Points of Impact in Generalized Linear Models with Functional Predictors

Dominik Poß[*], Dominik Liebl[*], Hedwig Eisenbarth[†], Tor D. Wager[‡], and Lisa Feldman Barrett[§]

September 19, 2018

Abstract

Generalized linear models with function-valued predictor variables and scalar outcomes belong to the well-established and widely used methodological toolbox in functional data analysis. In many applications, however, only specific locations or time-points of the functional predictors have an impact on the outcome. The selection of such "points of impact" constitutes a particular variable selection problem, since the high correlation in the functional predictors violates the basic assumptions of existing high-dimensional variable selection procedures. In this paper we introduce a generalized linear regression model with functional predictors evaluated at unknown points of impact which need to be estimated from the data alongside the model parameters. 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 estimators of the linear model parameters. The finite sample properties of our estimators are assessed by means of a simulation study. Our methodology is motivated by a psychological case study in which the participants were asked to continuously rate their emotional state while watching an affective online video on the persecution of African albinos. Accompanying supplementary materials are available online.

Keywords: functional data; variable selection; quasi-maximum likelihood; emotional stimuli; online video rating

[*]Department of Statistics, University of Bonn, Bonn, Germany

[†]Department of Psychology, University of Southampton, Southampton, United Kingdom

[‡]Department of Psychology and Neuroscience and Institute of Cognitive Science, University of Colorado Boulder, Boulder, Colorado, USA

[§]Department of Psychology, Northeastern University and Department of Psychiatry, Massachusetts General Hospital/Harvard Medical School, Boston, Massachusetts, USA; and Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, Massachusetts, USA

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

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

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. The values $\tau_1, \tau_2, \ldots, \tau_S$ are called points of impact and give specific locations at which the functional predictor 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 points of impact by analyzing a genetic data set, where they aim to determine a single genetic locus ($S = 1$) that allows for distinguishing 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. Their method, however, is restrictive in that it only allows for a single point of impact. A point of impact model, where $S = 1$ is assumed known, has also been studied in survival analysis for the Cox regression model (Zhang, 2012).

The special case where $g(x) = x$ is the identity function is considered in other research. McKeague and Sen (2010) consider a functional linear regression model with a single ($S = 1$) point of impact and derive the distribution of their estimates in the case where X is a fractional Brownian motion. Kneip et al. (2016) consider also a functional linear regression model with points of impact, but allow for an unknown number $S \geq 1$. Aneiros and Vieu (2014) consider a points of impact model, but their theory postulates the existence of some 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 special case where $g(x) = x$.

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 the most predictive design points minimizing a overall cross validation criterion. Floriello and Vitelli (2017) propose a method for sparse clustering of functional data. In a slightly different context, Park et al. (2016) focus on selecting predictive subdomains of the functional data. These papers, however, do not focus on parameter estimation, which is of central interest in our work. Related to this paper is also the work of Lindquist (2012) and Sobel and Lindquist (2014). Lindquist (2012) extends structural equation models to the functional data analysis setting and uses his methodology to select significantly impacting time-intervals in functional magnetic resonance imaging (fMRI) data. Sobel and Lindquist (2014) propose a mixed effects model which facilitates selecting significant impact regions in fMRI data by controlling for systematic measurement errors. 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).

In many applications only specific points of impact matter and the rest of the data is ignorable for some reason. For example, stock market events occur that can have long-lasting impact on prices. As another example, emotional experiences may be influenced by events at specific moments.

This paper is motivated by a case study from psychology in which participants were asked to continuously rate their emotional state (from very negative to very positive) while watching an affective documentary video on the persecution of African albinos (see Figure 1). After the video, the participants were asked to rate their overall emotional state. Psychologists are interested in understanding how such concluding overall ratings relate to the fluctuations of the emotional states while watching the video, as this has implications for the way emotional states are assessed in research using such material. 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 (see Section 6). This heuristic approach, however, results in non-interpretable and insignificant results. By contrast, our methodology allows to identify the crucial affective video scenes– the basic prerequisite to understand the emergence of emotional states in this kind of experiments.

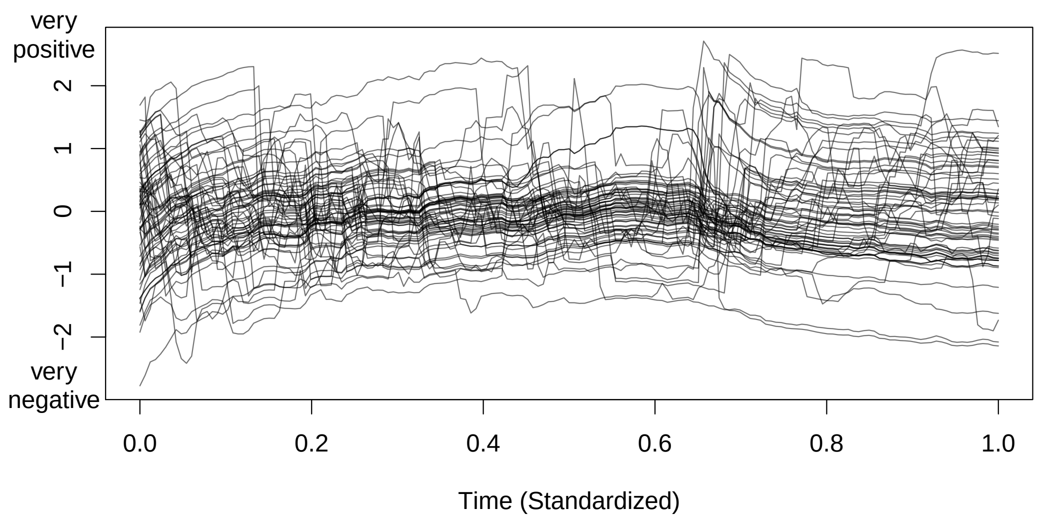

Figure 1: Continuously self-reported emotion trajectories X of $n = 67$ participants.

The rest of this work is structured as follows. Section 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 3. Section 4 proposes a practical data-driven implementation of the estimation procedure. Our simulation study is contained in Section 5 and Section 6 contains our real data application. All proofs and additional simulation results can be found in the supplement supporting this article. The accompanying R-package GFLMPOI contains implementations of the estimation procedures.

2 Estimating Points of Impact

The selection of points of impact constitutes a particular variable selection problem. Since in practice the functional predictor is observed over a densely discretized grid, one may tend to apply multivariate variable selection methods like Lasso or related procedures. The high correlation of the predictor at neighboring discretization points, however, violates the basic requirements of these multivariate variable selection procedures (see Kneip et al., 2016, Sec. 4 for more formal arguments). In this section we present the theoretical framework for estimating the points of impact $\tau_1, \dots, \tau_S$. 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), $i = 1, \dots, n$, 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 is a real valued random variable. It is assumed that the relationship between Y_i and the functional predictor X_i can be modeled as

$$Y_i = g\left(\alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r)\right) + \varepsilon_i, \tag{2}$$

in which the error term ε_i respects $\mathbb{E}(\varepsilon_i | X_i(t)) = 0$ for all $t \in [a, b]$. The number S, the points of impact $\tau_1, \dots, \tau_S$ and the parameters $\alpha, \beta_1, \dots, \beta_S$ are unknown and have to be estimated from the data. The points of impact $\tau_1, \dots, \tau_S$ indicate the locations at which the functional values $X_i(\tau_1), \dots, X_i(\tau_S)$ have a specific influence on Y_i. The inclusion of a constant α allows us to consider centered random functions X_i with $\mathbb{E}(X_i(t)) = 0$ for all $t \in [a, b]$. Denoting the linear predictor

$$\eta_i = \alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r) \tag{3}$$

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

In our application, we are primarily interested in 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 the points of impact τ_r and their number 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 mild regularity conditions. We emphasize that contrary to a generalized linear model setup we do not need to assume that the conditional distribution of Y_i given X_i belongs to the exponential family. We only require that the conditional expectation of Y_i given X_i is linked to the linear predictor η_i through some function g which extends the class of possible applications.

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 motion, Ornstein-Uhlenbeck processes, Poisson processes, etc. These processes also have important applications in fields such as finance, chemometrics, econometrics, and 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 on the covariance function of X_i 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 $s, t \in [a,b]$*

$$\sigma(s,t) = \omega(s,t,|s-t|^{\kappa}). \tag{4}$$

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 (4) with $\kappa = 2$, a very small value of κ will indicate a process with non-smooth sample paths. See Kneip et al. (2016) 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 $c_1 > 0$ and some exponent $H > 0$. 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 $c_3 > 0$. The Assumption 2.1 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 those 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. Stein, 1999, Ch. 2.7, p. 32). Assumption 2.1 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 $\sigma(s,t) = \omega(s,t,|s-t|) = 0.5\,\exp(-\theta|s-t|)\sigma_{OU}^2/\theta$, 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 following lemma is important for our later results. It facilitates estimating the points of impact τ_r under a large class of possibly unknown functions g.

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 for all $s \in [a,b]$, $\mathrm{Cov}\big(X_i(s) - \widetilde{\eta}_i \mathbb{E}(X_i(s)\widetilde{\eta}_i)/\mathbb{V}(\widetilde{\eta}_i), g(\eta_i)\big) = 0$. 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 in Lemma 2.1 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 (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. Lemma 2.1 simplifies our proofs, but the estimation of the points of impact may also be possible under more general situations; see also remark (iv) on Theorem 2.1.

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. Covariance functions $\sigma(s,t)$ satisfying this assumption are obviously not two times differentiable at the diagonal $s = t$, but two times differentiable for $s \neq t$. Under Assumption 2.1 and Lemma 2.1, the locations of the points of impact are uniquely identifiable from the cross-covariance $\mathbb{E}(X_i(s)Y_i)$. 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 $s = t$, the cross-covariance $f_{XY}(s)$ will not be two times differentiable at $s = \tau_r$, for all $r = 1, \ldots, S$, resulting in kink-like features at τ_r as depicted in the upper plot of Figure 2.

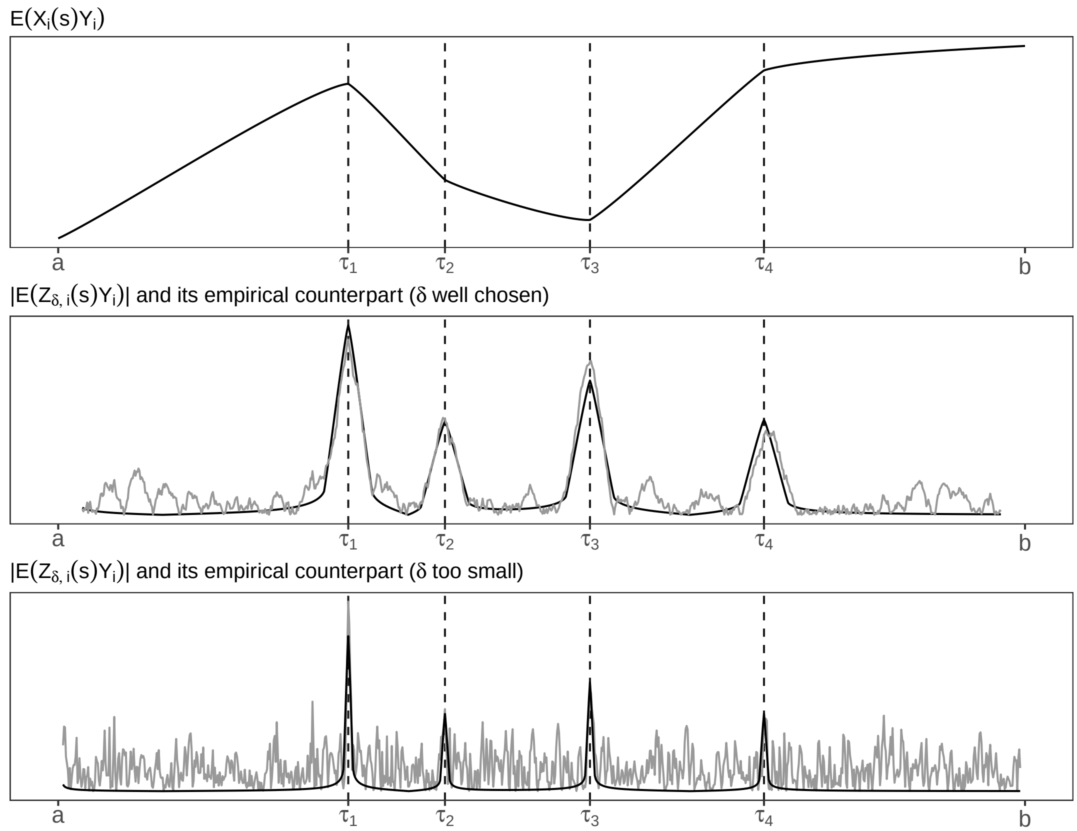

Figure 2: The upper panel shows $\mathbb{E}(X_i(s)Y_i)$ as a function of s with 4 kink-like features at the points of impact (dashed vertical lines). The two lower panels show $|\,\mathbb{E}(Z_{\delta,i}(s)Y_i)|$ (black) and their empirical counterparts (gray) for different values of δ.

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 $\delta > 0$:

$$f_{XY}(s) - \frac{1}{2}(f_{XY}(s+\delta) + f_{XY}(s-\delta)) \approx -\frac{1}{2}\delta^2 f''_{XY}(s). \tag{5}$$

Since $f''_{XY}(s)$ does not exist at $s = \tau_r$, the left hand side of (5) will decline more slowly 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. (2016) together with Lemma 2.1 that under Assumption 2.1 one obtains the following theoretical result justifying such an estimation strategy:

$$\mathbb{E}(Z_{\delta,i}(s)Y_i) = c_0\,\mathbb{E}(Z_{\delta,i}(s)\eta_i) = \begin{cases} c_0\beta_r c(\tau_r)\delta^\kappa + O(\max\{\delta^{2\kappa}, \delta^2\}) & \text{if } |s-\tau_r| \approx 0, \\ O(\max\{\delta^{\kappa+1}, \delta^2\}) & \text{if } \min_{r=1,\ldots,S} |s-\tau_r| \gg \delta, \end{cases} \tag{6}$$

where $c(\cdot)$ is as defined in Assumption 2.1.

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

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

Consequently, the identification of points of impacts requires a sensible choice of δ. For too small δ-values (e.g., $\delta^\kappa \sim n^{-1}$) the estimation noise will overlay the signal; this situation is depicted in the bottom plot of Figure 2. For too large δ-values, however, it will not be possible to distinguish between neighboring points of impact.

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, there exists a $m \geq 2$ times differentiable function $\widetilde{\sigma} : \mathbb{R} \to \mathbb{R}$ such that $\sigma(s,t) = \widetilde{\sigma}(|s-t|)$, where $\widetilde{\sigma}(|s-t|)$ 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 $|\widetilde{\sigma}''(0)| > |\widetilde{\sigma}''(|s-t|)|$, for $s \neq t$, and $\min_{r\neq k}|\tau_r - \tau_k|$ is large enough, then all points of impact might be identified as local extrema of $\mathbb{E}(Z_{\delta,i}(s)Y_i)$.

2.1 Estimation

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

Algorithm 2.1. (Estimating Points of Impact)

1. **Calculate:**
$$\widehat{f}_{XY}(t_j) := \frac{1}{n}\sum_{i=1}^n X_i(t_j)Y_i, \quad \textit{for each} \quad j = 1, \ldots, p$$

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

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

4. **Repeat:**

   Initiate *the repetition by setting* $l = 1$.

   Estimate *the* l*th point of impact candidate as*
$$\widehat{\tau}_l = \underset{t_j:\, j\in\mathcal{J}_\delta}{\arg\max}|\widehat{f}_{ZY}(t_j)|.$$

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

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

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 threshold $\lambda > 0$*.*

A practical choice of the threshold λ is discussed below of Theorem 2.1.

2.2 Asymptotic Results

In this section, 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) $X_1, \dots, X_n$ *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* $m = 1, 2, \dots$ *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 $m = 1, 2, \dots$. Condition b) also 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). If X_i satisfies condition a), then condition b) in particular holds if the errors ε_i are sub-Gaussian and if g is differentiable with a bounded derivative.

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

Theorem 2.1. *Under 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* $n\delta^\kappa / |\log \delta| \to \infty$ *as well as* $\delta^\kappa / n^{-\kappa+1} \to 0$*. We then obtain* *that*

$$\max_{r=1,\dots,\widehat{S}} \min_{s=1,\dots,S} |\widehat{\tau}_r - \tau_s| = O_P(n^{-1/\kappa}). \tag{7}$$

Moreover, there exists a constant $0 < D < \infty$ *such that when the Algorithm 2.1 is applied with threshold*

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

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

$$P(\widehat{S} = S) \ \rightarrow \ 1 \quad \textit{as} \quad n \rightarrow \infty. \tag{8}$$

Theorem 2.1 is related to Theorem 4 in Kneip et al. (2016), but differs in the choice the threshold λ. A threshold λ which performed well in our simulations is given by $\lambda = A\sqrt{\sqrt{\mathbb{E}(Y_i^4)}\log\big((b-a)/\delta\big)/n}$, where $\mathbb{E}(Y_i^4)$ is estimated by $\widehat{\mathbb{E}}(Y_i^4) = n^{-1}\sum_{i=1}^n Y_i^4$ and $A = \sqrt{2\sqrt{3}}$. This value is motivated by an argument using the central limit theorem in the derivations of the threshold for Theorem 2.1. See the remark after the proof of Lemma B.2 in Appendix B for 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 demand 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 not crucial. If, for instance, $\sigma(s,t)$ is $d > 2$ times continuously differentiable for $s \neq t$ and not being d times differentiable at $s = t$, one may look at the central difference approximation of at least the dth derivative of $\mathbb{E}(X_i(s)Y_i)$. For example, if $d = 4$ 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 (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, for instance, 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. In this case we can replace $c_0\,\mathbb{E}(Z_{\delta,i}(s)\eta_i)$ in (6) by $c_0(s)\,\mathbb{E}(Z_{\delta,i}(s)\eta_i)$ and points of impact can still be estimated, since the arguments for the estimation of the points of impact rely on $|\,\mathbb{E}(Z_{\delta,i}(s)\eta_i)|$.

3 Parameter Estimation

In the following we assume the existence of some consistent estimation procedure for the points of impact τ_r satisfying $P(\widehat{S} = S) \to 1$ and $\max_{r=1,\ldots,\widehat{S}} |\widehat{\tau}_r - \tau_r| = O_P(n^{-1/\kappa})$, where we use matched labels in the sense that $\tau_r = \arg\min_{s=1,\ldots,S} |\widehat{\tau}_r - \tau_s|$. These requirements are fulfilled by our estimation procedure described in Section 2.1, but may also be fulfilled for alternative procedures.

For estimating the parameters $\alpha, \beta_1, \ldots, \beta_S$ we impose the following additional assumptions for model (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) 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 3.1. *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*

$$\mathbb{E}\left(\left(g(\alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r)) - g(\alpha^* + \sum_{r=1}^{S^*} \beta_r^* X_i(\tau_r))\right)^2\right) > 0, \tag{9}$$

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 3.1, 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 notational 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

$$\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = \widehat{\mathbf{D}}_n(\boldsymbol{\beta})^T \widehat{\mathbf{V}}_n(\boldsymbol{\beta})^{-1}(\mathbf{Y}_n - \widehat{\boldsymbol{\mu}}_n(\boldsymbol{\beta})). \tag{10}$$

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

In the following, it will be convenient to define

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

By definition it holds that $\mathbb{E}(n^{-1}\,\mathbf{F}_n(\boldsymbol{\beta})) = [\mathbb{E}(g'(\eta_i(\boldsymbol{\beta}))^2/\sigma^2(g(\eta_i(\boldsymbol{\beta})))\, X_{ik}X_{il})]_{k,l}$ with $k, l = 1, \ldots, S+1$. Let $\eta(\boldsymbol{\beta})$ and X_j be generic copies of $\eta_i(\boldsymbol{\beta})$ and of 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 $\mathbb{E}(\mathbf{F}(\boldsymbol{\beta})) = [\mathbb{E}(g'(\eta(\boldsymbol{\beta}))^2/\sigma^2(g(\eta(\boldsymbol{\beta})))\, X_k X_l)]_{k,l}$, 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 $\mathbb{E}(\widehat{\mathbf{F}}(\boldsymbol{\beta})) = [\mathbb{E}(g'(\widehat{\eta}(\boldsymbol{\beta}))^2/\sigma^2(g(\widehat{\eta}(\boldsymbol{\beta})))\, \widehat{X}_k \widehat{X}_l)]_{k,l}$.

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

Assumption 3.1.

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

Condition 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 the supplementary Appendix C. Conditions 3.1 a) to c) hold, for example, in the important case of a functional logistic regression with points of impact. Condition 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 $h(x) = 1$.

Theorem 3.2. *Let $\widehat{S} = S$, $\max_{r=1,\dots,S} |\widehat{\tau}_r - \tau_r| = O_P(n^{-1/\kappa})$ and let X_i be a Gaussian process satisfying Assumption 2.1. Under Assumption 3.1 we then obtain*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N\big(\mathbf{0}, (\mathbb{E}(\mathbf{F}(\boldsymbol{\beta}_0)))^{-1}\big). \tag{11}$$

This result is remarkable; our estimator based on $\widehat{\tau}_r$ enjoys the same asymptotic efficiency properties as if the true points of impact τ_r were known. It achieves the same asymptotic efficiency properties under classical multivariate setups (cf. McCullagh, 1983). In practice one might 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 (48) and (74) in the supplementary Appendix C.

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. Estimating the points of impact τ_r relies on the choice of δ and a choice of the threshold parameter λ (see Section 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 generally of a limited use in practice. In the following we propose an alternative fully data-driven model selection approach.

For a given δ, our estimation procedure leads to a set of potential point of impact candidates $\{\widehat{\tau}_1, \widehat{\tau}_2, \dots, \widehat{\tau}_{M_\delta}\}$ (see Section 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),

$$\mathrm{BIC}_{\mathcal{X}}(\delta) = -2\log \mathcal{L}_{\mathcal{X}} + K_{\mathcal{X}} \log(n). \tag{12}$$

Here, $\log \mathcal{L}_{\mathcal{X}}$ is the log-likelihood of the model with intercept and predictor variables $\mathcal{X} \subseteq \{X_i(\widehat{\tau}_1), X_i(\widehat{\tau}_2), \ldots, X_i(\widehat{\tau}_{M_\delta})\}$, where $K_{\mathcal{X}} = 1 + |\mathcal{X}|$ denotes the number of predictors. Minimizing $\mathrm{BIC}_{\mathcal{X}}(\delta)$ 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 $\mathbb{V}(Y_i|X_i) = \sigma^2(g(\eta_i))$ are known, one may replace the deviance $-2\log \mathcal{L}_{\mathcal{X}}$ by the expression for the quasi-deviance $-2Q_{\mathcal{X}} = -2\sum_{i=1}^n \int_{y_i}^{g(\widehat{\eta}_{\mathcal{X},i})} (y_i - t)/(\sigma^2(t))\, dt$, where $\widehat{\eta}_{\mathcal{X},i}$ is the linear predictor with intercept and predictor variables $\mathcal{X}$.

In order to calculate $\mathrm{BIC}_{\mathcal{X}}(\delta)$, 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 $m = 1, 2, \ldots,$

$$\widehat{\boldsymbol{\beta}}_m = \widehat{\boldsymbol{\beta}}_{m-1} + \Big(\widehat{\mathbf{F}}_n(\widehat{\boldsymbol{\beta}}_{m-1})\Big)^{-1} \widehat{\mathbf{U}}_n(\widehat{\boldsymbol{\beta}}_{m-1}). \tag{13}$$

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. The procedure is implemented in the accompanying R package GFLMPOI. In compliance with our theory we allow for a large family of generalized linear models.

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, 2018), 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 4. The threshold estimator from Section 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 $\lambda = A\sqrt{\sqrt{\widehat{\mathbb{E}}(Y^4)} \log((b-a)/\delta)/n}$, where $A = \sqrt{2\sqrt{3}}$ and $\widehat{\mathbb{E}}(Y^4) = n^{-1} \sum_{i=1}^{n} Y_i^4$, as motivated below of 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 with error bars representing the 10% and 90% quantiles. The simulation study can be reproduced using the R-scripts provided as supplement supporting this article. Five data generating processes (DGP) are considered (see Table 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 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 covariance function 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 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 4 (denoted as POI) with the estimation procedure of Lindquist and McKeague (2009) (denoted as LMcK). Lindquist and McKeague (2009) consider situations where $S =$

DGP 1: Estimation errors for different sample sizes n

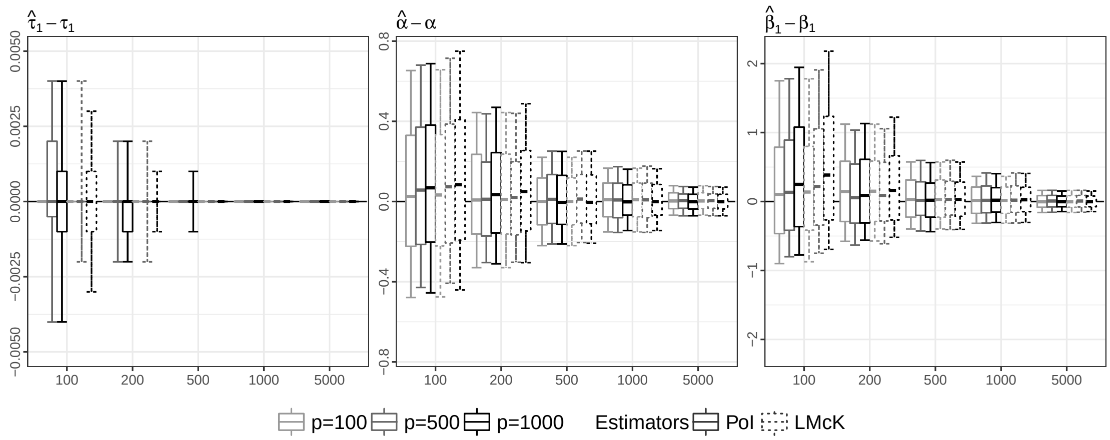

Figure 3: Comparison of the estimation errors from using our BIC-based method POI (solid lines) and the method of Lindquist and McKeague (2009) (dashed lines).

1 is known and propose estimating 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 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 $|n^{-1} \sum_{i=1}^{n} Z_{\delta,i}(s) Y_i|$ 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 3. The performance of our threshold based procedure is reported in detail for the remaining simulation studies (DGP 2-5).

DGP 2 is more complex than DGP 1, also because $S = 2$ is considered unknown. Figure 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,

DGP 2: Estimation errors for different sample sizes n

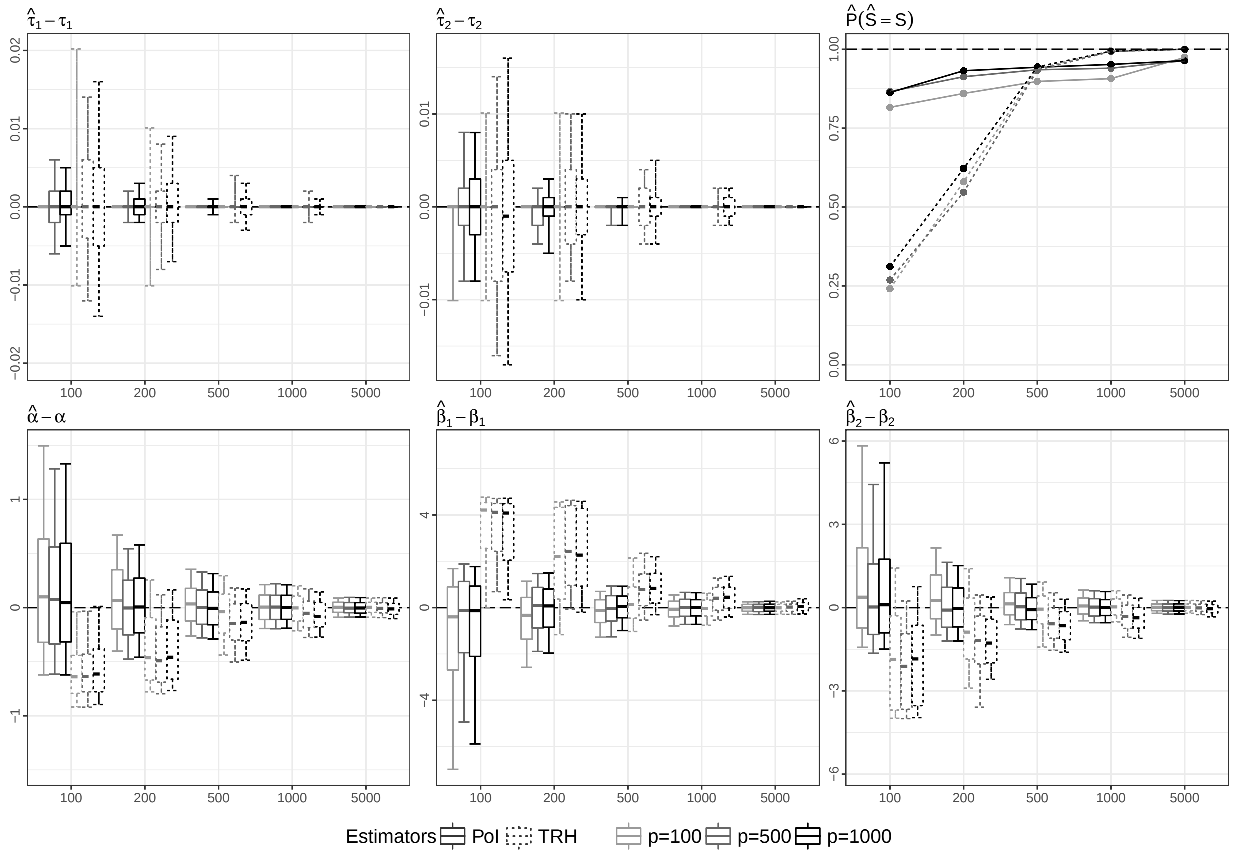

Figure 4: Comparison of the estimation errors from using our BIC-based method POI (solid lines) and our threshold-based method TRH (dashed lines).

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. For reasons of space, Figure 7 is referred to Appendix A. It shows that the qualitative results from DGP 2 still hold. For large n, the POI and TRH estimators both lead 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 2.1). For DGP 3, the choice of $c_\delta = 1.5$ results in a too large $\sqrt{\delta}$-neighborhood, such that the estimation procedure also eliminates 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 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 under this Assumption. From Figure 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 number of points of impact S even for large n, since the λ-threshold is 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

DGP 4: Estimation errors for different sample sizes n

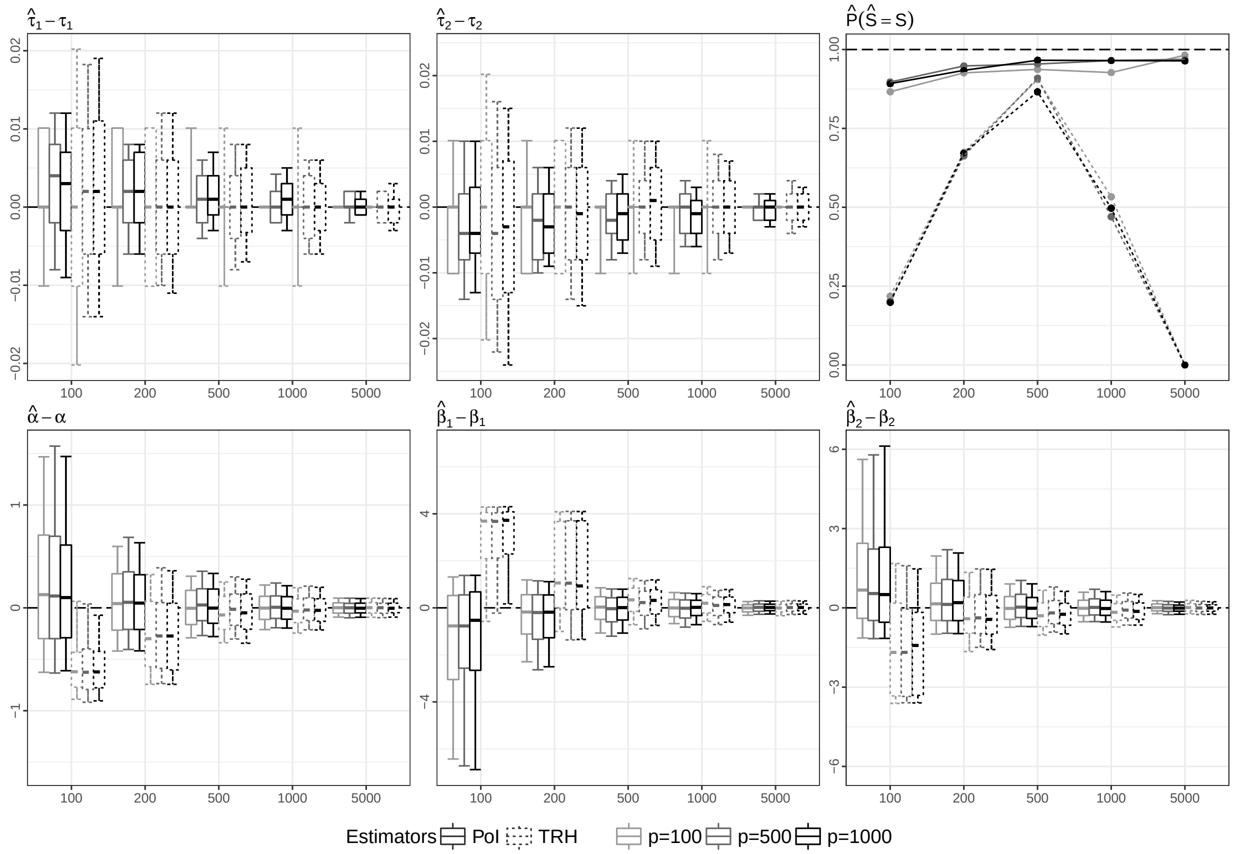

Figure 5: Comparison of the estimation errors from using our BIC-based method POI (solid lines) and our threshold-based method TRH (dashed lines).

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) 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 8 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 we keep 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 is a clear advantage when using the POI-estimator.

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 the overall rating in order to understand the emergence of emotional states and to make use of specific parts 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 vary strongly across participants, which prevents linking 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 an affective 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. A version of the video can be found online at YouTube (`www.youtube.com`)[2]. The first six data points (< 1 sec.) are removed from the trajectories as they contain some obviously erratic components.

Figure 1 shows the standardized continuously self-reported feeling trajectories $X_i(t_j)$, 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 final overall feeling. This overall rating was coded as a binary variable Y_i, where $Y_i = 0$ denotes "I feel negative" (48% of the participants) and $Y_i = 1$ denotes "I do not feel negative" (52% of the participants). 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 SoSci Survey (`www.soscisurvey.de`). The study was approved by the local institutional review board (IRB, University of Colorado Boulder). The documentary video is taken from the Interdisciplinary Affective Science Laboratory Movie Set (Feldman Barrett, L., unpublished). The real-data application can be reproduced using the R-scripts provided as supplement supporting this article.

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 $p_i^{\text{abs}} = \arg\max_t(|X_i(t)|)$

[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.

[2]Link to the video: `https://youtu.be/9F6UpuJIFaY`. The video clip used in the experiment corresponds to the first 112 sec.

Table 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

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 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 and PER-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 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 and therefore do not allow linking the overall ratings Y_i to specific common time points t in the continuous emotional stimuli. By contrast, the estimated points of impact $\widehat{\tau}_1$ and $\widehat{\tau}_2$ allow for such a link and point to two emotionally arousing text phrases spoken at those impact points: "even genitals" ($\widehat{\tau}_1$) and "selling his brother's body parts" ($\widehat{\tau}_2$).

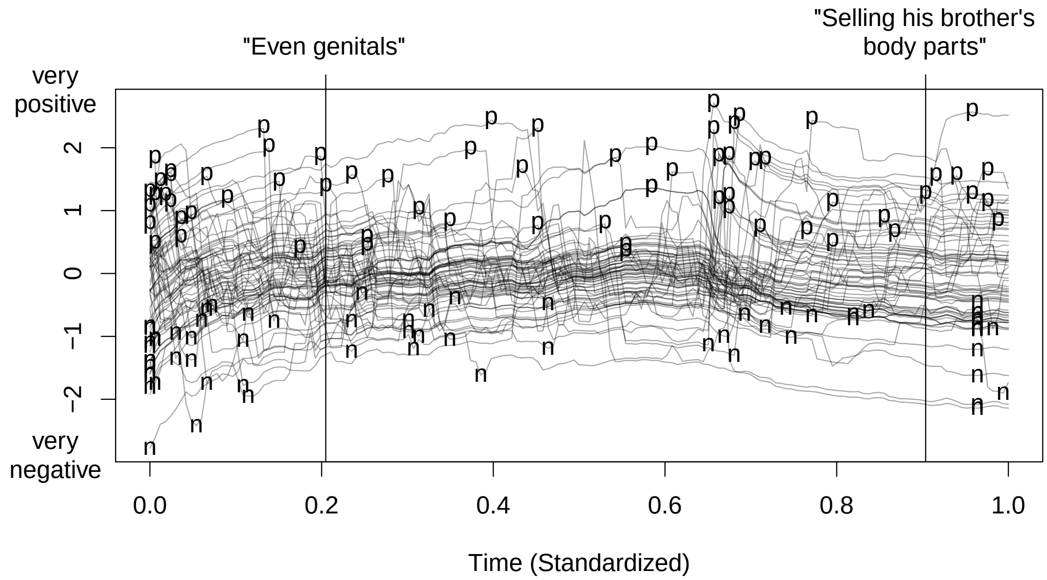

Figure 6: Visualization of the positive (p) and negative (n) peak intensity predictors $X_i(p_i^{\text{pos}})$ and $X_i(p_i^{\text{neg}})$ and the two impact points $\widehat{\tau}_1$ and $\widehat{\tau}_2$ (vertical lines) along with the corresponding text phrases from the video.

Supplementary Materials

The online supplementary material contain additional figures, proofs of the theoretical results, the R-package GFLMPOI and R-scripts for reproducing the empirical results.

Acknowledgments

The authors thank Alois Kneip (University of Bonn) for fruitful discussions and valuable comments which helped to improve this research work. The online rating tool for the data collection was kindly provided by Dominik Leiner (SoSci Survey, Germany).

Funding

Data collection was funded by the National Institutes of Health Director's Pioneer Award (DP1OD003312) to Lisa Feldman Barrett, and the National Institute on Drug Abuse grant

(R01DA035484) to Tor D. Wager. 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 (W5J9CQ-11-C-0046) to Lisa Feldman Barrett and Tor D. Wager. 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.

References

Aneiros, G. and P. Vieu (2014). Variable selection in infinite-dimensional problems. *Statistics & Probability Letters 94*, 12–20.

Berrendero, J. R., B. Bueno-Larraz, and A. Cuevas (2017). An rkhs model for variable selection in functional regression. *arXiv preprint arXiv:1701.02512*.

Brillinger, D. R. (2012a). A generalized linear model with "gaussian" regressor variables. In *Selected Works of David Brillinger*, pp. 589–606. Springer.

Brillinger, D. R. (2012b). The identification of a particular nonlinear time series system. In *Selected Works of David Brillinger*, pp. 607–613. Springer.

Calcagno, V. (2013). *glmulti: Model selection and multimodel inference made easy.* R package version 1.0.7.

Dagsvik, J. K. and S. Strøm (2006). Sectoral labour supply, choice restrictions and functional form. *Journal of Applied Econometrics 21*(6), 803–826.

Embrechts, P. and M. Maejima (2000). An introduction to the theory of self-similar stochastic processes. *International Journal of Modern Physics B 14*(12n13), 1399–1420.

Fahrmeir, L. and H. Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics 13*(1), 342–368.

Ferraty, F., P. Hall, and P. Vieu (2010). Most-predictive design points for functional data predictors. *Biometrika 97*(4), 807–824.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice* (1. ed.). Springer Series in Statistics. Springer.

Floriello, D. and V. Vitelli (2017). Sparse clustering of functional data. *Journal of Multivariate Analysis 154*, 1–18.

Fredrickson, B. L. (2000). Extracting meaning from past affective experiences: The importance of peaks, ends, and specific emotions. *Cognition & Emotion 14*(4), 577–606.

Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications*, Volume 200. Springer.

Hsing, T. and R. Eubank (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons.

Kneip, A., D. Poß, and P. Sarda (2016). Functional linear regression with points of impact. *The Annals of Statistics 44*(1), 1–30.

Kokoszka, P. and R. Matthew (2017). *Introduction to Functional Data Analysis* (1. ed.). Texts in Statistical Science. Chapman & Hall/CRC.

Lee, C. and M. J. Ready (1991). Inferring trade direction from intraday data. *The Journal of Finance 46*(2), 733–746.

Levina, E., A. Wagaman, A. Callender, G. Mandair, and M. Morris (2007). Estimating the number of pure chemical components in a mixture by maximum likelihood. *Journal of Chemometrics 21*(1-2), 24–34.

Lindquist, M. A. (2012). Functional causal mediation analysis with an application to brain connectivity. *Journal of the American Statistical Association 107*(500), 1297–1309.

Lindquist, M. A. and I. W. McKeague (2009). Logistic regression with brownian-like predictors. *Journal of the American Statistical Association 104*(488), 1575–1585.

Mauss, I. B., R. W. Levenson, L. McCarter, F. H. Wilhelm, and J. J. Gross (2005). The tie that binds? coherence among emotion experience, behavior, and physiology. *Emotion 5*(2), 175.

McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics 11*(1), 59–67.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models* (2 ed.). Monographs on Statistics & Applied Probability (37). Chapman and Hall/CRC.

McKeague, I. W. and B. Sen (2010). Fractals with point impact in functional linear regression. *The Annals of Statistics 38*(4), 2559–2586.

Park, A. Y., J. A. Aston, and F. Ferraty (2016). Stable and predictive functional domain selection with application to brain images. *arXiv preprint arXiv:1606.02186*.

R Core Team (2018). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis* (2. ed.). Springer Series in Statistics. Springer.

Rohlfs, R. V., P. Harrigan, and R. 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.

Schubert, E. (1999). Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology 51*(3), 154–165.

Sobel, M. E. and M. A. Lindquist (2014). Causal inference for fmri time series data with systematic errors of measurement in a balanced on/off study of social evaluative threat. *Journal of the American Statistical Association 109*(507), 967–976.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics 9*(6), 1135–1151.

Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer.

Trautmann, S. A., T. Fehr, and M. 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.

Wang, J.-L., J.-M. Chiou, and H.-G. Müller (2016). Functional data analysis. *Annual Review of Statistics and Its Application 3*, 257–295.

Zhang, Y. (2012). *Sparse selection in Cox models with functional predictors*. Ph. D. thesis.

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

Dominik Poß
Department of Statistics, University of Bonn

Dominik Liebl
Department of Statistics, University of Bonn

Hedwig Eisenbarth
Department of Psychology, University of Southampton

Tor D. Wager
Department of Psychology and Neuroscience and Institute of Cognitive Science, University of Colorado Boulder

and

Lisa Feldman Barrett
Department of Psychology, Northeastern University, Boston;
Department of Psychiatry, Massachusetts General Hospital/Harvard Medical School;
Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, Massachusetts

This online supplement to "Points of Impact in Generalized Linear Models with Functional Predictors" contains some additional figures for our simulations, proofs of the theoretical results and further derivations. Appendix A contains the simulation results for DGP 3 and DGP 5 from Section 5. In Appendix B proofs related to the estimation of the points of impact as presented in Section 2 can be found. Proofs for the parameter estimates from Section 3 are collected in Appendix C.

A Additional Simulation Results

This appendix contains two additional figures showing the remaining simulation results discussed in Section 5 of the main paper. While Figure 7 depicts the results from DGP 3, Figure 8 illustrates the results from DGP 5.

DGP 3: ESTIMATION ERRORS FOR DIFFERENT SAMPLE SIZES n

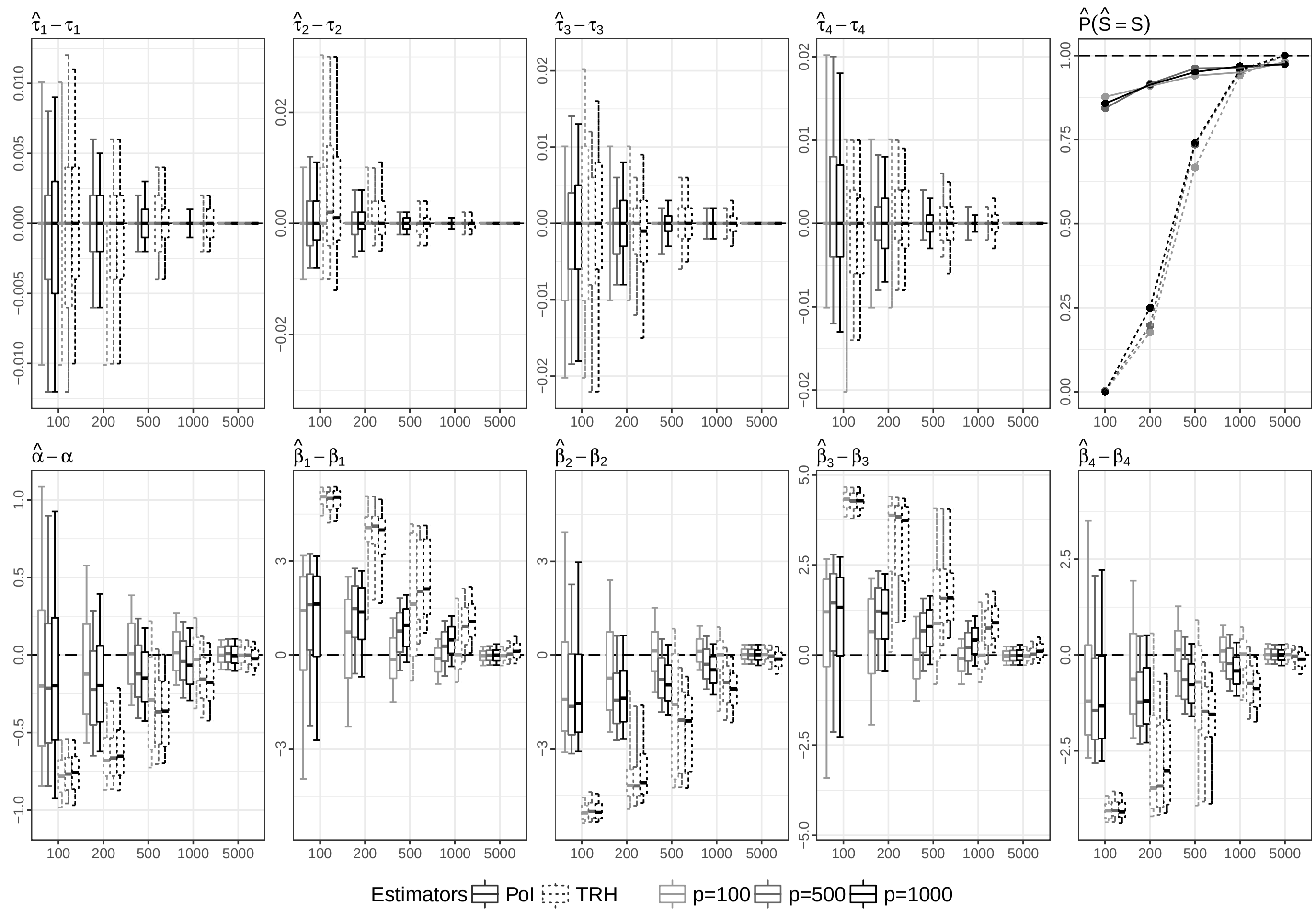

Figure 7: Comparison of the estimation errors from using our BIC-based method POI (solid lines) and our threshold-based method TRH (dashed lines).

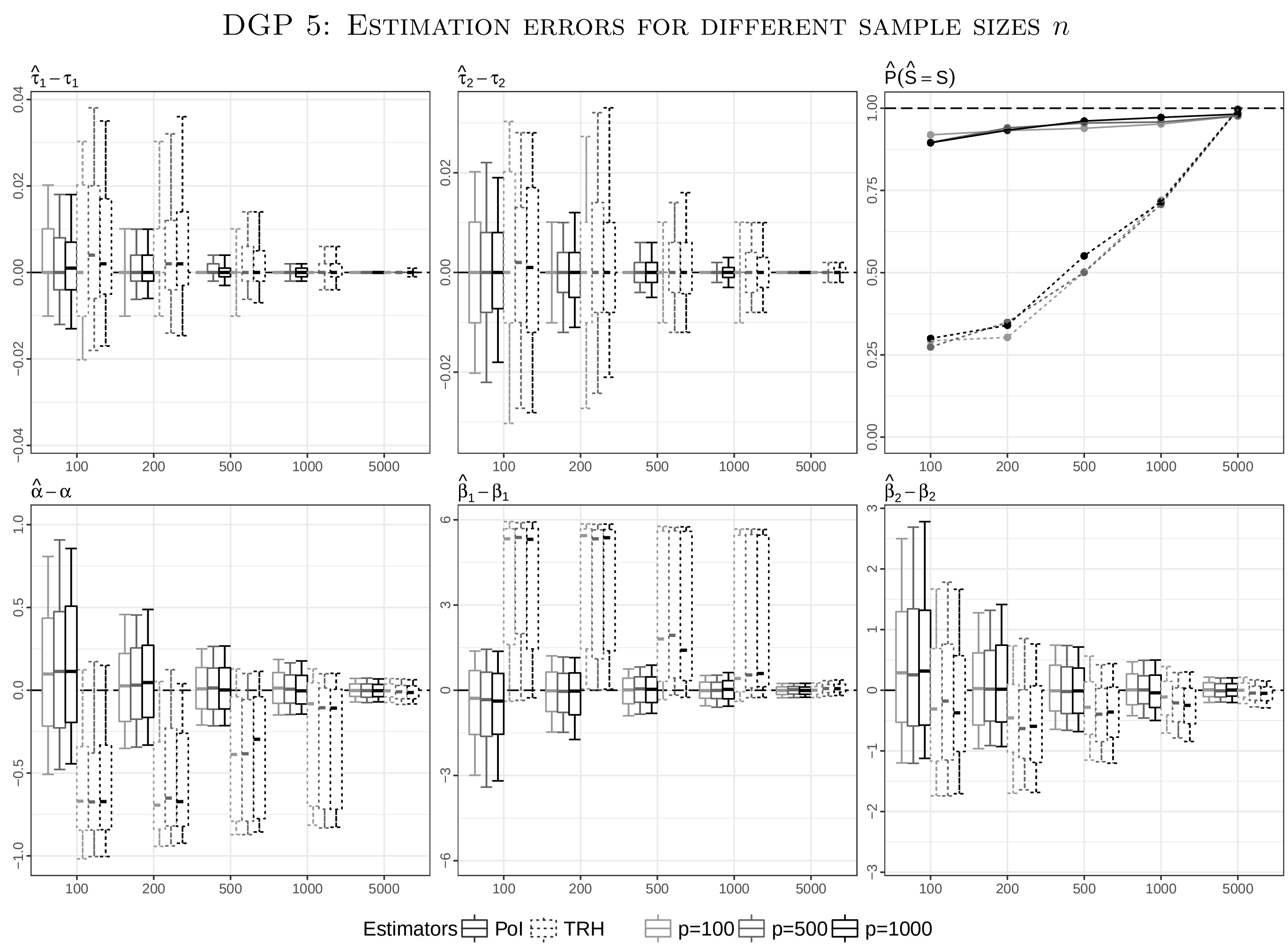

Figure 8: Comparison of the estimation errors from using our BIC-based method POI (solid lines) and our threshold-based method TRH (dashed lines).

B Proofs of the Theoretical Results from Section 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.

In this and the following appendix, $||X||_\Phi = \inf\{C > 0| \, \mathbb{E}(\Phi(|X|/C)) \leq 1\}$ 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.

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 (2012). 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

$$X_i(s) = \frac{\mathbb{E}(X_i(s)\widetilde{\eta}_i)}{\mathbb{V}(\widetilde{\eta}_i)}\widetilde{\eta}_i + (X_i(s) - \frac{\mathbb{E}(X_i(s)\widetilde{\eta}_i)}{\mathbb{V}(\widetilde{\eta}_i)}\widetilde{\eta}_i) = \frac{\mathbb{E}(X_i(s)\widetilde{\eta}_i)}{\mathbb{V}(\widetilde{\eta}_i)}\widetilde{\eta}_i + e_i(s), \tag{14}$$

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, 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 $\mathbb{E}(e_i(s)g(\eta_i)) = 0$, 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 an additional functional linear regression 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*

$$P\Big(\sup_{t-s\delta \leq u \leq t+s\delta} |\frac{1}{n}\sum_{i=1}^n [(Z_{\delta,i}(t) - Z_{\delta,i}(u))Y_i - \mathbb{E}((Z_{\delta,i}(t) - Z_{\delta,i}(u))Y_i)]| \leq zD_1 \sqrt{\frac{\delta^k s^\kappa}{n}} \Big) \\ \geq 1 - 2\exp(-z^2) \tag{15}$$

and

$$P\Big(\sup_{t-s\delta \leq u \leq t+s\delta} |\frac{1}{n}\sum_{i=1}^n [(Z_{\delta,i}(t)^2 - Z_{\delta,i}(u)^2) - \mathbb{E}(Z_{\delta,i}(t)^2 - Z_{\delta,i}(u)^2)]| \leq zD_2\delta^\kappa \sqrt{\frac{s^\kappa}{n}} \Big) \\ \geq 1 - 2\exp(-z^2). \tag{16}$$

Proof of Lemma B.1. Assertion (16) follows directly from Lemma 1 in Kneip et al. (2016b). For the proof of (15) 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 (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 $m \geq 2$, 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 derive the following analogue of Lemma 2 in Kneip et al. (2016b).

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*

$$0 < D_3\delta^\kappa \leq \inf_{t\in[a+\delta,b-\delta]} \mathbb{E}(Z_{\delta,i}(t)^2) \leq \sigma^2_{z,sup} := \sup_{t\in[a+\delta,b-\delta]} \mathbb{E}(Z_{\delta,i}(t)^2) \leq D_4\delta^\kappa \tag{17}$$

$$\lim_{n\to\infty} P\Bigg(\sup_{t\in[a+\delta,b-\delta]} |\frac{1}{n}\sum_{i=1}^{n}[Z_{\delta,i}(t)^2 - \mathbb{E}(Z_{\delta,i}(t)^2)]| \leq D_5\delta^\kappa \sqrt{\frac{1}{n}\log(\frac{b-a}{\delta})} \Bigg) = 1. \tag{18}$$

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

$$P\Bigg(\sup_{t\in[a+\delta,b-\delta]} (\frac{1}{n}\sum_{i=1}^{n} Z_{\delta,i}(t)^2)^{-\frac{1}{2}} |\frac{1}{n}\sum_{i=1}^{n}(Z_{\delta,i}(t)Y_i - \mathbb{E}(Z_{\delta,i}(t)Y_i))| \leq A^*\sqrt{\frac{\sigma^2_{|y|}}{n}\log(\frac{b-a}{\delta})} \Bigg) \to 1, \tag{19}$$

$$P\Bigg(\sup_{t\in[a+\delta,b-\delta]} |\frac{1}{n}\sum_{i=1}^{n}(Z_{\delta,i}(t)Y_i - \mathbb{E}(Z_{\delta,i}(t)Y_i))| \leq A^*\sqrt{\frac{\sigma^2_{|y|}D_4\delta^\kappa}{n}\log(\frac{b-a}{\delta})} \Bigg) \to 1. \tag{20}$$

Proof of Lemma B.2. Assertions (17) and (18) follow immediately from the proof of Lemma 2 in Kneip et al. (2016b) for any $\omega_2 > \omega_1 > 1$, where ω_2 and ω_1 are constants used in Kneip et al. (2016b). In order to show (19) one can follow the proof given in Kneip et al. (2016b) until assertion (C.17). It is then the crucial point to show that

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

Recall that it follows from (17) and (18) 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 (17) 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

$$\mathbb{E}(|W_i(s_j)|^m) \leq \frac{m!}{2}K^{m-2}K^2 \quad \text{for all} \quad m = 2, 3, \ldots$$

It immediately follows from an application of Corollary 1 in van de Geer and Lederer (2013) that there exists a constant $0 < L_{2,2} < \infty$ such that the Orlicz-Norm $||\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i(s_j)||_\Psi$ can be bounded by $L_{2,2} < \infty$. And hence we can infer that

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

It then follows from similar steps as in the proof of Lemma 1 in Kneip et al. (2016a) that there exists a constant $0 < L_{2,3} < \infty$ such that for all $0 < z \leq \sqrt{n}$ we obtain

$$\begin{aligned}
P&\left(|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i(s_j)| > zL_{2,3}\right) \\
&= P(\frac{|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)Y_i - E(Z_{\delta,i}(s_j)Y_i)|}{\sqrt{L_{2,1}\delta^\kappa}} > zL_{2,3}\sqrt{\frac{D_4\delta^\kappa\sigma^2_{|y|}}{nL_{2,1}\delta^\kappa}}\Bigg) \leq 2\exp(-z^2).
\end{aligned}$$

We may thus conclude that there exists a constant $0 < L_{2,4} < \infty$ such that

$$P\left(\frac{|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)Y_i - E(Z_{\delta,i}(s_j)Y_i)|}{\sqrt{L_{2,1}\delta^\kappa}} > zL_{2,4}\sqrt{\frac{\sigma^2_{|y|}}{n}}\right) \leq 2\exp(-z^2).$$

Finally, it follows from the union bound that

$$\begin{aligned}
P&\left(\sup_{j\in\{2,3,\ldots,N_{\omega_1}\}}\frac{|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)Y_i - E(Z_{\delta,i}(s_j)Y_i)|}{(L_{2,1}\delta^\kappa)^{\frac{1}{2}}} \leq zL_{2,4}\sqrt{\frac{\sigma^2_{|y|}}{n}}\right) \\
&\geq 1 - \sum_{j=1}^{N_{\omega_1}} P\left(\frac{|\frac{1}{n}\sum_{i=1}^n Z_{\delta,i}(s_j)Y_i - E(Z_{\delta,i}(s_j)Y_i)|}{(L_{2,1}\delta^\kappa)^{\frac{1}{2}}} > zL_{2,4}\sqrt{\frac{\sigma^2_{|y|}}{n}}\right) \\
&\geq 1 - N_{\omega_1}2\exp(-z^2) \\
&\geq 1 - 2(\frac{b-a}{\delta})^{\omega_1}\exp(-z^2).
\end{aligned}$$

Setting $z = \sqrt{\omega_2 \log\left(\frac{b-a}{\delta}\right)}$ for some $\omega_2 > \omega_1$ we have that

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

There now obviously exists a constant D with $0 < \sqrt{\omega_2} L_{2,4} = D < \infty$ for which assertion (19) will hold.

Finally, (20) 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. This is the price to pay for not assuming Gaussian Y_i.

Remarks to Lemma B.2 concerning the threshold λ:

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

We proceed to give more details about the motivation for the threshold used in the simulation:

2. Arguments for the applicability of the threshold λ in the proof of Theorem 2.1 follow from Lemma B.2. The crucial step for determining an operable threshold 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, the union bound in the proof of Lemma B.2 together with an elementary bound on the tails of the normal distribution 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 threshold $A\sqrt{\sqrt{\mathbb{E}(Y_i^4)}\log(\frac{b-a}{\delta})/n}$ for some $A \geq \sqrt{2\sqrt{3}}$ is then an immediate consequence.

Lemma 3 in Kneip et al. (2016b) remains unchanged and is repeated in the following 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*

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

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}^**

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

as well as

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

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

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

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 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). □

C Proofs of the Theoretical Results from Section 3

We begin with the proof of Theorem 3.1.

Theorem C.1. *Under our setup 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*

$$\mathbb{E}\left(\left(g(\alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r)) - g(\alpha^* + \sum_{r=1}^{S^*} \beta_r^* X_i(\tau_r))\right)^2\right) > 0, \tag{25}$$

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

Proof of Theorem C.1. 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

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

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

$$\mathbb{E}\left(\left((\alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r)) - (\alpha^* + \sum_{r=1}^{S^*} \beta_r^* X_i(\tau_r))\right)^2\right) > 0, \tag{26}$$

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

Now suppose

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

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

$$P\left(g(\alpha+\sum_{r=1}^{S}\beta_r X_i(\tau_r)) = g(\alpha^*+\sum_{r=1}^{S^*}\beta_r^* X_i(\tau_r))\right) = 1.$$

Since g is invertible we then have

$$P\left(g(\alpha+\sum_{r=1}^{S}\beta_r X_i(\tau_r)) = g(\alpha^*+\sum_{r=1}^{S^*}\beta_r^* X_i(\tau_r))\right) = 1$$

if and only if

$$P\left((\alpha+\sum_{r=1}^{S}\beta_r X_i(\tau_r)) = (\alpha^*+\sum_{r=1}^{S^*}\beta_r^* X_i(\tau_r))\right) = 1.$$

But by (26) we have

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

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

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

whenever $|\alpha-\alpha^*|>0$, or $\sup_{r=1,\dots,S}|\beta_r-\beta_r^*|>0$, or $\sup_{r=S+1,\dots,S^*}|\beta_r^*|>0$, which proves the assertion of the theorem. □

Theorem 3.1 now follows directly from Theorem C.1 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 $X_i=(X_i(t): t\in[a,b])$, $i=1,\dots,n$ 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 (7), 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* $r = 1, \dots, S$*:*

$$\frac{1}{n}\sum_{i=1}^{n}(X_i(\widehat{\tau}_r) - X_i(\tau_r))^2 = O_P(n^{-1}) \tag{27}$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i(\widehat{\tau}_r) - X_i(\tau_r))f(\eta_i^*) = O_P(n^{-\min\{1,\frac{1}{\kappa}\}}) \tag{28}$$

$$\frac{1}{n}\sum_{i=1}^{n}X_i(t^*)(X_i(\widehat{\tau}_r) - X_i(\tau_r))f(\eta_i^*) = O_P(n^{-\min\{1,\frac{1}{\kappa}\}}) \tag{29}$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i(\widehat{\tau}_r) - X_i(\tau_r))\varepsilon_i f(\eta_i^*) = O_P(n^{-1}) \tag{30}$$

$$\frac{1}{n}\sum_{i=1}^{n}X_i(t^*)(X_i(\widehat{\tau}_r) - X_i(\tau_r))\varepsilon_i f(\eta_i^*) = O_P(n^{-1}) \tag{31}$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i(\widehat{\tau}_r) - X_i(\tau_r))^4 = O_P(n^{-2}) \tag{32}$$

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

$$\begin{aligned}|\mathbb{E}((X_i(t+qs) - X_i(t))X_i(t^*))| &= |\omega(t+qs, t^*, |t+qs-t^*|^\kappa) - \omega(t, t^*, |t-t^*|^\kappa)| \\ &\leq L_{1,1}|qs|^{\min\{1,\kappa\}}.\end{aligned} \tag{33}$$

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

$$\begin{aligned}\sigma^2_{(X_i(\tau_r+sq_1)-X_i(\tau_r+sq_2))} &= \mathbb{E}\Big(((X_i(\tau_r+sq_1) - X_i(\tau_r+sq_2)))^2\Big) \\ &\leq L_{1,2}|q_1-q_2|^\kappa s^\kappa \leq L_{1,3}s^\kappa.\end{aligned} \tag{34}$$

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 (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 (2012) for a more precise statement.

We are now equipped with the tools to proof the different assertions of the proposition. Assertion (27) follows from Proposition 2 in Kneip et al. (2016a). In order to proof Assertion (28), choose any $0 < s$ sufficiently small and define for $q_1, q_2 \in [-1, 1]$

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

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 $m = 2, 3, \ldots$ we have

$$\mathbb{E}(|\frac{1}{s^{\frac{\kappa}{2}}}\chi_i(q_1, q_2)|^m) \leq \frac{m!}{2}(L_{1,4}|q_1 - q_2|^{\frac{\kappa}{2}})^m. \tag{35}$$

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 $0 < L_{1,5} < \infty$:

$$||\frac{1}{\sqrt{ns^\kappa}}\sum_{i=1}^n \chi_i(q_1, q_2)||_\Psi \leq L_{1,5}|q_1 - q_2|^{\frac{\kappa}{2}}. \tag{36}$$

By (36) 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

$$P\Bigg(\sup_{q_1,q_2\in[-1,1]}|\frac{1}{\sqrt{ns^\kappa}}\sum_{i=1}^n \chi_i(q_1, q_2)| < L_{1,7}x\Bigg) \geq 1 - 2\exp(-x^2). \tag{37}$$

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 (33) we can therefore conclude that there exists a constant $0 \leq L_{1,8} < \infty$ such that for all $q_1 \in [-1, 1]$:

$$|\mathbb{E}((X_i(\tau_r + sq_1) - X_i(\tau_r))f(\eta_i^*))| \leq L_{1,8}s^{\min\{1,\kappa\}} \tag{38}$$

Using (38) together with (37) we can conclude that for all $0 < x \leq \sqrt{n}$ we have:

$$P\left(\sup_{\tau_r - s \leq u_r \leq \tau_r + s} |\frac{1}{n}\sum_{i=1}^{n}(X_i(u_r) - X_i(\tau_r))f(\eta_i^*)| < L_{1,8}s^{\min\{1,\kappa\}} + L_{1,7}\frac{s^{\frac{\kappa}{2}}}{\sqrt{n}}x\right) \geq 1 - 2\exp(-x^2) \tag{39}$$

Assertion (28) then follows immediately from (7).
By the boundedness of f, the proof of (29) 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 (33) 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 (33) 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:

$$|\mathbb{E}(X_i(t^*)(X_i(\tau_r + sq_1) - X_i(\tau_r))f(\eta_i^*))| \leq L_{1,11}s^{\min\{1,\kappa\}}. \tag{40}$$

By (40) one can conclude similar to (39) 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 (29) then follows again immediately from (7).
In order to show assertion (30) 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

$$||s^{-\kappa/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i(\tau_r + sq_1) - X_i(\tau_r + sq_2))\varepsilon_i f(\eta_i^*)||_p \leq L_{1,13}|q_1 - q_2|^{\frac{\kappa}{2}}. \tag{41}$$

By assertion (41) 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

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

Assertion (30) then follows from (7) and our conditions on p. Moreover, assertion (31) follows from exactly the same steps.

It remains to proof (32). 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 (34) 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

$$\mathbb{E}\left(|s^{-2\kappa}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i(\tau_r + sq_1) - X_i(\tau_r))^4 - (X_i(\tau_r + sq_2) - X_i(\tau_r))^4|^p\right) \leq L_{1,15}^p|q_1 - q_2|^{\frac{p\kappa}{2}}. \tag{42}$$

At the same time (34) 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 (42) 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 (32) then follows once more from (7). $\square$

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

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

Similarly, 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

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

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

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

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 3.1 and under the results of Proposition C.1 we have*

$$\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0) = \frac{1}{n}\,\mathbf{U}_n(\boldsymbol{\beta}_0) + O_P(n^{-\min\{1,1/\kappa\}}). \tag{43}$$

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

$$\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = \frac{1}{n}\,\mathbf{U}_n(\boldsymbol{\beta}) + O_P(n^{-\frac{1}{2}}), \tag{44}$$

$$\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}) = \frac{1}{n}\,\mathbf{F}_n(\boldsymbol{\beta}) + O_P(n^{-1/2}), \tag{45}$$

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

Moreover, we have as $n \to \infty$

$$\mathbb{E}(\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta})) \to \mathbb{E}(\frac{1}{n}\,\mathbf{U}_n(\boldsymbol{\beta})), \tag{47}$$

$$\mathbb{E}(\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta})) \to \mathbb{E}(\frac{1}{n}\,\mathbf{F}_n(\boldsymbol{\beta})), \tag{48}$$

$$\mathbb{E}(\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta})) \to \mathbb{E}(\frac{1}{n}\,\mathbf{R}_n(\boldsymbol{\beta})). \tag{49}$$

Particularly,

$$\mathbb{E}(\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta}_0)) \to 0 \tag{50}$$

and

$$\mathbb{E}(\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0)) \to 0. \tag{51}$$

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

$$\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) = \frac{1}{n}\,\mathbf{U}_n(\boldsymbol{\beta}) + \mathbf{Rest}_n(\boldsymbol{\beta}), \tag{52}$$

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

$$\begin{aligned}\mathbf{Rest}_n(\boldsymbol{\beta}) &= \frac{1}{n}(\widehat{\mathbf{D}}_n^T(\boldsymbol{\beta})\widehat{\mathbf{V}}_n^{-1}(\boldsymbol{\beta}) - \mathbf{D}_n^T(\boldsymbol{\beta})\,\mathbf{V}_n^{-1}(\boldsymbol{\beta}))(\mathbf{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})) - \frac{1}{n}\widehat{\mathbf{D}}_n^T(\boldsymbol{\beta})\widehat{\mathbf{V}}_n^{-1}(\boldsymbol{\beta})(\widehat{\boldsymbol{\mu}}_n(\boldsymbol{\beta}) - \boldsymbol{\mu}_n(\boldsymbol{\beta})) \\ &= \mathbf{Rest}_1(\boldsymbol{\beta}) + \mathbf{Rest}_2(\boldsymbol{\beta}), \quad \text{say.}\end{aligned} \tag{53}$$

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

$$\begin{aligned} Rest_{j,1}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}(\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))}\widehat{X}_{ij} - \frac{g'(\eta_i(\boldsymbol{\beta}))}{\sigma^2(g(\eta_i(\boldsymbol{\beta})))}X_{ij})(y_i - g(\eta_i(\boldsymbol{\beta}))) \\ &= \frac{1}{n}\sum_{i=1}^{n}X_{ij}(\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))} - \frac{g'(\eta_i(\boldsymbol{\beta}))}{\sigma^2(g(\eta_i(\boldsymbol{\beta})))})(y_i - g(\eta_i(\boldsymbol{\beta}))) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}(\widehat{X}_{ij} - X_{ij})(\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))})(y_i - g(\eta_i(\boldsymbol{\beta}))) \\ &= R_{j,1,a}(\boldsymbol{\beta}) + R_{j,1,b}(\boldsymbol{\beta}), \quad \text{say.} \end{aligned} \tag{54}$$

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 $\boldsymbol{\beta} = \boldsymbol{\beta}_0$

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

Since $|h'(\cdot)| \leq M_h$ and $|h''(\cdot)| \leq M_h$, $1/nR_{j,1,a}(\boldsymbol{\beta}_0) = O_P(n^{-1})$ for $j = 1, \ldots, S+1$ follows immediately from (30) and (31) together with the Cauchy-Schwarz inequalitiy and (32). 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:

$$\mathbf{Rest}_1(\boldsymbol{\beta}_0) = O_P(n^{-1}). \tag{55}$$

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 (28), (29) as well as the Cauchy-Schwarz inequality together with (32), can now be used to conclude that or all $\boldsymbol{\beta}$ and $j = 1, \ldots, S+1$ we have

$$Rest_{j,2}(\boldsymbol{\beta}) = O_P(n^{-\min\{1,1/\kappa\}}). \tag{56}$$

Assertion (43) then follows from (54), (55) and (56). 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 (43), we thus have $\mathbb{E}(\mathbf{Rest}_n(\boldsymbol{\beta}_0)) \to \mathbf{0}$ implying (51), $\mathbb{E}(\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0)/n) \to \mathbf{0}$, since $\mathbb{E}(\mathbf{U}_n(\boldsymbol{\beta}_0)/n) = 0$.

In order to proof assertion (44) suppose $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ and note that we still have (56). 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 (44) 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 (27) now yield

$$
\frac{1}{n}\sum_{i=1}^{n} X_{ij}(h(\widehat{\eta}_i(\boldsymbol{\beta})) - h(\eta_i(\boldsymbol{\beta})))(y_i - g(\eta_i(\boldsymbol{\beta}_0))) = O_P(n^{-\frac{1}{2}}). \tag{57}
$$

While the Cauchy-Schwarz inequalitiy together with (27) yields

$$
\frac{1}{n}\sum_{i=1}^{n}(\widehat{X}_{ij} - X_{ij})h(\widehat{\eta}_i(\boldsymbol{\beta}))(y_i - g(\eta_i(\boldsymbol{\beta}_0))) = O_P(n^{-\frac{1}{2}}). \tag{58}
$$

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 (27) it can immediately seen that

$$
\frac{1}{n}\sum_{i=1}^{n} X_{ij}(h(\widehat{\eta}_i(\boldsymbol{\beta})) - h(\eta_i(\boldsymbol{\beta})))(g(\eta_i(\boldsymbol{\beta})) - g(\eta_i(\boldsymbol{\beta}_0))) = O_P(n^{-\frac{1}{2}}). \tag{59}
$$

Similar one may show that

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{X}_{ij}-X_{ij})(\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))})(g(\eta_i(\boldsymbol{\beta}))-g(\eta_i(\boldsymbol{\beta}_0)))=O_P(n^{-\frac{1}{2}}). \tag{60}$$

Assertion (44) then follows from (56) and (57)–(60). (47) follows again from a closer investigation of the existence and boundedness of moments of the involved remainder terms, leading to (56).

In order to proof (45), 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

$$\begin{aligned}
Rest_{jk}^{(F)}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}(\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))^2}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))}\widehat{X}_{ij}\widehat{X}_{ik}-\frac{g'(\eta_i(\boldsymbol{\beta}))^2}{\sigma^2(g(\eta_i(\boldsymbol{\beta})))}X_{ij}X_{ik}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))^2}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))}-\frac{g'(\eta_i(\boldsymbol{\beta}))^2}{\sigma^2(g(\eta_i(\boldsymbol{\beta})))})X_{ij}X_{ik} && (61)\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))^2}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))}(\widehat{X}_{ij}-X_{ij})X_{ik} && (62)\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))^2}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))}(\widetilde{X}_{ik}-X_{ik})X_{ij} && (63)\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\frac{g'(\widehat{\eta}_i(\boldsymbol{\beta}))^2}{\sigma^2(g(\widehat{\eta}_i(\boldsymbol{\beta})))}(\widehat{X}_{ik}-X_{ik})(\widehat{X}_{ij}-X_{ij}). && (64)
\end{aligned}$$

$Rest_{jk}^{(F)}(\boldsymbol{\beta})$ consists of the sum of four terms. We begin with (61).

Define $h_1(x)=g'(x)^2/\sigma^2(g(x))$ and note that $|h_1(x)|\le M_{h_1}$ as well as $|h_1'(x)|\le M_{h_1}$ for some constant $M_{h_1}<\infty$. With the help of the Cauchy-Schwarz inequality and (27), 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:

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

On the other hand, the Cauchy-Schwarz inequality together with the boundedness $|h_1(x)|$ and (27) implies that each of the other terms (62)–(64) is $O_P(n^{-1/2})$. Assertion (45) 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 (48) follows immediately.

In order so show (46), 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

$$
\begin{aligned}
Rest_{jk}^{(R)}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(h'(\widehat{\eta}_i(\boldsymbol{\beta})) - h'(\eta_i(\boldsymbol{\beta})))X_{ij}X_{ik}(y_i - g(\eta_i(\boldsymbol{\beta}))) \quad (65)\\
+ \frac{1}{n}\sum_{i=1}^{n} h'(\widehat{\eta}_i(\boldsymbol{\beta}))X_{ij}(\widehat{X}_{ik} - X_{ik})(y_i - g(\eta_i(\boldsymbol{\beta}))) \quad (66)\\
+ \frac{1}{n}\sum_{i=1}^{n} h'(\widehat{\eta}_i(\boldsymbol{\beta}))(\widehat{X}_{ij} - X_{ij})X_{ik}(y_i - g(\eta_i(\boldsymbol{\beta}))) \quad (67)\\
+ \frac{1}{n}\sum_{i=1}^{n} h'(\widehat{\eta}_i(\boldsymbol{\beta}))(\widehat{X}_{ij} - X_{ij})(\widehat{X}_{ik} - X_{ik})(y_i - g(\eta_i(\boldsymbol{\beta}))) \quad (68)\\
- \frac{1}{n}\sum_{i=1}^{n} h'(\widehat{\eta}_i(\boldsymbol{\beta}))\widehat{X}_{ij}\widehat{X}_{ik}(g(\widehat{\eta}_i(\boldsymbol{\beta})) - g(\eta_i(\boldsymbol{\beta}))). \quad (69)
\end{aligned}
$$

we will first show

$$\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta}_0) = \frac{1}{n}\mathbf{R}_n(\boldsymbol{\beta}_0) + O_P(n^{-\frac{1}{2}}). \quad (70)$$

For $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, since $|h''(\cdot)| \leq M_h$, a Taylor expansion together with the Cauchy-Schwarz inequality and (27) 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 (66)–(68) are $O_P(n^{-\frac{1}{2}})$ At the same time another Taylor expansion of (69) yields together with the Cauchy-Schwarz inequality and (27) for some $\xi_{i,5}$ between $\widehat{\eta}_i(\boldsymbol{\beta})$ and $\eta_i(\boldsymbol{\beta})$:

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

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 (70) and (50).

Now suppose $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ and take another look at (65):

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

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

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

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 (69) and (66) are, for all $\boldsymbol{\beta}$, $O_P(n^{-\frac{1}{2}})$.
Considerations for (67)–(68) are parallel to the case (66) assertion (46) follows immediately. (49) follows again from a closer investigation of the existence and boundedness of the moments of the rest terms used in the derivations (46). □

The proof of Theorem 3.2 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 3.2. 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 $p \geq 1$:

$$\mathbb{E}(\max_{\boldsymbol{\beta}\in N} |h_1(\widehat{\eta}(\boldsymbol{\beta}))\widehat{X}_j\widehat{X}_k|^p) \leq M_{1,1} \tag{71}$$

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 $p \geq 1$:

$$\mathbb{E}(\max_{\boldsymbol{\beta}\in N} |h'(\widehat{\eta}(\boldsymbol{\beta}))\widehat{X}_j\widehat{X}_k(g(\eta(\boldsymbol{\beta}_0)) - g(\widehat{\eta}(\boldsymbol{\beta})))|^p) \leq M_{1,2} \tag{72}$$

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 $1 \leq p \leq 2$:

$$\mathbb{E}(\max_{\boldsymbol{\beta}\in N} |h'(\widehat{\eta}(\boldsymbol{\beta}))\widehat{X}_j\widehat{X}_k\varepsilon|^p) \leq M_{1,3} \tag{73}$$

for some constant $M_{1,3} < \infty$, not depending on n. By (71), (72) and (73) a uniform law of large numbers for triangular arrays leads to

$$\max_{\boldsymbol{\beta}\in N} ||\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}) - \mathbb{E}(\widehat{\mathbf{F}}(\boldsymbol{\beta}))|| \xrightarrow{p} 0, \tag{74}$$

as well as

$$\max_{\boldsymbol{\beta}\in N} ||\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta}) - \mathbb{E}(\widehat{\mathbf{R}}(\boldsymbol{\beta}))|| \xrightarrow{p} 0. \tag{75}$$

Moreover, by (71), $\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0)/n$ 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 (50), $\mathbb{E}(\widehat{\mathbf{R}}(\boldsymbol{\beta}_0)) \to 0$ and $\mathbb{E}(\widehat{\mathbf{R}}(\boldsymbol{\beta}))$ is continuous in $\boldsymbol{\beta}$ we have for all $\epsilon > 0$, with probability converging to 1,

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

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

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

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

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

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

where $\widehat{\boldsymbol{\mathcal{V}}}_n(\boldsymbol{\beta}) = -\widehat{\mathbf{F}}_n^{-1/2}(\boldsymbol{\beta}_0)\widehat{\mathbf{H}}_n(\boldsymbol{\beta})\widehat{\mathbf{F}}_n^{-T/2}(\boldsymbol{\beta}_0)$ 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

$$P(-\widehat{\mathbf{H}}_n(\boldsymbol{\beta}) - c\widehat{\mathbf{F}}_n(\boldsymbol{\beta}_0) \text{ positive semidefinite for all } \boldsymbol{\beta} \in N_n(\delta_1)) \to 1 \tag{76}$$

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

$$P(\widehat{Q}_n(\boldsymbol{\beta}) - \widehat{Q}_n(\boldsymbol{\beta}_0) < 0 \text{ for all } \boldsymbol{\beta} \in \partial N_n(\delta_1)) \geq 1 - \zeta \tag{77}$$

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 $\widehat{\mathbf{U}}_n(\boldsymbol{\beta})$. (77) 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 $\boldsymbol{\beta}_0$:

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

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

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

Note that (76) implies that with probability converging to one we have

$$\lambda^2_{\min}\widehat{\boldsymbol{\mathcal{V}}}_n(\widetilde{\boldsymbol{\beta}}) \geq c^2.$$

Hence, with probability converging to one:

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

At the same time (43) and (45) 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

$$||\widehat{\mathbf{F}}_n^{-1/2}(\boldsymbol{\beta}_0)\widehat{\mathbf{U}}_n(\boldsymbol{\beta}_0)||^2 \leq ||\,\mathbf{F}_n^{-1/2}(\boldsymbol{\beta}_0)\,\mathbf{U}_n(\boldsymbol{\beta}_0)||^2 + \epsilon. \tag{78}$$

Since $\mathbb{E}(||\,\mathbf{F}_n^{-1/2}(\boldsymbol{\beta}_0)\,\mathbf{U}_n(\boldsymbol{\beta}_0)||^2) = p$, we may conclude from (78) that with probability converging to 1 we have for all sufficiently large n:

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

yielding (77) 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

$$\sqrt{n}(\widehat{\boldsymbol\beta}-\boldsymbol\beta_0)=\Bigg(\mathbf{I}_{S+1}-\Big(\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol\beta_0)\Big)^{-1}\Big(\frac{\widehat{\mathbf{H}}_n(\widetilde{\boldsymbol\beta})-\widehat{\mathbf{H}}_n(\boldsymbol\beta_0)}{n}\Big) \\ -\Big(\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol\beta_0)\Big)^{-1}\Big(\frac{\widehat{\mathbf{H}}_n(\boldsymbol\beta_0)+\widehat{\mathbf{F}}_n(\boldsymbol\beta_0)}{n}\Big)\Bigg)^{-1}\Big(\frac{\widehat{\mathbf{F}}_n(\boldsymbol\beta_0)}{n}\Big)^{-1}\frac{\widehat{\mathbf{U}}_n(\boldsymbol\beta_0)}{\sqrt{n}}. \tag{79}$$

By (45) and (46) 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 $\boldsymbol\beta\in\mathbf{R}^{S+1}$

$$\mathbb{E}(||\frac{\mathbf{H}_n(\boldsymbol\beta)+\mathbf{F}_n(\boldsymbol\beta)}{n}||_2^2)=\sum_{j=1}^{S+1}\sum_{k=1}^{S+1}\mathbb{E}\left((\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iX_{ij}X_{ik}h'(\eta_i(\boldsymbol\beta)))^2\right)=O(n^{-1}),$$

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 (74) and (75) we can conclude that for any compact neighborhood N around $\boldsymbol\beta_0$:

$$\max_{\boldsymbol\beta\in N}||\frac{1}{n}\widehat{\mathbf{H}}_n(\boldsymbol\beta)-\mathbb{E}(\widehat{\mathbf{H}}(\boldsymbol\beta))||\xrightarrow{p}0. \tag{80}$$

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 $\boldsymbol\beta$, (80) then implies that additionally we have

$$\max_{\widetilde{\boldsymbol\beta}\in N}||\frac{1}{n}\widehat{\mathbf{H}}_n(\widetilde{\boldsymbol\beta})-\mathbb{E}(\widehat{\mathbf{H}}(\boldsymbol\beta_0))||=o_P(1). \tag{81}$$

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 (79) can then be seen as

$$\sqrt{n}(\widehat{\boldsymbol\beta}-\boldsymbol\beta)\sim\left(\frac{\widehat{\mathbf{F}}_n(\boldsymbol\beta_0)}{n}\right)^{-1}\frac{\widehat{\mathbf{U}}_n(\boldsymbol\beta_0)}{\sqrt{n}}. \tag{82}$$

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 (45) 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 (43) 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. □

Corollary C.1. *Under the assumptions of Section 3. For any compact neighborhood N around $\boldsymbol{\beta}_0$ we have*

$$\max_{\boldsymbol{\beta} \in N} ||\frac{1}{n}\widehat{\mathbf{U}}_n(\boldsymbol{\beta}) - \frac{1}{n}\mathbf{U}_n(\boldsymbol{\beta})|| = o_P(1), \tag{83}$$

$$\max_{\boldsymbol{\beta} \in N} ||\frac{1}{n}\widehat{\mathbf{F}}_n(\boldsymbol{\beta}) - \frac{1}{n}\mathbf{F}_n(\boldsymbol{\beta})|| = o_P(1), \tag{84}$$

$$\max_{\boldsymbol{\beta} \in N} ||\frac{1}{n}\widehat{\mathbf{R}}_n(\boldsymbol{\beta}) - \frac{1}{n}\mathbf{R}_n(\boldsymbol{\beta})|| = o_P(1), \tag{85}$$

as well as

$$\max_{\boldsymbol{\beta} \in N} ||\frac{1}{n}\widehat{\mathbf{H}}_n(\boldsymbol{\beta}) - \frac{1}{n}\mathbf{H}_n(\boldsymbol{\beta})|| = o_P(1). \tag{86}$$

Proof of Corollary C.1: The proofs of Assertions 83-85 are very similar. We begin with the proof of Assertion 83. 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

$$\mathbb{E}(\max_{\boldsymbol{\beta} \in N} |h(\widehat{\eta}(\boldsymbol{\beta}))\widehat{X}_j(y - g(\widehat{\eta}(\boldsymbol{\beta})))|^p) \leq M_{1,1} \tag{87}$$

for a constant $0 \leq M_{1,1} < \infty$ not depending on n. By (87) we can apply a uniform law of large numbers for triangular arrays to conclude that

$$\max_{\boldsymbol{\beta} \in N} || \frac{1}{n} \widehat{\mathbf{U}}_n(\boldsymbol{\beta}) - \mathbb{E}(\widehat{\mathbf{U}}(\boldsymbol{\beta}))|| = o_P(1). \tag{88}$$

Similar considerations lead to

$$\max_{\boldsymbol{\beta} \in N} || \frac{1}{n} \mathbf{U}_n(\boldsymbol{\beta}) - \mathbb{E}(\mathbf{U}(\boldsymbol{\beta}))|| = o_P(1). \tag{89}$$

By the usual decomposition we have

$$\begin{aligned} || \frac{1}{n} \widehat{\mathbf{U}}_n(\boldsymbol{\beta}) - \frac{1}{n} \mathbf{U}_n(\boldsymbol{\beta})|| \leq\, & || \frac{1}{n} \widehat{\mathbf{U}}_n(\boldsymbol{\beta}) - \mathbb{E}(\widehat{\mathbf{U}}(\boldsymbol{\beta}))|| \\ & + || \frac{1}{n} \mathbf{U}_n(\boldsymbol{\beta}) - \mathbb{E}(\mathbf{U}(\boldsymbol{\beta}))|| + || \mathbb{E}(\widehat{\mathbf{U}}(\boldsymbol{\beta})) - \mathbb{E}(\mathbf{U}(\boldsymbol{\beta}))||. \end{aligned} \tag{90}$$

Assertion (83) then follows immediately from (88), (89), 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 (47).

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

$$\begin{aligned} & h(\widehat{\eta}(\boldsymbol{\beta}))\widehat{X}_j(y - g(\widehat{\eta}(\boldsymbol{\beta}))) - h(\widehat{\eta}(\widetilde{\boldsymbol{\beta}}))\widehat{X}_j(y - g(\widehat{\eta}(\widetilde{\boldsymbol{\beta}})))) \\ & = \widehat{X}_j y(h(\widehat{\eta}(\boldsymbol{\beta})) - h(\widehat{\eta}(\widetilde{\boldsymbol{\beta}}))) \\ & \quad + \widehat{X}_j h(\widehat{\eta}(\widetilde{\boldsymbol{\beta}}))(g(\widehat{\eta}(\widetilde{\boldsymbol{\beta}})) - g(\widehat{\eta}(\boldsymbol{\beta}))) \\ & \quad - \widehat{X}_j(h(\widehat{\eta}(\widetilde{\boldsymbol{\beta}})) - h(\widehat{\eta}(\boldsymbol{\beta})))g(\widehat{\eta}(\boldsymbol{\beta})). \end{aligned} \tag{91}$$

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 (91) that there exists a constant L, which is in particular independent of n such that for all $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\beta}} \in N$

$$|| \mathbb{E}(\widehat{\mathbf{U}}(\boldsymbol{\beta})) - \mathbb{E}(\widehat{\mathbf{U}}(\widetilde{\boldsymbol{\beta}}))|| \leq L || \boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}||. \tag{92}$$

Assertion (83) then follows from (90) together with (88), (89), (47) and (92).

In order to proof Assertion (84) 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

$$\mathbb{E}(\max_{\boldsymbol{\beta}\in N} ||h_1(\eta_i(\boldsymbol{\beta}))\, \mathbf{X}_i\, \mathbf{X}_i^T\, ||) < \infty. \tag{93}$$

By (93) we can apply a uniform law of large numbers to derive

$$\max_{\boldsymbol{\beta}\in N} ||\frac{1}{n}\, \mathbf{F}_n(\boldsymbol{\beta}) - \mathbb{E}(\mathbf{F}(\boldsymbol{\beta}))|| = o_P(1). \tag{94}$$

Assertion 84 then follows immediately from (94), (74) and (48), 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 85 can be proved in a similar manner using (47), (80) 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 (86), remember that $\widehat{\mathbf{H}}(\boldsymbol{\beta})/n = -\widehat{\mathbf{F}}_n(\boldsymbol{\beta})/n + \widehat{\mathbf{R}}_n(\boldsymbol{\beta})/n$, assertion (86) follows then immediately from (84) and (85). □

References

Brillinger, D. R. (2012). A generalized linear model with "gaussian" regressor variables. In *Selected Works of David Brillinger*, pp. 589–606. Springer.

Fahrmeir, L. and H. Kaufmann (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics 13*(1), 342–368.

Kneip, A., D. Poß, and P. Sarda (2016a). Functional linear regression with points of impact. *The Annals of Statistics 44*(1), 1–30.

Kneip, A., D. Poß, and P. Sarda (2016b). Supplement to "Functional linear regression with points of impact". *The Annals of Statistics*.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics 9*(6), 1135–1151.

van de Geer, S. and J. Lederer (2013). The bernstein-orlicz norm and deviation inequalities. *Probability Theory and Related Fields 157*(1), 225–250.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.