The Annals of Statistics
2012, Vol. 40, No. 6, 3108–3136
DOI: 10.1214/12-AOS1066

ESTIMATION IN FUNCTIONAL LINEAR QUANTILE REGRESSION[1]

BY KENGO KATO

Hiroshima University

This paper studies estimation in functional linear quantile regression in which the dependent variable is scalar while the covariate is a function, and the conditional quantile for each fixed quantile index is modeled as a linear functional of the covariate. Here we suppose that covariates are discretely observed and sampling points may differ across subjects, where the number of measurements per subject increases as the sample size. Also, we allow the quantile index to vary over a given subset of the open unit interval, so the slope function is a function of two variables: (typically) time and quantile index. Likewise, the conditional quantile function is a function of the quantile index and the covariate. We consider an estimator for the slope function based on the principal component basis. An estimator for the conditional quantile function is obtained by a plug-in method. Since the so-constructed plug-in estimator not necessarily satisfies the monotonicity constraint with respect to the quantile index, we also consider a class of monotonized estimators for the conditional quantile function. We establish rates of convergence for these estimators under suitable norms, showing that these rates are optimal in a minimax sense under some smoothness assumptions on the covariance kernel of the covariate and the slope function. Empirical choice of the cutoff level is studied by using simulations.

1. Introduction. Quantile regression, initially developed by the seminal work of Koenker and Bassett (1978), is one of the most important statistical methods in measuring the impact of covariates on dependent variables. An attractive feature of quantile regression is that it allows us to make inference on the entire conditional distribution by estimating several different conditional quantiles. Some basic materials on quantile regression and its applications are summarized in Koenker (2005).

This paper studies estimation in functional linear quantile regression in which the dependent variable is scalar while the covariate is a function, and the conditional quantile for each fixed quantile index is modeled as a linear functional of the covariate. The model that we consider is an extension of functional linear regression to the quantile regression case. Here we suppose that covariates are discretely observed and sampling points may differ across subjects, where the number of

Received March 2012; revised August 2012.

[1]Supported by the Grant-in-Aid for Young Scientists (B) (22730179) from the JSPS.

MSC2010 subject classifications. 62G20.

Key words and phrases. Functional data, nonlinear ill-posed problem, principal component analysis, quantile regression.

measurements per subject increases as the sample size. Also, we allow the quantile index to vary over a given subset of the open unit interval, so the slope function is a function of two variables: (typically) time and quantile index. Likewise, the conditional quantile function is a function of the quantile index and the covariate. We consider the problem of estimating the slope function as well as the conditional quantile function itself. The estimator we consider for the slope function is based on the principal component analysis (PCA). Expanding the covariate and the slope function in terms of the PCA basis, the model is transformed into a quantile regression model with an infinite number of regressors. Truncating the infinite sum by the first m (say) terms, we may apply a standard quantile regression technique to estimating the first m coefficients in the basis expansion of the slope function at each quantile index, where m diverges as the sample size. In practice, the population PCA basis is unknown, so it has to be replaced by a suitable estimate for it. Once the estimator for the slope function is available, an estimator for the conditional quantile function is obtained by a plug-in method. Since the so-constructed plug-in estimator not necessarily satisfies the monotonicity constraint with respect to the quantile index, we also consider a class of monotonized estimators for the conditional quantile function.

In summary, we have the following three types of estimators in mind:

(i) a PCA-based estimator for the slope function;
(ii) a plug-in estimator for the conditional quantile function;
(iii) monotonized estimators for the conditional quantile function.

We establish rates of convergence for these estimators under suitable norms, showing that these rates are optimal in a minimax sense under some smoothness assumptions on the covariance kernel of the covariate and the slope function.

In practice, we have to choose the cutoff level empirically. We suggest some criteria, namely, (integrated-)AIC, BIC and GACV, to choose the cutoff level. We study the performance of these criteria by using simulations. In our limited simulation experiments, although none of these criteria clearly dominated the others, (integrated-)BIC worked relatively stably.

Functional data have become increasingly important. For data collected on dense grids, a traditional multivariate analysis is not directly applicable since the number of grid points may be larger than the sample size and the correlation between distinct grid points is potentially high. *Functional data analysis* views such data as realizations of random functions and takes into account the functional nature of the data. We refer the reader to Ramsay and Silverman (2005) for a comprehensive treatment on functional data analysis. Earlier theoretical studies in functional data analysis have focused on functional linear mean regression models [see Cai and Hall (2006), Cardot, Ferraty and Sarda (1999, 2003), Crambes, Kneip and Sarda (2009), Delaigle and Hall (2012), Hall and Horowitz (2007), James, Wang and Zhu (2009), Yao, Müller and Wang (2005), Yuan and Cai (2010) and references cited in these papers]. Among them, Hall and Horowitz (2007) established fundamental results in functional linear mean regression, deriving sharp rates of convergence for a PCA-based estimator for the slope function under some smoothness assumptions. Note that they assumed that covariates are continuously observed. Other than functional linear mean regression, Müller and Stadtmüller (2005) developed estimation methods for generalized functional linear models using series expansions of covariates and slope functions.

While not many, there are some earlier papers on estimating conditional quantiles with function-valued covariates. Cardot, Crambes and Sarda (2005) studied smoothing splines estimators for functional linear quantile regression models, while their established rates are not sharp. Ferraty, Rabhi and Vieu (2005) considered nonparametric estimation of conditional quantiles when covariates are functions, which is a different topic than ours. Chen and Müller (2012) considered an "indirect" estimation of conditional quantiles. They modeled the conditional distribution as the composition of some (possibly unknown) link function and a linear functional of the covariate. They first estimated the conditional distribution function by adapting the method developed in Müller and Stadtmüller (2005) and then estimated the conditional quantile function by inverting the estimated conditional distribution function. In the quantile regression literature, there are two ways to estimate conditional quantiles. One is to directly model conditional quantiles and estimate unknown parameters by minimizing check functions. The other is to estimate conditional distribution functions and invert them to estimate conditional quantile functions. We refer to the former as a "direct" method while to the latter as an "indirect" method. The approach taken in this paper is classified into a direct method, while that of Chen and Müller (2012) is classified into an indirect method. Note that although their method is flexible, they only established consistency of the estimator.

Conditional quantile estimation offers a variety of fruitful applications for data containing function-valued covariates. A leading example in which conditional quantile estimation with function-valued covariates is useful appears in analysis of growth data [Chen and Müller (2012)]. Suppose that we have a growth data set of girls' heights between ages 1 and 18, say, where multiple measurements may occur at some ages. Use girl's growth history between ages 1 and 12 as a covariate, and her height at age 18 as a response. Then, conditional quantile estimation gives us an overall picture of the predictive distribution of girl's height at age 18 given her growth history between ages 1 and 12, which is more informative than just knowing the mean response. In addition to growth data, functional quantile regression has been applied in analysis of ozone pollution data [Cardot, Crambes and Sarda (2007)] and El Niño data [Ferraty, Rabhi and Vieu (2005)]. We believe that functional linear quantile regression modeling is a benchmark modeling in conditional quantile estimation when covariates are functions, just as linear quantile regression modeling is so when covariates are vectors.

Our estimator for the slope function (at a fixed quantile index) can be understood as a regularized solution to an empirical version of a *nonlinear* ill-posed inverse problem that corresponds to the "normal equation" in the quantile regression case, where the regularization is controlled by the cutoff level. The paper is thus in part related to the literature on statistical nonlinear inverse problems, which is still an ongoing research area [see Bissantz, Hohage and Munk (2004), Chen and Pouzo (2012), Gagliardini and Scaillet (2012), Horowitz and Lee (2007), Loubes and Ludeña (2010)]. On the other hand, in the mean regression case, the normal equation becomes a *linear* ill-posed inverse problem. Hall and Horowitz (2007) considered two regularized estimators for the slope function based on the normal equation in the mean regression case. Conceptually, the problems handled in our and their papers are different in their nature: linearity and nonlinearity.

From a technical point of view, establishing sharp rates of convergence for our estimators is challenging. Our proof strategy builds upon the techniques developed in the asymptotic analysis for M-estimators with diverging numbers of parameters [see, e.g., Belloni, Chernozhukov and Fernández-Val (2011), He and Shao (2000)]. However, the additional complication arises essentially because "regressors" here are estimated ones and the estimation error has to be taken into account, which requires some new techniques such as "conditional" maximal inequalities and some careful moment calculations. Additionally, the discretization error brings a further technical complication.

Finally, the setting here is similar to Section 3 of Crambes, Kneip and Sarda (2009): covariates are densely but discretely observed, and the discretization error is taken into account in the analysis. However, the paper does not cover the case in which covariates are discretely observed with measurement errors because of the technical complication. A formal theoretical analysis in such a case is left to a future work.

The remainder of the paper is organized as follows. Section 2 presents the model and estimators. Section 3 gives the main results in which we derive rates of convergence for the estimators. Section 4 discusses empirical choice of the cutoff level. A proof of Theorem 3.1 is given in Section 5. Some additional discussions, technical proofs, useful technical tools and simulation results are provided in the Appendix. Due to the space limitation, the Appendix is contained in the supplementary file [Kato (2013)].

NOTATION. Throughout the paper, we shall obey the following notation. For $z \in \mathbb{R}^k$, let $\|z\|_{\ell^2}$ denote the Euclidean norm of z. For any integer $k \geq 2$, let $\mathbb{S}^{k-1}$ denote the set of all unit vectors in $\mathbb{R}^k$: $\mathbb{S}^{k-1} = \{z \in \mathbb{R}^p : \|z\|_{\ell^2} = 1\}$. For any $y, z \in \mathbb{R}$, let $y \vee z = \max\{y, z\}$ and $y \wedge z = \min\{y, z\}$. Let $1(\cdot)$ denote the indicator function. For any given (random or nonrandom, scalar or vector) sequence $\{z_i\}_{i=1}^n$, we use the notation

$$\mathbb{E}_n[z_i] = \frac{1}{n} \sum_{i=1}^n z_i,$$

which should be distinguished from the population expectation $\mathbb{E}[\cdot]$. For any two sequences of positive constants r_n and s_n, we write $r_n \asymp s_n$ if the ratio r_n/s_n is bounded and bounded away from zero. Let $L_2[0,1]$ denote the usual L_2 space with respect to the Lebesgue measure for functions defined on $[0,1]$. Let $\|\cdot\|$ denote the L_2-norm: $\|f\|^2 = \int_0^1 f^2(t)\,dt$. For any finite set I, $\mathrm{Card}(I)$ denotes the cardinality of I.

2. Methodology.

2.1. *Functional linear quantile regression modeling.* Let (Y, X) be a pair of a scalar random variable Y and a random function $X = (X(t))_{t\in\mathcal{T}}$ on a bounded closed interval $\mathcal{T}$ in $\mathbb{R}$. Without loss of generality, we assume $\mathcal{T} = [0,1]$. By "random function," we mean that $X(t)$ is a random variable for each $t \in [0,1]$. For our purpose of formulating a functional linear quantile regression model, we need the existence of the regular conditional distribution of Y given X, to which end we assume a mild regularity condition on the path property of X. Let $D[0,1]$ denote the space of all càdlàg functions on $[0,1]$, equipped with the Skorohod metric [see Billingsley (1968)]. We assume that the map $t \mapsto X(t)$ is càdlàg almost surely. Equip $D[0,1]$ with the Borel σ-field. Then, X can be taken as a $D[0,1]$-valued random variable. Since $D[0,1]$ is a Polish space, and the product space $\mathbb{R}\times D[0,1]$ with the product metric is also Polish, the regular conditional distribution of Y given X exists [see, e.g., Dudley (2002), Theorem 10.2.2].

Let $Q_{Y|X}(\cdot|X)$ denote the conditional quantile function of Y given X. Let $\mathcal{U}$ be a given subset of $(0,1)$ that is away from 0 and 1, that is, for some small $\epsilon \in (0,1/2)$, $\mathcal{U} \subset [\epsilon, 1-\epsilon]$. For each $u \in \mathcal{U}$, we assume that $Q_{Y|X}(u|X)$ can be written as a linear functional of X, that is, for each $u \in \mathcal{U}$, there exist a scalar constant $a(u) \in \mathbb{R}$ and a scalar function $b(\cdot,u) \in L_2[0,1]$ such that

$$Q_{Y|X}(u|X) = a(u) + \int_0^1 b(t,u)X^c(t)\,dt, \qquad u \in \mathcal{U}, \tag{2.1}$$

where $X^c(t) = X(t) - \mathbb{E}[X(t)]$. Typical examples of $\mathcal{U} \subset (0,1)$ are as follows: (i) $\mathcal{U} = \{u\}$ (singleton); (ii) $\mathcal{U} = \{u_1,\dots,u_K\}$ with $0 < u_1 < \cdots < u_K < 1$ (finite set); (iii) $\mathcal{U} = [u_L, u_U]$ with $0 < u_L < u_U < 1$ (bounded closed interval). Formally, we allow for all these possibilities.

The model (2.1) is a natural extension of standard linear quantile regression models to function-valued covariates and was first formulated in Cardot, Crambes and Sarda (2005). In what follows, we consider estimating the slope function $(t,u) \mapsto b(t,u)$ and the conditional quantile function $(u,x) \mapsto Q_{Y|X}(u|x)$.

Before going to the next step, we briefly give two simple examples of data generating processes that admit the conditional quantile restriction (2.1).

EXAMPLE 2.1 (Linear location design). Suppose that (Y, X) obeys the linear location design

$$Y = c + \int_0^1 \varrho(t)X(t)\,dt + \varepsilon, \qquad \varepsilon \perp\!\!\!\perp X,$$

where c is a constant and $\varrho(t)$ is a function in $L_2[0, 1]$. Let F_ε denote the distribution function of ε and denote by F_ε^{-1} the quantile function of F_ε. Then (Y, X) obeys the conditional quantile restriction

$$Q_{Y|X}(u|X) = c + F_\varepsilon^{-1}(u) + \int_0^1 \varrho(t)X(t)\,dt, \qquad u \in (0, 1).$$

EXAMPLE 2.2 (Linear location-scale design). Suppose that (Y, X) obeys the linear location-scale design

$$Y = c_1 + \int_0^1 \varrho_1(t)X(t)\,dt + \sigma(X)\varepsilon, \qquad \sigma(X) = c_2 + \int_0^1 \varrho_2(t)X(t)\,dt, \varepsilon \perp\!\!\!\perp X,$$

where c_1, c_2 are constants and $\varrho_1(t), \varrho_2(t)$ are functions in $L_2[0, 1]$. Suppose that $\sigma(X) > 0$ almost surely. Denote by F_ε^{-1} the quantile function of the distribution of ε. Then (Y, X) obeys the conditional quantile restriction

$$Q_{Y|X}(u|X) = c_1 + c_2 F_\varepsilon^{-1}(u) + \int_0^1 \{\varrho_1(t) + \varrho_2(t)F_\varepsilon^{-1}(u)\}X(t)\,dt, \qquad u \in (0, 1).$$

In this design, the slope function depends on the quantile index.

2.2. *Estimation strategy.* We base our estimation strategy on the principal component analysis (PCA). To this end, we additionally assume here that $\int_0^1 \mathbb{E}[X^2(t)]\,dt < \infty$. Define the covariance kernel $K(s, t) = \mathrm{Cov}(X(s), X(t))$. Then, by the Hilbert–Schmidt theorem, $K(s, t)$ admits the spectral expansion

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

where $\kappa_1 \geq \kappa_2 \geq \cdots \geq 0$ are ordered eigenvalues, and $\{\phi_j\}_{j=1}^{\infty}$ is an orthonormal basis of $L_2[0, 1]$ consisting of eigenfunctions of the integral operator from $L_2[0, 1]$ to itself with kernel $K(s, t)$. We will later assume that κ_j are all positive and there are no ties in κ_j, that is, $\kappa_1 > \kappa_2 > \cdots > 0$. Since $\{\phi_j\}_{j=1}^{\infty}$ is an orthonormal basis of $L_2[0, 1]$, we have the following expansions in $L_2[0, 1]$:

$$X^c(t) = \sum_{j=1}^{\infty} \xi_j \phi_j(t), \qquad b(t, u) = \sum_{j=1}^{\infty} b_j(u)\phi_j(t),$$

where ξ_j and $b_j(u)$ are defined by

$$\xi_j = \int_0^1 X^c(t)\phi_j(t)\,dt, \qquad b_j(u) = \int_0^1 b(t, u)\phi_j(t)\,dt.$$

The ξ_j are called "principal scores" for X. The expansion for X^c is called the "Karhunen–Loève expansion." This leads to the expression $\int_0^1 b(t, u)X^c(t)\,dt =$

$\sum_{j=1}^{\infty} b_j(u)\xi_j$. Then, the model (2.1) is transformed into a quantile regression model with an infinite number of "regressors":

$$Q_{Y|X}(u|X) = a(u) + \sum_{j=1}^{\infty} b_j(u)\xi_j, \qquad u \in \mathcal{U}. \tag{2.2}$$

Note that $\mathbb{E}[\xi_j] = 0$, $\mathbb{E}[\xi_j^2] = \kappa_j$ and $\mathbb{E}[\xi_j \xi_k] = 0$ for all $j \neq k$.

We first consider estimating the slope function $(t, u) \mapsto b(t, u)$. To this end, we estimate the function $b(\cdot, u)$ for each $u \in \mathcal{U}$ and collect them to construct a final estimator for $(t, u) \mapsto b(t, u)$. To explain the basic idea, suppose for a while that (i) X were continuously observable; and (ii) the covariance kernel $K(s, t)$ were known. The problem then reduces to finding suitable estimates of the coefficients $b_j(u)$. Let $(Y_1, X_1), \dots, (Y_n, X_n)$ be independent copies of (Y, X). For each $i = 1, \dots, n$, let ξ_{ij} be the principal scores for X_i. Pick any $u \in \mathcal{U}$. Then, a plausible approach to estimating $b(\cdot, u)$ is to truncate $\sum_{j=1}^{\infty} b_j(u)\xi_j$ by $\sum_{j=1}^{m} b_j(u)\xi_j$ for some large m, and estimate only the first m coefficients $b_1(u), \dots, b_m(u)$ using a standard quantile regression technique. Let $m = m_n$ be the "cutoff" level such that $1 \le m \le n - 1$ and $m \to \infty$ as $n \to \infty$. Estimate $a(u)$ and the first m coefficients $b_1(u), \dots, b_m(u)$ of $b(\cdot, u)$ by

$$\bigl(\tilde{a}(u), \tilde{b}_1(u), \dots, \tilde{b}_m(u)\bigr) = \operatorname*{arg\,min}_{a, b_1, \dots, b_m} \mathbb{E}_n\left[\rho_u\left(Y_i - a - \sum_{j=1}^{m} b_j \xi_{ij}\right)\right], \tag{2.3}$$

where $\rho_u(y) = \{u - 1(y \le 0)\}y$ is the check function [Koenker and Bassett (1978)]. Note that for $u = 0.5$, $\rho_{0.5}(\cdot)$ is equivalent to the absolute value function. Here recall that $\mathbb{E}_n[z_i] = n^{-1}\sum_{i=1}^{n} z_i$ for any sequence $\{z_i\}_{i=1}^{n}$. The resulting estimator for the slope function $(t, u) \mapsto b(t, u)$ is given by

$$\tilde{b} : (t, u) \mapsto \tilde{b}(t, u), \qquad \tilde{b}(t, u) = \sum_{j=1}^{m} \tilde{b}_j(u)\phi_j(t), \qquad t \in [0, 1], u \in \mathcal{U}.$$

However, this "estimator" is infeasible since (i) X is usually discretely observed; and (ii) $K(s, t)$ is unknown. Because of (i), it is usually not possible to directly estimate the covariance kernel $K(s, t)$ by the empirical one (since $\mathbb{E}_n[(X_i(s) - \bar{X}(s))(X_i(t) - \bar{X}(t))]$ with $\bar{X}(t) = n^{-1}\sum_{i=1}^{n} X_i(t)$ is unavailable for some (s, t)). Similarly to Crambes, Kneip and Sarda (2009), we consider the following setting:

(1) For each $i = 1, \dots, n$, X_i is observed only at $L_i + 1$ discrete points $0 = t_{i1} < t_{i2} < \cdots < t_{i,L_i+1} = 1$, that is, we only observe $X_i(t_{ij})$, $j = 1, \dots, L_i + 1$. Typically, $\max_{1 \le i \le n} \max_{1 \le l \le L_i}(t_{i,l+1} - t_{il}) \to 0$ as $n \to \infty$ is assumed.

(2) Based on the discrete observations, for each $i = 1, \dots, n$, construct an interpolated function $\hat{X}_i = (\hat{X}_i(t))_{t \in [0,1]}$ for $X_i = (X_i(t))_{t \in [0,1]}$.

Here we shall use a simple interpolation rule (see also the later remark):

$$\hat{X}_i(t) = \sum_{l=1}^{L_i} X(t_{il}) 1\big(t \in [t_{il}, t_{i,l+1})\big), \qquad i = 1, \dots, n.$$

The observation time points $t_{i1}, \dots, t_{i,L_i+1}$ (and L_i) should be indexed by the sample size n, however, this is suppressed for the notational convenience. Suppose now that the interpolated functions $\hat{X}_1, \dots, \hat{X}_n$ are obtained. Then, we may estimate the covariance kernel $K(s,t)$ by

$$\hat{K}(s,t) = \mathbb{E}_n\big[\big(\hat{X}_i(s) - \bar{\hat{X}}(s)\big)\big(\hat{X}_i(t) - \bar{\hat{X}}(t)\big)\big],$$

where $\bar{\hat{X}}(t) = n^{-1}\sum_{i=1}^n \hat{X}_i(t)$. Let $\hat{K}(s,t) = \sum_{j=1}^\infty \hat{\kappa}_j \hat{\phi}_j(s)\hat{\phi}_j(t)$ be the spectral expansion of $\hat{K}(s,t)$, where $\hat{\kappa}_1 \ge \hat{\kappa}_2 \ge \cdots \ge 0$ are ordered eigenvalues and $\{\hat{\phi}_j\}_{j=1}^\infty$ is an orthonormal basis of $L_2[0,1]$ consisting of eigenfunctions of the integral operator from $L_2[0,1]$ to itself with kernel $\hat{K}(s,t)$. Each principal score ξ_{ij} is now estimated by

$$\hat{\xi}_{ij} = \int_0^1 \big(\hat{X}_i(t) - \bar{\hat{X}}(t)\big)\hat{\phi}_j(t)\,dt.$$

Then, the coefficients $a(u)$ and $b_1(u), \dots, b_m(u)$ are estimated by

$$\big(\hat{a}(u), \hat{b}_1(u), \dots, \hat{b}_m(u)\big) = \operatorname*{arg\,min}_{a, b_1, \dots, b_m} \mathbb{E}_n\left[\rho_u\left(Y_i - a - \sum_{j=1}^m b_j \hat{\xi}_{ij}\right)\right]. \tag{2.4}$$

The resulting estimator for the slope function $(t,u) \mapsto b(t,u)$ is given by

$$\hat{b} : (t,u) \mapsto \hat{b}(t,u), \qquad \hat{b}(t,u) = \sum_{j=1}^m \hat{b}_j(u)\hat{\phi}_j(t), \qquad t \in [0,1], u \in \mathcal{U}.$$

The optimization problem (2.4) can be transformed into a linear programming problem and can be solved by using standard statistical softwares. Once the estimator for the slope function is obtained, the conditional u-quantile of Y given $X = x$ for a given function $x = (x(t))_{t\in[0,1]} \in L_2[0,1]$ is estimated by a plug-in method:

$$\hat{Q}_{Y|X}(u|x) = \hat{a}(u) + \int_0^1 \hat{b}(t,u)\big(x(t) - \bar{\hat{X}}(t)\big)\,dt.$$

Empirical choice of the cutoff level will be discussed in Section 4.

The basis $\{\phi_j\}_{j=1}^\infty$ is called the (population) PCA basis. Alternatively, one may use other basis functions independent of the data, such as Fourier and wavelet bases, in which case the analysis becomes more tractable. A potential drawback of using such basis functions is that, as discussed in Delaigle and Hall (2012), using the "first" m basis functions is less motivated. The PCA basis is a benchmark basis in functional data analysis, which is the reason why the PCA basis is used in this paper. Other estimation methods such as smoothing splines [Crambes, Kneip and Sarda (2009)] and a reproducing kernel Hilbert space approach [Yuan and Cai (2010)] could be adapted in the quantile regression case, which is left as a future topic.

The interpolation rule used here may be replaced by any other reasonable interpolation rule. For example, a plausible alternative is to use

$$\hat{X}_i^{\mathrm{mid}}(t) = \sum_{l=1}^{L_i} \frac{X(t_{il}) + X(t_{i,l+1})}{2} 1\big(t \in [t_{il}, t_{i,l+1})\big), \qquad i = 1, \dots, n.$$

It is not hard to see that the theory below also applies to this interpolation rule. In practice, this interpolation rule may be more recommended since it uses all the discrete observations $X_i(t_{i1}), \dots, X_i(t_{i,L_i+1})$.

Finally, we note that for any fixed $u \in \mathcal{U}$, our estimator $\hat{b}(\cdot, u)$ can be understood as a regularized solution to an empirical version of a nonlinear ill-posed inverse problem that corresponds to the "normal equation" in the quantile regression case. Due to the space limitation, we shall discuss this connection to nonlinear ill-posed inverse problems in Appendix A in the supplementary file [Kato (2013)].

2.3. *Monotonization.* Suppose in this section that $\mathcal{U}$ is a bounded closed interval: $\mathcal{U} = [u_L, u_U]$ with $0 < u_L < u_U < 1$. The conditional quantile function $Q_{Y|X}(u|x)$ is monotonically nondecreasing in u. However, the plug-in estimator $\hat{Q}_{Y|X}(u|x)$ constructed is not necessarily so. To circumvent this problem, we may monotonize the map $u \mapsto \hat{Q}_{Y|X}(u|x)$ by one of the following three methods: (i) rearrangement [Chernozhukov, Fernández-Val and Galichon (2009)], (ii) isotonization [Barlow et al. (1972)], (iii) convex combination of (i) and (ii). Such methods are explained in Chernozhukov, Fernández-Val and Galichon (2009) in a general setup. We briefly explain their basic ideas. Pick any $x \in L_2[0, 1]$.

(i) Rearrangement. Define the function

$$\hat{F}_{\mathcal{U}}(y|x) = \frac{1}{u_U - u_L} \int_{\mathcal{U}} 1\{\hat{Q}_{Y|X}(u|x) \le y\}\, du.$$

Then the map $y \mapsto \hat{F}_{\mathcal{U}}(y|x)$ is a proper distribution function supported in $[\min_{u \in \mathcal{U}} \hat{Q}_{Y|X}(u|x), \max_{u \in \mathcal{U}} \hat{Q}_{Y|X}(u|x)]$. The rearranged estimator for $Q_{Y|X}(u|x)$ is defined by

$$\hat{Q}^*_{Y|X}(u|x) = \hat{F}_{\mathcal{U}}^{-1}\left(\frac{u - u_L}{u_U - u_L}\Big| x\right).$$

Clearly, the map $u \mapsto \hat{Q}^*_{Y|X}(u|x)$ is nondecreasing.

(ii) Isotonization. The isotonization is carried out by projecting the original estimate $u \mapsto \hat{Q}_{Y|X}(u|x)$ on the set of nondecreasing functions, typically by the "pool adjacent violators" algorithm. Denote by $\hat{Q}^I_{Y|X}(u|x)$ the isotonized estimator.

(iii) Convex combination of (i) and (ii). Take $\lambda \in [0, 1]$. Then, the convex combination $\hat{Q}^\lambda_{Y|X}(u|x) = \lambda \hat{Q}^*_{Y|X}(u|x) + (1-\lambda)\hat{Q}^I_{Y|X}(u|x)$ is nondecreasing in u.

By Chernozhukov, Fernández-Val and Galichon (2009), it is shown that any monotonized estimate [constructed by using one of (i)–(iii)] is at least as good as the initial estimate $\hat{Q}_{Y|X}(u|x)$ in the following sense: let $\hat{Q}^\dagger_{Y|X}(u|x)$ be any monotonized estimate for $Q_{Y|X}(u|x)$ given above. Then, we have for all $q \in [1, \infty]$,

$$\begin{aligned} &\left[\int_{\mathcal{U}} |\hat{Q}^\dagger_{Y|X}(u|x) - Q_{Y|X}(u|x)|^q \, du\right]^{1/q} \\ &\qquad \le \left[\int_{\mathcal{U}} |\hat{Q}_{Y|X}(u|x) - Q_{Y|X}(u|x)|^q \, du\right]^{1/q}, \end{aligned} \tag{2.5}$$

where the obvious modification is made when $q = \infty$.

3. Rates of convergence. In this section we derive rates of convergence for the estimators defined in the previous section and argue their optimality. We make the following assumptions. Let $C_1 > 1$ be some sufficiently large constant. First of all, we assume the following:

(A1) $\{(Y_i, X_i)\}_{i=1}^\infty$ is i.i.d. with (Y, X).
(A2) $\int_0^1 \mathbb{E}[X^4(t)]\,dt \le C_1$ and $\mathbb{E}[\xi_j^4] \le C_1 \kappa_j^2$ for all $j \ge 1$.

The i.i.d. assumption is conventional. It is beyond the scope of the paper to extend the theory to dependent data. Note that Hörmann and Kokoszka (2010) discussed weakly dependent functional data. Assumption (A2) is a mild moment restriction. Here no moment condition on Y is needed, so $\mathbb{E}[Y]$ may not exist.

(A3) For some $\alpha > 1$, $C_1^{-1} j^{-\alpha} \le \kappa_j \le C_1 j^{-\alpha}$ and $\kappa_j - \kappa_{j+1} \ge C_1^{-1} j^{-\alpha-1}$ for all $j \ge 1$.
(A4) For some $\beta > \alpha/2 + 1$, $\sup_{u \in \mathcal{U}} |b_j(u)| \le C_1 j^{-\beta}$ for all $j \ge 1$.
(A5) Let $F_{Y|X}(y|X)$ denote the conditional distribution function of Y given X. Then, for every realization of X, the map $y \mapsto F_{Y|X}(y|X)$ is twice continuously differentiable with $f_{Y|X}(y|X) = \partial F_{Y|X}(y|X)/\partial y$ and $f'_{Y|X}(y|X) = \partial f_{Y|X}(y|X)/\partial y$. Furthermore, $f_{Y|X}(y|X) \vee |f'_{Y|X}(y|X)| \le C_1$.
(A6) $\inf_{u \in \mathcal{U}} f_{Y|X}(Q_{Y|X}(u|X)|X) \ge C_1^{-1}$.

Assumptions (A3) and (A4) are adapted from (3.2) and (3.3) of Hall and Horowitz (2007). Assumption (A3) especially means that all κ_j are positive, which guarantees identification of the slope function. In assumption (A3), α measures the smoothness of the covariance kernel K, which also measures the difficulty of estimating the slope function $(t,u) \mapsto b(t,u)$. The second part of assumption (A3) is to require the spacings among κ_j not be too small, which ensures identifiability of eigenfunctions ϕ_j and thereby sufficient estimation accuracy of $\hat{\phi}_j$. Note that the lower inequality for κ_j follows essentially from the condition on the difference, as $\kappa_j \to 0$ $(j \to \infty)$ and hence $\kappa_j = \sum_{k=j}^{\infty}(\kappa_k - \kappa_{k+1})$. Assumption (A4) determines the "smoothness" of the function $t \mapsto b(t,u)$. The condition that $\beta > \alpha/2 + 1$ requires the function $t \mapsto b(t,u)$ to be sufficiently smooth relative to K uniformly in $u \in \mathcal{U}$. See Hall and Horowitz [(2007), page 74] for some related discussions on these assumptions. Assumptions (A5) and (A6) are specific to the quantile regression case. Both assumptions are standard in the quantile regression literature when X is a vector. The role of assumption (A6) is to guarantee that the conditional distribution function $F_{Y|X}(y|X)$ is not "flat" near quantiles of interest, which is essential to our asymptotic study.

(A7) For each $i = 1, \dots, n$, X_i is observed only at discrete points $0 = t_{i1} < t_{i2} < \cdots < t_{i,L_i+1} = 1$. Define $\Delta = \Delta_n = \max_{1 \le i \le n} \max_{1 \le l \le L_i}(t_{i,l+1} - t_{il})$. Then, $\Delta \to 0$ as $n \to \infty$.

(A8) There exists a constant $\gamma \in (0,2]$ such that $\mathbb{E}[(X(t) - X(s))^2] \le C_1|t - s|^{\gamma}$ for all $s, t \in [0,1]$.

Assumptions (A7) and (A8) are a set of sampling assumptions on X_i. A rate restriction will be imposed on Δ. Assumption (A7) at least requires that $\min_{1 \le i \le n} L_i \to \infty$ as $n \to \infty$, which means that each set of discrete points $t_{i1}, \dots, t_{i,L_i+1}$ has to be dense in $[0,1]$ as the sample size grows. Assumption (A8) is an additional assumption on the smoothness of the covariance kernel as well as the mean function $t \mapsto \mathbb{E}[X(t)]$. Suppose in this discussion $\mathbb{E}[X(t)] = 0$ for all $t \in [0,1]$. Then, for example, $\gamma = 1$ if $K(s,t)$ is Lipschitz continuous and $\gamma = 2$ if $K(s,t)$ is twice continuously differentiable. The value of γ controls the discretization error. Note that possible values of γ depend on the value of α. Typically, if α is sufficiently large, (A8) is satisfied with $\gamma = 2$ (and vice versa). The relation between the smoothness of the covariance (or more generally integral) kernel and the decay rate of the associated eigenvalues is studied in Ritter, Wasilkowski and Woźniakowski (1995) and Ferreira and Menegatto (2009) and the references therein. For example, the former paper shows that, if K verifies the "Sacks–Ylvisaker condition" of order r with r a nonnegative integer (see the original paper for the precise description of the conditions), in which case (A8) is satisfied with $\gamma = 2$, then $\kappa_j \asymp j^{-2r-2}$ as $j \to \infty$. Assumption (A8) is similar in spirit to (A2) of Crambes, Kneip and Sarda (2009), in which they directly assumed some smoothness of the random function $t \mapsto X(t)$ to deal with the discretization error [roughly speaking, their 2κ corresponds to our γ as (A2) of Crambes, Kneip and Sarda (2009) essentially assumes X to be κ-Hölder continuous and in that case $\mathbb{E}[(X(t) - X(s))^2] = O(|t - s|^{2\kappa})$].

Let $\mathcal{F} = \mathcal{F}(C_1, \alpha, \beta, \gamma)$ denote the set of all distributions of (Y, X) compatible with assumptions (A2)–(A6) and (A8) for given (admissible) values of C_1, α, β and γ (such that $\mathcal{F} \neq \varnothing$). The following theorem, which will be proved in Section 5 below, establishes rates of convergence for the slope estimator $(t, u) \mapsto \hat{b}(t, u)$.

THEOREM 3.1. *Suppose that assumptions* (A1)–(A8) *are satisfied. Take* $m = m_n \asymp n^{1/(\alpha+2\beta)}$. *Then, we have*

$$\lim_{M\to\infty} \limsup_{n\to\infty} \sup_{F\in\mathcal{F}} \mathbb{P}_F\left[\sup_{u\in\mathcal{U}} \int_0^1 \{\hat{b}(t, u) - b(t, u)\}^2 \, dt > Mn^{-(2\beta-1)/(\alpha+2\beta)}\right] = 0, \tag{3.1}$$

provided that $(n \vee (\log n)m^{3\alpha+3})\Delta^\gamma = O(1)$ *as* $n \to \infty$.

Inspection of the proof of Theorem 3.1 shows that, when X is continuously observable, under assumptions (A1)–(A6), the rate of convergence of the estimator based on the direct empirical covariance kernel will be $n^{-(2\beta-1)/(\alpha+2\beta)}$. The side condition that $(n \vee (\log n)m^{3\alpha+3})\Delta^\gamma = O(1)$ is assumed to make the discretization error negligible. This condition seems not quite restrictive. For example, if $\beta \geq \alpha + 3/2$ and $\gamma = 2$, it is satisfied as long as $\Delta = O((n \log n)^{-1/2})$, which seems to be mild in view of some applications in functional data analysis.

The following theorem, which will be proved in Appendix B in the supplementary file [Kato (2013)], establishes rates of convergence for $\hat{Q}_{Y|X}(u|x)$. For notational convenience, define

$$\mathcal{E}(\hat{Q}_{Y|X}, u) = \int \{\hat{Q}_{Y|X}(u|x) - Q_{Y|X}(u|x)\}^2 \, dP_X(x),$$

where P_X denotes the distribution of X (defined on $D[0, 1]$).

THEOREM 3.2. *Suppose that assumptions* (A1)–(A8) *are satisfied. Take* $m = m_n \asymp n^{1/(\alpha+2\beta)}$. *Then, we have*

$$\lim_{M\to\infty} \limsup_{n\to\infty} \sup_{F\in\mathcal{F}} \mathbb{P}_F\left[\sup_{u\in\mathcal{U}} \mathcal{E}(\hat{Q}_{Y|X}, u) > Mn^{-(\alpha+2\beta-1)/(\alpha+2\beta)}\right] = 0, \tag{3.2}$$

provided that $(n \vee (\log n)m^{3\alpha+3})\Delta^\gamma = O(1)$ *as* $n \to \infty$.

For monotonized estimators, the following corollary directly follows in view of (2.5) and Theorem 3.2.

COROLLARY 3.1. *Let* $\mathcal{U}$ *be a bounded closed interval in* $(0, 1)$. *Suppose that all the assumptions of Theorem 3.2 are satisfied. Let* $\hat{Q}^\dagger_{Y|X}(u|x)$ *be any monotonized estimator for* $Q_{Y|X}(u|x)$ *given in Section 2.3. Then, we have*

$$\lim_{M\to\infty} \limsup_{n\to\infty} \sup_{F\in\mathcal{F}} \mathbb{P}_F\left[\int_{\mathcal{U}} \mathcal{E}(\hat{Q}^\dagger_{Y|X}, u) \, du > Mn^{-(\alpha+2\beta-1)/(\alpha+2\beta)}\right] = 0.$$

Here note that the rate $n^{-(\alpha+2\beta-1)/(\alpha+2\beta)}$ attained in estimating $Q_{Y|X}(u|x)$ is faster than the rate $n^{-(2\beta-1)/(\alpha+2\beta)}$ attained in estimating $b(t,u)$.

In what follows, we discuss optimality of these rates.

PROPOSITION 3.1. *Suppose that assumptions* (A1)–(A6) *and* (A8) *are satisfied. Let* γ *be such that*

$$0<\gamma\begin{cases}<\alpha-1, & \text{if } \alpha\le 3,\\ \le 2, & \text{if } \alpha>3.\end{cases} \tag{3.3}$$

Then, there exists a constant $M>0$ *such that for* $\mathcal{F}=\mathcal{F}(C_1,\alpha,\beta,\gamma)$,

$$\liminf_{n\to\infty}\inf_{\bar b}\sup_{F\in\mathcal{F}}\mathbb{P}_F\left[\sup_{u\in\mathcal{U}}\int_0^1\{\bar b(t,u)-b(t,u)\}^2\,dt > Mn^{-(2\beta-1)/(\alpha+2\beta)}\right]>0, \tag{3.4}$$

where $\inf_{\bar b}$ *is taken over all estimators for the slope function* $(t,u)\mapsto b(t,u)$ *based on* $(Y_1,X_1),\dots,(Y_n,X_n)$. *Similarly, there exists a constant* $M>0$ *such that for* $\mathcal{F}=\mathcal{F}(C_1,\alpha,\beta,\gamma)$,

$$\liminf_{n\to\infty}\inf_{\bar Q_{Y|X}}\sup_{F\in\mathcal{F}}\mathbb{P}_F\Big[\sup_{u\in\mathcal{U}}\mathcal{E}(\bar Q_{Y|X},u)>Mn^{-(\alpha+2\beta-1)/(\alpha+2\beta)}\Big]>0,$$

where $\inf_{\bar Q_{Y|X}}$ *is taken over all estimators for the conditional quantile function* $Q_{Y|X}:(u,x)\mapsto Q_{Y|X}(u|x)$ *based on* $(Y_1,X_1),\dots,(Y_n,X_n)$. *When* $\mathcal{U}$ *is a bounded closed interval in* $(0,1)$, *there exists a constant* $M>0$ *such that for* $\mathcal{F}=\mathcal{F}(C_1,\alpha,\beta,\gamma)$,

$$\liminf_{n\to\infty}\inf_{\bar Q_{Y|X}}\sup_{F\in\mathcal{F}}\mathbb{P}_F\left[\int_{\mathcal{U}}\mathcal{E}(\bar Q_{Y|X},u)\,du>Mn^{-(\alpha+2\beta-1)/(\alpha+2\beta)}\right]>0,$$

where the previous convention applies.

The side condition (3.3) is a compatibility condition between assumptions (A3) and (A8). It is not addressed here whether this condition is tight. However, some restriction between α and γ is required in establishing lower bounds of rates of convergence to guarantee that the class $\mathcal{F}(C_1,\alpha,\beta,\gamma)$ is at least nonempty. Proposition 3.1 shows that under this side condition the rates established in Theorems 3.1, 3.2 and Corollary 3.1 are indeed optimal in the minimax sense. A proof of Proposition 3.1 is given in Appendix C in the supplementary file [Kato (2013)].

4. Empirical choice of the cutoff level. In this section we suggest three criteria to choose m, and investigate their performance by simulations.[2] We

[2]Application of a (weighted) Lasso type procedure could be an alternative in the m-selection. Also, a thresholding rule for the estimated eigenvalues is a possible option. Analysis of such procedures is left to a future work.

use a heuristic reasoning to derive selection criteria. Suppose that $\mathcal{U}$ is a singleton: $\mathcal{U} = \{u\}$. Suppose that there is no truncation bias, that is, $b(t,u) = \sum_{j=1}^{m} b_j(u)\phi_j(t)$ and $Q_{Y|X}(u|X) = a(u) + \sum_{j=1}^{m} b_j(u)\xi_j$. Then, the infeasible estimator $(\tilde{a}(u), \tilde{b}_1(u), \ldots, \tilde{b}_m(u))'$ defined by (2.3) can be regarded as a (conditional) maximum likelihood estimator when the conditional distribution of Y given X has the asymmetric Laplace density of the form

$$f(y|X,u,\sigma) = \frac{u(1-u)}{\sigma} \exp\left\{-\frac{1}{\sigma}\rho_u\left(y - a(u) - \sum_{j=1}^{m} b_j(u)\xi_j\right)\right\},$$

where $\sigma > 0$ is a scale parameter. This suggests the following analogues of AIC and BIC in the present context:

$$\texttt{AIC}(u) = \log\left[\frac{1}{n}\sum_{i=1}^{n}\rho_u\left(Y_i - \hat{a}(u) - \sum_{j=1}^{m}\hat{b}_j(u)\hat{\xi}_{ij}\right)\right] + \frac{(m+1)}{n},$$

$$\texttt{BIC}(u) = \log\left[\frac{1}{n}\sum_{i=1}^{n}\rho_u\left(Y_i - \hat{a}(u) - \sum_{j=1}^{m}\hat{b}_j(u)\hat{\xi}_{ij}\right)\right] + \frac{(m+1)\log n}{n}.$$

See also Koenker [(2005), Section 4.9.1] for some related discussion. According to Yuan (2006), we may also consider an analogue of GACV as follows:

$$\texttt{GACV}(u) = \frac{\sum_{i=1}^{n}\rho_u(Y_i - \hat{a}(u) - \sum_{j=1}^{m}\hat{b}_j(u)\hat{\xi}_{ij})}{n-(m+1)}.$$

When $\mathcal{U}$ is a bounded closed interval, define the integrated-AIC, BIC and GACV as follows:

$$\texttt{IAIC} = \int_{\mathcal{U}} \texttt{AIC}(u)\,du, \qquad \texttt{IBIC} = \int_{\mathcal{U}} \texttt{BIC}(u)\,du,$$

$$\texttt{IGACV} = \int_{\mathcal{U}} \texttt{GACV}(u)\,du.$$

When $\mathcal{U}$ is a set of finite grid points, each integral is replaced by the summation over the grid points.

We carried out a small Monte Carlo study to investigate the finite sample performance of these criteria. In all cases, the number of Monte Carlo repetitions was 1000. The numerical results obtained in this section were carried out by using the matrix language Ox [Doornik (2002)]. The Ox code for solving quantile regression problems supplied on Professor Koenker's website was used. See also Portnoy and Koenker (1997) for some computational aspects of quantile regression problems.

The simulation design is described as follows:

$$Y = \int_0^1 \varrho(t)X(t)\,dt + \varepsilon,$$

$$\varrho(t) = \sum_{j=1}^{50} \varrho_j \phi_j(t),$$

$$\varrho_1 = 0.3, \varrho_j = 4(-1)^{j+1} j^{-2}, j \geq 2, \phi_j(t) = 2^{1/2}\cos(j\pi t),$$

$$X(t) = \sum_{j=1}^{50} \gamma_j Z_j \phi_j(t),$$

$$\gamma_j = (-1)^{j+1} j^{-\alpha/2}, \alpha \in \{1.1, 2\}, Z_j \sim U[-3^{1/2}, 3^{1/2}],$$

$$\varepsilon \sim N(0,1) \text{ or Cauchy}, \qquad n \in \{100, 200, 500\}.$$

This α corresponds to α in assumption (A3). Each X_i is observed at 201 equally spaced grid points on [0, 1]. In this design, we have

$$Q_{Y|X}(u|X) = F_\varepsilon^{-1}(u) + \int_0^1 \varrho(t)X(t)\,dt,$$

where $F_\varepsilon^{-1}(\cdot)$ is the quantile function of the distribution of ε. Thus, $a(u) = F_\varepsilon^{-1}(u)$ and $b(t,u) \equiv \varrho(t)$ [$b(t,u)$ is independent of u]. We considered two cases for $\mathcal{U}$: (a) $\mathcal{U} = \{0.5\}$ and (b) $\mathcal{U} = \{0.15, 0.2, \dots, 0.85\}$. In each case, the performance was measured by

$$\texttt{QA-MISE} = \frac{1}{\mathrm{Card}(\mathcal{U})} \sum_{u\in\mathcal{U}} \mathbb{E}\left[\int_0^1 \{\hat{b}(t,u) - b(t,u)\}^2\,dt\right] \quad \text{or}$$

$$\texttt{QA-MISE} = \frac{1}{\mathrm{Card}(\mathcal{U})} \sum_{u\in\mathcal{U}} \mathbb{E}\left[\int \{\hat{Q}_{Y|X}(u|x) - Q_{Y|X}(u|x)\}^2\,dP_X(x)\right],$$

where P_X denotes the distribution of X and QA-MISE is the abbreviation of "quantile-averaged mean integrated squared error."

The simulation results for case (a) are summarized in Figures 1–4 and Table 1. Figures 1 and 2 show the performance of the selection criteria for the normal error case, while Figures 3 and 4 show that for the Cauchy error case. In each figure, "Fixed" refers to the performance of the estimator with fixed m. Table 1 shows the average numbers of m selected by three criteria. The table shows that as the sample size increases, all the criteria tend to choose larger m, and as α increases, they tend to choose smaller m, which is consistent with the theoretical requirement in the m-selection. Interestingly, it is observed that the error distribution affects the performance of the m-selection.

In the normal error case, in Figure 1, BIC worked better than the other two criteria. On the other hand, in the Cauchy error case, in Figure 2, AIC and GACV

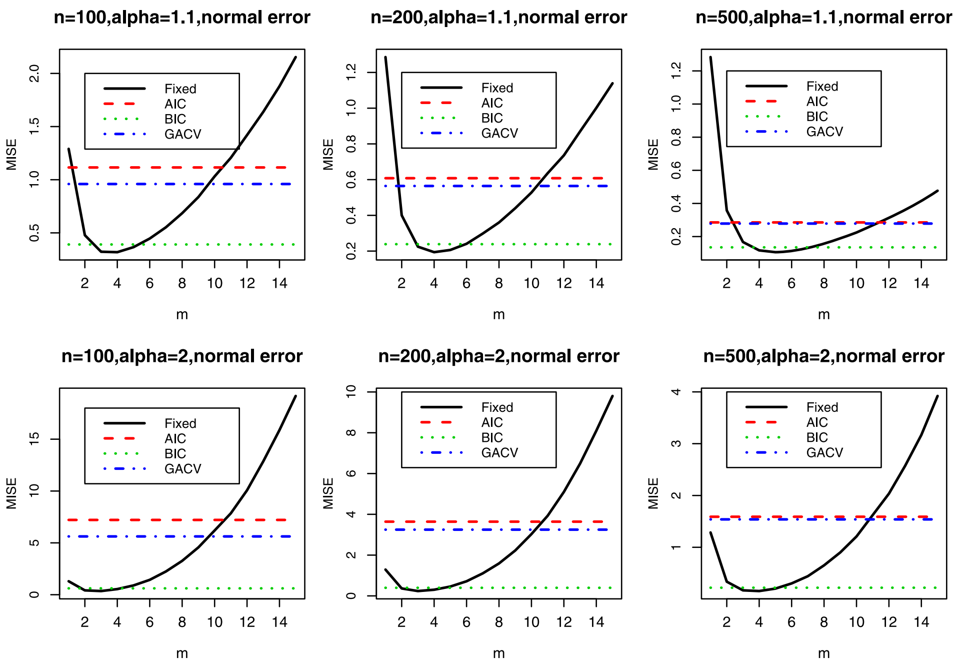

FIG. 1. *Performance of selection criteria. Case* (a). *Estimation of* $b(\cdot, 0.5)$.

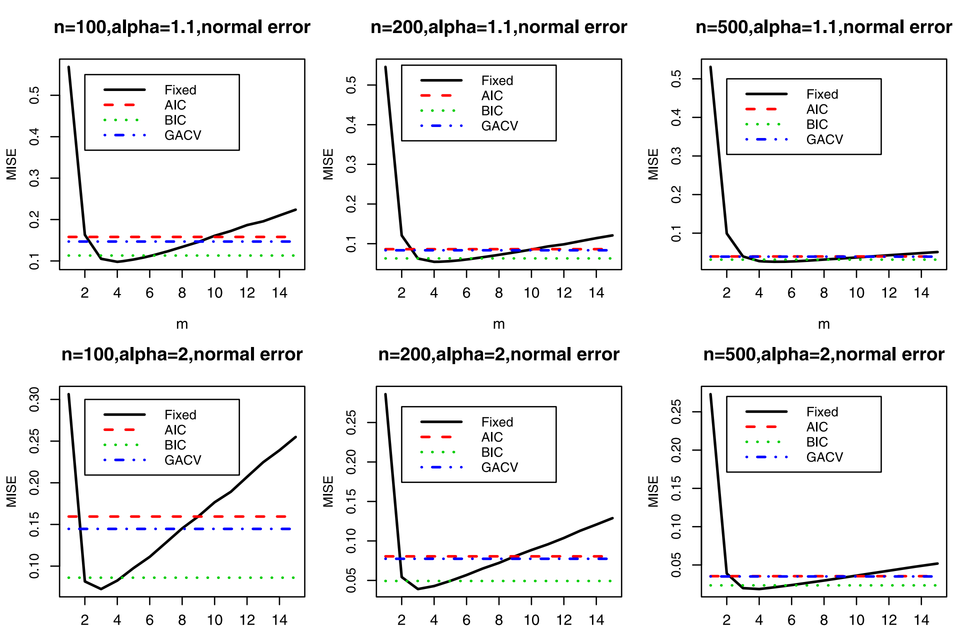

FIG. 2. *Performance of selection criteria. Case* (a). *Estimation of* $Q_{Y|X}(0.5|x)$.

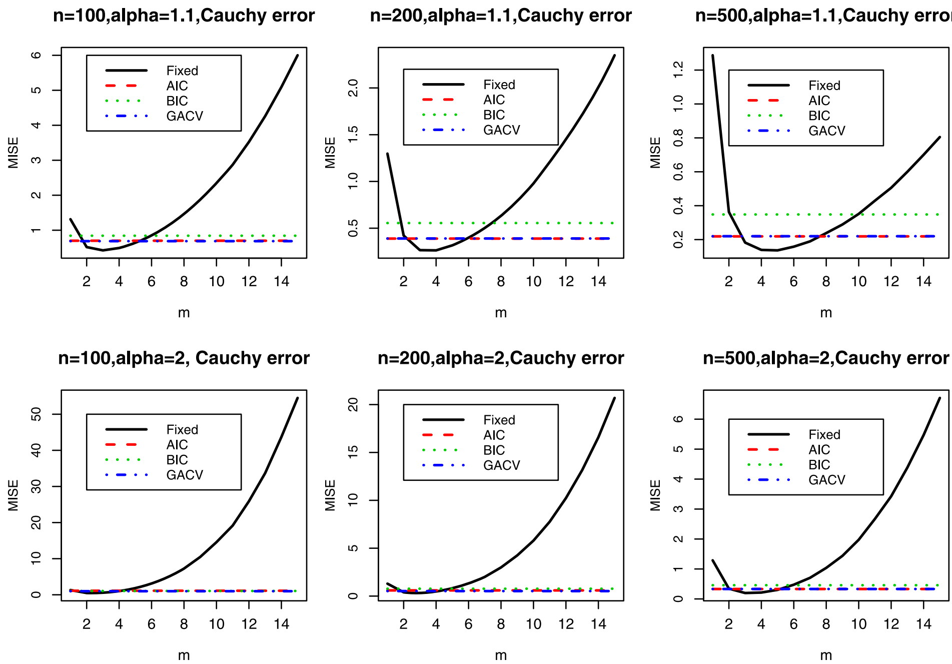

FIG. 3. *Performance of selection criteria. Case* (a). *Estimation of* $b(\cdot, 0.5)$.

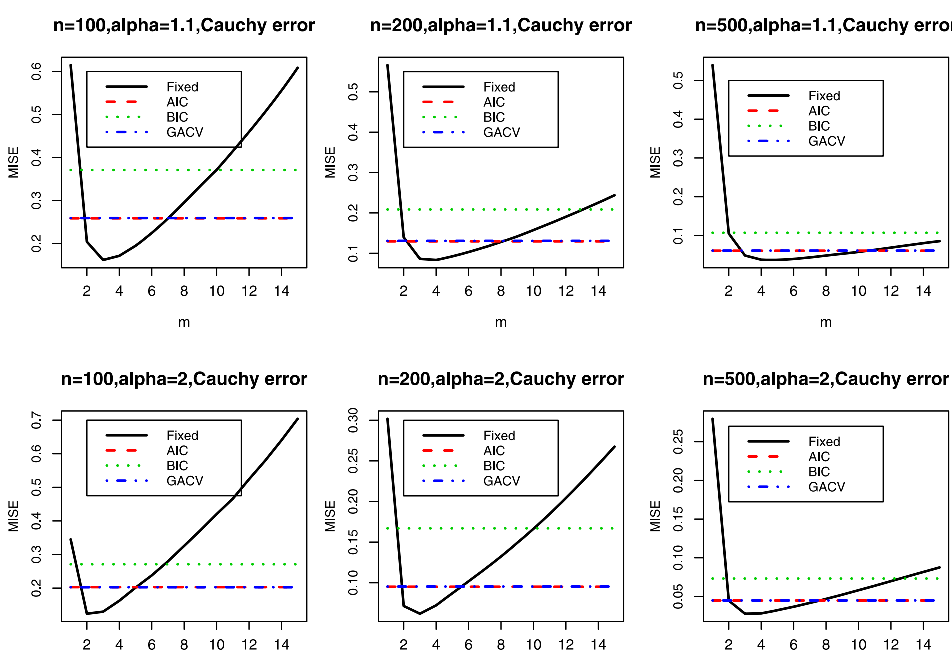

FIG. 4. *Performance of selection criteria. Case* (a). *Estimation of* $Q_{Y|X}(0.5|x)$.

TABLE 1
Average numbers of m selected by three criteria in case (a). *Standard deviations are given in parentheses*

n	α	error dist.	AIC	BIC	GACV
100	1.1	normal	7.40 (4.12)	3.09 (1.24)	6.58 (3.78)
200	1.1	normal	7.34 (3.87)	3.39 (1.14)	6.98 (3.67)
500	1.1	normal	8.13 (3.59)	3.94 (1.05)	7.97 (3.52)
100	2	normal	6.50 (4.23)	2.53 (0.93)	5.67 (3.84)
200	2	normal	6.33 (3.95)	2.67 (0.85)	5.96 (3.76)
500	2	normal	6.92 (3.85)	3.12 (0.77)	6.81 (3.80)
100	1.1	Cauchy	2.30 (1.18)	1.62 (0.61)	2.24 (1.08)
200	1.1	Cauchy	2.51 (0.88)	1.91 (0.64)	2.48 (0.84)
500	1.1	Cauchy	3.00 (0.87)	2.20 (0.51)	2.99 (0.86)
100	2	Cauchy	1.90 (0.94)	1.35 (0.51)	1.86 (0.81)
200	2	Cauchy	2.14 (0.76)	1.60 (0.53)	2.10 (0.60)
500	2	Cauchy	2.41 (0.67)	1.91 (0.38)	2.40 (0.67)

worked better than BIC. Looking closely at these figures, one finds that AIC and GACV performed quite badly in some cases (see the bottom half in Figure 1). Although none of these criteria clearly dominated the others, BIC worked relatively stably. These figures also show that as α increases from 1.1 to 2, the performance of $\hat{b}(\cdot, 0.5)$ becomes worse, while that of $\hat{Q}_{Y|X}(0.5|x)$ becomes better. This is consistent with the theoretical results in the previous section. Essentially similar comments apply to case (b), Figures 1–4 and Table 1 in Appendix F in the supplementary file [Kato (2013)].

5. Proof of Theorem 3.1. We divide the proof into three subsections. Some technical results are proved in Appendices D and E in the supplementary file [Kato (2013)]. To avoid the notational complication, uniformity in $F \in \mathcal{F}$ will be suppressed. Let $C > 0$ denote a generic constant of which the value may change from line to line. In most cases, qualification "almost surely" will be suppressed. In some parts of the proofs, we use empirical process techniques. We follow the basic notation used in van der Vaart and Wellner (1996).

5.1. *Reduction of the problem.* Let $b_0(u) = a(u), \hat{b}_0(u) = \hat{a}(u)$ and $\xi_{i0} = \hat{\xi}_{i0} = 1$. For any $v_0, \dots, v_m \in \mathbb{R}$, write $v^m = (v_0, \dots, v_m)'$. For $v^m \in \mathbb{R}^{m+1}$ and $w^m \in \mathbb{R}^{m+1}$, write $v^m \cdot w^m = \sum_{j=0}^m v_j w_j$. Then,

$$\hat{b}^m(u) = (\hat{b}_0(u), \hat{b}_1(u), \dots, \hat{b}_m(u))' = \operatorname*{arg\,min}_{b^m \in \mathbb{R}^{m+1}} \mathbb{E}_n[\rho_u(Y_i - \hat{\xi}_i^m \cdot b^m)].$$

We use a further re-parameterization. For $i = 1, \dots, n$ and $j \geq 1$, let $\eta_{ij} = \kappa_j^{-1/2}\xi_{ij}, \hat{\eta}_{ij} = \kappa_j^{-1/2}\hat{\xi}_{ij}, d_j(u) = \kappa_j^{1/2} b_j(u)$ and $\hat{d}_j(u) = \kappa_j^{1/2}\hat{b}_j(u)$. For $j = 0$, define $\eta_{i0} = \hat{\eta}_{i0} = 1$, $d_0(u) = b_0(u)$ and $\hat{d}_0(u) = \hat{b}_0(u)$. Note that $\mathbb{E}[\eta_{ij}] = 0, \mathbb{E}[\eta_{ij}^2] = 1$ for all $j \geq 1$, and $\mathbb{E}[\eta_{ij}\eta_{ik}] = 0$ for all $j, k \geq 0$ with $j \neq k$. Then,

$$\hat{d}^m(u) = (\hat{d}_0(u), \hat{d}_1(u), \dots, \hat{d}_m(u))' = \arg\min_{d^m \in \mathbb{R}^{m+1}} \mathbb{E}_n[\rho_u(Y_i - \hat{\eta}_i^m \cdot d^m)].$$

We first consider bounding $\sup_{u \in \mathcal{U}} \|\hat{d}^m(u) - d^m(u)\|_{\ell^2}$.

LEMMA 5.1. *Suppose that for all $\epsilon > 0$, there exist constants $c > 0$ and $M > 0$ possibly depending on ϵ such that*

$$\liminf_{n\to\infty} \mathbb{P}\{-\mathbb{E}_n[\{u - 1(Y_i \leq \hat{\eta}_i^m \cdot (d^m(u) + M\sqrt{m/n}h^m))\}(h^m \cdot \hat{\eta}_i^m)] > c\sqrt{m/n}\ \forall u \in \mathcal{U}, \forall h^m \in \mathbb{S}^m\} > 1 - \epsilon.$$

Then, we have

$$\limsup_{n\to\infty} \mathbb{P}\Big\{\sup_{u\in\mathcal{U}} \|\hat{d}^m(u) - d^m(u)\|_{\ell^2} > M\sqrt{m/n}\Big\} \leq \epsilon.$$

Hence, as soon as the hypothesis of Lemma 5.1 is verified, we will have $\sup_{u\in\mathcal{U}} \|\hat{d}^m(u) - d^m(u)\|_{\ell^2} = O_P(\sqrt{m/n})$, from which the desired conclusion follows in a relatively straightforward manner. We will verify the hypothesis of Lemma 5.1 in Section 5.2 and complete the proof of Theorem 3.1 in Section 5.3.

PROOF OF LEMMA 5.1. The proof is divided into three steps.

Step 1: $\|\mathbb{E}_n[\{u - 1(Y_i \leq \hat{\eta}_i^m \cdot \hat{d}^m(u))\}\hat{\eta}_i^m]\|_{\ell^2} \leq ((m+1)/n) \max_{1\leq i\leq n} \|\hat{\eta}_i^m\|_{\ell^2}$. The proof is based on the next lemma.

LEMMA 5.2. *Let $\{(y_i, z_i')'\}_{i=1}^n$ be a sequence of pairs of nonstochastic variables $(y_i, z_i')'$ where $y_i \in \mathbb{R}$ and $z_i \in \mathbb{R}^k$. Pick any $u \in (0, 1)$. Let $\check{d}(u) \in \mathbb{R}^k$ be any solution to the minimization problem*

$$\min_{d\in\mathbb{R}^k} \mathbb{E}_n[\rho_u(y_i - z_i'd)].$$

Then, we have

$$\|\mathbb{E}_n[\{u - 1(y_i \leq z_i'\check{d}(u))\}z_i]\|_{\ell^2} \leq n^{-1} \operatorname{Card}(\{i \in \{1, \dots, n\} : y_i = z_i'\check{d}(u)\}) \max_{1\leq i\leq n} \|z_i\|_{\ell^2}.$$

PROOF. The proof follows from a small modification of El-Attar, Vidyasagar and Dutta [(1979), Lemma 2.1]. □

Recall that $\hat{\eta}_i^m$ depends only on $X_1^n := \{X_1, \dots, X_n\}$ and not on $Y_1, \dots, Y_n$. Since the conditional distribution of $Y_1, \dots, Y_n$ given X_1^n is absolutely continuous, by Sard's theorem [see Milnor (1965), Section 3], $\sup_{u\in\mathcal{U}} \operatorname{Card}(\{i \in$

$\{1,\dots,n\}: Y_i = \hat{\eta}_i^m \cdot \hat{d}^m(u)\}) \le m+1$ almost surely. To be more precise, pick any subset $I \subset \{1,\dots,n\}$ such that $\mathrm{Card}(I) \ge m+2$. Conditional on X_1^n, consider the set

$$S_I = \{(\hat{\eta}_i^m \cdot \delta^m)_{i\in I} : \delta^m \in \mathbb{R}^{m+1}\} \subset \mathbb{R}^{\mathrm{Card}(I)}.$$

Then, S_I is a linear subspace of dimension at most $m+1$. Suppose that $Y_i = \hat{\eta}_i^m \cdot \hat{d}^m(u)$ for all $i \in I$ for some $u \in \mathcal{U}$. Then, $(Y_i)_{i\in I} \in S_I$, by which we have

$$\mathbb{P}\{Y_i = \hat{\eta}_i^m \cdot \hat{d}^m(u),\ \forall i \in I, \exists u \in \mathcal{U} | X_1^n\} \le \mathbb{P}\{(Y_i)_{i\in I} \in S_I | X_1^n\}. \tag{5.1}$$

However, by Sard's theorem, the Lebesgue measure of S_I in $\mathbb{R}^{\mathrm{Card}(I)}$ is zero, and by the absolute continuity of the conditional distribution of $(Y_i)_{i\in I}$ given X_1^n, the right side of (5.1) is zero. Thus, we conclude that

$$\begin{aligned}\mathbb{P}\Big\{&\sup_{u\in\mathcal{U}} \mathrm{Card}(\{i \in \{1,\dots,n\} : Y_i = \hat{\eta}_i^m \cdot \hat{d}^m(u)\}) \ge m+2 | X_1^n\Big\}\\ &\le \sum_{\substack{I\subset\{1,\dots,n\}\\ \mathrm{Card}(I)\ge m+2}} \mathbb{P}\{Y_i = \hat{\eta}_i^m \cdot \hat{d}^m(u)\ \forall i \in I, \exists u \in \mathcal{U} | X_1^n\} = 0,\end{aligned}$$

by which we have $\sup_{u\in\mathcal{U}} \mathrm{Card}(\{i \in \{1,\dots,n\} : Y_i = \hat{\eta}_i^m \cdot \hat{d}^m(u)\}) \le m+1$ almost surely. Then, the conclusion of Step 1 follows from an application of Lemma 5.2.

Step 2: We have

$$\max_{1\le i\le n} \|\hat{\eta}_i^m\|_{\ell^2} = o_P\{(\log n)^{-1}\sqrt{n/m}\}. \tag{5.2}$$

We defer the proof of (5.2) to Appendix D.

Step 3: Proof of the lemma.

Define

$$\hat{h}^m(u) = \begin{cases} \dfrac{\hat{d}^m(u) - d^m(u)}{\|\hat{d}^m(u) - d^m(u)\|_{\ell^2}}, & \text{if } \hat{d}^m(u) \ne d^m(u),\\ 0, & \text{otherwise.}\end{cases}$$

Then, by Steps 1 and 2, we have

$$\sup_{u\in\mathcal{U}} \big|\mathbb{E}_n[\{u - 1(Y_i \le \hat{\eta}_i^m \cdot \hat{d}^m(u))\}(\hat{h}^m(u) \cdot \hat{\eta}_i^m)]\big| = o_P(\sqrt{m/n}).$$

Define the event

$$\begin{aligned}\mathcal{E}_n := \{-\mathbb{E}_n[\{u - 1(Y_i \le \hat{\eta}_i^m \cdot (d^m(u) + M\sqrt{m/n}h^m))\}(h^m \cdot \hat{\eta}_i^m)]\\ > c\sqrt{m/n}\ \forall u \in \mathcal{U}, \forall h^m \in \mathbb{S}^m\}.\end{aligned}$$

Since the map

$$l \mapsto -\mathbb{E}_n[\{u - 1(Y_i \le \hat{\eta}_i^m \cdot (d^m(u) + l\sqrt{m/n}h^m))\}(h^m \cdot \hat{\eta}_i^m)]$$

is nondecreasing for all $u \in \mathcal{U}$ and $h^m \in \mathbb{S}^m$, the event $\mathcal{E}_n$ is also written as

$$\mathcal{E}_n = \bigl\{-\mathbb{E}_n\bigl[\bigl\{u - 1\bigl(Y_i \le \hat{\eta}_i^m \cdot \bigl(d^m(u) + l\sqrt{m/n}h^m\bigr)\bigr)\bigr\}\bigl(h^m \cdot \hat{\eta}_i^m\bigr)\bigr] > c\sqrt{m/n}\ \forall u \in \mathcal{U}, \forall l \ge M, \forall h^m \in \mathbb{S}^m\bigr\}.$$

Thus, as $n \to \infty$,

$$\begin{aligned}
&\mathbb{P}\Bigl\{\sup_{u\in\mathcal{U}}\|\hat{d}^m(u) - d^m(u)\|_{\ell^2} > M\sqrt{m/n}\Bigr\} \\
&\quad = \mathbb{P}\bigl\{\|\hat{d}^m(u) - d^m(u)\|_{\ell^2} > M\sqrt{m/n}, \exists u \in \mathcal{U}\bigr\} \\
&\quad \le \mathbb{P}\bigl[\bigl\{\|\hat{d}^m(u) - d^m(u)\|_{\ell^2} > M\sqrt{m/n}, \exists u \in \mathcal{U}\bigr\} \cap \mathcal{E}_n\bigr] + \mathbb{P}(\mathcal{E}_n^c) \\
&\quad \le \mathbb{P}\bigl\{-\mathbb{E}_n\bigl[\bigl\{u - 1\bigl(Y_i \le \hat{\eta}_i^m \cdot \hat{d}^m(u)\bigr)\bigr\}\bigl(\hat{h}^m(u) \cdot \hat{\eta}_i^m\bigr)\bigr] > c\sqrt{m/n}, \exists u \in \mathcal{U}\bigr\} \\
&\qquad + \mathbb{P}(\mathcal{E}_n^c) \\
&\quad \le o(1) + \bigl(1 + o(1)\bigr)\epsilon.
\end{aligned}$$

This completes the proof of Lemma 5.1. □

5.2. *Verification of the hypothesis of Lemma* 5.1. Pick any $h^m = (h_0, h_1, \dots, h_m)' \in \mathbb{S}^m$. For a given $M > 1$, let $\delta^m = (\delta_0, \delta_1, \dots, \delta_m)' = M\sqrt{m/n}h^m$. Then,

$$\begin{aligned}
&-\mathbb{E}_n\bigl[\bigl\{u - 1\bigl(Y_i \le \hat{\eta}_i^m \cdot \bigl(d^m(u) + \delta^m\bigr)\bigr)\bigr\}\bigl(h^m \cdot \hat{\eta}_i^m\bigr)\bigr] \\
&\quad = -\mathbb{E}_n\bigl[\bigl\{u - 1\bigl(Y_i \le Q_{Y|X}(u|X_i)\bigr)\bigr\}\bigl(h^m \cdot \hat{\eta}_i^m\bigr)\bigr] \\
&\qquad + \mathbb{E}_n\bigl[\bigl\{F_{Y|X}\bigl(\hat{\eta}_i^m \cdot \bigl(d^m(u) + \delta^m\bigr)|X_i\bigr) - F_{Y|X}\bigl(Q_{Y|X}(u|X_i)|X_i\bigr)\bigr\}\bigl(h^m \cdot \hat{\eta}_i^m\bigr)\bigr] \\
&\qquad + n^{-1/2}\mathbb{G}_{n|X}\bigl[\bigl\{1\bigl(Y_i \le \hat{\eta}_i^m \cdot \bigl(d^m(u) + \delta^m\bigr)\bigr) - 1\bigl(Y_i \le Q_{Y|X}(u|X_i)\bigr)\bigr\} \\
&\qquad\qquad \times \bigl(h^m \cdot \hat{\eta}_i^m\bigr)\bigr] \\
&\quad =: I + II + III,
\end{aligned}$$

where we have used the fact that $F_{Y|X}(Q_{Y|X}(u|X)|X) = u$ and the last term (*III*) is defined by

$$\begin{aligned}
&n^{-1/2}\mathbb{G}_{n|X}\bigl[\bigl\{1\bigl(Y_i \le \hat{\eta}_i^m \cdot \bigl(d^m(u) + \delta^m\bigr)\bigr) - 1\bigl(Y_i \le Q_{Y|X}(u|X_i)\bigr)\bigr\}\bigl(h^m \cdot \hat{\eta}_i^m\bigr)\bigr] \\
&\quad := \mathbb{E}_n\bigl[\bigl\{1\bigl(Y_i \le \hat{\eta}_i^m \cdot \bigl(d^m(u) + \delta^m\bigr)\bigr) - 1\bigl(Y_i \le Q_{Y|X}(u|X_i)\bigr) \\
&\qquad - F_{Y|X}\bigl(\hat{\eta}_i^m \cdot \bigl(d^m(u) + \delta^m\bigr)|X_i\bigr) + F_{Y|X}\bigl(Q_{Y|X}(u|X_i)|X_i\bigr)\bigr\}\bigl(h^m \cdot \hat{\eta}_i^m\bigr)\bigr].
\end{aligned}$$

We separately bound the terms I, II and III uniformly in $u \in \mathcal{U}$ and $h^m \in \mathbb{S}^m$. In what follows, stochastic orders are interpreted independent of M. Note that

$$\begin{aligned}
&\hat{\eta}_i^m \cdot \bigl(d^m(u) + \delta^m\bigr) - Q_{Y|X}(u|X_i) \\
&\quad = \sum_{j=0}^{m}\bigl(d_j(u) + \delta_j\bigr)\hat{\eta}_{ij} - \sum_{j=0}^{\infty} d_j(u)\eta_{ij}
\end{aligned} \tag{5.3}$$

$$= \hat{\eta}_i^m \cdot \delta^m + (\hat{\eta}_i^m - \eta_i^m) \cdot d^m(u) - \sum_{j=m+1}^{\infty} d_j(u)\eta_{ij}$$

$$=: \hat{\eta}_i^m \cdot \delta^m + \hat{r}_i(u).$$

Bounding I: observe that

$$I \geq -\|\mathbb{E}_n[\{u - 1(Y_i \leq Q_{Y|X}(u|X_i))\}\hat{\eta}_i^m]\|_{\ell^2}.$$

Using the relation

$$\begin{aligned} 1(Y_i \leq Q_{Y|X}(u|X_i)) &= 1(F_{Y|X}(Y_i|X_i) \leq u) \\ &= 1(U_i \leq u) \qquad \text{with } U_i = F_{Y|X}(Y_i|X_i), \end{aligned}$$

we have

$$\begin{aligned} &\sup_{u \in \mathcal{U}} \|\mathbb{E}_n[\{u - 1(Y_i \leq Q_{Y|X}(u|X_i))\}\hat{\eta}_i^m]\|_{\ell^2} \\ &\qquad \leq \sqrt{\sum_{j=0}^{m} \sup_{u \in \mathcal{U}} (\mathbb{E}_n[\{u - 1(U_i \leq u)\}\hat{\eta}_{ij}])^2}. \end{aligned}$$

Here $U_1, \dots, U_n$ are independent uniform random variables on (0, 1) independent of $X_1^n := \{X_1, \dots, X_n\}$. Pick and fix any $0 \leq j \leq m$. Let $\sigma_1, \dots, \sigma_n$ be independent Rademacher random variables independent of $(U_1, X_1), \dots, (U_n, X_n)$. Since $U_1, \dots, U_n$ are independent from X_1^n, applying the symmetrization inequality [see Lemma 2.3.6 of van der Vaart and Wellner (1996)] conditional on X_1^n, we have

$$\mathbb{E}\Big[\sup_{u \in \mathcal{U}} (\mathbb{E}_n[\{u - 1(U_i \leq u)\}\hat{\eta}_{ij}])^2 | X_1^n\Big] \leq 4\mathbb{E}\Big[\sup_{u \in \mathcal{U}} (\mathbb{E}_n[\sigma_i 1(U_i \leq u)\hat{\eta}_{ij}])^2 | X_1^n\Big].$$

We make use of Proposition E.2 in Appendix E to bound the right side. Consider the class of functions

$$\mathcal{G} = \{\mathbb{R} \times \mathbb{R} \ni (y, z) \mapsto 1(y \leq u)z : u \in \mathcal{U}\}.$$

Then, we have

$$\mathbb{E}\Big[\sup_{u \in \mathcal{U}} (\mathbb{E}_n[\sigma_i 1(U_i \leq u)\hat{\eta}_{ij}])^2 | X_1^n\Big] = \mathbb{E}\Big[\sup_{g \in \mathcal{G}} (\mathbb{E}_n[\sigma_i g(U_i, \hat{\eta}_{ij})])^2 | X_1^n\Big].$$

It is standard to see that $\mathcal{G}$ is a VC subgraph class with VC index ≤ 3. Thus, by Theorem 2.6.7 of van der Vaart and Wellner (1996), there exist universal constants $A \geq e$ and $W \geq 1$ such that, for envelope function $G(y, z) = |z|$,

$$N(\epsilon\|G\|_{L_2(P_n)}, \mathcal{G}, L_2(P_n)) \leq (A/\epsilon)^W \qquad 0 < \forall\epsilon \leq 1,$$

where P_n denotes the empirical distribution on $\mathbb{R} \times \mathbb{R}$ that assigns probability n^{-1} to each $(U_i, \hat{\eta}_{ij}), i = 1, \dots, n$. Therefore, by Proposition E.2, we conclude that

$$\mathbb{E}\Big[\sup_{u \in \mathcal{U}} (\mathbb{E}_n[\sigma_i 1(U_i \leq u)\hat{\eta}_{ij}])^2 | X_1^n\Big] \leq n^{-1} D\mathbb{E}_n[\hat{\eta}_{ij}^2],$$

where D is a universal constant. Since $0 \le j \le m$ is arbitrary, we have

$$\sup_{u \in \mathcal{U}} \|\mathbb{E}_n[\{u - 1(Y_i \le Q_{Y|X}(u|X_i))\}\hat{\eta}_i^m]\|_{\ell^2} = O_P\{(\mathbb{E}_n[\|\hat{\eta}_i^m\|_{\ell^2}^2])^{1/2} n^{-1/2}\}.$$

We shall show in Appendix D that

$$\mathbb{E}_n[\|\hat{\eta}_i^m\|_{\ell^2}^2] = O_P(m), \tag{5.4}$$

by which we have

$$\sup_{u \in \mathcal{U}} \|\mathbb{E}_n[\{u - 1(Y_i \le Q_{Y|X}(u|X_i))\}\hat{\eta}_i^m]\|_{\ell^2} = O_P(\sqrt{m/n}).$$

Bounding II: by Taylor's theorem, we have

$$\begin{aligned} &F_{Y|X}(Q_{Y|X}(u|X) + y|X) - F_{Y|X}(Q_{Y|X}(u|X)|X) \\ &\quad = f_{Y|X}(Q_{Y|X}(u|X)|X)y + y^2 \int_0^1 f'_{Y|X}(Q_{Y|X}(u|X) + \theta y|X)(1-\theta)\, d\theta \\ &\quad =: f_{Y|X}(Q_{Y|X}(u|X)|X)y + \frac{y^2}{2} R(u, y, X), \end{aligned}$$

by which we have, using (5.3),

$$\begin{aligned} II &= \mathbb{E}_n[f_{Y|X}(Q_{Y|X}(u|X_i)|X_i)(\hat{\eta}_i^m \cdot \delta^m + \hat{r}_i(u))(h^m \cdot \hat{\eta}_i^m)] \\ &\quad + \frac{1}{2}\mathbb{E}_n[(\hat{\eta}_i^m \cdot \delta^m + \hat{r}_i)^2 R(u, \hat{\eta}_i^m \cdot \delta^m + \hat{r}_i(u), X_i)(h^m \cdot \hat{\eta}_i^m)] \\ &\ge M\sqrt{m/n}\mathbb{E}_n[f_{Y|X}(Q_{Y|X}(u|X_i)|X_i)(h^m \cdot \hat{\eta}_i^m)^2] \\ &\quad - C\mathbb{E}_n[|\hat{r}_i(u)(h^m \cdot \hat{\eta}_i^m)|] - C\mathbb{E}_n[(\hat{\eta}_i^m \cdot \delta^m + \hat{r}_i(u))^2 |h^m \cdot \hat{\eta}_i^m|] \\ &\ge M\sqrt{m/n}\mathbb{E}_n[f_{Y|X}(Q_{Y|X}(u|X_i)|X_i)(h^m \cdot \hat{\eta}_i^m)^2] \\ &\quad - C(\mathbb{E}_n[\hat{r}_i^2(u)])^{1/2}(\mathbb{E}_n[(h^m \cdot \hat{\eta}_i^m)^2])^{1/2} \\ &\quad - CM^2(m/n)\Big(\max_{1 \le i \le n}\|\hat{\eta}_i^m\|_{\ell^2}\Big)\mathbb{E}_n[(h^m \cdot \hat{\eta}_i^m)^2] \\ &\quad - C\Big(\max_{1 \le i \le n}\|\hat{\eta}_i^m\|_{\ell^2}\Big)\mathbb{E}_n[\hat{r}_i^2(u)], \end{aligned} \tag{5.5}$$

where we have used the fact that $f_{Y|X}(y|X) \vee |f'_{Y|X}(y|X)| \le C$. By assumption (A5), there exists a small constant $c_1 > 0$ such that

$$C\mathbb{E}_n[f_{Y|X}(Q_{Y|X}(u|X_i))(h^m \cdot \hat{\eta}_i^m)^2] \ge c_1\mathbb{E}_n[(h^m \cdot \hat{\eta}_i^m)^2] \qquad \forall u \in \mathcal{U}, \forall h^m \in \mathbb{S}^m.$$

We shall show in Appendix D that

$$\sup_{h^m \in \mathbb{S}^m} |\mathbb{E}_n[(h^m \cdot \hat{\eta}_i^m)^2] - 1| = o_P(1), \tag{5.6}$$

$$\sup_{u \in \mathcal{U}} \mathbb{E}_n[\hat{r}_i^2(u)] = O_P(mn^{-1}). \tag{5.7}$$

Thus, by (5.2), (5.4), (5.6) and (5.7), we have

$$(5.5) \geq c_1 M\sqrt{m/n}\bigl(1 - o_P(1)\bigr) - O_P(\sqrt{m/n}) - M^2 o_P(\sqrt{m/n}),$$

where the stochastic orders are evaluated uniformly in $u \in \mathcal{U}$ and $h^m \in \mathbb{S}^m$.

Bounding *III*: Let $\sigma_1, \dots, \sigma_n$ be independent Rademacher random variables independent of the data $(Y_1, X_1), \dots, (Y_n, X_n)$. Applying the symmetrization inequality conditional on $X_1^n := \{X_1, \dots, X_n\}$, we have

$$\begin{aligned}
&\mathbb{E}\Bigl[\sup_{u\in\mathcal{U}, h^m\in\mathbb{S}^m} \bigl|n^{-1/2}\mathbb{G}_{n|X}\bigl[\bigl\{1\bigl(Y_i \leq \hat{\eta}_i^m \cdot \bigl(d^m(u) + \delta^m\bigr)\bigr) \\
&\qquad\qquad - 1\bigl(Y_i \leq Q_{Y|X}(u|X_i)\bigr)\bigr\}\bigl(h^m \cdot \hat{\eta}_i^m\bigr)\bigr]\bigr| \big| X_1^n\Bigr] \\
&\quad \leq 2\mathbb{E}\Bigl[\sup_{u\in\mathcal{U}, h^m\in\mathbb{S}^m} \bigl|\mathbb{E}_n\bigl[\sigma_i\bigl\{1\bigl(Y_i \leq \hat{\eta}_i^m \cdot \bigl(d^m(u) + \delta^m\bigr)\bigr) \\
&\qquad\qquad - 1\bigl(Y_i \leq Q_{Y|X}(u|X_i)\bigr)\bigr\}\bigl(h^m \cdot \hat{\eta}_i^m\bigr)\bigr]\bigr| \big| X_1^n\Bigr],
\end{aligned} \tag{5.8}$$

where δ^m is taken as $\delta^m = M\sqrt{m/n}h^m$ in the suprema. Note that the symmetrization inequality is applicable since the regular conditional distribution of $(Y_1, \dots, Y_n)'$ given X_1^n exists and conditional on X_1^n, $Y_1, \dots, Y_n$ are independent. Consider the class of functions

$$\begin{aligned}
\mathcal{G} = \bigl\{\mathbb{R} \times D[0,1] \times \mathbb{R}^{m+1} \ni \bigl(y, x, \eta^m\bigr) \mapsto \bigl\{&1\bigl(y \leq \eta^m \cdot \bigl(d^m(u) + \delta^m\bigr)\bigr) \\
&- 1\bigl(y \leq Q_{Y|X}(u|x)\bigr)\bigr\}\bigl(h^m \cdot \eta^m\bigr) : \\
&u \in \mathcal{U}, h^m \in \mathbb{S}^m, \delta^m = M\sqrt{m/n}h^m\bigr\}.
\end{aligned}$$

Then, we have

$$(5.8) = 2\mathbb{E}\Bigl[\sup_{g\in\mathcal{G}} \bigl|\mathbb{E}_n\bigl[\sigma_i g\bigl(Y_i, X_i, \hat{\eta}_i^m\bigr)\bigr]\bigr| \big| X_1^n\Bigr].$$

We apply Proposition E.1 in Appendix E to bound the right side. Note that $(X_i, \hat{\eta}_i^m)$ are measurable with respect to the σ-field generated by X_1^n, the regular conditional distribution of $(Y_1, \dots, Y_n)'$ given X_1^n exists, and conditional on X_1^n, $Y_1, \dots, Y_n$ are independent. Observe that

$$\sup_{g\in\mathcal{G}} \bigl|g\bigl(Y_i, X_i, \hat{\eta}_i^m\bigr)\bigr| \leq \max_{1\leq i\leq n} \|\hat{\eta}_i^m\|_{\ell^2} =: \hat{B},$$

and, by (5.3),

$$\begin{aligned}
&\sup_{g\in\mathcal{G}} \mathbb{E}_n\bigl[\mathbb{E}\bigl[g^2\bigl(Y_i, X_i, \hat{\eta}_i^m\bigr) | X_1^n\bigr]\bigr] \\
&\quad = \sup_{u\in\mathcal{U}, h^m\in\mathbb{S}^m} \mathbb{E}_n\bigl[\bigl|F_{Y|X}\bigl(\hat{\eta}_i^m \cdot \bigl(d^m(u) + \delta^m\bigr) | X_i\bigr) \\
&\qquad\qquad - F_{Y|X}\bigl(Q_{Y|X}(u|X_i) | X_i\bigr)\bigr|\bigl(h^m \cdot \hat{\eta}_i^m\bigr)^2\bigr]
\end{aligned}$$

$$
\begin{aligned}
&\leq \sup_{u\in\mathcal{U},h^m\in\mathbb{S}^m}\{CM\sqrt{m/n}\mathbb{E}_n[|h^m\cdot\hat{\eta}_i^m|^3]+C\mathbb{E}_n[|\hat{r}_i(u)|(h^m\cdot\hat{\eta}_i^m)^2]\} \\
&\leq CM\sqrt{m/n}\max_{1\leq i\leq n}\|\hat{\eta}_i^m\|_{\ell^2}\sup_{h^m\in\mathbb{S}^m}\mathbb{E}_n[(h^m\cdot\hat{\eta}_i^m)^2] \\
&\quad + C\max_{1\leq i\leq n}\|\hat{\eta}_i^m\|_{\ell^2}\Big(\sup_{u\in\mathcal{U}}\mathbb{E}_n[\hat{r}_i^2(u)]\Big)^{1/2}\Big(\sup_{h^m\in\mathbb{S}^m}\mathbb{E}_n[(h^m\cdot\hat{\eta}_i^m)^2]\Big)^{1/2} \\
&=: \hat{\tau}^2.
\end{aligned}
$$

We shall show in Appendix D that there exist some constants $c_2 \geq 1$ and $A' \geq 3\sqrt{e}$ such that

$$
N(\hat{B}\epsilon, \mathcal{G}, L_2(P_n')) \leq (A'/\epsilon)^{c_2 m}, \qquad 0 < \forall\epsilon \leq 1, \tag{5.9}
$$

where P_n' denotes the empirical distribution on $\mathbb{R}\times D[0,1]\times\mathbb{R}^{m+1}$ that assigns probability n^{-1} to each $(Y_i, X_i, \hat{\eta}_i^m)$, $i = 1,\dots,n$. Therefore, by Proposition E.1 in Appendix E, we conclude that

$$
\begin{aligned}
&\mathbb{E}\Big[\sup_{g\in\mathcal{G}}|\mathbb{E}_n[\sigma_i g(Y_i, X_i, \hat{\eta}_i^m)]||X_1^n\Big] \\
&\quad \leq 1(\hat{\tau} > 0)D'\Big[\sqrt{\frac{c_2 m\hat{\tau}^2}{n}}\sqrt{\log\frac{A'\hat{B}}{\hat{\tau}}}+\frac{c_2 m\hat{B}}{n}\log\frac{A'\hat{B}}{\hat{\tau}}\Big],
\end{aligned} \tag{5.10}
$$

provided that $\hat{\tau} \leq \hat{B}$, where D' is a universal constant.

By (5.2), (5.4) and (5.6), and the fact that $M > 1$, we have

$$
\hat{B} = o_P\{(\log n)^{-1}\sqrt{n/m}\}, \qquad \hat{\tau}^2 = Mo_P\{(\log n)^{-1}\},
$$

and there exists a small constant $c_3 > 0$ such that with probability approaching one

$$
\hat{\tau}^2 \geq c_3\hat{B}\sqrt{m/n}.
$$

Thus, replacing $\hat{\tau}$ by $\hat{\tau}\wedge\hat{B}$ if necessary, (5.10) $= M^{1/2}o_P(\sqrt{m/n})$, by which we conclude that

$$
\mathit{III} \geq -M^{1/2}o_P(\sqrt{m/n}),
$$

where the stochastic order is evaluated uniformly in $u\in\mathcal{U}$ and $h^m\in\mathbb{S}^m$.

Taking these together, we now conclude that

$$
I + \mathit{II} + \mathit{III} \geq c_1 M\sqrt{m/n}(1 - o_P(1)) - O_P(\sqrt{m/n}) - M^2 o_P(\sqrt{m/n}),
$$

where the stochastic orders are evaluated uniformly in $u\in\mathcal{U}$ and $h^m\in\mathbb{S}^m$. This immediately implies the hypothesis of Lemma 5.1.

5.3. *Completion of the proof.* We have shown that

$$\sup_{u\in\mathcal{U}}\|\hat{d}^m(u)-d^m(u)\|_{\ell^2}=O_P(\sqrt{m/n}).$$

Since $\|\hat{d}^m(u)-d^m(u)\|_{\ell^2}^2=\sum_{j=0}^m\kappa_j(\hat{b}_j(u)-b_j(u))^2\geq\kappa_m\sum_{j=0}^m(\hat{b}_j(u)-b_j(u))^2=\kappa_m\|\hat{b}^m(u)-b^m(u)\|_{\ell^2}^2$ (with $\kappa_0:=1$), we have

$$\begin{aligned}\sup_{u\in\mathcal{U}}\|\hat{b}^m(u)-b^m(u)\|_{\ell^2}^2&=O_P(\kappa_m^{-1}mn^{-1})=O_P(m^{\alpha+1}n^{-1})\\&=O_P(n^{-(2\beta-1)/(\alpha+2\beta)}).\end{aligned}$$

Observe that

$$\begin{aligned}&\hat{b}(t,u)-b(t,u)\\&\quad=\sum_{j=1}^m\hat{b}_j(u)\hat{\phi}_j(t)-\sum_{j=1}^m b_j(u)\hat{\phi}_j(t)+\sum_{j=1}^m b_j(u)\hat{\phi}_j(t)\\&\qquad-\sum_{j=1}^m b_j(u)\phi_j(t)-\sum_{j=m+1}^\infty b_j(u)\phi_j(t)\\&\quad=\sum_{j=1}^m(\hat{b}_j-b_j)(u)\hat{\phi}_j(t)+\sum_{j=1}^m b_j(u)(\hat{\phi}_j(t)-\phi_j(t))-\sum_{j=m+1}^\infty b_j(u)\phi_j(t),\end{aligned}$$

so that, uniformly in $u\in\mathcal{U}$,

$$\begin{aligned}&\int_0^1(\hat{b}(t,u)-b(t,u))^2\,dt\\&\qquad\leq 3\|\hat{b}^m(u)-b^m(u)\|_{\ell^2}^2+3m\sum_{j=1}^m b_j^2(u)\|\hat{\phi}_j-\phi_j\|^2+3\sum_{j=m+1}^\infty b_j^2(u)\\&\qquad=O_P(n^{-(2\beta-1)/(\alpha+2\beta)})+3m\sum_{j=1}^m b_j^2(u)\|\hat{\phi}_j-\phi_j\|^2+O(m^{-2\beta+1})\\&\qquad=O_P(n^{-(2\beta-1)/(\alpha+2\beta)})+3m\sum_{j=1}^m b_j^2(u)\|\hat{\phi}_j-\phi_j\|^2.\end{aligned}$$

By Lemmas D.1 and D.2 together with the computation in the proof of (5.7) in Appendix D, we see that

$$\begin{aligned}m\sum_{j=1}^m b_j^2(u)\|\hat{\phi}_j-\phi_j\|^2&\leq Cm\sum_{j=1}^m j^{-2\beta}\|\hat{\phi}_j-\phi_j\|^2\\&=O_P(m(n^{-1}+\Delta^\gamma+n^{-1}(\log n)m^{3\alpha+3}\Delta^\gamma))=O(mn^{-1})\\&=o_P(n^{-(2\beta-1)/(\alpha+2\beta)}).\end{aligned}$$

This completes the proof of Theorem 3.1.

Acknowledgments. Most of this work was done while the author was visiting the Department of Economics, MIT. He would like to thank Professors Victor Chernozhukov, Hidehiko Ichimura and Roger Koenker for their suggestions and encouragements. He also would like to thank the Editor, Professor T. Tony Cai, the Associate Editor and three anonymous referees for constructive comments that helped to improve the quality of the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Estimation in functional linear quantile regression" (DOI: 10.1214/12-AOS1066SUPP; .pdf). This supplementary file contains the additional discussion on the connection to nonlinear ill-posed inverse problems, technical proofs omitted in the main body, some useful technical tools and additional simulation results.

REFERENCES

BARLOW, R., BARTHOLENEW, J., BREMER, J. and BRUNK, H. (1972). *Statistical Inference Under Order Restrictions*. Wiley, New York.

BELLONI, A., CHERNOZHUKOV, V. and FERNÁNDEZ-VAL, I. (2011). Conditional quantile processes based on series or many regressors. Available at arXiv:1105.6154.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York. MR0233396

BISSANTZ, N., HOHAGE, T. and MUNK, A. (2004). Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise. *Inverse Problems* **20** 1773–1789. MR2107236

CAI, T. T. and HALL, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179. MR2291496

CARDOT, H., CRAMBES, C. and SARDA, P. (2005). Quantile regression when the covariates are functions. *J. Nonparametr. Stat.* **17** 841–856. MR2180369

CARDOT, H., CRAMBES, C. and SARDA, P. (2007). Ozone pollution forecasting using conditional mean and conditional quantiles with functional covariates. In *Statistical Methods for Biostatistics and Related Fields* (W. Härdle, Y. Mori and P. Vieu, eds.) 221–243. Springer, Berlin. MR2376412

CARDOT, H., FERRATY, F. and SARDA, P. (1999). Functional linear model. *Statist. Probab. Lett.* **45** 11–22. MR1718346

CARDOT, H., FERRATY, F. and SARDA, P. (2003). Spline estimators for the functional linear model. *Statist. Sinica* **13** 571–591. MR1997162

CHEN, K. and MÜLLER, H.-G. (2012). Conditional quantile analysis when covariates are functions, with application to growth data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74** 67–89. MR2885840

CHEN, X. and POUZO, D. (2012). Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica* **80** 277–321. MR2920758

CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and GALICHON, A. (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika* **96** 559–575. MR2538757

CRAMBES, C., KNEIP, A. and SARDA, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37** 35–72. MR2488344

DELAIGLE, A. and HALL, P. (2012). Methodology and theory for partial least squares applied to functional data. *Ann. Statist.* **40** 322–352.

DOORNIK, J. A. (2002). *Object-Oriented Matrix Programming Using Ox*, 3rd ed. Timberlake Consultants Press, London.

DUDLEY, R. M. (2002). *Real Analysis and Probability*. *Cambridge Studies in Advanced Mathematics* **74**. Cambridge Univ. Press, Cambridge. MR1932358

EL-ATTAR, R. A., VIDYASAGAR, M. and DUTTA, S. R. K. (1979). An algorithm for l_1-norm minimization with application to nonlinear l_1-approximation. *SIAM J. Numer. Anal.* **16** 70–86. MR0518685

FERRATY, F., RABHI, A. and VIEU, P. (2005). Conditional quantiles for dependent functional data with application to the climatic El Niño phenomenon. *Sankhyā* **67** 378–398. MR2208895

FERREIRA, J. C. and MENEGATTO, V. A. (2009). Eigenvalues of integral operators defined by smooth positive definite kernels. *Integral Equations Operator Theory* **64** 61–81. MR2501172

GAGLIARDINI, P. and SCAILLET, O. (2012). Nonparametric instrumental variables estimation of structural quantile effects. *Econometrica* **80** 1533–1562.

HALL, P. and HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269

HE, X. and SHAO, Q.-M. (2000). On parameters of increasing dimensions. *J. Multivariate Anal.* **73** 120–135. MR1766124

HÖRMANN, S. and KOKOSZKA, P. (2010). Weakly dependent functional data. *Ann. Statist.* **38** 1845–1884. MR2662361

HOROWITZ, J. L. and LEE, S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. *Econometrica* **75** 1191–1208. MR2333498

JAMES, G. M., WANG, J. and ZHU, J. (2009). Functional linear regression that's interpretable. *Ann. Statist.* **37** 2083–2108. MR2543686

KATO, K. (2013). Supplement to "Estimation in functional linear quantile regression." DOI:10.1214/12-AOS1066SUPP.

KOENKER, R. (2005). *Quantile Regression*. *Econometric Society Monographs* **38**. Cambridge Univ. Press, Cambridge. MR2268657

KOENKER, R. and BASSETT, G. JR. (1978). Regression quantiles. *Econometrica* **46** 33–50. MR0474644

LOUBES, J.-M. and LUDEÑA, C. (2010). Penalized estimators for non linear inverse problems. *ESAIM Probab. Stat.* **14** 173–191. MR2741964

MILNOR, J. W. (1965). *Topology from the Differentiable Viewpoint*. Univ. Press of Virginia, Charlottesville, VA. MR0226651

MÜLLER, H.-G. and STADTMÜLLER, U. (2005). Generalized functional linear models. *Ann. Statist.* **33** 774–805. MR2163159

PORTNOY, S. and KOENKER, R. (1997). The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Statist. Sci.* **12** 279–300. MR1619189

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993

RITTER, K., WASILKOWSKI, G. W. and WOŹNIAKOWSKI, H. (1995). Multivariate integration and approximation for random fields satisfying Sacks–Ylvisaker conditions. *Ann. Appl. Probab.* **5** 518–540. MR1336881

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York. MR1385671

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. MR2253106

YUAN, M. (2006). GACV for quantile smoothing splines. *Comput. Statist. Data Anal.* **50** 813–829. MR2207010

YUAN, M. and CAI, T. T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Statist.* **38** 3412–3444. MR2766857

DEPARTMENT OF MATHEMATICS
GRADUATE SCHOOL OF SCIENCE
HIROSHIMA UNIVERSITY
1-3-1 KAGAMIYAMA, HIGASHI-HIROSHIMA
HIROSHIMA 739-8526
JAPAN
E-MAIL: kkato@hiroshima-u.ac.jp