Semiparametric Model Averaging for Ultrahigh-Dimensional Conditional Quantile Prediction

Chao Hui GUO
School of Mathematical Sciences, Chongqing Normal University, Chongqing 401331, P. R. China
E-mail: guochaohu2010@126.com

Jing LV
School of Mathematics and Statistics, Southwest University, Chongqing 400715, P. R. China
E-mail: cqulv2015@126.com

Hu YANG Jing Wen TU
College of Mathematics and Statistics, Chongqing Normal University, Chongqing 401331, P. R. China
E-mail: yh@cqu.edu.cn 458711089@qq.com

Chen Xiao TIAN
College of Finance, University of International Business and Economics, Beijing 100020, P. R. China
E-mail: 202110220551@uibe.edu.cn

Abstract In this paper, we develop a flexible semiparametric model averaging marginal regression procedure to forecast the joint conditional quantile function of the response variable for ultrahigh-dimensional data. First, we approximate the joint conditional quantile function by a weighted average of one-dimensional marginal conditional quantile functions that have varying coefficient structures. Then, a local linear regression technique is employed to derive the consistent estimates of marginal conditional quantile functions. Second, based on estimated marginal conditional quantile functions, we estimate and select the significant model weights involved in the approximation by a nonconvex penalized quantile regression. Under some relaxed conditions, we establish the asymptotic properties for the nonparametric kernel estimators and oracle estimators of the model averaging weights. We further derive the oracle property for the proposed nonconvex penalized model averaging procedure. Finally, simulation studies and a real data analysis are conducted to illustrate the merits of our proposed model averaging method.

Keywords Kernel regression, model averaging, oracle property, penalized quantile regression, ultrahigh-dimensional data

MR(2010) Subject Classification 62G05, 62G20
1 Introduction

With the rapid development of scientific techniques, ultrahigh-dimensional data is frequently collected in many research areas such as genomics, tomography, economics and finance. The so-called ultrahigh-dimensional data represents that the number of predictors $p_n$ grows at an exponential rate of the sample size $n$, namely, $\log(p_n) = O(n^\rho)$ for $\rho > 0$. Statisticians may encounter many challenges for the analysis of ultrahigh-dimensional data, owing to a large number of variables and uncertain functional forms, and thus need to develop new statistical theories and methodologies. As a fundamental statistical modeling tool for high dimensional data, variable selection plays an important role in improving prediction accuracy and has gained much attention in the last two decades, including [9, 42, 48, 49, 53]. Recently, [23] studied the SCAD estimation in ultrahigh-dimensional linear regression models and proved that their proposed estimator still has the oracle property under the normal error distribution. [44] considered the nonconvex penalized quantile regression for ultrahigh-dimensional linear models and established its oracle property. Later, [39] extended their method to partially linear additive models with ultrahigh-dimensional linear covariates. Other related research results can refer to [3, 7, 10, 11, 17] and references therein.

Prediction is important, but the above regularization approaches select models based on some common criteria and ignore useful information, which will affect the prediction accuracy. Instead of selecting significant predictors to obtain a best model, the model averaging approach takes the weighted average of these estimators obtained from a set of potential candidate models. Thus, the main benefit of model averaging is that it effectively incorporates useful information from different candidate models and substantially reduces the risk of misspecification relative to the model selection. Therefore, model averaging effectively deals with the model uncertainty and makes more accuracy prediction. Due to these advantages, model averaging has gained considerable attention. See, [15] derived the Mallows model average estimate which can achieve the lowest possible squared error. [31] estimated the optimal weights for the parametric model averaging by minimizing the trace of an unbiased estimator of the model average estimator’s MSE. Recently, [1] selected the model averaging weights by the leave-subject-out cross-validation for ultrahigh-dimensional linear models. [2] introduced the model averaging for ultrahigh-dimensional generalized linear models. [32] developed a random splitting model averaging approach which can effectively avoid model overfitting and achieve stable prediction for ultrahigh-dimensional linear models.

Thus far, previous authors aim to study model averaging on a set of parametric models such as linear regression models. In practice, the response variable may rely on the covariate in a very complicated manner, and thus traditional parametric models may lead to biased predictions. As one of the most commonly used semiparametric models, varying coefficient models ([16]) not only can avoid the well-known “curse of dimensionality” but also depict dynamic behaviors between the response and predictors. Much works had been done on high-dimensional varying coefficient models, see, [6, 12, 36, 40, 47]. Recently, [27] developed a flexible varying coefficient model averaging procedure based on the Mallows criterion and derived asymptotic optimality for the proposed estimator under some mild conditions. Later, [29] introduced a new varying coefficient semiparametric model averaging prediction by considering only one varying
coefficient and taking the other variables as constant coefficient in each candidate sub-model. Based on the Mallows-type criterion, [52] studied the varying coefficient partially linear model averaging.

To the best of our knowledge, researchers focus on the mean regression for semiparametric model averaging, which may be somewhat limiting. In many situations of practical interest, we care about the conditional quantile of the interest variable rather than its conditional mean. For example, many economists focus on the high quantile of oil prices, because the oil price fluctuation may make significantly effects on economic activity, such as automobile production. Compared with the mean regression, quantile regression ([25]) provides us with a more complete description about the conditional distribution of the response variable. Another advantage of quantile regression is that it requires weaker conditions for the random error and is robust to outlier contamination and heavy tailed errors. Existing works on high dimensional quantile varying coefficient models have been largely limited to the estimation and variable selection, see [36, 40, 50]. However, an important practical interest is to predict the conditional quantile of the response variable in ultrahigh-dimensional data.

In this paper, without any specification of the error distribution, we focus on forecasting the joint conditional quantile of the response variable for ultrahigh-dimensional data by using some penalization devices and model averaging strategies. Compared with some existing works, this paper has several innovations as follows. First, we consider the problem of ultrahigh-dimensional conditional quantile prediction, which may face more difficulties and challenges to establish its asymptotic theory. Second, we do not need to specify the model structure of the joint conditional quantile function and only assume that approximated marginal conditional quantile functions have varying coefficient structures. This ensures that our method is more flexible than traditional varying coefficient models. Finally, each candidate submodel only involves two one-dimensional nonparametric functions which can be estimated by the popular local linear smoothing. At the same time, based on nonzero model weights, we combine useful marginal submodels in an effective manner to obtain a more accurate prediction for the joint conditional quantile function.

The rest of the article is organized as follows. In subsection 2.1, we approximate the joint conditional quantile function by an affine combination of marginal conditional quantile varying coefficient functions. Asymptotic properties of the estimated marginal varying coefficient functions and model weights obtained from the oracle model are provided in subsection 2.2. In Section 3, we propose an ultrahigh-dimensional model averaging procedure and derive its oracle property. In Section 4, we conduct simulation examples to illustrate the merit of the proposed method. In Section 5, a Boston housing data is analyzed to further illustrate our proposed methodology. Conclusions are given in Section 6. The technical proofs of theoretical results are presented in the Appendix.

2 Semiparametric Model Averaging Procedure with Diverging Number of Parameters

Let $Y$ be a response variable and $Z \triangleq (X^T, U)^T$ be a $(p_n + 1)$-dimensional predictor vector, where $X = (X_1, \ldots, X_{p_n})^T \in \mathbb{R}^{p_n}$ is a $p_n$-dimensional covariate and $U$ is an univariate observ-
able exposure variable. At a given quantile level \( \tau \in (0, 1) \), we consider the following quantile regression model

\[
m_\tau(Z) = Q_\tau(Y|Z),
\]

where \( Q_\tau(Y|Z) \) is the conditional quantile of \( Y \) given \( Z \). If the dimension of \( X \) is large, it is challenging to estimate \( m_\tau(Z) \). Although the model selection technique helps us to choose a single optimal model, it may ignore the information from the other models which are not selected, often resulting in a poor predictive performance. As an alternative to model selection, model averaging fits a set of candidate models and then weights them by some criterion. In this paper, we will combine the marginal varying coefficient models and SCAD penalty function to develop a new semiparametric model averaging procedure to predict \( m_\tau(Z) \) for ultrahigh-dimensional covariates.

2.1 Marginal Varying Coefficient Quantile Regression Models

[28] firstly proposed to approximate the multivariate regression function by a weighted average of marginal regression functions, that is

\[
E(Y|X) \approx w_0 + \sum_{j=1}^{p_n} w_j E(Y|X_j).
\]

Note that each marginal regression \( E(Y|X_j) \) can be treated as a “nonlinear candidate model” and \( w_j, j = 1, \ldots, p_n \) can be seen as the weights for the \( j \)th candidate model. The main merit of the above linear combination is that estimating \( E(Y|X) \), can be converted into estimating the weight \( w_j \) and the univariate function \( E(Y|X_j) \), which can effectively avoid the curse of dimensionality problem. \[4, 5\] extended the work of [28] to the dynamic portfolio choice and the ultrahigh-dimensional case, respectively. The main focus of this article is on the conditional quantile function \( m_\tau(Z) \). Motivated by [28], we approximate \( m_\tau(Z) \) by

\[
m_\tau(Z) \approx m_{\tau}^{\omega_n}(Z) \overset{\Delta}{=} w_0 + \sum_{j=1}^{p_n} w_j Q_\tau(Y|X_j,U),
\]

where \( Q_\tau(Y|X_j,U) \) is the marginal conditional quantile function of \( Y \) given \((X_j,U)\). The intercept \( w_0 \) (corresponding to a null model) can be regarded as the unconditional quantile of the response \( Q_\tau(Y) \). Here, we assume that the marginal conditional quantile function \( Q_\tau(Y|X_j,U) \) has a varying coefficient structure. That is, \( Q_\tau(Y|X_j,U) = \alpha_{0j\tau}(U) + \alpha_{j\tau}(U)X_j \). As (2.2) involves nonparametric components, we propose a two stage estimation procedure. In the first stage, we employ a local linear regression technique to derive the estimates of \( \alpha_{0j\tau}(\cdot) \) and \( \alpha_{j\tau}(\cdot) \) and obtain the estimators \( \hat{Q}_\tau(Y|X_j,U) \) of \( Q_\tau(Y|X_j,U) \), for \( j = 1, \ldots, p_n \). In the second stage, we estimate the model average weight vector \( \omega_n = (w_0, w_1, \ldots, w_{p_n})^T \) based on the estimated \( \hat{Q}_\tau(Y|X_j,U) \), for \( j = 1, \ldots, p_n \). Throughout this paper, for any differential function \( \hat{f}(x) \), we use \( \hat{f}(x) \) and \( \hat{f}(x) \) to denote its first-order and second-order derivative. In the rest of the study, we will drop the subscript \( \tau \) in \( \alpha_{0j\tau}(\cdot) \) and \( \alpha_{j\tau}(\cdot) \), \( j = 1, \ldots, p_n \) for simplicity.

Firstly, for a fixed \( j \) and \( U \) in the neighborhood of \( u \), taking Taylor’s expansion for \( \alpha_{0j}(U) \) and \( \alpha_{j}(U) \) at \( u \), we have

\[
\alpha_{0j}(U) \approx \alpha_{0j}(u) + \dot{\alpha}_{0j}(u)(U-u) \overset{\Delta}{=} a_j + b_j(U-u),
\]

and

\[
\alpha_{j}(U) \approx \alpha_{j}(u) + \dot{\alpha}_{j}(u)(U-u) \overset{\Delta}{=} c_j + d_j(U-u).
\]
Let \( \{\hat{a}_j, \hat{b}_j, \hat{c}_j, \hat{d}_j\} \) be the minimizer of the following local weighted quantile loss function

\[
\sum_{i=1}^{n} \rho_{\tau}(Y_i - a_j - b_j(U_i - u) - X_{ij}[c_j + d_j(U_i - u)]) K_h(U_i - u),
\]

where \( \rho_{\tau}(u) = \tau u - uI(u < 0) \) is check loss function for \( \tau \in (0, 1) \), \( K_h(\cdot) = K(\cdot/h) \) with a bandwidth \( h \) and \( K(\cdot) \) is a given kernel function. Then, we have \( \hat{a}_{0j}(u) = \hat{a}_j \), \( \hat{b}_j(u) = \hat{c}_j \) and \( \hat{Q}_\tau(Y_i|X_{ij}, U_i) = \hat{a}_{0j}(U_i) + \hat{a}_j(U_i)X_{ij} \) for \( i = 1, \ldots, n, j = 1, \ldots, p_n \).

Following the convention in variable selection, we assume that a small number of \( Q_\tau(Y|X_j, U) \) have effects on predicting \( m_\tau(Z) \). In other words, we assume that the model average weight vector is sparse, indicating that most of components are zero for ultrahigh-dimensional setting. Without loss of generality, we denote the optimal weights as \( \omega_{n0} = (\omega_{n0}^{(1)}, \omega_{m0}^{(2)})^T \), where \( \omega_{n0}^{(1)} = (\omega_{00}, \omega_{01}, \ldots, \omega_{0q_n})^T \) stands for nonzero weights with dimension \( (q_n + 1) \) and \( \omega_{n0}^{(2)} = (\omega_{0q_n+1}, \ldots, \omega_{0p_n})^T \) is zero weights with dimension \( (p_n - q_n) \). Before stating the theoretical results of the nonparametric estimates, we need to introduce some notations. Let \( \omega_{n0}^{(1)} = \tau \omega_{n0}^{(2)} \) be the optimal weights of the values in the semiparametric model averaging. In this paper, one-dimensional marginal regression functions are taken as the candidate models and the covariates are non overlapping, indicating that the optimal model weight vector \( \omega_{n0}^{(1)} \) is unique (see, [28]). Let \( \mathcal{A}^* = \{1 \leq j \leq p_n; \omega_{0j} \neq 0\} \) be the index set of nonzero coefficients and \( q_n = |\mathcal{A}^*| \) be the cardinality of \( \mathcal{A}^* \). The set \( \mathcal{A}^* \) is unknown and will be estimated. Define \( \xi_{ij} = Y_i - Q_\tau(Y_i|X_{ij}, U_i), \varepsilon_i = Y_i - \omega_{n0} - \sum_{j=1}^{q_n} \omega_{0j}Q_\tau(Y_i|X_{ij}, U_i) \), we have \( P(\xi_{ij} \leq 0 | X_{ij}, U_i) = \tau \) and \( P(\varepsilon_i \leq 0 | X_{ij}, U_i) = \tau \).

Let \( f_{ij}(\cdot|X_{ij}, U_i) \) and \( F_{ij}(\cdot|X_{ij}, U_i) \) be the marginal probability density and distribution functions of \( \xi_{ij} \) conditional on \( (X_{ij}, U_i) \), and \( f_i(\cdot|X_i, U_i) \) and \( F_i(\cdot|X_i, U_i) \) be the probability density and distribution functions of \( \varepsilon_i \) conditional on \( (X_i, U_i) \). We use the notations \( B_j(u) = E\{(1, X_j)^T(1, X_j) | U = u\}, D_j(u) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} E\{f_{ij}(0|X_{ij}, U_i)(1, X_{ij})^T(1, X_{ij}) | U_i = u\}, \mu_k = \int u^k K(u)du \) and \( \nu_k = \int u^k K^2(u)du \) for \( k = 0, 1, \ldots \).

In order to establish the asymptotic theory of the nonparametric kernel estimates, we introduce some regularity conditions.

\begin{itemize}
  \item[(C1)] There exists positive constant \( M_1 \) such that \( |X_{ij}| \leq M_1, \forall 1 \leq i \leq n, 1 \leq j \leq p_n \).
  \item[(C2)] The random variable \( U \) has bounded support \( \Omega \) and its marginal density function \( f_U(\cdot) \) is positive and has a continuous second derivative.
  \item[(C3)] The varying coefficients functions \( \alpha_{0j}(u) \) and \( \alpha_j(u) \) have continuous second derivatives for \( u \in \Omega \).
  \item[(C4)] We assume that the density functions \( f_{ij}(\cdot|X_{ij}, U_i) \) and \( f_i(\cdot|X_i, U_i) \) are uniformly bounded away from zero and infinity in a neighborhood of zero, and their first derivatives have uniform upper bound in a neighborhood of zero, for \( 1 \leq i \leq n, j = 1, \ldots, p_n \).
  \item[(C5)] The kernel function \( K(\cdot) \) is a Lipschitz continuous, symmetric and bounded probability density function with a compact support.
\end{itemize}

Condition (C1) assumes that the absolute values of the predictors are bounded, which is commonly assumed in high-dimensional analysis; see [26, 35, 39, 44]. This assumption can be relaxed to the moment condition given in [51] that there exists a positive constant \( t_0 \) such that \( \max_{1 \leq j \leq p_n} E\{\exp(tX_j)\} < \infty \) for \( 0 < t \leq t_0 \). This condition holds for a variety of distributions, including the normal distribution and the distributions with bounded support. In this case, our
Theoretical results still hold with some modification to the proofs. But, to mitigate notational complexity and facilitate mathematical derivations, we assume that covariates are bounded in this study. Conditions (C2) and (C3) make some smoothness restrictions on the marginal density function and marginal regression functions, which are similar to the conditions (C1) and (C2) in [22]. Condition (C4) is similar to the condition on the error distribution for quantile regression (see, [25]). Condition (C5) is a commonly used condition on the kernel function.

**Theorem 2.1** Under the regularity conditions (C1)-(C5), if \( h \to 0 \) and \( nh \to \infty \) as \( n \to \infty \), for a fixed \( j, j = 1, \ldots, p_n \). Then, we have

\[
\sqrt{nh} \left[ \left( \hat{\alpha}_0j(u) - \alpha_0j(u) \right) / \hat{\sigma}_j(u) - \alpha_j(u) \right] - \frac{1}{2} \mu_2h^2 \left( \hat{\sigma}_j(u) \right) \to N \left( 0, \frac{\nu_0}{f_U(u)} D_j^{-1}(u) B_j(u) D_j^{-1}(u) \right)
\]

2.2 Oracle Estimator and Asymptotic Property

We first study the oracle estimator of model averaging weights. This means that we know that the first \((q_n + 1)\) elements of \( \omega_{n0} \) are nonzero and the last \((p_n - q_n)\) elements of \( \omega_{n0} \) are all zero in advance. Furthermore, we assume that \( q_n \) diverges with the sample size and let

\[
\hat{\omega}^{(1)}_n = \arg \min_{\omega^{(1)}_n} \sum_{i=1}^n \rho_r \left\{ Y_i - \omega_0 - \sum_{j=1}^{q_n} \omega_j \hat{Q}_r(Y_i | X_{ij}, U_i) \right\}.
\]

Then, the oracle estimator of \( \omega_{n0} \) is \( \hat{\omega}_n = (\hat{\omega}^{(1)}_n)^T, 0_{T_{p_n - q_n}}^T \).

Before presenting the asymptotic property of the oracle estimator, some additional notations need to be introduced. Define \( \hat{\mathcal{M}}_i = (\hat{\mathcal{M}}_{i0}, \ldots, \hat{\mathcal{M}}_{ip_n})^T \) with \( \hat{\mathcal{M}}_{ij} = 1 \) for \( j = 0 \) and \( \hat{\mathcal{M}}_{ij} = \hat{Q}_r(Y_i | X_{ij}, U_i) \) for \( j = 1, \ldots, p_n \). Similarly, we can define \( \hat{\mathcal{M}}_i = (\hat{\mathcal{M}}_{i0}, \ldots, \hat{\mathcal{M}}_{ip_n})^T \) with \( \hat{\mathcal{M}}_{ij} = 1 \) for \( j = 0 \) and \( \hat{\mathcal{M}}_{ij} = Q_r(Y_i | X_{ij}, U_i) \) for \( j = 1, \ldots, p_n \). Let \( \psi_r(\xi_{ij}) = \tau - I(\xi_{ij} \leq 0) \) and \( \psi_r(\varepsilon_i) = \tau - I(\varepsilon_i \leq 0) \), where \( \xi_{ij} \) and \( \varepsilon_i \) are given in Subsection 2.1. For \( 0 \leq j \leq q_n, 1 \leq s \leq n \), let \( \hat{\phi}_{s} = \hat{\mathcal{M}}_{ij} \psi_r(\varepsilon_s), \phi_{s} = \sum_{k=1}^{q_n} \omega_{0k} \psi_r(\xi_{sk})E(\hat{f}_s(0 | X_s, U_s) | \hat{\mathcal{M}}_{sk} | U_s)D_k^{-1}(U_s)(1, X_{sk})^T U_s) \) and \( \hat{\phi}_s = \phi_s - \hat{\phi}_s = (\hat{\phi}_{s0} - \phi_{s0}, \hat{\phi}_{s1} - \phi_{s1}, \ldots, \hat{\phi}_{sq_n} - \phi_{sq_n})^T \). Define \( S_n = \text{Cov}(n^{-1/2} \sum_{s=1}^{n} (\hat{\phi}_s - \phi_s)), K_n = n^{-1} \sum_{s=1}^{n} f_s(0 | X_s, U_s) \hat{\mathcal{M}}_{iA} M_{iA}^T \) and \( \Sigma_n = K_n^{-1} S_n K_n^{-1} \). Then, \( \hat{\mathcal{M}}_i \) is a subvector of \( \hat{\mathcal{M}}_i \) consisting of the first \((q_n + 1)\) elements. Let \( \rho_{\min}(\mathbf{A}) \) and \( \rho_{\max}(\mathbf{A}) \) be the smallest eigenvalue and largest eigenvalue of a symmetric square matrix \( \mathbf{A} \), respectively.

We now provide theoretical justifications for the oracle estimator \( \hat{\omega}^{(1)}_n \) as \( q_n \) diverges. In addition to conditions (C1)-(C5), we need additional two conditions.

(C6) Let \( \Lambda_n = E(\hat{\mathcal{M}}_{iA} \hat{M}_{iA}^T) \). There exist two positive constants \( M_2 \) and \( M_3 \) such that

\[
0 < M_2 \leq \rho_{\min}(\Lambda_n) \leq \rho_{\max}(\Lambda_n) \leq M_3 < \infty.
\]

(C7) The dimension \( q_n \) satisfies \( q_n = O(n^{c_1}) \) for some \( 0 \leq c_1 < 1/2 \), and \( n^{1/2}(h^2 + 1/\sqrt{nh}) = o(1) \).

Condition (C6) is popularly adopted in high-dimensional regression, which is similar to the condition (A6) in [5]. Condition (C7) implies that undersmoothing is needed in our semiparametric procedure and \( q_n \) can only be divergent at a polynomial rate of \( n \), which is similar to the condition (A7) in [5]. Theorem 2.2 shows that \( \hat{\omega}^{(1)}_n \) is \( \sqrt{n/q_n} \) consistent and asymptotically normal under some mild regularity conditions.
Theorem 2.2 Suppose that conditions (C1)–(C7) are satisfied. Then for $\forall e_n \in \mathbb{R}^{q_n+1}$ such that $\|e_n\| = 1$, the oracle estimator $\hat{\omega}_n = (\hat{\omega}_n^{(1)T}, 0^T)^T$ has

(i) $\|\hat{\omega}_n^{(1)} - \omega_n^{(1)}\| = O_p(\sqrt{q_n/n})$;

(ii) $\sqrt{n}e_n^T \Sigma_n^{-1/2}(\hat{\omega}_n^{(1)} - \omega_n^{(1)}) \xrightarrow{d} N(0, 1)$.

3 Nonconvex Penalized Quantile Model Averaging Estimation

3.1 Nonconvex Penalized Estimation

Variable selection is a crucial step in high-dimensional setting, and various penalization methods have been proved to be powerful for performing variable selection. Let $\tilde{\omega}_n = (\tilde{\omega}_1, \ldots, \tilde{\omega}_{p_n})^T$ be the estimator of $\omega_n$ that minimizes the following penalized quantile loss function

$$Q_n(\omega_n) = n^{-1} \sum_{i=1}^n \rho_r(Y_i - \hat{\Lambda}_i^T \omega_n) + \sum_{j=1}^{p_n} p_\lambda(|\omega_j|),$$

(3.1)

where $p_\lambda(\cdot)$ is a commonly used SCAD penalty function proposed by [9] and its first order derivative is defined by

$$p_\lambda(x) = \lambda \left\{ I(x \leq \lambda) + \frac{(a\lambda - x)}{(a-1)\lambda} I(x > \lambda) \right\},$$

where $a > 2$, $\lambda$ is a nonnegative tuning parameter and $p_\lambda(0) = 0$. This is a complicated nonlinear programming problem, and we use the “cv.rq.pen” function in the package “rqPen” to find the minimum of (3.1) using numerical computing methods. Finally, for a future observation $z \triangleq (x^T, u)$ with $x = (x_1, \ldots, x_{p_n})^T$, we can predict $m_r(z)$ by $\bar{m}_r(z) = \bar{\omega}_0 + \sum_{j=1}^{p_n} \bar{\omega}_j \hat{Q}_r(Y|x_j, u)$.

3.2 Asymptotic Theory

Due to the nonsmoothness and nonconvexity of the penalized objective function $Q_n(\omega_n)$, we use the idea of [39, 41, 44] to investigate the asymptotic theory of the nonconvex estimator $\tilde{\omega}_n$. We first write the nonconvex objective function $Q_n(\omega_n)$ as the difference of two convex functions:

$$Q_n(\omega_n) = g(\omega_n) - h(\omega_n),$$

where $g(\omega_n) = n^{-1} \sum_{i=1}^n \rho_r(Y_i - \hat{\Lambda}_i^T \omega_n) + \lambda \sum_{j=1}^{p_n} |\omega_j|$ and $h(\omega_n) = \sum_{j=1}^{p_n} H_\lambda(\omega_j)$, where

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

Suppose that $\text{dom}(g) = \{\eta : g(\eta) < \infty\}$ is the effective domain of $g$ and

$$\partial g(\eta_0) = \{t : g(\eta) \geq g(\eta_0) + (\eta - \eta_0)^T t, \forall \eta\}$$

is the subdifferential of a convex function $g(\eta)$ at $\eta = \eta_0$. An optimality condition for $\omega_n^*$ to be a local minimizer of $g(\omega_n) - h(\omega_n)$ is that $\omega_n^*$ has neighborhood $\Gamma$ such that $\partial h(\omega_n) \cap \partial g(\omega_n^*) \neq \emptyset, \forall \omega_n \in \Gamma \cap \text{dom}(g)$.

To study the optimality condition for the convex differencing problem, we formally define $\partial g(\omega_n)$ and $\partial h(\omega_n)$, the subdifferential of $g(\omega_n)$ and $h(\omega_n)$, respectively. First, the subdifferential of $g(\omega_n)$ at $\omega_n$ is defined as

$$\partial g(\omega_n) = \left\{ \xi = (\xi_0, \xi_1, \ldots, \xi_{p_n})^T \in \mathbb{R}^{p_n+1} : \xi_j = -\tau n^{-1} \sum_{i=1}^n \hat{\Lambda}_{ij} I(Y_i - \hat{\Lambda}_i^T \omega_n > 0) \right\}.$$
Furthermore, we consider the subgradient of the unpenalized objective function, which is given by

\[\frac{\partial h}{\partial \omega_n} = \left\{ \zeta = (\zeta_0, \zeta_1, \ldots, \zeta_p)^T \in \mathbb{R}^{p+1} : \zeta_j = \frac{\partial h(\omega_n)}{\partial \omega_j} \right\},\]

where \(\frac{\partial h(\omega_n)}{\partial \omega_j} = 0\) for \(j = 0\). For \(1 \leq j \leq p_n\),

\[\frac{\partial h(\omega_n)}{\partial \omega_j} = \begin{cases} 0, & 0 \leq |\omega_j| < \lambda, \\ (\omega_j - \lambda \text{sgn}(\omega_j)) / (a - 1), & \lambda \leq |\omega_j| \leq a\lambda, \\ \lambda \text{sgn}(\omega_j), & |\omega_j| > a\lambda. \end{cases}\]

Next, we consider the subgradient of the unpenalized objective function, which is given by

\[s(\omega_n) = (s_0(\omega_n), s_1(\omega_n), \ldots, s_{p_n}(\omega_n))^T,\]

where

\[s_j(\omega_n) = -\tau n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} I(Y_i - \hat{M}_i^T \omega_n > 0) + (1 - \tau) n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} I(Y_i - \hat{M}_i^T \omega_n < 0) - n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} v_i,\]

for \(j = 0, 1, \ldots, p_n\), we have \(v_i = 0\) if \(Y_i - \hat{M}_i^T \omega_n \neq 0\) and \(v_i \in [\tau - 1, \tau]\) otherwise.

To obtain the theoretical results, we next present an additional condition on the smallest of optimal nonzero weights.

(C8) There exist positive constants \(c_2\) and \(M_4\) such that \(n^{(1-c_2)/2} \min_{1 \leq j \leq q_n} |\omega_{0j}| \geq M_4\) for \(2c_1 < c_2 < 1\), where \(c_1\) is given in the condition (C7).

Condition (C8) requires that the smallest of optimal nonzero weights should not decay too fast, which is similar to the condition (C5) in [44] and condition 6 in [39]. Proposition 3.1 states the properties of the oracle estimator \(\hat{\omega}_n = (\hat{\omega}_n^{(1)}T, 0^T)^T\) and \(s_j(\hat{\omega}_n)^T\),

**Proposition 3.1** Assume conditions (C1)–(C8) hold and \(\lambda = o(n^{-(1-c_2)/2}), q_n n^{-1/2} = o(\lambda), \log(p_n) = o(n \lambda^2)\) and \(n \lambda^2 \to \infty\). For the oracle estimator \(\hat{\omega}_n = (\hat{\omega}_n^{(1)}T, 0^T)^T\), there exist \(v_i^*\) with \(v_i^* = 0\) if \(Y_i - \hat{M}_i^T \omega_n \neq 0\) and \(v_i^* \in [\tau - 1, \tau]\) otherwise, such that for \(s_j(\hat{\omega}_n)\) with \(v_i = v_i^*\), with probability approaching one, we have

1. \(s_j(\hat{\omega}_n) = 0, j = 0, 1, \ldots, q_n\),
2. \(|\hat{\omega}_j| \geq (a + 1/2) \lambda, j = 1, \ldots, q_n\),
3. \(|s_j(\hat{\omega}_n)| \leq \lambda, j = q_n + 1, \ldots, p_n\).

Based on Proposition 3.1, we can prove (A.37) in the Appendix. Then, we can demonstrate that the oracle estimator \(\hat{\omega}_n = (\hat{\omega}_n^{(1)}T, 0^T)^T\) belongs to the set of local minimizer of \(Q_n(\omega_n)\). This leads to the following theorem, which is our main result.

\[+(1 - \tau) n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} I(Y_i - \hat{M}_i^T \omega_n < 0) - n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} v_i + \lambda_j \} , \]

where \(v_i = 0\) if \(Y_i - \hat{M}_i^T \omega_n \neq 0\) and \(v_i \in [\tau - 1, \tau]\) otherwise. \(l_0 = 0\) and for \(1 \leq j \leq p_n\), we have \(l_j = \text{sgn}(\omega_j)\) if \(\omega_j \neq 0\) and \(l_j \in [-1, 1]\) otherwise, where \(\text{sgn}(t) = I(t > 0) - I(t < 0)\). Furthermore, \(h(\omega_n)\) is differential everywhere. Hence, its subdifferential is simply the regular derivative. For any value of \(\omega_n\), we have

\[\frac{\partial h(\omega_n)}{\partial \omega_n} = \begin{cases} 0, & 0 \leq |\omega_j| < \lambda, \\ (\omega_j - \lambda \text{sgn}(\omega_j)) / (a - 1), & \lambda \leq |\omega_j| \leq a\lambda, \\ \lambda \text{sgn}(\omega_j), & |\omega_j| > a\lambda. \end{cases}\]
Theorem 3.2 Let $\mathcal{B}_n(\lambda)$ be the set of local minimizer of $\mathcal{Q}_n(\omega_n)$. Assume that conditions (C1)–(C8) hold and $\lambda = o(n^{-(1-c_2)/2})$, $q_n n^{-1/2} = o(\lambda)$, $\log(p_n) = o(n\lambda^2)$ and $n\lambda^2 \to \infty$. Then, the oracle estimator $\hat{\omega}_n = (\hat{\omega}_n^{(1)T}, 0^T)^T$ satisfies
\[
P(\hat{\omega}_n \in \mathcal{B}_n(\lambda)) \to 1, \quad \text{as } n \to \infty.
\]

Remark 3.3 As far as we know, various screening procedures (e.g., [5, 11, 12, 18, 30, 46, 51]) had been certified to effective tools to select the significant variables in the past decade. However, the results from the screening procedure are rather crude and often include some inactive variables. Moreover, it is intractable to determine the threshold parameter when carrying out an screening procedure. By contrast, this paper considers a one-step penalized model averaging approach to achieve accurate prediction for the condition quantile $m_{\tau}(Z)$. Furthermore, Theorem 3.2 indicates that the oracle estimator belongs to the set $\mathcal{B}_n(\lambda)$ with probability approaching one, as $n \to \infty$. Therefore, the oracle property for the sparse quantile model averaging holds in ultrahigh-dimensional data. How to develop a new screening approach remains an interesting question. However, fully addressing this question is very challenging, and beyond the scope of this study.

4 Simulation Studies

In this section, our main goal is to accomplish two tasks. On the one hand, simulation examples are conducted to verify the sparse and oracle properties of our proposed penalized estimation. On the other hand, compared with some existing ultrahigh-dimensional model selection and model averaging approaches, we aim to show the superiority of our proposed method in terms of the prediction accuracy.

4.1 Implementations

In practice, we need to choose several tuning parameters appropriately for achieving good numerical performances. First, we use Epanechnikov kernel when estimating the marginal conditional quantile functions, namely, $K(u) = 0.75(1-u^2)$. According to [22], we set the bandwidth as $h = h_{LS} \times \{\tau(1-\tau)/\phi^2(\Phi^{-1}(\tau))\}^{1/5}$, where $\phi$ (or $\Phi$) is the standard normal density (or distribution) function and $h_{LS}$ is the asymptotic optimal bandwidth for least squares estimator and can be selected by the rule of thumb. Second, the tuning parameter $\lambda$ controls the sparsity of $\bar{\omega}_n$ and can be selected by a data-driven method. Recently, [26] developed a modified BIC that is consistent in terms of variable selection for the high-dimensional linear quantile regression. Later, [39] applied it to the ultrahigh-dimensional partially linear additive quantile model. Motivated by these findings, we choose $\lambda$ by minimizing the following high dimensional BIC (HBIC) criterion
\[
\text{HBIC}(\lambda) = \log \left( \sum_{i=1}^{n} \rho_r \{ Y_i - \hat{M}_i^T \bar{\omega}_n \} \right) + df \frac{C_n \log(n)}{2n},
\]
where $C_n$ is a positive constant that diverges to infinity, $df$ is the number of nonzero coefficients in the estimated model weight vector $\bar{\omega}_n$ for a given $\lambda$. In simulation studies and real data analysis, we set $C_n$ to be $C_n = \log(p_n)$. Third, we compare our proposed ultrahigh-dimensional semiparametric model averaging conditional quantile prediction method (SMACQP) with the following popular model selection or model averaging methods:
(1) The penalized ultrahigh-dimensional quantile linear models ([44]), denoted by \( \text{qlm} \), which can be implemented by the R function “cv.rq.pen” in the package “rqPen”, where the tuning parameter \( \lambda \) is selected by the PBIC criterion, which can refer to the package “rqPen”.

(2) The ultrahigh-dimensional linear model averaging proposed by [1], denoted as \( \text{lmma} \). This method consists of two steps. First, we calculate the marginal correlation between co-variates and the response, and sort the correlation in order of absolute value from largest to smallest. Then, we choose the first \( M = \lceil n / \log(n) \rceil \) regressors, where \( \lceil a \rceil \) stands for the largest integer not greater than \( a \). Second, we consider \( J_n \) candidate sub-models and set the number of predictors \( p^* \) in each model to be the same. The delete-one cross-validation criterion is adopted to select the optimal \( J_n \) and \( p^* \) such that \( J_n p^* \leq M \). Finally, the model weights are also estimated by a delete-one cross-validation procedure.

(3) The quantile linear model averaging proposed by [33], denoted as \( \text{qlmma} \). Note that [33] only considers low-dimensional prediction problems. To deal with ultra-high dimension, the quantile-adaptive model-free feature screening procedure ([18]) is considered to screen out the most relevant \( \lceil 3n^{1/3} \rceil \) regressors. We use the R function “qaSIS” from the package “rqPen” to implement the screening procedure. Based on the selected regressors, the quantile linear model averaging is used to make predictions.

(4) The ultrahigh-dimensional semiparametric model averaging marginal regression proposed by [5], denoted by \( \text{sma} \). This method is a two-step forecasting approach. In the first step, the kernel sure independence screening technique is proposed to screen out \( \lceil 10n^{1/6} \rceil \) significant regressors. In the second step, the penalized semiparametric model averaging is developed to obtain sparse model weights, where the tuning parameter \( \lambda \) is selected by the delete-one cross-validation criterion.

(5) The ultrahigh-dimensional quantile varying coefficient models proposed by [46], denoted as \( \text{qvcm} \). This is also a two-step procedure. In the first step, the conditional quantile correlation sure independent screening is proposed to select \( \lceil n / \log(n) \rceil \) regressors based on B-spline approximation with the number of basis \( \lceil n^{1/5} \rceil + 1 \). Second, a group penalized variable selection method is used to identify important covariates and make predictions on the basis of MM algorithm ([20]). The tuning parameter is selected by the BIC criterion which can refer to section 4 of [46].

For the above approaches, we should realize the following three standpoints. First, \( \text{qlm} \) and \( \text{qvcm} \) are variable selection strategies, while \( \text{lmma} \), \( \text{qlmma} \), \( \text{sma} \) and SMACQP are model averaging procedures. Second, \( \text{lmma} \) and \( \text{sma} \) are used to forecast the conditional mean of the response variable but \( \text{qlm} \), \( \text{qlmma} \), \( \text{qvcm} \) and SMACQP can be utilized to predict the conditional quantile of the response variable.

4.2 Simulation Examples

**Example 4.1** We define the conditional quantile function \( Q_\tau(Y|X,U) \) by

\[
Q_\tau(Y|X,U) = \Phi^{-1}(\tau) + (\exp(-2U) + 0.5)X_1 + (2\tau \sin(0.5\pi U) + 0.5)X_2 + UX_3, \quad (4.2)
\]

where \( U \sim U(0,1) \), \( X = (X_1, \ldots, X_p) \) is from the multivariate normal distribution \( N(0, \Sigma) \) with \( \Sigma = (0.5^{j-k})_{j,k=1}^{p_n} \) and \( \Phi^{-1}(\tau) \) can be regarded as the intercept. According to (4.2), we have \( \omega_{0j} \neq 0 \) for \( j = 1, 2, 3 \) and \( \omega_{0j} = 0 \) otherwise. The coefficient \( 2\tau \sin(0.5\pi U) + 0.5 \) is a...
function of \( \tau \), suggesting that the covariate effects vary with the quantile level. Similarly to [14, 34], we generate \( Y_i \) as

\[
y_i = \Phi^{-1}(v_i) + (\exp(-2u_i) + 0.5)X_{i1} + (2v_i \sin(0.5\pi u_i) + 0.5)X_{i2} + U_i X_{i3},
\]

where \( V_i \) follows a uniform distribution \( U(0, 1) \).

**Example 4.2** As the reviewer’s suggestion, we consider a similar model set with [46]. Specifically, we generate the random samples from the following model

\[
Y = X_1\beta_1(U) + X_2\beta_2(U) + X_3\beta_3(U) + g(U)\epsilon,
\]

where \( \beta_1(u) = 3u, \beta_2(u) = 3\sqrt{|u|}, \beta_3(u) = \exp(u/2), g(u) = 0.2[\exp(u)/(1+\exp(u))+(u-1)^2+\sin(2\pi u)] \), indicating that the error has varying variance, the index variable \( U \) is drawn from a uniform distribution \( U(0, 2) \) and \( X = (X_1, \ldots, X_{p_n})^T \) comes from a multivariate normal distribution \( N(0, \Sigma) \) with \( \Sigma = (0.7^{j-k})_{j,k=1}^{p_n} \), which means that there is strong correlation among \( X \). In order to examine the robustness of the proposed procedure, we consider the following three different error distributions of \( \epsilon \): standard normal distribution \( (N(0,1)) \), \( t \)-distribution with three degrees of freedom \( (t_3) \), mixture normal distribution \( (MN) \) which is a mixture of \( N(0,1^2) \) and \( N(0,5^2) \) with the weights 95% and 5%, respectively.

Obviously, the conditional quantile function \( Q_\tau(Y|X, U) = X_1\beta_1(U) + X_2\beta_2(U) + X_3\beta_3(U) + g(U)Q_\tau(\epsilon) \), where \( Q_\tau(\epsilon) \) is the quantile function of \( \epsilon \). Note that \( Q_\tau(Y|X, U) \) is equal to the conditional mean function \( E(Y|X, U) \) for a fixed \( \tau = 0.5 \), because the median and mean of \( \epsilon \) are equal to zero under the three considered distributions. Therefore, it is fair to compare the conditional quantile regression prediction approaches (SMACQP, qlm, qlmma and qvcm) with the conditional mean regression prediction approaches (lmma and sma) under this setting. For \( \tau = 0.75 \), we only consider SMACQP, qlm, qlmma and qvcm due to \( Q_\tau(\epsilon) \neq 0 \). Based on the data generating mechanism of this example, we can easily understand that \( X_1, X_2 \) and \( X_3 \) have significant effects on the conditional quantile function \( Q_\tau(Y|X, U) \), indicating \( \omega_{0j} \neq 0 \) for \( j = 1, 2, 3 \) and \( \omega_{0j} = 0 \) otherwise.

**Example 4.3** In Examples 4.1 and 4.2, we generate data from the varying coefficient model structure, but our proposed model average method forecasts the response variable without structure assumption. Therefore, we generate the random samples from the following model

\[
Y = X_1\beta_1(U_1) + X_{11}\beta_{11}(U_2) + X_{21}\beta_{21}(U_2) + X_{31}\beta_{31}(U_1) + X_{41}\beta_{41} + 0.5\epsilon,
\]

where \( \beta_1(u) = 2u, \beta_{11}(u) = (u + 1)^2, \beta_{21}(u) = \exp(u), \beta_{31}(u) = 2\sqrt{|u|}, \beta_{41} = 2, U_1 \) and \( U_2 \) are independently drawn from \( N(0, 1) \) and \( U(0, 1) \), respectively. Other settings are the same as that in Example 4.2. We take \( U_1 \) as the index variable for the methods qvcm and SMACQP.

In all our simulation studies, we generate a training data set of sample size \( n_t \), to estimate unknown functions and model weights, then generate another 50 observations (a testing set) to calculate the out-of-sample quantile prediction error. To evaluate the performance of variable selection and prediction accuracy for our proposed method, we consider the following five criteria: **Size**: average number of selecting as the significant covariates from \( X_1, \ldots, X_{p_n} \); **IC**: average number of relevant covariates correctly included in the model; **IC**: average number of irrelevant covariates incorrectly included in the model; **P**: proportion of all relevant covariates...
correctly included in the model; **OQPE**: out-of-sample quantile prediction error, defined by \( \sum_{i \in I} \rho_\tau(Y_i - \hat{Q}_\tau(Y_i | Z_i))/|I| \), where \( I \) stands for an index set of the testing sample. We repeat the simulation 200 times and report the mean, median and standard deviations (SD) of **OQPE** in Tables 1–3.

| \( \tau \) | Method | Size   | C    | IC   | P     | OQPE               |
|-------|--------|--------|------|------|-------|--------------------|
|       |        | Mean   | Median | SD   |       |                    |
| 0.5   | qlm    | 2.895  | 2.895 | 0.000 | 0.895 | 0.4534 0.4487 0.0574 |
|       | lmma   |        |       |       |       | 0.4763 0.4743 0.0547 |
|       | qlmma  |        |       |       |       | 0.4568 0.4558 0.0532 |
|       | sma    | 22.420 | 3.000 | 19.420| 1.000 | 0.4963 0.4943 0.0588 |
|       | qvcm   | 3.035  | 2.990 | 0.045 | 0.990 | 0.4164 0.4101 0.1029 |
|       | SMACQP | 3.005  | 3.000 | 0.005 | 1.000 | 0.4272 0.4248 0.0551 |
| 0.75  | qlm    | 2.630  | 2.625 | 0.005 | 0.625 | 0.7523 0.7459 0.1055 |
|       | qlmma  |        |       |       |       | 0.7330 0.7314 0.0952 |
|       | qvcm   | 2.980  | 2.945 | 0.035 | 0.945 | 0.5354 0.5324 0.1155 |
|       | SMACQP | 2.975  | 2.975 | 0.000 | 0.975 | 0.6982 0.6939 0.1006 |

Table 1  Simulation results of various approaches in Example 4.1 with \( n_{tr} = 400 \) and \( p_n = 500 \)

| \( \tau \) | error | Method | Size   | C    | IC   | P     | OQPE               |
|-------|-------|--------|--------|------|------|-------|--------------------|
|       |       |        | Mean   | Median | SD   |       |                    |
| 0.5 N(0,1) | qlm    | 398.20 | 3     | 395.20 | 1   | 2.7706 2.7349 0.3961 |
|       | lmma   |        |       |       |       | 1.0212 1.0169 0.1467 |
|       | qlmma  |        |       |       |       | 0.9616 0.9531 0.1443 |
|       | sma    | 11.28  | 3     | 8.28  | 1   | 1.0156 1.0085 0.1630 |
|       | qvcm   | 3.00   | 3     | 0.00  | 1   | 0.1929 0.1702 0.0721 |
|       | SMACQP | 4.82   | 3     | 1.82  | 1   | 0.1784 0.1764 0.0233 |
| t\(_3\) | qlm    | 398.23 | 3     | 395.23| 1   | 2.7343 2.7708 0.3497 |
|       | lmma   |        |       |       |       | 1.0583 1.0297 0.1548 |
|       | qlmma  |        |       |       |       | 1.0036 0.9807 0.1574 |
|       | sma    | 11.88  | 3     | 8.88  | 1   | 1.0617 1.0367 0.1770 |
|       | qvcm   | 3.00   | 3     | 0.00  | 1   | 0.2237 0.2137 0.0760 |
|       | SMACQP | 4.09   | 3     | 1.09  | 1   | 0.2097 0.2081 0.0395 |
| MN    | qlm    | 398.20 | 3     | 395.20| 1   | 2.7319 2.7157 0.3604 |
|       | lmma   |        |       |       |       | 1.0622 1.0559 0.1611 |
|       | qlmma  |        |       |       |       | 1.0068 0.9937 0.1595 |
|       | sma    | 11.84  | 3     | 8.84  | 1   | 1.0673 1.0522 0.1769 |
|       | qvcm   | 3.00   | 3     | 0.00  | 1   | 0.2226 0.2045 0.0796 |
|       | SMACQP | 5.16   | 3     | 2.16  | 1   | 0.2031 0.2022 0.0468 |
| 0.75  | N(0,1) | qlm    | 3.19  | 2.97  | 0.22  | 0.97  | 1.4777 1.4587 0.2377 |
|       | qlmma  |        |       |       |       | 1.5013 1.4921 0.2294 |
|       | qvcm   | 3.27   | 2.93  | 0.34  | 0.96  | 0.2847 0.1971 0.4001 |
|       | SMACQP | 4.24   | 3.00  | 1.24  | 1.00  | 0.3146 0.3042 0.0774 |
| t\(_3\) | qlm    | 3.19   | 2.98  | 0.21  | 0.98  | 1.5249 1.5369 0.2341 |
|       | qlmma  |        |       |       |       | 1.5425 1.5267 0.2321 |
|       | qvcm   | 3.01   | 3.00  | 0.01  | 1.00  | 0.2757 0.2520 0.0968 |
|       | SMACQP | 4.22   | 3.00  | 1.22  | 1.00  | 0.3523 0.3398 0.0720 |
| MN    | qlm    | 3.14   | 2.94  | 0.20  | 0.94  | 1.5599 1.5483 0.2518 |
|       | qlmma  |        |       |       |       | 1.5717 1.5428 0.2561 |
|       | qvcm   | 3.00   | 3.00  | 0.00  | 1.00  | 0.2368 0.2235 0.0817 |
|       | SMACQP | 4.29   | 3.00  | 1.29  | 1.00  | 0.3132 0.3064 0.0661 |

Table 2  Simulation results of various approaches in Example 4.2 with \( n_{tr} = 400 \) and \( p_n = 500 \)
| \( \tau \) | error | Method | Size | C   | IC   | P     | OQPE   |
|-------|-------|--------|------|-----|------|-------|--------|
| 0.5   | N(0,1)| qlm    | 3.995| 3.985| 0.010| 0.000| 0.8854| 0.8849| 0.1312|
|       |       | lmma   | –    | –    | –    | –    | 0.9152| 0.9156| 0.1270|
|       |       | qlmma  | –    | –    | –    | –    | 1.6083| 1.6167| 0.1832|
|       |       | sma    | 7.740| 4.015| 3.725| 0.015| 0.9375| 0.9364| 0.1291|
|       |       | qvcm   | 16.830| 5.000| 11.830| 1.000| 0.7600| 0.6396| 0.3561|
|       |       | SMACQP | 6.165| 4.870| 1.295| 0.955| 0.6992| 0.6356| 0.2348|
| \( t_3 \) | qlm    | 3.995| 3.960| 0.035| 0.00 | 0.9112| 0.9026| 0.1267|
|       | lmma   | –    | –    | –    | –    | 0.9367| 0.9336| 0.1233|
|       | qlmma  | –    | –    | –    | –    | 1.6114| 1.6058| 0.2528|
|       | sma    | 7.965| 4.030| 3.935| 0.03 | 0.9592| 0.9596| 0.1247|
|       | qvcm   | 27.370| 5.000| 22.370| 1.00 | 1.0031| 0.8633| 0.4552|
|       | SMACQP | 5.985| 4.725| 1.260| 0.88 | 0.8066| 0.7076| 0.2957|
| MN    | qlm    | 4.000| 3.987| 0.013| 0.00 | 0.9056| 0.9129| 0.1333|
|       | lmma   | –    | –    | –    | –    | 0.9343| 0.9278| 0.1362|
|       | qlmma  | –    | –    | –    | –    | 1.5881| 1.6081| 0.1847|
|       | sma    | 8.747| 4.040| 4.707| 0.04 | 0.9684| 0.9514| 0.1463|
|       | qvcm   | 20.413| 5.000| 15.413| 1.00 | 0.8798| 0.7089| 0.4266|
|       | SMACQP | 5.787| 4.653| 1.133| 0.88 | 0.7835| 0.6842| 0.3519|

Table 3  Simulation results of various approaches in Example 4.3 with \( n_{tr} = 400 \) and \( p_n = 500 \)

From these tables, we may safely arrive at the following conclusions. First, we explain why variable selection results of lmma and qlmma are not listed, because these two methods utilize the screening technique to achieve dimension reduction and predict the response variable based on the screened out regressors. Second, SMACQP performs well in terms of variable selection performance, because it has large \( P \) and the values in the column labeled \( C \) are close to the true number of nonzero weights 3 (Examples 4.1 and 4.2) and 4 (Example 4.3), indicating that it can consistently identify significant regressors whose marginal quantile functions have important effects on the joint quantile function \( Q_\tau(Y|X,U) \). Third, for Examples 4.1 and 4.2, it is not uncommon to find that our proposed SMACQP slightly performs worse than qvcm, but their OQPE are very close. The main reason is that the true model structure is the varying coefficient model in Examples 4.1 and 4.2. However, SMACQP has the best prediction performance among all compared methods in Example 4.3, which shows that SMACQP has obvious advantages over some model-based methods prediction methods in terms of the out-of-sample quantile prediction error.

5  An Application

In this section, we illustrate our proposed model averaging method using an example based on Boston Housing data. This data set consists of 506 samples and is publicly available in the R package “mlbench” (http://cran.r-project.org/). Several authors had analyzed this data set by using varying coefficient models, see, [12, 19, 40, 43]. Their purpose is to identify the constant coefficients from varying ones and select relevant covariates which have significant effects on the response variable. However, our goal is to obtain the accurate predicted value for the conditional quantile of the response variable. Thus, we may make a different conclusion by analyzing this data set. Similarly to [19, 43], we regard medv (median value of owner-occupied
homes in 1,000USD) as the response variable $Y$, *lstat* (the percentage of lower status of the population) as the index variable $U$. Meanwhile, the other 11 variables are taken as covariates and their definitions can be found in Table 4. Figure 1 clearly shows that the distribution of the response is not normality, because there is significant difference between the red and blue curve and the scattered points of Q–Q plot is far away from the red dashed line. Thus, we may obtain more accuracy prediction by adopting quantile regression. Unlike previous studies, we do not assume the structure of the full model, and only take marginal varying coefficient models as our candidate models when analyzing this data set, which is more robust against model mis-specification.

Before implementing our prediction procedure, we standardize all covariates to eliminate the effects of dimensions and units. The selections of bandwidth $h$ and tuning parameter $\lambda$ are the same as those in simulation studies. We also consider two quantile levels $\tau = 0.5$ and 0.75. According to the approximate model (2.2), the estimated model weights in Table 4 show that marginal quantile functions of these variables ($zn$, $rm$ and $ptratio$) may make significant effects on estimating $Q_\tau(Y|X, U)$. In addition, we apply in-sample quantile error (IQE) to evaluate the in-sample performance of various methods, defined by $506^{-1} \sum_{i=1}^{506} \rho_\tau(Y_i - \hat{Y}_i)$. To embody the superiority of our proposal, we still compare SMACQP with existing approaches given in simulation studies. From Table 5, we can see that our proposed SMACQP and the traditional varying coefficient model prediction procedure *qvcm* outperform the other methods in terms of in-sample error.

![Histogram of $Y$](image1)

![Normal Q–Q Plot](image2)

Figure 1 The histogram (left panel) and Q–Q plot (right panel) of the response variable $Y$, where the red and blue curves represent estimated density curve of the response variable and its corresponding normal density curve
### Table 4: Regressors and estimated model weights at $\tau = 0.5$ and $\tau = 0.75$ for Boston Housing data

| Variable | Name         | Description                                      | $\tau = 0.5$          | $\tau = 0.75$          |
|----------|--------------|--------------------------------------------------|-----------------------|------------------------|
| $X_0$    | intercept    | the intercept                                    | 0.3980(0.5725)        | -0.9138(0.8870)        |
| $X_1$    | crim         | per capita crime rate by town                    | 0.6535(0.3720)        | 0.7710(0.4545)         |
| $X_2$    | zn           | proportion of residential land zoned for lots over 25,000 sq.ft | -1.3052(0.5897) | -1.5836(0.5364)        |
| $X_3$    | indus        | proportion of non-retail business acres per town | 0(0.3563)             | 0(0.3284)              |
| $X_4$    | nox          | nitric oxides concentration (parts per 10 million) | 0(0.1536)             | 0(0.2321)              |
| $X_5$    | rm           | average number of rooms per dwelling             | 0.8166(0.0742)        | 0.7921(0.0910)         |
| $X_6$    | age          | proportion of owner-occupied units built prior to 1940 | 0(0.1978)             | 0(0.1426)              |
| $X_7$    | dis          | weighted distances to five Boston employment centres | 0.4399(0.2486) | 0.3738(0.2260)         |
| $X_8$    | rad          | index of accessibility to radial highways        | 0(0.2796)             | 0(0.3564)              |
| $X_9$    | tax          | full-value property-tax rate per USD 10,000      | 0(0.3138)             | 0.3679(0.2744)         |
| $X_{10}$ | ptratio      | pupil-teacher ratio by town                      | 0.3587(0.1072)        | 0.3027(0.1275)         |
| $X_{11}$ | $b$          | $1000(B-0.63)^2$ where $B$ is the proportion of blacks by town | 0(0.2363)             | 0(0.2370)              |

### Table 5: The IQEs of various approaches for the Boston Housing data

| $\tau$ | qlm | lmma | qlmma | sma | qvcm | SMACQP |
|--------|-----|------|-------|-----|------|--------|
| 0.5    | 1.676 | 1.922 | 1.622 | 1.480 | 1.120 | 1.296  |
| 0.75   | 2.628 | –    | 2.559 | –   | 2.058 | 1.954  |

To evaluate the prediction ability of various methods, we randomly choose $n_{tr}$ observations as training set to estimate the marginal quantile functions and model weights, and calculate the prediction error $OQPE$ by using the rest $n_{te}$ observations. For eliminating randomness, we repeat this process over 200 times and corresponding results are displayed in Figure 2, indicating that SMACQP has obvious superiority in prediction accuracy.
Finally, it is necessary for us to verify whether SMACQP still has satisfactory prediction performance under a high-dimensional setting. To accomplish this, we expand the original data set by adding the following artificial covariates

\[ X_j = \xi_j + 2 \cdot \nu, \quad j = 12, \ldots, p_n \]

where \( \{\xi_{12}, \ldots, \xi_{p_n}\} \) are independent and identically distributed standard normal random variables and \( \nu \sim U(0, 1) \). Note that \( \{X_1, \ldots, X_{11}, U\} \) are variables in the original data set and \( \{X_{12}, \ldots, X_{p_n}\} \) are artificial predictors. Thus we know that these generated artificial variables obviously have no significant effects on predicting the housing price \( Y \). This approach of expanding the original data is originated from [12] and has been adopted by [30]. Here, we consider \( n_{te} = 106, p_n = 500 \) and repeat the process 200 times. Table 6 shows that SMACQP can effectively filter noise variables, owing to their small SNVs. Figure 3 shows that SMACQP achieves the best prediction performance among all the approaches.

In conclusion, the proposed SMACQP methodology is very useful in ultrahigh-dimensional scientific discoveries, which not only can effectively choose a parsimonious close-to-truth model but also performs more accuracy prediction for the joint conditional quantile function.

| \( p_n \) | qlm | sma | qvcm | SMACQP | qlm | qvcm | SMACQP |
|---------|-----|-----|------|--------|-----|------|--------|
| 500     | avg.Size | 3.870 | 22.72 | 6.595 | 3.035 | 2.855 | 11.11 | 3.240 |
|         |       | (0.804) | (3.470) | (3.373) | (0.932) | (0.613) | (6.522) | (0.681) |
| SNV     | 0    | 13.23 | 2.480 | 0.635 | 0    | 8.850 | 0.130 |
|         |      | (2.704) | (2.699) | (0.731) | (0) | (6.328) | (0.417) |

Notation: avg.Size: the averaged number of selected variables from all covariates; SNV: the averaged number of selected noise variables from artificial predictors

Table 6 The selected results and their standard errors (in parenthesis) of the Boston Housing data
6 Conclusions

In this paper, we developed a penalized quantile model averaging estimation for ultrahigh-dimensional covariates, and obtained good forecasting performances of the joint condition quantile function. This new approach can directly select those covariates whose marginal quantile functions make significant effects on the joint distribution function of the response.

Analysis of high-dimensional model averaging estimator has attracted much attention (e.g., [1, 5, 47]). Thus far, even recently, existing research findings on semiparametric model averaging have been based mostly on the continuous response variable. In practice, we are often faced with the discrete response such as the binary data or count data. Thus, it is interesting to study semiparametric model averaging estimation for the discrete response variable with the ultrahigh-dimensional covariates. This intriguing extension is left to future research.

Appendix

Let $C$ denote a positive constant that may be different at different place throughout this paper, $\theta = \sqrt{n h} \{a_j - \alpha_{0j}(u), c_j - \alpha_j(u), h(b_j - \dot{\alpha}_{0j}(u)), h(d_j - \dot{\alpha}_j(u))\}^T$, $K(u) = K(U_n, h)$, $X^*_{ij} = \{1, X_{ij}, \frac{U_i - u}{h}, X_{ij} \frac{U_i - u}{h}\}$ and $\Delta_{ij} = (X_{ij}^T \theta) / \sqrt{nh}$. Define $r_{ij} = \alpha_{0j}(U_i) - \alpha_{0j}(u) - \dot{\alpha}_{0j}(u)(U_i - u) + X_{ij} [\alpha_j(U_i) - \alpha_j(u) - \dot{\alpha}_j(u)(U_i - u)]$, $\eta_{ij} = I(\xi_{ij} + r_{ij} \leq 0) - \tau$ and $\eta_{ij} = I(\xi_{ij} \leq 0) - \tau$, where $\xi_{ij} = Y_i - Q_\tau(Y_i | X_{ij}, U_i)$.

Proof of Theorem 2.1 To apply the identity ([24])

$$\rho_\tau(u - v) - \rho_\tau(u) = v \{I(u \leq 0) - \tau\} + \int_0^v \{I(u \leq t) - I(u \leq 0)\} dt. \tag{A.1}$$

Minimizing (2.3) is equivalent to minimizing

$$L_{n,j}(\theta) = \sum_{i=1}^n [\rho_\tau \{\xi_{ij} + r_{ij} - \Delta_{ij}\} K_i(u) - \rho_\tau \{\xi_{ij} + r_{ij}\} K_i(u)].$$
Using identity (A.1) and with some straightforward calculations, it follows that
\[
L_{nj}(\theta) = \sum_{i=1}^{n} K_i(u) \eta_{ij}^{*} \Delta_{ij} + B_{nj}(\theta) = W_{nj}^{*T} \theta + B_{nj}(\theta),
\]  
(A.2)
where \( W_{nj}^{*} = \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} K_i(u) \eta_{ij} X_{ij}^{*} \) and
\[
B_{nj}(\theta) = \sum_{i=1}^{n} \int_{0}^{\Delta_{ij}} K_i(u) \{ I(\xi_{ij} + r_{ij} \leq t) - I(\xi_{ij} + r_{ij} \leq 0) \} dt.
\]

Let \( \mathcal{X}_j = (X_{ij}, \ldots, X_{nj})^T \) and \( \mathcal{U} = (U_1, \ldots, U_n)^T \). Since \( B_{nj}(\theta) \) is a summation of independent random variables of the kernel form, it follows that
\[
E(B_{nj}(\theta)|\mathcal{X}_j, \mathcal{U}) = \sum_{i=1}^{n} \int_{0}^{\Delta_{ij}} K_i(u) \{ F_{ij}(t - r_{ij}|X_{ij}, U_i) - F_{ij}(-r_{ij}|X_{ij}, U_i) \} dt
\]
\[
= \sum_{i=1}^{n} \int_{0}^{\Delta_{ij}} K_i(u) \{ f_{ij}(-r_{ij}|X_{ij}, U_i) t (1 + o(1)) \} dt
\]
\[
= \frac{1}{2} \sum_{i=1}^{n} f_{ij}(0|X_{ij}, U_i) K_i(u) \Delta_{ij}^2 + o_p(1)
\]
\[
\Delta \frac{1}{2} \theta^T S_{nj}^{*T} \theta + o_p(1),
\]  
(A.3)
where \( S_{nj}^{*} = \sum_{i=1}^{n} f_{ij}(0|X_{ij}, U_i) K_i(u) X_{ij}^{*} X_{ij}^{*T} / (nh) \), and
\[
\text{Var}(B_{nj}(\theta)|\mathcal{X}_j, \mathcal{U})
\]
\[
= \sum_{i=1}^{n} \text{Var} \left\{ \left( \int_{0}^{\Delta_{ij}} K_i(u) \{ I(\xi_{ij} + r_{ij} \leq t) - I(\xi_{ij} + r_{ij} \leq 0) \} dt \right) | \mathcal{X}_j, \mathcal{U} \right\}
\]
\[
\leq \sum_{i=1}^{n} E \left\{ \left( \int_{0}^{\Delta_{ij}} K_i(u) \{ I(\xi_{ij} + r_{ij} \leq t) - I(\xi_{ij} + r_{ij} \leq 0) \} dt \right)^2 | \mathcal{X}_j, \mathcal{U} \right\}
\]
\[
\leq \sum_{i=1}^{n} K_i^2(u) \int_{0}^{\Delta_{ij}} \int_{0}^{\Delta_{ij}} \{ F_{ij}((-|\Delta_{ij}| - r_{ij})|X_{ij}, U_i) - F_{ij}(-r_{ij}|X_{ij}, U_i) \} dt_1 dt_2
\]
\[
= o \left( \sum_{i=1}^{n} K_i^2(u) \Delta_{ij}^2 \right) = o_p(1).
\]  
(A.4)

Similar to Parzen [37], we have
\[
\frac{1}{nh} \sum_{i=1}^{n} K_i(u) \left( \frac{U_i - u}{h} \right)^k \overset{p}{\to} f_U(u) \mu_k,
\]
where \( \overset{p}{\to} \) stands for convergence in probability. Thus
\[
S_{nj}^{*} \overset{p}{\to} f_U(u) S_{j}^{*}(u),
\]
where \( S_{j}^{*} = \text{diag}\{D_j(u), \mu_2 D_j(u)\} \) is a quasi-diagonal matrix. This together with (A.2)–(A.4), leads to
\[
L_{nj}(\theta) = \frac{1}{2} f_U(u) \theta^T S_{j}^{*}(u) \theta + W_{nj}^{*T} \theta + o_p(1).
\]
By applying the convexity lemma ([38]) and the quadratic approximation lemma ([8]), the minimizer of \( L_{nj}(\theta) \) can be expressed as
\[
\hat{\theta} = -\frac{1}{f_U(u) S_{j}^{*-1}(u)} W_{nj}^{*} + o_p(1).
\]
Therefore, we have
\[
\sqrt{nh} \left( \hat{a}_j - \alpha_{0j}(u) \right) = -\frac{1}{f_U(u)} D_j^{-1}(u) W_{n_j}^{(1)} + o_p(1),
\]
\[\text{(A.5)}\]
where \( W_{n_j}^{(1)} = (w_{j1}^*, w_{j2}^*)^T = \frac{1}{\sqrt{nh}} \sum_{i=1}^n K_i(u) \eta_{ij}^* (1, X_{ij})^T \) with \( w_{j1}^* = \frac{1}{\sqrt{nh}} \sum_{i=1}^n K_i(u) \eta_{ij}^* \) and \( w_{j2}^* = \frac{1}{\sqrt{nh}} \sum_{i=1}^n K_i(u) \eta_{ij}^* X_{ij} \). Let \( W_{n_j}^{(1)} = (w_{j1}, w_{j2})^T = \frac{1}{\sqrt{nh}} \sum_{i=1}^n K_i(u) \eta_{ij} (1, X_{ij})^T \) with \( w_{j1} = \frac{1}{\sqrt{nh}} \sum_{i=1}^n K_i(u) \eta_{ij} \) and \( w_{j2} = \frac{1}{\sqrt{nh}} \sum_{i=1}^n K_i(u) \eta_{ij}^* X_{ij} \). By the results of [37], we can obtain
\[
(nh)^{-1} \sum_{i=1}^n K_i^2(u) \left( \frac{U_i - u}{h} \right)^k \to f_U(u) \nu_k.
\]
In addition, by some calculations, we have that \( E(\eta_{ij}) = 0 \) and \( \text{Var}(\eta_{ij}) = \tau(1 - \tau) \). By the Cramér–Wold theorem, the central limit theorem for \( W_{n_j}^{(1)} \) holds. Therefore,
\[\text{(W}_{n_j}^{(1)} | \mathcal{X}_j, \mathcal{U}) \overset{d}{\to} N(0, \tau(1 - \tau)v_0 f_U(u) B_j(u)). \]
\[\text{(A.6)}\]
Moreover, we have
\[
\text{Var}(w_{j1}^* - w_{j1} | \mathcal{X}_j, \mathcal{U}) = \frac{1}{nh} \sum_{i=1}^n K_i^2(u) \text{Var}(\eta_{ij}^* - \eta_{ij} | \mathcal{X}_j, \mathcal{U})
\leq \frac{1}{nh} \sum_{i=1}^n K_i^2(u) [F_{ij}(|r_{ij}||X_{ij}, U_i) - F_{ij}(0|X_{ij}, U_i)] = o_p(1),
\]
and
\[
\text{Var}(w_{j2}^* - w_{j2} | \mathcal{X}_j, \mathcal{U}) = \frac{1}{nh} \sum_{i=1}^n K_i^2(u) X_{ij}^2 \text{Var}(\eta_{ij}^* - \eta_{ij} | \mathcal{X}_j, \mathcal{U})
\leq \frac{1}{nh} \sum_{i=1}^n K_i^2(u) X_{ij}^2 [F_{ij}(|r_{ij}||X_{ij}, U_i) - F_{ij}(0|X_{ij}, U_i)] = o_p(1).
\]
Thus, we obtain \( \text{Var}((W_{n_j}^{(1)} - W_{n_j}^{(1)}) | \mathcal{X}_j, \mathcal{U}) = o_p(1) \). So by Slutsky’s theorem, conditioning on \( \mathcal{X}_j, \mathcal{U} \), we have
\[
W_{n_j}^{(1)} | \mathcal{X}_j, \mathcal{U} \overset{d}{\to} N(0, \tau(1 - \tau)v_0 f_U(u) B_j(u)).
\]
\[\text{(A.7)}\]
We now calculate the conditional mean of \( W_{n_j}^{(1)} \),
\[
\frac{1}{\sqrt{nh}} E[W_{n_j}^{(1)} | \mathcal{X}_j, \mathcal{U}]
= \frac{1}{nh} \sum_{i=1}^n K_i(u) [F_{ij}(-r_{ij}|X_{ij}, U_i) - F_{ij}(0|X_{ij}, U_i)](1, X_{ij})^T
= -\frac{1}{nh} \sum_{i=1}^n K_i(u) f_{ij}(0|X_{ij}, U_i) r_{ij} \{1 + o(1)\}(1, X_{ij})^T
= -\frac{1}{2} \mu_2 h^2 f_U(u) D_j(u) \begin{pmatrix} \breve{\alpha}_{0j}(u) \\ \breve{\alpha}_j(u) \end{pmatrix} + o_p(h^2).
\]
\[\text{(A.8)}\]
The proof is completed by combining (A.5)–(A.8). Therefore,

\[
\sqrt{nh} \left[ \begin{pmatrix} \tilde{a}_j - \alpha_0j(u) \\ \tilde{c}_j - \alpha_j(u) \end{pmatrix} - \frac{1}{2} \mu_2 h^2 \begin{pmatrix} \tilde{a}_0j(u) \\ \tilde{c}_j(u) \end{pmatrix} \right] \overset{d}{\to} N \left( 0, \frac{\tau(1-\tau)\nu_0}{f(u)} D_j^{-1}(u) B_j(u) D_j^{-1}(u) \right).
\]

This completes the proof. \qed

Proof of Theorem 2.2 Define \( \mathcal{M}_{i,A} \) as a subvector consisting of the first \((q_n + 1)\) element of \( \mathcal{M}_i \). Let \( L_n(\omega_n^{(1)}) = \sum_{i=1}^{q_n} \rho_i \{ Y_i - \hat{M}_{i,A}^T \omega_n^{(1)} \} \), \( \alpha_n = \sqrt{q_n/n} \) and \( u_n = \alpha_n^{-1}(\omega_n^{(1)} - \omega_n^{(0)}) \). In order to prove the convergence rate in Theorem 2.2 (i), our aim is to show that for any given \( \delta > 0 \), there is a large constant \( C > 0 \) such that, for a large \( n \), we have

\[
P \left\{ \inf_{\|u_n\| = C} L_n(\omega_n^{(1)} + \alpha_n u_n) > L_n(\omega_n^{(1)}) \right\} \geq 1 - \delta.
\]

(A.9)

Using Knight identity (A.1) and some straightforward calculations, it follows that

\[
G_n = L_n(\omega_n^{(1)} + \alpha_n u_n) - L_n(\omega_n^{(1)})
\]

\[
= \sum_{i=1}^{n} \alpha_n \hat{M}_{i,A}^T u_n [I \{ Y_i - \hat{M}_{i,A}^T \omega_n^{(1)} \leq 0 \} - \tau]
\]

\[
+ \sum_{i=1}^{n} \int_{0}^{n} \alpha_n \hat{M}_{i,A}^T u_n [I \{ Y_i - \hat{M}_{i,A}^T \omega_n^{(1)} \leq t \} - I \{ Y_i - \hat{M}_{i,A}^T \omega_n^{(1)} \leq 0 \}] dt
\]

\[
\Delta = I + II.
\]

(A.10)

Let \( \Delta_i = (\hat{M}_{i,A} - M_{i,A})^T \omega_n^{(1)} \). Then the first term \( I \) can be divided into the following three parts

\[
I = \sum_{i=1}^{n} \alpha_n \hat{M}_{i,A}^T u_n [I \{ \varepsilon_i \leq \Delta_i \} - \tau]
\]

\[
= \sum_{i=1}^{n} \alpha_n \hat{M}_{i,A}^T u_n [I \{ \varepsilon_i \leq 0 \} - \tau] + \sum_{i=1}^{n} \alpha_n \hat{M}_{i,A}^T u_n [F_i(\Delta_i|X_i, U_i) - F_i(0|X_i, U_i)]
\]

\[
+ \sum_{i=1}^{n} \alpha_n \hat{M}_{i,A}^T u_n [I \{ \varepsilon_i \leq \Delta_i \} - I \{ \varepsilon_i \leq 0 \}] - [F_i(\Delta_i|X_i, U_i) - F_i(0|X_i, U_i)]
\]

\[
= I_1 + I_2 + I_3.
\]

(A.11)

By Theorem 2.1, we have, for uniformly for \( u \) and \( j = 1, \ldots, q_n \), \( \hat{a}_j - \alpha_0j(u) = O(h^2 + 1/\sqrt{n}h) \) and \( \hat{c}_j - \alpha_j(u) = O(h^2 + 1/\sqrt{n}h) \). Observe that

\[
\hat{M}_{i,A} \hat{M}_{i,A}^T - M_{i,A} M_{i,A}^T = (\hat{M}_{i,A} - M_{i,A}) M_{i,A}^T + M_{i,A}(\hat{M}_{i,A} - M_{i,A})^T + (\hat{M}_{i,A} - M_{i,A})(\hat{M}_{i,A} - M_{i,A})^T.
\]

Then, for any \( \delta > 0 \), by Chebyshev’s inequality and following the proof of Lemma 8 in Fan and Peng [13], we have

\[
P \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \hat{M}_{i,A} M_{i,A}^T - A_n \right\|_F > \delta \right\} \leq \frac{1}{n^2 \delta^2} E \left[ \left\| \sum_{i=1}^{n} \hat{M}_{i,A} M_{i,A}^T - nA_n \right\|_F^2 \right] = O(q_n^2/n) = o(1)
\]

under Condition (C7), \( \|A\|_F \) stands for Frobenius norm for a real matrix \( A = (a_{ij})_{i,j}^{m,n} \), namely,
By direct calculation of the mean and variance, we can show that
\[
\|A\|_F = (\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2)^{1/2}
\] Hence, we have
\[
\left\| \frac{1}{n} \sum_{i=1}^{n} M_i A M_i^T - \Lambda_n \right\|_F = o_p(1).
\] (A.12)

Equation (A.12) and Condition (C6) imply that \( u_n^T (\sum_{i=1}^{n} M_i A M_i^T / n) u_n \) is asymptotically dominated by \( u_n^T A_n u_n \). As \( q_n(h^2 + 1/\sqrt{n}h) = o(n^{1/2}(h^2 + 1/\sqrt{n}h)) = o(1) \) from Condition (C7), we can easily prove that
\[
\left\| \frac{1}{n} \sum_{i=1}^{n} (\hat{M}_i A - M_i A) u_n^T \right\|_F \rightarrow 0, \quad \left\| \frac{1}{n} \sum_{i=1}^{n} (\hat{M}_i A - M_i A)^T u_n \right\|_F \rightarrow 0,
\]
and
\[
\left\| \frac{1}{n} \sum_{i=1}^{n} (\hat{M}_i A - M_i A)(\hat{M}_i A - M_i A)^T u_n \right\|_F \rightarrow 0.
\]
Thus
\[
\frac{1}{n} \sum_{i=1}^{n} \hat{M}_i A M_i^T = \frac{1}{n} \sum_{i=1}^{n} M_i A M_i^T \Lambda_n.
\] (A.13)

By condition \( P\{\varepsilon_i \leq 0 \mid X_i, U_i\} = \tau \), we have \( E(I_1) = 0 \) and by Condition (C6) and (A.13), we have
\[
\text{Var}(I_1) = \sum_{i=1}^{n} \text{Var}\{\alpha_n \hat{M}_i A u_n[I\{\varepsilon_i \leq 0\} - \tau]\}
\]
\[
= \alpha_n^2 \tau(1 - \tau) u_n^T \sum_{i=1}^{n} E(\hat{M}_i A M_i^T) u_n
\]
\[
= O(n\alpha_n^2 \|u_n\|^2).
\]
Thus
\[
I_1 = O_p(\sqrt{n\alpha_n^2 \|u_n\|}) = o_p(n\alpha_n^2 \|u_n\|).
\] (A.14)

In addition, by Condition (C4), taking Taylor’s explanation for \( F_i(\Delta_i \mid X_i, U_i) \) at 0 gives
\[
I_2 = \sum_{i=1}^{n} \alpha_n \hat{M}_i A u_n[f_i(0 \mid X_i, U_i) \Delta_i \{1 + o(1)\}] = o_p(n\alpha_n^2 \|u_n\|).
\] (A.15)

By direct calculation of the mean and variance, we can show that \( I_3 = o_p(n\alpha_n^2 \|u_n\|) \). This combined with (A.11), (A.14) and (A.15) lead to \( I = o_p(n\alpha_n^2 \|u_n\|) \).

Let \( X = (X_{11}, \ldots, X_{1p_n}, \ldots, X_{n1}, \ldots, X_{np_n})^T \), we have
\[
E(II \mid X, U) = \sum_{i=1}^{n} \int_{0}^{t} \alpha_n \hat{M}_i A u_n [F_i((t + \Delta_i) \mid X_i, U_i) - F_i(\Delta_i \mid X_i, U_i)] dt
\]
\[
= \sum_{i=1}^{n} \int_{0}^{t} \alpha_n \hat{M}_i A u_n [f_i(\Delta_i \mid X_i, U_i) t(1 + o(1))] dt
\]
\[
= \frac{1}{2} \sum_{i=1}^{n} f_i(0 \mid X_i, U_i) (\alpha_n \hat{M}_i A u_n)^2 (1 + o(1))
\]
\[
= O_p(n\alpha_n^2 \|u_n\|^2).
\] (A.16)
By the Schwarz inequality, it is not difficult to show that

\[ \text{Var}(II|X, U) \]

\[ \leq \sum_{i=1}^{n} E \left( \int_{0}^{\alpha_{n}N_{iA}^{T}u_{n}} |I\{\varepsilon_{i} \leq t + \Delta_{i}\} - I\{\varepsilon_{i} \leq \Delta_{i}\}|dt|X, U \right)^{2} \]

\[ \leq \sum_{i=1}^{n} \int_{0}^{\alpha_{n}N_{iA}^{T}u_{n}} \int_{0}^{\alpha_{n}N_{iA}^{T}u_{n}} |F_{i}(\{\alpha_{n}N_{iA}^{T}u_{n} + \Delta_{i}\}|X, U_{i}) - F_{i}(\Delta_{i}|X_{i}, U_{i})|dt_{1}dt_{2} \]

\[ = o_{p}(n\alpha_{n}^{2}\|u_{n}\|^{2}). \quad (A.17) \]

It follows from (A.16)–(A.17) and above results that \( G_{n} \) in (A.10) is dominated by \( II \) which is positive when a sufficiently large \( \|u_{n}\| = C \) is chosen. Therefore, (A.9) holds and this completes the proof of Theorem 2.2 (i).

(ii) It can be shown easily that there exists a \( \hat{\omega}^{(1)}_{n} \) in Theorem 2.2 (i) that is a \( \sqrt{n/n_{q}} \) consistent local minimizer of \( L_{n}(\omega^{(1)}_{n}) \), and satisfies the equations

\[ n^{-1} \sum_{i=1}^{n} \hat{M}_{iA}\psi_{\tau}\{Y_{i} - \hat{M}_{iA}^{T}\hat{\omega}^{(1)}_{n}\} = 0, \quad (A.18) \]

where \( \psi_{\tau}(u) = \tau - I(u \leq 0) \). We can write

\[ n^{-1} \sum_{i=1}^{n} \hat{M}_{iA}\psi_{\tau}\{Y_{i} - \hat{M}_{iA}^{T}\hat{\omega}^{(1)}_{n}\} = -z_{n1} + B_{n1} + B_{n2}, \quad (A.19) \]

where

\[ z_{n1} = n^{-1} \sum_{i=1}^{n} \hat{M}_{iA}\{I(\varepsilon_{i} \leq 0) - \tau\}, \]

\[ B_{n1} = n^{-1} \sum_{i=1}^{n} \hat{M}_{iA}\{F_{i}(0|X_{i}, U_{i}) - F_{i}(\zeta_{i}|X_{i}, U_{i})\}, \]

\[ B_{n2} = n^{-1} \sum_{i=1}^{n} (\hat{M}_{iA} - \hat{M}_{iA})^{T}\omega_{n}^{(1)}, \]

where \( \zeta_{i} = \hat{M}_{iA}^{T}(\hat{\omega}^{(1)}_{n} - \omega_{n}^{(1)}) + (\hat{M}_{iA} - \hat{M}_{iA})^{T}\omega_{n}^{(1)} \).

Taking Taylor’s explanation for \( F_{i}(\zeta_{i}|X_{i}, U_{i}) \) at 0 gives

\[ B_{n1} = -n^{-1} \sum_{i=1}^{n} f_{i}(0|X_{i}, U_{i})\hat{M}_{iA}\zeta_{i}\{1 + o(1)\} \]

\[ = -n^{-1} \left( \sum_{i=1}^{n} f_{i}(0|X_{i}, U_{i})\hat{M}_{iA}\hat{M}_{iA}^{T}\hat{\omega}^{(1)}_{n} - \omega_{n}^{(1)} \right) \]

\[ + \sum_{i=1}^{n} f_{i}(0|X_{i}, U_{i})\hat{M}_{iA}(\hat{M}_{iA} - \hat{M}_{iA})^{T}\omega_{n}^{(1)} \}

\[ \{1 + o(1)\}. \quad (A.20) \]

By direct calculation of the mean and variance, we can show, as in [21], that \( B_{n2} = o_{p}(\sqrt{n/n_{q}}) \).

This combined with (A.18)–(A.20) leads to

\[ \hat{\omega}^{(1)}_{n} - \omega_{n}^{(1)} \]

\[ = - \left[ n^{-1} \sum_{i=1}^{n} f_{i}(0|X_{i}, U_{i})\hat{M}_{iA}\hat{M}_{iA}^{T} \right]^{-1} n^{-1} \sum_{i=1}^{n} f_{i}(0|X_{i}, U_{i})\hat{M}_{iA}(\hat{M}_{iA} - \hat{M}_{iA})^{T}\omega_{n}^{(1)} \]

\[ - \left[ n^{-1} \sum_{i=1}^{n} f_{i}(0|X_{i}, U_{i})\hat{M}_{iA}\hat{M}_{iA}^{T} \right]^{-1} z_{n1}\{1 + o_{p}(1)\} \]
= \Pi_{n1} + \Pi_{n2}. \tag{A.21}

We first derive the leading term of \( \Pi_{n1} \). According to (A.5), for each \( j = 1, \ldots, q_n \), we have

\[
\left( \hat{a}_j(u) - \alpha_{0j}(u) \right) \left( \hat{c}_j(u) - \alpha_j(u) \right) = \frac{1}{nhf(U)} D_j^{-1}(u) \sum_{i=1}^{n} K_i(u) \psi_\tau(\xi_{ij}) (1, X_{ij})^T \{1 + o_p(1)\}.
\]

Thus

\[
\mathcal{M}_{ij} - M_{ij} = Q_\tau(Y_i | X_{ij}, U_i) - Q_\tau(Y_i | X_{ij}, U_i) = (1, X_{ij}) \left( \hat{a}_j(U_i) - \alpha_{0j}(U_i) \right) \left( \hat{c}_j(U_i) - \alpha_j(U_i) \right) = (1, X_{ij}) \frac{1}{nhf(U_i)} D_j^{-1}(U_i) \sum_{s=1}^{n} K \left( \frac{U_s - U_i}{h} \right) \psi_\tau(\xi_{sj}) (1, X_{sj})^T \{1 + o_p(1)\}. \tag{A.22}
\]

On the other hand, by (A.13), (A.21) and (A.22), we have

\[
\Pi_{n1} = -K_n^{-1} n^{-1} \Pi_{n3} \{1 + o_p(1)\}, \tag{A.23}
\]

where

\[
\Pi_{n3} = \left\{ \sum_{i=1}^{n} \mathcal{M}_{ij} f_i(0 | X_i, U_i) \sum_{k=1}^{q_n} \omega_{0k} \mathcal{M}_{ik} - M_{ik} \right\}^T_{j=0,1, \ldots, q_n}
\]

\[
= \left\{ \sum_{i=1}^{n} \mathcal{M}_{ij} f_i(0 | X_i, U_i) \sum_{k=1}^{q_n} \omega_{0k} \frac{1}{nhf(U_i)} (1, X_{ik}) D_k^{-1}(U_i) \right\}^T_{j=0,1, \ldots, q_n}
\]

\[
\times \sum_{s=1}^{n} K \left( \frac{U_s - U_i}{h} \right) \psi_\tau(\xi_{sk}) (1, X_{sk})^T \{1 + o_p(1)\} \}
\]

\[
\times \left\{ \sum_{s=1}^{n} \sum_{k=1}^{q_n} \omega_{0k} \psi_\tau(\xi_{sk}) \mathcal{M}_{ij} f_i(0 | X_i, U_i) K \left( \frac{U_s - U_i}{h} \right) (1, X_{ik}) D_k^{-1}(1, X_{sk})^T \right\}^T_{j=0,1, \ldots, q_n}
\]

\[
\times \{1 + o_p(1)\}
\]

\[
= \left\{ \sum_{s=1}^{n} \phi_{s0} \sum_{s=1}^{n} \phi_{s1}, \ldots, \sum_{s=1}^{n} \phi_{sq_n} \right\}^T \{1 + o_p(1)\}, \tag{A.24}
\]

where \( \phi_{sj} = \sum_{k=1}^{q_n} \omega_{0k} \psi_\tau(\xi_{sk}) \mathcal{E}(f_s(0 | X_s, U_s) M_{sj}(1, X_{sk}) D_k^{-1}(U_s)(1, X_{sk})^T | U_s) \). We next consider \( \Pi_{n2} \). Observe that

\[
\Pi_{n2} = \left[ n^{-1} \sum_{i=1}^{n} f_i(0 | X_i, U_i) \hat{M}_{iA} \hat{M}_{iA}^T \right]^{-1} n^{-1} \sum_{i=1}^{n} \hat{M}_{iA} \psi_\tau(\xi_i) \{1 + o_p(1)\}
\]

\[
= K_n^{-1} n^{-1} \sum_{i=1}^{n} \hat{M}_{iA} \psi_\tau(\xi_i) \{1 + o_p(1)\}
\]

\[
= K_n^{-1} n^{-1} \left[ \sum_{i=1}^{n} M_{iA} \psi_\tau(\xi_i) + \sum_{i=1}^{n} (\hat{M}_{iA} - M_{iA}) \psi_\tau(\xi_i) \right] \{1 + o_p(1)\}.\]
We can show that the leading term of $\Pi_{n2}$ is $K_n^{-1}n^{-1} \sum_{i=1}^n M_i \psi_\tau(\varepsilon_i)$. That is,
\[
\Pi_{n2} = K_n^{-1}n^{-1}\left(\sum_{s=1}^n \phi_s0, \sum_{s=1}^n \phi_s1, \ldots, \sum_{s=1}^n \phi_{sqn}\right)^T \{1 + o_p(1)\},
\]  
(A.25)
where $\phi_{sj} = M_{sj}\psi_\tau(\varepsilon_s)$. By (A.21), (A.23)–(A.25), we have
\[
\hat{\omega}^{(1)} - \omega^{(1)} = K_n^{-1}n^{-1} \sum_{s=1}^n (\phi_s - \phi_s).
\]
Let $\sqrt{n}e_n^T e_n^{-1} K_n^{-1} n^{-1} \sum_{s=1}^n (\phi_s - \phi_s) = \sum_{s=1}^n D_{ns}$, $D_{ns} = n^{-1/2}e_n^T \Sigma_n^{-1/2} K_n^{-1} (\phi_s - \phi_s)$. To verify asymptotic normality, we first note that $E(D_{ns}) = 0$ and
\[
\text{Var}\left(\sum_{s=1}^n D_{ns}\right) = e_n^T \Sigma_n^{-1/2} K_n^{-1} S_n K_n^{-1} \Sigma_n^{-1/2} e_n = 1.
\]
The proof is complete by checking the Lindeberg–Feller condition. For any $\epsilon > 0$ and using Conditions (C1)–(C4) and (C7), we have
\[
\sum_{s=1}^n E[D_{ns}^2 I(|D_{ns}| > \epsilon)]
\leq \epsilon^{-2} \sum_{s=1}^n E(D_{ns}^4)
\leq (n\epsilon)^{-2} \sum_{s=1}^n \rho_{\max}^2(e_n e_n^T) \rho_{\max}^2(\Sigma_n^{-1}) \rho_{\max}^2(K_n^{-2}) E((\phi_s - \phi_s)^T (\phi_s - \phi_s))^2
\leq C(n\epsilon)^{-2} \sum_{s=1}^n E(||\phi_s - \phi_s||^4)
= O(q_n^2/n) = o(1).
\]
We complete the proof. \hfill \Box

**Lemma 6.1** Define $g(\eta)$ and $h(\eta)$ as convex functions with subdifferential functions $\partial g(\eta)$ and $\partial h(\eta)$. Let $\eta^*$ be a point that has neighborhood $\Gamma$ such that $\partial h(\eta) \cap \partial g(\eta^*) \neq \emptyset$, $\forall \eta \in \Gamma \cap \text{dom}(g)$. Then $\eta^*$ is a local minimizer of $g(\eta) - h(\eta)$.

**Proof** The proof is available in [41]. \hfill \Box

**Lemma 6.2** Assume that Conditions (C1)–(C8) are satisfied and that $q_n n^{-1/2} = o(\lambda)$, $\log(p_n)$ = $o(n\lambda^2)$ and $n\lambda^2 \to \infty$, we have
\[
P\left\{\max_{q_n+1 \leq j \leq p_n} \frac{n}{n} \sum_{i=1}^n \hat{M}_{ij}^T [I(Y_i - \hat{M}_{iA}^T \omega_{n0}^{(1)} \leq 0) - I] > \lambda/2\right\} \to 0
\]
as $n \to \infty$.

**Proof** From the definition $\Delta_i = (\hat{M}_{iA} - M_{iA})^T \omega_{n0}^{(1)}$, there exists a positive constant $c$, we have
\[
{n^{-1} \sum_{i=1}^n \hat{M}_{ij} [I(Y_i - \hat{M}_{iA}^T \omega_{n0}^{(1)} \leq 0) - I(Y_i - M_{iA}^T \omega_{n0}^{(1)} \leq 0)]}
= \left| n^{-1} \sum_{i=1}^n \hat{M}_{ij} \{[F_i(\Delta_i | X_i, U_i) - F_i(0|X_i, U_i)] + |I(\varepsilon_i \leq \Delta_i) - I(\varepsilon_i \leq 0)]
\right|
\]
\[-[F_i(\Delta_i|X_i, U_i) - F_i(0|X_i, U_i)]\]
\[\leq n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} [F_i(\Delta_i|X_i, U_i) - F_i(0|X_i, U_i)]\]
\[+ n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} \{|I(\varepsilon_i \leq \Delta_i) - I(\varepsilon_i \leq 0)| - [F_i(\Delta_i|X_i, U_i) - F_i(0|X_i, U_i)]\}\]
\[\leq c_q n (h^2 + 1/\sqrt{nh})\]
\[+ n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} \{|I(\varepsilon_i \leq \Delta_i) - I(\varepsilon_i \leq 0)| - [F_i(\Delta_i|X_i, U_i) - F_i(0|X_i, U_i)]\}.\] (A.26)

Since \(I(Y_i - \mathcal{M}_{tA}^T \omega_n^{(1)} \leq 0), i = 1, \ldots, n\) are independent Bernoulli random variables with mean \(\tau\) and \(\hat{M}_{ij}, j = q_n + 1, \ldots, p_n\) are uniformly bounded, we have
\[P\left\{ n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} |I(Y_i - \mathcal{M}_{tA}^T \omega_n^{(1)} \leq 0) - \tau| > \lambda/4 \right\} \leq \exp(-C n \lambda^2)\] (A.27)
by Hoeffding’s inequality. Furthermore, \(I(\varepsilon_i \leq \Delta_i) - I(\varepsilon_i \leq 0), i = 1, \ldots, n\) are independent random variables and \(E\{|I(\varepsilon_i \leq \Delta_i) - I(\varepsilon_i \leq 0)| - [F_i(\Delta_i|X_i, U_i) - F_i(0|X_i, U_i)]\} = 0\), we have
\[P\left\{ n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} \{|I(\varepsilon_i \leq \Delta_i) - I(\varepsilon_i \leq 0)| - [F_i(\Delta_i|X_i, U_i) - F_i(0|X_i, U_i)]\} > \lambda/8 \right\}\]
\[\leq \exp(-C n \lambda^2)\] (A.28)
by Hoeffding’s inequality. By (A.26)–(A.28), we have
\[P\left\{ n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} |I(Y_i - \mathcal{M}_{tA}^T \omega_n^{(1)} \leq 0) - \tau| > \lambda/2 \right\}\]
\[\leq P\left\{ n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} |I(Y_i - \mathcal{M}_{tA}^T \omega_n^{(1)} \leq 0) - \tau| > \lambda/4 \right\}\]
\[+ P\left\{ n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} |I(Y_i - \mathcal{M}_{tA}^T \omega_n^{(1)} \leq 0) - I(Y_i - \mathcal{M}_{tA}^T \omega_n^{(1)} \leq 0)| > \lambda/4 \right\}\]
\[\leq P\left\{ n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} |I(Y_i - \mathcal{M}_{tA}^T \omega_n^{(1)} \leq 0) - \tau| > \lambda/4 \right\}\]
\[+ P\left\{ n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} \{|I(\varepsilon_i \leq \Delta_i) - I(\varepsilon_i \leq 0)| - [F_i(\Delta_i|X_i, U_i) - F_i(0|X_i, U_i)]\} > \lambda/4 - c_q n \left( h^2 + \frac{1}{\sqrt{nh}} \right) \right\}\]
\[\leq P\left\{ n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} |I(Y_i - \mathcal{M}_{tA}^T \omega_n^{(1)} \leq 0) - \tau| > \lambda/4 \right\}\]
\[+ P\left\{ n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} \{|I(\varepsilon_i \leq \Delta_i) - I(\varepsilon_i \leq 0)| - [F_i(\Delta_i|X_i, U_i) - F_i(0|X_i, U_i)]\} > \lambda/8 \right\}\]
\[\leq \exp(-C n \lambda^2),\]
where the third inequality follows from \( q_n(h^2 + \frac{1}{\sqrt{n}h}) = o(\lambda) \) under the conditions \( q_n n^{-1/2} = o(\lambda) \) and \( n^{1/2}(h^2 + 1/\sqrt{n}h) = o(1) \). Thus

\[
P \left\{ \max_{q_n+1 \leq j \leq p_n} \sum_{i=1}^{n} \hat{\mathcal{M}}_{ij} |I(Y_i - \hat{M}_{iA}^T \omega_n^{(1)} \leq 0) - \tau| > \lambda/2 \right\} \\
\leq p_n \exp(-C\lambda^2) = \exp(p_n - C\lambda^2) \to 0
\]

under the conditions \( \log(p_n) = o(\lambda^2) \) and \( n\lambda \to \infty \). \( \square \)

**Lemma 6.3** Assume Conditions (C1)–(C8) hold, \( n^{-1/2}q_n = o(\lambda) \), \( \log(p_n) = o(n\lambda^2) \) and \( n\lambda \to \infty \). Then for some positive constant \( \Delta \)

\[
P \left\{ \max_{q_n+1 \leq j \leq p_n} \sup_{\|\omega_n^{(1)} - \omega_n^{(0)}\| \leq \Delta \sqrt{q_n/n}} \sum_{i=1}^{n} \hat{\mathcal{M}}_{ij} |I(Y_i - \hat{M}_{iA}^T \omega_n^{(1)} \leq 0) - I(Y_i - \hat{M}_{iA}^T \omega_n^{(0)} \leq 0) - \lambda/2 \right\} \to 0
\]

**Proof** Based on the idea of [45] and [44], we cover the ball \( \{\omega_n^{(1)} : \|\omega_n^{(1)} - \omega_n^{(0)}\| \leq \Delta \sqrt{q_n/n}\} \) with a net of balls with radius \( \Delta \sqrt{q_n/n} \) with cardinality \( N \leq Cn^{q_n} \). Denote the \( N \) balls by \( B(t_1), \ldots, B(t_N) \), where the ball \( B(t_k) \) is centered at \( t_k, k = 1, \ldots, N \). To simplify the notation, let \( \kappa_i(\omega_n^{(1)}) = Y_i - \hat{M}_{iA}^T \omega_n^{(1)} \). Then

\[
P \left\{ \sup_{\|\omega_n^{(1)} - \omega_n^{(0)}\| \leq \Delta \sqrt{q_n/n}} \sum_{i=1}^{n} \hat{\mathcal{M}}_{ij} |I(\kappa_i(\omega_n^{(1)}) \leq 0) - I(\kappa_i(\omega_n^{(0)}) \leq 0) - \lambda/2 \right\} \leq \sum_{k=1}^{N} P \left\{ \sum_{i=1}^{n} \hat{\mathcal{M}}_{ij} |I(\kappa_i(t_k) \leq 0) - I(\kappa_i(\omega_n^{(0)}) \leq 0) - \lambda/2 \right\}
\]

\[
+ \sum_{k=1}^{N} P \left\{ \sup_{\|\omega_n^{(1)} - t_k\| \leq \Delta \sqrt{q_n/n}} \sum_{i=1}^{n} \hat{\mathcal{M}}_{ij} |I(\kappa_i(\omega_n^{(1)}) \leq 0) - I(\kappa_i(t_k) \leq 0) - \lambda/2 \right\}
\]

\[
\equiv I_{nj1} + I_{nj2}
\]

Let \( u_i = \hat{\mathcal{M}}_{ij} |I(\kappa_i(t_k) \leq 0) - I(\kappa_i(\omega_n^{(0)}) \leq 0) - \lambda/2 \right| \). Then the \( u_i \) are independent mean-zero random variables, and

\[
\text{Var}(u_i|X, U) = \hat{\mathcal{M}}_{ij}^2 [F_i(\hat{M}_{iA}^T(t_k - \omega_n^{(1)})|X_i, U_i) (1 - F_i(\hat{M}_{iA}^T(t_k - \omega_n^{(0)}|X_i, U_i)) + F_i(0|X_i, U_i) (1 - F_i(0|X_i, U_i)) - 2F_i(\min(\hat{M}_{iA}^T(t_k - \omega_n^{(1)}), 0)|X_i, U_i)
\]

\[
+ 2F_i(\hat{M}_{iA}^T(t_k - \omega_n^{(0)}|X_i, U_i) (1 - F_i(0|X_i, U_i)) \leq C|\hat{M}_{iA}^T(t_k - \omega_n^{(1)})|.
\]
Thus

\[
\text{Var}\left(\sum_{i=1}^{n} u_i \right) \leq nC \max_i E|\hat{\mathcal{M}}_{i,A}\|t_k - \omega_{n0}^{(1)}\| = nO(\sqrt{q/n})O(\sqrt{q/n})
\]

\[
= O(\sqrt{q/n}) = o(n\lambda)
\]

under the condition \(q_n n^{-1/2} = o(\lambda)\). By Conditions (C1) and (C3), we have

\[
\max_{i,j} |\hat{\mathcal{M}}_{ij}| [I(\kappa_i(t_k) \leq 0) - I(\kappa_i(\omega_{n0}^{(1)}) \leq 0)] \leq C.
\]

Applying Bernstein’s inequality,

\[
I_{nj1} \leq N \exp\left(-\frac{n^2 \lambda^2 / 4}{2\sqrt{nq/n} + Cn\lambda}\right) \leq N \exp(-Cn\lambda) \leq C \exp(4q_n \log n - Cn\lambda).
\] (A.30)

Now, we compute \(I_{nj2}\), note that \(I(x \leq s)\) is an increasing function about \(s\). Therefore,

\[
\left| -P(\kappa_i(\hat{\omega}_n^{(1)}) \leq 0) + P(\kappa_i(t_k) \leq 0) \right|
\]

\[
\leq \sum_{i=1}^{n} |\hat{\mathcal{M}}_{ij}| [I(\kappa_i(t_k) \leq \Delta \sqrt{q/n} n^5 ||\hat{\mathcal{M}}_{i,A}||) - I(\kappa_i(t_k) \leq 0)]
\]

\[
- P(\kappa_i(t_k) \leq \Delta \sqrt{q/n} n^5 ||\hat{\mathcal{M}}_{i,A}||) + P(\kappa_i(t_k) \leq 0)]
\]

\[
= \sum_{i=1}^{n} |\hat{\mathcal{M}}_{ij}| [I(\kappa_i(t_k) \leq \Delta \sqrt{q/n} n^5 ||\hat{\mathcal{M}}_{i,A}||) - I(\kappa_i(t_k) \leq 0)]
\]

\[
- P(\kappa_i(t_k) \leq \Delta \sqrt{q/n} n^5 ||\hat{\mathcal{M}}_{i,A}||) + P(\kappa_i(t_k) \leq 0)]
\]

\[
+ \sum_{i=1}^{n} |\hat{\mathcal{M}}_{ij}| [P(\kappa_i(t_k) \leq \Delta \sqrt{q/n} n^5 ||\hat{\mathcal{M}}_{i,A}||) - P(\kappa_i(t_k) \leq -\Delta \sqrt{q/n} n^5 ||\hat{\mathcal{M}}_{i,A}||)].
\] (A.31)

Note that

\[
\sum_{i=1}^{n} |\hat{\mathcal{M}}_{ij}| [P(\kappa_i(t_k) \leq \Delta \sqrt{q/n} n^5 ||\hat{\mathcal{M}}_{i,A}||) - P(\kappa_i(t_k) \leq -\Delta \sqrt{q/n} n^5 ||\hat{\mathcal{M}}_{i,A}||)]
\]

\[
= \sum_{i=1}^{n} |\hat{\mathcal{M}}_{ij}| [F_i((\Delta \sqrt{q/n} n^5 ||\hat{\mathcal{M}}_{i,A}|| + \hat{\mathcal{M}}_{i,A}^T (t_k - \omega_{n0}^{(1)} + \Delta_i)|X_i, U_i)
\]

\[
- F_i((-\Delta \sqrt{q/n} n^5 ||\hat{\mathcal{M}}_{i,A}|| + \hat{\mathcal{M}}_{i,A}^T (t_k - \omega_{n0}^{(1)} + \Delta_i)|X_i, U_i)]
\]

\[
\leq C \sum_{i=1}^{n} |\hat{\mathcal{M}}_{ij}| \sqrt{q/n} n^5 ||\hat{\mathcal{M}}_{i,A}|| \leq Cn \sqrt{q/n} n^5 \sqrt{q/n}
\]

\[
= Cn \sqrt{q/n} n^{-3/2} = o(n\lambda).
\] (A.32)

by Condition (C7) and \(n\lambda \to \infty\) Thus, by (A.31) and (A.32)

\[
I_{nj2} \leq \sum_{k=1}^{N} P\left(\sum_{i=1}^{n} |\hat{\mathcal{M}}_{ij}| [I(\kappa_i(t_k) \leq \Delta \sqrt{q/n} n^5 ||\hat{\mathcal{M}}_{i,A}||) - I(\kappa_i(t_k) \leq 0)]
\]

\[
- P(\kappa_i(t_k) \leq \Delta \sqrt{q/n} n^5 ||\hat{\mathcal{M}}_{i,A}||) + P(\kappa_i(t_k) \leq 0)] \geq n\lambda/2
\]
we have max

Thus (A.34) follows from the condition

So we have Var(

Proof of Proposition

Proof of (ii), it suffices to show that

and

|\kappa_i(t_k)| \leq \Delta \sqrt{q_n/n^5} ||\hat{\kappa}_i|| - I(\kappa_i \leq 0))

Let



Theorem 2.2 (i) and Conditions (C8). Then, by Theorem 2.2 (i) and Conditions (C7) and (C8), we have

Finally, by (A.30) and (A.33), we have that the probability in (A.29) is bounded by

under the assumptions

This completes the proof.

Proof of Proposition 3.1 Proof of (i). Since the unpenalized quantile loss objective function is convex. By the convex optimization theory, \(0 \in \partial \sum_{i=1}^{n} \rho_{\tau}(Y_i - \hat{\kappa}_i^{(1)})\). Therefore there exists v_i^* such that

s_j(\hat{\omega}_n) = 0 with v_i = v_i^* for \(j = 0, 1, \ldots, q_n\), and (i) holds.

Proof of (ii), it suffices to show that

as \(n \to \infty\). Note that

\[\min_{1 \leq j \leq q_n} |\hat{\omega}_j| \geq \min_{1 \leq j \leq q_n} |\omega_0|, \quad \max_{1 \leq j \leq q_n} |\hat{\omega}_j - \omega_0j| \geq Mn^{-1/(c_2-2)}\]

by Condition (C8). Then, by Theorem 2.2 (i) and Conditions (C7) and (C8), we have

Thus (A.34) follows from the condition \(\lambda = o(n^{-1/(c_2-2)})\).
Proof of (iii), we only need to show that
\[ P\{|s_j(\hat{\omega}_n)| > \lambda, \text{ for some } j = q_n + 1, \ldots, p_n \} \to 0 \quad (A.35) \]
as \( n \to \infty \). Let \( D = \{ i : Y_i - \hat{M}_{iA}^T \hat{\omega}_n^{(1)} = 0 \} \), then for \( j = q_n + 1, \ldots, p_n \)
\[ s_j(\hat{\omega}_n) = n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} [I(Y_i - \hat{M}_{iA}^T \hat{\omega}_n^{(1)} \leq 0) - \tau] - n^{-1} \sum_{i \in D} \hat{M}_{ij} v_i^* , \]
where \( v_i^* \in [\tau - 1, \tau] \) with \( i \in D \) satisfies \( s_j(\hat{\omega}_n) = 0 \) for \( j = 0, 1, \ldots, q_n \), when \( v_i = v_i^* \). Similar to the proof of Lemma 2.3 in [44], with probability one, we have
\[ \max_{q_n + 1 \leq j \leq p_n} \left| n^{-1} \sum_{i \in D} \hat{M}_{ij} v_i^* \right| = O(q_n/n) = o(\lambda) \]
under Condition (C1) and \( q_n n^{-1/2} = o(\lambda) \). Thus to prove (A.35), it suffices to show that
\[ P\left\{ n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} [I(Y_i - \hat{M}_{iA}^T \hat{\omega}_n^{(1)} \leq 0) - \tau] > \lambda, \text{ for some } j = q_n + 1, \ldots, p_n \right\} \to 0. \quad (A.36) \]
Note that
\[
P\left\{ \max_{q_n + 1 \leq j \leq p_n} \left| n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} [I(Y_i - \hat{M}_{iA}^T \hat{\omega}_n^{(1)} \leq 0) - \tau] \right| > \lambda \right\}
\leq P\left\{ \max_{q_n + 1 \leq j \leq p_n} \left| n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} [I(Y_i - \hat{M}_{iA}^T \hat{\omega}_n^{(1)} \leq 0) - I(Y_i - \hat{M}_{iA}^T \hat{\omega}_n^{(1)} \leq 0)] \right| > \lambda/2 \right\}
+ P\left\{ \max_{q_n + 1 \leq j \leq p_n} \left| n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} [I(Y_i - \hat{M}_{iA}^T \hat{\omega}_n^{(1)} \leq 0) - \tau] \right| > \lambda/2 \right\}
\leq P\left\{ \max_{q_n + 1 \leq j \leq p_n} \left| n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} [I(Y_i - \hat{M}_{iA}^T \hat{\omega}_n^{(1)} \leq 0) - I(Y_i - \hat{M}_{iA}^T \hat{\omega}_n^{(1)} \leq 0)] \right| > \lambda/2 \right\} + o_p(1)
\leq P\left\{ \max_{q_n + 1 \leq j \leq p_n} \sup_{\|\omega_n^{(1)} - \omega_n^{(1)}\| \leq \Delta \sqrt{q_n/n}} \left| n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} [I(Y_i - \hat{M}_{iA}^T \hat{\omega}_n^{(1)} \leq 0)
- I(Y_i - \hat{M}_{iA}^T \hat{\omega}_n^{(1)} \leq 0)] \right| > \lambda/4 \right\} + o_p(1)
+ P\left\{ \max_{q_n + 1 \leq j \leq p_n} \sup_{\|\omega_n^{(1)} - \omega_n^{(1)}\| \leq \Delta \sqrt{q_n/n}} \left| n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} [P(Y_i - \hat{M}_{iA}^T \hat{\omega}_n^{(1)} \leq 0)
- P(Y_i - \hat{M}_{iA}^T \hat{\omega}_n^{(1)} \leq 0)] \right| > \lambda/4 \right\} + o_p(1)
\]
where the second inequality follows from Lemma 6.2, and the last inequality follows from Lemma 6.3. Note that
\[
\max_{q_n + 1 \leq j \leq p_n} \sup_{\|\omega_n^{(1)} - \omega_n^{(1)}\| \leq \Delta \sqrt{q_n/n}} \left| n^{-1} \sum_{i=1}^{n} \hat{M}_{ij} [P(Y_i - \hat{M}_{iA}^T \hat{\omega}_n^{(1)} \leq 0) - P(Y_i - \hat{M}_{iA}^T \hat{\omega}_n^{(1)} \leq 0)] \right|
\]
Therefore, for sufficiently large $n$ by Proposition 3.1 (ii). Therefore, $P_l$ and $l$ and $1200$

This proves (A.36).

Proof of Theorem 3.2 For $g_j \in \partial g(\hat{\omega}_n)$, $g_j = s_j(\hat{\omega}_n) + \lambda l_j$ for $j = 0, 1, \ldots, p_n$, where $l_0 = 0$; for $1 \leq j \leq p_n$, $l_j = \text{sgn}(\omega_j)$ if $\omega_j \neq 0$ and $l_j \in [-1, 1]$ otherwise. By Proposition 3.1 (i), there exist $v_i, i = 1, \ldots, n$, such that the subgradient function $s_j(\hat{\omega}_n)$ defined with $v_i = v_i^*$ satisfies $P\{s_j(\hat{\omega}_n) = 0, j = 0, 1, \ldots, q_n\} \to 1$. Therefore, by the definition of the set $\partial g(\hat{\omega}_n)$, we have $P\{G \subseteq \partial g(\hat{\omega}_n)\} \to 1$ where 

$$G = \{\xi = (\xi_0, \xi_1, \ldots, \xi_{p_n})^T : \xi_0 = 0; \xi_j = \lambda \text{sgn}(\hat{\omega}_j), j = 1, \ldots, q_n;$$

and $l_j$ ranges over $[-1, 1]$ for $j = q_n + 1, \ldots, p_n$.

Consider any $\omega_n$ in a ball in $R^{p_n+1}$ with the center $\hat{\omega}_n$ and radius $\lambda/2$. By Lemma 6.1, to prove the theorem it is sufficient to show that there exists a vector $\xi^* = (\xi_0^*, \xi_1^*, \ldots, \xi_{p_n}^*)^T \in G$ such that

$$P\left\{\xi_j^* = \frac{\partial h(\omega_n)}{\partial \omega_j}, j = 0, 1, \ldots, p_n\right\} \to 1, \quad n \to \infty.$$  \hspace{1cm} (A.37)

By Proposition 3.1 (iii), we have $P\{|s_j(\hat{\omega}_n)| \leq \lambda, j = q_n + 1, \ldots, p_n\} \to 1$. Thus, we can always find $l_j^* \in [-1, 1]$ such that $s_j(\hat{\omega}_n) + \lambda l_j^* = 0$ for $j = q_n + 1, \ldots, p_n$. Let $\xi^*$ be the vector in $G$ with $l_j = l_j^*, j = q_n + 1, \ldots, p_n$. We next verify that $\xi^*$ satisfies (A.37).

For $j = 0$, we have $\xi_0^* = 0$. Since $\frac{\partial h(\omega_n)}{\partial \omega_0} = 0$, we have $\frac{\partial h(\omega_n)}{\partial \omega_j} = \xi_0^*.$

For $j = 1, \ldots, q_n$, we have $\xi_j^* = \lambda \text{sgn}(\hat{\omega}_j)$ for $\omega_j \neq 0$. Note that $\min_{1 \leq j \leq q_n} |\omega_j| \geq \min_{1 \leq j \leq q_n} |\hat{\omega}_j - \omega_j| \geq (a + 1/2)\lambda - \lambda/2 = a\lambda$ with probability approaching one by Proposition 3.1 (ii). Therefore, $P\{\frac{\partial h(\omega_n)}{\partial \omega_j} = \lambda \text{sgn}(\omega_j), j = 1, \ldots, q_n\} \to 1$. For any $j = 1, \ldots, q_n, |\hat{\omega}_j - \omega_j| \leq ||\omega_n^{(1)} - \omega_n^{(1)}|| = O(\sqrt{q_n/n}) = o(\lambda)$ by condition $n^{-1/2}q_n = o(\lambda)$. Therefore, for sufficiently large $n, \hat{\omega}_j$ and $\omega_j$ have the same sign. This implies

$$P\left\{\xi_j^* = \frac{\partial h(\omega_n)}{\partial \omega_j}, j = 1, \ldots, p_n\right\} \to 1$$

as $n \to \infty$.

For $j = q_n + 1, \ldots, p_n$, we have $\xi_j^* = s_j(\hat{\omega}_n) + \lambda l_j^* = 0$ following the definition of $\xi^*$. By definition of the oracle estimator, $\hat{\omega}_j = 0$ for $j = q_n + 1, \ldots, p_n$. Therefore, $|\omega_j| \leq |\hat{\omega}_j| +
\[|\hat{\omega}_j - \omega_j| \leq \lambda/2. \] For \(|\omega_j| < \lambda\), we have \(\partial h(\omega_n)/\partial \omega_j = 0\), \(j = q_n + 1, \ldots, p_n\) by the definition of \(h(\omega_n)\). Therefore, \(P\{\xi_j = \frac{\partial h(\omega_n)}{\partial \omega_j}, j = 1, \ldots, p_n\} \to 1\) as \(n \to \infty\). This completes the proof. □

Acknowledgements We appreciate the constructive suggestions from the referees and the editors.

References

[1] Ando, T., Li, K.-C.: A model-averaging approach for high-dimensional regression. J. Amer. Statist. Assoc., 109, 1–46 (2014)
[2] Ando, T., Li, K.-C.: A weight-relaxed model averaging approach for high-dimensional generalized linear models. Ann. Statist., 45, 2654–2679 (2017)
[3] Bradic, J., Fan, J., Wang, W.: Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. J. R. Stat. Soc. Ser. B Stat. Methodol., 73, 325–349 (2011)
[4] Chen, J., Li, D., Linton, O., et al.: Semiparametric dynamic portfolio choice with multiple conditioning variables. J. Econometrics, 194, 309–318 (2016)
[5] Chen, J., Li, D., Linton, O., et al.: Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. J. Amer. Statist. Assoc., 113, 919–932 (2018)
[6] Cheng, M., Honda, T., Zhang, J.: Forward variable selection for sparse ultra-high dimensional varying coefficient models. J. Amer. Statist. Assoc., 111, 1209–1221 (2016)
[7] Fan, J., Fan, Y., Barut, E.: Adaptive robust variable selection. Ann. Statist., 42, 324–351 (2014)
[8] Fan, J., Gijbels, I.: Local Polynomial Modelling and Its Applications, Chapman & Hall, London, 1996
[9] Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc., 96, 1348–1360 (2001)
[10] Fan, J., Lv, J.: A selective overview of variable selection in high dimensional feature space. Statist. Sinica, 20, 101–148 (2010)
[11] Fan, J., Lv, J.: Nonconcave penalized likelihood with NP-dimensionality. IEEE T. Inform. Theory, 57, 5467–5484 (2011)
[12] Fan, J., Ma, Y., Dai, W.: Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. J. Amer. Statist. Assoc., 109, 1270–1284 (2014)
[13] Fan, J., Peng, H.: On non-concave penalized likelihood with diverging number of parameters. Ann. Statist., 32, 928–961 (2004)
[14] Frumento, P., Bottai, M.: Parametric modeling of quantile regression coefficient functions. Biometrics, 72, 74–84 (2016)
[15] Hansen, B. E.: Least squares model averaging. Econometrica, 75, 1175–1189 (2007)
[16] Hastie, T., Tibshirani, R.: Varying-coefficient model. J. R. Stat. Soc. Ser. B Stat. Methodol., 55, 757–796 (1993)
[17] Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: the Lasso and Generalizations, Chemical Rubber Company Press, Taylor & Francis Group, 2015
[18] He, X., Wang, L., Hong, H. G.: Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. Ann. Statist., 41, 342–369 (2013)
[19] Hu, T., Xia, Y.: Adaptive semi-varying coefficient model selection. Statist. Sinica, 22, 575–599 (2012)
[20] Hunter, D., Lange, K.: Quantile regression via an MM algorithm. J. Comput. Graph. Statist., 9, 60–77 (2000)
[21] Jiang, J., Zhao, Q., Hui, Y. V.: Robust modelling of ARCH models. J. Forecasting, 20, 111–133 (2001)
[22] Kai, B., Li, R., Zou, H.: New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. Ann. Statist., 39, 305–332 (2011)
[23] Kim, Y., Choi, H., Oh, H. S.: Smoothly clipped absolute deviation on high dimensions. J. Amer. Statist. Assoc., 103, 1665–1673 (2008)
[24] Knight, K.: Limiting distributions for \(L_1\) regression estimators under general conditions. Ann. Statist., 26, 755–770 (1998)
[25] Koenker, R.: Quantile Regression, Cambridge University Press, Cambridge, 2005
[26] Lee, E. R., Noh, H., Park, B. U.: Model selection via Bayesian information criterion for quantile regression models. *J. Amer. Statist. Assoc.*, 109, 216–229 (2014)
[27] Li, C., Li, Q., Racine, J., et al.: Optimal model averaging of varying coefficient models. *Statist. Sinica*, 28, 2795–2809 (2018)
[28] Li, D., Linton, O., Lu, Z.: A flexible semiparametric forecasting model for time series. *J. Econometrics*, 187, 345–357 (2015)
[29] Li, J., Xia, X., Wong, W., et al.: Varying-coefficient semiparametric model averaging prediction. *Biometrics*, 74, 1417–1426 (2018)
[30] Li, X., Ma, X., Zhang, J.: Conditional quantile correlation screening procedure for ultrahigh-dimensional varying coefficient models. *J. Statist. Plann. Inference*, 197, 69–92 (2018)
[31] Liang, H., Zou, G., Wan, A.T.K., et al.: Optimal weight choice for frequentist model average estimators. *J. Amer. Statist. Assoc.*, 106, 1053–1066 (2011)
[32] Lin, B, Wang, Q., Zhang, J., et al.: Stable prediction in high-dimensional linear models. *Stat. Comput.*, 27, 1401–1412 (2017)
[33] Lu, X., Su, L.: Jackknife model averaging for quantile regressions. *J. Econometrics*, 188, 40–58 (2015)
[34] Ma, S., He, X.: Inference for single-index quantile regression models with profile optimization. *Ann. Statist.*, 44, 1234–1268 (2016)
[35] Ma, S., Li, R., Tsai, C.-L.: Variable screening via quantile partial correlation. *J. Amer. Statist. Assoc.*, 112, 650–663 (2017)
[36] Noh, H., Chung, K., Keilegom, I.: Variable selection of varying coefficient models in quantile regression. *Electron. J. Stat.*, 6, 1220–1238 (2012)
[37] Parzen, E.: On estimation of a probability density function and model. *Ann. Statist.*, 33, 1065–1076 (1962)
[38] Pollard, D.: Asymptotics for least absolute deviation regression estimators. *Economet. Theor.*, 7, 186–199 (1991)
[39] Sherwood, B., Wang, L.: Partially linear additive quantile regression in ultra-high dimension. *Ann. Statist.*, 44, 288–317 (2016)
[40] Tang, Y., Song, X., Wang, H., et al.: Variable selection in high-dimensional quantile varying coefficient models. *J. Multivariate Anal.*, 122, 115–132 (2013)
[41] Tao, P. D., An, L. T. H.: Convex analysis approach to d.c. programming: Theory, algorithms and applications. *Acta Math. Vietnam.*, 22, 289–355 (1997)
[42] Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 58, 267–288 (1996)
[43] Wang, H., Xia, Y.: Shrinkage estimation of the varying coefficient model. *J. Amer. Statist. Assoc.*, 104, 747–757 (2009)
[44] Wang, L., Wu, Y., Li, R.: Quantile regression for analyzing heterogeneity in ultra-high dimension. *J. Amer. Statist. Assoc.*, 107, 214–222 (2012)
[45] Welsh, A. H.: On M-processes and M-estimation. *Ann. Statist.*, 17, 337–361 (1989)
[46] Xia, X., Li, J., Fu, B.: Conditional quantile correlation learning for ultrahigh dimensional varying coefficient models and its application in survival analysis. *Statist. Sinica*, 29, 645–669 (2019)
[47] Xue, L., Qu, A.: Variable selection in high-dimensional varying-coefficient models with global optimality. *J. Mach. Learn. Res.*, 13, 1973–1998 (2012)
[48] Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68, 49–67 (2006)
[49] Zhang, C. H.: Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38, 894–942 (2010)
[50] Zhao, W., Jiang, X., Lian, H.: A principal varying-coefficient model for quantile regression: Joint variable selection and dimension reduction. *Comput. Statist. Data Anal.*, 127, 269–280 (2018)
[51] Zhu, L. P., Li, L., Li, R., et al.: Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.*, 106, 1464–1475 (2011)
[52] Zhu, R., Wan, T. K., Zhang, X., et al.: A Mallows-type model averaging estimator for the varying-coefficient partially linear model. *J. Amer. Statist. Assoc.*, 114, 882–892 (2019)
[53] Zou, H.: The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101, 1418–1429 (2006)