Journal of the Korean Statistical Society 47 (2018) 436–449

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Estimation and variable selection for partially functional linear models

Jiang Du [a], Dengke Xu [b], Ruiyuan Cao [a,*]

[a] College of Applied Sciences, Beijing University of Technology, Beijing, 100124, China
[b] Department of Statistics, Zhejiang Agriculture and Forestry University, Hangzhou, 311300, China

ARTICLE INFO

Article history:
Received 23 August 2017
Accepted 7 May 2018
Available online 28 May 2018

AMS 2000 subject classifications:
62G05
62G20

Keywords:
Partially functional linear regression model
Variable selection
Composite quantile regression
Oracle property

ABSTRACT

In this paper, a new estimation procedure based on composite quantile regression and functional principal component analysis (PCA) method is proposed for the partially functional linear regression models (PFLRMs). The proposed estimation method can simultaneously estimate both the parametric regression coefficients and functional coefficient components without specification of the error distributions. The proposed estimation method is shown to be more efficient empirically for non-normal random error, especially for Cauchy error, and almost as efficient for normal random errors. Furthermore, based on the proposed estimation procedure, we use the penalized composite quantile regression method to study variable selection for parametric part in the PFLRMs. Under certain regularity conditions, consistency, asymptotic normality, and Oracle property of the resulting estimators are derived. Simulation studies and a real data analysis are conducted to assess the finite sample performance of the proposed methods.

1. Introduction

With the rapid development of technology in computation and measurement, scientists usually possess data containing information about curves, surfaces, or something else varying with continuous variables. Data with such type of structure, termed as functional data, have attracted great interests in various fields, such as biomedical studies, finance, environmental science, and engineering. Therefore, many useful statistical models have been proposed by taking into account the functional feature of such data. For example, Yao, Müller, and Wang (2005) proposed a functional linear model (FLM) to analyze the primary biliary cirrhosis data. Müller and Stadtmüller (2005) presented a generalized functional regression linear model and applied it to discriminate between short-lived and long-lived medflies. Sentürk and Müller (2010) proposed a functional varying coefficient model which included a smooth history index function to capture the effects of the recent past of the predictor on current response. An extensive review of functional data analysis can be found in Greven and Scheipl (2017), Horváth and Kokoszka (2012), Morris (2015) and Ramsay and Silverman (2002, 2005). In recent years, partially functional linear models, as a special model for functional data, have been extensively investigated. Based on mean regression, Shin (2009) developed estimation methods for the finite parameter and functional slope parameter, and established the asymptotic properties of the resulting estimators. Further, Lu, Du, and Sun (2014) extended the result of Shin (2009) to quantile regression. Kong, Xue, Yao, and Zhang (2016) studied variable selection for partially function models. For more recent developments, the interested readers can refer to Reiss, Goldsmith, Shang, and Ogden (2017) and Wang, Chiou, and Müller (2016).

* Corresponding author.
E-mail address: caoruiyuan@bjut.edu.cn (R. Cao).

https://doi.org/10.1016/j.jkss.2018.05.002
1226-3192/

In addition, since first proposed by Koenker and Bassett (1978), quantile regression has emerged as an important statistical methodology. By estimating various conditional quantile functions, quantile regression differs from classical least squares regression in that, it shifts the focus away from analyzing the effects of covariates on the conditional mean to study their effects on location, scale, and shape of the distribution of the response variable. It has been used in wide range of applications including economics, biology and finance. A comprehensive review of the theory and most recent developments of quantile regression can be found in Koenker (2005). Intuitively, the composite quantile regression (CQR) should provide estimation efficiency gain over a single quantile regression, see Zou and Yuan (2008). A composite quantile regression model assumes that there exist common covariate effects in a range of quantiles such that the quantile levels only differ in terms of the intercept. From a more general regression perspective, composite quantile regression seeks to model a set of parallel regression curves, and thus it can be viewed as a compromise between a set of quantile regression curves with different intercepts and slopes and a single summary regression curve. Kai, Li, and Zou (2010) proposed the local polynomial CQR estimators for estimating the nonparametric regression function and its derivative. It is shown that the local CQR method can significantly improve the estimation efficiency of the local least squares estimator for commonly-used non-normal error distributions. Furthermore, Kai, Li, and Zou (2011) studied semiparametric CQR estimates for semiparametric varying-coefficient partially linear model. They compared CQR with least squares and quantile regression, and the results showed that CQR outperformed both least squares and quantile regression. Although there is a growing literature on CQR, it is relatively sparse for partially functional linear models. It is therefore our impetus for applying CQR method to partially functional linear regression models. Under mild conditions, we establish asymptotic normality for the estimators of the parametric components, and obtain the rate of convergence for the estimator of the slope function.

It is well known that variable selection is a crucial issue in regression analysis. In practice, a number of variables are available in an initial analysis, but many of them may not be significant and should be excluded from the final model to increase the accuracy of the prediction. Traditional variable selection methods such as stepwise regression and best subset selection are computationally infeasible when the number of predictors is large. Therefore, various shrinkage methods, such as the LASSO (Tibshirani, 1996), the adaptive LASSO (Zou, 2006) and the SCAD (Fan & Li, 2001), have gained much attention in recent years. Just as in ordinary data analysis, variable selection is also an important aspect in functional data analysis. The functional data suffers from high dimensionality and multicollinearity among functional predictors. This could lead us to wrong model selection and hence wrong scientific conclusions. Collinearity also gives rise to issues of over-fitting and model misidentification. So it is very important to perform variable selection for regression model with a functional covariate. With sparsity, variable selection effectively identifies the subset of significant predictors, which improves the estimation accuracy and enhances the model interpretability. However, in the presence of outliers which are curves deviating from the remaining functional data, identifying significant scalar covariates effectively and correctly becomes even more challenging. Most recently, there has been increasing research work in variable selection for functional predictors in functional regression models. For example, Matsui and Konishi (2011) considered applying group SCAD regularization to the functional regression model with functional predictors and a scalar response, accomplishing parameter estimation and model selection simultaneously. Lian (2013) proposed a regularization method for shrinkage estimation of multiple functional linear regression models, and showed that it is consistent in estimation and variable selection. Gertheiss, Maity, and Staicu (2014) considered a penalized likelihood approach that combines selection of the functional predictors with estimation of the smooth effects for the chosen subset of predictors. Huang, Zhao, Wang, and Wang (2016) developed robust variable selection for functional multiple linear model based on penalized the least absolute deviation loss function. Using group-Lasso, Pannu and Billor (2017) proposed variable selection for functional multiple linear model via group Lasso penality.

Aneiros, Ferraty, and Vieu (2015) considered the variable selection for partially functional linear models with diverging number of covariates, and obtained the oracle property of the resulting variable selection method as well as the rate of convergence of estimator for the nonparametric slope function. Furthermore, Kong et al. (2016) developed the variable selection for partially functional linear models with ultra-high dimensional scalar predictors, and established the consistency and oracle properties of the resulting method under proper choices of the penalty functions and the regularization parameter. In addition, Yao, Sue-Chee, and Wang (2017) proposed variable selection for partially functional quantile regression, and established the large sample properties of the proposed method. However, the effectiveness of these methods may be lessened in the presence of outliers. Since the above mentioned variable selection techniques except Yao et al. (2017) are all based on minimizing the penalized residual sum of squares, they are known to be non-robust in nature. Thus, there is a need for a robust and highly effective variable selection method for the partially functional linear regression model.

In this paper, we intend to develop an estimation and variable selection method for partially functional linear regression models based on the composite quantile. Under some regularity conditions, we obtain the optimal convergence rate of the functional coefficient and establish the asymptotic normality of parametric component. Deriving the asymptotic properties of the penalized estimator is very challenging as we need to simultaneously deal with the non-smooth loss function and the functional predictor. To tackle these challenges, we combine the quadratic approximation with the functional principal component analysis.

The remainder of this paper is organized as follows. In Section 2, we first introduce the partially functional linear regression model, and propose the estimation and variable selection procedures for PFLRMs. Then, the large sample properties of the proposed estimators are given. Section 3 contains the simulation studies. Section 4 presents a real data analysis. Final remarks are given in Section 5. All the conditions and technical proofs are deferred to the Appendix.

2. Estimation and variable selection

2.1. Partially functional linear regression model

Consider the following partially functional linear regression model:

$$Y = \int_{\mathcal{T}} \beta(t)X(t)dt + Z^T\theta + \varepsilon, \tag{1}$$

where Y is a real-valued random variable defined on a probability space $(\Omega, \mathcal{B}, P)$, $Z = (Z_1, \dots, Z_p)^T$ is a p-dimensional random vector, $\{X(t) : t \in \mathcal{T}\}$ is a zero mean, second-order (i.e., $E\big(X(t)\big) = 0$ and $E|X(t)|^2 < \infty$ for all $t \in \mathcal{T}$) stochastic process defined on $(\Omega, \mathcal{B}, P)$ with sample paths in $L_2(\mathcal{T})$, the Hilbert space containing square integrable functions defined on $\mathcal{T}$ with inner product $\langle x, y\rangle = \int_{\mathcal{T}} x(t)y(t)dt$, $\forall x, y \in L_2(\mathcal{T})$ and norm $\|x\| = \langle x, x\rangle^{1/2}$, $\beta(t)$ is a square integrable function on $\mathcal{T}$, and ε is a random error independent of Z and X. Without loss of generality, we suppose throughout that $\mathcal{T} = [0, 1]$.

2.2. Estimation method based on the composite quantile

Let $\{(X_i, Z_i, Y_i),\ i = 1, \dots, n\}$ be an independent and identically distributed sample which is generated from model (1). Define the covariance function and the empirical covariance function for functional predictor $X(t)$, respectively, as

$$K(s, t) = \mathrm{Cov}(X(t), X(s))$$

and

$$\hat{K}(s, t) = \frac{1}{n}\sum_{i=1}^{n} X_i(s)X_i(t).$$

It is well known that the covariance function K defines a linear operator which maps a function f to Kf given by $(Kf)(u) = \int K(u, v)f(v)dv$. We shall assume that the linear operator with kernel K is positive definite. Let $\lambda_1 > \lambda_2 > \cdots > 0$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \hat{\lambda}_n \geq \hat{\lambda}_{n+1} = \hat{\lambda}_{n+2} = \cdots = 0$ be the ordered eigenvalue sequences of the linear operators with kernels K and $\hat{K}$, $\{\phi_j\}$ and $\{\hat{\phi}_j\}$ be the corresponding orthonormal eigenfunction sequences, respectively. It is clear that the sequences $\{\phi_j\}$ and $\{\hat{\phi}_j\}$ each forms an orthonormal basis in $L^2([0, 1])$. Then, the spectral decompositions of the covariance functions K and $\hat{K}$ can be written as

$$K(s, t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(t)$$

and

$$\hat{K}(s, t) = \sum_{j=1}^{\infty} \hat{\lambda}_j \hat{\phi}_j(s)\hat{\phi}_j(t),$$

respectively.

According to the Karhunen–Loève representation, we have

$$X(t) = \sum_{i=1}^{\infty} \xi_i \phi_i(t)$$

and

$$\beta(t) = \sum_{i=1}^{\infty} \gamma_i \phi_i(t), \tag{2}$$

where the ξ_i are uncorrelated random variables with mean 0 and variance $E[\xi_i^2] = \lambda_i$, and $\gamma_i = \langle\beta, \phi_i\rangle$, for more details see Ramsay and Silverman (2005). Substituting (2) into model (1), we can get

$$Y = \sum_{j=1}^{\infty} \gamma_j \langle\phi_j, X\rangle + Z^T\theta + \varepsilon. \tag{3}$$

Therefore, the regression model in (3) can be well approximated by

$$Y \approx \sum_{j=1}^{m} \gamma_j \langle\phi_j, X\rangle + Z^T\theta + \varepsilon, \tag{4}$$

where $m \leq n$ is the truncation level that trades off approximation error against variability and typically diverges with n. Replace ϕ_j by $\hat{\phi}_j$ for $j = 1, \ldots, m$, model (4) can be rewritten as

$$Y \approx Z^T\theta + U^T\gamma + \varepsilon,$$

where $U = \{\langle X, \hat{\phi}_j\rangle\}_{j=1,\ldots,m}$, $\gamma = (\gamma_1, \ldots, \gamma_m)^T$.

Then, following Zou and Yuan (2008), we denote the $100\tau_k\%$ quantile of ε in model (1) by b_{τ_k}, and assume that $\tau_1 < \tau_2 < \cdots < \tau_M < 1$. In particular, we use the equally spaced quantiles, i.e., $\tau_k = k/(M+1)$ for $k = 1, 2, \ldots, M$. For brevity, assume that the density function of ε is non-vanishing anywhere. Hence, b_{τ_k} is uniquely defined for any $0 < \tau_k < 1$. Let $\rho_\tau(t) = t(\tau - I(t < 0))$ be the check function at $\tau \in (0, 1)$. Based on data (X_i, Z_i, Y_i), $i = 1, \ldots, n$ generated from model (1), we have (U_i, Z_i, Y_i), $i = 1, \ldots, n$ by functional principal component analysis. Using Eq. (4), we can estimate θ and γ by minimizing the following CQR loss function:

$$\sum_{k=1}^{M}\sum_{i=1}^{n}\rho_{\tau_k}(Y_i - b_{\tau_k} - Z_i^T\theta - U_i^T\gamma). \tag{5}$$

Denoted the solutions by $\hat{\theta}^{CQR}$ and $\hat{\gamma}$. Then, it can be obtained that

$$\hat{\beta}(\cdot) = \sum_{j=1}^{m}\hat{\gamma}_j\hat{\phi}_j(\cdot). \tag{6}$$

To establish asymptotic properties of the estimator for θ and β, the following assumptions are needed. Throughout this paper, the constant C may change in different situations for convenience.

C1: $E\|X\|^4 < C < \infty$.

C2: For each j, $E[U_j^4] \leq C\lambda_j$. For the eigenvalues λ_j and Fourier coefficients γ_j, it is required that $\lambda_j - \lambda_{j+1} \geq C^{-1}j^{-a-1}$ and $|\gamma_j| \leq Cj^{-b}$ for $j > 1$, $a > 1$ and $b > a/2 + 1$.

C3: The random vector Z is bounded in probability. One complicated issue for PFLRMs comes from the dependence between Z and X. Similar to Lu et al. (2014) and Shin (2009), let $Z = \eta + \langle g, X\rangle$, where $\eta = (\eta_1, \ldots, \eta_p)^T$ is zero-mean random vector, $g = (g_1, \ldots, g_p)^T$ with $g_j \in L_2([0, 1])$, $j = 1, \ldots, p$.

C4: $E[\eta] = 0$ and $E[\eta\eta^T] = \Sigma$. Furthermore, Σ is a positive definite matrix.

C5: ε has cumulative distribution function $F(\cdot)$ and density function $f(\cdot)$. For each p-vector $\boldsymbol{u}$,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\int_0^{u_0+\eta_i^T\boldsymbol{u}}[F(a + t/\sqrt{n}) - F(a)]dt = \frac{1}{2}f(a)(u_0, \boldsymbol{u}^T)\begin{bmatrix}1 & 0\\ 0 & \Sigma\end{bmatrix}(u_0, \boldsymbol{u}^T)^T,$$

where $\eta_i = Z_i - E(Z_i|X_i)$. Let $c_k = F^{-1}(\tau_k)$ and C^* be an $M \times M$ diagonal matrix with $C^*_{jj} = f(c_j)$. Write $c = C^*1$, $\varpi = \sum_{k,l=1}^{M}\min(\tau_k, \tau_l)(1 - \max(\tau_k, \tau_l))/(1^TC^*1)^2$.

Remark 1. Conditions C1 and C2 are very common in functional linear regression model (see Cai and Hall (2006), Hall and Horowitz (2007)). Condition C2 ensures that β is sufficiently smooth and can be approximated by functions, which is often assumed in asymptotic analysis of partially functional linear regression problems (see Lu et al. (2014) and Shin (2009)) and is required to achieve the optimal convergence rate of β. The boundness assumption in Condition C3 is made for convenience. It suffices to assume that Z has bounded fourth moment which will complicate the technical proof. Condition C4 controls the asymptotic properties of the estimator $\hat{\theta}$; for example, He, Zhu, and Fung (2002) and Lu et al. (2014). Condition C5 is very common in composite quantile regression, see Zou and Yuan (2008).

For convenience, we denote the true functional regression coefficient as β_0 and the true parametric regression coefficient as θ_0. The following theorem describes the asymptotic properties of the proposed estimators.

Theorem 1. *Suppose tuning parameter m satisfies $m \sim n^{1/(a+2b)}$. Under Assumptions C1–C5, it has the following:*

(i) $\|\hat{\beta} - \beta_0\|^2 = O_p\left(n^{-(2b-1)/(a+2b)}\right)$;

(ii) $\sqrt{n}(\hat{\theta}^{CQR} - \theta_0) \to N\left(0, \varpi\Sigma^{-1}\right)$.

Remark 2. For the estimator of slope function $\beta(t)$, the convergence rate is similar to Shin (2009). This result shows that the existence of random vector Z in the model does not change the rate of convergence of the estimator of slope function. For the parametric vector θ, the asymptotic normality of the estimator $\hat{\theta}^{CQR}$ is similar to the result of Theorem 1 in Zou and Yuan (2008). It indicates that the functional variable in the model does not influence the asymptotic behavior of $\hat{\theta}^{CQR}$.

2.3. Variable selection based on the composite quantile

In this subsection, we propose a variable selection procedure for composite quantile regression in PFLRMs based on the adaptive LASSO penalized method by Zou (2006), and then present the Oracle property. The adaptive LASSO penalized composite quantile regression estimator in PFLRMs, denoted by $\hat{\theta}^{ALCQR}$, is the minimizer of the following function:

$$\sum_{k=1}^{M}\sum_{i=1}^{n}\rho_{\tau_k}(Y_i - b_{\tau_k} - Z_i^T\theta - U_i^T\gamma) + \varrho\sum_{j=1}^{p}\frac{|\theta_j|}{|\hat{\theta}_j^{CQR}|^2}. \quad (7)$$

Then, obtain the estimator of $\beta(\cdot)$ as $\hat{\beta}(\cdot) = \sum_{j=1}^{m}\hat{\gamma}_j\hat{\phi}_j(\cdot)$, where $\hat{\gamma}_j$ is obtained by minimizing function (7).

In this paper, we focus on adaptive Lasso penalty function. However, the proposed procedure can be easily modified to accommodate other penalty functions, such as SCAD proposed by Fan and Li (2001) and MCP proposed by Zhang (2010). In addition, we show that our proposed variable selection procedure can identify the true model and estimate the functional coefficient and constant regression coefficients simultaneously. The following theorem gives the rate of convergence of the estimator $\hat{\beta}$ and the Oracle property of estimator $\hat{\theta}$. Denote by $\theta_0 = (\theta_{0I}^T, \theta_{0II}^T)^T$ the true value of the regression coefficients. Without loss of generality, assume that $\theta_{0II} = 0$ and elements in θ_{0I} are nonzero. According to the partition of θ_0, we write $\eta = (\eta_I^T, \eta_{II}^T)^T$. Let $E[\eta_{II}\eta_{II}^T] = \Sigma_{II}$.

Theorem 2. *Suppose tuning parameter m satisfies $m \sim n^{1/(a+2b)}$, $\varrho/\sqrt{n} \to 0$, and $\frac{\varrho}{m^{a+1}} \to \infty$. Denote $I = \{j : \theta_{0j} \neq 0\}$. Under conditions C1–C5, it has the following:*

(i) Convergence rate: $\|\hat{\beta} - \beta_0\|^2 = O_p\left(n^{-(2b-1)/(a+2b)}\right)$;

(ii) Sparsity: $P(\{j : \hat{\theta}_j^{ALCQR} \neq 0\} = I) \to 1$;

(iii) Asymptotic normality: $\sqrt{n}(\hat{\theta}_I^{ALCQR} - \theta_{0I}) \to N\left(0, \varpi\Sigma_{II}^{-1}\right)$.

3. Simulation studies

3.1. Performance of the proposed estimator

In this subsection, we investigate the finite sample properties of the proposed estimation procedure via Monte Carlo simulation. The data sets are generated from the following model:

$$Y = Z^T\theta + \int_0^1 X(t)\beta(t)dt + \varepsilon, \quad (8)$$

where Z follows multivariate normal distribution $N(0, \Sigma_Z)$ with $(\Sigma_Z)_{i,j} = 0.5^{|i-j|}$ for $i, j = 1, 2$, $\theta = (1, 2)^T$. For the functional linear component, we take the same form as Shin (2009), that is, $\beta(t) = \sqrt{2}\sin(\pi t/2) + 3\sqrt{2}\sin(3\pi t/2)$ and $X(t) = \sum_{j=1}^{100}\xi_j\phi_j(t)$, where ξ_js are distributed as independent normal with mean 0 and variance $\lambda_j = ((j-0.5)\pi)^{-2}$ and $\phi_j(t) = \sqrt{2}\sin((j-0.5)\pi t)$. As for the random error, the following six cases are considered:
Case 1. ε follows a standard normal distribution.
Case 2. ε follows a $t(3)$ distribution. This yields a model with heavy-tail.
Case 3. ε follows a standard Cauchy distribution. This yields a model in which the expectation of the response does not exist.
Case 4. ε follows a contaminated normal distribution with 1% outliers from the standard Cauchy distribution.
Case 5. ε follows a log-normal distribution.
Case 6. ε follows a contaminated normal distribution with 20% outliers from $t(3)$ distribution.

For comparison, we also present the performance of the least squares estimator (LS). Similar to Zou and Yuan (2008), we use the quantiles $\tau_k = \frac{k}{20}$ for $k = 1, 2, \ldots, 19$, in the composite quantile regression. Replicate the simulation for 500 times, each consisting of $n = 200$ and 400 random samples. In addition, we use the following SIC and BIC for choice of tuning parameters m for composite quantile regression and the least squares regression:

$$\mathrm{SIC}(m) = \log\left\{\sum_{k=1}^{M}\sum_{i=1}^{n}\rho_{\tau_k}\left(Y_i - \hat{b}_{\tau_k(m)} - Z_i^T\hat{\theta}_{(m)} - U_i^T\hat{\gamma}_{(m)}\right)\right\} + \frac{\log(n)}{2n}(m+p+M),$$

where $p = 2$, $\hat{b}_{\tau_k(m)}$, $\hat{\theta}_{(m)}$ and $\hat{\gamma}_{(m)}$ are the estimators obtained from minimizing (5) with m eigenfunctions; or

$$\mathrm{BIC}(m) = \log\left\{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - Z_i^T\hat{\theta}_{(m)} - U_i^T\hat{\gamma}_{(m)}\right)^2\right\} + \frac{\log(n)}{2n}(m+p),$$

here $\hat{\theta}_{(m)}$ and $\hat{\gamma}_{(m)}$ are the least squares estimator with m eigenfunctions; see He et al. (2002), Lu et al. (2014), Shin (2009), and Wang, Kai, and Li (2009) for a similar criterion to choose the tuning parameters. The accuracy of nonparametric estimators is examined by the mean squared errors (MSE) which is defined as follows:

$$\mathrm{MSE} = \frac{1}{500}\sum_{i=1}^{500}\left\{\frac{1}{N}\sum_{s=1}^{N}\left[\hat{\beta}_i(t_s) - \beta(t_s)\right]^2\right\},$$

Table 1
Finite sample performance of the proposed estimators.

Case	Method	$n = 200$					$n = 400$				
		θ_1		θ_2		β	θ_1		θ_2		β
		Bias	SSD	Bias	SSD	MSE	Bias	SSD	Bias	SSD	MSE
1	LS	−0.3	8.0	−0.5	8.1	0.9	−0.2	5.8	−0.3	6.0	0.4
	CQR	−0.3	8.4	−0.3	8.4	0.9	−0.2	6.0	−0.2	6.2	0.5
2	LS	−0.3	14.2	0.6	14.6	1.9	0.4	10.0	0.1	9.9	0.8
	CQR	0.3	10.4	0.3	10.6	1.1	0.4	7.2	0.0	7.5	0.5
3	LS	0.1	2.5	0.0	0.3	236.7	0.0	0.6	0.0	0.3	0.4
	CQR	−0.1	16.1	0.6	16.6	3.2	−0.1	10.9	0.3	11.0	1.6
4	LS	1.8	70.8	−0.5	109.4	2.4	−0.2	41.9	−2.4	57.4	70.0
	CQR	−0.1	9.0	−0.5	8.9	0.9	−0.1	6.1	0.1	6.0	0.5
5	LS	−1.1	17.7	1.2	18.5	2.9	−0.1	12.9	0.3	12.9	1.2
	CQR	−0.3	7.4	0.2	7.4	0.9	0.0	5.0	−0.2	5.1	0.4
6	LS	−0.6	12.2	0.0	12.0	1.4	0.1	8.4	0.0	8.7	0.6
	CQR	−0.6	9.4	0.3	9.3	0.9	0.0	6.1	0.1	6.5	0.4

The entries are multiplied by 100, except the values in case 3 with LS, which are multiplied by 10^{-2}.

where t_s, $s = 1, \ldots, N$, are the grid points at which the function $\hat{\beta}_i(t)$ is evaluated, and subscript i implies the ith replication. In this simulation, $N = 200$. For the parametric component, the estimated error is computed by sample standard deviation (SSD).

Table 1 presents the finite sample properties of the proposed estimators. Following observations can be concluded:

(1) The performance of our proposed procedure becomes better with increase of sample size for all cases considered. In general, the bias is reasonably small, and the SSD and MSE decrease as the sample size increases.
(2) For standard Cauchy random error, as expected, the bias, SSD and MSE are extremely large, this is because the finite constant variance assumption, needed in least squares regression, is not satisfied for classical Cauchy distribution. In fact, the standard Cauchy random variable has an infinite second moment.
(3) For contaminated normal distribution, i.e. Case 4 and Case 6, the proposed method performs better than LS, especially in Case 4. This shows that LS method is sensitive to the outliers, and the proposed method is relatively robust. In addition, Bias, SSD, and MSE of the LS in Case 4 are larger than those in Case 6. The reason lies in that the second moment of the random error in Case 4 does not exist.
(4) With the exception of Case 1, the proposed estimator performs better than LS estimator in terms of SSD and MSE as expected, which is consistent with the theoretical results in Section 2.

3.2. *Performance of the proposed variable selection method*

In this subsection, we explore the numerical performance of the proposed variable selection method. In the mean regression, the tuning parameters are chosen by the methods proposed by Kong et al. (2016). The data sets are generated from the following model:

$$Y = Z^T\theta + \int_0^1 X(t)\beta(t)dt + \varepsilon, \tag{9}$$

where Z follows a multivariate normal distribution $N(0, \Sigma_Z)$ with $(\Sigma_Z)_{i,j} = 0.25^{|i-j|}$ for $1 \le i, j \le 8$, $\theta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$, and the other settings are the same as Section 3.1. In this subsection, the mean estimated error is used to assess the accuracy of the parametric component estimators, which is defined as follows:

$$ME = E\left[(\hat{\theta} - \theta)^T \Sigma_Z(\hat{\theta} - \theta)\right].$$

It is well known that tuning parameter selection plays an important role in regularization methods. We use the following Schwarz-type information to select the regularization parameters m and ϱ.

$$\mathrm{SIC}(m, \varrho) = \log\left\{\sum_{k=1}^{M}\sum_{i=1}^{n}\rho_{\tau_k}\left(Y_i - \hat{b}_{\tau_k} - Z_i^T\hat{\theta} - U_i^T\hat{\gamma}\right)\right\} + \frac{\log(n)}{2n}(m + p_1 + M),$$

where p_1 is the number of the nonzero elements in $\hat{\theta}$, $\hat{b}_{\tau_k}$, $\hat{\theta}$ and $\hat{\gamma}$ are the estimators obtained by minimizing (7) with the regularization parameters m and ϱ.

Simulation results are presented in Table 2; the columns labeled by "C" list the average number of zero regression coefficients that are correctly estimated as zero, the columns labeled by "IC" depict the average number of non-zero

Table 2
Simulation results of variable selection ($\times 100$).

Case	Method	$n = 200$			$n = 400$		
		C	IC	ME	C	IC	ME
1	CQR-ALasso	488.4	0.0	2.42	494.8	0	1.12
	CQR-Oracle	500.0	0.0	1.79	500.0	0	0.81
	LS-ALasso	437.2	0.0	2.92	456.2	0	1.21
	LS-Oracle	500.0	0.0	1.67	500.0	0	0.75
2	CQR-ALasso	495.0	0.0	2.94	498.0	0	1.37
	CQR-Oracle	500.0	0.0	2.49	500.0	0	1.17
	LS-ALasso	416.4	0.0	8.92	443.0	0	4.08
	LS-Oracle	500.0	0.0	4.21	500.0	0	2.35
3	CQR-ALasso	498.6	0.0	5.38	499.8	0	2.28
	CQR-Oracle	500.0	0.0	5.07	500.0	0	2.55
	LS-ALasso	65.8	9.8	0.22	60.0	6	0.26
	LS-Oracle	500.0	0.0	0.11	500.0	0	0.10
4	CQR-ALasso	490.4	0.0	2.14	496.6	0	1.15
	CQR-Oracle	500.0	0.0	1.59	500.0	0	0.85
	LS-ALasso	431.0	0.0	25.54	433.2	0	159.43
	LS-Oracle	500.0	0.0	6.27	500.0	0	30.15
5	CQR-ALasso	496.8	0.0	1.77	499.0	0	0.88
	CQR-Oracle	500.0	0.0	1.23	500.0	0	0.60
	LS-ALasso	370.6	0.0	17.03	380.4	0	9.04
	LS-Oracle	500.0	0.0	7.47	500.0	0	4.27
6	CQR-ALasso	491.8	0.0	2.36	497.0	0	1.09
	CQR-Oracle	500.0	0.0	1.88	500.0	0	0.93
	LS-ALasso	432.4	0.0	5.83	450.8	0	2.72
	LS-Oracle	500.0	0.0	3.02	500.0	0	1.64

The entries are multiplied by 100, except the MEs in case 3 with LS methods, which are multiplied by 10^{-4}.

regression coefficients that are erroneously set to zero. Rows refer to methods, where "LS-Oracle" and "CQR-Oracle" stand for the oracle estimates based on the mean regression and the composite quantile regression, respectively. "LS-ALasso" and "CQR-ALasso" stand for variable selection via penalized the least square loss function and the composite quantile regression with adaptive Lasso, respectively. From Table 2, it can be concluded as follows:

(i) As expected, the performance of CQR-Oracle procedure performs well in all cases in terms of model errors. Furthermore, the performance of the proposed variable selection CQR-ALasso becomes more and more closer to those of CQR-Oracle procedure as the sample size increases. This confirms the theory in Section 2.
(ii) The results of variable selection are mildly sensitive to error distributions.
(iii) For the normal random error, the performances based on mean regression perform better than the composite quantile regression.
(iv) For the heavy-tailed random error, the proposed variable selection methods are superior to the existing methods, especially for standard Cauchy random error.
(v) It is well known that LS method is consistent with maximum likelihood method for normal error. Not surprisingly, CQR-ALasso outperforms LS-ALasso in Case one, this is partly because the sample size is relatively small. Another reason may be that CQR is almost as efficient as the least squares in this random error, see Zou and Yuan (2008).

3.3. High dimensional case

In this subsection, we consider assessing the performance of the proposed variable selection method in the high dimensional case. The data sets are generated from the following model:

$$Y = Z^T\theta_n + \int_0^1 X(t)\beta(t)dt + \varepsilon, \tag{10}$$

where Z follows multivariate normal distribution $N(0, \Sigma_Z)$ with $(\Sigma_Z)_{i,j} = \kappa^{|i-j|}$ for $1 \le i, j \le p_n$, θ_{n1} is a d_n-dimensional vector, θ_{n2} is a $(p_n - d_n)$-dimensional vector with all elements being 0, $\theta_n = (\theta_{n1}^T, \theta_{n2}^T)^T$, and the other settings are the same as Section 3.2. For sample size n, we consider two cases which are $n = 200$ and 400. When $n = 200$, let $p_n = 15$ and $\theta_{n1} = (3, 2, 1.5, 0.8, 1.2)$. When $n = 400$, let $p_n = 25$ and $\theta_{n1} = (3, 2, 1.5, 0.8, 1.2, -1, -0.75, 2.5)$. The simulation results are reported in Tables 3–4. Generally, the proposed method can select the important variables, and obtain the precise models. The Oracle method is better than the proposed method as expected. Estimates obtained from the proposed method get closer to those obtained from the Oracle procedure as the sample size increases. In addition, this subsection focuses on the high-dimensional setting, i.e. $p_n/n \to 0$, and how to extend the results to ultra-high dimensional setting, i.e $p_n \gg n$ or "Big p Small n", warrants a future investigation.

Table 3
Simulation results ($\times 100$) for high dimensional case with sample size $n = 200$.

Case	Method	$\kappa = 0.25$			$\kappa = 0.75$		
		C	IC	ME	C	IC	ME
1	ALasso	902	0	0.05	894	0	0.05
	Oracle	1000	0	0.03	1000	0	0.03
2	ALasso	912	0	0.08	888	0	0.06
	Oracle	1000	0	0.05	1000	0	0.04
3	ALasso	900	0	0.12	892	0.06	0.14
	Oracle	1000	0	0.10	1000	0	0.10
4	ALasso	910	0	0.04	906	0	0.04
	Oracle	1000	0	0.03	1000	0	0.03
5	ALasso	926	0	0.04	910	0	0.04
	Oracle	1000	0	0.03	1000	0	0.02
6	ALasso	906	0	0.04	922	0	0.05
	Oracle	1000	0	0.03	1000	0	0.03

Table 4
Simulation results ($\times 100$) for high dimensional case with sample size $n = 400$.

Case	Method	$\kappa = 0.25$			$\kappa = 0.75$		
		C	IC	ME	C	IC	ME
1	ALasso	1653	0.00	0.07	1629	0.01	0.08
	Oracle	1700	0.00	0.05	1700	0.00	0.05
2	ALasso	1669	0.00	0.10	1635	0.07	0.13
	Oracle	1700	0.00	0.07	1700	0.00	0.07
3	ALasso	1676	0.02	0.23	1673	0.70	0.43
	Oracle	1700	0.00	0.16	1700	0.00	0.16
4	ALasso	1656	0.00	0.07	1632	0.01	0.08
	Oracle	1700	0.00	0.05	1700	0.00	0.05
5	ALasso	1680	0.00	0.05	1671	0.01	0.07
	Oracle	1700	0.00	0.04	1700	0.00	0.04
6	ALasso	1662	0.00	0.13	1634	0.20	0.19
	Oracle	1700	0.00	0.09	1700	0.00	0.09

4. Real data example

In this section, we illustrate the proposed method via an analysis of biscuit dough data. The data was obtained from an experiment done to test the feasibility of near-infrared (NIR) spectroscopy to measure the composition of biscuit dough pieces. An NIR reflectance spectrum is recorded for each dough piece. The spectral data consist of 700 points measured from 1100 to 2498 nanometers (nm) in increment of 2 nm. The study focuses on predicting the percentage of each of the four constituents: fat, sucrose, dry flour, and water, based on the NIR reflectance spectrum. The data set is available from the R package "ppls". The data set contains outliers, see Brown, Fearn, and Vannucci (2001) and Mas and Pumo (2009). Therefore, we need robust method to fit the data set. In addition, we are interested to study the relationship between the fat and other ingredient in biscuit dough. In our analysis, we take the response y to be the fat; z_1 to the sucrose; z_2 to the dry flour; z_3 to the water; $X(t)$ to be the spectra. We first standardized the z-covariates. We are interested in whether there are any interaction effects and quadratic effects between these covariates. Therefore, we consider the following model:

$$y = z_1\theta_1 + z_2\theta_2 + z_3\theta_3 + z_1^2\theta_4 + z_2^2\theta_5 + z_3^2\theta_6 + z_1z_2\theta_7 + z_1z_3\theta_8 + z_2z_3\theta_9 + \int_0^1 X(t)\beta(t)dt + \varepsilon.$$

We use the composite quantile regression and mean regression to fit the data set, and use the adaptive Lasso to select significant variables. The resulting estimators are presented in Table 5. From Table 5, we can observe that z_1, z_2 and z_3 have negative effects on response based on the mean regression or the composite quantile regression. In addition, there are no interaction effects or quadratic effects between these covariates. The estimated slope function is presented in Fig. 1. In Fig. 1, the left panel (i.e. solid line) presents the estimator of slope function $\beta(t)$ based ALasso, and the right panel (i.e. dotted line) presents the estimator of slope function $\beta(t)$ based on full model. As one referee pointed out that the coefficient functions in Fig. 1 are very wiggly and thus hard to interpret. In partially functional linear models, the slope function is approximated by functional principal component analysis. We only assume that the slope function β belongs to $L^2(\mathcal{T})$, where $L^2(\mathcal{T})$ is the Hilbert space containing square integrable functions defined on $\mathcal{T}$. We do not impose any smooth conditions on the slope

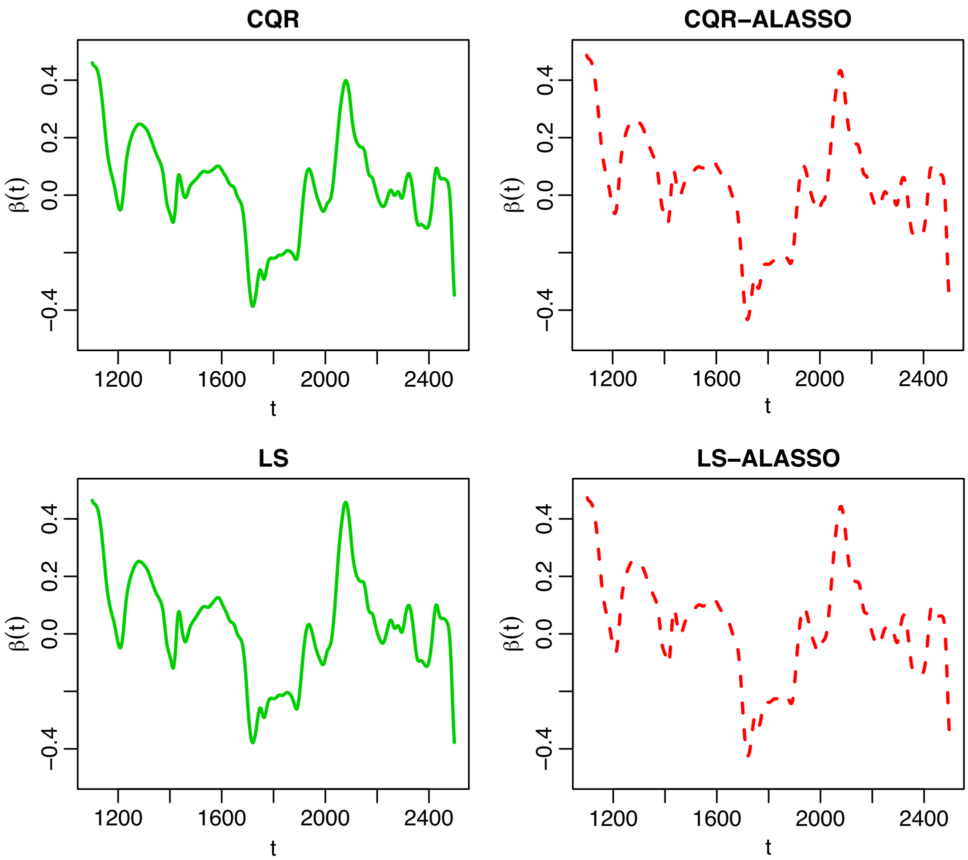

Fig. 1. Estimation of slope function $\beta(t)$.

Table 5
Estimated coefficients for the full model and the models selected by adaptive Lasso.

Method	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9
CQR	−4.66	−2.26	−2.69	0.85	−1.22	−0.01	0.54	−0.62	0.58
CQR-ALASSO	−5.24	−3.39	−2.08	0.00	0.00	0.00	0.00	0.00	0.00
LS	−4.87	−2.57	−2.51	0.68	−2.21	−0.06	−0.40	−0.38	1.15
LS-ALASSO	−5.28	−3.37	−1.92	0.00	0.00	0.00	0.00	0.00	0.00

function, thus the estimator is very wiggly. In addition, if some smooth conditions on β are imposed, we can use B-spline to approximate the slope function. Thus, the corresponding estimator is a smooth function. How to extend the results to such method is worthy of investigation.

5. Conclusion

In this paper, we develop the estimation and variable selection methods for partially functional linear regression models based on the composite quantiles. Under some regularity conditions, we obtain the optimal convergence rate of the functional coefficient and establish the asymptotic normality of parametric component coefficients. Some simulation studies and a real data analysis are conducted to illustrate the finite sample performance of the proposed methods.

Acknowledgments

Du's work is supported by the National Natural Science Foundation of China (Nos. 11501018, 11771032, and 11571340), Program for Rixin Talents in Beijing University of Technology (No. 006000514116003) and the Science and Technology Project of Beijing Municipal Education Commission (KM201710005032). Xu's work is supported by the National Statistical Science Research Project (No. 2016LZ06), and Startup Foundation for Talents in Zhejiang Agriculture and Forestry University (No. 2016FR030). Cao's work is partly supported by the National Natural Science Foundation of China (No. 11701020). The authors appreciate the associate editor, and the two referees for their helpful comments that led to an improvement of an early manuscript.

Appendix. Proofs of theorems

Proof of Theorem 1. Denote $H_m = \lambda_m^{-1} U^T U$, $P = U(U^T U)^{-1} U^T$, $Z^* = (I - P)Z$ and $\Sigma_n = Z^{*T} Z^*$.

Let

$$\xi \begin{pmatrix} b \\ \theta \\ \gamma \end{pmatrix} = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} \sqrt{n}(b - b_0) \\ \Sigma_n^{\frac{1}{2}}(\theta - \theta_0) \\ \lambda_m^{-1/2} H_m^{\frac{1}{2}}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_m) + \lambda_m^{1/2} H_m^{-\frac{1}{2}} UZ(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \end{pmatrix}.$$

Let $\hat{\boldsymbol{\xi}} = \xi(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}) = (\hat{\boldsymbol{\xi}}_1^T, \hat{\boldsymbol{\xi}}_2^T)^T$. Now, we show that $\|\hat{\boldsymbol{\xi}}\|^2 = O_p(m^{a+1})$. Denote $\boldsymbol{\Theta}_n = \{\boldsymbol{\xi} : \|\boldsymbol{\xi}\| = Lm^{(a+1)/2}\}$ for some sufficiently large constant L. Let $\tilde{Z}_i = \Sigma_n^{-\frac{1}{2}} Z_i$, $\tilde{U}_i = H_m^{-1} U_i$, $R_i = \sum_{j=1}^m \langle x_i, \hat{\phi}_j \rangle \gamma_{j0} - \int_0^1 \beta(t) x(t) dt$. Thus,

$$\begin{aligned} Q(\xi) &= \sum_{k=1}^M \sum_{i=1}^n \rho_{\tau_k}(y_i - b_{\tau_k} - Z_i^T \boldsymbol{\theta} - U_i^T \boldsymbol{\gamma}) \\ &= \sum_{k=1}^M \sum_{i=1}^n \rho_{\tau_k}\left(\varepsilon_{ik} - (b_{\tau_k} - b_{\tau_k 0}) - Z_i^T(\boldsymbol{\theta} - \boldsymbol{\theta}_0) - U_i^T(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) - R_i\right) \\ &= \sum_{k=1}^M \sum_{i=1}^n \rho_{\tau_k}\left(\varepsilon_{ik} - \nu_k^T \xi_1 - \tilde{Z}_i^T \boldsymbol{\xi}_2 - \tilde{U}_i^T \boldsymbol{\xi}_3 - R_i\right), \end{aligned}$$

where $\nu_k = e_k/\sqrt{n}$, $e_k = (0, 0, \ldots, 1, 0, \ldots, 0)^T$, i.e., e_k is a unit vector with 1 in its kth place, and $\varepsilon_{ik} = \varepsilon_i - b_{\tau_k 0}$. By the proof of Theorem 1 of Lu et al. (2014), one has $\|R_i\|^2 = O_p(n^{-\frac{2b+a-1}{a+2b}})$. Invoking Knight's identity in Knight (1998), one has

$$\begin{aligned} Q(\xi) - Q(0) &= \sum_{k=1}^M \sum_{i=1}^n \left[\rho_{\tau_k}\left(\varepsilon_{ik} - \nu_k^T \xi_1 - \tilde{Z}_i^T \boldsymbol{\xi}_2 - \tilde{U}_i^T \boldsymbol{\xi}_3 - R_i\right) - \rho_\tau(\varepsilon_{ik} - R_i)\right] \\ &= A_{n1} + A_{n2}, \end{aligned}$$

where

$$\begin{aligned} A_{n1} &= \sum_{k=1}^M \sum_{i=1}^n (\nu_k^T \xi_1 + \tilde{Z}_i^T \boldsymbol{\xi}_2 + \tilde{U}_i^T \boldsymbol{\xi}_3) \psi_\tau(\varepsilon_{ik} - R_i), \\ A_{n2} &= \sum_{k=1}^M \sum_{i=1}^n \int_0^{\nu_k^T \xi_1 + \tilde{Z}_i^T \boldsymbol{\xi}_2 + \tilde{U}_i^T \xi_3} [I(\varepsilon_{ik} - R_i \le s) - I(\varepsilon_{ik} - R_i \le 0)] ds. \end{aligned}$$

Denote $\mathcal{F}_n = \{(Z_1, X_1), \ldots, (Z_n, X_n)\}$. By $\|R_i\|^2 = O_p(n^{-\frac{2b+a-1}{a+2b}})$ and Conditions C4 and C5, we have

$$\begin{aligned} |E(A_{n1}|\mathcal{F}_n)|^2 &= \left|\sum_{k=1}^M \sum_{i=1}^n (\nu_k^T \xi_1 + \tilde{Z}_i^T \boldsymbol{\xi}_2 + \tilde{U}_i^T \boldsymbol{\xi}_3) E\psi_\tau(\varepsilon_{ik} - R_i)\right|^2 \\ &= \left|\sum_{k=1}^M \sum_{i=1}^n (\nu_k^T \xi_1 + \tilde{Z}_i^T \boldsymbol{\xi}_2 + \tilde{U}_i^T \boldsymbol{\xi}_3)[\tau - P(\varepsilon_{ik} - R_i < 0|\mathcal{F}_n)]\right|^2 \\ &\le \left(\sum_{i=1}^n \sum_{k=1}^M \left[\nu_k^T \xi_1 + \tilde{Z}_i^T \boldsymbol{\xi}_2 + \tilde{U}_i^T \boldsymbol{\xi}_3\right]^2\right) \left(\sum_{i=1}^n \sum_{k=1}^K [\tau_k - P(\varepsilon_{ik} - R_i < 0|\mathcal{F}_n)]^2\right) \\ &= O_p\left(m^{a+2}\right), \end{aligned}$$

where the first inequality holds by the Cauchy–Schwarz inequality.

$$\begin{aligned} \mathrm{Var}(A_{n1}|\mathcal{F}_n) &= E\left|\sum_{k=1}^M \sum_{i=1}^n (\nu_k^T \xi_1 + \tilde{Z}_i^T \boldsymbol{\xi}_2 + \tilde{U}_i^T \boldsymbol{\xi}_3)[\psi_\tau(\varepsilon_{ik} - R_i) - E\psi_\tau(\varepsilon_{ik} - R_i)]\right|^2 \\ &\le M \sum_{k=1}^M E\left|\sum_{i=1}^n (\nu_k^T \xi_1 + \tilde{Z}_i^T \boldsymbol{\xi}_2 + \tilde{U}_i^T \boldsymbol{\xi}_3)[\psi_\tau(\varepsilon_{ik} - R_i) - E\psi_\tau(\varepsilon_{ik} - R_i)]\right|^2 \end{aligned}$$

$$\leq M\sum_{k=1}^{M}\sum_{i=1}^{n}(\nu_k^T\xi_1+\tilde{Z}_i^T\boldsymbol{\xi}_2+\tilde{U}_i^T\boldsymbol{\xi}_3)^2\mathrm{Var}(\psi_\tau(\varepsilon_{ik}-R_i))$$
$$=O_p\left(m^{a+1}\right).$$

Therefore, one has $A_{n1}=O_p\left(m^{\frac{a+2}{2}}\right)+O_p\left(m^{\frac{a+1}{2}}\right)=O_p\left(m^{\frac{a+2}{2}}\right)$.

Next, consider A_{n2}. By elementary calculation, one has

$$\begin{aligned}E(A_{n2}|\mathcal{F}_n)&=\sum_{k=1}^{M}\sum_{i=1}^{n}\int_0^{\nu_k^T\xi_1+\tilde{Z}_i^T\xi_2+\tilde{U}_i^T\xi_3}E[I(\varepsilon_{ik}-R_i\leq s)-I(\varepsilon_{ik}-R_i\leq 0)]ds\\&=\sum_{k=1}^{M}\sum_{i=1}^{n}\int_0^{\nu_k^T\xi_1+\tilde{Z}_i^T\xi_2+\tilde{U}_i^T\xi_3}(f(b_k)+O(\|R_i\|))\,sds\\&=O_p\left(\|\boldsymbol{\xi}\|^2\right)\\&=O_p\left(m^{a+1}\right).\end{aligned}$$

By routine calculation, for large n, we have

$$\begin{aligned}\mathrm{Var}(A_{n2}|\mathcal{F}_n)&\leq 2\max_{1\leq i\leq n,1\leq k\leq M}\|\nu_k^T\xi_1+\tilde{Z}_i^T\boldsymbol{\xi}_2+\tilde{U}_i^T\boldsymbol{\xi}_3\|E\|I_{n2}\|\\&\leq o_p\left(\|\theta\|\right)E(I_{n2}).\end{aligned}$$

Considering the stochastic order of A_{n1} and A_{n2}, one can know that A_{n1} is dominated by A_{n2}. Thus, we have

$$P\left\{\inf_{\|\boldsymbol{\xi}\|\in\Theta_n}Q(\boldsymbol{\xi})>Q(0)\right\}\to 1.$$

On the other hand, since $Q(\boldsymbol{\xi})$ is minimized at $\hat{\boldsymbol{\xi}}$, we have $Q(\hat{\boldsymbol{\xi}})\leq Q(0)$. Combining this with the convexity of $Q(\boldsymbol{\xi})$, we obtain $\|\hat{\boldsymbol{\xi}}\|^2=O_p\left(m^{a+1}\right)$. By the definition of ξ, one has $\|\hat{\theta}-\theta_0\|^2=O_p\left(\frac{m^{a+1}}{n}\right)$ and $\|\hat{b}-b_0\|^2=O_p\left(\frac{m^{a+1}}{n}\right)$. By triangle inequality, one has

$$\begin{aligned}\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_m\|^2&\leq\|\lambda_m^{-1/2}H_m^{\frac{1}{2}}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_m)\|^2\\&\leq 2\|\hat{\boldsymbol{\xi}}_3\|^2+2\|\lambda_m^{1/2}H_m^{-\frac{1}{2}}UZ(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|^2\\&=O_p\left(\frac{m^{a+1}}{n}\right).\end{aligned}$$

Invoking the fact that $\|\hat{\phi}_j-\phi_j\|=O_p(j^2/n)$, we have

$$\begin{aligned}\|\hat{\beta}-\beta_0\|^2&=\left\|\sum_{l=1}^{m}\hat{\gamma}_l\hat{\phi}_l-\sum_{l=1}^{\infty}\gamma_{0l}\phi_l\right\|^2\\&\leq 2\left\|\sum_{l=1}^{m}\hat{\gamma}_l\hat{\phi}_l-\sum_{l=1}^{m}\gamma_{0l}\phi_l\right\|^2+2\left\|\sum_{j=m+1}^{\infty}\gamma_{0l}\phi_l\right\|^2\\&\leq 4\left\|\sum_{l=1}^{m}(\hat{\gamma}_l-\gamma_{0l})\hat{\phi}_l\right\|^2+4\left\|\sum_{l=1}^{m}\gamma_{0l}(\hat{\phi}_{jl}-\phi_{jl})\right\|^2+2\sum_{l=m+1}^{\infty}\gamma_{0l}^2\\&\leq 4\sum_{l=1}^{m}\|\hat{\gamma}_l-\gamma_{0l}\|^2+4m\sum_{l=1}^{m}\gamma_{0l}^2\|\hat{\phi}_{jl}-\phi_{jl}\|^2+2\sum_{l=m+1}^{\infty}\gamma_{0l}^2\\&\leq 4\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_m\|+O_p\left(\frac{m^{4-2b}}{n}\right)+2\sum_{l=m+1}^{\infty}l^{-2b}\\&=O_p\left(\frac{m^{a+1}}{n}\right)+O_p(m^{-a-4b+4})+O(m^{-2b+1})\end{aligned}$$

$$= O_p\left(\frac{m^{a+1}}{n}\right) + O_p(m^{-a-4b+4}) + O(m^{-2b+1})$$
$$= O_p\left(n^{-\frac{2b-1}{a+2b}}\right).$$

Now, we consider the asymptotic normality of $\hat{\theta}$. Let

$$\xi_2^* = (\sqrt{\varpi}\Sigma)^{-1}\sum_{i=0}^{n} Z_i^*[\sum_{k=1}^{M}[I(\varepsilon_i \le b_k) - \tau_k]].$$

By the central limit theorem and the law of large numbers, one has $\boldsymbol{\xi}_2^* \xrightarrow{L} N\left(0, \varpi\Sigma^{-1}\right)$. If we can show $\hat{\boldsymbol{\xi}}_2^* = \boldsymbol{\xi}_2^* + o_p(1)$, the asymptotic normality of $\hat{\theta}$ follows. Let

$$\tilde{\boldsymbol{\xi}}_2 = (\sqrt{\varpi}\Sigma)^{-1}\sum_{i=0}^{n} Z_i^*[\sum_{k=1}^{M}[I(\varepsilon_i - R_i \le b_k) - \tau_k]].$$

By similar arguments in Theorem 4.1 of Wei and He (2006), we have $\|\tilde{\xi}_2 - \xi_2^*\| = o_p(1)$. Therefore, we only need to show that $\|\hat{\boldsymbol{\xi}}_2 - \tilde{\boldsymbol{\xi}}_2\| = o_p(1)$, which can be proved by similar arguments used in Theorem 1 of Wang et al. (2009).

Proof of Theorem 2. We first prove the part (i): Convergence rate. Similar to Theorem 1, we have

$$\begin{aligned}
Q(\boldsymbol{\xi}) &= \sum_{k=1}^{M}\sum_{i=1}^{n}\rho_{\tau_k}(y_i - b_{\tau_k} - Z_i^T\theta - U_i^T\gamma) + \sum_{j=1}^{p}\varrho\frac{|\theta_j|}{|\hat{\theta}_j^{CQR}|^2} \\
&= \sum_{k=1}^{M}\sum_{i=1}^{n}\rho_{\tau_k}\left(\varepsilon_{ik} - (b_{\tau_k} - b_{\tau_k 0}) - Z_i^T(\theta-\theta_0) - U_i^T(\gamma-\gamma_0) - R_i\right) + \sum_{j=1}^{p}\varrho\frac{|\theta_j|}{|\hat{\theta}_j^{CQR}|^2} \\
&= \sum_{k=1}^{M}\sum_{i=1}^{n}\rho_{\tau_k}\left(\varepsilon_{ik} - \nu_k^T\xi_1 - \tilde{Z}_i^T\boldsymbol{\xi}_2 - \tilde{U}_i^T\boldsymbol{\xi}_3 - R_i\right) + \sum_{j=1}^{p}\varrho\frac{|\theta_j|}{|\hat{\theta}_j^{CQR}|^2},
\end{aligned}$$

where $\varepsilon_{ik} = \varepsilon_i - b_{k0}$. Then invoking Knight's identity in Knight (1998), we have

$$\begin{aligned}
Q(\boldsymbol{\xi}) - Q(0) &= \sum_{k=1}^{M}\sum_{i=1}^{n}\left[\rho_{\tau_k}\left(\varepsilon_{ik} - \nu_k^T\xi_1 - \tilde{Z}_i^T\boldsymbol{\xi}_2 - \tilde{U}_i^T\boldsymbol{\xi}_3 - R_i\right) - \rho_\tau(\varepsilon_{ik} - R_i)\right] \\
&\quad + \sum_{j=1}^{p}\varrho\frac{|\theta_j|}{|\hat{\theta}_j^{CQR}|^2} - \sum_{j=1}^{p}\varrho\frac{|\theta_{0j}|}{|\hat{\theta}_j^{CQR}|^2} \\
&= I_{n1} + I_{n2} + I_{n3},
\end{aligned}$$

where

$$\begin{aligned}
I_{n1} &= \sum_{k=1}^{M}\sum_{i=1}^{n}(\nu_k^T\xi_1 + \tilde{Z}_i^T\boldsymbol{\xi}_2 + \tilde{U}_i^T\boldsymbol{\xi}_3)\psi_\tau(\varepsilon_{ik} - R_i), \\
I_{n2} &= \sum_{k=1}^{M}\sum_{i=1}^{n}\int_0^{\nu_k^T\xi_1 + \tilde{Z}_i^T\boldsymbol{\xi}_2 + \tilde{U}_i^T\boldsymbol{\xi}_3}[I(\varepsilon_{ik} - R_i \le s) - I(\varepsilon_{ik} - R_i \le 0)]ds, \\
I_{n3} &= \sum_{j=1}^{p}\frac{\varrho}{|\hat{\theta}_j^{CQR}|^2}(|\theta_j| - |\theta_{0j}|).
\end{aligned}$$

Similar to A_{n1} and A_{n2}, one has $I_{n1} = O_p\left(m^{\frac{a+2}{2}}\right)$ and $E(I_{n2}) = O_p\left(m^{a+1}\right)$. In addition, by routine calculation, for large n, we have

$$\begin{aligned}
\mathrm{Var}(I_{n2}|\mathcal{F}_n) &\le 2\max_{1\le i\le n, 1\le k\le K}\|\nu_k^T\xi_1 + \tilde{Z}_i^T\boldsymbol{\xi}_2 + \tilde{U}_i^T\boldsymbol{\xi}_3\|E\|I_{n2}\| \\
&\le o_p(\|\theta\|)\,E(I_{n2}).
\end{aligned}$$

Now, we consider I_{n3}. By the condition $\varrho/\sqrt{n} \to 0$, as $n \to \infty$, one has

$$\begin{aligned} I_{n3} &= \sum_{j=1}^{p} \frac{\varrho}{|\hat{\theta}_j^{CQR}|^2}(|\theta_j| - |\theta_{0j}|) \\ &\geq -\sum_{j=1}^{p} \frac{\varrho}{|\hat{\theta}_j^{CQR}|^2}(|\theta_j - \theta_{0j}|) \\ &\geq -\frac{\varrho}{\sqrt{n}|\hat{\theta}_j^{CQR}|^2} O_p(\|\boldsymbol{\xi}_2\|) \\ &\geq o_p(\|\boldsymbol{\xi}_2\|). \end{aligned}$$

Considering the stochastic order of I_{n1}, I_{n2}, and I_{n3}, we can know that I_{n1} and I_{n3} are dominated by I_{n2}. Thus, we have

$$P\left\{\inf_{\|\boldsymbol{\xi}\| \in \Theta_n} Q(\boldsymbol{\xi}) > Q(0)\right\} \to 1.$$

On the other hand, since $Q(\boldsymbol{\xi})$ is minimized at $\hat{\boldsymbol{\xi}}$, we have $Q(\hat{\boldsymbol{\xi}}) \leq Q(0)$. Then combining this with the convexity of $Q(\boldsymbol{\xi})$, we obtain $\|\hat{\boldsymbol{\xi}}\| = O_p\left(m^{a+1}\right)$. By the definition of ξ, one has $\|\hat{\theta} - \theta_0\|^2 = O_p\left(\frac{m^{a+1}}{n}\right)$ and $\|\hat{b} - b_0\|^2 = O_p\left(\frac{m^{a+1}}{n}\right)$. Using similar argument in Theorem 1 to prove the rate of convergence of the functional slope parameter $\hat{\beta}$, we have $\|\hat{\beta} - \beta_0\|^2 = O_p\left(n^{-\frac{2b-1}{a+2b}}\right)$.

Next, we consider part (ii): Sparsity. Let

$$\hat{\boldsymbol{\xi}}^* = \begin{pmatrix} \sqrt{n}(\hat{b} - b_0) \\ \Sigma_n^{\frac{1}{2}}(\hat{\theta}^* - \theta_0) \\ \lambda_m^{-1/2} H_m^{\frac{1}{2}}(\hat{\gamma} - \gamma_m) + \lambda_m^{1/2} H_m^{-\frac{1}{2}} UZ(\hat{\theta}^* - \theta_0) \end{pmatrix},$$

where $\hat{\theta}^* = (\hat{\theta}_I^T, 0^T)^T$ and $\theta_0 = (\theta_{0I}^T, 0^T)^T$. In fact, $\hat{\theta}^*$ is obtained from $\hat{\theta}$ by constraining the insignificant components to zero.

By the definition of $\hat{\boldsymbol{\xi}}$, we have $Q(\hat{\boldsymbol{\xi}}^*) \geq Q(\hat{\boldsymbol{\xi}})$. The sparsity of $\hat{\theta}$ will be proved on the condition that if there exists some $\hat{\theta}_j \neq 0$ for $s+1 \leq j \leq p$, it must have $Q(\hat{\boldsymbol{\xi}}^*) < Q(\hat{\boldsymbol{\xi}})$ for enough large n. Invoking the fact that $\|\hat{\boldsymbol{\xi}}\| = O_P(m^{a+1})$, $\|\hat{\boldsymbol{\xi}}^*\| = O_p(m^{a+1})$ and the proof of Theorem 1, one has

$$\begin{aligned} Q(\hat{\boldsymbol{\xi}}^*) - Q(\hat{\boldsymbol{\xi}}) &= Q(\hat{\boldsymbol{\xi}}^*) - Q(0) + Q(0) - Q(\hat{\boldsymbol{\xi}}) \\ &= Q(\hat{\boldsymbol{\xi}}^*) - Q(0) - [Q(\hat{\boldsymbol{\xi}}) - Q(0)] \\ &= O_P(\|\hat{\boldsymbol{\xi}}^*\|^2 - \|\hat{\boldsymbol{\xi}}\|^2) - \sum_{j=s+1}^{p} \varrho \frac{|\hat{\theta}_j|}{|\hat{\theta}_j^{CQR}|^2} \\ &\leq m^{a+1}\left[O_p\left(m^{-a-1}\sum_{j=s+1}^{p}|\hat{\xi}_j|^2\right) - \frac{\varrho}{m^{a+1}}\sum_{j=s+1}^{p}|\hat{\gamma}_j|\right]. \end{aligned}$$

If $\sum_{j=s+1}^{p}|\hat{\theta}_j| > 0$, by the condition $\frac{\varrho}{m^{a+1}} \to \infty$ and $\sum_{j=s+1}^{p}|\hat{\xi}_j|^2 = O_p(m^{a+1})$, we have $Q(\hat{\boldsymbol{\xi}}^*) < Q(\hat{\boldsymbol{\xi}})$ for large n. This completes the proof.

Lastly, we should prove the asymptotic normality of $\hat{\theta}_I$. The strategy to prove this result is similar to Theorem 1, thus we omit it.

References

Aneiros, G., Ferraty, F., & Vieu, P. (2015). Variable selection in partial linear regression with functional covariate. *Statistics*, *49*(6), 1322–1347.

Brown, P. J., Fearn, T., & Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, *96*(454), 398–408.

Cai, T., & Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics*, *34*(5), 2159–2179.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Gertheiss, J., Maity, A., & Staicu, A.-M. (2014). Variable selection in generalized functional linear models. *Stat*, *2*, 86–101.

Greven, S., & Scheipl, F. (2017). A general framework for functional regression modelling. *Statistical Modelling*, *17*, 1–35.

Hall, P., & Horowitz, J. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, *35*, 70–91.

He, X., Zhu, Z. Y., & Fung, W. K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, *90*, 579–590.
Horváth, L., & Kokoszka, P. (2012). *Inference for functional data with applications*. New York: Springer.
Huang, L., Zhao, J., Wang, H., & Wang, S. (2016). Robust shrinkage estimation and selection for functional multiple linear model through LAD loss. *Computational Statistics & Data Analysis*, *103*, 384–400.
Kai, B., Li, R., & Zou, H. (2010). Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression. *Journal of the Royal Statistical Society. Series B.*, *72*, 49–69.
Kai, B., Li, R., & Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *The Annals of Statistics*, *39*, 305–332.
Knight, K. (1998). Limiting distributions for l_1 regression estimators under general conditions. *The Annals of Statistics*, *26*(2), 755–770.
Koenker, R. (2005). *Quantile regression*. Cambridge: Cambridge University Press.
Koenker, R., & Bassett, G. S. (1978). Regression quantiles. *Econometrica*, *46*, 33–50.
Kong, D., Xue, K., Yao, F., & Zhang, H. H. (2016). Partially functional linear regression in high dimensions. *Biometrika*, *103*, 147–159.
Lian, H. (2013). Shrinkage estimation and selection for multiple functional regression. *Statistica Sinica*, *23*, 51–74.
Lu, Y., Du, J., & Sun, Z. (2014). Functional partially linear quantile regression model. *Metrika*, *77*(3), 317–332.
Mas, A., & Pumo, B. (2009). *Linear processes for functional data*. (pp. 47–71). Oxford Univ Press.
Matsui, H., & Konishi, S. (2011). Variable selection for functional regression models via the L1 regularization. *Computational Statistics & Data Analysis*, *55*, 3304–3310.
Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, *2*, 321–359.
Müller, H., & Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, *33*, 774–805.
Pannu, J., & Billor, N. (2017). Robust group-Lasso for functional regression model. *Communications in Statistics. Simulation and Computation*, *46*, 3356–3374.
Ramsay, J. O., & Silverman, B. W. (2002). *Applied functional data analysis: Methods and case studies*. New York: Springer.
Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). New York: Springer.
Reiss, P. T., Goldsmith, J., Shang, H. L., & Ogden, R. T. (2017). Methods for scalar-on-function regression. *International Statistical Review*, *85*, 228–249.
Sentürk, D., & Müller, H. (2010). Functional varying coefficient models for longitudinal data. *Journal of the American Statistical Association*, *105*, 1256–1264.
Shin, H. (2009). Partial functional linear regression. *Journal of Statistical Planning and Inference*, *139*, 3405–3418.
Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B.*, *58*, 267–288.
Wang, J. L., Chiou, J. M., & Müller, H. G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, *3*, 257–295.
Wang, L., Kai, B., & Li, R. (2009). Local rank inference for varying coefficient models. *Journal of the American Statistical Association*, *104*(488), 1631–1645.
Wei, Y., & He, X. (2006). Conditional growth charts. *The Annals of Statistics*, *34*(5), 2069–2097.
Yao, F., Müller, H., & Wang, J. (2005). Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, *33*, 2873–2903.
Yao, F., Sue-Chee, S., & Wang, F. (2017). Regularized partially functional quantile regression. *Journal of Multivariate Analysis*, *156*, 39–56.
Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*, 894–942.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.
Zou, H., & Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *The Annals of Statistics*, *36*, 1108–1126.