Poß, Dominik et al.

Article — Published Version

Superconsistent estimation of points of impact in non-parametric regression with functional predictors

Journal of the Royal Statistical Society: Series B (Statistical Methodology)

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Poß, Dominik et al. (2020) : Superconsistent estimation of points of impact in non-parametric regression with functional predictors, Journal of the Royal Statistical Society: Series B (Statistical Methodology), ISSN 1467-9868, Wiley, Hoboken, NJ, Vol. 82, Iss. 4, pp. 1115-1140, https://doi.org/10.1111/rssb.12386

J. R. Statist. Soc. B (2020)
82, *Part* 4, *pp.* 1115–1140

Superconsistent estimation of points of impact in non-parametric regression with functional predictors

Dominik Poß, Dominik Liebl and Alois Kneip,

University of Bonn, Germany

Hedwig Eisenbarth,

Victoria University of Wellington, New Zealand

Tor D. Wager

Dartmouth College, Hanover, USA

and Lisa Feldman Barrett

Northeastern University, Boston, Massachusetts General Hospital and Harvard Medical School, Boston, and Massachusetts General Hospital, Charlestown, USA

[Received May 2019. Final revision May 2020]

Summary. Predicting scalar outcomes by using functional predictors is a classical problem in functional data analysis. In many applications, however, only specific locations or time points of the functional predictors have an influence on the outcome. Such 'points of impact' are typically unknown and must be estimated in addition to estimating the usual model components. We show that our points-of-impact estimator enjoys a superconsistent rate of convergence and does not require knowledge or pre-estimates of the unknown model components. This remarkable result facilitates the subsequent estimation of the remaining model components as shown in the theoretical part, where we consider the case of non-parametric models and the practically relevant case of generalized linear models. The finite sample properties of our estimators are assessed by means of a simulation study. Our methodology is motivated by data from a psychological experiment in which the participants were asked to rate their emotional state continuously while watching an affective video eliciting a varying intensity of emotional reactions.

Keywords: Emotional stimuli; Functional data analysis; Non-parametric regression; On-line video rating; Quasi-maximum-likelihood; Variable selection

1. Introduction

Identifying important time points in time continuous trajectories is a difficult but highly relevant problem. For instance, current psychological research on emotional experiences often includes time continuous stimuli such as videos to induce emotional states, say $X(t) \in \mathbb{R}$, with $t \in [a, b]$, where a denotes the start of the video and b the end (Fig. 1). The evaluation of such stimuli is based on asking participants whether the video made them feel negative, say $Y = 0$, or positive,

Address for correspondence: Dominik Liebl, Institute of Finance and Statistics and Hausdorff Center for Mathematics, University of Bonn, Adenauerallee 24–42, Bonn 53113, Germany.
E-mail: dliebl@uni-bonn.de

 1369–7412/20/821115

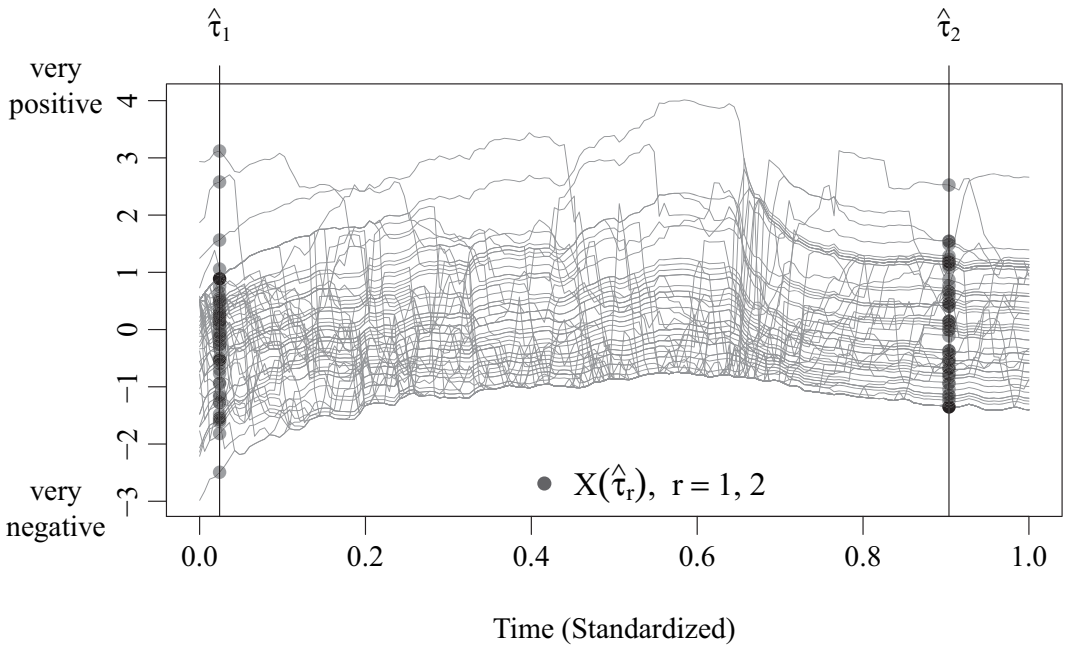

Fig. 1. Continuously self-reported emotion trajectories $\{X(t) : 0 \leqslant t \leqslant 1\}$ of $n = 65$ participants with two estimated points of impact $\hat{\tau}_1$ and $\hat{\tau}_2$; see the application in Section 5

say $Y = 1$. In this paper we consider a novel data set where participants were asked to report their emotional states continuously while watching an affective documentary video on the persecution of African albinos. After watching the video, the participants were asked to rate their final overall feeling. 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 that emotional states are assessed in research using such material. Because of a lack of appropriate statistical methods, existing approaches use heuristics such as the 'peak-and-end rule' (PER) approach to link the overall ratings with the continuous emotional stimuli (see Section 5). Such heuristic approaches, however, can produce results that do not accurately capture the summary rating and can be easily overinterpreted, as there is no unbiased formal inference about which time points contribute to the summary rating. By contrast, our new methodology enables us to identify the crucial affective video scenes—the basic prerequisite to understanding the emergence of emotional states in this kind of experiment.

The identification of 'influential' stimuli in a video corresponds to identifying corresponding time points $\tau \in (a, b)$. We aim to estimate such time points within the non-parametric model

$$Y = g\{X(\tau_1), \ldots, X(\tau_S)\} + \varepsilon, \tag{1}$$

where $\tau_1, \ldots, \tau_S \in (a, b)$ and their number $S \in \mathbb{N}$ are unknown and need to be estimated. The values $\tau_1, \ldots, \tau_S$ are called *points of impact* and provide specific locations at which the functional predictor $X \in L^2([a, b])$ influences the scalar outcome Y. In our real data application in Section 5, Y is a binary variable and the functional predictor X is evaluated at two estimated points of impact $\hat{\tau}_1$ and $\hat{\tau}_2$; see Fig. 5 in Section 4.1.

Our method builds on the work of Kneip *et al.* (2016); however, we consider the much more challenging case of estimating points of impact within a fully non-parametric function g. A

remarkable feature of our method is that identification and estimation of the points of impact $\tau_1, \ldots, \tau_S$ neither require knowledge about the non-parametric function g nor an estimate of g. The estimation of the points of impact $\tau_1, \ldots, \tau_S$ is thus robust to model misspecifications and is free of additional contaminating estimation errors. This result goes far beyond the special case of a functional linear model as considered by Kneip *et al.* (2016).

To the best of our knowledge, the problem of estimating $\tau_1, \ldots, \tau_S$ in model (1) has, so far, been considered by Ferraty *et al.* (2010) only, who proposed to estimate g non-parametrically for any combination of point-of-impact candidates $t_1^*, \ldots, t_{S^*}^* \in \{t_j : t_j = a + j(b-a)/p \text{ with } j = 1, \ldots, p\}$ and to select the best model by using cross-validation. This brute force method, however, becomes problematic in practice for $S \geqslant 2$ and large p. Furthermore, the non-parametric estimation of g implies that the points of impact $\tau_1, \ldots, \tau_S$ can be estimated at most with the non-parametric rate $n^{-2/(4+S)}$, where n denotes the sample size. Here the speed of convergence decreases dramatically for dimensions $S \geqslant 2$. By contrast, we can estimate the points of impact with a superconsistent rate of convergence, i.e. faster than the parametric rate $n^{-1/2}$, and our estimation algorithm is applicable in practice for any fixed S and large $p \gg n$.

The superconsistency result for our points-of-impact estimators is very beneficial for subsequent estimation problems and enables us to estimate the unknown function g as if the points of impact were known. We demonstrate this for a non-parametric model g as well as for the practically relevant case of generalized linear models with linear predictors, i.e.

$$g\{X(\tau_1), \ldots, X(\tau_S)\} = g\left\{\alpha + \sum_{r=1}^{S} \beta_r X(\tau_r)\right\}$$

with assumed known parametric link function g.

So far, the purely non-parametric framework has been considered by Ferraty *et al.* (2010) only. The case of a known g and linear predictor function $\alpha + \Sigma_{r=1}^{S} \beta_r X(\tau_r)$ had already been considered by previous studies; however, none of these studies provides a superconsistent estimation of points of impact independent of the model g. The term 'impact point' was coined by Lindquist and McKeague (2009) and McKeague and Sen (2010). Lindquist and McKeague (2009) considered a logistic regression framework and McKeague and Sen (2010) considered a linear regression framework. 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). Kneip *et al.* (2016) allowed for an unknown number $S \geqslant 0$ of points of impact augmenting the functional linear regression model. Liebl *et al.* (2020) proposed an improved estimation algorithm for the latter work. Aneiros and Vieu (2014) considered a linear regression framework with multiple points of impact postulating the existence of some consistent estimation procedure. Berrendero *et al.* (2019) considered a linear regression framework and proposed a reproducing kernel Hilbert space approach. Selecting sparse features from functional data X is also useful for clustering. For instance, Floriello and Vitelli (2017) proposed a method for sparse clustering of functional data. In a slightly different context, Park *et al.* (2016) focused on selecting predictive subdomains of the functional data. Related to this paper is also the work of Lindquist (2012) and Sobel and Lindquist (2014). Lindquist (2012) extended structural equation models to the functional data analysis setting and used his methodology to select significantly impacting time intervals in functional magnetic resonance imaging data. Sobel and Lindquist (2014) proposed a mixed effects model which facilitates selecting significant impact regions in functional magnetic resonance imaging data by controlling for systematic measurement errors.

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 independent of the model g. Subsequent estimation of the function g is discussed in Section 3. The simulation study and the real data application are in

Sections 4 and 5. All proofs and additional simulation results can be found in the appendices of the on-line supplementary paper supporting this paper. The R package `fdapoi` and the R scripts for reproducing our main empirical results are also available from

`https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets`.

2. Estimating points of impact

In what follows we present our theoretical framework (Section 2.1), the estimation algorithm (Section 2.2) and our asymptotic results (Section 2.3). The section concludes with a discussion of possibilities to generalize our theoretical results (Section 2.4).

2.1. Theoretical framework

In this section we present our theoretical framework for estimating the points of impact $\tau_1, \ldots, \tau_S$ without knowing or (pre-)estimating the possibly non-parametric model function g. The identification of points of impact constitutes a particular variable-selection problem. Since we consider the case where the functional predictor is observed over a densely discretized grid, one might be tempted to apply multivariate variable-selection methods like the lasso or related procedures. Note, however, that the high correlation of the predictor at neighbouring discretization points violates the basic requirements of these multivariate variable-selection procedures.

Suppose that we are given an independent and identically distributed sample of data (X_i, Y_i), $i = 1, \ldots, n$, where $X_i = \{X_i(t), t \in [a, b]\}$ is a stochastic process with $\mathbb{E}\{\int_a^b X_i(t)^2 \, \mathrm{d}t\} < \infty$, $[a, b]$ 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 modelled as

$$Y_i = g\{X_i(\tau_1), \ldots, X_i(\tau_S)\} + \varepsilon_i, \tag{2}$$

where ε_i denotes the statistical error term with $\mathbb{E}\{\varepsilon_i | X_i(t)\} = 0$ for all $t \in [a, b]$. The number $0 \leqslant S < \infty$ and the points of impact $\tau_1, \ldots, \tau_S$ are unknown and must be estimated from the data—without knowing the true model function g. The points of impact $\tau_1, \ldots, \tau_S$ indicate the locations at which the functional values $X_i(\tau_1), \ldots, X_i(\tau_S)$ have a specific influence on Y_i. Without loss of generality, we consider centred random functions X_i with $\mathbb{E}\{X_i(t)\} = 0$ for all $t \in [a, b]$.

Surprisingly, the unknown function g must fulfil only some very mild regularity conditions and does not have to be estimated to estimate the points of impact $\tau_1, \ldots, \tau_S$ (see theorem 1 later). Estimating points of impact, however, 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 (OUPs) and Poisson processes. 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 enables us to derive precise theoretical results.

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

$$\sigma(s,t)=\omega(s,t,|s-t|^{\kappa}). \tag{3}$$

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. Although a twice continuously differentiable covariance function will satisfy assumption (3) with $\kappa=2$, small values $0<\kappa<2$ indicate a process with non-smooth sample paths.

Assumption 1 covers several important classes of stochastic processes. Recall, for instance, that the class of self-similar (not necessarily centred) processes $X_i=\{X_i(t):t\geqslant 0\}$ has the property that $X_i(c_1t)=c_1^H X_i(t)$ 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 $c_2>0$; see theorem 1.2 in Embrechts and Maejima (2000). If $0<H<1$ such a process respects assumption 1 with $\kappa=2H$ and $c(t)=c_2$. A prominent example of a self-similar process is fractional Brownian motion.

Another class of processes is given when $X_i=\{X_i(t):t\geqslant 0\}$ 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 $c_3>0$. Assumption 1 then holds with $\kappa=1$ and $c(t)=c_3$. These conditions on X_i are, for instance, satisfied by second-order Lévy processes which include important processes such as Poisson processes, compound Poisson processes or jump diffusion processes.

A third important class of processes satisfying assumption 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

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

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 $\nu>m$ (see Stein (1999), chapter 2.7, page 32). Assumption 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 OUP which is given as

$$\sigma(s,t)=\omega(s,t,|s-t|)=0.5\exp(-\theta|s-t|)\sigma_{\mathrm{OU}}^2/\theta,$$

for some parameter $\theta>0$ and $\sigma_{\mathrm{OU}}>0$. Assumption 1 is then covered with $\kappa=1$ and $c(t)=0.5\sigma_{\mathrm{OU}}^2$.

The remarkable result that identification and estimation of the points of impact $\tau_1,\ldots,\tau_S$ require neither knowledge about the possibly non-parametric function g nor an estimate of g is based on the following theorem.

Theorem 1. Let X_i be a Gaussian process. For any function $g(x_1,\ldots,x_S)$ such that for all $r=1,\ldots,S$ the partial derivative $\partial g(x_1,\ldots,x_S)/\partial x_r$ is continuous almost everywhere and

$$0<\left|\mathbb{E}\left[\frac{\partial}{\partial x_r}g\{X_i(\tau_1),\ldots,X_i(\tau_S)\}\right]\right|<\infty,$$

we define

$$\vartheta_r = \mathbb{E}\left[\frac{\partial}{\partial x_r} g\{X_i(\tau_1), \ldots, X_i(\tau_S)\}\right].$$

Then the equation

$$\mathbb{E}\{X_i(s)Y_i\} = \sum_{r=1}^{S} \vartheta_r \sigma(s, \tau_r)$$

holds for all $s \in [a, b]$.

Theorem 1 enables us to decompose the cross-covariance $\mathbb{E}\{X_i(s)Y_i\}$ into coefficients ϑ_r, which depend on the unknown function g, and the covariance function σ, which depends on X_i only. Our estimation strategy for the points of impact τ_r works for unknown ϑ_r with $0 < |\vartheta_r| < \infty$. The latter imposes only mild regularity assumptions on g and is fulfilled, for instance, by any non-parametric single-index model, $g\{X_i(\tau_1), \ldots, X_i(\tau_S)\} \equiv g(\eta_i)$ with $\eta_i = \alpha + \Sigma_{r=1}^{S} \beta_r X_i(\tau_r)$, where $0 < |\mathbb{E}\{g'(\eta_i)\}| < \infty$. Of course, the class of possible functions g that is defined by theorem 1 also contains much more complex cases than single-index models.

The intention of our estimator for the points of impact τ_r is to exploit the covariance structure of processes that is described by assumption 1. Covariance functions $\sigma(s, t)$ satisfying assumption 1 are obviously not two-times differentiable at the diagonal $s = t$ but are two-times differentiable for $s \neq t$. Using theorem 1 in conjunction with assumption 1 enables us to identify uniquely the locations of the points of impact from the cross-covariance $\mathbb{E}\{X_i(s)Y_i\}$. We make this more precise by defining

$$f_{XY}(s) := \mathbb{E}\{X_i(s)Y_i\} = \sum_{r=1}^{S} \vartheta_r \sigma(s, \tau_r) \qquad \text{for } s \in [a, b].$$

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 Fig. 2(a).

A natural strategy for estimating 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 $\delta > 0$:

$$f_{XY}(s) - \tfrac{1}{2}\{f_{XY}(s + \delta) + f_{XY}(s - \delta)\}. \tag{4}$$

By defining the auxiliary process

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

we have the following equivalent moment expression for approximation (4):

$$\mathbb{E}\{Z_{\delta,i}(s)Y_i\} = f_{XY}(s) - \tfrac{1}{2}\{f_{XY}(s + \delta) + f_{XY}(s - \delta)\}. \tag{5}$$

At $s = \tau_r$, expression (5) will decline more slowly to 0 as $\delta \to 0$ than for $s \neq \tau_r$, $r = 1, \ldots, S$. For suitable values of δ, the points of impact τ_r can then be estimated by using the local extrema of the empirical counterpart of $|\mathbb{E}\{Z_{\delta,i}(s)Y_i\}|$ (see Fig. 2(b)).

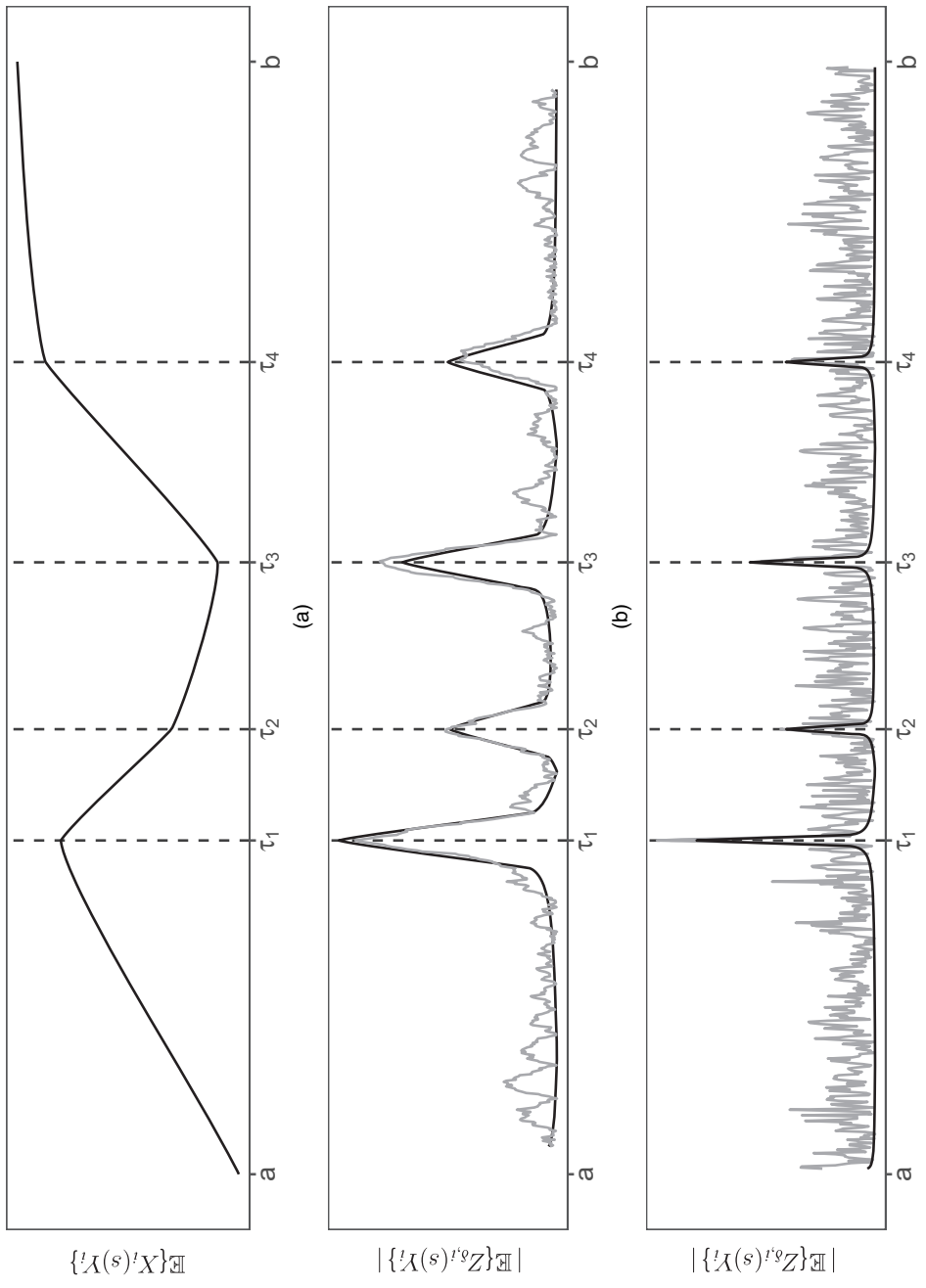

Fig. 2. (a) $\mathbb{E}\{X_i(s)Y_i\}$ as a function of s with four kink-like features at the points of impact ($\cdot$), and $|\mathbb{E}\{Z_{\delta,i}(s)Y_i\}|$, $a + \delta \leqslant s \leqslant b - \delta$ (———), and its empirical counterparts (———) for $\delta > 0$ (b) well chosen or (c) too small

More precisely, theorem 1 together with proposition C.1 and lemma C.4 in the on-line appendix C imply that as $\delta \to 0$

$$\mathbb{E}\{Z_{\delta,i}(s)Y_i\} = \begin{cases} \vartheta_r c(\tau_r)\delta^\kappa + o(\delta^\kappa) & \text{if } s \in \{\tau_1, \ldots, \tau_S\}, \\ O(\delta^2) & \text{if } s \notin \{\tau_1, \ldots, \tau_S\}, \end{cases}$$

where $0 < \kappa < 2$ and $c(\cdot) > 0$ are as defined in assumption 1.

Of course, $\mathbb{E}\{Z_{\delta,i}(s)Y_i\}$ is not known and we must rely on $n^{-1}\Sigma_{i=1}^n Z_{\delta,i}(s)Y_i$ as its estimate. Under our setting we shall have that the variance $\mathbb{V}\{Z_{\delta,i}(s)Y_i\} = O(\delta^\kappa)$, which implies that

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

Consequently, the identification of points of impact 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 Fig. 2(c). For too large δ-values, however, it will not be possible to distinguish between neighbouring points of impact.

Remark 1. Even if the covariance function $\sigma(s,t)$ does not satisfy assumption 1, the points of impact τ_r may still be estimated by using the local extrema of $\mathbb{E}\{Z_{\delta,i}(s)Y_i\}$. Suppose, for instance, that there is an $m \geqslant 2$ times differentiable function $\tilde{\sigma}: \mathbb{R} \to \mathbb{R}$ such that $\sigma(s,t) = \tilde{\sigma}(|s-t|)$, where $\tilde{\sigma}(|s-t|)$ decays sufficiently fast, as $|s-t|$ increases, such that $X_i(s)$ is essentially uncorrelated with $X_i(\tau_r)$ for $|\tau_r - s| \gg 0$. If $|\tilde{\sigma}''(0)| > |\tilde{\sigma}''(|s-t|)|$, for $s \neq t$, and $\min_{r \neq k} |\tau_r - \tau_k|$ is suffieciently large, then all points of impact might be identified as local extrema of $\mathbb{E}\{Z_{\delta,i}(s)Y_i\}$.

2.2. Estimation algorithm

In what follows 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 $|n^{-1}\Sigma_{i=1}^n Z_{\delta,i}(t_j)Y_i|$.

The procedure in algorithm 1 (Table 1) will result in estimates $\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_{M_\delta}$, where $M_\delta < \infty$ denotes the maximum number of possible repetitions. The estimator of S is

$$\hat{S} = \min\left\{ l \in \mathbb{N}_0 : \left| \frac{(1/n)\sum_{i=1}^n Z_{\delta,i}(\hat{\tau}_{l+1})Y_i}{\left\{(1/n)\sum_{i=1}^n Z_{\delta,i}(\hat{\tau}_{l+1})^2\right\}^{1/2}} \right| < \lambda \right\} \qquad \text{for some threshold } \lambda > 0.$$

An asymptotically valid choice of the threshold λ is presented in theorem 2 in Section 2.3 and a practical implementation of λ is discussed below theorem 2.

Remark 2. This estimation algorithm is made for the case of densely observed functional data. In practice this means functional data that are sampled at a high frequency such as in our real data example (Section 5). Unfortunately, we do not see a simple way to generalize our method to the case of irregularly or sparsely sampled functional data. Such a generalization would require a very different approach based on non-parametric smoothing procedures.

2.3. Asymptotic results

In this section, we consider asymptotics as $n \to \infty$ with $p \equiv p_n \geqslant Ln^{1/\kappa}$ for some constant $(b-a)/2 < L < \infty$. We introduce the following assumption.

Table 1. Algorithm 1 (estimating points of impact)

Step 1: calculate $\hat{f}_{XY}(t_j) := (1/n)\Sigma_{i=1}^{n} X_i(t_j)Y_i$, for each $j = 1, \ldots, p$
Step 2: choose $\delta > 0$ such that there exists some $k_\delta \in \mathbb{N}$ with $1 \leqslant k_\delta < (p-1)/2$ and $\delta = k_\delta(b-a)/(p-1)$
Step 3: calculate $\hat{f}_{ZY}(t_j) := \hat{f}_{XY}(t_j) - \frac{1}{2}\{\hat{f}_{XY}(t_j - \delta) + \hat{f}_{XY}(t_j + \delta)\}$, for all $j \in \mathcal{J}_\delta$, where $\mathcal{J}_\delta := \{k_\delta + 1, \ldots, p - k_\delta\}$
Step 4: repeat
initiate the repetition by setting $l = 1$
estimate the lth point-of-impact candidate as

$$\hat{\tau}_l = \arg\max_{t_j : j \in \mathcal{J}_\delta} |\hat{f}_{ZY}(t_j)|$$

update $\mathcal{J}_\delta$ by eliminating all points in $\mathcal{J}_\delta$ in an interval of size $\sqrt{\delta}$ around $\hat{\tau}_l$
set $l = l + 1$
end repetition if $\mathcal{J}_\delta = \emptyset$

Assumption 2.

(a) $X_1, \ldots, X_n$ are independent and identically distributed random functions distributed according to X. The process X is Gaussian with covariance function $\sigma(s,t)$.
(b) There is a $0 < \sigma_{|y|} < \infty$ such that for each $m = 1, 2, \ldots$ we have $\mathbb{E}(|Y_i|^{2m}) \leqslant 2^{m-1} m! \sigma_{|y|}^{2m}$.

The moment condition (b) is obviously fulfilled for bounded Y_i. For instance, in the functional logistic regression we have that $\mathbb{E}(|Y_i|^m) \leqslant 1$ for all $m = 1, 2, \ldots$. Condition (b) holds also for any centred sub-Gaussian Y_i, where a centring of Y_i can always be achieved by substituting $g\{X_i(\tau_1), \ldots, X_i(\tau_S)\} + \mathbb{E}[g\{X_i(\tau_1), \ldots, X_i(\tau_S)\}]$ for $g\{X_i(\tau_1), \ldots, X_i(\tau_S)\}$ in model (2). If X_i satisfies condition (a), then condition (b) in particular holds if the errors ε_i are sub-Gaussian and g is differentiable with bounded partial derivatives.

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

Theorem 2. Under assumptions 1 and 2, and the assumptions of theorem 1, let $\delta \equiv \delta_n \to 0$ as $n \to \infty$ such that $n\delta^\kappa / |\log(\delta)| \to \infty$ and $\delta^\kappa / n^{-\kappa+1} \to 0$. We then obtain that

$$\max_{r=1,\ldots,\hat{S}} \min_{s=1,\ldots,S} |\hat{\tau}_r - \tau_s| = O_P(n^{-1/\kappa}). \tag{6}$$

Moreover, there is a constant $0 < D < \infty$ such that when algorithm 1 is applied with threshold

$$\lambda \equiv \lambda_n = A\sqrt{\left\{ \frac{\sigma_{|y|}^2}{n} \log\left(\frac{b-a}{\delta} \right) \right\}}, \qquad A > D, \text{ and } \delta^2 = O(n^{-1}),$$

then

$$P(\hat{S} = S) \to 1 \qquad \text{as } n \to \infty. \tag{7}$$

Note that the rates of convergence in expression (6) are superconsistent, since $0 < \kappa < 2$. For instance, for OUPs or Brownian motions we have $\kappa = 1$, such that $\max_{r=1,\ldots,\hat{S}} \min_{s=1,\ldots,S} |\hat{\tau}_r - \tau_s| = O_P(n^{-1})$.

In principle, arbitrarily fast rates of convergence can be achieved for κ-values that are close to 0, because small κ-values correspond to rough processes X_i. Roughness means that the process has strong local variations also within small intervals $[\tau_r - \epsilon, \tau_r + \epsilon]$, $\epsilon > 0$, which facilitates

differentiating a point of impact τ_r, $r=1,\ldots,S$, from the neighbouring points $t\in[\tau_r-\epsilon,\tau_r+\epsilon]$. By contrast, for smooth processes (large κ-values) all values of $X_i(t)$ with $t\in[\tau_r-\epsilon,\tau_r+\epsilon]$ will be almost identical, which makes it difficult to identify the correct point of impact τ_r.

A practical and asymptotically valid threshold specification which performed well in our simulation studies is given by $\lambda=A[\mathbb{E}(Y_i^4)^{1/2}\log\{(b-a)/\delta\}/n]^{1/2}$, where $\mathbb{E}(Y_i^4)$ is estimated by $\hat{\mathbb{E}}(Y_i^4)=n^{-1}\Sigma_{i=1}^n Y_i^4$ 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. See the remark after the proof of lemma C.3 in the on-line appendix C for additional information.

The superconsistency result in theorem 2 is very general and does not require knowledge of g or a pre-estimate of g; only a set of mild and verifiably assumptions on g is postulated. Therefore, we expect that theorem 2 will be found to be useful for deriving inferential results for a broad variety of models g. In what follows we demonstrate the usefulness of theorem 2 for deriving inferential results for non-parametric models and parametric generalized linear models. Note that the related corollary 1 in Ferraty *et al.* (2010) requires the simultaneous estimation of the non-parametric model function g and the points of impact. This approach results in substantially slower non-parametric convergence rates and limits the applicability of their result considerably.

2.4. Generalizations

The above theoretical assumptions provide a tractable set-up that will be used also in the remaining parts of the paper. In this subsection, however, we show that the Gaussian assumption of theorem 1 and theorem 2 can be relaxed and that our approach for identifying and estimating the points of impact may also work for a large class of non-Gaussian processes (Section 2.4.1). Moreover, we outline how our estimation procedure can be adapted to a more general version of the covariance assumption 1 (Section 2.4.2).

2.4.1. Non-Gaussian processes

To generalize theorem 1 we can build on the framework of elliptical processes which includes the case of non-Gaussian, heavy-tailed distributions, i.e. we can consider processes X_i that depend on some latent random variable V_i such that the conditional distribution of X_i given $V_i=v$ is Gaussian. However, the (unconditional) distribution of X_i then additionally depends on the distribution of V_i and may be far from Gaussian.

Our conditions A and B in the on-line appendix B.2 define a general framework for such non-Gaussian processes X_i and proposition B.1 in appendix B.2 generalizes theorem 1 for this general framework. Here in this subsection, however, we focus on the arguably most important special case of our general framework—namely, the case of elliptically distributed processes. Elliptical distributions include the special case of a Gaussian distribution as considered in theorem 1, but also many important non-Gaussian distributions such as the t-distribution the Laplace distribution and the logistic distribution (see, for instance, Boente *et al.* (2014)).

Let X_i be a (centred) elliptical process, i.e. let $X_i(t)\stackrel{\mathrm{d}}{=}V_iX_i^*(t)$, $t\in[a,b]$, where $V_i>0$ is a strictly positive real-valued random variable and X_i^* is a zero-mean Gaussian process with covariance function $\sigma^*(s,t)$, and where V_i and X_i^* are independent of each other. Moreover, let the error term ε_i in expression (2) be independent of V_i and X_i and let $\mathbb{V}(V_i)<\infty$. Then the elliptically distributed random function X_i fulfils our conditions A and B in the on-line appendix B.2 and it follows by proposition B.1 in appendix B.2 that

$$\mathbb{E}\{X_i(s)Y_i\}=\sum_{r=1}^{S}\sigma^*(s,\tau_r)\mathbb{E}\{V_i^2\vartheta_r(V_i)\}=\sum_{r=1}^{S}\sigma(s,\tau_r)\frac{\mathbb{E}\{V_i^2\vartheta_r(V_i)\}}{\mathbb{V}(V_i)},$$

where

$$\vartheta_r(V_i) = \mathbb{E}\left[\frac{\partial}{\partial x_r} g\{X_i(\tau_1), \ldots, X_i(\tau_S)\} | V_i\right]$$

and $\sigma(s,t) = \mathbb{V}(V_i)\sigma^*(s,t)$ is the covariance function of the elliptically distributed process X_i. As in the case of theorem 1, the above result enables us to decompose the cross-covariance $\mathbb{E}\{X_i(s)Y_i\}$ into a scaling coefficient $\mathbb{E}\{V_i^2\vartheta_r(V_i)\}/\mathbb{V}(V_i)$ which depends on the unknown function g (via ϑ_r) and the covariance function $\sigma(s,\tau_r)$ which depends only on X_i. This result holds for elliptically distributed X_i and requires only mild regularity assumptions on g which are essentially equivalent to those imposed by theorem 2.1; see conditions A and B in appendix B.2.

As in the preceding section, the identification of the points of impact relies only on the structural covariance assumption 1 which holds for rough—Gaussian or non-Gaussian—processes X_i. Since $\sigma(s,t) = \mathbb{V}(V_i)\sigma^*(s,t)$, the requirements of assumption 1 may directly be applied to the covariance function $\sigma^*(s,t)$ of the Gaussian process component X_i^* of the elliptical process X_i. If σ^* satisfies assumption 1 for some $\omega^* : \Omega \to \mathbb{R}$, then proposition B.1 in the on-line appendix B.2 leads to

$$\mathbb{E}\{Z_{\delta,i}(s)Y_i\} = \begin{cases} C(\tau_r)\delta^\kappa + o(\delta^\kappa) & \text{if } s \in \{\tau_1, \ldots, \tau_r\}, \\ O(\delta^2) & \text{if } s \notin \{\tau_1, \ldots, \tau_r\} \end{cases}$$

as $\delta \to 0$ with $C(\tau_r) = c^*(\tau_r)\mathbb{E}\{V_i^2\vartheta_r(V_i)\}$, where

$$c^*(\tau_r) = -\frac{\partial}{\partial z}\omega^*(\tau_r, \tau_r, z)|_{z=0}, \qquad r = 1, \ldots, S.$$

Theorem 2 can also be generalized to the case that X_i is elliptically distributed. Then $Z_{\delta,i}(s)Y_i =^{\mathrm{d}} Z^*_{\delta,i}(s)Y_i^*$, where

$$Z^*_{\delta,i}(s) = X_i^*(s) - \tfrac{1}{2}\{X_i^*(s-\delta) + X_i^*(s+\delta)\},$$

for $s \in [a+\delta, b-\delta]$, and $Y_i^* = V_iY_i$. Therefore, estimating points of impact from data (X_i, Y_i) is equivalent to estimating points of impact from data (X_i^*, Y_i^*). Thus, theorem 2 remains valid if all conditions on X_i and Y_i in theorem 2.2 now apply to X_i^* and Y_i^*.

Our more general framework of conditions A and B in the on-line appendix B.2 includes even more complex cases than the elliptical processes discussed above. For instance, one may consider processes $X_i =^{\mathrm{d}} V_{i1}(t)X_i^*(t) + V_{i2}(t)$, where (V_{i1}, V_{i2}) is jointly independent of X_i^* and where V_{i1} and V_{i2} are almost surely twice continuously differentiable functions on $[a,b]$ (see appendix B.2 for more details).

2.4.2. Generalizing covariance assumption 1

Assumption 1 holds for non-smooth or rough processes X_i with covariance function $\sigma(s,t) = \omega(s,t,|s-t|^\kappa)$, where the requirement $0 < \kappa < 2$ excludes all smooth, twice continuously differentiable processes X_i, with $\kappa \geqslant 2$.

However, the degree of roughness of the processes X_i is actually not a necessary requirement for identifying and estimating points of impact. The crucial property is that the covariance function $\sigma(s,t)$ of X_i is less smooth at the diagonal than for $|t-s| > 0$. For instance, let $\sigma(s,t)$ be $d=4$ times continuously differentiable at all off-diagonal points, $s \neq t$, but *not* $d=4$ times differentiable at the diagonal points, $s=t$. This scenario corresponds to a generalization of assumption 1 with $0 < \kappa < d = 4$ which now excludes only all four-times continuously differentiable processes X_i, with $\kappa \geqslant d = 4$. In this case, we may look at the modified fourth central difference

approximation of the fourth derivative of $\mathbb{E}\{X_i(s)Y_i\}$ and replace $Z_{\delta,i}(s)$ by

$$\tilde{Z}^{(4)}_{\delta,i}(s) := X_i(s) - \tfrac{2}{3}\{X_i(s-\delta) + X_i(s+\delta)\} + \tfrac{1}{6}\{X_i(s-2\delta) + X_i(s+2\delta)\}.$$

Theoretical results may then be derived under a generalized version of assumption 1 demanding that there is a $d=4$ times differentiable function ω such that condition (3) holds for any $\kappa < d = 4$.

Equivalent generalizations can, for instance, be made for any $d \in \{2, 4, 6, 8, \ldots\}$, which would involve then a modified dth-order central difference processes $\tilde{Z}^{(d)}_{\delta,i}(s)$. This way, assumption 1 can be generalized to the requirement $0 < \kappa < d$ which also then includes smooth processes X_i. Deriving the estimation theory under this set-up would then lead to even more accurate points-of-impact estimators with an even faster superconsistent convergence rate. However, taking higher order differences in practice usually involves numerical instabilities.

3. Subsequent estimation of *g*

Given estimates of the points of impact $\tau_1, \ldots, \tau_S$ and their number S, one is typically interested in the subsequent estimation and inference regarding the remaining model components. The following section considers the case of a non-parametric model g. Section 3.2 considers the case of a generalized linear model, which is of particular practical relevance.

In what follows we assume the existence of some consistent estimation procedure for the points of impact satisfying $\max_{r=1,\ldots,\hat{S}} |\hat{\tau}_r - \tau_r| = O_P(n^{-1/\kappa})$ and $P(\hat{S} = S) \to 1$, where we use matched labels in the sense that $\tau_r = \arg\min_{s=1,\ldots,S} |\hat{\tau}_r - \tau_s|$. These requirements are fulfilled by our estimation procedure described in Section 2.2 but may also be fulfilled for alternative procedures.

3.1. Non-parametric estimation

Estimating the non-parametric function g in expression (2) is a non-standard estimation problem, since the unknown points of impact τ_r of the predictor variables $X_i(\tau_r)$ must be replaced by their estimates $\hat{\tau}_r$, i.e. for given estimates $\hat{\tau}_1, \ldots, \hat{\tau}_S$ we may estimate the unknown regression function g by the following Nadaraya–Watson-type estimator

$$\hat{g}_{\hat{\tau}}(x_1, \ldots, x_S) = \sum_{i=1}^{n} K\left\{\frac{X_i(\hat{\tau}_1) - x_1}{h_1}, \ldots, \frac{X_i(\hat{\tau}_S) - x_S}{h_S}\right\} Y_i \Big/ \sum_{i=1}^{n} K\left\{\frac{X_i(\hat{\tau}_1) - x_1}{h_1}, \ldots, \frac{X_i(\hat{\tau}_S) - x_S}{h_S}\right\}, \tag{8}$$

where K denotes a standard non-negative symmetric bounded second-order kernel function with $\int K(u)\mathrm{d}u = 1$, and where $h_1, \ldots, h_S$ denote the bandwidth parameters.

For the following result we make use of our superconsistency result in theorem 2. Note, however, that the rates of consistency for the point-of-impact estimators $\hat{\tau}_r$ of theorem 2 cannot be used directly to quantify the errors $|X_i(\hat{\tau}_r) - X_i(\tau_r)|$, $r = 1, \ldots, S$, since under assumption 1 we cannot make use of Taylor series expansions of X_i. Therefore, the following result is non-standard because of the additional error component

$$\hat{g}_{\hat{\tau}}(x_1, \ldots, x_S) - \hat{g}_{\tau}(x_1, \ldots, x_S) = O_p\left\{\sum_{r=1}^{S} \frac{1}{n^{\min\{1,1/\kappa\}}(h_1 \ldots h_S)h_r^2}\right\}$$

that is contained in equation (9) in the following theorem, where $\hat{g}_{\tau}$ is defined as in equation (8), but using the true predictor variables $X_i(\tau_1), \ldots, X_i(\tau_S)$.

Theorem 3. Let $\hat{S} = S$, $\max_{r=1,\ldots,S} |\hat{\tau}_r - \tau_r| = O_p(n^{-1/\kappa})$, and let assumptions 1 and 2 and the assumptions of theorem 2 hold. Moreover, let the kernel function $K : \mathbb{R}^S \to \mathbb{R}$ be a second-order kernel (i.e. a density function that is symmetric around zero) with continuous second-order partial derivatives and let the regression function g have continuous second-order partial derivatives. We then have for any points $x_1, \ldots, x_S$ in the interior of the support of X that

$$\hat{g}_{\hat{\tau}}(x_1, \ldots, x_S) - g(x_1, \ldots, x_S) = O_p\left\{ \sum_{r=1}^{S} h_r^2 + (nh_1 \ldots h_S)^{-1/2} + \sum_{r=1}^{S} \frac{1}{n^{\min\{1,1/\kappa\}}(h_1 \ldots h_S)h_r^2} \right\} \tag{9}$$

for $n \to \infty$, and $h_1, \ldots, h_S \to 0$ with $n^{\min\{1,1/\kappa\}}(h_1 \ldots h_S)h_r^2 \to \infty$, for each $r = 1, \ldots, S$.

If each bandwidth has the same order of magnitude and $0 < \kappa \leqslant 1$, the well-known optimal bandwidth choice $h_r \sim n^{-1/(S+4)}$, $r = 1, \ldots, S$, can be used to simplify theorem 3 as follows.

Corollary 1. Under the assumptions of theorem 3, let $0 < \kappa \leqslant 1$ and $h_r \sim n^{-1/(S+4)}$ for all $r = 1, \ldots, S$. Then

$$\hat{g}_{\hat{\tau}}(x_1, \ldots, x_S) - g(x_1, \ldots, x_S) = O_p(n^{-2/(S+4)}).$$

That is, under the conditions of corollary 1, we have the same optimal rates of convergence as in the case where the points of impact were known.

3.2. Parametric estimation

In this section it is assumed that the relationship between Y_i and the functional predictor X_i can be modelled by using the framework of generalized linear models with known parametric function g,

$$Y_i = g\left\{ \alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r) \right\} + \varepsilon_i, \tag{10}$$

in which the independent and identically distributed errors term ε_i respects $\mathbb{E}\{\varepsilon_i | X_i(t)\} = 0$ for all $t \in [a, b]$ and where $\mathbb{V}\{\varepsilon_i | X_i(t), t \in [a, b]\} = \sigma^2\{g(\eta_i)\} < \infty$ with strictly positive variance function $\sigma^2(\cdot)$ defined over the range of g. 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 is invertible (see, for instance, Müller and Stadtmüller (2005) for similar assumptions). The constant α enables us to consider centred random functions X_i with $\mathbb{E}\{X_i(t)\} = 0$ for all $t \in [a, b]$. Note that we do not assume that the conditional distribution of Y_i belongs to the exponential family of distributions. Denoting the linear predictor

$$\eta_i = \alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r) \tag{11}$$

enables us to write $\mathbb{E}(Y_i | X_i) = g(\eta_i)$ as well as $\mathbb{V}(Y_i | X_i) = \sigma^2\{g(\eta_i)\} < \infty$. Hence, this set-up of model (10) belongs to the broad class of quasi-likelihood models which can be seen as a generalization of a generalized linear model framework (see McCullagh and Nelder (1989), chapter 9).

Identifiability of the model parameters in model (10) is not obvious because of the functional predictor $X_i(\cdot)$, which, in principle, allows for infinitely many alternative model candidates. The following theorem 4 shows that any possible kind of model misspecification in α, β_r, τ_r, $r = 1, \ldots, S$, or S, will lead to a different model in the mean-squared error sense implying the identifiability of model (10).

Theorem 4. Let $g(\cdot)$ be invertible and assume that X_i satisfies assumptions 1 and 2. Then for all $S^* \geqslant S$, 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

$$\mathbb{E}\left(\left[g\left\{\alpha + \sum_{r=1}^{S} \beta_r X_i(\tau_r)\right\} - g\left\{\alpha^* + \sum_{r=1}^{S^*} \beta_r^* X_i(\tau_r)\right\}\right]^2\right) > 0, \tag{12}$$

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 the proof of theorem 4 does only require the existence of second moments and, therefore, may be generalized also to the case of non-Gaussian processes X_i.

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

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

$$\hat{\mathbf{U}}_n(\boldsymbol{\beta}) = \hat{\mathbf{D}}_n(\boldsymbol{\beta})^{\mathrm{T}}\hat{\mathbf{V}}_n(\boldsymbol{\beta})^{-1}(\mathbf{Y}_n - \hat{\boldsymbol{\mu}}_n(\boldsymbol{\beta})). \tag{13}$$

Note that these are non-classic score equations evaluated at the estimates $\hat{\tau}_r$ instead of τ_r.

In what follows, it will be convenient to define

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

and

$$\hat{\mathbf{F}}_n(\boldsymbol{\beta}) = \hat{\mathbf{D}}_n(\boldsymbol{\beta})^{\mathrm{T}}\hat{\mathbf{V}}_n(\boldsymbol{\beta})^{-1}\hat{\mathbf{D}}_n(\boldsymbol{\beta}).$$

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 $\mathbf{X}_i(\boldsymbol{\tau})$ respectively. This enables us to write $\mathbb{E}\{n^{-1}\mathbf{F}_n(\boldsymbol{\beta})\} = \mathbb{E}\{\mathbf{F}(\boldsymbol{\beta})\}$ with

$$\mathbb{E}\{\mathbf{F}(\boldsymbol{\beta})\} = [\mathbb{E}(g'\{\eta(\boldsymbol{\beta})\}^2/\sigma^2[g\{\eta(\boldsymbol{\beta})\}]X_kX_l)]_{k,l},$$

where we point out that $\mathbb{E}\{\mathbf{F}(\boldsymbol{\beta})\}$ is for all $\boldsymbol{\beta} \in \mathbf{R}^{S+1}$ a symmetric and strictly positive definite matrix with inverse $\mathbb{E}\{\mathbf{F}(\boldsymbol{\beta})\}^{-1}$. Indeed, suppose that $\mathbb{E}\{\mathbf{F}(\boldsymbol{\beta})\}$ were not strictly positive definite; we would then derive the contradiction

$$\mathbb{E}\left\{\left(\sum_{j=1}^{S+1} \frac{a_j X_j g'\{\eta(\boldsymbol{\beta})\}}{\sigma[g\{\eta(\boldsymbol{\beta})\}]}\right)^2\right\} = 0$$

for non-zero constants $a_1, \ldots, a_{S+1}$. A similar argument can be used to show that $\mathbb{E}\{\hat{\mathbf{F}}(\boldsymbol{\beta})\}$ is strictly positive definite, where

$$\mathbb{E}\{\hat{\mathbf{F}}(\boldsymbol{\beta})\} = \left[\mathbb{E}\left(\frac{g'\{\hat{\eta}(\boldsymbol{\beta})\}^2}{\sigma^2[g\{\hat{\eta}(\boldsymbol{\beta})\}]}\hat{X}_k\hat{X}_l\right)\right]_{k,l}.$$

The following additional set of assumptions is used to derive more precise theoretical statements.

Assumption 3.

(a) There is a constant $0 < M_\varepsilon < \infty$, such that $\mathbb{E}\{\varepsilon_i^p | X_i(t)\} \leqslant M_\varepsilon$, for all $t \in [a, b]$ and for some even p with $p \geqslant \max\{2/\kappa + \epsilon, 4\}$ and some $\epsilon > 0$.
(b) The function g is monotone, invertible, with two bounded derivatives $|g'(\cdot)| \leqslant c_g$, $|g''(\cdot)| \leqslant c_g$, for some constant $0 \leqslant c_g < \infty$.
(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. Although the condition on $p \geqslant 4$ and p being even simplifies the proofs, the condition $p > 2/\kappa$ is more crucial and is used in the proof of proposition D.2 in the on-line supplementary appendix D.2. Conditions (a)–(c) hold, for example, in the important case of a functional logistic regression with points of impact, where g is the standard logistic function. Condition (c) is satisfied, for instance, in the special case of generalized linear models with natural link functions. For that case, we have $\sigma^2\{g(x)\} = g'(x)$ such that $h(x) = 1$.

Theorem 5. Let $\hat{S} = S$, $\max_{r=1,\ldots,S} |\hat{\tau}_r - \tau_r| = O_P(n^{-1/\kappa})$ and let X_i be a Gaussian process satisfying assumption 1. Under assumption 3 we then obtain

$$\sqrt{n^{1/2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathrm{d}} N[\mathbf{0}, \mathbb{E}\{\mathbf{F}(\boldsymbol{\beta}_0)\}^{-1}]. \tag{14}$$

That is, our estimator based on $\hat{\tau}_r$ enjoys the same asymptotic efficiency properties as if the true points of impact τ_r were known. In fact, it achieves the same asymptotic efficiency properties as under classical multivariate set-ups (see McCullagh (1983)). In practice one might replace $\mathbb{E}\{\mathbf{F}(\boldsymbol{\beta}_0)\}$ with its consistent estimator $n^{-1}\hat{\mathbf{F}}_n(\hat{\boldsymbol{\beta}})$ to derive approximate results. This is a direct consequence of equations (129) and (155) in the on-line supplementary appendix D.2.

3.2.1. Parametric estimation: practical implementation

An implementation of our parametric 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 2.2). Asymptotic specifications are given in theorem 2; however, these determine the tuning parameters δ and λ up to constants only and are generally of a limited use in practice. In what follows 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 $\{\hat{\tau}_1, \hat{\tau}_2, \ldots, \hat{\tau}_{M_\delta}\}$ (see Section 2.2). 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 logistic regression) one may perform a best subset selection optimizing a standard model selection criterion such as the Bayesian information criterion (BIC):

$$\mathrm{BIC}_{\mathcal{X}}(\delta) = -2\log(\mathcal{L}_{\mathcal{X}}) + K_{\mathcal{X}}\log(n). \tag{15}$$

Here, $\log(\mathcal{L}_{\mathcal{X}})$ is the log-likelihood of the model with intercept and predictor variables $\mathcal{X} \subseteq \{X_i(\hat{\tau}_1), X_i(\hat{\tau}_2), \ldots, X_i(\hat{\tau}_{M_\delta})\}$, 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 more general case of quasi-likelihood models (see McCullagh and Nelder (1989), chapter 9) 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, we may replace the deviance $-2\log(\mathcal{L}_{\mathcal{X}})$ by the expression for the quasi-deviance

$$-2Q_{\mathcal{X}} = -2\sum_{i=1}^{n}\int_{y_i}^{g(\hat{\eta}_{\mathcal{X},i})}\frac{y_i - t}{\sigma^2(t)}\,\mathrm{d}t,$$

where $\hat{\eta}_{\mathcal{X},i}$ is the linear predictor with intercept and predictor variables $\mathcal{X}$.

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

$$\hat{\boldsymbol{\beta}}_m = \hat{\boldsymbol{\beta}}_{m-1} + \hat{\mathbf{F}}_n(\hat{\boldsymbol{\beta}}_{m-1})^{-1}\hat{\mathbf{U}}_n(\hat{\boldsymbol{\beta}}_{m-1}). \tag{16}$$

Iteration is executed until convergence and the final step of the procedure yields the estimate $\hat{\boldsymbol{\beta}}$. Here, $\hat{\mathbf{F}}_n(\boldsymbol{\beta})$ and $\hat{\mathbf{U}}_n(\boldsymbol{\beta})$ replace $\mathbf{F}_n(\boldsymbol{\beta})$ and $\mathbf{U}_n(\boldsymbol{\beta})$ in the usual Fisher scoring algorithm, since the unknown τ_r, $1 \leqslant r \leqslant S$, are replaced by their estimates $\hat{\tau}_r$. This replacement is justified asymptotically by our results in corollary D.1 and proposition D.3 in the on-line appendix D.2.

4. Simulation

We investigate the finite sample performance of our estimators by 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, 2020), where we use the R package `glmulti` (Calcagno, 2013) to implement the fully data-driven BIC-based best subset selection estimation procedure that was described in Section 3.2.1. The threshold estimator from Section 2.2 requires appropriate choices of $\delta = \delta_n$ and $\lambda = \lambda_n$. Theorem 2 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 $\lambda = A[\hat{\mathbb{E}}(Y^4)^{1/2}\log\{(b-a)/\delta\}/n]^{1/2}$, where $A = \sqrt{(2\sqrt{3})}$ and $\hat{\mathbb{E}}(Y^4) = n^{-1}\sum_{i=1}^{n} Y_i^4$, as motivated below theorem 2.

In what follows, we denote the BIC-based selection (see Section 3.2.1) of points of impact by POI and the threshold-based selection (algorithm 1) by TRH. Estimated points of impact candidates are related to the true impact points by the following matching rule: in the 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 $\hat{\tau}_l$ in interval I_j with the closest distance to τ_j is then taken as the estimate of τ_j.

The simulation results for our parametric estimation procedure (Section 3.2) are based on 1000 Monte Carlo iterations for each constellation of $n \in \{100, 200, 500, 1000, 3000\}$ and $p \in \{100, 500, 1000\}$. The results for our non-parametric estimation procedure (Section 3.1) are based on the same general set-up but consider the reduced set of sample sizes $n \in \{100, 200, 500\}$. Estimation errors for the parametric estimation procedure are illustrated by boxplots with error bars representing the 10% and 90% quantiles. The estimation errors for the non-parametric estimation procedure are quantified by the mean average squared error

$$\mathrm{MASE} = 1000^{-1}\sum_{r=1}^{1000} n^{-1}\sum_{i=1}^{n}[g(\eta_i) - \hat{g}^r_{\hat{\tau}}\{X_i^r(\hat{\tau}_1^r), \ldots, X_i^r(\hat{\tau}_{\hat{S}}^r)\}]^2,$$

Table 2. DGPs 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†	$\frac{1}{2}$				1	4			
DGP 2	OUP	2	$\frac{1}{3}$	$\frac{2}{3}$			1	−6	5		
DGP 3	OUP	4	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1	−6	6	−5	5
DGP 4	GCM	2	$\frac{1}{3}$	$\frac{2}{3}$			1	−6	5		
DGP 5	EBM	2	$\frac{1}{3}$	$\frac{2}{3}$			1	−6	5		

†$S=1$ is assumed known (only for DGP 1).

where the superscript r denotes the rth simulation run.

Five data-generating processes (DGPs) are considered (Table 2) using the following three processes $\{X_i(t) : 0 \leqslant t \leqslant 1\}$ covering a broad range of situations.

(a) OUP, 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 $\sigma_u^2 = 3.5$;
(b) Gaussian covariance model (GCM), a Gaussian process with covariance function $\sigma(s,t) = \sigma(|s-t|) = \exp\{-(|s-t|/d)^2\}$; we choose $d = 1/10$;
(c) exponential Brownian motion (EBM), 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 Brownian motion.

DGPs 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 many times differentiable, even at the diagonal where $s=t$, contradicting assumption 1, but fitting the remark underneath assumption 1. DGP 4 contradicts the Gaussian assumption 2.

4.1. Evaluation of the parametric estimation procedure

DGP 1 enables us to compare our data-driven BIC-based estimation procedure from Section 3.2.1 (denoted as POI) with the estimation procedure of Lindquist and McKeague (2009) (denoted as LMcK). Lindquist and McKeague (2009) considered situations where $S=1$ is known and proposed 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 comparability, we restrict the BIC-based model selection process to allow only for models containing one point-of-impact candidate. The simulation results are depicted in Fig. 3 and are virtually identical for both methods and show satisfying behaviour of the estimates. However, 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 $|n^{-1}\Sigma_{i=1}^n Z_{\delta,i}(s)Y_i|$ as possible point-of-impact candidates. Our threshold-based estimation procedure leads to similar qualitative results. We omit these results, however, to enable a clear display in Fig. 3. The performance of our threshold-based procedure is reported in detail for the remaining simulation studies (DGPs 2–5).

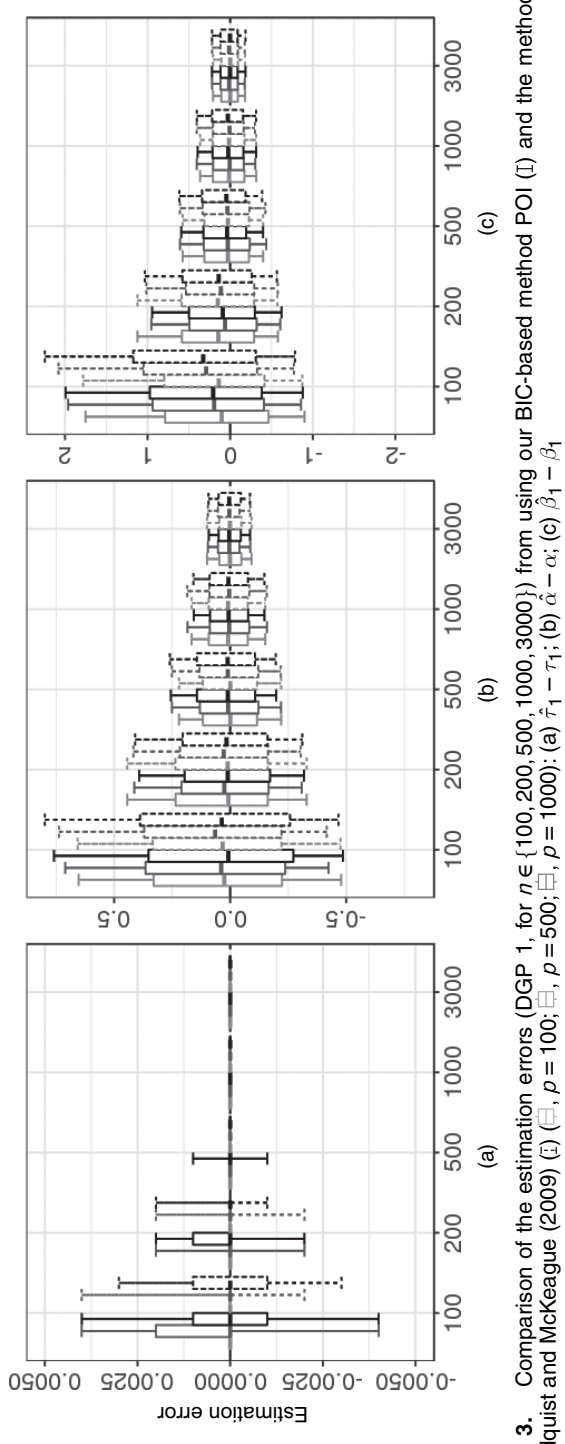

Fig. 3. Comparison of the estimation errors (DGP 1, for $n \in \{100, 200, 500, 1000, 3000\}$) from using our BIC-based method POI (I) and the method of Lindquist and McKeague (2009) (I), ($p = 100$; , $p = 500$; , $p = 1000$): (a) $\hat{\tau}_1 - \tau_1$; (b) $\hat{\alpha} - \alpha$; (c) $\hat{\beta}_1 - \beta_1$

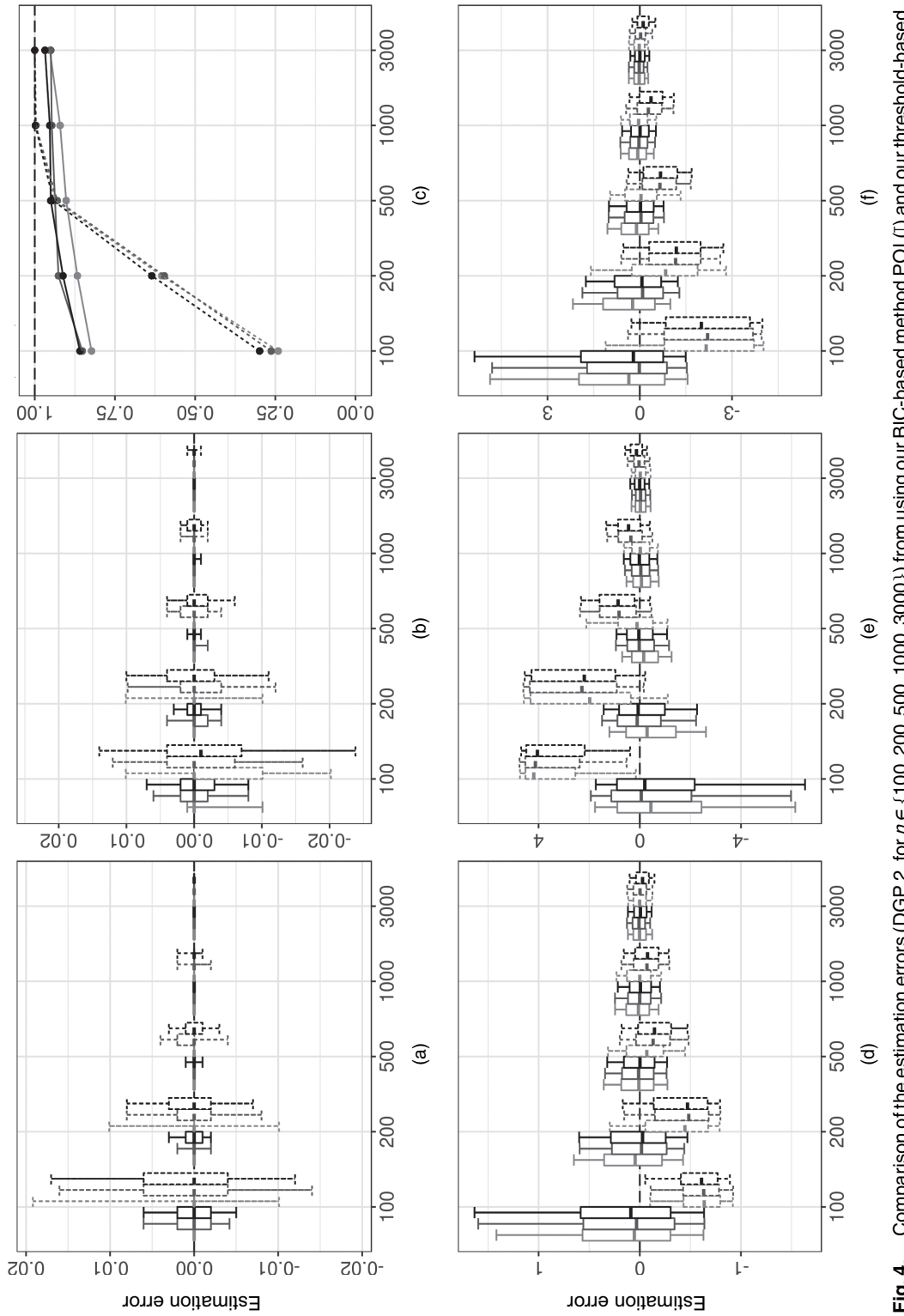

Fig. 4. Comparison of the estimation errors (DGP 2, for $n \in \{100, 200, 500, 1000, 3000\}$) from using our BIC-based method POI (□) and our threshold-based method TRH (□) (□, $p = 100$; □, $p = 500$; □, $p = 1000$): (a) $\hat{\tau}_1 - \tau_1$; (b) $\hat{\tau}_2 - \tau_2$; (c) $\hat{P}(\hat{S} = S)$; (d) $\hat{\alpha} - \alpha$; (e) $\hat{\beta}_1 - \beta_1$; (f) $\hat{\beta}_2 - \beta_2$

DGP 2 is more complex than DGP 1 since $S = 2$ and is considered unknown. Fig. 4 compares the estimation errors from using our BIC-based POI estimator with those from our threshold-based estimator (denoted as TRH). For smaller sample sizes n, the POI estimator seems to be preferable to the TRH estimator. Although estimating the locations of the points of impact τ_1 and τ_2 is quite accurate for both procedures, the number S is estimated correctly more often by using the POI estimator (see Fig. 4(c)). The more precise estimation of S when using the POI estimator results in essentially unbiased estimates of the parameters α, β_1 and β_2. By contrast, the less precise estimation of S by 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 $\hat{S}$ improves for the TRH estimator such that both estimators show eventually a similar performance.

DGP 3 with $S = 4$ unknown points of impact comprises an even more complex situation than DGP 2. For brevity, Fig. 7 has been deferred to the on-line 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 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 it may be inappropriate for small n, since the estimation procedure eliminates all points within a $\sqrt{\delta}$-neighbourhood around a chosen candidate $\hat{\tau}_r$ (see Section 2.2). For DGP 3, the choice of $c_\delta = 1.5$ results in a too-large $\sqrt{\delta}$-neighbourhood, such that the estimation procedure also eliminates true point-of-impact locations for small n. By contrast, the POI estimator can avoid such adverse eliminations as the BIC value is also minimized over δ.

DGP 4 takes up the general set-up of DGP 2, but the functional data X_i are simulated by using a GCM which is characterized by an infinitely many times differentiable covariance function. This setting contradicts our basic assumption 1 but fits our remark at the end of Section 2.1. From Fig. 5 it can be concluded that, even under the failure of assumption 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 1. Here the TRH estimator can estimate the true points of impact but additionally selects increasingly more redundant point-of-impact candidates as n becomes large, i.e. the TRH estimator becomes more a screening than a selection procedure, which can be problematic in practice. By contrast, the POI estimator can avoid such redundant selections of point-of-impact candidates, as the BIC selects point-of-impact candidates only if they result in a sufficiently large improvement in the model fit.

DGP 5 also takes up the set-up of DGP 2; however, the process X_i is simulated as an EBM violating assumption 2, but still satisfying assumption 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 Fig. 8 in the on-line 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.

4.2. Evaluation of the non-parametric estimation procedure

Table 3 contains the simulation results for our non-parametric estimation procedure described in Section 3.1. We focus on the more challenging DGPs 2–5, with at least two points of impact and

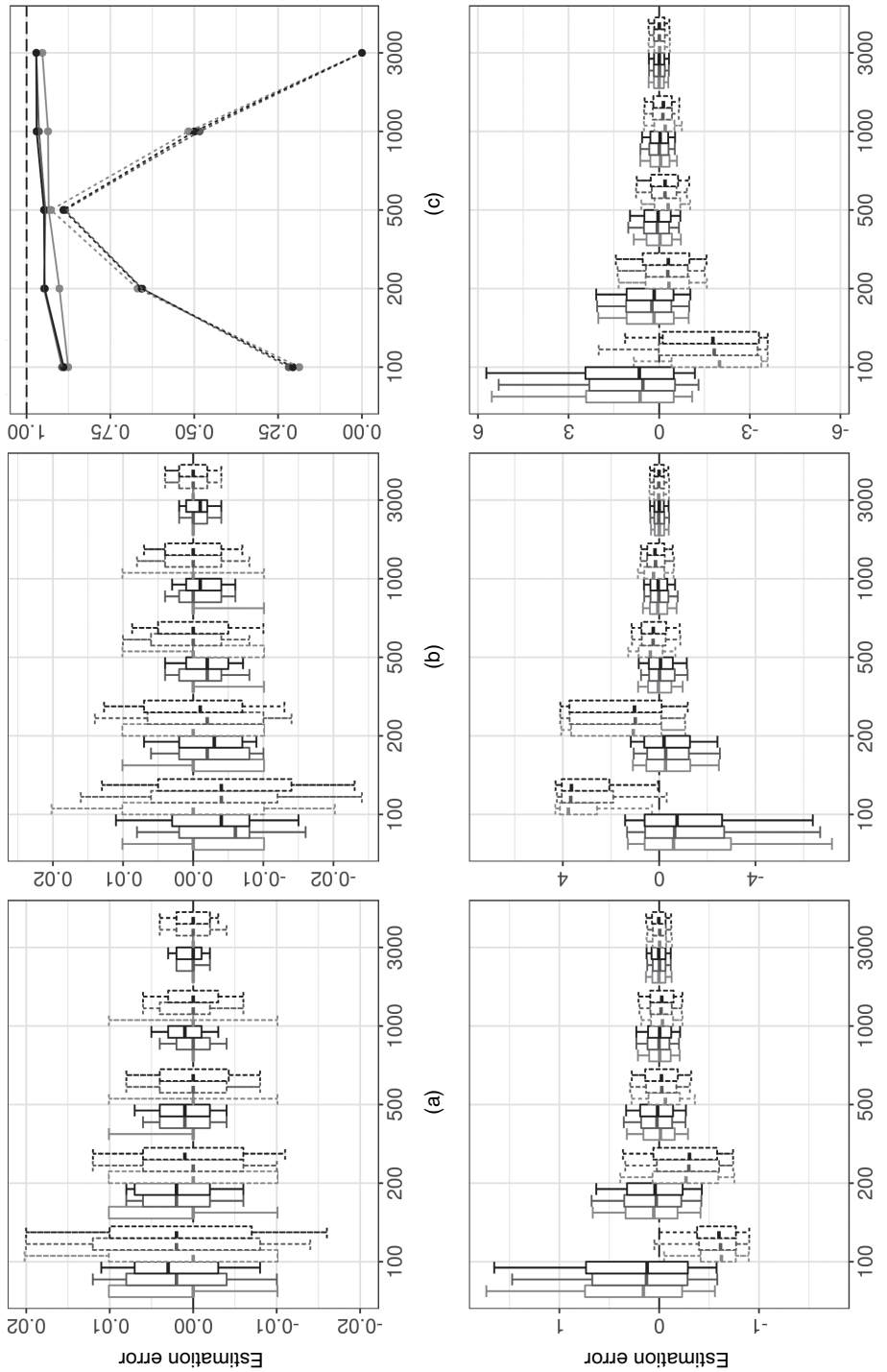

Fig. 5. Comparison of the estimation errors (DGP 4, for $n \in \{100, 200, 500, 1000, 3000\}$) from using our BIC-based method POI (□) and our threshold-based method TRH (□) (□, $p = 100$; □, $p = 500$; □, $p = 1000$): (a) $\hat{\tau}_1 - \tau_1$; (b) $\hat{\tau}_2 - \tau_2$; (c) $\hat{P}(\hat{S} = S)$; (d) $\hat{\alpha} - \alpha$; (e) $\hat{\beta}_1 - \beta_1$; (f) $\hat{\beta}_2 - \beta_2$

Table 3. MASE for the non-parametric estimator $\hat{g}_{\hat{\tau}}$

DGP	p	n	*MASE*		*DGP*	p	n	*MASE*	
			TRH	*MPDP*				*TRH*	*MPDP*
2	100	100	0.098	0.100	4	100	100	0.089	0.093
2	100	200	0.061	0.089	4	100	200	0.044	0.087
2	100	500	0.017		4	100	500	0.011	
2	500	100	0.097	0.098	4	500	100	0.085	0.096
2	500	200	0.064	0.092	4	500	200	0.045	0.087
2	500	500	0.023		4	500	500	0.010	
2	1000	100	0.094		4	1000	100	0.086	
2	1000	200	0.060		4	1000	200	0.045	
2	1000	500	0.022		4	1000	500	0.010	
3	100	100	0.155	0.175	5	100	100	0.096	0.180
3	100	200	0.105	0.156	5	100	200	0.089	0.177
3	100	500	0.058		5	100	500	0.069	
3	500	100	0.150	0.173	5	500	100	0.094	0.179
3	500	200	0.102	0.161	5	500	200	0.090	0.176
3	500	500	0.060		5	500	500	0.069	
3	1000	100	0.149		5	1000	100	0.092	
3	1000	200	0.100		5	1000	200	0.091	
3	1000	500	0.059		5	1000	500	0.066	

compare our non-parametric method with the most-predictive design point (MPDP) method of Ferraty *et al.* (2010). To the best of our knowledge, the MPDP method is the only comparable method in the literature. We tried hard to carry out the full simulation study for the MPDP method; however, Ferraty *et al.* (2010) used a brute force minimization approach based on cross-validation considering 2^p grid point combinations, which makes their method computationally extremely expensive. (Because of the high computational costs, the simulation study in Ferraty *et al.* (2010) is based on only 50 Monte Carlo replications. In a 'readme' file, provided at Frederic Ferraty's homepage, the authors report that one run with a data set of $n = 149$ curves and $p = 700$ grid points lasts about 30 min.) For the MPDP method, we, therefore, had to limit the number of Monte Carlo replications to 500, the number of grid points to $p \in \{100, 500\}$ and the sample sizes to $n \in \{100, 200\}$.

The results in Table 3 show that MASE decreases with increasing sample size n and that the effect of different numbers of grid points p is essentially negligible for both methods. The differences in the simulation results for the different DGPs are generally equivalent to those discussed for the parametric estimation procedure. DGP 3 with its four points of impact is the most challenging case and, therefore, produces the largest estimation errors. The MPDP method of Ferraty *et al.* (2010) has throughout larger estimation errors than does our non-parametric estimation results based on the TRH estimator (algorithm 1). The larger estimation errors in $\hat{g}$ of the MPDP method can be explained by its larger estimation errors when estimating the points of impact $\tau_1, \ldots, \tau_S$ (Table 4). In fact, our superconsistent points-of-impact estimator has substantially smaller estimation errors (factors of from 1/10 to 1/100) than does the MPDP method.

5. 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

Table 4. Average mean-squared errors AvgMSE for $\hat{\tau}_1, \ldots, \hat{\tau}_S$†

DGP	*p*	*n*	*AvgMSE*		*DGP*	*p*	*n*	*AvgMSE*	
			TRH	*MPDP*				*TRH*	*MPDP*
2	100	100	0.0002	0.0063	4	100	100	0.0003	0.0072
2	100	200	0.0001	0.0023	4	100	200	0.0001	0.0029
2	100	500	0.0000		4	100	500	0.0001	
2	500	100	0.0002	0.0084	4	500	100	0.0002	0.0062
2	500	200	0.0001	0.0013	4	500	200	0.0001	0.0023
2	500	500	0.0000		4	500	500	0.0000	
2	1000	100	0.0002		4	1000	100	0.0002	
2	1000	200	0.0001		4	1000	200	0.0001	
2	1000	500	0.0000		4	1000	500	0.0000	
3	100	100	0.0002	0.0186	5	100	100	0.0004	0.0111
3	100	200	0.0001	0.0036	5	100	200	0.0006	0.0025
3	100	500	0.0000		5	100	500	0.0002	
3	500	100	0.0002	0.0218	5	500	100	0.0004	0.0097
3	500	200	0.0001	0.0035	5	500	200	0.0006	0.0009
3	500	500	0.0000		5	500	500	0.0001	
3	1000	100	0.0002		5	1000	100	0.0004	
3	1000	200	0.0001		5	1000	200	0.0007	
3	1000	500	0.0000		5	1000	500	0.0002	

†$\text{AvgMSE} = S^{-1}\Sigma_{l=1}^{S}\text{MSE}(\hat{\tau}_l)$.

validity (Trautmann *et al.*, 2009). Asking participants to evaluate those stimuli is most often done by using an overall rating such as 'How positive or negative did this video make you feel?'. 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 to understand the emergence of emotional states and to make use of specific 'impacting' parts of the stimuli.

Because of a lack of appropriate statistical methods, existing approaches use heuristics such as the 'PER rule' to link the overall ratings with the continuous emotional stimuli. The PER rule states that people's evaluations can be well predicted by using just two characteristics: the moment of emotional peak intensity and the ending of the emotional stimuli (Fredrickson, 2000). Such a heuristic approach, however, is only of 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 specific moments in the continuous emotional stimuli that are of a *common* relevance.

Our case-study comprises data from $n = 65$ participants, who were asked to report their emotional state continuously (from very negative to very positive) while watching a documentary video (112 s) on the persecution of African albinos. A version of the video can be found on line in YouTube (`https://youtu.be/9F6UpuJIFaY`; the video clip that was used in the experiment corresponds approximately to the first 115 s of the video in YouTube). The first six data points (within 1 s) have been removed as they contain some obviously erratic components. Fig. 1 shows the standardized emotion trajectories $X_i(t_j)$, where t_j are equidistant grid points within the unit interval $0 = t_1 < \ldots < t_p = 1$ 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 $Y_i \in \{0, 1\}$, 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

Table 5. Estimation results, with standard errors in parentheses

Regressor	*POI coefficients*		*PER 1 coefficients*		*PER 2 coefficients*	
$X(\hat{\tau}_1)$	1.16†	(0.41)				
$X(\hat{\tau}_2)$	0.71‡	(0.32)				
$X(p^{\text{abs}})$			0.41	(0.36)		
$X(p^{\text{pos}})$					0.46	(0.29)
$X(p^{\text{neg}})$					0.54	(0.43)
$X(1)$			0.20	(0.26)	0.04	(0.28)
Constant	−0.29	(0.29)	−0.98	(0.71)	−0.29	(0.73)
Log-likelihood	−36.03		−43.48		−41.58	
Akaike information criterion	78.07		92.96		91.15	
McFadden pseudo-R^2	0.19		0.03		0.07	
Somers's D_{xy}	0.53		0.20		0.34	

†p-value <0.01.
‡p-value <0.05.

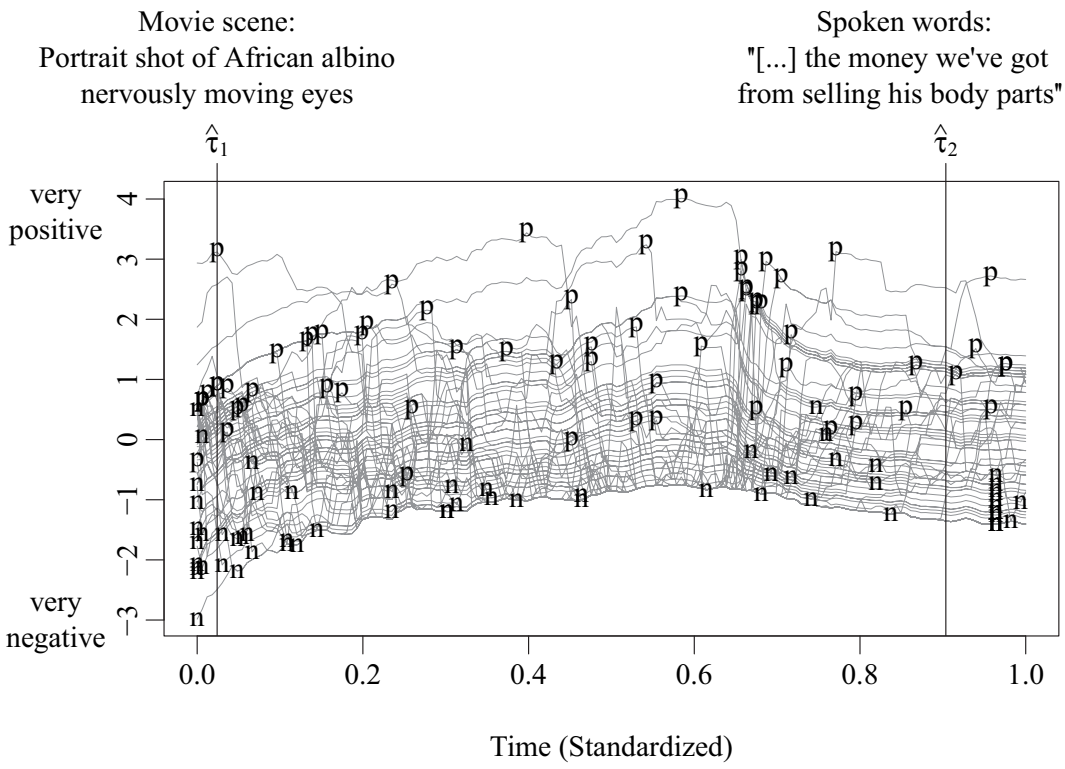

Fig. 6. Visualization of the impact points $\hat{\tau}_1$ and $\hat{\tau}_2$ (|): positive 'p' and negative 'n' peak intensity predictors $X_i(p_i^{\text{pos}})$ and $X_i(p_i^{\text{neg}})$

were recruited through Amazon Mechanical Turk (`www.mturk.com`) and received $1 for completing the ratings via the on-line survey platform SoSci Survey (`www.soscisurvey.de`). The study was approved by the local institutional review board (University of Colorado Boulder). The documentary video is taken from the Interdisciplinary Affective Science Laboratory movie set (L. Feldman Barrett, unpublished).

To analyse the data we use our parametric estimation procedure (Section 3.2) using a logit link function g and the BIC-based selection of points of impact (Section 3.2.1). We compare our estimation procedure with the performance of the following two logit regression models based on PER predictor variables:

(a) PER model 1, logit regression with peak intensity predictor $X_i(p_i^{\mathrm{abs}})$ and the end feeling predictor $X_i(1)$, where $p_i^{\mathrm{abs}} = \arg\max_t(|X_i(t)|)$;
(b) PER model 2, logit regression with peak intensity predictors $X_i(p_i^{\mathrm{pos}})$ and $X_i(p_i^{\mathrm{neg}})$ and end feeling predictor $X_i(1)$, where $p_i^{\mathrm{pos}} = \arg\max_t\{X_i(t)\}$ and $p_i^{\mathrm{neg}} = \arg\min_t\{X_i(t)\}$.

Table 5 shows the estimated coefficients and standard errors, as well as summary statistics for each of the three models, where our estimation procedure is denoted by POI. In comparison with 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's D_{xy}), where $D_{xy} = 0$ means that a model is making random predictions and $D_{xy} = 1$ means that a model discriminates perfectly.

Fig. 6 shows the positive p and negative n peak intensity predictors $X_i(p_i^{\mathrm{pos}})$ and $X_i(p_i^{\mathrm{neg}})$ for all participants; the absolute intensity predictors $X_i(p_i^{\mathrm{abs}})$ form a subset of these. 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 $t \in [0, 1]$ in the continuous emotional stimuli. By contrast, the estimated points of impact $\hat{\tau}_1$ and $\hat{\tau}_2$ allow for such a link and point to two emotionally arousing movie scenes:

(a) $\hat{\tau}_1$, a portrait shot of the traumatized African albino protagonist nervously moving eyes;
(b) $\hat{\tau}_2$, spoken words, '... the money we've got from selling his body parts'.

6. Supplementary materials

The on-line supplementary materials include the supplementary paper containing additional simulation results and the proofs of our theoretical results, the R package `fdapoi` and R scripts for reproducing our main empirical results.

Acknowledgements

The on-line rating tool for the data collection was kindly provided by Dominik Leiner (SoSci Survey, Germany). Many thanks go to the Joint Editor, the Associate Editor and two referees whose constructive comments helped us to improve our manuscript and motivated Section 2.4.

Data collection was funded by a National Institutes of Health Director's pioneer award (DP1OD003312) to Lisa Feldman Barrett, and National Institute on Drug Abuse grant (R01DA035484) to Tor D. Wager. The development of the Interdisciplinary Affective Science Laboratory movie set was supported by a grant from the US Army Research Institute for the Behavioral and Social Sciences (grant W5J9CQ-11-C-0046) to Lisa Feldman Barrett and Tor D. Wager. The views, opinions and/or findings that are contained in this paper are those of the authors and should not be construed as an official US Department of the Army position, policy or decision, unless so designated by other documents.

References

Aneiros, G. and Vieu, P. (2014) Variable selection in infinite-dimensional problems. *Statist. Probab. Lett.*, **94**, 12–20.

Berrendero, J. R., Bueno-Larraz, B. and Cuevas, A. (2019) An RKHS model for variable selection in functional linear regression. *J. Multiv. Anal.*, **170**, 25–45.
Boente, G., Barrera, M. S. and Tyler, D. E. (2014) A characterization of elliptical distributions and some optimality properties of principal components for functional data. *J. Multiv. Anal.*, **131**, 254–264.
Calcagno, V. (2013) glmulti: model selection and multimodel inference made easy. *R Package Version 1.0.7*.
Dagsvik, J. K. and Strøm, S. (2006) Sectoral labour supply, choice restrictions and functional form. *J. Appl. Econmetr.*, **21**, 803–826.
Embrechts, P. and Maejima, M. (2000) An introduction to the theory of self-similar stochastic processes. *Int. J. Mod. Phys.* B, **14**, 1399–1420.
Ferraty, F., Hall, P. and Vieu, P. (2010) Most-predictive design points for functional data predictors. *Biometrika*, **97**, 807–824.
Floriello, D. and Vitelli, V. (2017) Sparse clustering of functional data. *J. Multiv. Anal.*, **154**, 1–18.
Fredrickson, B. L. (2000) Extracting meaning from past affective experiences: the importance of peaks, ends, and specific emotions. *Cogn. Emtn*, **14**, 577–606.
Kneip, A., Poß, D. and Sarda, P. (2016) Functional linear regression with points of impact. *Ann. Statist.*, **44**, 1–30.
Lee, C. and Ready, M. J. (1991) Inferring trade direction from intraday data. *J. Finan.*, **46**, 733–746.
Levina, E., Wagaman, A., Callender, A., Mandair, G. and Morris, M. (2007) Estimating the number of pure chemical components in a mixture by maximum likelihood. *J. Chemometr.*, **21**, 24–34.
Liebl, D., Rameseder, S. and Rust, C. (2020) Improving estimation in functional linear regression with points of impact: insights into Google AdWords. *Preprint arXiv:1709.02166*.
Lindquist, M. A. (2012) Functional causal mediation analysis with an application to brain connectivity. *J. Am. Statist. Ass.*, **107**, 1297–1309.
Lindquist, M. A. and McKeague, I. W. (2009) Logistic regression with Brownian-like predictors. *J. Am. Statist. Ass.*, **104**, 1575–1585.
Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H. and Gross, J. J. (2005) The tie that binds?: Coherence among emotion experience, behavior, and physiology. *Emotion*, **5**, no. 2, 175–190.
McCullagh, P. (1983) Quasi-likelihood functions. *Ann. Statist.*, **11**, 59–67.
McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
McKeague, I. W. and Sen, B. (2010) Fractals with point impact in functional linear regression. *Ann. Statist.*, **38**, 2559–2586.
Müller, H.-G. and Stadtmüller, U. (2005) Generalized functional linear models. *Ann. Statist.*, **33**, 774–805.
Park, A. Y., Aston, J. A. and Ferraty, F. (2016) Stable and predictive functional domain selection with application to brain images. *Preprint arXiv:1606.02186*.
R Core Team (2020) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Rohlfs, R. V., Harrigan, P. and Nielsen, R. (2013) Modeling gene expression evolution with an extended Ornstein-Uhlenbeck process accounting for within-species variation. *Molec. Biol. Evoln*, **31**, 201–211.
Schubert, E. (1999) Measuring emotion continuously: validity and reliability of the two-dimensional emotion-space. *Aust. J. Psychol.*, **51**, 154–165.
Sobel, M. E. and Lindquist, M. A. (2014) Causal inference for fMRI time series data with systematic errors of measurement in a balanced on/off study of social evaluative threat. *J. Am. Statist. Ass.*, **109**, 967–976.
Stein, M. (1999) *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer.
Trautmann, S. A., Fehr, T. and Herrmann, M. (2009) Emotions in motion: dynamic compared to static facial expressions of disgust and happiness reveal more widespread emotion-specific activations. *Brain Res.*, **1284**, 100–115.
Zhang, Y. (2012) Sparse selection in Cox models with functional predictors. *PhD Thesis*.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplement to "Super-consistent estimation of points of impact in nonparametric regression with functional predictors"'.