Robust depth-based estimation of the functional autoregressive model

Israel Martínez-Hernández [a,*], Marc G. Genton [a], Graciela González-Farías [b]

[a] *Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia*
[b] *Centro de Investigación en Matemáticas, A.C. Jalisco S/N, Col. Valenciana 36240, Guanajuato, Mexico*

ARTICLE INFO

ABSTRACT

A robust estimator for functional autoregressive models is proposed, the Depth-based Least Squares (DLS) estimator. The DLS estimator down-weights the influence of outliers by using the functional directional outlyingness as a centrality measure. It consists of two steps: identifying the outliers with a two-stage functional boxplot, then down-weighting the outliers using the functional directional outlyingness. Theoretical properties of the DLS estimator are investigated such as consistency and boundedness of its influence function. Through a Monte Carlo study, it is shown that the DLS estimator performs better than estimators based on Principal Component Analysis (PCA) and robust PCA, which are the most commonly used. To illustrate a practical application, the DLS estimator is used to analyze a dataset of ambient CO_2 concentrations in California.

1. Introduction

The analysis of functional data has received sustained interest in recent years because data can be obtained easily and many phenomena can be thought of as the realization of a continuous stochastic process, e.g., yield curves, intraday price curves, temperature records, or electricity consumption; see Ramsay and Silverman (2005). There is special interest in finding relationships among different functional data. As in the analysis of data in the scalar case, regression models have been used to identify these possible relations. Most of the developed models in functional regression consider the response variable as a scalar (Cardot et al., 2003; Goldsmith et al., 2011; Attouch et al., 2012; Maronna and Yohai, 2013; Tang, 2014; Boente and Vahnovan, 2017). However, little research exists on the case of functional predictor with a functional response.

A particular case of a functional response with a functional predictor is when the functional data have a temporal dependence, i.e., we can predict the present using the past. One of the most popular models for functional data with a temporal dependence structure is the functional autoregressive model of order p: FAR(p). In practice, the autoregressive order p takes small values such as 1 or 2 (Kokoszka and Reimherr, 2013). Most research has assumed $p = 1$, and many analyses have been developed with FAR(1) (Antoch et al., 2010; Horváth et al., 2010; Kokoszka and Reimherr, 2013). We, too, assume that the FAR(1) model holds for our data. However, we also suppose that the observed data could be contaminated. This motivates us to propose a robust estimator for the FAR(1) model.

To define our proposed estimator, we first introduce the FAR(1) model. Let $\mathcal{H} = L^2(D, \mathcal{B}(D), \lambda)$ be the Hilbert space of square integrable functions defined on compact subsets $D \subset \mathbb{R}$, with inner product $\langle f, g \rangle = \int_D f(x)g(x)\,dx$, and λ the Lebesgue measure. Let $\mathcal{B}_{\mathcal{H}} = \{\rho : \mathcal{H} \to \mathcal{H}; \ \rho \text{ is bounded and linear}\}$ be the set of all bounded linear operators from $\mathcal{H}$ to $\mathcal{H}$. Let $L^q_{\mathcal{H}}(\Omega, \mathcal{F}, \mathbb{P}) = \{X : (\Omega, \mathcal{F}, \mathbb{P}) \to \mathcal{H}; \ \mathbb{E}(\|X\|^q) < \infty\}$ be the set of random variables in $\mathcal{H}$ with finite moments of order q.

---

* Corresponding author.
  E-mail address: israel.martinezhernandez@kaust.edu.sa (I. Martínez-Hernández).

Then, the FAR(1) process is defined as a sequence of functional random variables $\{X_n(s); s \in D, -\infty < n < \infty\}$ in $\mathcal{H}$ such that

$$(X_n - \mu)(s) = \rho(X_{n-1} - \mu)(s) + W_n(s), \tag{1}$$

where ρ is an operator in $\mathcal{B}_\mathcal{H}$, $\mathbb{E}(X_n) = \mu$, and $\{W_n\}$ is functional white noise (Bosq, 2000). Without loss of generality, we assume that $\mu = 0$. If $X_n \in L_\mathcal{H}^2$ the covariance operator at lag h is defined as an element of $\mathcal{B}_\mathcal{H}$ by $\Gamma_h(z) = \mathbb{E}(\langle X_0, z \rangle X_h) = \mathbb{E}\{(X_0 \otimes X_h)z\}$; the corresponding empirical version is defined as $\hat{\Gamma}_h = \frac{1}{N-h}\sum_{i=1}^{N-h} X_i \otimes X_{i+h}$, where $\otimes$ denotes the tensor product.

Under this framework, integral operators form an important class. Let $\beta(t, s)$ be a fixed measurable function on $\mathbb{R}^2$. The integral operator ρ with kernel β is defined as $\rho(z)(s) = \int \beta(t, s)z(t)\,dt$. A crucial issue is to estimate $\rho(z)$ or $\beta(t, s)$. If model (1) is satisfied with stationarity conditions, then

$$\Gamma_1 = \rho\Gamma_0, \tag{2}$$

and if Γ_0 is invertible, then $\rho = \Gamma_1\Gamma_0^{-1}$. However, Γ_0^{-1} does not exist (Bosq, 2000). There is a significant amount of research that focuses on avoiding this problem in order to obtain the estimator of ρ under this approach. The method used most often is the principal component analysis (PCA) approach, where only the first k eigenfunctions corresponding to the largest eigenvalues are taken into account, and the inverse operator of Γ_0 is approximated by $\Gamma_0^{-1} = \sum_{j=1}^{k}\lambda_j^{-1}v_j \otimes v_j$. Cardot et al. (1999) proposed projecting data into a space spanned by finite eigenfunctions of Γ_0 to get a bounded Γ_0^{-1}. Bosq (2000) used the same idea to predict a FAR(1) model. In general, regularization methods can be applied to compute the inverse, i.e., $(\Gamma_0 + \alpha_n)^{-1}$ can be computed instead of Γ_0^{-1} (Crambes and Mas, 2013).

Another approach for estimating ρ is through $\beta(t, s)$ with a least squares (LS) estimator, which was first studied by Ramsay and Dalzell (1991), who proposed a penalized LS (PeLS) method. Similarly to the regression model case in $\mathbb{R}$, $\hat{\beta}(t, s)$ can be obtained by minimizing errors in the L_2-norm. As before, we have an ill-posed problem that requires regularization to obtain the estimator. Ramsay and Silverman (2005) discussed several approaches, including the assumption of finite basis function and regularization. Wu and Müller (2011) computed the estimator using a weighted LS approach. Ivanescu et al. (2015) extended penalized functional regression considering P-splines for the regularization, and Scheipl and Greven (2016) discussed identifiability issues.

In addition to the estimation problem discussed above, real datasets might face the issue of contamination with different types of outliers, such as isolated outliers, level shifts in mean, variance change, which affect the parameter estimation. Several probability models for time series outliers can be defined, such as *additive* outliers, *innovations* outliers, and *replacement* outliers (see Maronna et al., 2006, Chapter 8). The effect of the outliers is different depending on the type of outliers. For example, innovation outliers have rather minor effects compared to additive outliers because of leverage points. The isolated outliers are a special case of the additive outliers model and appear in many phenomena. They could be associated with data quality, i.e., a mistake made in an observation record. Let X_n be a functional time series. Then Y_n is an observation with isolated outlier at time n if $Y_n(s) = X_n(s) + v(\zeta, s)$, where $v(\zeta, s)$ represents the outlier and ζ is the contamination size. Notice that the effect is only in the observed time and does not affect the subsequent observations. In this paper, we suppose that the observed functional time series has isolated outliers.

Detecting outliers could be a challenge; therefore, many methods have been developed to detect them. A common approach used to detect functional outliers is based on functional depth. Sun and Genton (2011) defined the *functional boxplot*, which is based on the modified band depth (MBD) (López-Pintado and Romo, 2009) or other curve rankings. The functional boxplot detects shift outliers efficiently, even in the case of temporal dependence (Sun and Genton, 2012). Other proposals for the detection of functional outliers can be found in Hyndman and Shang (2010), Hubert et al. (2015) and Arribas-Gil and Romo (2015). Huang and Sun (2016) proposed a concept of total variation depth to detect shape and magnitude outliers, and combined this concept with the functional boxplot. Dai and Genton (2018a) proposed a notion of directional outlyingness as a centrality measure of functional data, which allows data to be ranked according to magnitude and shape outlyingness. In this work, we use the two-stage functional boxplot (Dai and Genton, 2018b) to detect outliers, that is, the functional boxplot combined with an outlier detection procedure based on the functional directional outlyingness proposed in Dai and Genton (2018a). We also use the corresponding depth to down-weight the outliers and obtain a robust estimator of the kernel surface, $\beta(t, s)$.

To illustrate the effect of contamination with isolated outliers on a dataset, we simulate $\{X_i(s); s \in [0, 1], i = 1, \ldots, N\}$, using the FAR(1) model with the kernel function $\beta(t, s) = c\exp\{-(t^2 + s^2)/2\}$, where c is such that $\|\rho\|_\mathcal{S} = \{\iint \beta^2(t, s)\,dt\,ds\}^{1/2} = 0.5$. We consider a modified set of five fixed observations by adding an outlier with Model 2 defined in Section 4.2. Then, we estimate $\beta(t, s)$ for each simulation using the PeLS estimator, $\hat{\beta}_{\text{PeLS}}$ (defined in Section 4.1), and compute the Integrated Squared Error (ISE) (defined in (14)) as a function of the perturbation level ζ. Fig. 1 depicts how the error increases as the perturbation level increases.

A robust estimator could be achieved in terms of the FAR(1) model by performing a robust estimation of the eigenfunctions, then using the PCA approach. Examples of robust estimators for eigenfunctions can be found in Locantore et al. (1999), Gervini (2008), Bali et al. (2011), Lee et al. (2013), Boente and Salibian-Barrera (2015) and Dürre et al. (2016). However, a robust estimation of the eigenfunctions changes the order of the eigenvalues, which would change the approximation of the covariance structure, due to the truncation procedure (Dürre et al., 2016). Another option is
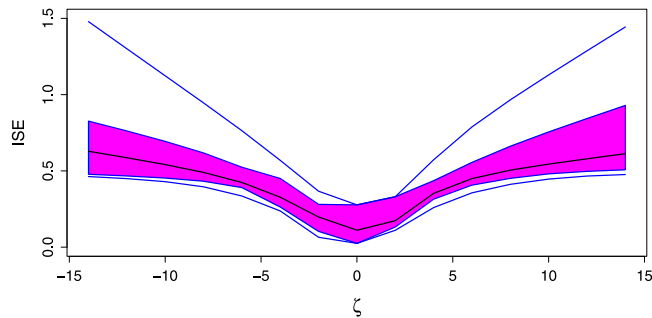
Fig. 1. Functional boxplot of ISE values as a function of perturbation level ζ contaminated with Model 2 (Section 4.2), on the estimated kernel surface $\beta(t, s)$.

the minimization approach, where we weight the function to be minimized in L_1-norm or L_2-norm. More generally, we could consider robust estimation methods such as M-estimation and S-estimation. Maronna and Yohai (2013) proposed a robust version of a spline-based estimate for a regression model with a scalar response, using an MM-estimate. Gervini (2012) proposed estimating β in a general framework of functional regression by expanding the data in terms of robust eigenfunctions and obtaining the estimator via M-estimation. In general, there are few works on robust parameter estimation in functional times series. Moreover, they tend to be difficult to apply to real datasets and computationally expensive for large datasets.

The aim of this paper is to propose a robust estimator of the kernel $\beta(t, s)$ under the presence of replacement outliers. Our method considers directional outlyingness as a centrality measure to define weights that are used in a PeLS estimation. Potential outliers in the data correspond to smaller weights. To identify the potential outliers, we use the two-stage functional boxplot, since this approach is robust to various types of outliers and takes into account the shape and the magnitude of the functional data.

The remainder of our paper is organized as follows. In Section 2, we define the weights based on the functional depth and the Depth-based Least Squares (DLS) estimator for the kernel $\beta(t, s)$, and theoretical properties of the proposed estimator are studied. In Section 3, we derive the theoretical influence function of the DLS estimator and we present a visualization tool for the empirical influential function. In Section 4, we compare the performance of our DLS estimator with those of the PeLS, PCA, and robust PCA estimators under different simulation settings. The performance is measured by the integrated squared error (ISE) and the empirical influence function (EIF). In Section 5, we analyze a dataset of hourly measurements of ambient CO_2 in California. Finally, in Section 6 we present some discussions and a direction for future work.

2. Robust estimation of the FAR(1) model

There are two main approaches to the study of functional data. The first approach assumes that data are collected without measurement errors. Under this assumption, each observed point can be used to make inferences about the process. However, if we have a dense array, then we could fall into a big data problem. The second approach assumes that each curve observed, $X_{obs}(s)$, is composed of a smooth underlying curve, $X(s)$, with measurement errors in each evaluation point, i.e., $X_{obs}(s) = X(s) + e(s)$. In this approach, inference is based on the underlying curve $X(s)$ and the assumption of some finite basis functions for $X(s)$; in other words, one selects a basis function and replaces the original data according to the best fit under this basis. We adopt the second approach.

2.1. Model setting

Let $\mathcal{B}_{\mathcal{H}}$ be the set of all bounded kernel operators defined as $\rho(z)(s) = \int_D \beta(t, s)z(t)\,dt$, with the norm operator $\|\rho\|_{\mathcal{S}} := \{\int_D \int_D \beta^2(t, s)\,dt\,ds\}^{1/2}$ and, without loss of generality, $D = [0, 1]$. Let X_n be a FAR(1) process with kernel $\beta(t, s)$ for the integral operator, that is,

$$X_n(s) = \int_D \beta(t, s)X_{n-1}(t)\,dt + W_n(s), \tag{3}$$

such that $\|\rho\|_{\mathcal{S}} < 1$. Let $\boldsymbol{\phi}(s) = \{\phi_1(s), \ldots, \phi_l(s)\}^T$ be a finite basis function defined on D. Assume that $X_n(s)$ can be approximated by $\hat{X}_n(s) = \mathbf{b}_n^T \boldsymbol{\phi}(s) = \sum_{i=1}^{l} b_{n,i}\phi_i(s)$ and assume that it is the true process. We use the notation X_n and $\hat{X}_n$ indistinctly.

Let $\mathbb{S}_1 = \text{span}\{v_1, \ldots, v_{k_1}\}$, $\mathbb{S}_2 = \text{span}\{\eta_1, \ldots, \eta_{k_2}\}$, where $\boldsymbol{v} = \{v_1, \ldots, v_{k_1}\}^T$ and $\boldsymbol{\eta} = \{\eta_1, \ldots, \eta_{k_2}\}^T$ are B-spline basis functions. The linear space spanned by $\{v_i(t)\eta_j(s)\}$ is denoted by $\mathbb{S}_1 \otimes \mathbb{S}_2$. We assume that $\beta(t, s) \in \mathbb{S}_1 \otimes \mathbb{S}_2$ has a tensor product surface $\beta(t, s) = \boldsymbol{v}^T(t)\Theta\boldsymbol{\eta}(s)$, with Θ being a $k_1 \times k_2$ coefficient matrix.

In addition, we suppose that the observed functional time series are generated by the model $Y_n = (1 - Z_n)X_n + Z_nV_n$, where Z_n is a Bernoulli process such that $\mathbb{P}(Z_n = 1) = 1 - \varepsilon, \varepsilon \geq 0, X_n(s)$ is a FAR(1) process, $V_n(\zeta, s)$ is an outlier-generating process with contamination size ζ such that $V_n(\zeta, s) = O(\zeta)$ and it is independent of Z_n.

As mentioned before, a regularization method can be applied to solve the inverse issue and the smoothness when estimating β. Therefore, we implement the PeLS method of using a tensor product (B-spline) surface to obtain our estimator. The inner product matrix of the basis is denoted by $\mathbf{J}_{\nu\eta} = \int \nu(s)\eta^T(s)\,ds$.

2.2. Penalized weighted optimization

We propose estimating $\beta(t, s)$ using a penalized, weighted LS estimator to reduce the impact of outliers. The weights are a function of the functional depth, since the use of depth is natural in this context. Depth reaches its maximum at the center of the data and decreases as it moves further away from the center. The curves near the center (or median curve) correspond to depth values approximately one and small values to extreme curves, hence, outlying curves are naturally down-weighted by the depth.

Let $d(X_i)$ be the depth of the curve X_i and w_i be the weight based on $(d(X_i), d(X_{i-1}))$. Then, we obtain the estimator from the solution of the following optimization problem:

$$\hat{\Theta} = \min_{\Theta} \left[\int_D \sum_{i=2}^N w_i \left\{ X_i(t) - \int_D X_{i-1}(s)\nu^T(s)\Theta\eta(t)\,ds \right\}^2 dt + P(\beta) \right], \tag{4}$$

in which $P(\beta)$ is a marginal quadratic penalty that can be expressed as

$$P(\beta) = \lambda_1 \text{vec}(\Theta)^T \mathbf{J}_{\eta\eta} \otimes \mathbf{P}_1 \text{vec}(\Theta) + \lambda_2 \text{vec}(\Theta)^T \mathbf{P}_2 \otimes \mathbf{J}_{\nu\nu}\text{vec}(\Theta), \tag{5}$$

where λ_1 and λ_2 are the tuning parameters and $\mathbf{P}_1$ and $\mathbf{P}_2$ are the fixed marginal penalty matrices for the first component and second component on the kernel surface's domain, respectively; see Wood (2006) for this tensor representation.

Let $\mathbf{B}^T$ be the coefficient matrix with rows $\mathbf{b}_j^T$, and let $\mathbf{B}_{-j}$ denote the matrix $\mathbf{B}$ without the column j. Let $\mathbf{W}$ be a $(N-1) \times (N-1)$ diagonal matrix with elements w_i for $i = 2, \ldots, N$. Using (2) and the empirical version of the covariance operator, $\hat{\Theta}$ must satisfy

$$\left(\mathbf{J}_{\eta\eta}^T \otimes \mathbf{J}_{\phi\nu}^T \mathbf{W}\mathbf{B}_{-N}^T \mathbf{J}_{\phi\nu} + \lambda_1 \mathbf{J}_{\eta\eta} \otimes \mathbf{P}_1 + \lambda_2 \mathbf{P}_2 \otimes \mathbf{J}_{\nu\nu}\right)\text{vec}(\hat{\Theta}) = \text{vec}\left(\mathbf{J}_{\phi\nu}^T \mathbf{B}_{-N}\mathbf{W}\mathbf{B}_{-1}^T \mathbf{J}_{\phi\eta}^T\right), \tag{6}$$

which is equivalent to $\hat{\rho}(\hat{\Gamma}_0^w + \varrho\tilde{P}) = \frac{N}{N-1}\hat{\Gamma}_1^w$, where $\hat{\Gamma}_h^w$ represents a weighted estimator of Γ_h; that is, $\hat{\Gamma}_h^w(z)(s) = \frac{1}{N-h}\sum_{i=1}^{N-h} w_{i+1}\langle X_i, z\rangle X_{i+h}(s)$. Thus, a robust estimator can be obtained from $\hat{\beta}(t, s) = \nu^T(t)\hat{\Theta}\eta(s)$, where $\hat{\Theta}$ is the solution of Eq. (6). We note that the estimator depends on the selection of the weight function and the selection of the type of depth.

To choose the weight function, we perform two steps. First, we identify the potential outliers using the two-stage functional boxplot; second, we down-weight the outliers using the corresponding functional depth defined from the directional Stahel–Donoho outlyingness.

The two-stage functional boxplot consists of the following procedure. For each s, let $\text{dSDO}(X_0(s), F_{X(s)}) = u(s) \cdot |X_0(s) - \tilde{X}(s)|/\text{MAD}\{X(s)\}$ be the directional Stahel–Donoho outlyingness (dSDO), where $u(s) = \{X_0(s) - \tilde{X}(s)\}/|X_0(s) - \tilde{X}(s)|, \tilde{X}$ is the median of X, and $\text{MAD}(X)$ is the Median Absolute Deviation of X. Let MO and VO be defined as

$$\text{MO}(X, F_X) = \int \text{dSDO}_X(s)\,ds \quad \text{and} \quad \text{VO}(X, F_X) = \int \|\text{dSDO}_X(s) - \text{MO}\|^2\,ds.$$

A curve X_i is considered as an outlier if its corresponding $(\text{MO}_i, \text{VO}_i)$ is an outlier with respect to the population distribution of (MO, VO), see Dai and Genton (2018a). Let S_1 be a subset of the indexes such that, for each $i \in S_1, X_i$ is an outlier. Then, we obtain new indexes of abnormal curves, S_2, applying the functional boxplot to the remaining non-outlying curves, $X_j, j \notin S_1$. Let $S = S_1 \cup S_2$ be the set of indexes of outlying curves.

We consider the functional directional outlyingness (FO) as $\text{FO}(X, F_X) = \int_D \text{dSDO}(X(s), F_{X(s)})^2\,ds$, and propose using the corresponding depth as follows:

$$\text{dO}(X, F_X) = \frac{1}{1 + \text{FO}(X, F_X)} \in [0, 1]. \tag{7}$$

Let dO_i be the depth of observation i, and let $c = \max\{\text{dO}_j; j \in S\}$. With this depth and the two-stage functional boxplot, we define the weights as follows.

Definition 1 (*Weights Based on* $\text{dO}(X, F_X)$). Let $\{X_i\}_{i=1}^N$ be a FAR(1) observed time series. Define the weight w_i as

$$w_i = W(d_i) = \mathbb{1}_{(c,1]}(d_i) + \text{dO}_i\text{dO}_{i-1}\mathbb{1}_{[0,c]}(d_i), \tag{8}$$

where $d_i = \min\{\text{dO}_i, \text{dO}_{i-1}\}$ and $\mathbb{1}_A(x)$ is the indicator function.

Remark 1. The parameter d_i indicates if X_i or X_{i-1} are detected as outliers. If $d_i \leq c$, then X_i and X_{i-1} are down-weighted with their corresponding depths values.

Observe that, if $V_n(\zeta, s) = O(\zeta)$ then $F(V_n, F_X)$ has the same order, $O(\zeta^2)$, than $V_n^2(\zeta, s)$ when $\zeta \to \infty$, thus the effect of the outlier in the optimization problem (4) is "removed" by the corresponding depth $dO(V_n, F_X)$.

Definition 2 (*Depth-based Least Squares (DLS) Estimator*)**.** Let $\{X_i\}_{i=1}^N$ be a functional time series such that the FAR(1) model holds. Then, the DLS estimator is $\hat{\beta}_{DLS}(t, s) = \boldsymbol{v}^T(t)\hat{\Theta}_{DLS}\boldsymbol{\eta}(s)$, where $\hat{\Theta}_{DLS}$ is the solution of the optimization problem (6) with weights w_i as in (8).

Algorithm 1 presents the steps followed to obtain the DLS estimator for the observed functional time series $\{X_i(s) ; s \in D, i = 1, \ldots, N\}$.

Algorithm 1.

---

1. For $i = 1, \ldots, N$, compute dO_i from (7).
2. Use the two-stage functional boxplot to obtain S and set $c = \max\{dO_j; X_j \text{ is an outlier}\}$.
3. Compute w_i and let $\hat{\Theta}$ be the solution to Eq. (6).

The next proposition shows the consistency of the corresponding operator $\hat{\rho}(z)(s) = \int \hat{\beta}_{DLS}(t, s)z(s)dt$. Its proof is in Appendix A. Let $\|z\|^2 = \langle z, z \rangle$ be the associated norm in $\mathcal{H}$.

Proposition 1. *Assume that $X_n(s) = \boldsymbol{b}_n^T \boldsymbol{\phi}(s)$ is a FAR(1) process, $\|X_0\| \leq M_1 < \infty$ a.s., and there exists a continuous differentiable function $\mathcal{W}$ on $(0, 1)$ that approximates W uniformly, i.e., for every $\epsilon > 0$ there exists $N_1 \in \mathbb{N}$ such that $|\mathcal{W}(d_i) - W(d_i)| < \epsilon$, $i \in \{1, \ldots, N\}$, for all $N > N_1$. Let $\lambda_1 \to 0$ and $\lambda_2 \to 0$ as $N \to \infty$. Then,*

$$\sup_{\|z\| \leq 1} \|\rho(z) - \hat{\rho}(z)\| = o_P(1),$$

if one of the following conditions holds:

(a) *The parameter c in Definition 1 is such that $c \to 0$ a.s. as $N \to \infty$, or*
(b) $\boldsymbol{b}_n$ *has a symmetric distribution about the origin* 0.

Remark 2. Condition (a) holds when the method to detect outliers is efficient, which is satisfied for the two-stage functional boxplot. Condition (b) holds, for example, when the functional data follow a Gaussian process.

The DLS estimator reduces the impact of the outliers on the estimation of the kernel surface $\beta(t, s)$, down-weighting the outliers similarly to the weighted LS estimator in the case of different variances in a univariate regression model. Indeed, $FO(X, F_X)$ can be decomposed as $FO(X, F_X) = MO(X, F_X)^2 + VO(X, F_X)$, where $MO(X, F_X)$ describes the position of X relative to the center curve; that is, $|MO(X, F_X)|$ represents the magnitude outlyingness (MO) and $VO(X, F_X)$ represents the shape outlyingness (VO), see Dai and Genton (2018a) for details.

3. Influence function of the DLS estimator

3.1. Theoretical influence function

To measure the robustness of an estimator, we use the influence function (IF) (Hampel et al., 1986). IF is a particular case of the Gâteaux derivative. We consider a random variable with distribution F on $\mathbb{R}$, where $x_0 \in \mathbb{R}$ and $T(F)$ is a functional defined on the space of distribution functions, $\mathcal{F} = \{F + \varepsilon(\delta_x - F)\}$ with $\varepsilon \in [0, 1]$. IF is defined as $IF(x_0; T, F) = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}[T\{(1 - \varepsilon)F + \varepsilon\delta_{x_0}\} - T(F)]$, where δ_x is the distribution that assigns mass one to x. This represents the rate of change in the functional T as a result of a small contamination by another distribution, $G = \delta_x$. However, in time series analysis, the form of contamination $(1 - \varepsilon)F + \varepsilon\delta_{x_0}$ is not realistic. Martin and Yohai (1986) introduced another definition of IF for a time series in which time series contamination is defined as $Y_n^\varepsilon = (1 - Z_n^\varepsilon)X_n + Z_n^\varepsilon V_n$, where X_n is a stationary process, Z_n is a Bernoulli process such that $\mathbb{P}(Z_n^\varepsilon = 1) = \varepsilon$, and V_n is an outlier-generating process. Then, the time series IF of T is defined as

$$IF(F_{X,Z,V}; T) = \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}\{T(F_Y^\varepsilon) - T(F_X)\}, \tag{9}$$

where F_Y^ε is the joint distribution of Y_n^ε, F_X is the joint distribution of X_n, and $F_{X,Z,V}$ is the joint distribution of X_n, Z_n^ε, and V_n. We note that (9) is valid in a general setting, even for random variables with values in the Hilbert space as defined in Section 1; that is, X_n and V_n are functional stochastic processes.

The next proposition shows that the IF corresponding to the DLS estimator is bounded. Let $T^{DLS} : \mathcal{F} \to \mathbb{S}_1 \otimes \mathbb{S}_2$ be the functional $T^{DLS}(F_X) = \beta_{DLS}$ where $\beta_{DLS}(t, s) = \boldsymbol{v}^T(t)\Theta_{DLS}\boldsymbol{\eta}(s)$. We note that the Θ_{DLS} determines the values of T^{DLS}, and, without confusion, we denote by $\text{vec}(T^{DLS})$ the corresponding $\text{vec}(\Theta_{DLS})$.
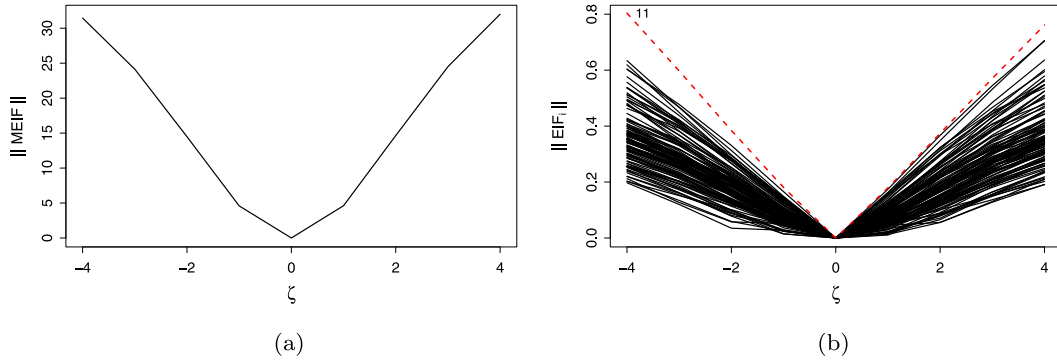
Fig. 2. Norm of mean empirical influence function (MEIF) (a) and functional hair-plot (b) for the $\hat{\beta}_{\text{PeLS}}$ kernel estimator with contamination Model 1. Observation 11 is indicated by the red dashed line.

Proposition 2. *Let $X_n(s) = \mathbf{b}_n^T \boldsymbol{\phi}(s)$ be a functional process such that the FAR(1) model holds and let $V_n(\zeta, s) = \mathbf{v}_n^T \boldsymbol{\phi}(s)$ be the outlier-generating process with $V_n(\zeta, s) = O(\zeta)$. The influence function of T^{DLS} is given by*

$$\text{vec}\{\text{IF}_n(F_{X,Z,V}; T^{\text{DLS}})\} = \left\{ \mathbb{E}_{F_X} \left(\mathbf{J}_{\eta\eta}^T \otimes \mathbf{J}_{\phi\nu}^T \mathbf{b}_n \mathbf{b}_n^T \mathbf{J}_{\phi\nu} + \lambda_1 \mathbf{J}_{\eta\eta} \otimes \mathbf{P}_1 + \lambda_2 \mathbf{P}_2 \otimes \mathbf{J}_{\nu\nu} \right)^{-1} \right\}$$
$$\times \left\{ \mathbb{E}_{F_{V_n, V_{n+1}}} (\gamma_n \mathbf{v}_n \mathbf{v}_{n+1}^T) - \mathbb{E}_{F_{V_n, V_{n+1}}} (\gamma_n \mathbf{v}_n \mathbf{v}_n^T) \text{vec}(\boldsymbol{\Theta}_{\text{DLS}}) \right\}, \tag{10}$$

and it is bounded, where $\gamma_n = \text{dO}(V_n, F_X) \text{dO}(X_{n+1}, F_X)$.

The proof is in Appendix A. We observe that $\text{vec}\{\text{IF}_n(F_{X,Z,V}; T^{\text{DLS}})\} = \text{IF}_n(F_{X,Z,V}; \text{vec}(T^{\text{DLS}}))$, so we conclude that Eq. (10) defines the IF of the T^{DLS}.

3.2. Empirical influence function

In practice, we use an empirical version of the IF (EIF) to quantify the influence of a contaminated observation i on the estimator $\hat{\beta}(t, s)$ of the kernel. Let $\mathbf{X} = \{X_i(s); \ s \in D, \ i = 1, \ldots, N\}$ be a FAR(1) time series of length N, and let $\hat{\beta}(\mathbf{X})$ be the kernel surface estimator using data $\mathbf{X}$. The EIF of $\hat{\beta}(t, s)$ from an observation i is defined as

$$\text{EIF}_i(\hat{\beta}, \mathbf{X}, \zeta) = N \left\{ \hat{\beta}(\mathbf{X}[i, \zeta]) - \hat{\beta}(\mathbf{X}) \right\}, \tag{11}$$

where $\mathbf{X}[i, \zeta]$ is equal to $\mathbf{X}$, except on the ith curve where it is replaced by the contaminated curve $V_i(\zeta, s)$, with contamination size ζ and $s \in D$; see Section 4.2 for contamination models.

If there is no temporal dependence, then the EIF is independent of the index i. In our case, we have temporal dependence. To measure the effect of contamination at all possible locations, we consider the mean EIF (MEIF):

$$\text{MEIF}_N(\hat{\beta}, \mathbf{X}, \zeta) = \sum_{i=1}^{N} \left\{ \hat{\beta}(\mathbf{X}[i, \zeta]) - \hat{\beta}(\mathbf{X}) \right\}. \tag{12}$$

To visualize the EIF we consider a hair-plot. A hair-plot for the scalar case was proposed by Genton and Ruiz-Gazen (2010) as a tool for visualizing influential observations in dependent data. To extend the hair-plot to the functional case, we plot the norm $\|\text{EIF}_i(\hat{\beta}, \mathbf{X}, \zeta)\|_{\mathcal{S}}$ versus ζ. We visualize the MEIF in a similar way, i.e., $\|\text{MEIF}(\hat{\beta}, \mathbf{X}, \zeta)\|_{\mathcal{S}}$ versus ζ. Fig. 2(a) shows the MEIF and Fig. 2(b) shows the functional hair-plot of $\hat{\beta}_{\text{PeLS}}(t, s)$ with $\beta(t, s) = c \exp\{-(t^2 + s^2)/2\}$, where c is such that $\|\rho\|_{\mathcal{S}} = 0.5$, and the model of contamination is $V_i(\zeta, s) = X_i(s) + \zeta$. From the MEIF, we see that if $|\zeta|$ increases, then the $\|\text{MEIF}\|_{\mathcal{S}}$ increases. From the hair-plot, we see that the $\|\text{EIF}_i\|_{\mathcal{S}}$ is different depending on the location; in this case, contamination in observation 11 results in a higher influential observation. A more exhaustive simulation study is performed in Section 4.4 to compare the IF of $\hat{\beta}_{\text{DLS}}(t, s)$ and the IF of $\hat{\beta}_{\text{PeLS}}(t, s)$ defined in Section 4.1.

4. Monte Carlo simulation study

4.1. Competing estimators

In this section, we investigate the performance of the DLS estimator (defined in Section 2). We consider different kernels and vary the contamination level with different types of outliers. We use the R package *fda* with the proposed weights to compare the DLS estimator with the following estimators:

1. PeLS estimator: $\hat{\beta}_{\text{PeLS}}(t,s) = \boldsymbol{v}^T(s)\boldsymbol{\Theta}\hat{}_{\text{PeLS}}\boldsymbol{\eta}(t)$, where $\hat{\boldsymbol{\Theta}}_{\text{PeLS}}$ is solution to (4) and $w_i = 1$ for all i.
2. PCA estimator:

$$\hat{\beta}_{\text{PCA}}(t,s) = \frac{1}{N-1} \sum_{l=1}^{N-1} \sum_{j=1}^{k} \sum_{i=1}^{k} \hat{\lambda}_j^{-1} \langle X_l, \hat{v}_j \rangle \langle X_{l+1}, \hat{v}_i \rangle \hat{v}_j(s) \hat{v}_i(t),$$

   where $\hat{v}_i(s)$ denotes the eigenfunctions of $\hat{\Gamma}_0$.
3. RPCA estimator: Similar to PCA estimator, but replaces $\hat{v}_i(s)$ with $\hat{v}_i^R(s)$, where $\hat{v}_i^R(s)$ denotes the robust eigenfunctions proposed by Hyndman and Ullah (2007).

4.2. Simulation setting

We simulate $\{X_i(s); \ s \in [0,1], \ i = 1, \ldots, N\}$ by using the FAR(1) model from Eq. (3) with two different types of kernel functions ($\beta(t,s)$) and two types of functional white noise, approximated on a fixed grid. The considered kernel functions are (a) the Gaussian kernel, $\beta_1(t,s) = c \ \exp\{-(t^2+s^2)/2\}$ and (b) the linear kernel, $\beta_2(t,s) = c \ t$, where c, in both cases, is such that $\|\rho\|_S = 0.5$. The error terms considered are (a) the Brownian motion $W(s), s \in [0,1]$, and (b) the Brownian bridge, $\tilde{W}(s) = W(s) - sW(1), s \in [0,1]$.

We consider the sample sizes $N = 200$ and $N = 300$, at $m = 100$ points equispaced on $[0,1]$, and $M = 1000$ replicates. We fit a cubic B-spline basis for each simulated curve for a total of 15 basis functions, i.e, $X_i(s) = \sum_{r=1}^{15} b_{r,i}\phi_r(s)$. Similarly, we take 15 basis functions for $\boldsymbol{v}$ and $\boldsymbol{\eta}$ on the tensor representation of $\beta(t,s)$. The considered penalty functions in (5) are as follows:

$$\mathbf{P}_1 = \int \left\{\frac{\partial^2}{\partial s^2}\boldsymbol{v}(s)\right\} \left\{\frac{\partial^2}{\partial s^2}\boldsymbol{v}^T(s)\right\} \ \mathrm{d}s \ \text{ and } \ \mathbf{P}_2 = \int \left\{\frac{\partial^2}{\partial s^2}\boldsymbol{\eta}(s)\right\} \left\{\frac{\partial^2}{\partial s^2}\boldsymbol{\eta}^T(s)\right\} \ \mathrm{d}s. \tag{13}$$

We determine the tuning parameters λ_1 and λ_2 through cross-validation. The same method is used to determine the parameter k on PCA and RPCA estimators.

We consider five contamination models to produce outliers into the model $Y_n^\varepsilon = (1-Z_n^\varepsilon)X_n + Z_n^\varepsilon V_n$, where Z_n^ε is a Bernoulli process such that $\mathbb{P}(Z_n^\varepsilon = 1) = \varepsilon$ (see Sun and Genton, 2011 for the first four models):

1. Model 1 (no contamination): $V_n(\zeta, s) = X_n(s)$;
2. Model 2 (shifted outlier): $V_n(\zeta, s) = X_n(s) + \sigma \zeta$, where ζ is the contamination size and σ has a Bernoulli distribution on $\{-1, 1\}$ with probability $1/2$;
3. Model 3 (partially contaminated): $V_n(\zeta, s) = X_n(s) + \sigma \zeta 1_{[U,1]}(s)$, with the same parameters as Model 2, where U has a uniform distribution on $[0, 1]$;
4. Model 4 (contaminated by peaks): $V_n(\zeta, s) = X_n(s) + \sigma \zeta 1_{[U,U+l]}(s)$, with the same parameters as Model 2, where U has a uniform distribution on $[0, 1-l]$; and
5. Model 5 (contaminated with functional white noise): $V_n(\zeta, s) = W_n(s)$, where W_n is functional white noise.

We fix $\zeta = 1.5$ and $l = 1/4$, and we vary the contamination levels (ε), $\varepsilon = 0.05$, $\varepsilon = 0.10$, and $\varepsilon = 0.15$. Fig. 3 shows a simulation set from each of these models with the Gaussian kernel and the Brownian bridge. To evaluate the performance, we compute ISE as:

$$\text{ISE} = \sqrt{\int_D \int_D \left\{\beta(t,s) - \hat{\beta}(t,s)\right\}^2 \mathrm{d}t \mathrm{d}s}. \tag{14}$$

4.3. Simulation results

We only present the Gaussian kernel results with white noise $\tilde{W}$, since the results from the linear kernel and W are similar (available in the online supplement). Table 1 presents a complete summary of the mean of ISE (MISE) values from all of the simulation settings. The presence of outliers in the data increases the ISE values, and the magnitude varies depending on the type of outliers, except for the RPCA estimator. The non-robust estimators, PeLS and PCA, produce larger ISE values than the others, especially for Model 2. The RPCA estimator presents constant ISE values in all simulation settings, indicating that it is robust, but its performance is still not satisfactory. It is worth mentioning that the RPCA outperforms the PeLS and PCA estimators when ζ increases (see the online supplement). In contrast, the ISE values obtained by applying our DLS method do not significantly increase in the presence of outliers and are smaller than the corresponding RPCA estimator. This result is consistent across all simulation settings, even when varying N and the error term.

The DLS estimator outperforms the others estimators in all cases, having the smallest MISE values. Moreover, we can notice that when there is no contamination, the DLS estimator is as accurate as the PeLS estimator.

The norm of the corresponding operator (ρ) defined by the kernel $\beta(t,s)$ must be estimated more accurately, since this value indicates the grade of persistence or dependency on the functional time series. Under the stationarity assumptions of
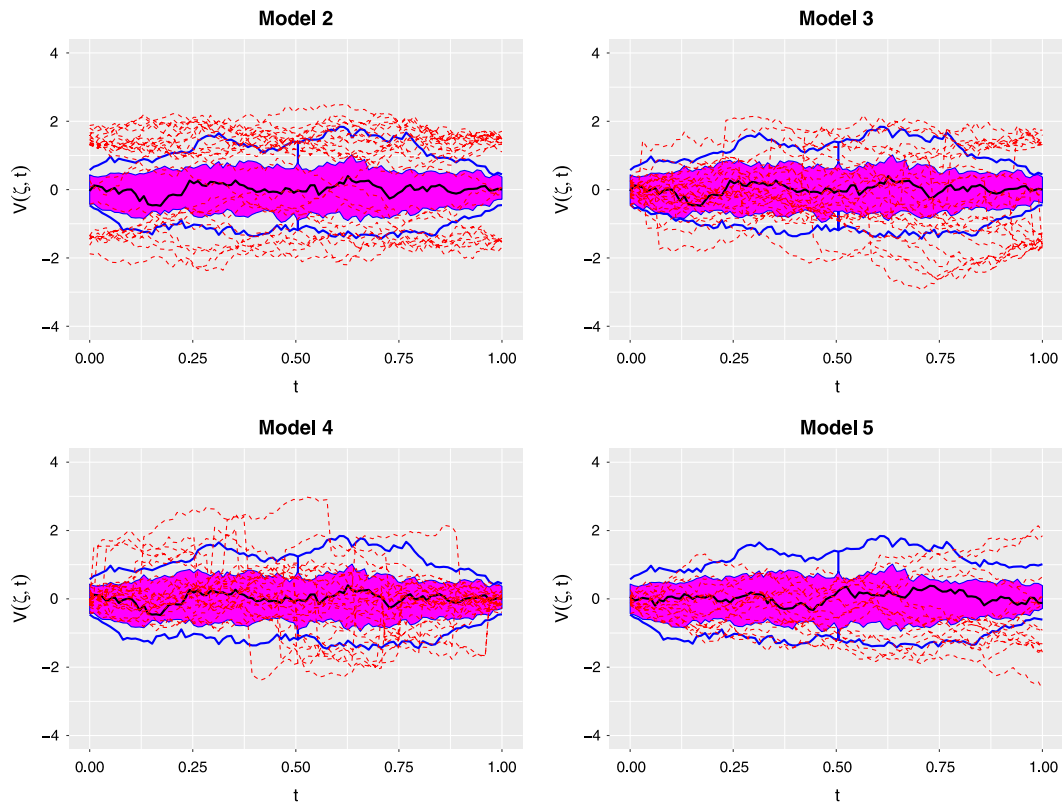
Fig. 3. Simulated samples from Models 2, 3, 4 and 5. The red curves are outliers detected by the two-stage functional boxplot. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the FAR(1) model (Bosq, 2000), the norm of the operator ρ should be less than one, i.e., $\|\rho\|_{\mathcal{S}} < 1$. In our simulation, we set $\|\rho\|_{\mathcal{S}} = 0.5$.

When contaminated by outliers, these norm values are affected. Similar to the ISE value results, the non-robust estimators, PeLS and PCA, are the most affected, especially by the shifted outlier in Model 2, producing norms that are much too large. In contrast, the norm values obtained from the DLS estimator are more accurate than those of the other estimators in all of the simulation settings. Boxplot results can be found in the supplementary online material.

4.4. Mean empirical influence function

We calculate the Mean Empirical Influence Function (MEIF) of the DLS estimator to determine its robustness. We also present the MEIF for the PeLS estimator for comparison. We use the MEIF to explore the influence of contamination at all possible locations, instead of the EIF, which depends on the individual observation index i. To illustrate the performance in terms of the MEIF, we simulate realizations from the FAR(1) process with the Gaussian kernel, $\beta(t, s) = c \exp\{-(t^2 + s^2)/2\}$, where c is such that $\|\rho\|_{\mathcal{S}} = 0.5$ with sample size $N = 200$. For each simulation, $\{X_i(s)\}_{i=1}^{N}$, we compute the MEIF from Eq. (12) with different models of contamination; that is, $V_n(\zeta, s) = $ Model 2, Model 3, and Model 4. We vary ζ from -14 to 14 and perform a Monte Carlo study with 100 replicates.

Fig. 4 shows the Hilbert–Schmidt norm of each MEIF ($\|\text{MEIF}\|_{\mathcal{S}}$) as a function of ζ for each simulation. We observe that the $\|\text{MEIF}\|_{\mathcal{S}}$ corresponding to the DLS estimator remains constant as $|\zeta|$ increases. On the other hand, the $\|\text{MEIF}\|_{\mathcal{S}}$ corresponding to the PeLS estimator increases as $|\zeta|$ increases.

In real applications, accurately estimating the kernel surface ($\beta(t, s)$) and the norm of the operator ρ is important. Indeed, to guarantee the stationarity assumption, we must have a norm less than one. Our method estimates $\|\rho\|_{\mathcal{S}}$ more accurately than the other estimators in the presence of outliers. In conclusion, the DLS estimator is a robust method to fit the FAR(1) model.

5. Data application

We consider daily curves of CO_2 measurements (ppm units) from the California Air Quality Data published by the California Air Resources Board. From Fig. 5, we expect some days with extreme measurements that affect the statistical

Table 1
MISE values from Monte Carlo simulation study with the Gaussian kernel ($\beta(t, s)$) and functional white noise $\tilde{W}$ (Brownian bridge). Models 1, 2, 3, 4, and 5 indicate the different types of contamination (defined in Section 4.2); the contamination level varies from 0.05 to 0.15 and sample sizes are $N = 200$ and $N = 300$. Bold font is used to highlight the best performance.

$N = 200$	Model 1	Model 2	Model 3	Model 4	Model 5
$\varepsilon = 0.05$					
PeLS	**0.162**	0.350	0.273	**0.166**	0.220
DLS	0.173	**0.186**	**0.175**	0.176	**0.184**
PCA	0.241	0.275	0.272	0.251	0.257
RPCA	0.415	0.415	0.415	0.415	0.415
$\varepsilon = 0.10$					
PeLS	**0.162**	0.452	0.337	**0.172**	0.276
DLS	0.173	**0.197**	**0.182**	0.177	**0.208**
PCA	0.241	0.342	0.314	0.261	0.285
RPCA	0.415	0.415	0.415	0.415	0.415
$\varepsilon = 0.15$					
PeLS	**0.162**	0.510	0.358	0.181	0.320
DLS	0.173	**0.223**	**0.191**	**0.180**	**0.231**
PCA	0.241	0.379	0.347	0.271	0.320
RPCA	0.415	0.415	0.415	0.415	0.415
$N = 300$	Model 1	Model 2	Model 3	Model 4	Model 5
$\varepsilon = 0.05$					
PeLS	**0.135**	0.323	0.260	**0.138**	0.200
DLS	0.144	**0.149**	**0.146**	0.142	**0.155**
PCA	0.237	0.272	0.268	0.246	0.252
RPCA	0.412	0.412	0.412	0.412	0.412
$\varepsilon = 0.10$					
PeLS	**0.135**	0.417	0.323	**0.146**	0.261
DLS	0.144	**0.163**	**0.151**	0.146	**0.178**
PCA	0.237	0.341	0.311	0.255	0.283
RPCA	0.412	0.412	0.412	0.412	0.412
$\varepsilon = 0.15$					
PeLS	**0.135**	0.470	0.350	0.157	0.300
DLS	0.144	**0.190**	**0.168**	**0.146**	**0.207**
PCA	0.237	0.380	0.348	0.267	0.316
RPCA	0.412	0.412	0.412	0.412	0.412

analysis. As in Section 4, we compare the PeLS, DLS, PCA, and RPCA estimations of the kernel surface ($\beta(t, s)$). We analyze hourly measurements of the CO_2 concentration (ppm units) in the ambient air in California, during the period of October 20, 2009, to February 14, 2011, for a total of 11,592 observations. These data can be downloaded from https://www.arb.ca.gov/aqd/aqdcd/aqdcddld.htm. Here we focus on 483 daily trajectories of the data, where each curve has 24 observations (hourly data); see Fig. 5. Due to the structure of the data, there is a natural dependence between the CO_2 concentration during the final hours of a day and the CO_2 concentration during the early hours of the next day. Also, there is a dependence between the CO_2 concentration on a given day at hour h_1 and the CO_2 concentration on the next day at the same hour h_1. We expect our model to capture both dependencies. There are 44 missing observations. Since the goal is to describe the underlying dynamic changes of a continuous curve, we replace missing observations with the average of their two closest neighbors.

In this analysis, our goal is to understand the dependency of the CO_2 daily concentration on the previous profile. We explore how the data analysis can be affected by the presence of outliers. The *p*-value of the test from Horváth et al. (2014) for log CO_2 data is 0.5, hence we conclude that the functional time series is stationary. The following dates are identified as outliers by the two-stage functional boxplot:

11/16/2009	11/25/2009	12/21/2009	12/23/2009	1/11/2010	1/12/2010
1/23/2010	2/14/2010	11/5/2010	11/11/2010	11/22/2010	11/23/2010
11/25/2010	11/27/2010	11/29/2010	12/5/2010	12/6/2010	12/7/2010
12/23/2010	12/24/2010	1/15/2011	1/24/2011	1/25/2011	1/26/2011
1/27/2011	1/28/2011.				

There are some curves that show fluctuation throughout the day, for example from 2/6/2011 to 2/13/2011, see Fig. 6(b). However, from the functional boxplot in Fig. 5, we know that this behavior is not the most common in the data.

Fig. 7 shows the estimated kernel surface for the CO_2 log curves. Fig. 7(b) represents the PeLS estimator. Despite the fact that the stationarity test indicates that the data are stationary, the norm of the operator with $\beta_{\text{PeLS}}(t, s)$ is greater than one: $\|\hat{\rho}_{\beta_{\text{PeLS}}}\|_{\mathcal{S}} \approx 2.43$. This means that the estimated FAR(1) model is non-stationary. Fig. 7(a) corresponds to the robust DLS estimator, which has a norm close to 0.39, which agrees with the test. Following Section 4, we can say that this value is more
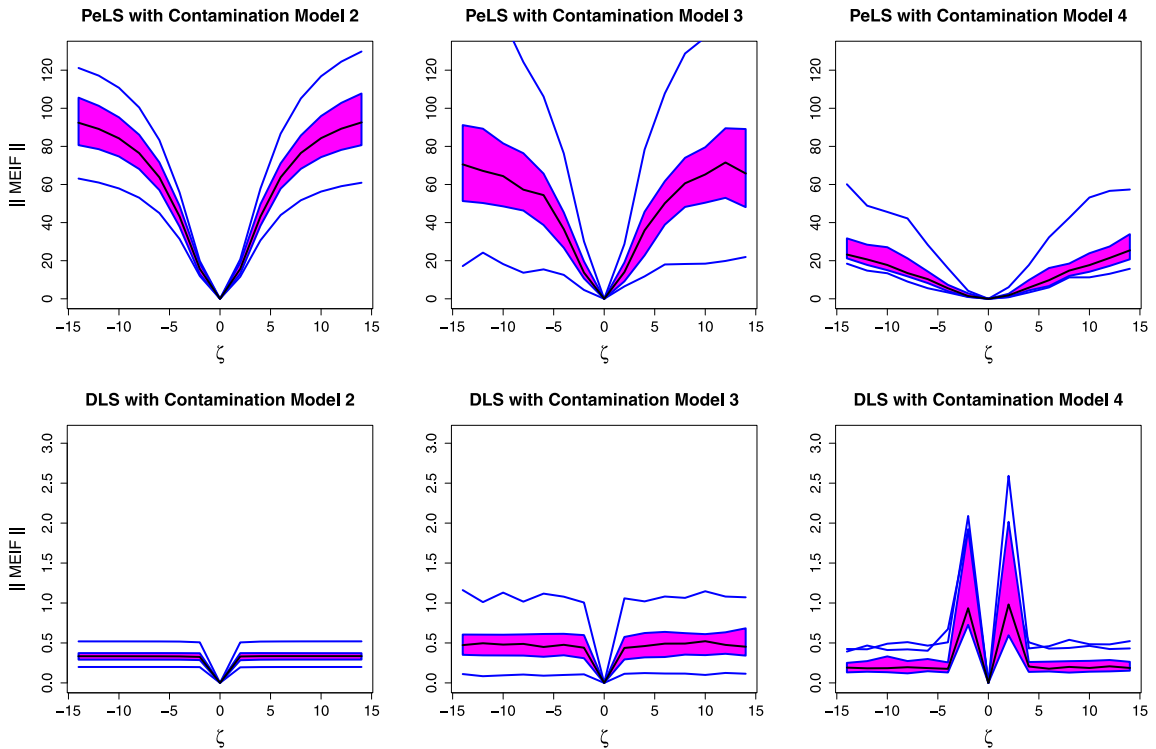
Fig. 4. Functional boxplot of the norm $\|MEIF\|_{\mathcal{S}}$ for each simulation of $\hat{\beta}_{PeLS}(t,s)$ (top) and $\hat{\beta}_{DLS}(t,s)$ (bottom). In this case $N = 200$, $\beta(t,s)$ is the Gaussian kernel, and W_n is the Brownian bridge.
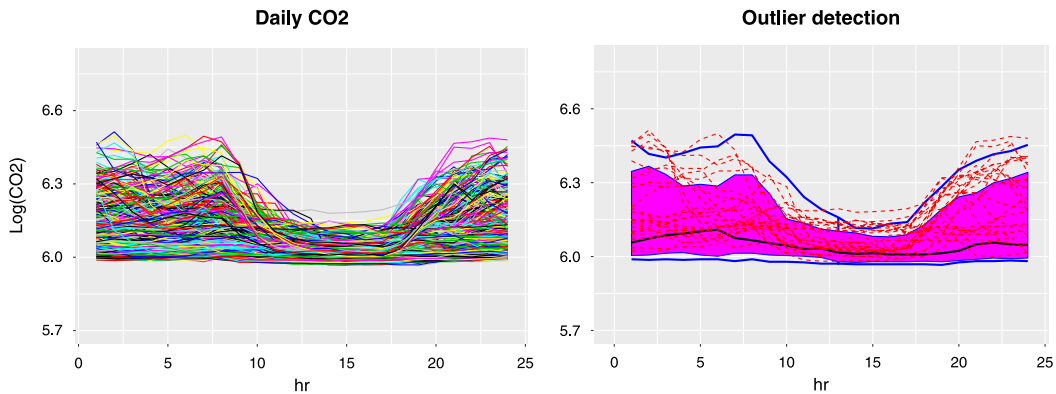

Fig. 5. Daily trajectories of the log of CO_2 data during the period of October 20, 2009, to February 14, 2011 (left), and two-stage functional boxplot of the log of CO_2 data (right) showing the outliers detected.

accurate. The operator norm estimated for PCA is ≈ 1.44 and for RPCA it is ≈ 0.01. These represent an example where the use of a non-robust method can affect the stationarity assumption.

The MISE values are 0.06, 0.15, 0.06, and 0.04 for the DLS, PeLS, RPCA, and PCA estimators, respectively. Although the MISE values of PCA are smaller than the others, this estimator corresponds to a non-stationary process. Therefore, we prefer to use the DLS or RPCA estimators. Regarding the shape of the estimated surface, the PCA and RPCA estimators produce similar shapes but on different scales. We observe some of the daily profiles that peak before decreasing to a plateau dominate the principal component. The PeLS estimator, $\hat{\beta}_{PeLS}(t,s)$, depends only on variable t, which means that the predicted response is a constant function. In addition, the PeLS estimator gives more weight to the last hour of the day, suggesting that the entire daily curve could be completely determined by that one hour. In contrast, the DLS estimator captures the intraday and intraperiod dependencies more accurately by giving more weight to final hours of the previous day, but also by giving more weight to data from the same hours of the day across different days. Considering the accurate estimation of the norm
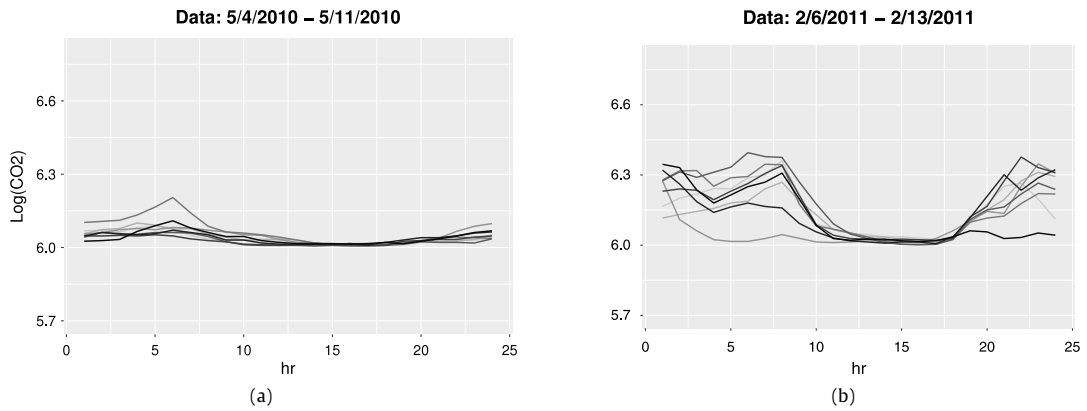
Data: 5/4/2010 − 5/11/2010

Data: 2/6/2011 − 2/13/2011

Fig. 6. Seven days of consecutive curves of the log of CO_2 data from 5/4/2010 to 5/11/2010 (a), and seven days of consecutive curves from 2/6/2011 to 2/13/2011 (b).
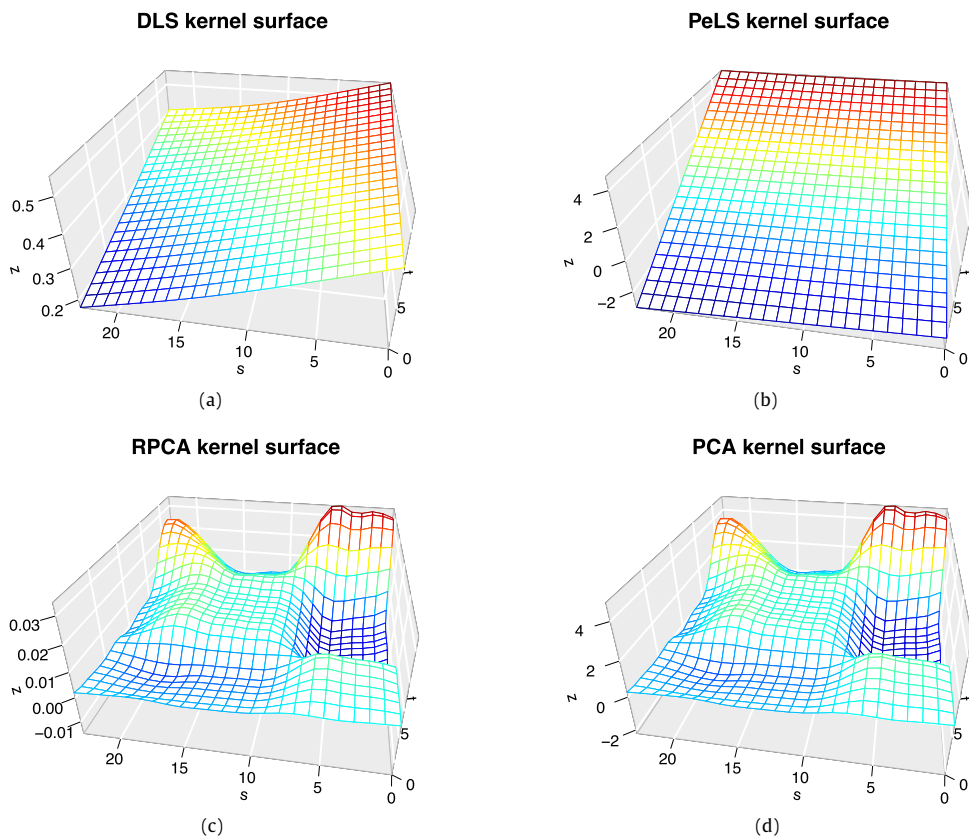
DLS kernel surface

PeLS kernel surface

RPCA kernel surface

PCA kernel surface

Fig. 7. Kernel surface estimator for functional time series of the log of ambient CO_2 concentrations in California. (a) DLS estimator, (b) PeLs estimator, (c) RCPA estimator with eigenfunctions $k = 3$, and (d) PCA estimator with eigenfunctions $k = 3$.

produced by the DLS estimator and the reasonable interpretability of the kernel surface, we conclude that the DLS estimator is superior.

6. Discussion

In this paper, we proposed the DSL estimator, a depth-based robust estimator of the kernel surface $\beta(t, s)$ in the FAR(1) model, using penalized, weighted least squares. We down-weighted possible outliers by defining the weights as a function

of outlyingness. To detect outliers, we used the two-stage functional boxplot. Proposition 2 showed that the influence function of the proposed estimator is bounded for replacement outliers, which is an attractive property. We also presented a visualization tool similar to the hair-plot proposed by Genton and Ruiz-Gazen (2010), to visualize influential observations for functional time series. In the Monte Carlo simulation study we demonstrated that the DSL estimator performs well with different simulation settings. Moreover, we demonstrated that, with the proposed robust estimator, the operator norm does not significantly change in the presence of outliers, a desirable property. Other notions of depths with the property $d(V_n, F_X) = O(1/\zeta^2)$ could be used in this framework, but we prefer to use the directional Stahel–Donoho outlyingness, because it performs well and provides a natural extension to the multivariate case for future work, including an extension to the FAR(p) model.

Acknowledgments

Appendix A

Proof of Proposition 1. Assume condition (a) holds. Denote by $\| \cdot \|_{\mathcal{L}}$ the uniform norm of operators, and observe that $\| \cdot \|_{\mathcal{L}} \leq \| \cdot \|_{\mathcal{S}}$. Let $\hat{\Gamma}_0^w$ be the corresponding depth-based estimator of Γ_0, i.e.,

$$\hat{\Gamma}_0^w(z)(t) = \frac{\sum_{i=1}^N w_i X_i \otimes X_i(z)(t)}{\sum_{i=1}^N w_i} = \frac{\mathbf{z}^T \mathbf{J}_{\phi\phi} \sum_{i=1}^N w_i \mathbf{b}_i \mathbf{b}_i^T \phi(t)}{\sum_{i=1}^N w_i},$$

for $z(t) = \mathbf{z}^T \phi(t)$. Let $\hat{\mathcal{W}}(\cdot)$ be the evaluation of $\mathcal{W}$ on the empirical d_i. Then,

$$\hat{\Gamma}_0^w = \frac{\sum_{i=1}^N w_i X_i \otimes X_i}{\sum_{i=1}^N w_i} = \frac{\sum_{i=1}^N \hat{\mathcal{W}}(d_i) X_i \otimes X_i + \sum_{i=1}^N \{w_i - \hat{\mathcal{W}}(d_i)\} X_i \otimes X_i}{\sum_{i=1}^N \hat{\mathcal{W}}(d_i) + \sum_{i=1}^N \{w_i - \hat{\mathcal{W}}(d_i)\}}.$$

By the stationarity assumption of X_t, we claim that

$$\frac{\sum_{i=1}^N \hat{\mathcal{W}}(d_i) X_i \otimes X_i(z)}{\sum_{i=1}^N \hat{\mathcal{W}}(d_i)} \xrightarrow[N\to\infty]{a.s.} \Gamma_0^w(z) := \frac{\mathbb{E}\{\mathcal{W}(d_0)\langle X_0, z\rangle X_0\}}{\mathbb{E}\{\mathcal{W}(d_0)\}},$$

$$\frac{\sum_{i=1}^N \{w_i - \hat{\mathcal{W}}(d_i)\} X_i \otimes X_i}{\sum_{i=1}^N \hat{\mathcal{W}}(d_i) + \sum_{i=1}^N \{w_i - \hat{\mathcal{W}}(d_i)\}} \xrightarrow[N\to\infty]{a.s.} \frac{\mathbb{E}[\{W(d_0) - \mathcal{W}(d_0)\} X_0 \otimes X_0]}{\mathbb{E}\{\mathcal{W}(d_0)\} + \mathbb{E}[\{W(d_0) - \mathcal{W}(d_0)\}]},$$

$$\left\| \frac{\mathbb{E}[\{W(d_0) - \mathcal{W}(d_0)\} X_0 \otimes X_0]}{\mathbb{E}\{\mathcal{W}(d_0)\} + \mathbb{E}[\{W(d_0) - \mathcal{W}(d_0)\}]} \right\|_{\mathcal{L}} \leq \frac{\epsilon M_1^2}{c_1} \xrightarrow[\epsilon\to 0]{a.s.} 0,$$

and $N^{-1}\sum_{i=1}^N \{w_i - \hat{\mathcal{W}}(d_i)\} \xrightarrow[N\to\infty]{} 0$. Using the above, it follows that $\|\hat{\Gamma}_0^w - \Gamma_0^w\|_{\mathcal{L}} \to 0$, and using the assumption $c \to 0$ as $N \to \infty$, we obtain that $\mathcal{W} \equiv 1$, and $\|\Gamma_0^w - \Gamma_0\|_{\mathcal{L}} \to 0$. Therefore $\|\hat{\Gamma}_0^w - \Gamma_0\|_{\mathcal{L}} \leq \|\hat{\Gamma}_0^w - \Gamma_0^w\|_{\mathcal{L}} + \|\Gamma_0^w - \Gamma_0\|_{\mathcal{L}} \to 0$. Adding the penalization term $\hat{\Gamma}_0^w + \varrho\tilde{P}$ we obtain a similar result,

$$\|\hat{\Gamma}_0^w + \varrho\tilde{P} - \Gamma_0\|_{\mathcal{L}} \to 0 \tag{15}$$

if $\varrho \to 0$ as $N \to \infty$. Using the same arguments in the covariance operator at lag one, $\hat{\Gamma}_1^w$, we obtain

$$\|\hat{\Gamma}_1^w - \Gamma_1\|_{\mathcal{L}} \to 0. \tag{16}$$

We are now in a position to show the consistency. Following Bosq (2000), and the fact that $\|\Gamma_0^{-1}\|$ is bounded, we have

$$\hat{\rho} - \rho = (\hat{\Gamma}_1^w - \Gamma_1)(\hat{\Gamma}_0^w + \varrho\tilde{P})^{-1} + \Gamma_1(\hat{\Gamma}_0^w + \varrho\tilde{P})^{-1}\{\Gamma_0 - (\hat{\Gamma}_0^w + \varrho\tilde{P})\}\Gamma_0^{-1},$$

consequently, $\|\hat{\rho} - \rho\|_{\mathcal{L}} \leq \|(\hat{\Gamma}_0^w + \varrho\tilde{P})^{-1}\|_{\mathcal{L}}\{\|\hat{\Gamma}_1^w - \Gamma_1\|_{\mathcal{L}} + \|\Gamma_1\|_{\mathcal{L}}\|\Gamma_0^{-1}\|_{\mathcal{L}}\|\hat{\Gamma}_0^w + \varrho\tilde{P} - \Gamma_0\|_{\mathcal{L}}\}$, and by using (15) and (16), the proof is complete.

Assume condition (b) holds. From Zuo and Cui (2005), it follows that

$$\hat{\sigma}_h^w(t, s) = \phi(t)^T \frac{\sum_{i=1}^N w_i \mathbf{b}_i \mathbf{b}_{i+h}^T}{\sum_{i=1}^N w_i} \phi(s),$$

is a consistent estimator of $\kappa \sigma_h(t, s)$. Since $\hat{\sigma}_h^w(t, s)$ and $\sigma_h(t, s)$ are the kernel of the covariance operator $\hat{\Gamma}_h^w$ and Γ_h, respectively, it follows that $\hat{\Gamma}_h^w$ is a consistent estimator of $\kappa \Gamma_h$, $h = 0, 1$. By similar arguments to the case (a), we have

$$\hat{\rho} - \rho = (\hat{\Gamma}_1^w - \kappa \Gamma_1)(\hat{\Gamma}_0^w + \varrho \tilde{P})^{-1} + \kappa \Gamma_1(\hat{\Gamma}_0^w + \varrho \tilde{P})^{-1}\{\kappa \Gamma_0 - (\hat{\Gamma}_0^w + \varrho \tilde{P})\}(\kappa \Gamma_0)^{-1}.$$

Therefore, $\|\hat{\rho} - \rho\|_{\mathcal{L}} \leq \|(\hat{\Gamma}_0^w + \varrho \tilde{P})^{-1}\|_{\mathcal{L}}\{\|\hat{\Gamma}_1^w - \kappa \Gamma_1\|_{\mathcal{L}} + \|\Gamma_1\|_{\mathcal{L}}\|\Gamma_0^{-1}\|_{\mathcal{L}}\|\hat{\Gamma}_0^w + \varrho \tilde{P} - \kappa \Gamma_0\|_{\mathcal{L}}\}$, and by using the consistency of $\hat{\Gamma}_h^w$ for $\kappa \Gamma_h$, the proof is complete. $\square$

Proof of Proposition 2. We represent $X_n(s)$ and $V_n(\zeta, s)$ with a finite basis function; $X_n(s) = \mathbf{x}_n^T \boldsymbol{\phi}(s)$ and for each ζ, $V_n(\zeta, s) = \mathbf{v}_n^T \boldsymbol{\phi}(s)$. Without confusion, we write Y instead of Y^ε. Let F_X and F_Y be the joint distributions of X_n and Y_n, respectively. Let $\text{vec}(\boldsymbol{\Theta}_{\text{DLS}})$ be such that $\beta_{\text{DLS}}(t, s) = \mathbf{v}^T(t)\boldsymbol{\Theta}_{\text{DLS}}\boldsymbol{\eta}(s)$. Define $\mathcal{X}_{n,n}, \mathcal{V}_{n,n}, \mathcal{X}_{n,n+1}, \mathcal{V}_{n,n+1}$, and $\mathcal{P}$ as

$$\mathcal{X}_{n,n} := \mathbf{J}_{\eta\eta}^T \otimes \mathbf{J}_{\phi v}^T \mathbf{x}_n \mathbf{x}_n^T \mathbf{J}_{\phi v},$$
$$\mathcal{V}_{n,n} := \mathbf{J}_{\eta\eta}^T \otimes \mathbf{J}_{\phi v}^T \mathbf{v}_n \mathbf{v}_n^T \mathbf{J}_{\phi v},$$
$$\mathcal{X}_{n,n+1} := \text{vec}\left(\mathbf{J}_{\phi v}^T \mathbf{x}_n \mathbf{x}_{n+1}^T \mathbf{J}_{\psi\eta}^T\right),$$
$$\mathcal{V}_{n,n+1} := \text{vec}\left(\mathbf{J}_{\phi v}^T \mathbf{v}_n \mathbf{v}_{n+1}^T \mathbf{J}_{\psi\eta}^T\right),$$
$$\mathcal{P} := \lambda_1 \mathbf{J}_{\eta\eta} \otimes \mathbf{P}_1 + \lambda_2 \mathbf{P}_2 \otimes \mathbf{J}_{vv}.$$

We know that $\text{vec}(\boldsymbol{\Theta}_{\text{DLS}})$ satisfies

$$\mathbb{E}_{F_X}(\mathcal{X}_{n,n} + \mathcal{P})\text{vec}(\boldsymbol{\Theta}_{\text{DLS}}) = \mathbb{E}_{F_X}(\mathcal{X}_{n,n+1}).$$

If we let $F_{Y_{n,n+1}}$ be the joint distribution of the bivariate process (Y_n, Y_{n+1}), then the bivariate contamination distribution is $F_{Y_{n,n+1}} = (1 - \varepsilon)F_{X_{n,n+1}} + \varepsilon F_{V_{n,n+1}}$. Denote by $\text{vec}(\boldsymbol{\Theta}_{\text{DLS}}^\varepsilon)$ the corresponding estimator using process Y_n. Therefore, the equation $\mathbb{E}_{F_Y}(\mathcal{X}_{n,n} + \mathcal{P})\text{vec}(\boldsymbol{\Theta}_{\text{DLS}}^\varepsilon) = \mathbb{E}_{F_{Y_{n,n+1}}}(\mathcal{X}_{n,n+1})$ is equivalent to

$$\left\{(1 - \varepsilon)\mathbb{E}_{F_{X_{n,n+1}}}(\mathcal{X}_{n,n} + \mathcal{P}) + \varepsilon\mathbb{E}_{F_{V_{n,n+1}}}(\gamma_t \mathcal{V}_{n,n})\right\}\text{vec}(\boldsymbol{\Theta}_{\text{DLS}}^\varepsilon) = (1 - \varepsilon)\mathbb{E}_{F_{X_{n,n+1}}}(\mathcal{X}_{n,n+1}) + \varepsilon\mathbb{E}_{F_{V_{n,n+1}}}(\gamma_n \mathcal{V}_{n,n+1}).$$

We differentiate with respect to ε and evaluate at $\varepsilon = 0$:

$$\mathbb{E}_{F_{X_{n,n+1}}}(\mathcal{X}_{n,n} + \mathcal{P})\frac{\partial}{\partial \varepsilon}\text{vec}(\boldsymbol{\Theta}_{\text{DLS}}^\varepsilon)\big|_{\varepsilon=0} + \mathbb{E}_{F_{V_{n,n+1}}}(\gamma_n \mathcal{V}_{n,n})\text{vec}(\boldsymbol{\Theta}_{\text{DLS}}) = \mathbb{E}_{F_{V_{n,n+1}}}(\gamma_t \mathcal{V}_{n,n+1}).$$

We observe that $\frac{\partial}{\partial \varepsilon}\text{vec}(\boldsymbol{\Theta}_{\text{DLS}}^\varepsilon)|_{\varepsilon=0} = \text{vec}(\frac{\partial}{\partial \varepsilon}\boldsymbol{\Theta}_{\text{DLS}}^\varepsilon|_{\varepsilon=0}) = \text{vec}(\text{IF}(F_{X,Z,V}; \boldsymbol{\Theta}_{\text{DLS}})) = \text{IF}(F_{X,Z,V}; \text{vec}(\boldsymbol{\Theta}_{\text{DLS}}))$. Thus,

$$\text{IF}(F_{X,Z,V}; \text{vec}(\boldsymbol{\Theta}_{\text{DLS}})) = \{\mathbb{E}_{F_{X_{n,n+1}}}(\mathcal{X}_{n,n} + \mathcal{P})\}^{-1}\{\mathbb{E}_{F_{V_{n,n+1}}}(\gamma_n \mathcal{V}_{n,n+1}) - \mathbb{E}_{F_{V_{n,n+1}}}(\gamma_n \mathcal{V}_{n,n})\text{vec}(\boldsymbol{\Theta}_{\text{DLS}})\}.$$

The contamination is only found in one component; i.e., the joint distribution $F_{V_{n,n+1}}$ is such that $F_{V_{n,n+1}} = \delta_{(V_t(\zeta,s),X_{n+1}(s))}$, where $\delta_{(v_1,v_2)}(x, y)$ is the distribution that assigns mass one to (v_1, v_2). To prove that the corresponding IF is bounded, we note that by definition $\gamma_n = \text{dO}(V_n, F_X)\text{dO}(X_{n+1}, F_X)$ and $\text{dO} = 1/(1 + \text{FO})$, therefore $\gamma_n = O(1/\zeta^2)$. Since $\mathcal{V}_{n,n} = \mathbf{J}_{\eta\eta}^T \otimes \mathbf{J}_{\phi v}^T \mathbf{v}_n \mathbf{v}_n^T \mathbf{J}_{\phi v}$, and $F_{V_{n,n+1}} = \delta_{(V_n(\zeta,s),X_{n+1}(s))}$, we have

$$\mathbb{E}_{F_{V_{n,n+1}}}(\gamma_n \mathcal{V}_{n,n}) = \gamma_n \mathbf{J}_{\eta\eta}^T \otimes \mathbf{J}_{\phi v}^T \mathbf{v}_n \mathbf{v}_n^T \mathbf{J}_{\phi v} = O(1/\zeta^2)O(\zeta^2) = O(1),$$

and

$$\mathbb{E}_{F_{V_{n,n+1}}}(\gamma_n \mathcal{V}_{n,n+1}) = \gamma_n \mathbf{J}_{\eta\eta}^T \otimes \mathbf{J}_{\phi v}^T \mathbf{v}_n \mathbf{x}_{n+1}^T \mathbf{J}_{\phi v} = O(1/\zeta^2)O(\zeta) = O(1/\zeta).$$

Consequently, $\mathbb{E}_{F_{V_{n,n+1}}}(\gamma_n \mathcal{V}_{n,n+1}) - \mathbb{E}_{F_{V_{n,n+1}}}(\gamma_n \mathcal{V}_{n,n})\text{vec}(\boldsymbol{\Theta}_{\text{DLS}}) = O(1)$, and we conclude that $\text{IF}(F_{X,Z,V}; \text{vec}(\boldsymbol{\Theta}_{\text{DLS}}))$ is bounded. $\square$

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2018.06.003.

References

Antoch, J., Prchal, L., De Rosa, M.R., Sarda, P., 2010. Electricity consumption prediction with functional linear regression using spline estimators. J. Appl. Stat. 37 (12), 2027–2041. http://dx.doi.org/10.1080/02664760903214395.

Arribas-Gil, A., Romo, J., 2015. Discussion of Multivariate functional outlier detection. Stat. Methods Appl. 24 (2), 263–267. http://dx.doi.org/10.1007/s10260-015-0328-5.

Attouch, M.K., Laksaci, A., Saïd, E.O., 2012. Robust regression for functional time series data. J. Japan Statist. Soc. 42 (2), 125–143. http://dx.doi.org/10.14490/jjss.42.125.

Bali, J.L., Boente, G., Tyler, D.E., Wang, J.L., 2011. Robust functional principal components: a projection-pursuit approach. Ann. Statist. 39 (6), 2852–2882. http://dx.doi.org/10.1214/11-AOS923.

Boente, G., Salibian-Barrera, M., 2015. S-estimators for functional principal component analysis. J. Amer. Statist. Assoc. 110 (511), 1100–1111. http://dx.doi.org/10.1080/01621459.2014.946991.

Boente, G., Vahnovan, A., 2017. Robust estimators in semi-functional partial linear regression models. J. Multivariate Anal. 154, 59–84. http://dx.doi.org/10.1016/j.jmva.2016.10.005.

Bosq, D., 2000. Linear Processes in Function Spaces. In: Lecture Notes in Statistics, vol. 149, Springer-Verlag, New York. http://dx.doi.org/10.1007/978-1-4612-1154-9. theory and applications.

Cardot, H., Ferraty, F., Sarda, P., 1999. Functional linear model. Statist. Probab. Lett. 45 (1), 11–22. http://dx.doi.org/10.1016/S0167-7152(99)00036-X.

Cardot, H., Ferraty, F., Sarda, P., 2003. Spline estimators for the functional linear model. Statist. Sinica 13 (3), 571–591.

Crambes, C., Mas, A., 2013. Asymptotics of prediction in functional linear regression with functional outputs. Bernoulli 19 (5B), 2627–2651. http://dx.doi.org/10.3150/12-BEJ469.

Dai, W., Genton, M.G., 2018a. Directional outlyingness for multivariate functional data. Comput. Statist. Data Anal..

Dai, W., Genton, M.G., 2018b. Functional boxplots for multivariate curves. Stat. in press.

Dürre, A., Fried, R., Vogel, D., 2016. The spatial sign covariance matrix and its application for robust correlation estimation. arXiv:1606.02274.

Genton, M.G., Ruiz-Gazen, A., 2010. Visualizing influential observations in dependent data. J. Comput. Graph. Statist. 19 (4), 808–825. http://dx.doi.org/10.1198/jcgs.2010.09101. supplementary materials available online.

Gervini, D., 2008. Robust functional estimation using the median and spherical principal components. Biometrika 95 (3), 587–600. http://dx.doi.org/10.1093/biomet/asn031.

Gervini, D., 2012. Functional robust regression for longitudinal data. arXiv:1211.7332.

Goldsmith, J., Bobb, J., Crainiceanu, C.M., Caffo, B., Reich, D., 2011. Penalized functional regression. J. Comput. Graph. Statist. 20 (4), 830–851. http://dx.doi.org/10.1198/jcgs.2010.10007.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions. In: Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York.

Horváth, L., Hušková, M., Kokoszka, P., 2010. Testing the stability of the functional autoregressive process. J. Multivariate Anal. 101 (2), 352–367. http://dx.doi.org/10.1016/j.jmva.2008.12.008.

Horváth, L., Kokoszka, P., Rice, G., 2014. Testing stationarity of functional time series. J. Econometrics 179 (1), 66–82. http://dx.doi.org/10.1016/j.jeconom.2013.11.002.

Huang, H., Sun, Y., 2016. Total variation depth for functional data. arXiv:1611.04913v1.

Hubert, M., Rousseeuw, P., Segaert, P., 2015. Multivariate functional outlier detection. Stat. Methods Appl. 24 (2), 269–277. http://dx.doi.org/10.1007/s10260-015-0327-6.

Hyndman, R.J., Shang, H.L., 2010. Rainbow plots, bagplots, and boxplots for functional data. J. Comput. Graph. Statist. 19 (1), 29–45. http://dx.doi.org/10.1198/jcgs.2009.08158.

Hyndman, R.J., Ullah, M.S., 2007. Robust forecasting of mortality and fertility rates: a functional data approach. Comput. Statist. Data Anal. 51 (10), 4942–4956. http://dx.doi.org/10.1016/j.csda.2006.07.028.

Ivanescu, A.E., Staicu, A.M., Scheipl, F., Greven, S., 2015. Penalized function-on-function regression. Comput. Statist. 30 (2), 539–568. http://dx.doi.org/10.1007/s00180-014-0548-4.

Kokoszka, P., Reimherr, M., 2013. Determining the order of the functional autoregressive model. J. Time Ser. Anal. 34 (1), 116–129. http://dx.doi.org/10.1111/j.1467-9892.2012.00816.x.

Lee, S., Shin, H., Billor, N., 2013. *M*-type smoothing spline estimators for principal functions. Comput. Statist. Data Anal. 66, 89–100. http://dx.doi.org/10.1016/j.csda.2013.03.022.

Locantore, N., Marron, J.S., Simpson, D.G., Tripoli, N., Zhang, J.T., Cohen, K.L., 1999. Robust principal component analysis for functional data. TEST 8 (1), 1–73. http://dx.doi.org/10.1007/BF02595862.

López-Pintado, S., Romo, J., 2009. On the concept of depth for functional data. J. Amer. Statist. Assoc. 104 (486), 718–734. http://dx.doi.org/10.1198/jasa.2009.0108.

Maronna, R.A., Martin, R.D., Yohai, V.J., 2006. Robust Statistics: Theory and Methods. In: Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd., Chichester. http://dx.doi.org/10.1002/0470010940.

Maronna, R.A., Yohai, V.J., 2013. Robust functional linear regression based on splines. Comput. Statist. Data Anal. 65, 46–55. http://dx.doi.org/10.1016/j.csda.2011.11.014. URL http://www.sciencedirect.com/science/article/pii/S0167947311004117.

Martin, R.D., Yohai, V.J., 1986. Influence functionals for time series. Ann. Statist. 14 (3), 781–855. http://dx.doi.org/10.1214/aos/1176350027.

Ramsay, J.O., Dalzell, C.J., 1991. Some tools for functional data analysis. J. Roy. Statist. Soc. Ser. B. 53 (3), 539–572. http://www.jstor.org/stable/2345586.

Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis, second ed.. In: Springer Series in Statistics, Springer, New York.

Scheipl, F., Greven, S., 2016. Identifiability in penalized function-on-function regression models. Electron. J. Stat. 10 (1), 495–526. http://dx.doi.org/10.1214/16-EJS1123.

Sun, Y., Genton, M.G., 2011. Functional boxplots. J. Comput. Graph. Statist. 20 (2), 316–334. http://dx.doi.org/10.1198/jcgs.2011.09224.

Sun, Y., Genton, M.G., 2012. Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. Environmetrics 23 (1), 54–64. http://dx.doi.org/10.1002/env.1136.

Tang, Q., 2014. Robust estimation for functional coefficient regression models with spatial data. Statistics 48 (2), 388–404. http://dx.doi.org/10.1080/02331888.2012.719520.

Wood, S.N., 2006. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. Biometrics 62 (4), 1025–1036. http://dx.doi.org/10.1111/j.1541-0420.2006.00574.x.

Wu, S., Müller, H.G., 2011. Response-adaptive regression for longitudinal data. Biometrics 67 (3), 852–860. http://dx.doi.org/10.1111/j.1541-0420.2010.01518.x.

Zuo, Y., Cui, H., 2005. Depth weighted scatter estimators. Ann. Statist. 33 (1), 381–413. URL https://doi.org/10.1214/009053604000000922.