Contributions to Functional Data Analysis

With a Focus on

Points of Impact in Functional Regression

Inaugural-Dissertation

zur Erlangung des Grades eines Doktors
der Wirtschafts- und Gesellschaftswissenschaften
durch die
Rechts- und Staatswissenschaftliche Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Dominik Johannes Poß

aus Koblenz

2018

Dekan:	Prof. Dr. Daniel Zimmer, LL.M.
Erstreferent:	Prof. Dr. Alois Kneip
Zweitreferent:	JProf. Dr. Dominik Liebl
Tag der mündlichen Prüfung:	03.05.2018

Contents

List of Figures

Page

List of Tables

Page

Acknowledgments

First of all I would like to thank my supervisor Prof. Dr. Alois Kneip for his excellent supervision, guidance and steady encouragement during my studies. While I have already profited as a diploma student from his outstanding ability to intuitively explain even highly theoretical mathematical topics, it was him who aroused my curiosity for functional data analysis during my Ph.D. studies. I have the deepest respect for his remarkable knowledge, experience and keen perception concerning statistical topics in general and functional data analysis in specific. I have greatly benefited from his advices, fruitful discussions and valuable comments.

I would also like to thank my second supervisor JProf. Dr. Dominik Liebl. During the process of this thesis, he teached me a lot about the art of writing scientific papers while he patiently listened and helped me with my smaller and larger concerns on and off this thesis. I thank him for sharing his knowledge, ideas, creativity and experience with me during the process. It is needless to say that it helped a lot.

Moreover I have to express my gratitude to Prof. Dr. Lorens Imhof, who helped me with his untouched mathematical precision on several occasions, impressively helping me to solve problems I have been literally carrying around with me for several months. I learned a lot from him and enjoyed our discussions on and off topic.

Also I have to thank Professor Pascal Sarda. I profited a lot from working with him on one of the papers. The short time at Toulouse finishing a first draft of the paper will be in good memory.

It was a long way leading to this thesis. On this way not only the past couple of years as a Ph.D. student mattered but the foundations were already laid out during my diploma studies here in Bonn. In this regards I would also like to thank Prof. Dr. Klaus Schürger. As one of my lecturer during the very first courses of my diploma studies he left a deep impression on me which lead me to specialize in statistics during my economics studies. I thank him for his unselfishly willingness to help, his open ears and his useful ideas on some of my mathematical problems I have encountered during my studies.

A thank you also goes out to Klaus Utikal, Ph.D. I greatly enjoyed his lectures and benefited in particular from them by introducing me to R.

Focusing on statistics as a member of the Bonn Graduate School of Economics makes you more or less unique with regards to this group. I am thankful to my (alumi) fellow doctoral students Maximilian Conze and Thomas Nebeling. They not only made the time a lot more enjoyable, but Max has also been so kind to lend me a basic .tex template.

Another thank you is reserved for my colleague and fellow BGSE member Daniel Becker. I enjoy working with him a lot and I will really miss not only our small question rounds but also our pretty useful discussions on the “obvious”.

From my former colleagues I would like to thank Heiko Wagner and Oualid Bada for their steady support and helpful discussions on and off topic.

Finally, I am deeply grateful to my family and friends for their support and encouragement. My special thanks are reserved for Barbara Ahrens who has always supported me. I am grateful to be together with you.

Introduction

For several years now functional linear regression has been a standard tool for analyzing the relationship between a dependent scalar variable Y and a functional regressor X by proposing a model of the form

$$Y = \alpha + \int_a^b X(t)\beta(t)\,dt + \varepsilon.$$

Being the pendant of the multiple regression framework for the case of functional data, the functional linear regression model certainly constitutes one of the most important tools used to analyze functional data.

Somewhat surprisingly there can exist specific points $\tau_1, \ldots, \tau_S$ at which the trajectory of X may have an additional effect on the outcome Y which can not be captured within $\int_a^b X(t)\beta(t)\,dt$. The points $\tau_1, \ldots, \tau_S$ are called "points of impact" and their estimation is the main focus of this thesis.

By generalizing both, the classical functional linear regression model as well as the generalized functional linear model by allowing each of them to capture the additional effect of points of impact, this thesis constitutes an important contribution to the current research on functional data analysis. The thesis not only opens and answers new question about the identification and estimation of the points of impact but also provides an overall satisfying and detailed theoretical framework for the estimation of all involved model components.

In more detail, Chapter 1, which is joined work with Alois Kneip and Pascal Sarda, is concerned about the functional linear regression model with points of impact. The underlying paper has been published in the Annals of Statistics (Kneip et al., 2016a). The chapter constitutes an exhaustive theoretical framework for both, identification of points of impact and estimation of points of impact and associated parameters. The first part of this chapter is concerned about the identification of points of impact. For the identification of points of impact a new concept of "specific local variation" is introduced. It is shown that specific local variation constitutes a sufficient condition for the identification of points of impact and all model

parameters. It is then shown that specific local variation is a result of a certain approximation property of the eigenfunctions of the covariance operator and hence, for instance, the actual degree of smoothness of the trajectories is incidental.

Theoretical results for an estimator of the points of impact are derived under the assumption that the covariance function of the functional regressor is less smooth at the diagonal than everywhere else. Having derived estimates for the points of impact, one might then be interested in the remaining model parameters. Rates of convergence for these parameters are derived using results from Hall and Horowitz (2007). The performance of the estimation procedure is captured within a simulation study and the method is illustrated in an application using weather data. The chapter is complemented by a supplement which contains most of the proofs and another application using NIR data.

Chapter 2 is joined work with Dominik Liebl in collaboration with Hedwig Eisenbarth, Lisa Feldman Barrett and Tor Wager. In this part of the thesis results from the previous chapter are extended to a generalized functional linear model framework in which a linear predictor is connected to a real valued outcome through some function g. We derive a holistic theoretical framework for our estimates of the points of impact as well as the corresponding parameters. Quite remarkable our parameter estimates enjoy the same asymptotic properties as in the case where the points of impact are known. The behavior of our estimates is illustrated in a simulation study and finally applied to our data set, a psychological case study where participants were asked to continuously rate their emotional state during watching an affective video on the persecution of African albinos. A supplement to this chapter provides proofs of the theoretical statements and graphical representations of additional simulation results.
While driven by our application, this chapter focuses on a simplified model with $\beta(t) \equiv 0$ although proofs for the points of impact estimates are already tailored to contain the case $\beta(t) \neq 0$. Allowing for $\beta(t) \neq 0$ hence only affects results on the parameter estimates. The last part of the supplement to Chapter 2 is dedicated to briefly capture this setting. In this part, results on two different parameter estimators are introduced. While the first one is related to the instrumental variables estimation the second one relies on a basic truncation approach. Asymptotic theory for the latter estimator follows from using results from Müller and Stadtmüller (2005). The excursion closes with another simulation study and further proofs.

Chapter 3 is joined work with Heiko Wagner. It is an applied work that resulted from the CTW: "Statistics of Time Warpings and Phase Variations" at the Ohio State University. The underlying paper has been published in the Electronic Journal of Statistics (Poß and Wagner, 2014). The chapter focuses on the registration and interpretation of juggling data. The work of Kneip and Ramsay (2008) was adjusted to fit the multivariate nature of the juggling data. The registered data is then analyzed by an functional principal component analysis and a further investigation of the principal scores is performed.

Chapter 1

Functional Linear Regression with Points of Impact

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", i.e. 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 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 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.

1.1 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*

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

where ε_i, $i = 1,\dots,n$ 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, 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", i.e. *specific* time points $\tau_1,\dots,\tau_S$ with the property that the corresponding functional values $X_i(\tau_1),\dots,X_i(\tau_S)$ 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, $r = 1,\dots,S$, 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, $\tau_r \in (a,b)$, $r = 1,\dots,S$. 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 et al. (1999), Cardot et al. (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 et al. (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 et al. (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 non-zero $\beta(t)$.

A different approach based on impact points is proposed by Ferraty et al. (2010). For a pre-specified $q \in \mathbb{N}$ they aim to identify a function g as well as those design points $\tau_1,\dots,\tau_q$ which are "most influential" in the sense that $g(X_i(\tau_1),\dots,X_i(\tau_q))$ provides a best possible prediction of Y_i. Nonparametric smoothing methods are used to estimate g, while $\tau_1,\dots,\tau_q$ 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 $Y_i = \beta X_i(\tau) + \epsilon_i$ 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 $n^{-\frac{1}{2H}}$. The coefficient β can be estimated with a parametric rate of convergence $n^{-\frac{1}{2}}$.

There also exists a link between our approach and the work of Hsing and Ren (2009) who for a given grid $t_1,\dots,t_p$ 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 $k = 1,2,3,\dots$ allows to determine a suitable element $(\widehat{c}_1,\dots,\widehat{c}_p)^T$ of the eigenspace spanned by the eigenvectors of the k leading eigenvalues of the empirical covariance matrix of $(X_i(t_1),\dots,X_i(t_p))^T$. They then show consistency of the resulting estimators $\widehat{m}(X_i)$ as $n,p \to \infty$ and then $k \to \infty$. Note that (1.1) necessarily implies that $Y_i = m(X_i) + \epsilon_i$, 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 $\widehat{m}(X_i)$ 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 $\mathscr{Y}_i(t)$, 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 et al., 2000) will then assume that $Y_i := \mathscr{Y}_i(\tau) = \int_a^b \beta_\tau(t) X_i(t)dt + \varepsilon_i$, 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 $Y_i := \mathscr{Y}_i(\tau) = \int_a^b \beta_\tau(t) X_i(t)dt + \beta_1 X_i(\tau) + \varepsilon_i$ 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 $X_1,\dots,X_n$ denote a stationary time series of random function, and $\mathscr{Y}(\tau) \equiv X_i(\tau)$ 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 non-smooth 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 non-smooth 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 non-smooth trajectories covers a wide range of applications. Examples are given in Section 7 and in the supplementary material (Kneip et al., 2016b), 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. Section 8 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.

1.2 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 $X_i(\tau_1),\dots,X_i(\tau_S)$ at the points of impact $\tau_1,\dots,\tau_S$. 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 1.1. *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 $f_{\epsilon,t}(s) := \frac{cov(X(s),\zeta_{\epsilon,t}(X))}{var(\zeta_{\epsilon,t}(X))}$ the following conditions are satisfied:*

i) $0 < var(\zeta_{\epsilon,t}(X)) < \infty$,
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

$$X(s) = X_{\epsilon,t}(s) + \zeta_{\epsilon,t}(X)f_{\epsilon,t}(s), \quad s \in [a,b], \tag{1.2}$$

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

$$\sigma(s,u) = \sigma_{\epsilon,t}(s,u) + var(\zeta_{\epsilon,t}(X))f_{\epsilon,t}(s)f_{\epsilon,t}(u), \quad s,u \in [a,b]. \tag{1.3}$$

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 $f_{\epsilon,t}(s) := \frac{cov(X(s),\zeta_{\epsilon,t}(X))}{var(\zeta_{\epsilon,t}(X))} = 1$ for

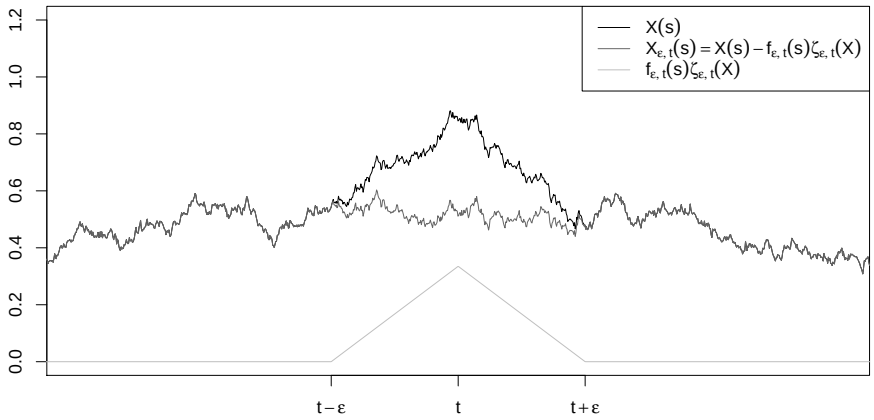

Figure 1.1: *The figure illustrates the decomposition of a trajectory from a Brownian motion X (black) in $X_{\epsilon,t}$ (grey) and $\zeta_{\epsilon,t}(X)f_{\epsilon,t}$ (light grey). The component $\zeta_{\epsilon,t}(X)f_{\epsilon,t}$ can be seen to quantify the local variation of X in an interval around t.*

$t = s$, while $f_{\epsilon,t}(s) = 0$ for all $s \in [a,b]$ with $|t-s| \geq \epsilon$. Figure 1.1 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

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

Theorem 1.1. *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*

$$\mathbb{E}\left(\left(m(X) - \int_a^b \beta^*(t)X(t)dt - \sum_{r=1}^{S^*} \beta_r^* X(\tau_r)\right)^2\right) > 0, \tag{1.4}$$

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 $\lambda_1 \geq \lambda_2 \geq \ldots$ to denote the non-zero 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

$$X(t) = \sum_{r=1}^{\infty} \langle X, \psi_r \rangle \psi_r(t), \tag{1.5}$$

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 $\mathscr{C}(t, \epsilon, [a,b])$ 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 1.2. *Let $\psi_1, \psi_2, \ldots$ be a system of orthonormal eigenfunctions corresponding to the non-zero eigenvalues of the covariance operator Γ of X. If for all $t \in (a,b)$ there exists an $\epsilon_t > 0$ such that*

$$\lim_{k \to \infty} \inf_{f \in \mathscr{C}(t, \epsilon, [a,b])} \sup_{s \in [a,b]} |f(s) - \sum_{r=1}^{k} \langle f, \psi_r \rangle \psi_r(s)| = 0 \quad \textit{for every } 0 < \epsilon < \epsilon_t, \tag{1.6}$$

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 eigen-

functions $\psi_r(t) = \sqrt{2}\sin((r-1/2)\pi t)$, $r = 1,2,\ldots$. In the Supplementary Appendix B it is verified that this system of orthonormal eigenfunctions satisfies (1.6). 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 1.2 states that any process of this type still has specific local variation.

1.3 Covariance functions which are non-smooth 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 non-smoothness. As pointed out in the introduction, we also believe that assuming non-smooth 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 non-smooth 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 e.g. 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 non-smooth sample paths which will be expressed in terms of a non-smooth diagonal of the corresponding covariance function $\sigma(t,s)$. It will be assumed that $\sigma(t,s)$ possesses non-smooth 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 (1.3) a non-smooth 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

$$\sigma(t,s) = \sigma(\frac{1}{2}(t+s+|t-s|), \frac{1}{2}(t+s-|t-s|)) =: \omega^*(t+s,|t-s|) \quad \text{for all } t,s \in [a,b].$$

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

$$\frac{\partial}{\partial y_+}\omega^*(2t,y)|_{y=0} = \frac{\partial}{\partial y}\sigma(t+\frac{y}{2}, t-\frac{y}{2})|_{y=0} = \frac{1}{2}(\frac{\partial}{\partial s}\sigma(s,t)|_{s=t} - \frac{\partial}{\partial s}\sigma(t,s)|_{s=t}) = 0. \tag{1.7}$$

In contrast, any process with $\frac{\partial}{\partial y_+}\omega^*(x,y)|_{y=0} \neq 0$ is non-smooth 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 1.1 below which, as will be shown in Theorem 1.3, 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 1.1. *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 $t,s \in [a,b]$*

$$\sigma(t,s) = \omega(t,s,|t-s|^\kappa). \tag{1.8}$$

Moreover,

$$0 < \inf_{t\in[a,b]} c(t), \quad \text{where } c(t) := -\frac{\partial}{\partial z}\omega(t,t,z)|_{z=0}. \tag{1.9}$$

One can infer from (1.7) that for every twice continuously differentiable covariance function σ there exists some function ω such that (1.8) holds with $\kappa = 2$. But note that formally introducing $|t-s|^\kappa$ as an extra argument establishes an easy way of capturing non-smooth behavior as $|t-s| \to 0$, since σ is not twice differentiable at the diagonal if $\kappa < 2$. In Assump-

tion 1.1 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 non-smooth 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

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

In this case Assumption 1.1 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

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

Then Assumption 1.1 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 1.3 below now states that any process respecting Assumption 1.1 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

$$Z_\delta(X,t) \;=\; X(t) - \frac{1}{2}\,(X(t-\delta) + X(t+\delta))\,. \tag{1.10}$$

Theorem 1.3. *Under our setup assume that the covariance function σ of X satisfies Assumption 1.1. 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 (1.10).*

1.4 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), $i = 1,\ldots,n$, 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 $t_j = a + \frac{j-1}{p-1}(b-a)$, $j = 1,\ldots,p$.

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 the-

oretical results developed in this section remain unchanged for this situation with however substantially longer proofs.

Determining $\tau_1, \ldots, \tau_S$ 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 $Var(X_i(t_j) - X_i(t_{j-m})) \to 0$ 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 et al. (2009) indicate that convergence of parameter estimates *at least* requires that $\sqrt{n/\log p}(Var(X_i(t_j) - X_i(t_{j-1}))) \to \infty$ as $n \to \infty$. This follows from the distribution version of the restricted eigenvalue assumption and Theorem 5.2 of Bickel et al. (2009) (see also Zhou et al. (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 preceeding 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, Lemma C.3 and Lemma C.4 in the Supplementary Appendix C show that as $\delta \to 0$ and $\min_{r\neq s}|\tau_s - \tau_r| \gg \delta$

$$\begin{aligned}
\mathbb{E}\left(Z_{\delta,i}(t_j)Y_i\right) &= \beta_r c(\tau_r)\delta^\kappa + O(\max\{\delta^{\kappa+1}, \delta^2\}) \quad \text{if } |t_j - \tau_r| \approx 0 \\
\mathbb{E}\left(Z_{\delta,i}(t_j)Y_i\right) &= O(\max\{\delta^{\kappa+1}, \delta^2\}) \qquad \text{if } \min_{r=1,\ldots,S}|t_j - \tau_r| \gg \delta.
\end{aligned}$$

Moreover, assuming that the process X possesses a Gaussian distribution, then, since it holds that $Var(Z_{\delta,i}(t_j)) = O(\delta^\kappa)$ (see (1.26) in the proof of Theorem 1.3), the Cauchy-Schwarz inequality leads to $Var(Z_{\delta,i}(t_j)Y_i) = O(\delta^\kappa)$, and hence

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

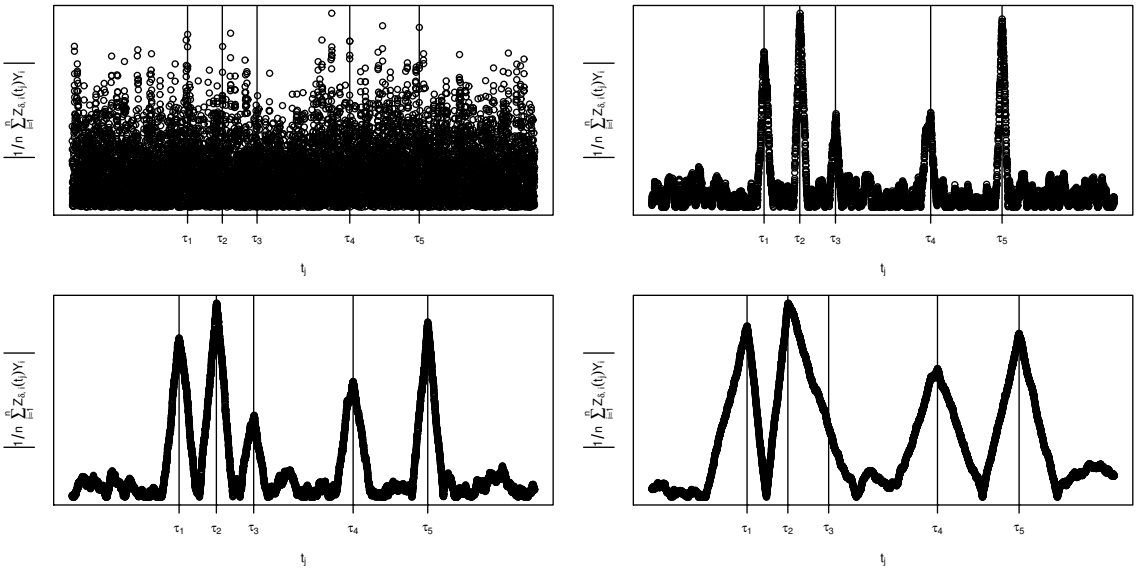

Figure 1.2: *The figure shows* $|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t_j)Y_i|$ *for different choices of* δ *in a point of impact model with* 5 *points of impact whose locations are indicated by vertical lines. The upper left panel corresponds to a very small* δ*, where the noise level overlays the 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.*

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 1.2. 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 Ornstein-Uhlenbeck process with $\theta = 5$ and $\sigma_u = 3.5$, evaluated over $p = 10001$ 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 $\delta = 10/10001 \approx 5/n$, $\delta = 142/10001 \approx 1/\sqrt{n}$, $\delta = 350/10001 \approx 2.47/\sqrt{n}$, and $\delta = 750/10001 \approx 5.3/\sqrt{n}$.
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 1.2 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 $1 \leq k_\delta < \frac{p-1}{2}$ and $\delta = k_\delta(b-a)/(p-1)$. In a first step determine for all $j \in \mathscr{J}_{0,\delta} := \{k_\delta + 1, \ldots, p - k_\delta\}$

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

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

- Determine
$$j_l = arg \max_{j \in \mathscr{J}_{l-1,\delta}} |\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t_j) Y_i|$$
and set $\widehat{\tau}_l := t_{j_l}$.

- Set $\mathscr{J}_{l,\delta} := \{j \in \mathscr{J}_{l-1,\delta} \,|\, |t_j - \widehat{\tau}_l| \geq \sqrt{\delta}/2\}$, i.e eliminate all points in an interval of size $\sqrt{\delta}$ around $\widehat{\tau}_l$. Stop iteration if $\mathscr{J}_{l,\delta} = \emptyset$.

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

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

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

A theoretical justification for this estimation procedure is given by Theorem 1.4. Its proof along with the proofs of Proposition 1.1 and 1.2 below can be found in the Supplementary Appendix C. Theory relies on an asymptotics $n \to \infty$ with $p \equiv p_n \geq L n^{1/\kappa}$ for some constant $0 < L < \infty$. It is based on the following additional assumption on the structure of X and Y.

Assumption 1.2.

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 1.4. *Under our setup and Assumptions 1.1 as well as 1.2 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*

$$\max_{r=1,\ldots,\widehat{S}} \min_{s=1,\ldots,S} |\widehat{\tau}_r - \tau_s| \;=\; O_P(n^{-\frac{1}{k}}). \tag{1.11}$$

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

$$\lambda \equiv \lambda_n = A\sqrt{\frac{Var(Y_i)}{n}\log(\frac{b-a}{\delta})}, \quad \textit{where } A > \sqrt{2}.$$

Then

$$P(\widehat{S} = S) \;\to\; 1 \quad \textit{as } n \to \infty. \tag{1.12}$$

The theorem of course implies that the rates of convergence of the estimated points of impact depend on κ. If $\kappa = 1$, as e.g. for the Brownian motion or the Ornstein-Uhlenbeck process, then $\max_{r=1,\dots,\widehat{S}} \min_{s=1,\dots,S} |\widehat{\tau}_r - \tau_s| = O_P(n^{-1})$. Arbitrarily fast rates of convergence can be achieved for very non-smooth 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 wich are not in $\sqrt{\delta}/2$-neighborhoods of any previously selected $\widehat{\tau}_1, \dots, \widehat{\tau}_{l-1}$. 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 (1.11) it suffices to eliminate all points in $\delta|\log\delta|$ neighborhoods of $\widehat{\tau}_1, \dots, \widehat{\tau}_{l-1}$ 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 $\widehat{\tau}_1, \dots, \widehat{\tau}_{M_\delta}$ 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 $X_i(\widehat{\tau}_1), \dots, X_i(\widehat{\tau}_{M_\delta})$, and by calculating the residual sum of squares RSS_s for each subset of size s. An estimate $\widehat{S}$ is obtained by minimizing

$$BIC_s = n\log(RSS_s/n) + s\log(n) \tag{1.13}$$

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 $\tilde{\tau}_1, \dots, \tilde{\tau}_S \in \{\widehat{\tau}_1, \dots, \widehat{\tau}_{M_\delta}\}$

one may consider an approximate relationship which resembles an "augmented model" as proposed by Kneip and Sarda (2011) in a different context:

$$Y_i \approx \sum_{r=1}^{k} \alpha_r \langle X_i, \widehat{\psi}_r \rangle + \sum_{r=1}^{S} \beta_r X_i(\tilde{\tau}_r) + \varepsilon_i^*. \tag{1.14}$$

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 $\widehat{\tau}_r$, $r = 1, 2, \ldots$. 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 (1.14). 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 1.1 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 $\mathbb{E}(Z_{\delta,i}(t_j)^2) = \delta^\kappa \left(2c(t_j) - \frac{2^\kappa}{2} c(t_j)\right) + o(\delta^\kappa)$ (see (1.26) in the proof of Theorem 1.3). 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

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

of κ. In Proposition 1.1 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 non-smooth diagonal.

Proposition 1.1. *Under the conditions of Theorem 1.4 we have*

$$\widehat{\kappa} = \kappa + O_P(n^{-1/2} + \delta^{\min\{2, 2/\kappa\}}). \tag{1.15}$$

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 $\tau_r = arg\min_{s=1,\ldots,S} |\widehat{\tau}_r - \tau_s|$, $r = 1, \ldots, S$.

Proposition 1.2. *Under the assumptions of Theorem 1.4 we obtain for every* $r = 1, \dots, S$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i(\tau_r) - X_i(\widehat{\tau}_r))^2 = O_p(n^{-1}), \tag{1.16}$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i(\tau_r) - X_i(\widehat{\tau}_r))\varepsilon_i = O_p(n^{-1}). \tag{1.17}$$

1.5 Parameter estimates

Recall that Assumption 1.1 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 (1.16) as well as (1.17) 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 encounter $Y_i = \sum_{r=1}^{S}\beta_r X_i(\tau_r) + \varepsilon_i$, $i = 1, \dots, n$, 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

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{r=1}^{\widehat{S}} b_r X_i(\widehat{\tau}_r))^2 \tag{1.18}$$

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

Let $\mathbf{X}_i(\boldsymbol{\tau}) := (X_i(\tau_1), \dots, X_i(\tau_S))^T$, 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 1.1 in particular implies that Σ_τ is invertible.

If $\widehat{S} = S$, then by (1.16) and (1.17) the differences between $\widehat{\tau}_r$ and τ_r, $r = 1, \dots, S$ 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:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to_D N(0, \sigma^2\Sigma_\tau^{-1}) \tag{1.19}$$

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 (1.14). Thus let $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \dots$ and $\widehat{\psi}_1, \widehat{\psi}_2, \dots$ denote eigenvalues and eigenfunctions of the empirical covariance operator of $X_1, \dots, X_n$. Given estimates $\widehat{\tau}_1, \dots, \widehat{\tau}_{\widehat{S}}$ 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

$$\sum_{i=1}^{n} \left(Y_i - \sum_{r=1}^{k} a_r \langle X_i, \widehat{\psi}_r \rangle - \sum_{r=1}^{\widehat{S}} b_r X_i(\widehat{\tau}_r) \right)^2 \tag{1.20}$$

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 (1.20) 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 $\widehat{\lambda}_j = 0$ for $j > n$).

Our theoretical analysis of the estimators defined by (1.20) 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 1.2 their results are additionally based on the following assumption on the eigendecompositions of X and β:

Assumption 1.3.

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 $|\alpha_j| \leq C_0 j^{-\nu}$ 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 $\boldsymbol{M}_k$ denote the $S \times S$ matrix with elements $\sigma^{[k]}(\tau_r, \tau_s)$, $r, s = 1, \ldots, S$. Furthermore, let $\lambda_{min}(\boldsymbol{M}_k)$ denote the smallest eigenvalue of the matrix $\boldsymbol{M}_k$.

Assumption 1.4.

a) $\sup_t \sup_j \psi_j(t)^2 \leq C_\psi$ for some $C_\psi < \infty$.

b) There exists some $0 < C_1 < \infty$ such that $\lambda_j \leq C_1 j^{-\mu}$ for all j.

c) There exists some $0 < D < \infty$ such that $\lambda_{min}(\boldsymbol{M}_k) \geq D k^{-\mu+1}$ for all k.

Condition a) is, for example, satisfied if $\psi_1, \psi_2, \ldots$ correspond to a Fourier-type basis. Note that Assumption 1.3 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 $\sum_{j=k+1}^{\infty} \lambda_j \leq C_2 k^{-\mu+1}$ as well as $|\sigma^{[k]}(t,s)| \leq C_2 C_\psi^2 k^{-\mu+1}$ for some $C_2 < \infty$ and all k. Condition c) therefore only introduces an additional regularity condition on the matrix $\boldsymbol{M}_k$. 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 1.5. *Under our setup and Assumptions 1.2 - 1.4 suppose that $\widehat{S} = S$ and that estimators $\widehat{\tau}_r$ satisfy (1.16) as well as (1.17) for all $r = 1,\ldots,S$. If additionally $k = O(n^{1/(\mu+2\nu)})$ and $n^{1/(\mu+2\nu)} = O(k)$ as $n \to \infty$, then*

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 = O_p(n^{-2\nu/(\mu+2\nu)}), \tag{1.21}$$

$$\int_a^b (\widehat{\beta}(t) - \beta(t))^2 dt = O_p(n^{-(2\nu-1)/(\mu+2\nu)}). \tag{1.22}$$

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, $r = 1,\ldots,S$, achieve a slightly faster rate of convergence.

1.6 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 1,000 repetitions of the simulation experiments. The corresponding R-code can be obtained from the authors upon request.

The data $X_1,\ldots,X_n$ 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 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 1.1, 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

Table 1.1: *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 $\beta(t) \equiv 0$											
1,001	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 $\beta(t) \neq 0$											
1,001	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

by performing best subset selection with the BIC-criterion via the R package bestglm on the augmented model (1.14)

$$Y_i \approx \sum_{r=1}^{k} \alpha_r \langle X_i, \widehat{\psi}_r \rangle + \sum_{r=1}^{\widetilde{S}} \beta_r X_i(\tilde{\tau}_r) + \varepsilon_i^*. \tag{1.23}$$

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 1.1 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 $MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2$ 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 $\lambda = 2\sqrt{\widehat{Var}(Y)/n \log\left(\frac{b-a}{\delta}\right)}$, where $\widehat{Var}(Y)$ denotes the estimated sample variance of Y_i. The cut-off criterion yields very reliable estimates $\widehat{S}$ of S for

$n = 5,000$, 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 $n = 5,000$ 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 partitioned into $I_1 = \left[0, \frac{1}{2}(\tau_1+\tau_2)\right[$ and $I_2 = \left[\frac{1}{2}(\tau_1+\tau_2), 1\right]$. 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 $n = 5,000$ and $p = 1,001$ 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 $p = 1,001$ 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 $p = 1,001$ 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.

1.7 Application 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 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 (1.11).

After estimating $\widetilde{S}$ possible candidates for the points of impact, the approximate model (1.14),

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

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 max-

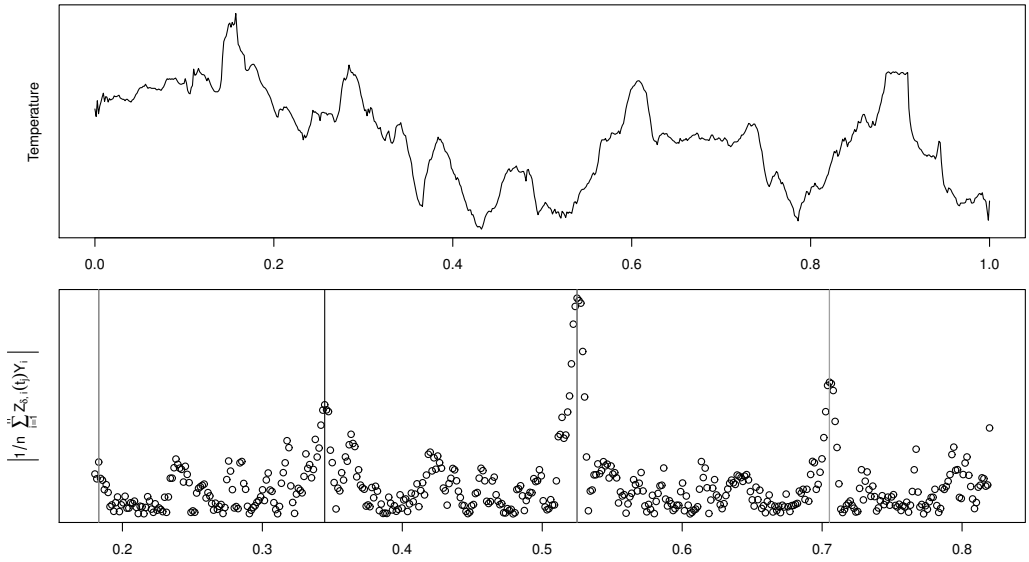

Figure 1.3: *The upper panel of this figure shows a trajectory from the observed temperature curves of the Canadian weather data. The lower panel shows* $|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t_j)Y_i|$ *during the selection procedure. Locations of selected points of impact in the augmented model are indicated by grey lines. The location of the remaining candidate is displayed by a black line.*

imum 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 non available 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 (1.14). 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

Table 1.2: *Estimated number of principal components k, points of impact S, prediction error and the median of* $(y_i - \widehat{y}_i)^2$ *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

tuning parameters δ, S, and/or k in a first step. The mean squared prediction error $MSPE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ 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 $(y_i - \widehat{y}_i)^2$, $i = 1, \ldots, n$, has been calculated as a more robust measure of the error. Depicted in the upper panel of Figure 1.3 is the observed temperature trajectory for the weather station "McTavish", showing a rather rough process. The lower panel of this figure show $|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t_j)Y_i|$ 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.

Table 1.2 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 $(y_i - \widehat{y}_i)^2$, 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 $\delta \approx 0.3$, i.e. 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 1.3.

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 insights. 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 non-smooth. 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".

1.8 Proofs of some theorems

Proof of Theorem 1.1. 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 (1.2) we obtain a decomposition into two uncorrelated components $X_{\epsilon,\tau_j}(\cdot)$ and $\zeta_{\epsilon,\tau_j}(X) f_{\epsilon,\tau_j}(\cdot)$:

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

By condition iv) we have

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

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

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

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 1.2. 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 $f \in \mathscr{C}(t,\epsilon,[a,b])$ 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 $\mathscr{C}(t,\epsilon,[a,b])$ then implies that $f_k(t) \geq 1 - \epsilon/3$ as well as

$$\begin{aligned} \sup_{s\in[a,b]} |f_k(s)| &\leq 1 + \frac{\epsilon}{3} \leq (1+\epsilon)(1-\frac{\epsilon}{3}) \leq (1+\epsilon) f_k(t), \\ \sup_{s\in[a,b], s\notin[t-\epsilon,t+\epsilon]} |f_k(s)| &\leq \frac{\epsilon}{3} \leq \epsilon(1-\frac{\epsilon}{3}) \leq \epsilon f_k(t). \end{aligned} \tag{1.24}$$

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 $var(\langle X, \psi_r\rangle) = \lambda_r$. By (1.5) we obtain

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

Furthermore, $var(\zeta_{\epsilon,t}(X)) = \sum_{r=1}^k \frac{\langle f,\psi_r\rangle^2}{\lambda_r} > 0$, and it thus follows from (1.24) 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 1.3. First note that Assumption 1.1 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 $M_1 < \infty$

$$\begin{aligned} \mathbb{E}(X(t)Z_\delta(X,t)) &= \sigma(t,t) - \frac{1}{2}\sigma(t,t-\delta) - \frac{1}{2}\sigma(t,t+\delta) \\ &= \omega(t,t,0) - \frac{1}{2}\omega(t,t-\delta,\delta^\kappa) - \frac{1}{2}\omega(t,t+\delta,\delta^\kappa) \\ &= \delta^\kappa c(t) + R_{1;\delta,t}, \quad \text{with} \sup_{t\in[a+\delta,b-\delta]} |R_{1;\delta,t}| \leq M_1 \delta^{\min\{2\kappa,2\}}. \end{aligned} \tag{1.25}$$

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

$$\begin{aligned} var(Z_\delta(X,t)) &= 2\omega(t,t,0) - \omega(t,t-\delta,\delta^\kappa) - \omega(t,t+\delta,\delta^\kappa) - \frac{1}{2}(\omega(t,t,0) - \omega(t+\delta,t-\delta,(2\delta)^\kappa)) \\ &\quad - \frac{1}{4}(2\omega(t,t,0) - \omega(t-\delta,t-\delta,0) - \omega(t+\delta,t+\delta,0)) \\ &= \delta^\kappa\left(2c(t) - \frac{2^\kappa}{2}c(t)\right) + R_{2;\delta,t}, \quad \text{with} \sup_{t\in[a+\delta,b-\delta]} |R_{2;\delta,t}| < M_2\delta^{\min\{2\kappa,2\}} \end{aligned} \tag{1.26}$$

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 $u \in [-c, c]$

$$\begin{aligned}\mathbb{E}(X(t+u\delta)Z_\delta(X,t)) &= \sigma(t+u\delta,t) - \frac{1}{2}\sigma(t+u\delta,t-\delta) - \frac{1}{2}\sigma(t+u\delta,t+\delta)\\ &= \omega(t,t,0) - \frac{1}{2}\omega(t,t-\delta,\delta^\kappa) - \frac{1}{2}\omega(t,t+\delta,\delta^\kappa)\\ &\quad - c(t)\delta^\kappa\left(|u|^\kappa - \frac{1}{2}(|u+1|^\kappa - 1) - \frac{1}{2}(|u-1|^\kappa - 1)\right) + R_{3;c,u,\delta,t} \qquad (1.27)\\ &= -c(t)\delta^\kappa\left(|u|^\kappa - \frac{1}{2}|u+1|^\kappa - \frac{1}{2}|u-1|^\kappa\right) + R_{4;c,u,\delta,t}, \qquad (1.28)\end{aligned}$$

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

$$\sup_{t\in[a+\delta,b-\delta]} R_{3;c,u,\delta,t} \le M_{3,c}(|u|^{1/2}\delta)^{\min\{2\kappa,2\}}, \qquad \sup_{t\in[a+\delta,b-\delta]} R_{4;c,u,\delta,t} \le M_{4,c}\delta^{\min\{2\kappa,2\}}$$

hold for all $u \in [-c, c]$. Finally, Assumption 1.1 implies that there exists a constant $M_5 < \infty$ such that for all $s \in [a, b]$ with $|t - s| \ge \delta$

$$\begin{aligned}|\mathbb{E}(X(s)Z_\delta(X,t))| &= |\omega(s,t,|s-t|^\kappa) - \frac{1}{2}\omega(s,t-\delta,|s-t+\delta|^\kappa) - \frac{1}{2}\omega(s,t+\delta,|s-t-\delta|^\kappa)|\\ &\le \begin{cases} M_5\frac{\delta^2}{|t-s|^{2-\kappa}} & \text{if } \kappa \ne 1\\ M_5\delta^2 & \text{if } \kappa = 1.\end{cases} \qquad (1.29)\end{aligned}$$

It follows from (1.25), (1.28), and (1.29) that for arbitrary $t \in (a, b)$ and any $\epsilon > 0$ there exist a $\delta_\epsilon > 0$ as well as a constant $a_\epsilon \ge 1$ such that for all $\delta \le \delta_\epsilon$

$$\begin{aligned}&|\mathbb{E}(X(s)Z_\delta(X,t))| \le (1+\epsilon)\mathbb{E}(X(t)Z_\delta(X,t)) \quad \text{for all } s \in [a,b], s \ne t\\ &|\mathbb{E}(X(s)Z_\delta(X,t))| \le \epsilon\cdot\mathbb{E}(X(t)Z_\delta(X,t)) \quad \text{for all } s \in [a,b], |s-t| \ge a_\epsilon\delta.\end{aligned}$$

Together with (1.26), the assertion of the theorem is an immediate consequence. □

Proof of Theorem 1.5. Let $\widehat{\theta}_{ij} := \langle X_i, \widehat{\psi}_j\rangle$, $\theta_{ij} := \langle X_i, \psi_j\rangle$, and $\tilde{\alpha}_j := \langle \beta, \widehat{\psi}_j\rangle$ for all i, j. Using empirical eigenfunctions we obtain $X_i = \sum_{j=1}^n \widehat{\theta}_{ij}\widehat{\psi}_j$ and $\int_a^b \beta(t)X_i(t)dt = \sum_{j=1}^n \tilde{\alpha}_j\widehat{\theta}_{ij}$. Therefore,

$$Y_i = \sum_{j=1}^n\left(\tilde{\alpha}_j + \sum_{r=1}^S \beta_r\widehat{\psi}_j(\tau_r)\right)\widehat{\theta}_{ij} + \varepsilon_i, \qquad (1.30)$$

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

$$\sum_{j=1}^k a_j\widehat{\theta}_{ij} + \sum_{r=1}^S b_r X_i(\widehat{\tau}_r) = \sum_{j=1}^k\left(a_j + \sum_{r=1}^S b_r\widehat{\psi}_j(\widehat{\tau}_r)\right)\widehat{\theta}_{ij} + \sum_{j=k+1}^n\sum_{r=1}^S b_r\widehat{\psi}_j(\widehat{\tau}_r)\widehat{\theta}_{ij} \qquad (1.31)$$

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

$$\widehat{\alpha}(\boldsymbol{b})_j = \tilde{\alpha}_j + \widehat{\lambda}_j^{-1}\frac{1}{n}\sum_{i=1}^n \widehat{\theta}_{ij}\varepsilon_i + \sum_{r=1}^S (\beta_r \widehat{\psi}_j(\tau_r) - b_r \widehat{\psi}_j(\widehat{\tau}_r)), \quad j = 1, \dots, k. \tag{1.32}$$

Note that $\tilde{\alpha}_j + \widehat{\lambda}_j^{-1}\frac{1}{n}\sum_{i=1}^n \widehat{\theta}_{ij}\varepsilon_i$ 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

$$\int_a^b \left(\beta(t) - \sum_{j=1}^k (\tilde{\alpha}_j + \widehat{\lambda}_j^{-1}\frac{1}{n}\sum_{i=1}^n \widehat{\theta}_{ij}\varepsilon_i)\widehat{\psi}_j(t)dt\right)^2 dt = O_p(n^{-(2\nu-1)/(\mu+2\nu)}). \tag{1.33}$$

Further analysis requires to analyze the differences between θ_{ij}, ψ_j and their empirical counterparts $\widehat{\theta}_{ij}, \widehat{\psi}_j$. By Assumptions 1.2 - 1.4 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 $q = 1, 2, 3, \dots$ there exists some $A_q, B_q < \infty$ such that

$$E\left(|\lambda_j - \widehat{\lambda}_j|^q\right) \leq A_q n^{-q/2}, \quad \sup_t E\left(|\widehat{\psi}_j(t) - \psi_j(t)|\right) \leq B_q n^{-q/2} j^{q(\mu+1)}, \quad j = 1, \dots, k+1 \tag{1.34}$$

for all sufficiently large n. Let $X_i^{[k]} := X_i - \sum_{j=1}^k \widehat{\theta}_{ij}\widehat{\psi}_j$. Recall that $\lambda_j = O(j^{-\mu})$ and note that by Assumptions 1.3 and 1.4, $n^{-1/2}n^{2/(\mu+2\nu)} = o(n^{(-\mu+1)/(\mu+2\nu)})$, while $n^{(-\mu+1)/(\mu+2\nu)} = O(\sigma^{[k]}(\tau_r, \tau_r))$. By (1.34) we thus obtain for all $t, s \in [a, b]$

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n X_i^{[k]}(t)X_i^{[k]}(s) &= \frac{1}{n}\sum_{i=1}^n X_i(t)X_i(s) - \sum_{j=1}^k \widehat{\lambda}_j \widehat{\psi}_j(t)\widehat{\psi}_j(s) \\
&= \sigma(t,s) - \sum_{j=1}^k \lambda_j \psi_j(t)\psi_j(s) + \sum_{j=1}^k \lambda_j(\psi_j(t)\psi_j(s) - \widehat{\psi}_j(t)\widehat{\psi}_j(s)) \\
&\quad + \sum_{j=1}^k (\lambda_j - \widehat{\lambda}_j)\widehat{\psi}_j(t)\widehat{\psi}_j(s) + O_P(n^{-1/2}) \\
&= \sigma^{[k]}(t,s) + O_P(n^{-1/2}n^{2/(\mu+2\nu)}) = \sigma^{[k]}(t,s) + o_P(n^{(-\mu+1)/(\mu+2\nu)}).
\end{aligned} \tag{1.35}$$

At the same time (1.16) leads to

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(X_i^{[k]}(\tau_r)-X_i^{[k]}(\widehat{\tau}_r))^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i(\tau_r)-X_i(\widehat{\tau}_r))^2-\sum_{j=1}^{k}\widehat{\lambda}_j(\widehat{\psi}_j(\tau_r)-\widehat{\psi}_j(\widehat{\tau}_r))^2\\ &\leq \frac{1}{n}\sum_{i=1}^{n}(X_i(\tau_r)-X_i(\widehat{\tau}_r))^2 = O_P(n^{-1}).\end{aligned} \tag{1.36}$$

for all $r=1,\ldots,S$. Expressions (1.35) and (1.36) together imply that for all r,s

$$\frac{1}{n}\sum_{i=1}^{n}X_i^{[k]}(\widehat{\tau}_r)X_i^{[k]}(\widehat{\tau}_s)=\sigma^{[k]}(\tau_r,\tau_s)+o_P(n^{(-\mu+1)/(\mu+2\nu)}). \tag{1.37}$$

Let $\mathbf{X}_i^{[k]} := (X_i^{[k]}(\widehat{\tau}_1),\ldots,X_i^{[k]}(\widehat{\tau}_S))^T$ and note that by (1.37) we have $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{[k]}(\mathbf{X}_i^{[k]})^T = \boldsymbol{M}_k + o_P(n^{(-\mu+1)/(\mu+2\nu)})$. By Assumption 1.4 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,

$$n^{(-\mu+1)/(\mu+2\nu)}(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{[k]}(\mathbf{X}_i^{[k]})^T)^{-1} = n^{(-\mu+1)/(\mu+2\nu)}(\boldsymbol{M}_k)^{-1}+o_P(1) \tag{1.38}$$

and hence by (1.30) - (1.32) the least squares estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ can be written in the form

$$\widehat{\boldsymbol{\beta}} = (\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{[k]}(\mathbf{X}_i^{[k]})^T)^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{[k]}\left(\sum_{r=1}^{S}\beta_r X_i^{[k]}(\tau_r)+\sum_{j=k+1}^{n}\tilde{\alpha}_j\widehat{\theta}_{ij}+\varepsilon_i\right). \tag{1.39}$$

By (1.36) and (1.37) we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{[k]}\sum_{r=1}^{S}\beta_r X_i^{[k]}(\tau_r)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i^{[k]}(\mathbf{X}_i^{[k]})^T\boldsymbol{\beta}+O_P(n^{(-\mu+1)/2(\mu+2\nu)}\cdot n^{-1/2}). \tag{1.40}$$

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

$$\begin{aligned}&|\frac{1}{n}\sum_{i=1}^{n}X_i^{[k]}(\widehat{\tau}_r)(\sum_{j=k+1}^{n}\tilde{\alpha}_j\widehat{\theta}_{ij})| = |\sum_{j=k+1}^{n}\tilde{\alpha}_j\widehat{\lambda}_j\widehat{\psi}_j(\widehat{\tau}_r)|\\ &\leq \sqrt{\sum_{j=k+1}^{n}\widehat{\lambda}_j\tilde{\alpha}_j^2}\sqrt{\sum_{j=k+1}^{n}\widehat{\lambda}_j\widehat{\psi}_j(\widehat{\tau}_r)^2} \leq \sqrt{\widehat{\lambda}_{k+1}\sum_{j=k+1}^{n}\tilde{\alpha}_j^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}X_i^{[k]}(\widehat{\tau}_r)^2}\\ &= O_P(n^{-(\mu+2\nu-1)/2(\mu+2\nu)}\cdot n^{(-\mu+1)/2(\mu+2\nu)})\end{aligned} \tag{1.41}$$

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

$$|\widehat{\psi}_j(\tau_r) - \widehat{\psi}_j(\widehat{\tau}_r)| = |\widehat{\lambda}_j^{-1} \frac{1}{n} \sum_{i=1}^n \widehat{\theta}_{ij}(X_i(\tau_r) - X_i(\widehat{\tau}_r))| \leq \widehat{\lambda}_j^{-1/2} \sqrt{\frac{1}{n} \sum_{l=1}^n (X_l(\tau_r) - X_l(\widehat{\tau}_r))^2}. \tag{1.42}$$

Now note that by the independence of $\widehat{\theta}_{ij}$ and ε_i we have $\widehat{\lambda}_j^{-1/2} \frac{1}{n} \sum_{i=1}^n \widehat{\theta}_{ij} \varepsilon_i = O_P(n^{-1/2})$. By (1.17) it therefore follows from (1.42) that

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

Using (1.35), it is immediately seen that $\frac{1}{n} \sum_{i=1}^n X_i^{[k]}(\tau_r) \varepsilon_i = O_P(n^{-1/2} n^{(-\mu+1)/2(\mu+2\nu)})$. Consequently,

$$\frac{1}{n} \sum_{i=1}^n X_i^{[k]}(\widehat{\tau}_r) \varepsilon_i = \frac{1}{n} \sum_{i=1}^n X_i^{[k]}(\tau_r) \varepsilon_i + \frac{1}{n} \sum_{i=1}^n (X_i^{[k]}(\widehat{\tau}_r) - X_i^{[k]}(\tau_r)) \varepsilon_i = O_P(n^{-1/2} n^{(-\mu+1)/2(\mu+2\nu)}) \tag{1.43}$$

By Assumption 1.4 c) we can infer from (1.38) 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 $\lambda_{max}((\frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^{[k]} (\mathbf{X}_i^{[k]})^T)^{-1}) = O_P(n^{(\mu-1)/(\mu+2\nu)})$. It therefore follows from (1.39) - (1.43) that

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

This proves (1.21). Using (1.32), it follows that the least squares estimators $\widehat{\alpha}_j$ of $\tilde{\alpha}_j$ are given by

$$\widehat{\alpha}_j = \tilde{\alpha}_j + \widehat{\lambda}_j^{-1} \frac{1}{n} \sum_{i=1}^n \widehat{\theta}_{ij} \varepsilon_i + \sum_{r=1}^S (\beta_r - \widehat{\beta}_r) \widehat{\psi}_j(\tau_r) - \sum_{r=1}^S \widehat{\beta}_r (\widehat{\psi}_j(\widehat{\tau}_r) - \widehat{\psi}_j(\tau_r)), \quad j = 1, \dots, k \tag{1.44}$$

But (1.34) and (1.21) imply that

$$\sum_{j=1}^k (\sum_{r=1}^S (\beta_r - \widehat{\beta}_r) \widehat{\psi}_j(\tau_r))^2 = O_P(k n^{-2\nu/(\mu+2\nu)}) = O_P(n^{-(2\nu-1)/(\mu+2\nu)}), \tag{1.45}$$

while by (1.34) and (1.42)

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

and therefore

$$\sum_{j=1}^{k}(\sum_{r=1}^{S}\widehat{\beta}_r(\widehat{\psi}_j(\widehat{\tau}_r)-\widehat{\psi}_j(\tau_r)))^2 = O_P(n^{-(2\nu-1)/(\mu+2\nu)}). \tag{1.46}$$

Assertion (1.22) now is an immediate consequence of (1.33) and (1.44) - (1.46). □

Supplement to: Functional Linear Regression with Points of Impact

This supplement to the chapter "Functional Linear Regression with Points of Impact" 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 1.6 in Theorem 1.2. Appendix C provides the proofs of Theorem 1.4, Proposition 1.1 as well as the proof of Proposition 1.2.

Appendix A Application to near infrared data

The estimation procedure is applied to a well known near infrared dataset from the Chambersburg Shoot-out 2002. This data consists of a series of NIR spectra of pharmaceutical tablets, which is measured over $p = 650$ equidistant wavelengths ranging from $600 - 1898$ nm. The data can for example be found in the R Package ChemometricsWithRData Wehrens (2011). We focus on the first calibration dataset, consisting of $n = 155$ spectra.

The response variable Y_i is chosen to be the weight of tablet i. All results for the augmented and point of impact model are based on minimizing the BIC-criterion over a fine grid of 132 values of δ between 0.05 and 0.45 in a first step. The number k of principal components for the functional linear regression model was estimated by using a 10-fold cross-validation but conclusions remained unchanged using the BIC. Moreover, results from the augmented as well as from the functional linear model remained stable even if we increased the maximum number of the first principal components initially allowed to enter the model up to at least 10.

The observations corresponding to the frequency of 1820 inhabited an anomaly high variance and has for each curve been replaced by the mean of their closest neighbors. The replaced point was mostly visible as an outlier in the "correlation" plots between Y_i and $Z_{\delta,i}(t)$ but did not change the overall conclusions as given in Table A.1. A typical curve from the Shoot-out dataset is shown in the upper panel of Figure A.1. It is common to NIR data that the observed

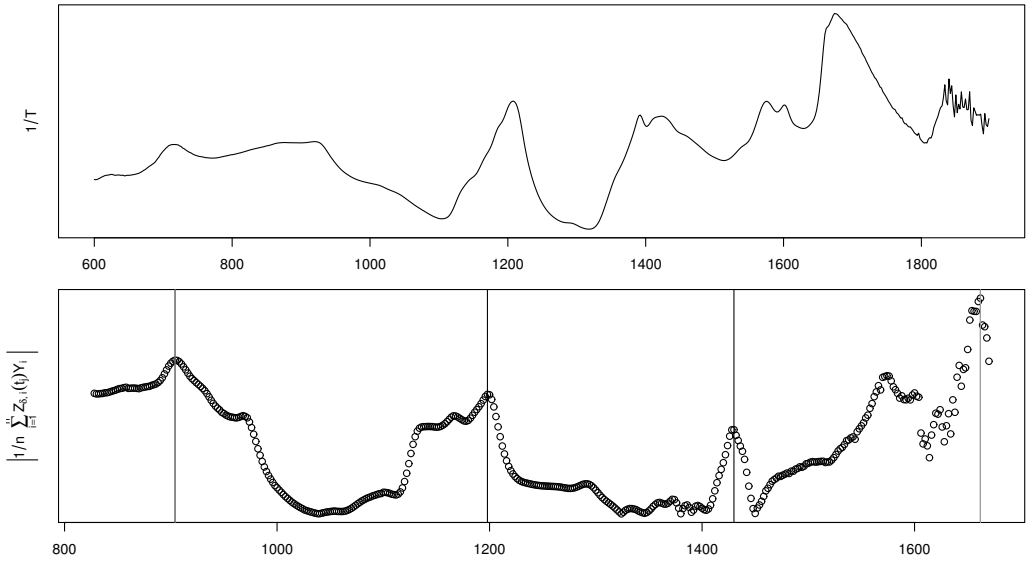

Figure A.1: *While the upper plot shows the trajectory of a typical data curve from the NIR data, the lower part of the figure show* $|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t_j)Y_i|$ *for the optimal value of* δ *as obtained by the best model fit of the augmented model. Selected points of impact candidates are tagged grey, remaining candidates are indicated by black vertical lines.*

curves are essentially smooth but show variability around some wavelengths and variability increases towards the end of the measured wavelengths. Remember that the smoothness of the observed sample paths is reflected by κ. In an idealized setting under Assumption 1 (Kneip et al. 2013) one would expect $\widehat{\kappa}$ to be independent from the particular choice of δ. In practice however covariance functions are often more complex as the smoothness of the observed process can change over the observed sample points. The estimate $\widehat{\kappa} = 1.37$ for κ was obtained for the midpoint of the considered values of δ and supports the impression that the curves in this example exhibit smoother sample paths when compared to the weather data. Figure A.1 also shows the "correlation" between Y_i and $Z_{\delta,i}(t)$ for the optimal value of δ being 0.176. The two points of impact were estimated at wavelengths 1662 and 904. The mean squared prediction error in the third column of Table A.1 was derived from a 10-fold cross-validation after having chosen points of impact and/or principal components by our selection criteria in a first step. The qualitative results derived from the MSPE are supported by the calculated

Table A.1: *Estimated number of principal components k, points of impact S, prediction error and the median of* $(y_i - \hat{y}_i)^2$ *for the NIR data.*

Model	$\widehat{k}$	$\widehat{S}$	**MSPE**	$median((y-\hat{y})^2)$
Augmented	0	2	1.487	0.329
Points of impact	0	2	1.487	0.329
FLR	3	0	1.616	0.428

values of the median value of $(y_i - \hat{y}_i)^2$ for each model given in the last column. In this example, the augmented model was estimated identical to the model consisting solely of points of impact and both models are slightly superior in terms of the prediction error when compared to the traditional functional linear regression model while, at the same time, being easier to be interpreted through their two points of impact.

Appendix B On the approximation properties of the eigenfunctions of a Brownian motion

In this Appendix it is shown that the eigenfunctions of a Brownian motion satisfy assertion (1.6) in Theorem 1.2.

For $t \in (a,b)$ and $\epsilon > 0$ let $\mathscr{C}(t,\epsilon,[a,b])$ 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 1.2. *Let $\psi_1, \psi_2, \ldots$ be a system of orthonormal eigenfunctions corresponding to the non-zero eigenvalues of the covariance operator Γ of X. If for all $t \in (a,b)$ there exists an $\epsilon_t > 0$ such that*

$$\lim_{k\to\infty} \inf_{f\in\mathscr{C}(t,\epsilon,[a,b])} \sup_{s\in[a,b]} |f(s) - \sum_{r=1}^{k} \langle f, \psi_r\rangle \psi_r(s)| = 0 \quad \textit{for every } 0 < \epsilon < \epsilon_t, \tag{1.6}$$

then the process X possesses specific local variation.

We have to show the following statement:

Lemma B.1. *Let $\phi_r(s) = \sqrt{2}\sin((r-\frac{1}{2})\pi s)$, $r = 1,2,\ldots$, be the system of orthonormal eigenfunctions derived from a Brownian motion on $[0,1]$. Then $\phi_r(s)$, $r = 1,2,\ldots$ satisfy assertion (1.6) in Theorem 1.2.*

Proof of Lemma B.1. Assertion (1.6) follows from elementary properties of Fourier series and sine functions. To see this, recall that the basis functions of a Fourier-Sine-Series to approximate a function with periodicity $P = 4$ are given by $\Phi_n(s) = \sin(\frac{1}{2}\pi n s)$ for $n = 1,2,\ldots$. It is now immediately clear, that the eigenfunctions $\phi_r(s) = \sqrt{2}\sin((r-\frac{1}{2})\pi s)$, $r = 1,2,\ldots$, are proportional to the functions of $\Phi_n(s)$, for $n = 2r-1$, i.e. when n is uneven.
Select an arbitrary $0 < t < 1$ and $0 < \epsilon \le \epsilon_t = \min(t, 1-t)$. Let $f_t \in \mathscr{C}(t,\epsilon,[0,1])$ be defined

as $f_t(s) := \max(1-|s-t|/\epsilon, 0)$, $s \in [0,1]$. This function can be easily extended on the whole real line to an odd function F_t with periodicity $P = 4$ by

$$F_t(s) = \begin{cases} f_t(s), & s \in [0,1] \\ f_t(2-s), & s \in [1,2] \\ -F_t(4-s), & s \in [2,4] \\ F_t(s-4m), & s \in [4m, 4(m+1)], m = 1, \dots \\ -F_t(s), & s \leq 0. \end{cases}$$

It is then known that the Fourier series of an odd periodic function reduces to a Fourier-Sine-Series. Moreover, from well known results about Fourier series, we have that the Fourier series of a continuous periodic function F, which is piecewise continuous differentiable, converges uniformly to F.
The only question left is why we can discard the sine basis functions $\Phi_n(s)$ for n even. This point is immediately seen if we look at $F_t(s)$ for $s \in [0,4]$ which is illustrated in Figure B.1 for a specific value of t and ϵ.

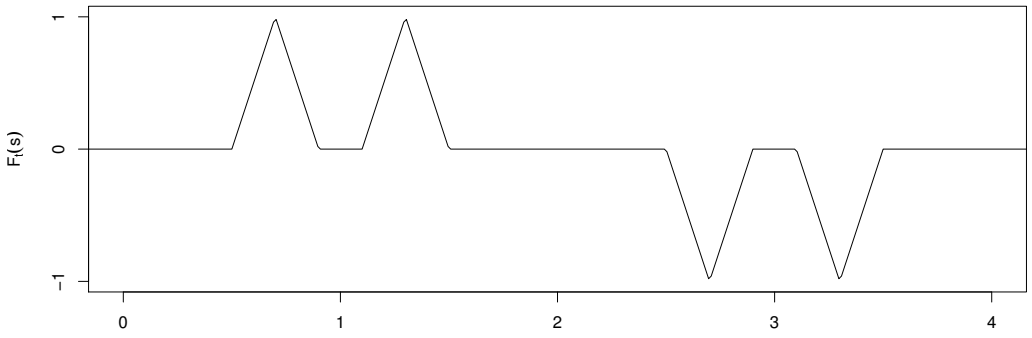

Figure B.1: *The figure shows $F_t(s)$ for $t = 0.7$ and $\epsilon = 0.2$ over the interval $[0,4]$. $F_t(s)$ is an odd function with periodicity 4. The function is point symmetric at $s = 2$ on $[0,4]$.*

First note that all even sine basis functions $\Phi_n(s)$, $n = 2,4,\dots$, have periodicity 2. But by construction $F_t(s)$ is point symmetric at $s = 2$ and hence, for $n = 2,4,6,\dots$, we have $\langle F_t, \Phi_n \rangle = 0$, i.e. the coefficients of all even sine basis functions in the Fourier-Sine-Series are 0. Uniform convergence on $[0,1]$ to $F_t(s)$ follows immediately for the system of the sine basis functions $\Phi_n(s)$ for n even and hence for the system of the eigenfunctions ϕ_r. Since t and ϵ were arbitrary, the system of eigenfunctions satisfy indeed assertion (1.6). □

Appendix C Additional proofs

In this appendix Theorem 1.4 and Propositions 1.1 and 1.2 are proven. We begin by repeating the two main assumptions and Theorem 1.4.

Assumption 1.1. *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 $t,s \in [a,b]$*

$$\sigma(t,s) = \omega(t,s,|t-s|^\kappa). \tag{1.8}$$

Moreover,

$$0 < \inf_{t\in[a,b]} c(t), \quad \text{where } c(t) := -\frac{\partial}{\partial z}\omega(t,t,z)|_{z=0}. \tag{1.9}$$

Assumption 1.2.

a) $X_1,\dots,X_n$ 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 $\varepsilon_1,\dots,\varepsilon_n$ are i.i.d $N(0,\sigma^2)$ r.v. which are independent of X_i.

Theorem 1.4. *Under our setup and Assumptions 1.1 as well as 1.2 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*

$$\max_{r=1,\dots,\widehat{S}} \min_{s=1,\dots,S} |\widehat{\tau}_r - \tau_s| \;=\; O_P(n^{-\frac{1}{k}}). \tag{1.11}$$

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

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

Then

$$P(\widehat{S} = S) \;\to\; 1 \quad \text{as } n \to \infty. \tag{1.12}$$

For the proof of Theorem 1.4 we need an additional proposition and several lemmata.

Proposition C.1. *Consider independent and identically distributed random vectors $\mathbf{V}_i \in \mathbb{R}^p$, $i = 1,\dots,n$, such that $\mathbf{V}_i \sim \mathcal{N}(\mathbf{0},\mathbf{\Sigma})$. Then for all $j,l \in \{1,\dots,p\}$, $j \neq l$ and for any $\epsilon \leq \frac{\mathbb{E}(V_{ij}^2)\mathbb{E}(V_{il}^2)+\mathbb{E}^2(V_{ij}V_{il})}{2(\mathbb{E}(V_{ij}^2)\mathbb{E}(V_{il}^2))^{1/2}}$*

$$P\left(|\frac{1}{n}\sum_{i=1}^n V_{ij}V_{il} - cov(V_{ij},V_{il})| \leq \epsilon\right) \geq 1 - \exp\left(\frac{-3n\epsilon^2}{20\mathbb{E}(V_{ij}^2)\mathbb{E}(V_{il}^2)}\right) \tag{C.1}$$

and for any $j \in \{1,\dots,p\}$ *and* $\epsilon \leq \frac{\mathbb{E}(V_{ij}^2)}{2}$,

$$P\left(|\frac{1}{n}\sum_{i=1}^{n} V_{ij}^2 - \mathbb{E}(V_{ij}^2)| \leq \epsilon\right) \geq 1 - \exp\left(\frac{-3n\epsilon^2}{16(\mathbb{E}(V_{ij}^2))^2}\right). \tag{C.2}$$

Proof of Proposition C.1. The proof follows some lines of the proof of Lemma 2.5 in Zhou et al. (2009). For $j,l \in \{1,\dots,p\}$, $j \neq l$, take $\epsilon \leq \frac{\mathbb{E}(V_{ij}^2)\mathbb{E}(V_{il}^2)+\mathbb{E}^2(V_{ij}V_{il})}{2(\mathbb{E}(V_{ij}^2)\mathbb{E}(V_{il}^2))^{1/2}}$. A direct application of Lemma 38 in Zhou et al. (2010) implies that

$$P\left(|\frac{1}{n}\sum_{i=1}^{n} V_{ij}V_{il} - cov(V_{ij},V_{il})| > \epsilon\right) \leq \exp(-c_{4,j,l}n\epsilon^2),$$

where $c_{4,j,l} = \frac{3}{20\psi_{2,j,l}}$ and $\psi_{2,j,l} = \frac{\mathbb{E}(V_{ij}^2)\mathbb{E}(V_{il}^2)+\mathbb{E}^2(V_{ij}V_{il})}{2}$. Now for all $j,l \in \{1,\dots,p\}$, $j \neq l$ the Cauchy-Schwarz inequality yields

$$\psi_{2,j,l} \;\leq\; \mathbb{E}(V_{ij}^2)\mathbb{E}(V_{il}^2),$$

so that $c_{4,j,l} \geq \frac{3}{20\mathbb{E}(V_{ij}^2)\mathbb{E}(V_{il}^2)}$, $j,l \in \{1,\dots,p\}$, $j \neq l$. Thus

$$P\left(|\frac{1}{n}\sum_{i=1}^{n} V_{ij}V_{il} - cov(V_{ij},V_{il})| > \epsilon\right) \leq \exp\left(\frac{-3n\epsilon^2}{20\mathbb{E}(V_{ij}^2)\mathbb{E}(V_{il}^2)}\right). \tag{C.3}$$

On another side, the large deviation bound (9.3) given in Zhou et al. (2009) implies that for $j \in \{1,\dots,p\}$ and $\epsilon \leq \frac{\mathbb{E}(V_{ij}^2)}{2}$

$$P\left(\frac{1}{n}\sum_{i=1}^{n} V_{ij}^2 - \mathbb{E}(V_{ij}^2) > \epsilon\right) \leq \exp\left(\frac{-3n\epsilon^2}{16(\mathbb{E}(V_{ij}^2))^2}\right).$$

□

Lemma C.1. *Under the assumptions of Theorem 1.4 there exist constants* $0 < D_1 < \infty$ *and* $0 < D_2 < \infty$, *such that for all n, all* $0 < \delta < (b-a)/2$, *all* $t \in [a+\delta, b-\delta]$, *all* $0 < s \leq 1/2$ *with* $\delta^\kappa s^\kappa \geq s\delta^2$, *and every* $0 < z \leq \sqrt{n}$ *we obtain*

$$\begin{aligned} P\Big(\sup_{t-s\delta\leq u\leq t+s\delta} |\frac{1}{n}\sum_{i=1}^{n}[(Z_{\delta,i}(t) - Z_{\delta,i}(u))Y_i - \mathbb{E}((Z_{\delta,i}(t) - Z_{\delta,i}(u))Y_i)]| \leq zD_1\sqrt{\frac{\delta^k s^\kappa}{n}}\Big) \\ \geq 1 - 2\exp(-z^2) \end{aligned} \tag{C.4}$$

and

$$P\Big(\sup_{t-s\delta\leq u\leq t+s\delta}|\frac{1}{n}\sum_{i=1}^{n}[(Z_{\delta,i}(t)^2-Z_{\delta,i}(u)^2)-\mathbb{E}(Z_{\delta,i}(t)^2-Z_{\delta,i}(u)^2)]|\leq zD_2\delta^{\kappa}\sqrt{\frac{s^{\kappa}}{n}}\Big) \geq 1-2\exp(-z^2). \tag{C.5}$$

Proof of Lemma C.1. Choose some arbitrary $0<\delta<(b-a)/2$, $t\in[a+\delta,b-\delta]$, as well as $0<s\leq 1/2$. For $q_1,q_2\in[-1,1]$, Taylor expansions then yield

$$\begin{aligned}
&\mathbb{E}((Z_{\delta,i}(t+q_1s\delta)-Z_{\delta,i}(t+q_2s\delta))^2)=\mathbb{E}((Z_{\delta,i}(t+q_1s\delta)-Z_{\delta,i}(t+q_1s\delta+(q_2-q_1)s\delta))^2)\\
=&c(t+q_1s\delta)s^{\kappa}\delta^{\kappa}2\Big(3/2|q_2-q_1|^{\kappa}-(|q_2-q_1+\frac{1}{s}|^{\kappa}+|q_2-q_1-\frac{1}{s}|^{\kappa}-2\frac{1}{s^{\kappa}})\Big)\\
&+\frac{1}{2}c(t+q_1s\delta)2^{\kappa}s^{\kappa}\delta^{\kappa}\Big(|\frac{q_2-q_1}{2}+\frac{1}{s}|^{\kappa}+|\frac{q_2-q_1}{2}-\frac{1}{s}|^{\kappa}-2\frac{1}{s^{\kappa}}\Big)+R_{5;a,\delta,t}\\
&\text{with } |R_{5;a,\delta,t}|\leq L_{1,1}||q_2-q_1|^{1/2}s^{1/2}\delta|^{\min\{2\kappa,2\}}
\end{aligned} \tag{C.6}$$

for some constant $L_{1,1}<\infty$.
Note that there exists a constant $L_{1,2}<\infty$ such that for all $0<s\leq 1/2$ we have $||q_2-q_1+\frac{1}{s}|^{\kappa}+|q_2-q_1-\frac{1}{s}|^{\kappa}-2\frac{1}{s^{\kappa}}|\leq L_{1,2}|q_2-q_1|^2$ as well as $||\frac{q_2-q_1}{2}+\frac{1}{s}|^{\kappa}+|\frac{q_2-q_1}{2}-\frac{1}{s}|^{\kappa}-2\frac{1}{s^{\kappa}}|\leq L_{0,2}|q_2-q_1|^2$. This in turn implies that there exists a constant $L_{1,3}<\infty$, which can be chosen independent of s and δ, such that for all $q_1,q_2\in[-1,1]$

$$\mathbb{E}((Z_{\delta,i}(t+q_1s\delta)-Z_{\delta,i}(t+q_2s\delta))^2)\leq L_{1,3}s^{\kappa}\delta^{\kappa}|q_1-q_2|^{\min\{1,\kappa\}}. \tag{C.7}$$

Define $Z^*_{\delta,i}(q):=\frac{1}{\sqrt{s^{\kappa}\delta^{\kappa}}}(Z_{\delta,i}(t+qs\delta)Y_i-\mathbb{E}(Z_{\delta,i}(t+qs\delta)Y_i))$ and $Z^*_{\delta}(q):=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Z^*_{\delta,i}(q)$. Under Assumption 1.2 it is easy to show that with $K=4L_{1,3}Var(Y_i)|q_1-q_2|^{\min\{1,\kappa\}}$ we have for all $q_1,q_2\in[-1,1]$ and all integers $m\geq 2$:

$$\mathbb{E}(|Z^*_{\delta,i}(q_1)-Z^*_{\delta,i}(q_2)|^m)\leq\frac{m!}{2}K^{m-2}K^2. \tag{C.8}$$

Now, an application of Corollary 1 of van de Geer and Lederer (2013) guarantees the existence of a constant $0<L_{0,4}<\infty$ such that with $\Psi(x)=\exp(\frac{n}{6}(\sqrt{1+\frac{2\sqrt{6}x}{\sqrt{n}}}-1)^2)-1$, the Orlicz norm of $Z^*_{\delta}(q_1)-Z^*_{\delta}(q_2)$ can be bounded, i.e. we have for all for all $q_1,q_2\in[-1,1]$:

$$\|Z^*_{\delta}(q_1)-Z^*_{\delta}(q_2)\|_{\Psi}\leq L_{1,4}|q_1-q_2|^{\min\{1,\kappa\}}. \tag{C.9}$$

The proof now follows from well known maximal inequalities of empirical process theory. In particular, by (C.9) one may apply theorem 2.2.4 of van der Vaart and Wellner (1996). It is

immediately seen that the covering integral appearing in this theorem is finite, and we can thus infer that there exists a constant $0 < D_{1,1} < \infty$ such that

$$\mathbb{E}\left(\exp\left(\sup_{q_1,q_2\in[-1,1]} n/6\left(\sqrt{1+2\sqrt{\frac{6}{nD_{1,1}^2}|Z_\delta^*(q_1)-Z_\delta^*(q_2)|}-1}\right)^2\right)\right) \leq 2.$$

For every $z > 0$, the Markov inequality then yields

$$\begin{aligned}
&P\Big(\sup_{q_1,q_2\in[-1,1]} |Z_\delta^*(q_1)-Z_\delta^*(q_2)| \geq z\frac{D_{1,1}}{2\sqrt{6}}\Big)\\
&\quad= P\Big(\exp\Big(\sup_{q_1,q_2\in[-1,1]} n/6\Big(\sqrt{1+2\sqrt{\frac{6}{nD_{1,1}^2}|Z_\delta^*(q_1)-Z_\delta^*(q_2)|}-1}\Big)^2\Big)\\
&\qquad\qquad \geq \exp\Big(n/6\Big(\sqrt{1+z/\sqrt{n}}-1\Big)^2\Big)\Big)\\
&\quad\leq 2\exp\Big(-n/6\Big(\sqrt{1+z/\sqrt{n}}-1\Big)^2\Big).
\end{aligned}$$

At the same time it follows from a Taylor expansion that for any $0 < z \leq \sqrt{n}$ there exists a constant $0 < D_{1,2} < \infty$ such that

$$\frac{n}{6}(\sqrt{1+z/\sqrt{n}}-1)^2 \geq D_{1,2}z^2. \tag{C.10}$$

Assertion (C.4) is an immediate consequence.

In order to prove (C.5) first note that $Z_{\delta,i}(t_1)^2 - Z_{\delta,i}(t_2)^2 = (Z_{\delta,i}(t_1)-Z_{\delta,i}(t_2))(Z_{\delta,i}(t_1)+Z_{\delta,i}(t_2))$. Equation (1.29) implies the existence of a constant $0 < L_{1,5} < \infty$ such that $\mathbb{E}((Z_{\delta,i}(t+q_1s\delta)+Z_{\delta,i}(t+q_2s\delta))^2) \leq L_{1,5}\delta^\kappa$ for all $q_1, q_2 \in [-1,1]$, and all n, t, s and δ. With $Z_\delta^{**}(q) = \frac{1}{\sqrt{\delta^{2\kappa}s^\kappa}}\frac{1}{\sqrt{n}}\sum_{i=1}^n (Z_{\delta,i}(t+qs\delta)^2 - \mathbb{E}(Z_{\delta,i}(t+qs\delta)^2))$, similar steps as above now imply the existence of a constant $0 < L_{1,6} < \infty$ such that

$$\|Z_\delta^{**}(q_1)-Z_\delta^{**}(q_2)\|_\Psi \leq L_{1,6}|q_1-q_2|^{\min\{1,\kappa\}}.$$

Using again maximal inequalities of empirical process theory and (C.10), Assertion (C.5) now follows from arguments similar to those used to prove (C.4). □

Lemma C.2. *Under the assumptions of Theorem 1.4 there exist constants $0 < D_3 < D_4 < \infty$ and $0 < D_5 < \infty$ such that*

$$0 < D_3\delta^\kappa \leq \inf_{t\in[a+\delta,b-\delta]} \mathbb{E}(Z_{\delta,i}(t)^2) \leq \sigma_{z,sup}^2 := \sup_{t\in[a+\delta,b-\delta]} \mathbb{E}(Z_{\delta,i}(t)^2) \leq D_4\delta^\kappa \tag{C.11}$$

$$\lim_{n\to\infty} P\Big(\sup_{t\in[a+\delta,b-\delta]} |\frac{1}{n}\sum_{i=1}^n [Z_{\delta,i}(t)^2 - \mathbb{E}(Z_{\delta,i}(t)^2)]| \leq D_5\delta^\kappa\sqrt{\frac{1}{n}\log(\frac{b-a}{\delta})}\Big) = 1. \tag{C.12}$$

Moreover, for any constant A^* with $\sqrt{2} < A^* \leq A$ we obtain as $n \to \infty$

$$P\Bigg(\sup_{t\in[a+\delta,b-\delta]} (\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)^2)^{-\frac{1}{2}} |\frac{1}{n}\sum_{i=1}^{n}(Z_{\delta,i}(t)Y_i - \mathbb{E}(Z_{\delta,i}(t)Y_i))| \leq A^* \sqrt{\frac{var(Y_i)}{n}\log(\frac{b-a}{\delta})}\Bigg) \to 1, \tag{C.13}$$

$$P\Bigg(\sup_{t\in[a+\delta,b-\delta]} |\frac{1}{n}\sum_{i=1}^{n}(Z_{\delta,i}(t)Y_i - \mathbb{E}(Z_{\delta,i}(t)Y_i))| \leq A^* \sqrt{\frac{var(Y_i)D_4\delta^\kappa}{n}\log(\frac{b-a}{\delta})}\Bigg) \to 1. \tag{C.14}$$

Proof of Lemma C.2. Obviously, Assertion C.11 is an immediate consequence of Assumption 1.1 and of equation (1.26).
Let $\mathscr{J}_\delta := \{j|\ t_j \in [a+\delta, b-\delta], j \in \{1,\dots,p\}\}$. Assumption 1.2 implies that all joint distributions of vectors with elements $\{Z_{\delta,i}(t_j)\}_{i=1,\dots,n, j\in\mathscr{J}_\delta}$ or $\{Y_i\}_{i=1,\dots,n}$ are multivariate normal. Choose some constants w_1, w_2 with $1 < w_1 < w_2 < \frac{A^*}{\sqrt{2}}$ and determine an equidistant grid $s_1 = a+\delta < s_2 < \dots < s_{N_{w_1}} = b-\delta$ of $N_{w_1} = [(\frac{b-a}{\delta})^{w_1}]$ points in $[a+\delta, b-\delta]$. Obviously, $\ell_{w_1} := |s_j - s_{j-1}| = O(\delta^{w_1})$, $j = 2,\dots,N_{w_1}$, as $\delta \to 0$. Then

$$\begin{aligned}\sup_{t\in[a+\delta,b-\delta]} |\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)^2 - \mathbb{E}(Z_{\delta,i}(t)^2)| \leq & \sup_{j\in\{2,\dots,N_{w_1}\}} |\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s_j)^2 - \mathbb{E}(Z_{\delta,i}(s_j)^2)| \\ & + \sup_{j\in\{2,\dots,N_{w_1}\}} \sup_{t\in[s_{j-1},s_j]} |\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)^2 - \mathbb{E}(Z_{\delta,i}(t)^2) - (Z_{\delta,i}(s_j)^2 - \mathbb{E}(Z_{\delta,i}(s_j)^2))|.\end{aligned}$$

When using (C.2) as well as $\sup_{t\in[a+\delta,b-\delta]} \mathbb{E}(Z_{\delta,i}(t)^2) \leq D_4\delta^\kappa$ it follows from the Bonferroni-inequality that as $n \to \infty$

$$\begin{aligned}P\Bigg(\sup_{j\in\{2,\dots,N_{w_1}\}} |\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s_j)^2 - \mathbb{E}(Z_{\delta,i}(s_j)^2)| \leq \sqrt{\frac{16}{3}w_2 D_4\delta^\kappa}\sqrt{\frac{1}{n}\log(\frac{b-a}{\delta})}\Bigg) \\ \geq 1 - N_{w_1}\cdot\exp\Bigg(-w_2\log(\frac{b-a}{\delta})\Bigg) \geq 1 - (\frac{b-a}{\delta})^{w_1-w_2} \to 1,\end{aligned}$$

while Lemma C.1 implies that as $n \to \infty$

$$P\Bigg(\sup_{j\in\{2,\dots,N_{w_1}\}} \sup_{t\in[s_{j-1},s_j]} |\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)^2 - \mathbb{E}(Z_{\delta,i}(t)^2) - (Z_{\delta,i}(s_j)^2 - \mathbb{E}(Z_{\delta,i}(s_j)^2))| \leq D_2\sqrt{w_2}\delta^\kappa\sqrt{\frac{\ell_{w_1}^\kappa}{\delta^\kappa n}\log(\frac{b-a}{\delta})}\Bigg) \to 1.$$

Recall that $\frac{\ell_{w_1}^{\kappa}}{\delta^{\kappa}} = O(\delta^{\kappa(w_1-1)})$ and hence $\sqrt{\frac{\ell_{w_1}^{\kappa}}{\delta^{\kappa} n}\log(\frac{b-a}{\delta})} = o(\sqrt{\frac{1}{n}\log(\frac{b-a}{\delta})})$. When combining the above arguments we thus obtain (C.12).
Consider (C.13) and note that

$$(\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)^2)^{\frac{1}{2}} - (\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s)^2)^{\frac{1}{2}} = \frac{\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)^2 - \frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s)^2}{(\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)^2)^{\frac{1}{2}} + (\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s)^2)^{\frac{1}{2}}}.$$

Some straightforward computations lead to

$$\begin{aligned}
&\sup_{t\in[a+\delta,b-\delta]} (\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)^2)^{-\frac{1}{2}} |\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)Y_i - \mathbb{E}(Z_{\delta,i}(t)Y_i)| \\
&\leq \sup_{j\in\{2,\dots,N_{w_1}\}} (\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s_j)^2)^{-\frac{1}{2}} |\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s_j)Y_i - \mathbb{E}(Z_{\delta,i}(s_j)Y_i)| \\
&+ \sup_{j\in\{2,\dots,N_{w_1}\}} \left(\sup_{t\in[s_{j-1},s_j]} \frac{|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)^2 - \frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s_j)^2|}{2\inf_{u\in[s_{j-1},s_j]} \frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(u)^2} \frac{|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s_j)Y_i - \mathbb{E}(Z_{\delta,i}(s_j)Y_i)|}{(\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s_j)^2)^{\frac{1}{2}}} \right) \\
&+ \sup_{j\in\{2,\dots,N_{w_1}\}} \sup_{t\in[s_{j-1},s_j]} \frac{|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)Y_i - \mathbb{E}(Z_{\delta,i}(t)Y_i) - (Z_{\delta,i}(s_j)Y_i - \mathbb{E}(Z_{\delta,i}(s_j)Y_i))|}{\inf_{u\in[s_{j-1},s_j]}(\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(u)^2)^{\frac{1}{2}}}, \qquad \text{(C.15)}
\end{aligned}$$

It follows from (C.11) and (C.12) that there exists a constant $0 < L_{2,1} < \infty$ such that $P(\inf_{u\in[a+\delta,b-\delta]} \frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(u)^2 \geq L_{2,1}\delta^{\kappa}) \to 1$ as $n \to \infty$. Furthermore,

$$\begin{aligned}
|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)^2 - \frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s_j)^2| &\leq |\mathbb{E}(Z_{\delta,i}(t)^2) - \mathbb{E}(Z_{\delta,i}(s_j)^2)| \\
&+ |\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)^2 - \frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s_j)^2 - (\mathbb{E}(Z_{\delta,i}(t)^2) - \mathbb{E}(Z_{\delta,i}(s_j)^2)|,
\end{aligned}$$

and it follows from (C.7) and (C.11) that there is a constant $0 < L_{2,2} < \infty$ such that for every $j \in \{2,\dots,N_{w_1}\}$

$$|\mathbb{E}(Z_{\delta,i}(t)^2) - \mathbb{E}(Z_{\delta,i}(s_j)^2)| \leq \sqrt{\mathbb{E}((Z_{\delta,i}(t) - Z_{\delta,i}(s_j))^2)\mathbb{E}((Z_{\delta,i}(t) + Z_{\delta,i}(s_j))^2)} \leq L_{2,2}\delta^{\kappa}\sqrt{\frac{\ell_{w_1}^{\kappa}}{\delta^{\kappa}}}$$

holds for all sufficiently large n. We can therefore infer from Lemma C.1 that for some constants $0 < L_{2,3} < \infty$, $0 < L_{2,4} < \infty$

$$P\left(\sup_{j\in\{2,\dots,N_{w_1}\}} \sup_{t\in[s_{j-1},s_j]} \frac{|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)^2 - \frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s_j)^2|}{2\inf_{u\in[s_{j-1},s_j]} \frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(u)^2} \leq L_{2,3}\sqrt{\frac{\ell_{w_1}^{\kappa}}{\delta^{\kappa}}} \right) \to 1, \qquad \text{(C.16)}$$

$$P\Bigg(\sup_{j\in\{2,\dots,N_{w_1}\}}\sup_{t\in[s_{j-1},s_j]}\frac{|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(t)Y_i-\mathbb{E}(Z_{\delta,i}(t)Y_i)-(Z_{\delta,i}(s_j)Y_i-\mathbb{E}(Z_{\delta,i}(s_j)Y_i))|}{\inf_{u\in[s_{j-1},s_j]}(\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(u)^2)^{\frac{1}{2}}} \\ \leq L_{2,4}\sqrt{\frac{\ell_{w_1}^\kappa}{\delta^\kappa n}\log(\frac{b-a}{\delta})}\Bigg)\to 1, \tag{C.17}$$

as $n\to\infty$.

For any $t\in[a+\delta,b-\delta]$ let $\gamma_\delta(t):=\frac{(\int_a^b\beta(t)\mathbb{E}(Z_{\delta,i}(t)X_i(s))ds+\sum_{r=1}^S\beta_r\mathbb{E}(Z_{\delta,i}(t)X_i(\tau_r)))}{\mathbb{E}Z_{\delta,i}(t)^2}$ be the coefficient of a linear regression from Y_i on $Z_{\delta,i}(t)$. There obviously is a constant $0<L_{2,5}<\infty$ such that $\sup_{t\in[a+\delta,b-\delta]}|\gamma_\delta(t)|\leq L_{2,5}$ for all sufficiently small $\delta>0$. With $e_{\delta,i}(t):=Y_i-Z_{\delta,i}(t)\gamma_\delta(t)$ we then obtain

$$Y_i=Z_{\delta,i}(t)\gamma_\delta(t)+e_{\delta,i}(t),\ \mathbb{E}(e_{\delta,i}(t))=0,\ var(e_{\delta,i}(t))\leq var(Y_i),\ \mathbb{E}(Z_{\delta,i}(t)e_{\delta,i}(t))=0,$$

and

$$\sup_{j\in\{2,\dots,N_{w_1}\}}(\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)^2)^{-\frac{1}{2}}|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)Y_i-\mathbb{E}(Z_{\delta,i}(s_j)Y_i)| \\ \leq \sup_{j\in\{2,\dots,N_{w_1}\}}(\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)^2)^{-\frac{1}{2}}|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)^2-\mathbb{E}(Z_{\delta,i}(s_j)^2)||\gamma_\delta(s_j)| \\ +\sup_{j\in\{2,\dots,N_{w_1}\}}(\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)^2)^{-\frac{1}{2}}|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)e_{\delta,i}(s_j)|, \tag{C.18}$$

while (C.11) and (C.12) imply that for some constant $0<L_{2,6}<\infty$

$$P\Bigg(\sup_{j\in\{2,\dots,N_{w_1}\}}(\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)^2)^{-\frac{1}{2}}|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)^2-\mathbb{E}(Z_{\delta,i}(s_j)^2)||\gamma_\delta(s_j)| \\ \leq L_{1,6}\sqrt{\frac{\delta^\kappa}{n}\log(\frac{b-a}{\delta})}\Bigg)\to 1, \tag{C.19}$$

as $n\to\infty$.

The joint distribution of $(Z_{\delta,i}(s_j),e_{\delta,i}(s_j))$ is multivariate normal, and hence uncorrelatedness of $Z_{\delta,i}(s_j)$ and $e_{\delta,i}(s_j)$ even implies independence. Consequently, for any realization of $Z_{\delta,i}(s_j)\neq 0$ the conditional distribution of $V_{\delta,i}(s_j):=(\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)^2)^{-\frac{1}{2}}$
$\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)e_{\delta,i}(s_j)$ given $Z_{\delta,i}(s_j)$ is equal to $N(0,\frac{var(e_{\delta,i}(s_j))}{n})$. Thus also the marginal distribution of $V_{\delta,i}(s_j)$ is equal to $N(0,\frac{var(e_{\delta,i}(s_j))}{n})$, and a well-known elementary bound on the tails of a normal distribution yields $P(|V_{\delta,i}(s_j)|\geq v\sqrt{\frac{var(e_{\delta,i}(s_j))}{n}})$
$\leq\exp(-\frac{v^2}{2})$ for all $v>0$. Therefore,

$$P\Bigg(\sup_{j\in\{2,\dots,N_{w_1}\}} (\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s_j)^2)^{-\frac{1}{2}} |\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s_j)e_{\delta,i}(s_j)| \le \sqrt{2}w_2\sqrt{\frac{var(Y_i)}{n}\log(\frac{b-a}{\delta})}$$
$$\ge 1 - N_{w_1}\cdot\exp\Big(-w_2\log(\frac{b-a}{\delta})\Big) \ge 1-(\frac{b-a}{\delta})^{w_1-w_2} \to 1, \tag{C.20}$$

as $n\to\infty$. Since $\sqrt{2}w_2 < A^*$, assertion (C.13) is an immediate consequence of (C.15) - (C.20). Finally, (C.14) follows from (C.11), (C.12) and (C.13). □

Lemma C.3. *Under the assumptions of Theorem 1.4 there exists a constant $0 < M_{sup} < \infty$ such that for all n, all $0 < \delta < (b-a)/2$ and every $t \in [a+\delta, b-\delta]$ we obtain*

$$\left|\mathbb{E}\left(Z_{\delta,i}(t)\int_a^b \beta(s)X_i(s)ds\right)\right| \le M_{sup}\delta^{\min\{2,\kappa+1\}}. \tag{C.21}$$

Proof of Lemma C.3. We have

$$\left|\mathbb{E}\left(Z_{\delta,i}(t)\int_a^b \beta(s)X_i(s)ds\right)\right| \le |\int_a^{t-\delta} \beta(s)\mathbb{E}(Z_{\delta,i}(t)X_i(s))ds| + |\int_{t-\delta}^{t+\delta} \beta(s)\mathbb{E}(Z_{\delta,i}(t)X_i(s))ds|$$
$$+ |\int_{t+\delta}^{b} \beta(s)\mathbb{E}(Z_{\delta,i}(t)X_i(s))ds|.$$

Since by assumption $|\beta(s)|$ is bounded by some constant $D<\infty$, assertions (1.25) and (1.28) imply that for some constant $M_{1,sup}<\infty$ we have $|\int_{t-\delta}^{t+\delta} \beta(s)\mathbb{E}(Z_{\delta,i}(t)X_i(s))ds| \le M_{1,sup}\delta^{\kappa+1}$. If $\kappa = 1$, then (C.21) is therefore an immediate consequence of assertion (1.29). If $\kappa \neq 1$ it follows from (1.29) that there exists a constant $M_{2,sup}<\infty$ such that

$$|\int_a^{t-\delta} \beta(s)\mathbb{E}(Z_{\delta,i}(t)X_i(s))ds| \le \delta^2|\int_\delta^{t-a} DM_5 s^{\kappa-2}ds| \le M_{2,sup}\delta^{\min\{2,\kappa+1\}}.$$

A similar bound can obviously be derived for $|\int_{t+\delta}^{b} \beta(s)\mathbb{E}(Z_{\delta,i}(t)X_i(s))ds|$. This proves the lemma. □

Lemma C.4. *Under the assumptions of Theorem 1.4 let $I_r := \{t\in[a,b]|\ |t-\tau_r| \le \min_{s\neq r}|t-\tau_s|\}$, $r=1,\dots,S$.*

If $S>0$, there then exist constants $0<Q_1<\infty$ and $0<Q_2<\infty$ such that for all sufficiently small $\delta>0$ and all $r=1,\dots,S$ we have

$$|\mathbb{E}(Z_{\delta,i}(t)Y_i)| \le Q_1\frac{\delta^2}{\max\{\delta,|t-\tau_r|\}^{2-\kappa}} + M_{sup}\delta^{\min\{2,\kappa+1\}} \quad \textit{for every } t\in I_r, \tag{C.22}$$

as well as

$$\sup_{t\in I_r,\ |t-\tau_r|\geq\frac{\delta}{2}} |\mathbb{E}(Z_{\delta,i}(t)Y_i)| \leq (1-Q_2)|\beta_r|c(\tau_r)\delta^\kappa, \tag{C.23}$$

and for any $u\in[-0.5,0.5]$

$$\begin{aligned}&|\mathbb{E}(Z_{\delta,i}(\tau_r)Y_i)-\mathbb{E}(Z_{\delta,i}(\tau_r+u\delta)Y_i)|\\ &\quad=|-\beta_r c(\tau_r)\delta^\kappa\left(|u|^\kappa-\frac{1}{2}(|u+1|^\kappa-1)-\frac{1}{2}(|u-1|^\kappa-1)\right)+R_{5;r}(u)|,\end{aligned} \tag{C.24}$$

where $|R_{5;r}(u)|\leq\widetilde{M}_r||u|^{1/2}\delta|^{\min\{2\kappa,2\}}$ *for some constants* $\widetilde{M}_r<\infty$, $r=1,\ldots,S$.

Proof of Lemma C.4. Our setup implies that for all $t\in[a+\delta,b-\delta]$

$$\mathbb{E}(Z_{\delta,i}(t)Y_i)=\int_a^b\beta(s)\mathbb{E}(Z_{\delta,i}(t)X_i(s))ds+\sum_{r=1}^S\beta_r\mathbb{E}(Z_{\delta,i}(t)X_i(\tau_r)). \tag{C.25}$$

Since $\tau_1,\ldots,\tau_S\in(a,b)$ are fixed, we have $\tau_r\in[a+\delta,b-\delta]$, $r=1,\ldots,S$, as well as $\delta\ll\frac{1}{2}\min_{r\neq s}|\tau_r-\tau_s|$ for all sufficiently small $\delta>0$. Using (C.25), assertions (C.22) and (C.23) are thus immediate consequences of (1.28) and (1.29). and Lemma C.3.
In order to prove (C.24) first note that similar to (1.27) and (1.29) straightforward Taylor expansions can be used to show that there exists a constant $L_{4,1}<\infty$ such that for all $t\in[a+\delta,b-\delta]$

$$\begin{aligned}&|\mathbb{E}\left(Z_{\delta,i}(\tau_r+u\delta)X(t)-Z_{\delta,i}(\tau_r)X(t)\right)|=\left|\omega(\tau_r+u\delta,t,|\tau_r-t+u\delta|^\kappa)-\omega(\tau_r,t,|\tau_r-t|^\kappa)\right.\\ &\quad-\left(\frac{1}{2}\omega(\tau_r+(u+1)\delta,t,|\tau_r-t+(u+1)\delta|^\kappa)-\frac{1}{2}\omega(\tau_r+\delta,t,|\tau_r-t+\delta|^\kappa)\right)\\ &\quad\left.-\left(\frac{1}{2}\omega(\tau_r+(u-1)\delta,t,|\tau_r-t+(u-1)\delta|^\kappa)-\frac{1}{2}\omega(\tau_r-\delta,t,|\tau_r-t-\delta|^\kappa)\right)\right|\\ &\quad\leq L_{4,1}\left(\frac{|u|\delta^2}{\max\{|u|\delta,|t-\tau_r|\}^{2-\kappa}}+||u|^{1/2}\delta|^{\min\{2\kappa,2\}}\right).\end{aligned}$$

Generalizing the arguments used to prove Lemma C.3 we thus obtain
$|\int_a^b\beta(t)(\mathbb{E}(Z_{\delta,i}(\tau_r)X_i(t)-Z_{\delta,i}(\tau_r+u\delta)X_i(t))dt|\leq L_{4,2}||u|^{1/2}\delta|^{\min\{2\kappa,2\}}$ for some constant $L_{4,2}<\infty$. Furthermore, $|\mathbb{E}(Z_{\delta,i}(\tau_r)X_i(\tau_s)-Z_{\delta,i}(\tau_r+u\delta)X_i(\tau_s))|\leq L_{4,3}||u|^{1/2}\delta|^{\min\{2\kappa,2\}}$ for some $L_{4,3}<\infty$ and all $r,s\in\{1,\ldots,S\}$, $r\neq s$. Assertion (C.24) then follows from equation (1.27). □

Proof of Theorem 1.4. Let $\lambda_n = A\sqrt{\frac{Var(Y_i)}{n}\log(\frac{b-a}{\delta})}$ and let $I_\delta := \{t_j |\ t_j \in [a+\delta, b-\delta], j \in \{1,\dots,p\}\}$. For any $t \in I_\delta$ we obviously have

$$\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)Y_i = \mathbb{E}(Z_{\delta,i}(t)Y_i) + \frac{1}{n}\sum_{i=1}^{n}(Z_{\delta,i}(t)Y_i - \mathbb{E}(Z_{\delta,i}(t)Y_i)). \tag{C.26}$$

First consider the case that there are no points of impact, i.e. $S = 0$. Then by Lemma C.3 we have $|\mathbb{E}(Z_{\delta,i}(t)Y_i)| = \left|\mathbb{E}\left(Z_{\delta,i}(t)\int_a^b \beta(s)X_i(s)ds\right)\right| \leq M_{sup}\delta^{\min\{2,\kappa+1\}}$. Since by assumption $\delta^{\min\{2,\kappa+1\}} = o(\sqrt{\frac{\delta^\kappa}{n}})$, Lemma C.2 implies that
$P\left(\sup_{t\in[a+\delta,b-\delta]} \frac{|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(t)Y_i|}{(\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(t)^2)^{1/2}} \leq \lambda_n\right) \to 1$, and hence $P(\widehat{S} = S) \to 1$, as $n \to \infty$.
Now consider the case that $S \geq 1$. Select some arbitrary $\alpha > 2$. As $n \to \infty$ we have $\delta \equiv \delta_n \to 0$. Therefore, $\tau_r \in [a+\delta, b-\delta]$, $r = 1,\dots,S$, as well as $\sqrt{\delta}/\alpha < \frac{1}{2}\min_{r\neq s}|\tau_r - \tau_s|$, provided that n is sufficiently large. Let $I_{r,\delta,\alpha} := \{t \in I_\delta \mid |t-\tau_r| \leq \sqrt{\delta}/\alpha\}$, $r = 1,\dots,S$, as well as $I_{\delta,\alpha} = \bigcup_{r=1}^S I_{r,\delta,\alpha}$ and $I_{\delta,\alpha}^C := I_\delta \backslash I_{\delta,\alpha}$.
By our assumptions on the sequence $\delta \equiv \delta_n$ we can infer from (C.26), (C.22), and (C.14) that there exist constants $0 < C_1 < \infty$ and $0 < C_2 < \infty$ such that the event

$$\sup_{t\in I_{\delta,\alpha}^C} |\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)Y_i| \leq C_1\sqrt{\frac{\delta^\kappa}{n}|\log\delta|} + C_2\alpha^{2-\kappa}\delta^{1+\kappa/2} \tag{C.27}$$

holds with probability tending to 1 as $n \to \infty$. Since by assumption $\frac{|\log\delta|}{n\delta^\kappa} \to 0$ and hence $\sqrt{\frac{\delta^\kappa}{n}|\log\delta|} = o(\delta^\kappa)$, (C.23) and (C.14) imply the existence of a constant $0 < C_3 < Q_2$ such that

$$\sup_{t\in I_{r,\delta,\alpha}, |t-\tau_r|\geq\delta/2} |\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)Y_i| \leq (1-C_3)|\beta_r|c(\tau_r)\delta^\kappa, \quad r = 1,\dots,S \tag{C.28}$$

hold with probability tending to 1 as $n \to \infty$.
For $r = 1,\dots,S$ let $j(r)$ be an index satisfying $|\tau_r - t_{j(r)}| = \min_{j=1,\dots,p}|\tau_r - t_j|$. Obviously $|\tau_r - t_{j(r)}| \leq \frac{b-a}{2(p-1)}$ and by (8.5) in Kneip et al. (2013) our conditions on $p \equiv p_n$ imply that there exists a constant $0 < C_4 < \infty$ such that

$$|\mathbb{E}(Z_{\delta,i}(t_{j(r)})X_i(\tau_r)) - c(\tau_r)\delta^\kappa| \leq C_4 n^{-1/\kappa},\ r = 1,\dots,S.$$

Using again (C.14) together with $\sqrt{\frac{\delta^\kappa}{n}|\log\delta|} = o(\delta^\kappa)$, we can thus conclude that there exists a sequence $\{\epsilon_n\}$ of positive numbers with $\lim_{n\to\infty}\epsilon_n \to 0$ such that

$$|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t_{j(r)})Y_i| \geq (1-\epsilon_n)|\beta_r|c(\tau_r)\delta^\kappa, \quad r = 1,\dots,S \tag{C.29}$$

holds with probability tending to 1 as $n \to \infty$. Now define

$$\widetilde{\tau}_r := arg \max_{t \in I_\delta:\ |t-\tau_r| \le \delta/2} |\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t) Y_i|. \tag{C.30}$$

Since $\delta^{1+\kappa/2} = o(\delta^\kappa)$ and since $\alpha > 2$, one can infer from (C.27) - (C.29) that the following assertions hold with probability tending to 1 as $n \to \infty$:

$$\widetilde{\tau}_r = arg \max_{t \in I_{r,\delta,\alpha}} |\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t) Y_i| = arg \max_{t \in I_{r,\delta,\alpha} \cup I^C_{\delta,\alpha}} |\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t) Y_i|, \quad r = 1, \dots, S, \tag{C.31}$$

as well as

$$I_{r,\delta,\alpha} \subset [\widetilde{\tau}_r - \sqrt{\delta}/2, \widetilde{\tau}_r + \sqrt{\delta}/2] \qquad r = 1, \dots, S. \tag{C.32}$$

But under (C.31) and (C.32) construction of the estimators $\widehat{\tau}_k$, $k = 1, \dots, S$, for the first S steps of our estimation procedure implies that $\{\widehat{\tau}_1, \dots, \widehat{\tau}_S\} = \{\widetilde{\tau}_1, \dots, \widetilde{\tau}_S\}$. Therefore,

$$P(\{\widehat{\tau}_1, \dots, \widehat{\tau}_S\} = \{\widetilde{\tau}_1, \dots, \widetilde{\tau}_S\}) \to 1 \tag{C.33}$$

$$P\left(I_\delta \setminus \bigcup_{r=1}^{S} [\widehat{\tau}_r - \sqrt{\delta}/2, \widehat{\tau}_r + \sqrt{\delta}/2] \subset I^C_{\delta,\alpha} \right) \to 1 \tag{C.34}$$

as $n \to \infty$.

By definition of $\widetilde{\tau}_r$, $r = 1, \dots, S$, in (C.30) it already follows from (C.33) that $\widehat{\tau}_1, \dots, \widehat{\tau}_S$ provide consistent estimators of the true points of impact. Some more precise approximations are, however, required to show Assertion (1.11).

Note that for all $r = 1, \dots, S$ and all $t \in (a, b)$

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t) Y_i =& \mathbb{E}(Z_{\delta,i}(t) Y_i) \\ &+ \frac{1}{n}\sum_{i=1}^{n} [(Z_{\delta,i}(t) - Z_{\delta,i}(\tau_r)) Y_i - \mathbb{E}((Z_{\delta,i}(t) - Z_{\delta,i}(\tau_r)) Y_i)] \\ &+ \frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(\tau_r) Y_i - \mathbb{E}(Z_{\delta,i}(\tau_r) Y_i). \end{aligned} \tag{C.35}$$

Recall that $|\tau_r - t_{j(r)}| \le \frac{b-a}{2(p-1)} = O(n^{-1/\kappa})$ and let $M_p := arg \max\{m \in \mathbb{N} |\ \frac{\delta}{2^m} \ge 2n^{-1/\kappa}\}$.

Our assumptions on the sequence $\delta \equiv \delta_n$ yield $\sup_{m=1,\dots,M_p} \frac{|2^{-m/2}\delta|^{\min\{2\kappa,2\}}}{2^{-\kappa m}\delta^\kappa} \to 0$. We can therefore infer from (C.24) that there are constants $0 < C_5 < C_6 < \infty$ such that for all $m = 1, 2, \dots, M_p$ and all sufficiently large n

$$\begin{aligned} \sup_{t\in I_{r,\delta,\alpha},\ |t-\tau_r|\geq \frac{\delta}{2^{m+1}}} |\mathbb{E}(Z_{\delta,i}(t)Y_i)| &\leq |\beta_r| c(\tau_r)\left(\delta^\kappa - C_6\frac{\delta^\kappa}{2^{\kappa m}}\right) \\ |\mathbb{E}(Z_{\delta,i}(t_{j(r)})Y_i)| &> |\beta_r| c(\tau_r)\left(\delta^\kappa - C_5\frac{\delta^\kappa}{2^{\kappa m}}\right) \end{aligned} \tag{C.36}$$

hold for every $r = 1, \dots, S$. On the other hand, the exponential inequality (C.4) obviously implies the existence of a constant $0 < C_7 < \infty$ such that for any $0 < q < \sqrt{n}$

$$P\left(\sup_{|t-\tau_r|\leq \delta/2^m} |\frac{1}{n}\sum_{i=1}^n [(Z_{\delta,i}(t) - Z_{\delta,i}(\tau_r))Y_i - \mathbb{E}((Z_{\delta,i}(t) - Z_{\delta,i}(\tau_r))Y_i)]| \leq C_7 q\sqrt{\frac{\delta^\kappa}{2^{\kappa m}n}}\right) \geq 1 - \frac{1}{q}, \tag{C.37}$$

holds for all $m = 1, 2, \dots$ and each $r = 1, \dots, S$.

For all $m = 1, 2, \dots$ and $r = 1, \dots, S$ let $\mathscr{A}(n, m, r)$ denote the event that

$$\sup_{|t-\tau_r|\leq \delta/2^m} |\frac{1}{n}\sum_{i=1}^n [(Z_{\delta,i}(\tau_r) - Z_{\delta,i}(t))Y_i - \mathbb{E}((Z_{\delta,i}(\tau_r) - Z_{\delta,i}(t))Y_i)]| < (C_6 - C_5)|\beta_r| c(\tau_r)\frac{\delta^\kappa}{2^{\kappa m}}.$$

Inequality (C.37) implies that with $C_8 := \frac{C_7}{(C_6-C_5)\min_{r=1,\dots,S}|\beta_r| c(\tau_r)}$ the complementary events $\mathscr{A}(n, m, r)^C$ can be bounded by

$$P\left(\mathscr{A}(n, m, r)^C\right) \leq C_8\sqrt{\frac{2^{\kappa m}}{\delta^\kappa n}} \tag{C.38}$$

for all $m = 1, 2, \dots$ and $r = 1, \dots, S$. If $m \leq M_p$, then (C.35) and (C.36) imply that under $\mathscr{A}(n, m, r)$ we have

$$\widetilde{\tau}_{r,m} := arg \sup_{t\in I_{r,\delta,\alpha}, |t-\tau_r|\leq \delta/2^m} |\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(t)Y_i| \ \in [\tau_r - \frac{\delta}{2^{m+1}}, \tau_r + \frac{\delta}{2^{m+1}}] \tag{C.39}$$

for each $r = 1, \dots, S$ and all sufficiently large n.

Choose an arbitrary $\epsilon > 0$ and set

$$m^*(\epsilon) := \min\left\{m = 1, 2, \dots |\ \epsilon \geq C_8\sqrt{\frac{2^{\kappa m}}{\delta^\kappa n}}\right\}$$

whenever there exists an integer $m > 0$ such that $\epsilon \geq C_8\sqrt{\frac{2^{\kappa m}}{\delta^\kappa n}}$ and set $m^*(\epsilon) := 1$ otherwise. Furthermore define

$$m(\epsilon) := \min\{m^*(\epsilon), M_p\}.$$

By our assumptions on $\delta \equiv \delta_n$ there then obviously exists a constant $A(\epsilon) < \infty$ such that for all sufficiently large n,

$$\frac{\delta}{2^{m(\epsilon)}} \leq A(\epsilon) n^{-1/\kappa}. \tag{C.40}$$

Now consider the event $\mathscr{A}(n,\epsilon) := \bigcup_{r=1}^{S} \bigcup_{m=1}^{m(\epsilon)} \mathscr{A}(n,m,r)$. By (C.38) the Bonferroni inequality implies that

$$P(\mathscr{A}(n,\epsilon)) \geq 1 - S \sum_{m=1}^{m(\epsilon)} C_8 \sqrt{\frac{2^{\kappa m}}{\delta^{\kappa} n}} \geq 1 - S \sum_{m=0}^{m(\epsilon)-1} (\frac{1}{2^{\kappa/2}})^{m(\epsilon)-m} \epsilon \geq 1 - \frac{S\epsilon}{1-(\frac{1}{2})^{\kappa/2}}. \tag{C.41}$$

But under event $\mathscr{A}(n,\epsilon)$ we can infer from (C.39) that

$$\widetilde{\tau}_{r,1} = \widetilde{\tau}_{r,2} = \cdots = \widetilde{\tau}_{r,m(\epsilon)}. \tag{C.42}$$

Additionally let $\mathscr{A}^*(n)$ denote the event that $\{\widehat{\tau}_1, \ldots, \widehat{\tau}_S\} = \{\widetilde{\tau}_1, \ldots, \widetilde{\tau}_S\}$. The definitions in (C.30) and (C.39) yield $\widetilde{\tau}_{1,s} = \widetilde{\tau}_s$, $s = 1, \ldots, S$, and we can thus conclude from (C.40) and (C.42) that under events $\mathscr{A}^*(n)$ and $\mathscr{A}(n,\epsilon)$ we have

$$\max_{r=1,\ldots,S} \min_{s=1,\ldots,S} |\widehat{\tau}_r - \tau_s| = \max_{r=1,\ldots,S} \min_{s=1,\ldots,S} |\widetilde{\tau}_{r,m(\epsilon)} - \tau_s| \leq \frac{\delta}{2^{m(\epsilon)+1}} \leq \frac{A(\epsilon)}{2} n^{-1/\kappa} \tag{C.43}$$

for all n sufficiently large.

Recall that $P(\mathscr{A}^*(n)) \to 1$ as $n \to \infty$. Since ϵ is arbitrary, (1.11) thus follows from (C.41) and (C.43).

It remains to prove Assertion (1.12). For some $\sqrt{2} < A^* < A$ define $\lambda_n^* < \lambda_n$ by $\lambda_n^* := A^* \sqrt{\frac{var(Y_i)}{n} \log(\frac{b-a}{\delta})}$. By (C.13) it is immediately seen that in addition to (C.27) also

$$\sup_{t \in I_{\delta,\alpha}^C} \frac{|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(t) Y_i|}{\sqrt{\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(t)^2}} \leq \lambda_n^* + C_2 \alpha^{2-\kappa} \frac{\delta^{1+\kappa/2}}{\inf_{t \in I_{\delta,\alpha}^C} \sqrt{\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(t)^2}}$$

holds with probability tending to 1 as $n \to \infty$. But (C.12) and our assumptions on the sequence $\delta \equiv \delta_n$ lead to $\frac{\delta^{1+\kappa/2}}{\inf_{t \in I_{\delta,\alpha}^C} \sqrt{\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(t)^2}} = o_P(\lambda_n^*)$. Using (C.34), the construction of the estimator $\widehat{\tau}_{S+1}$ therefore implies that as $n \to \infty$,

$$P\left(\frac{|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(\widehat{\tau}_{S+1}) Y_i|}{\sqrt{\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(\widehat{\tau}_{S+1})^2}} < \lambda_n \right) = P\left(\sup_{t \in I_{\delta,\alpha}^C} \frac{|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(t) Y_i|}{\sqrt{\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(t)^2}} \leq \lambda_n \right) \to 1,$$

while (C.29) together with (C.33), (C.11) and (C.12) yield

$$P\left(\min_{r=1,\dots,S}\frac{|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(\widehat{\tau}_r)Y_i|}{\sqrt{\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(\widehat{\tau}_r)^2}}>\lambda_n\right)\to 1.$$

By definition of our estimator $\widehat{S}$, (1.12) is an immediate consequence. □

Proposition 1.1. *Under the conditions of Theorem 1.4 we have*

$$\widehat{\kappa}=\kappa+O_P(n^{-1/2}+\delta^{\min\{2,2/\kappa\}}). \tag{1.15}$$

Proof of Proposition 1.1. Define $p_{k_\delta}:=p-2k_\delta$. By Proposition C.1 and arguments similar to those leading to (C.12) we obtain

$$\frac{1}{np_{k_\delta}}\sum_{i=1}^n\sum_{j\in\mathscr{J}_{0,\delta}}Z_{\delta^*,i}(t_j)^2=\frac{1}{p_{k_\delta}}\sum_{j\in\mathscr{J}_{0,\delta}}\mathbb{E}(Z_{\delta^*,i}(t_j)^2)+O_P(\frac{\delta^k}{\sqrt{n}}),\quad \text{for } \delta^*\in\{\delta,\frac{\delta}{2}\}$$

On the other hand, with $C_9:=\frac{1}{p_{k_\delta}}\sum_{j\in J_{0,\delta}}(2c(t_j)-\frac{2^\kappa}{2}c(t_j))>0$, (1.26) leads to

$$\frac{1}{p_{k_\delta}}\sum_{j\in\mathscr{J}_{0,\delta}}\mathbb{E}(Z_{\delta,i}(t_j)^2)=\delta^\kappa C_9+O_P(\delta^{\min\{2\kappa,2\}}),\ \frac{1}{p_{k_\delta}}\sum_{j\in\mathscr{J}_{0,\delta}}\mathbb{E}(Z_{\frac{\delta}{2},i}(t_j)^2)=\frac{\delta^\kappa}{2^\kappa}C_9+O_P(\delta^{\min\{2\kappa,2\}}).$$

When combining these results, (1.15) follows from elementary Taylor expansions. □

Proposition 1.2. *Under the assumptions of Theorem 1.4 we obtain for every $r=1,\dots,S$*

$$\frac{1}{n}\sum_{i=1}^n(X_i(\tau_r)-X_i(\widehat{\tau}_r))^2=O_p(n^{-1}), \tag{1.16}$$

$$\frac{1}{n}\sum_{i=1}^n(X_i(\tau_r)-X_i(\widehat{\tau}_r))\varepsilon_i=O_p(n^{-1}). \tag{1.17}$$

Proof of Proposition 1.2. Let $r\in\{1,\dots,S\}$. Choose some $t\in[a,b]$, some $q\in[-2,2]$ and some $s>0$ with $[t-2s,t+2s]\subset[a,b]$, and recall that $\sigma(t,s)=\sigma(s,t)$. Using Taylor

expansions, one can infer from Assumption 1.1 that there exist constants $0 < C_{10} < \infty$, $0 < C_{11} < \infty$ such that for all $s, t \in [a, b]$

$$\begin{aligned}&\mathbb{E}\big((X_i(t)-X_i(t+qs))^2\big)\\&\leq |\omega(t,t,0)-\omega(t,t+qs,0)-\omega(t+qs,t,0)+\omega(t+qs,t+qs,0)| + C_{10}|qs|^\kappa\\&\leq C_{11}|qs|^2 + C_{10}|qs|^\kappa.\end{aligned}$$

Therefore, for any sufficiently small $s > 0$ and all $q_1, q_2 \in [-1, 1]$ we have

$$\mathbb{E}\big((X_i(\tau_r + q_1 s) - X_i(\tau_r + q_2 s))^2\big) \leq (C_{10} + C_{11})|q_1 - q_2|^2 s^\kappa. \tag{C.44}$$

Obviously, $(X_i(\tau_r)-X_i(t_1))^2-(X_i(\tau_r)-X_i(t_2))^2 = (X_i(t_2))-X_i(t_1))(2X_i(\tau_r)-X_i(t_1)-X_i(t_2))$ and there exists a constant $0 < C_{12} < \infty$ such that $\mathbb{E}[(2X_i(\tau_r)-X_i(\tau_r+q_1 s)-X_i(\tau_r+q_2 s))^2] \leq C_{12}s^\kappa$ for all $q_1, q_2 \in [-1, 1]$, all n and all sufficiently small $s > 0$. When applying inequality (C.1) with $V_{ij} := X_i(t + q_1 s) - X_i(t + q_2 s)$ and $V_{il} := 2X_i(\tau_r) - X_i(t + q_1 s\delta) - X_i(t + q_2 s\delta)$, arguments similar to those used in the proof of Lemma C.1 show that maximal inequalities of empirical process theory can be used to bound the supremum of $\frac{1}{n}\sum_{i=1}^n (X_i(\tau_r)-X_i(t+q_1 s))^2 - \mathbb{E}((X_i(\tau_r)-X_i(\tau_r+q_1 s))^2) - \frac{1}{n}\sum_{i=1}^n (X_i(\tau_r)-X_i(t+q_1 s))^2 + \mathbb{E}((X_i(\tau_r)-X_i(\tau_r+q_2 s))^2)$ over $q_1, q_2 \in [-1, 1]$. Together with (C.44) we then arrive at

$$P\bigg(\sup_{\tau_r - s \leq u \leq \tau_r + s} \frac{1}{n}\sum_{i=1}^n (X_i(\tau_r) - X_i(u))^2 \leq s^\kappa (C_{13} + z D_3 n^{-1/2})\bigg) \geq 1 - 2\exp(-z^2) \tag{C.45}$$

for some constant $0 < D_3 < \infty$, $C_{13} := 2^2(C_{10} + C_{11})$, all n and all sufficiently small $s > 0$. Assertion (1.16) now is a straightforward consequence of (C.45) and (1.11).

Finally, by our assumptions on ε_i and (C.44) another application of maximal inequalities of empirical process theory leads to

$$P\bigg(\sup_{\tau_r - s \leq u \leq \tau_r + s} \frac{1}{n}\sum_{i=1}^n (X_i(\tau_r) - X_i(u))\varepsilon_i \leq z D_4 s^{\kappa/2} n^{-1/2})\bigg) \geq 1 - 2\exp(-z^2)$$

for some constant $0 < D_4 < \infty$, all n and all sufficiently small $s > 0$. Together with (1.16) we then arrive at (1.17). □

Chapter 2

Points of Impact in Generalized Linear Models with Functional Predictors

We introduce a generalized linear regression model with functional predictors. The predictor trajectories are evaluated at a finite set of unknown points of impact, which are treated as additional model parameters that need to be estimated from the data. We propose a threshold-based and a fully data-driven estimator, establish the identifiability of our model, derive the convergence rates of our point of impact estimators, and develop the asymptotic normality of the linear model parameter estimators. The finite sample properties of our estimators are assessed by means of a simulation study. Our methodology is motivated by a psychological case study, where participants were asked to continuously rate their emotional state while watching an affective video on the persecution of African albinos.

2.1 Introduction

In this paper it is assumed that an unknown number S of values $X(\tau_1), X(\tau_2), \ldots, X(\tau_S)$ of a functional random variable $X = \{X(t) : t \in [a,b] \subset \mathbb{R}\}$ are linked to a scalar valued dependent variable Y via

$$\mathbb{E}(Y|X) = g\left(\alpha + \sum_{r=1}^{S} \beta_r X(\tau_r)\right), \tag{2.1}$$

where $S \in \mathbb{N}$, $\tau_1, \tau_2, \ldots, \tau_S \in (a,b)$ as well as $\alpha, \beta_1, \beta_2, \ldots, \beta_S \in \mathbb{R}$ are unknown and need to be estimated from the data. The values $\tau_1, \tau_2, \ldots, \tau_S$ are called points of impact and give specific locations at which the functional regressor X influences the scalar outcome Y.

For estimating the points of impact τ_r and their number S, knowledge of g is not required. Estimation of the parameters α and β_r relies on quasi maximum likelihood estimation and, therefore, requires knowledge of g. Our statistical theory allows for a large family of mean functions g including the practically relevant case of a logistic regression model with points of impact where Y is binary and $g(x) = \exp(x)/(1+\exp(x))$.

Lindquist and McKeague (2009) convincingly demonstrate the importance of a logistic regression model with a single ($S = 1$) point of impact τ_1 by analyzing a genetic data set, where they aim to determine a single genetic locus that allows to distinguish between breast cancer patients and patients without breast cancer. They derive the limiting distribution of their estimate $\widehat{\tau}_1$ under the assumption that $X(\tau_1 + t) - X(\tau_1)$ follows a two-sided Brownian motion. A point of impact model, where $S = 1$ is assumed known, has also been studied in survival analysis for the Cox-Regression (Zhang, 2012).

The case where $g(x) = x$ is the identity function is considered in several works. McKeague and Sen (2010) consider a functional linear regression model with a single ($S = 1$) point of impact and derived the distribution of their estimates in the case where X is a fractional Brownian motion. Kneip et al. (2016a) consider also a functional linear regression model with points of impact, but allow for an unknown parameter $S \geq 1$. Aneiros and Vieu (2014) consider a points of impact model, but postulate the existence of a consistent estimation procedure, which is a non-trivial requirement for our more general case with $g(x) \neq x$. Berrendero et al. (2017) consider a general Reproducing Kernel Hilbert Space framework for the case $g(x) = x$.

Selecting sparse features from functional data X is also found useful in the literature on prediction models. For instance, Ferraty et al. (2010) aim to extract most predictive design points. Floriello and Vitelli (2017) propose a method for sparse clustering of functional data. Park et al. (2016) focus on selecting predictive subdomains of the functional data. These works, however, do not focus on parameter estimation which is of central interest in our work. Readers with a general interest in functional data analysis are referred to the textbooks of Ramsay and Silverman (2005), Ferraty and Vieu (2006), Horváth and Kokoszka (2012), Hsing and Eubank (2015), Kokoszka and Matthew (2017) and the overview article of Wang et al. (2016).

Our methodology is motivated by a case study from psychology, where participants were asked to continuously rate their emotional state (from very negative to very positive) during watching a documentary video on the persecution of African albinos. After the video, the participants were asked to rate their overall emotional state. Psychologists are interested in understanding how overall ratings of emotional states after such an emotion inducing video is related to fluctuations of emotional states during watching the video, as this has implications for the way emotional states are assessed in research using such material. Figure 2.1 shows the continuously self-reported emotion trajectories $X_1(t), \ldots, X_n(t)$, where t denotes standardized time $0 \leq t \leq 1$.

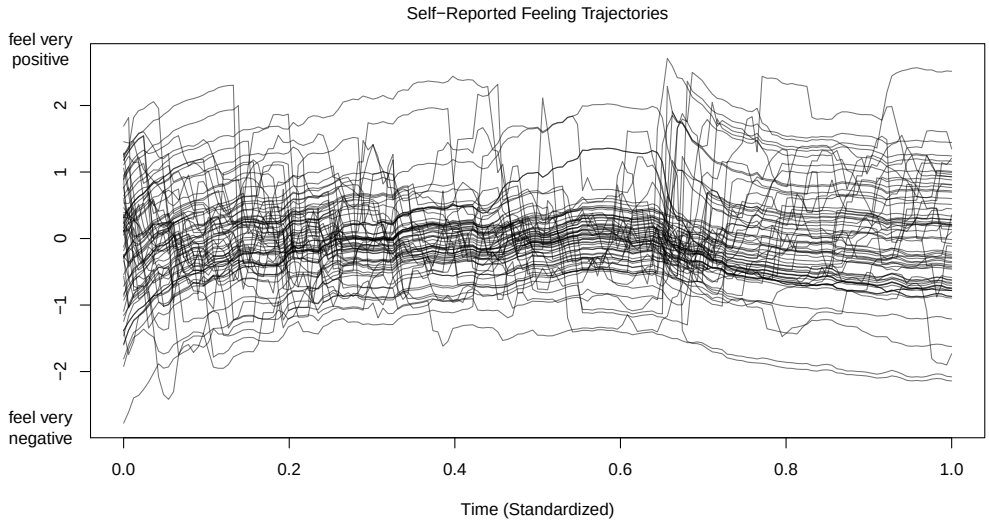

Figure 2.1: *Continuously self-reported emotion trajectories of $n = 67$ participants.*

The remainder of this work is structured as follows. Section 2.2 considers the estimation of the points of impact τ_r and their number S. The estimation of the slope coefficients β_r is described in Section 2.3. Section 2.4 proposes a practical data-driven implementation of the estimation procedure. Our simulation study is contained in Section 2.5 and Section 2.6 contains our real data application. All proofs and additional simulation results can be found in the supplementary appendices supporting this article.

2.2 Determining points of impact

In this section we present the theoretical framework for estimating the points of impact $\tau_1, \ldots, \tau_S$. We also give a more intuitive description of the general idea of the estimation process.

Suppose we are given an i.i.d. sample of data (X_i, Y_i), $i = 1, \ldots, n$, where $X_i = \{X_i(t), t \in [a, b]\}$ is a stochastic process with $\mathbb{E}(\int_a^b X_i(t)^2\, dt) < \infty$, $[a, b]$ is a compact subset of $\mathbb{R}$ and Y_i a real valued random variable. It is assumed that the relationship between Y_i and the functional regressor X_i can be modeled as

$$Y_i = g\left(\alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r)\right) + \varepsilon_i, \tag{2.2}$$

in which the error term ε_i respects $\mathbb{E}(\varepsilon_i | X_i(t)) = 0$ for all $t \in [a, b]$. The parameters S, $\tau_1, \ldots, \tau_S$ as well as $\alpha, \beta_1, \ldots, \beta_S$ are unknown and have to be estimated from the data. The inclusion of a constant α allows us to consider centered random functions X_i with $\mathbb{E}(X_i(t)) = 0$ for all $t \in [a, b]$. The specific locations $\tau_1, \ldots, \tau_S$ are called "points of impact" and indicate

the locations at which the corresponding functional values $X_i(\tau_1), \ldots, X_i(\tau_S)$ have a specific influence on Y_i. Denoting the linear regression function as

$$\eta_i = \alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r) \tag{2.3}$$

allows us to write $\mathbb{E}(Y_i|X_i) = g(\eta_i)$.

In our application, we are primarily interested with the logistic regression framework with points of impact where Y_i is a binary variable and $g(\eta_i) = \exp(\eta_i)/(1+\exp(\eta_i))$. It is important to note, however, that our main theoretical results on estimating S and the points of impact $\tau_1, \ldots, \tau_S$ are valid under much more general assumptions on g. Indeed, the functional form of g does not need to be known and has to fulfill only very mild regularity conditions.

Estimating points of impact τ_r necessarily depends on the structure of X_i. Motivated by our application we consider stochastic processes with rough sample paths such as (fractional) Brownian motions, Ornstein-Uhlenbeck processes, Poisson processes etc. These processes have also important applications in many fields such as finance, chemometrics, econometrics, or the analysis of gene expression data (Lee and Ready, 1991; Levina et al., 2007; Dagsvik and Strøm, 2006; Rohlfs et al., 2013). Common to these processes are covariance functions $\sigma(s,t) = \mathbb{E}(X_i(s)X_i(t))$ which are two times continuously differentiable for all points $s \neq t$, but not two times differentiable at the diagonal $s = t$. The following assumption describes a very large class of such stochastic processes and allows us to derive precise theoretical results:

Assumption 2.1. *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 $s, t \in [a,b]$*

$$\sigma(s,t) = \omega(s,t,|s-t|^\kappa). \tag{2.4}$$

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

The parameter κ quantifies the degree of smoothness of the covariance function σ at the diagonal. While a twice continuously differentiable covariance function will satisfy (2.4) with $\kappa = 2$, a very small value of κ will indicate a process with non-smooth sample paths. See Kneip et al. (2016a) for an estimator of κ which is applicable under our assumptions.

Assumption 2.1 covers important classes of stochastic processes. Recall, for instance, that the class of self similar (not necessarily centered) processes $X_i = \{X_i(t) : t \geq 0\}$ has the property that $X_i(c_1 t) = c_1^H X_i(t)$ for any constant $0 < c_1$ and some exponent $0 < H$. It is then

well known that the covariance function of any such process X_i with stationary increments and $0 < \mathbb{E}(X_i(1)^2) < \infty$ satisfies

$$\sigma(s,t) = \omega(s,t,|s-t|^{2H}) = (s^{2H} + t^{2H} - |s-t|^{2H})\, c_2$$

for some constant $0 < c_2$; see Theorem 1.2 in Embrechts and Maejima (2000). If $0 < H < 1$ such a process respects Assumption 2.1 with $\kappa = 2H$ and $c(t) = c_2$. A prominent example of a self similar process is the fractional Brownian motion.

Another class of processes is given when $X_i = \{X_i(t) : t \geq 0\}$ is a second order process with stationary and independent increments. In this case it is easy to show that $\sigma(s,t) = \omega(s,t,|s-t|) = (s + t - |s-t|)\, c_3$ for some constant $0 < c_3$. The assumption then holds with $\kappa = 1$ and $c(t) = c_3$. The latter conditions on X_i are, for instance, satisfied by second order Lèvy processes which include important processes such as Poisson processes, compound Poisson processes, or jump-diffusion processes.

A third important class of processes satisfying Assumption 2.1 are processes with a Matérn covariance function. For this class of processes the covariance function depends only on the distance between s and t through

$$\sigma(s,t) = \omega_\nu(s,t,|s-t|) = \frac{\pi\phi}{2^{\nu-1}\Gamma(\nu+1/2)\alpha^{2\nu}}(\alpha|s-t|)^\nu K_\nu\big(\alpha|s-t|\big),$$

where K_ν is the modified Bessel function of the second kind, and ρ, ν and α are non-negative parameters of the covariance. It is known that this covariance function is $2m$ times differentiable if and only if $\nu > m$ (cf. M. L. Stein, 1999, Ch. 2.7, p. 32). Our assumption is then satisfied for $\nu < 1$. For the special case where $\nu = 0.5$ one may derive the long term covariance function of an Ornstein-Uhlenbeck process which is given as $\sigma(s,t) = \omega(s,t,|s-t|) = 0.5\exp(-\theta|s-t|)\sigma_{OU}^2$, for some parameter $\theta > 0$ and $\sigma_{OU} > 0$. Assumption 2.1 is then covered with $\kappa = 1$ and $c(t) = 0.5\,\sigma_{OU}^2$.

The intention of our estimator for the points of impact τ_r is to exploit the covariance structure of the processes described by Assumption 2.1. Note that covariance functions $\sigma(s,t)$ satisfying this assumption are obviously not two times differentiable at the diagonal $s = t$, but two times differentiable for $s \neq t$. The following lemma is important for our later results by relating $\mathbb{E}(X_i(s)Y_i)$ to the covariance between $X_i(s)$ and $\eta_i = \alpha + \sum_{r=1}^S \beta_r X_i(\tau_r)$.

Lemma 2.1. *Let $\widetilde{\eta}_i = \eta_i - \alpha$ and let $g : \mathbb{R} \to \mathbb{R}$ be a function with $\mathbb{E}(|g(\eta_i)|) < \infty$, $\mathbb{E}(|\widetilde{\eta}_i g(\eta_i)|) < \infty$, $\mathbb{E}(\widetilde{\eta}_i g(\eta_i)) \neq 0$, $0 < \mathbb{V}(\eta_i) < \infty$, and $\mathrm{Cov}\big(X_i(s) - \widetilde{\eta}_i \mathbb{E}(X_i(s)\widetilde{\eta}_i)/\mathbb{V}(\widetilde{\eta}_i), g(\eta_i)\big) = 0$ for all $s \in [a,b]$. There then exists a constant $c_0 \neq 0$ which is independent of s, such that*

$$\mathbb{E}(X_i(s)Y_i) = c_0 \cdot \mathbb{E}(X_i(s)\eta_i) \quad \textit{for all} \quad s \in [a,b].$$

The only crucial assumption is $\mathrm{Cov}\big(X_i(s) - \widetilde{\eta}_i \mathbb{E}(X_i(s)\widetilde{\eta}_i)/\mathbb{V}(\widetilde{\eta}_i), g(\eta_i)\big) = 0$ for all $s \in [a,b]$. This assumption is, for instance, fulfilled for Gaussian processes X_i where the residuals $X_i(s) - \widetilde{\eta}_i \mathbb{E}(X_i(s)\widetilde{\eta}_i)/\mathbb{V}(\widetilde{\eta}_i)$ are independent from $g(\eta_i)$. Moreover, if X_i is a Gaussian process it follows from Stein's Lemma (C. M. Stein, 1981) that $c_0 = \mathbb{E}(g'(\eta_i))$ provided that g is differentiable and $\mathbb{E}(|g'(\eta_i)|) < \infty$. See also Brillinger (2012b) and Brillinger (2012a) for related results.

Under Assumption 2.1 and Lemma 2.1, the locations of the points of impact are uniquely identifiable from $\mathbb{E}(X_i(s)Y_i)$. Let us make this more precise by defining

$$f_{XY}(s) := \mathbb{E}(X_i(s)Y_i) = c_0\,\mathbb{E}(X_i(s)\eta_i) = c_0 \sum_{r=1}^{S} \beta_r \sigma(s,\tau_r).$$

Since $\sigma(s,t)$ is not two times differentiable at $s=t$, $f(s) = \mathbb{E}(X_i(s)Y_i)$ will not be two times differentiable at $s=\tau_r$, for all $r=1,\dots,S$, resulting in kink-like features at τ_r as depicted in the upper plot of Figure 2.2.

A natural strategy to estimate τ_r is to detect these kinks by considering the following modified central difference approximation of the second derivative of f at a point $s \in [a-\delta, b-\delta]$ for some $\delta > 0$:

$$f_{XY}(s) - \frac{1}{2}(f_{XY}(s+\delta) + f_{XY}(s-\delta)) \approx -\frac{1}{2}\delta^2 f''_{XY}(s). \tag{2.5}$$

Since $f''_{XY}(s)$ does not exist at $s=\tau_r$, the left hand side of (2.5) will tendentiously decline much slower to zero as $\delta \to 0$ for $|s-\tau_r| \approx 0$ than for s with $|s-\tau_r| \gg \delta$.

By defining

$$Z_{\delta,i}(s) := X_i(s) - \frac{1}{2}(X_i(s-\delta), +X_i(s+\delta)) \quad \text{for} \quad s \in [a+\delta, b-\delta]$$

we have that $\mathbb{E}(Z_{\delta,i}(s)Y_i) = f_{XY}(s) - (f_{XY}(s+\delta) + f_{XY}(s-\delta))/2$. The above discussion then suggests estimating the points of impact τ_r using the local extrema of $\mathbb{E}(Z_{\delta,i}(s)Y_i)$. Indeed, it follows by exactly the same arguments as in Kneip et al. (2016a) together with Lemma 2.1 that under Assumption 2.1 one obtains the following theoretical result justifying such an estimation strategy:

$$\mathbb{E}(Z_{\delta,i}(s)Y_i) = \begin{cases} c_0\beta_r c(\tau_r)\delta^{\kappa} + O(\max\{\delta^{2\kappa}, \delta^2\}) & \text{if } |s-\tau_r| \approx 0, \\ O(\max\{\delta^{\kappa+1}, \delta^2\}) & \text{if } \min\limits_{r=1,\dots,S} |s-\tau_r| \gg \delta. \end{cases}$$

Of course, $\mathbb{E}(Z_{\delta,i}(s)Y_i)$ is not known and we have to rely on $n^{-1}\sum_{i=1}^{n} Z_{\delta,i}(s)Y_i$ as its estimate. Under our setting we will have $\mathbb{V}(Z_{\delta,i}(s)Y_i) = O(\delta^\kappa)$, implying

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

As a consequence, as $n \to \infty$, δ can not be chosen to go arbitrarily fast to 0 otherwise the effect of the estimation noise will overlay the signal. This situation is depicted in the bottom plot of Figure 2.2.

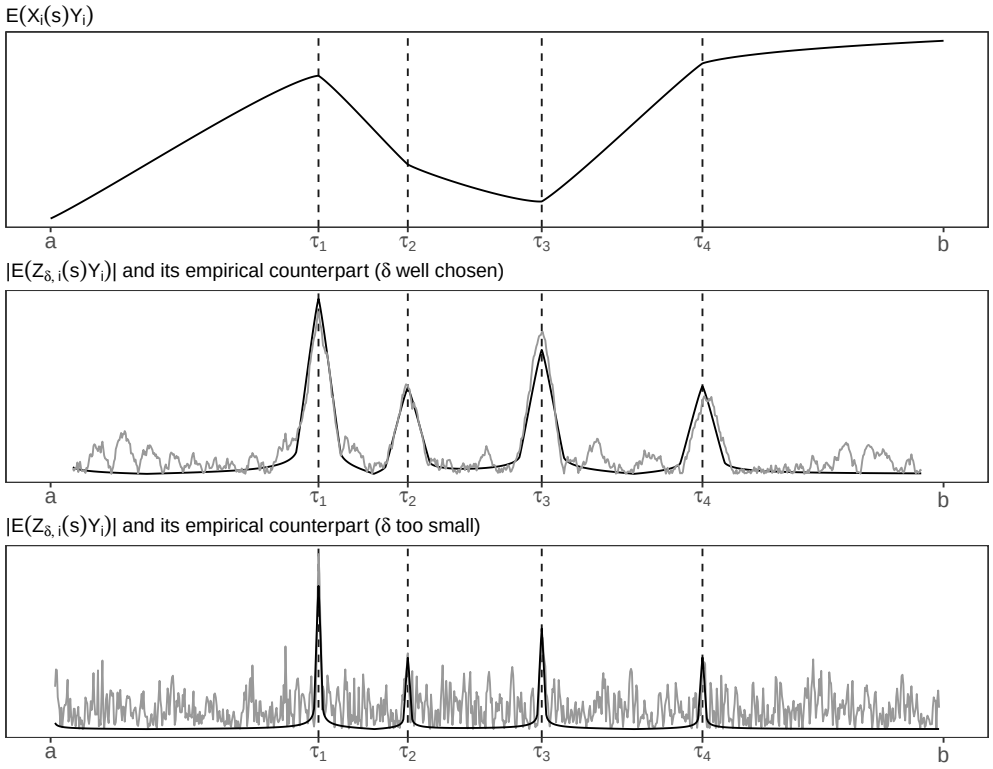

Figure 2.2: *The upper panel shows $\mathbb{E}(X_i(s)Y_i)$ as a function of s with 4 kink-like features at the points of impact (dashed vertical lines). The lower panels show $|\mathbb{E}(Z_{\delta,i}(s)Y_i)|$ (black) and their empirical counterparts (gray) for different values of δ.*

Remark Even if the covariance function $\sigma(s,t)$ does not satisfy Assumption 2.1, the points of impact τ_r may still be estimated using the local extrema of $\mathbb{E}(Z_{\delta,i}(s)Y_i)$. Suppose, for instance, that for some $m \geq 2$ there exists a m times differentiable function $\widetilde{\sigma} : \mathbb{R} \to \mathbb{R}$ such that $\sigma(s,t) = \widetilde{\sigma}(|s-t|)$. Moreover assume that $\widetilde{\sigma}(|s-t|)$ decays fast enough, as $|s-t|$ increases, such that $X_i(s)$ is essentially uncorrelated with $X_i(\tau_r)$ for $|\tau_r - s| \gg 0$. If $|\widetilde{\sigma}''(0)| > |\widetilde{\sigma}''(x)|$ and $\min_{r\neq k}|\tau_r - \tau_k|$ is large enough, then all points of impact might be identified as local extrema of $\mathbb{E}(Z_{\delta,i}(s)Y_i)$.

2.2.1 Estimation

In the following we consider the case where each X_i has been observed over p equidistant points $t_j = a + (j-1)(b-a)/(p-1)$, $j = 1,\dots,p$, where p may be much larger than n. Estimators for the points of impact are determined by sufficiently large local maxima of $\left|n^{-1}\sum_{i=1}^n Z_{\delta,i}(t_j)Y_i\right|$. This strategy is similar to Kneip et al. (2016a), however, in contrast to Kneip et al. (2016a), we avoid a direct computation of $Z_{\delta,i}(t_j)$ for every t_j and propose the following much more efficient estimation procedure:

Estimating points of impact:

1. **Calculate:**
$$\widehat{f}_{XY}(t_j) := \frac{1}{n}\sum_{i=1}^n X_i(t_j)Y_i, \qquad j = 1,\dots,p$$

2. **Choose:** $\delta > 0$ such that there exists some $k_\delta \in \mathbb{N}$ with $1 \le k_\delta < (p-1)/2$ and $\delta = k_\delta(b-a)/(p-1)$.

3. **Calculate:** For all $j \in \mathscr{J}_\delta := \{k_\delta + 1,\dots,p-k_\delta\}$
$$\widehat{f}_{ZY}(t_j) := \widehat{f}_{XY}(t_j) - \frac{1}{2}(\widehat{f}_{XY}(t_j-\delta) + \widehat{f}_{XY}(t_j+\delta))$$

4. **Repeat:**

   Initiate the repetition by setting $l = 1$.

   Estimate a point of impact candidate as
$$\widehat{\tau}_l = \underset{t_j : j\in\mathscr{J}_\delta}{\arg\max} |\widehat{f}_{ZY}(t_j)|.$$

   Update $\mathscr{J}_\delta$ by eliminating all points in $\mathscr{J}_\delta$ in an interval of size $\sqrt{\delta}$ around $\widehat{\tau}_l$.
   Set $l = l+1$.

   End repetition if $\mathscr{J}_\delta = \emptyset$.

The procedure will result in estimates $\widehat{\tau}_1, \widehat{\tau}_2, \dots, \widehat{\tau}_{M_\delta}$, where $M_\delta < \infty$ denotes the maximum number of repetitions. Finally, one then may estimate S as

$$\widehat{S} = \min\left\{ l \in \mathbb{N}_0 : \left| \frac{\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(\widehat{\tau}_{l+1})Y_i}{(\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(\widehat{\tau}_{l+1})^2)^{1/2}} \right| < \lambda \right\}$$

for some $\lambda > 0$.

2.2.2 Asymptotic results

For deriving our asymptotic results we need some further assumptions. We consider asymptotics as $n \to \infty$ with $p \equiv p_n \geq Ln^{1/\kappa}$ for some constant $0 < L < \infty$. Furthermore, we introduce the following assumption:

Assumption 2.2.

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

b) *There exists a $0 < \sigma_{|y|} < \infty$ such that for each $m = 1, 2, \ldots$ we have $E(|Y_i|^{2m}) \leq 2^{m-1} m! \sigma_{|y|}^{2m}$.*

The moment condition in b) is obviously fulfilled for bounded Y_i, for instance, in the functional logistic regression we have that $E(|Y_i|^m) \leq 1$ for all $m = 1, 2, \ldots$. Note that, condition b) holds for any centered sub-Gaussian Y_i, where a centering of Y_i can always be achieved by substituting $g(\eta_i) + \mathbb{E}(g(\eta_i))$ for $g(\eta_i)$ in model 2.2. If X_i satisfies condition a) of Assumption 2.2, then condition b) in particular holds if the errors ε_i are sub-Gaussian and g is assumed to have a bounded derivative.

The following result shows consistency of our estimators for the points of impact $\widehat{\tau}_r$ and the estimator of the total number of points of impact $\widehat{S}$:

Theorem 2.1. *Under our setup, Assumptions 2.1, 2.2, and the assumptions of Lemma 2.1, 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*

$$\max_{r=1,\ldots,\widehat{S}} \min_{s=1,\ldots,S} |\widehat{\tau}_r - \tau_s| \;=\; O_P(n^{-\frac{1}{k}}). \tag{2.6}$$

Moreover, there exists a constant $0 < D < \infty$ such that when the algorithm is applied with cut-off parameter

$$\lambda \equiv \lambda_n = A\sqrt{\frac{\sigma_{|y|}^2}{n} \log\left(\frac{b-a}{\delta}\right)}, \quad \text{where } A > D,$$

and $\delta^2 = O(n^{-1})$, then

$$P(\widehat{S} = S) \;\to\; 1 \quad \text{as } n \to \infty. \tag{2.7}$$

Theorem 2.1 is qualitatively the same as Theorem 4 in Kneip et al. (2016a), but differs in choice of the constant of the cut-off parameter λ. The constant D is derived from asymptotic considerations. A cut-off λ which performed well in our simulations is given by $\lambda = A\sqrt{\sqrt{\mathbb{E}(Y_i^4)} \log(\frac{b-a}{\delta})/n}$, where $A = \sqrt{2\sqrt{3}}$ and the unknown $\mathbb{E}(Y_i^4)$ is estimated by $\widehat{\mathbb{E}}(Y_i^4) = n^{-1} \sum_{i=1}^n Y_i^4$ in practice. Its value is motivated by an argument using the central limit theorem in the derivations of the cut off for Theorem 2.1. See the remark after the proof of Lemma B.2 in Appendix B for some additional information.

Remarks on Theorem 2.1: (i) The proof of Theorem 2.1 applies even to more general linear predictors η_i of the form $\eta_i = \beta_0 + \sum_{r=1}^{S} \beta_r X_i(\tau_r) + \int_a^b \beta(t) X_i(t)\, dt$, where $\beta(t)$ is a bounded and square integrable function over $[a, b]$. In this case $\int_a^b \beta(t) X_i(t)\, dt$ can be understood as a common effect of the whole trajectory X_i on Y_i.
(ii) The results of the theorem rely on Lemma 2.1. Note that for this lemma to hold the specific form of the function g does not need to be known nor does the lemma demands any smoothness assumptions on g. As a result Theorem 2.1 holds for any g satisfying Lemma 2.1.
(iii) Furthermore, Assumption 2.1 gives only a sufficient condition for estimating points of impact. The main argument for the estimation procedure of the points of impact is the property that $\sigma(s, t)$ is less smooth at the diagonal than for $|t - s| > 0$ while the actual degree of smoothness is negligible. For example, suppose $\sigma(s, t)$ is $d > 2$ times continuously differentiable for $s \neq t$ and not being d times differentiable at $s = t$ one may then look at the central difference approximation of at least the d-th derivative of $\mathbb{E}(X_i(s)Y_i)$. If for example $d = 4$ one may replace $Z_{\delta,i}(s)$ by

$$Z^*_{\delta,i}(s) := X_i(s) - \frac{2}{3}(X_i(s-\delta) + X_i(s+\delta)) + \frac{1}{6}(X_i(s-2\delta) + X_i(s+2\delta)).$$

Theoretical results then may be derived by modifying Assumption 2.1 by demanding that there exists now a d-times differentiable function ω such that (2.4) holds for $\kappa < d$.
(iv) In conjunction with the Gaussian assumption on X_i it is somewhat natural to rely on Lemma 2.1, see the discussion after this lemma. Estimation of points of impact is, however, still possible if the result from Lemma 2.1 does not hold, e.g., whenever X_i is not Gaussian but there exists a two times differentiable function $c_0(s)$ with $c_0(\tau_r) \neq 0$ and a bounded second derivative such that $\mathbb{E}(X_i(s)g(\eta_i)) = c_0(s)\,\mathbb{E}(X_i(s)\eta_i)$. In this case we have $\mathbb{E}(Z_{\delta,i}(s)Y_i) = c_0(s)\,\mathbb{E}(Z_{\delta,i}(s)\eta_i) + O(\delta^2)$. But the arguments for the estimation of the points of impact relied on $|\,\mathbb{E}(Z_{\delta,i}(s)\eta_i)|$, and hence, points of impact can still be estimated.

2.3 Parameter estimation

In the following it is assumed that the labels for the points of impact are ordered such that $\tau_r = \arg\min_{s=1,\dots,S} |\widehat{\tau}_r - \tau_s|$. Moreover we assume that S has been consistently estimated by $\widehat{S}$ and $\max_{r=1,\dots,\widehat{S}} |\widehat{\tau}_r - \tau_s| = O_P(n^{-\frac{1}{k}})$. For estimating the parameters $\alpha, \beta_1, \dots \beta_S$ we impose the following additional assumptions for model (2.2): Additional to $\mathbb{E}(\varepsilon_i | X_i(t), t \in [a, b]) = 0$ we assume that $\mathbb{V}(\varepsilon_i | X_i(t), t \in [a, b]) = \sigma^2(g(\eta_i)) < \infty$, where the variance function σ^2 is defined over the range of g and is strictly positive. For simplicity the function g is assumed to be a known strictly monotone and smooth function with bounded first and second order derivatives and hence invertible. Model (2.2) then implies $\mathbb{E}(Y_i | X_i) = g(\eta_i)$ as well as $\mathbb{V}(Y_i | X_i) = \sigma^2(g(\eta_i)) < \infty$ and, therefore, represents a quasi-likelihood model which

can be seen as a generalization of a generalized linear model framework (cf. McCullagh and Nelder, 1989, Ch. 9). The following result shows that our model is uniquely identified:

Theorem 2.2. *Let $g(\cdot)$ be invertible and assume that X_i satisfies Assumption 2.1. Then for all $S^* \geq S$, all $\alpha^*, \beta_1^*, \ldots, \beta_{S^*}^* \in \mathbb{R}$, 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*

$$\mathbb{E}\left(\left(g(\alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r)) - g(\alpha^* + \sum_{r=1}^{S^*} \beta_r^* X_i(\tau_r))\right)^2\right) > 0, \tag{2.8}$$

whenever

$|\alpha - \alpha^*| > 0$, *or* $\sup_{r=1,\ldots,S} |\beta_r - \beta_r^*| > 0$, *or* $\sup_{r=S+1,\ldots,S^*} |\beta_r^*| > 0$.

Note that it already follows from Theorem 2.1 that all points of impact τ_r are uniquely identifiable under the assumptions of the theorem. Invertibility of g additionally ensures that the coefficients $\alpha, \beta_1, \ldots, \beta_S$ are uniquely identified. Furthermore, it follows from the proof of Theorem 2.2, that under Assumption 2.1, $\mathbb{E}(\mathbf{X}_i(\boldsymbol{\tau})\mathbf{X}_i(\boldsymbol{\tau})^T)$ is invertible, where $\mathbf{X}_i(\boldsymbol{\tau}) = (1, X_i(\tau_1), \ldots, X_i(\tau_S))^T$.

Estimation of $\boldsymbol{\beta}_0 = (\alpha, \beta_1, \ldots, \beta_S)^T$ is performed by quasi-maximum likelihood. Define $\mathbf{X}_i(\widehat{\boldsymbol{\tau}}) = (1, X_i(\widehat{\tau}_1), \ldots, X_i(\widehat{\tau}_S))^T$ and denote the jth element of this vector as $\widehat{X}_{ij}$. For $\boldsymbol{\beta} \in \mathbf{R}^{S+1}$ let $\widehat{\eta}_i(\boldsymbol{\beta}) = \mathbf{X}_i(\widehat{\boldsymbol{\tau}})^T \boldsymbol{\beta}$, $\widehat{\boldsymbol{\mu}}_n(\boldsymbol{\beta}) = (g(\widehat{\eta}_1(\boldsymbol{\beta})), \ldots, g(\widehat{\eta}_n(\boldsymbol{\beta})))^T$, $\widehat{\mathbf{D}}_n(\boldsymbol{\beta})$ be the $n \times (S+1)$ matrix with entries $g'(\widehat{\eta}_i(\boldsymbol{\beta}))\widehat{X}_{ij}$, and let $\widehat{\mathbf{V}}_n(\boldsymbol{\beta})$ be a $n \times n$ diagonal matrix with elements $\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))$. Furthermore, denote the corresponding objects evaluated at the true points of impact τ_r by $\mathbf{X}_i(\tau)$, X_{ij}, $\eta_i(\boldsymbol{\beta})$, $\boldsymbol{\mu}_n(\boldsymbol{\beta})$, $\mathbf{D}_n(\boldsymbol{\beta})$, and $\mathbf{V}_n(\boldsymbol{\beta})$; this convention applies also to the below defined objects.

Then our estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0 = (\alpha, \beta_1, \ldots, \beta_S)^T$ is defined as the solution of the $S+1$ score equations $\widehat{\mathbf{U}}_n(\widehat{\boldsymbol{\beta}}) = 0$, where

$$\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = \widehat{\mathbf{D}}_n(\boldsymbol{\beta})^T \widehat{\mathbf{V}}_n(\boldsymbol{\beta})^{-1}(\mathbf{Y}_n - \widehat{\boldsymbol{\mu}}_n(\boldsymbol{\beta})). \tag{2.9}$$

Note that the score equations are evaluated at the estimates $\widehat{\tau}_r$ instead of τ_r.

In the following, it will be convenient to define

$$\mathbf{F}_n(\boldsymbol{\beta}) = \mathbf{D}_n(\boldsymbol{\beta})^T \mathbf{V}_n(\boldsymbol{\beta})^{-1} \mathbf{D}_n(\boldsymbol{\beta}) \quad \text{and} \quad \widehat{\mathbf{F}}_n(\boldsymbol{\beta}) = \widehat{\mathbf{D}}_n(\boldsymbol{\beta})^T \widehat{\mathbf{V}}_n(\boldsymbol{\beta})^{-1} \widehat{\mathbf{D}}_n(\boldsymbol{\beta}).$$

Observe that the $S+1 \times S+1$ matrix $\mathbb{E}(n^{-1}\mathbf{F}_n(\boldsymbol{\beta}))$ can be represented as $\mathbb{E}(n^{-1}\mathbf{F}_n(\boldsymbol{\beta})) = [\mathbb{E}(g'^2(\eta_i(\boldsymbol{\beta}))/\sigma^2(g(\eta_i(\boldsymbol{\beta})))X_{ik}X_{il})]_{k,l}$, where $k = 1, \ldots, S+1$ and $l = 1, \ldots, S+1$. Let $\eta(\boldsymbol{\beta})$ and X_j be generic copies of $\eta_i(\boldsymbol{\beta})$ and the jth component of $\mathbf{X}_i$, respectively. This allows us to write $\mathbb{E}(n^{-1}\mathbf{F}_n(\boldsymbol{\beta})) = \mathbb{E}(\mathbf{F}(\boldsymbol{\beta}))$ with $\mathbb{E}(\mathbf{F}(\boldsymbol{\beta})) = [\mathbb{E}(g'^2(\eta(\boldsymbol{\beta}))/\sigma^2(g(\eta(\boldsymbol{\beta})))X_kX_l)]_{k,l}$,

where we point out that $\mathbb{E}(\mathbf{F}(\boldsymbol{\beta}))$ is for all $\boldsymbol{\beta} \in \mathbf{R}^{S+1}$ a symmetric and strictly positive definite matrix with inverse $\mathbb{E}(\mathbf{F}(\boldsymbol{\beta}))^{-1}$. Indeed, suppose $\mathbb{E}(\mathbf{F}(\boldsymbol{\beta}))$ is not strictly positive definite, one would then derive the contradiction $\mathbb{E}((\sum_{j=1}^{S+1} a_j X_j g'(\eta(\boldsymbol{\beta}))/\sigma(g(\eta(\boldsymbol{\beta}))))^2) = 0$ for nonzero constants $a_1, \ldots, a_{S+1}$. A similar argument can be used to show that $\mathbb{E}(\widehat{\mathbf{F}}(\boldsymbol{\beta}))$ is strictly positive definite, where $\mathbb{E}(\widehat{\mathbf{F}}(\boldsymbol{\beta})) = [\mathbb{E}(g'^2(\widehat{\eta}(\boldsymbol{\beta}))/\sigma^2(g(\widehat{\eta}(\boldsymbol{\beta})))\widehat{X}_k\widehat{X}_l)]_{k,l}$.

In the rest of this section we assume X_i to be i.i.d. Gaussian distributed which covariance $\sigma(s,t)$ satisfying Assumption 2.1. The following additional set of assumptions are used to derive more precise theoretical statements:

Assumption 2.3.

a) There exists a constant $0 < M_\varepsilon < \infty$, such that $\mathbb{E}(\varepsilon_i^p | X_i) \leq M_\varepsilon$ for some even p with $p \geq \max\{2/\kappa + \epsilon, 4\}$ for some $\epsilon > 0$.

b) The link function g is monotone, invertible with two bounded derivatives $|g'(\cdot)| \leq c_g$, $|g''(\cdot)| \leq c_g$, for some constant $0 \leq c_g < \infty$.

c) $h(\cdot) := \frac{g'(\cdot)}{\sigma^2(g(\cdot))}$ is a bounded function with two bounded derivatives.

Assumption 2.3 a) states that some higher moments of ε_i exist. While the condition on $p \geq 4$ and p being even simplifies the proofs, the condition $p > 2/\kappa$ is a more crucial one and is used in the proof of Proposition C.1 in Appendix C. The Assumptions 2.3 a) and b) and c) hold, for example, in the important case of a functional logistic regression with points of impact. Assumption 2.3 c) is satisfied, for instance, in the special case of generalized linear models with natural link functions. For the latter case, we have $\sigma^2(g(x)) = g'(x)$ such that $h(x) = 1$. The boundedness conditions in b) and c) constitute a set of sufficient conditions needed to obtain our theoretical results.

Theorem 2.3. *Let $\widehat{S} = S$, $\max_{r=1,\ldots,S} |\widehat{\tau}_r - \tau_r| = O_p(n^{-1/\kappa})$ and let X_i be a Gaussian process satisfying Assumption 2.1. under Assumption 2.3 we then obtain*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, (\mathbb{E}(\mathbf{F}(\boldsymbol{\beta}_0)))^{-1}). \tag{2.10}$$

This result is remarkable; our estimator based on $\widehat{\tau}_r$ enjoys the same asymptotic efficiency properties as the infeasible estimator based on the true points of impact τ_r. It achieves the same asymptotic efficiency properties under classical multivariate setups (cf. McCullagh, 1983). In practice one might then replace $\mathbb{E}(\mathbf{F}(\boldsymbol{\beta}_0))$ by its consistent estimator $n^{-1}\widehat{\mathbf{F}}_n(\widehat{\boldsymbol{\beta}})$ in order to derive approximate results. This is a direct consequence of (C.24) and (C.50) in the supplementary Appendix C.

Parameter estimation under misspecified variance functions

So far, we have considered the case where $\sigma^2(g(\eta_i(\boldsymbol{\beta})))$ is specified using a (correct) model assumption. In the following, we consider situations where only the mean function $g(\eta_i(\boldsymbol{\beta}))$ can be specified, but where the functional form of $\sigma^2(\cdot)$ is unknown. By Theorem 2.2, an estimator $\widetilde{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$ minimizes the squared error

$$\widetilde{\beta} = \arg\min_{\boldsymbol{\beta}\in\mathbf{R}^{S+1}} \frac{1}{2n}\sum_{i=1}^{n}(y_i - g(\widehat{\eta}_i(\boldsymbol{\beta})))^2.$$

The estimator $\widetilde{\boldsymbol{\beta}}$ can then be described as the solution of the score functions $\widetilde{\mathbf{U}}_n(\boldsymbol{\beta}) = 0$, where

$$\widetilde{\mathbf{U}}_n(\boldsymbol{\beta}) = \widehat{\mathbf{D}}_n(\boldsymbol{\beta})^T(\mathbf{Y}_n - \widehat{\boldsymbol{\mu}}_n(\boldsymbol{\beta})). \tag{2.11}$$

Provided $|g'''(x)| \leq M_g$, we get the following corollary by following the same arguments as in the proof of Theorem 2.3:

Corrolary 2.1. *Under the Assumptions of Theorem 2.3, but with Assumption 2.3 c) replaced by the assumption that $|g'''(x)| \leq M_g$ for some $0 \leq M_g < \infty$, we have*

$$\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}), \tag{2.12}$$

where

$$\mathbf{A} = \mathbb{E}(g'(\eta(\boldsymbol{\beta}_0))^2\mathbf{X}\mathbf{X}^T) \text{ and } \mathbf{B} = cov(g'(\eta(\boldsymbol{\beta}_0))\mathbf{X}\varepsilon) = \mathbb{E}(g'(\eta(\boldsymbol{\beta}_0))^2\sigma^2(g(\eta(\boldsymbol{\beta}_0)))\mathbf{X}\mathbf{X}^T).$$

In practice one might replace the sandwich matrix in (2.12) by their estimators, i.e., replacing $\mathbb{E}(g'(\eta(\boldsymbol{\beta}_0))^2\mathbf{X}\mathbf{X}^T)$ by $n^{-1}\sum_{i=1}^{n} g'(\eta_i(\widetilde{\boldsymbol{\beta}}))\widehat{\mathbf{X}}_i\widehat{\mathbf{X}}_i^T$ and $cov(g'(\eta(\boldsymbol{\beta}_0))\mathbf{X}\varepsilon)$ by $n^{-1}\sum_{i=1}^{n} g'(\widehat{\eta}_i(\widetilde{\boldsymbol{\beta}}))^2(y_i - g(\widehat{\eta}_i(\widetilde{\boldsymbol{\beta}})))^2\widehat{\mathbf{X}}_i\widehat{\mathbf{X}}_i^T$.

The above case corresponds to situations where $\sigma^2(g(\eta_i(\boldsymbol{\beta})))$ is incorrectly specified by $\widetilde{\sigma}^2(g(\eta_i(\boldsymbol{\beta})))$ with $\widetilde{\sigma}^2(g(\eta_i(\boldsymbol{\beta}))) = 1$. More general misspecifications lead to similar sandwich estimators as in (2.12) provided $\widetilde{h}(\cdot) = g'(\cdot)/\widetilde{\sigma}^2(\cdot)$ is a bounded function with two bounded derivatives.

2.4 Practical implementation

An implementation of our estimation procedure comprises, first, the estimation of the points of impact τ_r and, second, the estimation of the parameters α and β_r associated with $X_i(\tau_r)$. Estimating the points of impact τ_r relies on the choice of δ and a choice of the cut-off parameter λ (see Section 2.2.1). Asymptotic specifications are given in Theorem 2.1, however, these determine the tuning parameters δ and λ only up to constants and are typically not useful in

practice. In the following we propose to select the tuning parameters using a fully data-driven model selection approach.

For a given δ, our estimation procedure leads to a set of potential point of impact candidates $\{\widehat{\tau}_1, \widehat{\tau}_2, \ldots, \widehat{\tau}_{M_\delta}\}$ (see Section 2.2.1). Selecting final point of impact estimates from this set of candidates corresponds to a classical variable selection problem. In the case where the distribution of $Y_i|X_i$ belongs to the exponential family (as in the logistic regression) one may perform a best subset selection optimizing a standard model selection criterion such as the Bayesian Information Criterion (BIC),

$$\mathrm{BIC}_{\mathscr{X}}(\delta) = -2\log \mathscr{L}_{\mathscr{X}} + K_{\mathscr{X}} \log(n). \tag{2.13}$$

Here, $\log \mathscr{L}_{\mathscr{X}}$ is the log-likelihood of the model with intercept and predictor variables $\mathscr{X} \subseteq \{X_i(\widehat{\tau}_1), X_i(\widehat{\tau}_2), \ldots, X_i(\widehat{\tau}_{M_\delta})\}$, where $K_{\mathscr{X}} = 1 + |\mathscr{X}|$ denotes the number of predictors. Minimizing $\mathrm{BIC}_{\mathscr{X}}(\delta)$ over $0 < \delta < (b-a)/2$ leads to the final model choice.

In the case where only the first two moments $\mathbb{E}(Y_i|X_i) = g(\eta_i)$ and $\mathbb{V}(Y_i|X_i) = \sigma^2(g(\eta_i)) < \infty$ are known, one may replace the deviance $-2\log \mathscr{L}_{\mathscr{X}}$ by the quasi deviance $-2Q_{\mathscr{X}} = -2\sum_{i=1}^{n} \int_{y_i}^{g(\widehat{\eta}_{\mathscr{X},i})} (y_i - t)/(\sigma^2(t))\, dt$, where $\widehat{\eta}_{\mathscr{X},i}$ is the linear predictor with intercept and predictor variables $\mathscr{X}$.

In order to calculate $\mathrm{BIC}_{\mathscr{X}}(\delta)$, we need the estimates $\widehat{\boldsymbol{\beta}}$ solving the estimation equations $\widehat{\mathbf{U}}_n(\widehat{\boldsymbol{\beta}}) = 0$. In practice these equations are solved iteratively, for instance, by the usual Newton-Raphson method with Fisher-type scoring. That is, for an arbitrary initial value $\widehat{\boldsymbol{\beta}}_0$ sufficiently close to $\widehat{\boldsymbol{\beta}}$ one generates a sequence of estimates $\widehat{\boldsymbol{\beta}}_m$, with $m = 1, 2, \ldots$,

$$\widehat{\boldsymbol{\beta}}_m = \widehat{\boldsymbol{\beta}}_{m-1} + \left(\widehat{\mathbf{F}}_n(\widehat{\boldsymbol{\beta}}_{m-1})\right)^{-1} \widehat{\mathbf{U}}_n(\widehat{\boldsymbol{\beta}}_{m-1}). \tag{2.14}$$

Iteration is executed until convergence and the final step of the procedure yields the estimate $\widehat{\boldsymbol{\beta}}$. Here, $\widehat{\mathbf{F}}_n(\boldsymbol{\beta})$ and $\widehat{\mathbf{U}}_n(\boldsymbol{\beta})$ replace $\mathbf{F}_n(\boldsymbol{\beta})$ and $\mathbf{U}_n(\boldsymbol{\beta})$ in the usual Fisher scoring algorithm, since the unknown τ_r are replaced by their estimates $\widehat{\tau}_r$. The latter is justified asymptotically by our results in Corollary C.1 and Proposition C.2 in Appendix C.

2.5 Simulation

We investigate the finite sample performance of our estimators using Monte Carlo simulations. After simulating a trajectory X_i over p equidistant grid points t_j, $j = 1, \ldots, p$, on $[a, b] = [0, 1]$, linear predictors of the form $\eta_i = \alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r)$ are constructed for some predetermined model parameters α, β_r, τ_r, and S, where a point of impact is implemented as the smallest observed grid point t_j closest to τ_r. The response Y_i is derived as a realization of a Bernoulli random variable with success probability $g(\eta_i) = \exp(\eta_i)/(1 + \exp(\eta_i))$,

resulting in a logistic regression framework with points of impact. The simulation study is implemented in R (R Core Team, 2017), where we use the R-package glmulti (Calcagno, 2013) in order to implement the fully data-driven BIC-based best subset selection estimation procedure described in Section 2.4. The threshold estimator from Section 2.2.1 requires appropriate choices of $\delta = \delta_n$ and $\lambda = \lambda_n$. Theorem 2.1 suggests that a suitable choice of δ is given by $\delta = c_\delta\, n^{-1/2}$ for some constant $c_\delta > 0$. Our simulation results are based on $c_\delta = 1.5$; similar qualitative results were derived for a broader range of values c_δ. For the threshold λ we use $\lambda = A\sqrt{\sqrt{\widehat{\mathbb{E}}(Y_i^4)}\log((b-a)/\delta)/n}$, where $A = \sqrt{2\sqrt{3}}$ and $\widehat{\mathbb{E}}(Y_i^4) = n^{-1}\sum_{i=1}^n Y_i^4$, as motivated in connection to Theorem 2.1. Estimated points of impact candidates are related to the true impact points by the following matching rule: In a first step the interval $[a,b]$ is partitioned into S subintervals of the form $I_j = [m_{j-1}, m_j)$, where $m_0 = a$, $m_S = b$ and $m_j = (\tau_j + \tau_{j+1})/2$ for $0 < j < S$. The candidate $\widehat{\tau}_l$ in interval I_j with the closest distance to τ_j is then taken as the estimate of τ_j. No impact point estimate in an interval results in an unmatched τ_j and a missing value when calculating statistics for the estimator. Results are based on 1000 Monte Carlo iterations for each constellation of $n \in \{100, 200, 500, 1000, 5000\}$ and $p \in \{100, 500, 1000\}$. Estimation errors are illustrated by boxplots, where the error bars at the end of the whiskers represent the 10% and 90% quantiles. Five data generating processes (DGP) are considered (see Table 2.1) using the following three processes $X_i(t)$ covering a broad range of situations:

OUP ORNSTEIN-UHLENBECK PROCESS. A Gaussian process with covariance function $\sigma(s,t) = \sigma_u^2/(2\theta)(\exp(-\theta|s-t|) - \exp(-\theta(s+t)))$. We choose $\theta = 5$ and $\sigma_u^2 = 3.5$.

GCM GAUSSIAN COVARIANCE MODEL. A Gaussian process with covariance function $\sigma(s,t) = \sigma(|s-t|) = \exp(-(|s-t|/d)^2)$. We choose $d = 1/10$.

EBM EXPONENTIAL BROWNIAN MOTION. A non Gaussian process with covariance function $\sigma(s,t) = \exp((s+t+|s-t|)/2) - 1$. It is defined by $X_i(t) = \exp(B_i(t))$, where $B_i(t)$ is a Brownian motion.

Table 2.1: *Data generating processes considered in the simulations*

Model		**Points of impact**					**Parameters**				
Label	**Process**	S	τ_1	τ_2	τ_3	τ_4	α	β_1	β_2	β_3	β_4
DGP 1	OUP	1*	1/2				1	4			
DGP 2	OUP	2	1/3	2/3			1	-6	5		
DGP 3	OUP	4	1/6	2/6	4/6	5/6	1	-6	6	-5	5
DGP 4	GCM	2	1/3	2/3			1	-6	5		
DGP 5	EBM	2	1/3	2/3			1	-6	5		

*Note: $S = 1$ is assumed known (only in DGP 1).

DGP 1-3 are increasingly complex, but satisfy our theoretical assumptions. The general setups of DGP 4 and DGP 5 are equivalent to DGP 2, but the processes X_i (GCM and EBM) violate our theoretical assumptions. The GCM covariance function of X_i in DGP 4 is infinitely differentiable, even at the diagonal where $s = t$, contradicting Assumption 2.1, but fitting the remark underneath this Assumption. The EBM process in DGP 4 contradicts the Gaussian Assumption 2.2.

DGP 1 allows us to compare our data-driven BIC-based estimation procedure from Section 2.4 (denoted as POI) with the estimation procedure of Lindquist and McKeague (2009) (denoted as LMcK). Lindquist and McKeague (2009) consider situations where $S = 1$ is known and propose to estimate the unknown parameters α, β_1 and τ_1 by simultaneously maximizing the likelihood over α, β_1 and the grid points t_j. Our estimation procedure does not require knowledge about S, but profits from a situation where $S = 1$ is known. Therefore, for reasons of comparability, we restrict the BIC-based model selection process to allow only for models containing at most one point of impact candidate. The simulation results are depicted in Figure 2.3 and are virtually identical for both methods and show a satisfying behavior of the estimates. It should be noted, however, that our estimator is computationally advantageous as it greatly thins out the number of possible point of impact candidates by allowing only the local maxima of $|1/N \sum_{i=1}^{n} Z_{\delta,i}(s)Y_i|$ as possible point of impact candidates. Our practically less relevant threshold based estimation procedure leads to similar qualitative results. These results are, however, omitted in order to allow for a clear display in Figure 2.3. The performance of our threshold based procedure is reported in detail for the remaining simulation studies (DGP 2-5).

DGP 1: ESTIMATION ERRORS FOR DIFFERENT SAMPLE SIZES n

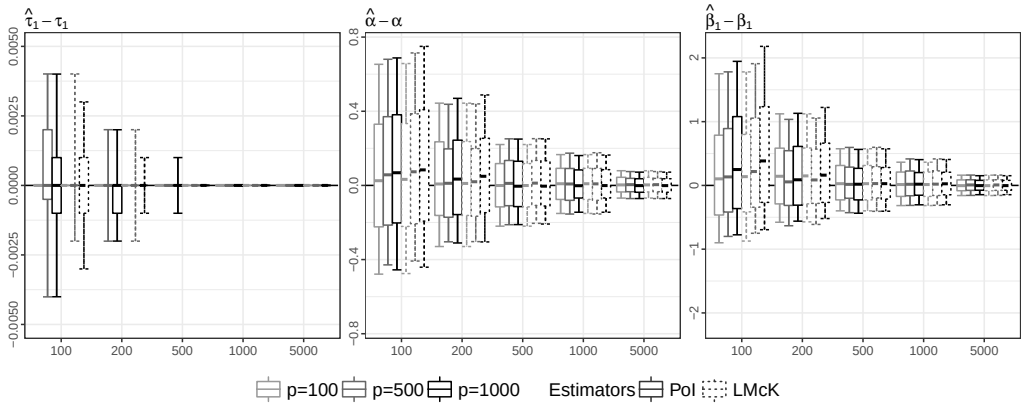

Figure 2.3: *Comparison of the estimation errors from using our BIC-based method POI (solid lines) and the method of Lindquist and McKeague (2009) (dashed lines). The error bars of the boxplots are set to the* 10% *and* 90% *quantiles.*

DGP 2: ESTIMATION ERRORS FOR DIFFERENT SAMPLE SIZES n

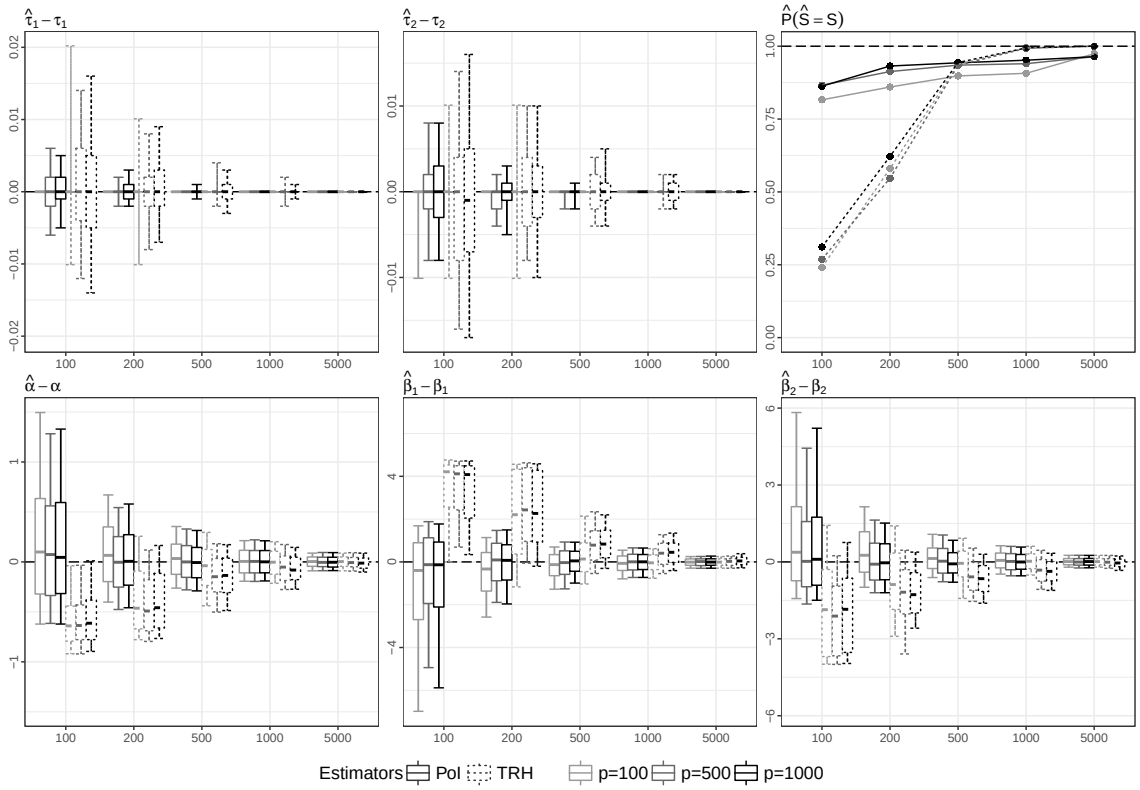

Figure 2.4: *Comparison of the estimation errors from using our BIC-based method POI (solid lines) and our threshold-based method TRH (dashed lines). The error bars of the boxplots are set to the* 10% *and* 90% *quantiles.*

DGP 2 is more complex than DGP 1, also because $S = 2$ is considered unknown. Figure 2.4 compares the estimation errors from using our BIC-based POI estimator with those from our threshold-based estimator (denoted as TRH). For smaller sample sizes n, the POI estimator seems to be preferable to the TRH estimator. Although, estimating the locations of the points of impact τ_1 and τ_2 is quite accurate for both procedures, the number S is more often estimated correctly using the POI estimator (see upper right panel). This more precise estimation of S results in essentially unbiased estimates of the parameters α, β_1, and β_2. By contrast, the less precise estimation of S using the TRH estimator leads to clearly visible omitted variable biases in the estimates of the parameters α, β_1, and β_2. As the sample size increases, however, the accuracy of estimating $\widehat{S}$ improves for the TRH estimator such that both estimators show a similar performance.

DGP 3 with $S = 4$ unknown points of impact comprises an even more complex situation than DGP 2. Figure A.1 in Appendix A shows that the qualitative results from DGP 2 still hold. For large n, the POI and TRH estimators lead both to accurate estimates of the model parameters for all choices of p. As already observed in DGP 2, however, the TRH estimator

leads to imprecise estimates of S for small n, which results in omitted variables biases in the estimates of the parameters α, β_1, β_2, β_3, and β_4. Because of the increased complexity of DGP 3, these biases are even more pronounced than in DGP 2. The reason for this is partly due to the construction of the TRH estimator, where we set the value of δ to $\delta = c_\delta n^{-1/2}$ with $c_\delta = 1.5$. Asymptotically, the choice of c_δ has a negligible effect, but may be inappropriate for small n, since the estimation procedure eliminates all points within a $\sqrt{\delta}$-neighborhood around a chosen candidate $\widehat{\tau}_r$ (see Section 2.2.1). For DGP 3, the choice of $c_\delta = 1.5$ results in a too large $\sqrt{\delta}$-neighborhood, such that the estimation procedure eliminates also true point of impact locations for small n. By contrast, the POI estimator is able to avoid such adverse eliminations as the BIC criterion is also minimized over δ.

DGP 4: ESTIMATION ERRORS FOR DIFFERENT SAMPLE SIZES n

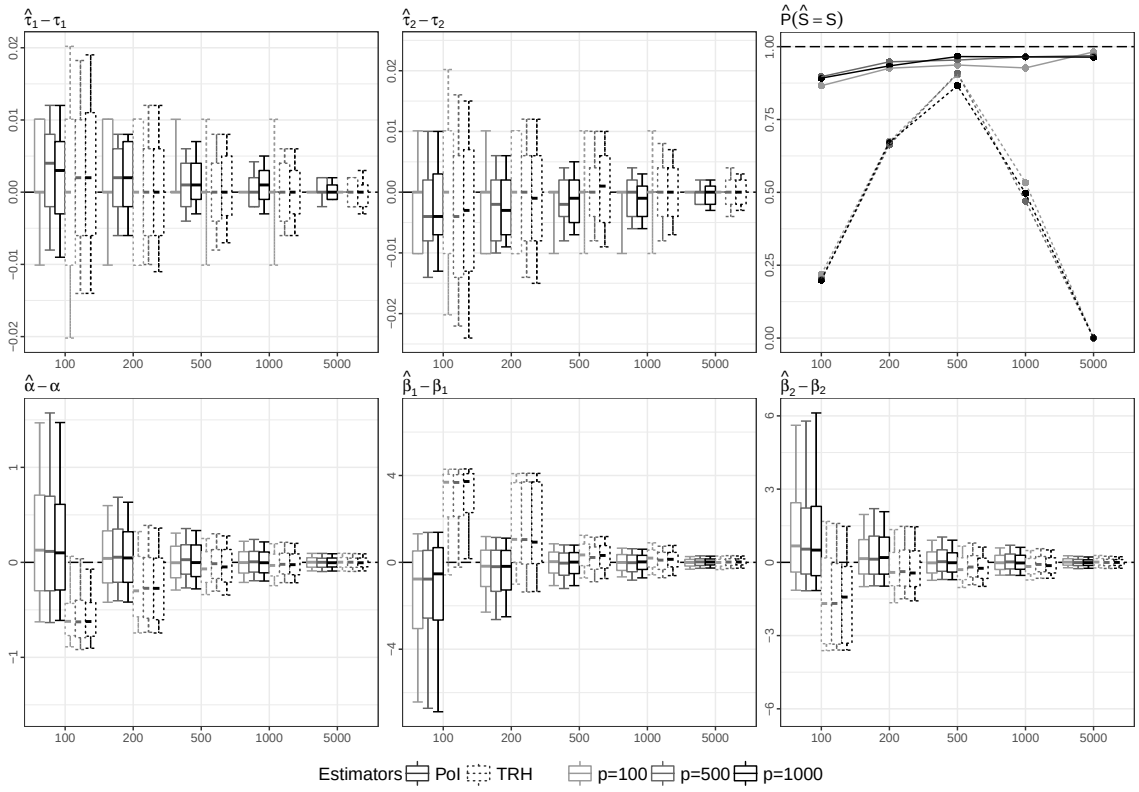

Figure 2.5: *Comparison of the estimation errors from using our BIC-based method POI (solid lines) and our threshold-based method TRH (dashed lines). The error bars of the boxplots are set to the* 10% *and* 90% *quantiles.*

DGP 4 takes up the general setup of DGP 2, but the functional data X_i are simulated using a Gaussian covariance model (GCM) which is characterized by an indefinite differentiable covariance function. This setting contradicts our basic Assumption 2.1, but fits its remark underneath this Assumption. From Figure 2.5 it can be concluded that even under the failure of Assumption 2.1, both estimation procedures are capable of consistently estimating the points

of impact and the model parameters. The TRH estimator, however, fails to estimate the parameter S even for large n, since the λ-threshold tailored for situations under Assumption 2.1. Here the TRH estimator is able to estimate the true points of impact, but additionally selects more and more redundant point of impact candidates as n becomes large. That is, the TRH estimator becomes more a screening than a selection procedure which can be problematic in practice. By contrast, the POI estimator is able to avoid such redundant selections of point of impact candidates as the BIC criterion only selects points of impact candidates if they result in a sufficiently large improvement of the model fit.

DGP 5 also takes up the setup of DGP 2, however, the process X_i is simulated as an exponential Brownian motion (EBM), which is non Gaussian, violating Assumption 2.2, but still satisfying Assumption 2.1. Here we set the asymptotically negligible tuning parameter c_δ of the TRH estimator equal to 3. The evolution of the estimation errors can be seen in Figure A.2 in Appendix A. The results are comparable with our previous simulations in DGP 2 and DGP 3, indicating that the estimation procedure is robust to at least some violations of Assumption 2.2.

Resume: Asymptotically both estimation procedures POI and TRH work well. The effect of increasing p is generally negligible for all considered sample sizes n. Estimates of τ_r are very accurate, especially if kept in mind that the distance between two successive grid points is given by approximately 0.01, 0.002 and 0.001 for our choices of p. In small samples and for violations of the model assumptions, however, there seems to be a clear advantage when using the POI-estimator.

2.6 Points of impact in continuous emotional stimuli

Current psychological research on emotional experiences increasingly includes continuous emotional stimuli such as videos to induce emotional states as an attempt to increase ecological validity (see, e.g., Trautmann et al., 2009). Asking participants to evaluate those stimuli is most often done after presenting the video using an overall rating such as "How positive or negative did this video make you feel?" or "Do you rate this video as positive or negative?". Such global overall ratings are guided by the participant's affective experiences while watching the video (Schubert, 1999; Mauss et al., 2005) which makes it crucial to identify the relevant parts of the stimulus impacting on the overall rating in order to understand the emergence of emotional states and to make use of specific part of such stimuli.

Due to a lack of appropriate statistical methods, existing approaches use heuristics such as the "peak-and-end rule" in order to link the overall ratings with the continuous emotional stimuli. The peak-and-end rule states that people's global evaluations can be well predicted using just two characteristics: the moment of emotional peak intensity and the ending of the emotional stimuli (see review by Fredrickson, 2000). Such a heuristic approach, however,

is only of a limited practical use. The peak intensity moment and the ending are not necessarily good predictors. Furthermore, the peak intensity moment can strongly vary across participants, which prevents to link the overall rating to moments in the continuous emotional stimuli which are of a common relevance. Both of these limitations are clearly visible in our real data application.

Our case study comprises data from $n = 67$ participants, who were asked to continuously report their emotional state (from very negative to very positive) while watching a documentary video (112 sec.) on the persecution of African albinos[1]. The video does not contain emotionally arousing visual material, but the spoken words contain some emotionally arousing descriptions.

Figure 2.1 shows the continuously self-reported feeling trajectories $X_1(t_j), \ldots, X_n(t_j)$, where t_j are equidistant grid points within the unit-interval $0 = t_1 < \cdots < t_p = 1$ with $p = 167$. After watching the video, the participants were asked to rate their overall feeling. This overall rating was coded as a binary variable Y, where $Y = 1$ denotes "I feel positive" and $Y = 0$ denotes "I feel negative". The data were collected in May 2013. Participants were recruited through Amazon Mechanical Turk (`www.mturk.com`) and received 1USD reimbursement for completing the ratings via the online survey platform Soscisurvey (`www.soscisurvey.de`). The study was approved by the local institutional review board (IRB, University of Colorado Boulder). The documentary video is provided by the Interdisciplinary Affective Science Laboratory (`www.affective-science.org`).

We compare our BIC-based POI estimation procedure with the performance of the following two logit regression models based on peak-and-end rule (PER) predictor variables:

PER-1 Logit regression with peak intensity predictor $X_i(p_i^{\text{abs}})$ and the end-feeling predictor $X_i(1)$, where $p_i^{\text{abs}} = \arg\max_t(|X_i(t)|)$

PER-2 Logit regression with peak intensity predictors $X_i(p_i^{\text{pos}})$ and $X_i(p_i^{\text{neg}})$ and end-feeling predictor $X_i(1)$, where $p_i^{\text{pos}} = \arg\max_t(X_i(t))$ and $p_i^{\text{neg}} = \arg\min_t(X_i(t))$

Table 2.2 shows the estimated coefficients, standard errors, as well as summary statistics for each of the three models. In comparison to our POI estimator, both benchmark models (PER 1&2) have significantly lower model fits (McFadden Pseudo R^2) and significantly lower predictive abilities (Somers' D_{xy}), where $D_{xy} = 0$ means that a model is making random predictions and $D_{xy} = 1$ means that a model discriminates perfectly.

Figure 2.6 shows the positive (p) and negative (n) peak intensity predictors $X_i(p_i^{\text{pos}})$ and $X_i(p_i^{\text{neg}})$ for all participants; the absolute intensity predictors $X_i(p_i^{\text{abs}})$ form a subset of these. It is striking that the peak intensity predictors are distributed across the total domain $[0, 1]$ and, therefore, do not allow to link the overall ratings Y_i to specific time points t in the continuous

[1]The persecution of African albinos primarily happens in East Africa, where still well-established witch doctors use albino body parts for good luck potions for which clients are willing to pay high prices.

Table 2.2: *Estimation results using emotional stimuli data. For each model the table contains the estimated parameters, their significance codes and their corresponding standard error. The overall model quality is evaluated using four different criteria.*

	POI		**PER-1**		**PER-2**	
Regressor	**Coefficient**	**(S.E.)**	**Coefficient**	**(S.E.)**	**Coefficient**	**(S.E.)**
$X(\widehat{\tau}_1)$	-1.862^{***}	(0.673)				
$X(\widehat{\tau}_2)$	-1.271^{**}	(0.521)				
$X(p^{\text{abs}})$			-0.396	(0.452)		
$X(p^{\text{pos}})$					-0.012	(0.463)
$X(p^{\text{neg}})$					0.434	(0.559)
$X(1)$			0.245	(0.287)	0.243	(0.289)
Constant	0.089	(0.265)	0.683	(0.720)	0.583	(0.690)
Log Likelihood	-41.053		-45.689		-45.671	
Akaike Inf. Crit.	88.106		97.377		99.343	
McFadden Pseudo-R^2	0.115		0.015		0.015	
Somers' D_{xy}	0.406		0.153		0.135	

Note: *pvalue<0.1; **pvalue<0.05; ***pvalue<0.01

emotional stimuli. By contrast, the estimated points of impact $\widehat{\tau}_1$ and $\widehat{\tau}_2$ allow for such a link and point to the following two emotionally arousing text phrases spoken at those impact points: "even genitals" ($\widehat{\tau}_1$) and "selling his brother's body parts" ($\widehat{\tau}_2$).

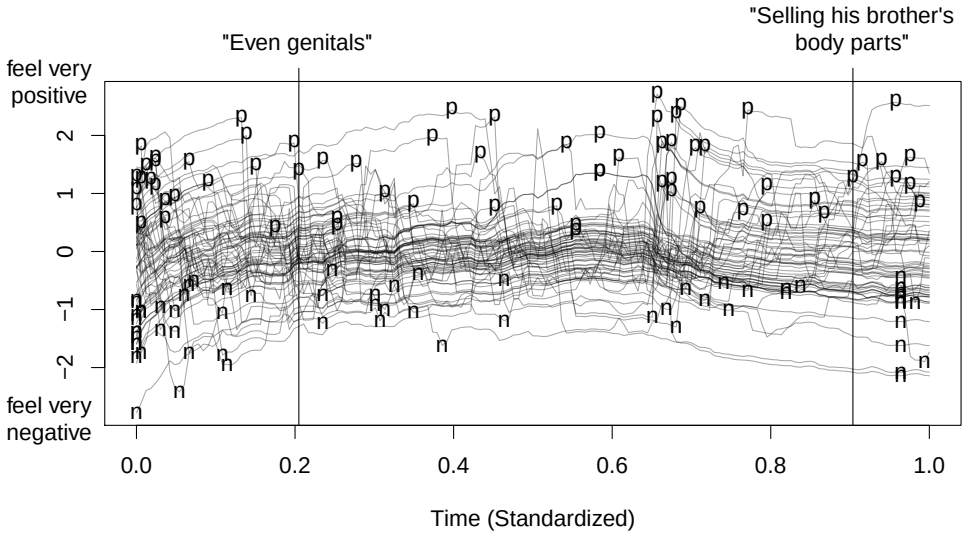

Figure 2.6: *Visualization of the positive (p) and negative (n) peak intensity predictors $X_i(p_i^{\text{pos}})$ and $X_i(p_i^{\text{neg}})$ and the two impact points $\widehat{\tau}_1$ and $\widehat{\tau}_2$ (vertical lines) along with the corresponding text phrases from the video.*

Acknowledgments: We thank Lisa Feldman Barrett (Northeastern University) and Tor D. Wager (University of Colorado Boulder) for sharing the data acquired with their funding: The development of the Interdisciplinary Affective Science Laboratory (IASLab) Movie Set was supported by a funds from the U.S. Army Research Institute for the Behavioral and Social Sciences (contract W5J9CQ-11-C-0046) to Lisa Feldman Barrett and Tor D. Wager, and a National Institutes of Health Director's Pioneer Award (DP1OD003312) to Lisa Feldman Barrett. The online rating tool for the data collection was provided by Dominik Leiner, soscisurvey, Germany). The views, opinions, and/or findings contained in this paper are those of the authors and shall not be construed as an official U.S. Department of the Army position, policy, or decision, unless so designated by other documents.

Supplement to: Points of Impact in Generalized Linear Models with Functional Predictors

This supplement to "Points of Impact in Generalized Linear Models with Functional Predictors" contains results from some additional simulations, proofs of our theoretical results and further derivations. It is divided into four appendices. Appendix A contains the simulation results for DGP 3 and DGP 5 from Section 2.5. In Appendix B proofs related to the estimation of the points of impact as presented in Section 2.2 can be found. Proofs for the parameter estimates from Section 2.3 are collected in Appendix C.

Appendix D is concerned about a situation in which the linear predictor is given by $\eta = \alpha + \sum_{r=1}^{S} \beta_r X(\tau_r) + \int_a^b \beta(t)X(t)\,dt$ and provides some additional theoretical results, another simulation study and additional proofs concerning this last part of the supplement.

In the following $||X||_\Phi = \inf\{C > 0 : \mathbb{E}(\Phi(|X|/C)) \leq 1\}$ refers to the Orlicz norm of a random variable X with respect to $\Phi(x) = \exp(n/6(\sqrt{1+2\sqrt{6}x/\sqrt{n}}-1)^2)-1$. Similar we use for $p \geq 1$ the Orlicz norm $||X||_p = \{\inf C > 0 : (\mathbb{E}(|X|^p))^{1/p} < C\}$ which corresponds to the usual L_p-norm.

Appendix A Additional simulation results

This appendix contains two additional figures showing the remaining simulation results discussed in Section 2.5 in the main paper. While Figure A.1 depicts the results from DGP 3, Figure A.2 illustrates the results from DGP 5.

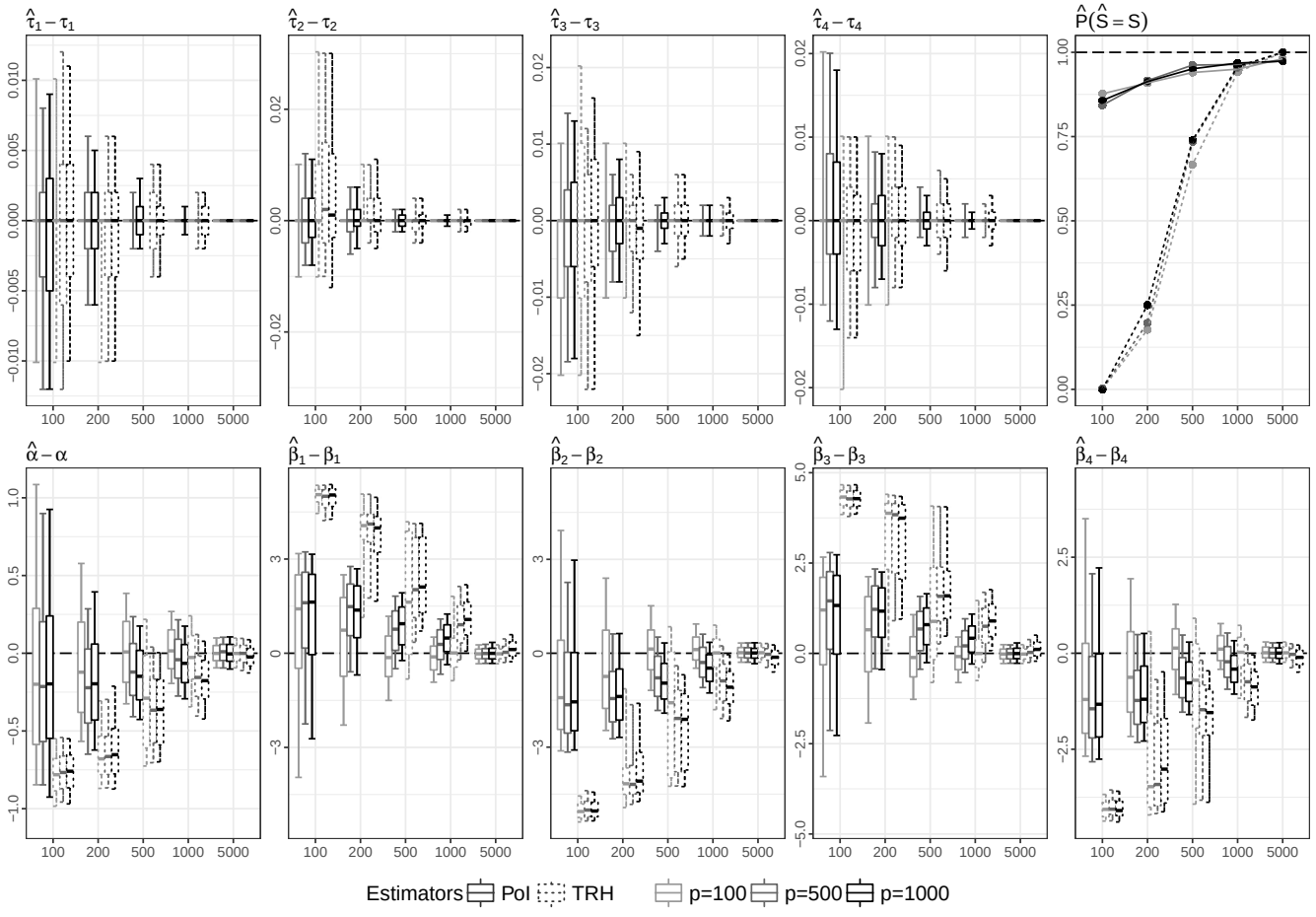

Figure A.1: *Comparison of the estimation errors from using our BIC-based method POI (solid lines) and our threshold-based method TRH (dashed lines). The error bars of the boxplots are set to the* 10% *and* 90% *quantiles.*

DGP 5: ESTIMATION ERRORS FOR DIFFERENT SAMPLE SIZES n

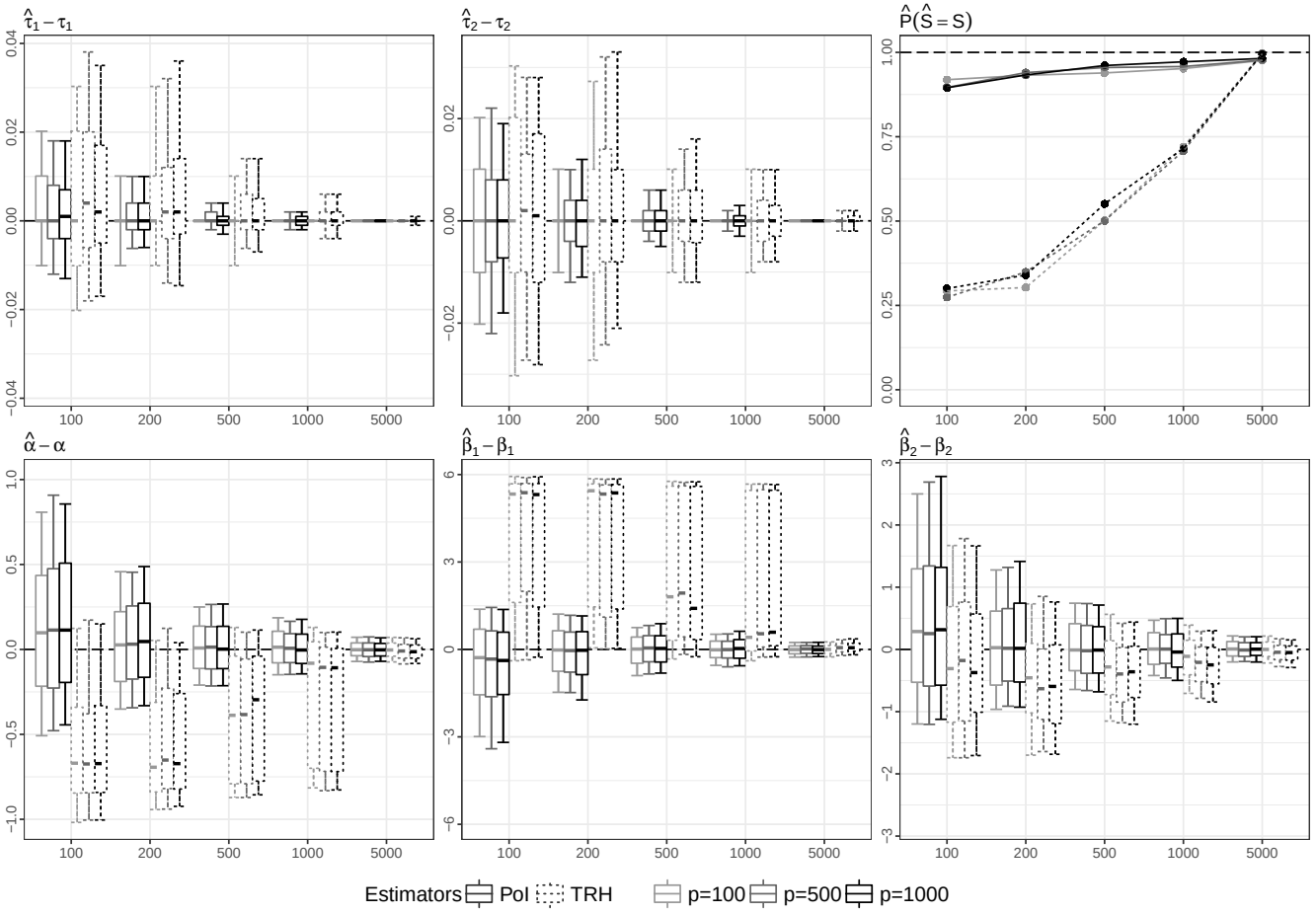

Figure A.2: *Comparison of the estimation errors from using our BIC-based method POI (solid lines) and our threshold-based method TRH (dashed lines). The error bars of the boxplots are set to the* 10% *and* 90% *quantiles.*

Appendix B Proofs of the theoretical results from Section 2.2

This appendix contains the proofs related to the estimation of the points of impact as presented in Section 2.2. The proof of Theorem 2.1 relies on the results in Kneip et al. (2016a); in fact, under our Assumption 2.1 in particularly Theorem 3 in Kneip et al. (2016a) holds. In order to proof Theorem 2.1 we need to adjust Lemma 1–4 from Kneip et al. (2016b) to our current setting (see Lemma B.1–B.4 below). Together with Lemma 2.1, Theorem 2.1 will then be an immediate consequence.

We will first proof Lemma 2.1.

Proof of Lemma 2.1. Note that $\mathbb{E}(X_i(s)Y_i)$ can be written as $\mathbb{E}(X_i(s)Y_i) = cov(X_i(s), g(\eta_i) + \varepsilon_i) = \mathbb{E}(X_i(s)g(\eta_i)) = cov(X_i(s), g(\eta_i))$. Moreover, if X_i is additionally assumed to be Gaussian, a direct proof of of Lemma 2.1 then follows already from the proof of Lemma 1 in Brillinger (2012a). We consider the case where X_i is not assumed to be Gaussian.

Under the assumptions of Lemma 2.1 we can decompose $X_i(s)$ by

$$X_i(s) = \frac{\mathbb{E}(X_i(s)\widetilde{\eta}_i)}{\mathbb{V}(\widetilde{\eta}_i)}\widetilde{\eta}_i + (X_i(s) - \frac{\mathbb{E}(X_i(s)\widetilde{\eta}_i)}{\mathbb{V}(\widetilde{\eta}_i)}\widetilde{\eta}_i) = \frac{\mathbb{E}(X_i(s)\widetilde{\eta}_i)}{\mathbb{V}(\widetilde{\eta}_i)}\widetilde{\eta}_i + e_i(s), \tag{B.1}$$

where $e_i(s) = (X_i(s) - \widetilde{\eta}_i\mathbb{E}(X_i(s)\widetilde{\eta}_i)/\mathbb{V}(\widetilde{\eta}_i))$ with $\mathbb{E}(e_i(s)\widetilde{\eta}_i) = 0$ as well as $\mathbb{E}(e_i(s)) = 0$ for all $s \in [a, b]$. We then have, since by assumption $\mathbb{E}(e_i(s)g(\eta_i)) = 0$,

$$\mathbb{E}(X_i(s)g(\eta_i)) = \frac{\mathbb{E}(X_i(s)\widetilde{\eta}_i)}{\mathbb{V}(\widetilde{\eta}_i)}\mathbb{E}(\widetilde{\eta}_i g(\eta_i)).$$

Setting $c_0 := \frac{\mathbb{E}(\widetilde{\eta}g(\eta))}{\mathbb{V}(\widetilde{\eta})}$ we arrive at

$$\mathbb{E}(X_i(s)g(\eta_i)) = c_0\mathbb{E}(X_i(s)\widetilde{\eta}_i).$$

Since c_0 is independent of s, the assertion of Lemma 2.1 follows immediately. □

Remarks on Lemma 2.1

1. If X_i is assumed to be a Gaussian process, then also the distribution of $e_i(s) = (X_i(s) - \widetilde{\eta}_i\mathbb{E}(X_i(s)\widetilde{\eta}_i)/\mathbb{V}(\widetilde{\eta}_i))$ is Gaussian. Moreover, $\widetilde{\eta}_i$ and $e_i(s)$ are jointly normal distributed and, since $\mathbb{E}(e_i(s)\widetilde{\eta}_i) = 0$, the residual $e_i(s)$ is also independent of $\eta_i = \widetilde{\eta}_i + \alpha$ and we may conclude that $\mathbb{E}(e_i(s) \cdot g(\eta_i)) = 0$, i.e., the main assumption in Lemma 2.1 holds. One then may conclude that Lemma 2.1 holds under the additional moment conditions given in this Lemma.

2. It is important to note that the assertion of the lemma does not depend on the concrete form of η_i and hence will also hold if η_i contains the additional part $\int_a^b \beta(t)X_i(t)\,dt$, where it is assumed that $\beta(t) \in L^2([a,b])$ with $|\beta(t)| \leq M_\beta$ for some constant $M_\beta < \infty$.

Following the remark, in the proofs leading to Theorem 2.1, we will assume that η_i is given by $\eta_i = \alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r) + \int_a^b \beta(t)X_i(t)\,dt$, where $\beta(t) \in L^2([a,b])$ with $|\beta(t)| \leq M_\beta$ for some constant $M_\beta < \infty$. Theorem 2.1 may then be recovered by letting $\beta(t) \equiv 0$.

We now focus on the lemmas needed to proof Theorem 2.1. Under the moment condition given in Assumption 2.2 one may adapt Lemma 1 and Lemma 2 from Kneip et al. (2016b) to our setting:

Lemma B.1. *Under Assumption 2.2 there exist constants $0 < D_1 < \infty$ and $0 < D_2 < \infty$, such that for all n, all $0 < \delta < (b-a)/2$, all $t \in [a+\delta, b-\delta]$, all $0 < s \leq 1/2$ with $\delta^\kappa s^\kappa \geq s\delta^2$, and every $0 < z \leq \sqrt{n}$ we obtain*

$$P\Bigg(\sup_{t-s\delta\leq u\leq t+s\delta} \Big|\frac{1}{n}\sum_{i=1}^{n}[(Z_{\delta,i}(t)-Z_{\delta,i}(u))Y_i - \mathbb{E}((Z_{\delta,i}(t)-Z_{\delta,i}(u))Y_i)]\Big| \leq zD_1\sqrt{\frac{\delta^k s^\kappa}{n}}\Bigg) \geq 1-2\exp(-z^2) \tag{B.2}$$

and

$$P\Bigg(\sup_{t-s\delta\leq u\leq t+s\delta} \Big|\frac{1}{n}\sum_{i=1}^{n}[(Z_{\delta,i}(t)^2-Z_{\delta,i}(u)^2) - \mathbb{E}(Z_{\delta,i}(t)^2-Z_{\delta,i}(u)^2)]\Big| \leq zD_2\delta^\kappa\sqrt{\frac{s^\kappa}{n}}\Bigg) \geq 1-2\exp(-z^2). \tag{B.3}$$

Proof of Lemma B.1. Assertion (B.3) follows directly from Lemma 1 in Kneip et al. (2016b). For the proof of (B.2) we follow the notation of Lemma 1 in Kneip et al. (2016b) and define $Z^*_{\delta,i}(q) := \frac{1}{\sqrt{s^\kappa\delta^\kappa}}(Z_{\delta,i}(t+qs\delta)Y_i - \mathbb{E}(Z_{\delta,i}(t+qs\delta)Y_i))$ as well as $Z^*_\delta(q) := \frac{1}{\sqrt{n}}\sum_{i=1}^n Z^*_{\delta,i}(q)$. Note that Y_i is not assumed Gaussian anymore. However by bounding the absolute moments of $E(|Y_i|^{2m})$ by Assumption 2.2 one can easily verify that for $K = 4\sqrt{L_{1,3}|q_2-q_1|^{\min\{1,\kappa\}}\sigma_{|y|}}$, where the constant $0 < L_{1,3} < \infty$ is taken from by (C.7) in Kneip et al. (2016b), the Bernstein Condition

$$E(|Z^*_{\delta,i}(q_1) - Z^*_{\delta,i}(q_2)|^m) \leq \frac{m!}{2}K^{m-2}K^2$$

holds for all $0 < s \leq 0.5$, all integers $m \geq 2$ and all $q_1, q_2 \in [-1,1]$ and all $0 < \delta < (b-a)/2$.

An application of Corollary 1 in van de Geer and Lederer (2013) then guarantees that the Orlicz norm of $Z^*(q_1) - Z^*(q_2)$ is bounded, i.e. one has for all $q_1, q_2 \in [-1, 1]$

$$||Z^*_\delta(q_1) - Z^*_\delta(q_2)||_\Phi \leq L_{1,4}|q_1 - q_2|^{\min\{\frac{1}{2}, \frac{1}{2}\kappa\}}$$

for some constant $0 < L_{1,4} < \infty$. The assertion then follows again by the same arguments as given in Kneip et al. (2016b). □

A slightly more difficult task is to get an analogue of Lemma 2 in Kneip et al. (2016b). We derive

Lemma B.2. *Under the assumptions of Theorem 2.1 there exist constants* $0 < D_3 < D_4 < \infty$ *and* $0 < D_5 < \infty$ *such that*

$$0 < D_3\delta^\kappa \leq \inf_{t\in[a+\delta,b-\delta]} \mathbb{E}(Z_{\delta,i}(t)^2) \leq \sigma^2_{z,sup} := \sup_{t\in[a+\delta,b-\delta]} \mathbb{E}(Z_{\delta,i}(t)^2) \leq D_4\delta^\kappa \tag{B.4}$$

$$\lim_{n\to\infty} P\Big(\sup_{t\in[a+\delta,b-\delta]} |\frac{1}{n}\sum_{i=1}^n [Z_{\delta,i}(t)^2 - \mathbb{E}(Z_{\delta,i}(t)^2)]| \leq D_5\delta^\kappa \sqrt{\frac{1}{n}\log(\frac{b-a}{\delta})} \Big) = 1. \tag{B.5}$$

Moreover, there exist a constant $0 < D < \infty$ *such that for any* A^* *with* $D < A^* \leq A$ *we obtain as* $n \to \infty$*:*

$$\begin{aligned} P\Big(\sup_{t\in[a+\delta,b-\delta]} (\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(t)^2)^{-\frac{1}{2}} |\frac{1}{n}\sum_{i=1}^n (Z_{\delta,i}(t)Y_i - \mathbb{E}(Z_{\delta,i}(t)Y_i))| \\ \leq A^* \sqrt{\frac{\sigma^2_{|y|}}{n}\log(\frac{b-a}{\delta})} \Big) \to 1, \end{aligned} \tag{B.6}$$

$$\begin{aligned} P\Big(\sup_{t\in[a+\delta,b-\delta]} |\frac{1}{n}\sum_{i=1}^n (Z_{\delta,i}(t)Y_i - \mathbb{E}(Z_{\delta,i}(t)Y_i))| \\ \leq A^* \sqrt{\frac{\sigma^2_{|y|} D_4\delta^\kappa}{n}\log(\frac{b-a}{\delta})} \Big) \to 1. \end{aligned} \tag{B.7}$$

Proof of Lemma B.2. Again we can follow the proof and the notation given in Kneip et al. (2016b). Assertions (B.4) and (B.5) follow immediately from the proof of Lemma 2 in Kneip et al. (2016b) for any $\omega_2 > \omega_1 > 1$. In order to show (B.6) one can follow the proof given in Kneip et al. (2016b) until assertion (C.17).

It is then the crucial point to show that

$$\lim_{n\to\infty} P\left(\sup_{j\in\{2,3,\ldots,N_{\omega_1}\}} \frac{|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)Y_i - \mathbb{E}(Z_{\delta,i}(s_j)Y_i)|}{(\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)^2)^{\frac{1}{2}}} \leq A^* \sqrt{\frac{\sigma^2_{|y|}}{n}\log(\frac{b-a}{\delta})} \right) = 1.$$

Recall that it follows from (B.4) and (B.5) that with probability 1 (as $n \to \infty$) there exists a constant $0 < L_{2,1} < \infty$ such that

$$\inf_{u\in[a+\delta,b-\delta]} \frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(u)^2 \geq L_{2,1}\delta^\kappa.$$

Hence, because of an event which happens with probability converging to 1 (as $n \to \infty$) it is sufficient to show that

$$\sup_{j\in\{2,3,\ldots,N_{\omega_1}\}} \frac{|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)Y_i - \mathbb{E}(Z_{\delta,i}(s_j)Y_i)|}{(L_{2,1}\delta^\kappa)^{\frac{1}{2}}} \leq A^* \sqrt{\frac{\sigma^2_{|y|}}{n}\log(\frac{b-a}{\delta})}$$

holds with probability converging to 1 (as $n \to \infty$).

Remember that by (B.4) there exists a constant $0 < D_4 < \infty$ such that for all sufficiently small $\delta > 0$ we have $\sup_{t\in[a+\delta,b-\delta]} E(Z_{\delta,i}(t)^2) \leq D_4\delta^\kappa$. Chose an arbitrary point s_j and define

$$W_i(s_j) := \frac{1}{\sqrt{D_4\delta^\kappa\sigma^2_{|y|}}}(Z_{\delta,i}(s_j)Y_i - \mathbb{E}(Z_{\delta,i}(s_j)Y_i)),$$

then $E(W_i(s_j)) = 0$ and it is easy to show that under Assumption 2.2 with $K = 4$, a constant which is independent of s_j, $W_i(s_j)$ satisfies the Bernstein condition in Corollary 1 of van de Geer and Lederer (2013), i.e., we have for all $m = 2, 3, \ldots$:

$$\mathbb{E}(|W_i(s_j)|^m) \leq \frac{m!}{2}K^{m-2}K^2.$$

It immediately follows from an application of Corollary 1 in van de Geer and Lederer (2013) that there exists a constant $0 < L_3 < \infty$ such that the Orlicz-Norm $||\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i(s_j)||_\Psi$ can be bounded by $L_3 < \infty$. And hence we can infer that

$$\mathbb{E}\left(\exp(\frac{n}{6}(\sqrt{1 + 2\sqrt{\frac{6}{L_3^2 n}}|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i|} - 1)^2) \right) \leq 2.$$

It then follows from similar steps as in the proof of Lemma 1 in Kneip et al. (2016a) that there exists a constant $0 < L_4 < \infty$ such that for all $0 < z \leq \sqrt{n}$ we obtain

$$P\Bigg(|\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i(s_j)| > z L_4 \Bigg)$$
$$= P\Big(\frac{|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s_j)Y_i - E(Z_{\delta,i}(s_j)Y_i)|}{\sqrt{L_{2,1}\delta^{\kappa}}} > z L_4 \sqrt{\frac{D_4 \delta^{\kappa} \sigma^2_{|y|}}{n L_{2,1}\delta^{\kappa}}}\Big) \leq 2\exp(-z^2).$$

We may thus conclude that there then exists a constant $0 < L_5 < \infty$ such that

$$P\Big(\frac{|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s_j)Y_i - E(Z_{\delta,i}(s_j)Y_i)|}{\sqrt{L_{2,1}\delta^{\kappa}}} > z L_5 \sqrt{\frac{\sigma^2_{|y|}}{n}}\Big) \leq 2\exp(-z^2).$$

Finally, it follows from the union bound that

$$P\Bigg(\sup_{j\in\{2,3,\ldots,N_{\omega_1}\}} \frac{|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s_j)Y_i - E(Z_{\delta,i}(s_j)Y_i)|}{(L_{2,1}\delta^{\kappa})^{\frac{1}{2}}} \leq z L_5 \sqrt{\frac{\sigma^2_{|y|}}{n}}\Bigg)$$
$$\geq 1 - \sum_{j=1}^{N_{\omega_1}} P\Bigg(\frac{|\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(s_j)Y_i - E(Z_{\delta,i}(s_j)Y_i)|}{(L_{2,1}\delta^{\kappa})^{\frac{1}{2}}} > z L_5 \sqrt{\frac{\sigma^2_{|y|}}{n}}\Bigg)$$
$$\geq 1 - N_{\omega_1} 2\exp(-z^2)$$
$$\geq 1 - 2(\frac{b-a}{\delta})^{\omega_1}\exp(-z^2).$$

Setting $z = \sqrt{\omega_2 \log(\frac{b-a}{\delta})}$ for some $\omega_2 > \omega_1$ we then have, for sufficiently large n, $z \leq \sqrt{n}$ and

$$1 - 2(\frac{b-a}{\delta})^{\omega_1}\exp(-z^2) \geq 1 - 2(\frac{b-a}{\delta})^{\omega_1-\omega_2} \to 1.$$

There now obviously exists a constant D with $0 < \sqrt{\omega_2} L_5 = D < \infty$ for which assertion (B.6) will hold.

(As a side note we mention here that in the special case of a logistic regression one may set $D = 4\sqrt{\frac{D_4}{L_{2,1}}}$ and chose ω_1 and ω_2 such that $1 < \omega_1 < \omega_2 < \frac{A^*}{D}$.
Indeed, it is easy to show that in this case $Z_{\delta,i}(s_j)Y_i - \mathbb{E}(Z_{\delta,i}(s_j)Y_i)$ is sub-Gaussian with pa-

rameter σ of at most $2^{3/2}\sqrt{\mathbb{E}(Z_{\delta,i}(s_j)^2)}$. It then follows from the Hoeffding bound that for all j we have

$$P\left(\frac{|\frac{1}{n}\sum_{i=1}^n (Z_{\delta,i}(s_j)Y_i - \mathbb{E}(Z_{\delta,i}(s_j)Y_i))|}{\sqrt{L_{2,1}\delta^\kappa}} \geq z\frac{4}{\sqrt{n}}\sqrt{\frac{D_4}{L_{2,1}}}\right) \leq 2\exp(-z^2)$$

and assertion (B.6) follows again from the union bound while now setting $z = \omega_2\sqrt{\log(\frac{b-a}{\delta})}$.)

Finally, (B.7) now follows again from similar steps as in Kneip et al. (2016b). □

The difference to Lemma 2 in Kneip et al. (2016b) is that we don't have $D = \sqrt{2}$ anymore but $D = \sqrt{\omega_2}L_5$ for some constant $\omega_2 > 1$ and L_5. This is the price to pay for not assuming Gaussian Y_i.

Remarks to Lemma B.2 concerning the cut-off λ:

1. Using a slight abuse of notation, first note that there is a close connection between $\lambda = A\sqrt{\sigma^2_{|y|}\log(\frac{b-a}{\delta})/n}$ for some $A > D$ given in Theorem 2.1 and $\widetilde{\lambda} := A\sqrt{\sqrt{\mathbb{E}(Y_i^4)}\log(\frac{b-a}{\delta})/n}$ for $A = \sqrt{2\sqrt{3}}$ as used in our simulations. Indeed, set $\sigma^2_{|y|} = \mathbb{E}(Y^2)$. Jensen's inequality implies that there exists a constant $0 < \widetilde{D} \leq 1$ such that $\mathbb{E}(Y_i^2)\widetilde{D} = \sqrt{\mathbb{E}(Y_i^4)}$. We can therefore rewrite the expression for $\widetilde{\lambda}$ in the form of λ presented in Theorem 2.1 as $A\sqrt{\sigma^2_{|y|}\log(\frac{b-a}{\delta})/n}$ with $A = \sqrt{2\sqrt{3\widetilde{D}}}$.

We proceed to give more details about the motivation for cut-off used in the simulations:

2. Arguments for the applicability of the cut-off λ in the proof of Theorem 2.1 follow from Lemma B.2. The crucial step for determining an operable cut-off λ is to derive useful bounds on
$$\sup_{j\in\{2,3,\ldots,N_{\omega_1}\}} \frac{|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)Y_i - E(Z_{\delta,i}(s_j)Y_i)|}{(\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)^2)^{\frac{1}{2}}}.$$
Define $V_\delta(t) := (1/n\sum_{i=1}^n Z_{\delta,i}(t)Y_i - \mathbb{E}(Z_{\delta,i}(t)Y_i))/(1/n\sum_{i=1}^n Z_{\delta,i}(t)^2)^{1/2}$. It is then easy to see that under our assumptions $\sqrt{n}(1/n\sum_{i=1}^n Z_{\delta,i}(t)Y_i - \mathbb{E}(Z_{\delta,i}(t)Y_i))$ satisfies the Lyapunov conditions. We hence can conclude that $\sqrt{n}V_\delta(t)$ converges for all t in distribution to $N(0, \mathbb{V}(Z_{\delta,i}(t)Y_i)/\mathbb{E}(Z_{\delta,i}(t)^2))$, while at the same time the Cauchy-Schwarz inequality implies $\mathbb{V}(Z_{\delta,i}(t)Y_i)/\mathbb{E}(Z_{\delta,i}(t)^2) \leq \sqrt{3\,\mathbb{E}(Y_i^4)}$.
If the convergence to the normal distribution is sufficiently fast, using again the union

bound as in the proof of Lemma B.2 now together with an elementary bound on the tails of the normal distribution then leads to

$$P\left(\sup_{j\in\{2,3,\ldots,N_{\omega_1}\}}\frac{|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)Y_i - E(Z_{\delta,i}(s_j)Y_i)|}{(\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)^2)^{\frac{1}{2}}} \leq A^*\sqrt{\frac{\sqrt{\mathbb{E}(Y_i^4)}}{n}\log(\frac{b-a}{\delta})}\right)\to 1.$$

For some $A^* \geq \sqrt{2\sqrt{3}}$. The choice $A\sqrt{\sqrt{\mathbb{E}(Y_i^4)}\log(\frac{b-a}{\delta})/n}$ for some $A \geq \sqrt{2\sqrt{3}}$ for the cut-off would then be an immediate consequence.

Lemma 3 in Kneip et al. (2016b) remains unchanged and is repeated for convenience.

Lemma B.3. *Under the assumptions of Theorem 2.1 there exists a constant $0 < M_{sup} < \infty$ such that for all n, all $0 < \delta < (b-a)/2$ and every $t \in [a+\delta, b-\delta]$ we obtain*

$$\left|\mathbb{E}\left(Z_{\delta,i}(t)\int_a^b \beta(s)X_i(s)ds\right)\right| \leq M_{sup}\delta^{\min\{2,\kappa+1\}}. \tag{B.8}$$

Note that this Lemma is trivial in the case where $\beta(t) \equiv 0$.

Due to Lemma 2.1, we obtain a slightly modified version of Lemma 4 in Kneip et al. (2016b):

Lemma B.4. *Under the assumptions of Theorem 2.1 let $I_r := \{t \in [a,b] |\ |t-\tau_r| \leq \min_{s\neq r}|t-\tau_s|\}$, $r = 1,\ldots,S$.*

If $S > 0$, there then exist constants $0 < Q_1^* < \infty$ and $0 < Q_2 < \infty$ as well as $0 < c < \infty$ such that for all sufficiently small $\delta > 0$ and all $r = 1,\ldots,S$ we have with M_{sup}^*

$$|\mathbb{E}(Z_{\delta,i}(t)Y_i)| \leq Q_1^*\frac{\delta^2}{\max\{\delta,|t-\tau_r|\}^{2-\kappa}} + M_{sup}^*\delta^{\min\{2,\kappa+1\}} \quad \textit{for every } t \in I_r, \tag{B.9}$$

as well as

$$\sup_{t\in I_r,\ |t-\tau_r|\geq\frac{\delta}{2}} |\mathbb{E}(Z_{\delta,i}(t)Y_i)| \leq (1-Q_2)c|\beta_r|c(\tau_r)\delta^\kappa, \tag{B.10}$$

and for any $u \in [-0.5, 0.5]$

$$\begin{aligned}&|\mathbb{E}(Z_{\delta,i}(\tau_r)Y_i) - \mathbb{E}(Z_{\delta,i}(\tau_r+u\delta)Y_i)|\\ &\quad = |-c\beta_r c(\tau_r)\delta^\kappa\left(|u|^\kappa - \frac{1}{2}(|u+1|^\kappa - 1) - \frac{1}{2}(|u-1|^\kappa - 1)\right) + R_{5;r}(u)|,\end{aligned} \tag{B.11}$$

where $|R_{5;r}(u)| \leq \widetilde{M}_r||u|^{1/2}\delta|^{\min\{2\kappa,2\}}$ for some constants $\widetilde{M}_r < \infty$, $r = 1,\ldots,S$.

Proof of Lemma B.4. Lemma 2.1 guarantees us the existence of a constant c_0 such that

$$\mathbb{E}(Z_{\delta,i}(t)Y_i) = c_0\left(\int_a^b \beta(s)\mathbb{E}(Z_{\delta,i}(t)X_i(s))ds + \sum_{r=1}^{S}\beta_r X(\tau_r)\right)$$

The proof then follows immediately from the same steps as in Kneip et al. (2016b) for $Q_1^* = cQ_1$ and $M_{sup}^* = cM_{sup}$, where $c = |c_0|$. □

Proof of Theorem 2.1. By Lemma 2.1 we have for some constant $c_0 \neq 0$ with $c_0 < \infty$:

$$\begin{aligned}\mathbb{E}(Z_{\delta,i}(t)Y_i) &= \mathbb{E}(X_i(t)Y_i) - 0.5\mathbb{E}(X_i(t-\delta)Y_i) - 0.5\mathbb{E}(X_i(t+\delta)Y_i)\\ &= c_0 \cdot \mathbb{E}\left(Z_{\delta,i}(t)\left(\sum_{r=1}^{S}\beta_r X_i(\tau_r) + \int_a^b \beta(s)X_i(s)\,ds\right)\right),\end{aligned}$$

From this it is immediately seen that one has to simply adjust some of the constants appearing in the proof Theorem 4 in Kneip et al. (2016b). In particular with $c = |c_0|$ one has to exchange the term $|\beta_r|c(\tau_r)$ by $c|\beta_r|c(\tau_r)$ whenever it appears. Since c is a constant, which is independent of s, and the assertions in our Lemma B.1–B.4 correspond exactly to the assertions of Lemma 1–4 in Kneip et al. (2016b), the proof of the Theorem then follows by the same steps as given in the proof of Theorem 4 in Kneip et al. (2016b). □

Appendix C Proofs of the theoretical results from Section 2.3

In this appendix the proofs leading to our theoretical results concerning the parameter estimates as discussed in Section 2.3 are given.

We begin with the proof of Theorem 2.2. But instead of proofing Theorem 2.2 directly, we proof a more general statement for future references, allowing again for a functional part $\int_a^b \beta(t)X_i(t)\,dt$ in the linear predictor η_i to be present:

Theorem 2.4. *Let $g(\cdot)$ be invertible and assume that X_i satisfies Assumption 2.1. Then for all $S^* \geq S$, all $\alpha^*, \beta_1^*, \ldots, \beta_{S^*}^* \in \mathbb{R}$, 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*

$$\mathbb{E}\left(\left(g(\alpha + \sum_{r=1}^{S}\beta_r X_i(\tau_r) + \int_a^b \beta(t)X_i(t)\,dt) - g(\alpha^* + \sum_{r=1}^{S^*}\beta_r^* X_i(\tau_r) + \int_a^b \beta^*(t)X_i(t)\,dt)\right)^2\right) > 0, \tag{C.1}$$

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

Proof of Theorem 2.4. Since X_i satisfies Assumption 2.1, Theorem 3 in Kneip et al. (2016a) implies that the assumptions of Theorem 1 in Kneip et al. (2016a) are met. Since

$$\begin{aligned}
&\mathbb{E}\left(\left((\alpha + \sum_{r=1}^{S}\beta_r X_i(\tau_r) + \int_a^b \beta(t)X_i(t)\,dt) - (\alpha^* + \sum_{r=1}^{S^*}\beta_r^* X_i(\tau_r) + \int_a^b \beta^*(t)X_i(t)\,dt)\right)^2\right) \\
&= (\alpha-\alpha^*)^2 + \mathbb{E}\left(\left(\sum_{r=1}^{S}\beta_r X_i(\tau_r) + \int_a^b \beta(t)X_i(t)\,dt - \sum_{r=1}^{S^*}\beta_r^* X_i(\tau_r) - \int_a^b \beta^*(t)X_i(t)\,dt\right)^2\right)
\end{aligned}$$

It follows from Theorem 1 in Kneip et al. (2016a) that

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

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

Now suppose

$$\mathbb{E}\left(\left(g(\alpha + \sum_{r=1}^{S}\beta_r X_i(\tau_r) + \int_a^b \beta(t)X_i(t)\,dt) - g(\alpha^* + \sum_{r=1}^{S^*}\beta_r^* X_i(\tau_r) + \int_a^b \beta^*(t)X_i(t)\,dt)\right)^2\right) = 0.$$

It then follows that $g(\alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r) + \int_a^b \beta(t) X_i(t)\, dt)$ and $g(\alpha^* + \sum_{r=1}^{S^*} \beta_r^* X_i(\tau_r) + \int_a^b \beta^*(t) X_i(t)\, dt)$ must be identical, i.e.

$$P\left(g(\alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r) + \int_a^b \beta(t) X_i(t)\, dt) = g(\alpha^* + \sum_{r=1}^{S^*} \beta_r^* X_i(\tau_r) + \int_a^b \beta^*(t) X_i(t)\, dt) \right) = 1.$$

Since g is invertible we then have

$$P\left(g(\alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r) + \int_a^b \beta(t) X_i(t)\, dt) = g(\alpha^* + \sum_{r=1}^{S^*} \beta_r^* X_i(\tau_r) + \int_a^b \beta(t) X_i(t)\, dt) \right) = 1$$

if and only if

$$P\left((\alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r) + \int_a^b \beta(t) X_i(t)\, dt) = (\alpha^* + \sum_{r=1}^{S^*} \beta_r^* X_i(\tau_r) + \int_a^b \beta(t) X_i(t)\, dt) \right) = 1.$$

But by (C.2) we have

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

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

$$P\left((\alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r) + \int_a^b \beta(t) X_i(t)\, dt) = (\alpha^* + \sum_{r=1}^{S^*} \beta_r^* X_i(\tau_r) + \int_a^b \beta^*(t) X_i(t)\, dt) \right) < 1,$$

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

Theorem 2.2 now follows directly from Theorem 2.4 by setting $\beta(t) = \beta^*(t) \equiv 0$.

The following Propostion (C.1) is instrumental to derive rates of convergence for the system of estimated score equations $\widehat{\mathbf{U}}_n$ and their derivatives.

Proposition C.1. *Let $X_i = (X_i(t) : t \in [a, b])$, $i = 1, \ldots, n$ be i.i.d. Gaussian processes with covariance function $\sigma(s, t)$ satisfying Assumption 2.1. Let $\mathbb{E}(\varepsilon_i | X_i) = 0$ with $\mathbb{E}(\varepsilon_i^p | X_i) \leq M_\varepsilon < \infty$ for some even p with $p > \frac{2}{\kappa}$ and let $\widehat{\tau}_r$ enjoy the property given by (2.6), i.e. $|\widehat{\tau}_r - \tau_r| = O_p(n^{-\frac{1}{\kappa}})$. We then have for any differentiable bounded function $f : \mathbb{R} \to \mathbb{R}$ with $|f(x)| \leq M_f < \infty$, any*

$t^* \in [a,b]$, any linear predictor $\eta_i^* = \beta_0^* + \sum_{r=1}^{S^*} \beta_r^* X_i(t_r^*)$, where $t_r^* \in [a,b]$, $\beta_r^* \in \mathbf{R}$ and S^* are arbitrary and any $r = 1,\ldots,S$:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i(\widehat{\tau}_r) - X_i(\tau_r))^2 = O_p(n^{-1}) \tag{C.3}$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i(\widehat{\tau}_r) - X_i(\tau_r))f(\eta_i^*) = O_p(n^{-\min\{1,\frac{1}{\kappa}\}}) \tag{C.4}$$

$$\frac{1}{n}\sum_{i=1}^{n}X_i(t^*)(X_i(\widehat{\tau}_r) - X_i(\tau_r))f(\eta_i^*) = O_p(n^{-\min\{1,\frac{1}{\kappa}\}}) \tag{C.5}$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i(\widehat{\tau}_r) - X_i(\tau_r))\varepsilon_i f(\eta_i^*) = O_p(n^{-1}) \tag{C.6}$$

$$\frac{1}{n}\sum_{i=1}^{n}X_i(t^*)(X_i(\widehat{\tau}_r) - X_i(\tau_r))\varepsilon_i f(\eta_i^*) = O_p(n^{-1}) \tag{C.7}$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i(\widehat{\tau}_r) - X_i(\tau_r))^4 = O_p(n^{-2}) \tag{C.8}$$

Proof of Proposition C.1. Before the different assertions are proven, note that it follows from a Taylor expansion that under Assumption (2.1) there exists a constant $0 < L_{1,1} < \infty$ such that for all sufficiently small $0 < s$, all $q \in [-1,1]$, all $t \in [a+s, b-s]$ and all $t^* \in [a,b]$ we have

$$\begin{aligned}|\mathbb{E}((X_i(t+qs) - X_i(t))X_i(t^*))| &= |\omega(t+qs, t^*, |t+qs-t^*|^\kappa) - \omega(t,t^*,|t-t^*|^\kappa)| \\ &\leq L_{1,1}|qs|^{\min\{1,\kappa\}}.\end{aligned} \tag{C.9}$$

On the other hand, recall that (C.44) in Kneip et al. (2016b) implies that there exists a constant $0 < L_{1,2} < \infty$ and $0 < L_{1,3} < \infty$ such that for all sufficiently small $0 < s$ and all $q_1, q_2 \in [-1,1]$ we have

$$\begin{aligned}\sigma^2_{(X_i(\tau_r+sq_1) - X_i(\tau_r+sq_2))} &= \mathbb{E}\Big(((X_i(\tau_r+sq_1) - X_i(\tau_r+sq_2)))^2\Big) \\ &\leq L_{1,2}|q_1-q_2|^\kappa s^\kappa \leq L_{1,3}s^\kappa.\end{aligned} \tag{C.10}$$

Moreover recall that Lemma 2.1 and its proof in particular imply that for any bivariate normal random variables (X_1, X_2) we have

$$cov(f(X_1), X_2) = \frac{cov(f(X_1), X_1)}{Var(X_1)} cov(X_1, X_2),$$

where by Stein's Lemma (C. M. Stein (1981)) $\frac{cov(f(X_1),X_1)}{Var(X_1)} = \mathbb{E}(f'(X_1))$ provided f is differentiable and $\mathbb{E}(|f'(X_1)|) < \infty$; see also Lemma 1 in Brillinger (2012a) for a more precise statement.

We are now equipped with the tools to proof the different assertions of the proposition. Assertion (C.3) follows from Proposition 2 in Kneip et al. (2016a). In order to proof Assertion (C.4), choose any $0 < s$ sufficiently small and define for $q_1, q_2 \in [-1,1]$

$$\begin{aligned}\chi_i(q_1,q_2) &:= (X_i(\tau_r + sq_1) - X_i(\tau_r))f(\eta_i^*) - (X_i(\tau_r + sq_2) - X_i(\tau_r))f(\eta_i^*) \\ &\quad - \mathbb{E}\Big((X_i(\tau_r + sq_1) - X_i(\tau_r))f(\eta_i^*) - (X_i(\tau_r + sq_2) - X_i(\tau_r))f(\eta_i^*)\Big) \\ &= (X_i(\tau_r + sq_1) - X_i(\tau_r + sq_2))f(\eta_i^*) - \mathbb{E}((X_i(\tau_r + sq_1) - X_i(\tau_r + sq_2))f(\eta_i^*)).\end{aligned}$$

We then have $\mathbb{E}(\chi_i(q_1,q_2)) = 0$ and it follows from some straightforward calculations, since $|f(\eta_i^*)| \leq M_f$, that there exists a constant $0 < L_{1,4} < \infty$ such that for $m = 2,3,\dots$ we have

$$\mathbb{E}(|\frac{1}{s^{\frac{\kappa}{2}}}\chi_i(q_1,q_2)|^m) \leq \frac{m!}{2}(L_{1,4}|q_1 - q_2|^{\frac{\kappa}{2}})^m. \tag{C.11}$$

Corollary 1 in van de Geer and Lederer (2013) now guarantees that there exists a constant $0 < L_{1,5} < \infty$ such that the Orlicz norm of $\frac{1}{\sqrt{ns^\kappa}}\sum_{i=1}^n(\chi_i(q_1,q_2))$ can be bounded, i.e., we have for some $0 < L_{1,5} < \infty$:

$$||\frac{1}{\sqrt{ns^\kappa}}\sum_{i=1}^n \chi_i(q_1,q_2)||_\Psi \leq L_{1,5}|q_1 - q_2|^{\frac{\kappa}{2}}. \tag{C.12}$$

By (C.12) one may apply Theorem 2.2.4 of van der Vaart and Wellner (1996). The covering integral in this theorem can easily be seen to be finite and one can thus infer that there exists a constant $0 < L_{1,6} < \infty$ such that

$$\mathbb{E}\Bigg(\exp\Big(\sup_{q_1,q_2\in[-1,1]} n/6\Big(\sqrt{1 + 2\sqrt{\frac{6}{nL_{1,6}^2}|\frac{1}{\sqrt{ns^\kappa}}\sum_{i=1}^n \chi_i(q_1,q_2)| - 1}}\Big)^2\Big)\Bigg) \leq 2.$$

For every $x > 0$, the Markov inequality then yields

$$P\Bigg(\sup_{q_1,q_2\in[-1,1]} |\frac{1}{\sqrt{ns^\kappa}}\sum_{i=1}^n \chi_i(q_1,q_2)| \geq x\frac{L_{1,6}}{2\sqrt{6}}\Bigg) \leq 2\exp\Big(-\frac{n}{6}(\sqrt{1 + x/\sqrt{n}} - 1)^2\Big).$$

Improving the readability, it then follows from a Taylor expansion of $\frac{n}{6}(\sqrt{1+x/\sqrt{n}}-1)^2$ that we may conclude that there exists a constant $0 < L_{1,7} < \infty$ such that for all $0 < x \leq \sqrt{n}$ we have

$$P\left(\sup_{q_1,q_2\in[-1,1]} |\frac{1}{\sqrt{n}s^{\kappa}} \sum_{i=1}^{n} \chi_i(q_1,q_2)| < L_{1,7}x \right) \geq 1-2\exp(-x^2). \tag{C.13}$$

Now, note that it follows from the proof of Lemma 2.1 that there exists a constant $|c_0| < \infty$, not depending on t^*, such that $\mathbb{E}(X(t^*)f(\eta_i^*)) = c_0\,\mathbb{E}(X(t^*)\eta_i^*)$ for all $t^* \in [a,b]$. Together with (C.9) we can therefore conclude that there exists a constant $0 \leq L_{1,8} < \infty$ such that for all $q_1 \in [-1,1]$:

$$|\mathbb{E}((X_i(\tau_r+sq_1)-X_i(\tau_r))f(\eta_i^*))| \leq L_{1,8}s^{\min\{1,\kappa\}} \tag{C.14}$$

Using (C.14) together with (C.13) we can conclude that for all $0 < x \leq \sqrt{n}$ we have:

$$\begin{aligned} P\left(\sup_{\tau_r-s\leq u_r\leq\tau_r+s} |\frac{1}{n}\sum_{i=1}^{n}(X_i(u_r)-X_i(\tau_r))f(\eta_i^*)| < L_{1,8}s^{\min\{1,\kappa\}} + L_{1,7}\frac{s^{\frac{\kappa}{2}}}{\sqrt{n}}x \right) \\ \geq 1-2\exp(-x^2) \end{aligned} \tag{C.15}$$

Assertion (C.4) then follows immediately from (2.6).
By the boundedness of f, the proof of (C.5) proceeds similar, but one now has to bound

$$|\mathbb{E}(X_i(t^*)(X_i(\tau_r+sq_1)-X_i(\tau_r))f(\eta_i^*))|.$$

For $X_i(t^*) = \eta_i^*$, Lemma 2.1 together with (C.9) already implies that there exists a constant L such that $|\mathbb{E}(X_i(t^*)(X_i(\tau_r+sq_1)-X_i(\tau_r))f(\eta_i^*))| \leq Ls^{\min\{1,\kappa\}}$. Let $X_i(t^*) \neq \eta_i^*$. Note that $(X_i(t^*),(X_i(\tau_r+sq_1)-X_i(\tau_r)),\eta_i^*)$ are multivariate normal. Hence also the conditional distribution of $((X_i(\tau_r+sq_1)-X_i(\tau_r)),\eta_i^*)$ given $X_i(t^*)$ is multivariate normal. To ease the notation set $X_1 = \eta_i^*$, $X_2 = (X_i(\tau_r+sq_1)-X_i(\tau_r))$ and $X_3 = X_i(t^*)$ and define by $\sigma_{i,j}$, $i,j \in \{1,2,3\}$ their associated covariance and variances. We then have by conditional expectation together with an application of Lemma 2.1 (c.f Brillinger (2012a, Lemma 1))

$$\begin{aligned} |\mathbb{E}(X_i(t^*)(X_i(\tau_r+sq_1)-X_i(\tau_r))f(\eta_i^*))| &= |\mathbb{E}(f(X_1)X_2X_3)| = |\mathbb{E}(X_3\mathbb{E}(f(X_1)X_2|X_3))| \\ &= |(\sigma_{12}-\frac{\sigma_{13}\sigma_{23}}{\sigma_{33}})\mathbb{E}(\frac{cov(f(X_1),X_1|X_3)}{\mathbf{V}(X_1|X_3)}X_3) + \frac{\sigma_{23}}{\sigma_{33}}\mathbb{E}(X_3^2\mathbb{E}(f(X_1)|X_3))|. \end{aligned}$$

Using (C.9) it is then easy to see that there exists a constant $0 < L_{1,9} < \infty$ such that $|\sigma_{12}| \leq L_{1,9}s^{\min\{1,\kappa\}}$, as well as $|\sigma_{23}| \leq L_{1,9}s^{\min\{1,\kappa\}}$. On the other hand Assumption 2.1 implies that there exists a constant $0 < L_{1,10} < \infty$ such that $|\sigma_{13}| \leq L_{1,10}$. Note that $cov(f(X_1),X_1|X_3) =$

$\mathbb{E}(f(X_1)X_1|X_3) - \mathbb{E}(f(X_1)|X_3)\,\mathbb{E}(X_1|X_3)$ and $\mathbf{V}(X_1|X_3) = \sigma_{11} - \frac{\sigma_{13}^2}{\sigma_{33}} > 0$. Moreover note that if f is assumed to be differentiable and $\mathbb{E}(|f'(X_1)||X_3) < \infty$, it follows from and Stein's Lemma that $cov(f(X_1), X_1|X_3)/\mathbf{V}(X_1|X_3)$ can be substituted by $\mathbb{E}(f'(X_1)|X_3)$.

Since f is bounded it then follows immediately that for all linear predictors η_i^* and all $t^* \in [a,b]$ there exists a constant $0 < L_{1,11} < \infty$ such that for all $q_1 \in [-1,1]$ and all sufficiently small s and all $r = 1,\ldots,S$ we have:

$$|\mathbb{E}(X_i(t^*)(X_i(\tau_r + sq_1) - X_i(\tau_r))f(\eta_i^*))| \leq L_{1,11}s^{\min\{1,\kappa\}}. \tag{C.16}$$

By (C.16) one can conclude similar to (C.15) that for all $0 < x \leq \sqrt{n}$ and for some constant $L_{1,12} < \infty$

$$P\left(\sup_{\tau_r - s \leq u \leq \tau_r + s} |\frac{1}{n}\sum_{i=1}^{n} X_i(t^*)(X_i(u) - X_i(\tau_r))f(\eta_i^*)| < s^{\min\{1,\kappa\}}L_{1,11} + L_{1,12}\frac{s^{\frac{\kappa}{2}}}{\sqrt{n}}x\right) \geq 1 - 2\exp(-x^2).$$

Assertion (C.5) then follows again immediately from (2.6).
In order to show assertion (C.6) we make use the Orlicz-norm $||X||_p$.
Choose some $p > \frac{2}{\kappa} = p_\kappa$, and let p be even. Note that $\mathbb{E}((X_i(\tau_r + sq_1) - X_i(\tau_r + sq_2))\varepsilon_i f(\eta_i^*)) = 0$. For all sufficiently small $0 < s$ and all $q_1, q_2 \in [-1,1]$ it is easy to show that there exists a constant $L_{1,13} < \infty$ such that

$$\mathbb{E}(|s^{-\kappa/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i(\tau_r + sq_1) - X_i(\tau_r + sq_2))\varepsilon_i f(\eta_i^*)|^p) \leq L_{1,13}^p |q_1 - q_2|^{\frac{p\kappa}{2}}.$$

We may conclude

$$||s^{-\kappa/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i(\tau_r + sq_1) - X_i(\tau_r + sq_2))\varepsilon_i f(\eta_i^*)||_p \leq L_{1,13}|q_1 - q_2|^{\frac{\kappa}{2}}. \tag{C.17}$$

By assertion (C.17) one may apply Theorem 2.2.4 in van der Vaart and Wellner (1996). Our condition on p ensures that the covering integral appearing in this theorem is finite. The maximum inequalities of empirical processes then imply:

$$||\sup_{q_1,q_2 \in [-1,1]} |s^{-\kappa/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i(\tau_r + sq_1) - X_i(\tau_r + sq_2))\varepsilon_i f(\eta_i^*)|||_{\Psi_p} \leq L_{1,14}$$

for some constant $L_{1,14} < \infty$. At the same time, the Markov inequality implies

$$P\left(\sup_{\tau_r - s \leq u \leq \tau_r + s} |\frac{1}{n} \sum_{i=1}^{n} (X_i(u) - X_i(\tau_r)) \varepsilon_i f(\eta_i^*)| > s^{\kappa/2} \frac{x}{\sqrt{n}} \right)$$
$$\leq P\left(| \sup_{q_1, q_2 \in [-1,1]} |s^{-\frac{\kappa}{2}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i(\tau_r + sq_1) - X_i(\tau_r + sq_2)) \varepsilon_i f(\eta_i^*)||^p > x^p \right) \leq \frac{L_{1,14}^p}{x^p}.$$

Assertion (C.6) then follows from (2.6) and our conditions on p. Moreover, assertion (C.7) follows from exactly the same steps.

It remains to proof (C.8). For real numbers x and y it obviously holds that $x^4 - y^4 = (x - y)(x + y)(x^2 + y^2)$. With the help of this decomposition and (C.10) it is easy to see that there exists a constant $L_{1,15}$ such that for all $p \geq 1$ for all sufficiently small s and $q_1, q_2 \in [-1, 1]$ and all $p \geq 1$ we now have

$$\mathbb{E}\left(|s^{-2\kappa} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i(\tau_r + sq_1) - X_i(\tau_r))^4 - (X_i(\tau_r + sq_2) - X_i(\tau_r))^4|^p \right) \leq L_{1,15}^p |q_1 - q_2|^{\frac{p\kappa}{2}}. \tag{C.18}$$

At the same time (C.10) implies that there exists a constant $L_{1,16} < \infty$ such that $|\mathbb{E}((X_i(\tau_r + sq_1) - X_i(\tau_r))^4)| \leq L_{1,16} S^{2\kappa}$. Choose some $p > \frac{2}{\kappa}$, by (C.18) and with the help of another application of the maximum inequalities for empirical processes we can then conclude that there exists a constant $L_{1,17} < \infty$ such that

$$P\left(\sup_{\tau_r - s \leq u \leq \tau_r + s} |\frac{1}{n} \sum_{i=1}^{n} (X_i(u) - X_i(\tau_r))^4| > L_{1,16} s^{2\kappa} + L_{1,17} \frac{s^{2\kappa}}{\sqrt{n}} x \right) \leq \frac{L_{1,17}^p}{x^p},$$

Assertion (C.8) then follows once more from (2.6). □

For the following proofs we introduce some additional notation. Let $h(x) = g'(x)/\sigma^2(g(x))$ and note that differentiating the estimation equation

$$\frac{1}{n} \widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = \frac{1}{n} \widehat{\mathbf{D}}_n(\boldsymbol{\beta}) \widehat{\mathbf{V}}_n^{-1}(\boldsymbol{\beta})(y - \boldsymbol{\mu}(\boldsymbol{\beta})) = \frac{1}{n} \sum_{i=1}^{n} h(\widehat{\eta}_i(\boldsymbol{\beta})) \widehat{\mathbf{X}}_i (y_i - g(\widehat{\eta}_i))$$

leads to

$$\frac{1}{n} \widehat{\mathbf{H}}(\boldsymbol{\beta}) = \frac{1}{n} \frac{\partial \widehat{\mathbf{U}}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{1}{n} \widehat{\mathbf{D}}_n(\boldsymbol{\beta})^T \widehat{\mathbf{V}}_n(\boldsymbol{\beta})^{-1} \widehat{\mathbf{D}}_n(\boldsymbol{\beta}) + \frac{1}{n} \sum_{i=1}^{n} h'(\widehat{\eta}_i) \widehat{\mathbf{X}}_i \widehat{\mathbf{X}}_i^T (y_i - g(\widehat{\eta}_i(\boldsymbol{\beta})))$$
$$= -\frac{1}{n} \widehat{\mathbf{F}}_n(\boldsymbol{\beta}) + \frac{1}{n} \widehat{\mathbf{R}}_n(\boldsymbol{\beta}) \quad \text{say.}$$

In a similar manner one obtains by replacing the estimates $\widehat{\tau}_r$ with their true counterparts τ_r:

$$\frac{1}{n}\mathbf{H}(\boldsymbol{\beta}) = \frac{1}{n}\frac{\partial\,\mathbf{U}_n(\boldsymbol{\beta})}{\partial\,\boldsymbol{\beta}} = -\frac{1}{n}\mathbf{F}_n(\boldsymbol{\beta}) + \frac{1}{n}\mathbf{R}_n(\boldsymbol{\beta}),$$

where

$$\frac{1}{n}\mathbf{F}_n(\boldsymbol{\beta}) = \frac{1}{n}\mathbf{D}_n(\boldsymbol{\beta})^T\,\mathbf{V}_n(\boldsymbol{\beta})^{-1}\,\mathbf{D}_n(\boldsymbol{\beta}),$$

and

$$\frac{1}{n}\mathbf{R}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} h'(\eta_i(\boldsymbol{\beta}))\,\mathbf{X}_i\,\mathbf{X}_i^T(y_i - g(\eta_i(\boldsymbol{\beta}))).$$

Now, let $\widehat{\eta}(\boldsymbol{\beta})$, $\widehat{\mathbf{X}}$ and y be generic copies of $\widehat{\eta}_i(\boldsymbol{\beta})$, $\widehat{\mathbf{X}}_i$ and y_i. We then have

$$\mathbb{E}(\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta})) = \mathbb{E}(\frac{g'(\widehat{\eta}(\boldsymbol{\beta}))^2}{\sigma^2(g(\widehat{\eta}(\boldsymbol{\beta}))}\widehat{\mathbf{X}}\widehat{\mathbf{X}}^T) =: \mathbb{E}(\widehat{\mathbf{F}}(\boldsymbol{\beta})),$$

as well as

$$\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta}) = \mathbb{E}(h'(\widehat{\eta}(\boldsymbol{\beta}))\widehat{\mathbf{X}}\widehat{\mathbf{X}}^T(y - g(\widehat{\eta}(\boldsymbol{\beta})))) =: \mathbb{E}(\widehat{\mathbf{R}}(\boldsymbol{\beta})).$$

In a similar manner $\mathbb{E}(\mathbf{F}(\boldsymbol{\beta})) = \mathbb{E}(n^{-1}\,\mathbf{F}_n(\boldsymbol{\beta}))$ and $\mathbb{E}(\mathbf{R}(\boldsymbol{\beta})) = \mathbb{E}(n^{-1}\,\mathbf{R}_n(\boldsymbol{\beta}))$ are defined.

The next proposition is crucial, as it tells us that the estimated score function and its derivative are sufficiently close to each other. Of particular importance are the facts that

$$\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0) = \frac{1}{n}\mathbf{U}_n(\boldsymbol{\beta}_0) + o_p(n^{-\frac{1}{2}}),$$

and

$$\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0) = \frac{1}{n}\mathbf{F}_n(\boldsymbol{\beta}_0) + O_p(n^{-\frac{1}{2}}),$$

which follow from this proposition.

Proposition C.2. *Let $X_i = (X_i(t) : t \in [a,b])$, $i = 1,...,n$ be i.i.d. Gaussian processes. Under Assumption 2.3 and under the results of Proposition C.1 we have*

$$\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0) = \frac{1}{n}\mathbf{U}_n(\boldsymbol{\beta}_0) + O_p(n^{-\min\{1,1/\kappa\}}). \tag{C.19}$$

Additionally, for all $\boldsymbol{\beta} \in \mathbf{R}^{S+1}$*:*

$$\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = \frac{1}{n}\mathbf{U}_n(\boldsymbol{\beta}) + O_p(n^{-\frac{1}{2}}), \tag{C.20}$$

$$\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}) = \frac{1}{n}\mathbf{F}_n(\boldsymbol{\beta}) + O_p(n^{-1/2}), \tag{C.21}$$

$$\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta}) = \frac{1}{n}\mathbf{R}_n(\boldsymbol{\beta}) + O_p(n^{-\frac{1}{2}}). \tag{C.22}$$

Moreover, we have

$$\mathbb{E}(\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta})) \to \mathbb{E}(\frac{1}{n}\mathbf{U}_n(\boldsymbol{\beta})), \tag{C.23}$$

$$\mathbb{E}(\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta})) \to \mathbb{E}(\frac{1}{n}\mathbf{F}_n(\boldsymbol{\beta})), \tag{C.24}$$

$$\mathbb{E}(\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta})) \to \mathbb{E}(\frac{1}{n}\mathbf{R}_n(\boldsymbol{\beta})). \tag{C.25}$$

Particularly,

$$\mathbb{E}(\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta}_0)) \to 0 \tag{C.26}$$

and

$$\mathbb{E}(\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0)) \to 0. \tag{C.27}$$

Proof of Proposition C.2. To ease notation we use $\boldsymbol{\beta}_0 = (\beta_0^{(0)}, \beta_1^{(0)}, \ldots, \beta_S^{(0)})^T$ to denote the true parameter vector. For instance, the intercept is given by $\beta_0^{(0)}$, while $\beta_r^{(0)}$ is the coefficient for the rth point of impact. Similar we denote the entries of $\boldsymbol{\beta}$ by $(\beta_0, \ldots, \beta_S)$. Write

$$\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = \frac{1}{n}\mathbf{U}_n(\boldsymbol{\beta}) + \mathbf{Rest}_n(\boldsymbol{\beta}), \tag{C.28}$$

then $\mathbf{Rest}_n(\boldsymbol{\beta})$ can be decomposed into two parts:

$$\begin{aligned}\mathbf{Rest}_n(\boldsymbol{\beta}) &= \frac{1}{n}(\widehat{\mathbf{D}}_n^T(\boldsymbol{\beta})\widehat{\mathbf{V}}_n^{-1}(\boldsymbol{\beta}) - \mathbf{D}_n^T(\boldsymbol{\beta})\mathbf{V}_n^{-1}(\boldsymbol{\beta}))(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) - \frac{1}{n}\widehat{\mathbf{D}}_n^T(\boldsymbol{\beta})\widehat{\mathbf{V}}_n^{-1}(\boldsymbol{\beta})(\widehat{\boldsymbol{\mu}}_n(\boldsymbol{\beta}) - \boldsymbol{\mu}_n(\boldsymbol{\beta})) \\ &= \mathbf{Rest}_1(\boldsymbol{\beta}) + \mathbf{Rest}_2(\boldsymbol{\beta}), \quad \text{say.}\end{aligned} \tag{C.29}$$

The first summand $\mathbf{Rest}_1(\boldsymbol{\beta})$ is given by:

$$\mathbf{Rest}_1(\boldsymbol{\beta}) = \frac{1}{n}(\widehat{\mathbf{D}}_n^T(\boldsymbol{\beta})\widehat{\mathbf{V}}_n^{-1}(\boldsymbol{\beta}) - \mathbf{D}_n^T(\boldsymbol{\beta})\mathbf{V}_n^{-1}(\boldsymbol{\beta}))(\mathbf{Y}_n - \boldsymbol{\mu}_n(\boldsymbol{\beta})).$$

The jth equation of $\mathbf{Rest}_1(\boldsymbol{\beta})$ can be written as

$$\begin{aligned}
Rest_{j,1}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))}\widehat{X}_{ij} - \frac{g'(\eta_i(\boldsymbol{\beta}))}{\sigma^2(g(\eta_i(\boldsymbol{\beta})))}X_{ij}\Big)(y_i - g(\eta_i(\boldsymbol{\beta})))\\
&= \frac{1}{n}\sum_{i=1}^{n}X_{ij}\Big(\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))} - \frac{g'(\eta_i(\boldsymbol{\beta}))}{\sigma^2(g(\eta_i(\boldsymbol{\beta})))}\Big)(y_i - g(\eta_i(\boldsymbol{\beta})))\\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(\widehat{X}_{ij} - X_{ij})\Big(\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))}\Big)(y_i - g(\eta_i(\boldsymbol{\beta})))\\
&= R_{j,1,a}(\boldsymbol{\beta}) + R_{j,1,b}(\boldsymbol{\beta}), \quad \text{say.}
\end{aligned} \tag{C.30}$$

With $h(x) = g'(x)/\sigma^2(g(x))$, a Taylor expansion implies the existence of some some $\xi_{i,1}$ between $\widehat{\eta}_i(\boldsymbol{\beta})$ and $\eta_i(\boldsymbol{\beta})$ such that for $\boldsymbol{\beta} = \boldsymbol{\beta}_0$

$$\begin{aligned}
R_{j,1,a}(\boldsymbol{\beta}_0) &= \frac{1}{n}\sum_{i=1}^{n}X_{ij}\Big(\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}_0))}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta}_0)))} - \frac{g'(\eta_i(\boldsymbol{\beta}_0))}{\sigma^2(g(\eta_i(\boldsymbol{\beta}_0)))}\Big)(y_i - g(\eta_i(\boldsymbol{\beta}_0)))\\
&= \sum_{r=2}^{S+1}\beta_{r-1}^{(0)}\frac{1}{n}\sum_{i=1}^{n}X_{ij}(\widehat{X}_{ir} - X_{ir})\varepsilon_i h'(\eta_i(\boldsymbol{\beta}_0))\\
&\quad + \frac{1}{n}\sum_{i=1}^{n}X_{ij}\varepsilon_i h''(\xi_{i,1})/2\Big(\sum_{l=2}^{S+1}\beta_{l-1}^{(0)}(X_{il} - \widehat{X}_{il})\Big)^2.
\end{aligned}$$

Since $|h'(\cdot)| \leq M_h$ and $|h''(\cdot)| \leq M_h$, $R_{j,1,a}(\boldsymbol{\beta}_0) = O_p(n^{-1})$ for $j = 1,\ldots,S+1$ follows immediately from (C.6) and (C.7) together with the Cauchy-Schwarz inequalitiy and (C.8). At the same time it follows from similar arguments that for all $j = 1,\ldots,S+1$ we have $R_{j,1,b}(\boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{X}_{ij} - X_{ij})h(\widehat{\eta}_i(\boldsymbol{\beta}_0))\varepsilon_i = O_p(n^{-1})$. The above arguments then imply:

$$\mathbf{Rest}_1(\boldsymbol{\beta}_0) = O_p(n^{-1}). \tag{C.31}$$

The jth equation of $\mathbf{Rest}_2(\boldsymbol{\beta})$ can be written as $Rest_{j,2}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}h(\widehat{\eta}_i(\boldsymbol{\beta}))\widehat{X}_{ij}(g(\eta_i(\boldsymbol{\beta})) - g(\widehat{\eta}_i(\boldsymbol{\beta})))$. Using again Taylor expansions together with assertions (C.4), (C.5) as well as the Cauchy-Schwarz inequality together with (C.8), can now be used to conclude that or all $\boldsymbol{\beta}$ and $j = 1,\ldots,S+1$ we have

$$Rest_{j,2}(\boldsymbol{\beta}) = O_p(n^{-\min\{1,1/\kappa\}}). \tag{C.32}$$

Assertion (C.19) then follows from (C.30), (C.31) and (C.32). Note that our assumptions in particular imply that $\mathbf{Rest}_1(\boldsymbol{\beta}_0)$ and $\mathbf{Rest}_2(\boldsymbol{\beta}_0)$ are uniform integrable. Additional to (C.19), we thus have $\mathbb{E}(\mathbf{Rest}_n(\boldsymbol{\beta}_0)) \to \mathbf{0}$ implying (C.27), $\mathbb{E}(\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0)/n) \to \mathbf{0}$, since $\mathbb{E}(\mathbf{U}_n(\boldsymbol{\beta}_0)/n) = 0$.

In order to proof assertion (C.20) suppose $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ and note that we still have (C.32). However, $\mathbf{Rest}_1(\boldsymbol{\beta})$ needs a closer investigation. Its jth row can be written as

$$\begin{aligned}
Rest_{j,1}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))}\widehat{X}_{ij} - \frac{g'(\eta_i(\boldsymbol{\beta}))}{\sigma^2(g(\eta_i(\boldsymbol{\beta})))}X_{ij}\Big)(y_i - g(\eta_i(\boldsymbol{\beta})))\\
&= \frac{1}{n}\sum_{i=1}^{n}X_{ij}(h(\widehat{\eta}_i(\boldsymbol{\beta})) - h(\eta_i(\boldsymbol{\beta})))(y_i - g(\eta_i(\boldsymbol{\beta}_0)))\\
&\quad - \frac{1}{n}\sum_{i=1}^{n}X_{ij}(h(\widehat{\eta}_i(\boldsymbol{\beta})) - h(\eta_i(\boldsymbol{\beta})))(g(\eta_i(\boldsymbol{\beta})) - g(\eta_i(\boldsymbol{\beta}_0)))\\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(\widehat{X}_{ij} - X_{ij})h(\widehat{\eta}_i(\boldsymbol{\beta}))(y_i - g(\eta_i(\boldsymbol{\beta}_0)))\\
&\quad - \frac{1}{n}\sum_{i=1}^{n}(\widehat{X}_{ij} - X_{ij})h(\widehat{\eta}_i(\boldsymbol{\beta}))(g(\eta_i(\boldsymbol{\beta})) - g(\eta_i(\boldsymbol{\beta}_0))).
\end{aligned}$$

To obtain (C.20) it is sufficient to use some rather conservative inequalities of each of the appearing terms. For instance, another Taylor expansion together with the Cauchy-Schwarz inequality and (C.3) now yield

$$\frac{1}{n}\sum_{i=1}^{n}X_{ij}(h(\widehat{\eta}_i(\boldsymbol{\beta})) - h(\eta_i(\boldsymbol{\beta})))(y_i - g(\eta_i(\boldsymbol{\beta}_0))) = O_p(n^{-\frac{1}{2}}). \tag{C.33}$$

While the Cauchy-Schwarz inequalitiy together with (C.3) yields

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{X}_{ij} - X_{ij})h(\widehat{\eta}_i(\boldsymbol{\beta}))(y_i - g(\eta_i(\boldsymbol{\beta}_0))) = O_p(n^{-\frac{1}{2}}). \tag{C.34}$$

It follows from additional Taylor expansions that there exists a $\xi_{i,2}$ between $\widehat{\eta}_i(\boldsymbol{\beta})$ and $\eta_i(\boldsymbol{\beta})$ as well as some $\xi_{i,3}$ between $\eta_i(\boldsymbol{\beta})$ and $\eta_i(\boldsymbol{\beta}_0)$ such that:

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}X_{ij}(h(\widehat{\eta}_i(\boldsymbol{\beta})) - h(\eta_i(\boldsymbol{\beta})))(g(\eta_i(\boldsymbol{\beta})) - g(\eta_i(\boldsymbol{\beta}_0)))\\
&\qquad = \sum_{r=2}^{S+1}\beta_{r-1}\sum_{l=1}^{S+1}(\beta_{l-1}^{(0)} - \beta_{l-1})\frac{1}{n}\sum_{i=1}X_{ij}(X_{ir} - \widehat{X}_{ir})X_{il}h'(\xi_{i,2})g'(\xi_{i,3}).
\end{aligned}$$

Again, with the help of the Cauchy-Schwarz inequality together with (C.3) it can immediately seen that

$$\frac{1}{n}\sum_{i=1}^{n}X_{ij}(h(\widehat{\eta}_i(\boldsymbol{\beta})) - h(\eta_i(\boldsymbol{\beta})))(g(\eta_i(\boldsymbol{\beta})) - g(\eta_i(\boldsymbol{\beta}_0))) = O_p(n^{-\frac{1}{2}}). \tag{C.35}$$

Similar one may show that

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{X}_{ij}-X_{ij})(\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))})(g(\eta_i(\boldsymbol{\beta}))-g(\eta_i(\boldsymbol{\beta}_0)))=O_p(n^{-\frac{1}{2}}). \tag{C.36}$$

Assertion (C.20) then follows from (C.32) and (C.33)–(C.36). (C.23) follows again from a closer investigation of the existence and boundedness of moments of the involved remainder terms, leading to (C.32).

In order to proof (C.21), note that the $(s+1)\times(s+1)$ matrix $\widehat{\mathbf{F}}(\boldsymbol{\beta})=\widehat{\mathbf{D}}^T(\boldsymbol{\beta})\widehat{\mathbf{V}}^{-1}(\boldsymbol{\beta})\widehat{\mathbf{D}}(\boldsymbol{\beta})$ may be written as

$$\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta})=\frac{1}{n}\mathbf{F}_n(\boldsymbol{\beta})+\mathbf{Rest}_n^{(F)}(\boldsymbol{\beta}).$$

$\mathbf{Rest}_n^{(F)}(\boldsymbol{\beta})$ has a typical element $Rest_{jk}^{(F)}(\boldsymbol{\beta})$ which is given by

$$\begin{aligned}
Rest_{jk}^{(F)}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}(\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))^2}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))}\widehat{X}_{ij}\widehat{X}_{ik}-\frac{g'(\eta_i(\boldsymbol{\beta}))^2}{\sigma^2(g(\eta_i(\boldsymbol{\beta})))}X_{ij}X_{ik}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))^2}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))}-\frac{g'(\eta_i(\boldsymbol{\beta}))^2}{\sigma^2(g(\eta_i(\boldsymbol{\beta})))})X_{ij}X_{ik} && \text{(C.37)} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))^2}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))}(\widehat{X}_{ij}-X_{ij})X_{ik} && \text{(C.38)} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))^2}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))}(\widetilde{X}_{ik}-X_{ik})X_{ij} && \text{(C.39)} \\
&+ \frac{1}{n}\sum_{i=1}^{n}\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))^2}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))}(\widehat{X}_{ik}-X_{ik})(\widehat{X}_{ij}-X_{ij}). && \text{(C.40)}
\end{aligned}$$

$Rest_{jk}^{(F)}(\boldsymbol{\beta})$ consists of the sum of four terms. We begin with (C.37).

Define $h_1(x)=g'(x)^2/\sigma^2(g(x))$ and note that $|h_1(x)|\leq M_{h_1}$ as well as $|h_1'(x)|\leq M_{h_1}$ for some constant $M_{h_1}<\infty$. With the help of the Cauchy-Schwarz inequality and (C.3), it follows from another Taylor expansion that there exists a $\xi_{i,4}$ between $\widehat{\eta}_i(\boldsymbol{\beta})$ and $\eta_i(\boldsymbol{\beta})$ such that:

$$\begin{aligned}
|\frac{1}{n}\sum_{i=1}^{n}(\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))^2}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))}-\frac{g'(\eta_i(\boldsymbol{\beta}))^2}{\sigma^2(g(\eta_i(\boldsymbol{\beta})))})X_{ij}X_{ik}| &= |\sum_{r=2}^{S+1}\beta_{r-1}\frac{1}{n}\sum_{i=1}^{n}(\widehat{X}_{ir}-X_{ir})X_{ij}X_{ik}h_1'(\xi_{i,4})| \\
&\leq \sum_{r=2}^{S+1}|\beta_{r-1}|\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{X}_{ir}-X_{ir})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_{ij}X_{ik}h_1'(\xi_{i,4}))^2}=O_p(n^{-\frac{1}{2}}).
\end{aligned}$$

On the other hand, the Cauchy-Schwarz inequality together with the boundedness $|h_1(x)|$ and (C.3) implies that each of the other terms (C.38)–(C.40) is $O_p(n^{-1/2})$. Assertion (C.21) is then

an immediate consequence. Moreover, since $h_1(x)$ is bounded, it can immediately be seen that $Rest_{jk}^{(F)}(\boldsymbol{\beta})$ is uniform integrable, providing additionally $\mathbb{E}(Rest_{jk}^{(F)}(\boldsymbol{\beta})) \to 0$. Assertion (C.24) follows immediately.
In order so show (C.22), note that $\widehat{\mathbf{R}}_n(\boldsymbol{\beta})/n = \mathbf{R}_n(\boldsymbol{\beta})/n + \mathbf{Rest}_n^{(R)}(\boldsymbol{\beta})$.a typical entry of $\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta})$ reads as

$$Rest_{jk}^{(R)}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(h'(\widehat{\eta}_i(\boldsymbol{\beta})) - h'(\eta_i(\boldsymbol{\beta})))X_{ij}X_{ik}(y_i - g(\eta_i(\boldsymbol{\beta}))) \tag{C.41}$$

$$+ \frac{1}{n}\sum_{i=1}^{n}h'(\widehat{\eta}_i(\boldsymbol{\beta}))X_{ij}(\widehat{X}_{ik} - X_{ik})(y_i - g(\eta_i(\boldsymbol{\beta}))) \tag{C.42}$$

$$+ \frac{1}{n}\sum_{i=1}^{n}h'(\widehat{\eta}_i(\boldsymbol{\beta}))(\widehat{X}_{ij} - X_{ij})X_{ik}(y_i - g(\eta_i(\boldsymbol{\beta}))) \tag{C.43}$$

$$+ \frac{1}{n}\sum_{i=1}^{n}h'(\widehat{\eta}_i(\boldsymbol{\beta}))(\widehat{X}_{ij} - X_{ij})(\widehat{X}_{ik} - X_{ik})(y_i - g(\eta_i(\boldsymbol{\beta}))) \tag{C.44}$$

$$- \frac{1}{n}\sum_{i=1}^{n}h'(\widehat{\eta}_i(\boldsymbol{\beta}))\widehat{X}_{ij}\widehat{X}_{ik}(g(\widehat{\eta}_i(\boldsymbol{\beta})) - g(\eta_i(\boldsymbol{\beta}))). \tag{C.45}$$

we will first show

$$\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta}_0) = \frac{1}{n}\mathbf{R}_n(\boldsymbol{\beta}_0) + O_p(n^{-\frac{1}{2}}). \tag{C.46}$$

For $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, since $|h''(\cdot)| \leq M_h$, a Taylor expansion together with the Cauchy-Schwarz inequality and (C.3) yield $\frac{1}{n}\sum_{i=1}^{n}(h'(\widehat{\eta}_i(\boldsymbol{\beta}_0)) - h'(\eta_i(\boldsymbol{\beta}_0)))X_{ij}X_{ik}\varepsilon_i = O_p(n^{-\frac{1}{2}})$. Similarly each of the assertions (C.42)–(C.44) are $O_p(n^{-\frac{1}{2}})$ At the same time another Taylor expansion of (C.45) yields together with the Cauchy-Schwarz inequality and (C.3) for some $\xi_{i,5}$ between $\widehat{\eta}_i(\boldsymbol{\beta})$ and $\eta_i(\boldsymbol{\beta})$:

$$\frac{1}{n}\sum_{i=1}^{n}h'(\widehat{\eta}_i(\boldsymbol{\beta}_0))\widehat{X}_{ij}\widehat{X}_{ik}(g(\widehat{\eta}_i(\boldsymbol{\beta}_0)) - g(\eta_i(\boldsymbol{\beta}_0)))$$
$$= \sum_{r=2}^{S+1}\beta_{r-1}\frac{1}{n}\sum_{i=1}^{n}h'(\widehat{\eta}_i(\boldsymbol{\beta}_0))g'(\xi_{i,5})\widehat{X}_{ij}\widehat{X}_{ik}(X_{ir} - \widehat{X}_{ir}) = O_p(n^{-\frac{1}{2}}).$$

We may conclude that

$$\widehat{\mathbf{R}}_n(\boldsymbol{\beta}_0) = \mathbf{R}_n(\boldsymbol{\beta}_0) + O_p(n^{-\frac{1}{2}}).$$

Moreover, our assumptions in particular imply that besides $Rest_{jk}^{(R)}(\boldsymbol{\beta}_0)/n = O_p(n^{-\frac{1}{2}})$ we have $\mathbb{E}(Rest_{jk}^{(R)}(\boldsymbol{\beta}_0)) \to 0$, proving assertions (C.46) and (C.26).

Now suppose $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ and take another look at (C.41):

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(h'(\widehat{\eta}_i(\boldsymbol{\beta}))-h'(\eta_i(\boldsymbol{\beta})))X_{ij}X_{ik}(y_i-g(\eta_i(\boldsymbol{\beta})))\\
&\quad=\frac{1}{n}\sum_{i=1}^{n}(h'(\widehat{\eta}_i(\boldsymbol{\beta}))-h'(\eta_i(\boldsymbol{\beta})))X_{ij}X_{ik}\varepsilon_i\\
&\quad\quad-\frac{1}{n}\sum_{i=1}^{n}(h'(\widehat{\eta}_i(\boldsymbol{\beta}))-h'(\eta_i(\boldsymbol{\beta})))X_{ij}X_{ik}(g(\eta_i(\boldsymbol{\beta}))-g(\eta_i(\boldsymbol{\beta}_0)))
\end{aligned}$$

Similar arguments as before, together with $\mathbb{E}(\varepsilon_i^4) < \infty$, can now be used to show that

$$\frac{1}{n}\sum_{i=1}^{n}(h'(\widehat{\eta}_i(\boldsymbol{\beta}))-h'(\eta_i(\boldsymbol{\beta})))X_{ij}X_{ik}\varepsilon_i = O_p(n^{-\frac{1}{2}}).$$

A Taylor expansion of $g(\eta_i(\boldsymbol{\beta}))$ leads for some $\xi_{i,6}$ between $\eta_i(\boldsymbol{\beta})$ and $\eta_i(\boldsymbol{\beta}_0)$ to

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(h'(\widehat{\eta}_i(\boldsymbol{\beta}))-h'(\eta_i(\boldsymbol{\beta})))X_{ij}X_{ik}(g(\eta_i(\boldsymbol{\beta}))-g(\eta_i(\boldsymbol{\beta}_0)))\\
&\quad=\sum_{r=1}^{S+1}(\beta_{r-1}-\beta_{r-1}^{(0)})\frac{1}{n}\sum_{i=1}^{n}(h'(\widehat{\eta}_i(\boldsymbol{\beta}))-h'(\eta_i(\boldsymbol{\beta})))X_{ij}X_{ik}X_{ir}g'(\xi_{i,6}).
\end{aligned}$$

Another Taylor expansion of $h'(\widehat{\eta}_i(\boldsymbol{\beta}))$ together with the Cauchy-Schwarz inequality and the boundedness of $|g'(x)|$ and $|h''(x)|$ leads for some $\xi_{i,7}$ between $\widehat{\eta}_i(\boldsymbol{\beta})$ and $\eta_i(\boldsymbol{\beta})$ to

$$\sum_{r=1}^{S+1}(\beta_{r-1}-\beta_{r-1}^{(0)})\sum_{l=2}^{S+1}\beta_{l-1}\frac{1}{n}\sum_{i=1}^{n}(X_{il}-\widehat{X}_{il})X_{ij}X_{ik}X_{ir}g'(\xi_{i,6})h''(\xi_{i,7}) = O_p(n^{-\frac{1}{2}}).$$

With similar arguments (C.45) and (C.42) are, for all $\boldsymbol{\beta}$, $O_p(n^{-\frac{1}{2}})$.
Considerations for (C.43)–(C.44) are parallel to the case (C.42) assertion (C.22) follows immediately. (C.25) follows again from a closer investigation of the existence and boundedness of the moments of the rest terms used in the derivations (C.22). □

The proof of Theorem 2.3 consists roughly of two steps. In a first step asymptotic existence and consistency of our estimator $\widehat{\boldsymbol{\beta}}$ is developed. In a second step we can then make use of the usual Taylor expansion of the estimation equation $\widehat{\mathbf{U}}_n(\boldsymbol{\beta})$. With the help of Proposition C.2 asymptotic normality of our estimator will follow.

Proof of Theorem 2.3. For a $q_1 \times q_2$ matrix $\mathbf{A}$ let $||\mathbf{A}|| = \sqrt{\sum_{i=1}^{q_1}\sum_{j=1}^{q_2} a_{ij}^2}$ its Frobenius norm. Moreover we denote by $\mathbf{A}^{1/2}$ ($\mathbf{A}^{T/2}$) the left (the corresponding right) square root of a positive definite matrix $\mathbf{A}$.

The proof generalizes the arguments used in Corollary 3 and Theorem 1 in Fahrmeir and Kaufmann (1985). For $\delta_1 > 0$ define the neighborhoods

$$N_n(\delta_1) = \{\boldsymbol{\beta} : ||\widehat{\mathbf{F}}_n^{1/2}(\boldsymbol{\beta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0)|| \leq \delta_1\},$$

and remember that with $h_1(x) = g'(x)^2/\sigma^2(g(x))$ we have:

$$\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} h_1(\widehat{\eta}_i(\boldsymbol{\beta}))\widehat{\mathbf{X}}_i\widehat{\mathbf{X}}_i^T.$$

The (j,k)-element of this random matrix is given by $1/n\sum_{i=1}^{n} h_1(\widehat{\eta}_i(\boldsymbol{\beta}))\widehat{X}_{ij}\widehat{X}_{ik}$ and constitutes a triangular array of row-wise independent and identical distributed random variables. Let $\widehat{\eta}(\boldsymbol{\beta})$, $\widehat{\mathbf{X}}$ and ε be generic copies of $\widehat{\eta}_i(\boldsymbol{\beta})$, $\widehat{\mathbf{X}}_i$ and ε_i. Since h_1 is bounded it is then easy to see that for any compact neighborhood N around $\boldsymbol{\beta}_0$ we have for all $p \geq 1$:

$$\mathbb{E}(\max_{\boldsymbol{\beta}\in N} |h_1(\widehat{\eta}(\boldsymbol{\beta}))\widehat{X}_j\widehat{X}_k|^p) \leq M_{1,1} \tag{C.47}$$

for some constant $M_{1,1} < \infty$, not depending on n. On the other hand the (j,k)-element of $\widehat{\mathbf{R}}_n(\boldsymbol{\beta})/n$ can be written as

$$\frac{1}{n}\sum_{i=1}^{n} h'(\widehat{\eta}_i(\boldsymbol{\beta}))\widehat{X}_{ij}\widehat{X}_{ik}(g(\eta_i(\boldsymbol{\beta}_0)) - g(\widehat{\eta}_i(\boldsymbol{\beta}))) + \frac{1}{n}\sum_{i=1}^{n} h'(\widehat{\eta}_i(\boldsymbol{\beta}))\widehat{X}_{ij}\widehat{X}_{ik}\varepsilon_i.$$

Using the boundedness of g' and h' it follows from a Taylor expansion that for all $p \geq 1$:

$$\mathbb{E}(\max_{\boldsymbol{\beta}\in N} |h'(\widehat{\eta}(\boldsymbol{\beta}))\widehat{X}_j\widehat{X}_k(g(\eta(\boldsymbol{\beta}_0)) - g(\widehat{\eta}(\boldsymbol{\beta})))|^p) \leq M_{1,2} \tag{C.48}$$

for some constant $M_{1,2} < \infty$, not depending on n. While the Cauchy-Schwarz inequality together with the assumption $\mathbb{E}(\varepsilon^4) < \infty$ implies that for $1 \leq p \leq 2$:

$$\mathbb{E}(\max_{\boldsymbol{\beta}\in N} |h'(\widehat{\eta}(\boldsymbol{\beta}))\widehat{X}_j\widehat{X}_k\varepsilon|^p) \leq M_{1,3} \tag{C.49}$$

for some constant $M_{1,3} < \infty$, not depending on n. By (C.47), (C.48) and (C.49) a uniform law of large numbers for triangular arrays leads to

$$\max_{\boldsymbol{\beta}\in N} ||\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}) - \mathbb{E}(\widehat{\mathbf{F}}(\boldsymbol{\beta}))|| \xrightarrow{p} 0, \tag{C.50}$$

as well as

$$\max_{\boldsymbol{\beta}\in N}||\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta})-\mathbb{E}(\widehat{\mathbf{R}}(\boldsymbol{\beta}))||\xrightarrow{p}0. \tag{C.51}$$

Moreover, by (C.47), $\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)/n$ converges a.s. to $\mathbb{E}(\widehat{\mathbf{F}}(\boldsymbol{\beta}_0))$, implying $\lambda_{min}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)\to\infty$ a.s., where $\lambda_{min}\mathbf{A}$ denotes the smallest eigenvalue of a matrix $\mathbf{A}$. Note that as a direct consequence the neighborhoods $N_n(\delta_1)$ shrink (a.s.) to $\boldsymbol{\beta}_0$ for all $\delta_1>0$. On the other hand, since by (C.26), $\mathbb{E}(\widehat{\mathbf{R}}(\boldsymbol{\beta}_0))\to 0$ and $\mathbb{E}(\widehat{\mathbf{R}}(\boldsymbol{\beta}))$ is continuous in $\boldsymbol{\beta}$ we have for all $\epsilon>0$, with probability converging to 1,

$$||\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta})||\le||\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta})-\mathbb{E}(\widehat{\mathbf{R}}(\boldsymbol{\beta}))||+||\mathbb{E}(\widehat{\mathbf{R}}(\boldsymbol{\beta}))-\mathbb{E}(\widehat{\mathbf{R}}(\boldsymbol{\beta}_0))||+||\mathbb{E}(\widehat{\mathbf{R}}(\boldsymbol{\beta}_0))||\le\epsilon$$

if $\boldsymbol{\beta}$ is sufficiently close to $\boldsymbol{\beta}_0$.

The usual decomposition then yields for all $\epsilon>0$, with probability converging to 1:

$$\begin{aligned}
&||-\frac{1}{n}\widehat{\mathbf{H}}_n(\boldsymbol{\beta})-\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)||\le||\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta})-\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)||+||\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta})||\\
&\le||\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta})-\mathbb{E}(\widehat{\mathbf{F}}(\boldsymbol{\beta}))||+||\mathbb{E}(\widehat{\mathbf{F}}(\boldsymbol{\beta}_0))-\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)||+||\mathbb{E}(\widehat{\mathbf{F}}(\boldsymbol{\beta}))-\mathbb{E}(\widehat{\mathbf{F}}(\boldsymbol{\beta}_0))||\\
&+||\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta})||\le\epsilon,
\end{aligned}$$

if $\boldsymbol{\beta}$ is sufficiently close to $\boldsymbol{\beta}_0$. Similar to the proof of Corollary 3 in Fahrmeir and Kaufmann (1985) we may infer from this inequality that for all $\delta_1>0$ we have

$$\max_{\boldsymbol{\beta}\in N_n(\delta_1)}||\widehat{\mathscr{V}}_n(\boldsymbol{\beta})-\mathbf{I}_{S+1}||\xrightarrow{p}0,$$

where $\widehat{\mathscr{V}}_n(\boldsymbol{\beta})=-\widehat{\mathbf{F}}_n^{-1/2}(\boldsymbol{\beta}_0)\widehat{\mathbf{H}}_n(\boldsymbol{\beta})\widehat{\mathbf{F}}_n^{-T/2}(\boldsymbol{\beta}_0)$ and $\mathbf{I}_p$ denotes the $p\times p$ identity matrix. Again, following the arguments in Fahrmeir and Kaufmann (1985, cf. Section 4.1), this in particular implies that for all $\delta_1>0$ we have

$$P(-\widehat{\mathbf{H}}_n(\boldsymbol{\beta})-c\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)\text{ positive semidefinite for all }\boldsymbol{\beta}\in N_n(\delta_1))\to 1 \tag{C.52}$$

for some constant $c>0$, c independent of δ_1.
Let $\widehat{Q}_n(\boldsymbol{\beta})$ be the quasi likelihood function evaluated at the points of impact estimates $\widehat{\tau}_r$. We aim to show that for any $\zeta>0$ there exists a $\delta_1>0$ such that

$$P(\widehat{Q}_n(\boldsymbol{\beta})-\widehat{Q}_n(\boldsymbol{\beta}_0)<0\text{ for all }\boldsymbol{\beta}\in\partial N_n(\delta_1))\ge 1-\zeta \tag{C.53}$$

for all sufficiently large n. Note that the event $\widehat{Q}_n(\boldsymbol{\beta})-\widehat{Q}_n(\boldsymbol{\beta}_0)<0$ for all $\boldsymbol{\beta}\in\partial N_n(\delta_1)$ implies that the there is a maximum inside of $N_n(\delta_1)$. Moreover, since $\widehat{\mathbf{R}}_n(\boldsymbol{\beta})/n$ is asymptotical

negligible in a neighborhood around $\boldsymbol{\beta}_0$, and at the same time $\widehat{\mathbf{F}}_n(\boldsymbol{\beta})/n$ converges in probability to a positive definite matrix, the maximum will, with probability converging to 1, be uniquely determined as a zero of the score function $\widehat{\mathbf{U}}_n(\boldsymbol{\beta})$. (C.53) then in particular implies that $P(\widehat{\mathbf{U}}_n(\widehat{\boldsymbol{\beta}}) = 0) \to 1$ and, together with the observation that $N_n(\delta_1)$ shrink (a.s.) to $\boldsymbol{\beta}_0$, it implies consistency of our estimator, i.e. $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$.
A Taylor expansion yields, with $\boldsymbol{\lambda} = \widehat{\mathbf{F}}_n^{T/2}(\boldsymbol{\beta}_0)(\boldsymbol{\beta}-\boldsymbol{\beta}_0)/\delta_1$, for some $\widetilde{\boldsymbol{\beta}}$ on the line segment between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$:

$$\widehat{Q}_n(\boldsymbol{\beta}) - \widehat{Q}_n(\boldsymbol{\beta}_0) = \delta_1 \boldsymbol{\lambda}' \widehat{\mathbf{F}}_n^{-1/2}(\boldsymbol{\beta}_0)\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0) - \delta_1^2 \boldsymbol{\lambda}' \widehat{\mathscr{V}}_n(\widetilde{\boldsymbol{\beta}})\boldsymbol{\lambda}/2, \quad \boldsymbol{\lambda}'\boldsymbol{\lambda} = 1.$$

Using for the next few lines the spectral norm one may argue similarly to (3.9) in Fahrmeir and Kaufmann (1985), that it suffices to show that for any $\zeta > 0$ we have

$$P(||\widehat{\mathbf{F}}_n^{-1/2}(\boldsymbol{\beta}_0)\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0)|| < \delta_1^2 \lambda_{\min}^2 \widehat{\mathscr{V}}_n(\widetilde{\boldsymbol{\beta}})/4) \geq 1-\zeta.$$

Note that (C.52) implies that with probability converging to one we have

$$\lambda_{\min}^2 \widehat{\mathscr{V}}_n(\widetilde{\boldsymbol{\beta}}) \geq c^2.$$

Hence, with probability converging to one:

$$P(||\widehat{\mathbf{F}}_n^{-1/2}(\boldsymbol{\beta}_0)\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0)||^2 < \delta_1^2 \lambda_{\min}^2 \widehat{\mathscr{V}}_n(\widetilde{\boldsymbol{\beta}})/4) \geq P(||\widehat{\mathbf{F}}_n^{-1/2}(\boldsymbol{\beta}_0)\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0)||^2 < (\delta_1 c)^2/4).$$

At the same time (C.19) and (C.21) can be used to derive

$$\widehat{\mathbf{F}}_n^{-1/2}(\boldsymbol{\beta}_0)\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0) = (\frac{1}{n}\widehat{\mathbf{F}}_n)^{-1/2}(\boldsymbol{\beta}_0)\frac{1}{\sqrt{n}}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0) = \mathbf{F}_n^{-1/2}(\boldsymbol{\beta}_0)\,\mathbf{U}_n(\boldsymbol{\beta}_0) + o_p(1).$$

By the continuous mapping theorem we then have for all $\epsilon > 0$ with probability converging to 1

$$||\widehat{\mathbf{F}}_n^{-1/2}(\boldsymbol{\beta}_0)\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0)||^2 \leq ||\,\mathbf{F}_n^{-1/2}(\boldsymbol{\beta}_0)\,\mathbf{U}_n(\boldsymbol{\beta}_0)||^2 + \epsilon. \tag{C.54}$$

Since $\mathbb{E}(||\,\mathbf{F}_n^{-1/2}(\boldsymbol{\beta}_0)\,\mathbf{U}_n(\boldsymbol{\beta}_0)||^2) = p$, we may conclude from (C.54) that with probability converging to 1 we have for all sufficiently large n:

$$\begin{aligned} P(||\widehat{\mathbf{F}}_n^{-1/2}(\boldsymbol{\beta}_0)\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0)||^2 < \delta_1^2 \lambda_{\min}^2 \widehat{\mathscr{V}}_n(\widetilde{\boldsymbol{\beta}})/4) &\geq P(||\widehat{\mathbf{F}}_n^{-1/2}(\boldsymbol{\beta}_0)\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0)||^2 < (\delta_1 c)^2/4) \\ &\geq P(||\,\mathbf{F}_n^{-1/2}(\boldsymbol{\beta}_0)\,\mathbf{U}_n(\boldsymbol{\beta}_0)||^2 < (\delta_1 c)^2/8) \\ &\geq 1 - 8p/(\delta_1 c)^2 = 1-\zeta, \end{aligned}$$

yielding (C.53) for $\delta_1^2 = 8p/(c^2\zeta)$. Asymptotic existence and consistency of our estimator are immediate consequences.

Remember that we have

$$\begin{aligned}\frac{1}{n}\widehat{\mathbf{H}}_n(\boldsymbol{\beta}) &= \frac{1}{n}\frac{\partial \widehat{\mathbf{U}}_n(\boldsymbol{\beta})}{\partial\, \boldsymbol{\beta}} \\ &= -\frac{1}{n}\widehat{\mathbf{D}}_n(\boldsymbol{\beta})^T\widehat{\mathbf{V}}_n(\boldsymbol{\beta})^{-1}\widehat{\mathbf{D}}_n(\boldsymbol{\beta}) + \frac{1}{n}\sum_{i=1}^{n} h'(\widehat{\eta}_i)\widehat{\mathbf{X}}_i\widehat{\mathbf{X}}_i^T(y_i - g(\widehat{\eta}_i(\boldsymbol{\beta}))) \\ &= -\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}) + \frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta}).\end{aligned}$$

Now, a Taylor expansion of $\widehat{\mathbf{U}}_n(\widehat{\boldsymbol{\beta}})$ around $\boldsymbol{\beta}_0$ yields for some $\widetilde{\boldsymbol{\beta}}$ between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$ (note that $\widetilde{\boldsymbol{\beta}}$ obviously differs from element to element):

$$\begin{aligned}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0) &= \widehat{\mathbf{U}}_n(\widehat{\boldsymbol{\beta}}) - \widehat{\mathbf{H}}_n(\widetilde{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = -\widehat{\mathbf{H}}_n(\widetilde{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \\ &= -\Big(-\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + (\widehat{\mathbf{H}}_n(\widetilde{\boldsymbol{\beta}}) - \widehat{\mathbf{H}}_n(\boldsymbol{\beta}_0))(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + (\widehat{\mathbf{H}}_n(\boldsymbol{\beta}_0) + \widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0))(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\Big).\end{aligned}$$

With some straightforward calculations this leads to

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \Bigg(\mathbf{I}_{S+1} &- \Big(\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)\Big)^{-1}\Big(\frac{\widehat{\mathbf{H}}_n(\widetilde{\boldsymbol{\beta}}) - \widehat{\mathbf{H}}_n(\boldsymbol{\beta}_0)}{n}\Big) \\ &- \Big(\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)\Big)^{-1}\Big(\frac{\widehat{\mathbf{H}}_n(\boldsymbol{\beta}_0) + \widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)}{n}\Big)\Bigg)^{-1}\Big(\frac{\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)}{n}\Big)^{-1}\frac{\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0)}{\sqrt{n}}.\end{aligned} \tag{C.55}$$

By (C.21) and (C.22) in Proposition C.2 we have

$$\frac{\widehat{\mathbf{H}}_n(\boldsymbol{\beta}_0) + \widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)}{n} = \frac{\mathbf{H}_n(\boldsymbol{\beta}_0) + \mathbf{F}_n(\boldsymbol{\beta}_0)}{n} + o_p(1).$$

But since h' is bounded we have for all $\boldsymbol{\beta} \in \mathbf{R}^{S+1}$

$$\mathbb{E}(||\frac{\mathbf{H}_n(\boldsymbol{\beta}) + \mathbf{F}_n(\boldsymbol{\beta})}{n}||_2^2) = \sum_{j=1}^{S+1}\sum_{k=1}^{S+1}\mathbb{E}((\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_{ij}X_{ik}h'(\eta_i(\boldsymbol{\beta})))^2) = O(\frac{1}{n}),$$

implying $(\mathbf{H}_n(\boldsymbol{\beta}_0) + \mathbf{F}_n(\boldsymbol{\beta}_0))/n = O_p(n^{-\frac{1}{2}})$ and hence also

$$||\Big(\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)\Big)^{-1}\Big(\frac{\widehat{\mathbf{H}}(\boldsymbol{\beta}_0) + \widehat{\mathbf{F}}(\boldsymbol{\beta}_0)}{n}\Big)||_2 = o_p(1).$$

By using (C.50) and (C.51) we can conclude that for any compact neighborhood N around $\boldsymbol{\beta}_0$:

$$\max_{\boldsymbol{\beta}\in N}||\frac{1}{n}\widehat{\mathbf{H}}_n(\boldsymbol{\beta})-\mathbb{E}(\widehat{\mathbf{H}}(\boldsymbol{\beta}))||\xrightarrow{p}0. \tag{C.56}$$

Obviously, $\widetilde{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_0$, since $\widehat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}_0$. We may conclude that $\widetilde{\boldsymbol{\beta}}$ will be in some compact neighborhood N around $\boldsymbol{\beta}_0$ with probability converging to 1. Moreover, since $\mathbb{E}(\widehat{\mathbf{H}}(\boldsymbol{\beta}))$ is continuous in $\boldsymbol{\beta}$, (C.56) then implies that additionally we have

$$\max_{\widetilde{\boldsymbol{\beta}}\in N}||\frac{1}{n}\widehat{\mathbf{H}}_n(\widetilde{\boldsymbol{\beta}})-\mathbb{E}(\widehat{\mathbf{H}}(\boldsymbol{\beta}_0))||=o_p(1). \tag{C.57}$$

The above arguments can then be used to show that

$$||\Big(\frac{\widehat{\mathbf{H}}_n(\widetilde{\boldsymbol{\beta}})-\widehat{\mathbf{H}}_n(\boldsymbol{\beta}_0)}{n}\Big)||\leq||\frac{\widehat{\mathbf{H}}_n(\widetilde{\boldsymbol{\beta}})}{n}-\mathbb{E}(\widehat{\mathbf{H}}(\boldsymbol{\beta}_0))||+||\frac{\widehat{\mathbf{H}}_n(\boldsymbol{\beta}_0)}{n}-\mathbb{E}(\widehat{\mathbf{H}}(\boldsymbol{\beta}_0))||=o_p(1).$$

Hence it also holds that

$$||\Big(\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)\Big)^{-1}\Big(\frac{\widehat{\mathbf{H}}_n(\widetilde{\boldsymbol{\beta}})-\widehat{\mathbf{H}}_n(\boldsymbol{\beta}_0)}{n}\Big)||=o_p(1).$$

The asymptotic prevailing term in (C.55) can then be seen as

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\sim\left(\frac{\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)}{n}\right)^{-1}\frac{\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0)}{\sqrt{n}}. \tag{C.58}$$

It is easy to see that our assumptions on $h(x)=g'(x)/\sigma^2(g(x))$ imply that $\mathbb{E}(||\mathbf{F}_n(\boldsymbol{\beta}_0)/n||^2)=O(\frac{1}{n})$. Together with (C.21) we thus have $\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)/n=\mathbf{F}_n(\boldsymbol{\beta}_0)/n+O_p(n^{-\frac{1}{2}})=\mathbb{E}(\mathbf{F}(\boldsymbol{\beta}_0))+O_p(n^{-\frac{1}{2}})$ as well as $(\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)/n)^{-1}=(\mathbb{E}(\mathbf{F}(\boldsymbol{\beta}_0)))^{-1}+O_p(n^{-\frac{1}{2}})$.

On the other hand, the Lindeberg-Lévy central limit theorem implies that $\frac{1}{\sqrt{n}}U(\boldsymbol{\beta}_0)\xrightarrow{d}N(\mathbf{0},\mathbb{E}(\mathbf{F}(\boldsymbol{\beta}_0)))$. Together with (C.19) we then obtain

$$\left(\frac{\widehat{\mathbf{F}}(\boldsymbol{\beta}_0)}{n}\right)^{-1}\frac{\widehat{\mathbf{U}}(\boldsymbol{\beta}_0)}{\sqrt{n}}\xrightarrow{d}N(\mathbf{0},(\mathbb{E}(\mathbf{F}(\boldsymbol{\beta}_0))^{-1}),$$

which proves the theorem. □

Corrolary C.1. *Under the assumptions of Section 2.3. For any compact neighborhood N around* $\boldsymbol{\beta}_0$ *we have*

$$\max_{\boldsymbol{\beta}\in N}||\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta})-\frac{1}{n}\mathbf{U}_n(\boldsymbol{\beta})||=o_p(1), \tag{C.59}$$

$$\max_{\boldsymbol{\beta}\in N}||\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta})-\frac{1}{n}\mathbf{F}_n(\boldsymbol{\beta})||=o_p(1), \tag{C.60}$$

$$\max_{\boldsymbol{\beta}\in N}||\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta})-\frac{1}{n}\mathbf{R}_n(\boldsymbol{\beta})||=o_p(1), \tag{C.61}$$

as well as

$$\max_{\boldsymbol{\beta}\in N}||\frac{1}{n}\widehat{\mathbf{H}}_n(\boldsymbol{\beta})-\frac{1}{n}\mathbf{H}_n(\boldsymbol{\beta})||=o_p(1). \tag{C.62}$$

Proof of Corollary C.1: The proofs of Assertions C.59-C.61 are very similar. We begin with the proof of Assertion C.59. Using again generic copies of $\widehat{\eta}_i$, $\widehat{\mathbf{X}}_i$ and y_i we have with $h(x)=g'(x)/\sigma^2(g(x))$:

$$\mathbb{E}(n^{-1}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}))=\mathbb{E}(\widehat{\mathbf{U}}(\boldsymbol{\beta}))=\mathbb{E}(h(\widehat{\eta})\widehat{\mathbf{X}}(y-g(\widehat{\eta}(\boldsymbol{\beta})))).$$

The j-th equation of $\widehat{\mathbf{U}}(\boldsymbol{\beta})$ can be rewritten as

$$h(\widehat{\eta}(\boldsymbol{\beta}))\widehat{X}_j(y-g(\widehat{\eta}(\boldsymbol{\beta})))=h(\widehat{\eta}(\boldsymbol{\beta}))\widehat{X}_j\epsilon+h(\widehat{\eta}(\boldsymbol{\beta}))\widehat{X}_j(g(\eta(\boldsymbol{\beta}_0))-g(\widehat{\eta}(\boldsymbol{\beta}))).$$

Choose an arbitrary compact neighborhood N around $\boldsymbol{\beta}_0$. Since $|h(\cdot)|\leq M_h$, $\mathbb{E}(\epsilon^4)<\infty$ and $|g'(\cdot)|<M_g$, it follows from a Taylor expansion that for $1\leq p\leq 2$ we have

$$\mathbb{E}(\max_{\boldsymbol{\beta}\in N}|h(\widehat{\eta}(\boldsymbol{\beta}))\widehat{X}_j(y-g(\widehat{\eta}(\boldsymbol{\beta})))|^p)\leq M_{1,1} \tag{C.63}$$

for a constant $0\leq M_{1,1}<\infty$ not depending on n. By (C.63) we can apply a uniform law of large numbers for triangular arrays to conclude that

$$\max_{\boldsymbol{\beta}\in N}||\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta})-\mathbb{E}(\widehat{\mathbf{U}}(\boldsymbol{\beta}))||=o_p(1). \tag{C.64}$$

Similar considerations lead to

$$\max_{\boldsymbol{\beta}\in N}||\frac{1}{n}\mathbf{U}_n(\boldsymbol{\beta})-\mathbb{E}(\mathbf{U}(\boldsymbol{\beta}))||=o_p(1). \tag{C.65}$$

By the usual decomposition we have

$$\begin{aligned}||\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta})-\frac{1}{n}\mathbf{U}_n(\boldsymbol{\beta})|| \leq ||\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta})-\mathbb{E}(\widehat{\mathbf{U}}(\boldsymbol{\beta}))|| \\ + ||\frac{1}{n}\mathbf{U}_n(\boldsymbol{\beta})-\mathbb{E}(\mathbf{U}(\boldsymbol{\beta}))|| + ||\,\mathbb{E}(\widehat{\mathbf{U}}(\boldsymbol{\beta}))-\mathbb{E}(\mathbf{U}(\boldsymbol{\beta}))||.\end{aligned} \tag{C.66}$$

Assertion (C.59) then follows immediately from (C.64), (C.65), if we can show that $\mathbb{E}(\widehat{\mathbf{U}}(\boldsymbol{\beta}))$ converges uniformly to $\mathbb{E}(\mathbf{U}(\boldsymbol{\beta}))$ and not only pointwise as given in (C.23).

It is well known that pointwise convergence of a sequence of functions f_n on a compact set N can be extended to uniform convergence over N, if f_n is an equicontinuous sequence. Remember that a sufficient condition for equicontinuity is that there exists a common Lipschitz constant. We aim to show that there exists a constant $L < \infty$ where L does not depend on n, such that for all $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\beta}}$ in N we have $||\,\mathbb{E}(\widehat{\mathbf{U}}(\boldsymbol{\beta}))-\mathbb{E}(\widehat{\mathbf{U}}(\widetilde{\boldsymbol{\beta}}))|| \leq L||\,\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}||$. Remember that the jth equation of $\mathbb{E}(\widehat{\mathbf{U}}(\boldsymbol{\beta}))$ is given by $\mathbb{E}(h(\widehat{\eta}(\boldsymbol{\beta}))\widehat{X}_j(y-g(\widehat{\eta}(\boldsymbol{\beta}))))$. Note that

$$\begin{aligned}&h(\widehat{\eta}(\boldsymbol{\beta}))\widehat{X}_j(y-g(\widehat{\eta}(\boldsymbol{\beta})))-h(\widehat{\eta}(\widetilde{\boldsymbol{\beta}}))\widehat{X}_j(y-g(\widehat{\eta}(\widetilde{\boldsymbol{\beta}})))\\ &\quad=\widehat{X}_j y(h(\widehat{\eta}(\boldsymbol{\beta}))-h(\widehat{\eta}(\widetilde{\boldsymbol{\beta}})))\\ &\qquad+\widehat{X}_j h(\widehat{\eta}(\widetilde{\boldsymbol{\beta}}))(g(\widehat{\eta}(\widetilde{\boldsymbol{\beta}}))-g(\widehat{\eta}(\boldsymbol{\beta})))\\ &\qquad-\widehat{X}_j(h(\widehat{\eta}(\widetilde{\boldsymbol{\beta}}))-h(\widehat{\eta}(\boldsymbol{\beta})))g(\widehat{\eta}(\boldsymbol{\beta})).\end{aligned} \tag{C.67}$$

Since for a $J \times K$ matrix A we have $||A|| = \sqrt{\sum_{j,k} a_{jk}^2} \leq \sum_{j,k}|a_{jk}|$ and since h, h' and g' are bounded and N is compact, our assumptions on X then in particularly imply together with (C.67) that there exists a constant L, which is in particular independent of n such that for all $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\beta}} \in N$

$$||\,\mathbb{E}(\widehat{\mathbf{U}}(\boldsymbol{\beta}))-\mathbb{E}(\widehat{\mathbf{U}}(\widetilde{\boldsymbol{\beta}}))|| \leq L||\,\boldsymbol{\beta}-\widetilde{\boldsymbol{\beta}}||. \tag{C.68}$$

Assertion (C.59) then follows from (C.66) together with (C.64), (C.65), (C.23) and (C.68).

In order to proof Assertion (C.60) we can use the decomposition

$$\begin{aligned}||\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta})-\frac{1}{n}\mathbf{F}_n(\boldsymbol{\beta})|| \leq ||\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta})-\mathbb{E}(\widehat{\mathbf{F}}(\boldsymbol{\beta}))|| \\ + ||\frac{1}{n}\mathbf{F}_n(\boldsymbol{\beta})-\mathbb{E}(\mathbf{F}(\boldsymbol{\beta}))|| + ||\,\mathbb{E}(\widehat{\mathbf{F}}(\boldsymbol{\beta}))-\mathbb{E}(\mathbf{F}(\boldsymbol{\beta}))||.\end{aligned}$$

Let $h_1(x) = g'(x)^2/\sigma^2(g(x))$ and remember that $|h_1(\cdot)| \leq M_{h_1}$ for some constant $M_{h_1} < \infty$. It immediately follows that for any compact neighborhood N arround $\boldsymbol{\beta}_0$ we have

$$\mathbb{E}(\max_{\boldsymbol{\beta}\in N}||h_1(\eta_i(\boldsymbol{\beta}))\mathbf{X}_i\mathbf{X}_i^T||) < \infty. \tag{C.69}$$

By (C.69) we can apply a uniform law of large numbers to derive

$$\max_{\boldsymbol{\beta}\in N} || \frac{1}{n} \mathbf{F}_n(\boldsymbol{\beta}) - \mathbb{E}(\mathbf{F}(\boldsymbol{\beta}))|| = o_p(1). \tag{C.70}$$

Assertion C.60 then follows immediately from (C.70), (C.50) and (C.24), which holds uniformly on N by similar steps as above by noting that a typical element of $\widehat{\mathbf{F}}(\boldsymbol{\beta}) - \widehat{\mathbf{F}}(\widetilde{\boldsymbol{\beta}})$ can be written as $\widehat{X}_j \widehat{X}_k (h_1(\widehat{\eta}(\boldsymbol{\beta})) - h_1(\widehat{\eta}(\widetilde{\boldsymbol{\beta}})))$.
Assertion C.61 can be proved in a similar manner using (C.23), (C.56) and the uniform convergence of $|| \mathbf{R}_n(\boldsymbol{\beta})/n - \mathbb{E}(\mathbf{R}(\boldsymbol{\beta}))||$, which is easy to establish using $h_1(\eta(\boldsymbol{\beta}))\mathbf{X}\mathbf{X}^T(y - g(\eta(\boldsymbol{\beta}))) = h_1(\eta(\boldsymbol{\beta}))\mathbf{X}\mathbf{X}^T \varepsilon + h_1(\eta(\boldsymbol{\beta}))\mathbf{X}\mathbf{X}^T(g(\eta(\boldsymbol{\beta}_0)) - g(\eta(\boldsymbol{\beta})))$ and the assumption that $|h_1'(\cdot)| \leq M_{h_1}$.

To proof assertion (C.62), remember that $\widehat{\mathbf{H}}(\boldsymbol{\beta})/n = -\widehat{\mathbf{F}}_n(\boldsymbol{\beta})/n + \widehat{\mathbf{R}}_n(\boldsymbol{\beta})/n$, assertion (C.62) follows then immediately from (C.60) and (C.61). □

Appendix D Extending the linear predictor by $\int_a^b \beta(t)X_i(t)\,dt$

In this supplementary appendix the case where the linear predictor is given by

$$\eta_i = \alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r) + \int_a^b \beta(t)X_i(t)\,dt \tag{D.1}$$

is considered. The slope function $\beta(t)$ appearing in (D.1) is assumed to be bounded and square integrable over $[a,b]$. Adding $\int_a^b \beta(t)X_i(t)\,dt$ to the linear predictor allows us to capture a common effect of the whole trajectory X_i on Y_i. Since the proofs in Appendix B are already tailored for the linear predictor (D.1), theoretical results of Section 2.2 concerning the estimation of points of impact τ_r remain valid. Moreover identifiability of all model components is still guaranteed by Theorem 2.4 in Appendix C given the covariance function of X_i satisfies Asssumption 2.1 and g is invertible.

By introducing $\int_a^b \beta(t)X_i(t)\,dt \approx 1/p \sum_{j=1}^{p} X_i(t_j)\beta(t_j)$ into the model each observed grid point t_j has a potential influence on the outcome Y_i through the functional value $X_i(t_j)$. Quite obviously, estimating points of impact by using simple model selection criteria like the BIC as in Section 2.4 can not work anymore since these procedures do not account for the integral part of the linear predictor. However by Theorem 2.1 points of impact can, for example, still be estimated as described in Section 2.2 using a suitable cut-off criterion. But even if the points of impact are consistently estimated in a first step, the corresponding parameter estimates $\widehat{\beta}_r$ will be biased whenever the functional part is not considered during their estimation procedure. On the other hand, also the effect of the points of impact can not be neglected by solely using a generalized functional linear model formulation since the regression function $\beta(t)$ alone is not able to adequately capture the specific effects of the trajectory X_i at the points of impact. To see this, first note that by the sifting property of the Dirac's delta function $\delta(x)$ it holds that $\beta_r X_i(\tau_r) = \beta_r \int_a^b X_i(t)\delta(t-\tau_r)\,dt$. On the other hand it is well known that the Dirac delta function is not an element of the function space L^2 and hence can not be captured adequately by means of a standard generalized functional regression model as given by Müller and Stadtmüller (2005). Thus, by incorporating $\int_a^b \beta(t)X_i(t)\,dt$ into the linear predictor some additional considerations for deriving theoretical results and practical guidance for estimators of $\beta_0, \beta_1, \ldots, \beta_S$ and $\beta(t)$ are needed.

In this appendix two additional results for parameter estimates are presented covering two cases. In the first case one might be only interested in the estimates for $\beta_1, \ldots, \beta_S$. In Theorem 2.5 it will be shown that estimation of these parameters up to a common constant is possible using a very simple and computational efficient estimation procedure. This case is particularly important if one is not interested in the actual values of the estimated coefficients but in their relative importance. The second approach generalizes the results from Müller and

Stadtmüller (2005) to the current setting and is based on a basis expansion approach of $\beta(t)$. This approach is concerned about a comprehensive estimation of all involved model parameters $\beta_0, \beta_1, \ldots, \beta_S$ as well as the slope function $\beta(t)$. After having derived some theoretical results, this appendix is concluded with some additional simulation results.

Throughout this appendix we hold on to the basic assumptions of Section 2.3 by assuming that X_i is Gaussian and satisfies Assumption 2.1, S has been consistently estimated (i.e. $\widehat{S} = S$) and that the points of impact are ordered such that $\tau_r = \arg\min_{s=1,\ldots,S} |\widehat{\tau}_r - \tau_s|$, $r = 1, \ldots, S$ and $|\widehat{\tau}_r - \tau_r| = O_p(n^{-1/\kappa})$.

D.1 Estimating model parameters in the extended model: instrumental variables estimation approach

Consider the case where $\beta(t) \equiv 0$ such that $\eta_i = \alpha + \sum_{r=1}^{S} \beta_r X(\tau_r)$. If additionally the points of impact τ_r are known, we are then in a standard setting of a generalized linear model with Gaussian regressors. With a slight abuse of notation, it follows in this case from Lemma 2.1 that the ordinary least squares estimator for β_r derived from fitting the simplified model (denoted as OLS)

$$\widetilde{Y}_i = \alpha + \sum_{r=1}^{S} \beta_r X(\tau_r) + \varepsilon_i^*$$

is proportional to β_r from the correctly specified model $Y_i = g(\alpha + \sum_{r=1}^{S} \beta_r X(\tau_r)) + \varepsilon_i$ (cf. Brillinger (2012a)). Put differently, by neglecting the function g in our model (2.2), the ordinary least squares estimator for the unknown parameter β_r will still yield an image of the relative importance of the points of impact.

Given one is primarily interested in this relative importance one may want to estimate the unknown coefficients by OLS even if $\beta(t) \neq 0$. But one then estimates β_r in the model

$$\widetilde{Y}_i = \alpha + \sum_{r=1}^{S} \beta_r X(\tau_r) + \widetilde{\varepsilon}_i,$$

where the error term $\widetilde{\epsilon}_i = \int_a^b \beta(t) X_i(t)\, dt + \varepsilon_i^*$ will now be correlated with $X_i(\tau_r)$. In this setting an instrumental variable estimation approach for estimating the coefficients $\beta_1, \ldots, \beta_r$ can be considered.

Interestingly, $Z_{\delta,i}(s)$ evaluated at $s \approx t$ behaves similar to an instrumental variable for $X_i(t)$. Indeed it follows from Theorem 3 in Kneip et al. (2016a), that for $|s - t| \approx 0$ we have

$$E(Z_{\delta,i}(s) X_i(t)) = \delta^\kappa c(t) + O(\max\{\delta^{2\kappa}, \delta^2\}),$$

while for $|s - t| > \delta$ we have $E(Z_{\delta,i}(s) X_i(t)) = o(\delta^\kappa)$. For instance, $Z_{\delta,i}(t)$ is highly correlated with $X_i(t)$ but essentially uncorrelated with the rest of its trajectory outside a small neigh-

borhood around the point t. As a direct consequence it is shown in Kneip et al. (2016a) that $\mathbb{E}(\int_a^b \beta(t)Z_{\delta,i}(s)X_i(t)\,dt) = o(\delta^\kappa)$ (cf. Lemma B.3). Hence $Z_{\delta,i}(t)$ is also essentially uncorrelated with $\widetilde{\epsilon}_i$ and $Z_{\delta,i}(t)$ can indeed be seen as an "approximate" instrumental variable for $X_i(t)$.

After having obtained points of impact candidates $\widehat{\tau}_1, \widehat{\tau}_2, \ldots, \widehat{\tau}_{\widehat{S}}$, one may thus use a simple instrumental variable estimator to obtain estimates of a multiple of the coefficients β_r:

$$\widetilde{\beta}_r = \frac{\sum_{i=1}^n Z_{\delta,i}(\widehat{\tau}_r)Y_i}{\sum_{i=1}^n Z_{\delta,i}(\widehat{\tau}_r)X_i(\widehat{\tau}_r)}. \tag{D.2}$$

The theoretical justification for this estimator is given by the following theorem.

Theorem 2.5. *Let X_i be a Gaussian process satisfying Assumption 2.1 and let $\max_{r=1,\ldots,S} |\widehat{\tau_r} - \tau_r| = O_p(n^{-1/\kappa})$. Under the assumptions of Lemma 2.1 suppose $\delta \to 0$, $\delta n^\kappa \to \infty$. There then exists a constant M with $0 < |M| < \infty$ such that*

$$\widetilde{\beta}_r = M\beta_r + O_P\left(\delta^{\min\{1,2-\kappa\}} + \frac{1}{\sqrt{\delta^\kappa n}}\right) \tag{D.3}$$

for all $r = 1, \ldots, S$ as $n \to \infty$.

Theorem D.3 states that β_r can be consistently estimated up to a constant M. The constant M is given by Lemma 2.1. While the constant M is unknown, Theorem 2.5 might still be of particular importance since the estimator does not assume any concrete knowledge about the functional form of g and $\beta(t)$ and hence will be robust against model misspecifications. In the case of a functional logistic regression with points of impact the constant M obeys by Stein's Lemma (C. M. Stein (1981)) the bound $0 < M = \mathbb{E}(\exp(\eta_i)/(1+\exp(\eta_i))) \leq 0.25$, which gives some further information on the actual value of β_r. Rates of convergence are nonparametric and depend on κ and δ. If κ is known one might adapt the value of δ in order to achieve a best possible rate. If for example $\kappa = 1$, one might then choose $\delta \sim n^{-1/3}$ leading to $|\widetilde{\beta}_r - M\beta_r| = O_p(n^{-1/3})$. Note that an estimator for κ is available from Proposition 1 in Kneip et al. (2016a).

D.2 Estimating model parameters in the extended model: comprehensive approach

For a comprehensive estimation of the coefficients $\beta_0, \beta_1, \ldots, \beta_S$ as well as the slope function $\beta(t)$ a basis expansion approach is used. For this let γ_j, $j = 1, 2, \ldots$ be an orthonormal basis of the function space $L^2([a,b])$. It then follows that $\int_a^b \beta(t)X_i(t)$ can be expressed as

$$\int_a^b \beta(t)X_i(t) = \sum_{j=1}^\infty \alpha_j\theta_{ij},$$

where $\alpha_j = \int_a^b \beta(t)\gamma_j(t)\,dt$ are unknown and $\theta_{ij} = \int_a^b X_i(t)\gamma_j(t)\,dt$. Note that θ_{ij} is Gaussian distributed, since X_i is assumed to be Gaussian.

In the case $\beta(t) \neq 0$ it is more convenient to work with standardized errors ε_i', such that model (2.2) can be rewritten as

$$Y_i = g\left(\alpha + \sum_{r=1}^{S}\beta_r X_i(\tau_r) + \sum_{j=1}^{\infty}\alpha_j\theta_{ij}\right) + \varepsilon_i'\sigma(g(\alpha + \sum_{r=1}^{S}\beta_r X_i(\tau_r) + \sum_{j=1}^{\infty}\alpha_j\theta_{ij})), \tag{D.4}$$

where $\mathbb{E}(\varepsilon_i'|X_i) = 0$ and $\mathbb{E}(\varepsilon_i'^2|X) = 1$, implying $\mathbb{E}(\varepsilon') = 0$ and $\mathbb{E}(\varepsilon'^2) = 1$.
Following the arguments given in Müller and Stadtmüller (2005, Sec. 2), it is then sufficient to analyze the P truncated models

$$Y_i^{(P)} = g\left(\alpha + \sum_{r=1}^{S}\beta_r X_i(\tau_r) + \sum_{j=1}^{P}\alpha_j\theta_{ij}\right) + \varepsilon_i'\sigma(\alpha + \sum_{r=1}^{S}\beta_r X_i(\tau_r) + \sum_{j=1}^{P}\alpha_j\theta_{ij}), \tag{D.5}$$

where $P = P_n \to \infty$ as $n \to \infty$.

In the following, the notation from section 2.3 is augmented in a straight forward way. For example, the objects $\mathbf{X}_i$ and $\widehat{\mathbf{X}}_i$ are expanded by the additional components $\theta_{i1},\ldots,\theta_{iP}$, to $\mathbf{X}_i = (1, X_i(\tau_1),\ldots,X_i(\tau_S),\theta_{i1},\ldots,\theta_{iP})^T$ and $\widehat{\mathbf{X}}_i = (1, X_i(\widehat{\tau}_1),\ldots,X_i(\widehat{\tau}_S),\theta_{i1},\ldots,\theta_{iP})^T$. Similary the parameter vector $\boldsymbol{\beta}_0$ is expanded to $(\alpha,\beta_1,\ldots,\beta_S,\alpha_1,\ldots,\alpha_1,\ldots,\alpha_P)$. The linear predictor of the P truncated model can then be written as $\eta_i(\boldsymbol{\beta}_0) = \mathbf{X}_i^T\boldsymbol{\beta}_0$. Using again generic copies of η_i and $\mathbf{X}_i$ we note that the now $(P+S+1)\times(P+S+1)$-matrix $\mathbb{E}(\mathbf{F}(\boldsymbol{\beta})) = \mathbb{E}(g'(\eta(\boldsymbol{\beta}))^2/\sigma^2(g(\eta(\boldsymbol{\beta})))\mathbf{X}\mathbf{X}^T)$ is strictly positive definite and hence invertible. We denote the (j,k)th element of the inverse $\mathbb{E}(\mathbf{F}(\boldsymbol{\beta}_0))^{-1}$ by $\xi_{j,k}$.

The following set of assumptions is needed to derive theoretical results for the estimator of $\boldsymbol{\beta}_0$.

Assumption D.1.

a) *There exists a constant* $0 < M_\varepsilon < \infty$, *such that* $\mathbb{E}((\varepsilon_i')^p|X_i) \leq M_\varepsilon$ *for some even* p *with* $p \geq \max\{2/\kappa + \epsilon, 4\}$ *for some* $\epsilon > 0$.

b) *The known link function* g *is monotone, invertible with two bounded derivatives* $|g'(\cdot)| \leq M_g$, $|g''(\cdot)| \leq M_g$, *for some constant* $0 \leq M_g < \infty$.

c) $h(\cdot) := \frac{g'(\cdot)}{\sigma^2(g(\cdot))}$ *is a bounded function with two bounded derivatives.*

d) *The variance function* $\sigma^2(\cdot)$ *has a continuous bounded derivative and* $|h(\cdot)\sigma(g(\cdot))|$ *as well as* $|h'(\cdot)\sigma(g(\cdot))|$ *are bounded.*

e) $\sup_{t\in[a,b]}\sup_j|\gamma_j(t)| \leq M_\gamma$ *for some* $M_\gamma < \infty$ *as well as* $\sum_{r=1}^{\infty}|\alpha_r| < M_\alpha$ *for some* $M_\alpha < \infty$.

f) *The number of basis functions used in the* p*-truncated model satisfy* $P = P_n \to \infty$, *with* $P_n n^{-1/4} \to 0$, *as* $n \to \infty$.

g) It holds that

$$\sum_{k_1,k_2,k_3,k_4=1}^{P_n+S+1} \mathbb{E}\left(X_{k_1}X_{k_2}X_{k_3}X_{k_4} \frac{g'^4(\eta(\boldsymbol{\beta}_0))}{\sigma^4(g(\eta(\boldsymbol{\beta}_0)))} \right) \xi_{k_1,k_2}\xi_{k_3,k_4} = o(n/P_n^2).$$

h) It holds that

$$\sum_{k_1,\ldots,k_8=1}^{P_n+S+1} \mathbb{E}\left(\frac{g'^4(\eta(\boldsymbol{\beta}_0))}{\sigma^4(g(\eta(\boldsymbol{\beta}_0)))} X_{k_1}X_{k_3}X_{k_5}X_{k_7} \right) \times \mathbb{E}\left(\frac{g'^4(\eta(\boldsymbol{\beta}_0))}{\sigma^4(g(\eta(\boldsymbol{\beta}_0)))} X_{k_2}X_{k_4}X_{k_6}X_{k_8} \right) \xi_{k_1,k_2}\xi_{k_3,k_4}\xi_{k_5,k_6}\xi_{k_7,k_8} = o(n^2P_n^2).$$

Assumptions D.1 a)-c) correspond exactly to Assumption 2.3. Assumption D.1 a) is adjusted to easily take standardized errors of the truncated model into account. The condition on γ_j in Assumption D.1 e) is for example fulfilled if the basis functions are taken from a Fourier-type basis. Assumptions f)-h) match the asymptotic assumptions (M2)-(M4) in Müller and Stadtmüller (2005) for the case where, besides the intercept, S additional covariates are present. Moreover Assumptions D.1 c) - d) replace assumption (M1) in Müller and Stadtmüller (2005) and are adjusted to allow for the important case of a functional logistic regression with points of impact.

Our estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$ is still defined as the solution of the, now $S+1+P$, score equations $\widehat{\mathbf{U}}_n(\widehat{\boldsymbol{\beta}}) = 0$ as given in Section 2.3, where

$$\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = \widehat{\mathbf{D}}_n(\boldsymbol{\beta})^T \widehat{\mathbf{V}}_n(\boldsymbol{\beta})^{-1}(\mathbf{Y}_n - \widehat{\boldsymbol{\mu}}_n(\boldsymbol{\beta})). \tag{D.6}$$

Under Assumption D.1 we will asymptotically still have $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} (\frac{1}{n}\mathbf{F}(\boldsymbol{\beta}_0))\frac{\mathbf{U}(\boldsymbol{\beta}_0)}{\sqrt{n}}$, i.e. the estimator $\widehat{\boldsymbol{\beta}}$ follows the same distribution as an estimator where the points of impact are known. From this we can generalize the results from Müller and Stadtmüller (2005) to derive the following theorem:

Theorem 2.6. *Let $\widehat{S} = S$, $\max_{r=1,\ldots,S} |\widehat{\tau_r} - \tau_r| = O_p(n^{-1/\kappa})$ and let X_i be a Gaussian process satisfying Assumption 2.1. Under Assumption D.1 we then obtain*

$$\frac{n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \mathbb{E}(\mathbf{F}(\boldsymbol{\beta}_0))(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - (P_n + S + 1)}{\sqrt{2(P_n + S + 1)}} \xrightarrow{d} N(0,1), \tag{D.7}$$

Theorem 2.6 is qualitatively the same as Theorem 4.1 in Müller and Stadtmüller (2005). Using Theorem 2.6, confidence bands for $\beta(t)$ and further theoretical assertions might be derived similarly as in Müller and Stadtmüller (2005).

In practice P_n can be chosen by using an information criteria like the AIC or BIC (cf. Müller and Stadtmüller (2005)). In the practical relevant case where S is unknown, this step will be performed after having estimated the points of impact and forcing $X_i(\widehat{\tau}_1), \ldots, X_i(\widehat{\tau}_{\widehat{S}})$ as additional covariates into the model.

D.3 Simulation study for the extended model

Additional Monte Carlo simulations are performed to measure the finite sample performance of our estimators given in the two previous sections. Observational data (X_i, Y_i) are constructed for different sample sizes of n and p by simulating a functional logistic regression model with two points of impact. For this, the curves $X_1, \ldots, X_n$ are generated as independent Ornstein-Uhlenbeck processes with parameters $\theta = 5$ and $\sigma_u = 3.5$ at p equidistant grid points $t_j = (j-1)/(p-1)$ over the interval $[0,1] = [a,b]$. The function g is chosen to be the logit link with $g(x) = \exp(x)/(1+\exp(x))$. The two points of impact τ_1 and τ_2 are chosen to be the smallest observed grid points closest to $1/3$ and $2/3$ respectively. Associated coefficients are set to $\beta_1 = -4$ and $\beta_2 = 5$ while the constant α is set to 1. For the slope function $\beta(t)$ the two cases $\beta(t) \equiv 0$ and $\beta(t) = 5\sum_{j=1}^{3} \alpha_j \gamma_j(t)$ are considered. In the latter case, the coefficients α_j are set to $\alpha_j = 1/j$ and $\gamma_j(t) = \sqrt{2}\sin(\pi j t)$ is chosen to belong to a Fourier-type basis.

The estimator from Section D.2 (denoted as TRH-BIC) is implemented using a two step procedure: (i) In a first step the cut-off procedure as described in Section 2.2.1 is applied to get points of impact candidates $\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{S}}$. The cut-off procedure relies on the choice of $\delta = \delta_n$ and the choice of a threshold $\lambda = \lambda_n$. Following Theorem 2.1, δ was set to $c_\delta/\sqrt{n}$ for $c_\delta = 1.5$, but similar qualitatively results were obtained for other choices of c_δ. As a cut-off we set $\lambda = A\sqrt{\log((b-a)/\delta)/n\sqrt{\widehat{\mathbb{E}}(Y^4)}}$ with $A = \sqrt{2\sqrt{3}}$, imitating the lower bound of the cut-off as it was derived from a central limit theorem driven motivation while replacing $\mathbb{E}(Y^4)$ by its estimator $\widehat{\mathbb{E}}(Y^4) = 1/n\sum_{i=1}^{n} Y_i^4$ (cf. the remark after the proof of Lemma B.2).
(ii) After the estimation of the number and locations of the points of impact $\widehat{S}$ and $\widehat{\tau}_1, \ldots, \widehat{\tau}_{\widehat{S}}$ the BIC is used to select the number of basis functions $\gamma_1(t), \ldots, \gamma_P(t)$ in the representation of $\beta(t)$ in a second step. In particularly, additionally to the fixed set of covariates $X_i(\widehat{\tau}_1), \ldots, X_i(\widehat{\tau}_{\widehat{S}})$, we allow a maximum of up to $5 = P_{max}$ basis functions to enter the model through the scores $\theta_{ij} \approx 1/p\sum_{j=1}^{p} X_i(t)\gamma_j(t)$. The number P of basis function to enter in the finally selected model is then chosen by minimizing the BIC while also allowing for $P = 0$. In the latter case no score enters the model, resulting in an estimated slope function of $\widehat{\beta}(t) \equiv 0$.

All simulations are performed in **R**. Table D.1 contains the results of the simulation study from 5000 repetitions for each combination of n and p. The first columns contain the mean absolute error of our TRH-BIC estimates for the points of impact τ_1, τ_2, their associated coefficients β_1 and β_2 as well as the mean absolute error of the intercept α. To measure the quality of our slope estimator the mean error of $\int_a^b |\beta(t) - \beta(t)|\, dt$ is calculated. The table also con-

tains the average value of $\widehat{S}$ as well as the relative frequency of the event $\widehat{S} = S$. Estimation errors for the estimator $\widetilde{\beta}_r$ from Section D.1 (denoted as IV) are contained in the last two columns of the table. In these columns the finite sample behavior of the mean absolute errors of $\widetilde{\beta}_r/\widehat{M} - \beta_r$ is depicted where an estimator $\widehat{M}$ for the constant M appearing in Theorem 2.5 is used. Estimation of the constant M by $\widehat{M}$ is performed by using 100 repetitions of another Monte Carlo study and setting $N = 50000$ as well as $p = 1000$.

In order to match an estimate $\widehat{\tau}_j$ to a point of impact τ_1 or τ_2, the interval $[0, 1]$ is divided into $I_1 = [0, 0.5]$ and $I_2 = (0.5, 1]$. The estimate $\widehat{\tau}_j$ 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" point of impact τ_r and a missing value when computing averages. For $n = 100$ and $n = 200$ a trimmed mean is applied using a small trim level of 0.025 and 0.001 respectively.

It can be seen from Table D.1 that the performance of our estimates is essentially independent of p and all estimation errors decrease swiftly as the sample size n increases. While estimators for the points of impact are quite accurate for all sample sizes one needs to keep in mind that simulations were set up in a way such that there exists a j_r such that $t_{j_r} = \tau_r$. Hence, especially for larger sample sizes, there is a fairly high probability that $\widehat{\tau}_r = \tau_r$. The estimator for τ_2 performs slightly better than the estimate for τ_1. This can be seen as a consequence of the fact that $|\beta_1| < |\beta_2|$, which leads to a weaker signal for estimating τ_1.

Overall performance of our parameter estimates is slightly better in the case of $\beta(t) \equiv 0$ for all sample sizes n. However, it is important to note that the performance of the parameter estimators for α, β_1 and β_2 and the slope function $\beta(t)$ can, for example, not be interpreted independently from the estimated number of points of impact $\widehat{S}$ and the relative frequency of the event $\widehat{S} = S$. Independent of the accuracy of the estimate for the points of impact, simulations which resulted in $\widehat{S} < S$ clearly lead to an omitted variable bias in the corresponding parameter estimates and in particularly lead to parameter estimates which try to mitigate the effect of this omission. As sample size increases we have $\widehat{P}(\widehat{S} = S) \to 1$ and all estimators converge to their true value. Note that by construction, the instrumental variable estimator $\widehat{\beta}_r$ would not be affected from an omission bias. However, the estimator $\widehat{\beta}_r$ has several downsides. First by converging to $M\beta_r$ it is an biased estimate for β_r where (besides this simulation study) M is unknown. Secondly, as it can be supported from Table D.1, convergence of $\widehat{\beta}_r$ to $M\beta_r$ is much slower than convergence of $\widehat{\beta}_r$ to β_r which is in consensus with the theory.

Table D.1: *Estimation errors for different sample sizes for the simulation study. (OU-process,* $\tau_1 = 1/3$, $\tau_2 = 0.2/3$, $\beta_1 = -4$, $\beta_2 = 5$*).*

Sample Sizes		**Parameter Estimates**									
p	n	$\lvert\widehat{\tau}_1-\tau_1\rvert$	$\lvert\widehat{\tau}_2-\tau_2\rvert$	$\lvert\widehat{\alpha}-\alpha\rvert$	$\lvert\widehat{\beta}_1-\beta_1\rvert$	$\lvert\widehat{\beta}_2-\beta_2\rvert$	$\int\lvert\widehat{\beta}-\beta\rvert$	$\widehat{S}$	$\widehat{P}(\widehat{S}=S)$	$\lvert\widetilde{\beta}_1/\widehat{M}-\beta_1\rvert$	$\lvert\widetilde{\beta}_2/\widehat{M}-\beta_2\rvert$
Simulation results if $\beta(t)\equiv 0$											
100	100	0.0078	0.0058	0.24	1.35	1.69	5.06	0.97	0.20	1.37	0.80
	200	0.0045	0.0031	0.14	0.94	1.36	2.80	1.48	0.51	0.83	0.57
	500	0.0013	0.0006	0.06	0.48	0.68	0.65	1.88	0.88	0.47	0.44
	1000	0.0003	0.0001	0.02	0.28	0.35	0.07	1.99	0.99	0.38	0.34
	2500	0.0000	0.0000	0.01	0.16	0.19	0.00	2.00	1.00	0.30	0.27
	5000	0.0000	0.0000	0.00	0.12	0.14	0.00	2.00	1.00	0.21	0.19
	10000	0.0000	0.0000	0.00	0.08	0.10	0.00	2.00	1.00	0.18	0.17
500	100	0.0095	0.0071	0.23	1.25	1.64	4.68	1.04	0.23	1.39	0.78
	200	0.0060	0.0044	0.14	0.93	1.39	2.78	1.48	0.51	0.85	0.57
	500	0.0027	0.0018	0.06	0.55	0.77	0.60	1.89	0.89	0.46	0.44
	1000	0.0013	0.0007	0.03	0.35	0.44	0.09	1.99	0.99	0.38	0.35
	2500	0.0003	0.0001	0.01	0.19	0.23	0.01	2.00	1.00	0.30	0.27
	5000	0.0001	0.0000	0.00	0.12	0.15	0.00	2.00	1.00	0.25	0.22
	10000	0.0000	0.0000	0.00	0.08	0.10	0.00	2.00	1.00	0.20	0.18
1000	100	0.0098	0.0073	0.25	1.31	1.66	4.63	1.08	0.25	1.38	0.79
	200	0.0061	0.0048	0.14	0.93	1.40	2.66	1.50	0.53	0.85	0.58
	500	0.0029	0.0019	0.06	0.56	0.78	0.54	1.90	0.91	0.49	0.43
	1000	0.0015	0.0009	0.03	0.36	0.45	0.07	1.99	0.99	0.40	0.36
	2500	0.0005	0.0002	0.01	0.20	0.24	0.01	2.00	1.00	0.31	0.27
	5000	0.0002	0.0001	0.00	0.13	0.15	0.00	2.00	1.00	0.25	0.22
	10000	0.0000	0.0000	0.00	0.08	0.10	0.00	2.00	1.00	0.21	0.19
Simulation results if $\beta(t)\neq 0$											
100	100	0.0097	0.0061	0.26	4.94	1.81	6.97	0.58	0.03	1.94	0.77
	200	0.0043	0.0033	0.18	1.24	1.67	5.66	1.04	0.18	1.07	0.52
	500	0.0013	0.0007	0.07	0.63	1.00	2.95	1.63	0.64	0.48	0.39
	1000	0.0004	0.0001	0.03	0.36	0.48	1.09	1.93	0.93	0.42	0.32
	2500	0.0000	0.0000	0.01	0.19	0.23	0.43	2.00	1.00	0.37	0.25
	5000	0.0000	0.0000	0.00	0.13	0.16	0.29	2.00	1.00	0.26	0.18
	10000	0.0000	0.0000	0.00	0.10	0.11	0.21	2.00	1.00	0.22	0.15
500	100	0.0103	0.0078	0.26	1.55	1.79	6.69	0.66	0.05	1.84	0.73
	200	0.0058	0.0050	0.18	1.17	1.72	5.55	1.07	0.20	1.15	0.51
	500	0.0028	0.0020	0.07	0.69	1.11	2.86	1.67	0.67	0.50	0.38
	1000	0.0014	0.0008	0.03	0.43	0.57	1.17	1.93	0.93	0.42	0.32
	2500	0.0004	0.0002	0.01	0.23	0.26	0.47	2.00	1.00	0.37	0.25
	5000	0.0001	0.0000	0.00	0.14	0.16	0.30	2.00	1.00	0.30	0.20
	10000	0.0000	0.0000	0.00	0.09	0.11	0.21	2.00	1.00	0.24	0.17
1000	100	0.0103	0.0080	0.26	1.48	1.82	6.69	0.67	0.06	1.97	0.72
	200	0.0063	0.0051	0.17	1.21	1.71	5.49	1.08	0.21	1.16	0.52
	500	0.0030	0.0022	0.07	0.69	1.10	2.81	1.68	0.68	0.52	0.39
	1000	0.0016	0.0011	0.03	0.44	0.60	1.18	1.93	0.93	0.43	0.32
	2500	0.0006	0.0003	0.01	0.24	0.27	0.47	2.00	1.00	0.36	0.25
	5000	0.0002	0.0001	0.00	0.15	0.18	0.31	2.00	1.00	0.30	0.20
	10000	0.0000	0.0000	0.00	0.10	0.11	0.21	2.00	1.00	0.25	0.17

Graphical illustration of the results from the simulation study

In addition to Table D.1, more detailed information about the results of the simulation study can be derived from Figure D.1 (for the case $\beta(t) \equiv 0$) and Figure D.2 (for the case $\beta(t) \neq 0$). The figures capture the estimation error of our estimates for the different constellations of n and p via boxplots. While estimators derived using the basis truncation approach from Section D.2 are summarized by TRH-BIC, estimators derived from the method of instrumental variables estimation from Section D.1 are assigned the abbreviation IV. For clarity of the representation, the whiskers of the boxplots were chosen to represent the 10% and 90% quantiles. Both figures highlight the findings from the results of Table D.1, illustrating the well behaved asymptotical convergence of all estimators for both cases of the slope function $\beta(t)$ and all choices of p.

In the upper right panel of Figure D.2 one might wonder about the lower ends of the boxplots for the estimation error of $\int_a^b \widehat{\beta}(t) - \beta(t)\, dt$ for smaller sample sizes which have an approximate value of -5. But since in our setting we have $\int_0^1 \beta(t)\, dt \approx -5$, the lower ends of the boxplots then can be seen to correspond to the case in which the TRH-BIC procedure falsely selected a model consisting only of points of impact (for instance, a model with P set to 0).

Finally, Figure D.1 and Figure D.2 illustrate that the small sample bias of $\widetilde{\beta}_r/\widehat{M}$ and $\widehat{\beta}_r$ points in different directions. For instance $\widetilde{\beta}_1/\widehat{M}$ underestimates β_1 while $\widehat{\beta}_1$ tends to overestimates β_1. For larger sample sizes, the instrumental variable estimator $\widetilde{\beta}_r$ suffers from a much larger variation when compared to $\widehat{\beta}_r$ while for smaller sample sizes it seems to be less variable than $\widetilde{\beta}_r$. But keep in mind that the estimation error of $|\widetilde{\beta}_r/\widehat{M} - \beta_r|$ is illustrated. In practice however an estimator $\widehat{M}$ for the unknown constant M is to the best of our knowledge not available and a misspecification of M might result in a huge bias.

CASE $\beta(t) \equiv 0$: ESTIMATION ERRORS FOR DIFFERENT SAMPLE SIZES n.

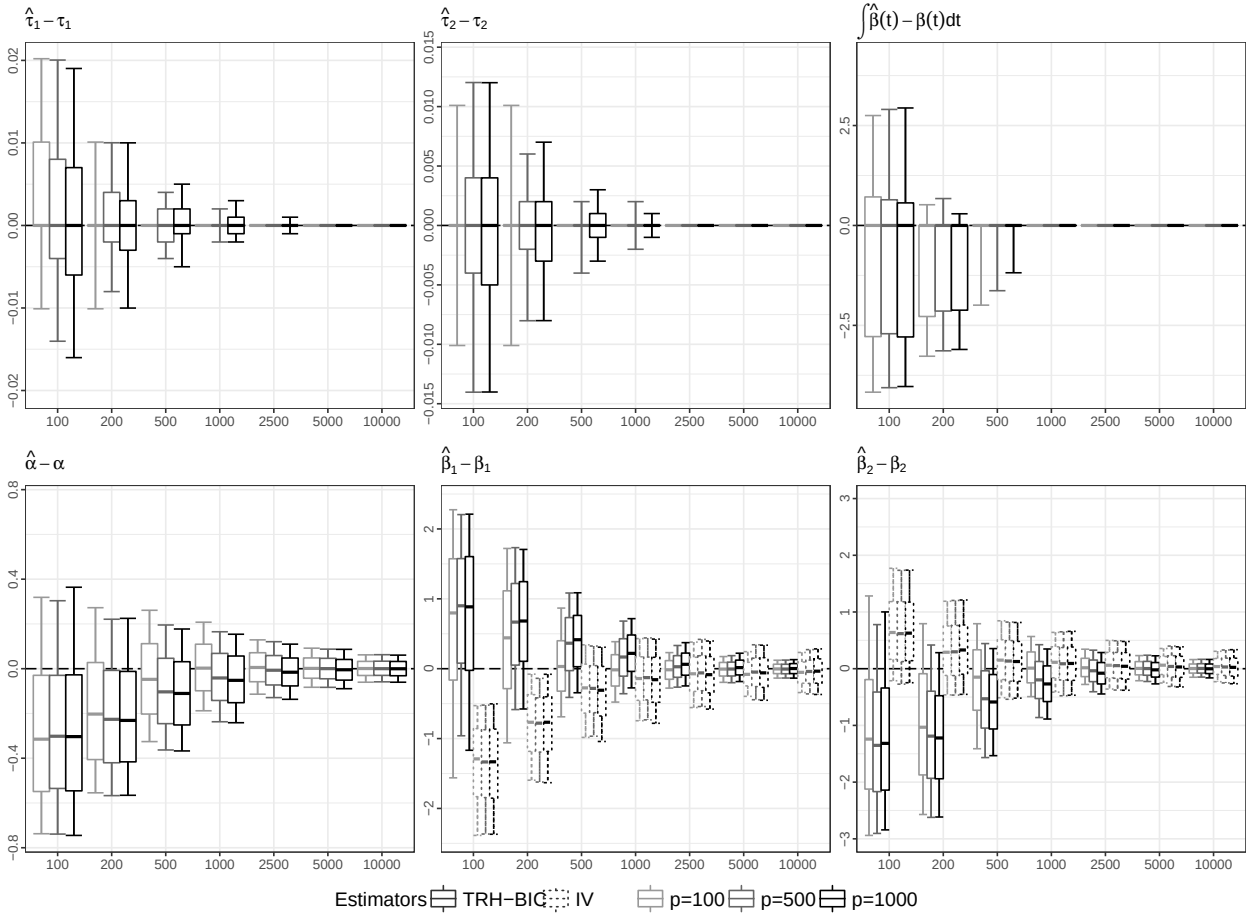

Figure D.1: *Comparison of the estimation errors in a model with $\beta(t) \equiv 0$ from using the basis truncation approach TRH-BIC (solid lines) and our instrumental variables method IV (dashed lines). The error bars of the boxplots are set to the* 10% *and* 90% *quantiles.*

CASE $\beta(t) \neq 0$: ESTIMATION ERRORS FOR DIFFERENT SAMPLE SIZES n.

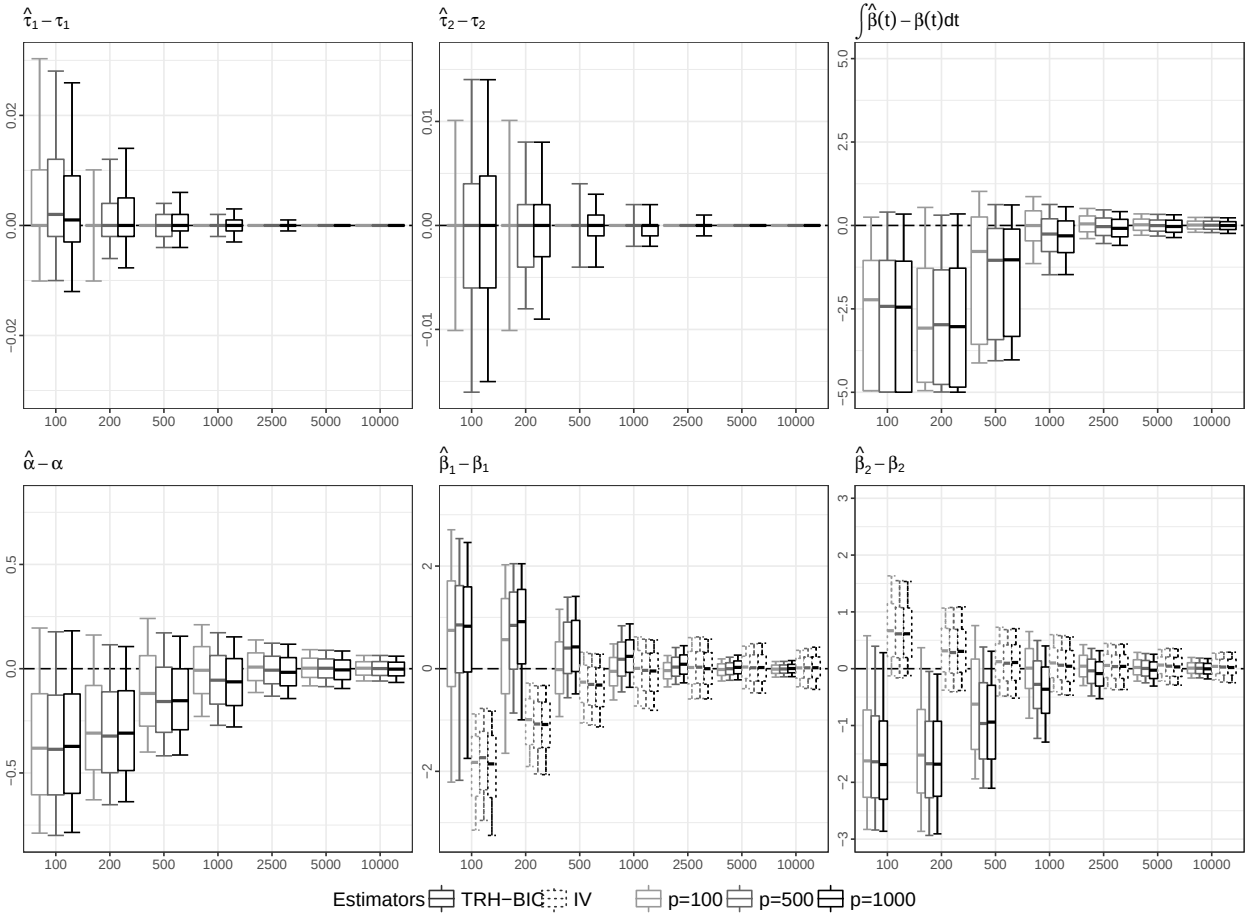

Figure D.2: *Comparison of the estimation errors in a model with $\beta(t) \neq 0$ from using the basis truncation approach TRH-BIC (solid lines) and our instrumental variables method IV (dashed lines). The error bars of the boxplots are set to the* 10% *and* 90% *quantiles.*

D.4 Proofs of the theoretical results from Appendix D

This section covers the additional proofs of Theorem 2.5 and Theorem 2.6 from Appendix D.

Proof of Theorem 2.5. Let $r \in \{1,\dots,S\}$ and $Y_i^{(r)} := Y_i - M\beta_r X_i(\tau_r)$, where the constant M is given by Lemma 2.1. By the Definition of $\widetilde{\beta}_r$ we have

$$\widetilde{\beta}_r = M\beta_r + \frac{M\beta_r \sum_{i=1}^n Z_{\delta,i}(\widehat{\tau}_r)(X_i(\tau_r) - X_i(\widehat{\tau}_r))}{\sum_{i=1}^n Z_{\delta,i}(\widehat{\tau}_r)X_i(\widehat{\tau}_r)} + \frac{\sum_{i=1}^n Z_{\delta,i}(\widehat{\tau}_r)Y_i^{(r)}}{\sum_{i=1}^n Z_{\delta,i}(\widehat{\tau}_r)X_i(\widehat{\tau}_r)}. \tag{D.8}$$

Remember that by (C.3) we obtain for every $r = 1,\dots,S$:

$$\frac{1}{n}\sum_{i=1}^n (X_i(\tau_r) - X_i(\widehat{\tau}_r))^2 = O_p(n^{-1}). \tag{D.9}$$

Note that by (B.4) and (B.5) there exist a constants $0 < L_1 < \infty$ such that (with probability converging to 1) $|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(\widehat{\tau}_r)^2| \le L_1\delta^\kappa$. Using (D.9) the Cauchy-Schwarz inequality then yields:

$$\begin{aligned} |\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(\widehat{\tau}_r)(X_i(\tau_r) - X_i(\widehat{\tau}_r))| &\le \sqrt{(\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(\widehat{\tau}_r)^2)(\frac{1}{n}\sum_{i=1}^n (X_i(\tau_r) - X_i(\widehat{\tau}_r))^2)} \\ &= O_p(\delta^{\kappa/2} n^{-\frac{1}{2}}). \end{aligned} \tag{D.10}$$

Similar arguments used to derive (B.4) and (B.5) may now be used to show that for some constants $0 < L_2 < \infty$ and $0 < L_3 < \infty$ with probability converging to 1

$$0 < L_2\delta^\kappa \le \inf_{t\in[a+\delta,b-\delta]} |\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(t)X_i(t)| < \sup_{t\in[a+\delta,b-\delta]} |\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(t)X_i(t)| \le L_3\delta^\kappa. \tag{D.11}$$

At the same time, it follows from the definition of $Y_i^{(r)}$, Lemma 2.1, Lemma B.3 and by (A.6) in Kneip et al. (2016a) that there exists a constant $0 < L_4 < \infty$ such that

$$\sup_{t\in[\tau_r-\delta/2,\tau_r+\delta/2]} |\mathbb{E}(Z_{\delta,i}(t)Y_i^{(r)})| \le L_4\delta^{\min\{2,\kappa+1\}}. \tag{D.12}$$

Moreover, since $|\widehat{\tau_r} - \tau_r| = O_p(n^{-1/\kappa})$ and $\delta^\kappa n \to \infty$, we have $P(\widehat{\tau}_r \in [\tau_r - \delta/2, \tau_r + \delta/2]) \to 1$ as $n \to \infty$.
Furthermore, $\frac{1}{n}\sum_{i=1}^n (Z_{\delta,i}(\tau_r)Y_i^{(r)} - \mathbb{E}(Z_{\delta,i}(\tau_r)Y_i^{(r)})) = O_p(\sqrt{\frac{\delta^\kappa}{n}})$. Finally, arguments similar as in the proof of Propostion (C.1), can be used to show that we have

$$\sup_{t-\delta/2\le u\le t+\delta/2} |\frac{1}{n}\sum_{i=1}^n [(Z_{\delta,i}(t) - Z_{\delta,i}(u))Y_i^{(r)} - \mathbb{E}((Z_{\delta,i}(t) - Z_{\delta,i}(u))Y_i^{(r)})]| = O_p(\sqrt{\frac{\delta^\kappa}{n}}).$$

When combining these arguments with (D.12), we can conclude that

$$\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(\widehat{\tau}_r)Y_i^{(r)} = O_p(\sqrt{\frac{\delta^\kappa}{n}} + \delta^{\min\{2,\kappa+1\}}). \tag{D.13}$$

The assertion of the Theorem now follows from (D.8) - (D.13). □

We note that the proof is similar to a proof which was already given for a similar assertion in an early and unpublished version of Kneip et al. (2016a) for the case $g(x) = x$, i.e. for the case of a functional linear regression model with points of impact.

Proof of Theorem 2.6. Note that the boundedness of $|\sigma^{2'}(\cdot)|$ together with the boundedness of $|g'(\cdot)|$ and the Gaussian assumption on $X_i(t)$ in particular implies that $\mathbb{E}(|\sigma^2(g(\eta(\boldsymbol{\beta}_0)))|^p) < \infty$ for all $p \geq 1$. The linear predictor $\eta_i(\boldsymbol{\beta})$ is given by $\beta_0 + \sum_{r=1}^{S}\beta_r X_i(\tau_r) + \sum_{j=1}^{P}\alpha_j\theta_{ij}$, where $\theta_{ij} = \int_a^b X_i(t)\gamma_j(t)\,dt$. Generalizing the arguments used in the proof of Propostion (C.1) we see that additional to (C.6) and (C.7) we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i(t^*)(X_i(\widehat{\tau}_r) - X_i(\tau_r))\varepsilon_i' f(\eta_i(\boldsymbol{\beta}_0)) = O_p(n^{-1}),$$

as well as

$$\frac{1}{n}\sum_{i=1}^{n} (X_i(\widehat{\tau}_r) - X_i(\tau_r))\varepsilon_i' f(\eta_i(\boldsymbol{\beta}_0)) = O_p(n^{-1}).$$

for any bounded function f and all $t^* \in [a,b]$. Moreover, Assumption D.1 e) together with (C.9) now guarantees that there exists a constant $L_1 < \infty$, which is independent of $P = P_n$, such that $|\mathbb{E}((X_i(\tau_r + sq) - X_i(\tau_r))\eta_i(\boldsymbol{\beta}_0))| \leq L_1 s^{\min\{1,\kappa\}}$. With this observation one can derive similar to (C.4) and (C.5) that

$$\frac{1}{n}\sum_{i=1}^{n} (X_i(\widehat{\tau}_r) - X_i(\tau_r))f(\eta_i(\boldsymbol{\beta}_0)) = O_p(n^{-\min\{1,1/\kappa\}})$$

as well as

$$\frac{1}{n}\sum_{i=1}^{n} X_i(t^*)(X_i(\widehat{\tau}_r) - X_i(\tau_r))f(\eta_i(\boldsymbol{\beta}_0)) = O_p(n^{-\min\{1,1/\kappa\}})$$

for any bounded function f and all $t^* \in [a,b]$. By replacing ε with $\varepsilon'\sigma(g(\eta(\boldsymbol{\beta}_0)))$ and since $|h(\eta_i(\boldsymbol{\beta}_0))\sigma(g(\eta_i(\boldsymbol{\beta}_0)))|$ as well as $|h'(\eta_i(\boldsymbol{\beta}_0))\sigma(g(\eta_i(\boldsymbol{\beta}_0)))|$ are bounded, it then follows from the same steps as in the proof of Proposition C.2 that (C.19), remains still valid, i.e. we have

$$\frac{1}{n}\widehat{\mathbf{U}}(\boldsymbol{\beta}_0) = \frac{1}{n}\mathbf{U}(\boldsymbol{\beta}_0) + O_p(n^{-\min\{1,1/\kappa\}}). \tag{D.14}$$

While another examination of the steps used in the proof of Proposition C.2 leads to

$$\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0) = \frac{1}{n}\mathbf{F}_n(\boldsymbol{\beta}_0) + O_p(n^{-1/2}), \tag{D.15}$$

as well as $\mathbb{E}(\mathbf{R}(\boldsymbol{\beta}_0)) \to 0$. The lines of the proof of Theorem 2.3 together with the assumptions on p_n can now be used to show that assertion (C.58) still holds: i.e. the asymptotically prevailing term is given by

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \left(\frac{\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)}{n}\right)^{-1} \frac{\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0)}{\sqrt{n}}.$$

But by D.14 and D.15 we have

$$\left(\frac{\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)}{n}\right)^{-1} \frac{\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0)}{\sqrt{n}} = \left(\frac{\mathbf{F}_n(\boldsymbol{\beta}_0)}{n}\right)^{-1} \frac{\mathbf{U}_n(\boldsymbol{\beta}_0)}{\sqrt{n}} + o_p(1).$$

Implying that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \left(\frac{\mathbf{F}_n(\boldsymbol{\beta}_0)}{n}\right)^{-1} \frac{\mathbf{U}_n(\boldsymbol{\beta}_0)}{\sqrt{n}}. \tag{D.16}$$

Since Assumption D.1 f) - h) corresponds exactly to assumptions (M2)-(M4) from Müller and Stadtmüller (2005) (accounted for S additional covariates), by (D.16) one may then continue to follow the proof given in Müller and Stadtmüller (2005, Section 7). (Note however the slightly different definition of the matrix $\mathbf{D}$). Indeed, by (D.16) we are in a generalized functional linear model setup where besides the intercept α, a fixed number of S of additional covariates have entered the model. The arguments from Müller and Stadtmüller (2005) then lead to

$$\frac{n(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \mathbb{E}(\mathbf{F}(\boldsymbol{\beta}_0))(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - (P_n + 1)}{\sqrt{2(P_n + 1)}} \to N(0,1), \tag{D.17}$$

where $P_n + 1 = S + P + 1$ denotes the total number of parameters to be estimated. □

Chapter 3

Analysis of juggling data: Registering data to principal components to explain amplitude variation

The paper considers an analysis of the juggling dataset based on registration. An elementary landmark registration is used to extract the juggling cycles from the data. The resulting cycles are then registered to functional principal components. After the registration step the paper then lays its focus on a functional principal component analysis to explain the amplitude variation of the cycles. More results about the behavior of the juggler's movements of the hand during the juggling trials are obtained by a further investigation of the principal scores.

3.1 Introduction

Functional Principal Component Analysis (FPCA) approximates a sample curve $f(t)$ as a linear combination of orthogonal basis functions $\gamma_j(t)$ with coefficients θ_j:

$$f(t) \approx \sum_{j=1}^{L} \gamma_j(t)\theta_j. \tag{3.1}$$

The principal components γ_j have the best basis property: for any fixed number L of orthogonal basis functions, the expected total squared lose is minimized. The choice of L is up to the operator, depending what accuracy is needed. It is often possible to describe the essential parts of the variations of functional data by looking only at a usually very small set of principal components and the corresponding principal scores θ_j.

However, if the curves have phase variation, even the most elementary tools of any data analysis like the pointwise mean or variance will not be able to describe the data adequately

Ramsay and Silverman (2005). In such a case not only are more principal components needed to describe the same amount of variation in the data, but also further analysis based on principal components will become more difficult to interpret. In order to analyze the juggling data, we use a registration procedure introduced by Kneip and Ramsay (2008) in which the principal components are the features which are aligned. The juggling data is a nice application, because the data set contains many problems that have to be solved using different strategies.

After registering the data in Section 3.2, we perform a FPCA on the individual juggling cycles in Section 3.2.1. In Section 3.2.2 we examine the evolution of the scores of the juggling cycles over the trials where we additionally take the information from the warping functions into account. Section 3.3 summarizes our findings.

3.2 Registering the juggling data

During our analysis we are especially interested in the juggling cycles. We will use the following notation: for $t \in [0,1]$ let $f(t) = (f_x(t), f_y(t), f_z(t))$ be the spatial coordinates of a typical juggling cycle, $\mu(t) = \mathbb{E}(f(t))$ their structural mean and $\gamma_j(t) = (\gamma_{x,j}(t), \gamma_{y,j}(t), \gamma_{z,j}(t))$ be a typical principal component. We refer to chapter 8.5 of Ramsay and Silverman (2005) for an instruction on how to calculate the principal components in our multivariate case in practice. Referred to Ramsay et al. (2014), a juggling cycle is observed on the "clock time scale" which is the "juggling time" t transformed by a warping function h. As usual, we assume h to be an element of the space $\mathscr{H}$ of strictly increasing continuous functions. We hence observe

$$f[h(t)] = \mu[h(t)] + \sum_{j=1}^{\infty} \gamma_j[h(t)]\theta_j, \tag{3.2}$$

where $\theta_j = \int_0^1 \gamma_{x,j}(u)f_x(u) + \gamma_{y,j}(u)f_y(u) + \gamma_{z,j}(u)f_z(u)\,du$.

Note that by stating equation (3.2), we met the natural assumption that time and therefore also the warping function has to be the same in all three directions by introducing a common h function for all three spatial dimensions. In contrast to Ramsay et al. (2014) where the tangential velocity function is used to avoid the problem of facing three spatial dimensions at once, we will work in the original three dimensional coordinate system. By doing so we hope to find effects which are only observable within the raw data. We approach the registration of the cycles with a two stage procedure by performing what we call "macro" and "micro" warping. By macro warping we mean a very basic registration. The purpose of this registration step is to normalize the overall juggling speed such that we can properly extract the cycles from each trial. We adjusted the data for the different numbers of cycles per trial by trimming each trial down to the first 10 juggling cycles. In order to preserve as much information of the

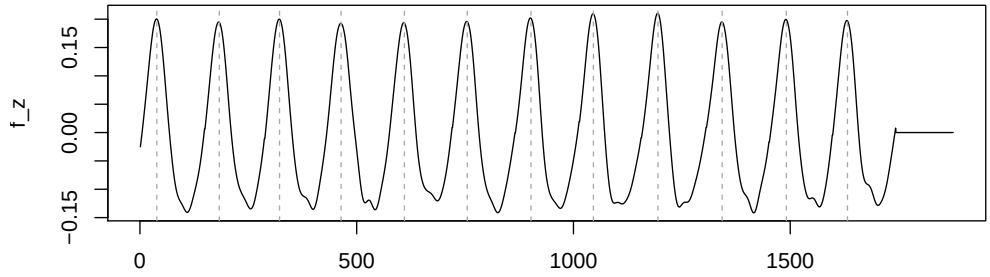

Figure 3.1: *A random trial along the z direction together with the chosen landmarks.*

cycles as possible for further analysis, we chose the simplest possible landmark registration which consists only of one landmark per cycle located at the local maxima occurring along the z-direction and a linear interpolation of the h function between. Since we only select one landmark per cycle, identifying it can be done very quickly.

The next step is to cut of all cycles at the landmarks such that we end up with a set of data consisting of a total of 100 cycles. This cropping implies that each of the cycles starts when one of the balls leaves the hand of the juggler to go up in the air in a high arc as seen in Figure 3.1. During the "micro" step, we register all 100 cycles simultaneously. By doing this we perform a very precise warping on the cycles. This is in fact a more difficult task than the "macro" warping part, because a lot of different features in the cycle curves have to be taken into account. To clarify this point we displayed a random sample of 20 cycles in Figure 3.2. It is seen from Figure 3.2 that the data needs more than just one principal component to be explained accurately. For example, by looking at the first half of this random sample along the x direction (left plot in the figure), we see variation which is obviously not induced by phase variation. Also a closer look at the middle part in the z direction (right plot) reveals a lot of variation which can not be explained by amplitude variation of a single component. Situations where we encounter more complex amplitude variations are well suited for the registration

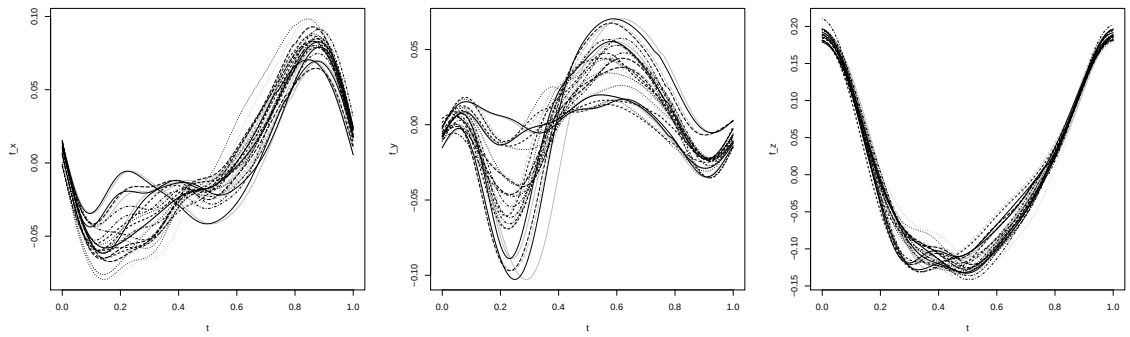

Figure 3.2: *The figure shows a random sample of* 20 *cycles for the x, y and z direction. Registered curves are displayed black, corresponding unregistered curves grey.*

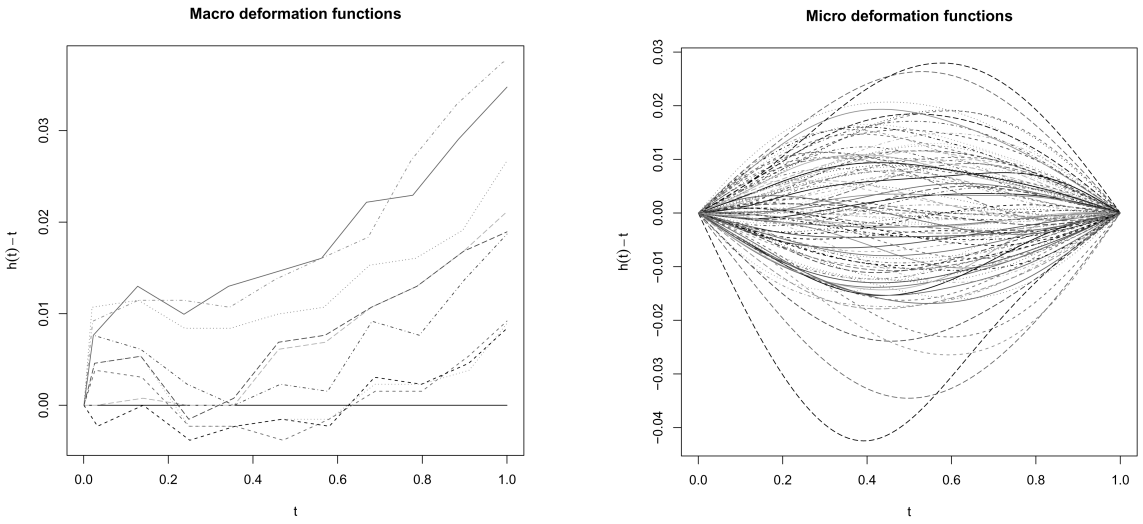

Figure 3.3: *The deformation functions estimate during the macro- and microwarping.*

method presented in Kneip and Ramsay (2008). This procedure has another advantage because it allows to control the intensity of the micro warping due to the smoothing parameter in equation (16) of Kneip and Ramsay (2008).

The method can be easily adapted to the multivariate case. Let D be the derivative operator, then a straightforward modification of equation (15) of Kneip and Ramsay (2008) now becomes

$$SSE(\widetilde{h}) = \int_0^1 \sum_{k=(x,y,z)} \{f_k(u) - f_k[h^{-1}(u)] - Df_k[h^{-1}(u)]\widetilde{h}(u)\}^2 \, du \tag{3.3}$$

which has to be minimized over $\widetilde{h} \in \mathscr{H}$. Finding a common warping function for multivariate data can easily be handled by using (3.3) for the SSE part occurring in the procedure of Kneip and Ramsay (2008).

The result of our alignment is shown as the black curves in Figure 3.2 where we registered the curves to 3 principal components. We observe that after the warping procedure the main features along all directions are well aligned. By looking at the first half of the left plot of Figure 3.2 one can observe the complexity of the juggling cycles along the x direction: If the cycles would belong to a one dimensional space (i.e. all cycles were random shifts from a mean curve), then all features would have been aligned. However, a more complex model underlies the data along this direction and any attempt to force the data to fit in a simpler model will destroy the intrinsic features of the data; the alleged shift we are observing after the registration is in fact a part of the data. The warping functions for our alignment are displayed in Figure 3.3 through the deformation functions $h(t) - t$ obtained from the macro and micro step. Note that the deformation functions for the macro step do not end at a value

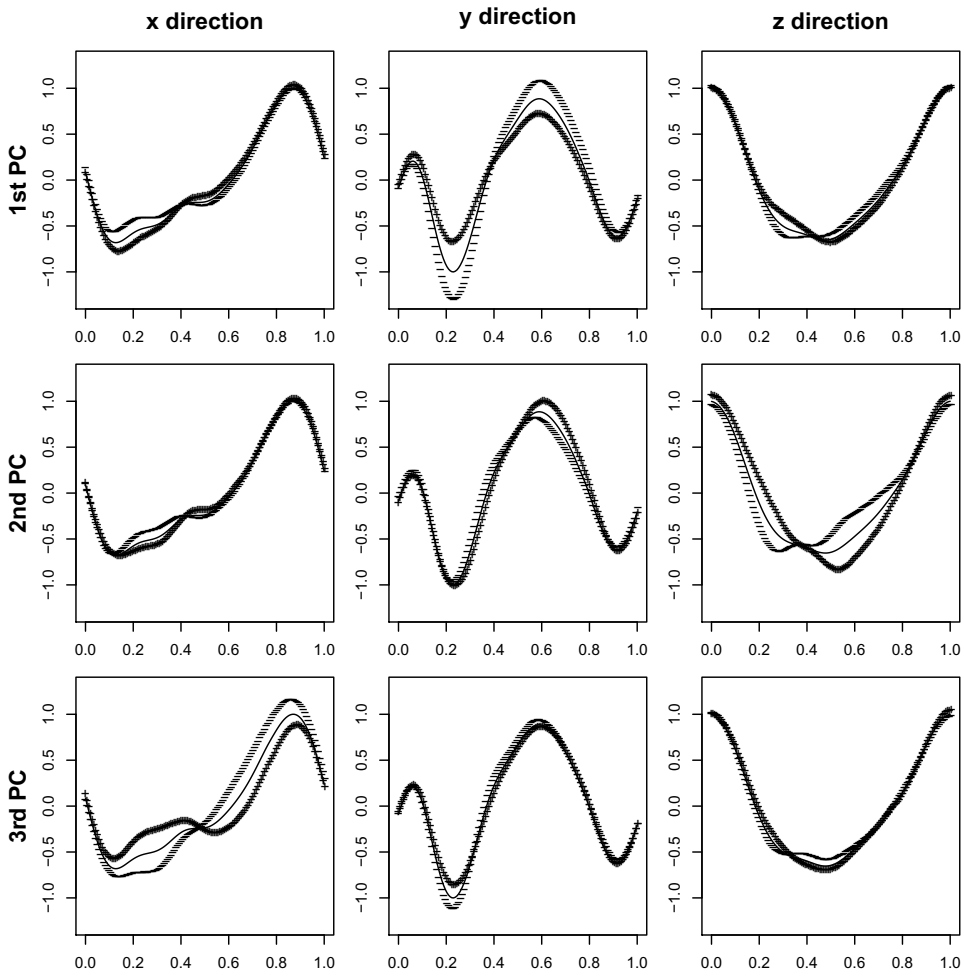

Figure 3.4: *The Figure shows the effect of adding or subtracting a multiple of each of the principal components to the scaled mean curves. The columns are the spatial directions x, y, z and the rows represent the first, second and third principal component respectively.*

of 0 since we only displayed the part of the warping functions corresponding to the first 10 cycles within the trials.

3.2.1 Analyzing the principal components

After the preprocessing steps we get suitable data to perform a FPCA. We chose to use three components to represent the data, which explain more than 80 percent of the total variance. The impact of the three principal components on each of the spatial directions of the data is displayed in Figure 3.4 where we also pictured the effect of adding and subtracting a multiple of each of the principal components to max-normalized mean curves. A closer look at Figure 3.4 reveals that the first component mainly explains the amplitude variation of the y direction while the second component explains mainly the z direction and the third compo-

Table 3.1: *Variation of the j-th principal component due to the l-th spatial direction*

	Spatial direction		
Principal Component	x	y	z
1st	0.117	**0.793**	0.091
2nd	0.053	0.185	**0.762**
3rd	**0.851**	0.100	0.049

nent the x direction. While the effect of the first component of the movement of the jugglers hand along the x and z direction only accounts for a small shift in the beginning of the movement (the catch phase) it has an important impact for the variation across the y direction. By looking at the impact of the first component along the y direction we can see that, if the ball coming in at low arch during the catch phase is juggled right in front of the juggler, then he will overcompensate for this movement by throwing the next ball from a much greater distance to himself. Such an compensation effect can also be seen for the second component along the z direction and for the the third component along the x direction. While for the y direction the latter two components mainly adjust for the two bumps, which are influenced by the first component, individually.

The importance of the components for the three directions is summarized in Table 3.1, where we capture the variability in the j-th principal component which is accounted for by the variation in the l-th direction. More formally: for a typical principal component γ we necessarily have $\int_0^1 \gamma_x^2(u)\,du + \int_0^1 \gamma_y^2(u)\,du + \int_0^1 \gamma_z^2(u)\,du = 1$. And hence each of the summands can be interpreted to give the proportion of the variability of the component which is accounted for by the spatial direction. It is seen from the table that the y direction contributes 80% of the variation of the first component while the z and x direction can be accounted for the variation of the second and third component respectively. These values reveal that the directions are somewhat independent in the way that each principal component represents mainly a single direction. These observations where only possible by keeping the data multivariate and not analyzing the tangential velocity function.

3.2.2 Analyzing the principal scores

If we perform activities like juggling several times, we expect something like a learning effect to happen. For a juggler this effect could be measured by the behavior of his hands along the directions, i.e. as the juggler gets more and more used to the juggling, one would expect the movements to be more efficient or at least the executions of the movements should become more homogeneous. By performing a FPCA we prepare our data for further statistical analysis which support us to answer such claims. This analysis will be performed on the scores.

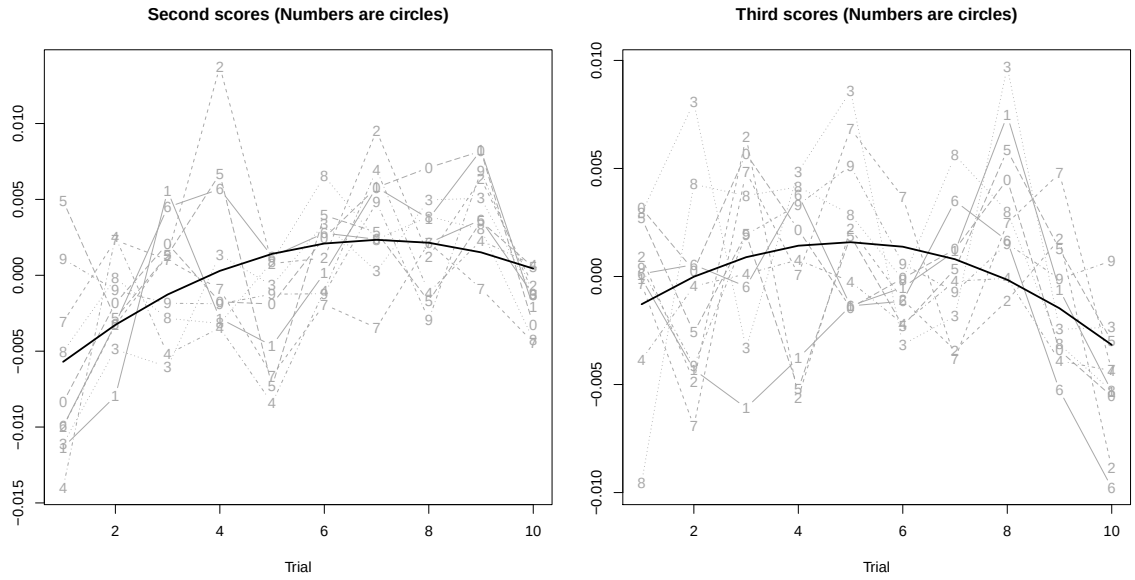

Figure 3.5: *The figure shows the evolution of the scores for the cycles corresponding to the second and third principal component over the ten trials. The solid black line represents the estimated regression function when we impose a quadratic model.*

Figure 3.5 shows the evolution of the scores corresponding to the second and third principal component over the ten trials. A typical principal score θ can be modeled as a function depending on trial $k = 1, \ldots, 10$ and number of cycle $i = 1, \ldots, 10$. Figure 3.5 suggests that a polynomial regression model can capture the main message of the data. i.e. we assume

$$\theta(i,k) = \alpha_0 + \alpha_1 k + \alpha_2 k^2 + \epsilon_i. \tag{3.4}$$

Table 3.2 contains the coefficients resulting from this regression. Before we interpret the results, recall that the first component explains mostly the y direction which is on one hand less complex in terms of its variability and on the other hand is less important for a juggler. Indeed, one could imagine a perfect juggling machine which would keep this direction constant such that a juggling cycle could be described by looking solely at the x and z directions. Now, the non-significant coefficients in the first row of Table 3.2 indicate that the movement across the y direction can not be explained by the trials. This is reasonable as one would expect that an experienced juggler mainly focuses about the movement in the other two directions and any variation of his movement along the y direction from a constant value should be random.

By the significance of the coefficients of the regressions for the scores corresponding to the second and third principal component, we can conclude that there exists indeed an evolution of the scores over the trials which can essentially be described by our regression. This evolution can be regarded as some kind of a "learning effect". For example, in Figure 3.5 we can see that the scores will have a small value at the peak of our regression function, implying that in this

Table 3.2: *Least squares coefficients obtained from a quadratic regression of the scores on the trials. Significance codes are added in parentheses where 0 '***'; 0.001 '**'; 0.01 '*'; 1 ' '*

	Parameter estimates		
Scores	α_0	α_1	α_2
1st	-0.0040 ( )	0.0014 ( )	-0.0001 ( )
2nd	-0.0086 (***)	0.0031 (***)	-0.0002 (***)
3rd	-0.0029 (*)	0.0018 (**)	-0.0002 (***)

area the variation of the movement of the jugglers hand is not very high and has to be close to the mean curve. This can be seen as an improvement in his juggling skills. Interestingly, the slope of the regression function decreases at the end. While this effect is subsidiary for the second principal score and could be seen as a nuisance from the simple quadratic model, it is apparent in the evolution of the scores corresponding to the third component.

Recall that the second component mainly quantifies the variation of the jugglers hand movement along the z-direction, which captures the up- and downwards movement of his hand. A negative score in the beginning of the trials indicates that he lunges out too far before throwing the ball up in the air. As the regression function for the scores of the second component approaches values close to zero, the "learning effect" becomes visible: getting used to the juggling in the later trials, he performs almost identical movements along this direction.

If we take a more precise look at the regression function of the scores corresponding to the third component, an interpretation is somewhat more complicated as we experience a significant downward slope at the last trials. Maybe the juggler gets fatigued or the behavior is caused by some kind of a psychological effect, i.e. the concentration of the juggler decreases as he knows that he only has to perform a few more trials and gets more impatient.

Taking a look at the time frame around 0.2–0.5 of the the bottom left panel of Figure 3.4, we see that a particular small value of the third component implies that his hand for catching the ball coming in from a low arch is comparable moved towards the other hand. Possibly e is learning to simplify the process of catching the ball coming in from low arch. Unfortunately this implies that he has to wind up more in order to throw the ball leaving in high arch.

We were further interested in an analysis of the warping functions themselves which was the reason to perform only a very basic "macro" warping. In this special kind of data set it is not reasonable to assume that the warping function is only a nuisance parameter because the speed of juggling might have an effect on the manner of the juggling.

To check this hypothesis we performed some further analysis on the warping functions. Note that we can not perform a FPCA on the warping functions directly, because we can not guarantee that the resulting curves are still elements of $\mathscr{H}$, i.e. strictly monotonic functions.

Instead we pursue the following way out. It is well known from Ramsay and Silverman (2005) that any function $h \in \mathscr{H}$ can be represented as

$$h(t) = \int_0^t e^{W(u)}\, du,$$

where $W(t) = log[Dh(t)]$ itself is an unrestricted function. In order to analyze the warping functions h appropriately, we can use the unrestricted functions $W(t)$. We approximate $W(t)$ by using the first two principal components which explain more than 95 Percent of the variations in $W(t)$ and define by $\theta_{W,1}$, $\theta_{W,1}$ a typical scores corresponding to these two components.

In Table 3.3 we computed the correlation between the scores of W and θ. We can determine that the speed a juggling cycle is performed with has nearly no influence on the first component of a cycle. But this speed does have an effect on the second and third component which explain mostly the x and z direction. Obviously, this effect is occurs mainly through the first component of W.

Table 3.3: *The table shows the correlation between the scores corresponding to the first two components of W and the scores corresponding to the first three components of the juggling cycles*

	Scores of the cycles		
Scores of W	$\boldsymbol{\theta_1}$	$\boldsymbol{\theta_2}$	$\boldsymbol{\theta_3}$
$\theta_{W,1}$	-0.0120	0.3044	-0.2351
$\theta_{W,2}$	-0.0122	0.0355	0.0013

Another interesting result occurs by computing the correlation between the scores of the principal components of W and and the residuals resulting from the polynomial regression in (3.4). It reveals a significant amount of correlation between these variables, i.e. a not negligible part of the residuals from (3.4) can be explained by the juggling speed of the cycles. Moreover, running a regression of the scores of the warping function W on the trials showed no significant coefficient. From this we can conclude that, what we identified as a learning effect, has no significant impact on the warping for a specific cycle. We hence can identify two effects which influence the scores of a juggling cycle. The first is due to learning and the second is a result which is related to the specific warping. The effects are modeled by augmenting equation (3.4) by

$$\theta(i,k) = \alpha_0 + \alpha_1 k + \alpha_2 k^2 + \beta_1 \theta_{W,1,i} + \beta_2 \theta_{W,2,i} + \epsilon_i, \tag{3.5}$$

where $\theta_{W,j,i}$ is the score of the i-th cycle corresponding to the j-th principal component of the function W. Estimated coefficients are given in Table 3.4, from where it can be seen that neither the speed the juggling cycles are performed with, nor the trials have an impact on the

movement of the jugglers hand along the y direction. Moreover, it can be seen that there is a connection between the scores of a juggling cycles and the speed of the juggling.

Table 3.4: *The table shows the results from a regression of the cycle scores on the trial number, squared trial number as well as the scores from W with corresponding coefficients β_1 and β_2. Significance codes are added in parentheses where 0 '***'; 0.001 '**'; 0.01 '*'; 1 ' '*

	Parameter estimates				
Scores	α_0	α_1	α_2	β_1	β_2
1st	-0.0042 ()	0.0014 ()	-0.0001 ( )	-0.0009 ()	0.0000()
2nd	-0.0081 (***)	0.0030 (***)	-0.0002 (***)	0.0034 (**)	0.0025 ()
3rd	-0.0033 (*)	0.0019 (***)	-0.0002 (***)	-0.0027 (*)	-0.0009 ()

3.3 Summary

We analyzed the juggling data by combining two registration methods. First we used an elementary landmark registration in order to crop the individual juggling cycles, which were the focus of our analysis. In order to perform a refined warping of the juggling cycles in a second step, we generalized the registration method from Kneip and Ramsay (2008) to the multivariate nature of the data. We analyze the registered data by performing a FPCA using three principal components where we observed that each of the components essentially quantified the variation across a single spatial direction.

More specific information about the behavior of the juggler is contained in the scores which we studied in dependence on the trials. By doing so, we were able to identify some kind of learning effect over the trials. The movement of the jugglers hand for throwing a ball up in the air levels out over the trials. After applying an alignment procedure one should not forget about the warping functions. Interpreting the warping functions can not only be a very interesting task for itself, but they can contain important additional information which can be helpful to analyze the data.

Bibliography

Aneiros, Germán and Philippe Vieu (2014). "Variable selection in infinite-dimensional problems." *Statistics & Probability Letters*, 94, 12–20. [54]

Berrendero, José R., Beatriz Bueno-Larraz, and Antonio Cuevas (2017). "An RKHS model for variable selection in functional regression." *arXiv preprint arXiv:1701.02512*. [54]

Bickel, Peter J., Ya'acov Ritov, and Alexandre B. Tsybakov (2009). "Simultaneous analysis of lasso and Dantzig selector." *The Annals of Statistics*, 37 (4), 1705–1732. [13]

Bosq, Denis (2000). *Linear processes in function spaces*. Vol. 149. Lecture Notes in Statistics. Theory and applications. Springer. [4, 5]

Brillinger, David R. (2012a). "A generalized linear model with "Gaussian" regressor variables." In *Selected Works of David Brillinger*. Springer, 589–606. [58, 78, 89, 90, 109]

Brillinger, David R. (2012b). "The identification of a particular nonlinear time series system." In *Selected Works of David Brillinger*. Springer, 607–613. [58]

Cai, T. Tony and Peter Hall (2006). "Prediction in functional linear regression." *The Annals of Statistics*, 34 (5), 2159–2179. [4]

Calcagno, Vincent (2013). *glmulti: Model selection and multimodel inference made easy*. R package version 1.0.7. [67]

Cardot, Hervé, Frédéric Ferraty, and Pascal Sarda (1999). "Functional linear model." *Statistics & Probability Letters*, 45 (1), 11–22. [4]

Cardot, Hervé and Jan Johannes (2010). "Thresholding projection estimators in functional linear models." *Journal of Multivariate Analysis*, 101 (2), 395–408. [4]

Cardot, Hervé, André Mas, and Pascal Sarda (2007). "CLT in functional linear regression models." *Probability Theory and Related Fields*, 138 (3-4), 325–361. [4]

Comte, Fabienne and Jan Johannes (2012). "Adaptive functional linear regression." *The Annals of Statistics*, 40 (6), 2765–2797. [4]

Crambes, Christophe, Alois Kneip, and Pascal Sarda (2009). "Smoothing splines estimators for functional linear regression." *The Annals of Statistics*, 37 (1), 35–72. [4]

Dagsvik, John K. and Steinar Strøm (2006). "Sectoral labour supply, choice restrictions and functional form." *Journal of Applied Econometrics*, 21 (6), 803–826. [56]

Delaigle, Aurore and Peter Hall (2012). "Methodology and theory for partial least squares applied to functional data." *The Annals of Statistics*, 40 (1), 322–352. [4]

Embrechts, Paul and Makoto Maejima (2000). "An introduction to the theory of self-similar stochastic processes." *International Journal of Modern Physics B*, 14 (12n13), 1399–1420. [57]

Fahrmeir, Ludwig and Heinz Kaufmann (1985). "Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models." *The Annals of Statistics*, 13 (1), 342–368. [100–102]

Ferraty, Frédéric, Peter Hall, and Philippe Vieu (2010). "Most-predictive design points for functional data predictors." *Biometrika*, 97 (4), 807–824. [4, 54]

Ferraty, Frédéric and Philippe Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. 1. Springer Series in Statistics. Springer. [54]

Floriello, Davide and Valeria Vitelli (2017). "Sparse clustering of functional data." *Journal of Multivariate Analysis*, 154, 1–18. [54]

Frank, Ildiko E. and Jerome H. Friedman (1993). "A Statistical View of Some Chemometrics Regression Tools." *Technometrics*, 35, 109–135. [4]

Fredrickson, Barbara L. (2000). "Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions." *Cognition & Emotion*, 14 (4), 577–606. [71]

Gillespie, Daniel T. (1996). "Exact numerical simulation of the Ornstein-Uhlenbeck process and its integral." *Phys. Rev. E*, 54 (2), 2084–2091. [20]

Hall, Peter and Joel L. Horowitz (2007). "Methodology and convergence rates for functional linear regression." *The Annals of Statistics*, 35 (1), 70–91. [2, 4, 19, 28, 29]

Hall, Peter and Mohammad Hosseini-Nasab (2006). "On properties of functional principal components analysis." *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 68 (1), 109–126. [28]

He, Guozhong, Hans-Georg Müller, and Jane-Ling Wang (2000). "Extending correlation and regression from multivariate to functional data." *Asymptotics in statistics and probability*, 301–315. [5]

Horváth, Lajos and Piotr Kokoszka (2012). *Inference for Functional Data with Applications*. Vol. 200. Springer. [54]

Hsing, Tailen and Randall Eubank (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons. [54]

Hsing, Tailen and Haobo Ren (2009). "An RKHS formulation of the inverse regression dimension-reduction problem." *The Annals of Statistics*, 37 (2), 726–755. [5]

James, Gareth M., Jing Wang, and Ji Zhu (2009). "Functional linear regression that's interpretable." *The Annals of Statistics*, 37 (5A), 2083–2108. [4]

Kneip, Alois, Dominik Poß, and Pascal Sarda (2016a). "Functional linear regression with points of impact." *The Annals of Statistics*, 44 (1), 1–30. [1, 54, 56, 58, 60, 61, 78, 82, 86, 89, 109, 110, 119, 120]

Kneip, Alois, Dominik Poß, and Pascal Sarda (2016b). "Supplement to: Functional linear regression with points of impact." *The Annals of Statistics*. [6, 78–80, 83–85, 88]

Kneip, Alois and James O. Ramsay (2008). "Combining Registration and Fitting for Functional Models." *Journal of the American Statistical Association*, 103 (483), 1155–1165. [2, 124, 126, 132]

Kneip, Alois and Pascal Sarda (2011). "Factor models and variable selection in high-dimensional regression analysis." *The Annals of Statistics*, 39 (5), 2410–2447. [17]

Kokoszka, Piotr and Reimherr Matthew (2017). *Introduction to Functional Data Analysis*. 1. Texts in Statistical Science. Chapman & Hall/CRC. [54]

Lee, Charles and Mark J. Ready (1991). "Inferring trade direction from intraday data." *The Journal of Finance*, 46 (2), 733–746. [56]

Levina, Elizaveta, Amy S. Wagaman, Andrew F. Callender, Gurjit S. Mandair, and Michael D. Morris (2007). "Estimating the number of pure chemical components in a mixture by maximum likelihood." *Journal of Chemometrics*, 21 (1-2), 24–34. [56]

Lindquist, Martin A. and Ian W. McKeague (2009). "Logistic regression with Brownian-like predictors." *Journal of the American Statistical Association*, 104 (488), 1575–1585. [54, 68]

Mauss, Iris B., Robert W. Levenson, Loren McCarter, Frank H. Wilhelm, and James J. Gross (2005). "The tie that binds? Coherence among emotion experience, behavior, and physiology." *Emotion*, 5 (2), 175. [71]

McCullagh, Peter (1983). "Quasi-likelihood functions." *The Annals of Statistics*, 11 (1), 59–67. [64]

McCullagh, Peter and John A. Nelder (1989). *Generalized Linear Models*. 2nd ed. Monographs on Statistics & Applied Probability (37). Chapman and Hall/CRC. [63]

McKeague, Ian W. and Bodhisattva Sen (2010). "Fractals with point impact in functional linear regression." *The Annals of Statistics*, 38 (4), 2559–2586. [4–6, 10, 54]

Müller, Hans-Georg and Ulrich Stadtmüller (2005). "Generalized functional linear models." *The Annals of Statistics*, 33 (2), 774–805. [2, 4, 108, 111–113, 121]

Park, Ah Yeon, John A. D. Aston, and Frédéric Ferraty (2016). "Stable and predictive functional domain selection with application to brain images." *arXiv preprint arXiv:1606.02186*. [54]

Poß, Dominik and Heiko Wagner (2014). "Analysis of juggling data: Registering data to principal components to explain amplitude variation." *Electronic Journal of Statistics*, 8 (2), 1825–1834. [2]

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. [67]

Ramsay, James O., Paul Gribble, and Sebastian Kurtek (2014). "Description and processing of functional data arising from juggling trajectories." *Electronic Journal of Statistics*, 8 (2), 1811–1816. [124]

Ramsay, James O. and Bernard W. Silverman (2005). *Functional Data Analysis*. 2. Springer Series in Statistics. Springer. [54, 124, 131]

Rohlfs, Rori V., Patrick Harrigan, and Rasmus Nielsen (2013). "Modeling gene expression evolution with an extended Ornstein-Uhlenbeck process accounting for within-species variation." *Molecular biology and evolution*, 31 (1), 201–211. [56]

Schubert, Emery (1999). "Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space." *Australian Journal of Psychology*, 51 (3), 154–165. [71]

Stein, Charles M. (1981). "Estimation of the mean of a multivariate normal distribution." *The Annals of Statistics*, 9 (6), 1135–1151. [58, 89, 110]

Stein, Michael L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer. [57]

Trautmann, Sina Alexa, Thorsten Fehr, and Manfred Herrmann (2009). "Emotions in motion: Dynamic compared to static facial expressions of disgust and happiness reveal more widespread emotion-specific activations." *Brain Research*, 1284, 100–115. [71]

Van de Geer, Sara and Johannes Lederer (2013). "The Bernstein-Orlicz norm and deviation inequalities." *Probability Theory and Related Fields*, 157 (1), 225–250. [39, 80, 81, 89]

Van der Vaart, Aad W. and Jon A. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer. [39, 89, 91]

Wang, Jane-Ling, Jeng-Min Chiou, and Hans-Georg Müller (2016). "Functional data analysis." *Annual Review of Statistics and Its Application*, 3, 257–295. [54]

Wehrens, Ron (2011). *Chemometrics With R: Multivariate Data Analysis in the Natural Sciences and Life Sciences*. Springer. [33]

Zhang, Yulei (2012). "Sparse selection in Cox models with functional predictors." PhD thesis. [54]

Zhou, Shuheng, John Lafferty, and Larry Wasserman (2010). "Time varying undirected graphs." *Machine Learning*, 80 (2-3), 295–319. [38]

Zhou, Shuheng, Sara van de Geer, and Peter Bühlmann (Mar. 2009). "Adaptive Lasso for High Dimensional Regression and Gaussian Graphical Modeling." *ArXiv e-prints*. arXiv: 0903.2515. [13, 38]