Statistica Sinica **26** (2016), 255-272
doi:http://dx.doi.org/10.5705/ss.202014.0063

AN RKHS APPROACH TO ROBUST FUNCTIONAL LINEAR REGRESSION

Hyejin Shin and Seokho Lee

Samsung Software R&D Center and Hankuk University of Foreign Studies

Abstract: We investigate the theoretical properties of robust estimators for the regression coefficient function in functional linear regression. A robust procedure is provided in which we use outlier-resistant loss functions including non-convex loss functions. Their robust estimates are computed using an iteratively reweighted penalized least-squares algorithm. Using a pseudo data approach, we are able to show that our robust estimators also achieve the same convergence rate for prediction and estimation as the penalized least squares estimator does in the classical functional linear regression. Theoretical developments are demonstrated using numerical studies with various types of robust loss, illustrating the merit of the method.

Key words and phrases: M-type smoothing splines, outlier-resistant loss function, reproducing kernel Hilbert space, robust functional linear regression.

1. Introduction

Regression problems with functional predictors are increasingly encountered in many applications, and several statistical models and methods have been developed for them. It is frequently the case that a functional predictor is linked to a scalar response variable. In such cases, the most popular regression model for modeling their relationship is the functional linear model. This assumes that the scalar response Y depends on a square integrable random function X through the relationship

$$Y = \alpha_0 + \int_{\mathcal{T}} X(t)\beta_0(t)dt + \sigma\varepsilon, \tag{1.1}$$

where α_0 is the intercept, β_0 is a square integrable function on the compact interval $\mathcal{T}$ representing the slope function, and ε is a random error with zero mean and unit variance. Recent work on functional linear regression includes, among others, Cardot, Ferraty, and Sarda (2003), Yao, Müller, and Wang (2005), Cai and Hall (2006), Hall and Horowitz (2007), Li and Hsing (2007), Crambes, Kneip, and Sarda (2009), Yuan and Cai (2010), Cardot and Johannes (2010), and Shin and Hsing (2012).

Since the estimation of the slope function β_0 is an infinite dimensional problem, regularization through dimension reduction or penalization is necessary. The popular methods are least squares estimation with dimension reduction by functional principal component analysis (e.g., Yao, Müller, and Wang (2005); Hall and Horowitz (2007)), and penalized least squares with smoothness-inducing penalty on β_0 (e.g., see Cardot, Ferraty, and Sarda (2003); Crambes, Kneip, and Sarda (2009); Yuan and Cai (2010)). However, the majority of methods in the literature are least-squares type estimators with squared error loss function and, as a result, the presence of outliers has a serious effect on the resulting estimators. There have been some proposals to address robustness in the functional linear regression by adopting outlier-resistant loss functions. Maronna and Yohai (2011) proposed a robust version of smoothing spline estimator based on MM estimation with the biweight loss function. Gervini (2012) proposed a GM estimation for FPCA-based functional linear regression, where both the predictor and response variables are random functions, by considering an outlier-resistant loss function corresponding to t-distribution. Maronna and Yohai (2011) and Gervini (2012) suggested robust procedures and demonstrated robust properties under numerical studies, but asymptotic properties of their estimators were not studied. Yuan and Cai (2010) studied a general form of the estimator for β_0 with any convex loss function by assuming that β_0 resides in a reproducing kernel Hilbert space (RKHS). Although their estimator includes M-type estimators, their theoretical work did not go beyond the least-squares type estimator. Our goal here is to extend the applicability of the RKHS approach to robust functional linear regression by adopting an outlier-resistant loss function.

Suppose we observe data (x_i, y_i), $1 \le i \le n$, consisting of n independent copies of (X, Y) in the model (1.1). Suppose that β_0 is in a Hilbert space $\mathcal{H}$. For estimating α_0 and β_0, consider the general problem

$$\min_{\alpha\in\mathbb{R},\beta\in\mathcal{H}} \left[\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{y_i - \alpha - \int_{\mathcal{T}} x_i(t)\beta(t)dt}{\hat{\sigma}} \right) + \lambda J(\beta) \right], \tag{1.2}$$

where ρ is a loss function, $\hat{\sigma}$ is a preliminary scale estimate of errors, $J(\beta)$ is a penalty functional on β, and $\lambda > 0$ is a regularization parameter. Most penalized least-squares approaches to functional linear regression take $\mathcal{H} = W_2^m = \{\beta : \beta, \beta^{(1)}, \ldots, \beta^{(m-1)}$ are absolutely continuous and $\beta^{(m)} \in \mathcal{L}_2\}$ and $J(\beta) = \int_{\mathcal{T}} [\beta^{(m)}(t)]^2 dt$. In that case, if $\rho(r) = r^2$, the solution to (1.2) is the smoothing spline estimator for functional linear regression (Crambes, Kneip, and Sarda (2009)). However, this solution is known to be highly sensitive to outlying observations. Thus, it is natural to consider an outlier-resistant loss function in order to robustify the estimators of α_0 and β_0. By replacing the squared-error loss by a non-convex ρ-function, Maronna and Yohai (2011) proposed a robust version of the smoothing spline estimator. Further, Cardot, Crambes, and Sarda (2005) studied quantile regression by considering the L_1-type loss function defining quantiles of the regression. For a convex ρ-function, Yuan and Cai (2010) derived the explicit form of the minimizer over f of (1.2) by the representer theorem (Kimeldorf and Wahba (1971)). They then proposed the penalized least squares estimators for α_0 and β_0 associated with $\rho(r) = r^2$ and focused on their asymptotic properties. In this paper, we extend the scope of the approach in Yuan and Cai (2010), practically and theoretically, by considering outlier-resistant loss functions that are not necessarily convex. Specifically, we show that our robust estimators achieve the same convergence rate for prediction and estimation as do the least-squares type estimators in the regular functional linear regression. To the best of our knowledge, this is the first work that provides a theoretical background for robust functional linear regression.

The remainder of the paper is organized as follows. Section 2 introduces M-type smoothing spline estimators for functional linear regression and its estimating algorithm, and Section 3 investigates the asymptotic properties of the proposed estimator. Sections 4 and 5 then provides simulation studies and a data example to demonstrate the performance of the proposed method. Proofs of the main results in Section 3 are provided in the online supplementary note.

2. Robust Functional Linear Regression

As (1.1) we take the slope function β_0 to be in an RKHS $\mathcal{H}$, which is a subspace of the Hilbert space of square integrable functions on $\mathcal{T}$, and suppose X satisfies $E\left(\int_{\mathcal{T}} |X(t)|^2 dt\right) < \infty$. Suppose that $J(\beta) = \|P_1\beta\|_{\mathcal{H}}^2$, where P_1 is the orthogonal projection of β in $\mathcal{H}$ onto a subspace $\mathcal{H}_1$, and $\mathcal{H}$ has a decomposition $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0 = \{\beta \in \mathcal{H} : J(\beta) = 0\}$ is a finite dimensional linear subspace of $\mathcal{H}$ with $\dim(\mathcal{H}_0) = L \leq n$.

Let K be the reproducing kernel of $\mathcal{H}$ and K_1 the reproducing kernel of $\mathcal{H}_1$. If K is continuous and square integrable on $\mathcal{T} \times \mathcal{T}$, then $\eta_i(t) := \int_{\mathcal{T}} x_i(u)K(u,t)du$ are in $\mathcal{H}$, $i = 1, \ldots, n$ (Cucker and Smale (2001)). Since $\beta(u) = \langle \beta(\cdot), K(u,\cdot)\rangle_{\mathcal{H}}$ by the reproducing property, the penalized least squares (PLS) criterion in (1.2) is

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{y_i - \alpha - \langle \eta_i, \beta\rangle_{\mathcal{H}}}{\hat{\sigma}}\right) + \lambda\|P_1\beta\|_{\mathcal{H}}^2. \tag{2.1}$$

Using the representer theorem, it can be shown that the minimizer over β of (2.1) has the form

$$\beta_\lambda(t) = \sum_{l=1}^{L} d_l\theta_l(t) + \sum_{i=1}^{n} c_i\xi_i(t), \tag{2.2}$$

where the $\theta_l, 1 \le l \le L$, are an orthonormal basis of $\mathcal{H}_0$ and $\xi_i(t) = P_1\eta_i(t) = \int_{\mathcal{T}} x_i(u)K_1(u,t)du$. This is because for $\beta = \beta_\lambda + \varrho$ with ϱ an element in $\mathcal{H}_1$ perpendicular to $\xi_1, \ldots, \xi_n$, we observe $\langle \eta_i, \beta\rangle_{\mathcal{H}} = \langle \eta_i, \beta_\lambda\rangle_{\mathcal{H}}$ and $\|P_1\beta\|^2_{\mathcal{H}} = \|P_1\beta_\lambda\|^2_{\mathcal{H}} + \|\varrho\|^2_{\mathcal{H}}$. Thus, (2.1) is

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{y_i - \alpha - \langle \eta_i, \beta_\lambda\rangle_{\mathcal{H}}}{\hat{\sigma}}\right) + \lambda(\|P_1\beta_\lambda\|^2_{\mathcal{H}} + \|\varrho\|^2_{\mathcal{H}}),$$

which is minimized when $\varrho = 0$, and its minimizer over β_0 has the form (2.2). Since $(P_1\beta_\lambda)(\cdot) = \sum_{i=1}^n c_i\xi_i(\cdot)$ and $\|P_1\beta_\lambda\|^2_{\mathcal{H}} = \sum_{i=1}^n\sum_{j=1}^n c_ic_j\langle \xi_i, \xi_j\rangle_{\mathcal{H}}$, minimizing (2.1) is reduced to

$$\begin{aligned}\min_{\alpha,\boldsymbol{d},\boldsymbol{c}} \Bigg[\frac{1}{n}\sum_{i=1}^{n} \rho\Big(&\frac{y_i - \alpha - \sum_{l=1}^{L} d_l \int_{\mathcal{T}} x_i(t)\theta_l(t)dt - \sum_{j=1}^{n} c_j\langle \xi_i, \xi_j\rangle_{\mathcal{H}}}{\hat{\sigma}}\Big) \\ &+\lambda \sum_{i=1}^{n}\sum_{j=1}^{n} c_ic_j\langle \xi_i, \xi_j\rangle_{\mathcal{H}}\Bigg]\end{aligned} \tag{2.3}$$

with $\boldsymbol{d} = (d_1, \ldots, d_L)^T$ and $\boldsymbol{c} = (c_1, \ldots, c_n)^T$.

For illustration, suppose that $\mathcal{H} = W_2^m[0,1]$. If we use the inner product

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{k=0}^{m-1} f^{(k)}(0)g^{(k)}(0) + \int_0^1 f^{(m)}(t)g^{(m)}(t)dt,$$

then the reproducing kernel of $\mathcal{H}$ is

$$K(s,t) = \sum_{k=0}^{m-1} \frac{s^kt^k}{(k!)^2} + \int_0^1 \frac{(s-u)_+^{m-1}(t-u)_+^{m-1}}{\{(m-1)!\}^2}du$$

with $u_+ = \max(u, 0)$. If P_1 is the orthogonal projection onto the subspace $\mathcal{H}_1 = \{f \in \mathcal{H} : f^{(k)}(0) = 0,\ 0 \le k \le m-1\}$, then $J(f) = \|P_1 f\|^2_{\mathcal{H}} = \int_0^1 [f^{(m)}(t)]^2dt$. The reproducing kernel of $\mathcal{H}_1$ is $K_1(s,t) = \{(m-1)!\}^{-2}\int_0^1 (s-u)_+^{m-1}(t-u)_+^{m-1}du$ and $\theta_k(t) = t^{k-1}/(k-1)!, 1 \le k \le m$, are an orthonormal basis of $\mathcal{H}_0$. In case $m = 2$, $\beta_\lambda(t) = d_1 + d_2t + \sum_{i=1}^n c_i \int_{\mathcal{T}} x_i(s)K_1(s,t)ds$ with $\theta_1(t) = 1, \theta_2(t) = t$ and $K_1(s,t) = \int_0^1 (s-u)_+(t-u)_+du$.

If ρ is differentiable, there is a variational equation for obtaining a minimizer $\hat{\beta}_{n\lambda}$ over β of (2.1).

Proposition 1. *Suppose that $\psi = \rho'$ exists everywhere. Then, a minimizer $\hat{\beta}_{n\lambda}$ of (2.1) satisfies*

$$-\frac{1}{n}\sum_{i=1}^{n} \eta_i\psi\left(\frac{y_i - \alpha - \langle \eta_i, \beta\rangle_{\mathcal{H}}}{\hat{\sigma}}\right)\frac{1}{\hat{\sigma}} + 2\lambda P_1\beta = 0. \tag{2.4}$$

We now present the estimation algorithm with details. Taking derivatives (2.3) with respect to α, $\boldsymbol{c}$, and $\boldsymbol{d}$ and setting them equal to 0, it can be shown that the solution to the resulting estimating equations is the minimizer of the penalized weighted least-squares criterion

$$\frac{1}{n}\sum_{i=1}^{n} w_i\Big(y_i - \alpha - \sum_{l=1}^{L} d_l \int_{\mathcal{T}} x_i(t)\theta_l(t)dt - \sum_{j=1}^{n} c_j\langle \xi_i, \xi_j\rangle_{\mathcal{H}}\Big)^2 + 2\lambda \boldsymbol{c}^T\Sigma\boldsymbol{c}, \tag{2.5}$$

where $\Sigma = \{\Sigma_{ij}\}$ is an $n\times n$ matrix with $\Sigma_{ij} = \langle \xi_i, \xi_j\rangle_{\mathcal{H}} = \int_{\mathcal{T}}\int_{\mathcal{T}} x_i(s)K_1(s,t)x_j(t)dsdt$ and

$$w_i = \frac{1}{\hat{\sigma}^2} w\left(\frac{y_i - \alpha - \sum_{l=1}^{L} d_l \int_{\mathcal{T}} x_i(t)\theta_l(t)dt - \sum_{j=1}^{n} c_j\langle \xi_i, \xi_j\rangle_{\mathcal{H}}}{\hat{\sigma}}\right)$$

with $w(r) = \psi(r)/r$. Letting $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $T = \{T_{il}\}$ an $n \times L$ matrix with $T_{il} = \int_{\mathcal{T}} x_i(t)\theta_l(t)dt$, and $W = \mathrm{diag}(w_1, \ldots, w_n)$, (2.5) is written in the matrix form

$$\frac{1}{n}(\boldsymbol{y} - \alpha\mathbf{1} - T\boldsymbol{d} - \Sigma\boldsymbol{c})^T W(\boldsymbol{y} - \alpha\mathbf{1} - T\boldsymbol{d} - \Sigma\boldsymbol{c}) + 2\lambda\boldsymbol{c}^T\Sigma\boldsymbol{c}. \tag{2.6}$$

Letting $Z = [\mathbf{1}, T]$ and $\boldsymbol{b} = (\alpha, d_1, \ldots, d_L)^T$, the minimizer of (2.6) is given by

$$\begin{aligned} \widehat{\boldsymbol{b}} &= (Z^TM^{-1}Z)^{-1}Z^TM^{-1}\boldsymbol{y}, \\ \widehat{\boldsymbol{c}} &= M^{-1}(I_n - Z(Z^TM^{-1}Z)^{-1}Z^TM^{-1})\boldsymbol{y} \end{aligned} \tag{2.7}$$

with $M = \Sigma + 2n\lambda W^{-1}$. Here the matrix M is not well defined when $w_i = 0$ for some i. This can happen with a loss function (e.g., biweight loss) whose ψ function takes 0 value for some domain region. In such case we can easily show that the minimizer of (2.6) is

$$\begin{aligned} \widehat{\boldsymbol{b}} &= (Z_2^TM_2^{-1}Z_2)^{-1}Z_2^TM_2^{-1}\boldsymbol{y}_2, \\ \widehat{\boldsymbol{c}}_2 &= M_2^{-1}(I_{n_2} - Z_2(Z_2^TM_2^{-1}Z_2)^{-1}Z_2^TM_2^{-1})\boldsymbol{y}_2, \\ \widehat{\boldsymbol{c}}_1 &= \mathbf{0}_{n_1} \end{aligned} \tag{2.8}$$

with $M_2 = \Sigma_{22} + 2n\lambda W_2^{-1}$. Here, $n_1 = \#\{i : w_i = 0\}$ and $n_2 = \#\{i : w_i \neq 0\}$ with $n = n_1 + n_2$, $\boldsymbol{y}_2$, $\boldsymbol{c}_2$, W_2, Z_2, Σ_{22} are redefined appropriately after removing from $\mathbf{y}$, $\boldsymbol{c}$, W, Z, Σ the rows and rows/columns corresponding to $\{i : w_i = 0\}$, and $\boldsymbol{c}_1$ is the sub-vector of $\boldsymbol{c}$, having the entries corresponding $\{i : w_i = 0\}$.

The minimizer of (2.5) is obtained by an iteratively reweighted least squares (IRWLS) procedure. If ρ is a convex loss function having monotone ψ, then (2.5) has a unique minimum. However, when non-convex loss functions (e.g., biweight loss or t loss) are used, the objective function (2.1) is non-convex and may have multiple local minima. Consequently, when ρ is non-convex, it is important to start the IRWLS algorithm with a robust, consistent initial fit. In our implementation, we consider L_1 loss function with $\rho(r) = |r|$ at (2.1) and use the resulting quantile regression fit as an initial fit for the iterative algorithm. L_1 estimation does not require the scale parameter estimation and its estimators have high breakdown point (Maronna, Martin, and Yohai (2006)).

The optimization problem (1.2) involves an auxiliary scale estimate $\hat{\sigma}$ that is required to make the estimate $\hat{\beta}_{n\lambda}$ scale invariant. The preliminary scale estimate can be computed using the residuals r_i^0 from the initial L_1 fit. Since the robustness properties of M-type estimators depend on the robustness of the auxiliary scale estimator, we consider a robust scale estimate. A popular robust scale estimator is an M-estimator. Given the residuals r_i^0, an M-scale estimate $\tilde{\sigma}$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{r_i^0}{\tilde{\sigma}}\right) = \delta \tag{2.9}$$

with $\delta \in (0,1)$, where ρ_0 is a loss function. A frequently used scale estimate is obtained when ρ_0 is the biweight function with $k = 1$ and $\delta = 0.5$. To get a consistent scale estimate at the normal distribution, we use the M-scale estimate as $\hat{\sigma} = \tilde{\sigma}/1.56$. Another simple and robust choice of an M-scale estimate is the normalized median absolute deviation (MAD), $\hat{\sigma} = \mathrm{med}_i(|r_i^0|)/0.6745$. See Maronna, Martin, and Yohai (2006) for more details.

The choice of the smoothing parameter is crucial in performance of the regularized estimators for most smoothing methods. Commonly used practical strategies for choosing the smoothing parameter are cross validation (CV) and generalized cross validation (GCV). Since leave-one-out CV or k-fold CV are computationally burdensome, we propose to use GCV as follows. Based on the fact that the fitted value is a linear predictor of the response as $\hat{\boldsymbol{y}} = H_\lambda \boldsymbol{y}$, we select the smoothing parameter λ as a minimizer of the weighted version of GCV score:

$$\mathrm{GCV}(\lambda) = \frac{1}{n}\frac{(\hat{\boldsymbol{y}}-\boldsymbol{y})^T W(\hat{\boldsymbol{y}}-\boldsymbol{y})}{\{1-\mathrm{tr}(H_\lambda)/n\}^2}, \tag{2.10}$$

where the hat matrix H_λ has the form of $H_\lambda = \{\Sigma + 2n\lambda W^{-1}M^{-1}Z(Z^TM^{-1}Z)^{-1} Z^T\}M^{-1}$. If some $w_i = 0$ exist with the biweight ρ-function, the hat matrix is modified as $H_\lambda = \{\Sigma_{22} + 2n\lambda W_2^{-1}M_2^{-1}Z_2(Z_2^TM_2^{-1}Z_2)^{-1}Z_2^T\}M_2^{-1}$ with notations defined as in (2.8). We tested CV and GCV with extensive simulations, and found little difference between them. We prefer GCV and use it for the smoothing parameter selection in simulations.

3. Asymptotic Properties

In this section, we show that the asymptotic properties of the penalized least squares estimator for β_0 (e.g., Crambes, Kneip, and Sarda (2009); Yuan and Cai

(2010)), hold for penalized M-estimators for β_0. For simplicity, we assume that $EX(\cdot) = 0$ and $EY = 0$. Then, β_0 can be estimated by

$$\hat{\beta}_{n\lambda} = \arg\min_{\beta \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} \rho\Big(\frac{y_i - \langle \eta_i, \beta \rangle_{\mathcal{H}}}{\hat{\sigma}} \Big) + \lambda \|P_1 \beta\|_{\mathcal{H}}^2 \right].$$

All our results are applicable to the more general setting when $EX(\cdot) \neq 0$ and $EY \neq 0$.

In nonparametric regression, Cox (1983) showed that the asymptotic properties of the least squares smoothing spline hold for general M-type smoothing splines. He tackled this by approximating a nonlinear M-type smoothing spline by a linear smoothing spline acting on some unobservable pseudo data. For functional linear regression, we use a similar approach. For this, we define pseudo data

$$\tilde{y}_i = \int_{\mathcal{T}} x_i(t)\beta_0(t)dt + \sigma \frac{\psi(\varepsilon_i)}{E\psi'} = \langle \eta_i, \beta_0 \rangle_{\mathcal{H}} + \sigma \frac{\psi(\varepsilon_i)}{E\psi'}$$

and let $\tilde{\beta}_{n\lambda}$ be the minimizer of

$$\frac{1}{n} \sum_{i=1}^{n} (\tilde{y}_i - \langle \eta_i, \beta \rangle_{\mathcal{H}})^2 + 2 \frac{\lambda \sigma^2}{E\psi'} \|P_1 \beta\|_{\mathcal{H}}^2. \tag{3.1}$$

Now define operators on $S_n = \mathrm{span}\{\eta_1, \ldots, \eta_n\}$ by

$$\Phi_{n\lambda}(\beta, \sigma) = -\frac{1}{n} \sum_{i=1}^{n} \eta_i \psi\left(\frac{y_i - \langle \eta_i, \beta \rangle_{\mathcal{H}}}{\sigma} \right) \frac{1}{\sigma} + 2\lambda P_1 \beta,$$

and

$$\Psi_{n\lambda}\beta = -\frac{1}{n} \sum_{i=1}^{n} \tilde{y}_i \eta_i + \mathcal{G}_{n\lambda}\beta$$

with

$$\mathcal{G}_{n\lambda}\beta = \frac{1}{n} \sum_{i=1}^{n} \langle \eta_i, \beta \rangle_{\mathcal{H}} \eta_i + 2(\frac{\lambda \sigma^2}{E\psi'}) P_1 \beta.$$

An estimator $\hat{\beta}_{n\lambda}$ is a solution of $\Phi_{n\lambda}(\beta, \hat{\sigma}) = 0$ from Proposition 1, and $\tilde{\beta}_{n\lambda}$ is the solution of $\Psi_{n\lambda}\beta = 0$. Since $\Psi_{n\lambda}\tilde{\beta}_{n\lambda} = 0$, equivalently, $\mathcal{G}_{n\lambda}\tilde{\beta}_{n\lambda} = n^{-1} \sum_{i=1}^{n} \tilde{y}_i \eta_i$, we have $\tilde{\beta}_{n\lambda} = \mathcal{G}_{n\lambda}^{-1} \left(n^{-1} \sum_{i=1}^{n} \tilde{y}_i \eta_i \right)$. It can be shown that $\mathcal{G}_{n\lambda}$ is invertible as in Cox (1983). Here $\mathcal{G}_{n\lambda}\beta = 0$ is the equation for obtaining a least squares smoothing spline for the regression coefficient function β_0 with identically zero $\tilde{y}_i$'s, and its unique solution is $\beta = 0$.

We need assumptions for our theoretical development.

(A1) The random errors ε_i are independent of the covariates x_i.

(A2) $\psi = \rho'$ satisfies (i) $\psi \in C^2(-\infty, \infty)$, (ii) $\sup_t |\psi^{(j)}(t)| < \infty, j = 0, 1, 2$, (iii) $E\psi = 0, E\psi' \neq 0, \mathrm{Var}(\psi^{(j)}) < \infty, j = 0, 1$.

(A3) $\hat{\sigma} - \sigma = O_p(n^{-1/2})$.

(A4) The eigenvalues κ_k of the reproducing kernel K_1 of $\mathcal{H}_1$ satisfy $\kappa_k \sim k^{-2r}$ for $r > 1/2$.

(A5) The eigenvalues π_k of the covariance operator Γ for X satisfy $\pi_k \sim k^{-2s}$ for $s > 1/2$.

(A6) $\lambda \sim n^{-(2r+2s)/(2r+2s+1)}$.

(A7) For any square integrable function f, there exists some constant $C \geq 1$ such that

$$E\left(\int_{\mathcal{T}} f(t)X(t)dt\right)^4 \leq C\left\{E\left(\int_{\mathcal{T}} f(t)X(t)dt\right)^2\right\}^2.$$

(A8) $\gamma_k \sim \pi_k\kappa_k$ with $\nu_k = (1+\gamma_k^{-1})^{-1}$ the kth largest eigenvalue of $R^{1/2}\Gamma R^{1/2}$, where Γ is the covariance operator associated with the covariance function Γ of the process X and R is the operator associated with the reproducing kernel R of an RKHS with the norm

$$\|f\|_R^2 = \int_{\mathcal{T}}\int_{\mathcal{T}} f(s)\Gamma(s,t)f(t)dsdt + J(f).$$

We use the same notation for a nonnegative bivariate function and an integral operator with kernel having that function: $(Rf)(\cdot) = \int_{\mathcal{T}} R(\cdot,t)f(t)dt$.

The assumption (A2) is common in general M-type smoothing splines as in, for example, Cox (1983) and Cunningham, Eubank, and Hsing (1991). A special case of (A4) has $\mathcal{H} = W_2^m[0,1]$ and $\mathcal{H}_1 = \{f \in \mathcal{H} : f^{(k)}(0) = 0,\ 0 \leq k \leq m-1\}$, so $K_1(s,t) = [(m-1)!]^{-2}\int_0^1 (s-u)_+^{m-1}(t-u)_+^{m-1}du$ and $\kappa_k \sim k^{-2m}$. If X is Gaussian then $\int_{\mathcal{T}} f(t)X(t)dt$ is a normal random variable and (A7) follows; the constant C in (A7) is at least 1 since $\left\{E\left(\int_{\mathcal{T}} f(t)X(t)dt\right)^2\right\}^2 \leq E\left(\int_{\mathcal{T}} f(t)X(t)dt\right)^4$ by Lyapunov's inequality. Assumptions (A4)$-$(A8) are borrowed from Yuan and Cai (2010); if one uses the theoretical results in Crambes, Kneip, and Sarda (2009) instead of Yuan and Cai (2010), (A4)$-$(A8) need to be replaced by the corresponding ones there.

For $\|f\|_\Gamma^2 = \int_{\mathcal{T}}\int_{\mathcal{T}} f(s)\Gamma(s,t)f(t)dsdt$ with the covariance function Γ of X, the M-type smoothing spline estimator $\hat{\beta}_{n\lambda}$ behaves similarly to the least squares smoothing spline estimator $\tilde{\beta}_{n\lambda}$ based on pseudo data.

Theorem 1. *Let* $C_n = E\|\tilde{\beta}_{n\lambda} - \beta_0\|_\Gamma^2$. *If* (A1)$-$(A8) *hold, for any* $\delta > 0$ *and some constant* $M > 0$*, there is an* n_0 *such that for all* $n \geq n_0$*,*

$$P[\textit{there is a solution } \hat{\beta}_{n\lambda} \textit{ to } \Phi_{n\lambda}(\beta,\hat{\sigma}) = 0$$

satisfying

$$\|\hat{\beta}_{n\lambda} - \tilde{\beta}_{n\lambda}\|_\Gamma^2 \le \delta^{-2} M C_n] \ge 1 - \delta.$$

Thus, with high probability $\hat{\beta}_{n\lambda}$ and $\tilde{\beta}_{n\lambda}$ are much closer than $\tilde{\beta}_{n\lambda}$ and β_0, so our robust estimator $\hat{\beta}_{n\lambda}$ enjoys the same asymptotics as the least squares estimator $\tilde{\beta}_{n\lambda}$. Since the $v_i = \sigma\psi(\varepsilon_i)/E\psi'$ have zero mean and constant variance, it follows from Yuan and Cai (2010) that, under (A4)-(A8),

$$C_n = E\|\tilde{\beta}_{n\lambda} - \beta_0\|_\Gamma^2 = O(n^{-(2r+2s)/(2r+2s+1)}). \tag{3.2}$$

If there is a unique solution to $\Phi_{n\lambda}(f, \hat{\sigma}) = 0$, then

$$\|\hat{\beta}_{n\lambda} - \beta_0\|_\Gamma^2 \le 2\|\hat{\beta}_{n\lambda} - \tilde{\beta}_{n\lambda}\|_\Gamma^2 + 2\|\tilde{\beta}_{n\lambda} - \beta_0\|_\Gamma^2,$$

with $\|\hat{\beta}_{n\lambda} - \tilde{\beta}_{n\lambda}\|_\Gamma^2 = O_p(C_n)$ from Theorem 1 and $\|\tilde{\beta}_{n\lambda} - \beta_0\|_\Gamma^2 = O_p(C_n)$ from (3.2), gives the following theorem.

Theorem 2. *If* $\psi' > 0$ *and* (A1)$-$(A8) *hold,*

$$\|\hat{\beta}_{n\lambda} - \beta_0\|_\Gamma^2 = O_p(n^{-(2r+2s)/(2r+2s+1)}).$$

The condition $\psi' > 0$ here ensures that $\Phi_{n\lambda}(f, \hat{\sigma}) = 0$ has a unique solution, though it is not necessary to ensure uniqueness. For example, Huber's ψ is not strictly increasing, but the corresponding estimate is unique unless there is a large gap in the middle of the data. When there are multiple solutions to (2.4) when, for example, the loss function ρ is non-convex (e.g., biweight loss or Cauchy loss), Theorem 2 remains valid for a solution of (2.4) that is close enough to $\tilde{\beta}_{n\lambda}$. Thus, the initial value plays a crucial role in the IRWLS algorithm to get the estimator $\hat{\beta}_{n\lambda}$ which shares the asymptotic properties of the penalized least squares estimator $\tilde{\beta}_{n\lambda}$; with a robust and consistent initial fit, we could get a solution $\hat{\beta}_{n\lambda}$ sufficiently close to $\tilde{\beta}_{n\lambda}$ so that Theorem 2 holds.

Here $\|\hat{\beta}_{n\lambda} - \beta_0\|_\Gamma^2$ measures the prediction error for any random function X^* possessing the same distribution as X and independent of $x_1, \ldots, x_n$, since

$$\|\hat{\beta}_{n\lambda} - \beta_0\|_\Gamma^2 = E\Big[\Big(\int_{\mathcal{T}} \hat{\beta}_{n\lambda}(t) X^*(t)dt - \int_{\mathcal{T}} \beta_0(t) X^*(t)dt\Big)^2 \Big| x_i, y_i,\ 1 \le i \le n\Big].$$

Remark 1. One can obtain a similar rate to the smoothing spline estimators in Crambes, Kneip, and Sarda (2009) under slightly different assumptions. Using those results, $C_n = O(\lambda + (n\lambda^{1/(2m+2q+1)})^{-1})$ when the penalty functional $J(\beta) = \int_{\mathcal{T}} [\beta^{(m)}(t)]^2 dt$ and q quantifies the decaying rate of the eigenvalues of the covariance function Γ by $\sum_{r=k+1}^{\infty} \pi_r = O(k^{-2q})$. Here q is related to s by $2q = 2s - 1$. The values q and s explain the structure of the distribution of X. On the other hand, m and r explain the smoothness of the regression coefficient function β_0. Thus, the convergence rate for M-type smoothing spline estimators in functional linear regression model depends on the smoothness of the sample path of X and the regression coefficient function β_0 as the least squares smoothing spline estimator does.

Write $\|f\|^2 = \int_{\mathcal{T}}[f(t)]^2 dt$ as the standard norm for $\mathcal{L}_2$. Let ϕ_k be the eigenfunction of the covariance operator Γ corresponding to its eigenvalue π_k, and φ_k be the eigenfunction of the reproducing kernel K_1 corresponding to its eigenvalue κ_k. When Γ and K_1 have the same set of eigenfunctions, we can derive the convergence rate for the estimation error of an estimator for β_0, as follows.

Theorem 3. *Assume that $\phi_k = \varphi_k$ for all $k \geq 1$. If $2r > 2s+1$, then under the assumptions of Theorem 2, $\|\hat{\beta}_{n\lambda} - \beta_0\|^2 = O_p(n^{-2r/(2r+2s+1)})$.*

As mentioned in Yuan and Cai (2010), ϕ_k and φ_k differ in general, but they are the same when Γ and K_1 share a common set of eigenfunctions. The setting $\phi_k = \varphi_k$, $k = 1, 2, \ldots$, is commonly adopted in FPCA-based functional linear regression (Cai and Hall (2006); Hall and Horowitz (2007)). The condition $2r > 2s+1$ indicates that β_0 is smoother than the sample path of X.

Remark 2. The estimation error behaves differently from the prediction error. Theorem 3 has estimation error larger when the eigenvalues of the covariance operator Γ decay faster, while Theorem 2 has prediction error smaller when the eigenvalues of the covariance operator Γ decay faster.

As to assumption (A3), an initial estimator $\hat{\beta}^0$ is not $\sqrt{n}$-consistent so that the residuals $r_i^0 = y_i - \int_{\mathcal{T}} x_i(t)\hat{\beta}^0(t)dt$ differ from the true errors $\epsilon_i = \sigma\varepsilon_i$ by more than the order $n^{-1/2}$. Nevertheless, we can show that an M-scale estimator based on the residuals is $\sqrt{n}$-consistent for σ under appropriate conditions. For this, we consider a leave-one-out estimator in a manner similar to Müller, Schick, and Wefelmeyer (2004).

Theorem 4. *Suppose ρ_0 at (2.9) is twice differentiable and satisfies $\sup_t |\rho_0''(t)| < \infty$, $E\rho_0' = 0$, and $Var(\rho_0') < \infty$. Let $\hat{\beta}_{-i}$ be a leave-one-out estimator for β and set $\hat{\beta}_{ij} = E[\hat{\beta}_{-i}|(x_k, y_k), 1 \leq k \leq n, k \neq j]$. If*

$$E\|\hat{\beta}_{-i} - \beta_0\|_\Gamma^2 = o(n^{-1/2}), \tag{3.3}$$

$$\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n E\|\hat{\beta}_{-i} - \hat{\beta}_{ij}\|_\Gamma^2 = o(1), \tag{3.4}$$

then $\sqrt{n}(\hat{\sigma} - \sigma) = O_p(1)$.

In our implementation, we use the L_1 fit for initial fitting so $\hat{\beta}^0$ is the minimizer over β of (2.1) with $\rho(r) = |r|$. Cardot, Ferraty, and Sarda (2003) derived the convergence rate of the prediction error, but (3.3) is not achieved with their rate. Kato (2012) showed that the convergence rate for the prediction error in FPCA-based functional linear quantile regression is the same as that in FPCA-based functional linear regression (Cai and Hall (2006)), where the prediction error always satisfies (3.3). Although the convergence rate for the quantile smoothing spline estimator is not well-studied in literature, the convergence rate of the prediction error with smoothing spline estimator, faster than the order $n^{-1/2}$, can be derived in parallel with FPCA-based functional linear quantile regression. From (2.2) and (2.7), a leave-one-out estimator $\hat{\beta}_{-i}$ is given in the form of $\hat{\beta}_{-i}(\cdot) = \sum_{j=1}^{n} W_{ij}(\cdot)y_j$ with $W_{ii}(\cdot) = 0$, where $W_{ij}(\cdot)$ depends only on $x_k, 1 \le k \le n, k \ne i$. Since $(\hat{\beta}_{-i} - \hat{\beta}_{ij})(t) = \sigma W_{ij}(t)\varepsilon_j$, the sufficient condition for (3.4) is $n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} E\|W_{ij}\|^2 = o(1)$. One can show that $\sum_{j=1}^{n} E\|W_{ij}\|^2 = O(n^{-1}\lambda^{-1})$ for the least squares smoothing spline estimator. Analogously, one can derive the same order for quantile smoothing spline estimator under some conditions so that (3.4) is met.

4. Simulation Study

In this section, we provide the numerical performance of the proposed estimators. Several outlier-resistant loss functions are considered, as is squared error. The ρ and ψ functions of the loss functions are:

(Huber) $\rho_H(x)=x^2+(2k|x| - k^2-x^2)I(|x|>k),\ \psi_H(x)=\min(|x|,k)\cdot \mathrm{sign}(x)$;

(Logistic) $\rho_L(x)=x + 2\log(1+\exp(-x)),\ \psi_L(x)=\dfrac{1-\exp(-x)}{1+\exp(-x)}$;

(Biweight) $\rho_B(x)=1-\{1-(\frac{x}{k})^2\}^3 I(|x|\le k),\ \psi_B(x) = x\{1-(\frac{x}{k})^2\}^2 I(|x|\le k)$;

(Cauchy) $\rho_C(x)=\log(1+x^2),\ \psi_C(x)=\dfrac{x}{1+x^2}$;

(Squared error) $\rho_S(x)=x^2,\quad \psi_S(x)=x$.

The Huber and biweight loss functions have an additional tuning parameter, k, that determines the robustness and efficiency of the resulting estimator. We use $k = 1.4$ for Huber and $k = 4.68$ for biweight, respectively, corresponding to 95% efficiency (Maronna, Martin, and Yohai (2006)). The Huber and logistic ρ-functions are convex, while the biweight and Cauchy ρ-functions are not.

We adopted the simulation setting of Hall and Horowitz (2007) and Yuan and Cai (2010) with modifications for additive error, in order to contain some outliers. The true slope function β_0 defined on $\mathcal{T} = [0,1]$ was $\beta_0 = \sum_{j=1}^{50} 4(-1)^{j+1} j^{-2}\phi_j$ with $\phi_1(t) = 1$ and $\phi_{j+1}(t) = \sqrt{2}\cos(j\pi t)$ for $j \geq 1$. The random function X was generated as $X = \sum_{j=1}^{50} \pi_j Z_j \phi_j$ with independent samples Z_j from $U(-\sqrt{3}, \sqrt{3})$ and $\pi_j = j^{-2s}$. We took $s = 0.6$, 1, and 2, to regulate the decaying rate of eigenvalues of the covariance function of X. We took $\mathcal{H} = W_2^2[0,1]$ with the associated inner product, reproducing kernel, and penalty term as defined in Section 2.1. Several additive random errors in the linear model were considered to represent the outlier-prone situations: Gaussian: $\epsilon \sim N(0,1)$ (no outliers); t with 3 degree of freedom: $\epsilon \sim t_3$; t with 10 degree of freedom: $\epsilon \sim t_{10}$; and mixture Gaussian: $\epsilon \sim (1-p)N(0,1) + pN(10,1)$ with $p = 0.1$. The scale parameter was set by $\sigma = 1$. We considered $n = 50$, 100, 200, and 500. As in Yuan and Cai (2010), we measured the estimation accuracy by integrated squared error $\|\hat{\beta}_{n\lambda} - \beta_0\|^2$ and prediction error $\|\hat{\beta}_{n\lambda} - \beta_0\|_\Gamma^2$. For each configuration, we repeated the experiment 1,000 times.

In Figures 1 through 4, we provide the prediction and estimation errors, averaged over 1,000 simulation runs, with λ chosen by GCV. Results were obtained under the different types of additive errors and loss functions. All panels in the figures show prediction and estimation errors linearly decreasing in the logarithmic scale as the sample size n increases. For prediction error this agrees with the results in Theorem 2, and for estimation errors this agrees with the results in Theorem 3. Here s affects prediction and estimation errors in opposite directions, as was shown under square-loss in Yuan and Cai (2010); it is the same for all four outlier-resistant loss functions under the existence of outliers (see Figures 2 through 4) as shown in Theorems 2 and 3.

As to performance of the loss functions considered, their qualities for prediction are rather different. Table 1 lists the averages and standard deviations of prediction errors over 1,000 simulation runs. There is no real difference under Gaussian additive errors, but prediction from the squared loss case is outperformed by other outlier-resistant loss functions when additive errors follow mixture Gaussian, t_3, and t_{10}. The non-convex loss functions (biweight and Cauchy) clearly outperform convex loss functions (Huber and logistic) under severe outlyingness (mixture Gaussian), while all four outlier-resistant loss functions perform comparably under mild outlyingness (t_3 and t_{10}). The same pattern can be observed in estimation. Numerical evidence suggests the use of outlier-resistant loss in functional linear regression and the preference of non-convex loss in the existence of strong outlying observations.

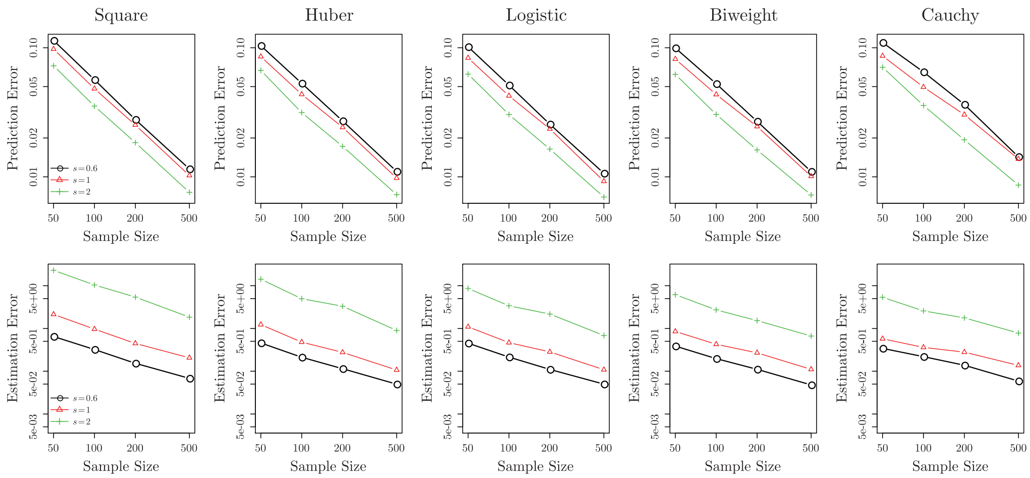

Figure 1. Prediction and estimation errors with different loss functions when the additive error was generated from Gaussian distribution (no outliers). All axes are in log scale.

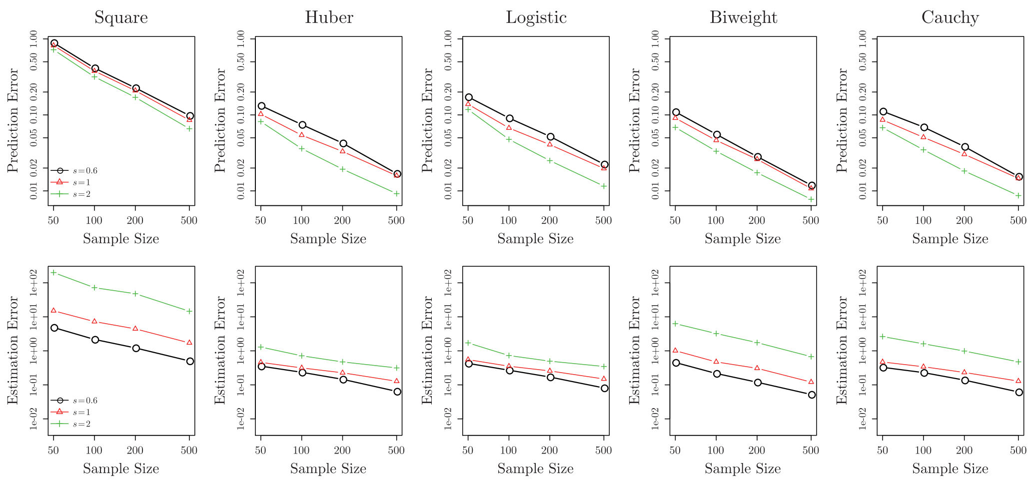

Figure 2. Prediction and estimation errors with different loss functions when the additive error was generated from mixture Gaussian distribution. All axes are in log scale.

5. Ozone Pollution Data Example

We applied our methodologies to ozone data from the California Environmental Protection Agency website (`http://www.arb.ca.gov/aqd/aqdcd/aqdcddld.htm`), where ozone observed in California during 1980 and 2011 was recorded. We took hourly ozone levels in the city of Sacramento between June and August of 2005 as our data set, a total of 92 days. Ozone levels at 4 A.M. were not recorded

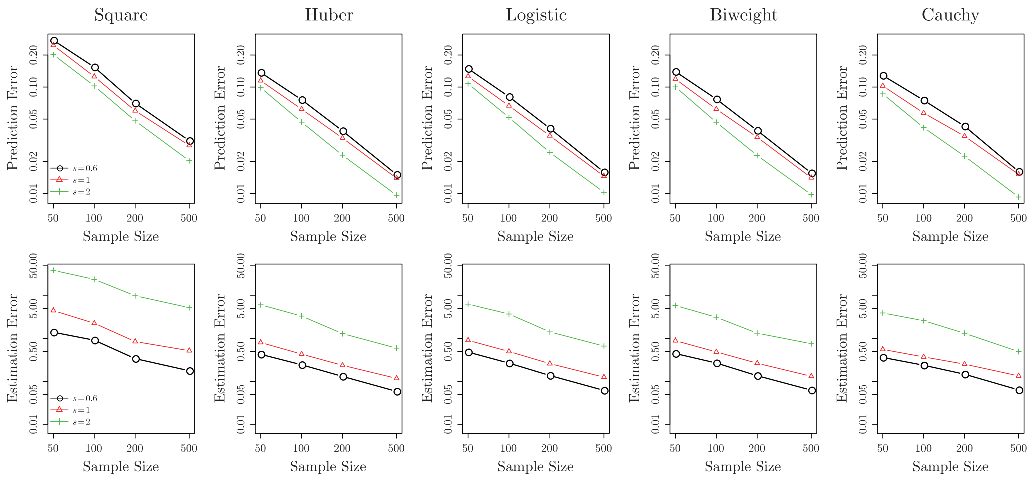

Figure 3. Prediction and estimation errors with different loss functions when the additive error was generated from t-distribution with $df = 3$. All axes are in log scale.

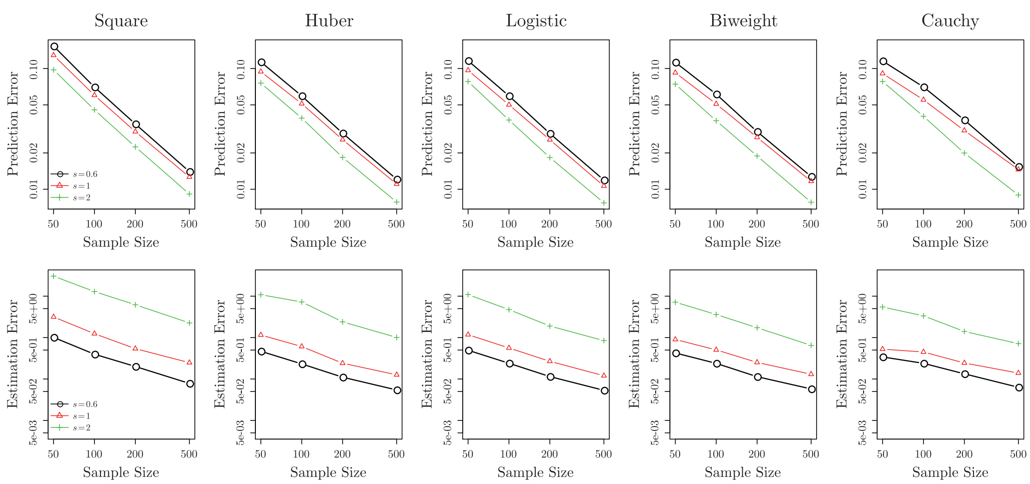

Figure 4. Prediction and estimation errors with different loss functions when the additive error was generated from t-distribution with $df = 10$. All axes are in log scale.

during the period, and two days of the period had some missing observations. We focused on the prediction of the daily maximum ozone level based on the ozone profile observed on the previous day. Thus we used hourly profile of ozone level as random covariate function evaluated on discrete time points and maximum ozone of the following day as response variable. Ozone levels were subjected to square-root transformation.

Table 1. Average (standard error) of prediction errors over 1,000 simulation runs.

s	n	Loss function	Distribution: Gaussian	mixture Gaussian	t_3	t_{10}
0.6	50	Square	0.1138 (0.128)	0.8896 (1.144)	0.2751 (0.404)	0.1533 (0.186)
		Huber	0.1038 (0.101)	0.1327 (0.092)	0.1370 (0.120)	0.1133 (0.117)
		Logistic	0.1017 (0.099)	0.1727 (0.141)	0.1491 (0.130)	0.1161 (0.114)
		Biweight	0.0995 (0.086)	0.1097 (0.107)	0.1398 (0.121)	0.1126 (0.097)
		Cauchy	0.1096 (0.070)	0.1118 (0.072)	0.1285 (0.093)	0.1155 (0.077)
	100	Square	0.0565 (0.050)	0.4144 (0.483)	0.1540 (0.279)	0.0702 (0.069)
		Huber	0.0530 (0.038)	0.0747 (0.043)	0.0760 (0.049)	0.0593 (0.047)
		Logistic	0.0513 (0.037)	0.0910 (0.055)	0.0811 (0.055)	0.0594 (0.047)
		Biweight	0.0526 (0.036)	0.0557 (0.040)	0.0770 (0.055)	0.0613 (0.047)
		Cauchy	0.0650 (0.039)	0.0692 (0.040)	0.0752 (0.044)	0.0703 (0.043)
	200	Square	0.0278 (0.023)	0.2268 (0.239)	0.0705 (0.063)	0.0347 (0.031)
		Huber	0.0271 (0.020)	0.0426 (0.027)	0.0387 (0.026)	0.0291 (0.022)
		Logistic	0.0256 (0.018)	0.0522 (0.033)	0.0409 (0.028)	0.0290 (0.022)
		Biweight	0.0269 (0.020)	0.0283 (0.020)	0.0391 (0.027)	0.0300 (0.023)
		Cauchy	0.0364 (0.025)	0.0383 (0.025)	0.0428 (0.028)	0.0375 (0.026)
	500	Square	0.0115 (0.009)	0.0981 (0.093)	0.0313 (0.031)	0.0140 (0.008)
		Huber	0.0110 (0.007)	0.0169 (0.011)	0.0150 (0.010)	0.0121 (0.010)
		Logistic	0.0106 (0.007)	0.0224 (0.015)	0.0159 (0.011)	0.0119 (0.007)
		Biweight	0.0110 (0.007)	0.0119 (0.008)	0.0155 (0.011)	0.0127 (0.008)
		Cauchy	0.0143 (0.010)	0.0155 (0.011)	0.0161 (0.012)	0.0154 (0.011)
1	50	Square	0.0976 (0.116)	0.8189 (0.988)	0.2468 (0.359)	0.1287 (0.153)
		Huber	0.0854 (0.093)	0.1023 (0.089)	0.1148 (0.113)	0.0942 (0.099)
		Logistic	0.0834 (0.088)	0.1384 (0.131)	0.1258 (0.122)	0.0966 (0.097)
		Biweight	0.0817 (0.079)	0.0907 (0.090)	0.1189 (0.118)	0.0922 (0.086)
		Cauchy	0.0864 (0.071)	0.0859 (0.071)	0.1025 (0.094)	0.0909 (0.075)
	100	Square	0.0482 (0.042)	0.3778 (0.439)	0.1254 (0.144)	0.0600 (0.061)
		Huber	0.0436 (0.034)	0.0541 (0.035)	0.0620 (0.046)	0.0511 (0.045)
		Logistic	0.0424 (0.033)	0.0672 (0.048)	0.0667 (0.052)	0.0501 (0.043)
		Biweight	0.0434 (0.032)	0.0463 (0.035)	0.0620 (0.048)	0.0509 (0.041)
		Cauchy	0.0496 (0.033)	0.0507 (0.034)	0.0570 (0.039)	0.0550 (0.041)
	200	Square	0.0252 (0.021)	0.2086 (0.235)	0.0599 (0.058)	0.0300 (0.026)
		Huber	0.0242 (0.018)	0.0330 (0.019)	0.0331 (0.021)	0.0257 (0.019)
		Logistic	0.0235 (0.018)	0.0407 (0.025)	0.0348 (0.023)	0.0257 (0.020)
		Biweight	0.0245 (0.018)	0.0261 (0.018)	0.0337 (0.022)	0.0269 (0.019)
		Cauchy	0.0303 (0.020)	0.0303 (0.018)	0.0345 (0.021)	0.0307 (0.020)
	500	Square	0.0102 (0.008)	0.0854 (0.090)	0.0282 (0.028)	0.0126 (0.010)
		Huber	0.0098 (0.007)	0.0158 (0.010)	0.0139 (0.010)	0.0110 (0.008)
		Logistic	0.0092 (0.007)	0.0197 (0.012)	0.0145 (0.011)	0.0106 (0.008)
		Biweight	0.0101 (0.007)	0.0107 (0.008)	0.0140 (0.010)	0.0116 (0.008)
		Cauchy	0.0137 (0.010)	0.0146 (0.010)	0.0151 (0.011)	0.0145 (0.010)

Table 1. (Continuous)

s	n	Loss function	Distribution			
			Gaussian	mixture Gaussian	t_3	t_{10}
2	50	Square	0.0725 (0.083)	0.7203 (0.785)	0.2017 (0.270)	0.0975 (0.106)
		Huber	0.0668 (0.077)	0.0816 (0.089)	0.0986 (0.110)	0.0756 (0.081)
		Logistic	0.0625 (0.069)	0.1177 (0.135)	0.1076 (0.117)	0.0781 (0.083)
		Biweight	0.0622 (0.065)	0.0685 (0.070)	0.1003 (0.111)	0.0742 (0.077)
		Cauchy	0.0706 (0.074)	0.0679 (0.072)	0.0863 (0.098)	0.0780 (0.082)
	100	Square	0.0353 (0.035)	0.3169 (0.330)	0.1025 (0.112)	0.0455 (0.046)
		Huber	0.0315 (0.032)	0.0360 (0.034)	0.0468 (0.046)	0.0388 (0.040)
		Logistic	0.0304 (0.030)	0.0477 (0.046)	0.0518 (0.052)	0.0374 (0.037)
		Biweight	0.0304 (0.029)	0.0333 (0.032)	0.0467 (0.046)	0.0370 (0.036)
		Cauchy	0.0357 (0.035)	0.0348 (0.034)	0.0415 (0.041)	0.0402 (0.039)
	200	Square	0.0184 (0.018)	0.1704 (0.177)	0.0482 (0.053)	0.0225 (0.022)
		Huber	0.0172 (0.017)	0.0193 (0.017)	0.0229 (0.020)	0.0184 (0.018)
		Logistic	0.0164 (0.015)	0.0250 (0.023)	0.0243 (0.021)	0.0183 (0.017)
		Biweight	0.0162 (0.015)	0.0174 (0.015)	0.0228 (0.020)	0.0189 (0.018)
		Cauchy	0.0194 (0.018)	0.0183 (0.017)	0.0224 (0.020)	0.0200 (0.019)
	500	Square	0.0076 (0.006)	0.0658 (0.068)	0.0203 (0.021)	0.0091 (0.007)
		Huber	0.0073 (0.006)	0.0092 (0.007)	0.0096 (0.008)	0.0078 (0.006)
		Logistic	0.0069 (0.005)	0.0115 (0.009)	0.0102 (0.009)	0.0077 (0.006)
		Biweight	0.0072 (0.005)	0.0077 (0.006)	0.0097 (0.008)	0.0078 (0.006)
		Cauchy	0.0086 (0.007)	0.0087 (0.007)	0.0092 (0.007)	0.0089 (0.007)

We first applied functional linear regression with the square loss. The QQ-plot using the resulting scaled residuals is presented in Figure 5; it hints at the existence of outliers. We applied four types of robust functional linear regression to the same data, and present QQ-plot of the scaled residuals from Huber loss case in Figure 5. Since robust regression is less likely affected by outliers, outlying observations are more likely highlighted in the resulting residual QQ-plot, as shown in Figure 5. Other robust regressions using the different robust losses yielded the similar residual QQ-plots.

To verify the prediction enhancement for an independent data set, we set up the test data for the same period of the year 2006. Using the model built on the 2005-year data, we predicted the daily maximum ozone level based on the previous hourly ozone profile. In Table 2, root mean squared error (RMSE) and RMSE with upper 10% trimming (RMSE(0.9)) are presented, demonstrating that all types of robust functional linear regression provide considerable improvement in prediction.

Supplementary Materials

The supplementary material contains the proofs of Proposition 1, Theorems 1, 3, and 4.

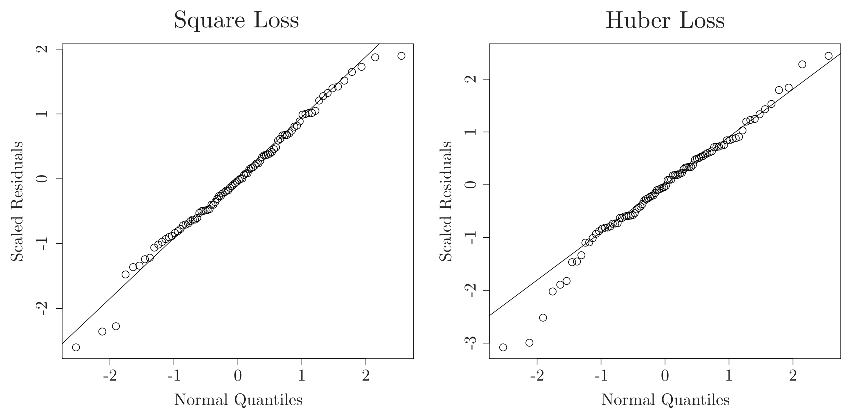

Figure 5. QQ-plots of scaled residuals for functional linear regression using square loss (left panel) and Huber loss (right panel).

Table 2. RMSE and RMSE(0.9) (RMSE with upper 10% trimming) for test data set.

	Square	Huber	Logistic	Biweight	Cauchy
RMSE	0.0294	0.0256	0.0281	0.0255	0.0261
RMSE (0.9)	0.0224	0.0206	0.0217	0.0207	0.0209

Table 3. Caption of the table.

Bandwidth		$h=3$			$h=4$			$h=5$		
Estimates		$\hat{N}_H^{LC}$	$\hat{N}_{\bar{H}}^{LC}$	$\hat{N}_{\bar{H}}^{LL}$	$\hat{N}_H^{LC}$	$\hat{N}_{\bar{H}}^{LC}$	$\hat{N}_{\bar{H}}^{LL}$	$\hat{N}_H^{LC}$	$\hat{N}_{\bar{H}}^{LC}$	$\hat{N}_{\bar{H}}^{LL}$
$beta(10,10)$	BIAS	-22.5	-14.8	14.0	-13.3	-6.9	12.5	-8.2	-4.7	11.5
$\bar{p}=0.500$	S.E.	13.8	14.9	12.1	12.6	14.6	11.7	15.1	19.3	15.5
$cv=0.218$	RMSE	26.4	21.0	18.5	18.3	16.2	17.1	17.2	19.9	19.3
$beta(5,5)$	BIAS	-32.2	-19.9	5.4	-21.9	-11.8	4.9	-15.6	-8.4	5.1
$\bar{p}=0.500$	S.E.	15.8	16.2	12.3	14.3	19.8	12.9	16.4	19.8	14.6
$cv=0.302$	RMSE	35.9	25.7	13.4	26.1	15.9	11.9	22.6	21.5	15.4
$beta(4,8)$	BIAS	-53.7	-29.1	-10.8	-42.0	-19.3	-8.4	-34.4	-13.9	-7.6
$\bar{p}=0.333$	S.E.	20.8	19.4	18.1	18.5	18.8	15.6	19.2	20.6	16.4
$cv=0.392$	RMSE	57.6	34.9	21.1	45.9	26.9	17.8	39.4	24.9	18.1
$beta(3,5)$	BIAS	-57.4	-32.4	-15.5	-45.6	-22.8	-13.1	-37.4	-17.1	-11.4
$\bar{p}=0.375$	S.E.	21.2	19.9	17.1	18.9	18.8	15.0	19.6	21.6	15.6
$cv=0.430$	RMSE	61.2	38.1	23.1	49.4	29.5	19.9	42.3	27.5	19.3

Acknowledgements

We thank two anonymous referees and the Editor/AE for their constructive comments and helpful suggestions. S. Lee was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2014R1A1A1002779).

References

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34**, 2159-2179.

Cardot, H., Crambes, C. and Sarda, P. (2005). Quantile regression when the covariates are functions. *J. Nonparametr. Stat.* **17**, 841-856.

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

Cardot, H. and Johannes, J. (2010). Thresholding projection estimators in functional linear models. *J. Multivariate Anal.* **101**, 395-408.

Cox, D. D. (1983). Asymptotics for M-Type Smoothing Splines, *Ann. Statist.* **11**, 530-551.

Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37**, 35-72.

Cucker, F. and Smale, S. (2001). On the mathematical foundations of learning. *Bull. Amer. Math. Soc.* **39**, 1-49.

Cunningham, J. K., Eubank, R. L. and Hsing, T. (1991). M-type smoothing splines with auxiliary scale estimation. *Comput. Statist. Data Anal.* **11**, 43-51.

Gervini, D. (2012). Functional robust regression for longitudinal data. ArXiv 1211.7332.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35**, 70-91.

Kato, K. (2012). Estimation in functional linear quantile regression. *Ann. Statist.* **40**, 3108-3136.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33**, 82-95.

Li, Y. and Hsing, T. (2007). On rates of convergence in functional linear regression. *J. Multivariate Anal.* **98**, 1782-1804.

Maronna, R. A., Martin, R. D. and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. John Wiley and Sons, New York.

Maronna, R. A. and Yohai, V. J. (2011). Robust functional linear regression based on splines. *Comput. Statist. Data Anal.* **65**, 46-55.

Müller, U. U., Schick, A. and Wefelmeyer, W. (2004). Estimating linear functionals of the error distribution in nonparametric regression. *J. Statist. Plann. Inference* **119**, 75-93.

Shin, H. and Hsing, T. (2012). Linear prediction in functional data analysis. *Stochastic Process. Appl.* **122**, 3680-3700.

Yao, F., Müller, H. G. and Wang, J. L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33**, 2873-2903.

Yuan, M. and Cai, T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Statist.* **38**, 3412-3444.

Frontier CS Lab, Samsung Software R&D Center, Suwon, 443-732, Korea.

E-mail: hjshin112@gmail.com

Department of Statistics, Hankuk University of Foreign Studies, Yongin, 130-791, Korea.

E-mail: lees@hufs.ac.kr

(Received September 2014; accepted February 2015)