widehat{S}$ is obtained by minimizing

$$BIC_s = n\log(RSS_s/n) + s\log(n) \tag{4.3}$$

over all possible values of s.

If $\int_a^b \beta(t)X_i(t)\,dt \neq 0$, this approach will of course lead to biased results, since part of the influence of this component on the response variable Y_i may be approximated by adding additional artificial "points of impact." But an obvious idea is then to incorporate estimates of the linear functional by relying on functional principal components. Recall the Karhunen–Loève decomposition already discussed in Section 2, and note that $\int_a^b \beta(t)X_i(t)\,dt = \sum_{r=1}^{\infty} \alpha_r \langle X, \psi_r\rangle$ with $\alpha_r = \langle \psi_r, \beta\rangle$. For $k, S \in \mathbb{N}$, estimates $\widehat{\psi}_r$ of ψ_r and a subset $\tilde{\tau}_1, \ldots, \tilde{\tau}_S \in \{\widehat{\tau}_1, \ldots, \widehat{\tau}_{M_\delta}\}$ one may consider an approximate relationship which resembles an "augmented model" as proposed by Kneip and Sarda (2011) in a different context:

$$Y_i \approx \sum_{r=1}^{k} \alpha_r \langle X_i, \widehat{\psi}_r\rangle + \sum_{r=1}^{S} \beta_r X_i(\tilde{\tau}_r) + \varepsilon_i^*. \tag{4.4}$$

Based on corresponding least-squares estimates of the coefficients α_r and β_r, the number S and an optimal value of k may then be estimated by the BIC criterion.

This approach also offers a way to select a sensible value of $\delta = Cn^{-1/2}$ for a suitable range of values $C \in [C_{\min}, C_{\max}]$. For finite n, different choices of C (and δ) may of course lead to different candidate values $\hat{\tau}_r$, $r = 1, 2, \ldots$. A straightforward approach is then to choose the value of δ, where the respective estimates of impact points lead to the best fitting augmented model (4.4). In addition to estimating S and an optimal value of k, BIC may thus also be used to approximate an optimal value of C (and δ).

Recall that the above approach is applicable if Assumption 1 holds for some $\kappa < 2$. In a practical application, one may thus want to check the applicability of the theory by estimating the value of κ from the data. We have $\mathbb{E}(Z_{\delta,i}(t_j)^2) = \delta^\kappa(2c(t_j) - \frac{2^\kappa}{2}c(t_j)) + o(\delta^\kappa)$ [see (A.3) in the proof of Theorem 3]. Consequently,

$\frac{\mathbb{E}(Z_{\delta,i}(t_j)^2)}{\mathbb{E}(Z_{\delta/2,i}(t_j)^2)} = 2^\kappa + o(1)$ as $\delta \to 0$. Without restriction assume that k_δ is an even number. The above arguments motivate the estimator

$$\widehat{\kappa} = \log_2\Bigg(\frac{(1/(p-2k_\delta))\sum_{j\in\mathcal{J}_{0,\delta}}\sum_{i=1}^n Z_{\delta,i}(t_j)^2}{(1/(p-2k_\delta))\sum_{j\in\mathcal{J}_{0,\delta}}\sum_{i=1}^n Z_{\delta/2,i}(t_j)^2}\Bigg)$$

of κ. In Proposition 1 below, it is shown that $\widehat{\kappa}$ is a consistent estimator of κ as $n \to \infty$, $\delta \to 0$. In practice, an estimate $\widehat{\kappa} \ll 2$ will indicate a process whose covariance function possesses a nonsmooth diagonal.

PROPOSITION 1. *Under the conditions of Theorem 4, we have*

$$\widehat{\kappa} = \kappa + O_P\big(n^{-1/2} + \delta^{\min\{2,2/\kappa\}}\big). \tag{4.5}$$

A final theoretical result concerns the distance between $X_i(\widehat{\tau}_r)$ and $X_i(\tau_r)$. It will be of crucial importance in the next section on parameter estimation. Without restriction, we will in the following assume that points of impact are ordered in such a way that $\tau_r = \arg\min_{s=1,\dots,S}|\widehat{\tau}_r - \tau_s|$, $r = 1,\dots,S$.

PROPOSITION 2. *Under the assumptions of Theorem 4, we obtain for every* $r = 1,\dots,S$

$$\frac{1}{n}\sum_{i=1}^n\big(X_i(\tau_r) - X_i(\widehat{\tau}_r)\big)^2 = O_p\big(n^{-1}\big), \tag{4.6}$$

$$\frac{1}{n}\sum_{i=1}^n\big(X_i(\tau_r) - X_i(\widehat{\tau}_r)\big)\varepsilon_i = O_p\big(n^{-1}\big). \tag{4.7}$$

5. Parameter estimates. Recall that Assumption 1 is only a sufficient, not a necessary condition of identifiability. Even if this assumption is violated and the covariance function $\sigma(t,s)$ is very smooth, there may exist alternative procedures leading to sensible estimators $\widehat{\tau}_r$. In the following, we will thus only assume that the points of impacts are estimated by some procedure such that $P(\widehat{S} = S) \to 1$ as $n \to \infty$ and such that (4.6) as well as (4.7) hold for all $r = 1,\dots,S$. Note that this assumption is trivially satisfied if analysis is based on pre-specified points of impact as discussed in the Introduction.

In situations where it can be assumed that $\int_a^b \beta(t)X_i(t)\,dt = 0$ a.s., we have $Y_i = \sum_{r=1}^S \beta_r X_i(\tau_r) + \varepsilon_i$, $i = 1,\dots,n$, and the regression coefficient may be obtained by least squares when replacing the unknown points of impact τ_r by their estimates $\widehat{\tau}_r$. More precisely, in this case an estimator $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1,\dots,\widehat{\beta}_{\widehat{S}})^T$ of $\boldsymbol{\beta} = (\beta_1,\dots,\beta_S)^T$ is determined by minimizing

$$\frac{1}{n}\sum_{i=1}^n\Bigg(Y_i - \sum_{r=1}^{\widehat{S}} b_r X_i(\widehat{\tau}_r)\Bigg)^2 \tag{5.1}$$

over all possible values $b_1, \ldots, b_{\widehat{S}}$.

Let $\mathbf{X}_i(\boldsymbol{\tau}) := (X_i(\tau_1), \ldots, X_i(\tau_S))^T$, and let $\Sigma_\tau := \mathbb{E}(\mathbf{X}_i(\boldsymbol{\tau})\mathbf{X}_i(\boldsymbol{\tau})^T)$. Note that identifiability of the regression model as stated in Theorem 1 in particular implies that Σ_τ is invertible.

If $\widehat{S} = S$, then by (4.6) and (4.7) the differences between $\widehat{\tau}_r$ and τ_r, $r = 1, \ldots, S$ are asymptotically negligible, and the asymptotic distribution of $\widehat{\boldsymbol{\beta}}$ coincides with the asymptotic distribution the least squares estimator to be obtained if points of impact were known:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_D N\big(0, \sigma^2 \Sigma_\tau^{-1}\big) \tag{5.2}$$

as $n \to \infty$. A proof is straightforward, and thus omitted.

In the general case with $\beta(t) \neq 0$ for some t, we propose to rely on the augmented model (4.4). Thus, let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots$ and $\widehat{\psi}_1, \widehat{\psi}_2, \ldots$ denote eigenvalues and eigenfunctions of the empirical covariance operator of $X_1, \ldots, X_n$. Given estimates $\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{S}}$ and a suitable cut-off parameter k estimates $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_{\widehat{S}})^T$ of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_S)^T$ and $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_k$ of $\alpha_1, \ldots, \alpha_k$ are determined by minimizing

$$\sum_{i=1}^{n} \left(Y_i - \sum_{r=1}^{k} a_r \langle X_i, \widehat{\psi}_r \rangle - \sum_{r=1}^{\widehat{S}} b_r X_i(\widehat{\tau}_r) \right)^2 \tag{5.3}$$

over all a_r, b_s, $r = 1, \ldots, k$, $s = 1, \ldots, \widehat{S}$. Based on the estimated coefficients $\widehat{\alpha}_1, \ldots, \widehat{\alpha}_k$, and estimator of the slope function β is then given by $\widehat{\beta}(t) := \sum_{r=1}^{k} \widehat{\alpha}_k \widehat{\psi}_r(t)$.

In the following we will rely on a slight change of notation in the sense that Y_i, X_i (and ϵ_i) are centered data obtained for each case by subtracting sample means. As pointed out in the remark, we argue that theoretical results stated in Section 4 remain unchanged for this situation. In the context of (5.3) centering ensures that $X_i, i = 1, \ldots, n$, can be *exactly* represented by $X_i = \sum_{j=1}^{n} \langle X_i, \widehat{\psi}_r \rangle \widehat{\psi}_r$ (necessarily $\hat{\lambda}_j = 0$ for $j > n$).

Our theoretical analysis of the estimators defined by (5.3) relies on the work of Hall and Horowitz (2007) who derive rates of convergence of the estimator $\widehat{\beta}(t)$ in a standard functional regression model with $S = 0$. Under our Assumption 2 their results are additionally based on the following assumption on the eigendecompositions of X and β.

ASSUMPTION 3. (a) There exist some $\mu > 1$ and some $\sigma^2 < C_0 < \infty$ such that $\lambda_j - \lambda_{j+1} \geq C_0^{-1} j^{-\mu-1}$ for all $j \geq 1$.

(b) $\beta(t) = \sum_{j=1}^{\infty} \alpha_j \psi(t)$ for all t, and $|\alpha_j| \geq C_0 j^{-\nu}$ for some $\nu > 1 + \frac{1}{2}\mu$.

Hall and Horowitz (2007) show that if $S = 0$ and $k = O(n^{1/(\mu+2\nu)})$, then $\int_a^b (\widehat{\beta}(t) - \beta(t))^2 \, dt = O_p(n^{-(2\nu-1)/(\mu+2\nu)})$. This is known to be an optimal rate of convergence under the standard model.

When dealing with points of impact, some additional conditions are required. Note that $\sigma(t,s) = \sum_{j=1}^{\infty} \lambda_j \psi_j(t)\psi_j(s)$. Let $\sigma^{[k]}(t,s) := \sum_{j=k+1}^{\infty} \lambda_j \psi_j(t)\psi_j(s)$, and let $\mathbf{M}_k$ denote the $S \times S$ matrix with elements $\sigma^{[k]}(\tau_r, \tau_s)$, $r,s = 1,\dots,S$. Furthermore, let $\lambda_{\min}(\mathbf{M}_k)$ denote the smallest eigenvalue of the matrix $\mathbf{M}_k$.

ASSUMPTION 4. (a) $\sup_t \sup_j \psi_j(t)^2 \le C_\psi$ for some $C_\psi < \infty$.
(b) There exists some $0 < C_1 < \infty$ such that $\lambda_j \le C_1 j^{-\mu}$ for all j.
(c) There exists some $0 < D < \infty$ such that $\lambda_{\min}(\mathbf{M}_k) \ge D k^{-\mu+1}$ for all k.

Condition (a) is, for example, satisfied if $\psi_1, \psi_2, \dots$ correspond to a Fourier-type basis. Note that Assumption 3(a) already implies that λ_j must not be less than a constant multiple of $j^{-\mu}$, and thus condition (b) requires that $j^{-\mu}$ is also an upper bound for the rate of convergence of λ_j. This in turn implies that $\sum_{j=k+1}^{\infty} \lambda_j \le C_2 k^{-\mu+1}$ as well as $|\sigma^{[k]}(t,s)| \le C_2 C_\psi^2 k^{-\mu+1}$ for some $C_2 < \infty$ and all k. Condition (c) therefore only introduces an additional regularity condition on the matrix $\mathbf{M}_k$. For the Brownian motion discussed in Section 3, it is easily seen that these requirements are necessarily fulfilled with $\mu = 2$.

We now obtain the following theorem.

THEOREM 5. *Under our setup and Assumptions 2–4 suppose that $\widehat{S} = S$ and that estimators $\widehat{\tau}_r$ satisfy (4.6) as well as (4.7) for all $r = 1,\dots,S$. If additionally $k = O(n^{1/(\mu+2\nu)})$ and $n^{1/(\mu+2\nu)} = O(k)$ as $n \to \infty$, then*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 = O_p\big(n^{-2\nu/(\mu+2\nu)}\big), \tag{5.4}$$

$$\int_a^b \big(\widehat{\beta}(t) - \beta(t)\big)^2 dt = O_p\big(n^{-(2\nu-1)/(\mu+2\nu)}\big). \tag{5.5}$$

In the presence of points of impact the slope function $\beta(t)$ can thus be estimated with the same rate of convergence as in the standard model with $S = 0$. The estimators $\widehat{\beta}_r$ of β_r, $r = 1,\dots,S$, achieve a slightly faster rate of convergence.

6. Simulation study. We proceed by studying the finite sample performance of our estimation procedure described in the preceding sections. For different values of n, p, observations (X_i, Y_i) are generated according to the points of impact model (1.1) where $\varepsilon_i \sim N(0,1)$ are independent error terms. The algorithms are implemented in R, and all tables are based on 1000 repetitions of the simulation experiments. The corresponding R-code can be obtained from the authors upon request.

The data $X_1,\dots,X_n$ are generated as independent Ornstein–Uhlenbeck processes ($\kappa = 1$) with parameters $\theta = 5$ and $\sigma_u = 3.5$ at p equidistant grid points over the interval $[0,1]$. Simulated trajectories are determined by using exact updating formulas as proposed by Gillespie (1996). The simulation study is based

TABLE 1

Estimation errors for different sample sizes for the simulation study (OU-process, $\tau_1 = 0.25$, $\tau_2 = 0.75$, $\beta_1 = 2$, $\beta_2 = 1$). The column containing the estimate $\widehat{P}(\widehat{S} = S)$ contains two numbers: the estimate derived from the BIC followed by its value derived from the cut-off procedure

Sample sizes		**Parameter estimates**									
p	n	$\lvert\widehat{\tau}_1 - \tau_1\rvert$	$\lvert\widehat{\tau}_2 - \tau_2\rvert$	$\lvert\widehat{\beta}_1 - \beta_1\rvert$	$\lvert\widehat{\beta}_2 - \beta_2\rvert$	$\widehat{S}$	$\widehat{P}(\widehat{S} = S)$	$\widehat{k}$	$\int(\widehat{\beta} - \beta)^2$	MSE	$\widehat{\kappa}$
Simulation results if $\beta(t) \equiv 0$											
1001	50	0.0130	0.0357	0.393	0.353	1.74	0.65/0.34	1.33	6.82	1.21	0.89
	100	0.0069	0.0226	0.274	0.249	1.96	0.77/0.40	1.05	3.43	1.21	0.94
	250	0.0027	0.0099	0.129	0.145	2.14	0.83/0.61	0.67	1.11	1.13	0.97
	500	0.0012	0.0061	0.070	0.097	2.15	0.86/0.73	0.45	0.51	1.08	0.98
	5000	0.0000	0.0004	0.012	0.012	2.04	0.96/0.98	0.03	0.00	1.00	1.00
20,001	50	0.0118	0.0333	0.393	0.350	1.71	0.64/0.35	1.78	6.91	1.19	0.89
	100	0.0068	0.0246	0.279	0.276	1.94	0.76/0.46	1.46	3.81	1.19	0.94
	250	0.0025	0.0108	0.121	0.144	2.15	0.83/0.62	0.74	1.02	1.12	0.97
	500	0.0013	0.0063	0.064	0.092	2.14	0.88/0.75	0.48	0.40	1.08	0.98
	5000	0.0001	0.0005	0.013	0.012	2.06	0.94/0.94	0.04	0.00	1.01	1.00
Simulation results if $\beta(t) \neq 0$											
1001	50	0.0150	0.0423	0.465	0.499	1.54	0.49/0.30	2.10	10.82	1.27	0.88
	100	0.0097	0.0317	0.376	0.400	1.86	0.63/0.34	2.06	5.93	1.27	0.94
	250	0.0039	0.0151	0.206	0.234	2.25	0.68/0.46	1.83	2.21	1.17	0.97
	500	0.0015	0.0083	0.107	0.164	2.30	0.72/0.59	1.69	0.90	1.10	0.99
	5000	0.0000	0.0006	0.036	0.027	2.25	0.79/0.97	2.01	0.05	1.01	1.00
20,001	50	0.0166	0.0399	0.467	0.465	1.52	0.47/0.29	2.14	11.19	1.29	0.89
	100	0.0099	0.0286	0.370	0.378	1.90	0.64/0.36	2.08	5.95	1.26	0.94
	250	0.0037	0.0171	0.185	0.263	2.27	0.67/0.49	1.90	2.19	1.15	0.97
	500	0.0018	0.0104	0.118	0.177	2.32	0.71/0.62	1.78	1.11	1.11	0.99
	5000	0.0002	0.0007	0.038	0.028	2.23	0.82/0.95	2.03	0.05	1.02	1.00

on $S = 2$ points of impact located at $\tau_1 = 0.25$ and $\tau_2 = 0.75$ with corresponding coefficients $\beta_1 = 2$ as well as $\beta_2 = 1$. Results are reported in Table 1, where the upper part of the table refers to the situation with $\beta(t) \equiv 0$, while the lower part represents a model with $\beta(t) = 3.5t^3 - 5.5t^2 + 3t + 0.5$.

In both cases, estimation of the points of impact relies on setting $\delta = C\frac{1}{\sqrt{n}}$ for $C = 1$, but similar results could be obtained for a wide range of values C. The results are then obtained by performing best subset selection with the BIC-criterion via the R package bestglm on the augmented model (4.4)

$$Y_i \approx \sum_{r=1}^{k} \alpha_r \langle X_i, \widehat{\psi}_r \rangle + \sum_{r=1}^{\widetilde{S}} \beta_r X_i(\tilde{\tau}_r) + \varepsilon_i^*. \tag{6.1}$$

Here, $\widetilde{S}$ is the number of all possible candidates for the points of impact and k is initially set to 6 principal components, but tendencies remain unchanged for a broad range of values k.

For different sample sizes n and p, Table 1 provides the average absolute errors of our estimates, the frequency of $\widehat{S} = S$, as well as average values of $\widehat{S}$, $\widehat{k}$, the prediction error $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2$ and $\widehat{\kappa}$. The column containing $\widehat{P}(\widehat{S} = S)$ consists of two values. The first one being the frequency of $\widehat{S} = S$ resulting from the BIC. For the second one, S was estimated by the cut-off procedure using $\lambda = 2\sqrt{\widehat{\mathrm{Var}}(Y)/n \log(\frac{b-a}{\delta})}$, where $\widehat{\mathrm{Var}}(Y)$ denotes the estimated sample variance of Y_i. The cut-off criterion yields very reliable estimates $\widehat{S}$ of S for $n = 5000$, but showed a clear tendency to underestimate S for smaller sample sizes. The BIC-criterion however proves to possess a much superior behavior in this regards for small n but is outperformed by the cut-off criterion for $n = 5000$ in the case $\beta(t) \neq 0$.

In order to match $\{\widehat{\tau}_s\}_{s=1,\ldots,\widehat{S}}$ and $\{\tau_r\}_{r=1,2}$ the interval $[0, 1]$ is divided into $I_1 = [0, \frac{1}{2}(\tau_1 + \tau_2)]$ and $I_2 = [\frac{1}{2}(\tau_1 + \tau_2), 1]$. The estimate $\widehat{\tau}_s$ in interval I_r with the minimal distance to τ_r is then used as an estimate for τ_r. No point of impact candidate in interval I_r results in an "unmatched" τ_r, $r = 1, \ldots, S$ and a missing value when computing averages.

The table shows that estimates of points of impact are generally quite accurate even for smaller sample sizes. The error decreases rapidly as n increases, and this improvement is essentially independent of p. As expected, since $\beta_2 < \beta_1$, the error of the absolute distance between the second point of impact and its estimate is larger than the error for the first point of impact.

Moreover, due to the common effect of the trajectory $X_i(\cdot)$ on Y_i, the overall estimation error in the case where $\beta(t) \neq 0$ is slightly higher than in the first case. At a first glance, one may be puzzled by the fact that for $n = 5000$ and $p = 1001$ the average error $|\widehat{\tau}_r - \tau_r|$ is considerably smaller than the distance $\frac{1}{p-1} = \frac{1}{1000}$ between two adjacent grid points. But note that our simulation design implies that $\tau_r \in \{t_j | j = 1, \ldots, p\}$, $r = 1, \ldots, S$, for $p = 1001$ as well as $p = 20{,}001$. For medium to large sample sizes, there is thus a fairly high probability that $\widehat{\tau}_r = \tau_r$. The case $p = 1001$ particularly profits from this situation. Finally, it can be seen that estimates for $\widehat{\kappa}$ tend to slightly underestimate the true value $\kappa = 1$ for small values of n.

7. Applications to real data. In this section, the algorithm from Section 4 is applied to a dataset consisting of Canadian weather data. In this dataset, we relate the mean relative humidity to hourly temperature data. In the Supplementary Appendix A [Kneip, Poss and Sarda (2015)], a further application can be found. We there analyze spectral data which play an important role in spectrophotometry and different applied scientific fields.

In both examples, the algorithm is applied to centered observations and the estimation procedure from Section 4 is modified by eliminating all points in an interval of size $\delta|\log\delta|$ around a point of impact candidate $\widehat{\tau}_j$, which is still sufficient to establish assertion (4.1).

After estimating $\widetilde{S}$ possible candidates for the points of impact, the approximate model (4.4),

$$Y_i \approx \sum_{r=1}^{k} \alpha_r \langle X_i, \widehat{\psi}_r \rangle + \sum_{r=1}^{\widetilde{S}} \beta_r X_i(\tilde{\tau}_r) + \varepsilon_i^*,$$

is used, where initially $k = 6$ is chosen. Over a fine grid of different values of δ, points of impact and principal components are selected simultaneously by best subset selection with the BIC-criterion and the model corresponding to the minimal BIC is then chosen. The maximum number of variables selected by the BIC-criterion is set to 6 and all curves have been transformed to be observed over $[0, 1]$ when applying the algorithm from Section 4. The performance of the model is then measured by means of a cross-validated prediction error.

In the Canadian weather dataset, the hourly mean temperature and relative humidity from the 15 closest weather stations in an area around 100 km from Montreal was obtained for each of the 31 days in December 2013. The data was compiled from http://climate.weather.gc.ca. Weather stations with more than ten missing observations on the temperature or relative humidity were discarded from the dataset. The remaining stations had their nonavailable observations replaced by the mean of their closest observed predecessor and successor. After preprocessing a total of $n = 13$ weather stations remained and for each station $p = 744$ equidistant hourly observations of the temperature were observed. The response variable Y_i was taken to be the mean over all observed values of the relatively humidity at station i.

A cross-validated prediction error was calculated for three competing regression models based on (4.4). In the first model, the mean relative humidity for each station was explained by using the approximate model which combines the points of impacts with a functional part. The second and third model describe the cases $k = 0$ and $\widetilde{S} = 0$ in the approximate model, consisting only of points of impact and the functional part, respectively. For the first two models, points of impact were determined by considering a total of 146 equidistant values of δ between 0.10 and 0.49. In all models BIC was used to approximate the optimal values of the respective tuning parameters δ, S and/or k in a first step. The mean squared prediction error $MSPE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ was then calculated by means of a leave one out cross-validation based on the chosen points of impact and/or principal components from the first step. Additionally, the median of $(y_i - \hat{y}_i)^2$, $i = 1, \ldots, n$, has been calculated as a more robust measure of the error. Depicted in the upper panel of Figure 3 is the observed temperature trajectory for the weather station "McTavish," showing a rather rough process. The lower panel of this figure shows $|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t_j) Y_i|$ for the optimal value of $\delta = 0.18$ as obtained by the best model fit of the approximate model. While orange lines represent the locations of the points of impact which were actually selected with the help of the BIC-criterion, the location of the remaining candidates are indicated by black vertical lines.

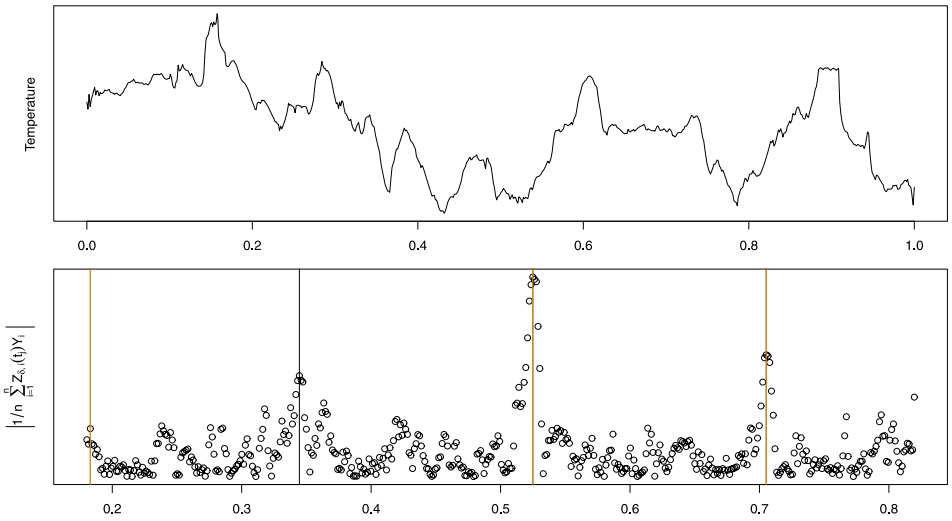

FIG. 3. *The upper panel of this figure shows a trajectory from the observed temperature curves of the Canadian weather data. The lower panel shows* $|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t_j)Y_i|$ *during the selection procedure. Locations of selected points of impact in the augmented model are indicated by orange lines. The location of the remaining candidate is displayed by a black line.*

Table 2 provides the empirical results when fitting the three competing models. In terms of the prediction error, it can clearly be seen from the table that the frequently applied functional linear regression model is outperformed by the model consisting solely of points of impact as well as the augmented (approximate) model. This impression is supported by the last column of the table which gives the median value of $(y_i - \hat{y}_i)^2$, showing additionally that, typically, the augmented model performs even better than the plain points of impact model.

An estimate $\widehat{\kappa} = 0.14$ for κ was obtained for $\delta \approx 0.3$, that is, the midpoint of the chosen values of δ. The estimated value of $\kappa = 0.14$ corresponds to rather rough sample paths as shown in the upper plot of Figure 3.

In view of the small sample size results have to be interpreted with care, and we therefore do not claim that this application provides important substantial in-

TABLE 2

Estimated number of principal components k*, points of impact* S*, prediction error and the median of* $(y_i - \hat{y}_i)^2$ *for the Canadian weather data*

Model	$\widehat{k}$	$\widehat{S}$	**MSPE**	**median**$((y-\hat{y})^2)$
Augmented	3	3	2.314	0.251
Points of impact	0	3	1.714	0.974
FLR	6	0	5.346	1.269

sights. Its main purpose is to serve as illustration for classes of problems where our approach may be of potential importance. It clearly shows that some relevant processes observed in practice are nonsmooth. With contemporary technical tools temperatures can be measured very accurately, leading to a negligible measurement error. But temperatures, especially in Canada, can vary rapidly over time. The rough sample paths thus must be interpreted as an intrinsic feature of temperature processes and cannot be explained by any type of "error."

APPENDIX: PROOFS OF THEOREMS

This appendix provides the proofs of some of the main results. Remaining proofs can be found in the supplementary material. Some of them rely on results from van de Geer and Lederer (2013), van der Vaart and Wellner (1996) as well as Zhou, Lafferty and Wasserman (2008).

PROOF OF THEOREM 1. Set $\beta_r := 0$ for $r = S+1, \dots, S^*$, and consider an arbitrary $j \in \{1, \dots, S^*\}$. Choose $0 < \epsilon < \min_{r,s\in\{1,\dots,S^*\}, r\neq s} |\tau_r - \tau_s|$ small enough such that conditions (i)–(iv) of Definition 1 are satisfied. Using (2.1), we obtain a decomposition into two uncorrelated components $X_{\epsilon,\tau_j}(\cdot)$ and $\zeta_{\epsilon,\tau_j}(X) f_{\epsilon,\tau_j}(\cdot)$:

$$
\begin{aligned}
&\mathbb{E}\left(\left(\int_a^b (\beta(t) - \beta^*(t)) X(t)\, dt + \sum_{r=1}^{S^*} (\beta_r - \beta_r^*) X(\tau_r)\right)^2\right) \\
&\quad = \mathbb{E}\left(\left(\int_a^b (\beta(t) - \beta^*(t)) X_{\epsilon,\tau_j}(t)\, dt + \sum_{r=1}^{S^*} (\beta_r - \beta_r^*) X_{\epsilon,\tau_j}(\tau_r)\right)^2\right) \\
&\qquad + \mathbb{E}\left(\left(\int_a^b (\beta(t) - \beta^*(t)) \zeta_{\epsilon,\tau_j}(X) f_{\epsilon,\tau_j}(t)\, dt \right.\right. \\
&\qquad \left.\left. + \sum_{r=1}^{S^*} (\beta_r - \beta_r^*) \zeta_{\epsilon,\tau_j}(X) f_{\epsilon,\tau_j}(\tau_r)\right)^2\right) \\
&\quad \geq \mathbb{E}\left(\left(\int_a^b (\beta(t) - \beta^*(t)) \zeta_{\epsilon,\tau_j}(X) f_{\epsilon,\tau_j}(t)\, dt \right.\right. \\
&\qquad \left.\left. + \sum_{r\neq j} (\beta_r - \beta_r^*) \zeta_{\epsilon,\tau_j}(X) f_{\epsilon,\tau_j}(\tau_r) + (\beta_j - \beta_j^*) \zeta_{\epsilon,\tau_j}(X) f_{\epsilon,\tau_j}(\tau_j)\right)^2\right) \\
&\quad \geq 2 \operatorname{var}(\zeta_{\epsilon,\tau_j}(X)) (\beta_j - \beta_j^*) f_{\epsilon,\tau_j}(\tau_j) \\
&\qquad \times \left(\int_a^b (\beta(t) - \beta^*(t)) f_{\epsilon,\tau_j}(t)\, dt + \sum_{r\neq j} (\beta_r - \beta_r^*) f_{\epsilon,\tau_j}(\tau_r)\right) \\
&\qquad + \operatorname{var}(\zeta_{\epsilon,\tau_j}(X)) (\beta_j - \beta_j^*)^2 f_{\epsilon,\tau_j}(\tau_j)^2.
\end{aligned}
$$

By condition (iv), we have

$$\left|\sum_{r\neq j}(\beta_r-\beta_r^*)f_{\epsilon,\tau_j}(\tau_r)\right|\leq \epsilon S^*\max_{r\neq j}|\beta_r-\beta_r^*||f_{\epsilon,\tau_j}(\tau_j)|,$$

while boundedness of $\beta(\cdot)$ and $\beta^*(\cdot)$ implies that there exits a constant $0\leq D<\infty$ such that for all sufficiently small $\epsilon>0$

$$\begin{aligned}\left|\int_a^b(\beta(t)-\beta^*(t))f_{\epsilon,\tau_j}(t)\,dt\right|&\leq \epsilon\int_{[a,b]\setminus[\tau_j-\epsilon,\tau_j+\epsilon]}D\big|f_{\epsilon,\tau_j}(\tau_j)\big|\,dt\\&\quad+\int_{\tau_j-\epsilon}^{\tau_j+\epsilon}(1+\epsilon)D\big|f_{\epsilon,\tau_j}(\tau_j)\big|\,dt\\&\leq \epsilon\big(b-a+2(1+\epsilon)\big)D\big|f_{\epsilon,\tau_j}(\tau_j)\big|.\end{aligned}$$

When combining these inequalities, we can conclude that for all sufficiently small ϵ we have $\mathbb{E}(\int_a^b(\beta(t)-\beta^*(t))X(t)\,dt+\sum_{r=1}^{S^*}(\beta_r-\beta_r^*)X(\tau_r))^2>0$ if $\beta_j-\beta_j^*\neq 0$. Since $j\in\{1,\dots,S^*\}$ is arbitrary, the assertion of the theorem is an immediate consequence. □

PROOF OF THEOREM 2. Choose some arbitrary $t\in(a,b)$ and some $0<\epsilon<1$ with $\epsilon\leq\epsilon_t$. By assumption, there exists a $k\in\mathbb{N}$ as well as some $f\in\mathcal{C}(t,\epsilon,[a,b])$ such that $|\langle f,\psi_r\rangle|>0$ for some $r\in\{1,\dots,k\}$ and $\sup_{s\in[a,b]}|f_k(s)-f(s)|\leq\epsilon/3$, where $f_k(s)=\sum_{r=1}^k\langle f,\psi_r\rangle\psi_r(s)$. The definition of $\mathcal{C}(t,\epsilon,[a,b])$ then implies that $f_k(t)\geq 1-\epsilon/3$ as well as

$$\begin{aligned}\sup_{s\in[a,b]}\big|f_k(s)\big|&\leq 1+\frac{\epsilon}{3}\leq(1+\epsilon)\left(1-\frac{\epsilon}{3}\right)\leq(1+\epsilon)f_k(t),\\\sup_{s\in[a,b],s\notin[t-\epsilon,t+\epsilon]}\big|f_k(s)\big|&\leq\frac{\epsilon}{3}\leq\epsilon\left(1-\frac{\epsilon}{3}\right)\leq\epsilon f_k(t).\end{aligned}\tag{A.1}$$

Now define the functional $\zeta_{\epsilon,t}$ by $\zeta_{\epsilon,t}(X):=\sum_{r=1}^k\frac{\langle f,\psi_r\rangle}{\lambda_r}\langle X,\psi_r\rangle$. Recall that the coefficients $\langle X,\psi_r\rangle$ are uncorrelated and $\operatorname{var}(\langle X,\psi_r\rangle)=\lambda_r$. By (2.4), we obtain

$$\begin{aligned}f_{\epsilon,t}(s)&:=\frac{\mathbb{E}(X(s)\zeta_{\epsilon,t}(X))}{\operatorname{var}(\zeta_{\epsilon,t}(X))}\\&=\frac{\mathbb{E}((\sum_{j=1}^\infty\langle X,\psi_j\rangle\psi_j(s))(\sum_{r=1}^k(\langle f,\psi_r\rangle/\lambda_r)\langle X,\psi_r\rangle))}{\operatorname{var}(\zeta_{\epsilon,t}(X))}\\&=\frac{\sum_{r=1}^k\langle f,\psi_r\rangle\psi_r(s)}{\operatorname{var}(\zeta_{\epsilon,t}(X))}=\frac{f_k(s)}{\operatorname{var}(\zeta_{\epsilon,t}(X))}.\end{aligned}$$

Furthermore, $\operatorname{var}(\zeta_{\epsilon,t}(X))=\sum_{r=1}^k\frac{\langle f,\psi_r\rangle^2}{\lambda_r}>0$, and it thus follows from (A.1) that the functional $\zeta(t,X)$ satisfies conditions (i)–(iv) of Definition 1. Since $t\in(a,b)$ and ϵ are arbitrary, X thus possesses specific local variation. □

PROOF OF THEOREM 3. First note that Assumption 1 implies that the absolute values of all first and second-order partial derivatives of $\omega(t,s,z)$ are uniformly bounded by some constant $M<\infty$ for all (t,s,z) in the compact subset $[a,b]^2\times[0,b-a]$ of Ω.

By definition of Z_δ, it thus follows from a Taylor expansion of ω that for $t\in(a,b)$, any sufficiently small $\delta>0$ and some constant $M_1<\infty$

$$\begin{aligned}\mathbb{E}\big(X(t)Z_\delta(X,t)\big)&=\sigma(t,t)-\frac{1}{2}\sigma(t,t-\delta)-\frac{1}{2}\sigma(t,t+\delta)\\&=\omega(t,t,0)-\frac{1}{2}\omega\big(t,t-\delta,\delta^\kappa\big)-\frac{1}{2}\omega\big(t,t+\delta,\delta^\kappa\big)\\&=\delta^\kappa c(t)+R_{1;\delta,t}\qquad\text{with}\sup_{t\in[a+\delta,b-\delta]}|R_{1;\delta,t}|\le M_1\delta^{\min\{2\kappa,2\}}.\end{aligned}\tag{A.2}$$

For the variance of $Z_\delta(X,t)$, we obtain by similar arguments

$$\begin{aligned}\operatorname{var}\big(Z_\delta(X,t)\big)&=2\omega(t,t,0)-\omega\big(t,t-\delta,\delta^\kappa\big)-\omega\big(t,t+\delta,\delta^\kappa\big)\\&\quad-\frac{1}{2}\big(\omega(t,t,0)-\omega\big(t+\delta,t-\delta,(2\delta)^\kappa\big)\big)\\&\quad-\frac{1}{4}\big(2\omega(t,t,0)-\omega(t-\delta,t-\delta,0)-\omega(t+\delta,t+\delta,0)\big)\\&=\delta^\kappa\left(2c(t)-\frac{2^\kappa}{2}c(t)\right)+R_{2;\delta,t}\\&\qquad\text{with }\sup_{t\in[a+\delta,b-\delta]}|R_{2;\delta,t}|<M_2\delta^{\min\{2\kappa,2\}}\end{aligned}\tag{A.3}$$

for some constant $M_2<\infty$. Moreover, for any $0<c<\infty$ Taylor expansions of ω yield that for any sufficiently small $\delta>0$ and all $u\in[-c,c]$

$$\begin{aligned}&\mathbb{E}\big(X(t+u\delta)Z_\delta(X,t)\big)\\&\quad=\sigma(t+u\delta,t)-\tfrac{1}{2}\sigma(t+u\delta,t-\delta)-\tfrac{1}{2}\sigma(t+u\delta,t+\delta)\\&\quad=\omega(t,t,0)-\tfrac{1}{2}\omega\big(t,t-\delta,\delta^\kappa\big)-\tfrac{1}{2}\omega\big(t,t+\delta,\delta^\kappa\big)\\&\qquad-c(t)\delta^\kappa\big(|u|^\kappa-\tfrac{1}{2}\big(|u+1|^\kappa-1\big)-\tfrac{1}{2}\big(|u-1|^\kappa-1\big)\big)+R_{3;c,u,\delta,t}\qquad\text{(A.4)}\\&\quad=-c(t)\delta^\kappa\big(|u|^\kappa-\tfrac{1}{2}|u+1|^\kappa-\tfrac{1}{2}|u-1|^\kappa\big)+R_{4;c,u,\delta,t},\qquad\text{(A.5)}\end{aligned}$$

where for some constants $M_{3,c}<\infty$ and $M_{4,c}<\infty$

$$\sup_{t\in[a+\delta,b-\delta]}R_{3;c,u,\delta,t}\le M_{3,c}\big(|u|^{1/2}\delta\big)^{\min\{2\kappa,2\}},$$

$$\sup_{t\in[a+\delta,b-\delta]}R_{4;c,u,\delta,t}\le M_{4,c}\delta^{\min\{2\kappa,2\}}$$

hold for all $u \in [-c, c]$. Finally, Assumption 1 implies that there exists a constant $M_5 < \infty$ such that for all $s \in [a, b]$ with $|t - s| \geq \delta$

$$\begin{aligned}
&\big|\mathbb{E}\big(X(s)Z_\delta(X,t)\big)\big| \\
&\quad = \big|\omega\big(s,t,|s-t|^\kappa\big) - \tfrac{1}{2}\omega\big(s,t-\delta,|s-t+\delta|^\kappa\big) \\
&\qquad - \tfrac{1}{2}\omega\big(s,t+\delta,|s-t-\delta|^\kappa\big)\big| \\
&\quad \leq \begin{cases} M_5 \dfrac{\delta^2}{|t-s|^{2-\kappa}}, & \text{if } \kappa \neq 1, \\ M_5\delta^2, & \text{if } \kappa = 1. \end{cases}
\end{aligned} \tag{A.6}$$

It follows from (A.2), (A.5) and (A.6) that for arbitrary $t \in (a, b)$ and any $\epsilon > 0$ there exist a $\delta_\epsilon > 0$ as well as a constant $a_\epsilon \geq 1$ such that for all $\delta \leq \delta_\epsilon$

$$\big|\mathbb{E}\big(X(s)Z_\delta(X,t)\big)\big| \leq (1+\epsilon)\mathbb{E}\big(X(t)Z_\delta(X,t)\big) \qquad \text{for all } s \in [a,b],\ s \neq t,$$

$$\big|\mathbb{E}\big(X(s)Z_\delta(X,t)\big)\big| \leq \epsilon \cdot \mathbb{E}\big(X(t)Z_\delta(X,t)\big) \qquad \text{for all } s \in [a,b],\ |s-t| \geq a_\epsilon\delta.$$

Together with (A.3), the assertion of the theorem is an immediate consequence. □

PROOF OF THEOREM 5. Let $\hat{\theta}_{ij} := \langle X_i, \widehat{\psi}_j\rangle$, $\theta_{ij} := \langle X_i, \psi_j\rangle$, and $\tilde{\alpha}_j := \langle \beta, \widehat{\psi}_j\rangle$ for all i, j. Using empirical eigenfunctions, we obtain $X_i = \sum_{j=1}^n \hat{\theta}_{ij}\widehat{\psi}_j$ and $\int_a^b \beta(t)X_i(t)\,dt = \sum_{j=1}^n \tilde{\alpha}_j\hat{\theta}_{ij}$. Therefore,

$$Y_i = \sum_{j=1}^n \left(\tilde{\alpha}_j + \sum_{r=1}^S \beta_r \widehat{\psi}_j(\tau_r)\right)\hat{\theta}_{ij} + \varepsilon_i, \tag{A.7}$$

and for all possible values $b_1, \ldots, b_S$ and all $a_1, \ldots, a_k$

$$\begin{aligned}
&\sum_{j=1}^k a_j\hat{\theta}_{ij} + \sum_{r=1}^S b_r X_i(\widehat{\tau}_r) \\
&\qquad = \sum_{j=1}^k \left(a_j + \sum_{r=1}^S b_r \widehat{\psi}_j(\widehat{\tau}_r)\right)\hat{\theta}_{ij} + \sum_{j=k+1}^n \sum_{r=1}^S b_r \widehat{\psi}_j(\widehat{\tau}_r)\hat{\theta}_{ij}
\end{aligned} \tag{A.8}$$

for all $i = 1, \ldots, n$. By definition, $\hat{\lambda}_j = \frac{1}{n}\sum_{i=1}^n \hat{\theta}_{ij}^2$, $j = 1, \ldots, n$, and for $j \neq l$ the coefficients $\hat{\theta}_{ij}$ and $\hat{\theta}_{il}$ are empirically uncorrelated, that is, $\sum_{i=1}^n \hat{\theta}_{ij}\hat{\theta}_{il} = 0$. It follows that for any given values $b_1, \ldots, b_S$ the values $\hat{\alpha}(\mathbf{b})_j$, $j = 1, \ldots, k$, minimizing $\sum_{i=1}^n (Y_i - \sum_{j=1}^k a_j\hat{\theta}_{ij} - \sum_{r=1}^S b_r X_i(\widehat{\tau}_r))^2$ over all $a_1, \ldots, a_k$ are given by

$$\hat{\alpha}(\mathbf{b})_j = \tilde{\alpha}_j + \hat{\lambda}_j^{-1}\frac{1}{n}\sum_{i=1}^n \hat{\theta}_{ij}\varepsilon_i + \sum_{r=1}^S \big(\beta_r\widehat{\psi}_j(\tau_r) - b_r\widehat{\psi}_j(\widehat{\tau}_r)\big), \qquad j = 1, \ldots, k. \tag{A.9}$$

Note that $\tilde{\alpha}_j + \hat{\lambda}_j^{-1}\frac{1}{n}\sum_{i=1}^n \hat{\theta}_{ij}\varepsilon_i$ is identical to the estimate of α_j to be obtained in a standard functional linear regression model with no points of impact. Theorem 1

of Hall and Horowitz (2007) thus implies that

$$\begin{aligned}\int_a^b \left(\beta(t) - \sum_{j=1}^{k}\left(\tilde{\alpha}_j + \hat{\lambda}_j^{-1}\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{ij}\varepsilon_i\right)\widehat{\psi}_j(t)\,dt\right)^2 dt \\ = O_p\big(n^{-(2\nu-1)/(\mu+2\nu)}\big).\end{aligned} \tag{A.10}$$

Further analysis requires to analyze the differences between θ_{ij}, ψ_j and their empirical counterparts $\hat{\theta}_{ij}$, $\widehat{\psi}_j$. By Assumptions 2–4 and $k = O(n^{1/(\mu+2\nu)})$, Theorems 1 and 2 together with equation (2.8) of Hall and Hosseini-Nasab (2006) imply that for any $q = 1, 2, 3, \ldots$ there exists some $A_q, B_q < \infty$ such that

$$\begin{aligned}E\big(|\lambda_j - \hat{\lambda}_j|^q\big) &\le A_q n^{-q/2}, \\ \sup_t E\big(\big|\widehat{\psi}_j(t) - \psi_j(t)\big|\big) &\le B_q n^{-q/2} j^{q(\mu+1)}, \qquad j = 1, \ldots, k+1\end{aligned} \tag{A.11}$$

for all sufficiently large n. Let $X_i^{[k]} := X_i - \sum_{j=1}^{k}\hat{\theta}_{ij}\widehat{\psi}_j$. Recall that $\lambda_j = O(j^{-\mu})$ and note that by Assumptions 3 and 4, $n^{-1/2}n^{2/(\mu+2\nu)} = o(n^{(-\mu+1)/(\mu+2\nu)})$, while $n^{(-\mu+1)/(\mu+2\nu)} = O(\sigma^{[k]}(\tau_r, \tau_r))$. By (A.11), we thus obtain for all $t, s \in [a, b]$

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} X_i^{[k]}(t)X_i^{[k]}(s) \\ &\quad= \frac{1}{n}\sum_{i=1}^{n} X_i(t)X_i(s) - \sum_{j=1}^{k}\hat{\lambda}_j\widehat{\psi}_j(t)\widehat{\psi}_j(s) \\ &\quad= \sigma(t,s) - \sum_{j=1}^{k}\lambda_j\psi_j(t)\psi_j(s) + \sum_{j=1}^{k}\lambda_j\big(\psi_j(t)\psi_j(s) - \widehat{\psi}_j(t)\widehat{\psi}_j(s)\big) \\ &\qquad+ \sum_{j=1}^{k}(\lambda_j - \hat{\lambda}_j)\widehat{\psi}_j(t)\widehat{\psi}_j(s) + O_P\big(n^{-1/2}\big) \\ &\quad= \sigma^{[k]}(t,s) + O_P\big(n^{-1/2}n^{2/(\mu+2\nu)}\big) \\ &\quad= \sigma^{[k]}(t,s) + o_P\big(n^{(-\mu+1)/(\mu+2\nu)}\big).\end{aligned} \tag{A.12}$$

At the same time, (4.6) leads to

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\big(X_i^{[k]}(\tau_r) - X_i^{[k]}(\widehat{\tau}_r)\big)^2 \\ &\quad= \frac{1}{n}\sum_{i=1}^{n}\big(X_i(\tau_r) - X_i(\widehat{\tau}_r)\big)^2 - \sum_{j=1}^{k}\hat{\lambda}_j\big(\widehat{\psi}_j(\tau_r) - \widehat{\psi}_j(\widehat{\tau}_r)\big)^2 \\ &\quad\le \frac{1}{n}\sum_{i=1}^{n}\big(X_i(\tau_r) - X_i(\widehat{\tau}_r)\big)^2 = O_P\big(n^{-1}\big)\end{aligned} \tag{A.13}$$

for all $r = 1, \ldots, S$. Expressions (A.12) and (A.13) together imply that for all r, s

$$\frac{1}{n}\sum_{i=1}^{n} X_i^{[k]}(\widehat{\tau}_r) X_i^{[k]}(\widehat{\tau}_s) = \sigma^{[k]}(\tau_r, \tau_s) + o_P\big(n^{(-\mu+1)/(\mu+2\nu)}\big). \tag{A.14}$$

Let $\mathbf{X}_i^{[k]} := (X_i^{[k]}(\widehat{\tau}_1), \ldots, X_i^{[k]}(\widehat{\tau}_S))^T$ and note that by (A.14) we have $\frac{1}{n} \times \sum_{i=1}^n \mathbf{X}_i^{[k]}(\mathbf{X}_i^{[k]})^T = \mathbf{M}_k + o_P(n^{(-\mu+1)/(\mu+2\nu)})$. By Assumption 4(b), we can conclude that with probability tending to 1 as $n \to \infty$ the matrix $\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^{[k]}(\mathbf{X}_i^{[k]})^T$ is invertible,

$$\begin{aligned} &n^{(-\mu+1)/(\mu+2\nu)}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{[k]}(\mathbf{X}_i^{[k]})^T\right)^{-1} \\ &\quad = n^{(-\mu+1)/(\mu+2\nu)}(\mathbf{M}_k)^{-1} + o_P(1) \end{aligned} \tag{A.15}$$

and hence by (A.7)–(A.9) the least squares estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ can be written in the form

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{[k]}(\mathbf{X}_i^{[k]})^T\right)^{-1} \\ &\quad \times \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{[k]}\left(\sum_{r=1}^{S}\beta_r X_i^{[k]}(\tau_r) + \sum_{j=k+1}^{n}\tilde{\alpha}_j\hat{\theta}_{ij} + \varepsilon_i\right). \end{aligned} \tag{A.16}$$

By (A.13) and (A.14), we obtain

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{[k]}\sum_{r=1}^{S}\beta_r X_i^{[k]}(\tau_r) \\ &\quad = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{[k]}(\mathbf{X}_i^{[k]})^T\boldsymbol{\beta} + O_P\big(n^{(-\mu+1)/2(\mu+2\nu)}\cdot n^{-1/2}\big). \end{aligned} \tag{A.17}$$

The results of Hall and Horowitz (2007) imply that $\sum_{j=k+1}^n \tilde{\alpha}_j^2 = O_P(n^{-(2\nu-1)/(\mu+2\nu)})$. The Cauchy–Schwarz inequality thus leads to

$$\begin{aligned} &\left|\frac{1}{n}\sum_{i=1}^{n}X_i^{[k]}(\widehat{\tau}_r)\left(\sum_{j=k+1}^{n}\tilde{\alpha}_j\hat{\theta}_{ij}\right)\right| \\ &\quad = \left|\sum_{j=k+1}^{n}\tilde{\alpha}_j\hat{\lambda}_j\widehat{\psi}_j(\widehat{\tau}_r)\right| \\ &\quad \le \sqrt{\sum_{j=k+1}^{n}\hat{\lambda}_j\tilde{\alpha}_j^2}\sqrt{\sum_{j=k+1}^{n}\hat{\lambda}_j\widehat{\psi}_j(\widehat{\tau}_r)^2} \le \sqrt{\hat{\lambda}_{k+1}\sum_{j=k+1}^{n}\tilde{\alpha}_j^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^{[k]}(\widehat{\tau}_r)^2} \\ &\quad = O_P\big(n^{-(\mu+2\nu-1)/2(\mu+2\nu)}\cdot n^{(-\mu+1)/2(\mu+2\nu)}\big) \end{aligned} \tag{A.18}$$

for all $r = 1, \ldots, S$. Furthermore, $\widehat{\psi}_j(t) = \hat{\lambda}_j^{-1} \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{ij} X_i(t)$, and hence the Cauchy–Schwarz inequality yields

$$\begin{aligned} |\widehat{\psi}_j(\tau_r) - \widehat{\psi}_j(\widehat{\tau}_r)| &= \left| \hat{\lambda}_j^{-1} \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{ij} \big(X_i(\tau_r) - X_i(\widehat{\tau}_r)\big) \right| \\ &\leq \hat{\lambda}_j^{-1/2} \sqrt{\frac{1}{n} \sum_{l=1}^n \big(X_l(\tau_r) - X_l(\widehat{\tau}_r)\big)^2}. \end{aligned} \tag{A.19}$$

Now note that by the independence of $\hat{\theta}_{ij}$ and ε_i we have $\hat{\lambda}_j^{-1/2} \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{ij} \varepsilon_i = O_P(n^{-1/2})$. By (4.7), it therefore follows from (A.19) that

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \big(X_i^{[k]}(\widehat{\tau}_r) - X_i^{[k]}(\tau_r)\big)\varepsilon_i \\ &\quad = \frac{1}{n} \sum_{i=1}^n \big(X_i(\widehat{\tau}_r) - X_i(\tau_r)\big)\varepsilon_i - \sum_{j=1}^k \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{ij} \varepsilon_i \big(\widehat{\psi}_j(\widehat{\tau}_r) - \widehat{\psi}_j(\tau_r)\big) \\ &\quad = O_P\big((k+1)n^{-1}\big) = O_P\big(n^{-(\mu+2\nu-1)/(\mu+2\nu)}\big). \end{aligned}$$

Using (A.12), it is immediately seen that $\frac{1}{n} \sum_{i=1}^n X_i^{[k]}(\tau_r)\varepsilon_i = O_P(n^{-1/2} \times n^{(-\mu+1)/2(\mu+2\nu)})$. Consequently,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n X_i^{[k]}(\widehat{\tau}_r)\varepsilon_i &= \frac{1}{n} \sum_{i=1}^n X_i^{[k]}(\tau_r)\varepsilon_i + \frac{1}{n} \sum_{i=1}^n \big(X_i^{[k]}(\widehat{\tau}_r) - X_i^{[k]}(\tau_r)\big)\varepsilon_i \\ &= O_P\big(n^{-1/2} n^{(-\mu+1)/2(\mu+2\nu)}\big). \end{aligned} \tag{A.20}$$

By Assumption 4(c), we can infer from (A.15) that the maximal eigenvalue of the matrix $(\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^{[k]} (\mathbf{X}_i^{[k]})^T)^{-1}$ can be bounded by $\lambda_{\max}((\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^{[k]} \times (\mathbf{X}_i^{[k]})^T)^{-1}) = O_P(n^{(\mu-1)/(\mu+2\nu)})$. It therefore follows from (A.16)–(A.20) that

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + O_P\big(n^{(\mu-1)/(\mu+2\nu)} \cdot n^{(-\mu+1)/2(\mu+2\nu)} \cdot n^{-(\mu+2\nu-1)/2(\mu+2\nu)}\big) \\ &= \boldsymbol{\beta} + O_P\big(n^{-\nu/(\mu+2\nu)}\big). \end{aligned}$$

This proves (5.4). Using (A.9), it follows that the least squares estimators $\widehat{\alpha}_j$ of $\tilde{\alpha}_j$ are given by

$$\begin{aligned} \widehat{\alpha}_j = \tilde{\alpha}_j &+ \hat{\lambda}_j^{-1} \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{ij} \varepsilon_i + \sum_{r=1}^S (\beta_r - \widehat{\beta}_r) \widehat{\psi}_j(\tau_r) \\ &- \sum_{r=1}^S \widehat{\beta}_r \big(\widehat{\psi}_j(\widehat{\tau}_r) - \widehat{\psi}_j(\tau_r)\big), \qquad j = 1, \ldots, k. \end{aligned} \tag{A.21}$$

But (A.11) and (5.4) imply that

$$\text{(A.22)}\quad \sum_{j=1}^{k}\Bigg(\sum_{r=1}^{S}(\beta_r-\widehat{\beta}_r)\widehat{\psi}_j(\tau_r)\Bigg)^2 = O_P\big(kn^{-2\nu/(\mu+2\nu)}\big) = O_P\big(n^{-(2\nu-1)/(\mu+2\nu)}\big),$$

while by (A.11) and (A.19)

$$\sum_{j=1}^{k}\big(\widehat{\psi}_j(\tau_r)-\widehat{\psi}_j(\widehat{\tau}_r)\big)^2 \le \frac{k}{\lambda_k}\frac{1}{n}\sum_{i=1}^{n}\big(X_i(\tau_r)-X_i(\widehat{\tau}_r)\big)^2 = O_P\big(n^{-(2\nu-1)/(\mu+2\nu)}\big),$$

and therefore

$$\text{(A.23)}\qquad \sum_{j=1}^{k}\Bigg(\sum_{r=1}^{S}\widehat{\beta}_r\big(\widehat{\psi}_j(\widehat{\tau}_r)-\widehat{\psi}_j(\tau_r)\big)\Bigg)^2 = O_P\big(n^{-(2\nu-1)/(\mu+2\nu)}\big).$$

Assertion (5.5) now is an immediate consequence of (A.10) and (A.21)–(A.23). □

SUPPLEMENTARY MATERIAL

Supplement to "Functional linear regression with points of impact" (DOI: 10.1214/15-AOS1323SUPP; .pdf). The supplementary document by Kneip, Poss and Sarda (2015) contains three Appendices. An application to NIR data can be found in Appendix A. In Appendix B, it is shown that the eigenfunctions of a Brownian motion satisfy assertion 2.5 in Theorem 2. Appendix C provides the proofs of Theorem 4 and Propositions 1 and 2.

REFERENCES

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

BOSQ, D. (2000). *Linear Processes in Function Spaces*: *Theory and Applications*. *Lecture Notes in Statistics* **149**. Springer, New York. MR1783138

CAI, T. T. and HALL, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179. MR2291496

CARDOT, H., FERRATY, F. and SARDA, P. (1999). Functional linear model. *Statist. Probab. Lett.* **45** 11–22. MR1718346

CARDOT, H. and JOHANNES, J. (2010). Thresholding projection estimators in functional linear models. *J. Multivariate Anal.* **101** 395–408. MR2564349

CARDOT, H., MAS, A. and SARDA, P. (2007). CLT in functional linear regression models. *Probab. Theory Related Fields* **138** 325–361. MR2299711

COMTE, F. and JOHANNES, J. (2012). Adaptive functional linear regression. *Ann. Statist.* **40** 2765–2797. MR3097959

CRAMBES, C., KNEIP, A. and SARDA, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37** 35–72. MR2488344

DELAIGLE, A. and HALL, P. (2012). Methodology and theory for partial least squares applied to functional data. *Ann. Statist.* **40** 322–352. MR3014309

FERRATY, F., HALL, P. and VIEU, P. (2010). Most-predictive design points for functional data predictors. *Biometrika* **97** 807–824. MR2746153

FRANK, I. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–135.

GILLESPIE, D. T. (1996). Exact numerical simulation of the Ornstein–Uhlenbeck process and its integral. *Phys. Rev. E* (3) **54** 2084–2091.

HALL, P. and HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269

HALL, P. and HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 109–126. MR2212577

HE, G., MÜLLER, H. G. and WANG, J. L. (2000). Extending correlation and regression from multivariate to functional data. In *Asymptotics in Statistics and Probability* 301–315. VSP, Leiden.

HSING, T. and REN, H. (2009). An RKHS formulation of the inverse regression dimension-reduction problem. *Ann. Statist.* **37** 726–755. MR2502649

JAMES, G. M., WANG, J. and ZHU, J. (2009). Functional linear regression that's interpretable. *Ann. Statist.* **37** 2083–2108. MR2543686

KNEIP, A., POSS, D. and SARDA, P. (2015). Supplement to "Functional linear regression with points of impact." DOI:10.1214/15-AOS1323SUPP.

KNEIP, A. and SARDA, P. (2011). Factor models and variable selection in high-dimensional regression analysis. *Ann. Statist.* **39** 2410–2447. MR2906873

MCKEAGUE, I. W. and SEN, B. (2010). Fractals with point impact in functional linear regression. *Ann. Statist.* **38** 2559–2586. MR2676898

MÜLLER, H.-G. and STADTMÜLLER, U. (2005). Generalized functional linear models. *Ann. Statist.* **33** 774–805. MR2163159

VAN DE GEER, S. and LEDERER, J. (2013). The Bernstein–Orlicz norm and deviation inequalities. *Probab. Theory Related Fields* **157** 225–250. MR3101846

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer, New York. MR1385671

ZHOU, S., LAFFERTY, J. and WASSERMAN, L. (2008). Time varying undirected graphs. In *Proceedings of the* 21*st Annual Conference on Computational Learning Theory (COLT 2008).* Available at http://arxiv.org/abs/0802.2758.

ZHOU, S., VAN DE GEER, S. and BÜHLMANN, P. (2009). Adaptive lasso for high dimensional regression and Gaussian graphical modeling. Available at http://arxiv.org/abs/0903.2515.

A. KNEIP
INSTITUT FÜR FINANZMARKTÖKONOMIK UND STATISTIK
DEPARTMENT OF ECONOMICS AND HAUSDORFF
CENTER FOR MATHEMATICS
UNIVERSITÄT BONN
ADENAUERALLEE 24-26
53113 BONN
GERMANY
E-MAIL: akneip@uni-bonn.de

D. POSS
BONN GRADUATE SCHOOL OF ECONOMICS
DEPARTMENT OF ECONOMICS
INSTITUT FÜR FINANZMARKTÖKONOMIK
UND STATISTIK
UNIVERSITÄT BONN
ADENAUERALLEE 24-26
53113 BONN
GERMANY
E-MAIL: dposs@uni-bonn.de

P. SARDA
INSTITUT DE MATHÉMATIQUES DE TOULOUSE; UMR 5219
UNIVERSITÉ DE TOULOUSE AND CNRS
118, ROUTE DE NARBONNE
31062 TOULOUSE CEDEX
FRANCE
E-MAIL: Pascal.Sarda@math.ups-tlse.fr