PARTIALLY LINEAR ADDITIVE QUANTILE REGRESSION IN ULTRA-HIGH DIMENSION

BY BEN SHERWOOD AND LAN WANG[1]

Johns Hopkins University and University of Minnesota

We consider a flexible semiparametric quantile regression model for analyzing high dimensional heterogeneous data. This model has several appealing features: (1) By considering different conditional quantiles, we may obtain a more complete picture of the conditional distribution of a response variable given high dimensional covariates. (2) The sparsity level is allowed to be different at different quantile levels. (3) The partially linear additive structure accommodates nonlinearity and circumvents the curse of dimensionality. (4) It is naturally robust to heavy-tailed distributions. In this paper, we approximate the nonlinear components using B-spline basis functions. We first study estimation under this model when the nonzero components are known in advance and the number of covariates in the linear part diverges. We then investigate a nonconvex penalized estimator for simultaneous variable selection and estimation. We derive its oracle property for a general class of nonconvex penalty functions in the presence of ultra-high dimensional covariates under relaxed conditions. To tackle the challenges of nonsmooth loss function, nonconvex penalty function and the presence of nonlinear components, we combine a recently developed convex-differencing method with modern empirical process techniques. Monte Carlo simulations and an application to a microarray study demonstrate the effectiveness of the proposed method. We also discuss how the method for a single quantile of interest can be extended to simultaneous variable selection and estimation at multiple quantiles.

1. Introduction. In this article, we study a flexible partially linear additive quantile regression model for analyzing high dimensional data. For the ith subject, we observe $\{Y_i, \mathbf{x}_i, \mathbf{z}_i\}$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip_n})'$ is a p_n-dimensional vector of covariates and $\mathbf{z}_i = (z_{i1}, \ldots, z_{id})'$ is a d-dimensional vector of covariates, $i = 1, \ldots, n$. The τth $(0 < \tau < 1)$ conditional quantile of Y_i given $\mathbf{x}_i, \mathbf{z}_i$ is defined as $Q_{Y_i|\mathbf{x}_i, \mathbf{z}_i}(\tau) = \inf\{t : F(t|\mathbf{x}_i, \mathbf{z}_i) \geq \tau\}$, where $F(\cdot|\mathbf{x}_i, \mathbf{z}_i)$ is the conditional distribution function of Y_i given $\mathbf{x}_i$ and $\mathbf{z}_i$. The case $\tau = 1/2$ corresponds to the conditional median. We consider the following semiparametric model for the conditional quantile function

$$(1.1) \qquad Q_{Y_i|\mathbf{x}_i, \mathbf{z}_i}(\tau) = \mathbf{x}_i'\boldsymbol{\beta}_0 + g_0(\mathbf{z}_i),$$

where $g_0(\mathbf{z}_i) = g_{00} + \sum_{j=1}^{d} g_{0j}(z_{ij})$, with $g_{00} \in \mathcal{R}$. It is assumed that g_{0j} satisfy $E(g_{0j}(z_{ij})) = 0$ for identification purposes. Let $\varepsilon_i = Y_i - Q_{Y_i|\mathbf{x}_i, \mathbf{z}_i}(\tau)$, then ε_i satisfies $P(\varepsilon_i \le 0|\mathbf{x}_i, \mathbf{z}_i) = \tau$ and we may also write $Y_i = \mathbf{x}_i' \boldsymbol{\beta}_0 + g_0(\mathbf{z}_i) + \varepsilon_i$. In the rest of the paper, we will drop the dependence on τ in the notation for simplicity.

Modeling conditional quantiles in high dimension is of significant importance for several reasons. First, it is well recognized that high dimensional data are often heterogeneous. How the covariate influence the center of the conditional distribution can be very different from how they influence the tails. As a result, focusing on the conditional mean function alone can be misleading. By estimating conditional quantiles at different quantile levels, we are able to gain a more complete picture of the relationship between the covariates and the response variable. Second, in the high dimensional setting, the quantile regression framework also allows a more realistic interpretation of the sparsity of the covariate effects, which we refer to as quantile-adaptive sparsity. That is, we assume a small subset of covariates influence the conditional distribution. However, when we estimate different conditional quantiles, we allow the subsets of active covariates to be different [Wang, Wu and Li (2012); He, Wang and Hong (2013)]. Furthermore, the conditional quantiles are often of direct interest to the researchers. For example, for the birth weight data we analyzed in Section 5, low birth weight, which corresponds to the low tail of the conditional distribution, is of direct interest to the doctors. Another advantage of quantile regression is that it is naturally robust to outlier contamination associated with heavy-tailed errors. For high dimensional data, identifying outliers can be difficult. The robustness of quantile regression provides a certain degree of protection.

Linear quantile regression with high dimensional covariates was investigated by Belloni and Chernozhukov [(2011), Lasso penalty] and Wang, Wu and Li [(2012), nonconvex penalty]. The partially linear additive structure we consider in this paper is useful for incorporating nonlinearity in the model while circumventing the curse of dimensionality. We are interested in the case p_n is of a similar order of n or much larger than n. For applications in microarray data analysis, the vector $\mathbf{x}_i$ often contains the measurements on thousands of genes, while the vector $\mathbf{z}_i$ contains the measurements of clinical or environment variables, such as age and weight. For example, in the birth weight example of Section 5, mother's age is modeled nonparametrically as exploratory analysis reveals a possible nonlinear effect. In general, model specification can be challenging in high dimension; see Section 7 for some further discussion.

We approximate the nonparametric components using B-spline basis functions, which are computationally convenient and often accurate. First, we study the asymptotic theory of estimating the model (1.1) when p_n diverges. In our setting, this corresponds to the oracle model, that is, the one we obtain if we know which covariates are important in advance. This is along the line of the work of Welsh (1989), Bai and Wu (1994) and He and Shao (2000) for M-regression with diverging number of parameters and possibly nonsmooth objective functions,

which, however, were restricted to linear regression. Lam and Fan (2008) derived the asymptotic theory of profile kernel estimator for general semiparametric models with diverging number of parameter while assuming a smooth quasi-likelihood function. Second, we propose a nonconvex penalized regression estimator when p_n is of an exponential order of n and the model has a sparse structure. For a general class of nonsmooth penalty functions, including the popular SCAD [Fan and Li (2001)] and MCP [Zhang (2010)] penalty, we derive the oracle property of the proposed estimator under relaxed conditions. An interesting finding is that solving the nonconvex penalized estimator can be achieved via solving a series of weighted quantile regression problems, which can be conveniently implemented using existing software packages.

Deriving the asymptotic properties of the penalized estimator is very challenging as we need to simultaneously deal with the nonsmooth loss function, the nonconvex penalty function, approximation of nonlinear functions and very high dimensionality. To tackle these challenges, we combine a recently developed convex-differencing method with modern empirical process techniques. The method relies on a representation of the penalized loss function as the difference of two convex functions, which leads to a sufficient local optimality condition [Tao and An (1997), Wang, Wu and Li (2012)]. Empirical process techniques are introduced to derive various error bounds associated with the nonsmooth objective function which contains both high dimensional linear covariates and approximations of nonlinear components. It is worth pointing out that our approach is different from what was used in the recent literature for studying the theory of high dimensional semiparametric mean regression and is able to considerably weaken the conditions required in the literature. In particular, we do not need moment conditions for the random error and allow it to depend on the covariates.

Existing work on penalized semiparametric regression has been largely limited to mean regression with fixed p; see, for example, Bunea (2004), Liang and Li (2009), Wang and Xia (2009), Liu, Wang and Liang (2011), Kai, Li and Zou (2011) and Wang et al. (2011). Important progress in the high dimensional p setting has been recently made by Xie and Huang [(2009), still assumes $p < n$] for partially linear regression, Huang, Horowitz and Wei (2010) for additive models, Li, Xue and Lian [(2011), $p = o(n)$] for semivarying coefficient models, among others. When p is fixed, the semiparametric quantile regression model was considered by He and Shi (1996), He, Zhu and Fung (2002), Wang, Zhu and Zhou (2009), among others. Tang et al. (2013) considered a two-step procedure for a nonparametric varying coefficients quantile regression model with a diverging number of nonparametric functional coefficients. They required two separate tuning parameters and quite complex design conditions.

The rest of this article is organized as follows. In Section 2, we present the partially linear additive quantile regression model and discuss the properties of the oracle estimator. In Section 3, we present a nonconvex penalized method for simultaneous variable selection and estimation and derive its oracle property. In Section 4, we assess the performance of the proposed penalized estimator via Monte

Carlo simulations. We analyze a birth weight data set while accounting for gene expression measurements in Section 5. In Section 6, we consider an extension to simultaneous estimation and variable selection at multiple quantiles. Section 7 concludes the paper with a discussion of related issues. The proofs are given in the Appendix. Some of the technical details and additional numerical results are provided in online supplementary material [Sherwood and Wang (2015)].

2. Partially linear additive quantile regression with diverging number of parameters.
For high dimensional inference, it is often assumed that the vector of coefficients $\boldsymbol{\beta}_0 = (\beta_{01}, \beta_{02}, \ldots, \beta_{0p_n})'$ in model (1.1) is sparse, that is, most of its components are zero. Let $A = \{1 \leq j \leq p_n : \beta_{0j} \neq 0\}$ be the index set of nonzero coefficients and $q_n = |A|$ be the cardinality of A. The set A is unknown and will be estimated. Without loss of generality, we assume that the first q_n components of $\boldsymbol{\beta}_0$ are nonzero and the remaining $p_n - q_n$ components are zero. Hence, we can write $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}', \mathbf{0}_{p_n-q_n}')'$, where $\mathbf{0}_{p_n-q_n}$ denotes the $(p_n - q_n)$-vector of zeros. Let X be the $n \times p_n$ matrix of linear covariates and write it as $X = (X_1, \ldots, X_{p_n})$. Let X_A be the submatrix consisting of the first q_n columns of X corresponding to the active covariates. For technical simplicity, we assume x_i is centered to have mean zero; and $z_{ij} \in [0, 1]$, $\forall i, j$.

2.1. *Oracle estimator.* We first study the estimator we would obtain when the index set A is known in advance, which we refer to as the oracle estimator. Our asymptotic framework allows q_n, the size of A, to increase with n. This resonates with the perspective that a more complex statistical model can be fit when more data are collected.

We use a linear combination of B-spline basis functions to approximate the unknown nonlinear functions $g_0(\cdot)$. To introduce the B-spline functions, we start with two definitions.

DEFINITION. Let $r \equiv m + v$, where m is a positive integer and $v \in (0, 1]$. Define $\mathcal{H}_r$ as the collection of functions $h(\cdot)$ on $[0, 1]$ whose mth derivative $h^{(m)}(\cdot)$ satisfies the Hölder condition of order v. That is, for any $h(\cdot) \in \mathcal{H}_r$, there exists some positive constant C such that

$$(2.1) \qquad |h^{(m)}(z') - h^{(m)}(z)| \leq C|z' - z|^v \qquad \forall 0 \leq z', z \leq 1.$$

Assume for some $r \geq 1.5$, the nonparametric component $g_{0k}(\cdot) \in \mathcal{H}_r$. Let $\boldsymbol{\pi}(t) = (b_1(t), \ldots, b_{k_n+l+1}(t))'$ denote a vector of normalized B-spline basis functions of order $l + 1$ with k_n quasi-uniform internal knots on $[0, 1]$. Then $g_{0k}(\cdot)$ can be approximated using a linear combination of B-spline basis functions in $\boldsymbol{\Pi}(\mathbf{z}_i) = (1, \boldsymbol{\pi}(z_{i1})', \ldots, \boldsymbol{\pi}(z_{id})')'$. We refer to Schumaker (1981) for details of the B-spline construction, and the result that there exists $\boldsymbol{\xi}_0 \in \mathcal{R}^{L_n}$, where $L_n = d(k_n + l + 1) + 1$, such that $\sup_{\mathbf{z}_i} |\boldsymbol{\Pi}(\mathbf{z}_i)'\boldsymbol{\xi}_0 - g_0(\mathbf{z}_i)| = O(k_n^{-r})$. For ease of notation and simplicity of proofs, we use the same number of basis functions for all

nonlinear components in model (1.1). In practice, such restrictions are not necessary.

Now consider quantile regression with the oracle information that the last $(p_n - q_n)$ elements of $\boldsymbol{\beta}_0$ are all zero. Let

$$(2.2) \qquad (\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\xi}}) = \underset{(\boldsymbol{\beta}_1, \boldsymbol{\xi})}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau (Y_i - \mathbf{x}'_{A_i} \boldsymbol{\beta}_1 - \boldsymbol{\Pi}(\mathbf{z}_i)' \boldsymbol{\xi}),$$

where $\rho_\tau(u) = u(\tau - I(u < 0))$ is the quantile loss function and $\mathbf{x}'_{A_1}, \ldots, \mathbf{x}'_{A_n}$ denote the row vectors of X_A. The oracle estimator for $\boldsymbol{\beta}_0$ is $(\hat{\boldsymbol{\beta}}'_1, \mathbf{0}'_{p_n - q_n})'$. Write $\hat{\boldsymbol{\xi}} = (\hat{\xi}_0, \hat{\boldsymbol{\xi}}'_1, \ldots, \hat{\boldsymbol{\xi}}'_d)'$ where $\hat{\xi}_0 \in \mathcal{R}$ and $\hat{\boldsymbol{\xi}}_j \in \mathcal{R}^{k_n + l + 1}$, $j = 1 \ldots, d$. The estimator for the nonparametric function g_{0j} is

$$\hat{g}_j(z_{ij}) = \boldsymbol{\pi}(z_{ij})' \hat{\boldsymbol{\xi}}_j - n^{-1} \sum_{i=1}^{n} \boldsymbol{\pi}(z_{ij})' \hat{\boldsymbol{\xi}}_j,$$

for $j = 1, \ldots, d$; for g_{00} is $\hat{g}_0 = \hat{\xi}_0 + n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{d} \boldsymbol{\pi}(z_{ij})' \hat{\boldsymbol{\xi}}_j$. The centering of $\hat{g}_j$ is the sample analog of the identifiability condition $E[g_{0j}(\mathbf{z}_i)] = 0$. The estimator of $g_0(\mathbf{z}_i)$ is $\hat{g}(\mathbf{z}_i) = \hat{g}_0 + \sum_{j=1}^{d} \hat{g}_j(z_{ij})$.

2.2. *Asymptotic properties.* We next present the asymptotic properties of the oracle estimators as q_n diverges.

DEFINITION. Given $\mathbf{z} = (z_1, \ldots, z_d)'$, the function $g(\mathbf{z})$ is said to belong to the class of functions $\mathcal{G}$ if it has the representation $g(\mathbf{z}) = \alpha + \sum_{k=1}^{d} g_k(\mathbf{z}_k)$, $\alpha \in \mathcal{R}$, $g_k \in \mathcal{H}_r$ and $E[g_k(\mathbf{z}_k)] = 0$.

Let

$$h_j^*(\cdot) = \underset{h_j(\cdot) \in \mathcal{G}}{\arg\inf} \sum_{i=1}^{n} E[f_i(0)(x_{ij} - h_j(\mathbf{z}_i))^2],$$

where $f_i(\cdot)$ is the probability density function of ε_i given $(\mathbf{x}_i, \mathbf{z}_i)$. Let $m_j(\mathbf{z}) = E[x_{ij}|\mathbf{z}_i = \mathbf{z}]$, then it can be shown that $h_j^*(\cdot)$ is the weighted projection of $m_j(\cdot)$ into $\mathcal{G}$ under the L_2 norm, where the weights $f_i(0)$ are included to account for the possibly heterogeneous errors. Furthermore, let $x_{A_{ij}}$ be the (i, j)th element of X_A. Define $\delta_{ij} \equiv x_{A_{ij}} - h_j^*(\mathbf{z}_i)$, $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{iq_n})' \in \mathcal{R}^{q_n}$ and $\Delta_n = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n)' \in \mathcal{R}^{n \times q_n}$. Let H be the $n \times q_n$ matrix with the (i, j)th element $H_{ij} = h_j^*(\mathbf{z}_i)$, then $X_A = H + \Delta_n$.

The following technical conditions are imposed for analyzing the asymptotic behavior of $\hat{\boldsymbol{\beta}}_1$ and $\hat{g}$.

CONDITION 1 (Conditions on the random error). The random error ε_i has the conditional distribution function F_i and continuous conditional density function f_i, given $\mathbf{x}_i$, $\mathbf{z}_i$. The f_i are uniformly bounded away from 0 and infinity in a neighborhood of zero, its first derivative f_i' has a uniform upper bound in a neighborhood of zero, for $1 \leq i \leq n$.

CONDITION 2 (Conditions on the covariates). There exist positive constants M_1 and M_2 such that $|x_{ij}| \leq M_1$, $\forall 1 \leq i \leq n$, $1 \leq j \leq p_n$ and $E[\delta_{ij}^4] \leq M_2$, $\forall 1 \leq i \leq n$, $1 \leq j \leq q_n$. There exist finite positive constants C_1 and C_2 such that with probability one

$$C_1 \leq \lambda_{\max}(n^{-1} X_A X_A') \leq C_2, \qquad C_1 \leq \lambda_{\max}(n^{-1} \Delta_n \Delta_n') \leq C_2.$$

CONDITION 3 (Condition on the nonlinear functions). For $r = m + v > 1.5$ $g_0 \in \mathcal{G}$.

CONDITION 4 (Condition on the B-spline basis). The dimension of the spline basis k_n has the following rate $k_n \approx n^{1/(2r+1)}$.

CONDITION 5 (Condition on model size). $q_n = O(n^{C_3})$ for some $C_3 < \frac{1}{3}$.

Condition 1 is considerably more relaxed than what is usually imposed on the random error for the theory of high dimensional mean regression, which often requires Gaussian or sub-Gaussian tail condition. Condition 2 is about the behavior of the covariates and the design matrix under the oracle model, which is not restrictive. Condition 3 is typical for the application of B-splines. Stone (1985) showed that B-splines basis functions can be used to effectively approximate functions satisfying Hölder's condition. Condition 4 provides the rate of k_n needed for the optimal convergence rate of $\hat{g}$. Condition 5 is standard for linear models with diverging number of parameters.

The following theorem summarizes the asymptotic properties of the oracle estimators.

THEOREM 2.1. *Assume Conditions 1–5 hold. Then*

$$\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}\| = O_p(\sqrt{n^{-1} q_n}),$$

$$n^{-1} \sum_{i=1}^{n} (\hat{g}(\mathbf{z}_i) - g_0(\mathbf{z}_i))^2 = O_p(n^{-1}(q_n + k_n)).$$

An interesting observation is that since we allow q_n to diverge with n, it influences the rates for estimating both $\boldsymbol{\beta}$ and g. As q_n diverges, to investigate the asymptotic distribution of $\hat{\boldsymbol{\beta}}_1$, we consider estimating an arbitrary linear combination of the components of $\boldsymbol{\beta}_{01}$.

THEOREM 2.2. *Assume the conditions of Theorem* 2.1 *hold. Let* A_n *be an* $l \times q_n$ *matrix with* l *fixed and* $A_n A_n' \to G$, *a positive definite matrix, then*

$$\sqrt{n} A_n \Sigma_n^{-1/2}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) \to N(\mathbf{0}_l, G)$$

in distribution, where $\Sigma_n = K_n^{-1} S_n K_n^{-1}$ *with* $K_n = n^{-1} \Delta_n' B_n \Delta_n$, $S_n = n^{-1} \tau (1 - \tau) \Delta_n' \Delta_n$, *and* $B_n = \text{diag}(f_1(0), \ldots, f_n(0))$ *is an* $n \times n$ *diagonal matrix with* $f_i(0)$ *denoting the conditional density function of* ε_i *given* $(\mathbf{x}_i, \mathbf{z}_i)$ *evaluated at zero.*

If we consider the case where q is fixed and finite, then we have the following result regarding the behavior of the oracle estimator.

COROLLARY 1. *Assume* q *is a fixed positive integer,* $n^{-1} \Delta_n' B_n \Delta_n \to \Sigma_1$ *and* $n^{-1} \tau (1 - \tau) \Delta_n' \Delta_n \to \Sigma_2$, *where* Σ_1 *and* Σ_2 *are positive definite matrices. If Conditions* 1–4 *hold, then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) \xrightarrow{d} N(\mathbf{0}_q, \Sigma_1^{-1} \Sigma_2 \Sigma_1^{-1}),$$

$$n^{-1} \sum_{i=1}^{n} (\hat{g}(\mathbf{z}_i) - g_0(\mathbf{z}_i))^2 = O_p(n^{-2r/(2r+1)}).$$

In the case q_n is fixed, the rates reduce to the classical $n^{-1/2}$ rate for estimating $\boldsymbol{\beta}$ and $n^{-2r/(2r+1)}$ for estimating g, the latter which is consistent with Stone (1985) for the optimal rate of convergence.

3. Nonconvex penalized estimation for partially linear additive quantile regression with ultra-high dimensional covariates.

3.1. *Nonconvex penalized estimator.* In real data analysis, we do not know which of the p_n covariates in $\mathbf{x}_i$ are important. To encourage sparse estimation, we minimize the following penalized objective function for estimating $(\boldsymbol{\beta}_0, \boldsymbol{\xi}_0)$,

$$(3.1) \qquad Q^P(\boldsymbol{\beta}, \boldsymbol{\xi}) = n^{-1} \sum_{i=1}^{n} \rho_\tau (Y_i - \mathbf{x}_i' \boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)' \boldsymbol{\xi}) + \sum_{j=1}^{p_n} p_\lambda(|\beta_j|),$$

where $p_\lambda(\cdot)$ is a penalty function with tuning parameter λ. The L_1 penalty or Lasso [Tibshirani (1996)] is a popular choice for penalized estimation. However, the L_1 penalty is known to over-penalize large coefficients, tends to be biased and requires strong conditions on the design matrix to achieve selection consistency. This is usually not a concern for prediction, but can be undesirable if the goal is to identify the underlying model. In comparison, an appropriate nonconvex penalty function can effectively overcome this problem [Fan and Li (2001)]. In this paper, we consider two such popular choices of penalty functions: the SCAD [Fan and

Li (2001)] and MCP [Zhang (2010)] penalty functions. For the SCAD penalty function,

$$p_\lambda(|\beta|) = \lambda|\beta|I(0 \le |\beta| < \lambda) + \frac{a\lambda|\beta| - (\beta^2 + \lambda^2)/2}{a - 1}I(\lambda \le |\beta| \le a\lambda)$$

$$+ \frac{(a + 1)\lambda^2}{2}I(|\beta| > a\lambda) \qquad \text{for some } a > 2,$$

and for the MCP penalty function,

$$p_\lambda(|\beta|) = \lambda\left(|\beta| - \frac{\beta^2}{2a\lambda}\right)I(0 \le |\beta| < a\lambda) + \frac{a\lambda^2}{2}I(|\beta| \ge a\lambda) \qquad \text{for some } a > 1.$$

For both penalty functions, the tuning parameter λ controls the complexity of the selected model and goes to zero as n increases to ∞.

3.2. *Solving the penalized estimator.* We propose an effective algorithm to solve the above penalized estimation problem. The algorithm is largely based on the idea of the local linear approximation (LLA) [Zou and Li (2008)]. We employ a new trick based on the observation $|\beta_j| = \rho_\tau(\beta_j) + \rho_\tau(-\beta_j)$ to transform the approximated objective function to a quantile regression objective function based on an augmented data set, so that the penalized estimator can be obtained by iteratively solving unpenalized weighted quantile regression problems.

More specifically, we initialize the algorithm by setting $\boldsymbol{\beta} = 0$ and $\boldsymbol{\xi} = 0$. Then for each step $t \ge 1$, we update the estimator by

$$(3.2) \quad (\hat{\boldsymbol{\beta}}^t, \hat{\boldsymbol{\xi}}^t) = \underset{(\boldsymbol{\beta}, \boldsymbol{\xi})}{\operatorname{argmin}}\left\{n^{-1}\sum_{i=1}^{n}\rho_\tau(Y_i - \mathbf{x}_i'\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)'\boldsymbol{\xi}) + \sum_{j=1}^{p_n}p_\lambda'(|\hat{\beta}_j^{t-1}|)|\beta_j|\right\},$$

where $\hat{\beta}_j^{t-1}$ is the value of β_j at step $t - 1$.

By observing that we can write $|\beta_j|$ as $\rho_\tau(\beta_j) + \rho_\tau(-\beta_j)$, the above minimization problem can be framed as an unpenalized weighted quantile regression problem with $n + 2p_n$ augmented observations. We denote these augmented observations by $(Y_i^*, \mathbf{x}_i^*, \mathbf{z}_i^*)$, $i = 1, \ldots, (n + 2p_n)$. The first n observations are those in the original data, that is, $(Y_i^*, \mathbf{x}_i^*, \mathbf{z}_i^*) = (Y_i, \mathbf{x}_i, \mathbf{z}_i)$, $i = 1, \ldots, n$; for the next p_n observations, we have $(Y_i^*, \mathbf{x}_i^*, \mathbf{z}_i^*) = (0, 1, 0)$, $i = n + 1, \ldots, n + p_n$; and the last p_n observations are given by $(Y_i^*, \mathbf{x}_i^*, \mathbf{z}_i^*) = (0, -1, 0)$, $i = n + p_n + 1, \ldots, n + 2p_n$. We fit weighted linear quantile regression model with the observations $(Y_i^*, \mathbf{x}_i^*, \mathbf{z}_i^*)$ and corresponding weights w_i^{t*}, where $w_i^{t*} = 1$, $i = 1, \ldots, n$; $w_{n+j}^{t*} = p_\lambda'(|\hat{\beta}_j^{t-1}|)$, $j = 1, \ldots, p_n$; and $w_{n+p_n+j}^{t*} = -p_\lambda'(|\hat{\beta}_j^{t-1}|)$, $j = 1, \ldots, p_n$.

The above new algorithm is simple and convenient, as weighted quantile regression can be implemented using many existing software packages. In our simulations, we used the quantreg package in R and continue with the iterative procedure until $\|\hat{\boldsymbol{\beta}}^t - \hat{\boldsymbol{\beta}}^{t-1}\|_1 < 10^{-7}$.

3.3. *Asymptotic theory.* In addition to Conditions 1–5, we impose an additional condition on how quickly a nonzero signal can decay, which is needed to identify the underlying model.

CONDITION 6 (Condition on the signal). There exist positive constants C_4 and C_5 such that $2C_3 < C_4 < 1$ and $n^{(1-C_4)/2} \min_{1 \leq j \leq q_n} |\beta_{0j}| \geq C_5$.

Due to the nonsmoothness and nonconvexity of the penalized objective function $Q^P(\boldsymbol{\beta}, \boldsymbol{\xi})$, the classical KKT condition is not applicable to analyzing the asymptotic properties of the penalized estimator. To investigate the asymptotic theory of the nonconvex estimator for ultra-high dimensional partially linear additive quantile regression model, we explore the necessary condition for the local minimizer of a convex differencing problem [Tao and An (1997); Wang, Wu and Li (2012)] and extend it to the setting involving nonparametric components.

Our approach concerns a nonconvex objective function that can be expressed as the difference of two convex functions. Specifically, we consider objective functions belonging to the class

$$\mathbf{F} = \{q(\boldsymbol{\eta}) : q(\boldsymbol{\eta}) = k(\boldsymbol{\eta}) - l(\boldsymbol{\eta}), k(\cdot), l(\cdot) \text{ are both convex}\}.$$

This is a very general formulation that incorporates many different forms of penalized objective functions. The subdifferential of $k(\boldsymbol{\eta})$ at $\boldsymbol{\eta} = \boldsymbol{\eta}_0$ is defined as

$$\partial k(\boldsymbol{\eta}_0) = \{t : k(\boldsymbol{\eta}) \geq k(\boldsymbol{\eta}_0) + (\boldsymbol{\eta} - \boldsymbol{\eta}_0)'t, \forall \boldsymbol{\eta}\}.$$

Similarly, we can define the subdifferential of $l(\boldsymbol{\eta})$. Let $\mathrm{dom}(k) = \{\boldsymbol{\eta} : k(\boldsymbol{\eta}) < \infty\}$ be the effective domain of k. A necessary condition for $\boldsymbol{\eta}^*$ to be a local minimizer of $q(\boldsymbol{\eta})$ is that $\boldsymbol{\eta}^*$ has a neighborhood U such that $\partial l(\boldsymbol{\eta}) \cap \partial k(\boldsymbol{\eta}^*) \neq \varnothing, \forall \boldsymbol{\eta} \in U \cap \mathrm{dom}(k)$ (see Lemma 7 in the Appendix).

To appeal to the above necessary condition for the convex differencing problem, it is noted that $Q^P(\boldsymbol{\beta}, \boldsymbol{\xi})$ can be written as

$$Q^P(\boldsymbol{\beta}, \boldsymbol{\xi}) = k(\boldsymbol{\beta}, \boldsymbol{\xi}) - l(\boldsymbol{\beta}, \boldsymbol{\xi}),$$

where the two convex functions $k(\boldsymbol{\beta}, \boldsymbol{\xi}) = n^{-1} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{x}_i'\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)'\boldsymbol{\xi}) + \lambda \sum_{j=1}^{p_n} |\beta_j|$, and $l(\boldsymbol{\beta}, \boldsymbol{\xi}) = \sum_{j=1}^{p_n} L(\beta_j)$. The specific form of $L(\beta_j)$ depends on the penalty function being used. For the SCAD penalty function,

$$L(\beta_j) = \left[(\beta_j^2 + 2\lambda|\beta_j| + \lambda^2)/(2(a-1))\right]I(\lambda \leq |\beta_j| \leq a\lambda)$$
$$+ \left[\lambda|\beta_j| - (a+1)\lambda^2/2\right]I(|\beta_j| > a\lambda);$$

while for the MCP penalty function,

$$L(\beta_j) = \left[\beta_j^2/(2a)\right]I(0 \leq |\beta_j| < a\lambda) + \left[\lambda|\beta_j| - a\lambda^2/2\right]I(|\beta_j| \geq a\lambda).$$

Building on the convex differencing structure, we show that with probability approaching one that the oracle estimator $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\xi}}')'$, where $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \mathbf{0}_{p_n-q_n}')'$, is

a local minimizer of $Q^P(\boldsymbol{\beta}, \boldsymbol{\xi})$. To study the necessary optimality condition, we formally define $\partial k(\boldsymbol{\beta}, \boldsymbol{\xi})$ and $\partial l(\boldsymbol{\beta}, \boldsymbol{\xi})$, the subdifferentials of $k(\boldsymbol{\beta}, \boldsymbol{\xi})$ and $l(\boldsymbol{\beta}, \boldsymbol{\xi})$, respectively. First, the function $l(\boldsymbol{\beta}, \boldsymbol{\xi})$ does not depend on $\boldsymbol{\xi}$ and is differentiable everywhere. Hence, its subdifferential is simply the regular derivative. For any value of $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$,

$$\partial l(\boldsymbol{\beta}, \boldsymbol{\xi}) = \left\{ \mu = (\mu_1, \mu_2, \ldots, \mu_{p_n+L_n})' \in \mathbb{R}^{p_n+L_n} : \right.$$

$$\left. \mu_j = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j}, 1 \leq j \leq p_n; \mu_j = 0, p_n + 1 \leq j \leq p_n + L_n \right\}.$$

For $1 \leq j \leq p_n$, for the SCAD penalty function,

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \begin{cases} 0, & 0 \leq |\beta_j| < \lambda, \\ (\beta_j - \lambda \operatorname{sgn}(\beta_j))/(a - 1), & \lambda \leq |\beta_j| \leq a\lambda, \\ \lambda \operatorname{sgn}(\beta_j), & |\beta_j| > a\lambda, \end{cases}$$

while for the MCP penalty function,

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \begin{cases} \beta_j/a, & 0 \leq |\beta_j| < a\lambda, \\ \lambda \operatorname{sgn}(\beta_j), & |\beta_j| \geq a\lambda. \end{cases}$$

On the other hand, the function $k(\boldsymbol{\beta}, \boldsymbol{\xi})$ is not differentiable everywhere. Its subdifferential at $(\boldsymbol{\beta}, \boldsymbol{\xi})$ is a collection of $(p_n + L_n)$-vectors:

$$\partial k(\boldsymbol{\beta}, \boldsymbol{\xi}) = \left\{ \kappa = (\kappa_1, \kappa_2, \ldots, \kappa_{p_n+L_n})' \in \mathbb{R}^{p_n+L_n} : \right.$$

$$\kappa_j = -\tau n^{-1} \sum_{i=1}^{n} x_{ij} I(Y_i - \mathbf{x}_i'\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)'\boldsymbol{\xi} > 0)$$

$$+ (1 - \tau)n^{-1} \sum_{i=1}^{n} x_{ij} I(Y_i - \mathbf{x}_i'\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)'\boldsymbol{\xi} < 0)$$

$$- n^{-1} \sum_{i=1}^{n} x_{ij} a_i + \lambda l_j, \text{ for } 1 \leq j \leq p_n;$$

$$\kappa_j = -\tau n^{-1} \sum_{i=1}^{n} \Pi_{j-p_n}(\mathbf{z}_i) I(Y_i - \mathbf{x}_i'\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)'\boldsymbol{\xi} > 0)$$

$$+ (1 - \tau)n^{-1} \sum_{i=1}^{n} \Pi_{j-p_n}(\mathbf{z}_i) I(Y_i - \mathbf{x}_i'\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)'\boldsymbol{\xi} < 0)$$

$$\left. - n^{-1} \sum_{i=1}^{n} \Pi_{j-p_n}(\mathbf{z}_i) a_i, \text{ for } p_n + 1 \leq j \leq p_n + L_n \right\},$$

where we write $\mathbf{\Pi}(\mathbf{z}_i) = (1, \Pi_1(\mathbf{z}_i), \ldots, \Pi_{L_n}(\mathbf{z}_i))'$; $a_i = 0$ if $Y_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{\Pi}(\mathbf{z}_i)'\boldsymbol{\xi} \neq 0$ and $a_i \in [\tau - 1, \tau]$ otherwise; for $1 \leq j \leq p_n$, $l_j = \text{sgn}(\beta_j)$ if $\beta_j \neq 0$ and $l_j \in [-1, 1]$ otherwise.

In the following, we analyze the subgradient of the unpenalized objective function, which plays an essential role in checking the condition of the optimality condition. The subgradient $s(\boldsymbol{\beta}, \boldsymbol{\xi}) = (s_1(\boldsymbol{\beta}, \boldsymbol{\xi}), \ldots, s_{p_n}(\boldsymbol{\beta}, \boldsymbol{\xi}), \ldots, s_{p_n+L_n}(\boldsymbol{\beta}, \boldsymbol{\xi}))'$ is given by

$$s_j(\boldsymbol{\beta}, \boldsymbol{\xi}) = -\frac{\tau}{n} \sum_{i=1}^n x_{ij} I(Y_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{\Pi}(\mathbf{z}_i)'\boldsymbol{\xi} > 0)$$

$$+ \frac{1-\tau}{n} \sum_{i=1}^n x_{ij} I(Y_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{\Pi}(\mathbf{z}_i)'\boldsymbol{\xi} < 0)$$

$$- \frac{1}{n} \sum_{i=1}^n x_{ij} a_i \qquad \text{for } 1 \leq j \leq p_n,$$

$$s_j(\boldsymbol{\beta}, \boldsymbol{\xi}) = -\frac{\tau}{n} \sum_{i=1}^n \Pi_{j-p_n}(\mathbf{z}_i) I(Y_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{\Pi}(\mathbf{z}_i)'\boldsymbol{\xi} > 0)$$

$$+ \frac{1-\tau}{n} \sum_{i=1}^n \Pi_{j-p_n}(\mathbf{z}_i) I(Y_i - \mathbf{x}_i'\boldsymbol{\beta} - \mathbf{\Pi}(\mathbf{z}_i)'\boldsymbol{\xi} < 0)$$

$$- \frac{1}{n} \sum_{i=1}^n \Pi_{j-p_n}(\mathbf{z}_i) a_i \qquad \text{for } p_n + 1 \leq j \leq p_n + L_n,$$

where a_i is defined as before. The following lemma states the behavior of $s_j(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})$ when being evaluated at the oracle estimator.

LEMMA 1. *Assume Conditions 1–6 are satisfied, $\lambda = o(n^{-(1-C_4)/2})$, $n^{-1/2}q_n = o(\lambda)$, $n^{-1/2}k_n = o(\lambda)$ and $\log(p_n) = o(n\lambda^2)$. For the oracle estimator $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})$ there exists a_i^* with $a_i^* = 0$ if $Y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}} - \mathbf{\Pi}(\mathbf{z}_i)'\hat{\boldsymbol{\xi}} \neq 0$ and $a_i^* \in [\tau - 1, \tau]$ otherwise, such that for $s_j(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})$ with $a_i = a_i^*$, with probability approaching one*

$$(3.3) \qquad s_j(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}) = 0, \qquad j = 1, \ldots, q_n \text{ or } j = p_n + 1, \ldots, p_n + L_n,$$

$$(3.4) \qquad |\hat{\beta}_j| \geq (a + 1/2)\lambda, \qquad j = 1, \ldots, q_n,$$

$$(3.5) \qquad |s_j(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})| \leq c\lambda \qquad \forall c > 0, j = q_n + 1, \ldots, p_n.$$

REMARK. Note that for $\kappa_j \in \partial k(\boldsymbol{\beta}, \boldsymbol{\xi})$ and l_j as defined earlier

$$\kappa_j = s_j(\boldsymbol{\beta}, \boldsymbol{\xi}) + \lambda l_j \qquad \text{for } 1 \leq j \leq p_n \quad \text{and}$$

$$\kappa_j = s_j(\boldsymbol{\beta}, \boldsymbol{\xi}) \qquad \text{for } p_n + 1 \leq j \leq p_n + L_n.$$

Thus, Lemma 1 provides important insight on the asymptotic behavior of $\boldsymbol{\kappa} \in \partial k(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})$. Consider a small neighborhood around the oracle estimator $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})$ with radius $\lambda/2$. Building on Lemma 1, we prove in the Appendix that with probability tending to one, for any $(\boldsymbol{\beta}, \boldsymbol{\xi}) \in \mathbb{R}^{p_n+L_n}$ in this neighborhood, there exists $\boldsymbol{\kappa} = (\kappa_1, \dots, \kappa_{p_n}, \mathbf{0}'_{L_n})' \in \partial k(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})$ such that

$$\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \beta_j} = \kappa_j, \qquad j = 1, \dots, p_n \quad \text{and}$$

$$\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}_j} = \kappa_{p_n+j}, \qquad j = 1, \dots, L_n.$$

This leads to the main theorem of the paper. Let $\mathcal{E}_n(\lambda)$ be the set of local minima of $Q^P(\boldsymbol{\beta}, \boldsymbol{\xi})$. The theorem below shows that with probability approaching one, the oracle estimator belongs to the set $\mathcal{E}_n(\lambda)$.

THEOREM 3.1. *Assume Conditions 1–6 are satisfied. Consider either the SCAD or the MCP penalty function with tuning parameter λ. Let $\hat{\boldsymbol{\eta}} \equiv (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})$ be the oracle estimator. If $\lambda = o(n^{-(1-C_4)/2})$, $n^{-1/2}q_n = o(\lambda)$, $n^{-1/2}k_n = o(\lambda)$ and $\log(p_n) = o(n\lambda^2)$, then*

$$P\big(\hat{\boldsymbol{\eta}} \in \mathcal{E}_n(\lambda)\big) \to 1 \qquad \text{as } n \to \infty.$$

REMARK. The conditions for λ in the theorem are satisfied for $\lambda = n^{-1/2+\delta}$ where $\delta \in (\max(1/(2r+1), C_3), C_4)$. The fastest rate of p_n allowed is $p_n = \exp(n^\alpha)$ with $0 < \alpha < 1/2 + 2\delta$. Hence, we allow for the ultra-high dimensional setting.

REMARK. The selection of the tuning parameter λ is important in practice. Cross-validation is a common approach, but is known to often result in overfitting. Lee, Noh and Park (2014) recently proposed high dimensional BIC for linear quantile regression when p is much larger than n. Motivated by their work, we choose λ that minimizes the following high dimensional BIC criterion:

(3.6)
$$\begin{aligned} \text{QBIC}(\lambda) = &\log\left(\sum_{i=1}^{n} \rho_\tau\big(Y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_\lambda - \mathbf{\Pi}(\mathbf{z}_i)' \hat{\boldsymbol{\xi}}_\lambda\big)\right) \\ &+ \nu_\lambda \frac{\log(p_n) \log(\log(n))}{2n}, \end{aligned}$$

where p_n is the number of candidate linear covariates and ν_λ is the degrees of freedom of the fitted model, which is the number of interpolated fits for quantile regression.

4. Simulation. We investigate the performance of the penalized partially linear additive quantile regression estimator in high dimension. We focus on the SCAD penalty and referred to the new procedure as Q-SCAD. An alternative popular nonconvex penalty function is the MCP penalty [Zhang (2010)], the simulation results for which are found to be similar and reported in the online supplementary material [Sherwood and Wang (2015)]. The Q-SCAD is compared with three alternative procedures: partially linear additive quantile regression estimator with the LASSO penalty (Q-LASSO), partially linear additive mean regression with SCAD penalty (LS-SCAD) and LASSO penalty (LS-LASSO). It worth noting that for the mean regression case, there appears to be no theory in the literature for the ultra-high dimensional case.

We first generate $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_{p+2})'$ from the $N_{p+2}(\mathbf{0}_{p+2}, \Sigma)$ multivariate normal distribution, where $\Sigma = (\sigma_{jk})_{(p+2) \times (p+2)}$ with $\sigma_{jk} = 0.5^{|j-k|}$. Then we set $X_1 = \sqrt{12}\Phi(\tilde{X}_1)$ where $\Phi(\cdot)$ is distribution function of $N(0, 1)$ distribution and $\sqrt{12}$ scales X_1 to have standard deviation one. Furthermore, we let $Z_1 = \Phi(\tilde{X}_{25})$, $Z_2 = \Phi(\tilde{X}_{26})$, $X_i = \tilde{X}_i$ for $i = 2, \ldots, 24$ and $X_i = \tilde{X}_{i-2}$ for $i = 27, \ldots, p+2$. The random responses are generated from the regression model

$$(4.1) \quad Y_i = X_{i6}\beta_1 + X_{i12}\beta_2 + X_{i15}\beta_3 + X_{i20}\beta_4 + \sin(2\pi Z_{i1}) + Z_{i2}^3 + \varepsilon_i,$$

where $\beta_j \sim U[0.5, 1.5]$ for $1 \leq j \leq 4$. We consider three different distributions of the error term ε_i: (1) standard normal distribution; (2) t distribution with 3 degrees of freedom; and (3) heteroscedastic normal distribution $\varepsilon_i = \tilde{X}_{i1}\zeta_i$ where $\zeta_i \sim N(0, \sigma = 0.7)$ are independent of the X_i's.

We perform 100 simulations for each setting with sample size $n = 300$, and $p = 100, 300, 600$. Results for additional simulations with sample sizes of 50, 100 and 200 are provided in the online supplementary material [Sherwood and Wang (2015)]. For the heteroscedastic error case, we model $\tau = 0.7$ and 0.9; otherwise, we model the conditional median. Note that at $\tau = 0.7$ or 0.9, when the error has the aforementioned heteroscedastic distribution, X_1 is part of the true model. At these two quantiles, the true model consists of 5 linear covariates. In all simulations, the number of basis functions is set to three, which we find to work satisfactorily in a variety of settings. For the LASSO method, we select the tuning parameters λ by using five-fold cross validation. For the Q-SCAD model, we select λ that minimizes (3.6) while for LS-SCAD we use a least squares equivalent. The tuning parameter a in the SCAD penalty function is set to 3.7 as recommended in Fan and Li (2001). To assess the performance of different methods, we adopt the following criteria:

1. False Variables (FV): average number of nonzero linear covariates incorrectly included in the model.
2. True Variables (TV): average number of nonzero linear covariates correctly included in the model.
3. True: proportion of times the true model is exactly identified.
4. P: proportion of times X_1 is selected.

TABLE 1
Simulation results comparing quantile ($\tau = 0.5$) and mean regression using SCAD and LASSO penalty functions for $\varepsilon \sim N(0, 1)$

Method	n	p	FV	TV	True	P	AADE	MSE
Q-SCAD	300	100	0.20	4.00	0.88	0.00	0.16	0.03
Q-LASSO	300	100	12.88	4.00	0.00	0.13	0.16	0.13
LS-SCAD	300	100	0.32	4.00	0.85	0.00	0.13	0.02
LS-LASSO	300	100	11.63	4.00	0.00	0.12	0.13	0.07
Q-SCAD	300	300	0.04	4.00	0.96	0.00	0.15	0.02
Q-LASSO	300	300	15.93	4.00	0.00	0.07	0.16	0.14
LS-SCAD	300	300	0.33	4.00	0.78	0.00	0.12	0.02
LS-LASSO	300	300	15.00	4.00	0.00	0.04	0.13	0.09
Q-SCAD	300	600	0.06	4.00	0.94	0.00	0.15	0.02
Q-LASSO	300	600	21.86	4.00	0.01	0.06	0.16	0.16
LS-SCAD	300	600	2.57	4.00	0.69	0.01	0.13	0.06
LS-LASSO	300	600	17.11	4.00	0.00	0.04	0.13	0.09

5. AADE: average of the *average absolute deviation* (ADE) of the fit of the nonlinear components, where the ADE is defined as $n^{-1} \sum_{i=1}^{n} |\hat{g}(\mathbf{z}_i) - g_0(\mathbf{z}_i)|$.

6. MSE: average of the mean squared error for estimating $\boldsymbol{\beta}_0$, that is, the average of $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$ across all simulation runs.

The simulation results are summarized in Tables 1–4. Tables 1 and 2 correspond to $\tau = 0.5$, $N(0, 1)$ and T_3 error distribution, respectively. Tables 3 and 4 are for the

TABLE 2
Simulation results comparing quantile ($\tau = 0.5$) and mean regression using SCAD and LASSO penalty functions for $\varepsilon \sim T_3$

Method	n	p	FV	TV	True	P	AADE	MSE
Q-SCAD	300	100	0.07	4.00	0.95	0.00	0.16	0.03
Q-LASSO	300	100	13.09	4.00	0.01	0.17	0.17	0.15
LS-SCAD	300	100	1.08	3.99	0.45	0.02	0.19	0.11
LS-LASSO	300	100	10.15	3.94	0.02	0.08	0.19	0.31
Q-SCAD	300	300	0.05	4.00	0.97	0.00	0.17	0.03
Q-LASSO	300	300	18.42	4.00	0.00	0.08	0.18	0.18
LS-SCAD	300	300	1.22	4.00	0.46	0.00	0.20	0.11
LS-LASSO	300	300	15.15	3.99	0.01	0.08	0.21	0.26
Q-SCAD	300	600	0.06	3.98	0.94	0.00	0.16	0.04
Q-LASSO	300	600	20.81	4.00	0.01	0.03	0.18	0.23
LS-SCAD	300	600	1.33	4.00	0.45	0.00	0.19	0.14
LS-LASSO	300	600	17.40	4.00	0.01	0.01	0.20	0.28

TABLE 3
Simulation results comparing quantile ($\tau = 0.7$) and mean regression using SCAD and LASSO penalty functions for heteroscedastic errors

Method	n	p	FV	TV	True	P	AADE	MSE
Q-SCAD	300	100	0.21	4.84	0.70	0.84	0.17	0.05
Q-LASSO	300	100	13.86	4.97	0.00	0.97	0.24	0.15
LS-SCAD	300	100	1.09	4.06	0.01	0.06	0.16	0.69
LS-LASSO	300	100	11.48	4.13	0.00	0.13	0.17	0.78
Q-SCAD	300	300	0.20	4.77	0.61	0.77	0.20	0.06
Q-LASSO	300	300	18.54	4.97	0.00	0.97	0.27	0.18
LS-SCAD	300	300	3.28	4.00	0.00	0.00	0.16	0.68
LS-LASSO	300	300	15.85	4.08	0.00	0.08	0.16	0.79
Q-SCAD	300	600	0.16	4.59	0.48	0.59	0.26	0.08
Q-LASSO	300	600	23.26	4.89	0.00	0.89	0.31	0.24
LLS-SCAD	300	600	6.31	4.02	0.00	0.02	0.16	0.69
LS-LASSO	300	600	18.50	4.09	0.00	0.09	0.16	0.83

heteroscedastic error, $\tau = 0.7$ and 0.9, respectively. Least squares based estimates of $\hat{\boldsymbol{\beta}}$ for $\tau = 0.7$ or 0.9 are obtained by assuming $\varepsilon_i \sim N(0, \sigma)$, with estimates of σ being used in each simulation. An extension of Table 3 for $p = 1200$ and 2400 is included in the online supplementary material [Sherwood and Wang (2015)]. We observe that the method with the SCAD penalty tends to pick a smaller and more accurate model. The advantages of quantile regression can be seen by its stronger performance at the presence of heavy-tailed distribution or heteroscedastic errors.

TABLE 4
Simulation results comparing quantile ($\tau = 0.9$) and mean regression using SCAD and LASSO penalty functions for heteroscedastic errors

Method	n	p	FV	TV	True	P	AADE	MSE
Q-SCAD	300	100	0.06	4.93	0.91	0.98	0.24	0.30
Q-LASSO	300	100	12.94	5.00	0.00	1.00	0.49	0.73
LS-SCAD	300	100	1.09	4.06	0.01	0.06	0.16	4.72
LS-LASSO	300	100	11.48	4.13	0.00	0.13	0.17	4.73
Q-SCAD	300	300	0.26	5.00	0.81	1.00	0.19	0.24
Q-LASSO	300	300	16.33	5.00	0.00	1.00	0.62	0.92
LS-SCAD	300	300	3.28	4.00	0.00	0.00	0.16	4.63
LS-LASSO	300	300	15.85	4.08	0.00	0.08	0.16	4.67
Q-SCAD	300	600	0.34	4.94	0.77	1.00	0.21	0.29
Q-LASSO	300	600	19.79	4.97	0.00	1.00	0.74	1.15
LS-SCAD	300	600	6.31	4.02	0.00	0.02	0.16	4.64
LS-LASSO	300	600	18.50	4.09	0.00	0.09	0.16	4.74

For the latter case, the least squared based methods perform poorly in identifying the active variables in the dispersion function. Estimation of the nonlinear terms is similar across different error distributions and different values of p.

5. An application to birth weight data. Votavova et al. (2011) collected blood samples from peripheral blood, cord blood and the placenta from 20 pregnant smokers and 52 pregnant women without significant exposure to smoking. Their main objective was to identify the difference in transcriptome alterations between the two groups. Birth weight of the baby (in kilograms) was recorded along with age of the mother, gestational age, parity, measurement of the amount of cotinine, a chemical found in tobacco, in the blood and mother's BMI. Low birth weight is known to be associated with both short-term and long-term health complications. Scientists are interested in which genes are associated with low birth weight [Turan et al. (2012)].

We consider modeling the 0.1, 0.3 and 0.5 conditional quantiles of infant birth weight. We use the genetic data from the peripheral blood sample which include 64 subjects after dropping those with incomplete information. The blood samples were assayed using HumanRef-8 v3 Expression BeadChips with 24,539 probes. For each quantile, the top 200 probes are selected using the quantile-adaptive screening method [He, Wang and Hong (2013)]. The gene expression values of the 200 probes are included as linear covariates for the semiparametric quantile regression model. The clinical variables parity, gestational age, cotinine level and BMI are also included as linear covariates. The age of the mother is modeled nonparametrically as exploratory analysis reveals potential nonlinear effect.

We consider the semiparametric quantile regression model with the SCAD and LASSO penalty functions. Least squares based semiparametric models with the SCAD and LASSO penalty functions are also considered. Results for the MCP penalty are reported in the online supplementary material [Sherwood and Wang (2015)]. The tuning parameter λ is selected by minimizing (3.6) for the SCAD estimator and by five-fold cross validation for LASSO as discussed in Section 4. The third column of Table 5 reports the number of nonzero elements, "Original NZ," for each model. As expected, the LASSO method selects a larger model than the SCAD penalty does. The number of nonzero variables varies with the quantile level, providing evidence that mean regression alone would provide a limited view of the conditional distribution.

Next, we compare different models on 100 random partitions of the data set. For each partition, we randomly select 50 subjects for the training data and 14 subjects for the test data. The fourth column of Table 5 reports the prediction error evaluated on the test data, defined as $14^{-1} \sum_{i=1}^{14} \rho_\tau (Y_i - \hat{Y}_i)$; while the fifth column reports the average number of linear covariates included in each model (denoted by "Randomized NZ"). Standard errors for the prediction error is reported in parentheses. We note that the SCAD method produces notably smaller models than the Lasso method does without sacrificing much prediction accuracy.

Quantile ($\tau = 0.1$, 0.3 and 0.5) and mean regression analysis of birth weight based on the original data and the random partitioned data

τ	Method	Original NZ	Prediction error	Randomized NZ
0.10	Q-SCAD	2	0.07 (0.03)	2.27
0.10	Q-LASSO	10	0.08 (0.02)	3.09
0.30	Q-SCAD	7	0.18 (0.04)	6.74
0.30	Q-LASSO	22	0.16 (0.03)	12.39
0.50	Q-SCAD	5	0.21 (0.04)	5.80
0.50	Q-LASSO	6	0.20 (0.04)	14.25
Mean	LS-SCAD	12	0.20 (0.04)	5.43
Mean	LS-LASSO	12	0.20 (0.04)	3.77

Model checking in high dimension is challenging. In the following, we consider a simulation-based diagnostic plot to help visually assess the overall lack-of-fit for the quantile regression model [Wei and He (2006)] to assess the overall lack-of-fit for the quantile regression model. First, we randomly generate $\tilde{\tau}$ from the uniform [0, 1] distribution. Then we fit the proposed semiparametric quantile regression model using the SCAD penalty for the quantile $\tilde{\tau}$. Next, we generate a response variable $\tilde{Y} = \mathbf{x}'\hat{\boldsymbol{\beta}}(\tilde{\tau}) + \hat{g}(z, \tilde{\tau})$, where $(\mathbf{x}, z)$ is randomly sampled from the set of observed covariates, with z denoting mother's age and $\mathbf{x}$ denoting the vector of other covariates. The process is repeated 100 times and produces a sample of 100 simulated birth weights based on the model. Figure 1 shows the QQ plot comparing the simulated and observed birth weights. Overall, the QQ plot is close to the 45 degree line and does not suggest gross lack-of-fit. Figure 2 displays the estimated nonlinear effects of mother's age $\hat{g}(z)$ at the three quantiles [standardized to satisfy the constraint $\sum_{i=1}^{n} \hat{g}(z_i) = 0$]. At the 0.1 and 0.3 quantiles, the estimated mother's age effects are similar except for some deviations at the tails of the mother's age distribution. At these two quantiles, after age 30, mother's age is observed to have a positive effect. The effect of mother's age at the median is nonmonotone: the effect is first increasing (up to age 25), then decreasing (to about age 33), and increasing again.

We observe that different models are often selected for different random partitions. Table 6 summarizes the variables selected by Q-SCAD for $\tau = 0.1$, 0.3 and 0.5 and the frequency these variables are selected in the 100 random partitions. Probes are listed by their identification number along with corresponding gene in parentheses. The SCAD models tend to produce sparser models while the LASSO models provide slightly better predictive performance.

Gestational age is identified to be important with high frequency at all three quantiles under consideration. This is not surprising given the known important
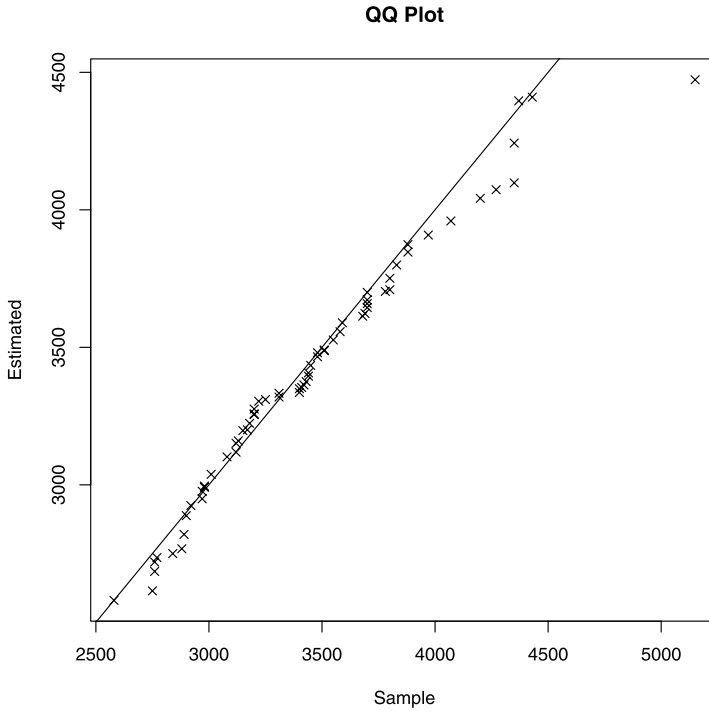
QQ Plot

FIG. 1. *Lack-of-fit diagnostic QQ plot for the birth weight data example.*

relationship between birth weight and gestational age. Premature birth is often strongly associated with low birth weight. The genes selected at the three different quantiles are not overlapping. This is an indication of the heterogeneity in the data. The variation in frequency is likely due to the relatively small sample size. However, examining the selected genes does provide some interesting insights. The gene SOGA1 is a suppressor of glucose, which is interesting because maternal gestational diabetes is known to have a significant effect on birth weight [Gilliam et al. (2003)]. The genes OR2AG1, OR5P2 and DEPDC7 are all located on chromosome 11, the chromosome with the most selected genes. Chromosome 11 also contains PHLDA2, a gene that has been reported to be highly expressed in mothers that have children with lower birth weight [Ishida et al. (2012)].

6. Estimation and variable selection for multiple quantiles. Motivated by referees' suggestions, we consider an extension for simultaneous variable selection at multiple quantiles. Let $\tau_1 < \tau_2 < \cdots < \tau_M$ be the set of quantiles of interest, where $M > 0$ is a positive integer. We assume that

$$(6.1) \qquad Q_{Y_i|\mathbf{x}_i,\mathbf{z}_i}(\tau_m) = \mathbf{x}_i'\boldsymbol{\beta}_0^{(m)} + g_0^{(m)}(\mathbf{z}_i), \qquad m = 1,\ldots,M,$$
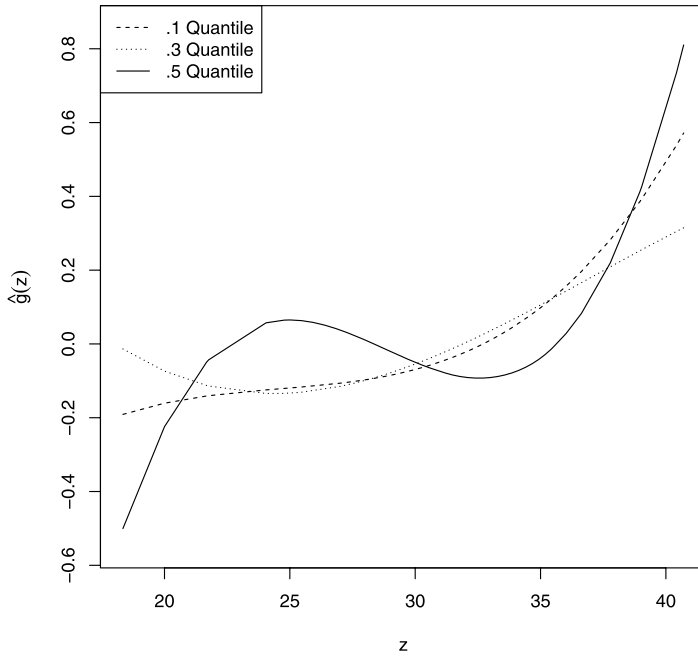
FIG. 2. *Estimated nonlinear effects of mother's age (denoted by z) at three different quantiles.*

where $g_0^{(m)}(\mathbf{z}_i) = g_{00}^{(m)} + \sum_{j=1}^{d} g_{0j}^{(m)}(z_{ij})$, with $g_{00}^{(m)} \in \mathcal{R}$. We assume that functions $g_{0j}^{(m)}$ satisfy $E[g_{0j}^{(m)}(z_{ij})] = 0$ for the purpose of identification. The nonlinear functions are allowed to vary with the quantiles. We are interested in the high dimensional case where most of the linear covariates have zero coefficients across all M quantiles, for which group selection will help us combine information across quantiles.

TABLE 6
Frequency of covariates selected at three quantiles among 100 *random partitions*

Q-SCAD 0.1		Q-SCAD 0.3		Q-SCAD 0.5	
Covariate	**Fre-quency**	**Covariate**	**Fre-quency**	**Covariate**	**Fre-quency**
Gestational age	82	Gestational age	86	Gestational age	69
1,687,073 (SOGA1)	24	1,804,451 (LEO1)	33	2,334,204 (ERCC6L)	57
		1,755,657 (RASIP1)	27	1,732,467 (OR2AG1)	52
		1,658,821 (SAMD1)	23	1,656,361 (LOC201175)	31
		2,059,464 (OR5P2)	14	1,747,184 (PUS7L)	5
		2,148,497 (C20orf107)	6		
		2,280,960 (DEPDC7)	3		

We write $\boldsymbol{\beta}_0^{(m)} = (\beta_{01}^{(m)}, \beta_{02}^{(m)}, \ldots, \beta_{0p_n}^{(m)})'$, $m = 1, \ldots, M$. Let $\bar{\boldsymbol{\beta}}^{0j}$ be the M-vector $(\beta_{0j}^{(1)}, \ldots, \beta_{0j}^{(M)})'$, $1 \leq j \leq p_n$. Let $\bar{A} = \{j : \|\bar{\boldsymbol{\beta}}^{0j}\| \neq 0, 1 \leq j \leq p_n\}$ be the index set of variables that are active at least one quantile level of interest, where $\| \cdot \|$ denotes the L_2 norm. Let $\bar{q}_n = |\bar{A}|$ be the cardinality of A. Without loss of generality, we assume $\bar{A} = \{1, \ldots, \bar{q}_n\}$. Let $X_{\bar{A}}$ and $\mathbf{x}'_{\bar{A}_1}, \ldots, \mathbf{x}_{\bar{A}_n}$ be defined as before. By the result of Schumaker (1981), there exists $\boldsymbol{\xi}_0^{(m)} \in \mathcal{R}^{L_n}$, where $L_n = d(k_n + l + 1) + 1$, such that $\sup_{\mathbf{z}_i} |\boldsymbol{\Pi}(\mathbf{z}_i)' \boldsymbol{\xi}_0^{(m)} - g_0^{(m)}(\mathbf{z}_i)| = O(k_n^{-r})$, $m = 1, \ldots, M$.

We write the (Mp_n)-vector $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)'}, \ldots, \boldsymbol{\beta}^{(M)'})'$, where for $k = 1, \ldots, M$, $\boldsymbol{\beta}^{(k)} = (\beta_1^{(k)}, \ldots, \beta_{p_n}^{(k)})'$; and we write the (ML_n)-vector $\boldsymbol{\xi} = (\boldsymbol{\xi}^{(1)'}, \ldots, \boldsymbol{\xi}^{(M)'})$. Let $\bar{\boldsymbol{\beta}}^j$ be the M-vector $(\beta_j^{(1)}, \ldots, \beta_j^{(M)})'$, $1 \leq j \leq p_n$. For simultaneous variable selection and estimation, we estimate $(\boldsymbol{\beta}_0^{(m)}, \xi_0^{(m)})$, $m = 1, \ldots, M$, by minimizing the following penalized objective function

$$
\begin{aligned}
\bar{Q}^P(\boldsymbol{\beta}, \boldsymbol{\xi}) = {} & n^{-1} \sum_{i=1}^n \sum_{m=1}^M \rho_{\tau_m}\big(Y_i - \mathbf{x}_i' \boldsymbol{\beta}^{(m)} - \boldsymbol{\Pi}(\mathbf{z}_i)' \boldsymbol{\xi}^{(m)}\big) \\
& + \sum_{j=1}^{p_n} p_\lambda(\|\bar{\boldsymbol{\beta}}^j\|_1),
\end{aligned}
\tag{6.2}
$$

where $p_\lambda(\cdot)$ is a penalty function with tuning parameter λ, $\| \cdot \|_1$ denotes the L_1 norm, which was used in Yuan and Lin (2006) for group penalty; see also Huang, Breheny and Ma (2012). The penalty function encourages group-wise sparsity and forces the covariates that have no effect on any of the M quantiles to be excluded together. Similarly penalty functions have been used in Zou and Yuan (2008), Liu and Wu (2011) for variable selection at multiple quantiles. The above estimator can be computed similarly as in Section 3.2.

In the oracle case, the estimator would be obtained by considering the unpenalized part of (6.2), but with $\mathbf{x}_i$ replaced by $\mathbf{x}_{\bar{A}_i}$. That is, we let

$$
\begin{aligned}
& \big\{\hat{\boldsymbol{\beta}}_1^{(m)}, \hat{\xi}^{(m)} : 1 \leq m \leq M\big\} \\
& \quad = \underset{\boldsymbol{\beta}_1^{(m)}, \boldsymbol{\xi}^{(m)}, 1 \leq m \leq M}{\operatorname{argmin}} \ n^{-1} \sum_{i=1}^n \sum_{m=1}^M \rho_{\tau_m}\big(Y_i - \mathbf{x}_{\bar{A}_i}' \boldsymbol{\beta}_1^{(m)} - \boldsymbol{\Pi}(\mathbf{z}_i)' \boldsymbol{\xi}^{(m)}\big).
\end{aligned}
\tag{6.3}
$$

The oracle estimator for $\boldsymbol{\beta}_0^{(m)}$ is $\hat{\boldsymbol{\beta}}^{(m)} = (\hat{\boldsymbol{\beta}}_1^{(m)'}, \mathbf{0}'_{p_n-q_n})'$, and across all quantiles is $\hat{\bar{\boldsymbol{\beta}}} = (\hat{\boldsymbol{\beta}}^{(1)}, \ldots, \hat{\boldsymbol{\beta}}^{(M)})$ and $\hat{\bar{\boldsymbol{\xi}}} = (\hat{\boldsymbol{\xi}}^{(1)}, \ldots, \hat{\boldsymbol{\xi}}^{(M)})$. The oracle estimator for the non-parametric function $g_{0j}^{(m)}$ is $\hat{g}_j^{(m)}(z_{ij}) = \boldsymbol{\pi}(z_{ij})' \hat{\boldsymbol{\xi}}_j^{(m)} - n^{-1} \sum_{i=1}^n \pi(z_{ij})' \hat{\boldsymbol{\xi}}_j^{(m)}$ for $j = 1, \ldots, d$; for $g_{00}^{(m)}$ is $\hat{g}_0^{(m)} = \hat{\xi}_0^{(m)} + n^{-1} \sum_{i=1}^n \sum_{j=1}^d \pi(z_{ij})' \hat{\boldsymbol{\xi}}_j^{(m)}$. The oracle es-

timator of $g_0^{(m)}(\mathbf{z}_i)$ is $\hat{g}^{(m)}(\mathbf{z}_i) = \hat{g}_0^{(m)} + \sum_{j=1}^d \hat{g}_j^{(m)}(z_{ij})$. As the next theorem suggests, Theorem 3.1 can be extended to the multiple quantile case. To save space, we present the regularity conditions and the technical derivations in the online supplementary material [Sherwood and Wang (2015)].

THEOREM 6.1. *Assume Conditions* B1–B6 *in the online supplementary material* [*Sherwood and Wang* (2015)] *are satisfied. Let* $\bar{\mathcal{E}}_n(\lambda)$ *be the set of local minima of the penalized objective function* $\bar{Q}^P(\boldsymbol{\beta}, \gamma)$. *Consider either the SCAD or the MCP penalty function with tuning parameter* λ. *Let* $\hat{\bar{\boldsymbol{\eta}}} \equiv (\hat{\bar{\boldsymbol{\beta}}}, \hat{\bar{\boldsymbol{\xi}}})$ *be the oracle estimator that solves* (6.3). *If* $\lambda = o(n^{-(1-C_4)/2})$, $n^{-1/2}\bar{q}_n = o(\lambda)$, $n^{-1/2}k_n = o(\lambda)$ *and* $\log(p_n) = o(n\lambda^2)$, *then*

$$P(\hat{\bar{\boldsymbol{\eta}}} \in \bar{\mathcal{E}}_n(\lambda)) \to 1 \qquad as\ n \to \infty.$$

A numerical example. To assess the multiple quantile estimator, we ran 100 simulations using the setting presented in Section 4 with $\varepsilon_i \sim T_3$, and consider $\tau = 0.5$, 0.7 and 0.9. We compare the variable selection performance of the multiple-quantile estimator (denoted by Q-group) in this section with the method that estimates each quantile separately (denoted by Q-ind). For both approaches, we use the SCAD penalty function. Results for the MCP penalty are included in the online supplementary material [Sherwood and Wang (2015)]. We also report results from the multiple-quantile oracle estimator (denotes by Q-oracle) which assumes the knowledge of the underlying model and serves as a benchmark.

Table 7 summarizes the simulation results for $n = 50$, $p = 300$ and 600. As in Zou and Yuan (2008), when evaluating the Q-ind method, at quantile level τ_m, we define $A_m = \{j : \hat{\beta}_j^{(m)} \neq 0\}$ be the index set of estimated nonzero coefficients at this quantile level. Let $\bigcup_{m=1}^M A_m$ be the set of the selected variables using Q-ind. As the simulations results in Section 4, we report FV, TV and TRUE. We also report the error for estimating the linear coefficients (L_2 error), which is defined as the average of $M^{-1} \sum_{m=1}^M (\hat{\boldsymbol{\beta}}^{(m)} - \boldsymbol{\beta}_0^{(m)})^2$ over all simulation runs. The results

TABLE 7
Comparison of group and individual penalty functions for multiple quantile estimation with $\varepsilon \sim T_3$

Method	p	FV	TV	True	L_2 error
Q-group-SCAD	300	1.01	4	0.49	0.14
Q-ind-SCAD	300	0.98	4	0.45	0.17
Q-oracle	300	0	4	1	0.06
Q-group-SCAD	600	1.2	4	0.56	0.15
Q-ind-SCAD	600	1.51	3.99	0.34	0.17
Q-oracle	600	0	4	1	0.07

demonstrate that comparing with Q-ind, the new method Q-group has lower false discovery rate, higher probability of identifying the true underlying model and smaller estimation error.

7. Discussion. We considered nonconvex penalized estimation for partially linear additive quantile regression models with high dimensional linear covariates. We derive the oracle theory under mild conditions. We have focused on estimating a particular quantile of interest and also considered an extension to simultaneous variable selection at multiple quantiles.

A problem of important practical interest is how to identify which covariates should be modeled linearly and which covariates should be modeled nonlinearly. Usually, we do not have such prior knowledge in real data analysis. This is a challenging problem in high dimension. Recently, important progresses have been made by Zhang, Cheng and Liu (2011); Huang, Wei and Ma (2012); Lian, Liang and Ruppert (2015) for semiparametric mean regression models. We plan on addressing this question for high dimensional semiparametric quantile regression in our future research.

Another relevant problem of practical interest is to estimate the conditional quantile function itself. Given $\mathbf{x}^*$, $\mathbf{z}^*$, we can estimate $Q_{Y_i|\mathbf{x}^*,\mathbf{z}^*}(\tau)$ by $\mathbf{x}^{*\prime}\hat{\boldsymbol{\beta}}_1 + \hat{g}(\mathbf{z}^*)$, where $\hat{\boldsymbol{\beta}}$ and $\hat{g}$ are obtained from penalized quantile regression. We conjecture that the consistency of estimating the conditional quantile function can be derived under somewhat weaker conditions in the current paper, as motivated by the results on *persistency* for linear mean regression in high dimension [Greenshtein and Ritov (2004)]. The details will also be further investigated in the future.

APPENDIX

Throughout the appendix, we use C to denote a positive constant which does not depend on n and may vary from line to line. For a vector $\mathbf{x}$, $\|\mathbf{x}\|$ denotes its Euclidean norm. For a matrix A, $\|A\| = \sqrt{\lambda_{\max}(A'A)}$ denotes its spectral norm. For a function $h(\cdot)$ on $[0, 1]$, $\|h\|_\infty = \sup_x |h(x)|$ denotes the uniform norm. Let I_n denote an $n \times n$ identity matrix.

A.1. Derivation of the results in Section 2.

A.1.1. *Notation.* To facilitate the proof, we will make use of the theoretically centered B-spline basis functions similar to the approach used by Xue and Yang (2006). More specifically, we consider the B-spline basis functions $b_j(\cdot)$ in Section 2.1 and let $B_j(z_{ik}) = b_{j+1}(z_{ik}) - \frac{E[b_{j+1}(z_{ik})]}{E[b_1(z_{ik})]} b_1(z_{ik})$ for $j = 1, \ldots, k_n + l$. Then $E(B_j(z_{ik})) = 0$. For a given covariate z_{ik}, let $\mathbf{w}(z_{ik}) = (B_1(z_{ik}), \ldots, B_{k_n+l}(z_{ik}))'$ be the vector of basis functions, and $\mathbf{W}(\mathbf{z}_i)$ denote the J_n-dimensional vector $(k_n^{-1/2}, \mathbf{w}(z_{i1})', \ldots, \mathbf{w}(z_{id})')'$, where $J_n = d(k_n + l) + 1$.

By the result of Schumaker [(1981), page 227], there exists a vector $\boldsymbol{\gamma}_0 \in \mathcal{R}^{J_n}$ and a positive constant C_0, such that $\sup_{t \in [0,1]^d} |g_0(\mathbf{t}) - \mathbf{W}(\mathbf{t})'\boldsymbol{\gamma}_0| \leq C_0 k_n^{-r}$. Let

$$(A.1) \qquad (\hat{\mathbf{c}}_1, \hat{\boldsymbol{\gamma}}) = \operatorname*{argmin}_{(\mathbf{c}_1, \boldsymbol{\gamma})} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau (Y_i - \mathbf{x}'_{A_i} \mathbf{c}_1 - \mathbf{W}(\mathbf{z}_i)'\boldsymbol{\gamma}).$$

We write $\boldsymbol{\gamma} = (\gamma_0, \boldsymbol{\gamma}'_1, \ldots, \boldsymbol{\gamma}'_d)'$, where $\gamma_0 \in \mathcal{R}$, $\boldsymbol{\gamma}_j \in \mathcal{R}^{k_n+l}$, $j = 1, \ldots, d$; and we write $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_0, \hat{\boldsymbol{\gamma}}'_1, \ldots, \hat{\boldsymbol{\gamma}}'_d)'$ the same fashion. It can be shown that (see the supplemental material) $\hat{\mathbf{c}}_1 = \hat{\boldsymbol{\beta}}_1$. So the change of the basis functions for the nonlinear part does not alter the estimator for the linear part. Let $\tilde{g}_j(\mathbf{z}_i) = w(z_{ij})'\hat{\boldsymbol{\gamma}}_j$ be the estimator of g_{0j}, $j = 1, \ldots, d$. The estimator for g_{00} is $\tilde{g}_0 = k_n^{-1/2}\hat{\boldsymbol{\gamma}}_0$. The estimator for $g_0(\mathbf{z}_i)$ is $\tilde{g}(\mathbf{z}_i) = \mathbf{W}(\mathbf{z}_i)'\hat{\boldsymbol{\gamma}} = \tilde{g}_0 + \sum_{j=1}^{d} \tilde{g}_j(\mathbf{z}_i)$. It can be derived that (see the supplemental material) $\hat{g}_j(\mathbf{z}_i) = \tilde{g}_j(\mathbf{z}_i) - n^{-1} \sum_{i=1}^{n} \tilde{g}_j(\mathbf{z}_i)$ and $\hat{g}_0 = \tilde{g}_0 + n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{d} \tilde{g}_j(\mathbf{z}_i)$. Hence, $\hat{g} = \hat{g}_0 + \sum_{j=1}^{d} \hat{g}_j = \tilde{g}$. Later, we will show $n^{-1} \sum_{i=1}^{n} (\tilde{g}(\mathbf{z}_i) - g_0(\mathbf{z}_i))^2 = O_p(n^{-1}(q_n + d J_n))$.

Throughout the proof, we will also use the following notation:

$$\psi_\tau(\varepsilon_i) = \tau - I(\varepsilon_i < 0),$$

$$W = \big(\mathbf{W}(\mathbf{z}_1), \ldots, \mathbf{W}(\mathbf{z}_n)\big)' \in \mathbb{R}^{n \times J_n},$$

$$P = W\big(W' B_n W\big)^{-1} W' B_n \in \mathbb{R}^{n \times n},$$

$$X^* = \big(\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*\big)' = (I_n - P) X_A \in \mathbb{R}^{n \times q_n},$$

$$W_B^2 = W' B_n W \in \mathbb{R}^{J_n \times J_n},$$

$$\boldsymbol{\theta}_1 = \sqrt{n}(\mathbf{c}_1 - \boldsymbol{\beta}_{10}) \in \mathbb{R}^{q_n},$$

$$\boldsymbol{\theta}_2 = W_B(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) + W_B^{-1} W' B_n X_A(\mathbf{c}_1 - \boldsymbol{\beta}_{10}) \in \mathbb{R}^{J_n},$$

$$\tilde{\mathbf{x}}_i = n^{-1/2} \mathbf{x}_i^* \in \mathbb{R}^{q_n},$$

$$\tilde{\mathbf{W}}(\mathbf{z}_i) = W_B^{-1} \mathbf{W}(\mathbf{z}_i) \in \mathbb{R}^{J_n},$$

$$\tilde{\mathbf{s}}_i = \big(\tilde{\mathbf{x}}_i', \tilde{\mathbf{W}}(\mathbf{z}_i)\big)' \in \mathbb{R}^{q_n + J_n},$$

$$u_{ni} = \mathbf{W}(\mathbf{z}_i)'\boldsymbol{\gamma}_0 - g_0(\mathbf{z}_i).$$

Notice that

$$n^{-1} \sum_{i=1}^{n} \rho_\tau (Y_i - \mathbf{x}'_{A_i} \mathbf{c}_1 - \mathbf{W}(\mathbf{z}_i)'\boldsymbol{\gamma}) = n^{-1} \sum_{i=1}^{n} \rho_\tau (\varepsilon_i - \tilde{\mathbf{x}}_i'\boldsymbol{\theta}_1 - \tilde{\mathbf{W}}(\mathbf{z}_i)'\boldsymbol{\theta}_2 - u_{ni}).$$

Define the minimizers under the transformation as

$$(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \arg\min_{(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} n^{-1} \sum_{i=1}^{n} \rho_\tau (\varepsilon_i - \tilde{\mathbf{x}}_i'\boldsymbol{\theta}_1 - \tilde{\mathbf{W}}(\mathbf{z}_i)'\boldsymbol{\theta}_2 - u_{ni}).$$

Let a_n be a sequence of positive numbers and define

$$Q_i(a_n) \equiv Q_i(a_n\boldsymbol{\theta}_1, a_n\boldsymbol{\theta}_2) = \rho_\tau(\varepsilon_i - a_n\tilde{\mathbf{x}}_i'\boldsymbol{\theta}_1 - a_n\tilde{\mathbf{W}}(\mathbf{z}_i)'\boldsymbol{\theta}_2 - u_{ni}),$$

$$E_s[Q_i] = E[Q_i|\mathbf{x}_i, \mathbf{z}_i].$$

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$. Define

(A.2)
$$D_i(\boldsymbol{\theta}, a_n) = Q_i(a_n) - Q_i(0) - E_s[Q_i(a_n) - Q_i(0)]$$
$$+ a_n(\tilde{\mathbf{x}}_i'\boldsymbol{\theta}_1 + \tilde{\mathbf{W}}(\mathbf{z}_i)'\boldsymbol{\theta}_2)\psi_\tau(\varepsilon_i).$$

Noting that $\rho_\tau(u) = \frac{1}{2}|u| + (\tau - \frac{1}{2})u$, we have

(A.3)
$$Q_i(a_n) - Q_i(0) = \frac{1}{2}[|\varepsilon_i - a_n\tilde{\mathbf{x}}_i'\boldsymbol{\theta}_1 - a_n\tilde{\mathbf{W}}(\mathbf{z}_i)'\boldsymbol{\theta}_2 - u_{ni}| - |\varepsilon_i - u_{ni}|]$$
$$- a_n(\tau - \frac{1}{2})(\tilde{\mathbf{x}}_i'\boldsymbol{\theta}_1 + \tilde{\mathbf{W}}(\mathbf{z}_i)'\boldsymbol{\theta}_2).$$

Define

$$Q_i^*(a_n) = \frac{1}{2}[|\varepsilon_i - \tilde{\mathbf{x}}_i'\boldsymbol{\theta}_1 a_n - \tilde{\mathbf{W}}(\mathbf{z}_i)'\boldsymbol{\theta}_2 a_n - u_{ni}| - |\varepsilon_i - u_{ni}|].$$

Then by combining (A.2) and (A.3),

(A.4) $$D_i(\boldsymbol{\theta}, a_n) = Q_i^*(a_n) - E_s[Q_i^*(a_n)] + a_n(\tilde{\mathbf{x}}_i'\boldsymbol{\theta}_1 + \tilde{\mathbf{W}}(\mathbf{z}_i)'\boldsymbol{\theta}_2)\psi_\tau(\varepsilon_i).$$

A.1.2. *Some technical lemmas.* The proofs of Lemmas 2–4 below are given in the supplemental material [Sherwood and Wang (2015)].

LEMMA 2. *We have the following properties for the spline basis vector:*

(1) $E(\|\mathbf{W}(\mathbf{z}_i)\|) \leq b_1$, $\forall i$, *for some positive constant b_1 for all n sufficiently large.*

(2) *There exists positive constant b_2 and b_2^* such that for all n sufficiently large $E(\lambda_{\min}(\mathbf{W}(\mathbf{z}_i)\mathbf{W}(\mathbf{z}_i)^T)) \geq b_2 k_n^{-1}$ and $E(\lambda_{\max}(\mathbf{W}(\mathbf{z}_i)\mathbf{W}(\mathbf{z}_i)^T)) \leq b_2^* k_n^{-1}$.*

(3) $E(\|W_B^{-1}\|) \geq b_3\sqrt{k_n n^{-1}}$, *for some positive constant b_3, for all n sufficiently large.*

(4) $\max_i \|\tilde{\mathbf{W}}(\mathbf{z}_i)\| = O_p(\sqrt{\frac{k_n}{n}})$.

(5) $\sum_{i=1}^n f_i(0)\tilde{\mathbf{x}}_i\tilde{\mathbf{W}}(\mathbf{z}_i)' = \mathbf{0}$.

LEMMA 3. *If Conditions 1–5 are satisfied, then:*

(1) *There exists a positive constant C such that $\lambda_{\max}(n^{-1}X^{*'}X^*) \leq C$, with probability one.*

(2) $n^{-1/2}X^* = n^{-1/2}\Delta_n + o_p(1)$. *Furthermore, $n^{-1}X^{*'}B_nX^* = K_n + o_p(1)$, where B_n and K_n are defined as in Theorem 2.2.*

LEMMA 4. *If Conditions 1–5 hold, then $n^{-1}\sum_{i=1}^n(\tilde{g}(\mathbf{z}_i) - g_0(\mathbf{z}_i))^2 = O_p(d_n/n)$.*

LEMMA 5. *Assume Conditions 1–5 hold. Let* $\tilde{\boldsymbol{\theta}}_1 = \sqrt{n}(X^{*\prime}B_n X^*)^{-1} \times X^{*\prime}\psi_\tau(\varepsilon)$, *where* $\psi_\tau(\varepsilon) = (\psi_\tau(\varepsilon_1), \ldots, \psi_\tau(\varepsilon_n))^\prime$, *then:*

(1) $\|\tilde{\boldsymbol{\theta}}_1\| = O_p(\sqrt{q_n})$.

(2) $A_n \Sigma_n^{-1/2} \tilde{\boldsymbol{\theta}}_1 \overset{d}{\to} N(0, G)$, *where* A_n, Σ_n *and* G *are defined in Theorem* 2.2.

PROOF.   (1) The result follows from the observation that, by Lemma 3,

$$\tilde{\boldsymbol{\theta}}_1 = (K_n + o_p(1))^{-1}\big[n^{-1/2}\Delta_n^\prime \psi_\tau(\varepsilon) + n^{-1/2}(H - PX_A)\psi_\tau(\varepsilon)\big],$$

and $n^{-1/2}\|H - PX_A\| = o_p(1)$.

(2)

$$A_n \Sigma_n^{-1/2} \tilde{\boldsymbol{\theta}}_1 = A_n \Sigma_n^{-1/2} K_n^{-1}\big[n^{-1/2}\Delta_n^\prime \psi_\tau(\varepsilon)\big](1 + o_p(1))$$
$$+ A_n \Sigma_n^{-1/2} K_n^{-1}\big[n^{-1/2}(H - PX_A)\big]\psi_\tau(\varepsilon)(1 + o_p(1)),$$

where the second term is $o_p(1)$ because $n^{-1/2}\|H - PX_A\| = o(1)$. We write $A_n \Sigma_n^{-1/2} K_n^{-1}[n^{-1/2}\Delta_n^\prime \psi_\tau(\varepsilon)] = \sum_{i=1}^n D_{ni}$, where

$$D_{ni} = n^{-1/2} A_n \Sigma_n^{-1/2} K_n^{-1} \delta_i \psi_\tau(\varepsilon_i).$$

To verify asymptotic normality, we first note that $E(D_{ni}) = 0$ and

$$\sum_{i=1}^n E(D_{ni}D_{ni}^\prime) = A_n \Sigma_n^{-1/2} K_n^{-1} S_n K_n^{-1} \Sigma_n^{-1/2} A_n^\prime = A_n A_n^\prime \to G.$$

The proof is complete by checking the Lindeberg–Feller condition. For any $\varepsilon > 0$ and using Conditions 1, 2 and 5

$$\sum_{i=1}^n E\big[\|D_{ni}\|^2 I\big(\|D_{ni}\| > \varepsilon\big)\big]$$

$$\leq \varepsilon^{-2} \sum_{i=1}^n E\|D_{ni}\|^4$$

$$\leq (n\varepsilon)^{-2} \sum_{i=1}^n E\big(\psi_\tau^4(\varepsilon_i)\big(\delta_i^\prime K_n^{-1} \Sigma_n^{-1/2} A_n^\prime A_n \Sigma_n^{-1/2} K_n^{-1} \delta_i\big)^2\big)$$

$$\leq C n^{-2}\varepsilon^{-2} \sum_{i=1}^n E\big(\|\delta_i\|^4\big) = O_p(q_n^2/n) = o_p(1),$$

where the last inequality follows by observing that $\lambda_{\max}(A_n^\prime A_n) = \lambda_{\max}(A_n A_n^\prime) \to c$ for some finite positive constant c.   □

LEMMA 6.   *If Conditions 1–5 hold, then*

$$\|\hat{\boldsymbol{\theta}}_1 - \tilde{\boldsymbol{\theta}}_1\| = o_p(1).$$

PROOF. Proof provided in online supplementary material [Sherwood and Wang (2015)]. □

A.1.3. *Proof of Theorems* 2.1, 2.2 *and Corollary* 1. By the observation $\hat{g} = \tilde{g}$, Lemma 4 implies the second result of Theorem 2.1. The first result of Theorem 2.1 follows by observing $\hat{\mathbf{c}}_1 = \hat{\boldsymbol{\beta}}_1$ and Lemmas 5 and 6. The proof of Theorem 2.2 follows from Lemmas 5 and 6. Set $A_n = I_q$, then the proof of Corollary 1 follows from the fact that q being constant and Theorems 2.1 and 2.2.

A.2. Derivation of the results in Section 3.3.

LEMMA 7. *Consider the function* $k(\eta) - l(\eta)$ *where both* k *and* l *are convex with subdifferential functions* $\partial k(\eta)$ *and* $\partial l(\eta)$. *Let* η^* *be a point that has neighborhood* U *such that* $\partial l(\eta) \cap \partial k(\eta^*) \neq \varnothing, \forall \eta \in U \cap \mathrm{dom}(k)$. *Then* η^* *is a local minimizer of* $k(\eta) - l(\eta)$.

PROOF. The proof is available in Tao and An (1997). □

A.2.1. *Proof of Lemma* 1.

PROOF OF (3.3). By convex optimization theory $\mathbf{0} \in \partial \sum_{i=1}^{n} \rho_\tau (Y_i - \mathbf{x}_i' \boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)' \boldsymbol{\xi})$. Thus, there exists a_j^* as described in the lemma such that with the choice $a_j = a_j^*$, we have $s_j(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}) = 0$ for $j = 1, \dots, q_n$ or $j = p_n + 1, \dots, p_n + J_n$. □

PROOF OF (3.4). It is sufficient to show $P(|\hat{\beta}_j| \geq (a + 1/2)\lambda, \text{ for } j = 1, \dots, q_n) \to 1$ as $n, p \to \infty$. Note that

(A.5) $$\min_{1 \leq j \leq q_n} |\hat{\beta}_j| \geq \min_{1 \leq j \leq q_n} |\beta_{0j}| - \max_{1 \leq j \leq q_n} |\hat{\beta}_j - \beta_{0j}|.$$

By Condition 6, $\min_{1 \leq j \leq q_n} |\beta_{0j}| \geq C_5 n^{-(1-C_4)/2}$. By Theorem 2.1 and Conditions 5 and 6, $\max_{1 \leq j \leq q_n} |\hat{\beta}_j - \beta_{0j}| = O_p(\sqrt{\frac{q_n}{n}}) = o_p(n^{-(1-C_4)/2})$. (3.3) holds by noting $\lambda = o(n^{-(1-C_4)/2})$. □

PROOF OF (3.5). Proof provided in the online supplementary material [Sherwood and Wang (2015)]. □

A.2.2. *Proof of Theorem* 3.1. Recall that for $\kappa_j \in \partial k(\boldsymbol{\beta}, \boldsymbol{\xi})$

$$\kappa_j = s_j(\boldsymbol{\beta}, \boldsymbol{\xi}) + \lambda l_j \qquad \text{for } 1 \leq j \leq p_n,$$
$$\kappa_j = s_j(\boldsymbol{\beta}, \boldsymbol{\xi}) \qquad \text{for } p_n + 1 \leq j \leq p_n + J_n.$$

Define the set

$$\mathcal{G} = \{ \boldsymbol{\kappa} = (\kappa_1, \kappa_2, \ldots, \kappa_{p_n + J_n})' : \kappa_j = \lambda \operatorname{sgn}(\hat{\beta}_j), j = 1, \ldots, q_n;$$

$$\kappa_j = s_j(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}) + \lambda l_j, j = q_n + 1, \ldots, p_n;$$

$$\kappa_j = 0, j = p_n + 1, \ldots, p_n + J_n, \},$$

where l_j ranges over $[-1, 1]$ for $j = q_n + 1, \ldots, p_n$. By Lemma 1, we have $P(\mathcal{G} \subset \partial k(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}})) \to 1$.

Consider any $(\boldsymbol{\beta}', \boldsymbol{\xi}')'$ in a ball with the center $(\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\xi}}')'$ and radius $\lambda/2$. By Lemma 7, to prove the theorem it is sufficient to show that there exists $\boldsymbol{\kappa}^* = (\kappa_1^*, \ldots, \kappa_{p_n + J_n}^*)' \in \mathcal{G}$ such that

(A.6)
$$P\left(\kappa_j^* = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \beta_j}, j = 1, \ldots, p_n \right) \to 1;$$

(A.7)
$$P\left(\kappa_{p_n + j}^* = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \xi_j}, j = 1, \ldots, J_n \right) \to 1.$$

Since $\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \xi_j} = 0$ for $j = 1, \ldots, J_n$, (A.7) is satisfied by Lemma 1. We outline how κ_j^* can be selected to satisfy (A.6).

1. For $1 \le j \le q_n$, we have $\kappa_j^* = \lambda \operatorname{sgn}(\hat{\beta}_j)$ for $\beta_j \ne 0$. For either SCAD or MCP penalty function, $\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \beta_j} = \lambda \operatorname{sgn}(\beta_j)$ for $|\beta_j| > a\lambda$. By Lemma 1, we have

$$\min_{1 \le j \le q_n} |\beta_j| \ge \min_{1 \le j \le q_n} |\hat{\beta}_j| - \max_{1 \le j \le q_n} |\hat{\beta}_j - \beta_j| \ge (a + 1/2)\lambda - \lambda/2 = a\lambda,$$

with probability approaching one. Thus, $P(\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \beta_j} = \lambda \operatorname{sgn}(\beta_j)) \to 1$. For any $1 \le j \le q_n$, $\|\hat{\beta}_j - \beta_{0j}\| = O_p(n^{-1/2} q_n^{1/2}) = o(\lambda)$. Therefore, for sufficiently large n, $\hat{\beta}_j$ and β_j have the same sign. This implies $P(\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \beta_j} = \kappa_j^*, 1 \le j \le q_n) \to 1$ as $n \to \infty$.

2. For $j = q_n + 1, \ldots, p_n$, $\hat{\beta}_j = 0$ by the definition of the oracle estimator and $\kappa_j = \lambda l_j$ with $l_j \in [-1, 1]$. Therefore,

$$|\beta_j| \le |\hat{\beta}_j| + |\hat{\beta}_j - \beta_j| < \lambda/2.$$

For $|\beta_j| < \lambda$, $\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \beta_j} = 0$ for the SCAD penalty and $\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \beta_j} = \beta_j/a$ for MCP, $j = q_n + 1, \ldots, p_n$. Note that for both penalty functions, we have $|\frac{l(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \beta_j}| \le \lambda$, $j = q_n + 1, \ldots, p_n$. By Lemma 1, $|s_j(\hat{\beta}_j)| \le \lambda/2$ with probability approaching one for $j = q_n + 1, \ldots, p_n$. Therefore, for both penalty functions, there exists $l_j^* \in [-1, 1]$ such that $P(s_j(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}) + \lambda l_j^* = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\xi})}{\beta_j}, j = q_n + 1, \ldots, p_n) \to 1$. Define $\kappa_j^* = s_j(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}) + \lambda l_j^*$. Then $P(\frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \beta_j} = \kappa_j^*, q_n + 1 \le j \le p_n) \to 1$ as $n \to \infty$.

This completes the proof.

Acknowledgments. We thank the Editor, the Associate Editor and the anonymous referees for their careful reading and constructive comments which have helped us to significantly improve the paper.

SUPPLEMENTARY MATERIAL

REFERENCES

BAI, Z. D. and WU, Y. (1994). Limiting behavior of M-estimators of regression coefficients in high-dimensional linear models. I. Scale-dependent case. *J. Multivariate Anal.* **51** 211–239. MR1321295

BELLONI, A. and CHERNOZHUKOV, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39** 82–130. MR2797841

BUNEA, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *Ann. Statist.* **32** 898–927. MR2065193

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

GILLIAM, M., RIFAS-SHIMAN, S., BERKEY, C., FIELD, A. and COLDITZ, G. (2003). Maternal gestational diabetes, birth weight and adolescent obesity. *Pediatrics* **111** 221–226.

GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039

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

HE, X. and SHI, P. (1996). Bivariate tensor-product B-splines in a partly linear model. *J. Multivariate Anal.* **58** 162–181. MR1405586

HE, X., WANG, L. and HONG, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Statist.* **41** 342–369. MR3059421

HE, X., ZHU, Z.-Y. and FUNG, W.-K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* **89** 579–590. MR1929164

HUANG, J., BREHENY, P. and MA, S. (2012). A selective review of group selection in high-dimensional models. *Statist. Sci.* **27** 481–499. MR3025130

HUANG, J., HOROWITZ, J. L. and WEI, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38** 2282–2313. MR2676890

HUANG, J., WEI, F. and MA, S. (2012). Semiparametric regression pursuit. *Statist. Sinica* **22** 1403–1426. MR3027093

ISHIDA, M., MONK, D., DUNCAN, A. J., ABU-AMERO, S., CHONG, J., RING, S. M., PEMBREY, M. E., HINDMARSH, P. C., WHITTAKER, J. C., STANIER, P. and MOORE, G. E. (2012). Maternal inheritance of a promoter variant in the imprinted PHLDA2 gene significantly increases birth weight. *Am. J. Hum. Genet.* **90** 715–719.

KAI, B., LI, R. and ZOU, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *Ann. Statist.* **39** 305–332. MR2797848

LAM, C. and FAN, J. (2008). Profile-kernel likelihood inference with diverging number of parameters. *Ann. Statist.* **36** 2232–2260. MR2458186

LEE, E. R., NOH, H. and PARK, B. U. (2014). Model selection via Bayesian information criterion for quantile regression models. *J. Amer. Statist. Assoc.* **109** 216–229. MR3180558

LI, G., XUE, L. and LIAN, H. (2011). Semi-varying coefficient models with a diverging number of components. *J. Multivariate Anal*. **102** 1166–1174. MR2805656

LIAN, H., LIANG, H. and RUPPERT, D. (2015). Separation of covariates into nonparametric and parametric parts in high-dimensional partially linear additive models. *Statist. Sinica* **25** 591–607.

LIANG, H. and LI, R. (2009). Variable selection for partially linear models with measurement errors. *J. Amer. Statist. Assoc*. **104** 234–248. MR2504375

LIU, X., WANG, L. and LIANG, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statist. Sinica* **21** 1225–1248. MR2827522

LIU, Y. and WU, Y. (2011). Simultaneous multiple non-crossing quantile regression estimation using kernel constraints. *J. Nonparametr. Stat*. **23** 415–437. MR2801302

SCHUMAKER, L. L. (1981). *Spline Functions*: *Basic Theory*. Wiley, New York. MR0606200

SHERWOOD, B. and WANG, L. (2015). Supplement to "Partially linear additive quantile regression in ultra-high dimension." DOI:10.1214/15-AOS1367SUPP.

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist*. **13** 689–705. MR0790566

TANG, Y., SONG, X., WANG, H. J. and ZHU, Z. (2013). Variable selection in high-dimensional quantile varying coefficient models. *J. Multivariate Anal*. **122** 115–132. MR3189311

TAO, P. D. and AN, L. T. H. (1997). Convex analysis approach to d.c. programming: Theory, algorithms and applications. *Acta Math. Vietnam*. **22** 289–355. MR1479751

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TURAN, N., GHALWASH, M., KATARIL, S., COUTIFARIS, C., OBRADOVIC, Z. and SAPIENZA, C. (2012). DNA methylation differences at growth related genes correlate with birth weight: A molecular signature linked to developmental origins of adult disease? *BMC Medical Genomics* **5** 10.

VOTAVOVA, H., DOSTALOVA MERKEROVA, M., FEJGLOVA, K., VASIKOVA, A., KREJCIK, Z., PASTORKOVA, A., TABASHIDZE, N., TOPINKA, J., VELEMINSKY, M., JR., SRAM, R. J. and BRDICKA, R. (2011). Transcriptome alterations in maternal and fetal cells induced by tobacco smoke. *Placenta* **32** 763–770.

WANG, L., WU, Y. and LI, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *J. Amer. Statist. Assoc*. **107** 214–222. MR2949353

WANG, H. and XIA, Y. (2009). Shrinkage estimation of the varying coefficient model. *J. Amer. Statist. Assoc*. **104** 747–757. MR2541592

WANG, H. J., ZHU, Z. and ZHOU, J. (2009). Quantile regression in partially linear varying coefficient models. *Ann. Statist*. **37** 3841–3866. MR2572445

WANG, L., LIU, X., LIANG, H. and CARROLL, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. *Ann. Statist*. **39** 1827–1851. MR2893854

WEI, Y. and HE, X. (2006). Conditional growth charts. *Ann. Statist*. **34** 2069–2131. With discussions and a rejoinder by the authors. MR2291494

WELSH, A. H. (1989). On *M*-processes and *M*-estimation. *Ann. Statist*. **17** 337–361. MR0981455

XIE, H. and HUANG, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *Ann. Statist*. **37** 673–696. MR2502647

XUE, L. and YANG, L. (2006). Additive coefficient modeling via polynomial spline. *Statist. Sinica* **16** 1423–1446. MR2327498

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **68** 49–67. MR2212574

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist*. **38** 894–942. MR2604701

ZHANG, H. H., CHENG, G. and LIU, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models. *J. Amer. Statist. Assoc*. **106** 1099–1112. MR2894767

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

ZOU, H. and YUAN, M. (2008). Regularized simultaneous model selection in multiple quantiles regression. *Comput. Statist. Data Anal.* **52** 5296–5304. MR2526595

DEPARTMENT OF BIOSTATISTICS
JOHNS HOPKINS UNIVERSITY
BALTIMORE, MARYLAND 21205
USA
E-MAIL: bsherwo2@jhu.edu

SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MINNESOTA 55455
USA
E-MAIL: wangx346@umn.edu