Biometrika (2017), **xx**, x, *pp*. 1–12

Printed in Great Britain

Projection-based outlier detection in functional data

BY HAOJIE REN

Institute of Statistics, Nankai University, No.94 Weijin Road, Tianjin, China, 300071
haojieren@gmail.com

NAN CHEN

Department of Industrial & Systems Engineering, National University of Singapore, 1 Engineering Drive 2, Singapore 117576
isecn@nus.edu.sg

AND CHANGLIANG ZOU

Institute of Statistics, Nankai University, No.94 Weijin Road, Tianjin, China, 300071
nk.chlzou@gmail.com

SUMMARY

The presence of outliers poses serious adverse effects to the modeling and prediction in data analysis. To detect outliers in functional data, we have developed a procedure based on a high-breakdown mean function estimator. The robust estimator is obtained from a clean subset of observations, excluding potential outliers, by minimizing the least trimmed squares of the projection coefficients after functional principal component analysis. A threshold rule is constructed based on the asymptotic distribution of the functional-score-based distance. The thresholding robustly controls the false positive rate and detects outliers effectively. Further power improvement is proposed by adding a one-step reweighting procedure. The finite sample performance of our method is evaluated through simulations and demonstrates satisfactory false positive and false negative rates compared with that of available outlier detection methods for functional data.

Some key words: Functional principal component analysis; Least trimmed squares estimator; Masking; Reweighting; Robustness; Swamping

1. INTRODUCTION

The analysis of functional data has received substantial attention in recent years due to the increasing availability and collection of observations in the form of functions. Data of this type arises in many disciplines, including growth curves in biology (Ramsay & Silverman, 2005), temperature changes in geosciences (Berkes et al., 2009), and profile monitoring in manufacturing processes (Qiu et al., 2010). For more examples, see Ramsay et al. (2009).

The presence of outlying observations poses serious adverse effects on data analysis, resulting in biased estimation, misleading inference, and poor prediction. Thus, we are concerned about identifying outlying functions among a set of functional observations. More specifically, we assume that N independent observations $\mathcal{X} = \{X_i(t), i = 1, \ldots, N\}$ are collected from the model

$$X_i(t) = \begin{cases} \mu_i(t) + \varepsilon_i(t), & i \in \mathcal{O}_N \\ \mu_0(t) + \varepsilon_i(t), & i \notin \mathcal{O}_N, \end{cases} \tag{1}$$

where $\mathcal{O}_N$ is the outlier set, and the mean function $\mu_i(t)$ of an outlying observation differs from that of the majority, $\mu_0(t)$. $\varepsilon_i(t)$ is the stochastic error with $E\{\varepsilon_i(t)\} = 0$ and $t \in \mathcal{T} = [a, b], -\infty < a < b < \infty$. A straightforward approach to identifying $\mathcal{O}_N$ is to apply multivariate outlier detection procedures when all functions $X_i(t)$ are measured at grid points $t_1, \ldots, t_p$; see Cerioli et al. (2009) and Cerioli (2010) for Mahalanobis distance methods. However, functional data are infinite-dimensional; hence, in practice, the number of grid points is always larger than that of samples, rendering that methods based on Mahalanobis distance are typically inapplicable. Filzmoser et al. (2008), Fritsch et al. (2011) and Ro et al. (2015) designed outlier detection procedures specially for high-dimensional data. However, these methods are not effective for detecting outliers in (1) because they ignore the smoothness of $\mu_i(t)$ and $\mu_0(t)$, and hence would suffer from the curse of dimensionality.

A more viable direction is to detect outliers through functional data analysis. Hyndman & Ullah (2007) used a method based on robust principal component analysis and the integrated squared error of a linear model. Febrero et al. (2008) suggested an approach using the likelihood ratio test and smoothed bootstrapping to identify functional outliers. However, the method is rather time-consuming, especially when N is large. López-Pintado & Romo (2009) introduced a notion of band depth which yields a center-outward ordering of functional data within a sample set. Some authors have also developed graphical tools for visualizing functional data and identifying functional outliers, e.g., Hyndman & Shang (2010), Sun & Genton (2011, 2012), and Genton & Hall (2016). Such graphical approaches are essentially based on the extension of classical ranks to functional data. However, it is not easy to control the false positive rates of these approaches, because they lack systematic distributional results. A closely related work is Yu et al. (2012) which introduced a test founded on functional principal component analysis to detect functional outliers. They derived the null distribution of their test statistic and demonstrated its advantages over some others in terms of outlier detection accuracy. Their procedure focused on the stepwise use of single-case diagnostics, that is, iteratively detecting one outlier and removing it from the remaining set. It is well known from the literature of outlier detection, however, that single-case diagnostics suffer from masking and swamping, and that tests based on them lose power in the presence of multiple outliers (Hawkins, 1980; Barnett & Lewis, 1984).

The goal of this paper is to develop a functional outlier detection procedure based on formal tests with good performance in controlling false positive rates and also with good power for detecting outliers. Our procedure improves on currently available methodologies in two aspects. First, we propose a high-breakdown estimator of the mean function $\mu_0(t)$, obtained from a clean subset of observations, which is assumed to contain only non-outlying observations. The subset can be found by minimizing the least trimmed square of the functional principal component scores. A computationally efficient procedure analogous to the fast minimum covariance determinant algorithm in Rousseeuw & Driessen (1999) is adapted to identify the subset. Second, a threshold rule is constructed based on the asymptotic distribution of the functional-score-based distance. We also propose an additional one-step reweighting procedure, which allows us to control the false positive rate more accurately and to increase the probability of finding outliers.

2. METHODS AND PROPERTIES

2·1. *Least trimmed square of functional scores*

Consider functional observations $\mathcal{X} = \{X_i(t) : i = 1, \ldots, N\}$ that follow model (1). Because of the infinite-dimensional nature of functional data, it is usually important to perform dimension reduction in functional data analysis. Functional principal component analysis is an essential technique to extract a few major and typical features from functional data. Similar to Hyndman & Shang (2010) and Yu et al. (2012), we work on the projections of the functions onto their principal components (Ramsay

& Silverman, 2005). Assume that $\varepsilon(t)$ is a square integrable random function with covariance function $c(t,s) = E\{\varepsilon(t)\varepsilon(s)\}$ for $t, s \in \mathcal{T} = [a,b]$. The spectral decomposition of $c(t,s)$ is $\sum_{k=1}^{\infty} \lambda_k v_k(t) v_k(s)$, where $\lambda_1 \geq \lambda_2 \geq \cdots$ are ordered nonnegative eigenvalues and $v_k(\cdot)$'s are the corresponding orthogonal eigenfunctions with unit L_2 norms, satisfying $\int_{\mathcal{T}} c(t,s) v_k(s) ds = \lambda_k v_k(t)$ $(t \in \mathcal{T};\ k = 1, 2, \ldots)$. Then, the random function $X(t)$ can be written as $X(t) = \mu(t) + \sum_{k=1}^{\infty} \eta_k v_k(t)$, where $\eta_k = \int_{\mathcal{T}} \{X(t) - \mu(t)\} v_k(t) dt$ are uncorrelated random variables, known as functional principal component scores or loadings, with mean 0 and variance λ_k. In practice, most important features from a sample of random functions can usually be explained by the few eigenfunctions with largest eigenvalues.

Given the kth eigenfunction and $\mu_0(t)$, detecting outliers amounts to performing N hypothesis tests with the null being $E(\eta_{ik}) = 0$, where $\eta_{ik} = \int_{\mathcal{T}} \{X_i(t) - \mu_0(t)\} v_k(t) dt$ $(i = 1, \ldots, N)$. In practice, $\mu_0(t)$ is often unknown. Yu et al. (2012) used pooled observations to estimate $\mu_0(t)$ by $\hat{\mu}_{\mathcal{X}}(t) = N^{-1} \sum_{i=1}^{N} X_i(t)$, but this estimator may break down in the presence of multiple outliers, so the corresponding detection procedure suffers from masking and swamping. In robust statistics, a general approach to outlier detection is to choose a clean subset of the dataset that is presumably free from outliers, and then use this subset to obtain robust estimators and test the outlyingness of the remaining observations (Hadi & Simonoff, 1993). To this end, we first propose a clean subset identification method based on functional principal component scores and accordingly obtain a high breakdown estimator of the mean function.

Our method can be viewed as an extension of least trimmed squares (Rousseeuw, 1984) to functional data. The least trimmed squares method searches for the subset of size h with the smallest sum of squared residuals in a linear model with p covariates, where $(N + p + 1)/2 < h < N$. It has several good features and is widely used as a starting point for two-step estimators (Yohai, 1987; Simpson et al., 1992). In a similar spirit, our approach identifies a subset of size h that has the smallest sum of h squared functional principal component scores there.

Let $\mathcal{H} = \{H \subset \{1, \ldots, N\} : |H| = h\}$ denote the set containing all index subsets of size h, where $|H|$ is the cardinality of H. For any $H \in \mathcal{H}$, let $\hat{\mu}_H(t)$ denote the sample mean function of the dataset $\{X_i(t) : i \in H\}$. Assume that $\hat{\lambda}_{kR}$ and $\hat{v}_{kR}(t)$ are appropriate initial estimators of λ_k and $v_k(t)$. We propose a distance to measure the outlyingness

$$D_i(H) = \sum_{k=1}^{d} \hat{\eta}_{ik}^2(H) / \hat{\lambda}_{kR} \quad (i = 1, \ldots, N),$$

based on the functional principal component scores corresponding to the d largest positive eigenvalues, where $\hat{\eta}_{ik}(H) = \int_{\mathcal{T}} \{X_i(t) - \hat{\mu}_H(t)\} \hat{v}_{kR}(t) dt$ $(k = 1, \ldots, d)$. If H is a clean set and the ith observation is an outlier, $D_i(H)$ is expected to be large. To find a clean set, we can choose H which minimizes the sum of the distance.

DEFINITION 1. *The least trimmed functional scores set is defined by*

$$H_{\mathrm{LTFS}} = \arg\min_{H \in \mathcal{H}} \sum_{i=1}^{h} D_{(i)}(H), \tag{2}$$

where $D_{(1)}(H) \leq \cdots \leq D_{(h)}(H)$ are the smallest h values among $\{D_i(H) : i = 1, \ldots, N\}$.

The mean estimator based on the least trimmed functional scores set is accordingly defined by $\hat{\mu}_{H_{\mathrm{L}}}(t) = h^{-1} \sum_{i \in H_{\mathrm{LTFS}}} X_i(t)$. When the observation is non-functional, where $v_k(t) \equiv 1$, H_{LTFS} corresponds to the subset with the smallest variance, and accordingly $\hat{\mu}_{H_{\mathrm{L}}}$ reduces to the least trimmed squares estimator (Rousseeuw & Leroy, 1987).

We compute the finite-sample breakdown point of $\hat{\mu}_{H_{\mathrm{L}}}(t)$ to evaluate its robustness. Generalizing the definition of finite-sample breakdown point (Donoho & Huber, 1983) to the functional principal component analysis framework, the breakdown point of a mean estimator $\hat{\mu}_{\mathcal{X}}(\cdot)$ is defined as

$$b_{N,v}(\hat{\mu}, \mathcal{X}) = \min_{1\le j\le N} \left\{ j/N : \sup_{1\le k\le d} \sup_{\mathcal{X}'} \| \hat{\mu}_{\mathcal{X}} - \hat{\mu}_{\mathcal{X}'} \|_{v_k} = \infty \right\},$$

where the supremum is taken over all sets of $\mathcal{X}'$ obtained by arbitrarily changing j functional observations in $\mathcal{X}$, and $\|\mu\|_{v_k} = |\int_a^b \mu(t) v_k(t) dt|$.

THEOREM 1. *For any dataset $\mathcal{X}$, $b_{N,\hat{v}_R}(\hat{\mu}_{H_{\mathrm{L}}}, \mathcal{X}) = \min(N - h + 1, h)/N$.*

To make the procedure as robust as possible, we take the subset size $h = [N/2] + 1$ because it yields the maximal breakdown point, 50%, where $[n]$ denotes the integer part of n. In infinite-dimensional settings, norms are usually not equivalent. The norm $\|\cdot\|_{v_k}$ is chosen here due to the definition of $D_i(H)$. This result may also hold for other norms if we modify the distance $D_i(H)$ accordingly.

To facilitate the search for H_{LTFS}, we propose the following construction, which guarantees a decrease of the objective function in (2).

THEOREM 2. *Let $H_1 \in \mathcal{H}$. Compute the distance based on H_1, $D_i(H_1)$ for $i = 1, \ldots, N$. If we take H_2 such that $\{D_i(H_1) : i \in H_2\} = \{D_{(1)}(H_1), \ldots, D_{(h)}(H_1)\}$, where $D_{(1)}(H_1) \le \cdots \le D_{(h)}(H_1)$ are the ordered distances, and compute $D_i(H_2)$ based on H_2, then $\sum_{i=1}^h D_{(i)}(H_2) \le \sum_{i=1}^h D_{(i)}(H_1)$. Equality holds if and only if $H_1 = H_2$.*

The fast least trimmed squares algorithm (Rousseeuw & Van Driessen, 2006) can be adapted to find H_{LTFS}, by replacing the squared residuals with $D_i(H)$. Our algorithm is as follows:

Algorithm 1. Least trimmed functional scores

Step 1. Construct a random subset H_0 with $|H_0| = 2$.

Step 2. Set $h = [N/2] + 1$. Based on H_0, obtain $\hat{\mu}_{H_0}(t)$ and compute $D_i(H_0)$ for $i = 1, \ldots, N$. Take H_1 such that $\{D_i(H_0) : i \in H_1\} = \{D_{(1)}(H_0), \ldots, D_{(h)}(H_0)\}$.

Step 3. From H_1, apply the construction in Theorem 2 until convergence and obtain the subset H_2.

Step 4. Repeat steps 1–3 m times, say with $m = 100$, and select the subset with the smallest $\sum_{i=1}^h D_{(i)}(H_2)$ among all m subsets H_2 to be H_{LTFS}.

This is a greedy procedure to find a local minimum from any initial subset. Multiple starting points are used to obtain improved solution to the global minimum H_{LTFS}. The algorithm starts from obtaining a robust estimate of $\mu_0(t)$ for computing $D_i(H_0)$, so we take $|H_0|$ to be 2. The initial eigenfunctions $\hat{v}_{kR}(\cdot)$'s must be specified before the algorithm executes. From Conditions A2 and A3 in the Appendix, we may expect that the estimation accuracy of λ_k and $v_k(\cdot)$ would not affect the performance of H_{LTFS} significantly; see Theorem 3. We also observe in our simulations that the number of outlying observations included in H_{LTFS} changes mildly over a wide range of $v_k(\cdot)$'s estimation. Hence, we suggest applying Ro et al. (2015)'s minimum diagonal product to find an initial subset of h observations, H_{MDP}, and obtain estimators $\hat{\mu}_{H_{\mathrm{MDP}}}(t) = h^{-1} \sum_{i\in H_{\mathrm{MDP}}} X_i(t)$ and

$$\hat{c}(t, s) = h^{-1} \sum_{i\in H_{\mathrm{MDP}}} \{X_i(t) - \hat{\mu}_{H_{\mathrm{MDP}}}(t)\}\{X_i(s) - \hat{\mu}_{H_{\mathrm{MDP}}}(s)\}.$$

The corresponding estimators of λ_k and $v_k(\cdot)$, satisfying $\int_a^b \hat{c}(t,s)\hat{v}_{kR}(s)ds = \hat{\lambda}_{kR}\hat{v}_{kR}(t)$, are used as $\hat{\lambda}_{kR}$ and $\hat{v}_{kR}(\cdot)$. In fact, the method proposed by Filzmoser et al. (2008) to estimate $v_k(\cdot)$ would also yield reasonably good results. See some simulation results in Section 3.

In practice, each function $X_i(t)$ is measured at a set of grid points $\{t_{ij} : j = 1, \ldots, p_i\}$. If these sampling points are the same across different functions, i.e., $t_{ij} = t_j$ and $p_i = p$, then the minimum diagonal product (Ro et al., 2015) can be applied directly on the $N \times p$ observation matrix. If the sampling grid is sparse or the sampling points are not aligned, the functional data can be smoothed first by applying appropriate smoothing techniques such as the spline method or kernel regression, and then use the interpolated values on an equally-spaced grid of points. With respect to the choice of the number d of eigenfunctions $v_k(\cdot)$ used for projection, there are several approaches in the literature (Yao et al., 2005). It seems difficult to select d by cross-validation or using an information criterion since there is no response in outlier detection. Hence, we include all the functional principal components that can explain a predetermined percentage of total variation, such as 90%. Simulation results show that this choice delivers reasonably good detection. The performances using percentages from 75% to 95% are quite similar; see the supplementary material.

2·2. *Threshold rule for outlier detection*

Outlier detection can be cast as N hypothesis tests, of

$$\mathbb{H}_{0i} : E\{X_i(t)\} = \mu_0(t) \text{ versus } \mathbb{H}_{1i} : E\{X_i(t)\} \neq \mu_0(t) \ (i = 1, \ldots, N).$$

To decide whether an observation is an outlier, we propose a threshold rule based on the asymptotic distribution of the distance

$$T_i(\hat{\mu}, \hat{v}, \hat{\lambda}) = \sum_{k=1}^{d} \left[\int_a^b \{X_i(t) - \hat{\mu}(t)\}\hat{v}_k(t)dt \right]^2 / \hat{\lambda}_k,$$

where $\hat{\mu}(\cdot)$, $\hat{\lambda}_k$ and $\hat{v}_k(\cdot)$ are estimates of $\mu_0(\cdot)$, λ_k and $v_k(\cdot)$, and d is fixed.

PROPOSITION 1. *Suppose that Conditions A1–A2 hold and $\hat{\mu}(\cdot)$ is a consistent estimator of $\mu_0(\cdot)$. Then for $i \notin \mathcal{O}_N$, as $N \to \infty$, $T_i(\hat{\mu}, \hat{v}, \hat{\lambda}) \to \chi^2_d$ in distribution.*

We can expect that this result would roughly hold for the distance constructed based on H_{LTFS}, from which reliable approximations can be obtained. This can be partially explained by the following theorem which concerns asymptotic properties of H_{LTFS}.

THEOREM 3. *Suppose Conditions A1 and A3 hold, $|\mathcal{O}_N|/N \to \rho < 1/2$ as $N \to \infty$ and d is fixed.*

(i) Assume that the true eigenfunctions and eigenvalues are used in $D_i(H)$. If the subset H contains $m_N \leq |\mathcal{O}_N|$ outliers, there exists a subset H' which contains no outlier such that

$$\mathrm{pr}\left\{ \sum_{i=1}^{h} D_{(i)}(H') > \sum_{i=1}^{h} D_{(i)}(H) \right\} \leq \exp(-Cm_N),$$

where C is a constant depending only on d, λ_k and the outlying magnitudes are given in Condition A3.

(ii) Suppose $\hat{v}_{kR}$ and $\hat{\lambda}_{kR}$ satisfy Condition A2. Then, $\mathrm{pr}\left\{\sum_{i=1}^{h} D_{(i)}(H') < \sum_{i=1}^{h} D_{(i)}(H)\right\} \to 1$ as $(m_N, N) \to \infty$, where H and H' are defined in (i). Furthermore, if the outlying magnitudes satisfy Condition A4, then

$$\mathrm{pr}\,(H_{\mathrm{LTFS}} \textit{ includes } m_N \textit{ outliers}) \to 0, \quad m_N, N \to \infty. \tag{3}$$

In Theorem 3-(i), if the true eigenfunctions are used, we could obtain a non-asymptotic bound on the probability that the measure of a subset including m_N outliers is smaller than that of a clean subset, which is exponentially small in m_N. That probability would not be large if well-behaved estimates are used in $D_i(H)$, as shown in Theorem 3-(ii). If the subset H contains more than m_N outliers, we can find a clean subset which is better than H with the probability tending to one. Asymptotically speaking, such H cannot be H_{LTFS}. The result (3) in (ii) tells us that the probability that $\sum_{i=1}^{h} D_{(i)}(H') > \sum_{i=1}^{h} D_{(i)}(H)$ could be uniformly small in H if some stronger conditions are imposed. That is, with high probability, H_{LTFS} cannot contain too many outliers.

After obtaining the clean subset H_{LTFS}, we update the estimation of $c(t,s)$ by

$$\hat{c}_{H_{\mathrm{L}}}(t,s) = h^{-1} \sum_{i \in H_{\mathrm{LTFS}}} \{X_i(t) - \hat{\mu}_{H_{\mathrm{L}}}(t)\}\{X_i(s) - \hat{\mu}_{H_{\mathrm{L}}}(s)\}.$$

The corresponding number of principal components, eigenvalues and eigenfunctions are respectively denoted by d_{H_L}, $\hat{\lambda}'_{kH_{\mathrm{L}}}$ and $\hat{v}_{kH_{\mathrm{L}}}(\cdot)$. Because the least trimmed functional scores algorithm searches for a clean subset which has the smallest total variation, $\hat{\lambda}'_{kH_{\mathrm{L}}}$ would underestimate λ_k (Pison et al., 2002). We may adjust the estimate by $\hat{\lambda}_{kH_{\mathrm{L}}} = \theta\hat{\lambda}'_{kH_{\mathrm{L}}}$, where the scaling constant θ depending on the values of $h,\ N$, serves to render $\hat{\lambda}_{kH_{\mathrm{L}}}$ consistent. By Proposition 1, in probability

$$\mathrm{median}_{1\le i\le N} T_i(\hat{\mu}_{H_{\mathrm{L}}}, \hat{v}_{H_{\mathrm{L}}}, \hat{\lambda}_{H_{\mathrm{L}}}) \to \chi^2_{d_{H_{\mathrm{L}}}}(0.5), \ \text{ as } N \to \infty,$$

and thus we take

$$\theta = \frac{\mathrm{median}_{1\le i\le N} T_i(\hat{\mu}_{H_{\mathrm{L}}}, \hat{v}_{H_{\mathrm{L}}}, \hat{\lambda}'_{H_{\mathrm{L}}})}{\chi^2_{d_{H_{\mathrm{L}}}}(0.5)}, \tag{4}$$

where $\chi^2_d(\alpha)$ is the α upper quantile of the χ^2_d distribution.

Proposition 1 suggests that the ith observation is identified as an outlier if

$$T_i(\hat{\mu}_{H_{\mathrm{L}}}, \hat{v}_{H_{\mathrm{L}}}, \hat{\lambda}_{H_{\mathrm{L}}}) > \chi^2_{d_{H_{\mathrm{L}}}}(\alpha), \tag{5}$$

where α is a pre-specified significance level. This threshold rule enables us to control the false positive rate, i.e, the percentage of non-outlying samples incorrectly identified as outliers. Similar to Proposition 1, we can show that under the alternative $\mathbb{H}_{1i}$, $T_i(\hat{\mu}_{H_{\mathrm{L}}}, \hat{v}_{H_{\mathrm{L}}}, \hat{\lambda}_{H_{\mathrm{L}}})$ asymptotically has a non-central chi-squared distribution with d degrees of freedom and noncentrality parameter $\sum_{k=1}^{d}[\int_{\mathcal{T}}\{\mu_i(t) - \mu_0(t)\}v_k(t)]^2/\lambda_k$.

2·3. *One-step refined procedure*

A one-step reweighting scheme is usually able to enhance the efficiency of outlier detection procedures (Cerioli, 2010). We refine the outlier detection rule after using the threshold rule (5). Let H_{R} be the subset of observations $X_i(t)$ for which $T_i(\hat{\mu}_{H_{\mathrm{L}}}, \hat{v}_{H_{\mathrm{L}}}, \hat{\lambda}_{H_{\mathrm{L}}}) < \chi^2_{d_{H_{\mathrm{L}}}}(\alpha')$, where α' is a predetermined significance level. Based on H_{R}, we update the estimation of $c(t,s)$ by $\hat{c}_{H_{\mathrm{R}}}(t,s)$ and $\mu_0(t)$ by $\hat{\mu}_{H_{\mathrm{R}}}$. Let $d_{H_{\mathrm{R}}}$, $\hat{\lambda}'_{kH_{\mathrm{R}}}$ and $\hat{v}_{kH_{\mathrm{R}}}(\cdot)$ be the number of principal components, eigenvalue and eigenfunction estimators, respectively. Similar to (4), a consistent estimate of λ_k can be obtained by

$$\hat{\lambda}_{kH_{\mathrm{R}}} = \hat{\lambda}'_{kH_{\mathrm{R}}} \frac{\mathrm{median}_{i\in H_{\mathrm{R}}} T_i(\hat{\mu}_{H_{\mathrm{R}}}, \hat{v}_{H_{\mathrm{R}}}, \hat{\lambda}'_{H_{\mathrm{R}}})}{\chi^2_{d_{H_{\mathrm{R}}}}(0.5)}. \tag{6}$$

Then, a refined distance can be constructed as $T_i(\hat{\mu}_{H_{\mathrm{R}}}, \hat{v}_{H_{\mathrm{R}}}, \hat{\lambda}_{H_{\mathrm{R}}})$. In (6), we use the median of the distance $T_i(\hat{\mu}_{H_{\mathrm{R}}}, \hat{v}_{H_{\mathrm{R}}}, \hat{\lambda}'_{H_{\mathrm{R}}})$ in the subset H_{R} rather than the median based on all observations as in (4).

When outliers are present, using the median among all samples would yield a larger scaling constant and in turn overestimate the λ_k. With the help of H_{R}, a more accurate estimate $\hat{\lambda}_{kH_{\mathrm{R}}}$ is achieved, and real false alarm rate can be better controlled. The refined procedure fits naturally in our setting as it involves only one additional step, and leads to an increase in the probability of correct identification of outlying observations. Finally, our proposed method is outlined as follows:

Algorithm 2. Refined least trimmed functional scores

Step 1. Compute the subset H_{LTFS} with $|H_{\mathrm{LTFS}}| = [N/2] + 1$ based on Algorithm 1.

Step 2. Set a significance level α. Compute the distance $T_i(\hat{\mu}_{H_{\mathrm{L}}}, \hat{v}_{H_{\mathrm{L}}}, \hat{\lambda}_{H_{\mathrm{L}}})$ and identify H_{R} based on the threshold rule (5) with α', e.g., $\alpha' = \alpha/2$.

Step 3. Compute the refined distances, and the ith observation is identified as an outlier if $T_i(\hat{\mu}_{H_{\mathrm{R}}}, \hat{v}_{H_{\mathrm{R}}}, \hat{\lambda}_{H_{\mathrm{R}}}) > \chi^2_{d_{H_{\mathrm{R}}}}(\alpha)$.

The entire outlier detection algorithm runs fast. For instance, when a sample has $N = 500$ observations with 10% outliers, it takes about only 7.6 seconds to run on an Intel i5 CPU using R (R Development Core Team, 2012). The R code is available in the supplementary material.

3. SIMULATION

In this section, we evaluate the performance of our proposed outlier detection procedure through a simulation study. We first study the accuracy of the subset H_{LTFS}, and then compare the false positive and false negative rates of Algorithm 2 with some existing methods. All the results in this section are obtained from 1,000 replicated simulations.

Without loss of generality, we fix $\mu_0(t) \equiv 0$ and consider the following three different models for $\varepsilon_i(t)$, namely the trajectories of the Brownian motion, the autoregressive model and the moving average model. All three processes are realised on a grid of $p = 100$ or $p = 500$ equispaced points in $\mathcal{T} = [0, 1]$. The independent increment of the Brownian motion is generated by $\varepsilon_i(t_{j+1}) - \varepsilon_i(t_j) \sim N(0, 0.2)$. The autoregressive model is $\varepsilon_i(t_j) = \varepsilon_i(t_{j-1}) - 0.9\varepsilon_i(t_{j-2}) + \epsilon(t_j)$ and the moving average model follows $\varepsilon_i(t_j) = \epsilon(t_j) + 0.5\epsilon(t_{j-1}) + 0.3\epsilon(t_{j-2})$, where $t_j = j/p$ $(j = 1, \ldots, p)$ and $\epsilon(t_j) \sim N(0, 1)$. Two classes of outlying functions are investigated: (a) $\mu(t) = \gamma \sin(2\pi t) I_{[1/3,1/2]}(t)$; and (b) $\mu(t) = \gamma t I_{[a_1/p, a_2/p]}(t)$, where $I(\cdot)$ is the indicator function and the locations a_1 and a_2 are randomly sampled from 1 to p. We generate outliers from a mixture of these two classes, i.e., in one outlier set, the proportions of outlying functions from (a) and (b) are ω and $1 - \omega$, respectively. Two cases are explored: (I) $\omega = 75\%$; (II) $\omega = 25\%$. We choose the sample size N to be $100, 200, 500$ and 1000 and the outlier ratio ρ in the sample to be $0.02, 0.04, 0.1$ and 0.2.

Following the basis function method introduced in Ramsay & Silverman (2005), all the observations in our simulation study are smoothed by 15 Fourier basis functions. Our simulation results indicate that our procedure is not affected much by the type of basis or the number of the basis functions used for smoothing. In practice, the local polynomial smoothing or spline can be used as well. We find that with 5 to 15 Fourier basis functions, the level of the proposed test can be maintained within an acceptable range. To provide a better protection against outliers with oscillating mean functions, we use a relatively large number of bases. In each replication, the number of eigenfunctions, d, is chosen such that 90% of the total variation can be explained by the first d principal components.

We first show that the subset H_{LTFS} determined by Algorithm 1 is relatively reliable and clean in finite samples. We compare the numbers of outlying functions contained in H_{LTFS} and H_{MDP}, where H_{MDP} is the subset chosen by Ro et al. (2015)'s minimum diagonal product algorithm. Figure 1 presents the numbers of outliers in H_{LTFS} and H_{MDP} under Case (I) with $\gamma = 2$, $N = 1000$ and $p = 100, 500$. It

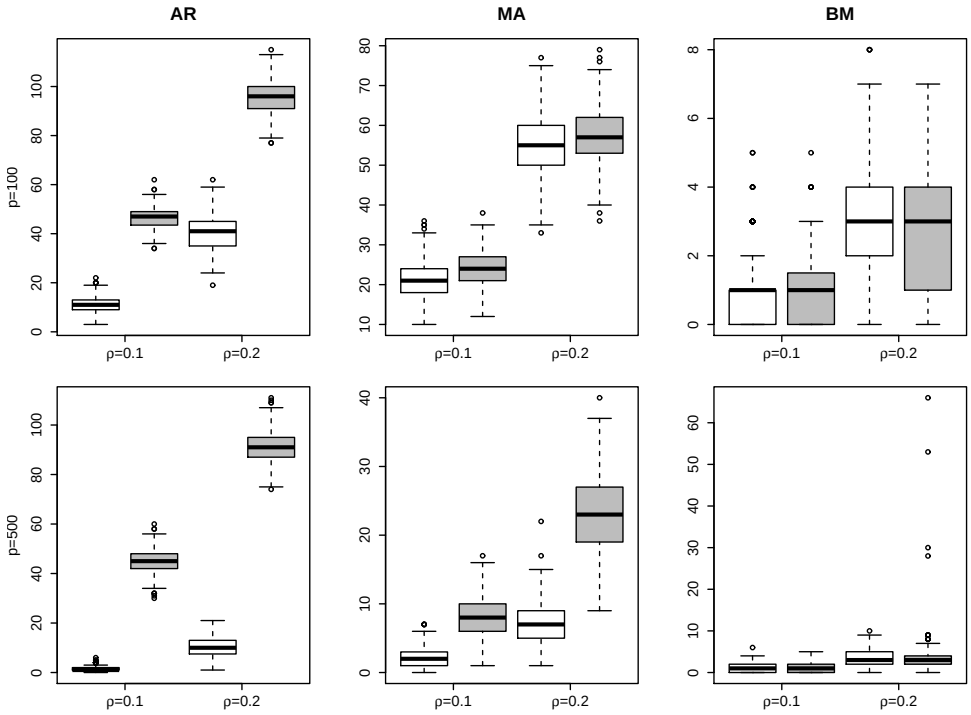

Fig. 1. Boxplots of the numbers of outliers contained in H_{LTFS} (white boxes) and H_{MDP} (grey boxes) under Case (I) when $N = 1000$ and $\gamma = 2$.

is evident that H_{LTFS} generally outperforms H_{MDP}, and its advantage becomes more pronounced for larger p or ρ. The performance of H_{LTFS} is affected by different choices of $\hat{v}_k(\cdot)$, but the influence is not significant; see the supplementary material.

Table 1. *Empirical false positive rates (%) under Case (I) for different outlier ratios* ρ *with* $\alpha = 1\%, 5\%, 10\%$*, when* $N = 100, 200, 500, 1000$*,*$\gamma = 2$ *and* $p = 500$

$\varepsilon_i(t)$	N	$\rho = 0.04$			$\rho = 0.1$			$\rho = 0.2$		
		1%	5%	10%	1%	5%	10%	1%	5%	10%
AR	100	0.9	4.6	10.4	0.7	4.2	9.5	0.6	3.9	9.0
	200	0.9	4.8	10.6	0.8	4.5	10.1	0.7	4.2	9.4
	500	0.9	4.9	10.6	0.8	4.9	10.4	0.8	4.7	9.9
	1000	1.0	5.2	10.9	0.9	5.0	10.6	0.9	4.8	10.2
MA	100	0.7	4.7	10.2	0.6	4.2	9.6	0.5	3.4	8.5
	200	0.8	4.8	10.5	0.7	4.5	9.9	0.6	4.0	9.2
	500	0.9	4.9	10.6	0.8	4.7	10.3	0.7	4.3	9.4
	1000	0.9	5.1	10.9	0.9	4.9	10.5	0.7	4.4	9.7
BM	100	1.5	7.3	14.3	1.5	6.9	13.5	1.5	6.2	12.1
	200	1.2	6.0	12.8	1.1	5.8	12.0	1.2	5.5	11.0
	500	1.1	6.0	12.3	1.1	5.7	11.7	1.1	5.3	10.9
	1000	1.1	5.8	12.1	1.0	5.5	11.5	1.0	5.2	10.8

AR stands for autoregressive, MA for moving average, and BM for Brownian motion.

Table 2. *Empirical false positive* (α) *and false negative* (β) *rates (%) under Case (I) for different outlier ratios* ρ *with nominal size* $\alpha = 5\%$*, when* $\gamma = 2$*,* $N = 200$ *and* $p = 500$

$\varepsilon_i(t)$	ρ	ReLTFS		DFOD		SFOD		RMDP		PCOut	
		α	β	α	β	α	β	α	β	α	β
AR	0.02	5.1	0.1	13.2	86.0	2.7	0.0	4.5	94.3	6.9	89.2
	0.04	4.8	2.0	9.8	89.8	2.6	2.1	4.2	94.7	7.1	89.0
	0.1	4.5	3.0	9.0	91.3	2.3	23.3	4.1	94.7	7.1	90.0
	0.2	4.2	6.7	9.9	89.7	0.5	80.9	4.1	95.1	7.2	90.5
MA	0.02	5.0	8.9	58.3	18.5	2.6	12.3	4.9	45.8	6.7	83.9
	0.04	4.9	11.3	61.0	17.1	2.6	15.6	4.7	46.2	6.8	49.7
	0.1	4.5	16.1	60.1	15.6	2.2	53.7	4.0	49.4	5.8	14.1
	0.2	4.0	29.7	23.6	61.0	1.2	82.0	3.2	58.3	2.7	35.6
BM	0.02	6.6	1.6	0.2	0.0	0.0	14.3	1.6	23.8	10.2	0.0
	0.04	6.4	3.8	0.1	10.5	0.0	12.3	1.5	27.7	9.7	0.1
	0.1	6.0	5.7	0.1	71.1	0.0	89.5	1.1	39.3	8.0	1.6
	0.2	5.7	6.3	0.0	85.8	0.0	92.0	0.8	62.9	6.1	11.3

ReLTFS: our refined least trimmed functional scores method; DFOD: the depth-based functional outlier detection procedure introduced by Febrero et al. (2008); SFOD: the stepwise functional outlier detection procedure proposed by Yu et al. (2012); RMDP: refined minimum diagonal product procedure by Ro et al. (2015); PCOut: principal component outlier detection procedure by Filzmoser et al. (2008)

To evaluate outlier detection performance, we compare the false positive and false negative rates, which correspond to the swamping and masking ratios, respectively. Simulation results, not reported here, show that our one-step refined procedure is generally better at controlling the false positive rate than the threshold rule (5), consistent with our analysis in Section 2.3, so in this section we focus on studying the performance of the refined procedure. Table 1 presents the false positive rates of the refined least trimmed functional scores method under Case (I) with $\gamma = 2$ for various combinations of ρ and N. In most cases, the empirical false positive rates are close to the nominal levels. Results for Case (II) can be found in the supplementary material.

We next compare the proposed outlier detection procedure with existing methods in Febrero et al. (2008), Yu et al. (2012), Ro et al. (2015) and Filzmoser et al. (2008). The approaches of Febrero et al. (2008) and Filzmoser et al. (2008) are implemented by the R packages fda.usc and mvoutlier, respectively.

Simulation results with $\alpha = 5\%$ are reported in Tables 2 and 3. In most cases, our method attains the nominal false positive rates and also yields smaller false negative rates than other methods. Yu et al. (2012)'s method does not perform well, as its false positive rates are overly conservative and the false negative rates increase rapidly when the outlier ratio ρ increases. This is not surprising, since the estimators in Yu et al. (2012) are calculated using all the observations, so the detection procedure suffers severely from masking, in which case, the removal of any single outlier may have little or no effect since other outliers remain. Accordingly, in the presence of multiple outliers, the estimators of the mean function and the covariance function will be distorted. Filzmoser et al. (2008)'s method approximately attains a false positive rate of 5% under the autoregressive and moving average models, but is generally outperformed by our method in terms of the false negative rates. Under the Brownian motion model, Filzmoser et al. (2008)'s method has performance comparable with our method in terms of false negative rates but has higher false positive rates. Ro et al. (2015)'s procedure performs similarly to Filzmoser et al. (2008)'s under the autoregressive and moving average models but has much more conservative false positive rates under the Brownian motion model. The performances of Febrero et al. (2008)'s method in false positive rates are not stable: it appears to be liberal under the autoregressive model but conservative

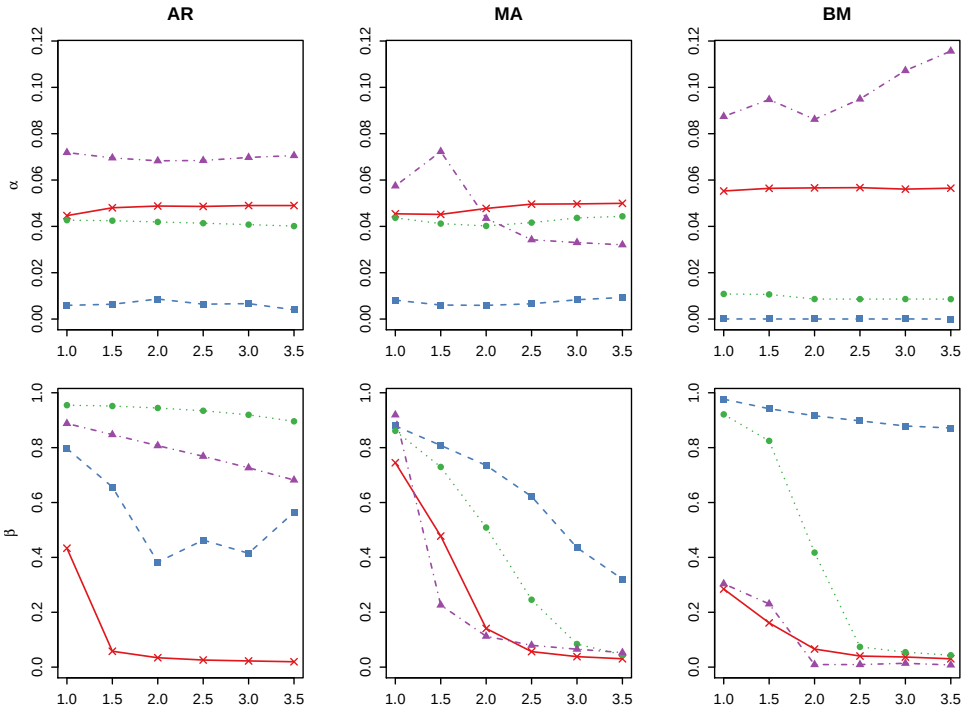

Fig. 2. Empirical false positive rates (α, top) and false negative rates (β, bottom) of four methods as follows: our refined least trimmed functional scores method (asterisks solid red line); Yu et al. (2012) (square dash blue line); Ro et al. (2015) (circle dot green line) and Filzmoser et al. (2008) (triangle dot-dash purple line) under Case (I) for various values of γ, when $\rho = 0.1$, $N = 500$ and $p = 500$.

under the Brownian motion model. For a fixed sample size, its false negative rates increase as the ratio of the outliers increases, showing that this approach suffers considerably from masking.

Empirical false positive rates and false negative rates under Case (I) are presented in Fig. 2 for $\rho = 0.1$, $\alpha = 5\%$, $N = 500$ and $p = 500$. Our method has stable false positive rates close to the nominal level, even when γ varies. It is able to successfully identify true outliers as γ increases and generally performs better than the other competitors. Similar conclusions can be drawn from additional simulation results given in the supplementary material.

ACKNOWLEDGMENT

The authors are grateful to the referees, Associate Editor, and Editor for comments that have significantly improved the article.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of theorems, additional simulation results, case studies and R code.

Table 3. *Empirical false positive* (α) *and false negative* (β) *rates (%) under Case(II) for different outlier ratios* ρ *with nominal size* $\alpha = 5\%$*, when* $\gamma = 3.5$*,* $N = 200$ *and* $p = 500$

$\varepsilon_i(t)$	ρ	ReLTFS		DFOD		SFOD		RMDP		PCOut	
		α	β	α	β	α	β	α	β	α	β
AR	0.02	5.1	7.7	11.1	87.5	2.7	8.5	4.4	92.8	7.4	83.1
	0.04	4.9	9.4	10.1	88.5	2.6	10.3	4.4	93.6	6.9	84.7
	0.1	4.6	10.2	9.6	88.7	2.8	11.0	4.4	93.9	6.7	88.5
	0.2	4.2	10.6	12.5	85.5	3.1	12.3	4.2	93.9	6.7	89.9
MA	0.02	4.8	16.6	64.0	13.3	2.5	19.0	4.8	42.3	7.2	74.7
	0.04	4.8	18.4	51.6	19.3	2.5	21.1	4.8	40.3	7.3	56.8
	0.1	4.6	20.9	11.5	66.4	2.6	24.0	4.2	41.7	5.2	43.7
	0.2	4.0	23.9	1.9	83.2	2.5	33.4	3.4	44.5	2.9	49.8
BM	0.02	6.6	9.1	0.2	13.5	0.0	26.8	1.6	28.1	10.7	1.1
	0.04	6.4	12.6	0.0	31.8	0.0	27.0	1.4	31.3	10.6	1.8
	0.1	6.1	14.0	0.0	62.6	0.0	25.9	1.1	36.6	10.2	2.0
	0.2	5.7	13.9	0.0	74.3	0.0	47.7	0.9	45.2	8.1	2.6

APPENDIX

We require the following technical conditions.

*Condition A*1. The mean $\mu_0(t)$ and $\mu_i(t)$ $(i \in \mathcal{O}_N)$ are all square-integrable, i.e. they lie in $\mathcal{L}^2(\mathcal{T})$. The $\varepsilon_i(t)$ are independent and identically distributed, following a Gaussian process with mean 0 and a square-integrable covariance function $c(t, s)$.

*Condition A*2. For each $k = 1, \ldots, d$,

$$\limsup_{N\to\infty} \left\{\zeta_N E\|\hat{c}_k v_k - \hat{v}_k\|^2\right\} < \infty, \quad \limsup_{N\to\infty} \left\{\zeta_N E|\lambda_k - \hat{\lambda}_k|^2\right\} < \infty,$$

where $\zeta_N^{-1/2} \log N \to 0$, $\hat{c}_k = \text{sgn}\left\{\int_{\mathcal{T}} v_k(t)\hat{v}_k(t)dt\right\}$, and v_k and λ_k are orthonormal eigenfunctions and non-negative eigenvalues of some square-integrable covariance function.

*Condition A*3. For any $i \in \mathcal{O}_N$, there exists at least one $v_k(\cdot)$ so that its outlying magnitude $\Delta_{ik}^2 \equiv \left[\int_{\mathcal{T}}\{\mu_i(t) - \mu_0(t)\}v_k(t)dt\right]^2 > 0$. Moreover, assume that the largest and smallest magnitudes are bounded, i.e.,

$$\delta_L \le \min_{i\in\mathcal{O}_N} \max_{1\le k\le d} \Delta_{ik}^2 \le \max_{i\in\mathcal{O}_N} \max_{1\le k\le d} \Delta_{ik}^2 \le \delta_U.$$

*Condition A*4. The outlying magnitudes satisfy

$$\delta_L > \frac{2d(\delta_U \lambda_d^{-1})^{1/2} + \max_{0<q<1}\{\psi_G(q') - \psi_G(q)\}}{(1 - 2\rho)\lambda_1^{-1}}, \tag{A1}$$

where $\psi_G(q) = \int_0^q G_d^{-1}(z)dz$, $q' = \{1 - q^{-1} + (2\rho q)^{-1}\}^{-1}$, $G_d(\cdot)$ is the cumulative distribution function of χ_d^2.

Condition A1 is quite standard in the literature. It implies the following expansions

$$c(t, s) = \sum_{k=1}^{\infty} \lambda_k v_k(t)v_k(s), \;\; \varepsilon_i(t) = \sum_{k=1}^{\infty} \eta_{ik} v_k(t), \tag{A2}$$

where the sequences $\{\eta_{ik} : i = 1, \ldots, N, k = 1, 2, \ldots\}$ are independent identically distributed normal random variables with mean 0 and variance λ_k. Condition A2 is to guarantee that $\hat{\lambda}_k$ and $\hat{v}_k(\cdot)$ converge as $N \to \infty$. If the data do not contain outliers and $\hat{\lambda}_k$ and $\hat{v}_k(\cdot)$ are the sample estimators, Lemmas 4.3 of Bosq (2000) implies that this assumption holds for $\zeta_N = N$. Condition A3 is a fairly common technical assumption in order to help us distinguish the outliers from normal data. We consider that the outlying magnitude of each outlier is fixed and does

not depend on N. Because our method relies on the projections, it is mild to assume that for each $i \in \mathcal{O}_N$ there is at least one v_k so that $\Delta_{ik} \neq 0$. Condition A4 requires that the smallest magnitude should be sufficiently large to guarantee identifiability. It is easy to show that when $\rho \leq 0.25$, $\max_{0<q<1}\{\psi_G(q') - \psi_G(q)\} = 0$ and accordingly Condition A4 reduces to $\delta_L > C\delta_U^{1/2}$ for some constant C. When all the Δ_{ik}'s are of similar magnitudes, Condition A4 is easily satisfied.

REFERENCES

BARNETT, V. & LEWIS, T. (1984). *Outliers in Statistical Data (2nd ed.)*. Wiley: New York.

BERKES, I., GABRYS, R., HORVÁTH, L. & KOKOSZKA, P. (2009). Detecting changes in the mean of functional observations. *J. R. Statist. Soc. B* **71**, 927–946.

BOSQ, D. (2000). *Linear Processes in Function Spaces*. Springer: New York.

CERIOLI, A. (2010). Multivariate outlier detection with high-breakdown estimators. *J. Am. Statist. Assoc.* **105**, 147–156.

CERIOLI, A., RIANI, M. & ATKINSON, A. C. (2009). Controlling the size of multivariate outlier tests with the mcd estimator of scatter. *Statist. Comput.* **19**, 341–353.

DONOHO, D. L. & HUBER, P. J. (1983). The notion of breakdown point. In *A festschrift for Erich L. Lehmann*. Ed. P. J. Bickel, K. A. Doksum and J. L. Hodges, pp. 157–184. Belmont: Wadsworth.

FEBRERO, M., GALEANO, P. & GONZÁLEZ-MANTEIGA, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal nox levels. *Environmetrics* **19**, 331–345.

FILZMOSER, P., MARONNA, R. & WERNER, M. (2008). Outlier identification in high dimensions. *Comput. Statist. Data Anal.* **52**, 1694–1711.

FRITSCH, V., VAROQUAUX, G., THYREAU, B., POLINE, J.-B. & THIRION, B. (2011). Detecting outlying subjects in high-dimensional neuroimaging datasets with regularized minimum covariance determinant. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Ed. G. Fichtinger, A. Martel and T. Peters, pp. 264–271. Springer: Heidelberg.

GENTON, M. G. & HALL, P. (2016). A tilting approach to ranking influence. *J. R. Statist. Soc. B* **78**, 77–97.

HADI, A. S. & SIMONOFF, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *J. Am. Statist. Assoc.* **88**, 1264–1272.

HAWKINS, D. M. (1980). *Identification of Outliers*, vol. 11. Chapman & Hall: London.

HYNDMAN, R. J. & SHANG, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *J. Comput. Graph. Statist.* **19**, 29–45.

HYNDMAN, R. J. & ULLAH, M. S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Comput. Statist. Data Anal.* **51**, 4942–4956.

LÓPEZ-PINTADO, S. & ROMO, J. (2009). On the concept of depth for functional data. *J. Am. Statist. Assoc.* **104**, 718–734.

PISON, G., VAN AELST, S. & WILLEMS, G. (2002). Small sample corrections for lts and mcd. *Metrika* **55**, 111–123.

QIU, P., ZOU, C. & WANG, Z. (2010). Nonparametric profile monitoring by mixed effect modeling (with discussions). *Technometrics* **52**, 265–277.

R DEVELOPMENT CORE TEAM (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

RAMSAY, J. O., HOOKER, G. & GRAVES, S. (2009). *Functional Data Analysis with R and MATLAB*. Springer: New York.

RAMSAY, J. O. & SILVERMAN, B. W. (2005). *Functional Data Analysis*. Springer: New York.

RO, K., ZOU, C., WANG, Z. & YIN, G. (2015). Outlier detection for high-dimensional data. *Biometrika* **102**, 589–599.

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Am. Statist. Assoc.* **79**, 871–880.

ROUSSEEUW, P. J. & DRIESSEN, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41**, 212–223.

ROUSSEEUW, P. J. & LEROY, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley: New York.

ROUSSEEUW, P. J. & VAN DRIESSEN, K. (2006). Computing lts regression for large data sets. *Data Mining and Knowledge Discovery* **12**, 29–45.

SIMPSON, D. G., RUPPERT, D. & CARROLL, R. J. (1992). On one-step gm estimates and stability of inferences in linear regression. *J. Am. Statist. Assoc.* **87**, 439–450.

SUN, Y. & GENTON, M. G. (2011). Functional boxplots. *J. Comput. Graph. Statist.* **20**, 316–334.

SUN, Y. & GENTON, M. G. (2012). Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. *Environmetrics* **23**, 54–64.

YAO, F., MÜLLER, H.-G. & WANG, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33**, 2873–2903.

YOHAI, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* , 642–656.

YU, G., ZOU, C. & WANG, Z. (2012). Outlier detection in functional observations with applications to profile monitoring. *Technometrics* **54**, 308–318.