Journal of Business & Economic Statistics

Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: https://amstat.tandfonline.com/loi/ubes20

Semiparametric Estimation of Additive Quantile Regression Models by Two-Fold Penalty

Heng Lian

To cite this article: Heng Lian (2012) Semiparametric Estimation of Additive Quantile Regression Models by Two-Fold Penalty, Journal of Business & Economic Statistics, 30:3, 337-350, DOI: 10.1080/07350015.2012.693851

To link to this article: https://doi.org/10.1080/07350015.2012.693851

Published online: 20 Jul 2012.

Semiparametric Estimation of Additive Quantile Regression Models by Two-Fold Penalty

Heng Lian
Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore (*henglian@ntu.edu.sg*)

In this article, we propose a model selection and semiparametric estimation method for additive models in the context of quantile regression problems. In particular, we are interested in finding nonzero components as well as linear components in the conditional quantile function. Our approach is based on spline approximation for the components aided by two Smoothly Clipped Absolute Deviation (SCAD) penalty terms. The advantage of our approach is that one can automatically choose between general additive models, partially linear additive models, and linear models in a single estimation step. The most important contribution is that this is achieved without the need for specifying which covariates enter the linear part, solving one serious practical issue for models with partially linear additive structure. Simulation studies as well as a real dataset are used to illustrate our method.

KEY WORDS: Oracle property; Partially linear additive models; SCAD penalty; Schwartz-type information criterion.

1. INTRODUCTION

Additive models have received considerable attention since their introduction by Hastie and Tibshirani (1990). They are more parsimonious than fully nonparametric models (Chaudhuri 1991) that are difficult to fit when the number of predictors is medium to large, and more flexible than linear models by not constraining the relationships to be linear. At a given quantile $\tau \in (0, 1)$, the additive model for quantile regression has the following form

$$Y_i = \mu_\tau + \sum_{j=1}^{p} m_{\tau,j}(X_{ij}) + \epsilon_{\tau,i}, \quad i = 1, \ldots, n, \tag{1}$$

where (Y_i, X_i) are independent and identically distributed with the same distribution as (Y, X), $X = (X_1, \ldots, X_p)^T$ is the p-dimensional predictor and $\epsilon_{\tau,i}$ is the random error that satisfies $P(\epsilon_{\tau,i} \leq 0|X_i) = \tau$. Note that no distribution for $\epsilon_{\tau,i}$ needs to be specified and heterogeneous errors are not excluded since $\epsilon_{\tau,i}$ can depend on the covariates. The value of quantile regression has been demonstrated by a rapidly expanding literature in econometrics, social sciences, and biomedical studies [see the original article by Koenker and Bassett Jr (1978) or refer to Koenker (2005) for a comprehensive introduction]. As a useful supplement to mean regression, it produces a more complete description of the conditional response distribution and is more robust to heavy-tailed random errors. Quantile regression for additive models has previously been treated in De Gooijer and Zerom (2003), Horowitz and Lee (2005), and quantile regression with varying coefficients represents a class of closely related models (Kim 2007; Wang, Zhu, and Zhou 2009).

The additive form in the conditional quantile above circumvents the curse of dimensionality problem in nonparametric quantile regression. However, some covariates do not have effects on the responses and we desire to find those covariates either for efficiency reasons or to make the model more easily interpretable. Such zero components in additive models can be found by performing tests or by optimizing a penalized likelihood. To further reduce the worry of overfitting, one can try to find those parametric (i.e., linear) components. Such components typically converge faster than a truly nonparametric component and the resulting model, a semiparametric additive model (or partially linear additive model), is more parsimonious than general additive models. Such advantages of a semiparametric additive model are emphasized in Opsomer and Ruppert (1999). However, it is generally difficult to determine which covariates should enter as nonparametric components and which should enter as linear components. The commonly adopted strategy in practice is just to consider continuous covariates entering as nonparametric components and discrete covariates entering as parametric. For example, in studying the impact of various possible determinants on the intention of East Germans to migrate to West Germany 1 year after the Germany unification (Härdle, Mammen, and Müller 1998), some covariates may affect the response variable linearly, for instance, discrete covariates, while other covariates enter nonlinearly. This generalization allows richer and more flexible model structures than linear models and general additive models. This is a reasonable approach but there is always the healthy skepticism for its efficiency if some continuous covariates have linear effects. It would be nice to have a statistical tool that can perform model selection (finding both zero and linear components) and estimation simultaneously.

Both zero components and linear components could be found by performing some hypothesis testing. However, this might be cumbersome to perform in practice whether there are more than just a few predictors to test. Given the success of the penalization approach for selecting a sparse model in Huang, Horowitz, and Wei (2010), it is highly desirable that the same approach could be applied to select parametric components as well as zero components. To the best of our knowledge, the present article is the first

© 2012 American Statistical Association
Journal of Business & Economic Statistics
July 2012, Vol. 30, No. 3
DOI: 10.1080/07350015.2012.693851

to consider a penalization approach for variable selection and parametric component selection in additive quantile regression. Huang, Horowitz, and Wei (2010) considered variable selection for mean regression in additive models when $p \gg n$. Due to technical difficulties in dealing with quantiles, we restrict our study to the fixed p case and leave the high-dimensional setting for the future. However, we note that several recent works (Belloni and Chernozhukov 2011; Kato 2011) have considered variable selection for linear quantile regression in high dimensions. We use a two-fold Smoothly Clipped Absolute Deviation (SCAD) penalty, originally introduced in Fan and Li (2001), one for finding zero components and the other for finding parametric components. The technical difficulty lies in dealing with the penalty for finding parametric components. In terms of computation, the difficulty lies in that the problem cannot be reduced to linear programming, which is the major approach to solving quantile regression problems in the literature. We note that when using only one penalty for finding zero components, our investigation is reduced to that of sparse additive models in quantile regression, which is of interest in itself since quantile regression and mean regression are sufficiently different.

In the next section, we will propose the two-fold SCAD penalization procedure, and present its theoretical properties. In particular, we show that the procedure can select the true model with probability approaching one, the usual one-dimensional nonparametric convergence rate is achieved on the component functions and furthermore the slope parameters for the linear parametric components actually converge faster at the root-n rate and are asymptotically normal. These results together show that our estimator has the oracle property. We also propose to use a Schwartz-type information criterion (SIC; also adapted to our context that includes two penalty terms) to choose the regularization parameters. In Section 3, some simulations are carried out to assess the performance of the proposed method, and we also apply the method to a real dataset as illustrations. The technical proofs for the main theoretical results are provided in the Appendix.

2. TWO-FOLD SCAD PENALIZATION

2.1 Spline-Based Estimation

For the ease of presentation, we will omit τ in the expressions wherever clear from the context. Without loss of generality, we assume the distribution of X_j, $1 \le j \le p$ is supported on [0, 1], and also impose the condition $Em_j(X_j) = 0$ which is required for identifiability.

At the start of the analysis, we do not know which component functions in (1) are linear or actually zero. We use polynomial splines to approximate the components. Let $t_0 = 0 < t_1 < \cdots < t_{K'} < t_{K'+1} = 1$ partition [0, 1] into subintervals $[t_k, t_{k+1})$, $k = 0, \ldots, K'$ with K' internal knots. We only restrict our attention to uniform (equally spaced) knots, although quasi-uniform or data-driven choices can be considered. A polynomial spline of order q is a function whose restriction to each subinterval is a polynomial of degree $q - 1$ and globally $q - 2$ times continuously differentiable on [0, 1]. The collection of splines with a fixed sequence of knots has a normalized B-spline basis $\{B_1(x), \ldots, B_{\tilde{K}}(x)\}$ with $\tilde{K} = K' + q$. Because of the centering constraint $Em_j(X_j) = 0$, we instead focus on the subspace of spline functions $S_j^0 := \{s : s = \sum_{k=1}^{K} b_{jk} B_{jk}(x), \sum_{i=1}^{n} s(X_{ij}) = 0\}$ with basis $\{B_{jk}(x) = B_k(x) - \sum_{i=1}^{n} B_k(X_{ij})/n, k = 1, \ldots, K = \tilde{K} - 1\}$ (the subspace is $K = \tilde{K} - 1$ dimensional due to the empirical version of the constraint). Using spline expansions, we can approximate the components by $m_j(x) \approx g_j(x) = \sum_{k=1}^{K} b_{jk} B_{jk}(x)$. Note that it is possible to specify a different K for each component but we assume that they are the same for simplicity.

Our main goal is to find both zero components (i.e., $m_j \equiv 0$) and linear components (i.e., m_j is a linear function). The former can be achieved by shrinking $\|g_j\|$ to zero. For the latter, we want to shrink the second derivative $\|g_j''\|$ to zero instead. This suggests the following minimization problem

$$(\hat{\mu}, \hat{b}) = \arg\min_{\mu, b} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \left(Y_i - \mu - \sum_j \sum_k b_{jk} B_{jk}(X_{ij}) \right) + \sum_{j=1}^{p} p_{\lambda_1}(\|g_j\|) + \sum_{j=1}^{p} p_{\lambda_2}(\|g_j''\|), \tag{2}$$

where $\rho_\tau(u) = u(\tau - I_{\{u \le 0\}})$ is the quantile loss function (also called the check function), p_{λ_1} and p_{λ_2} are two penalties used to find zero and linear coefficients, respectively, with two regularization parameters λ_1 and λ_2, and $g_j = b_j^T B_j$ with $b_j = (b_{j1}, \ldots, b_{jK})^T$, $B_j = (B_{j1}, \ldots, B_{jK})^T$. When the penalty function is chosen appropriately (Tibshirani 1996; Wang and Xia 2009), in the resulting estimates, some $\|g_j\|$ will be exactly zero and some $\|g_j''\|$ will be exactly zero. The former obviously corresponds to the zero components, while the latter will correspond to the linear components, since a function has a second derivative identically zero if and only if it is a linear function. The estimated component functions are $\hat{m}_j = \hat{b}_j^T B_j$. Note that since $\|g_j\|^2 = \|b_j^T B_j\|^2 = \int (\sum_k b_{jk} B_{jk}(x))(\sum_{k'} b_{jk'} B_{jk'}(x)) dx$ and $\|g_j''\|^2 = \int (\sum_k b_{jk} B_{jk}''(x))(\sum_{k'} b_{jk'} B_{jk'}''(x)) dx$, $\|g_j\|$ and $\|g_j''\|$ can be equivalently written as $\sqrt{b_j^T D_j b_j}$ and $\sqrt{b_j^T E_j b_j}$, respectively, with the (k, k') entry of D_j being $\int_0^1 B_k(x) B_{k'}(x) dx$ and the (k, k') entry of E_j being $\int_0^1 B_k''(x) B_{k'}''(x) dx$. Penalizing the second derivative is commonly used in smoothing spline estimation (Wahba 1990) as well as functional linear regression (Ramsay and Silverman 2005). However the purpose there is to encourage smoothness of the estimated nonparametric function and no model selection as we aim for here can be achieved. Accordingly, in the smoothing spline literature, the *square* of $\|g_j''\|$ is used as the penalty, which is quite different from using the SCAD penalty as done here.

For convenience, we define

$$Z_i = (B_{11}(X_{i1}), B_{12}(X_{i1}), \ldots, B_{1K}(X_{i1}), \ldots B_{pK}(X_{ip}))^T, \quad \mathbf{Z} = (Z_1, \ldots, Z_n)^T.$$

The minimization problem above can be written as

$$(\hat{\mu}, \hat{b}) = \arg\min_{\mu, b} Q(\mu, b) := \frac{1}{n} \sum_i \rho_\tau \left(Y_i - \mu - Z_i^T b \right) + \sum_{j=1}^{p} p_{\lambda_1}(\|g_j\|) + \sum_{j=1}^{p} p_{\lambda_2}(\|g_j''\|). \tag{3}$$

There is more than one way to specify the penalty function and here we only focus on the SCAD penalty function (Fan and Li 2001), defined by its first derivative

$$p'_\lambda(x) = \lambda\left\{I(x \le \lambda) + \frac{(a\lambda - x)_+}{(a-1)\lambda} I(x > \lambda)\right\},$$

with $a > 2$ and $p_\lambda(0) = 0$. We will use $a = 3.7$ as suggested in Fan and Li (2001). The SCAD penalty is motivated by three desirable properties of a penalty function: unbiased for large signals, resulting in sparse estimates due to singularity at zero, and producing estimators continuous in data. Other choices of penalty, such as the adaptive lasso (Zou 2006) or the minimax concave penalty (Zhang 2010), are expected to produce similar results in both theory and practice.

2.2 MM Algorithm

Quantile regression problems are usually solved by reformulating them into linear programming problems and efficient solvers for linear programming are then applied. However, due to the use of penalties, reduction to linear programming is no longer possible. Instead, we use the majorization-minimization (MM) algorithm to solve (3), which is a general technique for solving complicated optimization problems [see Hunter and Lange (2004) for a nice review]. We present the method here briefly with more details on the idea found in the article by Hunter and Lange (2000).

First, the loss function $\rho_\tau(u)$ is approximated by its perturbation for some small $\epsilon > 0$,

$$\rho_\tau^\epsilon(u) := \rho_\tau(u) - \frac{\epsilon}{2}\ln(\epsilon + |u|).$$

Then the function

$$\tilde\rho(u|u^k) = \frac{1}{4}\left[\frac{u^2}{\epsilon + |u^k|} + (4\tau - 2)u + c\right]$$

can be shown to majorize $\rho_\tau^\epsilon(u)$ at u^k [which simply means that $\tilde\rho(u|u^k) \ge \rho_\tau^\epsilon(u)$ for all u and $\tilde\rho(u^k|u^k) = \rho_\tau^\epsilon(u^k)$] for an appropriately chosen constant c. Without the penalty, at iteration $k+1$, the MM algorithm works by minimizing the majorizer

$$\frac{1}{n}\sum_i \tilde\rho_\tau\left(u_i|u_i^k\right)$$

with respect to μ, b, where $u_i = Y_i - \mu - Z_i^T b$, and $u_i^k = Y_i - \mu^k - Z_i^T b^k$ is the residual at iteration k. The minimizer is the new estimate μ^{k+1}, b^{k+1}.

With the two-fold penalty, the implementation is only slightly more complicated. Similar to the loss function, the two penalties can be approximated by

$$p_{\lambda_1}\left(\|g_j^{(k)}\|\right) + \frac{1}{2}\frac{p'_{\lambda_1}\left(\|g_j^{(k)}\|\right)}{\|g_j^{(k)}\| + \epsilon}\left\{\|g_j\|^2 - \|g_j^{(k)}\|^2\right\},$$

and

$$p_{\lambda_2}\left(\|g_j^{(k)''}\|\right) + \frac{1}{2}\frac{p'_{\lambda_2}\left(\|g_j^{(k)''}\|\right)}{\|g_j^{(k)''}\| + \epsilon}\left\{\|g_j^{''}\|^2 - \|g_j^{(k)''}\|^2\right\}.$$

Note that this is similar to the majorizer for the loss function when $\tau = 0.5$, and is just the same as local quadratic approximation advocated in Fan and Li (2001).

After these approximations, the minimization problem in each iteration is a quadratic function and can be solved in closed form. In our implementation, we set $\epsilon = 10^{-8}$. Besides, if $\|g_j\|$ or $\|g_j''\|$ falls below a small number (10^{-6} in our implementation), we take it to be zero.

Note that Hunter and Li (2005) had already shown that the above quadratic approximation for the penalty function actually majorizes a function that converges to the penalty function as $\epsilon \to 0$. Thus the iterative algorithm is an MM algorithm with solution converging to the minimizer of a functional (denoted by, say, Q_ϵ) that closely approximates Q defined in Equation (3). Thus by the general property of the MM algorithm, our algorithm has a descent property, in the sense that in each iteration, it decreases the value of Q_ϵ. Based on proposition 3.2 in Hunter and Li (2005) and proposition 5 in Hunter and Lange (2000), Q_ϵ is close to Q when $\epsilon \approx 0$. The reader can refer to these two articles for more details.

Finally, we note that if $\|\hat m_j''\| = 0$, then $\hat m_j$ is a linear function and implicitly we actually get an estimate $\hat\beta_j$ for the slope parameter.

The MM algorithm is very easy to implement since the solution has a closed-form expression at each iteration. For linear quantile regression, Wu and Liu (2009) proposed to use the difference convex algorithm (DCA) to solve the optimization problem with the SCAD penalty and showed that it is much faster than the MM algorithm. While it is certainly interesting to consider ways to speed up the numerical calculations in a similar fashion, we note that DCA cannot be directly applied in our case due to the appearance of $\|g_j\|$ and $\|g_j''\|$, which makes linear programming inapplicable even after writing the penalty as the difference of two convex functions.

2.3 Tuning Parameter Selection

In practice, to achieve good numerical performance, we need to choose several parameters appropriately. We fix the spline order to be $q = 4$, that is, we use cubic splines in all our numerical examples. For the number of basis functions K, we first fit the additive model without any penalization and use 10-fold cross-validation to select K.

With K determined, we propose to use an SIC to select the regularization parameters λ_1 and λ_2 simultaneously. In our context, a natural SIC is defined by

$$\mathrm{SIC}_\lambda = \log\left\{\sum_i \rho_\tau\left(Y_i - \hat\mu_\lambda - Z_i^T \hat b_\lambda\right)\right\} + (d_1 K + d_2)\frac{\log n}{2n},$$

where $\hat\mu_\lambda, \hat b_\lambda$ is the regularized estimate when $\lambda = (\lambda_1, \lambda_2)$ are used as the smoothing parameters, d_1 is the number of components estimated as nonparametric, and d_2 is the number of components estimated as parametric. We will demonstrate that SIC performs well in our numerical examples.

2.4 Asymptotic Results

For convenience, we assume that m_j is truly nonparametric for $1 \le j \le p_1$, is linear for $p_1 + 1 \le j \le s = p_1 + p_2$, and is zero for $s + 1 \le j \le p$. The true components are denoted by m_{0j}, $1 \le j \le p$ and the true slope parameters for the parametric components are denoted by $\beta_0 = (\beta_{0,p_1+1}, \ldots, \beta_{0s})^T$. Let F_i be the cumulative distribution function and f_i be the density function of ϵ_i conditional on $X_{i1}, \ldots, X_{is}$ (with the corresponding random element denoted by f, which depends on X). Denote $\mathbf{f} = \text{diag}\{f_1(0), \ldots, f_n(0)\}$.

Let $X^{(1)} = (X_1, \ldots, X_{p_1})^T$ and $X^{(2)} = (X_{p_1+1}, \ldots, X_s)^T$. Let $\mathcal{A}$ denote the subspace of functions on R^{p_1} that take an additive form

$$\mathcal{A} := \big\{h\big(x^{(1)}\big) : h(x^{(1)}) = h_1(x_1) + \cdots + h_{p1}(x_{p1}), \\ Eh_j(X_j)^2 < \infty \text{ and } Eh_j(X_j) = 0\big\},$$

and for any random variable W with $E(W^2) < \infty$, let $E_{\mathcal{A}}(W)$ denote the projection of W onto $\mathcal{A}$ in the sense that

$$E\{f(0)(W - E_{\mathcal{A}}(W))(W - E_{\mathcal{A}}(W))\} \\ = \inf_{h\in\mathcal{A}} E\big\{f(0)\big(W - h\big(X^{(1)}\big)\big)\big(W - h(X^{(1)})\big)\big\}.$$

The definition of $E_{\mathcal{A}}(W)$ trivially extends to the case where W is a random vector by componentwise projection.

Let $h(X^{(1)}) = E_{\mathcal{A}}(X^{(2)})$. Each component of $h(X^{(1)}) = (h_{(1)}(X^{(1)}), \ldots, h_{(p_2)}(X^{(1)}))^T$ can be written in the form $h_{(s)}(x) = \sum_{j=1}^{p_1} h_{(s)j}(x_j)$ for some $h_{(s)j} \in S_j^0$. Denote $\Xi_1 = Ef(0)\{(X^{(2)} - h(X^{(1)}))(X^{(2)}-h(X^{(1)}))^T\}$, $\Xi_2 = E\tau(1-\tau)\{(X^{(2)} - h(X^{(1)}))(X^{(2)} - h(X^{(1)}))^T\}$. These definitions are similar to those in Wang, Zhu, and Zhou (2009) for varying-coefficient models. In the case of quadratic loss, when there is only one nonparametric component (i.e., the true model is a partially linear model), we can simply define the projection $E(X^{(2)}|X_1)$, and working with $X^{(2)} - E(X^{(2)}|X_1)$ to "profile out" the nonparametric part is the basic strategy used for partially linear models in investigating the asymptotic properties of the linear part. Our definitions above can thus be regarded as an extension for dealing with additive partially linear quantile regression models.

The following standard regularity assumptions are used.

(A1) The covariate vector X has a continuous density supported on $[0, 1]^p$. Furthermore, the marginal densities for X_j, $1 \le j \le p$ are all bounded from below and above by two fixed positive constants, respectively.

(A2) $F_i(0) = \tau$, and f_i is bounded away from zero and has a continuous and uniformly bounded derivative.

(A3) $Em_j(X_j) = 0$, $1 \le j \le s$. $m_j(x)$ is linear in x for $p_1 + 1 \le j \le s$, and $m_j \equiv 0$ for $j > s$.

(A4) For $g = m_j$, $1 \le j \le p_1$ or $g = h_{(s)j}$, $1 \le s \le p_2$, $1 \le j \le p_1$, g satisfies a Lipschitz condition of order $d > 1/2$: $|g^{(\lfloor d\rfloor)}(t) - g^{(\lfloor d\rfloor)}(s)| \le C|s - t|^{d-\lfloor d\rfloor}$, where $\lfloor d\rfloor$ is the biggest integer strictly smaller than d and $g^{(\lfloor d\rfloor)}$ is the $\lfloor d\rfloor$th derivative of g. The order of the B-spline used satisfies $q \ge d + 2$.

(A5) The matrices Ξ_1 and Ξ_2 are both positive definite.

Theorem 1. Assume (A1)–(A5), and that $K \sim n^{1/(2d+1)}$, $\lambda_1, \lambda_2 \to 0$, we have the rate of convergence

$$\|m_{0j} - \hat{m}_j\|^2 = O\big(n^{-\frac{2d}{2d+1}}\big), 1 \le j \le p,$$

where $\hat{m}_j = \hat{b}_j^T B_j$ is the estimated component function.

Remark 1. Although not explicit in the convergence rate, it would be clear from the proof in the Appendix that the convergence rate can be written as $K/n + K^{-2d}$ for a range of values of K. The first term represents the stochastic error (as K increases, the dimension increases), while the second term corresponds to the approximation error (as K increases, the spline functions are more flexible). Thus there is a bias–variance trade-off in the choice of K. We will show in the next section that K chosen by 10-fold cross-validation works well in practice.

The next theorem shows that when λ_1, λ_2 are appropriately specified, we can select the true partially linear additive model with high probability.

Theorem 2. In addition to the assumptions in Theorem 1, we assume $n^{d/(2d+1)} \min\{\lambda_1, \lambda_2\} \to \infty$. Then with probability approaching 1,

(a) $\hat{m}_j \equiv 0$, $s + 1 \le j \le p$,
(b) $\hat{m}_j$ is a linear function for $p_1 + 1 \le j \le s$.

Next, we show that for the linear components, the estimator for the slope parameter is asymptotically normal (this estimator $\hat{\beta}_j$ is implicitly defined by $\hat{m}_j$ when $\hat{m}_j$ represents a linear function). We note that the asymptotic variance is the same as when the true model is known beforehand, thus our estimator has the so-called oracle property.

Theorem 3. (Asymptotic Normality) Under the same set of assumptions as in Theorem 2, we have $\sqrt{n}(\hat{\beta} - \beta_0) \to N(0, \Xi_1^{-1}\Xi_2\Xi_1^{-1})$ in distribution.

Remark 2. With regard to the consistency and asymptotic normality results presented above, we should also mention the recent work by Pötscher and coauthors (Leeb and Pötscher 2006, 2008; Pötscher and Schneider 2009), who have studied the (uniform) consistency of estimates of distribution functions of the penalized estimators including the SCAD estimator. In particular, they have shown that if an estimator is consistent in model selection, it is impossible to possess the oracle property (asymptotic normality) uniformly in a neighborhood of zero coefficients. These results are of great interest, but uniform convergence is more relevant when many of the true regression parameters are very close to zero, which commonly arises in cases where the number of parameters (in our context, the number of parameters in the linear part) diverges with sample size. Although we acknowledge that taking the parameters in our model to be fixed is partly subjective and in some cases modeling the parameters as changing with sample size is more appropriate, the oracle property for fixed parameter is still an important issue.

Finally, in the same spirit of that of Wang, Li, and Tsai (2007), we come to the question of whether the SIC can identify the true model in our setting.

Theorem 4. Under assumptions (A1)–(A5) and that $K \sim n^{1/(2d+1)}$, as assumed in Theorem 1, the parameters $\hat{\lambda}_1$, $\hat{\lambda}_2$ selected by SIC can select the true model with probability approaching 1.

3. NUMERICAL EXAMPLES

3.1 Simulation Studies

We conducted Monte Carlo studies for the following iid and heteroscedastic error model

$$Y_i = \sum_{j=1}^{p} m_j(X_{ij}) + (1 + |X_{i1}|)\big(e_i - F_e^{-1}(\tau)\big), \tag{4}$$

with $m_1(x) = 3\sin(2\pi x)/(2 - \sin(2\pi x))$, $m_2(x) = 6x(1-x)$, $m_3(x) = 2x, m_4(x) = x, m_5(x) = -x, p = 10$ so that five components are actually zero, and $F_e(.)$ denotes the distribution function of the mean zero error e_i. Thus in our generating model, the number of nonparametric components is $p_1 = 2$ and the number of nonzero linear components is $p_2 = 3$. Several simulation scenarios are considered. For sample size, we set $n = 100$ or 200, for τ we consider $\tau = 0.25$ and 0.5, for e_i we consider a normal distribution with standard deviation 0.2, and a Student's t distribution with scale parameter 0.2 and degrees of freedom 2. To generate the covariates, we first let X_{ij} be marginally standard normal with correlations given by $\mathrm{cov}(X_{ij_1}, X_{ij_2}) = (1/2)^{|j_1 - j_2|}$, and then apply the cumulative distribution function of the standard normal distribution to transform X_{ij} to be marginally uniform on [0, 1]. When generating data from (4) with $\tau = 0.5$, we use the quadratic loss function (mean regression) as well as $\rho_{0.5}$ to perform estimation for comparison. For any loss function used, we compute four estimators, including the oracle estimator where the nonparametric and the parametric components are correctly specified with no penalty used, our estimator with the two-fold SCAD penalty, the sparse additive estimator where only the first SCAD penalty is used (thus it performs variable selection but does not try to find the parametric components), and finally the linear quantile regression model with a SCAD penalty. We use 10-fold cross-validation to select K and use SIC to select the regularization parameters in both our estimator and the sparse additive estimator with a single penalty. The implementations are carried out on an HP workstation xw8600 in R software.

For all scenarios, 100 datasets are generated and the results are summarized in Tables 1–7. In Table 1, we investigate the model selection results for both our estimator and the sparse additive estimator with one single penalty, when the errors are Gaussian. For both estimators, we see SIC successfully selected the nonzero components, with increased sample size resulting in slightly better performance. Our estimator can further distinguish between nonparametric and parametric components. Furthermore, when data are generated from (4) with $\tau = 0.5$, the results from quantile regression (median regression) and least squares regression are similar. Table 2 reports the model selection results for errors with t distribution.

Table 1. Model selection results for our estimator and the sparse additive estimator, both using SIC for model selection, when errors are Gaussian

n	#pen	loss	NN	NNT	NL	NLT
100	1	$\rho_{0.25}$	$6.10_{2.012}$	$4.98_{0.141}$	0_0	0_0
		$\rho_{0.5}$	$5.82_{1.480}$	$4.96_{0.197}$	0_0	0_0
		lsq	$5.46_{1.034}$	5_0	0_0	0_0
	2	$\rho_{0.25}$	$2.66_{1.061}$	2_0	$4.06_{1.973}$	$2.70_{0.788}$
		$\rho_{0.5}$	$2.72_{1.278}$	2_0	$3.46_{1.643}$	$2.72_{0.814}$
		lsq	$2.82_{1.137}$	2_0	$2.86_{1.578}$	$2.54_{1.002}$
200	1	$\rho_{0.25}$	$5.16_{0.509}$	5_0	0_0	0_0
		$\rho_{0.5}$	$5.26_{0.527}$	5_0	0_0	0_0
		lsq	$5.10_{0.303}$	5_0	0_0	0_0
	2	$\rho_{0.25}$	$2.34_{0.658}$	2_0	$3.28_{1.143}$	$2.90_{0.614}$
		$\rho_{0.5}$	$2.32_{0.652}$	2_0	$3.16_{1.149}$	$2.90_{0.646}$
		lsq	$2.26_{0.564}$	2_0	$3.00_{0.857}$	$2.94_{0.564}$

NOTE: NN: average number of nonparametric components selected; NNT: average number of nonparametric components selected that are truly nonparametric (or truly nonzero for sparse additive estimator with one penalty); NL: average number of linear components selected; NLT: average number of linear components selected that are truly linear. The numbers in smaller fonts are the corresponding standard errors; #pen being 1 indicates the sparse additive estimator with one single penalty and #pen being 2 indicates our estimator with two-fold penalty; $\rho_{0.25}$ and $\rho_{0.5}$ denote the quantile regressions (with check function as the loss function) and "lsq" denotes the least squares regression (mean regression).

Table 2. Model selection results when random errors are distributed as t_2

n	#pen	los	NN	NNT	NL	NLT
100	1	$\rho_{0.25}$	$5.96_{2.303}$	$4.68_{0.793}$	0_0	0_0
		$\rho_{0.5}$	$5.76_{2.015}$	$4.72_{0.881}$	0_0	0_0
		lsq	$5.76_{2.065}$	$4.48_{0.838}$	0_0	0_0
	2	$\rho_{0.25}$	$2.80_{1.245}$	2_0	$3.40_{1.851}$	$2.48_{0.990}$
		$\rho_{0.5}$	$2.48_{1.199}$	$1.96_{0.197}$	$4.26_{1.987}$	$2.78_{0.882}$
		lsq	$3.28_{1.616}$	$1.90_{0.364}$	$2.44_{2.051}$	$1.94_{1.274}$
200	1	$\rho_{0.25}$	$5.40_{1.178}$	$4.92_{0.395}$	0_0	0_0
		$\rho_{0.5}$	$5.42_{0.758}$	5_0	0_0	0_0
		lsq	$5.14_{1.212}$	$4.68_{0.712}$	0_0	0_0
	2	$\rho_{0.25}$	$2.52_{0.579}$	2_0	$3.04_{1.105}$	$2.70_{0.580}$
		$\rho_{0.5}$	$2.30_{0.462}$	2_0	$3.06_{0.818}$	$2.92_{0.453}$
		lsq	$2.72_{0.969}$	$1.98_{0.141}$	$2.72_{1.616}$	$2.40_{0.968}$

Tables 3 and 4, for Gaussian random errors and Student's t errors, respectively, contain root mean squared errors (RMSE) for the first six component functions (note the sixth component is actually zero), which is defined by

$$\mathrm{RMSE}_j = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(\hat{m}_j(t_i) - m_j(t_i))^2},$$

on a fine grid $(t_1, \ldots, t_T)$ consisting of 500 points equally spaced on [0, 1]. We also show the RMSE for the regression function $m = \sum_{j=1}^{10} m_j$. For data generated from (4) with $\tau = 0.5$, we also compare the performance of median regression with least squares regression. On the nonparametric components (m_1 and m_2), the errors for estimators with a single penalty (sparse additive model) and double penalties are similar, and both are qualitatively close to those of the oracle estimator. However, for the parametric components, our estimator with the two-fold penalty is obviously more efficient, leading to about 40%∼50% reduction in RMSE. The estimation results for the 0.25 quantile are slightly worse than those for the 0.5 quantile. As expected, least squares regression is better than median regression in estimation when the errors are Gaussian, and worse than median regression when the errors are t_2. We also performed simulations for the Cauchy random errors (Student's t with degree of freedom 1), and in this case, the advantage of median regression

Table 3. Root mean squared errors for $m_1, \dots, m_6, m$ for the four different estimators, when errors are Gaussian

n	L		Oracle	Our estimator	Sparse additive	Linear
100	$\rho_{0.25}$	m_1	$0.2578_{0.0373}$	$0.2881_{0.0557}$	$0.2623_{0.0303}$	$1.1745_{0.0016}$
		m_2	$0.1030_{0.0331}$	$0.1111_{0.0488}$	$0.1157_{0.0494}$	$0.5794_{0.0158}$
		m_3	$0.0421_{0.0302}$	$0.0431_{0.0336}$	$0.1094_{0.0445}$	$0.2891_{0.0400}$
		m_4	$0.0414_{0.0325}$	$0.0500_{0.0371}$	$0.1193_{0.0531}$	$0.1020_{0.1374}$
		m_5	$0.0417_{0.0293}$	$0.0519_{0.0433}$	$0.1221_{0.0515}$	$0.1077_{0.1389}$
		m_6	0_0	$0.0148_{0.0231}$	$0.0375_{0.0555}$	$0.0113_{0.0561}$
		m	$0.2915_{0.0526}$	$0.3476_{0.0881}$	$0.3515_{0.0744}$	$1.2109_{0.0578}$
	$\rho_{0.5}$	m_1	$0.2479_{0.0329}$	$0.2562_{0.0256}$	$0.2482_{0.0234}$	$1.1665_{0.0011}$
		m_2	$0.0884_{0.0298}$	$0.0989_{0.0366}$	$0.1052_{0.0394}$	$0.5820_{0.0200}$
		m_3	$0.0391_{0.0248}$	$0.0429_{0.0359}$	$0.1091_{0.0377}$	$0.2891_{0.0400}$
		m_4	$0.0407_{0.0329}$	$0.0552_{0.0418}$	$0.1289_{0.0559}$	$0.0793_{0.1285}$
		m_5	$0.0386_{0.0304}$	$0.0536_{0.0388}$	$0.1301_{0.0552}$	$0.1360_{0.1430}$
		m_6	0_0	$0.0176_{0.0349}$	$0.0153_{0.0440}$	$0.0283_{0.0859}$
		m	$0.2749_{0.0456}$	$0.3040_{0.0489}$	$0.3222_{0.0667}$	$1.2446_{0.0962}$
	lsq	m_1	$0.2335_{0.0172}$	$0.2369_{0.0166}$	$0.2335_{0.0157}$	$1.1747_{0.0012}$
		m_2	$0.0756_{0.0274}$	$0.0805_{0.0271}$	$0.0802_{0.0294}$	$0.5847_{0.0231}$
		m_3	$0.0324_{0.0237}$	$0.0412_{0.0306}$	$0.0777_{0.0282}$	$0.2891_{0.0400}$
		‘m_4	$0.0299_{0.0212}$	$0.0450_{0.0321}$	$0.0873_{0.0319}$	$0.0850_{0.1312}$
		m_5	$0.0251_{0.0224}$	$0.0396_{0.0319}$	$0.0882_{0.0343}$	$0.0793_{0.1285}$
		m_6	0_0	$0.0099_{0.0263}$	$0.0065_{0.0213}$	$0.0396_{0.0993}$
		m	$0.2566_{0.0327}$	$0.2775_{0.0403}$	$0.3190_{0.0971}$	$1.2201_{0.0851}$
200	$\rho_{0.25}$	m_1	$0.2419_{0.0183}$	$0.2398_{0.0162}$	$0.2417_{0.0207}$	$1.1732_{0.0008}$
		m_2	$0.0700_{0.0280}$	$0.0779_{0.0211}$	$0.0778_{0.0254}$	$0.5780_{0.0130}$
		m_3	$0.0274_{0.0198}$	$0.0324_{0.0279}$	$0.0688_{0.0298}$	$0.2834_{0.0037}$
		m_4	$0.0208_{0.0173}$	$0.0382_{0.0261}$	$0.0713_{0.0302}$	$0.0226_{0.0776}$
		m_5	$0.0291_{0.0243}$	$0.0414_{0.0323}$	$0.0786_{0.0416}$	$0.0421_{0.0923}$
		m_6	0_0	$0.0047_{0.0152}$	$0.0009_{0.0049}$	$0.0170_{0.0680}$
		m	$0.2549_{0.0246}$	$0.2599_{0.0291}$	$0.2679_{0.0338}$	$1.2011_{0.0492}$
	$\rho_{0.5}$	m_1	$0.2327_{0.0159}$	$0.2324_{0.0132}$	$0.2311_{0.0133}$	$1.1704_{0.0066}$
		m_2	$0.0652_{0.0216}$	$0.0734_{0.0234}$	$0.0675_{0.0189}$	$0.5833_{0.0216}$
		m_3	$0.0269_{0.0231}$	$0.0377_{0.0299}$	$0.0684_{0.0306}$	$0.2834_{0.0039}$
		m_4	$0.0290_{0.0236}$	$0.0409_{0.0303}$	$0.0726_{0.0307}$	$0.0283_{0.0859}$
		m_5	$0.0270_{0.0177}$	$0.0374_{0.0262}$	$0.0691_{0.0230}$	$0.0282_{0.0828}$
		m_6	0_0	$0.0058_{0.0191}$	$0.0045_{0.0162}$	$0.0396_{0.0993}$
		m	$0.2499_{0.0268}$	$0.2617_{0.0360}$	$0.2754_{0.0403}$	$1.2160_{0.0815}$
	lsq	m_1	$0.2249_{0.0080}$	$0.2270_{0.0091}$	$0.2253_{0.0082}$	$1.1748_{0.0062}$
		m_2	$0.0543_{0.0151}$	$0.0602_{0.0177}$	$0.0560_{0.0164}$	$0.5820_{0.0201}$
		m_3	$0.0223_{0.0176}$	$0.0236_{0.0199}$	$0.0530_{0.0194}$	$0.2834_{0.0041}$
		m_4	$0.0222_{0.0177}$	$0.0266_{0.0202}$	$0.0557_{0.0244}$	$0.0049_{0.0256}$
		m_5	$0.0196_{0.0134}$	$0.0237_{0.0155}$	$0.0540_{0.0166}$	$0.0170_{0.0680}$
		m_6	0_0	$0.0029_{0.0118}$	$0.0045_{0.0183}$	$0.0283_{0.0859}$
		m	$0.2392_{0.0160}$	$0.2487_{0.0202}$	$0.2586_{0.086}$	$1.2193_{0.0659}$

NOTE: The numbers in smaller fonts are the corresponding standard errors.

is much more obvious (not reported here). The linear estimators obviously do not perform well for data simulated from a semiparametric model.

Now we use the case $n = 100$, $\tau = 0.5$ with normally distributed errors to study the sensitivity of the estimation results to the tuning parameters, and demonstrate that both 10-fold cross-validation (for choosing K) and SIC (for choosing λ_1 and λ_2) work well empirically. Figure 1 shows the RMSE for $K = 4, 5, \dots, 10$ as well as for K selected by 10-fold cross-validation. Here λ_1 and λ_2 are chosen by SIC in all cases. It is seen that the RMSE does not change much for K ranging from 5 to 8. For larger K, the effect of overfitting begins to appear. The 10-fold cross-validation works very well in this simulation. Table 5 shows that in terms of model selection, $K = 4$ and $K = 5$ are the best with 10-fold cross-validation still achieving a good accuracy. Next we use K chosen by 10-fold cross-validation and consider the sensitivity of the results to λ_1 and λ_2. Let $(\hat{\lambda}_1, \hat{\lambda}_2)$ be the tuning parameters found by SIC, and in Figure 2, we compare the RMSE when using tuning parameters $(c\hat{\lambda}_1, c\hat{\lambda}_2)$, for $c \in \{0.1, 0.25, 0.5, 0.75, 1, 1.5, 2\}$ (note that $c = 1$ just produces the original results when tuning parameters are selected by SIC). We see that SIC also has a reasonably good performance. Using $c \geq 1.5$ results in much larger RMSE. There is little overfitting when c is small, which is not surprising since when p is small compared with the sample size as in our simulations, overfitting is not expected to be severe even if no penalization is used. On the other hand, as Table 6 shows, tuning parameters selected by SIC produce reasonable model selection accuracy, while using c too small is not advisable.

Table 4. Root mean squared errors when errors are t_2

n	Loss		Oracle	Our estimator	Sparse additive	Linear
100	$\rho_{0.25}$	m_1	$0.2974_{0.1352}$	$0.3401_{0.1372}$	$0.3108_{0.1388}$	$1.1865_{0.0477}$
		m_2	$0.1349_{0.0555}$	$0.1636_{0.0781}$	$0.1809_{0.0900}$	$0.5807_{0.0180}$
		m_3	$0.0619_{0.0523}$	$0.0943_{0.0750}$	$0.1958_{0.0918}$	$0.3344_{0.1100}$
		m_4	$0.0584_{0.0523}$	$0.0842_{0.0717}$	$0.1718_{0.0697}$	$0.1417_{0.1431}$
		m_5	$0.0551_{0.0387}$	$0.0824_{0.0735}$	$0.1790_{0.0603}$	$0.1587_{0.1421}$
		m_6	0_0	$0.0162_{0.0427}$	$0.0428_{0.0777}$	$0.0340_{0.0930}$
		m	$0.3453_{0.1609}$	$0.4208_{0.1460}$	$0.4236_{0.1639}$	$1.2096_{0.0889}$
	$\rho_{0.5}$	m_1	$0.2544_{0.0244}$	$0.2682_{0.0342}$	$0.2641_{0.0315}$	$1.1785_{0.0281}$
		m_2	$0.1141_{0.0488}$	$0.1449_{0.1066}$	$0.1617_{0.1104}$	$0.5807_{0.0180}$
		m_3	$0.0542_{0.0419}$	$0.0832_{0.1127}$	$0.1664_{0.1070}$	$0.3118_{0.0859}$
		m_4	$0.0438_{0.0362}$	$0.0741_{0.0659}$	$0.1608_{0.0684}$	$0.1360_{0.1430}$
		m_5	$0.0499_{0.0368}$	$0.0832_{0.0638}$	$0.1672_{0.0715}$	$0.1303_{0.1427}$
		m_6	0_0	$0.0214_{0.0420}$	$0.0191_{0.0496}$	$0.0170_{0.0680}$
		m	$0.3004_{0.0592}$	$0.3492_{0.1182}$	$0.3766_{0.1257}$	$1.1925_{0.0606}$
	lsq	m_1	$0.2558_{0.0255}$	$0.3059_{0.1377}$	$0.2776_{0.0538}$	$1.1744_{0.0022}$
		m_2	$0.1409_{0.0955}$	$0.1702_{0.1432}$	$0.1660_{0.1160}$	$0.5794_{0.0158}$
		m_3	$0.0586_{0.0408}$	$0.1300_{0.1159}$	$0.1809_{0.0840}$	$0.3061_{0.0776}$
		m_4	$0.0561_{0.0470}$	$0.1352_{0.0944}$	$0.1773_{0.0769}$	$0.1587_{0.1421}$
		m_5	$0.0720_{0.0544}$	$0.1369_{0.0929}$	$0.1697_{0.0810}$	$0.1474_{0.1430}$
		m_6	0_0	$0.0142_{0.0419}$	$0.0393_{0.0745}$	$0.0170_{0.0680}$
		m	$0.3300_{0.0854}$	$0.4055_{0.1897}$	$0.4353_{0.1331}$	$1.2045_{0.0759}$
200	$\rho_{0.25}$	m_1	$0.2546_{0.0281}$	$0.2697_{0.0402}$	$0.2574_{0.0251}$	$1.1759_{0.0022}$
		m_2	$0.0927_{0.0375}$	$0.0980_{0.0429}$	$0.1040_{0.0403}$	$0.5780_{0.0130}$
		m_3	$0.0453_{0.0290}$	$0.0466_{0.0381}$	$0.1021_{0.0425}$	$0.2834_{0.0069}$
		m_4	$0.0352_{0.0302}$	$0.0564_{0.0422}$	$0.1159_{0.0516}$	$0.0453_{0.1049}$
		m_5	$0.0403_{0.0267}$	$0.0507_{0.0331}$	$0.1106_{0.0601}$	$0.0623_{0.1186}$
		m_6	0_0	$0.0038_{0.0137}$	$0.0069_{0.0320}$	$0.0113_{0.0561}$
		m	$0.2815_{0.0465}$	$0.3046_{0.0743}$	$0.3165_{0.0607}$	$1.2120_{0.0618}$
	$\rho_{0.5}$	m_1	$0.2335_{0.0138}$	$0.2460_{0.0274}$	$0.2329_{0.0141}$	$1.1723_{0.0013}$
		m_2	$0.0825_{0.0265}$	$0.0949_{0.0352}$	$0.0855_{0.0326}$	$0.5833_{0.0216}$
		m_3	$0.0271_{0.0191}$	$0.0319_{0.0217}$	$0.0791_{0.0263}$	$0.2891_{0.0400}$
		m_4	$0.0342_{0.0233}$	$0.0452_{0.0346}$	$0.0821_{0.0399}$	$0.0453_{0.1049}$
		m_5	$0.0304_{0.0262}$	$0.0391_{0.0308}$	$0.0779_{0.0333}$	$0.0566_{0.1145}$
		m_6	0_0	$0.0034_{0.0148}$	$0.0046_{0.0242}$	$0.0170_{0.0680}$
		m	$0.2589_{0.0245}$	$0.2856_{0.0477}$	$0.2782_{0.0328}$	$1.2127_{0.0704}$
	lsq	m_1	$0.2704_{0.1228}$	$0.2916_{0.1221}$	$0.2770_{0.1194}$	$1.1785_{0.0281}$
		m_2	$0.1359_{0.1013}$	$0.1558_{0.1087}$	$0.1501_{0.1076}$	$0.5807_{0.0181}$
		m_3	$0.0612_{0.1020}$	$0.0827_{0.1038}$	$0.1491_{0.1174}$	$0.2948_{0.0561}$
		m_4	$0.0499_{0.0538}$	$0.0927_{0.0843}$	$0.1496_{0.0897}$	$0.0850_{0.1312}$
		m_5	$0.0533_{0.0555}$	$0.0986_{0.0872}$	$0.1515_{0.0806}$	$0.0850_{0.1312}$
		m_6	0_0	$0.0109_{0.0429}$	$0.0125_{0.0529}$	$0.0396_{0.0993}$
		m	$0.3372_{0.2257}$	$0.3733_{0.1431}$	$0.3809_{0.1260}$	$1.2185_{0.0829}$

Table 5. Model selection results when K changes from 4 to 10, compared with the results when K is chosen by 10-fold cross-validation (CV)

K	NN	NNT	NL	NLT
4	$2.86_{1.143}$	2_0	$3.48_{1.403}$	$2.52_{0.706}$
5	$2.66_{1.205}$	2_0	$3.78_{1.887}$	$2.46_{0.930}$
6	$3.68_{1.910}$	2_0	$2.86_{1.784}$	$1.90_{1.015}$
7	$3.88_{2.066}$	2_0	$2.46_{2.042}$	$1.76_{1.187}$
8	$5.06_{3.046}$	2_0	$2.06_{2.333}$	$1.34_{1.394}$
9	$8.90_{2.296}$	2_0	$0.28_{1.050}$	$0.14_{0.571}$
10	$9.86_{0.989}$	2_0	$0.06_{0.424}$	$0.06_{0.424}$
CV	$2.72_{1.278}$	2_0	$3.46_{1.643}$	$2.52_{0.814}$

Table 6. Model selection results when using $(c\hat{\lambda}_1, c\hat{\lambda}_2)$ as the regularization parameters in the penalties, where $(\hat{\lambda}_1, \hat{\lambda}_2)$ are the parameters found from SIC

c	NN	NNT	NL	NLT
0.1	$9.22_{1.233}$	2_0	$0.50_{0.735}$	$0.20_{0.451}$
0.25	$7.40_{1.564}$	2_0	$1.88_{1.271}$	$0.92_{0.922}$
0.5	$4.48_{1.631}$	2_0	$3.44_{1.774}$	$1.84_{0.997}$
0.75	$3.00_{1.087}$	2_0	$4.02_{1.911}$	$2.46_{0.838}$
1	$2.72_{1.278}$	2_0	$3.46_{1.643}$	$2.52_{0.814}$
1.5	$2.28_{0.671}$	2_0	$3.76_{1.623}$	$2.82_{0.522}$
2	$2.10_{0.462}$	2_0	$3.68_{1.463}$	$2.86_{0.495}$

Table 7. Prediction errors for the three estimators on independently simulated data. RMSE and AD are the root mean squared prediction errors and the absolute deviation prediction errors, respectively, as defined in the main text

			Oracle		Our estimator		Sparse additive	
Error	n	Loss	RMSE	AD	RMSE	AD	RMSE	AD
Normal	100	$\rho_{0.5}$	0.357	0.140	0.389	0.157	0.445	0.178
		lsq	0.337	0.134	0.358	0.145	0.383	0.153
	200	$\rho_{0.5}$	0.317	0.122	0.327	0.129	0.342	0.134
		lsq	0.300	0.118	0.309	0.123	0.321	0.127
Student's t	100	$\rho_{0.5}$	0.383	0.155	0.460	0.188	0.540	0.217
		lsq	0.410	0.166	0.544	0.219	0.583	0.235
	200	$\rho_{0.5}$	0.318	0.123	0.339	0.134	0.355	0.139
		lsq	0.382	0.153	0.450	0.180	0.484	0.194

Finally in Table 7, we compare the prediction errors for three estimators, when data are generated from (4) with $\tau = 0.5$. Both median regression and least squares regression are considered. We use two prediction accuracy measures including RMSE and absolute deviation, which are defined by

$$\sqrt{\frac{1}{n'}\sum_{i=1}^{n'}(Y_i - \hat{Y}_i)^2}$$

and

$$\frac{1}{n'}\sum_{i=1}^{n'}|Y_i - \hat{Y}_i|,$$

respectively, where $Y_i, i = 1, \ldots, n' = 200$ are the responses of an independently simulated testing dataset and $\hat{Y}_i$ are the fitted values for the different methods. Comparing our estimator with the sparse additive estimator, our estimator has better performance, which is probably due to the more efficient estimates for the parametric part. Comparing median regression with least squares regression, the relative performances depend on the error distribution as before.

3.2 Real Data

Now we illustrate the methodology with real data used by Yafeh and Yosha (2003) to investigate the relationship between shareholder concentration and several indices for managerial moral hazard in the form of expenditure with scope for private benefit. The dataset includes a variety of variables describing more than 100 Japanese industrial chemical firms listed on the Tokyo stock exchange. (The dataset is available online through the Economic Journal at *http://www.res.org.uk*.) This dataset was also used in Horowitz and Lee (2005) as an application of additive models for median regression which is more flexible than the linear model used in Yafeh and Yosha (2003). We use the general sales and administrative expenses deflated by sales

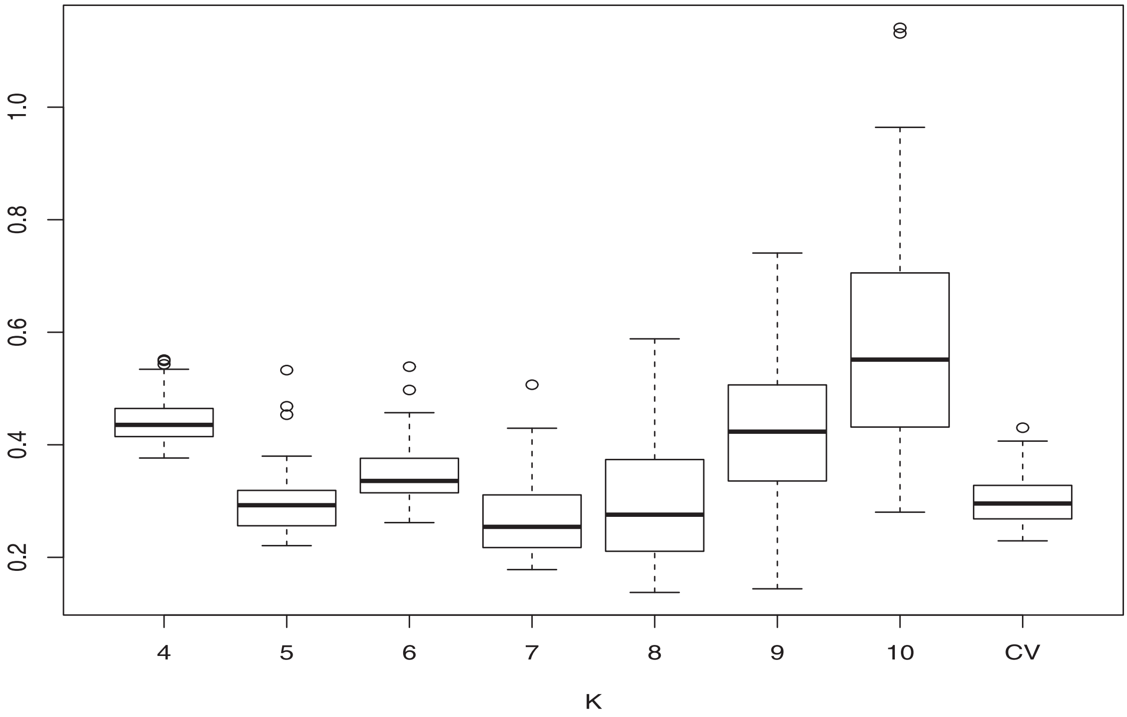

Figure 1. RMSE changes when K changes from 4 to 10, and K chosen by 10-fold cross-validation (CV) works well in terms of RMSE. This simulation is based on $n = 100$, $\tau = 0.5$ with normal errors.

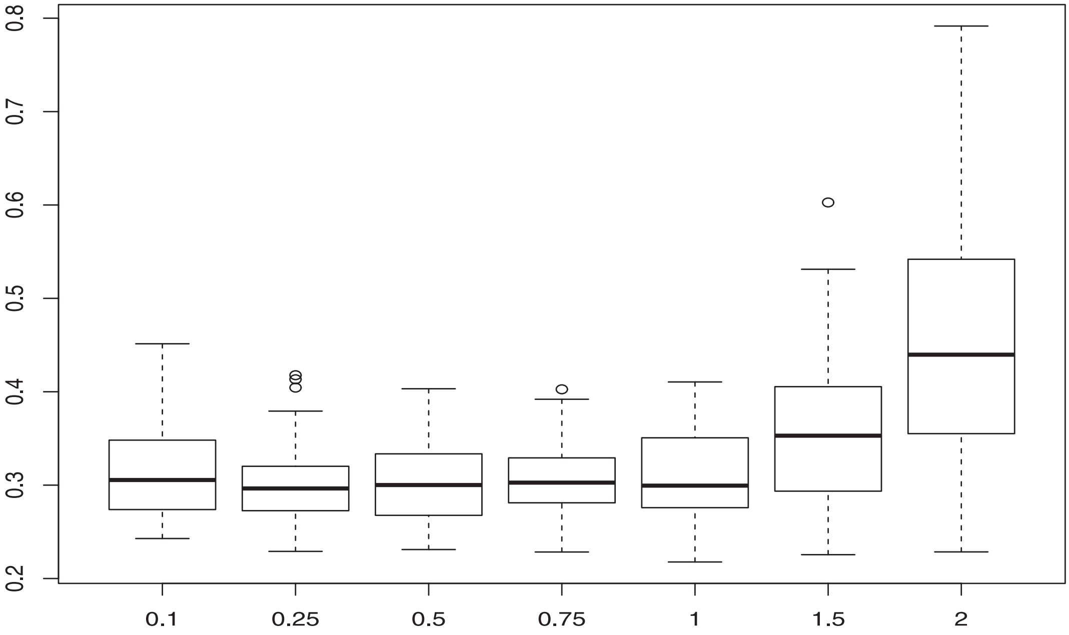

Figure 2. RMSE changes when λ_1 and λ_2 change. The label on the x-axis is the multiplicative factor on both λ_1 and λ_2 chosen by SIC. For example, the left-most boxplot is obtained by using a small λ_1 and λ_2 obtained by multiplying the tuning parameters found by SIC by a factor of 0.1, while the boxplot labeled by "1" is the results using tuning parameters obtained from SIC.

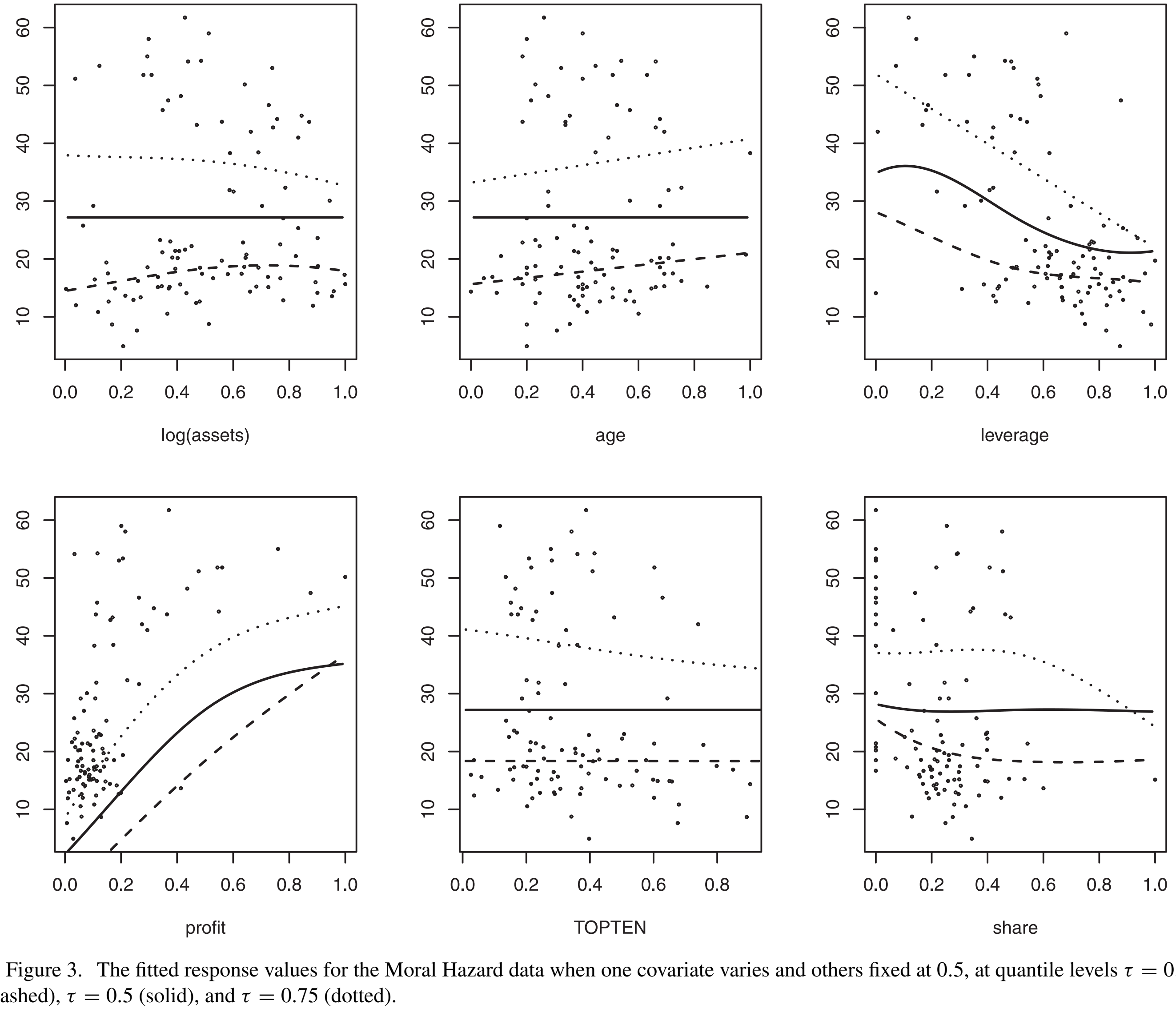

Figure 3. The fitted response values for the Moral Hazard data when one covariate varies and others fixed at 0.5, at quantile levels $\tau = 0.25$ (dashed), $\tau = 0.5$ (solid), and $\tau = 0.75$ (dotted).

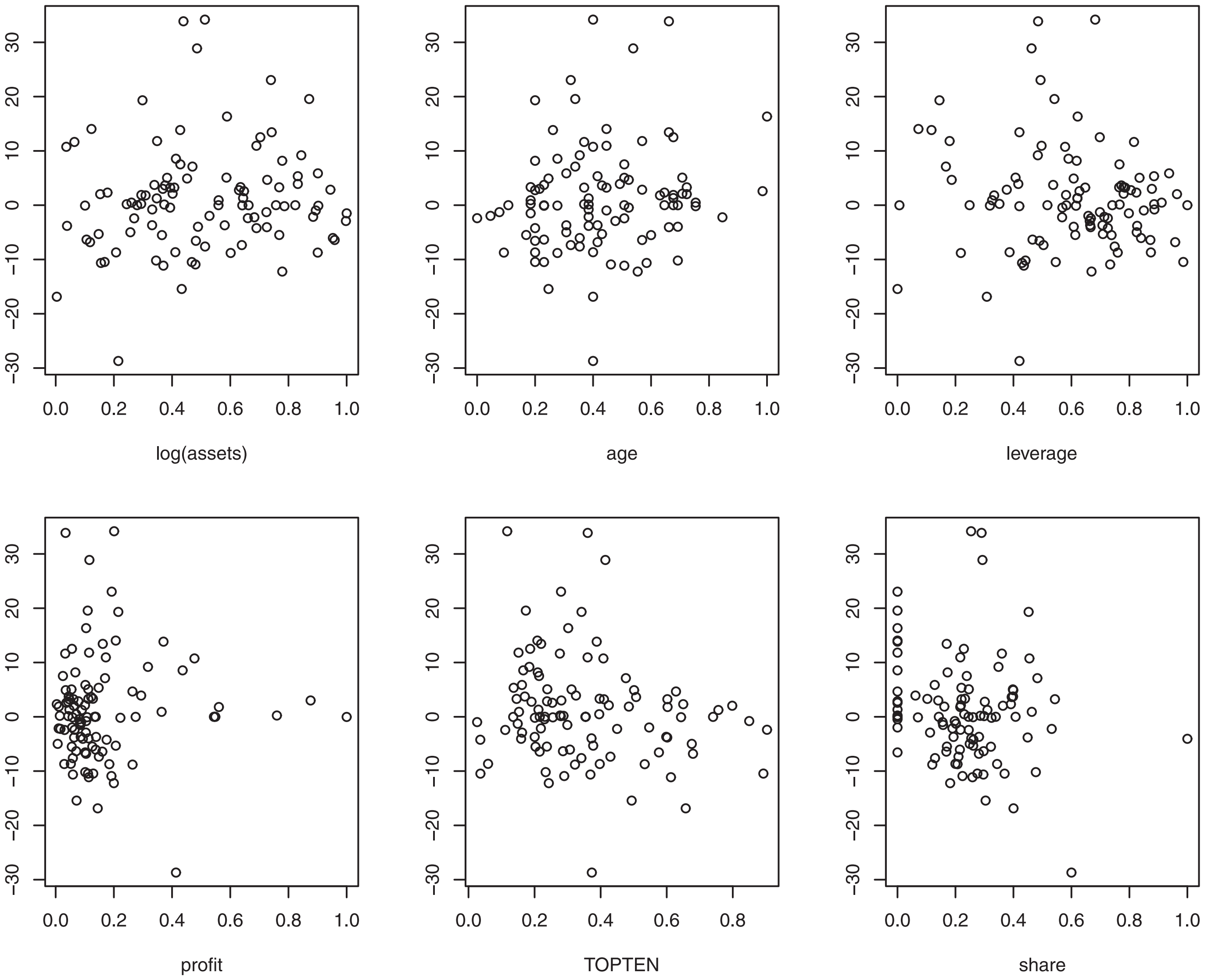

Figure 4. Scatterplots for the estimated residuals versus the covariates for $\tau = 0.5$. The assumption that the errors have median zero is visually reasonable for the estimated errors.

(denoted by MH5) as the response variable Y which is one of five measures of activities with a scope for managerial moral hazard in Yafeh and Yosha (2003). We use more covariates, a total of six, than those used in Horowitz and Lee (2005) though, including log(assets), the age of the firm, leverage (ratio of debt to total assets), profit (variance of operating profitability of firms between 1977 and 1986), TOPTEN (the percentage of ownership held by the 10 largest shareholders), and share (share of the largest creditor in total debt). All covariates are normalized by a linear transformation to lie in [0,1] before analysis. The sample size turns out to be 114 after removing firms with some covariates missing.

In Figure 3, we show the fitted values of the responses as one covariate varies while others are fixed at 0.5, at quantile levels 0.25, 0.5, and 0.75. At these three levels, we identified 5, 3, and 6 nonzero components, and 2, 0, and 3 linear (nonzero) components, respectively. The marginal relationship between each covariate and MH5 may differ depending on the quantile of interest. It is interesting to note that although some covariates are not related to responses for median regression, they have some effects at lower and/or upper quantiles. In particular, the covariate TOPTEN, which is the main covariate of interest in the study of moral hazard, is only estimated to have nonzero effect at the upper quantile. In other words, firms with a more concentrated ownership structure spend less on activities with a scope for managerial moral hazard, but this effect seems to be only restricted to the upper tail of the expenses distribution. Variability of firm performance has an increasing effect on MH5, which was nevertheless not found in Yafeh and Yosha (2003). Inspection of the data scatterplots in Figure 3 reveals that these features captured in the estimates are driven by the data and are not simply an artifact of the model. In contrast to Horowitz and Lee (2005) who found that log(asset), age, leverage, and TOPTEN are all significantly nonlinear based on confidence intervals without bias correction, three of these (except leverage) are estimated as zero in median regression in our method. Note that leverage is indeed the most significant predictor as in Horowitz and Lee (2005). We also fitted the unpenalized additive model and found that the resulting estimates were similar to those reported in Horowitz and Lee (2005) and thus the differences are mainly attributed to the use of penalties. Although Horowitz and Lee (2005) claimed that all four predictors are significant based on pointwise 90%

confidence intervals, while we find three of these (except leverage) are zeros, we note that based on their confidence intervals, indeed a major portion of the confidence intervals for these three component functions overlaps zero and deviations from zero mainly appear at the very beginning or the very end of the estimated component functions. The penalization approach is obviously different from the testing approach. In particular, the penalization approach explicitly gives a selected model by trading off model fitting with model parsimony, while when one uses testing, one needs to combine the marginal testing results somehow to get a final model. Furthermore, we also think that for testing the significance of the component functions, simultaneous confidence bands would be more appropriate than pointwise confidence intervals, if possible to obtain, which will be wider than the pointwise confidence intervals. And thus one would conjecture that those three predictors might not be significant based on simultaneous confidence bands.

Although it is not clear what caused these differences, the scatterplots show that our estimates are at least reasonable. We also think penalized estimates are more stable. The residual plots shown in Figure 4 suggest also the fitting is good. The results are very similar when λ_1 and λ_2 are multiplied by a scalar factor in the range from 0.7 to 1.3, and K in the range from 5 to 7.

Finally, we note that in Figure 3, there is some crossing of the quantile curves for covariate values close to 0 and 1. Such crossing problems are well known in the quantile regression literature, and could be addressed by the algorithm proposed in, say, Chernozhukov, Fernández-Val, and Galichon (2010). This however requires that the quantile estimators be obtained on a fine grid of quantile levels. We do not pursue this idea further in this work.

4. CONCLUSION

In this article, we show that it is possible to select both zero components and linear components, as well as perform estimation, in a single step for additive quantile regression models. The difficulty of correctly specifying a partially linear additive model seems to be the main obstacle for the wide application of the semiparametric model, despite its many advantages. Thus, we believe that our work has made some progress in arguing for its usefulness and applicability.

Although we demonstrated that using SIC to select λ_1 and λ_2 resulted in consistent model selection, we do not have similar theoretical results for selecting K based on cross-validation. We note that theories for cross-validation were mainly developed for linear models as well as in the context of kernel regression and density estimation. In the context of quantile regression, or additive models, or nonparametric estimation based on B-splines, we cannot find similar results for cross-validation in the literature. Thus, we mainly rely on our simulation studies to demonstrate that the data-driven choice works well in practice. Alternatively, as suggested by a referee, one can select K based on cross-validation with least squares loss, for which theoretical guarantees are available (Li 1987; Li and Racine 2007). Whether this would work well in our model, especially for relatively large or small values of τ, remains to be seen.

After the first version of the article was submitted, we learned that model selection using two penalties to simultaneously determine the zero and linear components was studied in Zhang, Cheng, and Liu (2011), for mean regression (with quadratic loss function). Besides the difference that they used smoothing splines to approximate the nonparametric components, they did not use a penalty shrinking the second derivative of the component function. Instead their method is based on an explicit decomposition of a general nonlinear function into a linear part and a nonlinear part and then shrinking the nonlinear part to zero. Furthermore, they did not satisfactorily demonstrate the model selection consistency of their approach due to the difficulty in dealing with smoothing splines (they only proved consistency for the special case where the nonparametric components are periodic functions and conjectured that it is generally true). In terms of computation, using smoothing splines, the number of basis coefficients is proportional to the sample size, while for polynomial splines, the number of basis coefficients is proportional to K, which is typically chosen to be less than 10 in the literature. We will report our detailed study on mean regression based on polynomial splines in another article (Lian, Chen, and Yang 2011).

APPENDIX

In the proofs, C denotes a generic constant that might assume different values at different places. Let $b_0 = (b_{01}^T, \ldots, b_{0p}^T)^T$ be a pK dimensional vector that satisfies $\|m_{0j} - b_{0j}^T B_j\| = O(K^{-d})$, $1 \le j \le p_1$ and $m_{0j} = b_{0j}^T B_j$, $j > p_1$, and denote $R_i = \sum_{j=1}^p b_{0j}^T B_j(X_{ij}) - \sum_{j=1}^p m_{0j}(X_{ij})$. For convenience, let $a_0 = (\sqrt{K}\mu_0, b_0^T)^T$ be the parametric vector that includes the intercept (appropriately normalized), and similarly $a := (\sqrt{K}\mu, b^T)^T$. Let $\tilde{Z}_i = (1/\sqrt{K}, Z_i^T)^T$, $\tilde{\mathbf{Z}} = (\tilde{Z}_1, \ldots, \tilde{Z}_n)^T$. That is, $\tilde{\mathbf{Z}}$ is obtained by adding $(1/\sqrt{K}, \ldots, 1/\sqrt{K})^T$ as the first column of $\mathbf{Z}$. Denote $\psi(u) = \tau - I_{\{u \le 0\}}$.

We note that, based on well-known properties of B-spline, D_j has eigenvalues of order $1/K$, E_j is of rank $K-1$, and all its positive eigenvalues are of order $1/K$.

The main idea of our proofs is summarized as follows. To show the asymptotic results, we first show that since the tuning parameters λ_1 and λ_2 are sufficiently small, the convergence rate is basically same as the estimator without penalty. On the other hand, since we assume λ_1 and λ_2 cannot converge to zero too fast, if the correct model is not selected, we can construct an estimator that achieves a smaller value on the objective function which will lead to a contradiction. Finally, since we show that the correct model is selected with high probability, the asymptotic normality of the linear components follows from that of the oracle estimator.

Proof of Theorem 1. For any $a = (\sqrt{K}\mu, b^T)^T$ satisfying $\|a - a_0\| = L(K/\sqrt{n} + 1/K^{d-1/2} + (\lambda_1 + \lambda_2)\sqrt{K})$, we have

$$
\begin{aligned}
& nQ(a) - nQ(a_0) \\
&\quad = \sum_i \rho(Y_i - \tilde{Z}_i a) - \sum_i \rho(Y_i - \tilde{Z}_i a_0) \\
&\qquad + n \sum_j p_{\lambda_1}\left(\sqrt{b_j^T D_j b_j}\right) - n \sum_j p_{\lambda_1}\left(\sqrt{b_{0j}^T D_j b_{0j}}\right) \\
&\qquad + n \sum_j p_{\lambda_2}\left(\sqrt{b_j^T E_j b_j}\right) - n \sum_j p_{\lambda_2}\left(\sqrt{b_{0j}^T E_j b_{0j}}\right) \\
&\quad \ge \sum_i \rho(Y_i - \tilde{Z}_i a) - \sum_i \rho(Y_i - \tilde{Z}_i a_0) - Cn(\lambda_1 + \lambda_2)
\end{aligned}
$$

$$\begin{aligned}
&\times \sum_j \|b_j - b_{0j}\|/\sqrt{K}\\
&= \sum_i \rho(\epsilon_i - \tilde{Z}_i(a-a_0) - R_i) - \sum_i \rho(\epsilon_i - R_i)\\
&\quad - Cn(\lambda_1+\lambda_2)\sum_j \|b_j - b_{0j}\|/\sqrt{K}, \qquad \text{(A.1)}
\end{aligned}$$

where in the inequality above, we used $|p_\lambda(|s|) - p_\lambda(|t|)| \le \lambda|s-t|$ and that eigenvalues of D_j and E_j are of order $O(1/K)$.

Let $G_i = \rho(\epsilon_i - \tilde{Z}_i(a-a_0) - R_i) - \rho(\epsilon_i - R_i) + \tilde{Z}_i^T(a-a_0)\psi(\epsilon_i)$. By the same arguments used for lemma 3.2 in He and Shi (1994) [which was also used in lemma 6.3 in He and Shi (1996) and lemma 8.1 in Wei and He (2006)], we have

$$\sup_{\|a-a_0\|=L(K/\sqrt{n}+1/K^{d-1/2}+(\lambda_1+\lambda_2)\sqrt{K})} \left|\sum_i (G_i - E[G_i])\right| = o_p(K). \qquad \text{(A.2)}$$

[We note that He and Shi (1994) dealt with the case where ϵ_i is independent of X_i but the arguments go through without change under our heterogeneous setup.]

Furthermore,

$$\begin{aligned}
\sum_i E[G_i] &= \sum_i E\int_{R_i}^{\tilde{Z}_i^T(a-a_0)+R_i} (F_i(t) - F_i(0))dt\\
&= \sum_i f_i(0)\frac{1}{2}\{(\tilde{Z}_i^T(a-a_0)+R_i)^2 - R_i^2\}(1+o_p(1))\\
&= \left\{\frac{1}{2}(a-a_0)^T\tilde{\mathbf{Z}}^T\mathbf{f}\tilde{\mathbf{Z}}(a-a_0)\right.\\
&\quad \left. + \sum_i f_i(0)R_i\tilde{Z}_i^T(a-a_0)\right\}(1+o_p(1)).
\end{aligned}$$

Thus

$$\begin{aligned}
&\sum_i \rho(\epsilon_i - \tilde{Z}_i(a-a_0) - R_i) - \sum_i \rho(\epsilon_i - R_i)\\
&\ge -\sum_i \tilde{Z}_i^T(a-a_0)\psi(\epsilon_i) + o_p(K)\\
&\quad + \left\{\frac{1}{2}(a-a_0)^T\tilde{\mathbf{Z}}\mathbf{f}\tilde{\mathbf{Z}}(a-a_0)\right.\\
&\quad \left. + \sum_i f_i(0)R_i\tilde{Z}_i^T(a-a_0)\right\}(1+o_p(1))\\
&\ge -\sum_i \tilde{Z}_i^T(a-a_0)\psi(\epsilon_i) + \frac{c_f}{4}\|\tilde{\mathbf{Z}}(a-a_0)\|^2\\
&\quad - O_p(nK^{-2d}), \qquad \text{(A.3)}
\end{aligned}$$

where $c_f := \min\{f_1(0), \ldots, f_n(0)\} > 0$ and we used that $\sum_i f_i(0)^2R_i^2 = O(nK^{-2d})$, and the Cauchy–Schwartz inequality

$$\begin{aligned}
&\left|\sum_i f_i(0)R_i\tilde{Z}_i^T(a-a_0)\right|\\
&\le \left(\sum_i f_i^2(0)R_i^2\right)^{1/2}\left(\sum_i |\tilde{Z}_i^T(a-a_0)|^2\right)^{1/2}\\
&\le \frac{1}{c_f}\sum_i f_i^2(0)R_i^2 + \frac{c_f}{4}\|\tilde{\mathbf{Z}}(a-a_0)\|^2\\
&= O(nK^{-2d}) + \frac{c_f}{4}\|\tilde{\mathbf{Z}}(a-a_0)\|^2.
\end{aligned}$$

Again by the Cauchy–Schwartz inequality, the first term in (A.3) is bounded by

$$\begin{aligned}
&\left|\sum_i \tilde{Z}_i^T(a-a_0)\psi(\epsilon_i)\right|\\
&\le \|\tilde{\mathbf{Z}}(a-a_0)\| \cdot \|P_{\tilde{\mathbf{Z}}}\psi(\epsilon)\|\\
&\le \frac{c_f}{8}\|\tilde{\mathbf{Z}}(a-a_0)\|^2 + \frac{2}{c_f}\|P_{\tilde{\mathbf{Z}}}\psi(\epsilon)\|^2,
\end{aligned}$$

where $\psi(\epsilon) = (\psi(\epsilon_1), \ldots, \psi(\epsilon_n))^T$ and $P_{\tilde{\mathbf{Z}}} = \tilde{\mathbf{Z}}(\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}^T$ is a projection matrix. Since $\|\tilde{\mathbf{Z}}(a-a_0)\| = O_p(\sqrt{n/K}\|a-a_0\|)$ and $\|P_{\tilde{\mathbf{Z}}}\psi(\epsilon)\|^2 = O_p(tr(P_{\tilde{\mathbf{Z}}}P_{\tilde{\mathbf{Z}}}^T)) = O_p(K)$, we have

$$-\sum_i \tilde{\mathbf{Z}}_i^T(a-a_0)\psi(\epsilon_i) \ge -O_p(K) - \frac{c_f}{8}\|\tilde{\mathbf{Z}}(a-a_0)\|^2.$$

Then (A.1) can be continued as

$$\begin{aligned}
nQ(a) - nQ(a_0) &\ge -|O_p(K + n/K^{2d})| - \frac{c_f}{8}\|\tilde{\mathbf{Z}}(a-a_0)\|^2\\
&\quad + \frac{c_f}{4}\|\tilde{\mathbf{Z}}(a-a_0)\|^2 - Cn(\lambda_1+\lambda_2)\sum_j \|b_j - b_{0j}\|/\sqrt{K}.
\end{aligned}$$

Since $\|\tilde{\mathbf{Z}}(a-a_0)\|^2 \sim (n/K)\|a-a_0\|^2$, the above would be positive if $\|a-a_0\| = L(K/\sqrt{n} + 1/K^{d-1/2} + (\lambda_1+\lambda_2)\sqrt{K})$ for L sufficiently large.

The above convergence rate can be further improved to $\|\hat{a} - a_0\|^2 = O_p(K^2/n + 1/K^{2d-1})$ as follows. First note that since the model is fixed as $n \to \infty$, we can find a constant $C > 0$ such that $b_{0j}^T D_j b_{0j} > C$ when $j \le s$ and $b_{0j}^T E_j b_{0j} > C$ when $j \le p_1$. Since $\|\hat{b}_j - b_{0j}\|^2 = o_p(K)$ by the proved convergence rates above and that $\lambda_1, \lambda_2 = o(1)$, we have

$$P\left(p_{\lambda_1}\left(\sqrt{b_{0j}^T D_j b_{0j}}\right) = p_{\lambda_1}\left(\sqrt{\hat{b}_j^T D_j \hat{b}_j}\right)\right) \to 1$$

if $j \le s$. Similarly

$$P\left(p_{\lambda_2}\left(\sqrt{b_{0j}^T E_j b_{0j}}\right) = p_{\lambda_2}\left(\sqrt{\hat{b}_j^T E_j \hat{b}_j}\right)\right) \to 1$$

if $j \le p_1$. These facts imply that

$$n\sum_{j=1}^p p_{\lambda_1}\left(\sqrt{\hat{b}_j^T D_j \hat{b}_j}\right) - n\sum_{j=1}^p p_{\lambda_1}\left(\sqrt{b_{0j}^T D_j b_{0j}}\right) \ge 0,$$

and

$$n\sum_{j=1}^p p_{\lambda_2}\left(\sqrt{\hat{b}_j^T E_j \hat{b}_j}\right) - n\sum_{j=1}^p p_{\lambda_2}\left(\sqrt{b_{0j}^T E_j b_{0j}}\right) \ge 0,$$

with probability tending to 1. Removing the regularizing terms in (A.1) and following the same reasoning as before, the rates are improved to $\|\hat{a} - a_0\|^2 = O_p(K^2/n + 1/K^{2d-1})$.

The rates of convergence for $\|\hat{b}_j - b_{0j}\|^2$ immediately imply the rates for $\|\hat{m}_j - m_{0j}\|$ by the property (De Boor 2001)

$$C_1 K \big\|\hat{b}_j^T B_j - b_{0j}^T B_j\big\|^2 \le \big\|\hat{b}_j - b_{0j}\big\|^2 \le C_2 K \big\|\hat{b}_j^T B_j - b_{0j}^T B_j\big\|^2$$

for some constants $C_1, C_2 > 0$. □

Proof of Theorem 2. We only show part (b) as an illustration and part (a) is similar. Suppose for some $p_1 < j \le s$, $\hat{b}_j^T B_j$ does not represent a linear function. Define $\hat{b}^*$ to be the same as $\hat{b}$ except that $\hat{b}_j$ is replaced by its projection onto the subspace $\{b_j : b_j^T B_j$ represents a linear function$\}$. Similarly as before, we let $\hat{a}^* = (\sqrt{K}\hat{\mu}, \hat{b}^{*T})^T$. We have

$$\begin{aligned} &nQ(\hat{a}) - nQ(\hat{a}^*) \\ &\quad = \rho(Y_i - \tilde{Z}_i\hat{a}) - \rho(Y_i - \tilde{Z}_i\hat{a}^*) + np_{\lambda_1}\Big(\sqrt{\hat{b}_j^T D_j \hat{b}_j}\Big) \\ &\quad - np_{\lambda_1}\Big(\sqrt{\hat{b}_j^{*T} D_j \hat{b}_j^*}\Big) + np_{\lambda_2}\Big(\sqrt{\hat{b}_j^T E_j \hat{b}_j}\Big) - np_{\lambda_2}\Big(\sqrt{\hat{b}_j^{*T} E_j \hat{b}_j^*}\Big). \end{aligned}$$

As in the proof of Theorem 1, we have

$$P\Big(p_{\lambda_1}\Big(\sqrt{\hat{b}_j^T D_j \hat{b}_j}\Big) = p_{\lambda_1}\Big(\sqrt{\hat{b}_j^{*T} D_j \hat{b}_j^*}\Big)\Big) \to 1$$

and thus with probability approaching 1 (since $\hat{b}_j^{*T} E_j \hat{b}_j^* = 0$),

$$\begin{aligned} nQ(\hat{a}) - nQ(\hat{a}^*) = \rho(Y_i - \tilde{Z}_i\hat{a}) - \rho(Y_i - \tilde{Z}_i\hat{a}^*) \\ + np_{\lambda_2}\Big(\sqrt{\hat{b}_j^T E_j \hat{b}_j}\Big). \end{aligned} \tag{A.4}$$

Let $\tilde{G}_i = \rho(Y_i - \tilde{Z}_i\hat{a}) - \rho(Y_i - \tilde{Z}_i\hat{a}^*) + \tilde{Z}_i^T(\hat{a} - \hat{a}^*)\psi(\epsilon_i)$, similar to (A.2), we have $|\sum_i(\tilde{G}_i - E[\tilde{G}_i])| = o_P(K)$, and furthermore,

$$\begin{aligned} &\Big|\sum_i E[\tilde{G}_i]\Big| \\ &= \Big|\sum_i E\int_{R_i+\tilde{Z}_i(\hat{a}^*-a_0)}^{-\tilde{Z}_i(\hat{a}^*-\hat{a})+R_i+\tilde{Z}_i(\hat{a}^*-a_0)} (F_i(t) - F_i(0))dt\Big| \\ &= \Big|\sum_i \frac{1}{2} f_i(0)\big[\|\tilde{Z}_i(\hat{a} - \hat{a}^*)\|^2 - 2\big(\tilde{Z}_i^T(\hat{a}^* - a_0) + R_i\big) \\ &\quad \times \big(\tilde{Z}_i^T(\hat{a}^* - \hat{a})\big)\big]\Big|(1 + o_p(1)) \\ &= O_p\big(\sqrt{n}\|\hat{a} - \hat{a}^*\|\big), \end{aligned}$$

using that $\|\hat{a} - \hat{a}^*\| = \|\hat{b}_j - \hat{b}_j^*\| \le \|\hat{b}_j - b_{0j}\| + \|\hat{b}_j^* - b_{0j}\| \le 2\|\hat{b}_j - \hat{b}_{0j}\| = O(K/\sqrt{n})$. Also similar to the proof of Theorem 1, we have $|\sum_i \tilde{Z}_i^T(\hat{a} - \hat{a}^*)\psi(\epsilon_i)| \le \|P_{\tilde{\mathbf{Z}}}\psi(\epsilon)\| \cdot \|\tilde{\mathbf{Z}}(\hat{a} - \hat{a}^*)\| = O_p(\sqrt{n}\|\hat{a} - \hat{a}^*\|)$ and thus

$$nQ(\hat{a}) - nQ(\hat{a}^*) = O_p(\sqrt{n}\|\hat{a} - \hat{a}^*\|) + np_{\lambda_2}\Big(\sqrt{\hat{b}_j^T E_j \hat{b}_j}\Big).$$

On the other hand, since

$$\begin{aligned} \sqrt{\hat{b}_j^T E_j \hat{b}_j} &= \sqrt{(\hat{b}_j - b_{0j})^T E_j(\hat{b}_j - b_{0j})} = O_p((K/n)^{1/2}) \\ &= o(\lambda_2) \text{ by } \|\hat{b}_j - b_{0j}\| = O(K/\sqrt{n}), \end{aligned}$$

we have

$$p_{\lambda_2}\Big(\sqrt{\hat{b}_j^T E_j \hat{b}_j}\Big) = \lambda_2\sqrt{\hat{b}_j^T E_j \hat{b}_j}, \text{ with probability tending to 1,} \tag{A.5}$$

by the definition of the SCAD penalty function. What is left is only to show $\|\hat{a} - \hat{a}^*\| = \|\hat{b}_j - \hat{b}_j^*\| = O_p(\sqrt{K\hat{b}_j^T E_j \hat{b}_j})$. In fact, if this is true, $np_{\lambda_2}(\sqrt{\hat{b}_j^T E_j \hat{b}_j}) = n\lambda_2\sqrt{\hat{b}_j^T E_j \hat{b}_j}$ will be asymptotically larger than $\sqrt{n}\|\hat{a} - \hat{a}^*\|$, making $nQ(\hat{a}) - nQ(\hat{a}^*) > 0$.

To show $\|\hat{b}_j - \hat{b}_j^*\| = O_p(\sqrt{K\hat{b}_j^T E_j \hat{b}_j})$, we first note that $\hat{b}_j^T E_j \hat{b}_j = (\hat{b}_j - \hat{b}_j^*)^T E_j(\hat{b}_j - \hat{b}_j^*)$ since $\hat{b}_j^{*T} E_j \hat{b}_j^* = 0$. Furthermore, since $\hat{b}_j^*$ is the projection of $\hat{b}_j$ onto $\{b_j : b_j^T E_j b_j = 0\}$, $\hat{b}_j - \hat{b}_j^*$ is orthogonal to this space. Besides, the space $\{b_j : b_j^T E_j b_j = 0\}$ is just the eigenspace of E_j associated with the zero eigenvalue. Thus by the characterization of eigenvalues in terms of Rayleigh quotient, $(\hat{b}_j - \hat{b}_j^*)^T E_j(\hat{b}_j - \hat{b}_j^*)/\|\hat{b}_j - \hat{b}_j^*\|^2$ lies between the minimum and the maximum positive eigenvalues of E_j, which is of order $1/K$. □

Proof of Theorem 3. We note that because of Theorem 2, we only need to consider a correctly specified partially linear additive model without regularization terms. This is very similar to partially linear varying-coefficient models studied in Wang, Zhu, and Zhou (2009) and the arguments used there can be followed line by line here, showing the asymptotic normality of the slope parameter. The reason is that in their arguments, for example, they only used the properties of covariate matrices, such as that eigenvalues of $\tilde{\mathbf{Z}}^T\tilde{\mathbf{Z}}$ are of order $O_p(n/K)$, which are also true here for additive models. □

Proof of Theorem 4. For any regularization parameters $\lambda = (\lambda_1, \lambda_2)$, we denote the corresponding minimizer of (3) by $\hat{a}_\lambda = (\sqrt{K}\hat{\mu}_\lambda, \hat{b}_\lambda)$ and denote by $\hat{a} = (\sqrt{K}\hat{\mu}, \hat{b})$ the minimizer when the optimal sequence of regularization parameters is chosen such that $\hat{b}$ represents the correct model with optimal convergence rates. There are four separate cases to consider.

Case 1, $\hat{b}_{\lambda j}^T B_j$ represents a linear component for some $j \le p_1$. Similar to the calculations performed in the proof of Theorems 1 and 2 [see Equations (A.2), (A.3) and the arguments following (A.4)], we have

$$\begin{aligned} &\frac{1}{n}\sum_i \rho(Y_i - \tilde{Z}_i\hat{a}_\lambda) - \frac{1}{n}\sum_i \rho(Y_i - \tilde{Z}_i\hat{a}) \\ &\qquad \ge -\big|O_p(1/K^{2d} + K/n)\big| + C\|\tilde{\mathbf{Z}}(\hat{a}_\lambda - \hat{a})\|^2/n. \end{aligned}$$

Since the true m_j is not linear and $\hat{b}_j$ is consistent in model selection, $\|\hat{a} - \hat{a}_\lambda\|^2/K$ is bounded away from zero and thus $\|\tilde{\mathbf{Z}}(\hat{a} - \hat{a}_\lambda)\|^2/n$ is at least of order $O_p(1)$ (i.e., bounded away from zero). Note that this lower bound is uniform for all λ in Case 1. Thus [noting $\frac{1}{n}\sum_i \rho(Y_i - \tilde{Z}_i\hat{a})$ is bounded away from zero]

$$\begin{aligned} &\log\Big\{\frac{1}{n}\sum_i \rho(Y_i - \tilde{Z}_i\hat{a}_\lambda)\Big\} - \log\Big\{\frac{1}{n}\sum_i \rho(Y_i - \tilde{Z}_i\hat{a})\Big\} \\ &\qquad > (C_1 K + C_2)\frac{\log n}{2n}, \end{aligned}$$

for any $0 \leq C_1, C_2 \leq p$, with probability tending to 1 and the SIC cannot select such λ.

Case 2, $\hat{b}_{\lambda j}$ is zero for some $1 \leq j \leq s$. The above case, as well as Case 1, results in underfitted models and the proof is very similar and therefore omitted.

Case 3, $\hat{b}_{\lambda j}^T B_j$ represents a nonlinear component for some $p_1 < j \leq s$. Here when considering Case 3, we implicitly exclude all previous cases, that is, no underfitting occurs. We define $\hat{a}^*$ as the unregularized estimator that minimizes (3) without the two penalty terms, but constrained to represent the same model as $\hat{a}_\lambda$. By definition, we immediately have $\sum_i \rho(Y_i - \tilde{Z}_i\hat{a}_\lambda) \geq \sum_i \rho(Y_i - \tilde{Z}_i\hat{a}^*)$ and then (by the same arguments concerning G_i and $\tilde{G}_i$ above)

$$\begin{aligned}
&\frac{1}{n}\sum_i \rho(Y_i - \tilde{Z}_i\hat{a}_\lambda) - \frac{1}{n}\sum_i \rho(Y_i - \tilde{Z}_i\hat{a}) \\
&\geq \frac{1}{n}\sum_i \rho(Y_i - \tilde{Z}_i\hat{a}^*) - \frac{1}{n}\sum_i \rho(Y_i - \tilde{Z}_i\hat{a}) \\
&\geq -|O_p(1/K^{2d} + K/n)|,
\end{aligned}$$

and thus $\log(\sum_i \rho(Y_i - \tilde{Z}_i\hat{a}_\lambda)/n) - \log(\sum_i \rho(Y_i - \tilde{Z}_i\hat{a})/n) \geq -|O_p(1/K^{2d} + K/n)|$ and

$$\begin{aligned}
&\log\left\{\frac{1}{n}\sum_i \rho(Y_i - \tilde{Z}_i\hat{a}_\lambda)\right\} + \frac{K\log(n)}{2n} \\
&> \log\left\{\frac{1}{n}\sum_i \rho(Y_i - \tilde{Z}_i\hat{a})\right\} + p\frac{\log n}{2n},
\end{aligned}$$

with probability tending to 1 (also uniformly for all such λ).

Case 4, $\hat{b}_{\lambda j}$ is nonzero for some $j \geq s$. The above case is similar to Case 3 and thus the proof is omitted. □

ACKNOWLEDGMENTS

The author thanks Professor Hirano, an associate editor, and two referees for their careful reading and constructive comments that greatly improved the manuscript.

[Received March 2011. Revised March 2012.]

REFERENCES

Belloni, A., and Chernozhukov, V. (2011), "L1-Penalized Quantile Regression in High-Dimensional Sparse Models," *The Annals of Statistics*, 39(1), 82–130. [338]

Chaudhuri, P. (1991), "Nonparametric Estimates of Regression Quantiles and Their Local Bahadur Representation," *The Annals of Statistics*, 19(2), 760–777. [337]

Chernozhukov, V., Fernández-Val, I., and Galichon, A. (2010), "Quantile and Probability Curves Without Crossing," *Econometrica*, 78(3), 1093–1125. [347]

De Boor, C. (2001), *A Practical Guide to Splines* (rev. ed.), New York: Springer-Verlag. [349]

De Gooijer, J. G., and Zerom, D. (2003), "On Additive Conditional Quantiles With High-Dimensional Covariates," *Journal of the American Statistical Association*, 98(461), 135–146. [337]

Fan, J. Q., and Li, R. Z. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96(456), 1348–1360. [338,339,339]

Härdle, W., Mammen, E., and Müller, M. (1998), "Testing Parametric Versus Semiparametric Modeling in Generalized Linear Models," *Journal of the American Statistical Association*, 93(444), 1461–1474. [337]

Hastie, T., Tibshirani, R. (1990), "Monographs on Statistics and Applied Probability," *Generalized Additive Models* (1st ed.), London: Chapman and Hall. [337]

He, X., and Shi, P. (1994), "Convergence Rate of B-Spline Estimators of Nonparametric Conditional Quantile Functions," *Journal of Nonparametric Statistics*, 3(3), 299–308. [348]

——— (1996), "Bivariate Tensor-Product B-Splines in a Partly Linear Model," *Journal of Multivariate Analysis*, 58(2), 162–181. [348]

Horowitz, J. L., and Lee, S. (2005), "Nonparametric Estimation of an Additive Quantile Regression Model," *Journal of the American Statistical Association*, 100(472), 1238–1249. [337,344,346]

Huang, J., Horowitz, J. L., and Wei, F. (2010), "Variable Selection in Nonparametric Additive Models," *The Annals of Statistics*, 38(4), 2282–2313. [337]

Hunter, D., and Li, R. (2005), "Variable Selection Using MM Algorithms," *The Annals of Statistics*, 33(4), 1617–1642. [339]

Hunter, D. R., and Lange, K. (2000), "Quantile Regression via an MM Algorithm," *Journal of Computational and Graphical Statistics*, 9(1), 60–77. [339]

——— (2004), "A Tutorial on MM Algorithms," *The American Statistician*, 58(1), 30–37. [339]

Kato, K. (2011), "Group Lasso for High Dimensional Sparse Quantile Regression Models," Arxiv preprint arXiv:1103.1458. [338]

Kim, M. (2007), "Quantile Regression With Varying Coefficients," *The Annals of Statistics*, 35, 92–108. [337]

Koenker, R. (2005), "Econometric Society Monographs," in *Quantile Regression*, Cambridge: Cambridge University Press. [337]

Koenker, R., and Bassett Jr, G. (1978), "Regression Quantiles," *Econometrica: Journal of the Econometric Society*, 46(1), 33–50. [337]

Leeb, H., and Pötscher, B. (2006), "Performance Limits for Estimators of the Risk or Distribution of Shrinkage-Type Estimators, and Some General Lower Risk-Bound Results," *Econometric Theory*, 22(01), 69–97. [340]

——— (2008), "Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator," *Journal of Econometrics*, 142(1), 201–211. [340]

Li, K. (1987), "Asymptotic Optimality for C_p, C_l, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15(3), 958–975. [347]

Li, Q., and Racine, J. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton NJ: Princeton University Press. [347]

Lian, H., and Chen, X., and Yang, J. (2011), "Identification of Partially Linear Structure in Additive Models With an Application to Gene Expression Prediction From Sequences," Biometrics. Available at *http://onlinelibrary.wiley.com/doi/10.1111/j.1541-0420.2011.01672.x/abstract* [347]

Opsomer, J. D., and Ruppert, D. (1999), "A Root-n Consistent Backfitting Estimator for Semiparametric Additive Modeling," *Journal of Computational and Graphical Statistics*, 8(4), 715–732. [337]

Pötscher, B., and Schneider, U. (2009), "On the Distribution of the Adaptive LASSO Estimator," *Journal of Statistical Planning and Inference*, 139(8), 2775–2790. [340]

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis* (2nd ed., Springer Series in Statistics), New York: Springer. [338]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58(1), 267–288. [338]

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia PA: Society for Industrial and Applied Mathematics. [338]

Wang, H., Li, R., and Tsai, C. L. (2007), "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," *Biometrika*, 94(3), 553–568. [340]

Wang, H. J., Zhu, Z., and Zhou, J. (2009), "Quantile Regression in Partially Linear Varying Coefficient Models," *The Annals of Statistics*, 37(6B), 3841–3866. [337,340,349]

Wang, H. S., and Xia, Y. C. (2009), "Shrinkage Estimation of the Varying Coefficient Model," *Journal of the American Statistical Association*, 104(486), 747–757. [338]

Wei, Y., and He, X. (2006), "Conditional Growth Charts," *The Annals of Statistics*, 34(5), 2069–2097. [348]

Wu, Y., and Liu, Y. (2009), "Variable Selection in Quantile Regression," *Statistica Sinica*, 19(2), 801–817. [339]

Yafeh, Y., and Yosha, O. (2003), "Large Shareholders and Banks: Who Monitors and How?," *The Economic Journal*, 113(484), 128–146. [344,346]

Zhang, C. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38(2), 894–942. [339]

Zhang, H. H., Cheng, G., and Liu, Y. (2011), "Linear of Nonlinear? Automatic Structure Discovery for Partially Linear Models," *Journal of the American Statistical Association*, 106(495), 1099–1112. [347]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101(476), 1418–1429. [339]