**Noname manuscript No.**
(will be inserted by the editor)

# Functional Single-index Quantile Regression Models

**Peijun Sang · Jiguo Cao**

Received: date / Accepted: date

**Abstract** It is known that functional single index regression models can achieve better prediction accuracy than functional linear models or fully nonparametric models, when the target is to predict a scalar response using a function-valued covariate. However, the performance of these models may be adversely affected by extremely large values or skewness in the response. In addition, they are not able to offer a full picture of the conditional distribution of the response. Motivated by using trajectories of $PM_{10}$ concentrations of last day to predict the maximum $PM_{10}$ concentration of the current day, a functional single-index quantile regression model is proposed to address those issues. A generalized profiling method is employed to estimate the model. Simulation studies are conducted to investigate the finite sample performance of the proposed estimator. We apply the proposed framework to predict the maximal value of $PM_{10}$ concentrations based on the intraday $PM_{10}$ concentrations of the previous day.

**Keywords** Functional data analysis · Generalized profiling · Quantile regression · Robustness · Single-index model

## 1 Introduction

With the development of technology, it is increasingly common to obtain data of complex structures like growth curves, medical images and spectral data. These data can be viewed as functions of time or spatial points or some other continua, which are called functional data. Numerous approaches for analyzing functional data have been rapidly developed (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006). One important topic in functional data analysis is to use functional covariates to predict a scalar response. For example, we are interested in prediction of the maximal concentration of an atmospheric particulate matter called $PM_{10}$ given the intraday $PM_{10}$ profile of the previous day.

In recent years, many investigators have proposed regression models with functional covariates and a scalar response, so-called scalar-on-function regression models. Reiss *et al.* (2017) provides a comprehensive overview of these models. The simplest one within the family is the functional linear model, which assumes that a scalar response depends linearly on a functional covariate. But this linear assumption is susceptible to model misspecification, and thus the prediction performance of this linear model may be adversely affected in applications. This concern has been raised by numerous researchers (see Chen *et al.*, 2011; Müller *et al.*, 2013; Müller and Yao, 2008, etc.). Fully nonparametric models, implemented, for example, using functional Nadayara-Watson (Ferraty and Vieu, 2006) and local linear estimators (Ferraty *et al.*, 2010) have been proposed. One concern of fully nonparametric models is that it is difficult to interpret the effect of the functional covariate (Horowitz, 2009). Moreover, Chen *et al.* (2011) argued that the prediction performance of these models is unsatisfactory and thus seeking methods can achieve better trade-off between bias and variance is indispensable to enhance the performance in prediction.

A critical idea for progress in this direction is to connect a response to a linear transformation of a functional covariate with some link functions; it is in a similar spirit to the generalized linear model. James (2002) assumed a known link function, while Müller and Stadtmüller (2005) considered a monotone but un-

Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada, V5A1S6. E-mail: jiguo_cao@sfu.ca

konwn link function. However, as pointed out by Chen *et al.* (2011), these link-function models often lack sufficient flexibility due to restrictive assumptions imposed on the link function. Instead, they proposed extensions of these models to allow for more flexible link functions. A faster convergence rate than that of fully nonparametric models was established in Chen *et al.* (2011). Moreover, they found that the proposed flexible link-function model enjoyed better trade-off between bias and variance than the models with restricted link functions.

The aforementioned models mainly focus on how the conditional mean of a scalar response is related to a given functional covariate. Thus they may not be effective tools if other features of the conditional distribution of a scalar response given a functional covariate are of primary interest. Conditional quantile regression models have been proposed as a useful alternative since they are able to provide a detailed description of the conditional distribution. For instance, Chen and Müller (2012) considered using a monotone link function to connect the conditional distribution with a linear functional of a functional covariate. This is an indirect way to model the conditional quantile function (Kato, 2012). Several methods, though not many, have been proposed in literature to directly model the conditional quantile function. Kato (2012) introduced a approach based on functional principal component analysis to fit a functional linear quantile regression model, which he considered as a benchmark model in estimating conditional quantiles given functional covariates. This model was also fitted with smoothing splines in Cardot *et al.* (2005). Ferraty *et al.* (2005) modelled the conditional quantile in a fully nonparametric manner. A more comprehensive review on the quantile regression can be found in Koenker (2005).

In light of the aforementioned issues of linear models and full nonparametric models in estimation of the conditional mean function of a scalar response given a functional covariate, we aim to seek appropriate models to achieve better variance-bias trade-off when the primary interest becomes the conditional quantile function of the scalar response. Since the link-function regression for modeling the conditional mean function achieves good performance in both theory and applications, as shown in Chen *et al.* (2011), we propose to connect the conditional quantile function with a linear transformation of a functional covariate using a nonparametric link function. To the best of our knowledge, this is the first framework based on nonparametric link functions to directly model the conditional quantile function for a given functional covariate. To fit this model, we propose a generalized profiling method, where two nested levels of optimization with different target functions are implemented to estimate the linear transformation and link function, respectively. Simulation studies show that the proposed estimator, compared with its counterpart of the conditional mean function, is more robust to extreme outcomes. Furthermore, when applying these two models to predict the maximal value of $PM_{10}$ concentrations in the real application, we find that the estimator of the conditional median outperforms its counterpart in prediction accuracy.

The remainder of the paper is organized as follows. In Section 2 we introduce the functional single index quantile regression model and the generalized profiling method to fit the model. Simulation results for evaluating finite sample performances of the proposed estimator are reported in Section 3. We apply the new model to the $PM_{10}$ data and compare its prediction accuracy with that of the functional single index mean regression model proposed by Chen *et al.* (2011) in Section 4. In Section 5, we provide concluding remarks.

## 2 Model and Estimation

### 2.1 Model

Suppose that $(X_1, Y_1), (X_2, Y_2), \dots (X_n, Y_n)$ are identical and independent copies of $(X, Y)$, where the covariate $X$ is a random function defined on a compact interval $\mathcal{I} = [0, T]$ and the response $Y$ is a scalar variable. We further postulate that $X$ belongs to $L^2[0, T]$, a collection of all square integrable functions defined over $[0, T]$. A general model for characterizing the relationship between $X$ and $Y$ can be expressed as follows:

$$Y = m(X) + \epsilon, \tag{1}$$

where $m$ is a functional defined on $L^2[0, T]$ and $\epsilon$ denotes the random error. In the literature of scalar-on-function regression, the most commonly used model assumes that $m$ is linear. In other words, $m(X) = a + \int_{\mathcal{I}} X(t)\beta(t)dt$ for some $a \in \mathbb{R}$ and $\beta \in L^2[0, T]$. As with multiple linear models, functional linear models may suffer from inadequate fit and misspecification. These issues might make it restrictive for use in practice.

Numerous alternative models have been proposed to address the issues of functional linear models. For instance, Ferraty and Vieu (2006) considered a functional Nadayara-Watson estimator of $m$ in model (1) and Ferraty *et al.* (2010) proposed a functional local linear estimator. These fully nonparametric models and estimators can avoid model misspecifications and thus can be widely applied to predict the responses using the functional covariate. However, the interpretation of such models is difficult in practice. In addition, their performance in prediction can be poor in small samples since convergence rates of them are slow (Chen *et al.*, 2011). To enhance interpretability and prediction accuracy of nonparametric models, Ait-Saïdi *et al.* (2008) and Chen *et al.* (2011) proposed functional single index models:

$$Y = f\left\{\int_{\mathcal{I}} X(t)\beta(t)dt\right\} + \epsilon,$$

where $f$ is an unknown link function and $\beta$ is an unknown index function satisfying $\int_{\mathcal{I}} \beta^2(t)dt < \infty$.

The conditional mean structure of $Y$ is the target of the models mentioned above. Therefore, they are not effective tools to characterize the entire conditional distribution of $Y$ given the functional covariate $X$. To achieve this goal, we propose a functional single index model for the conditional quantile of $Y$. For any $x \in L^2[0,T]$ and $\tau \in (0,1)$, the $\tau$th conditional quantile of $Y$ is defined as

$$f_\tau(x) = \arg\min_a \mathrm{E}\{\rho_\tau(Y-a)|X=x\},$$

where $\rho_\tau(s) = s[\tau - I(s<0)]$ is the quantile loss function. The conditional quantile function is assumed to be $f_\tau\left\{\int_{\mathcal{I}} X(t)\beta_\tau(t)dt\right\}$ for some $\beta_\tau \in L^2(\mathcal{I})$. To ease notation, we write for every $\tau \in (0,1)$

$$Q_\tau(Y|X=x) = f_\tau\left\{\int_{\mathcal{I}} x(t)\beta_\tau(t)dt\right\}, \tag{2}$$

where $Q_\tau$ denotes the conditional $\tau$th quantile of $Y$ given $X = x$. This quantile regression model is essentially an extension of the functional single index model proposed by Chen *et al.* (2011). Moreover, this quantile regression model is able to accommodate both nonlinearity in the functional covariate and heteroscedasticity in the error term; these properties are desirable for models to achieve good prediction performance in wide applications. For instance, assuming that $Y = \left\{\int_{\mathcal{I}} X(t)\beta(t)dt\right\}^2\epsilon$, where $\epsilon \sim N(0,1)$ is independent of the random function $X$, the conditional $\tau$th quantile of $Y$ given $X$ is $\left\{\int_{\mathcal{I}} X(t)\beta(t)dt\right\}^2\Phi^{-1}(\tau)$, where $\Phi^{-1}$ denotes the quantile function of the standard normal distribution.

To address the issue of model identifiability, we assume that $\beta_\tau(0) = 1$. The performance of quantile regression is more robust to skewness and outliers in the response, in contrast to that of least squares regression. This robust property of model (2) will be investigated in both simulation studies and a real application.

2.2 Estimation

Fitting the model (2) requires estimating both the index function $\beta_\tau$ and the link function $f_\tau$. Our basic strategy is to approximate the infinite dimensional space where $\beta_\tau$ lies by a finite dimensional space growing with respect to sample size in estimation of $\beta_\tau$, and to apply the representer theorem to covert an infinite dimensional minimization problem to a finite dimensional minimization problem in estimation of $f_\tau$. A generalized profiling approach is then implemented to connect estimates of these two functions. The details are as follows.

We first treat estimation of $\beta_\tau$. B-spline basis functions are employed for this purpose. We provide a brief review of B-spline functions. Let $0 = t_0 < t_1 < \cdots < t_N = T$ be a partition1 of the interval [0, T] into $N$ subintervals. Let $\mathscr{S}_n$ denote the space of polynomial splines of degree $l$ on $[0,T]$; it consists of functions $s$ satisfying: (i) $s$ is a polynomial of degree $l$ in each subinterval $[\tau_m, \tau_{m+1}], m = 0,\ldots,N-1$; (ii) for $0 \le l^\star \le l-1$, the $l^\star$-th order derivative of $s$ is continuous in $[0,T]$. Then there exist $M = N + l$ normalized B-spline basis functions $\{B_k, 1 \le k \le M\}$ bounded by 0 and 1 in [0, T], such that any $g \in \mathscr{S}_n$ can be written as

$$g(t) = \sum_{j=1}^{M} b_j B_k(t)$$

for $t \in [0,T]$. According to Corollary 1 of Stone (1985), the index function $\beta_\tau(t)$ can be approximated by functions in $\mathscr{S}_n$ arbitrarily well under some mild conditions. In particular, there exists a vector $\boldsymbol{\theta}_\tau \in \mathbb{R}^M$ such that $\beta_\tau(t) \approx \boldsymbol{\theta}_\tau^\top \boldsymbol{B}(t)$, where $\boldsymbol{B}(t) = (B_1(t),\ldots,B_M(t))^\top$. To ensure that $\beta_\tau(0) = 1$, we just need to take the first component of $\boldsymbol{\theta}_\tau$ to be 1.

Now we move to estimate the link function $f_\tau$. In this article we assume that the link function $f_\tau$ is an element of $\mathscr{F}_K$, a reproducing kernel Hilbert space (RKHS) generated by a positive definite kernel $K(\cdot,\cdot)$. The estimators of the spline coefficient vector $\boldsymbol{\theta}_\tau$ and the unknown function $f_\tau$ are defined by

$$(\hat{\boldsymbol{\theta}}_\tau, \hat{f}_\tau) = \mathop{\arg\min}_{\boldsymbol{\theta}_\tau \in \{1\}\times\mathbb{R}^{M-1}, f_\tau \in \mathscr{F}_K} n^{-1}\sum_{i=1}^{n} \rho_\tau\{Y_i - f_\tau(\boldsymbol{\theta}_\tau^\top \boldsymbol{u}_i)\} + \lambda J(f_\tau), \tag{3}$$

where $\boldsymbol{u}_i = \left\{\int_0^T X_i(t)B_1(t)dt,\ldots,\int_0^T X_i(t)B_M(t)dt\right\}^\top$ and $\lambda > 0$ denotes a smoothing parameter. Note that the right side of (3) consists of two terms: the first term concerns fidelity to the data and the second term $J(f_\tau)$ assesses the complexity of $f_\tau$. The smoothing parameter $\lambda$ controls the balance between these two parts. A special example is the Sobolev space $\mathscr{F} =$

$\{h \mid h^{(\nu)}$ is absolutely continuous for $\nu = 0, 1;\ h^{(2)} \in L^2[0,1]\}$ equipped with the norm

$$||h||^2_{\mathscr{F}} = \left(\int_0^1 h(x)dx\right)^2 + \left(\int_0^1 h'(x)dx\right)^2 + \int_0^1 \{h^{(2)}(x)\}^2 dx.$$

For this reproducing kernel Hilbert space, a possible choice for the functional $J$ is

$$J(f_\tau) = \int_0^1 \{f_\tau^{(2)}(t)\}^2 dt.$$

Actually this specific Sobolev space has been suggested by Nychka *et al.* (1995) to model the conditional quantile of $Y$ given a vector-valued covariate $\boldsymbol{x}$:

$$\min_{f_\tau \in \mathscr{F}} \sum_{i=1}^n \rho_\tau(Y_i - f_\tau(\boldsymbol{x}_i)) + \lambda \int_0^1 \{f_\tau^{(2)}(t)\}^2 dt.$$

This is referred to as a quantile smoothing spline model (Koenker *et al.*, 1994) in literature.

A generalized profiling approach is adopted to combine these two estimation tasks. Estimation of $\boldsymbol{\theta}_\tau$ and estimation of $f_\tau$ are two nested levels of optimization with different target functions. The essential difference between the generalized profiling approach and the ordinary profiling approach is that the latter approach can only accommodate the same objective functions in two levels of opotimization. To our best knowledge, this approach is firstly proposed by Ramsay *et al.* (2007), which is also called "parameter cascades".

In the inner level, we estimate the spline coefficient vector $\boldsymbol{\theta}_\tau$ given an estimate of $f_\tau$ by minimizing

$$\ell(\boldsymbol{\theta}_\tau | f_\tau) = n^{-1} \sum_{i=1}^n \rho_\tau\{Y_i - f_\tau(\boldsymbol{\theta}_\tau^\top \boldsymbol{u}_i)\}. \quad (4)$$

This implies that the estimate of $\boldsymbol{\theta}_\tau$ is an implicit function of $f_\tau$; it is denoted by $\widehat{\boldsymbol{\theta}_\tau}(f_\tau)$. We implement the *nlm* function in R (R Core Team, 2017) to minimize $\ell$ with respect to $\boldsymbol{\theta}_\tau$ to obtain this function. In the outer level, we search a minimizer of a regularized empirical risk function from a RKHS in (3), when $\boldsymbol{\theta}_\tau$ is given. Thus we resort to the representer theorem (Kimeldorf and Wahba, 1971) to convert this infinite dimensional minimization problem to a finite dimensional problem. This theorem is widely used in support vector machines when reproducing kernels are accounted for (Hastie *et al.*, 2009). Let $||f||_{\mathscr{F}_K}$ denote the norm of an element $f$ in the RKHS $\mathscr{F}_K$. In particular, for a given $\boldsymbol{\theta}_\tau$, there exist $\alpha_0 \in \mathbb{R}$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\top \in \mathbb{R}^n$ such that

$$\widehat{f}_\tau(t) = \alpha_0 + \sum_{i=1}^n \alpha_i K(t, \boldsymbol{\theta}_\tau^\top \boldsymbol{u}_i), \quad (5)$$

if we take $J(f_\tau) = ||f_\tau||^2_{\mathscr{F}_K}$. The representation in (5) further suggests that we only need to consider minimizing target functions with respect to $\boldsymbol{\theta}_\tau$, $\alpha_0$ and $\boldsymbol{\alpha}$. More specifically, using the profiling method we find that the minimizer of $\ell$ in (4), denoted by $\widehat{\boldsymbol{\theta}_\tau}$, is an implicit function of $\alpha_0$ and $\boldsymbol{\alpha}$. Then we we are able to solve the minimization problem of (3) by plugging (5) and replacing $\boldsymbol{\theta}_\tau$ with this implicit function.

## 2.3 Tuning Parameter Selection

There are two tuning parameters in estimation of $\beta_\tau$ and $f_\tau$: the number of B-spline basis functions, $M$, and the smoothing parameter $\lambda$ in (3). A small $M$ may not be able to provide an adequate approximation of the index function $\beta_\tau$, while a large $M$ may lead to an excessively wiggly estimate of $\beta_\tau$. The smoothing parameter $\lambda$ controls the trade off between fidelity to the data and complexity of the link function $f_\tau$. From the perspective of balance between bias and variance, a large $M$ and a small $\lambda$ result in an estimator of the conditional quantile with a low bias but a high variance. In contrast, an estimator with a high bias but a low variance would be generated if a small $M$ and a large $\lambda$ are chosen.

In our numerical studies, we adopt the Sobolev space as our choice of the RKHS for implementing this generalized profiling approach to fit the single index model. For linear fitting methods, we know that the generalized cross validation (GCV) score can be regarded as an approximation of the leave-one-out cross validation score (Hastie *et al.*, 2009). Wahba (1985) considered the GCV method in smoothing spline problems and established its nice properties in choosing an appropriate tuning parameter. In her framework, if the estimated response is given by $\hat{\boldsymbol{Y}} = S(\lambda)\boldsymbol{Y}$ for some matrix $S(\lambda)$, where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$, the GCV estimate of $\lambda$ is defined by

$$V(\lambda) = \frac{1/n||\boldsymbol{Y} - S(\lambda)\boldsymbol{Y}||^2}{[(1/n)\{n - \operatorname{tr}(S(\lambda))\}]^2}.$$

For the quantile smoothing spline model, the leave-one-out cross validation score is given by

$$RCV(\lambda) = \frac{1}{n}\sum_{i=1}^n \rho_\tau\{Y_i - \hat{f_\tau}^{[-i]}(\boldsymbol{x}_i)\},$$

where $\hat{f_\tau}^{[-i]}$ denotes the estimated $\tau$th conditional quantile of $Y$ when the $i$th observation $(\boldsymbol{x}_i, Y_i)$ is removed from the training set. Yuan (2006) showed that in this case, the following score

$$\frac{\sum_{i=1}^n \rho_\tau\{Y_i - \hat{f_\tau}(\boldsymbol{x}_i)\}}{n - \operatorname{tr}(H)},$$

where $H$ is a $n \times n$ matrix with the $(i,j)$th entry $\partial \hat{f}_\tau(\boldsymbol{x}_i)/\partial Y_j$, can approximate $RCV(\lambda)$ very well. Thus he called this score the generalized approximate cross-validation (GACV).

We employ the GACV to choose these two tuning parameters when fitting this single index quantile model. For each pair $(M, \lambda)$, the GACV is defined by

$$GACV(M,\lambda) = \frac{\sum_{i=1}^n \rho_\tau\{Y_i - \widehat{f}_\tau(\widehat{\boldsymbol{\theta}_\tau}^\top \boldsymbol{u}_i)\}}{n - \operatorname{tr}(H)}, \tag{6}$$

where $(\widehat{f}_\tau, \widehat{\boldsymbol{\theta}_\tau})$ denotes the solution of (3) by using the generalized profiling approach, and $H$ denotes a hat matrix with the $(i,j)$th entry $\partial \widehat{f}_\tau(\widehat{\boldsymbol{\theta}_\tau}^\top \boldsymbol{u}_i)/\partial Y_j$. We select the pair of $M$ and $\lambda$ which can minimize the GACV as the "optimal" tuning parameters.

## 3 Simulation Studies

In this section we aim to study the finite sample performance of the estimator of the functional single index quantile regression model obtained from solving (3).

We generated the functional covariate in a similar way to that adopted in Chen *et al.* (2011). More specifically, the functional covariates were identically and independently generated as:

$$X_i(t) = t + \sum_{k=1}^4 \phi_k(t)\xi_{ik}, \quad i = 1,\ldots,n,$$

where $\phi_1(t) = \sin(2\pi t)/\sqrt{2}$, $\phi_2(t) = \cos(2\pi t)/\sqrt{2}$, $\phi_3(t) = \sin(4\pi t)/\sqrt{2}$, $\phi_4(t) = \cos(4\pi t)/\sqrt{2}$ and $\xi_{ik} \sim N(0, \lambda_k)$ are independent with $\lambda_k = 0.5^{k-1}$, $k = 1,2,3,4$. These covariates are sampled at 100 equally spaced points between 0 and 1. The responses are generated as follows:

$$Y_i = f\left\{\int_0^1 X_i(t)\beta(t)dt\right\} + \epsilon_i, \quad i = 1,\ldots,n, \tag{7}$$

where $\beta(t) = 2t^2 + 0.25t + 1$ and $f(x) = 0.66\exp(x^2)$. The random errors, $\epsilon_i$'s, are generated in five scenarios: (i) $\epsilon_i \sim \text{Laplace}(\mu = 0, b = 1)$, (ii) $\epsilon_i \sim N(0,2)$, (iii) $\epsilon_i \sim 0.85 \cdot t_{\nu=3}$, (iv) $\epsilon_i \sim 0.58 \cdot (\chi^2(3) - 2.366)$ and (v) $\epsilon_i \sim \exp\{\int_0^1 X_i(t)(t+1)dt\}N(0,1)$ for better assessment of the finite sample performance of the proposed estimator in different scenarios. The first three designs consider symmetric distributions for random errors while the fourth one for asymmetric random errors and the last one for conditional heteroskedastic errors. The signal-to-noise ratios are all around 2 in the first four designs.

To assess the performance of the proposed estimator in prediction, we compare it with the estimator for the functional single index model proposed in Chen *et al.* (2011) and $k$-nearest neighbour method for functional data proposed by Burba *et al.* (2009) as an example of fully nonparametric regression models. Our estimator models the conditional quantiles of $Y_i$'s, while the estimator by Chen *et al.* (2011) and the nearest neighbour estimator concern the conditional mean. Therefore, we denote our estimator by FSiQ, the estimator based on the functional single index mean regression by FSiM and the $k$-nearest neighbour method by FKnn throughout the paper. In each simulation replicate we randomly generate $n = 100$ or 500 independent copies of $(X_i, Y_i)$ for training and $N = 100$ copies for testing. To better assess prediction performance of these two models, we repeat the simulation procedure 300 times for each of these three designs. We considered 20 candidate values for $\lambda$ and 4 for $M$, thus there were 80 combinations of these two tuning parameters. It turns out that each simulation replicate takes around 7.5 minutes in a laptop with an Intel core i7 4.2 GHz processor to implement the proposed estimation approach including selecting two tuning parameters.

In the first three designs, the true conditional median of $Y_i$ is $f\left\{\int_0^1 X_i(t)\beta(t)dt\right\}$, which is also the true conditional mean of $Y_i$. In contrast, the true conditional median of $Y_i$ is still $f\left\{\int_0^1 X_i(t)\beta(t)dt\right\}$ while the true conditional mean of $Y_i$ is $f\left\{\int_0^1 X_i(t)\beta(t)dt\right\} + 0.368$ in the fourth design. In the fifth design, the true conditional mean of $Y_i$ is still $f\left\{\int_0^1 X_i(t)\beta(t)dt\right\}$ while the true conditional $\tau$th quantile of $Y_i$ is $f\left\{\int_0^1 X_i(t)\beta(t)dt\right\} + \Phi^{-1}(\tau)\exp\{\int_0^1 X_i(t)(t+1)dt\}$, where $\Phi^{-1}(\tau)$ denotes the $\tau$th quantile of $N(0,1)$. We compare these three estimators in terms of prediction accuracy. Reich *et al.* (2011) suggested the mean squared error was an appropriate measure for comparing prediction performance of quantile models. More specifically, after estimating both the index function and the link function using the training data, we will then compute the mean squared errors on the test data:

$$\text{MSE} = \frac{1}{N}\sum_i \left[f_\tau\left\{\int_0^1 X_i(t)\beta_\tau(t)dt\right\} - \hat{f}_\tau\left\{\int_0^1 X_i(t)\hat{\beta}_\tau(t)dt\right\}\right]^2.$$

In the expression above, $f_\tau\left\{\int_0^1 X_i(t)\beta_\tau(t)dt\right\}$ and $\hat{f}_\tau\left\{\int_0^1 X_i(t)\hat{\beta}_\tau(t)dt\right\}$ denote the true and the estimated conditional quantiles of the $i$th response, respectively. The summation is taken on the test set and $N$ denotes its size. For the third estimator, FKnn, $\hat{f}\left\{\int_0^1 X_i(t)\hat{\beta}(t)dt\right\}$ denotes the predicted response using the $k$-nearest neighbor trained by the training set.

In addition, we also compare the first two estimators in terms of performance in estimation of the index function $\beta$, for which the integrated squared error is adopted in Chen *et al.* (2011) as a criterion. In each simulation replicate, the root integrated squared error for $\hat{\beta}$ is defined by:

$$\text{ISE} = \sqrt{\int_0^1 \{\beta(t) - \hat{\beta}(t)\}^2 dt},$$

in which $\beta$ and $\hat{\beta}$ denote the true and the estimated index functions, respectively.

Table 1 summarizes the comparison of the averages and the standard errors of MSEs and ISEs of these three estimators in all five designs. Note that Table 1 focuses on the predictive performance. That's why we only consider $\tau = 0.5$ in this table. We find that the conditional distribution of the response has a negligible effect on estimation of the index function for both relatively small samples ($n = 100$) and large samples ($n = 500$). This conclusion is supported by the right part of Table 1 and Figures S2 and S4 in the Supplementary Material.

However, there exist remarkable differences in the comparison of predictions between these three methods for different designs, especially when the size of the training set is large ($n = 500$). First of all, when the true model is given in the form of (2), mean squared errors of FKnn are more than threes times of those given by FSiQ and/or FSiM. In terms of the comparison between the first two estimators, when sample size is relatively small, say $n = 100$, we can hardly figure out any difference between these two estimators in terms of prediction accuracy regardless of the conditional distribution of the response. As presented in the top of Table 1 and also Figure S1 in the Supplementary Material, the difference in the average of mean squared errors across these 300 trials does not exceed 0.07. However, for relatively larges samples ($n = 500$), the conditional distribution of the response indeed plays a critical role in determining prediction performance of these two estimators. More specially, if the conditional distribution is light tailed, there is little difference in predictions between these two estimators, as can be seen from the row corresponding to "Normal" in Table 1 and Figure S3 in the Supplementary Material. But obvious distinctions can be seen in Figure S3 when the conditional distribution becomes heavy tailed. Moreover, the first, second and fourth rows in Table 1 indicate that the average of mean squared errors from FSiM is around twice as large as those based on FSiQ when tails of random errors are heavier. These distinctions imply that when extremely large values occur frequently in the response, the estimator from the functional single index mean regression model is less accurate than its counterpart from the quantile regression model. In other words, the functional single index quantile regression model appears more robust to extremely large outcomes. Regarding the last design where heteroskedastic errors are invovled, there's little difference in prediction performance between the first two methods.

In addition to the robustness property of the quantile regression model, we are also interested in whether the proposed estimation scheme is able to provide a reasonable estimate of conditional quantiles. Table 2 considers seven quantile levels in these five designs. More specifically, we estimate the $\tau$th conditional quantile for each design using a training set of size $n$ (500 or 100), and then compare the estimated quantiles with the underlying true quantiles. The mean squared errors of these estimates are summarized in Table 2. Obviously the proposed method is able to provide reasonable estimates of conditional quantiles when $\tau$ is neither too small (0.05 say) nor too large (0.95 say) regardless of the conditional distribution of the response. The estimation accuracy is adversely affected when the size of the training sample decreases quickly.

Figure 1 displays the estimated conditional quantiles in the first three designs from one randomly selected simulation replicate when $n = 500$ independent curves are used for training. Three quantile levels ($\tau = 0.25, 0.5, 0.75$) are considered in Figure 1; the estimated conditional quantile functions are reasonably close to the true ones in these designs. To provide a direct comparison of the two models, we present the estimated conditional mean functions in the middle row of Figure 1. Fitting these two models yields consistent estimators when the conditional distribution is light tailed. The estimator of FSiQ, however, is considerably closer to the true conditional median (mean) function in the Student $t$ case. This suggests that the performance of FSiM might be impaired when extremely large values are frequently observed in the response.

To study whether how the functional covariate is generated has an impact on performance of the proposed estimator, we also consider a non-Gaussian functional covariate, as suggested in Delaigle and Hall (2012) and Dai *et al.* (2017). More specifically, $\xi_{ik}$'s are independently generated from a centered exponential distribution with variance $\lambda_k$. The above procedure are repeated then and the corresponding numerical results are shown in the Supplementary Material.

## 4 Real Data Illustration

$PM_{10}$ is defined to be subtypes of atmospheric particles, such as suspended particulate matter, thoracic and res-

$n = 100$

| Distribution | MSE | | | ISE | |
|---|---|---|---|---|---|
| | FSiQ | FSiM | FKnn | FSiQ | FSiM |
| Laplace | .281 (.248) | .329 (.246) | 1.946 (.930) | .045 (.071) | .062 (.069) |
| Normal | .365(.232) | .310 (.214) | 1.983 (.924) | .055 (.070) | .058 (.055) |
| Student $t$ | .263 (.233) | .338 (.266) | 1.881 (.878) | .055(.093) | .071 (.089) |
| Chi-square | .357 (.160) | .308 (.226) | 1.997 (1.011) | .034(.039) | .046 (.036) |
| Hetero | .858 (.563) | .783 (.382) | 3.055 (2.264) | .092(.131) | .123 (.115) |

$n = 500$

| Distribution | MSE | | | ISE | |
|---|---|---|---|---|---|
| | FSiQ | FSiM | FKnn | FSiQ | FSiM |
| Laplace | .052 (.054) | .094 (.089) | 1.595 (.987) | .019 (.016) | .019 (.014) |
| Normal | .098 (.079) | .095 (.091) | 1.214 (.581) | .023 (.019) | .019 (.013) |
| Student $t$ | .047 (.037) | .115 (.144) | 1.558 (.965) | .022 (.021) | .020 (.015) |
| Chi-square | .058 (.061) | .094 (.086) | 1.620 (.985) | .014 (.010) | .019 (.010) |
| Hetero | .216 (.147) | .226 (.120) | 2.104 (1.179) | .026 (.022) | .030 (.023) |

Table 1: Summary of the averages and the standard errors (in brackets) of MSEs and ISEs across the 300 replicates in all four designs when the size of training data is $n = 100$ and 500. Hetero refers to the design of heteroskedastic errors.

$n = 100$

| Distribution | $\tau$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | .05 | 0.15 | 0.25 | 0.5 | 0.75 | 0.85 | 0.95 |
| Laplace | 4.754 | 2.307 | 1.814 | .313 | 1.232 | 2.168 | 4.539 |
| Normal | 4.453 | 2.728 | 1.812 | 0.360 | 1.745 | 2.771 | 4.821 |
| Student $t$ | 3.638 | 1.981 | 1.434 | 0.265 | 1.373 | 2.132 | 3.791 |
| Chi-square | 1.783 | 1.405 | 0.674 | 0.358 | 2.125 | 2.824 | 6.568 |
| Hetero | 5.190 | 3.252 | 2.599 | 0.504 | 1.547 | 2.487 | 4.201 |

$n = 500$

| Distribution | $\tau$ | | | | | | |
|---|---|---|---|---|---|---|---|
| | .05 | 0.15 | 0.25 | 0.5 | 0.75 | 0.85 | 0.95 |
| Laplace | 1.640 | .368 | .107 | .053 | .080 | .368 | 1.598 |
| Normal | .476 | .098 | .075 | .051 | .078 | .311 | 1.623 |
| Student $t$ | 1.906 | .237 | .086 | .047 | .058 | .262 | 2.026 |
| Chi-square | .158 | .050 | .053 | .070 | .166 | .866 | 2.647 |
| Hetero | 1.296 | .416 | .207 | .093 | .118 | .282 | 1.108 |

Table 2: The mean squared errors of the estimated conditional quantiles in all five designs when the size of training data is $n = 100$ and 500. Hetero refers to the design of heteroskedastic errors.

pirable particles and inhalable coarse particles, with a diameter greater than 2.5 and less than 10 micrometers. Perez and Reyes (2002) argued that these subtypes are so small that they are able to penetrate the respiratory tract of humans. As a result, normal functions of the respiratory system might be impaired by them, with adverse effects on the health of individuals. According to Perez and Reyes (2002), if the maximal value of the 24-hour moving average of $PM_{10}$ concentrations is above 240 $\mu$g/m$^3$, there may be a serious risk to human health. Thus air quality prediction is vital to management of human health, especially the respiratory system. There has been extensive research on prediction of the average of daily $PM_{10}$ concentrations (see Chaloulakou *et al.*, 2003; Hooyberghs *et al.*, 2005; Osowski and Garanty, 2007 for instance). However, considerably less attention has been drawn to prediction of the maximum of daily $PM_{10}$ concentrations, which was used to classify days into different categories of air quality in Santiago, the capital of Chile (Perez and Reyes, 2006).

Hörmann *et al.* (2015) carried out a dynamic functional principal component analysis of $PM_{10}$ data collected in Graz, Austria. This data set consists of measurements of $PM_{10}$ concentrations (in $\mu$g/m$^3$) collected every half hour, from October 1st, 2010 to March 31st, 2011, and it is available in the *ftsa* package (Shang, 2019). They suggested performing a square-root transformation to stabilize the variance and mitigate heavy-

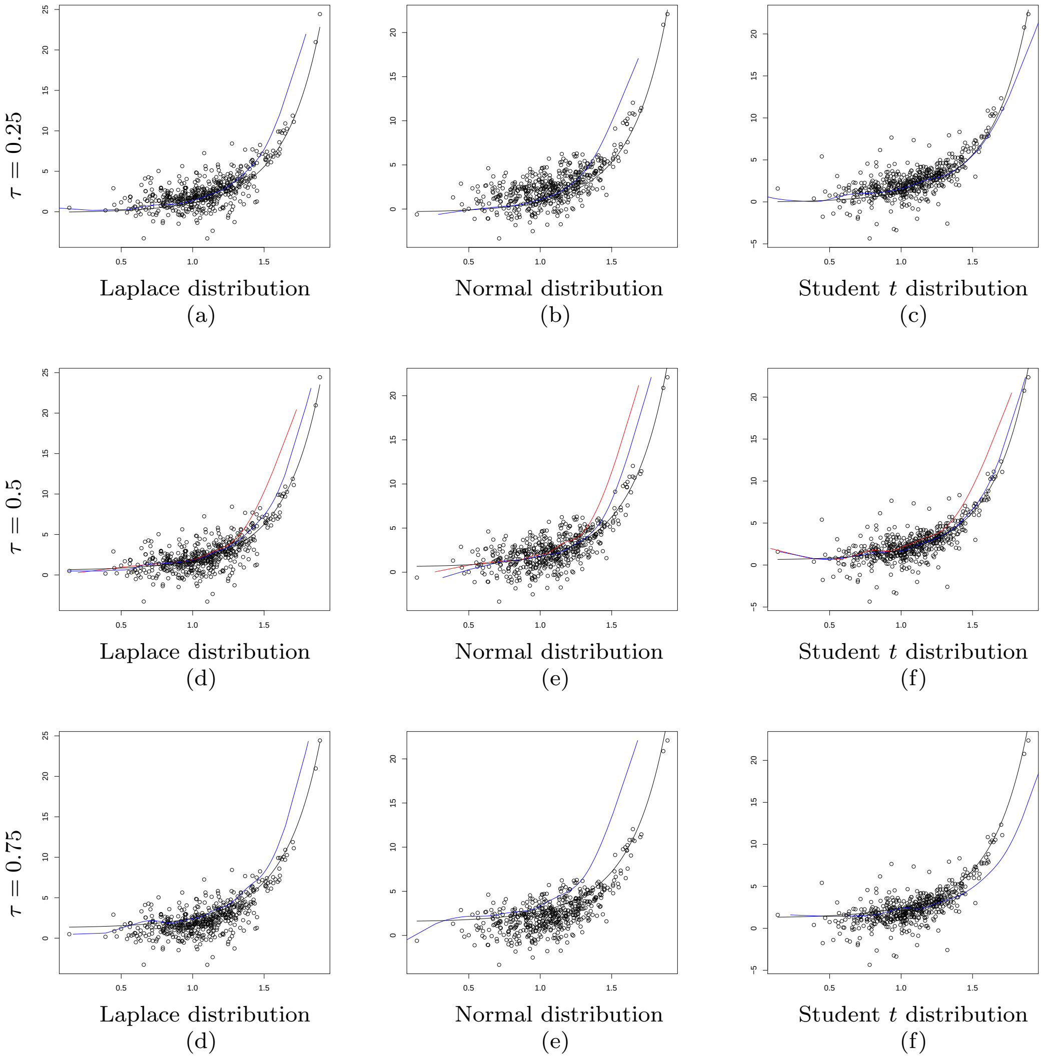
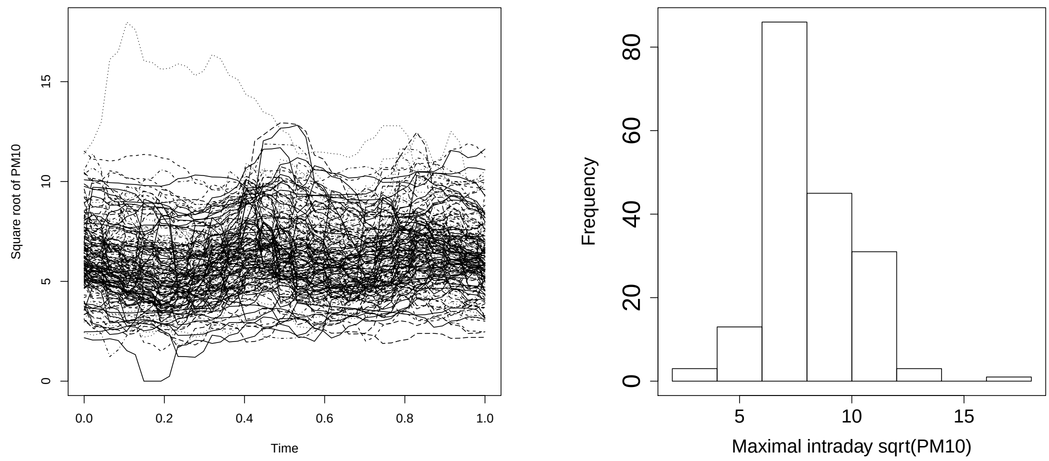



Fig. 1: Estimators of the link function for $\tau = 0.25, 0.5, 0.75$ in the three designs, respectively. In each panel, the black line denotes the true link function while the blue line represents the estimator from FsiQ. In the middle row, the red lines denote the estimated link function from FSiM.

tailed observations. Actually this is a commonly used technique to stabilize the variance when handling air pollutant concentrations (Sahu *et al.*, 2007; Berrocal *et al.*, 2010), and adopted by Reich *et al.* (2011) in the context of quantile regression. The profiles of measurements of $PM_{10}$ concentrations after the square-root transformation are displayed in the left panel of Figure 2, where the time scale has been transformed to [0, 1]. Heavy-tailed observations still exist after the transformation. This is further justified by the right panel of Figure 2; it depicts the histogram of maximal values of square root of intraday $PM_{10}$ concentrations, sqrt($PM_{10}$) in short. Our primary interest here is to predict the maximum sqrt($PM_{10}$) using the $PM_{10}$ trajectory of the last day. 15 Fourier basis functions are employed to transform the raw observations into functional data, as suggested by Hörmann *et al.* (2015). This transformation can be implemented with the *Data2fd* function in the **fda** package.

We use the GACV proposed in Section 2.3 to select both the number of B-spline basis functions and the smoothing parameter $\lambda$ in (3) to fit the functional single index quantile regression. The estimators of the index function and the link function are obtained by solving (3) with the proposed generalized profiling approach. To obtain a comprehensive understanding of the proposed methodology in estimating conditional quantiles, we consider seven quantile levels which are the same as in the simulation studies. The estimated index functions and link functions for $\tau = 0.5$ and $0.75$ are shown in Fig-




(a) (b)

Fig. 2: (a) Profiles of half-hourly square root of $PM_{10}$ concentrations from October 1st, 2010 to March 31st, 2011. (b)Histogram of maximal values of square root of intraday $PM_{10}$ concentrations.

ure 3, while the estimated link functions for $\tau = 0.05$, 0.15, 0.25, 0.85, and 0.95 are shown in Figure S5. To explore the variability of the estimated index functions, we employ the bootstrap method to generate a 95% pointwise confidence interval. To be more specific, we first use the whole data set to obtain the estimated link and index functions for each $\tau$. Let $\hat{u}_i = \int X_i(t)\hat{\beta}(t)$ denote the estimate projection for each subject $i$. We resample $(X_i(t), y_i)$'s with replacement, and implement the proposed method on this sample to obtain estimated conditional $\tau$th quantiles at $\hat{u}_i$'s. This enables us to acquire both the 0.05 and 0.975 quantiles of estimated conditional $\tau$th quantile at $\hat{u}_i$'s. We then use these quantiles to construct the pointwise confidence intervals for the estimated link function. We also employ the bootstrap method to quantify the uncertainty of the estimated index function, but now the time points $t$'s are fixed for pointwise evaluation.

From Figure 3 we find that the shape of the index function for $\tau = 0.5$ is different from that for $\tau = 0.75$. This suggests that the functional single index mean regression is inadequate to characterize the conditional distribution of the maximal value of intraday sqrt($PM_{10}$) concentrations. Additionally, the bumpy shapes of index functions suggest that the contribution made by the $PM_{10}$ concentrations of the last day in prediction of the maximum sqrt($PM_{10}$) varies over time. The linear shape of the estimated index functions suggests that there is a linear relationship between the $PM_{10}$ profile of the last day and the quantiles of the maximal $PM_{10}$ concentrations. Last but not least, Figure 3 and Figure S5 suggest that the proposed method manages to generate a reasonable estimate for moderate levels of conditional quantiles, but the standard errors for estimated extreme quantiles like $\tau = 0.05$ and 0.95 are way too large.

As shown in simulation studies, FSiQ compares favorably with FSiM in estimation of the target function when the scalar response is heavy-tailed distributed. We further compare these two methods in terms of prediction accuracy in this real example. Five-fold cross-validation is adopted to compute the mean squared errors (MSPEs) defined by

$$\text{MSPE} = \frac{\sum_{i \in \text{test set}} \left[ Y_i - \hat{f}\left\{ \int_0^1 X_i(t)\hat{\beta}(t)dt \right\} \right]^2}{|\text{test set}|},$$

where $|\text{test set}|$ denotes the cardinal number of a test set and $\hat{f}$ and $\hat{\beta}$ are obtained by taking $\tau = 0.5$ in FSiQ. We take $\tau = 0.5$ since the estimated conditional median, like the conditional mean, is a reasonable estimate of the center of the response. The whole data set is randomly splitted 100 times and the distributions of the corresponding MSPEs obtained from FSiM and FSiQ are displayed in Figure 4. Obviously FSiQ outperforms FSiM in predicting the maximal intraday sqrt($PM_{10}$) of the next day with the profile of the current day. This further implies that the functional single index mean regression is less flexible in predicting the maximal intraday sqrt($PM_{10}$) than the quantile regression model. A possible reason for this remarkable gap in prediction accuracy is that the distribution of the maximal intraday sqrt($PM_{10}$) is skewed and some extremely large values exist in the response; these can be found in the right panel of Figure 2.

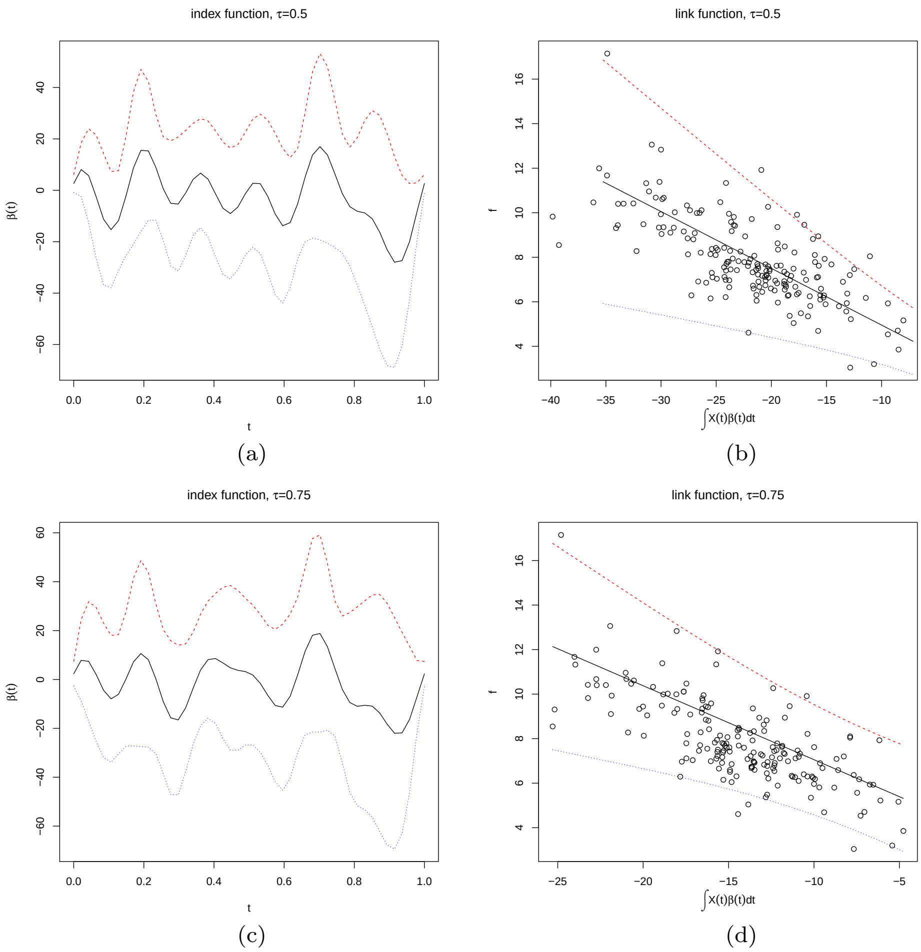



Fig. 3: The estimated index functions and link functions. Left two panels: the estimated index functions for the quantile $\tau = 0.5$ and $\tau = 0.75$ and their 95% pointwise confidence intervals. Right two panels: the estimated link functions for $\tau = 0.5$ and $\tau = 0.75$ and their 95% pointwise confidence intervals.

## 5 Conclusion

Scalar-on-function regression models play a significant role in characterizing the relationship between a scalar response and a functional covariate. There has been an extensive research on modelling the conditional mean structure of a scalar response given a functional covariate. These models such as functional single index regression models and fully nonparametric functional models are widely applied in practice for prediction. However, they are not able to provide an adequate characterization of the conditional distribution of the scalar response. Moreover, their prediction accuracy might be impaired when the distribution of the scalar response is skewed or heavy tailed.

In the paper, we have proposed a functional single index quantile regression model which concerns the conditional quantile of a scalar response. It, therefore, enables us to investigate the conditional distribution of the response in details. A generalized profiling method is proposed to fit this model. This method consists of two nested levels of optimization with different target functions. In the inner level, we use B-spline basis functions to approximate the index function and represent the estimated coefficient vector in terms of the link function. In the outer level, we resort to the representer theorem to simplify the minimization task that searches a minimizer of a regularized empirical risk function from an infinite dimensional functional space. Simulation studies are conducted to compare the finite sample performances of the proposed estimator and that for the functional single index mean regression model. We find that the performance of the proposed estimator is less susceptible to extremely large outcomes compared with its counterpart. Their performances in prediction accuracy is further compared in an real example, which concerns using the current intraday $PM_{10}$ concentrations to predict the maximal value of the next day. The

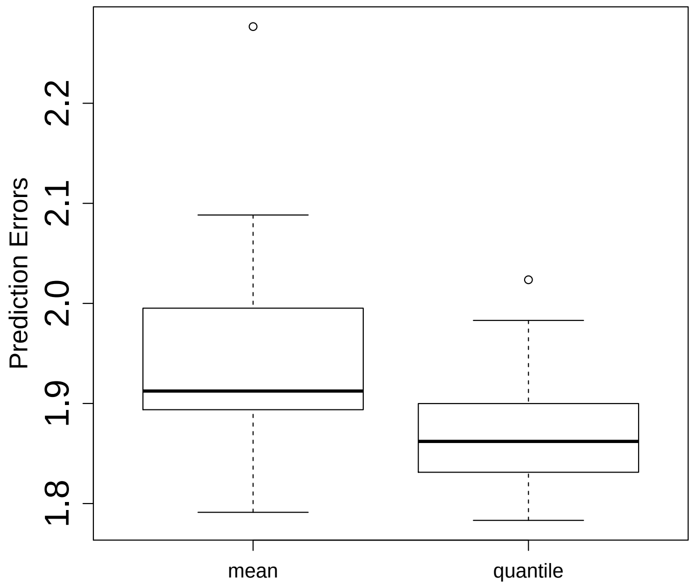



Fig. 4: Boxplots of mean squared prediction errors (MSPE) of FSiM and FSiQ across the 100 random splits. The left side and the right side represent MSPEs for FSiM and FSiQ, respectively. The FSiQ model uses the 50% quantile to predict the response.

comparison result demonstrates that the proposed estimator is superior to its counterpart in terms of prediction accuracy when the scalar response is skewed distributed and contains numerous extremely large values. Beside the real example shown in the paper, this model can also be applied to study the relationship between proportions of contents like protein and fat and Near-infrared spectra in Tecator data, which are available at `http://lib.stat.cmu.edu/datasets/tecator` and the relationship between the ADHD index and functional magnetic resonance imaging (FMRI) signals. Both of these two problems are of critical importance in practice.

It would be desirable to show that the estimated link function and index function enjoy some nice properties such as consistency and asymptotic normality. The theoretical framework proposed in Li *et al.* (2007) could be a useful tool to prove such properties. However, since a single index model is considered in our framework and the estimation of the two functions in the model is interdependent, it is difficult to establish asymptotic properties for these two estimators, respectively. We will develop tools to study asymptotic properties of these estimators in the future.

**Supplementary Material**

A supplementary document includes additional simulation study results and real data results. The R codes for the application and simulation studies are also available at `https://github.com/caojiguo/FunSIQ`.

**References**

Ait-Saïdi, A., Ferraty, F., Kassa, R., and Vieu, P. (2008). Cross-validated estimations in the single-functional index model. *Statistics*, **42**(6), 475–494.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics*, **15**(2), 176–197.

Burba, F., Ferraty, F., and Vieu, P. (2009). k-Nearest Neighbour method in functional nonparametric regression. *Journal of Nonparametric Statistics*, **21**(4), 453–469.

Cardot, H., Crambes, C., and Sarda, P. (2005). Quantile regression when the covariates are functions. *Nonparametric Statistics*, **17**(7), 841–856.

Chaloulakou, A., Grivas, G., and Spyrellis, N. (2003). Neural network and multiple regression models for

$PM_{10}$ prediction in Athens: a comparative assessment. *Journal of the Air & Waste Management Association*, **53**(10), 1183–1190.

Chen, D., Hall, P., and Müller, H.-G. (2011). Single and multiple index functional regression models with nonparametric link. *The Annals of Statistics*, **39**(3), 1720–1747.

Chen, K. and Müller, H.-G. (2012). Conditional quantile analysis when covariates are functions, with application to growth data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**(1), 67–89.

Dai, X., Müller, H.-G., and Yao, F. (2017). Optimal Bayes classifiers for functional data and density ratios. *Biometrika*, **104**(3), 545–560.

Delaigle, A. and Hall, P. (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**(2), 267–286.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. New York: Springer.

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

Ferraty, F., Keilegom, I. V., and Vieu, P. (2010). On the validity of the bootstrap in non-parametric functional regression. *Scandinavian Journal of Statistics*, **37**(2), 286–306.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction, 2nd Edition*. New York: Springer.

Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., and Brasseur, O. (2005). A neural network forecast for daily average $PM_{10}$ concentrations in Belgium. *Atmospheric Environment*, **39**(18), 3279–3289.

Hörmann, S., Kidziński, Ł., and Hallin, M. (2015). Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **77**(2), 319–348.

Horowitz, J. L. (2009). *Semiparametric and nonparametric methods in econometrics*. New York: Springer.

James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(3), 411–432.

Kato, K. (2012). Estimation in functional linear quantile regression. *The Annals of Statistics*, **40**(6), 3108–3136.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of mathematical analysis and applications*, **33**(1), 82–95.

Koenker, R. (2005). *Quantile Regression*. Cambridge: Cambridge University Press.

Koenker, R., Ng, P., and Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, **81**(4), 673–680.

Li, Y., Liu, Y., and Zhu, J. (2007). Quantile regression in reproducing kernel Hilbert spaces. *Journal of the American Statistical Association*, **102**(477), 255–268.

Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, **33**(2), 774–805.

Müller, H.-G. and Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association*, **103**(484), 1534–1544.

Müller, H.-G., Wu, Y., and Yao, F. (2013). Continuously additive models for nonlinear functional regression. *Biometrika*, **100**(3), 607–622.

Nychka, D., Gray, G., Haaland, P., Martin, D., and O'connell, M. (1995). A nonparametric regression approach to syringe grading for quality improvement. *Journal of the American Statistical Association*, **90**(432), 1171–1178.

Osowski, S. and Garanty, K. (2007). Forecasting of the daily meteorological pollution using wavelets and support vector machine. *Engineering Applications of Artificial Intelligence*, **20**(6), 745–755.

Perez, P. and Reyes, J. (2002). Prediction of maximum of 24-h average of PM10 concentrations 30 h in advance in Santiago, Chile. *Atmospheric Environment*, **36**(28), 4555–4561.

Perez, P. and Reyes, J. (2006). An integrated neural network model for PM10 forecasting. *Atmospheric Environment*, **40**(16), 2845–2851.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York, second edition.

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society Series B*, **69**(5), 741–796.

Reich, B. J., Fuentes, M., and Dunson, D. B. (2011). Bayesian spatial quantile regression. *Journal of the American Statistical Association*, **106**(493), 6–20.

Reiss, P. T., Goldsmith, J., Shang, H. L., and Ogden, R. T. (2017). Methods for scalar-on-function regression. *International Statistical Review*, **85**(2), 228–249.

Sahu, S. K., Gelfand, A. E., and Holland, D. M. (2007). High-resolution space–time ozone modeling for assessing trends. *Journal of the American Statistical*

*Association*, **102**(480), 1221–1234.

Shang, H. (2019). *ftsa: Functional Time Series Analysis*. R package version 5.5.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, **13**(2), 689–705.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics*, **13**(4), 1378–1402.

Yuan, M. (2006). GACV for quantile smoothing splines. *Computational statistics & data analysis*, **50**(3), 813–829.