Ann Inst Stat Math (2019) 71:577–598
https://doi.org/10.1007/s10463-018-0656-y

M-based simultaneous inference for the mean function of functional data

Italo R. Lima[1] · **Guanqun Cao**[1] · **Nedret Billor**[1]

Received: 14 June 2017 / Revised: 20 November 2017 / Published online: 13 March 2018

Abstract Estimating and constructing a simultaneous confidence band for the mean function in the presence of outliers is an important problem in the framework of functional data analysis. In this paper, we propose a robust estimator and a robust simultaneous confidence band for the mean function of functional data using M-estimation and B-splines. The robust simultaneous confidence band is also extended to the difference of mean functions of two populations. Further, the asymptotic properties of the M-based mean function estimator, such as the asymptotic consistency and asymptotic normality, are studied. The performance of the proposed robust methods and their robustness are demonstrated with an extensive simulation study and two real data examples.

Keywords Confidence band · Functional data analysis · Robust statistics · Spline smoothing · M-estimator · Pseudo-data

1 Introduction

Due to the advancements in computer technology, experimenters collect complex, high-dimensional data sets, such as curves, 2D or 3D images and other objects living in a functional space. This type of dataset so-called functional data is nowadays seen

Cao's research is supported in part by the Simons Foundation under Grant #354917 and the National Science Foundation under Grants DMS 1736470. We thank the Associate Editor and two anonymous referees for their helpful and constructive comments, which lead to significant improvement in this paper.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-018-0656-y) contains supplementary material, which is available to authorized users.

✉ Guanqun Cao
gzc0009@auburn.edu

[1] Department of Mathematics and Statistics, Auburn University, Auburn, AL 36849, USA

in almost all scientific fields such as geophysics (Ferraty et al. 2005), environmetrics (Febrero et al. 2008), ecology (Embling et al. 2012), chemometrics (Daszykowski et al. 2007). A general overview of the functional data analysis (FDA) and an extensive list of references can be found in the seminal work of Silverman and Ramsay (2005) and Ferraty (2011). In the last two decades, many FDA techniques have been developed as extensions of multivariate data analysis techniques. In all these techniques, as in the multivariate data analysis, it is assumed that functional data are homogeneous which are free of outliers. Although robust multivariate statistical techniques have been studied heavily in multivariate setting, this phenomenon has not been studied widely in functional data setting. Only recently some robust FDA techniques have been proposed. Among the recent literature, Gervini (2008) proposed a robust estimator for the location parameter of contaminated datasets, by extending the notion of median to functional datasets, and also proposed a robust alternative for the functional principal components analysis (FPCA) based on the spherical principal components defined in Locantore et al. (1999). Bali et al. (2011) and Lee et al. (2013) proposed robust estimators for the functional principal components by adapting the projection pursuit approach and based on MM estimation, respectively. Kraus and Panaretos (2012) used a different approach, instead estimating the dispersion operator containing influential observations. Maronna and Yohai (2013) established a robust version of spline-based estimators for a linear functional regression model. Further, Shin and Lee (2016) proposed a robust procedure based on convex and non-convex loss functions in functional linear regression models and the theoretical developments of this estimator by using numerical studies with various types of robust loss are provided.

In the presence of outliers, which are the observed curves behaving differently from the remaining curves, the estimate of the functional mean, and therefore simultaneous confidence bands (SCBs) for the mean function, may be affected badly which would yield misleading statistical conclusions. The inherent complexity of infinite-dimensional functional spaces allows for a wide variety of possible outlier behaviors, adding further complication to estimating functional mean and constructing SCB for the functional mean. Therefore, in this study we aim at developing outlying-resistant methods for the mean function in the functional setting that can provide valid statistical inference even in the presence of a significant proportion of outlier curves.

More recently, Lima et al. (2017) proposed LAD-based estimation of the mean function, and the asymptotic consistency of the LAD estimator was studied. However, the variance of the LAD-based estimator of the mean function was not derived due to the lack of an explicit form of the variance function of the LAD estimator and, consequently, the asymptotic normality of the mean function estimator was not provided. In Lima et al. (2017), this problem was overcome by using a spherical principal component based robust estimation of the covariance function of the LAD estimator (Gervini 2008) and simulated correction factors which resulted in SCBs with a large width. In this study, we have overcome these circumstances by developing M-based estimator. To the best of our knowledge, our proposal is the first publicly available implementation that allows such a robust estimator and confidence band for FDA, while prior work either limits the SCB to homogeneous functional data, such as Gu et al. (2014) and Cao et al. (2012) or to the LAD estimator without the theoretical support for the asymptotic normality of the estimator and RSCB (Lima et al. 2017).

The problem of simultaneous inference for functional data as well as the inclusion of robustness is non-trivial. First, the greatest technical difficulty is to formulate SCB for a mean function of infinite-dimensional functional response and establish their theoretical properties. Second, unlike the scenarios considered in the classical FDA literature, in our settings, the outlier structure considered is complex.

In this paper, we have two main contributions. First, we develop the asymptotic theory for a class of robust estimators for the mean function of functional data, based on M-estimation and B-splines smoothing techniques. Second, we propose a method for constructing SCB that is resilient to outlying curves. We further extend the robust simultaneous inference to the two-sample case and evaluate the equality of mean functions from two groups when atypical curves exist. Our numerical simulation results show that the proposed bands are superior to existing classical methods which do not account for atypical curves.

The paper is organized as follows. We begin by establishing the model of our analysis and then present our proposed methods in Sect. 2. In this section, we first define a robust estimator for the mean function when the dataset contains outliers, extending the M-estimation techniques and B-splines smoothing to functional data and asymptotic properties of the B-spline smoothed M-estimator are studied. Then we construct a RSCB for the mean function based on the proposed robust estimator for the mean function, and a modified definition of *pseudo-data*. In addition, we extend this method to form a RSCB for the difference of mean functions of two populations in the same section. In Sects. 3 and 4 , the performance of the proposed methods and their robustness are demonstrated with an extensive simulation study and real data examples. Finally, we conclude our paper with discussion and conclusion. "Appendix" contains technical proofs and further simulation study for the estimation of the variance of pseudo-data.

2 Proposed methods

2.1 Model

A functional dataset can be defined as a collection of i.i.d. random samples, $\{Y_i(x)\}_{i=1}^n$, where i is the subject index, from a smooth and square integrable random function $Y(x) \in L^2$, with unknown mean function, $\mathbb{E}[Y(x)] = m(x)$. The model is

$$Y_i(x) = m(x) + \epsilon_i(x), \quad 1 \leq i \leq n$$

where $\{\epsilon_i(\cdot)\}_{i=1}^n$ are independent random noise and without loss of generality, we assume $x \in [0, 1]$. In this paper, we also assume an equally spaced dense design, i.e., each random curve $Y_i(\cdot)$ is measured at the points $X_{ij} = j/N, 1 \leq j \leq N, 1 \leq i \leq n$, where N is the number of recorded data for each curve. Then, the j-th observation for the i-th subject can be written as

$$Y_{ij} = Y_i(X_{ij}) = m\left(X_{ij}\right) + e_{ij}, \tag{1}$$

where $e_{ij} = \epsilon_i(j/N)$. Notice that if $\mathbb{E}(\epsilon_i) = 0$, then $m(\cdot)$ is the mean function in the traditional statistical sense. If we weaken this assumption, only requiring that

distribution of ϵ_i is symmetric, although we cannot guarantee the existence of $\mathbb{E}[Y(\cdot)]$, $m(\cdot)$ can be viewed as a center function of the functional data. We discuss the specific assumptions in Sect. 2.3, but, for simplicity, we will use the term *mean function* when we refer to the function $m(\cdot)$.

2.2 B-spline smoothed M-estimator for the mean function

To describe the spline function, we first introduce a sequence of equally space points, $\{t_J\}_{J=1}^{N_m}$, called interior knots which divide the interval $[0, 1]$ into $(N_m + 1)$ equal subintervals $I_J = [t_J, t_{J+1})$, $J = 0, \ldots, N_m - 1$, $I_{N_m} = [t_{N_m}, 1]$. For any positive integer p, introduce left boundary knots $t_{1-p}, \ldots, t_0$, and right boundary knots $t_{N_m+1}, \ldots, t_{N_m+p}$, satisfying $t_{1-p} = \ldots t_0 = 0 < t_1 < \cdots < t_{N_m} < 1 = t_{N_m+1} = \cdots = t_{N_m+p}$, where $t_J = J/(N_m + 1)$, $0 \leq J \leq N_m + 1$. Denote by $\mathcal{H}^{(p-2)}$ the p-th order spline space, i.e., $p - 2$ times continuously differentiable functions on $[0, 1]$ that are polynomials of degree $p - 1$ on $[t_J, t_{J+1}]$, $J = 0, \ldots, N_m$. Then $\mathcal{H}^{(p-2)} = \{\sum_{J=1-p}^{N_m} \beta_J B_J(x), \beta_J \in \mathbb{R}, x \in [0, 1]\}$, where B_J is the J-th B-spline basis of order p.

Define the *B-spline smoothed M-estimator* of the mean function by

$$\hat{m}(x) = \underset{g(\cdot)\in\mathcal{H}^{(p-2)}}{\arg\min} \sum_{i=1}^{n}\sum_{j=1}^{N} \rho\left(Y_{ij} - g\left(j/N\right)\right) = \sum_{J=1-p}^{N_m} \hat{\boldsymbol{\beta}}_J B_J(x),$$

where

$$\left\{\hat{\boldsymbol{\beta}}_{1-p}, \ldots, \hat{\boldsymbol{\beta}}_{N_m}\right\}^{\mathrm{T}} = \operatorname{argmin}_{\{\boldsymbol{\beta}_{1-p},\ldots,\boldsymbol{\beta}_{N_m}\}\in\mathbb{R}^{N_m+p}} \sum_{i=1}^{n}\sum_{j=1}^{N} \rho \times \left(Y_{ij} - \sum_{J=1-p}^{N_m} \boldsymbol{\beta}_J B_J\left(j/N\right)\right), \tag{2}$$

and ρ is a suitably chosen loss function. In this work, we focus on convex loss functions, which guarantee that Eq. (2) has a unique solution. Different choices of ρ would lead to different estimation properties of the mean function m. For example, if $\rho(x) = x^2$, we obtain the ordinary least-squares (OLS) estimator which was studied in Cao et al. (2012). If $\rho(x) = |x|$, we obtain the least absolute deviation (LAD) estimator which was studied by Lima et al. (2017). The robust properties of different choices of ρ functions have been extensively studied in the literature, see Maronna et al. (2006).

Define the function $\psi(x) = \rho'(x)$, then the estimated coefficients in (2) can also be obtained by the following system of equations

$$\sum_{i=1}^{n}\sum_{j=1}^{N} \psi\left(Y_{ij} - \sum_{J=1-p}^{N_m} \hat{\boldsymbol{\beta}}_J B_J(j/N)\right) B_k(j/N) = 0, \quad \text{for any } 1 - p \leq k \leq N_m.$$

A closed form solution of the above equation usually does not exist, but an approximate solution can be obtained using an iteratively re-weighted least-squares fitting algorithm. This algorithm is implemented in the function `rlm` from the MASS R-package, Venables and Ripley (2002). Notice that, since we assume that ρ is a convex loss function, the uniqueness of the solution is guaranteed.

2.3 Asymptotic properties of the B-spline smoothed M-estimator

In this section we will explore the asymptotic properties of the proposed B-spline smoothed M-estimator for the mean function of functional data. Before stating the first result, we need to introduce some notations. For $0 < r < 1$, denote by $\mathcal{C}^{p,r}$ the Hölder function space, that is, the space of functions with continuous derivatives up to order p, and with r-Hölder continuous p-derivative. For a real-valued function, f, denote by $\|f\|_2$ the standard L^2 space norm, that is, $\|f\|_2^2 = \int_0^1 |f(x)|^2 dx$. Similarly, for any vector $\boldsymbol{V} = (V_1, \ldots, V_k) \in \mathbb{R}^k$, let $\|\boldsymbol{V}\|_2^2 = \sum_{i=1}^k |V_i|^2$ and for a matrix $\mathbb{A}$, $\|\mathbb{A}\|_2 = \sup_{\boldsymbol{V} \neq 0} \|\mathbb{A}\boldsymbol{V}\|_2 / \|\boldsymbol{V}\|_2$. Let $\lambda_{\max}(\mathbb{A})$ and $\lambda_{\min}(\mathbb{A})$ be the largest and smallest eigenvalues of matrix $\mathbb{A}$, respectively. Note that $\|\mathbb{A}\|_2 = \lambda_{\max}(\mathbb{A})$ and if matrix $\mathbb{A}$ is non-singular, $\|\mathbb{A}^{-1}\|_2 = \lambda_{\min}(\mathbb{A})^{-1}$. Throughout this section, C denotes a uniform positive constant. We need the following assumptions for the asymptotic consistency and the asymptotic normality of the proposed estimator:

(A1) Let $p > 1$ be the order of the smoothing splines, N_m be the number of interior knots and assume that $(n/N)^{1/2} \log^{1/2}(n) \ll N_m \ll \min(n^{1/3}, n/N)$ and $N_m \log N_m \ll N$.

(A2) The function $m(x)$ satisfies $m \in \mathcal{C}^{p,r}$.

(A3) Define $\psi(x) = \rho'(x)$, and $\psi(x)$ be continuous, non-decreasing and uniformly bounded, $|\psi(x)| < C, \forall x \in \mathbb{R}$. Also, $\rho(\cdot)$ is a *convex function*.

(A4) $\mathbb{E}\psi(e_{ij}) = 0$.

(A5) There exists a bounded function $\delta(x)$ satisfying $0 < \inf_{x\in[0,1]} \delta(x) < \sup_{x\in[0,1]} \delta(x) < \infty$, such that $\left|\mathbb{E}[\psi(e_{ij} + u)] - \delta\left(j/N\right) \cdot u\right| \leq Cu^2$, where $|u| < C$.

(A6) $\mathbb{E}\left[\psi(e_{ij} + u) - \psi(e_{ij})\right]^2 \leq C|u|$, and $|\psi(u + v) - \psi(v)| < C$, for $|u| < C$, and $v \in \mathbb{R}$.

(A7) Define $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{iN})^{\mathrm{T}}$, $\boldsymbol{\psi}(\boldsymbol{e}_i) = (\psi(e_{i1}), \ldots, \psi(e_{iN}))^{\mathrm{T}}$ and $\mathbb{G}_i = \mathbb{E}\left(\boldsymbol{\psi}(\boldsymbol{e}_i) \cdot \boldsymbol{\psi}(\boldsymbol{e}_i)^{\mathrm{T}}\right)$, $1 \leq i \leq n$. Also, $\min_{1\leq i\leq n} \lambda_{\min}(\mathbb{G}_i) \geq \lambda_0$.

Assumption (A2) is standard in B-spline approximation, see Huang et al. (2004) and Cao et al. (2012) for example, and allows for arbitrarily good approximations of $m(x)$ by spline functions. Assumption (A3) guarantees the existence of the solution of the optimization problem in (2). One notable example for the loss function ρ is the Huber loss function, i.e.,

$$\rho_k(x) = \begin{cases} x^2/2, & |x| \leq k, \\ k(|x| - k/2), & |x| > k, \end{cases} \tag{3}$$

where $k > 0$ is a constant. Note that ρ_k allows it to combine much of the sensitivity of the mean-unbiased, minimum-variance estimator of the mean (OLS) and the robust-

ness of the median-unbiased estimator (LAD). The parameter k controls a trade-off between the resistance to outlying observations and efficiency of the estimator (Huber 1964). The boundedness of ψ is a technical assumption needed for the proof of the consistency of the estimator. It doesn't pose a large restriction, since most of the ψ functions chosen in practice satisfy this condition, such as the Huber loss function. Assumption (A4) states that the function ψ adds some regularity to the errors e_{ij}. Assumptions (A5) and (A6) are regularity conditions on the function ψ. Assumption (A7) is needed for the asymptotic normality of the proposed estimator. Assumptions (A4)–(A7) on the score function ψ are also standard conditions in M-estimation literature (Wei and He 2006; Tang and Cheng 2012). For error terms e_{ij} satisfying these assumptions we can cite, for example, e_{ij} following a zero mean Gaussian process or mixture Normal–Cauchy distribution. We provide more detailed examples for e_{ij} in Sect. 3.1.

2.3.1 Asymptotic consistency

Theorem 1 (Asymptotic consistency) *Under Assumptions (A1)–(A6) we have*

$$\|\hat{m} - m\|_2^2 = O_P\left(n^{-1}N_m + N_m^{-2p}\right). \tag{4}$$

Remark 1 As a consequence from Theorem 1, the L^2 distance between $\hat{m}$ and m has an order of magnitude bounded by the maximum of $n^{-1}N_m$ and N_m^{-2p}. Choosing $N_m = O(n^{1/(2p+1)})$ produces an optimal convergence rate equal to $O_P(n^{-2p/(2p+1)})$ (Stone 1985).

2.3.2 Asymptotic normality

We prove that $\hat{m}(\cdot)$ converges to a normal distribution. Besides the convergence in distribution, this theorem provides an estimator for the variance of $\hat{m}(\cdot)$, which is of fundamental importance when constructing the robust simultaneous confidence band in Sect. 2.4. Note also that the convergence provided in Theorem 2 is only point-wise; therefore, this result cannot be directly used to construct a simultaneous confidence band.

Before presenting the second theorem, we need some additional notations. Let

$$\mathbb{W}_n = nN^{-1}\sum_{j=1}^{N}\delta\left(j/N\right)\boldsymbol{B}\left(j/N\right)\boldsymbol{B}^{\mathrm{T}}\left(j/N\right), \tag{5}$$

where $\delta(\cdot)$ is defined in the Assumption (A5).

Theorem 2 (Asymptotic normality) *Under Assumptions (A1)–(A7) and* $n^{(2p-1)/(2p+1)} \gg N$, *we have*

$$\frac{\hat{m}(x) - m(x)}{\sqrt{D_n(x)}} \xrightarrow{d} N(0,1), \quad 0 \le x \le 1, \tag{6}$$

where $D_n(x) = \boldsymbol{B}(x)^{\mathrm{T}} \mathbb{W}_n^{-1} \left(\sum_{i=1}^{n} N^{-2} \mathbb{B}^{\mathrm{T}} \mathbb{G}_i \mathbb{B}\right) \mathbb{W}_n^{-1} \boldsymbol{B}(x)$, *where* $\mathbb{G}_i$ *was defined in the Assumption (A7), and* $\mathbb{B}$ *is a* $N \times (N_m + p)$ *matrix with columns* $\boldsymbol{B}(j/N)$, $j = 1, \ldots, N$.

Remark 2 By adding the condition $n^{(2p-1)/(2p+1)} \gg N$, the bias term of $\hat{m}(x)$ is asymptotically negligible. The result of Theorem 2 can be used to construct asymptotic confident intervals.

Remark 3 In practice, we first estimate error terms as $\hat{e}_{ij} = Y_{ij} - \hat{m}(j/N)$, so the $\mathbb{G}_i$ is estimated as $\widehat{\mathbb{G}}_i \equiv n^{-1} \sum_{i=1}^{n} \left(\boldsymbol{\psi}(\widehat{\boldsymbol{e}}_i) \cdot \boldsymbol{\psi}(\widehat{\boldsymbol{e}}_i)^{\mathrm{T}}\right)$. The $\delta(\cdot)$ function defined in (A5) needs to be estimated based on the chosen ρ function. For example, in the case of the Huber loss function 3, by applying using the Taylor expansion on $\psi(x)$, we find that $\delta(j/N) = P(|e_{ij}| \leq k) \approx n^{-1} \sum_{i=1}^{n} I(|e_{ij}| \leq k)$, where constant k is given in (3). The variance term can be estimated as $\hat{D}_n(x) = \boldsymbol{B}(x)^{\mathrm{T}} \mathbb{W}_n^{-1} \left(\sum_{i=1}^{n} N^{-2} \mathbb{B}^{\mathrm{T}} \widehat{\mathbb{G}}_i \mathbb{B}\right) \mathbb{W}_n^{-1} \boldsymbol{B}(x)$.

The proofs of Theorems 1 and 2 are given in the Supplementary Material.

2.4 Robust simultaneous confidence band

Obtaining SCB for *homogeneous* functional datasets has been discussed in previous literature, such as Cao et al. (2012), in which the SCB is obtained by first estimating the covariance function of the functional process, and then calculating the quantile of a Gaussian process with the same covariance structure. This procedure, though, is very sensitive to outliers, as discussed in Sect. 3. The result of Theorem 2 alone is not enough to provide a RSCB, but with the help of a modified pseudo-data transformation, we can translate the calculation of the RSCB to the simpler problem of obtaining a SCB for homogenous functional datasets.

2.4.1 Pseudo-data

The objective of this section is first to modify the contaminated dataset into a homogeneous dataset to mitigate the effect of outliers, and then the resulting homogeneous dataset is used to construct SCB for the mean function by using the classical (non-robust) method to calculate the SCB.

To accomplish the first objective, we transform each curve in the original dataset into the new homogeneous curves borrowing the idea of the *pseudo-data*, which was first introduced in Cox (1983). The concept of *pseudo-data* suggests an equivalence between a robust estimator and a more conventional least-squares method. In our study, we modify the definition of the *pseudo-data* to allow for heteroscedasticity of the random errors considered in our model. In particularly, we generate *pseudo-data* Z_{ij} based on the original dataset Y_{ij}, while Z_{ij} share the similar information as Y_{ij}, for example, both of them have the same mean functions, but Z_{ij} is free of outliers. In this work, we define the *pseudo-data* derived from the dataset Y_{ij} as

$$Z_{ij} = m(j/N) + \sqrt{2nD_n(j/N)}\,\psi\left(e_{ij}/\sqrt{2nD_n(j/N)}\right), \quad i = 1, \ldots, n, \quad j = 1, \ldots, N \tag{7}$$

where $D_n(\cdot)$ is the variance of $\hat{m}(\cdot)$ obtained from Theorem 2.

To understand the multiplication by $\sqrt{2nD_n(x)}$ in (7), a simple example is helpful. Assume that $\{Y_{ij}\}_{i,j=1}^{n,N}$ is homogeneous, i.e., errors $\{e_{ij}\}_{j=1}^{N}$ are i.i.d. and $\mathrm{Var}(e_{ij}) < \infty$. Also, consider the least-squares loss function, $\rho_{LS}(x) = x^2/2, x \in \mathbb{R}$. Then, a direct calculation results in that $D_n(x)$ is the B-spline smoothing of $\{Y_{ij}\}_{i,j=1}^{n,N}$. That is, the *pseudo-data* defined in (7) is the original homogeneous dataset, $Z_{ij} \equiv Y_{ij}$.

In the general case, with outlier contaminated datasets, and robust loss functions, the *pseudo-data* Z_{ij} has a similar behavior as the original dataset, excluding the influence of the outlier curves. For each $j = 1, \ldots, N$, the variance of Z_{ij} is close to the variance of the original dataset, after excluding the influence of the outlier curves. A simulation to support this claim is presented in "Appendix."

Recently, Lim and Oh (2015) used the concept of *pseudo-data* to obtain SCB for regression function with i.i.d. data. Here we will extend this idea to work with functional data and transform the estimation of the RSCB for the mean function of contaminated functional data to the estimation of a SCB for the mean function of homogeneous functional data.

More precisely, we will apply the estimation method for the SCB for the mean function developed in Cao et al. (2012) to the *pseudo-data* Z_{ij}. Since neither the mean function nor the random error is directly observable, we calculate the empirical pseudo-data as

$$\hat{Z}_{ij} = \hat{m}(j/N) + \sqrt{2nD_n(j/N)}\,\psi\left(\hat{e}_{ij}/\sqrt{2nD_n(j/N)}\right), \quad i = 1, \ldots, n, \quad j = 1, \ldots, N,$$

where $\hat{m}(\cdot)$ is the B-spline M-estimator of the mean function defined in Sect. 2.2 and $\hat{e}_{ij}$ is the empirical estimator of the random error defined as $\hat{e}_{ij} = Y_{ij} - \hat{m}(j/N)$. Once we have empirical pseudo-data, $\hat{Z}_{ij}$, we can easily construct the SCB for the mean function based on the non-robust SCB method by Cao et al. (2012). The method to calculate the RSCB is provided in the following algorithm.

Algorithm 1: Robust simultaneous confidence band for the mean function

Input: Empirical Pseudo-Data, $\hat{Z}_{ij}$;

1 Estimate the mean function of the empirical pseudo-data using B-spline Smoothed least-square; $\hat{m}^{pd}$;

2 Estimate the sample covariance function of empirical pseudo-data using B-spline smoothing; $\hat{G}^{pd}(x, x'), x, x' \in [0, 1]$;

3 Calculate the empirical quantile of Gaussian process with the same covariance structure with empirical pseudo-data: $\hat{Q}^{pd}_{1-\alpha}$ [the same algorithm proposed in Cao et al. (2012)];

Output: Robust simultaneous confidence band: $\hat{m}^{pd}(x) \pm \hat{Q}^{pd}_{1-\alpha} n^{-1/2} \sqrt{\hat{G}^{pd}(x, x)}, x \in [0, 1]$.

2.5 Robust simultaneous confidence band for the difference of the two mean functions

The framework proposed here to obtain a RSCB for the mean function can be extended to construct a RSCB for the difference of the mean functions of two populations. Denote $d = 1, 2$ representing the samples coming from each population, satisfying the model defined in (1) and

$$Y_{dij} = m_d\left(X_{ij}\right) + e_{dij}, \quad 1 \le i \le n_d,\ 1 \le j \le N.$$

Define the ratio of two-sample sizes as $\hat{r} = n_1/n_2$ and assume that $\lim_{n_1\to\infty} \hat{r} = r > 0$. For each group, we obtain the M-estimator for the mean function as described in Sect. 2.2.

Following the procedure in Sect. 2.4, we can obtain empirical pseudo-samples $\hat{Z}_{d,ij}$ for each population. We can then use the empirical pseudo-samples to obtain the RSCB for the difference of the mean functions. First, we obtain estimators for the covariance function of each group, $\hat{G}_d^{pd}(\cdot,\cdot)$, then compute the empirical quantile, $\hat{Q}_{12,1-\alpha}$, of a Gaussian process having covariance structure defined by

$$\frac{\hat{G}_1^{pd}(x,x') + \hat{r}\hat{G}_2^{pd}(x,x')}{\left\{\hat{G}_1^{pd}(x,x) + \hat{r}\hat{G}_2^{pd}(x,x)\right\}^{1/2}\left\{\hat{G}_1^{pd}(x',x') + \hat{r}\hat{G}_2^{pd}(x',x')\right\}^{1/2}}, \quad x, x' \in [0,1].$$

The RSCB for $m_1(x) - m_2(x)$ is then given as

$$\left(\hat{m}_1(x) - \hat{m}_2(x)\right) \pm n_1^{1/2}\left[\hat{G}_1^{pd}(x,x) + \hat{r}\hat{G}_2^{pd}(x,x)\right]^{1/2}\hat{Q}_{12,1-\alpha}.$$

The confidence band for the difference of the mean functions can be used to perform a hypothesis test of the form $H_0 : m_1(x) \equiv m_2(x),\ \forall x \in [0,1]$ versus $H_A : m_1(x) \neq m_2(x), \exists x \in [0,1]$. The test can be performed by calculating the appropriate $(1-\alpha) \times 100\%$ RSCB and checking if the horizontal line $y = 0$ is fully contained in the RSCB. Although the p-value for the test cannot be calculated directly, it can be estimated by finding the largest α for which H_0 is rejected.

3 Simulation

In this section, we perform a numerical study to analyze the finite-sample performance of methods proposed in this paper. We investigate the consistency of the B-Spline smoothed M-estimator for the mean function, and the empirical coverage and band area of the RSCB. We use the SCB (non-robust) proposed in Cao et al. (2012) as a baseline for comparison. Since outlier curves often have different types of outlying behaviors in functional data, we consider several typical types of outliers for the assessment of the performance of the RSCB.

3.1 Simulation setting

Based on the model proposed in Cao et al. (2012), we generate the functional data from

$$Y_{ij} = m(j/N) + e_{ij} \quad 1 \leq j \leq N, \ 1 \leq i \leq n,$$

where $e_{ij} = \sum_{k=1}^{2} \xi_{ik}\phi_k(j/N) + \epsilon_{ij}$, $m(x) = 10 + \sin\{2\pi(x - 1/2)\}$, $\phi_1(x) = -2\cos\{\pi(x - 1/2)\}$, $\phi_2(x) = \sin\{\pi(x - 1/2)\}$ and $\xi_{ik} \sim N(0, 1)$, $k = 1, 2$ and $\epsilon_{ij} \sim N(0, 0.25)$. The random component of the sample curves, e_{ij}, is decomposed into the between-subject variation in the functional sample and the within-subject variation. We generate the number of observations $N = \lfloor 1.22n^{(2p-2)/(2p+1)} \log(n) \rfloor$ for each sample. The number of knots is taken as $N_m = \lfloor 0.3n^{1/p} \log(n) \rfloor$, where $p = 4$ (cubic spline).

Under this functional model, we introduce outlier curves, Y_{ij}^o, to the generated functional sample by contaminating a subset, I_o, of the original functional sample. The contamination proportion varies from 0.00 to 0.20, at 0.05 increment. In addition, we consider two-heavy-tailed model error distributions to assess the performance of the proposed method. All of these settings are described as in the following.

1. *Peak outliers* To simulate an outlier with a punctual influence, each curve was contaminated at a single measurement point, j^*/N, by adding a random value s_i taken from a uniform distribution on $[-s_u, -s_l] \cup [s_l, s_u]$, that is,

$$Y_{ij^*}^o = Y_{ij^*} + s_i, \quad i \in I_o, \ \ j^* = \lfloor 0.05N \rfloor.$$

   This produces a peak outlier curve with a *peak* at the point j^*/N. The parameters s_l and s_u control the strength of outliers.
2. *Bump outliers* This type is an extension of the peak outliers and the contamination occurs in an interval, $[b_0, b_1]$, rather than at a single point, that is,

$$Y_{ij^*}^o = Y_{ij^*} + s_i, \quad i \in I_o, \, j^*/N \in [b_0, b_1].$$

   In the simulation, the interval is chosen as $[b_0, b_1] = [0.5, 0.53]$.
3. *Step outliers* A further extension of the bump outliers is created by contaminating the curve in the interval $[c_i, 1]$, where c_i is randomly chosen from $[0.5, 1]$ for each outlying curve, i.e.,

$$Y_{ij^*}^o = Y_{ij^*} + s_i, \quad i \in I_o, \ \ j^*/N \geq c_i.$$

4. *Mixture Normal–Slash* To simulate outliers with heavy-tailed distribution, we consider for the distribution of the within-subject variation a mixture of a normal distribution $N(0, 0.25)$ and a Slash distribution with location 0 and scale 0.5.
5. *Mixture Normal–Cauchy* Similar to previous item, but using a mixture of a normal distribution $N(0, 0.25)$ and a Cauchy distribution with location 0 and scale 0.5.

3.2 An illustrative example

We first illustrate the performance of the proposed RSCB using a toy example. We generate a functional sample of $n = 50$ curves from the model defined in Sect. 3.1 and contaminate the data by using all types of outliers defined previously, using a contamination proportion of 20%, $s_l = 10$ and $s_u = 20$ for the contamination type of outliers, types *1* to *3*, and using a mixture proportion of 20% for the mixture model outlier, types *4* and *5*. We construct the 95% confidence band using the proposed RSCB (black) and (red) SCB for the mean function for each scenario listed in Sect. 3.1. We also construct the SCB and RSCB for the mean function when the sample does not have outlier curves to assess the consistency of the proposed RSCB. Figure 1 depicts the effects of outlier curves on the SCB and RSCB methods.

The first graph (top left in Fig. 1) for no outlier case shows that the proposed RSCB behaves the same as the SCB when there are no outlier curves in the data. For peak, bump and step outliers (top right and second row in Fig. 1), the width of the non-robust SCB is strongly deformed around the outlier location, resulting in a non-informative SCB. The RSCB is less influenced by the outliers, resulting in a SCB with similar characteristics to the SCB for clean dataset.

For the mixture outliers (bottom row in Fig. 1), the influence of the outliers is most notable, with a strong deformation of the non-robust mean estimation and the SCB for mixture Normal–Slash and mixture Normal–Cauchy distributions. The RSCB is not much influenced by the presence of outliers, albeit a slight increase in the bandwidth when compared to the clean dataset.

This illustrative example provides evidence that when there are outlier curves in a functional dataset, estimates of the mean function and the SCB are both strongly affected badly while the proposed RSCB based on the R mean estimator performs well for different types of outlier curves and heavy-tailed error distributions.

3.3 Asymptotic consistency of the mean function estimator

To evaluate the performance of the proposed estimator for the mean function, we generated functional sample from the model in Sect. 3.1 for sample sizes $n = 50$, 100 and 200, with $s_l = 5$ and $s_u = 7$ for outlier types peak, bump and step. Each simulation is repeated 500 times.

The average L^2 distance between the true mean function and the B-Spline smoothed M-estimator was calculated. As a baseline comparison, the least-squares method used in Cao et al. (2012) was also calculated. The results are presented in Tables 1, 2 and 3.

The B-spline smoothed M-estimator has a comparable or faster convergence rate, as measured by the average L^2 distance between the estimator and the true mean function, than that of the non-robust method for all outliers, with a smaller or similar standard deviation. For localized outliers, the results are similar, but robust method shows better results for large contamination, as can be seen in Table 1, for *step* outliers, with $n = 200$ and 20% contamination proportion, the average L^2 distance between the robust estimator and the real mean function is 0.120, while the non-robust is 0.133. The standard deviation of the L^2 distance is also smaller for the robust (0.058), when

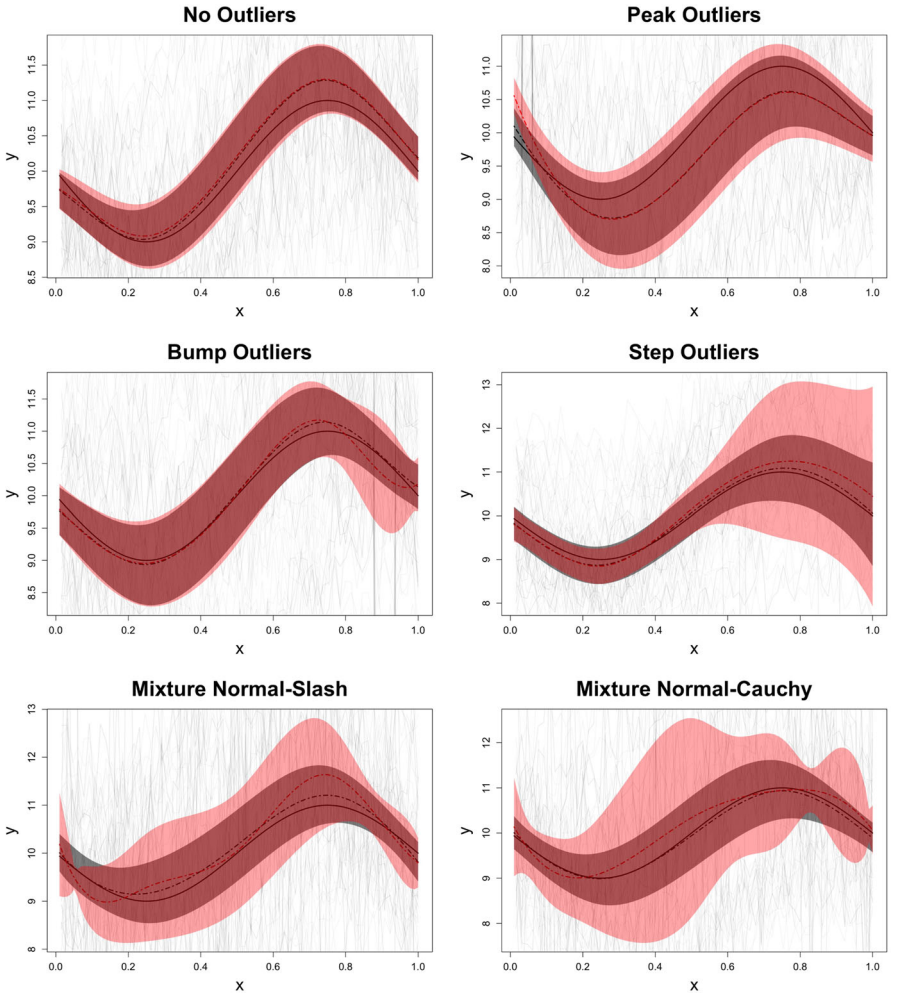

Fig. 1 Comparison between RSCB (black) and non-robust SCB (red) for a simulated dataset. Sample size $n = 50$, contamination proportion is 20% (colour figure online)

compared to the non-robust (0.063). The improvement of the robust method is made clearer for mixtures of heavy-tailed distributions, such as the mixture Normal–Slash and mixture Normal–Cauchy, when the convergence of the non-robust estimator is most influenced by the outlying curves. From Table 2, the average L^2 distance for *mixture Normal–Cauchy* outliers, with $n = 200$ and 20% contamination is 0.100, with a standard deviation of 0.060, while for the non-robust the average L^2 distance is 0.915, with a standard deviation of 2.322, an increase of approximately 9 times for the average distance, and 38 times for the standard deviation. In Table 3, notice that the results for the robust case are very similar to the results for the *clean dataset*, for which the average distance is 0.094 and the standard deviation 0.054. This further indicates

Table 1 Average (SD) of L^2 distance between the real mean function and the estimated mean function with contamination outliers

Outlier type	n	Method	Contamination prop.			
			0.05	0.10	0.15	0.20
Peak	50	R	0.200 (0.114)	0.193 (0.109)	0.193 (0.105)	0.196 (0.110)
		NR	0.198 (0.114)	0.189 (0.103)	0.192 (0.104)	0.199 (0.107)
	100	R	0.141 (0.081)	0.126 (0.076)	0.133 (0.078)	0.129 (0.072)
		NR	0.140 (0.083)	0.139 (0.079)	0.138 (0.079)	0.140 (0.076)
	200	R	0.094 (0.054)	0.099 (0.057)	0.103 (0.060)	0.098 (0.053)
		NR	0.097 (0.053)	0.097 (0.059)	0.096 (0.056)	0.092 (0.053)
Bump	50	R	0.200 (0.114)	0.195 (0.113)	0.200 (0.107)	0.200 (0.112)
		NR	0.200 (0.112)	0.197 (0.112)	0.202 (0.106)	0.202 (0.111)
	100	R	0.138 (0.081)	0.145 (0.082)	0.137 (0.076)	0.143 (0.080)
		NR	0.139 (0.081)	0.146 (0.082)	0.138 (0.076)	0.145 (0.079)
	200	R	0.096 (0.053)	0.098 (0.055)	0.099 (0.056)	0.102 (0.056)
		NR	0.096 (0.052)	0.099 (0.055)	0.100 (0.055)	0.103 (0.056)
Step	50	R	0.207 (0.113)	0.217 (0.110)	0.224 (0.114)	0.249 (0.119)
		NR	0.208 (0.109)	0.232 (0.114)	0.253 (0.127)	0.272 (0.134)
	100	R	0.146 (0.077)	0.148 (0.078)	0.166 (0.081)	0.176 (0.088)
		NR	0.153 (0.075)	0.165 (0.078)	0.174 (0.085)	0.187 (0.088)
	200	R	0.104 (0.052)	0.110 (0.055)	0.113 (0.053)	0.120 (0.058)
		NR	0.111 (0.056)	0.114 (0.052)	0.121 (0.055)	0.133 (0.063)

that the B-spline smoothed M-estimator is successfully diminishing the influence of the outliers in the estimation of the mean function.

3.4 Simulation of the SCB for the mean function and the difference of two mean functions

3.4.1 Case I: SCB for the mean function, $m(x)$

To evaluate the performance of the proposed RSCB method for the mean function, we calculate the empirical coverage rate. We generate functional samples from the model in Sect. 3.1 for sample sizes $n = 50$, 100 and 200, with $s_l = 5$ and $s_u = 7$ for outlier types *1* to *3*. Each simulation is repeated 500 times.

The empirical coverage rates for contamination proportions varying from 0.05 to 0.20 are presented in Tables 4, 5 and 6. The results for clean datasets are similar for robust and non-robust methods. When the empirical coverage for both methods approaching 95%, the area of RSCB is smaller than the area of non-robust SCB. For outlier contaminated datasets, the advantage of the RSCB becomes clear, with a breakdown point of around 20%, while the non-robust SCB has a breakdown point at 5–10%. The precision in the RSCB is also greater, with the area of the RSCB smaller

Table 2 Average (SD) of L^2 distance between the real mean function and the estimated mean function with mixture model outliers

Outlier type	n	Method	Contamination prop.			
			0.05	0.10	0.15	0.20
Mixture Normal–Slash	50	R	0.192 (0.110)	0.193 (0.105)	0.193 (0.107)	0.200 (0.109)
		NR	0.362 (0.684)	1.128 (7.103)	1.796 (13.37)	5.940 (109.9)
	100	R	0.140 (0.081)	0.141 (0.078)	0.137 (0.078)	0.144 (0.082)
		NR	0.458 (2.896)	5.060 (60.09)	6.983 (98.27)	1.510 (5.676)
	200	R	0.095 (0.054)	0.101 (0.059)	0.095 (0.057)	0.100 (0.051)
		NR	0.545 (3.406)	0.616 (1.676)	0.948 (2.655)	1.602 (10.99)
Mixture Normal–Cauchy	50	R	0.199 (0.111)	0.200 (0.116)	0.194 (0.112)	0.205 (0.111)
		NR	0.534 (3.067)	0.529 (1.213)	1.820 (13.74)	1.147 (5.073)
	100	R	0.132 (0.080)	0.140 (0.077)	0.143 (0.079)	0.143 (0.079)
		NR	0.416 (1.615)	1.561 (15.47)	0.786 (2.189)	3.621 (56.44)
	200	R	0.095 (0.057)	0.095 (0.056)	0.097 (0.054)	0.100 (0.060)
		NR	0.750 (8.707)	0.633 (3.040)	0.823 (2.279)	0.915 (2.322)

Table 3 Average (SD) of L^2 distance between the real mean function and the estimated mean function with clean data

Method	n		
	50	100	200
Clean dataset			
R	0.195 (0.109)	0.135 (0.076)	0.094 (0.054)
NR	0.198 (0.114)	0.128 (0.077)	0.098 (0.057)

than the non-robust SCB. Peak outliers produce similar empirical convergence rate for large sample size, but for $n = 50$, the advantage of the RSCB is apparent. For the less localized outliers, bump and step, the robust method has better empirical coverage and maintains the area of the RSCB reasonably constant for varying contamination levels. The non-robust method shows a quicker decay of the empirical coverage, for bump and at the same time has wider SCB. The mixture models more clearly show better results of the robust method, with the heavy-tailed mixture models, Slash and Cauchy, showing the significant advantage of the RSCB over the non-robust SCB. The precision of the RSCB is kept at the same level as all other outlier types and the clean dataset. This provides strong evidence that the proposed RSCB is less sensitive to the presence of outliers in the dataset, maintaining both a good confidence level and precision.

3.4.2 Case II: SCB for the difference of two mean functions, $m_1(x) - m_2(x)$

We also conducted a simulation to evaluate the performance of the RSCB method for the difference between two mean functions, by testing the hypotheses described in Sect. 2.5, $H_0 : m_1(x) = m_2(x),\ \forall x \in [0, 1]$ versus $H_A : m_1(x) \neq m_2(x), \exists x \in$

Table 4 Robust (R) and non-robust (NR) empirical coverage rates (average area) of 95% SCB for contamination outliers

Outlier type	n	Method	Contamination prop.			
			0.05	0.10	0.15	0.20
Peak	50	R	0.902 (1.053)	0.924 (1.045)	0.890 (1.053)	0.880 (1.049)
		NR	0.898 (1.055)	0.912 (1.047)	0.878 (1.056)	0.866 (1.054)
	100	R	0.926 (0.743)	0.934 (0.748)	0.914 (0.746)	0.934 (0.748)
		NR	0.928 (0.743)	0.934 (0.749)	0.914 (0.747)	0.936 (0.750)
	200	R	0.952 (0.530)	0.938 (0.529)	0.904 (0.528)	0.958 (0.531)
		NR	0.954 (0.530)	0.938 (0.530)	0.904 (0.528)	0.958 (0.531)
Bump	50	R	0.910 (1.047)	0.876 (1.036)	0.884 (1.054)	0.884 (1.048)
		NR	0.906 (1.050)	0.870 (1.040)	0.872 (1.059)	0.872 (1.056)
	100	R	0.898 (0.745)	0.886 (0.743)	0.908 (0.746)	0.874 (0.744)
		NR	0.894 (0.747)	0.886 (0.747)	0.916 (0.751)	0.876 (0.751)
	200	R	0.938 (0.529)	0.914 (0.529)	0.902 (0.530)	0.876 (0.529)
		NR	0.930 (0.530)	0.898 (0.531)	0.886 (0.533)	0.838 (0.533)
Step	50	R	0.888 (1.080)	0.878 (1.165)	0.892 (1.232)	0.878 (1.324)
		NR	0.906 (1.101)	0.896 (1.223)	0.874 (1.278)	0.868 (1.357)
	100	R	0.872 (0.772)	0.924 (0.832)	0.894 (0.899)	0.890 (0.943)
		NR	0.914 (0.794)	0.932 (0.873)	0.918 (0.928)	0.910 (0.965)
	200	R	0.902 (0.545)	0.910 (0.592)	0.914 (0.637)	0.908 (0.669)
		NR	0.890 (0.560)	0.926 (0.625)	0.936 (0.659)	0.928 (0.686)

Table 5 Robust (R) and non-robust (NR) empirical coverage rates (average area) of 95% SCB for mixture outliers

Outlier type	n	Method	Contamination prop.			
			0.05	0.10	0.15	0.20
Mixture Normal–Slash	50	R	0.908 (1.049)	0.888 (1.042)	0.890 (1.046)	0.822 (1.042)
		NR	0.676 (2.397)	0.470 (4.326)	0.394 (5.299)	0.338 (6.205)
	100	R	0.920 (0.745)	0.904 (0.744)	0.912 (0.740)	0.870 (0.739)
		NR	0.604 (2.805)	0.436 (10.192)	0.384 (5.332)	0.292 (7.718)
	200	R	0.944 (0.529)	0.910 (0.526)	0.918 (0.525)	0.888 (0.523)
		NR	0.490 (2.613)	0.386 (3.929)	0.300 (5.919)	0.326 (7.606)
Mixture Normal–Cauchy	50	R	0.912 (1.047)	0.856 (1.046)	0.884 (1.043)	0.862 (1.042)
		NR	0.680 (2.137)	0.556 (3.063)	0.442 (4.958)	0.348 (5.208)
	100	R	0.920 (0.744)	0.906 (0.746)	0.910 (0.740)	0.910 (0.737)
		NR	0.630 (4.167)	0.440 (5.504)	0.366 (4.401)	0.348 (18.69)
	200	R	0.912 (0.528)	0.944 (0.529)	0.926 (0.526)	0.924 (0.524)
		NR	0.494 (2.446)	0.414 (3.552)	0.304 (4.417)	0.318 (5.617)

Table 6 Robust (R) and non-robust (NR) empirical coverage rates (average area) of 95% SCB for clean data

Method	n		
	50	100	200
Clean dataset			
R	0.922 (1.046)	0.932 (0.746)	0.946 (0.529)
NR	0.906 (1.050)	0.922 (0.747)	0.922 (0.530)

[0, 1]. We employ the same model in Sect. 3.1 for the one sample case. In this simulation setup, $m_1(x) = m_2(x) = 10 + \sin\{2\pi(x - 1/2)\}, 0 \leq x \leq 1$, $n_1 = 100$ and $n_2 = 130$ correspond to the sample sizes for the first and the second population, respectively. $N = 100$ is number of measurement points for both samples, and outlier curves are introduced to the first population.

The results of the simulation are presented in Table 7 for all types of outliers. For peak outliers, the empirical type I error for the robust method is kept close to the nominal value, $\alpha = 0.05$, decreasing with the increase in the contamination proportion. For the non-robust test, the empirical type I error is much smaller than the nominal value, which indicates that the non-robust SCB is inflated. For less localized outliers, step and bump, the robust method produces an empirical type I error closer to the nominal value than the non-robust method for all contamination proportions. The mixture models with heavy-tailed distributions, mixture Normal–Slash and Normal–Cauchy highlight the most advantage of the robust method. The non-robust method has the empirical type I error 1.0, while the proposed method keeps the empirical type I errors similar to the non-contaminated dataset results.

In order to estimate the power of the RSCB method for the difference between two mean functions, we modify the mean function of the second population by adding $0.7\sin(x)$, that is, $m_2^*(x) = 10 + \sin\{2\pi(x - 1/2)\} + 0.7\sin(x)$, $0 \leq x \leq 1$. This setting is similar to the simulation performed by Cao et al. (2012). The remaining parameters are kept the same, and the first population is contaminated by the outlier curves. The results of the estimation of the empirical power are presented in Table 8. Similar to the previous results, the RSCB-based test has higher empirical power when compared to the SCB-based test, with the heavy-tailed distributions providing the highest separation between the two methods.

4 Applications

We illustrate our approach on two datasets: Octane dataset for the one sample case and ground-level ozone concentration dataset for the two-sample case.

4.1 Octane dataset

This dataset consists of 39 near infrared (NIR) spectra of gasoline sample, obtained from Esbensen et al. (1996). It is known that 6 of the samples contain added ethanol, which corresponds to an upward translation on the upper wavelength, 1390 onward,

Table 7 Empirical type I error. Nominal type I error set at $\alpha = 0.05$. Hypothesis test for the difference of mean functions. Sample size fixed at $n_1 = 100$, $n_2 = 130$, with $N = 100$ measurement points for both groups

Outlier type	Method	Contamination prop.			
		0.05	0.10	0.15	0.20
Peak	R	0.044	0.038	0.032	0.036
	NR	0.036	0.030	0.024	0.032
Bump	R	0.050	0.044	0.050	0.052
	NR	0.040	0.034	0.042	0.046
Step	R	0.064	0.050	0.066	0.098
	NR	0.044	0.060	0.082	0.096
Mixture Slash	R	0.038	0.056	0.046	0.042
	NR	1.000	1.000	1.000	1.000
Mixture Cauchy	R	0.030	0.042	0.058	0.062
	NR	1.000	1.000	1.000	1.000
Clean dataset	Method	Empirical error			
	R	0.048			
	NR	0.052			

Table 8 Empirical power. Nominal level $\alpha = 0.05$. Hypothesis test for the difference of mean functions. Sample size fixed at $n_1 = 100$, $n_2 = 130$, with $N = 100$ measurement points for both groups. Difference between the population mean functions is $0.7 \sin(x)$

Outlier type	Method	Contamination prop.			
		0.05	0.10	0.15	0.20
Peak	R	0.944	0.938	0.958	0.956
	NR	0.946	0.934	0.942	0.946
Bump	R	0.950	0.958	0.968	0.958
	NR	0.944	0.958	0.954	0.942
Step	R	0.870	0.686	0.542	0.472
	NR	0.726	0.530	0.456	0.442
Mixture Slash	R	0.950	0.942	0.972	0.944
	NR	0.000	0.000	0.000	0.000
Mixture Cauchy	R	0.958	0.942	0.958	0.952
	NR	0.000	0.000	0.000	0.000
Clean dataset	Method	Empirical error			
	R	0.958			
	NR	0.938			

interval of the spectrum. These outlying curves have a behavior similar to the step outliers described in Sect. 3.

The robust estimation of the mean and the 95% RSCB, as well as the mean estimator and confidence band using the method in Cao et al. (2012), are calculated. The results are presented in Fig. 2, showing the full spectrum measure (left panel) and magnified on the second half of the spectrum to display the differences more apparently between the non-robust and robust SCBs (right panel).

We observe that the robust mean estimator remains close to the non-outlying curves, while the non-robust estimate of the mean function is heavily influenced by the outliers,

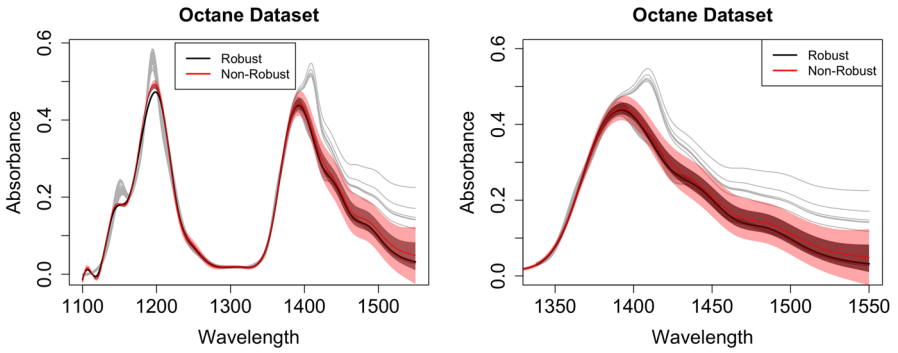

Fig. 2 95% SCB comparison for the octane dataset. non-robust (red) versus robust (black) methods. Left: full spectrum. Right: magnified on the second half of the spectrum (colour figure online)

resulting in an upward shift. The non-robust SCB is also heavily influenced by the outliers, translating in a very wide band on the second half of the spectrum. However, the proposed RSCB shows a smaller variation of the width across the spectrum.

4.2 Ground-level ozone concentration dataset

This dataset consists of hourly average measurements of ground-level ozone (O_3) concentrations from a monitoring station in Richmond, BC, Canada, from the years of 2004–2012. The presence of ozone at ground level is highly undesirable and considered a serious air pollutant. Since the concentration of ground-level ozone typically peaks at summer months, only the month of August is analyzed, resulting in 31 samples, with 24 measurement points for each sample. The same dataset was studied in Boente and Salibian-Barrera (2015). They proposed S-estimators for the principal components and examined the data for potential outliers by looking at the scores of each point on the estimated principal eigenvectors. The presence of outliers was detected in the year of 2005. For illustrative purposes, we take the ozone levels for the year of 2005 as one sample and the ozone levels for the year of 2007, which has no outlier curves (Boente and Salibian-Barrera 2015), as the other sample. The plot of the ground-level O_3 concentration for years 2005 and 2007 is presented in Fig. 3, top panel, with the year of 2005 in gray/black, and the year of 2007 in red. The outliers detected by Boente and Salibian-Barrera (2015) are highlighted.

We set up our hypotheses for testing if there is a difference between the ozone mean functions of the years 2005 and 2007 in Richmond, Canada. The outliers in the dataset are similar to the bump outliers described in Sect. 3, but they are asymmetrical, localized only in the upper portion of the dataset. The 95% SCB of the difference between the mean functions of the ground-level O_3 concentration in years of 2005 and 2007 is presented in the bottom left panel in Fig. 3. We also calculate the 95% SCB for the difference between the mean functions with the outliers kept for the RSCB, and excluding the outliers for the non-robust SCB. This is presented in the bottom right panel in Fig. 3. This plot provides a comparison of the SCB between

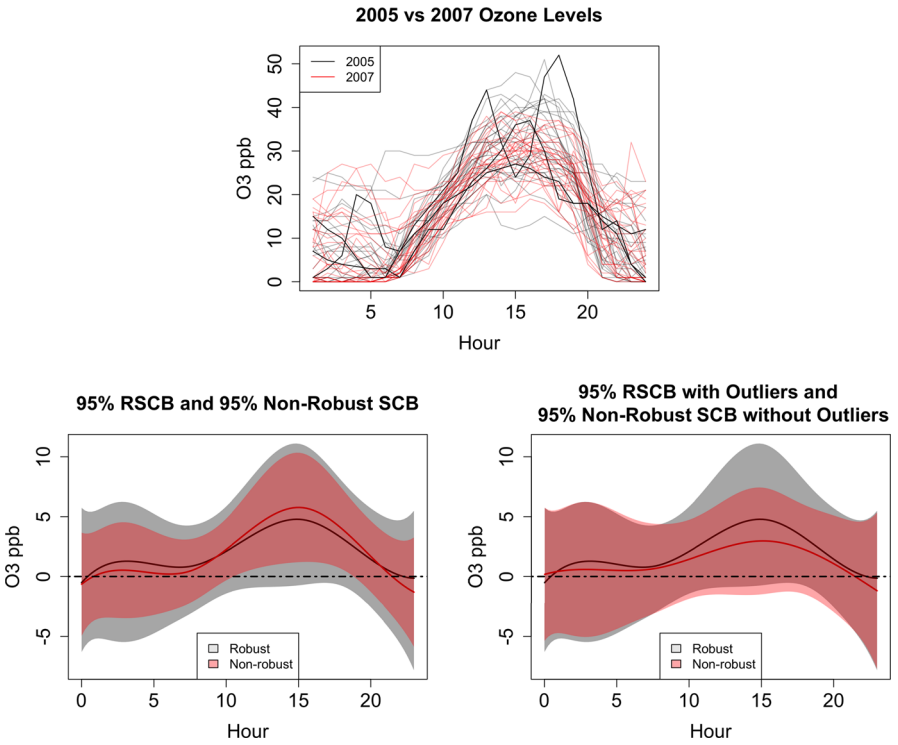

Fig. 3 Top: O_3 Levels in years of 2005 (gray and black) and 2007 (red) in Richmond. Black lines are the outliers which are determined in Boente and Salibian-Barrera (2015). Bottom left: 95% non-robust SCB (red) and RSCB (gray) for the difference between the mean functions of the 2 years. Bottom right: 95% non-robust SCB (red) and RSCB (gray) for the difference between the mean functions of the 2 years, keeping outliers for RSCB, excluding outliers for non-robust SCB (colour figure online)

the robust and non-robust methods. The non-robust SCB has a smaller width around the location of the outliers, but due to the asymmetrical disposition of the outlying curves, the estimated difference of mean functions is shifted slightly upwards by the non-robust method.

Notice that the robust method does not reject the null hypothesis at a significance level $\alpha = 0.05$, while the non-robust test rejects the null hypothesis (bottom left in Fig. 3). The result for the non-robust test is contradictory with the result of the hypothesis test using the dataset with outlier curves removed, which does not reject the null hypothesis at $\alpha = 0.05$ (bottom right in Fig. 3).

5 Discussion

In this paper we consider M-estimator and simultaneous inference for functional data observed in a dense functional design. We propose a robust way to estimate the mean function in the presence of outlier curves which is unlike the existing literature on this topic (Cao et al. 2012). It does not rely on strict assumptions about the mea-

surement error distribution. The method is applicable to contaminated as well as non-contaminated responses. We have derived the variance function for the M-based estimator of the mean function and shown the asymptotic normality of the proposed estimator. A RSCB is also proposed based on the M-estimator and *pseudo-data* (transformed data). This robust band can directly accommodate outlying curves observed on functional designs, which is the key advantage of our approach over available FDA methods. Because of this closed form for variance function and normality, we obtain more precise confidence bands based on M-estimator than the ones in Lima et al. (2016).

Numerical results show that the estimation performance of our approach is superior to existing classical approaches when the outlier curves are indeed present, and is very competitive with non-contaminated dataset. In spite of the robustness, the robust confidence band produces similar empirical convergence rate but with a narrower bandwidth compared to the classical one. Further, we applied our proposed robust estimator and the RSCB to the datasets in food science and climatology and found that the RSCB is resilient to outlier curves and can maintain type I error. An `R`-package for implementation of the proposed framework is posted publicly at: http://www.auburn.edu/~gzc0009/software.html.

Appendix: Variance of pseudo-data

In order to evaluate the efficiency of *pseudo-data* method, we compare the sample variance of the *pseudo-data* with the real variance of the uncontaminated model defined in Sect. 3.1. To further emphasize the influence of outliers in the calculation of the variance, we also compared the results with the sample variance of the outlier contaminated dataset, using the least-squares method as the estimator for the mean function. We generate a functional dataset from the model in Sect. 3.1 for sample size $n = 200$, with $s_l = 5$ and $s_u = 7$ for peak outlier case. Each simulation is repeated 500 times. The results are presented in Fig. 4. We only present the results for peak outliers, as all other cases have similar results. The results show that the variance of the *pseudo-data* is very close to the true sample variance computed from the clean dataset, while the non-robust estimation of the variance of the contaminated dataset is strongly affected by the outlier curves.

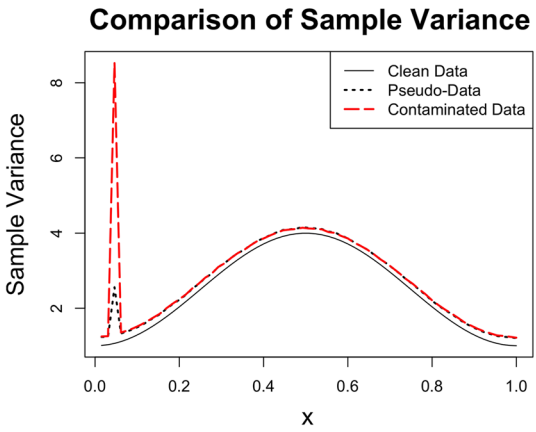

Fig. 4 Comparison of the real variance (solid), sample variance of pseudo-data (dotted) and sample variance of contaminated data (dashed). Peak outlier with 5% contamination

References

Bali, J. L., Boente, G., Tyler, D. E., Wang, J. L. (2011). Robust functional principal components: A projection-pursuit approach. *Annals of Statistics*, *39*(6), 2852–2882.

Boente, G., Salibian-Barrera, M. (2015). S-estimators for functional principal component analysis. *Journal of the American Statistical Association*, *110*(511), 1100–1111.

Cao, G., Yang, L., Todem, D. (2012). Simultaneous inference for the mean function based on dense functional data. *Journal of Nonparametric Statistics*, *24*(2), 359–377.

Cox, D. D. (1983). Asymptotics for m-type smoothing splines. *Annals of Statistics*, *11*, 530–551.

Daszykowski, M., Kaczmarek, K., Vander Heyden, Y., Walczak, B. (2007). Robust statistics in data analysis—A review: Basic concepts. *Chemometrics and Intelligent Laboratory Systems*, *85*(2), 203–219.

Embling, C. B., Illian, J., Armstrong, E., van der Kooij, J., Sharples, J., Camphuysen, K. C., Scott, B. E. (2012). Investigating fine-scale spatio-temporal predator-prey patterns in dynamic marine ecosystems: A functional data analysis approach. *Journal of Applied Ecology*, *49*(2), 481–492.

Esbensen, K., Schönkopf, S., Midtgaard, T., Guyot, D. (1996). *Multivariate analysis in practice*: *A training package*. Trondheim: Camo As.

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*(4), 331–345.

Ferraty, F. (2011). *Recent advances in functional data analysis and related topics*. Berlin: Springer.

Ferraty, F., Rabhi, A., Vieu, P. (2005). Conditional quantiles for dependent functional data with application to the climatic "el niño" phenomenon. *Sankhyā*: *The Indian Journal of Statistics*, *67*(2), 378–398.

Gervini, D. (2008). Robust functional estimation using the median and spherical principal components. *Biometrika*, *95*(3), 587–600.

Gu, L., Wang, L., Härdle, W. K., Yang, L. (2014). A simultaneous confidence corridor for varying coefficient regression with sparse functional data. *Test*, *23*(4), 806–843.

Huang, J. Z., Wu, C. O., Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, *14*, 763–788.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101.

Kraus, D., Panaretos, V. M. (2012). Dispersion operators and resistant second-order functional data analysis. *Biometrika*, *99*(4), 813–832.

Lee, S., Shin, H., Billor, N. (2013). M-type smoothing spline estimators for principal functions. *Computational Statistics & Data Analysis*, *66*, 89–100.

Lim, Y., Oh, H. S. (2015). Simultaneous confidence interval for quantile regression. *Computational Statistics*, *30*(2), 345–358.

Lima, I. R., Cao, G., Billor, N. (2017). Robust simultaneous inference for the mean function of functional data. Ph.D. dissertation. Auburn University.

Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., Cohen, K., Boente, G., Fraiman, R., Brumback, B., Croux, C. (1999). Robust principal component analysis for functional data. *Test*, *8*(1), 1–73.

Maronna, R., Martin, D., Yohai, V. (2006). *Robust statistics: Theory and methods*. Wiley series in probability and statistics. Chichester: Wiley.

Maronna, R. A., Yohai, V. J. (2013). Robust functional linear regression based on splines. *Computational Statistics & Data Analysis*, *65*, 46–55.

Shin, H., Lee, S. (2016). An RKHS approach to robust functional linear regression. *Statistica Sinica*, *26*, 255–272.

Silverman, B., Ramsay, J. (2005). *Functional data analysis* (2nd ed.). New York: Springer.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, *13*, 689–705.

Tang, Q., Cheng, L. (2012). M-estimation and b-spline approximation for varying coefficient models with longitudinal data. *Journal of Nonparametric Statistics*, *20*, 611–625.

Venables, W. N., Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). New York: Springer.

Wei, Y., He, X. (2006). Conditional growth charts. *Annals of Statistics*, *34*(5), 2069–2097.