Journal of Machine Learning Research 21 (2020) 1-25 Submitted 2/19; Revised 4/20; Published 11/20

Ultra-High Dimensional Single-Index Quantile Regression

Yuankun Zhang ZHANGYK@MAIL.UC.EDU
Department of Mathematical Sciences
University of Cincinnati
Cincinnati, OH 45221, USA

Heng Lian HENGLIAN@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
Hong Kong SAR

Yan Yu YAN.YU@UC.EDU
Department of Operations, Business Analytics, and Information Systems
University of Cincinnati
Cincinnati, OH 45221, USA

Editor: Massimiliano Pontil

Abstract

We consider a flexible semiparametric single-index quantile regression model where the number of covariates may be ultra-high dimensional, and the number of the relevant covariates is potentially diverging. The approach is particularly appealing to uncover the complex heterogeneity in high-dimensional data, incorporate nonlinearity and potential interaction, avoid the curse of dimensionality, and allow different variables to be included at different quantile levels. We estimate the unknown function via polynomial splines nonparametrically and adopt a nonconvex penalty function to identify the sparse variable set. We further extend it to partially linear single-index quantile model where both the single-index components in the nonparametric term and the partially linear components can be in ultra-high dimension. However, a number of major challenges arise in developing both theory and computation: (a) The model is highly nonlinear in single-index coefficients because the high-dimensional single-index covariates are embedded inside the unknown flexible function. (b) The data are ultra-high dimensional where the dimension of the single-index covariates (p_n) is diverging or even in the exponential order of sample size n. (c) The objective function is non-smooth for quantile regression. (d) Nonconvex variable selection such as SCAD is adopted for regularization. (e) The extended partially linear single-index quantile models may include both ultra-high dimensional (p_n) single-index covariates and ultra-high dimensional (q_n) partially linear covariates. We develop a novel approach using empirical process techniques in establishing the theoretical properties of the nonconvex penalized estimators for partially linear single-index quantile models and show those estimators indeed possess the oracle property in ultra-high dimensional setting. We propose an efficient algorithm to circumvent the computational challenges. The results of Monte Carlo simulations and an application to gene expression data demonstrate the effectiveness of the proposed models and estimation method.

Keywords: High-dimensional data, nonparametric, oracle property, variable selection, SCAD.

1. Introduction

With the rapid development of computing technologies, high-dimensional complex data have often emerged in many fields from social sciences to various scientific research areas (see a discussion in Fan et al. (2009) and Fan et al. (2011b) among others). For example, the microarray gene data used in our empirical example measure more than 22,000 gene expression levels on 60 laboratory mice with obesity or diabetes.

Quantile regression models have shown great promise since the seminal work by Koenker and Bassett Jr (1978), especially for such high-dimensional complex data. In practice, high-dimensional data often exhibit heterogeneity, which may be scientifically important but tend to be ignored by popular mean regression. By investigating conditional quantiles at various quantile levels, we can display a better picture about the heterogeneity in the conditional distribution of the response variable. Also, it is more reasonable to assume the so-called quantile-adaptive sparsity (Sherwood and Wang (2016)) for high-dimensional data to allow different relevant variables for different quantiles. Quantile models have also demonstrated to be more resistant to outliers and heavy tailed errors.

We propose flexible regularized single-index quantile regression models for high-dimensional data. For the observed data $\{y_i, z_{i1}, \ldots, z_{ip_n}\}, i = 1, \ldots, n$, $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip_n})$ is a p_n-dimensional predictor vector. A single-index model for the τ-th conditional quantile of y_i given the covariates $\mathbf{z}_i$ can be formulated as

$$\mathcal{Q}_\tau(y_i|\mathbf{z}_i) = \eta(z_1\beta_1 + z_2\beta_2 + \ldots + z_{p_n}\beta_{p_n}) = \eta(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_0), \tag{1}$$

that is, $P\{y_i \leq \eta(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_0)|\mathbf{z}_i\} = \tau$. Here $0 < \tau < 1$, the term $\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_0$ is called the single index and $\eta(\cdot)$ is an unknown univariate function, modeled by B-splines nonparametrically in this paper.

We further extend (1) to partially linear single-index quantile regression when $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq_n})$, a q_n-dimensional predictor vector may enter the model through a partially linear term

$$\mathcal{Q}_\tau(y_i|\mathbf{z}_i, \mathbf{x}_i) = \eta(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_0) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha}_0. \tag{2}$$

Identifiability of the model above has been established previously. In particular, as a special case of Theorem 2 of Lin and Kulasekera (2007), we have that under the assumptions: (i) the support of $(\mathbf{x}, \mathbf{z})$ is a bounded convex set with at least one interior point; (ii) $\|\boldsymbol{\beta}_0\| = 1$ with first nonzero element positive; (iii) η is continuous and non-constant, then $\eta, \boldsymbol{\beta}_0, \boldsymbol{\alpha}_0$ can be identified.

We allow both partially linear covariates and single-index covariates within the nonparametric function diverging and even in ultra-high dimension, that is, both $p_n \gg n$ and $q_n \gg n$ and even at the order of $exp(n^{1/2})$. To achieve the sparsity of the underlying model structure, where we assume only a small but potentially diverging number of important covariates will affect the response variable, we take the popular regularization approach with a non-convex penalty such as SCAD (Fan and Li (2001)) and MCP (Zhang (2010)).

To the best of our knowledge, this paper appears to be the first to tackle the challenging ultra-high dimensional semiparametric quantile regression where both single-index covariates in the nonparametric part and partially linear terms are potentially ultra-high dimensional. The past literature on single-index models mostly focused on mean regressions in the fixed low-to-moderate dimensional settings, see, for instance, Carroll et al.

(1997), Yu and Ruppert (2002), Xia et al. (2002), Liang et al. (2010), Cui et al. (2011), among others. Fixed-dimensional single-index quantile models have been considered in Wu et al. (2010), Wu and Yu (2014), Ma and He (2016), and Zhang et al. (2017). The studies of high-dimensional mean regression for single-index models include Radchenko (2015), Wang and Wang (2015), and Zhang et al. (2012). In addition, Alquier and Biau (2013) explored a PAC-Bayesian approach to select important variables in a sparse single-index model. Neykov et al. (2016) proposed a covariance screening procedure with LASSO penalization for Gaussian designs to uncover the sparsity pattern. A few papers study the high-dimensional linear quantile models; for example, Belloni and Chernozhukov (2011) considered L_1-penalized linear quantile regression for the high-dimensional model. Recently Wang et al. (2012) made an important theoretical advancement by imposing substantially weaker assumptions for ultra-high-dimensional linear quantile regression. Sherwood and Wang (2016) worked on a nice partially linear additive quantile model in which only the number of partially linear components is allowed to increase with the sample size while holding the fixed-dimensional additive component a priori.

Adopting single-index models (e.g. Ichimura (1993), Powell et al. (1989), Härdle and Stoker (1989)) for high-dimensional quantile regression is particularly appealing for its flexibility and interpretability. Single-index models generalize linear models by replacing the linear component $\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_0$ with a nonparametric component $\eta(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_0)$. It is flexible to accommodate possible nonlinearity while circumventing the so-called "curse of dimensionality." Unlike additive models, single-index models can also naturally incorporate some interactions among covariates. This feature becomes highly attractive in the high-dimensional setting because a high volume of covariates collected are more likely to exhibit some interaction effects. In fact, single-index models form the basis of more complicated models such as projection pursuit regression and deep neural networks (see Yang et al. (2017)). In addition, single-index coefficients retain easy interpretability. If η is monotonic, then the single-index coefficient $\boldsymbol{\beta}_0$ can have similar interpretation as in linear models. Often times, the index itself may be of particular interest (Ma et al. (2017); Guo et al. (2017)). Finally, partially linear single-index models introduce additional flexibility while maintaining the easy interpretability.

While the proposed ultra-high dimensional partially linear single-index quantile models enjoy many appealing features for complex heterogeneous data, there are a number of major challenges in developing both theory and computation. (i) Model (1) is highly nonlinear in single-index coefficients $\boldsymbol{\beta}$ because the high-dimensional single-index covariates $\mathbf{z}$ are embedded inside the unknown flexible function $\eta(\cdot)$ through a linear projection, more specifically, via $\eta(z_1\beta_1 + z_2\beta_2 + \ldots + z_{p_n}\beta_{p_n})$. (ii) The data are ultra-high dimensional where the dimension of single-index covariates p_n are diverging or even in the exponential order of n. The dimension of important variables are also potentially diverging. (iii) The objective function is non-smooth for quantile regression. (iv) Nonconvex variable selection such as SCAD is adopted for regularization. (v) The extended partially linear single-index quantile models (2) include both ultra-high dimensional (p_n) single-index covariates $\mathbf{z}$ and ultra-high dimensional (q_n) partially linear covariates $\mathbf{x}$.

Overall, in combination of (i) through (v), we have to deal with not only the the nonconvex and non-smooth objective function associated with quantile regression in (ultra-)high dimension, but also the highly nonlinear structure in the single-index coefficients.

Note that this work is very different from Sherwood and Wang (2016) which is essentially a high-dimensional linear quantile regression problem as Wang et al. (2012), because only the partially linear covariates are potentially in high dimension while the number of nonparametric additive terms holds fixed. In our paper, both the single-index covariates in the nonparametric part and partially linear covariates are allowed to be high-dimensional. Furthermore, because the single-index covariates are embedded inside the unknown function as stated in (i), it yields much greater challenges than those in the current literature.

We develop efficient algorithms and establish desirable theoretical properties for the proposed partially linear single-index models for quantile regression in ultra-high dimension. Computationally, to tackle the ultra-high dimensional single-index parameters inside an unknown function and the nonconvex non-smooth objective function, we develop an efficient estimation method. In particular, we make linear approximations to the nonparametric function and the nonconvex penalty function, which turn the nonconvex optimization into a convex optimization problem. For the ultra-high dimensional modeling, a pre-screening process such as Fan and Lv (2008) or the penalized linear quantile regression estimates with the screened variables via distance correlation in Zhong et al. (2016) can be used to help reduce dimensionality and produce initial values for the iterative estimation algorithm. Moreover, we adopt a technique that could create more augmented data and make use of the weighted quantile regression so that we are able to handle the ultra-high dimensional model fitting in quantile regression directly and obtain the penalized estimates with large p_n.

In theory, we establish the theoretical properties of the penalized estimators of high-dimensional partially linear single-index models for quantile regression. The high dimensionality combined with non-smooth loss function makes establishing the theoretical results challenging. With high dimensional parameters, we need to control the size of various quantities in terms of p_n and q_n explicitly. Furthermore, the high dimensionality of the single-index covariates in the nonparametric part gives rise to an extra challenge, that is, the spline basis functions are now defined on an interval potentially with diverging support whose properties are different from the standard case with fixed support. Finally, for the penalized estimator, existing theoretical investigations of linear and semiparametric quantile models with a nonconvex penalty, such as Wang et al. (2012) and Sherwood and Wang (2016), used the results in Tao and An (1997) which require writing the objective function as a difference of convex functions. This theoretical tool is nevertheless not applicable for single-index models due to the existence of the link function that makes the model highly nonlinear in single-index coefficients. Thus, we develop a novel approach in establishing the theoretical results, which directly compares the objective function values in a sufficiently small neighborhood of the oracle estimator, and bounds the differences in objective function values using empirical process techniques. We hope our new approach to prove the theoretical properties for high-dimensional nonparametric models could invite more work in investigating models with ultra-high dimensional covariates within possible complex nonparametric components.

The rest of this article is organized as follows. In Section 2, we present the methodology and establish the theoretical properties for the oracle estimators. In Section 3, we show the estimation algorithm and theoretical properties of the penalized estimators with a nonconvex penalty and establish their oracle property. In Section 4, we conduct Monte

Carlo simulations to evaluate the performance of the proposed method and apply the semi-parametric model and the penalized estimation to a gene expression data set. We conclude and discuss future research opportunities in Section 5. We relegate all lemmas and proofs to the supplementary material.

2. Single-Index Quantile Regression with Diverging Number of Relevant Covariates

2.1 The Methodology

Let us first consider the single-index quantile regression models (1). We allow the number of variables p_n inside the unknown function $\eta(\cdot)$ to be potentially ultra-high dimensional. It is commonly assumed that the model is sparse in the sense that among all parameters $\beta_{0j}, j = 1, \ldots, p_n$, many of them are zero. Let $B_1 = \{1 \le j \le p_n : \beta_{0j} \neq 0\}$ be the index set of nonzero coefficients and $p_{1n} = |B_1|$ be the cardinality of B_1. We need to estimate the set B_1. Without loss of generality, we assume the first p_{1n} components of $\boldsymbol{\beta}_0$ are nonzero, denoted by $\boldsymbol{\beta}_{10} = (\beta_{01}, \ldots, \beta_{0p_{1n}})^{\mathrm{T}}$, and the rest $p_n - p_{1n}$ elements of $\boldsymbol{\beta}_0$ zeros. Specifically, $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^{\mathrm{T}}, \mathbf{0}_{p_n - p_{1n}}^{\mathrm{T}})^{\mathrm{T}}$.

Throughout the paper, we assume that $\boldsymbol{\beta}_0$ belongs to the parameter space $\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\| = 1$ and the first component positive$\}$ for identifiability purposes; see more detailed discussion in Yu and Ruppert (2002), **?** and Zhang et al. (2017) among others.

We adopt polynomial splines to estimate the unknown function $\eta(\cdot)$ nonparametrically. Assume the support of $\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0$ is $[a, b]$. Given the focus of the high-dimensional setting in the current study, the size of the support, $b-a$, can be potentially diverging with n. In practice, the minimal and maximal values of $\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}$ with a given $\boldsymbol{\beta}$ serve as a and b, respectively, when B-spline basis functions are constructed. Let $a = \nu_0 < \nu_1 < \ldots < \nu_{N'} < \nu_{N'+1} = b$ be a partition of $[a, b]$ into N' subintervals $I_{k'} = [\nu_{k'}, \nu_{k'+1}), k' = 0, \ldots, N'$, where $N' \equiv N'_n$ increases with sample size n in the order $O(n^h)$ such that $\max_{0 \le k' \le N'} \|\nu_{k'+1} - \nu_{k'}\| = O(n^{-h})$ with $h \in (0, 0.5)$.

Any function $f(\cdot)$ from the space of polynomial splines of order $d \ge 2$ satisfies: (i) on each $I_{k'}$, $1 \le k' \le N'$, $f(\cdot)$ is a polynomial of degree $d-1$; (ii) $f(\cdot)$ is globally $d-2$ times continuously differentiable on $[0, 1]$. See the definition in Schumaker (1981) or Stone (1985). The collection of splines on $[0, 1]$ with a fixed sequence of knots has a B-spline basis $\widetilde{\boldsymbol{\Pi}}(s) := \{\widetilde{\Pi}_1(s), \ldots, \widetilde{\Pi}_N(s)\}$ with $N \equiv N_n = N'_n + d$. We assume the basis is scaled to have $\sum_{k=1}^{N} \widetilde{\Pi}_k(s) = \sqrt{N}$. Such normalization is not essential, but adopted to simplify some expressions in theoretical deductions later[1]. To define the basis on $[a, b]$, we let $\boldsymbol{\Pi}(s) := (\Pi_1(s), \ldots, \Pi_N(s))^{\mathrm{T}} = (l^{-1/2}\widetilde{\Pi}_1(l^{-1}(s-a)), \ldots, l^{-1/2}\widetilde{\Pi}_N(l^{-1}(s-a)))^{\mathrm{T}}$, where $l = b - a$, which makes sure that the eigenvalues of $E\boldsymbol{\Pi}(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\Pi}(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)^{\mathrm{T}}$ are bounded away from zero and infinity under mild assumptions (e.g. Eqn. (S.2) in Lemma 3 of Wang et al. (2011)). Given the single index $s_i = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}$, the unknown function $\eta(s_i)$ can be estimated

1. In addition, we will have the eigenvalues of $\int_0^1 \widetilde{\boldsymbol{\Pi}}(s)\widetilde{\boldsymbol{\Pi}}^{\mathrm{T}}(s)ds$ bounded away from zero and infinity, while if using the basis with $\sum_{k=1}^{N} \widetilde{\Pi}_k(s) = 1$ we would have the eigenvalues being of order $O(N^{-1})$.

nonparametrically by a linear combination of B-spline bases, i.e.,

$$\eta(s_i) \approx \sum_{k=1}^{N} \mathbf{\Pi}_k(s_i)^{\mathrm{T}}\theta_k,$$

where each $\mathbf{\Pi}_k$ is defined on $[a, b]$.

An extension to the high-dimensional single-index models is to incorporate partially linear components into quantile regression. We are interested in the high-dimensional partially linear single-index models (2) (PLSIMs) in which the number of covariates in both single-index part p_n and partially linear part q_n may be ultra-high dimensional. In the same spirit, the model is considered sparse. Let $A_1 = \{1 \le j \le q_n : \alpha_{0j} \neq 0\}$ be the index set of nonzero coefficients and $q_{1n} = |A_1|$ be the cardinality of A_1. Without loss of generality, we assume the first q_{1n} components of $\boldsymbol{\alpha}_0$ are nonzero denoted by $\boldsymbol{\alpha}_{10} = (\alpha_{01}, \ldots, \alpha_{0q_{1n}})^{\mathrm{T}}$.

For notational convenience, we suppress from now on the subscript n in p_n, q_n, p_{1n}, q_{1n}, and N_n, which are allowed to diverge as the sample size n grows.

2.2 Oracle Estimator and Asymptotic Properties

We start with the oracle estimator when the relevant variables of dimensions p_1 in the nonparametric component and q_1 in the partially linear component are known in advance. In the asymptotic properties we establish below, we allow both p_1 and q_1 to diverge with the sample size to accommodate more complex data in high dimensions. With the spline smoothing and $\eta(\cdot)$ being approximated by $\mathbf{\Pi}(\cdot)^{\mathrm{T}}\boldsymbol{\theta}$, we minimize

$$\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{\Pi}(\mathbf{z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1)^{\mathrm{T}}\boldsymbol{\theta} - \mathbf{x}_{1i}^{\mathrm{T}}\boldsymbol{\alpha}_1), \tag{3}$$

with the constraint $\|\boldsymbol{\beta}_1\| = 1$ and the first element positive for the identifiability purposes. Here $\rho_\tau(v) = \tau v - vI(v < 0)$ is the check loss function for quantile regression. $\mathbf{z}_{1i}$ and $\mathbf{x}_{1i}$ denote the i-th row vector of the corresponding important covariates in the nonparametric and parametric components respectively. The oracle estimators for $\boldsymbol{\beta}_0$ and $\boldsymbol{\alpha}_0$ are $(\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \mathbf{0}_{p_n-p_{1n}}^{\mathrm{T}})^{\mathrm{T}}$ and $(\widehat{\boldsymbol{\alpha}}_1^{\mathrm{T}}, \mathbf{0}_{q_n-q_{1n}}^{\mathrm{T}})^{\mathrm{T}}$ respectively.

We adopt the "delete-one-component" method (Yu and Ruppert (2002); Cui et al. (2011)) to satisfy the identifiability constraint on nonzero single-index parameters $\boldsymbol{\beta}_1$. Denote $\boldsymbol{\beta}_1 = ((1 - \|\boldsymbol{\beta}_1^{(-1)}\|^2)^{1/2}, \beta_2, \ldots, \beta_{p_1})^{\mathrm{T}}$ where $\boldsymbol{\beta}_1^{(-1)} = (\beta_2, \ldots, \beta_{p_1})^{\mathrm{T}}$ is a sub-vector of $\boldsymbol{\beta}_1$ without the first component. Thus $\boldsymbol{\beta}_1$ is a function of $\boldsymbol{\beta}_1^{(-1)}$. The $p_1 \times (p_1 - 1)$ Jacobian matrix is

$$\widetilde{\mathbf{J}} = \frac{\partial \boldsymbol{\beta}_1}{\partial \boldsymbol{\beta}_1^{(-1)}} = \begin{pmatrix} -\dfrac{\boldsymbol{\beta}_1^{(-1)}}{(1-\|\boldsymbol{\beta}_1^{(-1)}\|^2)^{1/2}} \\ \mathbf{I}_{(p_1-1)\times(p_1-1)} \end{pmatrix},$$

where $\mathbf{I}_{(p_1-1)\times(p_1-1)}$ is the $(p_1 - 1) \times (p_1 - 1)$ identity matrix. Let $\mathbf{J} = \mathrm{diag}\{\widetilde{\mathbf{J}}, \mathbf{I}_{q_1\times q_1}\}$. Equivalently, we regard $\boldsymbol{\beta}_1$ as a function of $\boldsymbol{\beta}_1^{(-1)}$ and optimize (3) over $(\boldsymbol{\beta}_1^{(-1)}, \boldsymbol{\alpha}_1, \boldsymbol{\theta})$.

For the proofs of convergence rate and asymptotic normality, we need to orthogonalize the parametric part with respect to the nonparametric part using the following projection.

Let $\mathcal{M} = \{m : m(\mathbf{z}) = f(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}), Em^2(\mathbf{z}) < \infty\}$ be the space of single-index functions. In this paper, the projection of any random variable Ψ onto $\mathcal{M}$, denoted by $E_{\mathcal{M}}[W]$, is defined as $m(\mathbf{z})$, with m being the minimizer of

$$E[f(0|\mathbf{z}, \mathbf{x})(\Psi - m(\mathbf{z}))^2], \tag{4}$$

with the constraint $m \in \mathcal{M}$. This definition can be extended trivially to the case where $\boldsymbol{\Psi} = (\Psi_1, \ldots, \Psi_p)^{\mathrm{T}}$ is a random vector by $E_{\mathcal{M}}(\boldsymbol{\Psi}) = (E_{\mathcal{M}}(\Psi_1), \ldots, E_{\mathcal{M}}(\Psi_p))^{\mathrm{T}}$.

We impose the following conditions.

(C1) The components of $\mathbf{x}$, $\mathbf{z}$ are bounded random variables. The support of $\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0$ is $[a, b]$ with length $l := b - a$ satisfying $c_1 \leq l \leq c_2 {p_1}^{1/2}$, and the density of $\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0$, say h, satisfies $c_3 < lh(t) < c_4$, for some positive constants $c_1, \ldots, c_4$.

(C2) Let $f(.|\mathbf{z}, \mathbf{x})$ be the conditional density of ε. We assume $f(.|\mathbf{z}, \mathbf{x})$ is bounded and bounded away from zero in a neighborhood of zero, uniformly over the support of $\mathbf{z}, \mathbf{x}$. The derivative of $f(.|\mathbf{z}, \mathbf{x})$ is uniformly bounded in a neighborhood of zero over the support of $\mathbf{z}, \mathbf{x}$.

(C3) The function η is in the Hölder space of order $r \geq 2$. That is $|\eta^{(u)}(x) - \eta^{(u)}(y)| \leq C|x - y|^v$ for $r = u + v$ and u is the largest integer strictly smaller than r, where $\eta^{(u)}$ is the u-th derivative of η. $\|\eta\|_\infty + \|\eta^{(1)}\|_\infty + \|\eta^{(2)}\|_\infty \leq c_5$ for some positive constant c_5, where $\|\cdot\|_\infty$ is the supremum norm for a bounded function.

(C4) Suppose $E_{\mathcal{M}}[z_j \eta'(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)] = f_j(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)$, $1 \leq j \leq p_1$, where η' is the first derivative of $\eta(\cdot)$ for the notational simplicity. The functions f_j are bounded and in the Hölder space of order $r' \geq 1$. The order of the B-spline used satisfies $d \geq \max\{r, r'\} + 1$. The same smoothness condition is satisfied by the component functions of $E_{\mathcal{M}}[\mathbf{x}]$.

(C5) $E[\mathbf{z}_1\mathbf{z}_1^{\mathrm{T}}]$, $E[\mathbf{x}_1\mathbf{x}_1^{\mathrm{T}}]$ and $E\left[f(0|\mathbf{z}_1, \mathbf{x}_1)\begin{pmatrix} \widetilde{\mathbf{J}}^{\mathrm{T}}\mathbf{z}_1\eta'(\mathbf{z}_1^{\mathrm{T}}\boldsymbol{\beta}_{10}) - E_{\mathcal{M}}[\widetilde{\mathbf{J}}^{\mathrm{T}}\mathbf{z}_1\eta'(\mathbf{z}_1^{\mathrm{T}}\boldsymbol{\beta}_{10})] \\ \mathbf{x}_1 - E_{\mathcal{M}}[\mathbf{x}_1] \end{pmatrix}^{\otimes 2}\right]$ are positive definite matrices with eigenvalues bounded away from zero and infinity, where for any matrix $\mathbf{A}$, $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$.

Boundedness of $\mathbf{x}$ is assumed mainly for convenience of proof, which can possibly be replaced by moment conditions with lengthier arguments. Boundedness of $\mathbf{z}$ is tied to our estimation approach, which is typically assumed when using regression splines, since the basis functions are defined on a compact interval. Given that components of $\mathbf{z}$ are bounded and $\boldsymbol{\beta}_0$ has unit norm, it automatically follows that $l = O({p_1}^{1/2})$. Since $l = b - a$ is diverging, the usual condition that $h(t)$ is bounded away from zero and infinity is herein replaced by that $lh(t)$ is bounded away from zero and infinity. Assumption (C2) on conditional density is commonly used in quantile regression (He and Shi (1994); Wang et al. (2009a)). Smoothness of η is required for the proof of convergence rate. Although it is not usually explicitly stated that η, η' and η'' are bounded functions when the dimension of $\boldsymbol{\beta}$ is fixed, here we add this assumption in the context of high dimensional models, since the function η is not considered fixed in the high-dimensional setting (the support of η is diverging). For (C4), we note that smoothness of functions in the representation of $E_{\mathcal{M}}[X_j \eta'(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_0)]$ is

usually used in semiparametric models to show the asymptotic normality of the parametric part. Finally (C5) can be regarded as an identifiability assumption for semiparametric models, also see Carroll et al. (1997); Li (2000); Wei and He (2006); Wang et al. (2011).

Theorem 1 *Under conditions (C1)-(C5) and that* $N \to \infty$, $N^3 p_1/n \to 0$, $(N+p_1+q_1)^2 log^2(n)/n \to 0$, $(N+p_1+q_1)^{3/2} N^r log(n)/n \to 0$, $(p_1+q_1)/N^{r'} \to 0$, *there is a local minimizer of* (3) *with*

$$\|\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}\| + \|\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_{10}\| + \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(\sqrt{(N+p_1+q_1)/n} + N^{-r}).$$

In particular, $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(\sqrt{(N+p_1+q_1)/n} + N^{-r})$ *implies that* $\|\widehat{\eta} - \eta\| = O_p(\sqrt{(N+p_1+q_1)/n} + N^{-r})$, *with* $\widehat{\eta}(\cdot) = \boldsymbol{\Pi}(\cdot)^{\mathrm{T}} \widehat{\boldsymbol{\theta}}$.

The convergence rate above takes a familiar form as in nonparametric regression with the two terms corresponding to bias and variance respectively. When p_1, q_1 are fixed (or $p_1, q_1 = O(n^{1/(2r+1)})$), the optimal choice of N is obviously $N \sim n^{1/(2r+1)}$. Under stronger assumptions on the choice of N and smoothness of nonparametric functions, we have the asymptotic normality of the parameters $\boldsymbol{\beta}_{10}$ and $\boldsymbol{\alpha}_{10}$. Note that when r' is large enough (for example $r' = r$), $N \sim n^{1/(2r+1)}$ is still contained in the permissible range.

Theorem 2 *Under conditions for Theorem 1 and in addition that* $(N^4 + p_1^6 + q_1^6) log(n^2)/n \to 0$, $(p_1^5 + q_1^5) log(n^2)/N^{2r-2} \to 0$, $\sqrt{n}(p_1+q_1)/N^{2r-1} \to 0$, $\sqrt{n(p_1+q_1)}/N^{r+r'} \to 0$, *for any unit vector* $\mathbf{a}$, *we have*

$$\sqrt{n}\mathbf{a}^{\mathrm{T}} \mathbf{W}^{-1/2} (\mathbf{J}^{\mathrm{T}}\mathbf{J})^{-1} \mathbf{J}^{\mathrm{T}} \left(\begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\alpha}}_1 \end{pmatrix} - \begin{pmatrix} \boldsymbol{\beta}_{10} \\ \boldsymbol{\alpha}_{10} \end{pmatrix} \right) \xrightarrow{d} N(0,1),$$

where

$$\mathbf{W} = (\mathbf{J}^{\mathrm{T}}\boldsymbol{\Omega}\mathbf{J})^{-1} \mathbf{J}^{\mathrm{T}} \boldsymbol{\Delta} \mathbf{J} (\mathbf{J}^{\mathrm{T}}\boldsymbol{\Omega}\mathbf{J})^{-1}),$$

$$\boldsymbol{\Omega} = E\left[f(0|\mathbf{z}_1, \mathbf{x}_1) \begin{pmatrix} \eta'(\mathbf{z}_1^{\mathrm{T}}\boldsymbol{\beta}_{10})\mathbf{z}_1 - E_{\mathcal{M}}[\eta'(\mathbf{z}_1^{\mathrm{T}}\boldsymbol{\beta}_{10})\mathbf{z}_1] \\ \mathbf{x}_1 - E_{\mathcal{M}}[\mathbf{x}_1] \end{pmatrix}^{\otimes 2} \right],$$

$$\boldsymbol{\Delta} = \tau(1-\tau) E\left[\begin{pmatrix} \eta'(\mathbf{z}_1^{\mathrm{T}}\boldsymbol{\beta}_{10})\mathbf{z}_1 - E_{\mathcal{M}}[\eta'(\mathbf{z}_1^{\mathrm{T}}\boldsymbol{\beta}_{10})\mathbf{z}_1] \\ \mathbf{x}_1 - E_{\mathcal{M}}[\mathbf{x}_1] \end{pmatrix}^{\otimes 2} \right],$$

and the Jacobian matrix $\mathbf{J}$ *is evaluated at the true* $\boldsymbol{\beta}_{10}$.

Remark 1 *In the statement of the above theorems, there are many constraints on* N, p_1, q_1. *Assuming the functions are sufficiently smooth* ($r = r'$ *being sufficiently large) and* $N \sim n^{1/(2r+1)}$ *such that* N^{2r-2}, $N^{r+r'}$ *are close to* n, *we see that these conditions roughly constrain* p_1, q_1 *to be of order* $n^{1/6}$.

3. Penalized Estimation for Single-index Quantile Models with Ultra-High Dimensional Covariates

3.1 Asymptotic Properties in Ultra-High Dimension

In reality, we do not have all the relevant covariates known a priori. In order to identify the important covariates both in the nonparametric part and the parametric part in an ultra-high dimensional setting, we propose to estimate parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ by minimizing the penalized loss function via a non-convex penalty:

$$Q(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{\Pi}(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\theta} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha}) + n\sum_{j=1}^{p_n} p_{\lambda_1}(|\beta_j|) + n\sum_{j=1}^{q_n} p_{\lambda_2}(|\alpha_j|). \quad (5)$$

Here $p_\lambda(\cdot)$ is a penalty function and λ_1 and λ_2 are the associated tuning parameters that control the amount of shrinkage in parameters for the nonparametric and partially linear components respectively. Many penalty functions are available in the literature including the adaptive Lasso (Zou, 2006), SCAD penalty (Fan and Li, 2001), and minimax concave penalty (MCP) (Zhang, 2010). In this paper, we consider the commonly used penalty function in high-dimensional models, the SCAD function, which is defined as

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

for some $a > 2$.

Now we consider penalized estimators in which p and q can be ultra-high dimensional while the dimension of important covariates p_1 and q_1 are diverging. Recall that the nonzero components of single-index parameters $\boldsymbol{\beta}$ and linear parameters $\boldsymbol{\alpha}$ are represented by $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_{p_1})^{\mathrm{T}}$ and $\boldsymbol{\alpha}_1 = (\alpha_1, \ldots, \alpha_{q_1})^{\mathrm{T}}$, respectively. The following theorem presents the oracle property (Fan and Li, 2001) of the penalized estimator of (5). That is, the asymptotic normality property is the same as when the nonzero components in $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are known in advance.

Theorem 3 *Under the same conditions assumed for Theorem 2, and that* $log(p) + log(q) = o(n^c)$ *for some* $c \in (0, 1/2)$, $\sqrt{N + p_1 + q_1}\cdot\xi_n << \lambda_1 << \min_{j \le p_1} |\beta_{0j}|$, $\sqrt{N + p_1 + q_1}\cdot\xi_n << \lambda_2 << \min_{j \le q_1} |\alpha_{0j}|$, *where* $\xi_n = \sqrt{(N + p_1 + q_1)/n} + N^{-r}$, *there is a* ξ_n*-consistent local minimizer of (5), say* $(\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\theta}})$, *such that for any unit vector* $\mathbf{a}$,

(i)

$$\sqrt{n}\mathbf{a}^{\mathrm{T}}\mathbf{W}^{-1/2}(\mathbf{J}^{\mathrm{T}}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{T}}\left(\begin{pmatrix} \widetilde{\boldsymbol{\beta}}_1 \\ \widetilde{\boldsymbol{\alpha}}_1 \end{pmatrix} - \begin{pmatrix} \boldsymbol{\beta}_{10} \\ \boldsymbol{\alpha}_{10} \end{pmatrix}\right) \xrightarrow{d} N(0, 1),$$

where $\mathbf{W}$ *and* $\mathbf{J}$ *are defined as in Theorem 2.*

(ii) $\widetilde{\beta}_{p_1+1} = \cdots = \widetilde{\beta}_p = \widetilde{\alpha}_{q_1+1} = \cdots = \widetilde{\alpha}_q = 0$ *with probability approaching one.*

3.2 An Efficient Algorithm

The estimator of parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ is defined to be the minimizer of the objective function (5). In computation, we face a number of challenges such as the non-smooth and

nonconvex optimization problem in addition to the high-dimensional model setting. More importantly, one main challenge for single-index modeling is the high nonlinearity due to the fact that the high-dimensional parameters are embedded inside the unknown function $\eta(\cdot)$ estimated nonparametrically. Hence, the straightforward "one-step" estimation by minimizing $Q(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$ over all parameters may not work, especially for the high-dimensional problem.

We propose an efficient algorithm with an iterative approach. The key idea is to utilize a linear approximation of $\eta(\cdot)$ to turn the problem into essentially a penalized linear quantile problem so that the algorithms for linear quantiles in the literature can be readily adopted. To be more specific, we adopt the linear approximation of $\eta(\cdot)$ in estimating the single-index parameters $\boldsymbol{\beta}$, that is, expanding $\eta(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta})$ to its first order around the value $\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_0$:

$$\eta(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}) \cong \eta(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_0) + \eta'(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_0)\mathbf{z}_i^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0).$$

We then employ polynomial splines with B-spline basis to estimate the terms associated with the link function η. Specifically given $\boldsymbol{\beta}$, $\eta(\cdot)$ and $\eta'(\cdot)$ will be estimated respectively by $\boldsymbol{\Pi}(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_0)^{\mathrm{T}}\boldsymbol{\theta}$ and $\boldsymbol{\Pi}'(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}_0)^{\mathrm{T}}\boldsymbol{\theta}$, in which $\boldsymbol{\Pi}'$ is the first derivative of B-spline basis functions. With the proposed linear approximation to the unknown function $\eta(\cdot)$, we obtain an appealing explicit form where the single-index parameters $\boldsymbol{\beta}$ shows up in the second term. This will essentially benefit computing the penalized estimation from the objective function (5). In particular, computationally the linear approximation effectively turns the highly nonlinear semiparametric quantile problem into equivalently a linear quantile problem.

Then for given the parameters $(\boldsymbol{\beta}, \boldsymbol{\alpha})$, the estimates of spline coefficients $\boldsymbol{\theta}$ can be obtained by minimizing $\sum_{i=1}^{n} \rho_\tau \left(y_i - \boldsymbol{\Pi}(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{\theta} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha}\right)$, where B-spline basis $\boldsymbol{\Pi}(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta})$ is analogous to the design matrix that can be essentially viewed as a linear quantile regression to obtain spline coefficients $\boldsymbol{\theta}$.

Lastly, we need to deal with the nonconvex SCAD penalty. Although the nonconvex type of penalty functions, such as SCAD penalty, makes the shrinkage more effective compared with the direct L_1 penalty, the nonconvexity causes extra computational burden in the high-dimensional optimization problem. To tackle this challenge, we adopt a local linear approximation algorithm (LLA) (Zou and Li, 2008) for the penalty terms $p_{\lambda_1}(|\beta_j|), 1 \le j \le p_n$ and $p_{\lambda_2}(|\alpha_j|), 1 \le j \le q_n$:

$$p_{\lambda_1}(|\beta_j|) \approx p_{\lambda_1}(|\widehat{\beta}_j^{(0)}|) + p'_{\lambda_1}(|\widehat{\beta}_j^{(0)}|)(|\beta_j| - |\widehat{\beta}_j^{(0)}|), for\ \beta_j \approx \widehat{\beta}_j^{(0)},$$

$$p_{\lambda_2}(|\alpha_j|) \approx p_{\lambda_2}(|\widehat{\alpha}_j^{(0)}|) + p'_{\lambda_2}(|\widehat{\alpha}_j^{(0)}|)(|\alpha_j| - |\widehat{\alpha}_j^{(0)}|), for\ \alpha_j \approx \widehat{\alpha}_j^{(0)},$$

where $\widehat{\beta}_j^{(0)}$ and $\widehat{\alpha}_j^{(0)}$ are given initial values. Under the linear approximation to $\eta(\cdot)$, the modified penalized objective function $\widetilde{Q}(\boldsymbol{\beta}, \boldsymbol{\alpha})$ regarding $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ with the estimated spline coefficients $\widehat{\boldsymbol{\theta}}$ and some initial values $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}})$ can be formulated as

$$\sum_i \rho_\tau(y_i - \boldsymbol{\Pi}(\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\widehat{\boldsymbol{\theta}} - \boldsymbol{\Pi}'(\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\widehat{\boldsymbol{\theta}}\mathbf{z}_i^{\mathrm{T}}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha}) + n\sum_{j=1}^{p_n} p'_{\lambda_1}(|\widehat{\beta}_j|)|\beta_j| + n\sum_{j=1}^{q_n} p'_{\lambda_2}(|\widehat{\alpha}_j|)|\alpha_j|. \tag{6}$$

Define $\widetilde{y}_i = y_i - \mathbf{\Pi}(\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\widehat{\boldsymbol{\theta}} + \mathbf{\Pi}'(\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\widehat{\boldsymbol{\theta}}\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$ and $\widetilde{\mathbf{z}}_i = \mathbf{\Pi}'(\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\widehat{\boldsymbol{\theta}}\mathbf{z}_i$, then we can rewrite (6) as

$$\widetilde{Q}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \rho_\tau(\widetilde{y}_i - \widetilde{\mathbf{z}}_i^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha}) + n\sum_{j=1}^{p_n} p'_{\lambda_1}(|\widehat{\beta}_j|)|\beta_j| + n\sum_{j=1}^{q_n} p'_{\lambda_2}(|\widehat{\alpha}_j|)|\alpha_j|.$$

We note that (6) is a convex problem, in fact, it is essentially a penalized linear quantile regression.

In summary, the iterative algorithm we propose to use can be carried out as follows:
Step 0. Initialize $(\widehat{\boldsymbol{\beta}}^{(0)}, \widehat{\boldsymbol{\alpha}}^{(0)})$.
Step 1. Given $\widehat{\boldsymbol{\beta}}^{(k-1)}$, construct B-spline basis functions $\mathbf{\Pi}(\mathbf{z}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(k-1)})$, then the spline coefficient estimates are obtained from $\widehat{\boldsymbol{\theta}}^{(k)} = \arg\min \sum_{i=1}^{n} \rho_\tau\big(y_i - \mathbf{\Pi}(\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(k-1)})^{\mathrm{T}}\boldsymbol{\theta} - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\alpha}}^{(k-1)}\big)$.
Step 2. Given the estimated spline coefficients $\widehat{\boldsymbol{\theta}}^{(k)}$, the kth-step penalized estimator of the single-index parameters $\widehat{\boldsymbol{\beta}}^{(k)}$ and partially linear parameters $\widehat{\boldsymbol{\alpha}}^{(k)}$ will be achieved by the minimization of

$$\widetilde{Q}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \rho_\tau(\widetilde{y}_i - \widetilde{\mathbf{z}}_i^{\mathrm{T}}\boldsymbol{\beta} - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha}) + n\sum_{j=1}^{p_n} p'_{\lambda_1}(|\widehat{\beta}_j^{(k-1)}|)|\beta_j| + n\sum_{j=1}^{q_n} p'_{\lambda_2}(|\widehat{\alpha}_j^{(k-1)}|)|\alpha_j|, \quad (7)$$

where $\widetilde{y}_i = y_i - \mathbf{\Pi}(\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(k-1)})^{\mathrm{T}}\widehat{\boldsymbol{\theta}}^{(k)} + \mathbf{\Pi}'(\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(k-1)})^{\mathrm{T}}\widehat{\boldsymbol{\theta}}^{(k)}\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(k-1)}$ and
$\widetilde{\mathbf{z}}_i = \mathbf{\Pi}'(\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(k-1)})^{\mathrm{T}}\widehat{\boldsymbol{\theta}}^{(k)}\mathbf{z}_i$.
Repeat Steps 1 and 2 until convergence.

To initialize the iterative algorithm, one may use estimates from penalized linear regression or the linear quantile model, such as $\mathcal{Q}_\tau(y_i) = \mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha}$. In this article, we adopted the estimator from single-index mean regression with candidates selected by iterative sure independence screening (Fan and Lv, 2008). We normalize $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta}\| = 1$, and its first nonzero element is positive for identifiability. These types of initial values work reasonably well both in the simulation studies and in the real-data application to the gene expression data. Again, trying different starting values is necessary in any optimization problem in general.

Finally, we adopt a data augmentation technique by introducing pseudo-observations for Step 2. This is based on two facts that $c\rho_\tau(v) = \rho_\tau(cv)$, for $c > 0$ and $|\beta_j|$ can be written as $\rho_\tau(\beta_j) + \rho_\tau(-\beta_j)$ (Wu and Liu, 2009; Wang et al., 2012; Sherwood and Wang, 2016). Let $n \cdot p'_{\lambda_1}(|\widehat{\beta}_j^{(k-1)}|) = c_j$. Then we can rearrange the first penalty part in (7) as $\sum_{j=1}^{p_n}(\rho_\tau(-c_j\beta_j) + \rho_\tau(c_j\beta_j))$. Furthermore, an "unpenalized" linear quantile regression can be implemented by using the extra pseudo-observations. Denote by $\widetilde{y}_i^+$ the new response $\widetilde{y}_i^+ = \widetilde{y}_i$, $i = 1, \ldots, n$; $\widetilde{y}_i^+ = 0$, $i = n+1, \ldots, n+2p_n$ and by $\widetilde{\mathbf{z}}_i^+ \in R^{p_n}$ the new response $\widetilde{\mathbf{z}}_i^+ = \widetilde{\mathbf{z}}_i$, $i = 1, \ldots, n$; $\widetilde{\mathbf{z}}_i^+ = (0, \ldots, 0, c_i, 0, \ldots, 0)$, $i = n+1, \ldots, n+p_n$; $\widetilde{\mathbf{z}}_i^+ = (0, \ldots, 0, -c_i, 0, \ldots, 0)$, $i = n+p_n+1, \ldots, n+2p_n$. Similarly, the extra $2q_n$ pseudo-observations will be created in the same fashion for the associated penalty terms $np'_{\lambda_2}(|\widehat{\alpha}_j^{(k-1)}|)|\alpha_j|$ in Step 2. Note $\widetilde{y}_i^+ = 0$ and $\widetilde{\mathbf{z}}_i^+ = (0, \ldots, 0)$, $i = n+2p_n+1, \ldots, n+2p_n+2q_n$, while the augmented data for linear covariates are denoted by $\widetilde{\mathbf{x}}^+$. With all the above, equation (7) reduces to a familiar form as $\sum_{i=1}^{n+2p_n+2q_n} \rho_\tau(\widetilde{y}_i^+ - \widetilde{\mathbf{z}}_i^{+\mathrm{T}}\boldsymbol{\beta} - \widetilde{\mathbf{x}}_i^{+\mathrm{T}}\boldsymbol{\alpha})$, which is

indeed a linear quantile regression and can be solved by many existing statistical software packages, for instance, the R *quantreg* package by Roger Koenker et. al.

3.3 Local convergence of the computable estimator in moderately-high dimension

The obvious theoretical gap in the previous subsection is that we only established asymptotic property of a certain local estimator and there is no guarantee that we can obtain such a local estimator. In general, for nonconvex optimization functions there may be many stationary points or local minimizers and one does not expect all local minimizers will have desirable asymptotic properties. The statistics, optimization and machine learning literature contains some success stories for some specific nonconvex optimization problems combined with specific algorithms, for example alternating minimization or projected/proximal gradient methods (Jain et al., 2014; Bhatia et al., 2015; Gu et al., 2016; Chen et al., 2019). Many such results are aimed at showing the convergence of iterative methods to the stationary point, local minimizer, or some cases even global minimizer of the optimization problem. It often requires some special structure of the problem such as restricted isometry property or incoherence property. Focusing on the statistical properties, Fan et al. (2014a) showed that for the high-dimensional parametric regression (both strongly smooth and strongly convex conditions are satisfied for the loss), linearization for the SCAD penalty or MCP leads to consistent and efficient estimator in only two iterations. A good initial estimator is very easy to obtain in their setting even in ultra-high dimensions (using LASSO, for example). Our approach here is similar to this work. Agarwal et al. (2012) considered proximal gradient method which relaxed strongly convex and strongly smooth condition of the loss function in high dimensions (but when reduced to low dimensions, it still requires strong convexity and strong smoothness). Our case is more complicated due to the nonconvex single-index term even if we restrict to the fixed-dimensional case and thus a lot of such techniques in the literature do not easily apply here. Xu and Yin (2017) has an interesting convergence result for very general convex loss function, but their Kurdyka-Lojasiewicz condition is hard to verify for specific problems, and furthermore such convergence results do not provide statistical properties of the estimators.

Partially motivated by Fan et al. (2014a), here we show that when the dimension (p, q) is diverging although not larger than n, and an initial estimator is available that is in a $O(1/\sqrt{N(p+q)})$ neighborhood of $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}_0)$, then the iterative algorithm we use will produce estimators that are guaranteed to have the same statistical properties as stated previously, and the number of iterations required is of order $O(\log n)$. We note that when $p > n$, there are several screening methods that can be used to reduce the dimension to $p < n$ with some theoretical guarantees (Fan and Lv, 2008; Fan et al., 2011a). This requires further assumptions of course.

As mentioned above, to establish our result, an initial estimator with convergence rate $O_p(1/\sqrt{N(p+q)})$ is necessary. In general, a good initial estimator $(\widehat{\boldsymbol{\beta}}^{(0)}, \widehat{\boldsymbol{\alpha}}^{(0)})$ is hard to obtain for single-index models. Fortunately, in some situations, a lot of existing proposed methods in sufficient dimension reduction (SDR) can be utilized (Li, 1991; Li and Nachtsheim, 2006). Technically, this often requires ellipticity condition on the distribution of the covariates (which guarantees the so-called linearity condition in the SDR literature),

but is empirically observed to work even when this is violated. In particular, both mean regression and quantile regression (Duan and Li, 1991; Lian et al., 2019) can extract the directions when the model has the form $y = f(\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}, \mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}, \varepsilon)$ with ε representing the noise and f is a general nonparametric function. When the dimension is diverging, it is a standard result that linear regression or quantile linear regression produces an initial estimator with $\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0\| + \|\widehat{\boldsymbol{\alpha}}^{(0)} - \boldsymbol{\alpha}_0\| = O_p(\sqrt{(p+q)/n})$ (Portnoy, 1985; Fan et al., 2014b) which is indeed $O_p(1/\sqrt{N(p+q)})$ under our assumptions. In practice we find that such simple parametric regression works well without the complication of using other more complicated SDR approaches.

From these discussions, the following high-level assumptions regarding the initial estimators are used.

(B1) $\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0\| + \|\widehat{\boldsymbol{\alpha}}^{(0)} - \boldsymbol{\alpha}_0\| = O_p(\frac{1}{\sqrt{N(p+q)}})$, $\frac{N^2(p+q)^2}{n} \to 0$, and $\frac{p+q}{N^{2(r-1)}} \to 0$.

Note that as in Theorem 2 we still require $p_1 + q_1$ is smaller than $n^{1/6}$, but p and q can be close to $n^{1/2}$ if r is large enough.

Theorem 4 *Suppose we use the iterative algorithm as explained in Section 3.2. Under the conditions of Theorem 3 as well as (B1), if* $\sqrt{N + p_1 + q_1} \cdot (\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0\| + \|\widehat{\boldsymbol{\alpha}}^{(0)} - \boldsymbol{\alpha}_0\| + N^{-r}) << \lambda_1 << \min_{j \le p_1} |\beta_{0j}|$, $\sqrt{N + p_1 + q_1} \cdot (\|\widehat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0\| + \|\widehat{\boldsymbol{\alpha}}^{(0)} - \boldsymbol{\alpha}_0\| + N^{-r}) << \lambda_2 << \min_{j \le q_1} |\alpha_{0j}|$, *then we have with probability approaching one* $\widehat{\beta}_j^{(t)} = 0$ *for* $j > p_1$ *and* $\widehat{\alpha}_j^{(t)} = 0$ *for* $j > q_1$ *for all* $t \ge 1$ *(here* t *indicates the iteration number). Furthermore, for all* $t \ge C \log n$ *with some constant* $C > 0$, $\widehat{\boldsymbol{\beta}}^{(t)}, \widehat{\boldsymbol{\alpha}}^{(t)}$ *has the same asymptotic distribution as* $\widetilde{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\alpha}}$ *in Theorem 3.*

3.4 Tuning Parameter Selection

The tuning parameter λ is important in practice. BIC is a common effective criterion to select λ in the fixed or low-to-moderate dimensional models. We adopt high dimensional BIC when p_n and q_n are potentially ultra-high dimensional (Wang et al., 2009b; Lee et al., 2014). We select λ that will minimize the following high-dimensional BIC criterion:

$$HBIC(\lambda_1, \lambda_2) = \log\left(\sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{\Pi}(\mathbf{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\lambda_1})^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_{\lambda_1} - \mathbf{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\alpha}}_{\lambda_2})\right) + d_\lambda \frac{\log(n)}{2n} C_n, \qquad (8)$$

where d_λ is the total number of non-zero parameters in both single-index part and partially linear terms. As suggested in the literature, C_n is taken to be $\log(\log(p_n + q_n))$ in empirical studies, where p_n is the number of candidate single-index covariates and q_n is the number of candidate partially linear covariates. One nice feature is that we can allow different levels of penalization through different λ_1 and λ_2 for the single-index parameters and linear parameters respectively. In practice, a grid of λ values will be used and for each given set of (λ_1, λ_2) the resulting penalized parameter estimates are associated with these λ values. Then the according $HBIC$ criterion can be calculated, and we will choose the pair of λ values that can minimize the $HBIC$ as in (8). Alternatively, a two-step grid searching algorithm can be used (Ruppert and Carroll, 2000). First, $HBIC$ can be minimized with

a common tuning parameter λ. Then λ_1 can be selected with λ_2 fixed and vice versa. This can save the computational cost by reducing the two-dimensional grid searching to a one-dimensional case.

4. Numerical Studies

We investigate the performance of the proposed single-index quantile models in high dimension. We focus on the non-convex SCAD penalty for selecting important variables. We have also implemented another popular nonconvex penalty, the MCP penalty ((Zhang, 2010)), and find that the simulation results for MCP are similar to SCAD results. For the penalty function part, we select the tuning parameters λ by minimizing the high-dimensional BIC defined in (8). In addition, tuning parameter a in SCAD penalty is set to be 3.7 as suggested by Fan and Li (2001). To estimate the unknown function nonparametrically, cubic B-spline bases with equally-spaced knots are adopted throughout the empirical studies, and the number of interior knots is taken to be two, since we found this works well in our examples. Other data-driven type of knots, for instance, placing knots at sample quantiles of the single-index values, may also be suitable. Other ways to choose an optimal number of knots could be using cross validation or BIC type of criterion, such as in He et al. (2002), Wang et al. (2009a) or Zhang et al. (2017).

We assess the performance of the proposed ultra-high dimensional models through the following criteria:
1. True Negative(TN): the average number of zero covariates that are exactly set to zero.
2. False Negative(FN): the average number of non-zero covariates that are incorrectly set to zero.
3. Correct(%) or C(%): the percentage of times that the true model with exact non-zero covariates is correctly identified.
4. MSE: the average of the mean squared error for estimators of $\boldsymbol{\alpha}_0$ or $\boldsymbol{\beta}_0$, i.e., the average of $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||^2$ for index parameters or the average of $||\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0||^2$ for partially linear parameters over a number of replications.

4.1 Example 1.

(Sine-bump Models.) We generate 100 random samples from the following sine-bump model:

$$y_i = \sin\left\{\frac{(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta} - a)\pi}{b-a}\right\} + 0.1\varepsilon_i,\ i = 1, \ldots, n,$$

where a and b are two constants taking $\sqrt{3}/2 - 1.645/\sqrt{12}$ and $\sqrt{3}/2 + 1.645/\sqrt{12}$ respectively. The true parameter vector is $\boldsymbol{\beta}_0 = (1, 3, 1.5, 0.5, 0, \ldots, 0)_p^{\mathrm{T}}/\sqrt{12.5}$. This model is widely used in the semiparametric modeling literature, for instance, Carroll et al. (1997) and Liang et al. (2010). $\mathbf{z} = (z_1, z_2, \ldots, z_p)$ are independent and uniformly distributed from $U(0, 1)$, and the error term ε is generated from $N(0, 1)$. We take the sample size $n = 200$ while the number of covariates $p = p_n$ varies from 50 to 100 to 1000. Estimation and variable selection results are also shown at different quantile levels, such as median ($\tau = 0.5$), the first quartile ($\tau = 0.25$) and the third quartile ($\tau = 0.75$).

The detailed estimation results for each of the non-zero parameters are shown in Table 1. The sample mean ("mean"), bias ("bias") and standard error ("se") of the parameter

Table 1: Summary of Parameter Estimates for High-dimensional SIMs in Example 1. The true non-zero value of $\boldsymbol{\beta}_0$ is $(0.2828, 0.8485, 0.4243, 0.1414)^{\mathrm{T}}$. The sample mean ("mean"), bias ("bias") and standard error ("se") of the parameter estimates are calculated over 100 simulations when sample size $n = 200$. The numbers of input variables are increasing from 50 to 100 to 1000.

		$\boldsymbol{p = 50}$			$\boldsymbol{p = 100}$			$\boldsymbol{p = 1000}$		
τ	par.	mean	bias	se	mean	bias	se	mean	bias	se
0.25	β_1	0.2842	0.0014	0.0121	0.2873	0.0044	0.0149	0.2865	0.0037	0.0123
	β_2	0.8497	0.0012	0.0067	0.8470	-0.0016	0.0082	0.8471	-0.0014	0.0070
	β_3	0.4206	-0.0037	0.0136	0.4234	-0.0008	0.0145	0.4250	0.0007	0.0116
	β_4	0.1403	-0.0011	0.0163	0.1418	0.0004	0.0160	0.1385	-0.0029	0.0146
0.5	β_1	0.2822	-0.0007	0.0122	0.2844	0.0016	0.0128	0.2839	0.0010	0.0122
	β_2	0.8488	0.0003	0.0075	0.8487	0.0002	0.0076	0.8483	-0.0002	0.0067
	β_3	0.4239	-0.0004	0.0131	0.4216	-0.0026	0.0129	0.4237	-0.0005	0.0110
	β_4	0.1400	-0.0014	0.0141	0.1427	0.0012	0.0134	0.1402	-0.0012	0.0130
0.75	β_1	0.2818	-0.0010	0.0139	0.2831	0.0002	0.0130	0.2853	0.0025	0.0135
	β_2	0.8486	0.0001	0.0076	0.8490	0.0005	0.0078	0.8476	-0.0009	0.0071
	β_3	0.4244	0.0001	0.0140	0.4215	-0.0028	0.0131	0.4244	0.0002	0.0120
	β_4	0.1406	-0.0009	0.0145	0.1440	0.0026	0.0151	0.1392	-0.0022	0.0151

estimates are calculated over 100 simulations for each quantile level and different numbers of input variables p. We find the estimation for non-zero index parameters is reliable when p increases or becomes larger than sample size n, considering the average of parameter estimates are close to the true values and standard errors are relatively small. To evaluate how the penalized method works for the high dimensional case, we report the summary of variable selection results in Table 2. TN by definition is $p - 4$ for each setting. Table 2 shows that our model captures most of zero covariates as indicated by Correct(%) reaching 100% in most cases. On the other hand, our high-dimensional single-index quantile model successfully retains all relevant covariates since there is no relevant covariates being set to zero, i.e. FN is 0 for all runs. Overall, MSEs are very small because zero covariates are identified in most of the runs, and estimates for non-zero parameters are quite accurate as shown in Table 1.

Furthermore, we plot the fitted curves for function $\eta(\cdot)$ in single-index median regression in Figure 1. On the left panel, we randomly select a simulated sample and plot the single fit as the dot-dash line. On the right panel, we plot the average fitted curve and corresponding 95% pointwise confidence bands over the 100 simulations. From both the single fit and the average fit, the fitted curves virtually overlay with the true sine function curve, which indicates that our nonparametric spline approximation to the unknown function works well.

Table 2: Summary of Variable Selection Results for High-dimensional SIMs in Example 1. "hdsim" represents the variable selection performance using the proposed penalized estimation for high-dimensional single-index models, while "oracle" indicates the fitted models only use the 4 relevant variables.

$n = 200$	quantile	Model	TN	FN	Correct(%)	MSE(10^{-4})
	0.25	hdsim	46	0	100	6.53
$\boldsymbol{p = 50}$		oracle	46	0	100	6.67
	0.5	hdsim	46	0	100	5.72
		oracle	46	0	100	5.51
	0.75	hdsim	46	0	100	6.50
		oracle	46	0	100	6.58
	0.25	hdsim	96	0	100	7.70
$\boldsymbol{p = 100}$		oracle	96	0	100	7.03
	0.5	hdsim	95.8	0	85	6.71
		oracle	96	0	100	5.56
	0.75	hdsim	96	0	100	6.39
		oracle	96	0	100	5.83
	0.25	hdsim	996	0	100	5.66
$\boldsymbol{p = 1000}$		oracle	996	0	100	5.03
	0.5	hdsim	995.89	0	90	5.26
		oracle	996	0	100	5.64
	0.75	hdsim	996	0	100	6.11
		oracle	996	0	100	5.51

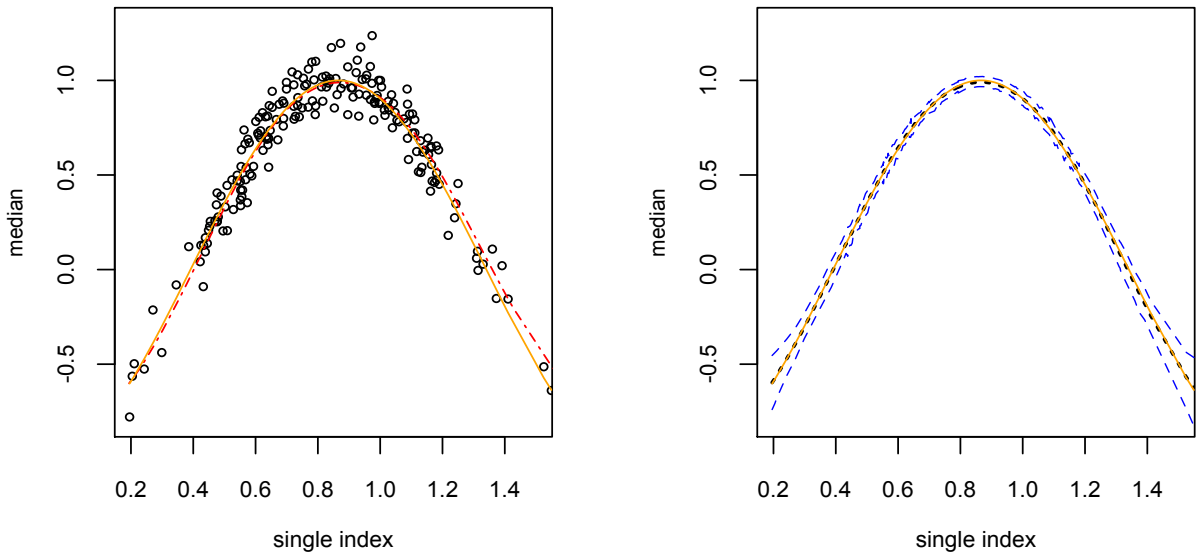

Figure 1: Fitted Curves for $\eta(\cdot)$ in Example 1 for HDSIM Median Regression when $p_n = 1000$. The orange solid lines in both plots are the curve of the true sine function. The dot-dash line on the left panel is the fitted curve from one simulated sample. On the right panel the dotted lines are the average fitted curves while the dashed curves are the corresponding 2.5% and 97.5% confidence bands over 100 simulations.

4.2 Example 2.

(Heteroscedastic Models) In this example, we consider the following heteroscedastic regression model ((Wang and Wang, 2015)). 100 simulations are generated from:

$$y_i = \sin(\frac{\pi}{4}\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta}) + \sigma\frac{5-\exp(||\mathbf{z}||/\sqrt{d})}{5+\exp(||\mathbf{z}||/\sqrt{d})}\varepsilon_i,\ i = 1,\ldots,n,$$

where $\sigma = 0.2$, $d = 5$ for 5 non-zero single-index parameters and the true parameter is $\boldsymbol{\beta} = (1,1,1,1,1,0,\ldots,0)^{\mathrm{T}}/\sqrt{5}$. $\mathbf{z}_1,\ldots,\mathbf{z}_{p_n},\varepsilon$ are independently and identically distributed as $N(0,1)$. We consider scenarios with various p and sample size n.

The simulation results at three quantile levels, that is, 0.25, 0.5, and 0.75 are presented in Table 3 for different n and p. We compare the variable selection and estimation results of the high dimensional single-index model as "hdsim" with those of the "oracle" model in which the exact important covariates are used to fit the single-index quantile model. For all sizes of the heteroscedastic models, there are no significant covariates being excluded by the SCAD penalty since the number of false negative(FN) are all zero. Occasionally, an extra one or two irrelevant covariates are kept in the model when the true negative(TN) is calculated. Overall, the MSEs of "hdsim" are very close to the MSEs of oracle models, which further indicates that our proposed penalized estimators for high-dimensional heteroscedastic models are consistent with oracle estimators.

4.3 Example 3.

(PLSIM Models.) We generate 100 random samples from the following partially linear single-index model:

$$y_i = \sin\left\{(\mathbf{z}_i^{\mathrm{T}}\boldsymbol{\beta})\pi\right\} + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha} + 0.1\varepsilon_i,\ i = 1,\ldots,n.$$

The true value for the single-index parameter vector is $\boldsymbol{\beta}_0 = (1,1,1,1,1,0,\ldots,0)^{\mathrm{T}}/\sqrt{5}$ and the true linear parameter vector is $\boldsymbol{\alpha}_0 = (3,2,0,1,0,0,0,-1,0,\ldots,0)$. Here the number of important covariates $p_1 = 5$ and $q_1 = 4$ respectively. All covariates including $\mathbf{z}$ and $\mathbf{x}$ are independent and uniformly distributed from $U(0,1)$, and the error terms ε are generated from $N(0,1)$. We again consider different numbers of parameters p_n when the sample size $n = 200$. Estimation and variable selection results are also shown at different quantile levels, namely, median ($\tau = 0.5$), the first quartile ($\tau = 0.25$) and the third quartile ($\tau = 0.75$).

Table 4 displays the variable selection results for high dimensional partially linear single-index models. Note that ideally TN is $p_n - 5$ for single-index components and $q_n - 4$ for partially linear components. Most of the zero covariates in both parts are found, except that in some simulation replicates, one or two zeroes in single-index part or linear part is occasionally not excluded. All relevant covariates in both parts are correctly identified as indicated by FN being 0 for all settings. The MSEs are also very small for all parameter estimates. To save space, we relegate the parameter estimation results to the supplementary material. We note that not only the magnitudes of parameter estimates are close to the true values, but also the sign of the negative parameter embedded in the partially linear part is correctly identified.

Table 3: Summary of Variable Selection Results for the High-dimensional Heteroscedastic Models in Example 2. "hdsim" represents the variables selection performance using the proposed penalized estimation, while "oracle" indicates only true variables are used in model fitting.

	quantile	Model	TN	FN	Correct(%)	MSE(10^{-4})
			$\boldsymbol{n = 100}$			
	0.25	hdsim	44.97	0	97	7.46
$\boldsymbol{p = 50}$		oracle	45	0	100	6.78
	0.5	hdsim	44.85	0	88	6.18
		oracle	45	0	100	4.88
	0.75	hdsim	45	0	100	5.62
		oracle	45	0	100	5.26
	0.25	hdsim	94.97	0	98	5.95
$\boldsymbol{p = 100}$		oracle	95	0	100	5.48
	0.5	hdsim	94.86	0	89	4.89
		oracle	95	0	100	4.45
	0.75	hdsim	94.95	0	95	6.52
		oracle	95	0	100	5.29
			$\boldsymbol{n = 200}$			
	0.25	hdsim	44.95	0	95	3.02
$\boldsymbol{p = 50}$		oracle	45	0	100	2.72
	0.5	hdsim	44.9	0	91	3.21
		oracle	45	0	100	2.36
	0.75	hdsim	45	0	100	3.21
		oracle	45	0	100	3.20
	0.25	hdsim	95	0	100	3.23
$\boldsymbol{p = 100}$		oracle	95	0	100	3.30
	0.5	hdsim	94.92	0	95	2.95
		oracle	95	0	100	2.89
	0.75	hdsim	95	0	100	3.75
		oracle	95	0	100	3.51
	0.25	hdsim	994.97	0	92	2.91
$\boldsymbol{p = 1000}$		oracle	995	0	100	2.81
	0.5	hdsim	994.86	0	92	3.33
		oracle	995	0	100	2.76
	0.75	hdsim	994.95	0	92	3.42
		oracle	995	0	100	3.29

Table 4: Summary of Variable Selection Results for Partially Linear Single-index Models in Example 3. $p_n + q_n$ is the total number of parameters in the partially linear single-index models when sample size $n = 200$.

			single-index components				linear components			
τ	$p_n + q_n$	C(%)	p_n	TN	FN	MSE(10^{-3})	q_n	TN	FN	MSE(10^{-3})
0.25	300	91	100	94.91	0	0.68	200	196	0	3.28
0.5	300	94	100	94.94	0	0.62	200	196	0	3.26
0.75	300	93	100	94.92	0	0.67	200	196	0	3.66
0.25	500	92	200	194.92	0	0.92	300	296	0	4.25
0.5	500	88	200	194.86	0	0.75	300	296	0	4.12
0.75	500	91	200	194.89	0	0.96	300	296	0	4.56
0.25	1000	93	500	495	0	0.61	500	495.93	0	4.72
0.5	1000	90	500	494.88	0	0.58	500	496	0	3.46
0.75	1000	94	500	494.93	0	0.62	500	496	0	4.13

4.4 Gene expression data.

In this section, we apply the proposed penalized single-index quantile regression to a real polymerase chain reaction high-dimensional dataset, which is from the experiment conducted by Lan et al. (2006). They examined the genetics of two inbred mouse populations segregating for obesity and diabetes. The sample that was used to monitor the expression level of a total of 22,575 genes consists of a total of 60 subjects with approximately half male mice and half female mice. The gene expression data and phenotype data can be found at GEO (http://www.nci.nih.gov/geo; accession number GSE3330). Some physiological phenotypes are also measured in the real-time polymerase chain reaction data, including the numbers of Phosphoenolpyruvate carboxykinase (PEPCK). We are interested in the level of PEPCK for it is evidenced by laboratory mice that the overexpression of PEPCK-C in mouse's liver results in their contracting diabetes mellitus type 2 according to existing PEPCK researches. Song and Liang (2015) studied this data on the linear relationship between PEPCK and gene expression levels by the reciprocal L_1-regularized mean regression.

To have a broader view of the relationship between PEPCK and the gene expression levels in the sample of size $n = 60$, we study the conditional quantiles of PEPCK. We start with $p = 1000$ genes as covariates that have the highest marginal correlation with PEPCK by the single-index models. The single-index models allow nonlinear relationship between the response PEPCK and all covariates, and the relevant genes will be selected with penalization at the same time. All 1000 covariates are standardized to have mean 0 and variance 1. We consider four types of models with different penalties, namely, the penalized single-index models with SCAD penalty (sim-scad), single-index model with MCP (sim-mcp), linear quantile model with Lasso penalty (linear-lasso) and linear quantile regression

Table 5: Variable Selection and Prediction Comparison on Gene data. "Model Type" shows 4 different combinations of the modeling and variable selection techniques for 3 quantile levels. The prediction errors ("PE") based on the check loss function are computed for the 5 testing data sets for each model with only the selected genes and the mean prediction errors ("Mean PE") are also calculated for each model type. The number of the selected genes for each model is also displayed in "Model Size".

quantile	Model	Predicted Error (PE)					Mean	Model
τ	Type	test1	test2	test3	test4	test5	PE	Size
0.25	linear lasso	0.177	0.160	0.198	0.208	0.391	0.227	60
	linear scad	0.076	0.228	0.129	0.142	0.214	0.158	20
	sim scad	0.059	0.124	0.061	0.072	0.089	0.081	7
	sim mcp	0.073	0.138	0.049	0.087	0.084	0.086	7
0.50	linear lasso	0.173	0.244	0.204	0.152	0.200	0.195	60
	linear scad	0.265	0.273	0.301	0.254	0.292	0.277	18
	sim scad	0.087	0.081	0.092	0.106	0.206	0.114	11
	sim mcp	0.093	0.168	0.092	0.152	0.199	0.141	11
0.75	linear lasso	0.166	0.230	0.248	0.117	0.153	0.183	60
	linear scad	0.126	0.170	0.217	0.190	0.224	0.186	20
	sim scad	0.077	0.090	0.034	0.077	0.142	0.084	9
	sim mcp	0.078	0.090	0.044	0.086	0.069	0.073	9

with SCAD penalty (linear-scad). We fit each type of model on the gene expression data at three quantile levels, i.e. $\tau = 0.25$, $\tau = 0.5$, $\tau = 0.75$, respectively. To compare the performance of different methods, we randomly split the original dataset into 5 testing data sets without replacement. Each testing data set contains 12 observations. The prediction errors (PE) based on the check loss function are computed for the 5 testing data sets for each model. Table 5 reports the prediction errors from different models for all testing data sets across various quantile levels. From Table 5, prediction errors from single-index quantile models are smaller than those of linear quantile models while the single-index quantile regression with either the SCAD penalty or MCP penalty produces very similar results as expected. The model size in Table 5 shows the number of selected genes by each model using the full data set. Overall, the penalized single-index quantile regression tends to produce sparser models than the penalized linear quantile models on this gene data set.

5. Conclusion

We investigate semiparametric partially linear single-index quantile regression models with ultra-high dimensional covariates both in the single-index part and partially linear part. Single-index models possess appealing flexibility and interpretability and can be viewed as

a close relative to deep neural networks. We tackle the challenges of high dimensionality of the single-index covariates in the nonparametric part combined with the non-smooth loss function and nonconvex penalty. We develop a novel proof using empirical process techniques when approaches in the existing literature are not applicable and establish the oracle theory. We propose an efficient yet very simple iterative algorithm, and show the success through numerical studies and an application to a gene expression data set.

We originally focus this work on ultra-high-dimensional single-index quantile regression and then find further success by extending to partially linear single-index quantile regression. Hence, we include both models in this paper. Naturally one may question in practice which variables should be included in the single-index terms and which variables should enter partially linearly. There are some recent developments in the literature such as Lian et al. (2015). However, it is still a very challenging question in the literature especially for conditional quantiles in the high-dimensional setting (see a discussion in Sherwood and Wang (2016)). We hope to address this question in future research.

Supplementary Materials The online supplementary materials contain the proofs of all lemmas and theorems and display the estimation results for Example 3.

References

A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Annals of Statistics*, 40:2452–2482, 2012.

Pierre Alquier and Gerard Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14(1): 243–280, 2013.

Alexandre Belloni and Victor Chernozhukov. ℓ1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 02 2011.

K. Bhatia, P. Jain, and P. Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, 2015.

R. J. Carroll, J.Q. Fan, I. Gijbels, and M. P. Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92:477–489, 1997.

H. Chen, G. Raskutti, and M. Yuan. Non-convex projected gradient descent for generalized low-rank tensor regression. *Journal of Machine Learning Research*, 20:1–37, 2019.

Xia Cui, Wolfgang Karl Haerdle, and Lixing Zhu. The EFM approach for single-index models. *The Annals of Statistics*, 39(3):1658–1688, 2011.

N. Duan and K.C. Li. Slicing regression: a link-free regression method. *The Annals of Statistics*, 19(2): 505–530, 1991.

J. Q. Fan, L. Xue, and H. Zou. Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics*, 42:819–849, 2014a.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

Jianqing Fan, Richard Samworth, and Yichao Wu. Ultrahigh dimensional feature selection: Beyond the linear model. *Journal of Machine Learning Research*, 10:2013–2038, December 2009. ISSN 1532-4435.

Jianqing Fan, Yang Feng, and Rui Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494):544–557, 2011a.

Jianqing Fan, Jinchi Lv, and Lei Qi. Sparse high-dimensional models in economics. *Annual Review of Economics*, 3(1):291–317, 2011b.

J.Q. Fan, Y. Fan, and E. Barut. Adaptive robust variable selection. *Annals of Statistics*, 42(1):324–351, 2014b.

Q. Gu, Z. Wang, and H. Liu. Low-rank and sparse structure pursuit via alternating minimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, pages 600–609, 2016.

H. Guo, C. Wu, and Yan Yu. Time-varying beta and the value premium. *Journal of Financial and Quantitative Analysis*, 52(4):1551–1576, 2017.

Wolfgang Härdle and Thomas M. Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408):986–995, 1989.

X. He and P. Shi. Convergence rate of B-spline estimators of nonparametric conditional quantile functions. *Journal of Nonparametric Statistics*, 3:299–308, 1994.

X. He, Z-Y. Zhu, and W-K Fung. Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, 89(3):579–590, 2002.

H. Ichimura. Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics*, 58(1):71–120, 1993.

P. Jain, A. Tewari, and P. Kar. On iterative hard thresholding methods for high-dimensional M-estimation. In *Advances in Neural Information Processing Systems*, 2014.

R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica*, 1(46):33–50, 1978.

Hong Lan, M. Chen, J. Flowers, B. Yandell, D. Stapleton, C. Mata, E. Mui, M. Flowers, K. Schueler, K. Manly, R. Williams, K. Kendziorski, and A. D. Attie. Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics*, 2:1–11, 2006.

Eun Ryung Lee, Hohsuk Noh, and Byeong U. Park. Model selection via bayesian information criterion for quantile regression models. *Journal of the American Statistical Association*, 109(505):216–229, 2014.

K.C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991.

Lexin Li and Christopher J Nachtsheim. Sparse sliced inverse regression. *Technometrics*, 48(4):503–510, 2006.

Q. Li. Efficient estimation of additive partially linear models. *International Economic Review*, 41(4):1073–1092, 2000.

H. Lian, Hua Liang, and D. Ruppert. Separation of covariates into nonparametric and parametric parts in high-dimensional partially linear additive models. *Statistica Sinica*, 25:591–607, 2015.

H. Lian, W. Zhao, and Y. Ma. Multiple quantile modeling via reduced-rank regression. *Statistica Sinica*, 29:1439–1464, 2019.

H. Liang, X. Liu, R. Li, and C. L. Tsai. Estimation and testing for partially linear single-index models. *The Annals of Statistics*, 38(6):3811–3836, 2010.

W. Lin and K. B. Kulasekera. Identifiability of single-index models and additive-index models. *Biometrika*, 94(2):496–501, 2007.

S. Ma and X. He. Inference for single-index quantile regression models with profile optimization. *The Annals of Statistics*, 44(3):1234–1268, 2016.

Shujie Ma, Yanyuan Ma, Yanqing Wang, Eli S. Kravitz, and Raymond J. Carroll. A semiparametric single-index risk score across populations. *Journal of the American Statistical Association*, 112(520):1648–1662, 2017.

Matey Neykov, Jun S. Liu, and Tianxi Cai. L1-regularized least squares for support recovery of high dimensional single index models with gaussian designs. *Journal of Machine Learning Research*, 17(1): 2976–3012, 2016.

S. Portnoy. Asymptotic behavior of m-estimators of p regression parameters when p^2/n is large. ii. normal approximation. *The Annals of Statistics*, 13(4):1403–1417, 1985.

J.L. Powell, J.H. Stock, and T.M. Stoker. Semiparametric estimation of weighted average derivatives. *Econometrica*, 57(6):1403–1430, 1989.

Peter Radchenko. High dimensional single index models. *Journal of Multivariate Analysis*, 139:266–282, 2015.

David Ruppert and Raymond J Carroll. Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics*, 42(2):205–223, 2000.

L. Schumaker. *Spline Functions: Basic Theory*. Wiley, New York, 1981.

Ben Sherwood and Lan Wang. Partially linear additive quantile regression in ultra-high dimension. *The Annals of Statistics*, 44(1):288–317, 2016.

Qifan Song and Faming Liang. High-dimensional variable selection with reciprocal l1-regularization. *Journal of the American Statistical Association*, 110(512):1607–1620, 2015.

Charles J Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13:689–705, 1985.

Pham Dinh Tao and Le Thi Hoai An. Convex analysis approach to dc programming: Theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1):289–355, 1997.

Guannan Wang and Li Wang. Spline estimation and variable selection for single-index prediction models with diverging number of index parameters. *Journal of Statistical Planning and Inference*, 162:1–19, 2015.

H. J. Wang, Z. Zhu, and J. Zhou. Quantile regression in partially linear varying coefficient models. *The Annals of Statistics*, 37(6B):3841–3866, 2009a.

H. S. Wang, B. Li, and C. L. Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:671–683, 2009b.

L. Wang, X. Liu, H. Liang, and R.J. Carroll. Estimation and variable selection for generalized additive partial linear models. *The Annals of Statistics*, 39(4):1827–1851, 2011.

L. Wang, Y. Wu, and R. Li. Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107(497):214–222, 2012.

Y. Wei and X. He. Conditional growth charts. *The Annals of Statistics*, 34(5):2069–2097, 2006.

C. Wu and Y. Yu. Partially linear modeling of conditional quantiles using penalized splines. *Computational Statistics & Data Analysis*, 77(C):170–187, 2014.

Yichao Wu and Yufeng Liu. Variable selection in quantile regression. *Statistica Sinica*, 19(2):801–817, 2009.

Z. Wu, K. Yu, and Y. Yu. Single-index quantile regression. *J. Multivariate Anal.*, 101(7):1607–1621, 2010.

Yingcun Xia, Howell Tong, W. K. Li, and Li-Xing Zhu. An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:363–410, 2002.

Y. Xu and W. Yin. A Globally convergent algorithm for nonconvex optimization based on block coordinate update. *Journal of Scientific Computing*, 72:pages700–734, 2017.

Zhuoran Yang, Krishnakumar Balasubramanian, and Han Liu. High-dimensional non-gaussian single index models via thresholded score function estimation. *Proceedings of the 34th International Conference on Machine Learning*, 70:3851–3860, 2017.

Yan Yu and David Ruppert. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460):1042–1054, 2002.

C.H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

Jun Zhang, Tao Wang, Lixing Zhu, and Hua Liang. A dimension reduction based approach for estimation and variable selection in partially linear single-index models with high-dimensional covariates. *Electron. J. Statist.*, 6:2235–2273, 2012.

Yuankun Zhang, Heng Lian, and Yan Yu. Estimation and variable selection for quantile partially linear single-index models. *Journal of Multivariate Analysis*, 162:215–234, 2017.

Wei Zhong, Liping Zhu, Runze Li, and Hengjian Cui. Regularized quantile regression and robust feature screening for single index models. *Statistica Sinica*, 26(1):69–95, 2016.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101 (476):1418–1429, 2006.

H. Zou and R. Z. Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533, 2008.