Appl. Math. J. Chinese Univ.
2025, 40(1): 53-77

Semiparametric expectile regression for high-dimensional heavy-tailed and heterogeneous data

ZHAO Jun[1,2] YAN Guan-ao[3] ZHANG Yi[4,*]

Abstract. High-dimensional heterogeneous data have acquired increasing attention and discussion in the past decade. In the context of heterogeneity, semiparametric regression emerges as a popular method to model this type of data in statistics. In this paper, we leverage the benefits of expectile regression for computational efficiency and analytical robustness in heterogeneity, and propose a regularized partially linear additive expectile regression model with a nonconvex penalty, such as SCAD or MCP, for high-dimensional heterogeneous data. We focus on a more realistic scenario where the regression error exhibits a heavy-tailed distribution with only finite moments. This scenario challenges the classical sub-gaussian distribution assumption and is more prevalent in practical applications. Under certain regular conditions, we demonstrate that with probability tending to one, the oracle estimator is one of the local minima of the induced optimization problem. Our theoretical analysis suggests that the dimensionality of linear covariates that our estimation procedure can handle is fundamentally limited by the moment condition of the regression error. Computationally, given the nonconvex and nonsmooth nature of the induced optimization problem, we have developed a two-step algorithm. Finally, our method's effectiveness is demonstrated through its high estimation accuracy and effective model selection, as evidenced by Monte Carlo simulation studies and a real-data application. Furthermore, by taking various expectile weights, our method effectively detects heterogeneity and explores the complete conditional distribution of the response variable, underscoring its utility in analyzing high-dimensional heterogeneous data.

Received: 2020-07-12. Revised: 2023-12-01.
MR Subject Classification: 62H12, 62G08.
Keywords: expectile regression, heterogeneity, heavy tail, partially linear additive model.
Digital Object Identifier(DOI): https://doi.org/10.1007/s11766-025-4215-z.
Supported by the Hangzhou Joint Fund of the Zhejiang Provincial Natural Science Foundation of China(LHZY24A010002), the MOE Project of Humanities and Social Sciences(21YJCZH235).
*Corresponding author.
†Zhao and Yan contributed equally to this study.

§1 Introduction

The past two decades have witnessed the rapid development of high-dimensional statistical analysis, most of which usually assume homogeneity in the data structure. However, the National Research Council (2013) highlights that multi-source data collection technologies and error accumulation in data preprocessing contribute to an opposing characteristic in high-dimensional data: heterogeneity. Evidence for heterogeneity in high-dimensional data is substantial. For instance, Daya, Chen and Li (2012) identified heteroscedasticity in eQTLs data, commonly linked to gene expression variations, underscoring the importance of incorporating heteroscedasticity in modeling process. Wang, Wu and Li (2012) applied regularized quantile regression to investigate genetic variations related to human eye disease, also observing heterogeneity in this genetic data.

Buja et al. (2014) highlighted that the nonlinear effects of certain covariates on the response can result in misconceptions in data mining when linear approximations are used indiscriminately. From this perspective, incorporating nonparametric effects into the modeling process becomes necessary under certain conditions. Semiparametric regression, combining the simplicity of linear models with the flexibility of nonparametric approaches, is widely adopted for modeling heterogeneous data in statistics and econometrics. Meanwhile, additivity is frequently assumed in the nonparametric component to avoid the curse of dimensionality, as discussed by Hastie and Tibshirani (1990). Consequently, a partially linear additive model is commonly employed as an intermediate strategy, enhancing both the reliability and flexibility of the analysis. Within this framework, Sherwood and Wang (2016) introduced the regularized partially linear additive quantile regression for analyzing high-dimensional heterogeneous data.

Expectile regression proposed by Newey and Powell (1987) assigns distinct weights to squared error loss based on positive and negative errors, and is an alternative approach to address heterogeneity. Expectile regression uses the asymmetric squares loss function $\phi_\alpha(\cdot)$,

$$\phi_\alpha(r) = |\alpha - \mathbb{I}(r < 0)|r^2 = \begin{cases} \alpha r^2, & r \geq 0, \\ (1-\alpha)r^2, & r < 0. \end{cases} \tag{1.1}$$

And the α-th expectile of random variable y is denoted by $m_\alpha(y) = \arg\min_{m\in\mathbb{R}} \mathbb{E}\phi_\alpha(y-m)$. Note that the 1/2-th expectile corresponds precisely to the mean. Expectile regression exhibits several advantageous properties over quantile regression. First, its differentiable loss function reduces the computational burden and facilitates a more manageable theoretical process in high-dimensional settings. Second, Waltrup et al.(2015) concluded from their simulations that expectile regression is less susceptible to crossing problems compared to quantile regression, offering robustness in nonparametric approximations. Owing to these favorable properties, expectile regression holds significant potential for analyzing heterogeneity within a semiparametric framework. Sobotka et al. (2013) introduced geoadditive expectile regression with P-spline approximation and established an asymptotic distribution for constructing confidence intervals in classical dimension settings. In the literature related to expectile, most studies typically assume that regression errors conform to Gaussian or sub-Gaussian distributions. For example,

Gu and Zou (2016), under this assumption, developed a linear expectile regression model to analyze heteroscedasticity in high-dimensional data. However, this assumption faces growing skepticism, especially in fields like genetics and finance. Specifically, regression errors in such analyses usually do not exhibit exponentially decreasing tail rates (Fan, Li, and Wang, 2017), and may even possess heavy tails with only finite moments (Zhao, Chen, and Zhang, 2018).

This article presents the methodology and theory for partially linear additive expectile regression with a general nonconvex penalty, tailored for high-dimensional heterogeneous data. To address heterogeneity, we embrace variance heterogeneity from Rigby and Stasinopoulos (1996), accommodating regression errors with either nonconstant or varying covariate effects. Crucially, our framework diverges from the classical Gaussian or sub-Gaussian error distributions by adopting a more realistic assumption: regression errors have only finite moments. Theoretically, we derive the asymptotic oracle properties of the estimator in our proposed framework and determine how heavy-tailed moment conditions influence the dimensionality of covariates our model can manage. Unlike Spiegel et al. (2017), our approach involves model selection and coefficient estimation within a nonconvex regularized framework, as opposed to relying on selection criteria. Computationally, given the nonconvex and nonsmooth nature of the optimization problem, we exploit the structure of our formulated problem, divide it into penalized linear and nonparametric segments, and introduce a two-step algorithm.

This article is structured as follows. Section 2 introduces our penalized partially linear additive expectile regression model with nonconvex penalty functions such as SCAD or MCP, along with an efficient algorithm for addressing the optimization problem. Section 3 presents the oracle estimator as a benchmark and establishes its relationship with the optimization problem we formulated, referred to as the oracle property. Section 4 involves conducting Monte Carlo simulations to evaluate the performance of our proposed method in a heteroscedasticity context. Section 5 applies our model to a genetic microarrays dataset, analyzing potential factors contributing to low infant birth weights. All proofs of theoretical results and necessary lemmas are provided in the Appendix.

§2 Methodology

2.1 Partially linear additive expectile regression

Consider a high-dimensional data sample $\{Y_i, \mathbf{x}_i, \mathbf{z}_i\}_{i=1}^n$ with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ representing independent and identically distributed p-dimensional covariates with a common mean 0 and $\mathbf{z}_i = (z_{i1}, \ldots, z_{id})$ as d-dimensional covariates. The data conform to the following high-dimensional partially linear model,

$$Y_i = \mu_0 + \sum_{k=1}^{p} \beta_k^* x_{ik} + \sum_{j=1}^{d} g_{0j}(z_{ij}) + \epsilon_i = \mathbf{x}_i^T \boldsymbol{\beta}^* + g_0(\mathbf{z}_i) + \epsilon_i, \tag{2.1}$$

with $g_0(\mathbf{z}_i) = \mu_0 + \sum_{j=1}^{d} g_{0j}(z_{ij})$ and $\{\epsilon_i\}_{i=1}^n$ mutually independent. Each g_{0j} is assumed zero mean, i.e., $g_0(\mathbf{z})$ belongs to $\mathcal{G} = \{g(\mathbf{z}) : g(z) = \mu + \sum_{j=1}^{d} g_j(z_j), \mathbb{E}[g_j(z_j)] = 0,\ j = 1, \ldots, d\}$.

For a specific α where $m_\alpha(\epsilon_i|\mathbf{x}_i,\mathbf{z}_i)=0$, the α-th conditional expectile of Y_i as per the model (2.1) is $m_\alpha(Y_i|\mathbf{x}_i,\mathbf{z}_i)=\mathbf{x}_i^T\boldsymbol{\beta}^*+g_0(\mathbf{z}_i)$. Thus $\boldsymbol{\beta}^*$ and $g_0(\mathbf{z})$ minimize the population risk

$$(\boldsymbol{\beta}^*,g_0(\mathbf{z}))=\underset{\boldsymbol{\beta}\in\mathbb{R}^p,g\in\mathcal{G}}{\arg\min}\,\mathbb{E}[\phi_\alpha(Y_i-\mathbf{x}^T\boldsymbol{\beta}-g(\mathbf{z}))]. \tag{2.2}$$

Considering data heterogeneity, ϵ_i can be represented as $\epsilon_i=\sigma(\mathbf{x}_i,\mathbf{z}_i)\eta_i$, where $\sigma(\mathbf{x}_i,\mathbf{z}_i)$ may vary as nonconstant, linear (Gu and Zou, 2016), or nonparametric (Rigby and Stasinopoulos, 1996). Generally, given data heterogeneity, $\boldsymbol{\beta}^*$ and $g_0(\mathbf{z})$may vary across different expectile levels. For example, analogous to Gu and Zou's model, $\sigma(\mathbf{x}_i,\mathbf{z}_i)$ can be specified as $\sigma(\mathbf{x}_i,\mathbf{z}_i)=\mathbf{x}_i^T\boldsymbol{\gamma}_1+\mathbf{z}_i^T\boldsymbol{\gamma}_2$. This special model is intuitive and for a different level $\tau\neq\alpha$, i.e.,$m_\tau(\epsilon_i|\mathbf{x}_i,\mathbf{z}_i)\neq 0$, the τ-th expectile $m_\tau(Y_i|\mathbf{x}_i,\mathbf{z}_i)=\mathbf{x}_i^T\left(\boldsymbol{\beta}+m_\tau(\epsilon_i|\mathbf{x}_i,\mathbf{z}_i)\cdot\boldsymbol{\gamma}_1\right)+g_0(\mathbf{z}_i)+\mathbf{z}_i^T\boldsymbol{\gamma}_2\cdot m_\tau(\epsilon_i|\mathbf{x}_i,\mathbf{z}_i)=\mathbf{x}_i^T\boldsymbol{\beta}^\tau+g_{0,\tau}(\mathbf{z}_i)$, i.e., $\boldsymbol{\beta}^\tau$ is different from $\boldsymbol{\beta}^\alpha$, and so is $g_{0,\tau}(\mathbf{z}_i)$.

2.2 The nonconvex regularized framework

The dimensionality of the nonparametric covariates $\mathbf{z}$ is held constant in the model (2.1). Define $\boldsymbol{\pi}(t)=(b_1(t),\ldots,b_{k_n+l+1}(t))^T$ as a vector of normalized B-spline basis functions of order $l+1$ with k_n quasi-uniform internal knots on $[0,1]$. Subsequently, $g_{0j}(\cdot),j=1,\ldots,d$ are approximated by a linear combination of these B-spline basis functions, represented as $\boldsymbol{\Pi}(\mathbf{z}_i)=(1,\boldsymbol{\pi}(z_{i1})^T,\ldots,\boldsymbol{\pi}(z_{id})^T)^T\in\mathbb{R}^{D_n}$, where $D_n=d(k_n+l+1)+1$. For notational simplicity, the same number of basis functions is used for all nonlinear components in the model (2.1). However, this is not a necessary restriction in practical applications.

The dimensionality of $\mathbf{x}$, p, is much larger than n and the true parameter $\boldsymbol{\beta}^*=(\beta_1^*,\ldots,\beta_p^*)$ is assumed sparse. Let $A=\{j:\beta_j^*\neq 0,1\leq j\leq p\}$ be the active index set and $q=q(n)=|A|$. Without loss of generality, we rewrite $\boldsymbol{\beta}^*=((\boldsymbol{\beta}_A^*)^T,\mathbf{0}^T)^T$ where $\boldsymbol{\beta}_A^*\in\mathbb{R}^q$ and $\mathbf{0}$ denotes a $(p-q)$ dimensional vector of zero. Let $\mathbf{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^T$ be the $n\times p$ matrix of covariates. Denote $\mathbf{X}_j$ the jth column of $\mathbf{X}$ and define $\mathbf{X}_A$ the submatrix of $\mathbf{X}$ that consists of its first q columns and denote by $\mathbf{X}_{A_i}$ the i th row of $\mathbf{X}_A$. With sparsity, the regularized framework has become pivotal in analyzing high-dimensional data over the past two decades. The L_1 penalty or Lasso (Tibshirani, 1996) is favored in penalized estimation as it leads to a convex optimization problem. However, the L_1 penalty has drawbacks, such as over-penalizing large coefficients, introducing bias, and necessitating strong irrepresentable conditions on the design matrix for selection consistency. In contrast, an appropriate nonconvex penalty function, as discussed by Fan and Li (2001), can effectively address these issues. Therefore, in this paper, we assume the regularizer $P_\lambda(t)$ to be a general folded concave penalty function, such as the well-known SCAD (Fan and Li, 2001) or MCP (Zhang, 2010).

To the end, the proposed estimators are obtained by

$$(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}})=\underset{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\xi}\in\mathbb{R}^{D_n}}{\arg\min}\;L(\boldsymbol{\beta},\boldsymbol{\xi}), \tag{2.3}$$

where the penalized expectile loss function $L(\boldsymbol{\beta},\boldsymbol{\xi})$ for our model is

$$L(\boldsymbol{\beta},\boldsymbol{\xi})=\frac{1}{n}\sum_{i=1}^n\phi_\alpha(y_i-\mathbf{x}_i^T\boldsymbol{\beta}-\boldsymbol{\Pi}(\mathbf{z}_i)^T\boldsymbol{\xi})+\sum_{j=1}^p P_\lambda(|\beta_j|). \tag{2.4}$$

Denote by $\hat{\boldsymbol{\xi}} = (\hat{\xi}_0, \hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_d)$, then the estimator of $g_0(\mathbf{z}_i)$ is $\hat{g}(\mathbf{z}_i) = \hat{\mu} + \sum_{j=1}^{d} \hat{g}_j(z_{ij})$, where

$$\hat{\mu} = \hat{\xi}_0 + n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{d} \boldsymbol{\pi}(z_{ij})^T \hat{\boldsymbol{\xi}}_j, \quad \hat{g}_j(z_{ij}) = \boldsymbol{\pi}(z_{ij})^T \hat{\boldsymbol{\xi}}_j - n^{-1} \sum_{i=1}^{n} \boldsymbol{\pi}(z_{ij})^T \hat{\boldsymbol{\xi}}_j.$$

The centering above is just the sample analog of the identifiability assumption $\mathbb{E}[g_{0j}(\mathbf{z}_j)] = 0$.

2.3 Algorithm

For the optimization problem (2.3), we derive a two-step algorithm as follows,

Algorithm 1 The iterative two-step algorithm for the nonconvex optimization problem (2.3)

1: Initialize $\boldsymbol{\beta}^{(0)} = \boldsymbol{\beta}^{\text{initial}}$.

2: For $k = 1, 2, \ldots$, repeat the following two steps (a) and (b) until convergence

(a) **The nonparametric part**: At k-th iteration, based on the previous solution $\boldsymbol{\beta}^{(k-1)}$,

$$\boldsymbol{\xi}^{(k)} = \arg\min_{\boldsymbol{\xi} \in \mathbb{R}^{D_n}} \frac{1}{n} \sum_{i=1}^{n} \phi_\alpha(y_i - \mathbf{x}_i^T \boldsymbol{\beta}^{(k-1)} - \boldsymbol{\Pi}(\mathbf{z}_i)^T \boldsymbol{\xi}).$$

(b) **The linear part**: Then at k-th iteration, $\boldsymbol{\beta}^{(k)}$ is obtained by the following procedure,

(b.1) Calculate the corresponding weights based on $\boldsymbol{\beta}^{(k-1)} = (\beta_1^{(k-1)}, \ldots, \beta_p^{(k-1)})^T$

$$\boldsymbol{\omega}^{(k)} = (\omega_1^{(k)}, \ldots, \omega_p^{(k)})^T = (P'_\lambda(|\beta_1^{(k-1)}|), \ldots, P'_\lambda(|\beta_p^{(k-1)}|))^T.$$

(b.2) The local linear approximation of $L(\boldsymbol{\beta}, \boldsymbol{\xi}^{(k)})$, denoted by $L(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k-1)}, \boldsymbol{\xi}^{(k)})$, is

$$L(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k-1)}, \boldsymbol{\xi}^{(k)}) = \frac{1}{n} \sum_{i=1}^{n} \phi_\alpha(y_i - \mathbf{x}_i^T \boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)^T \boldsymbol{\xi}^{(k)}) + \sum_{j=1}^{p} \omega_j^{(k)} |\beta_j|.$$

(b.3) $\boldsymbol{\beta}^{(k)} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} L(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k-1)}, \boldsymbol{\xi}^{(k)})$.

In Algorithm 1, instead of taking $(\boldsymbol{\beta}, \boldsymbol{\xi})$ as the whole optimization parameters, we split the optimization problem into two parts: the fixed-dimensional unpenalized nonparametric part and the high-dimensional penalized linear part. Specifically, in the first step, we determine the nonlinear part's parameters by minimizing an unpenalized objective function in D_n dimension, using the linear part's parameters set to their values from the previous iteration. Note by Lemma 6.1, the expectile loss function $\phi_\alpha(\cdot)$ is differentiable and strongly convex, thus facilitating the optimization process via convex analysis. Following the resolution of the nonparametric part, the second step involves determining the linear part's parameters by minimizing a penalized expectile loss function. Given the nonconvex nature of the penalty, this results in a nonconvex optimization problem in a high-dimensional space. We employ the Local Linear Approximation (LLA) strategy (Zou and Li, 2008) to transform the penalized optimization problem into a convex one, leveraging its computational efficiency and favorable statistical

properties, as discussed in Fan, Xue, and Zou (2014). The initial value $\boldsymbol{\beta}^{(0)}$ can be chosen as the estimator from the following pseudo-linear penalized expectile regression,

$$\boldsymbol{\beta}^{(0)} = \arg\min_{\mu,\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n} \phi_\alpha(y_i - \mu - \mathbf{x}_i'\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1.$$

In each step, the involved optimization subproblems in the algorithm above are convex after modification, and taking full advantage of expectile regression with its differentiability, there are many powerful programs to solve them. For example, to solve the problem (b.3), we can apply the proximal gradient method, and use CVX, a Matlab package for specifying and solving convex programs; see Michael and Stephen (2013).

§3 Asymptotic theory

3.1 Oracle study

Following Fan and Li (2001), we introduce the oracle estimator, denoted by $(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\xi}}^*)$ with $\hat{\boldsymbol{\beta}}^* = (\hat{\boldsymbol{\beta}}_A^{*T}, \mathbf{0}_{p-q}^T)^T$, as a performance benchmark for the partially linear additive model,

$$(\hat{\boldsymbol{\beta}}_A^*, \hat{\boldsymbol{\xi}}^*) = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^q, \boldsymbol{\xi}\in\mathbb{R}^{D_n}} \frac{1}{n}\sum_{i=1}^{n} \phi_\alpha(y_i - \mathbf{x}_{A_i}^T\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)^T\boldsymbol{\xi}). \tag{3.5}$$

The cardinality q_n of the index set A is allowed to change with n, which violates the classical scenario where the cardinality is fixed, for example, Hastie and Tibshirani (1990).

The derivative of $\phi_\alpha(r)$ is $\psi_\alpha(r) = 2|\alpha - \mathbb{I}(r<0)|r$. An analog second-order derivative is defined as $\varphi_\alpha(r)$ for $c_1 \triangleq \min\{\alpha, 1-\alpha\}$ and $c_2 \triangleq \max\{\alpha, 1-\alpha\}$,

$$\varphi_\alpha(r) = \begin{cases} 2|\alpha - \mathbb{I}(r<0)|, & r \neq 0, \\ \in 2[c_1, c_2], & r = 0. \end{cases}$$

Let $w_i = \mathbb{E}[\varphi_\alpha(\epsilon_i)|\mathbf{x}_i, \mathbf{z}_i]$, then w_i is uniformly bounded away from zero. Denote $\mathcal{H}_r$ the collection of functions $h(\cdot)$ on $[0,1]$ whose r_0-th derivative $h^{(r_0)}(\cdot)$ satisfies the Hölder condition of order ν, i.e. $|h^{(r_0)}(z_1) - h^{(r_0)}(z)| \le C|z_1 - z|^\nu,\ \forall\ 0 \le z_1, z \le 1$. Consider the weighted projection from $\mathbf{x}$ onto $\mathbf{z}$, $h_j^*(\cdot) = \arg\inf_{h_j(\cdot)\in\mathcal{G}\cap\mathcal{H}_r} \sum_{i=1}^n \mathbb{E}[w_i \cdot (x_{ij} - h_j(\mathbf{z}_i))^2]$. This projection strategy is commonly used in the semiparametric analysis, see Robinson (1988). Define $m_j(\mathbf{z}) = \mathbb{E}[x_{ij}|\mathbf{z}_i = \mathbf{z}]$. Then, $h_j^*(\mathbf{z})$ is a weighted projection from $m_j(\mathbf{z})$ to $\mathcal{G}\cap\mathcal{H}_r$ under L_2 norm. Then we define $H = (h_j^*(\mathbf{z}_i))_{n\times q_n}$, $\delta_{ij} = x_{A_{ij}} - h_j^*(\mathbf{z}_i)$, the vector $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{iq_n})^T \in \mathbb{R}^{q_n}$ and the matrix $\Delta_n = (\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_n)^T \in \mathbb{R}^{n\times q_n}$. Thus, $X_A = H + \Delta_n$.

Condition 3.1. $\mathbb{E}\left(\epsilon_i^{2k}|x_i, \mathbf{z}_i\right) < C < \infty$ *for all* i *and some* $k \ge 1$, *constant* $C > 0$.

Condition 3.2. *There exist constants* M_1 *and* M_2 *such that* $|x_{ij}| \le M_1, \forall 1 \le i \le n, 1 \le j \le p_n$ *and* $\mathbb{E}\left(\delta_{ij}^4\right) \le M_2, \forall 1 \le i \le n, 1 \le j \le q_n$. *There exist finite positive constants* C_1 *and* C_2 *such that with probability one,* $C_1 \le \lambda_{max}\left(n^{-1}X_A X_A^T\right) \le C_2$, $C_1 \le \lambda_{max}\left(n^{-1}\Delta_n\Delta_n^T\right) \le C_2$.

Condition 3.3. $g_0(\cdot) \in \mathcal{G}\cap\mathcal{H}_r$ *with* $r = r_0 + v > 1.5$, *and* k_n *satisfies* $k_n \approx n^{1/(2r+1)}$.

Condition 3.4. $q_n = O(n^{C_3})$ *for some* $C_3 < \frac{1}{2}$.

In Condition 3.1, the imposed moment condition on the error sequences is more relaxed than the classical Gaussian or sub-Gaussian tail condition. This condition is also used in Kim, Choi and Oh (2008). Condition 3.3 is required for the B-splines approximation accuracy and convergence rate of $\hat{g}(\cdot)$. As Stone (1985) and Schumaker (2007) pointed out, if $g_{0j}(\cdot) \in \mathcal{H}_r$, and $r \geq 1.5$, there exists $\boldsymbol{\xi}_0 = (\xi_{00}, \boldsymbol{\xi}_{01}, \ldots, \boldsymbol{\xi}_{0d}) \in \mathbb{R}^{D_n}$ such that $\sup_{\mathbf{z}_i} |\boldsymbol{\Pi}(\mathbf{z}_i)^T \boldsymbol{\xi}_0 - g_0(\mathbf{z}_i)| = O(k_n^{-r})$.

Theorem 3.1. *Suppose conditions 3.1-3.4 hold. Then the oracle estimator satisfies*

$$\| \hat{\boldsymbol{\beta}}_A^* - \boldsymbol{\beta}_A^* \| = O_p(\sqrt{n^{-1}q_n}), \quad n^{-1}\sum_{i=1}^{n}(\hat{g}(\mathbf{z}_i) - g_0(\mathbf{z}_i))^2 = O_p(n^{-1}(q_n + k_n)). \tag{3.6}$$

3.2 Differencing convex procedure

Note that $L(\boldsymbol{\beta}, \boldsymbol{\xi})$ is nonconvex, and $\phi_\alpha(r)$ is not smooth due to the non-existence of its second order derivative at $r = 0$. So the KKT condition appears not applicable to the optimization problem (2.3). Suppose SCAD penalty is used, then $L(\boldsymbol{\beta}, \boldsymbol{\xi})$ can be decomposed into the difference of two convex functions, $L(\boldsymbol{\beta}, \boldsymbol{\xi}) = k(\boldsymbol{\beta}, \boldsymbol{\xi}) - l(\boldsymbol{\beta}, \boldsymbol{\xi})$, with

$$k(\boldsymbol{\beta}, \boldsymbol{\xi}) = \frac{1}{n}\sum_{i=1}^{n}\phi_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)^T\boldsymbol{\xi}) + \lambda\sum_{j=1}^{p}|\beta_j|, \quad l(\boldsymbol{\beta}, \boldsymbol{\xi}) = \sum_{j=1}^{p}H_\lambda(\beta_j),$$

$$H_\lambda(\theta) = [(\theta^2 - 2\lambda|\theta| + \lambda^2)/(2(a-1))]\mathbb{I}(\lambda \leq |\theta| \leq a\lambda) + [\lambda|\theta| - (a+1)^2/2]\mathbb{I}(|\theta| > a\lambda).$$

Tao and An (1997) provided sufficient conditions for such type of non-convex optimization problem, see Lemma 6.10 for detail. The unpenalized empirical loss function $L_n(\boldsymbol{\beta}, \boldsymbol{\xi})$ is differentiable with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$.

$$L_n(\boldsymbol{\beta}, \boldsymbol{\xi}) = \frac{1}{n}\sum_{i=1}^{n}\phi_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)^T\boldsymbol{\xi}). \tag{3.7}$$

Denote $\boldsymbol{\Pi}(\mathbf{z}_i) = (1, \Pi_1(z_{i1}), \ldots, \Pi_{L_n}(z_{id}))$ the basis function at $\mathbf{z}_i$, for $j = 1, \ldots, p$

$$\begin{aligned} s_j(\boldsymbol{\beta}, \boldsymbol{\xi}) &= \frac{\partial}{\partial\beta_j}\Big(\frac{1}{n}\sum_{i=1}^{n}\phi_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)^T\boldsymbol{\xi})\Big) \\ &= -\frac{2}{n}\sum_{i=1}^{n}\alpha\mathbf{x}_{ij}(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)^T\boldsymbol{\xi})\mathbb{I}(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)^T\boldsymbol{\xi} \geq 0) \\ &\quad -\frac{2}{n}\sum_{i=1}^{n}(1-\alpha)\mathbf{x}_{ij}(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)^T\boldsymbol{\xi})\mathbb{I}(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)^T\boldsymbol{\xi} < 0), \end{aligned}$$

and for $j = p + l,\ l = 1, \ldots, D_n$,

$$\begin{aligned} s_j(\boldsymbol{\beta}, \boldsymbol{\xi}) &= \frac{\partial}{\partial\xi_l}\Big(\frac{1}{n}\sum_{i=1}^{n}\phi_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)^T\boldsymbol{\xi})\Big) \\ &= -\frac{2}{n}\sum_{i=1}^{n}\alpha\boldsymbol{\Pi}_l(\mathbf{z}_i)(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)^T\boldsymbol{\xi})\mathbb{I}(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \boldsymbol{\Pi}(\mathbf{z}_i)^T\boldsymbol{\xi} \geq 0) \end{aligned}$$

$$-\frac{2}{n}\sum_{i=1}^{n}(1-\alpha)\mathbf{\Pi}_l(\mathbf{z}_i)(y_i-\mathbf{x}_i^T\boldsymbol{\beta}-\mathbf{\Pi}(\mathbf{z}_i)^T\boldsymbol{\xi})\mathbb{I}(y_i-\mathbf{x}_i^T\boldsymbol{\beta}-\mathbf{\Pi}(\mathbf{z}_i)^T\boldsymbol{\xi}<0).$$

Lemma 6.12 shows that under the Conditions 3.1-3.4 and the following so-called Beta-min condition, for the oracle estimator $(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\xi}}^*)$,

$$s_j(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\xi}}^*) = 0, \quad j=1,\ldots,q_n \text{ or } j=p+1,\ldots,p+D_n, \tag{3.8}$$

$$|s_j(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\xi}}^*)| \leq \lambda, \quad j=q_n+1,\ldots,p. \tag{3.9}$$

Condition 3.5 (Beta-min condition)**.** *There exist positive constants C_4 and C_5 such that for $C_3 < C_4 < 1$, $n^{(1-C_4)/2}\min_{1\leq j\leq q_n}|\boldsymbol{\beta}_j^*| \geq C_5$.*

Theorem 3.2. *Assume Conditions 3.1-3.5 are satisfied. Denote $\mathcal{E}(\lambda)$ be the set of local minima of $L(\boldsymbol{\beta},\boldsymbol{\xi})$ with the tuning parameter λ. If the tuning parameter $\lambda = o\left(n^{-(1-C_4)/2}\right)$, $q_n = o(n\lambda^2)$, $k_n = o(n\lambda^2)$ and $p = o\big((n\lambda^2)^k\big)$, then we have that with probability tending to one, the oracle estimator $(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\xi}}^*)$ lies in the set $\mathcal{E}(\lambda)$ consisting of local minima of $L(\boldsymbol{\beta},\boldsymbol{\xi})$, i.e.,*

$$\mathbb{P}((\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\xi}}^*) \in \mathcal{E}(\lambda)) \to 1, \text{ as } n\to\infty. \tag{3.10}$$

By the constraints on λ, we can infer that $p = o(n^{C_4 k})$. So the moment condition and the signal strength directly influence the dimensionality our proposed method can manage. Note that if the regression error has only finite moments, p can be at most a certain polynomial power of n. If ϵ_i has all the moments, this asymptotic result holds when $p = O(n^\tau)$ for any $\tau > 0$ since $\mathbb{E}(\epsilon_i^{2k}|\mathbf{x}_i) < \infty$ for all $k > 0$. What's more, if the error ϵ follows a gaussian or sub-gaussian distribution, it can be shown that our method can be applied to an ultra-high dimension.

§4 Simulation

This section evaluates the finite sample performance of the proposed regularized expectile regression. We utilize the SCAD penalty as an example for the general folded concave penalty function $P_\lambda(t)$. We refer to the penalized partially linear additive expectile regression with the SCAD penalty as E-SCAD. This approach is also applicable with the MCP penalty or other general folded nonconvex penalty functions, though further details are omitted for brevity.

We employ a high-dimensional partially linear additive model from Sherwood and Wang (2016). In the data generation process, quasi-covariates $\tilde{\mathbf{x}} = (\tilde{x}_1,\ldots,\tilde{x}_{p+2})^T$ are generated from a multivariate normal distribution $N_{p+2}(\mathbf{0},\Sigma)$ with $\Sigma = (\sigma_{ij})_{(p+2)\times(p+2)}$ and $\sigma_{ij} = 0.5^{|i-j|}$ for $i,j = 1,\ldots,p+2$. Then we set $x_1 = \sqrt{12}\Phi(\tilde{x}_1)$ where $\Phi(\cdot)$ represents the cumulative distribution function of the standard normal distribution and $\sqrt{12}$ ensures that x_1 has a standard deviation of 1. Additionally, we set $z_1 = \Phi(\tilde{x}_{25})$ and $z_2 = \Phi(\tilde{x}_{26})$, $x_i = \tilde{\mathbf{x}}_i$ for $i = 2,\ldots,24$ and $x_i = \tilde{x}_{i+2}$ for $i = 25,\ldots,p$. The response variable y is then generated from the following sparse model,

$$y = x_6\beta_6 + x_{12}\beta_{12} + x_{15}\beta_{15} + x_{20}\beta_{20} + \sin(2\pi z_1) + z_2^3 + \epsilon, \tag{4.11}$$

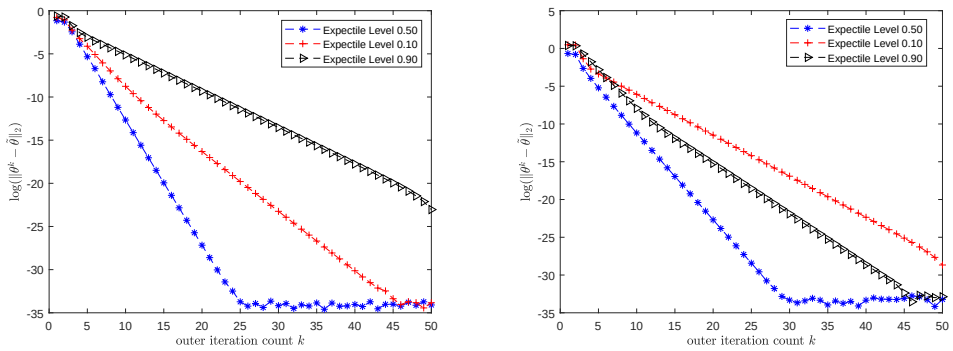
with $\beta_j = 1$ for $j = 6, 12, 15, 20$ and ϵ independent of the covariates $\mathbf{x}$. To figure out how the proposed method performs when the error ϵ shares heavy-tailed distributions, we consider two

scenarios: (1) $N(0,1)$ and (2) Standard t-distribution with degrees of freedom 5, t_5.

We focus on the heteroscedastic scenario where $\epsilon = 0.70x_1\varsigma$ with ς independent of x_1 and following the previously mentioned distributions. The data generation procedure reveals that the true coefficients $\boldsymbol{\beta}^*$ are sparse, comprising four informative variables. Additionally, x_1 should be considered significant due to its pivotal role in the conditional distribution of y and its contribution to heteroscedasticity. For comparative purposes, this simulation also examines the performance of Lasso-type regularized expectile regression, abbreviated as E-Lasso,

$$\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p,\boldsymbol{\xi}\in\mathbb{R}^{D_n}} \frac{1}{n}\sum_{i=1}^{n}\phi_\alpha(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - \mathbf{\Pi}(\mathbf{z}_i)^T\boldsymbol{\xi}) + \lambda\sum_{j=1}^{p}|\beta_j|. \tag{4.12}$$

Additionally, we introduce the oracle estimator (3.5) as the benchmark of estimation accuracy.



(a) ϵ follows $N(0,1)$ (b) ϵ follows t_5

Figure 1. Convergence Rate of Algorithm 1 with $n = 300$ and $p = 600$.

Table 1. Simulation results when $n = 300$, $p = 400$.

	Criteria	$N(0,1)$			t_5		
		E-SCAD	E-Lasso	Oracle	E-SCAD	E-Lasso	Oracle
$\alpha = 0.10$	AE	0.76 (0.27)	1.94 (0.42)	0.83 (0.17)	1.21 (0.77)	3.01 (1.22)	1.13 (0.33)
	SE	0.47 (0.20)	0.59 (0.11)	0.56 (0.11)	0.62 (0.32)	0.83 (0.21)	0.72 (0.18)
	ADE	0.54 (0.10)	0.67 (0.10)	0.26 (0.08)	0.62 (0.10)	0.73 (0.15)	0.36 (0.13)
	Size	6.79 (1.73)	24.05 (5.51)	-	8.46 (3.63)	28.25 (8.82)	-
	F,F1	100, 87	100, 97	-	100, 73	100, 94	-
$\alpha = 0.50$	AE	0.31 (0.14)	1.26 (0.30)	0.28 (0.11)	0.47 (0.16)	1.49 (0.23)	0.41 (0.13)
	SE	0.18 (0.08)	0.41 (0.09)	0.16 (0.06)	0.25 (0.10)	0.50 (0.09)	0.22 (0.07)
	ADE	0.38 (0.24)	0.37 (0.24)	0.18 (0.05)	0.44 (0.18)	0.43 (0.18)	0.32 (0.12)
	Size	4.64 (0.78)	21.78 (4.86)	-	6.49 (1.42)	20.43 (3.01)	-
	F,F1	100, 0	100, 8	-	100, 0	100, 8	-
$\alpha = 0.90$	AE	0.74 (0.27)	1.76 (0.36)	0.82 (0.16)	1.07 (0.52)	2.94 (1.23)	1.12 (0.31)
	SE	0.47 (0.20)	0.59 (0.09)	0.56 (0.11)	0.59 (0.26)	0.84 (0.19)	0.71 (0.18)
	ADE	0.49 (0.14)	0.76 (0.22)	0.25 (0.09)	0.52 (0.37)	0.68 (0.13)	0.38 (0.13)
	Size	6.21 (1.39)	19.54 (4.91)	-	7.74 (3.19)	26.06 (9.19)	-
	F,F1	100, 88	100, 96	-	100, 76	100, 87	-

The sample size is set at $n = 300$ with p being either 400 or 600. To detect heteroscedasticity more effectively, we consider three expectile weight levels: $\alpha = 0.10,\ 0.50,\ 0.90$. For the tuning parameter λ, we create an additional tuning data set of size $10n$ and select the optimal λ that minimizes the prediction expectile loss error on this data set. Regarding the nonparametric components, we utilize cubic B-splines with 3 basis functions for each nonparametric function. Figure 1 depicts $\log(\|\boldsymbol{\theta}^k - \tilde{\boldsymbol{\theta}}\|_2)$ (where $\boldsymbol{\theta}$ represents the entire parameter set and $\tilde{\boldsymbol{\theta}}$ the solution) against iteration count k in one single solution process under our simulation setting, which demonstrates that the proposed Algorithm 1 converges fast.

We evaluate the performance in terms of the following criteria based on 100 repetitions:

- AE: the average absolute estimation error defined by $\sum_i^p |\hat{\beta}_j - \beta_j^*|$.
- SE: the average square estimation error defined by $\sqrt{\sum_i^p |\hat{\beta}_j - \beta_j^*|^2}$.
- ADE: the average of the average absolute deviation defined by $\frac{1}{n}\sum_{i=1}^n |\hat{g}(\mathbf{z}_i) - g_0(\mathbf{z}_i)|$
- Size: given the role of x_1, the true size of our data generation model is supposed to be 5.
- F: the frequency that $x_6, x_{12}, x_{15}, x_{20}$ are selected during the 100 repetitions.
- F1: the frequency that x_1 is selected during the 100 repetitions.

Table 2. Simulation results when $n = 300$, $p = 600$.

	Criteria	$N(0,1)$			t_5		
		E-SCAD	E-Lasso	Oracle	E-SCAD	E-Lasso	Oracle
$\alpha = 0.10$	AE	0.97 (0.27)	2.11 (0.44)	0.86 (0.17)	1.36 (0.85)	3.74 (1.16)	1.14 (0.25)
	SE	0.54 (0.14)	0.64 (0.10)	0.57 (0.10)	0.66 (0.29)	0.86 (0.17)	0.72 (0.15)
	ADE	0.50 (0.27)	0.62 (0.07)	0.24 (0.07)	0.81 (0.51)	0.95 (0.29)	0.38 (0.12)
	Size	9.68 (2.78)	26.55 (5.60)	-	10.67 (4.71)	42.53 (9.50)	-
	F,F1	100, 96	100, 97	-	100, 77	100, 86	-
$\alpha = 0.50$	AE	0.31 (0.13)	1.50 (0.36)	0.35 (0.11)	0.63 (0.27)	1.85 (0.52)	0.43 (0.13)
	SE	0.17 (0.07)	0.43 (0.09)	0.18 (0.06)	0.32 (0.12)	0.55 (0.10)	0.23 (0.07)
	ADE	0.38 (0.23)	0.50 (0.19)	0.18 (0.04)	0.18 (0.02)	0.20 (0.06)	0.23 (0.06)
	Size	5.61 (1.55)	29.16 (6.05)	-	7.36 (3.16)	26.74 (7.11)	-
	F,F1	100, 1	100, 10	-	100, 3	100, 5	-
$\alpha = 0.90$	AE	0.82 (0.27)	1.80 (0.39)	0.83 (0.15)	1.12 (0.53)	3.18 (1.64)	1.16 (0.28)
	SE	0.49 (0.19)	0.64 (0.10)	0.56 (0.10)	0.60 (0.28)	0.88 (0.23)	0.72 (0.15)
	ADE	0.40 (0.21)	0.58 (0.10)	0.24 (0.09)	0.33 (0.04)	1.09 (0.54)	0.40 (0.22)
	Size	7.72 (2.14)	17.60 (4.67)	-	8.32 (3.36)	28.80 (10.98)	-
	F,F1	100, 88	100, 94	-	99, 79	100, 83	-

Table 1 and Table 2 present the simulation results for $p = 400$ and $p = 600$ respectively. Generally, E-SCAD demonstrates superior estimation accuracy compared to E-Lasso, tends to select smaller models, and its performance is closer to that of the oracle estimator. Regarding E-SCAD's further performances, note that in our simulation setting $m_{\alpha=0.5}(\epsilon|\mathbf{x}, \mathbf{z}) = 0$. According to Theorem 3.2, at $\alpha = 0.5$, E-SCAD outperforms other weight levels ($\alpha = 0.1$ or 0.9), in terms of estimation accuracy and model selection, justified by the AE, ADE and SE results in Table 1 and Table 2. However, variance heterogeneity is not evident in $m_{\alpha=0.50}(y|\mathbf{x}, \mathbf{z})$ so

in this situation E-SCAD can not pick x_1, the active variable resulting in heteroscedasticity. Expectile regression with different weights can address this issue effectively. It is observed that at $\alpha = 0.1$ and 0.9, x_1 is frequently identified as the active variable. Additionally, as seen from Table 1 to Table 2, an increase in p slightly worsens E-SCAD's performance, with this effect being more pronounced in the t_5 case. It must be noted that the dimensionality our proposed method can handle is influenced by the heavy-tailed characteristics of the error.

§5 Real Data Application

For public health interventions, significant research has been conducted on the determinants of low birth weight. Turan (2012) identified genes associated with low birth weight by analyzing gene promoter-specific DNA methylation levels, with cord blood and placenta samples collected from each newborn. Votavova et al. (2011) gathered peripheral blood, placenta, and cord blood samples from pregnant smokers and non-smokers , aiming to pinpoint tobacco smoke-related defects, particularly the transcriptomic alterations in genes affected by smoke exposure. We chose this dataset to investigate the potential factors for low infant birth weights since it provides a comprehensive measurements of infants, including birth weights, maternal age, gestational age, parity, maternal blood cotinine level, and BMI. This genetic dataset comprises $n = 65$ observations. Gene expression profiles for $24,526$ gene transcripts were assayed using the Illumina Expression Beadchip v3.

We include normalized genetic data and clinical variables such as parity, gestational age, maternal blood cotinine level, and BMI as linear covariates in the proposed partially linear additive model. Additionally, following the approach of Votavova et al. (2011), we incorporate maternal age as a nonparametric component to address nonlinear effects. To dissect the cause of low infant birth weight, the analysis is carried out under three different expectile levels $\alpha = 0.1, 0.3$ and 0.5. In each scenario, feature screening methods are utilized to select the top 200 relevant gene probes, as per Fan and Lv (2008). For comparative analysis, we apply the two methods, E-SCAD and E-Lasso in our data analysis. Regarding the tuning parameter λ, we use a five-fold cross-validation strategy to determine its value for both E-SCAD and E-Lasso.

Table 3. Numeric results at three expectile levels.

	Criteria	$\alpha = 0.1$		$\alpha = 0.3$		$\alpha = 0.5.$	
		E-SCAD	E-LASSO	E-SCAD	E-LASSO	E-SCAD	E-LASSO
All Data	L_1	0.66	0.67	0.60	0.53	0.38	0.34
	L_2	0.12	0.11	0.10	0.09	0.06	0.05
	$\hat{\mathcal{A}}_\alpha$	7.00	8.00	9.00	19.00	14.00	20.00
	$\hat{\mathcal{A}}_\alpha \bigcap \hat{\mathcal{A}}_{0.5}$	1	1	3	3	-	-
Random Partition	L_1	0.90 (0.21)	0.74 (0.17)	0.81 (0.17)	0.59 (0.13)	0.89 (0.20)	0.41 (0.08)
	L_2	0.30 (0.07)	0.26 (0.06)	0.27 (0.06)	0.19 (0.04)	0.30 (0.06)	0.13 (0.02)
	$\hat{\mathcal{A}}_\alpha$	5.72 (1.91)	8.19 (2.74)	9.00 (2.72)	13.94 (3.03)	4.72 (1.83)	20.25 (2.67)
	$\hat{\mathcal{A}}_\alpha \bigcap \hat{\mathcal{A}}_{0.5}$	3.86 (1.6433)	1.16 (0.39)	2.27 (1.13)	3.25 (1.18)	-	-

Initially, we apply the E-SCAD method to the entire data set at three distinct expectile lev-

Table 4. Top 6 Covariates Selected at Three Expectile Weight Levels among 100 Partitions.

E-SCAD $\alpha = 0.1$		E-SCAD $\alpha = 0.3$		E-SCAD $\alpha = 0.5$	
Variables	Frequency	Variables	Frequency	Variables	Frequency
PTPN3	34	GPR50	46	PTPN3	33
FXR1	40	FXR1	49	GPR50	40
GPR50	43	EPHA3	50	FXR1	41
LEO1	43	LEO1	59	LEO1	44
SLCO1A2	63	LOC388886	65	SLCO1A2	65
Gestational age	79	Gestational age	97	Gestational age	83

els. For each level, the set of variables selected in the linear component of our model is denoted as $\hat{\mathcal{A}}\alpha$, with its cardinality represented by $|\hat{\mathcal{A}}\alpha|$. Acknowledging potential heteroscedasticity, we also present the number of variables selected at multiple expectile levels, indicated by $|\hat{\mathcal{A}}_{0.1} \cap \hat{\mathcal{A}}_{0.5}|$ and $|\hat{\mathcal{A}}_{0.3} \cap \hat{\mathcal{A}}_{0.5}|$. The quantities of selected and overlapped variables are detailed in Table 3. Subsequently, the data set is randomly partitioned into a training set of 50 observations and a test set of 15 observations. E-SCAD is then applied to the training set to derive regression coefficients $\hat{\boldsymbol{\beta}}$, which are subsequently used to predict responses for the 15 individuals in the test set. This random splitting process is repeated 100 times. Variable selection results from the random partitioning scenario are also displayed in Table 3. Additionally, we report the mean absolute error, $L_1 = \frac{1}{24}\sum_{i\in test\ set}|y_i - x_i^T\hat{\boldsymbol{\beta}}|$, and the mean squared error, $L_2 = \frac{1}{24}\sqrt{\sum_{i\in test\ set}(y_i - x_i^T\hat{\boldsymbol{\beta}})^2}$ for predictions. Table 3 demonstrates that the models selected at expectile levels $\alpha = 0.1, 0.3, 0.5$ all result in relatively small prediction errors.

Table 3 reveals that different genes are selected at various weight levels, suggesting that diverse genetic factors influence different levels of birth weight, indicative of data heterogeneity. Table 4 provides further insights into this observation. Gestational age, consistently selected across all scenarios, corroborates the established understanding that premature birth often correlates with low birth weight. Interestingly, the scenarios $\alpha = 0.1$ and $\alpha = 0.5$ show similar performance, whereas $\alpha = 0.3$ exhibits distinct characteristics. SLCO1A2 is more frequently selected at $\alpha = 0.1$. SLCO1A2 is associated with drug resistance, and Votavova et al. (2011) link exposure to toxic compounds in tobacco smoke to low birth weight, a potential role for this gene. EPHA3 is notably more frequently selected in the $\alpha = 0.3$ scenario compared to the other two. Kudo et al. (2005) found that EPHA3 expression at the mRNA and protein levels is crucial during the development of the mammalian newborn forebrain. These findings support our analysis across different expectile levels and draw particular attention to the $\alpha = 0.3$ case, due to its unique results and potential biomedical significance.

§6 Appendix

6.1 Notation

B-spline basis functions $b_j(\cdot)$ can be centered as $B_j(z_{ik}) = b_{j+1}(z_{ik}) - \frac{\mathbb{E}[b_{j+1}(z_{ik})]}{\mathbb{E}[b_1(z_{ik})]}b_1(z_{ik})$, $j = 1, \ldots, k_n + l$, satisfying $\mathbb{E}[B_j(z_{ik})] = 0$. Denote $\mathbf{w}(z_{ik}) = (B_1(z_{ik}), \ldots, B_{k_n+l}(z_{ik}))^T$,

the J_n-dimensional vector $\mathbf{W}(\mathbf{z}_i) = (k_n^{-1/2}, \mathbf{w}(z_{i1})^T, \dots \mathbf{w}(z_{id})^T))^T$, where $J_n = d(k_n + l) + 1$ and the $n \times J_n$ matrix $W = (\mathbf{W}(\mathbf{z}_1), \dots, \mathbf{W}(\mathbf{z}_n))^T$. Define a new pair of minimizer $(\hat{\mathbf{c}}_A, \hat{\boldsymbol{\gamma}})$ as $(\hat{\mathbf{c}}_A^*, \hat{\boldsymbol{\gamma}}) = \underset{(\mathbf{c}_A, \boldsymbol{\gamma})}{\arg\min} \frac{1}{n} \sum_{i=1}^n \phi_\alpha(y_i - \mathbf{x}_{Ai}^T \mathbf{c}_A - \mathbf{W}(\mathbf{z}_i)^T \boldsymbol{\gamma})$. Denote $\boldsymbol{\gamma} = (\gamma_0, \boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_d^T)^T$ with $\gamma_0 \in \mathbb{R}$ and $\boldsymbol{\gamma}_j \in \mathbb{R}^{k_n + l}$, for $j = 1, \dots, d$. Thus, the estimator for $g_0(\mathbf{z}_i)$ is $\tilde{g}(\mathbf{z}_i) = \mathbf{W}(\mathbf{z}_i)^T \hat{\boldsymbol{\gamma}} = \tilde{\mu} + \sum_{j=1}^d \tilde{g}_j(z_{ij}) = k_n^{-1/2} \hat{\gamma}_0 + \sum_{j=1}^d \mathbf{w}(z_{ij})^T \hat{\boldsymbol{\gamma}}_j$. Note that $\hat{\mathbf{c}}_A^* = \hat{\boldsymbol{\beta}}_A^*$. Also, the original estimator of nonparametric functions can be derived from the new ones as $\hat{\mu} = \tilde{\mu} + \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^d \tilde{g}_j(z_{ij})$ and $\hat{g}_j(z_{ij}) = \tilde{g}_j(z_{ij}) - \frac{1}{n} \sum_{i=1}^n \tilde{g}_j(z_{ij})$. Thus, $\hat{g}(\mathbf{z}_i) = \tilde{g}(\mathbf{z}_i)$. Throughout the proofs, we denote C a positive constant which does not depend on n and may vary from line to line. For a vector $\boldsymbol{\theta}$, $||\boldsymbol{\theta}||$ refers to its L_2 norm. For a matrix X, $||X|| = \sqrt{\lambda_{max}(X^T X)}$ denotes its spectral norm. Furthermore, we have following notations throughout the appendix,

$$
\begin{aligned}
&B_n = \mathrm{diag}(w_1, \dots, w_n) \in \mathbb{R}^{n \times n}, \quad W_n = n^{-1} \sum_i w_i \delta_i \delta_i^T \in \mathbb{R}^{n \times n}, \\
&P = W(W^T B_n W)^{-1} W^T B_n \in \mathbb{R}^{n \times n}, \quad W_B^2 = W^T B_n W \in \mathbb{R}^{J_n \times J_n}, \\
&X^* = (\mathbf{x}_1^*, \dots, \mathbf{x}_n^*)^T = (I_n - P) X_A \in \mathbb{R}^{n \times q_n}, \quad \tilde{\mathbf{x}}_i = n^{-1/2} \mathbf{x}_i^* \in \mathbb{R}^{q_n}, \\
&\tilde{\mathbf{W}}(\mathbf{z}_i) = W_B^{-1} \mathbf{W}(\mathbf{z}_i) \in \mathbb{R}^{J_n}, \quad \tilde{s}_i = (\tilde{\mathbf{x}}_i^T, \tilde{\mathbf{W}}(\mathbf{z}_i)^T)^T \in \mathbb{R}^{J_n + q_n}, \quad u_{ni} = \mathbf{W}(\mathbf{z}_i)^T \boldsymbol{\gamma}_0 - g_0(\mathbf{z}_i) \\
&\boldsymbol{\theta}_1 = \sqrt{n}(\mathbf{c}_A - \beta_A^*) \in \mathbb{R}^{q_n}, \quad \boldsymbol{\theta}_2 = W_B(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) + W_B^{-1} W^T B_n X_A (\mathbf{c}_A - \boldsymbol{\beta}_A^*) \in \mathbb{R}^{q_n}.
\end{aligned}
$$

Under the new notation system, the objective loss function can be displayed as

$$
\frac{1}{n} \sum_{i=1}^n \phi_\alpha(y_i - \mathbf{x}_{Ai}^T \mathbf{c}_A - \mathbf{W}(\mathbf{z}_i)^T \boldsymbol{\gamma}) = \frac{1}{n} \sum_{i=1}^n \phi_\alpha(\epsilon_i - u_{ni} - \tilde{\mathbf{x}}_i^T \boldsymbol{\theta}_1 - \tilde{\mathbf{W}}(\mathbf{z}_i)^T \boldsymbol{\theta}_2),
$$

and the minimizers are $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \underset{(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\arg\min} \frac{1}{n} \sum_{i=1}^n \phi_\alpha(\epsilon_i - u_{ni} - \tilde{\mathbf{x}}_i^T \boldsymbol{\theta}_1 - \tilde{\mathbf{W}}(\mathbf{z}_i)^T \boldsymbol{\theta}_2)$.

6.2 Lemmas and theoretical proof of theorems

Lemma 6.1 (Properties of Loss Function)**.** *For the asymmetric squares loss function $\phi_\alpha(\cdot)$, the derivative of $\phi_\alpha(r)$ as $\psi_\alpha(r)$, the second-order derivative of $\phi_\alpha(r)$ as $\varphi_\alpha(r)$,*

(1) $\phi_\alpha(\cdot)$ is continuous differentiable. Moreover, for any $r, r_0 \in \mathbb{R}$, we have

$$
c_1 \cdot (r - r_0)^2 \le \phi_\alpha(r) - \phi_\alpha(r_0) - \psi_\alpha(r_0) \cdot (r - r_0) \le c_2 \cdot (r - r_0)^2. \tag{6.1}
$$

(2) $\psi_\alpha(\cdot)$ is Lipschitz continuous, which means for any $r, r_0 \in \mathbb{R}$, we have

$$
2c_1 |r - r_0| \le |\psi_\alpha(r) - \psi_\alpha(r_0)| \le 2c_2 |r - r_0|. \tag{6.2}
$$

(3) $\varphi_\alpha(r) \le 2c_2$ for any r.

Proof. (1) and (2) can be found in Gu and Zou (2016), (3) is from the definition of $\varphi_\alpha(u)$. □

Lemma 6.2. *Properties of centered spline basis functions:*

(1) $\max_i \mathbb{E}||\mathbf{W}(\mathbf{z}_i)|| \le m_1$, for some positive constant m_1 for a sufficiently large n;

(2) There exists positive constants m_2 and $m_2^{'}$ such that for n sufficiently large

$$
m_2 k_n^{-1} \le \mathbb{E}[\lambda_{min}(\mathbf{W}(\mathbf{z}_i)\mathbf{W}(\mathbf{z}_i)^T)] \le \mathbb{E}[\lambda_{max}(\mathbf{W}(\mathbf{z}_i)\mathbf{W}(\mathbf{z}_i)^T)] \le m_2^{'} k_n^{-1};
$$

(3) There exists positive constant m_3 such that for n sufficiently large $\mathbb{E}||W_B^{-1}|| \le m_3 \sqrt{k_n n^{-1}}$;

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

Proof. The proof for (1) and (2) can be found in Sherwood and Wang (2016). For (3), we need to show that $\mathbb{E}[\lambda_{min}(W_B^2)] > Cnk_n^{-1}$ since $||W_B^{-1}|| = \lambda_{max}(W_B^{-1}) = \lambda_{min}^{-1/2}(W_B^2)$. Through the definition of W_B, we have

$$\lambda_{min}(W_B^2) = \lambda_{min}(\sum_{i=1}^{n} \mathbb{E}[\varphi_\alpha(\epsilon_i)]\mathbf{W}(\mathbf{z}_i)\mathbf{W}(\mathbf{z}_i)^T) \geq C\sum_{i=1}^{n} \lambda_{min}(\mathbf{W}(\mathbf{z}_i)\mathbf{W}(\mathbf{z}_i)^T) \geq Ck_n^{-1}n.$$

For (4), $||\tilde{\mathbf{W}}(\mathbf{z}_i)||^2 = \mathbf{W}(\mathbf{z}_i)^T W_B^{-2}\mathbf{W}(\mathbf{z}_i) \leq ||\mathbf{W}(\mathbf{z}_i)||^2 \cdot \lambda_{max}^2(W_B^{-1}) = O_p(\frac{k_n}{n})$. □

Lemma 6.3. *Properties of the after-projection design matrix X^*:*
(1) $\lambda_{max}(n^{-1}{X^*}^T X^*) \leq C$ with probability one, where C is a positive constant;
(2) $n^{-1/2}X^* = n^{-1/2}\Delta_n^* + o_p(1)$. Also, $n^{-1}{X^*}^T B_n X^* = W_n + o_p(1)$;
(3) $\sum_{i=1}^n w_i \tilde{\mathbf{x}}_i \tilde{\mathbf{W}}(\mathbf{z}_i)^T = \mathbf{0}$.

Proof. Denote $\nu_j = \arg\min_{\nu \in \mathbb{R}^{J_n}} \sum_{i=1}^n \mathbb{E}[\varphi_\alpha(\epsilon_i)](X_{Aij} - \mathbf{W}(\mathbf{z}_i)^T\nu)^2$, then

$$\{\tilde{h}_j(\mathbf{z}_i)\}_{n\times q_n} = W(W^T B_n W)^{-1} W^T B_n X_A.$$

Recall $\frac{1}{\sqrt{n}}X^* = \frac{1}{\sqrt{n}}(I-P)X_A = \frac{1}{\sqrt{n}}\Delta_n + \frac{1}{\sqrt{n}}(H - PX_A)$, and

$$\frac{1}{n}\lambda_{max}\big((H-PX_A)^T(H-PX_A)\big) \leq \frac{1}{n} trace\big((H-PX_A)^T(H-PX_A)\big)$$
$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{q_n}(\tilde{h}_j(\mathbf{z}_i) - \hat{h}_j(\mathbf{z}_i))^2 = O_p(q_n n^{-2r/(2r+1)}) = o_p(1),$$

where the second last equality follows from Stone (1985). Then the following proof is similar to Lemma 3 in Sherwood and Wang (2016). □

Lemma 6.4 (Bernstein Inequality)**.** *Let $\xi_1, \xi_2, \ldots, \xi_n$ be independent mean-zero random variables, with uniform bounds $|\xi_i| \leq M$ and $v \geq \mathbb{V}ar(\sum_{i=1}^n \xi_i)$. Then for every positive t,*

$$\mathbb{P}\left(|\sum_{i=1}^{n}\xi_i| > t\right) \leq 2\exp\left(-\frac{t^2}{2(v+Mt/3)}\right).$$

Lemma 6.5. *Let $d_n = q_n + k_n$ and if Conditions 3.1-3.4 hold, then*

$$\frac{1}{n}\sum_{i=1}^{n}(\tilde{g}(z_i) - g_0(z_i))^2 = o_p(\frac{d_n}{n}).$$

Proof. Define $\Phi_i(a_n) \triangleq \Phi_i(a_n\theta_1, a_n\theta_2) = \phi_\alpha(\epsilon_i - a_n\tilde{\mathbf{x}}_i^T\theta_1 - a_n\tilde{W}(z_i)^T\theta_2 - u_{ni})$. Here we first show that $\forall \eta > 0$, there exists an $L > 0$ such that

$$\mathbb{P}\Big(\inf_{||\theta||=L}\frac{1}{d_n}\sum_{i=1}^{n}\big(\Phi_i(\sqrt{d_n}) - \Phi_i(0)\big) > 0\Big) > 1-\eta, \tag{6.3}$$

which implies with probability at least $1-\eta$ that there exists a local minimizer in the ball $\{\sqrt{d_n}\theta : ||\theta|| \leq L\}$, that means the local minimizer $\hat{\theta}$ satisfies $||\hat{\theta}|| = O_p(\sqrt{d_n})$.

Denote $\theta = (\theta_1^T, \theta_2^T)^T$ and by Taylor expansion, we have

$$\Phi_i(a_n) = \phi_\alpha(\epsilon_i - a_n\tilde{s}_i^T\theta - u_{ni}) = \phi_\alpha(\epsilon_i - u_{ni}) - \psi_\alpha(\epsilon_i - u_{ni})a_n\tilde{s}_i^T\theta + r_i(a_n)$$

where $r_i(a_n) = \frac{1}{2}\varphi_\alpha(\epsilon_i - u_{ni} - \xi_i a_n\tilde{s}_i^T\theta)(a_n\tilde{s}_i^T\theta)^2$ for some $0 < \xi_i < 1$. By Lemma 6.1,

$$|\psi_\alpha(\epsilon_i - u_{ni}) - \psi_\alpha(\epsilon_i)| \leq 2c_2|u_{ni}| = O_p(k_n^{-r}) = o_p(1).$$

Notify $\mathbb{E}_s[\cdot] \triangleq \mathbb{E}[\cdot|x_i, z_i]$. Thus, we have that $\frac{1}{d_n}\sum_{i=1}^n \Phi_i(\sqrt{d_n}) - \Phi_i(0)$ equals

$$\begin{aligned} & \frac{1}{d_n}\sum_{i=1}^n \mathbb{E}_s[\Phi_i(\sqrt{d_n}) - \Phi_i(0)] + \frac{1}{d_n}\sum_{i=1}^n \Big(\Phi_i(\sqrt{d_n}) - \Phi_i(0) - \mathbb{E}_s[\Phi_i(\sqrt{d_n}) - \Phi_i(0)]\Big) \\ = \quad & \Sigma_1 + \Sigma_2 + \Sigma_3 \end{aligned}$$

For Σ_3, define $D_i(a_n) = r_i(a_n) - \mathbb{E}_s[r_i(a_n)]$, then $\sup_{||\theta||\le L} d_n^{-1}\sum_{i=1}^n |D_i(\sqrt{d_n})| = o_p(1)$. Set F_{n1} as the event $\{\max_i ||\tilde{s}_i|| \le \alpha_1\sqrt{d_n/n}\}$. By Lemma 6.2 and the fact $\max_i ||\tilde{\mathbf{x}}_i|| = O(\sqrt{q_n/n})$, it implies that $\mathbb{P}(F_{n1}) \to 1$. Set F_{n2} as the event $\{\max_i |u_{ni}| \le \alpha_2 k_n^{-r}\}$, then according to Schumaker (2007), $\mathbb{P}(F_{n2}) \to 1$. Simplify $D_i(\sqrt{d_n})$ as D_i and are independent random variables satisfying $\mathbb{E}[D_i] = 0$ and $\max_i |D_i|\mathbb{I}(F_1 \bigcap F_2) \le \max_i C d_n |\tilde{s}_i^T\theta|^2 \mathbb{I}(F_1 \bigcap F_2) \le CL\frac{d_n^2}{n}$. Thus,

$$\begin{aligned} & \mathbb{V}ar(D_i\mathbb{I}(F_{n1}\bigcap F_{n2})|x_i, z_i) \le \mathbb{E}_s[r_i^2(\sqrt{d_n})\mathbb{I}(F_{n1}\bigcap F_{n2})] \\ \le \quad & \mathbb{E}_s[\frac{1}{2}\varphi_\alpha(\epsilon_i - u_{ni} - \xi_i a_n\tilde{s}_i^T\theta)(a_n\tilde{s}_i^T\theta)^2]^2 \le C d_n^2 ||\tilde{s}_i^T||^4||\theta||^4 \le C d_n^4/n^2. \end{aligned}$$

Applying Bernstein Inequality in Lemma 6.4, for any positive constant ϵ,

$$\mathbb{P}(|\sum_{i=1}^n D_i| > d_n\epsilon, F_{n1}\bigcap F_{n2}|x_i, z_i) \le 2\exp(\frac{-d_n^2\epsilon^2}{2(C\frac{d_n^4}{n} + C\frac{d_n^3 L\epsilon}{3n})}) \le C\exp(-\frac{n}{d_n^2}) \to 0,$$

which implies that $\sup_{||\theta||\le L} d_n^{-1}\sum_{i=1}^n |D_i(\sqrt{d_n})| = o_p(1)$.

For any $\epsilon > 0$, $\mathbb{P}(|\varphi_\alpha(\epsilon_i - u_{ni}) - \varphi_\alpha(\epsilon_i)| > \epsilon) = \max\{\mathbb{P}(u_{ni} < \epsilon_i < 0), \mathbb{P}(0 < \epsilon_i < u_{ni})\}$,

$$\max\{\mathbb{P}(u_{ni} < \epsilon_i < 0), \mathbb{P}(0 < \epsilon_i < u_{ni})\} \le C\cdot|u_{ni}| = O_p(k_n^{-r}) \to 0$$

which implies $|\varphi_\alpha(\epsilon_i - u_{ni}) - \varphi_\alpha(\epsilon_i)| = o_p(1)$. Then, for Σ_1, through Taylor expansion,

$$\begin{aligned} & \frac{1}{d_n}\sum_{i=1}^n \mathbb{E}_s\big[\Phi_i(\sqrt{d_n}) - \Phi_i(0))\big] \\ = \quad & \frac{1}{d_n}\sum_{i=1}^n \mathbb{E}_s\Big[-\psi_\alpha(\epsilon_i - u_{ni})\sqrt{d_n}\tilde{s}_i^T\theta + \frac{1}{2}\varphi_\alpha(\epsilon_i - u_{ni})d_n(\tilde{s}_i^T\theta)^2(1 + o(1))\Big] \\ = \quad & \sum_{i=1}^n \mathbb{E}_s\Big[\frac{1}{2}\varphi_\alpha(\epsilon_i)(\tilde{\mathbf{x}}_i^T\theta_1)^2(1 + o_p(1))\Big] + \sum_{i=1}^n \mathbb{E}_s\Big[\frac{1}{2}\varphi_\alpha(\epsilon_i)(\tilde{W}_i^T\theta_2)^2(1 + o_p(1))\Big] \\ & + \sum_{i=1}^n \mathbb{E}_s\Big[\frac{1}{2}\varphi_\alpha(\epsilon_i)\theta_1^T\tilde{\mathbf{x}}_i\tilde{W}_i^T\theta_2(1 + o_p(1))\Big] \\ = \quad & C\cdot\theta_1^T W_n\theta_1(1 + o_p(1)) + C\cdot||\theta_2||^2(1 + o_p(1)) = O_p(||\theta||^2), \end{aligned}$$

where $C > 0$ is some constant and the last second equation from Lemma 6.2 and 6.3.

As for Σ_2, $\mathbb{E}_s[\psi_\alpha(\epsilon_i)\tilde{s}_i\theta] = 0$ and $\mathbb{E}[\psi_\alpha(\epsilon_i)]^2 = \mathbb{E}\big[2\epsilon_i|\alpha - \mathbb{I}(\epsilon_i < 0)|\big]^2 \le C\cdot\mathbb{E}[\epsilon_i]^2 \le C$. Thus,

$$\mathbb{V}ar(\psi_\alpha(\epsilon_i)\tilde{s}_i\theta|x_i, z_i) = \mathbb{E}[\psi_\alpha(\epsilon_i)]^2(\tilde{s}_i\theta)^2 \le C\frac{d_n}{n}||\theta||^2,$$

which means $\mathbb{V}ar(\Sigma_2|x_i, z_i) \le \frac{1}{d_n}\sum_{i=1}^n \mathbb{V}ar(\psi_\alpha(\epsilon_i)\tilde{s}_i\theta|x_i, z_i) = O(||\theta||^2)$, i.e., $|\Sigma_2| = O_p(||\theta||)$. Hence, Σ_2 is dominated by Σ_1 for sufficiently large L, thus (6.3) holds.

From $||\hat{\theta}|| = O_p(\sqrt{d_n})$ and the definition of θ, $||W_B(\hat{\gamma} - \gamma_0)|| = O_p(\sqrt{d_n})$. Notice that

$\mathbb{E}[\varphi_\alpha(\epsilon_i)]$ is uniformly bounded away from zero,then

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\varphi_\alpha(\epsilon_i)](\tilde{g}(z_i)-g_0(z_i))^2 \le \frac{1}{n}C(\hat{\gamma}-\gamma_0)^T W_B^2(\hat{\gamma}-\gamma_0)+O_p(k_n^{-2r})=O_p(\frac{d_n}{n}).$$

Lemma 6.6. *Set* $\tilde{\theta}_1 = n^{-1/2}(X^{*^T}B_nX^*)^{-1}X^{*^T}\psi_\alpha(\epsilon)$ *where* $\psi_\alpha(\epsilon)=(\psi_\alpha(\epsilon_1),\ldots,\psi_\alpha(\epsilon_n))^T$. *Assume Conditions 3.1-3.4 hold, then* $||\tilde{\theta}_1|| = O_p(\sqrt{q_n})$.

Proof. From the definition of $\tilde{\theta}_1$ and Lemma 6.3,

$$\begin{aligned}\tilde{\theta}_1 &= \frac{1}{\sqrt{n}}(W_n+o_p(1))^{-1}(\Delta_n^T+o_p(1))\psi_\alpha(\epsilon)=\frac{1}{\sqrt{n}}W_n^{-1}(\Delta_n^T\psi_\alpha(\epsilon)(1+o_p(1))\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}W_n^{-1}\delta_i\psi_\alpha(\epsilon_i)\triangleq\sum_{i=1}^{n}D_{n,i}.\end{aligned}$$

And $D_{n,i}$ are independent random variables satisfying $\mathbb{E}[\tilde{\theta}_1]=0$ and

$$\mathbb{E}[||\tilde{\theta}_1||^2]=\sum_{i=1}^{n}\mathbb{E}[||D_{n,i}||^2]=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\psi_\alpha^2(\epsilon_i)(\delta_i^TW_n^{-2}\delta_i)]\le\frac{C}{n}\sum_{i=1}^{n}\mathbb{E}[||\delta_i||^2]=O_p(q_n),$$

where the second last inequality follows from the fact

$$\mathbb{E}[\psi_\alpha^2(\epsilon_i)]=\mathbb{E}[2|\alpha-\mathbb{I}(\epsilon_i<0)|\epsilon_i]^2\le C\mathbb{E}[\epsilon_i^2]=O(1).$$

Thus, we have $||\tilde{\theta}_1||=O_p(\sqrt{q_n})$.

Lemma 6.7. *Assume Condition 3.1-3.4 hold, then for any positive constant* C,

$$\sup_{||\theta_2||\le C\sqrt{d_n}}\frac{1}{n}\sum_{i=1}^{n}|\mathbb{I}(\epsilon_i<u_{ni}+\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)-\mathbb{I}(\epsilon_i<0)|=o_p(1).$$

Proof. Partition the left side of the equation into two parts. For $||\theta_2||\le C\sqrt{d_n}$

$$\begin{aligned}&\sup\frac{1}{n}\sum_{i=1}^{n}|\mathbb{I}(\epsilon_i<u_{ni}+\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)-\mathbb{I}(\epsilon_i<0)|\\ \le\ &\sup\frac{1}{n}\sum_{i=1}^{n}|\mathbb{I}(\epsilon_i<u_{ni}+\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)-\mathbb{I}(\epsilon_i<0)-\mathbb{P}(\epsilon_i<u_{ni}+\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)+\mathbb{P}(\epsilon_i<0)|\\ &+\sup\frac{1}{n}\sum_{i=1}^{n}|\mathbb{P}(\epsilon_i<u_{ni}+\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)-\mathbb{P}(\epsilon_i<0)|\triangleq I_1+I_2.\end{aligned}$$

$$\begin{aligned}I_2 &= \sup_{||\theta_2||\le C\sqrt{d_n}}\frac{1}{n}\sum_{i=1}^{n}|F_i(u_{ni}+\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)-F_i(0)|\\ &\le \sup_{||\theta_2||\le C\sqrt{d_n}}\frac{C}{n}\sum_{i=1}^{n}|u_{ni}+\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2|\le C\sup_{||\theta_2||\le\sqrt{d_n}}\max_i|u_{ni}|+\max_i\|\tilde{\mathbf{W}}(\mathbf{z}_i)^T\|\cdot\|\theta_2\|\\ &= O_p(k_n^{-r}+\sqrt{\frac{k_nd_n}{n}})=o_p(1)\end{aligned}$$

where the second last equation follows from Lemma 6.2 and $\max_i|u_{ni}|=O(k_n^{-r})$.

As for I_1, $v_i=\mathbb{I}(\epsilon_i<u_{ni}+\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)-\mathbb{I}(\epsilon_i<0)-\mathbb{P}(\epsilon_i<u_{ni}+\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)+\mathbb{P}(\epsilon_i<0)$, which are independent mean-zero random variables satisfying $|v_i|\le 2$. Note that $\mathbb{V}ar(v_i)=\mathbb{E}[\mathbb{I}(\epsilon_i<u_{ni}+\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)-\mathbb{I}(\epsilon_i<0)]^2$ and $\mathbb{I}(\epsilon_i<u_{ni}+\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)-\mathbb{I}(\epsilon_i<0)$ is nonzero

only when $0 < \epsilon_i < u_{ni} + \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2$ or $0 > \epsilon_i > u_{ni} + \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2$, depending on the sign of $u_{ni} + \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2$. Thus, under the set $\{\theta_2 : ||\theta_2|| \le C\sqrt{d_n}\}$,

$$\begin{aligned}\sum_{i=1}^n \mathbb{V}ar(v_i) &\le \sum_{i=1}^n \mathbb{P}\Big(0 < \epsilon_i < u_{ni} + \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2\Big) + \mathbb{P}\Big(0 > \epsilon_i > u_{ni} + \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2\Big)\\ \le \quad \sum_{i=1}^n \mathbb{P}\Big(|\epsilon_i| < |u_{ni} + \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2|\Big) &\le \sum_{i=1}^n C|u_{ni} + \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2| = O_p(nk_n^{-r} + \sqrt{nk_nd_n}),\end{aligned}$$

Then apply Bernstein Inequality in Lemma 6.4, for any $\epsilon > 0$,

$$\sup_{||\theta_2|| \le C\sqrt{d_n}} \mathbb{P}(|\sum_{i=1}^n v_i| > n\epsilon) \le 2\exp(-\frac{n^2\epsilon^2}{2(nk_n^{-r} + \sqrt{nd_nk_n} + 2n\epsilon/3)}) \le 2\exp(-n\epsilon) \to 0.$$

Lemma 6.8. *Assume Condition 3.1-3.4 hold, then for any finite positive constants M and C,*

$$\sup_{||\theta_1 - \tilde{\theta}_1|| \le M, ||\theta_2|| \le C\sqrt{d_n}} \left| \frac{1}{n}\sum_{i=1}^n \Big(\psi_\alpha(\epsilon_i - u_{ni} - \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2) - \psi_\alpha(\epsilon_i)\Big)\tilde{\mathbf{x}}_i^T(\theta_1 - \tilde{\theta}_1)\right| = o_p(1)$$

$$\sup_{||\theta_1 - \tilde{\theta}_1|| \le M, ||\theta_2|| \le C\sqrt{d_n}} \left| \frac{1}{n}\sum_{i=1}^n \Big(\varphi_\alpha(\epsilon_i - u_{ni} - \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2) - \varphi_\alpha(\epsilon_i)\Big)((\tilde{\mathbf{x}}_i^T\theta_1)^2 - (\tilde{\mathbf{x}}_i^T\tilde{\theta}_1)^2)\right| = o_p(1).$$

Proof. For the first part, notice that for $||\theta_1 - \tilde{\theta}_1|| \le M, ||\theta_2|| \le C\sqrt{d_n}$,

$$\begin{aligned}&\sup\left| \frac{1}{n}\sum_{i=1}^n \Big(\psi_\alpha(\epsilon_i - u_{ni} - \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2) - \psi_\alpha(\epsilon_i)\Big)\tilde{\mathbf{x}}_i^T(\theta_1 - \tilde{\theta}_1)\right|\\ \le \quad &\sup\max_i \Big|\psi_\alpha(\epsilon_i - u_{ni} - \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2) - \psi_\alpha(\epsilon_i)\Big| \cdot \max_i ||\tilde{\mathbf{x}}_i^T|| \cdot ||\theta_1 - \tilde{\theta}_1||\\ \le \quad &\sup\max_i C \cdot \Big|u_{ni} + \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2\Big| \cdot \max_i ||\tilde{\mathbf{x}}_i^T|| \cdot ||\theta_1 - \tilde{\theta}_1||\\ \le \quad &O_p(k_n^{-r} + \sqrt{d_nk_n/n}) \cdot O_p(\sqrt{q_n/n}) = o_p(1),\end{aligned}$$

where the last inequality follows from the fact $\max_i ||\tilde{\mathbf{x}}_i|| = O(\sqrt{q_n/n})$.

Consider the second part, for $||\theta_1 - \tilde{\theta}_1|| \le M, ||\theta_2|| \le C\sqrt{d_n}$,

$$\begin{aligned}&\sup\left| \frac{1}{n}\sum_{i=1}^n \Big(\varphi_\alpha(\epsilon_i - u_{ni} - \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2) - \varphi_\alpha(\epsilon_i)\Big)((\tilde{\mathbf{x}}_i^T\theta_1)^2 - (\tilde{\mathbf{x}}_i^T\tilde{\theta}_1)^2)\right|\\ \le \quad &\sup\frac{2}{n}\sum_{i=1}^n \Big|\mathbb{I}(\epsilon_i < u_{ni} + \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2) - \mathbb{I}(\epsilon_i < 0)\Big| \cdot \max_i \Big|(\tilde{\mathbf{x}}_i^T\theta_1)^2 - (\tilde{\mathbf{x}}_i^T\tilde{\theta}_1)^2\Big|.\end{aligned}$$

Notice that $\max_i ||\tilde{\mathbf{x}}_i|| = O(\sqrt{q_n/n})$ and use Lemma 6.6,

$$\begin{aligned}&\max_i \Big|((\tilde{\mathbf{x}}_i^T\theta_1)^2 - (\tilde{\mathbf{x}}_i^T\tilde{\theta}_1)^2)\Big| = \max_i \Big|(\theta_1 + \tilde{\theta}_1)^T\tilde{x}_i\tilde{x}_i^T(\theta_1 - \tilde{\theta}_1)\Big|\\ \le \quad &\max_i ||\tilde{x}_i||^2 \cdot ||\theta_1 - \tilde{\theta}_1|| \cdot \big(||\theta_1 - \tilde{\theta}_1|| + 2||\tilde{\theta}_1||\big) \le C \cdot \frac{q_n^{3/2}}{n} = o_p(1)\end{aligned}$$

Thus, apply Lemma 6.7, we have for $||\theta_1 - \tilde{\theta}_1|| \le M, ||\theta_2|| \le C\sqrt{d_n}$

$$\sup\left| \frac{1}{n}\sum_{i=1}^n \Big(\varphi_\alpha(\epsilon_i - u_{ni} - \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2) - \varphi_\alpha(\epsilon_i)\Big)((\tilde{\mathbf{x}}_i^T\theta_1)^2 - (\tilde{\mathbf{x}}_i^T\tilde{\theta}_1)^2)\right| = o_p(1).$$

Lemma 6.9. *Assume Condition 3.1-3.4 hold, then* $||\hat{\theta}_1 - \tilde{\theta}_1|| = o_p(1)$.

Proof. Define $\tilde{Q}_i(\theta_1,\tilde{\theta}_1,\theta_2)=\phi_\alpha(\epsilon_i-u_{ni}-\tilde{\mathbf{x}}_i^T\theta_1-\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)-\phi_\alpha(\epsilon_i-u_{ni}-\tilde{\mathbf{x}}_i^T\tilde{\theta}_1-\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)$. We first show that for any positive constants C and M,

$$\mathbb{P}\left(\inf_{||\theta_1-\tilde{\theta}_1||>M,||\theta_2||\le C\sqrt{d_n}}\frac{1}{n}\sum_{i=1}^n\tilde{Q}_i(\theta_1,\tilde{\theta}_1,\theta_2)>0\right)\to 1. \tag{6.4}$$

Notice $\tilde{Q}_i(\theta_1,\tilde{\theta}_1,\theta_2)=(\tilde{Q}_i(\theta_1,\tilde{\theta}_1,\theta_2)-\mathbb{E}[\tilde{Q}_i(\theta_1,\tilde{\theta}_1,\theta_2)])+\mathbb{E}[\tilde{Q}_i(\theta_1,\tilde{\theta}_1,\theta_2)]$. Now we prove that

$$\sup_{||\theta_1-\tilde{\theta}_1||\le M,||\theta_2||\le C\sqrt{d_n}}|\frac{1}{n}\sum_{i=1}^n\mathbb{E}[\tilde{Q}_i(\theta_1,\tilde{\theta}_1,\theta_2)]-\frac{1}{2n}\Big(\theta_1^TW_n\theta_1-\tilde{\theta}_1^TW_n\tilde{\theta}_1\Big)|=o_p(1).$$

Through Taylor Expansion,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n\mathbb{E}[\tilde{Q}_i(\theta_1,\tilde{\theta}_1,\theta_2)] &= \frac{1}{n}\sum_{i=1}^n\mathbb{E}\Big[-\psi_\alpha(\epsilon_i-u_{ni}-\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)\tilde{\mathbf{x}}_i^T(\theta_1-\tilde{\theta}_1)\\ &\quad+\frac{1}{2}\varphi_\alpha(\epsilon_i-u_{ni}-\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)((\tilde{\mathbf{x}}_i^T\theta_1)^2-(\tilde{\mathbf{x}}_i^T\tilde{\theta}_1)^2)(1+o(1))\Big]\end{aligned}$$

Then applying Lemma 6.8, we have

$$\frac{1}{n}\sum_{i=1}^n\mathbb{E}[\tilde{Q}_i(\theta_1,\tilde{\theta}_1,\theta_2)] = \frac{1}{n}\sum_{i=1}^n\mathbb{E}\Big[\frac{1}{2}\varphi_\alpha(\epsilon_i)((\tilde{\mathbf{x}}_i^T\theta_1)^2-(\tilde{\mathbf{x}}_i^T\tilde{\theta}_1)^2)\Big](1+o_p(1))$$

under the set $\{(\theta_1,\theta_2):||\theta_1-\tilde{\theta}_1||\le M,||\theta_2||\le C\sqrt{d_n}\}$, for any positive constants M and C, where the last equation follows the fact that $\mathbb{E}[\phi_\alpha(\epsilon_i)]=0,\ \ \forall i=1,\dots,n$. And this implies

$$\sup_{||\theta_1-\tilde{\theta}_1||\le M,||\theta_2||\le C\sqrt{d_n}}|\frac{1}{n}\sum_{i=1}^n\mathbb{E}[\tilde{Q}_i(\theta_1,\tilde{\theta}_1,\theta_2)]-\frac{1}{2n}\Big(\theta_1^TW_n\theta_1-\tilde{\theta}_1^TW_n\tilde{\theta}_1\Big)|=o_p(1).$$

Next we introduce

$$\begin{aligned}R_{i,n}(\theta_1) &= \phi_\alpha(\epsilon_i-u_{ni}-\tilde{\mathbf{x}}_i^T\theta_1-\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)-\phi_\alpha(\epsilon_i-u_{ni}-\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)\\ &\quad+\psi_\alpha(\epsilon_i-u_{ni}-\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)\tilde{\mathbf{x}}_i^T\theta_1.\end{aligned}$$

Through Taylor Expansion, there exits $0<\xi_{i,\theta_1},\ \xi_{i,\tilde{\theta}_1}<1$ satisfying

$$\begin{aligned}R_{i,n}(\theta_1) &= \frac{1}{2}\varphi_\alpha(\epsilon_i-u_{ni}-\xi_{i,\theta_1}\tilde{\mathbf{x}}_i^T\theta_1-\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)(\tilde{\mathbf{x}}_i^T\theta_1)^2\\ R_{i,n}(\tilde{\theta}_1) &= \frac{1}{2}\varphi_\alpha(\epsilon_i-u_{ni}-\xi_{i,\tilde{\theta}}\tilde{\mathbf{x}}_i^T\tilde{\theta}-\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)(\tilde{\mathbf{x}}_i^T\tilde{\theta}_1)^2.\end{aligned}$$

Notice $\frac{1}{2}\varphi_\alpha(\epsilon_i-u_{ni}-\xi_{i,\theta_1}\tilde{\mathbf{x}}_i^T\theta_1-\tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)\le 1$, thus

$$\begin{aligned}&\sup_{||\theta_1-\tilde{\theta}_1||\le M,||\theta_2||\le C\sqrt{d_n}}\Big|\frac{1}{n}\sum_{i=1}^nR_{i,n}(\theta_1)-R_{i,n}(\tilde{\theta}_1)-\mathbb{E}[R_{i,n}(\theta_1)-R_{i,n}(\tilde{\theta}_1)]\Big|\\ \le&\sup_{||\theta_1-\tilde{\theta}_1||\le M,||\theta_2||\le C\sqrt{d_n}}2\Big|\frac{1}{n}\sum_{i=1}^n((\tilde{\mathbf{x}}_i^T\theta_1)^2-(\tilde{\mathbf{x}}_i^T\tilde{\theta}_1)^2)\Big|\\ \le&\sup_{||\theta_1-\tilde{\theta}_1||\le M,||\theta_2||\le C\sqrt{d_n}}C\max_i||\tilde{x}_i||^2\cdot||\theta_1-\tilde{\theta}_1||\cdot||\tilde{\theta}_1||\le C\frac{q_n^{3/2}}{n}=o_p(1)\end{aligned}$$

where last inequality follows from the fact $\max_i||\tilde{\mathbf{x}}_i||=O(\sqrt{q_n/n})$ and Lemma 6.6.

Hence, under the set $\{(\theta_1,\theta_2):||\theta_1-\tilde{\theta}_1||\le M,||\theta_2||\le C\sqrt{d_n}\}$, by Lemma 6.8 and the fact

$\mathbb{E}[\psi_\alpha(\epsilon_i)] = 0,\ \forall i = 1, \ldots, n$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\tilde{Q}_i(\theta_1, \tilde{\theta}_1, \theta_2) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\tilde{Q}_i(\theta_1, \tilde{\theta}_1, \theta_2)] \\
= \ &\frac{1}{n}\sum_{i=1}^{n}\Bigg(- \psi_\alpha(\epsilon_i - u_{ni} - \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)\tilde{\mathbf{x}}_i^T(\theta_1 - \tilde{\theta}_1) + R_{i,n}(\theta_1) - R_{i,n}(\tilde{\theta}_1) \\
&+\mathbb{E}[\psi_\alpha(\epsilon_i - u_{ni} - \tilde{\mathbf{W}}(\mathbf{z}_i)^T\theta_2)\tilde{\mathbf{x}}_i^T(\theta_1 - \tilde{\theta}_1)\ - \mathbb{E}[R_{i,n}(\theta_1) - R_{i,n}(\tilde{\theta}_1)]\Bigg) \\
= \ &\frac{1}{n}\sum_{i=1}^{n}\Bigg(- \psi_\alpha(\epsilon_i)\tilde{\mathbf{x}}_i^T(\theta_1 - \tilde{\theta}_1)(1 + o_p(1)) + R_{i,n}(\theta_1) - R_{i,n}(\tilde{\theta}_1) - \mathbb{E}[R_{i,n}(\theta_1) - R_{i,n}(\tilde{\theta}_1)]\Bigg).
\end{aligned}$$

And this shows

$$\sup_{||\theta_1 - \tilde{\theta}_1|| \le M, ||\theta_2|| \le C\sqrt{d_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\Big(\tilde{Q}_i(\theta_1, \tilde{\theta}_1, \theta_2) - \mathbb{E}[\tilde{Q}_i(\theta_1, \tilde{\theta}_1, \theta_2)] + \psi_\alpha(\epsilon_i)\tilde{\mathbf{x}}_i^T(\theta_1 - \tilde{\theta}_1)\Big)\right| = o_p(1).$$

Notice the definition of $\tilde{\theta}_1$ and Lemma 6.3, it follows that

$$\sum_{i=1}^{n}\psi_\alpha(\epsilon_i)\tilde{\mathbf{x}}_i^T(\theta_1 - \tilde{\theta}_1) = (\theta_1 - \tilde{\theta}_1)^T n^{-1/2}X^{*T}\psi_\alpha(\epsilon) = (\theta_1 - \tilde{\theta}_1)^T W_n\tilde{\theta}_1(1 + o_p(1)).$$

Then, by using the result

$$\sup_{||\theta_1 - \tilde{\theta}_1|| \le M, ||\theta_2|| \le C\sqrt{d_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\tilde{Q}_i(\theta_1, \tilde{\theta}_1, \theta_2)] - \frac{1}{2n}\Big(\theta_1^T W_n\theta_1 - \tilde{\theta}_1^T W_n\tilde{\theta}_1\Big)\right| = o_p(1),$$

we have for $||\theta_1 - \tilde{\theta}_1|| \le M, ||\theta_2|| \le C\sqrt{d_n}$

$$\sup\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{Q}_i(\theta_1, \tilde{\theta}_1, \theta_2) - \frac{1}{2n}\Big(\theta_1^T W_n\theta_1 - \tilde{\theta}_1^T W_n\tilde{\theta}_1\Big) + \frac{1}{n}(\theta_1 - \tilde{\theta}_1)^T W_n\tilde{\theta}_1\right| = o_p(1),$$

which means

$$\sup_{||\theta_1 - \tilde{\theta}_1|| \le M, ||\theta_2|| \le C\sqrt{d_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\tilde{Q}_i(\theta_1, \tilde{\theta}_1, \theta_2) - \frac{1}{2n}(\theta_1 - \tilde{\theta}_1)^T W_n(\theta_1 - \tilde{\theta}_1)\right| = o_p(1).$$

Thus from Condition 3.2, for any θ_1 satisfying $||\theta_1 - \tilde{\theta}_1|| \ge M > 0$, we have

$$\frac{1}{2n}(\theta_1 - \tilde{\theta}_1)^T W_n(\theta_1 - \tilde{\theta}_1) > 0,$$

which implies (6.4) holds. Thus, the result follows. □

Proof of Theorem 3.1.

(1) For the first part of (3.4), from the results of Lemma 6.6 and Lemma 6.9, $||\hat{\mathbf{c}}_A^* - \boldsymbol{\beta}_A^*|| = O_p(\sqrt{q_n/n})$. Hence through the fact that $\hat{c}_A^* = \hat{\boldsymbol{\beta}}_A^*$, our first result can be proved.

(2) For the second part of (3.4), $\hat{g}(\mathbf{z}_i) = \tilde{g}(\mathbf{z}_i)$, thus by Lemma 6.5, the second result holds. □

Lemma 6.10 (Tao and An, 1997)**.** *Suppose the objective function $L(\boldsymbol{\theta})$ can be decomposed as the difference of two convex functions $k(\boldsymbol{\theta})$ and $l(\boldsymbol{\theta})$, i.e., $L(\boldsymbol{\theta}) = k(\boldsymbol{\theta}) - l(\boldsymbol{\theta})$, with the corresponding subdifferential functions $\partial k(\boldsymbol{\theta})$ and $\partial l(\boldsymbol{\theta})$ respectively. Let $dom(k) = \{\boldsymbol{\theta} : k(\boldsymbol{\theta}) < \infty\}$ be the effective domain of $k(\boldsymbol{\theta})$ and $\boldsymbol{\theta}^*$ be a point that has neighbourhood U such that $\partial l(\boldsymbol{\theta}) \cap \partial k(\boldsymbol{\theta}^*) \ne \emptyset$, $\forall\ \boldsymbol{\theta} \in U \cap dom(k)$. Then $\boldsymbol{\theta}^*$ is a local minimizer of $f(\boldsymbol{\theta})$.*

Lemma 6.11 (Chung, 2008)**.** *Denote $\{X_i\}_{i=1}^n$ a sequence of independent real valued random variables with $\mathbb{E}[X_i]=0$ and $S_n=\sum_{i=1}^n X_i$. Then for $r\geq 2$, the following inequality holds:*

$$\mathbb{E}[|S_n|^r]\leq C_r n^{r/2-1}\sum_{i=1}^n \mathbb{E}[|X_i|^r],$$

where C_r is some constant only depending on r.

Lemma 6.12. *Assume Conditions 3.1-3.5p are satisfied. The parameter $\lambda=o\left(n^{-(1-C_4)/2}\right)$, $q_n=o(n\lambda^2)$, $k_n=o(n\lambda^2)$ and $p=o\big((n\lambda^2)^k\big)$. For the oracle estimator $(\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\xi}}^*)$, with probability tending to one,*

$$\begin{aligned} s_j(\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\xi}}^*) &= 0, \quad j=1,\ldots,q_n \text{ or } j=p+1,\ldots,p+D_n, && (6.5)\\ |\hat{\beta}_j^*| &\geq (a+1/2)\lambda, \quad j=1,\ldots,q_n, && (6.6)\\ |s_j(\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\xi}}^*)| &\leq \lambda, \quad j=q_n+1,\ldots,p. && (6.7)\end{aligned}$$

Proof.

(1) Proof for (6.5). For $j=1,\ldots,q_n$ or $j=p+1,\ldots,p+D_n$, by the first order necessary condition of the optimal solution,

$$s_j(\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\xi}}^*)=\frac{\partial}{\partial\beta_j}\Big(\frac{1}{n}\sum_{i=1}^n\phi_\alpha(y_i-x_i^T\boldsymbol{\beta}-\boldsymbol{\Pi}(z_i)^T\boldsymbol{\xi})\Big)\mid_{(\boldsymbol{\beta},\boldsymbol{\xi})=(\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\xi}}^*)}=0.$$

(2) It's sufficient for Inequality (6.6) that $\mathbb{P}\left(\min\limits_{1\leq j\leq q_n}|\hat{\beta}_j^*|\geq(a+1/2)\lambda)\right)\to 1,\text{as}\quad n\to\infty$. Notice that $\min\limits_{1\leq j\leq q_n}|\hat{\beta}_j^*|\geq\min\limits_{1\leq j\leq q_n}|\beta_j^*|-\max\limits_{1\leq j\leq q_n}|\beta_j^*-\hat{\beta}_j^*|$. By Theorem 3.1 and Condition 3.4, $\|\hat{\boldsymbol{\beta}}_A^*-\boldsymbol{\beta}_A^*\|=O_p(\sqrt{\frac{q_n}{n}})=O_p(n^{-(1-C_3)/2})$, then, $\max\limits_{1\leq j\leq q_n}|\beta_j^*-\hat{\beta}_j^*|=O_p(n^{-(1-C_3)/2})=o_p(n^{-(1-C_4)/2})$. Also, Condition 3.5 shows $n^{(1-C_4)/2}\min\limits_{1\leq j\leq q_n}|\boldsymbol{\beta}_j^*|\geq C_5$. Thus, inequality (24) can hold by setting $\lambda=o(n^{-(1-C_4)/2})$.

(3) Proof for Inequality (6.7). For $j=q_n+1,\ldots,p$, recall the definition of $s_j(\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\xi}}^*)$ as

$$\begin{aligned} & s_j(\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\xi}}^*)=\frac{\partial}{\partial\beta_j}\Big(\frac{1}{n}\sum_{i=1}^n\phi_\alpha(y_i-x_i^T\boldsymbol{\beta}-\boldsymbol{\Pi}(z_i)^T\boldsymbol{\xi})\Big)\mid_{(\boldsymbol{\beta},\boldsymbol{\xi})=(\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\xi}}^*)}\\ =\;& -\frac{2}{n}\sum_{i=1}^n x_{ij}(y_i-x_i^T\hat{\boldsymbol{\beta}}^*-\boldsymbol{\Pi}(z_i)^T\hat{\boldsymbol{\xi}}^*)|\mathbb{I}(y_i-x_i^T\hat{\boldsymbol{\beta}}^*-\boldsymbol{\Pi}(z_i)^T\hat{\boldsymbol{\xi}}^*<0)-\alpha|\\ =\;& -\frac{2}{n}\sum_{i=1}^n x_{ij}(y_i-x_{Ai}^T\hat{\boldsymbol{\beta}}_A^*-\boldsymbol{\Pi}(z_i)^T\hat{\boldsymbol{\xi}}^*)|\mathbb{I}(y_i-x_{Ai}^T\hat{\boldsymbol{\beta}}_A^*-\boldsymbol{\Pi}(z_i)^T\hat{\boldsymbol{\xi}}^*<0)-\alpha|,\end{aligned}$$

where the last equality follows by the definition of oracle estimator $(\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\xi}}^*)$.

To prove the result, first we need to show that for $q_n+1\leq j\leq p$,

$$\mathbb{P}\left(\max_j\Big|\frac{2}{n}\sum_{i=1}^n x_{ij}(y_i-x_i^T\hat{\boldsymbol{\beta}}^*-\boldsymbol{\Pi}(z_i)^T\hat{\boldsymbol{\xi}}^*)|\mathbb{I}(y_i-x_i^T\hat{\boldsymbol{\beta}}^*-\boldsymbol{\Pi}(z_i)^T\hat{\boldsymbol{\xi}}^*<0)-\alpha|\Big|>\lambda\right)\to 0,$$

which is equivalent to show that

$$\mathbb{P}\left(\max_j\Big|\frac{2}{n}\sum_{i=1}^n x_{ij}(y_i-x_i^T\hat{\boldsymbol{\beta}}^*-\boldsymbol{W}(z_i)^T\hat{\boldsymbol{\gamma}})|\mathbb{I}(y_i-x_i^T\hat{\boldsymbol{\beta}}^*-\boldsymbol{W}(z_i)^T\hat{\boldsymbol{\gamma}}<0)-\alpha|\Big|>\lambda\right)\to 0.$$

Set

$$\epsilon_i(\boldsymbol{\beta}_A,\boldsymbol{\gamma})=y_i-x_{Ai}^T\boldsymbol{\beta}_A-\boldsymbol{W}(z_i)^T\boldsymbol{\gamma},\quad \epsilon_i^*=y_i-x_{Ai}^T\boldsymbol{\beta}_A^*-g_0(z_i)$$
$$\hat{\epsilon}_i=\epsilon_i(\hat{\boldsymbol{\beta}}_A^*,\hat{\boldsymbol{\gamma}})=y_i-x_{Ai}^T\hat{\boldsymbol{\beta}}_A^*-\boldsymbol{W}(z_i)^T\hat{\boldsymbol{\gamma}}=y_i-x_{Ai}^T\hat{\boldsymbol{\beta}}_A^*-\hat{g}(z_i).$$

We also set $I_j=\frac{2}{n}\sum_{i=1}^n x_{ij}\hat{\epsilon}_i|\mathbb{I}(\hat{\epsilon}_i\le 0)-\alpha|=I_{j1}+I_{j2}$, with

$$I_{j1}=\frac{2}{n}\sum_{i=1}^n x_{ij}(\hat{\epsilon}_i-\epsilon_i^*)|\mathbb{I}(\hat{\epsilon}_i\le 0)-\alpha|,\ I_{j2}=\frac{2}{n}\sum_{i=1}^n x_{ij}\epsilon_i^*|\mathbb{I}(\hat{\epsilon}_i\le 0)-\alpha|.$$

Let's first consider $\mathbb{P}\left(\max_{q_{n+1}\le j\le p}|I_{j1}|>\lambda/2\right)\to 0$. It follows from the proof of Lemma 6.5 that$\|\boldsymbol{\gamma}_0-\hat{\boldsymbol{\gamma}}\|_2=O_p(\sqrt{\frac{k_n d_n}{n}})$. Note that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n|\hat{\epsilon}_i-\epsilon_i^*|=\frac{1}{n}\sum_{i=1}^n|x_{Ai}^T(\hat{\boldsymbol{\beta}}_A^*-\boldsymbol{\beta}_A^*)+\boldsymbol{W}(z_i)^T(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)+u_{ni}|\\
\le\ &\frac{1}{n}\sum_{i=1}^n(|x_{Ai}^T(\hat{\boldsymbol{\beta}}_A^*-\boldsymbol{\beta}_A^*)|+|\boldsymbol{W}(z_i)^T(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)|+|u_{ni}|)\\
\le\ &(\frac{1}{n}\sum_{i=1}^n|x_{Ai}^T(\hat{\boldsymbol{\beta}}_A^*-\boldsymbol{\beta}_A^*)|^2)^{1/2}+(\frac{1}{n}\sum_{i=1}^n|\boldsymbol{W}(z_i)^T(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)|^2)^{1/2}+\sup_{1\le i\le n}|u_{ni}|\\
\le\ &\lambda_{max}^{1/2}(\frac{1}{n}X_AX_A^T)\|\boldsymbol{\beta}_A^*-\hat{\boldsymbol{\beta}}_A^*\|_2+(\frac{1}{n}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)^T\boldsymbol{W}^2(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0))^{1/2}+\sup_{1\le i\le n}|u_{ni}|\\
=\ &O_p(\sqrt{q_n/n}+\sqrt{d_n/n}+k_n^{-r}).
\end{aligned}$$

where the second inequality follows Jensen inequality and the last inequality applies Condition 3.2. Thus, by using Condition 3.2 again and Condition 3.3 and 3.4,

$$\begin{aligned}
\max_{q_{n+1}\le j\le p}|I_{j1}|\le\ &2\cdot\max_{q_{n+1}\le j\le p}|x_{ij}|\cdot\frac{1}{n}\sum_{i=1}^n|\hat{\epsilon}_i-\epsilon_i^*|\\
\le\ &C\cdot\frac{1}{n}\sum_{i=1}^n|\hat{\epsilon}_i-\epsilon_i^*|=O_p(\sqrt{q_n/n}+\sqrt{d_n/n}+k_n^{-r})=o_p(\lambda),
\end{aligned}$$

which means $\mathbb{P}\left(\max_{q_{n+1}\le j\le p}|I_{j1}|>\lambda/2\right)\to 0$. As for I_{j2},

$$I_{j2}\le\frac{2}{n}\sum_{i=1}^n x_{ij}\epsilon_i^*|\mathbb{I}(\hat{\epsilon}_i\le 0)-\mathbb{I}(\epsilon_i^*\le 0)|+\frac{2}{n}\sum_{i=1}^n x_{ij}\epsilon_i^*|\alpha-\mathbb{I}(\epsilon_i^*\le 0)|=I_{j21}+I_{j22}.$$

To evaluate I_{j21}, note that

$$|\mathbb{I}(\hat{\epsilon}_i\le 0)-\mathbb{I}(\epsilon_i^*\le 0)|=|\mathbb{I}(\epsilon_i^*\le\epsilon_i^*-\hat{\epsilon}_i)-\mathbb{I}(\epsilon_i^*\le 0)|\le\mathbb{I}(|\epsilon_i^*|\le|\epsilon_i^*-\hat{\epsilon}_i|).$$

So, we have for $j = q_{n+1}, \ldots, p$,

$$\max_{q_{n+1}\le j\le p}|I_{j21}| \le \max_{q_{n+1}\le j\le p} 2\times\frac{1}{n}\sum_{i=1}^{n}|x_{ij}|\times|\epsilon_i^*|\times\mathbb{I}(|\epsilon_i^*|\le|\epsilon_i^*-\hat{\epsilon}_i|)$$

$$\le \max_{q_{n+1}\le j\le p} C\times\frac{1}{n}\sum_{i=1}^{n}|\epsilon_i^*-\hat{\epsilon}_i| = O_p(\sqrt{q_n/n}+\sqrt{d_n/n}+k_n^{-r}) = o_p(\lambda).$$

Thus, we can show that $\mathbb{P}\left(\max_{q_{n+1}\le j\le p}|I_{j21}|>\lambda/4\right)\to 0$.

Now Consider I_{j22}, we define $\eta_i = \epsilon_i^*|\mathbb{I}(\epsilon_i^*\le 0)-\alpha|$, which is independent and satisfies $\mathbb{E}[\eta_i|x_i] = 0$. By Condition 3.1 we have $\mathbb{E}\left[\eta_i^{2k}|x_i\right] < \infty$. Also, x_{ij} is bounded because of Condition 3.2, thus by Lemma 6.11, we have $\mathbb{E}[I_{j22}^{2k}] = O(n^{-k})$. Therefore, by Markov Inequality, we have $\mathbb{P}(|I_{j22}|>\lambda) \le \frac{\mathbb{E}[I_{j22}^{2k}]}{(\lambda^{2k})} = O((n\lambda^2)^{-k})$, which contains $\mathbb{P}\left(\max_{q_n+1\le j\le p}|I_{j22}|>\lambda/4\right)\le\sum_{q_n+1}^{p}\mathbb{P}(|I_{j22}|>\lambda/4) = O(p(n\lambda^2)^{-k})\to 0$. Overall,

$$\begin{aligned}
&\mathbb{P}\left(\max_{q_n+1\le j\le p}|I_j|>\lambda\right)\le\mathbb{P}\left(\max_{q_n+1\le j\le p}|I_{j1}|>\lambda/2\right)+\mathbb{P}\left(\max_{q_n+1\le j\le p}|I_{j2}|>\lambda/2\right)\\
\le\;&\mathbb{P}\left(\max_{q_n+1\le j\le p}|I_{j1}|>\lambda/2\right)+\mathbb{P}\left(\max_{q_n+1\le j\le p}|I_{j22}|>\lambda/4\right)+\mathbb{P}\left(\max_{q_n+1\le j\le p}|I_{j21}|>\lambda/4\right)\\
\to\;&0.
\end{aligned}$$

we calculate the derivative of $H_\lambda(\theta)$ and define the subdifferential of $k(\boldsymbol{\theta})$ introduced in Section 3.2 is needed in the proof of Theorem 3.2. Notice that $H_\lambda(\theta)$ is differentiable everywhere,

$$H'_\lambda(\theta) = [(\theta-\lambda\mathrm{sgn}(\theta))/(a-1)]\mathbb{I}(\lambda\le|\theta|\le a\lambda)+\lambda\mathrm{sgn}(\theta)\mathbb{I}(|\theta|>a\lambda).$$

Define the subdifferential of $k(\boldsymbol{\theta})$ at $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ as follows: $\partial k(\boldsymbol{\theta}_0) = \{t : k(\boldsymbol{\theta})\ge k(\boldsymbol{\theta}_0)+t'(\boldsymbol{\theta}-\boldsymbol{\theta}_0), \forall\boldsymbol{\theta}\}$. Therefore, $\partial k(\boldsymbol{\beta},\boldsymbol{\xi}) = \left\{\boldsymbol{\kappa} = (\kappa_1,\ldots,\kappa_{p+D_n})^T\in\mathbb{R}^{p+D_n}\right\}$ has the following expression,

$$\kappa_j = \begin{cases} s_j(\boldsymbol{\beta},\boldsymbol{\xi})+\lambda l_j, & j=1,2,\ldots,p,\\ s_j(\boldsymbol{\beta},\boldsymbol{\xi}), & j=p+1,\ldots,p+D_n;\end{cases}$$

where $l_j = \mathrm{sgn}(\beta_j)$ if $\beta_j\neq 0$ or l_j takes value in $[-1,1]$ if $\beta_j = 0$. Similarly, $\partial l(\boldsymbol{\beta},\boldsymbol{\xi}) = \left\{\boldsymbol{\mu} = (\mu_1,\ldots,\mu_{p+D_n})^T\in\mathbb{R}^{p+D_n}\right\}$ has the following expression,

$$\mu_j = \begin{cases} H'_\lambda(\beta_j), & j=1,2,\ldots,p,\\ 0, & j=p+1,\ldots,p+D_n.\end{cases}$$

Proof of Theorem 3.2. We make use of Lemma 6.10 to prove our theorem. Consider any $(\boldsymbol{\beta}^T,\boldsymbol{\xi}^T)^T$ in a ball $\mathcal{B}(\lambda)$ with the center $(\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\xi}}^*)$ and radius $\lambda/2$. It is sufficient to show that for any $(\boldsymbol{\beta}^T,\boldsymbol{\xi}^T)^T\in\mathcal{B}(\lambda)$, with probability tending to one,

$$\partial l(\boldsymbol{\beta},\boldsymbol{\xi})\cap\partial k(\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\xi}}^*)\neq\emptyset.$$

Define the event $\mathcal{E}_1 = \{|\hat{\beta}_j^*|\ge(a+1/2)\lambda,\ 1\le j\le q_n\}$. Then by Lemma 6.12, $\mathbb{P}(\mathcal{E}_1)\to 1$, as $n\to\infty$. For $j=1,\ldots,q_n$, on the event $\mathcal{E}_1$, for any $(\boldsymbol{\beta}^T,\boldsymbol{\xi}^T)^T\in\mathcal{B}(\lambda)$,

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

So $H'_\lambda(\beta_j) = \lambda\mathrm{sgn}(\beta_j)$, i.e., $\mu_j = \frac{\partial l(\boldsymbol{\beta},\boldsymbol{\xi})}{\partial\beta_j} = \lambda\mathrm{sgn}(\beta_j)$. $\kappa_j = s_j(\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\xi}}^*)+\lambda l_j$, for $j=1,2,\ldots,q_n$.

By convex optimization theory or Lemma 6.12, $s_j(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\xi}}^*) = 0$. Then on the event $\mathcal{E}_1$, if $\mathrm{sgn}(\hat{\beta}_j^*) = \mathrm{sgn}(\beta_j)$, we have that $\kappa_j = \frac{\partial k(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\xi}}^*)}{\partial \beta_j} = \mu_j$. In fact, if $\mathrm{sgn}(\hat{\beta}_j^*) \neq \mathrm{sgn}(\beta_j)$, then on the event $\mathcal{E}_1$, $|\hat{\beta}_j^* - \beta_j| = |\hat{\beta}_j{}^*| + |\beta_j| \geq (a + 1/2)\lambda$, which causes contradition with that $(\boldsymbol{\beta}^T, \boldsymbol{\xi}^T)^T \in \mathcal{B}(\lambda)$.

Define the event $\mathcal{E}_2 = \{|s_j(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\xi}}^*)| \leq \lambda,\ q_n + 1 \leq j \leq p\}$. For $j = q_n + 1, \ldots, p$, by the construction of the oracle estimator, we have $\hat{\beta}_j^* = 0$. Then for any $(\boldsymbol{\beta}^T, \boldsymbol{\xi}^T)^T \in \mathcal{B}(\lambda)$,

$$\max_{q_n+1 \leq j \leq p} |\beta_j| \leq \max_{q_n+1 \leq j \leq p} |\hat{\beta}_j^*| + \max_{q_n+1 \leq j \leq p} |\hat{\beta}_j^* - \beta_j| \leq \lambda/2.$$

So in this situation, $\mu_j = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \beta_j} = 0$. On the other hand, $\kappa_j = s_j(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\xi}}^*) + \lambda l_j$ with $l_j \in [-1, 1]$. On the event $\mathcal{E}_2 = \{|s_j(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\xi}}^*)| \leq \lambda,\ q_n + 1 \leq j \leq p\}$, there exists $l_j^*,\ j = q_n + 1, \ldots, p$ such that $s_j(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\xi}}^*) + \lambda l_j^* = 0 = \mu_j$.

For $j = p+l,\ l = 1, \ldots, D_n$, by convex optimization theory or Lemma 6.12, $\kappa_j = s_j(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\xi}}^*) = 0$. Note that $\mu_j = \frac{\partial l(\boldsymbol{\beta}, \boldsymbol{\xi})}{\partial \xi_l} = 0$. So for $j = p + l,\ l = 1, \ldots, D_n$, $\kappa_j = \mu_j$.

Combined with all results above, on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have for any $(\boldsymbol{\beta}^T, \boldsymbol{\xi}^T)^T \in \mathcal{B}(\lambda)$,

$$\partial l(\boldsymbol{\beta}, \boldsymbol{\xi}) \cap \partial k(\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\xi}}^*) \neq \emptyset.$$

Note that $\mathbb{P}(\mathcal{E}_1 \cap \mathcal{E}_2) \geq 1 - \mathbb{P}(\bar{\mathcal{E}_1}) - \mathbb{P}(\bar{\mathcal{E}_2}) \to 1$, as $n \to \infty$. The proof has been completed. □

Declarations

Conflict of interest The authors declare no conflict of interest.

References

[1] A Buja, R Berk, L Brown, et al. *Models as approximations, part I: A conspiracy of nonlinearity and random regressors in linear regression*, 2014, arXiv preprint arXiv:1404.1578.

[2] K Chung. *The strong law of large numbers*, Selected Works of Kai Lai Chung, 2008, 47-52.

[3] Z J Daye, J Chen, H Li. *High−Dimensional Heteroscedastic Regression with an Application to eQTL Data Analysis*, Biometrics, 2012, 68(1): 316-326.

[4] J Fan, Q Li, Y Wang. *Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions*, Journal of the Royal Statistical Society Series B (Statistical Methodology), 2017, 79(1): 247-265.

[5] J Fan, R Li. *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of the American Statistical Association, 2001, 96(456): 1348-1360.

[6] J Fan, J Lv. *Sure independence screening for ultrahigh dimensional feature space*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2008, 70(5): 849-911.

[7] J Fan, L Xue, H Zou. *Strong oracle optimality of folded concave penalized estimation*, The Annals of Statistics, 2014, 42(3): 819-849.

[8] Y Gu, H Zou. *High-dimensional generalizations of asymmetric least squares regression and their applications*, The Annals of Statistics, 2016, 44(6): 2661-2694.

[9] T J Hastie, R J Tibshirani. *Generalized Additive Models*, Chapman & Hall, 1990.

[10] Y Kim, H Choi, H S Oh. *Smoothly clipped absolute deviation on high dimensions*, Journal of the American Statistical Association, 2008, 103(484): 1665-1673.

[11] C Kudo, I Ajioka, Y Hirata, et al. *Expression profiles of $EphA_3$ at both the RNA and protein level in the developing mammalian forebrain*, Journal of Comparative Neurology, 2005, 487(3): 255-269.

[12] M C Grant, S P Boyd. *CVX: Matlab software for disciplined convex programming, version 2.0 beta*, 2013, `http://cvxr.com/cvx`.

[13] M C Grant, S P Boyd. *Graph implementations for nonsmooth convex programs*, Recent Advances in Learning and Control, 2008, 95-110.

[14] National Research Council. *Frontiers in massive data analysis*, National Academies Press, 2013.

[15] W K Newey, J L Powell. *Asymmetric least squares estimation and testing*, Econometrica: Journal of the Econometric Society, 1987, 55(4): 819-847.

[16] R A Rigby, D M Stasinopoulos. *A semi-parametric additive model for variance heterogeneity*, Statistics and Computing, 1996, 6(1): 57-65.

[17] P M Robinson. *Root-N-consistent semiparametric regression*, Econometrica, 1988, 56(4): 931-954.

[18] L Schumaker. *Spline Functions: Basic Theory*, Cambridge University Press, 2007.

[19] B Sherwood, L Wang. *Partially linear additive quantile regression in ultra-high dimension*, The Annals of Statistics, 2016, 44(1): 288-317.

[20] F Sobotka, G Kauermann, L S Waltrup, et al. *On confidence intervals for semiparametric expectile regression*, Statistics and Computing, 2013, 23: 135-148.

[21] E Spiegel, F Sobotka, T Kneib. *Model selection in semiparametric expectile regression*, Electronic Journal of Statistics, 2017, 11(2): 3008-3038.

[22] C J Stone. *Additive regression and other nonparametric models*, The Annals of Statistics, 1985, 13(2): 689-705.

[23] P D Tao, L T H An. *Convex analysis approach to dc programming: Theory, algorithms and applications*, Acta Mathematica Vietnamica, 1997, 22(1): 289-355.

[24] R Tibshirani. *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society, Series B (Methodological), 1996, 58(1): 267-288.

[25] N Turan, M F Ghalwash, S Katari, et al. *DNA methylation differences at growth related genes correlate with birth weight: a molecular signature linked to developmental origins of adult disease?*, BMC Medical Genomics, 2012, 5: 1-21.

[26] A W van der Vaart, J A Wellner. *Weak Convergence and Empirical Processes: with applications to statistics*, Journal of the Royal Statistical Society−Series A Statistics in Society, 1997, 160(3): 596-608.

[27] H Votavova, M D Merkerova, K Fejglova, et al. *Transcriptome alterations in maternal and fetal cells induced by tobacco smoke*, Placenta, 2011, 32(10): 763-770.

[28] L S Waltrup, F Sobotka, T Kneib, et al. *Expectile and quantile regression−David and Goliath?*, Statistical Modelling, 2015, 15(5): 433-456.

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

[30] J Zhao, Y Chen, Y Zhang. *Expectile regression for analyzing heteroscedasticity in high dimension*, Statistics & Probability Letters, 2018, 137: 304-311.

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

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

[1]Institute of Digital Finance, Hangzhou City University, Hangzhou 310015, China.
[2]Department of Statistics, Hangzhou City University, Hangzhou 310015, China.
[3]Department of Statistics, University of California, Los Angeles, CA 90095, USA.
[4]School of Mathematical Sciences, Zhejiang University, Hangzhou 310027, China.
Email: zhangyi63@zju.edu.cn