Penalized Model Averaging for High Dimensional Quantile Regressions*†

Haowen Bao[a,b,c], Zongwu Cai[d], Yuying Sun[a,b,c,†], Shouyang Wang[a,b,c,e]

[a]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China

[b]Center for Forecasting Science, Chinese Academy of Sciences, China

[c]School of Economics and Management, University of Chinese Academy of Sciences, China

[d]Department of Economics, University of Kansas, USA

[d]School of Entrepreneurship and Management, ShanghaiTech University, China

Abstract: This paper proposes a new penalized model averaging method for high dimensional quantile regressions based on quasi-maximum likelihood estimation, which determines optimal combination weights and yields sparseness from various potential covariates simultaneously. The proposed weight choice criterion is based on the Kullback-Leibler loss with penalties, which could reduce to Mallows-type criterion for asymmetric Laplace density, and both the dimension of covariates and the number of possibly misspecified candidate models are allowed to be diverging with the sample size. Also, the asymptotic optimality and convergence rate of the selected weights are derived under time series framework, even when all candidate models are misspecified. We further extend our concern to the ultra-high dimensional scenarios and establish the corresponding asymptotic optimality. Simulation studies and empirical application to stock returns forecasting illustrate that the proposed method outperforms existing methods.

Keywords: High-dimensional quantile regressions; Optimality; Consistency; Kullback-Leibler loss; Model averaging

JEL Classification: C32, C52.

*Contact information: Haowen Bao (bhw@amss.ac.cn); Zongwu Cai (caiz@ku.edu); Yuying Sun (sunyuying@amss.ac.cn; Shouyang Wang (sywang@amss.ac.cn).

†*Address for correspondence:* Yuying Sun, Center for Forecasting Science, Chinese Academy of Sciences, Beijing, Beijing 100190, China.

1 Introduction

Quantile regression has been widely recognized as a pivotal data analysis technique in many applied fields such as biology, ecology, economics, environmental sciences, finance, medical sciences, and psychology. In contrast to the conventional mean regressions that concentrate on the conditional expectation of the model, quantile regressions pay more attention to the distribution of the response variables. Especially, quantile regression model is much more suitable to capture the structural features of the data when it follows a skewed or heavy tailed distribution. As a result, modeling conditional quantiles has a wide range of applications in many applied fields. For example, in finance and economics, value-at risk (VaR) is defined as the quantile of future portfolio values conditional on current information, which has become a standard measure of market risk as in Engle and Manganelli (2004). Besides, VaR and other distribution information, such as variance, skewness, expected shortfall of returns and other variables in financial market, appeal to many investors and decision makers as argued in Linton and Xiao (2017). In genetic and genomic sciences, Briollais and Durrieu (2014) provided a comprehensive review on real applications of quantile regression to the fields of genetic and the emerging genomic studies and emphasized the motivation and benefits for using quantile regression in genetic and genomic applications. Finally, in ecology and the environmental sciences, quantile regression can provide a more complete view of possible causal relationships between variables in ecological processes as described in Cade and Noon (2003).

Since the seminal framework of quantile regression, proposed by Koenker and Bassett (1978), there has been a large body of literature on estimating quantile regression models and examining their econometric properties and applications, to name just a few, for example, Cai (2002), Koenker and Xiao (2004), Koenker and Xiao (2006), Honda (2013), Bertsimas and Mazumder (2014), Chen and Wang (2023), Cai et al. (2023), and references therein. Especially, Xiao and Koenker (2009) studied the conditional quantiles estimation for the GARCH type model and proposed a two-step approach of quantile regression estimation for

linear GARCH time series, while Cai and Xu (2009) proposed nonparametric quantile estimations for dynamic smooth coefficient models. Furthermore, Cai and Xiao (2012) extended quantile regression to dynamic models with partially varying coefficients and estimated both parametric and nonparametric functional coefficients. More recently, Nguyen et al. (2020) applied a LASSO based quantile regression to investigate the tail-risk dependence in the cryptocurrency markets.

With the advent of sophisticated information systems to collect a huge amount of data, we face a large number of candidate conditional quantile models. A common approach to deal with this issue is to apply a model selection approach by determining an optimal model for prediction. There are several model selection strategies available for quantile regression in the literature, like Bayesian information criterion in Lee et al. (2014), composite quantile regression with adaptive LASSO studied by Zou and Yuan (2008), weighted composite quantile regression with the smoothly clipped absolute deviation (SCAD) penalty proposed in Jiang et al. (2012), and the Schwarz-like criterion as in Koenker (2011). However, model selection may ignore some useful information, as addressed by Liao et al. (2019), and may consequently yield poor forecast performance.

Different from model selection, model averaging provides a weighted average of all potential candidate models, which is a sensible approach to reducing model uncertainty. This strategy avoids to select a very poor candidate model and usually leads to a lower risk of model misspecification as elaborated by Li et al. (2022). There are two main classes of model averaging methods, i.e., Bayesian model averaging (BMA) and frequentist model averaging (FMA). Given prior information, BMA assigns weights to the models by posterior probability; see more details in Hoeting et al. (1999) and Brown et al. (2002). On the other hand, FMA focuses on determining the weights of candidate models from frequentist perspective. Most literature on FMA focuses on the strategies of determining the weights and the asymptotic properties of estimators. For example, various weight choice schemes have been proposed, like Mallows criterion as in Hansen (2007) and Wan et al. (2010), mean squared error minimization in Liang et al. (2011), leave-one-out cross-validation procedure

used by Hansen and Racine (2012), Kullback-Leibler (KL) type measures employed by Zhang et al. (2016), weight choice criterion for linear measurement error models by Zhang et al. (2019), time-varying local jackknife criterion proposed by Sun et al. (2021), and penalized leave-h-out forward validation criterion investigated by Sun et al. (2023). Recently, for ultra-high dimensional regressions with continuous responses, Ando and Li (2014, 2017) considered a two-step model averaging approach with model screening in the first step. For divergent-dimensional varying-coefficient multinomial logistic models, Li et al. (2022) proposed a least squares criterion with the AdaBoost algorithm. Besides modeling the conventional point-valued time series, Sun et al. (2022) proposed a model averaging procedure for interval-valued data and proved that this procedure yielded predictors of mid-points and ranges with an optimally asymptotic property.

However, the aforementioned literature focuses basically on the model averaging estimation for conditional mean models. To the best of our knowledge, there are only two strategies designed explicitly for quantile regression model averaging. First, Lu and Su (2015) constructed a jackknife model averaging (JMA) scheme for quantile regression with low dimensional covariates, while Wang et al. (2023) further extended JMA to a two-step process to accommodate high-dimensional quantile regression, where the marginal quantile utility is used to screen the covariates and construct candidate models in the first step. However, there are still several unsolved issues. First, model averaging can be considered as a mega model with a high-dimensional set of predictors, some of which may be highly multi-collinear to various extents or even redundant. But, the aforementioned approaches are unable to eliminate these redundant predictors, which may result in unstable forecasts. Second, the weight choice criteria of these approaches are essentially based on the classical quantile regression estimators, which is equivalent to a special case of quasi likelihood estimation (QMLE) with asymmetric Laplace density. Furthermore, these literature only focused on the asymptotic optimality of the estimated weights, while whether the selected weights is asymptotically consistent is unaddressed. Thus, it is highly desirable to construct a parsimonious and general model averaging method for quantile regression to improve forecast

accuracy and select important predictors simultaneously.

Our attempt in this paper is to propose a new model averaging strategy for high-dimensional quantile regressions, for which we immediately face three challenges in contrast to the existing literature. First, we need to develop a more general and parsimonious weight choice criterion, which selects optimal combination weights and yields sparseness from various potential covariates simultaneously. It has been shown that the aforementioned weight choice criteria for quantile regressions focus on the selected weights and ignore deleting redundant predictors. Also, the proposed criterion could reduce to some classic criteria in the existing literature such as Lu and Su (2015). Second, it is expected to prove the resultant model averaging estimators are asymptotically optimal when both the number of candidate models and the number of predictors are divergent. Note that considering the high-dimensional candidate models and relaxing the the conventional assumption of the weights summing to one enhance the difficulty of proofs. Furthermore, this article is the first to prove the asymptotic consistency of model averaging estimators for quantile regression, which fills the gap in the quantile regression literature. Also, note that under time series framework, the proof of the asymptotic consistency is much more difficult than the situation when the model is correctly specified, where a commonly used tool is Theorem 1 of Fan and Peng (2004). Specifically, the converging rate of this theorem can not been applied directly due to the misspecified candidate models and the positive restriction of weights.

To address the first challenge, we construct a weight choice criterion based on the KL loss with penalties, which provides a parsimonious model by obtaining the estimated weights and selecting powerful predictors simultaneously. Moreover, it is worth noting that our criterion could reduce to a Mallows-type criterion as in Lu and Su (2015)[1] for asymmetric Laplace density. To obtain the asymptotic optimality of the estimated weights, Lemma 1 later provides the rate of estimators in candidate models with a diverging number of

[1]Lu and Su (2015) also proposed a jackknife model averaging (JMA) for quantile regression. If the leave-one-out cross-validation method is introduced to our QMLE-based weight choice criterion with the asymmetric Laplace density, the proposed criterion in this paper also reduces to the JMA for quantile regression as in Lu and Su (2015).

predictors converging to the well-defined limits. Also, we provide some standard assumptions to indicate the relationships among the number of candidate models, the dimension of covariates, the sample size, and the tuning parameters, which is helpful to deal with the problems caused by high dimension and the relaxation of $\sum_{m=1}^{M} w_m = 1$, where $\{w_m\}_{m=1}^{M}$ are the weights. Finally, we extend the consistent theory developed by Komunjer (2005) and Fan and Peng (2004) to show that the model averaging estimators for high-dimensional quantile regression are asymptotically consistent whether the model is correctly specified or misspecified.

Besides, we also extend our methodology to the ultra-high dimensional scenarios by introducing a model screening process before model averaging, and the related asymptotic optimality is established. Furthermore, simulation studies are conducted to investigate the finite sample properties of our method under both homoscedastic and heteroscedastic settings, to demonstrate that the proposed method is promising. Finally, an empirical application to forecast S&P 500 stock returns in comparison with existing methods is examined to highlight the merits of our proposed method.

The remainder of this paper is organized as follows. Section 2 introduces the model averaging estimation. Section 3 derives the asymptotic theories of the proposed method. Section 4 extends our method to ultra-high dimensional scenarios and shows the corresponding asymptotic optimality. Section 5 gives the simulation. Section 6 applies our method to the empirical case. Section 7 concludes the paper. Mathematical proofs are relegated to Appendix C.

2 Model Averaging Estimation

2.1 Model Framework

Let $\{y_i, \mathbf{x}_i\}_{i=1}^{n}$ be a random sample (a stationary process), where y_i is a scalar dependent variable and $\mathbf{x}_i = (x_{i1}, x_{i2}, ...)'$ is a covariate vector with countable infinite dimension. Here,

$\mathbf{x}_i$ is allowed to be a finite dimensional vector, say, a $p \times 1$ vector of all possible predictors. The τth $(0 < \tau < 1)$ conditional quantile of y_i given $\mathbf{x}_i$ is defined by

$$q_\tau(y_i|\mathbf{x}_i) = \mu_{i,\tau} = \sum_{j=1}^{\infty} \theta_{j,\tau} x_{ij}, \tag{1}$$

where $\theta_{j,\tau}$ is an unknown coefficient. Alternatively, (1) can be re-expressed as

$$y_i = \sum_{j=1}^{\infty} \theta_{j,\tau} x_{ij} + \varepsilon_{i,\tau}, \tag{2}$$

where $\varepsilon_{i,\tau} = y_i - \mu_{i,\tau}$ is an unobservable error satisfying $P(\varepsilon_{i,\tau} \leq 0|\mathbf{x}_i) = \tau$.

In this paper, we consider M candidate models to estimate the quantile regression model (2), and M is allowed to be diverging as the sample size goes to infinity. Suppose there are k_m regressors contained in the mth candidate model, where k_m is also allowed to grow to infinity at some slower rates than the sample size n. As a result, the mth candidate model is expressed as

$$y_i = \mu_{i,\tau}^{(m)} + \varepsilon_{i,\tau}^{(m)} = \sum_{j=1}^{k_m} \theta_{j,\tau}^{(m)} x_{ij(m)} + \varepsilon_{i,\tau}^{(m)} = \boldsymbol{\theta}_\tau^{(m)\prime} \mathbf{x}_{i(m)} + \varepsilon_{i,\tau}^{(m)},$$

where $\mu_{i,\tau}^{(m)}$ is the conditional quantile of y_i given $\mathbf{x}_{i(m)}$, $\boldsymbol{\theta}_\tau^{(m)} = (\theta_{1,\tau}^{(m)}, \cdots, \theta_{k_m,\tau}^{(m)})'$ is a $k_m \times 1$ vector of coefficients, $\mathbf{x}_{i(m)} = (x_{i1(m)}, \cdots, x_{ik_m(m)})'$ is a k_m-dimensional sub-vector of $\mathbf{x}_i$, and $\varepsilon_{i,\tau}^{(m)} = y_i - \boldsymbol{\theta}_\tau^{(m)\prime} \mathbf{x}_{i(m)}$ is the error term of the mth candidate model.

To estimate each candidate quantile regression model, the estimator of $\boldsymbol{\theta}_\tau^{(m)}$ is given by the tick-exponential family, termed as QMLE as in Komunjer (2005), denoted as $\widehat{\boldsymbol{\theta}}_\tau^{(m)}$. For the mth candidate model, suppose that the true conditional cumulative distribution function (cdf) of y_i is $F_0(y|\mathbf{x}_{i(m)})$, and its probability density function (pdf) is denoted by $f_0(y|\mathbf{x}_{i(m)}) \equiv dF_0/dy$. Notice that $F_0(\cdot)$ and $f_0(\cdot)$ are usually unknown.

Remark 1. *The definition of quantile implies that the conditional quantile function of y_i given $\mathbf{x}_{i(m)}$ and the cumulative distribution function (cdf) are inverse of each other, i.e., $q_\tau(y_i|\mathbf{x}_{i(m)}) = F^{-1}_{y_i|\mathbf{x}_{i(m)}}(\tau)$. Let $\{q_\tau(\mathbf{x}_{i(m)}, \boldsymbol{\theta}_\tau^{(m)})\}$ denote a model for the τth quantile of y_i given*

$\mathbf{x}_{i(m)}$, *and* $q_\tau(\mathbf{x}_{i(m)}, \boldsymbol{\theta}_\tau^{(m)})$ *is allowed to be nonlinear. To estimate the parameters by QMLE, Komunjer (2005) indicated that we could assume the density function* $f(y_i, q_\tau(\mathbf{x}_{i(m)}, \boldsymbol{\theta}_\tau^{(m)}), \tau)$ *belongs to the tick-exponential family of densities, which leads to the estimator* $\widehat{\boldsymbol{\theta}}_\tau^{(m)}$ *with a well-defined limit.*

Then, the logarithm likelihood function of the mth candidate model can be expressed as

$$L_n(\boldsymbol{\theta}_\tau^{(m)}) = \sum_{i=1}^{n} \ln f(y_i, q_\tau(\mathbf{x}_{i(m)}, \boldsymbol{\theta}_\tau^{(m)}), \tau), \tag{3}$$

and $f(\cdot)$ takes the following form

$$f(y, \eta, \tau) = \exp\{-(1-\tau)[a(\eta) - b(y)]\mathbf{1}(y \le \eta) + \tau[a(\eta) - c(y)]\mathbf{1}(y > \eta)\}, \tag{4}$$

where $a(\eta)$ is continuously differentiable, $b(y)$ and $c(y)$ are continuous, and $a(\eta)$, $b(\eta)$ and $c(\eta)$ are such functions to satisfy that $\int_{-\infty}^{\infty} f(y, \eta, \tau)dy = 1$ and $\int_{-\infty}^{\eta} f(y, \eta, \tau)dy = \tau$. A common member of the tick-exponential family densities is the asymmetric Laplace density (see more details in Section 2.4), i.e.,

$$f(y, \mu, \tau) = \tau(1-\tau)\exp\{(1-\tau)(y-\mu)\mathbf{1}(y \le \mu) - \tau(y-\mu)\mathbf{1}(y > \mu)\}.$$

Besides, for a given $\alpha \in \mathbb{N}^*$, let $a(\eta) = \frac{1}{\tau(1-\tau)}\text{sgn}(\eta)\ln(1+|\eta|^\alpha)$ and $b(y) = c(y) = \frac{1}{\tau(1-\tau)}\text{sgn}(y)\ln(1+|y|^\alpha)$. Then, (3) becomes

$$\begin{aligned}\max \sum_{i=1}^{n} &\frac{1}{\tau}[\text{sgn}(y_i)\ln(1+|y_i|^\alpha) - \text{sgn}(q_\tau(\mathbf{x}_{i(m)}, \boldsymbol{\theta}_\tau^{(m)}))\ln(1+|q_\tau(\mathbf{x}_{i(m)}, \boldsymbol{\theta}_\tau^{(m)})|^\alpha)]\mathbf{1}(y_i \le q_\tau(\mathbf{x}_{i(m)}, \boldsymbol{\theta}_\tau^{(m)})) \\ &-\frac{1}{1-\tau}[\text{sgn}(y_i)\ln(1+|y_i|^\alpha) - \text{sgn}(q_\tau(\mathbf{x}_{i(m)}, \boldsymbol{\theta}_\tau^{(m)}))\ln(1+|q_\tau(\mathbf{x}_{i(m)}, \boldsymbol{\theta}_\tau^{(m)})|^\alpha)]\mathbf{1}(y_i > q_\tau(\mathbf{x}_{i(m)}, \boldsymbol{\theta}_\tau^{(m)})),\end{aligned}$$

which develops a new class of QMLEs for quantile regression.

Furthermore, the estimated τth conditional quantile of y_i given $\mathbf{x}_{i(m)}$ and $\widehat{\boldsymbol{\theta}}_\tau^{(m)}$ in the

mth candidate model can be expressed as:

$$\widehat{\mu}_{i,\tau}^{(m)} = \mathbf{x}'_{i(m)}\widehat{\boldsymbol{\theta}}_{\tau}^{(m)}, \quad m = 1, ..., M, \text{ and } i = 1, ..., n.$$

Denote the model averaging weight $\mathbf{w} = (w_1, \ldots, w_M)' \in \mathcal{W}$, where $\mathcal{W} = \{\mathbf{w} \in \mathbb{R}^M : 0 \leq w_m \leq 1, m = 1, ..., M\}$ is the weight vector space. Here, we relax the restriction that the sum of weights should be one, unlike the weight in Lu and Su (2015). The estimation and asymptotic properties given in Section 3 are not required this constraint. Then, the model averaging estimator of the τth quantile is given by

$$\widehat{\mu}_{i,\tau}(\mathbf{w}) = \sum_{m=1}^{M} w_m \widehat{\mu}_{i,\tau}^{(m)} = \sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\widehat{\boldsymbol{\theta}}_{\tau}^{(m)} = \mathbf{w}^{\top}\widehat{\boldsymbol{\mu}}_{i,\tau}.$$

where $\widehat{\boldsymbol{\mu}}_{i,\tau} = \left(\widehat{\mu}_{i,\tau}^{(1)}, ..., \widehat{\mu}_{i,\tau}^{(M)}\right)^{\top}$, which can be regarded as an approximate of the τth quantile of y_i given $\mathbf{x}_i$; that is $q_\tau(y_i|\mathbf{x}_i) \approx \mathbf{w}^{\top}\widehat{\boldsymbol{\mu}}_{i,\tau}$.

Now, $\widehat{\boldsymbol{\theta}}_{\tau}$ denotes the set of estimator in all candidate models, i.e., $\{\widehat{\boldsymbol{\theta}}_{\tau}^{(1)}, ..., \widehat{\boldsymbol{\theta}}_{\tau}^{(M)}\}$ and the estimated conditional quantile is $\widehat{\boldsymbol{\mu}}_{\tau}(\mathbf{w}) = (\widehat{\mu}_{1,\tau}(\mathbf{w}), ..., \widehat{\mu}_{n,\tau}(\mathbf{w}))'$. In addition, the model averaging estimators for quantile regression is defined as

$$\widehat{\boldsymbol{\theta}}_{\tau}(\mathbf{w}) = \sum_{m=1}^{M} w_m \boldsymbol{\Pi}^{(m)'}\widehat{\boldsymbol{\theta}}_{\tau}^{(m)},$$

where $\boldsymbol{\Pi}^{(m)'}$ is a matrix of $p \times k_m$ projecting $\widehat{\boldsymbol{\theta}}_{\tau}^{(m)}$ to $\widehat{\boldsymbol{\theta}}_{\tau}$, and p is the number of all predictors in the M candidate models. For instance, if all candidate models are nested, $\boldsymbol{\Pi}^{(m)} = (\mathbf{I}_{k_m}, \mathbf{0}_{k_m \times (p-k_m)})$ and $\mathbf{I}_{k_m}$ is a $k_m \times k_m$-identity matrix.

2.2 Weight Choice Criterion

In this section, we propose a penalized model averaging criterion to choose the optimal $\mathbf{w}$ based on the KL loss, which is defined as:

$$\mathrm{KL}(\mathbf{w}) = \sum_{i=1}^{n} \mathbb{E}_0[\ln(f_0(y_i|\mathbf{x}_i)) - \ln(f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau))|\mathbf{x}_i],$$

where y_i is a realization from the true conditional process $f_0(\cdot|\mathbf{x}_i)$, $f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)$ is a given conditional pdf of y_i, and $\mathbb{E}_0[\cdot|\mathbf{x}_i]$ is the conditional expectation with respect to $f_0(\cdot|\mathbf{x}_i)$. If the true conditional pdf $f_0(y_i|\mathbf{x}_i)$ is known, we could obtain a weight by minimizing $\mathrm{KL}(\mathbf{w})$. However, in most cases, it is infeasible to minimize the KL loss for the unknown pdf $f_0(y_i|\mathbf{x}_i)$ and the misspecification of candidate models. Denote the optimal weight $\mathbf{w}^* = \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}}\mathrm{KL}(\mathbf{w})$.

Remark 2. *There are two types of misspecification in our settings. One is the misspecification of the set of regressors. This type of misspecification has been considered in the literature and appears to be the focus of the model averaging literature (Sun et al., 2023; Lu and Su, 2015). The second type refers to misspecify the conditional distribution of y_i. These motivate us to employ QMLE methods to estimate the coefficients and weights. We also consider the asymptotic properties of model averaging for quantile regression under these misspecification settings. For example, Lemma 1 indicates that, for each candidate model, the difference between QMLE $\widehat{\boldsymbol{\theta}}_\tau^{(m)}$ and the pseudo-true value $\boldsymbol{\theta}_\tau^{(m)*}$ converges in probability to zero whether the regressors or the distribution is misspecified. The consistency of estimated weights is shown in Theorem 2.*

By the properties of expectation, we have

$$\mathrm{KL}(\mathbf{w}) = \sum_{i=1}^{n} \mathbb{E}_0[\ln(f_0(y_i|\mathbf{x}_i))] - \sum_{i=1}^{n} \mathbb{E}_0[\ln(f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau))|\mathbf{x}_i],$$

in which, the first term has no effect on the choice of $\mathbf{w}$. Then, we propose a feasible double

penalized weight choice criterion as follows:

$$G_n(\mathbf{w}) = -\sum_{i=1}^{n} \ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau) + \lambda_{n,1}\mathbf{w}'\mathbf{k} + \sum_{j=1}^{p} p_{\lambda_{n,2}}(|\widehat{\theta}_{j,\tau}(\mathbf{w})|), \tag{5}$$

where $\lambda_{n,1}\mathbf{w}'\mathbf{k}$ and $p_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w})|)$ are penalty terms, $\lambda_{n,i}(i = 1, 2)$ are the tuning parameters, $\mathbf{k} = (k_1, ..., k_M)'$ with k_m being the number of columns of $\mathbf{x}$ used in the mth candidate model,

$$\widehat{\theta}_{j,\tau}(\mathbf{w}) = \sum_{m=1}^{M} w_m(\boldsymbol{\Pi}^{(m)'}\widehat{\boldsymbol{\theta}}_\tau^{(m)})_j,$$

and $(\cdot)_j$ denotes the jth member of the vector $\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w})$.

Remark 3. *Note that $\lambda_{n,1}\mathbf{w}'\mathbf{k}$ is the penalty for the model complexity, which is widely used in Zhang et al. (2016) and Zhang et al. (2020). Intuitively, the selected weights of over-fitted (or some sufficiently large) candidate models may be shrunk to zero, if $\lambda_{n,1}$ is going to infinity when the sample size is going to infinity. On the other hand, following Sun et al. (2023), we employ $p_{\lambda_{n,2}}(|\widehat{\theta}_{j,\tau}(\mathbf{w})|)$ as a penalty for predictor sparsity, which can remove the redundant predictors. $p_{\lambda_{n,2}}(|\widehat{\theta}_{j,\tau}(\mathbf{w})|)$ could be a class of penalties, including the LASSO penalty as in Tibshirani (1996), the SCAD penalty as in Fan and Li (2001), and the adaptive LASSO penalty as in Zou (2006). For example, suppose all the candidate models are univariate models. Then, $\lambda_{n,1}\mathbf{w}'\mathbf{k}$ provides the same penalty for each candidate model, while $p_{\lambda_{n,2}}(|\widehat{\theta}_{j,\tau}(\mathbf{w})|)$ can select important predictors and construct a parsimonious mega model after model averaging. In addition, one can also use penalization methods at the estimation stage of candidate models, especially when $k_m > n$. However, combining the M candidate models still results in a mega model when estimating the weights. Thus, it is necessary to employ penalties at the averaging stage. To sum up, with these two penalties, the proposed weight choice criterion determines the optimal combination weights and yields sparseness for various potential covariates simultaneously.*

Remark 4. *Furthermore, the proposed weight choice criterion $G_n(\mathbf{w})$ could reduce to some classical criteria in the existing literature. For instance, when all elements of $\mathbf{w}$ are 0 or 1 and $\lambda_{n,2} = 0$, suppose $\lambda_{n,1} = 2$, then our proposed criterion $G_n(\mathbf{w})$ is equivalent to the*

AIC; and suppose $\lambda_{n,1} = \ln n$*, then* $G_n(\mathbf{w})$ *is the BIC. When* y_i *follows the asymmetric Laplace distribution and* $\lambda_{n,2} = 0$*, suppose* $\lambda_{n,1} = 2\sigma^2 = 2\text{var}(\varepsilon_i|\mathbf{x}_i)$*, then* $G_n(\mathbf{w})$ *reduces to a Mallows-type criterion for quantile regression in Lu and Su (2015).*

As a result, the weight vector is obtained from

$$\widehat{\mathbf{w}} = \text{argmin}_{\mathbf{w}} G_n(\mathbf{w}), \mathbf{w} \in \mathcal{W}, \tag{6}$$

where $G_n(\mathbf{w})$ is defined in (5). Then, the model averaging estimator for quantile regression is given by

$$\widehat{\boldsymbol{\theta}}_\tau(\widehat{\mathbf{w}}) = \sum_{m=1}^{M} \widehat{w}_m \boldsymbol{\Pi}^{(m)'} \widehat{\boldsymbol{\theta}}_\tau^{(m)},$$

and the estimated conditional quantile based on $\widehat{\boldsymbol{\theta}}_\tau(\widehat{\mathbf{w}})$ is defined as

$$\widehat{\boldsymbol{\mu}}_\tau(\widehat{\mathbf{w}}) = (\widehat{\mu}_{1,\tau}(\widehat{\mathbf{w}}), ..., \widehat{\mu}_{n,\tau}(\widehat{\mathbf{w}}))',$$

where $\widehat{\mu}_{i,\tau}(\widehat{\mathbf{w}}) = \sum_{m=1}^{M} \widehat{w}_m \widehat{\mu}_{i,\tau}^{(m)} = \sum_{m=1}^{M} \widehat{w}_m \mathbf{x}'_{i(m)} \widehat{\boldsymbol{\theta}}_\tau^{(m)}$.

2.3 Implementation Algorithm

In this section, we suggest an algorithm for our proposed model averaging process for an easy implementation, see more details in Algorithm 1. Note that we follow Fan and Li (2001), Zou (2006) and Meier et al. (2008) to determine the tuning parameters $\lambda_{n,1}$ and $\lambda_{n,2}$ in the proposed weight choice criterion by k-fold cross-validation. Specifically, we first randomly divide the sample set into k mutually exclusive subsets of the same size. Then, each time one subset is selected as the validation set, the remaining $k-1$ subsets are used as the training set. This process is repeated for k times until all subsets are taken as the validation set. Finally, the optimal tuning parameters are determined by minimizing the weight choice criterion on the validation sets.

Algorithm 1 Penalized model averaging for quantile regression algorithm

Step 1. Construct potential candidate models.

Step 2. Calculate the estimated conditional quantile of each candidate model.

For $m = 1$ to M:

(a) Determine the conditional density function $f(y, \eta)$ of quantile regression by the prior information, including asymmetric Laplace density and other densities in the tick-exponential family (4);

(b) Estimate the parameter $\boldsymbol{\theta}_\tau^{(m)}$ of each candidate models by QMLE;

(c) Set $\widehat{\mu}_{i,\tau}^{(m)} = \mathbf{x}'_{i(m)}\widehat{\boldsymbol{\theta}}_\tau^{(m)}$.

Step 3. Calculate the model averaging weight based on the penalized KL criterion.

(a) For $j = 1$ to p: choose a proper kind of penalty $p_{\lambda_{n,2}}$ for the predictors;

(b) Set $f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau) = f(y_i, \sum_{m=1}^M w_m \widehat{\mu}_{i,\tau}^{(m)})$;

(c) Construct the feasible double penalized weight choice criterion as:

$$G_n(\mathbf{w}) = -\sum_{i=1}^{n} \ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau) + \lambda_{n,1}\mathbf{w}'\mathbf{k} + \sum_{j=1}^{p} p_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w})|).$$

(d) Determine the tuning parameters $\lambda_{n,1}$ and $\lambda_{n,2}$;

(e) Solve the above penalized minimum negative likelihood problem and obtain the model averaging weight:

$$\widehat{\mathbf{w}} = \min_{\mathbf{w}\in\mathcal{W}} G_n(\mathbf{w});$$

Step 4. Compute the model averaging estimator:

The estimated conditional quantile $\widehat{\boldsymbol{\mu}}_\tau(\widehat{\mathbf{w}}) = (\widehat{\mu}_{1,\tau}(\widehat{\mathbf{w}}), ..., \widehat{\mu}_{n,\tau}(\widehat{\mathbf{w}}))'$, and

$$\widehat{\mu}_{i,\tau}(\widehat{\mathbf{w}}) = \sum_{m=1}^{M} \widehat{w}_m \widehat{\mu}_{i,\tau}^{(m)} = \sum_{m=1}^{M} \widehat{w}_m \mathbf{x}'_{i(m)}\widehat{\boldsymbol{\theta}}_\tau^{(m)}.$$

2.4 Application Based on Asymmetric Laplace Distribution

In this section, our intention is to study the proposed criterion for the popular case that y_i is supposed to follow the asymmetric Laplace distribution. As pointed out by Yu and Moyeed (2001), minimizing the quantile loss function proposed by Koenker and Bassett (1978) is equivalent to maximizing the parametric likelihood under the asymmetric Laplace error distribution, and also, Komunjer (2005) stated that QMLE with asymmetric Laplace density reduces to the classical quantile regression estimators. Based on the asymmetric

Laplace distribution, the probability density function of dependent variable y_i is given by

$$f(y, \mu, \tau, \sigma) = \frac{\tau(1-\tau)}{\sigma} \exp\left(-\rho_\tau(y-\mu)/\sigma\right),$$

where $\tau \in (0,1)$, μ is the location parameter, σ is the scale parameter, and ρ_τ is the check function as $\rho_\tau(y-\mu) = (y-\mu)(\tau \mathbf{1}(y-\mu \geq 0) - (1-\tau)\mathbf{1}(y-\mu \leq 0))$ with $\mathbf{1}(\cdot)$ denoting the indicative function. Then, the logarithmic likelihood function can be expressed as:

$$L_n(\mathbf{y}, \boldsymbol{\mu}, \tau, \sigma) = n \ln(\tau(1-\tau)/\sigma) - \sum_{i=1}^{n} \frac{1}{\sigma} \rho_\tau(y_i - \mu_i),$$

where $\mathbf{y} = (y_1, ..., y_n)$, $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)$, and $\mu_i = \sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)} \widehat{\boldsymbol{\theta}}_\tau^{(m)}$. For a given $\tau \in (0,1)$, the selected weight vector is obtained as follows,

$$\widehat{\mathbf{w}} = \text{argmin}_{\mathbf{w} \in \mathcal{W}} \left\{ \sum_{i=1}^{n} \frac{1}{\sigma} \rho_\tau \left(y_i - \sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)} \boldsymbol{\theta}^{(m)} \right) + \lambda_{n,1} \mathbf{w}' \mathbf{k} + \sum_{j=1}^{p} p_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w})|) \right\}. \quad (7)$$

Without any penalties in this criterion, i.e., $\lambda_{n,1} = \lambda_{n,2} = 0$, (7) is equivalent to the loss function in Lu and Su (2015) and Wang et al. (2023). Also, using the jackknife technique, our method reduces to Jacknife model averaging as in Lu and Su (2015), and further combining with a covariate screening process before model averaging, our method becomes to that in Wang et al. (2023). Thus, our proposed method is more general than the existing approaches with respect to the form of weight choice loss function.

3 Asymptotic Theories

In this section, we investigate the asymptotic properties of the penalized model averaging for quantile regression under time series framework. Specifically, the asymptotic optimality and consistency of weight $\mathbf{w}$ estimated by (6) are shown in Theorems 1 and 2, and Theorem 3 states that the model averaging estimator $\widehat{\boldsymbol{\theta}}_\tau(\widehat{\mathbf{w}})$ asymptotically converges to a well-defined limit, even if all candidate models are misspecified. Let $\boldsymbol{\theta}_\tau^{*(m)}$ be the parameter vector which minimizes the KL divergence between the true model with density $f_0(\cdot)$ and the mth

candidate model. First, the following assumptions are imposed although they might not be the weakest possible.

Assumption 1. *(i) Let* $\{y_i, \mathbf{x}_i, \mathscr{F}_i\}$ *be an adapted stochastic sequence such that* $\{y_i, \mathbf{x}_i\}$ *are stationary and ergodic process, where* $\mathscr{F}_i$ *is a* σ*-algebra, and also, the sequence* $\{y_i, \mathbf{x}_i\}$ *is strong mixing with* α_0 *size of* $-r_0/(r_0-2)$ *with* $r_0 > 2$*; (ii) the stochastic process* $\{\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)\}$ *satisfies that* $\text{var}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)] = \sigma_f^2 < \infty$, $\mathbb{E}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau) - \mathbb{E}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)]|\mathscr{F}_{-r}] \overset{q.m.}{\to} 0$ *in quadratic mean (q.m.) as* $r \to \infty$*, and* $\sum_{j=1}^{\infty} \text{var}(\mathcal{R}_{ij})^{1/2} < \infty$*, where* $\mathcal{R}_{ij} = \mathbb{E}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)|\mathscr{F}_{i-j}] - \mathbb{E}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)|\mathscr{F}_{i-j-1}]$.

Assumption 2. *(i)* $\ln f(y_i|\boldsymbol{\theta}, \mathbf{x}_i, \tau)$ *is continuously differentiable with probability 1 with respect to* θ*; (ii)* $\mathbb{E}_0[\|\frac{\partial \ln f(y_i|\boldsymbol{\theta}, \mathbf{x}_i, \tau)}{\partial \boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}(\mathbf{w})}\||\mathbf{x}_i]$ *is finite for every* $\tilde{\boldsymbol{\theta}}(\mathbf{w})$ *between* $\widehat{\boldsymbol{\theta}}(\mathbf{w})$ *and* $\boldsymbol{\theta}^*(\mathbf{w})$*, where* $\boldsymbol{\theta}^*(\mathbf{w}) = \sum_{m=1}^{M} w_m \boldsymbol{\theta}_\tau^{*(m)}$.

Assumption 3. $M^2 np \zeta_n^{-2} = o(1)$*, where* $\zeta_n = \inf_{\mathbf{w}\in\mathcal{W}} KL^*(\mathbf{w})$ *and*

$$KL^*(\mathbf{w}) = \sum_{i=1}^{n} \int [f_0(y_i|\mathbf{x}_i) \ln f_0(y_i|\mathbf{x}_i) - f_0(y_i|\mathbf{x}_i) \ln f(y_i|\boldsymbol{\theta}_\tau^*(\mathbf{w}), \mathbf{x}_i, \tau)] dy.$$

Assumption 4. *(i)* $\lambda_{n,1} = O(n^{1/2}p^{-1/2})$*; (ii) For all* $j = 1, ..., p$, $p'_{\lambda_{n,2}}(\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w})) = O(n^{1/2}p^{-1/2})$ *and* $p''_{\lambda_{n,2}}(\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w})) = o_p(np^{-1})$.

Assumption 5. *For some positive finite constants* C_1 *and* C_2*, the conditional matrix*

$$I(\mathbf{w}^*) = \mathbb{E}_0\left[\frac{\partial \ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}^*), \mathbf{x}_i, \tau)}{\partial \mathbf{w}} \frac{\partial \ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}^*), \mathbf{x}_i, \tau)}{\partial \mathbf{w}'}\right]$$

satisfies $0 < C_1 < \lambda_{min}\{I(\mathbf{w}^*)\} \leq \lambda_{max}\{I(\mathbf{w}^*)\} < C_2$ *for almost all* $\mathbf{x}_i$ *with probability* 1*, where* λ_{min} *and* λ_{max} *denote the minimum and maximum eigenvalue of the matrix. In addition, for* $m, s = 1, 2, \ldots, M$ *and some positive finite constants* C_3 *and* C_4*,*

$$\mathbb{E}_0\left[\frac{\partial \ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}^*), \mathbf{x}_i, \tau)}{\partial w_m} \frac{\partial \ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}^*), \mathbf{x}_i, \tau)}{\partial w_s}\right]^2 < C_3$$

and $\mathbb{E}_0\left[\partial^2 \ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}^*), \mathbf{x}_i, \tau)/\partial w_m \partial w_s\right]^2 < C_4$*, for almost all* $\mathbf{x}_i$ *with probability* 1.

Assumption 6. *For all* $\mathbf{w} \in \mathcal{W}$, $l, m, s = 1, ..., M$, *and some positive constant* C_5, *there exist functions* M_{lms} *such that*

$$\left| \frac{\partial^3 \ln f(y_i|\widehat{\boldsymbol{\theta}}(\mathbf{w}), \mathbf{x}_i, \tau)}{\partial w_l \partial w_m \partial w_s} \right| \leq M_{lms}(y_i|\mathbf{x}_i)$$

and $\mathbb{E}_0[M^2_{lms}(y_i|\mathbf{x}_i)|\mathbf{x}_i] < C_5 < \infty$, *for almost all* $\mathbf{x}_i$ *with probability* 1.

Assumption 7. *(i)* $M^{1/2}n^{-\delta/2} = o(1)$; *(ii)* $M^{3/2}\xi_n^{1/2}n^{-1/2+\delta/2} = o(1)$; *(iii)* $M^{1/2}p^{1/2}\xi_n^{-1/2}n^{-\delta/2} = o(1)$, *where* δ *is a positive a positive constant and*

$$\xi_n = \sum_{i=1}^{n} \mathbb{E}_0 \left[\left(\frac{f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}^*), \mathbf{x}_i, \tau)}{f_0(y_i|\mathbf{x}_i)} - 1 \right)^2 \right].$$

Remark 5 (Discussions of Assumptions)**.** *Assumption 1 includes some regularity conditions to derive some asymptotic theories for dependent identically distributed process, see, for example, Chapter 5 in White (1984) and Condition (A4) in Komunjer (2005). Among various mixing conditions used in literature, strong mixing is reasonably weak and is known to be fulfilled for many stochastic processes, including many time series models, see, for example, the paper by Cai (2002) for some examples. Assumption 2 is about the first derivative of the density function* $f(\cdot)$ *with respect to* $\boldsymbol{\theta}$, *which can be verified from original conditions in the existing literature such as White (1982), Komunjer (2005) and Zhang et al. (2016). The continuous differentiability of* $f(\cdot)$ *is, for example, satisfied when for every* θ, $\partial \ln f(y_i|\boldsymbol{\theta}, \mathbf{x}_i, \tau)/\partial \boldsymbol{\theta}'$ *exists and is continuously for almost all* $(y_i, \mathbf{x}_i)$. *Assumption 3 requires the infimum of* $KL^*(\mathbf{w})$ *on* $\mathcal{W}$ *grows to infinity at a faster rate than* $M\sqrt{np}$. *Similar assumptions can be found in the existing literature, like Condition (C.7) in Liao et al. (2019) and Condition (C.3) in Zhang et al. (2016). Assumption 4 provides constraints on the diverging rate of penalties, which are commonly used in the literature about penalized regression, like Tibshirani (1996), Fan and Peng (2004) and Zou (2006). The slower diverging rate of* p *than* n *implies* $\lambda_{n,1} \to \infty$ *as* $n \to \infty$. *When the second penalty is taken to be LASSO penalty, we have* $p''_{\lambda_{n,2}}(\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w})) = 0$ *and (ii) of Assumption 4 can be specified as* $n^{-1/2}p^{1/2}\lambda_{n,2} = O(1)$. *Assumption 5 requires that the eigenvalues of* $I(\mathbf{w}^*)$, *each term of* $I(\mathbf{w}^*)$ *and the expecta-*

tion of $\partial^2 \ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}^*), \mathbf{x}_i, \tau)/\partial w_m \partial w_s$ *are uniformly bounded. Assumption 6 is about the third derivative of the log likelihood function. Assumptions 5 and 6 impose some restrictions on the derivative of log likelihood functions, and similar assumptions can be found in Fan and Peng (2004). When all the candidate models are misspecified, Assumption 7 is used to prove that* $\|\widehat{\mathbf{w}} - \mathbf{w}^*\|$ *converges to zero at certain rate. It implies the relationships among* ξ_n, n, p, *and* M, *where* ξ_n *measures the difference between the density function with the model averaging estimator* $\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}^*)$ *and the true conditional pdf. It is similar to Condition (C.10) in Liao et al. (2019) and Condition (C.6) in Zhang et al. (2016). For example, suppose the* m_0*th candidate model is the correctly specified model,* $f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}^*), \mathbf{x}_i, \tau) = \tau(1-\tau)\exp(-\rho_\tau(y - \widehat{\boldsymbol{\theta}}_\tau^{(m_0)}\mathbf{x}'_{i,m_0}))$ *and* $f_0(y_i|\mathbf{x}_i) = \tau(1-\tau)\exp(-\rho_\tau(y - \boldsymbol{\theta}_0\mathbf{x}'_{i,m_0}))$. *Then, we have* $\xi_n = O(n)O_p([\exp(1/\sqrt{n}) - 1]^2) = O(1)$ *for* $\|\widehat{\boldsymbol{\theta}}_\tau^{(m_0)} - \boldsymbol{\theta}_0\| = O_p(1/\sqrt{n})$. *Thus, it is obvious to verify that* ξ_n *satisfies Assumption 7.*

For the mth quantile regression with a diverging number of regressors, we establish the consistency of the parameter estimators based on QMLE with its detailed proof given in Appendix C.

Lemma 1. *Suppose Conditions (A.1)-(A.5) in Appendix A are satisfied. Then, for any fixed* $\varepsilon > 0$, *there exists a constant* $\delta_\varepsilon > 0$, *such that for any sufficiently large* n *and* k_m, *we have*

$$P\left(\sqrt{n}k_m^{-1/2}\|\widehat{\boldsymbol{\theta}}_\tau^{(m)} - \boldsymbol{\theta}_\tau^{(m)*}\| \leq \delta_\varepsilon\right) \geq 1 - \varepsilon. \tag{8}$$

Remark 6. *Lemma 1 implies that* $\|\widehat{\boldsymbol{\theta}}_\tau^{(m)} - \boldsymbol{\theta}_\tau^{(m)*}\| = O_p(\sqrt{k_m}n^{-1/2})$ *when* k_m *is diverging. Based on this, the convergence rate of model averaging estimator* $\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w})$ *can be obtained as* $\|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}) - \boldsymbol{\theta}^*(\mathbf{w})\| = \|\sum_{m=1}^M w_m(\widehat{\boldsymbol{\theta}}_\tau^{(m)} - \boldsymbol{\theta}_\tau^{(m)*})\| \leq O_p(M\bar{k}^{1/2}n^{-1/2})$, *where* $\bar{k} = \max\{k_1, ..., k_M\}$. *It is obviously that* $\bar{k} \leq p$. *When all* k_m *for* $m = 1, \cdots, M$ *and* M *are fixed, (8) reduces to* $\|\widehat{\boldsymbol{\theta}}_\tau^{(m)} - \boldsymbol{\theta}_\tau^{(m)*}\| = O_p(n^{-1/2})$, *and thus we have* $\|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}) - \boldsymbol{\theta}^*(\mathbf{w})\| = O_p(n^{-1/2})$.

Next, Theorem 1 gives the asymptotic optimality of the estimated weight $\mathbf{w}$ in the penalized model averaging for quantile regression with its detailed proof relegated to Appendix C.

Theorem 1. *With Assumptions 1-4 and Equation (8), $\widehat{\mathbf{w}}$ is asymptotically optimal in the sense that*

$$\frac{KL(\widehat{\mathbf{w}})}{\inf_{\mathbf{w}\in\mathcal{W}} KL(\mathbf{w})} \xrightarrow{p} 1$$

as $n \to \infty$.

Theorem 1 implies that the proposed model averaging estimator is asymptotically optimal in the sense that the KL loss obtained using the weight vector by (6) is asymptotically equivalent to that of the infeasible best possible model averaging estimator for quantile regression with its detailed proof presented in Appendix C.

Theorem 2. *If Assumptions 1-7 and Equation (8) are satisfied, there is a local minimizer $\widehat{\mathbf{w}}$ of $G_n(\mathbf{w})$ such that*

$$\|\widehat{\mathbf{w}} - \mathbf{w}^*\| = O_p\left(\xi_n^{1/2} n^{-1/2+\delta/2}\right),$$

where ξ_n is defined in Assumption 7.

Theorem 2 shows that $\widehat{\mathbf{w}}$ converges to the optimal weight $\mathbf{w}^*$ at the rate $\xi_n^{-1/2} n^{(-1+\delta)/2}$. It is worth noticing that $\xi_n \to \infty$ if all candidate models are misspecified, and the slower the rate of $\xi_n \to \infty$, the faster the rate $\widehat{\mathbf{w}} \to \mathbf{w}^*$. When the candidate models include correctly specified model, $\widehat{\mathbf{w}}$ converges to the optimal weight $\mathbf{w}^*$ in probability at the rate $n^{-1/2+\delta/2}$.

With Lemma 1 and Theorem 2, the consistency of the model averaging estimators in misspecified models is given as follows, with its detailed proof depicted in Appendix C.

Theorem 3. *Under the assumptions of Theorem 2 and $pM\xi_n n^{-1+\delta} \to 0$, the model averaging estimator $\widehat{\boldsymbol{\theta}}_\tau(\widehat{\mathbf{w}})$ satisfies*

$$\|\widehat{\boldsymbol{\theta}}_\tau(\widehat{\mathbf{w}}) - \boldsymbol{\theta}_\tau^*(\mathbf{w}^*)\| = O_p\left(p^{1/2} M^{1/2} \xi_n^{1/2} n^{-1/2+\delta/2} + M p^{1/2} n^{-1/2}\right),$$

where ξ_n is defined in Assumption 7.

Theorem 3 indicates that the model averaging estimator $\widehat{\boldsymbol{\theta}}_\tau(\widehat{\mathbf{w}})$ converges to a well-defined limit $\boldsymbol{\theta}^*(\mathbf{w}^*)$, even if all candidate models may be misspecified.

4 Model Averaging for Ultra-High Dimensional Data

Section 3 has discussed the asymptotic properties of our method when p and M are diverging. When modeling the ultra-high dimensional data, a large number of candidate models will result in a heavy computational burden. Furthermore, the theoretical properties in Section 3 will break down if p and M are larger than n. To solve these problems, we introduce a two-step process to the model averaging for quantile regression in this section. Inspired by Yuan and Yang (2005) and Zhang et al. (2013), we could select the candidate models by penalized regression. And an intuitional method mentioned in Zhang et al. (2016) for mean regression is threshold model screening, which is also suitable for quantile regression.

In the first step, we employ a model screening process to reduce the dimension of candidate models and the complexity of our model. Therefore, we intend to eliminate the poor candidate models by some rules and make the dimension of candidate models to be smaller than n before doing the model averaging procedure for quantile regression. We provide three model screening strategies to screen the candidate models. Ordering model screening method follows the spirit of Ando and Li (2014, 2017).

In addition, if the number of predictors in some candidate models is also larger than n, we could employ some variable selection method to estimate the candidate models, like LASSO, elastic network, SCAD, and principal component analysis. Note that quantile regression with different τ may lead to different screening results. Thus, we consider the following three algorithms for model screening methods in practice.

Algorithm 2 Ordering model screening algorithm

Step 1. Employs some criteria to order the covariates and divide them into $M+1$ groups.

(a) Employ the marginal quantile utility proposed by He et al. (2013) to calculate the marginal utility between each predictor variable and the response variable when τ varies from 0 to 1.

(b) Partition the marginal utility of predictors into $M+1$ groups by their absolute values. The first group has the highest values and the $M+1$ group has values closest to zero.

(c) Construct one candidate model for each group and drop the $M+1$ group.

Step 2. Estimate the weights of penalized model averaging on the M candidate models by Algorithm 1.

Algorithm 3 Top m model screening algorithm

Step 1. Screen the candidate model by regularized penalization.

(a) Construct a penalized likelihood estimation with LASSO (or adaptive LASSO, SCAD, etc.) to estimate the full model.

(b) For any fixed τ, solve the solution path of the penalized model in the previous step as tuning parameter changes.

(c) Select different regressors by M different tuning parameters on the solution path, and construct M candidate models by the selected regressors.

Step 2. Estimate the weights of penalized model averaging on the M candidate models by Algorithm 1.

Algorithm 4 Threshold model screening algorithm

Step 1. Screen the candidate model by a given threshold.

(a) Estimate the weight vector $\widehat{\mathbf{w}}$ with all candidate models.

(b) Remove the models with weights smaller than a given threshold constant.

Step 2. Estimate the weights of penalized model averaging on the M candidate models by Algorithm 1.

Having reduced the dimension of our model, the next step is to estimate the model averaging weight with the screened candidate models. The weights of these eliminated models will be restricted to zero. Define the subspace of $\mathcal{W}$ generated in the first step as $\mathcal{W}^* = \{\mathbf{w} \in \mathcal{W} : w_m = 0, \text{if } m \notin \mathcal{M}^*\}$, where $\mathcal{M}^*$ is a subset of $\{1, ..., M\}$. Then, the model averaging process in section 2 can be implemented based on this new space. The resultant weight $\widehat{\mathbf{w}}^*$ of our method is solved by

$$\widehat{\mathbf{w}}^* = \text{argmin}_{\mathbf{w}} G_n(\mathbf{w}), \mathbf{w} \in \mathcal{W}^*.$$

In the following, we provide an additional assumption required for proving the asymptotic optimality.

Assumption 8. *(i) There exists a nonnegative constant series ν_n such that $\zeta_n^{-1}\nu_n \to 0$, where ζ_n is defined in Assumption 3;*
(ii) There exists a weight vector series $\mathbf{w}_n \in \mathcal{W}$ satisfying $\inf_{\mathbf{w}\in\mathcal{W}} KL(\mathbf{w}) = KL(\mathbf{w}_n) - \nu_n$ and $P(\mathbf{w}_n \in \mathcal{W}^*) \to 1$, as $n \to \infty$.

The next theorem can be seen as a generalization of Theorem 1. Assumption 8 and the assumptions of Theorem 1 ensure that the weight of model averaging after model screening is also asymptotically optimal. Note that the detailed proof of the following theorem is given in Appendix C.

Theorem 4. *Under Assumption 8 and the assumptions of Theorem 1, the two-step weight* $\widehat{\mathbf{w}}^*$ *is asymptotically optimal in the sense that*

$$\frac{KL(\widehat{\mathbf{w}}^*)}{\inf_{\mathbf{w}\in\mathcal{W}} KL(\mathbf{w})} \xrightarrow{p} 1$$

as $n \to \infty$.

Theorem 4 concludes that the model averaging estimator obtained based on subspace $\mathcal{W}^*$ yields the KL loss that is asymptotically identical to that of the estimator obtained from the whole space.

5 Monte Carlo Simulations

In this section, we aim to investigate the finite sample performance of the proposed model averaging method (namely, QMA) in comparison with existing methods, including jackknife model averaging (JMA) as in Lu and Su (2015), Mallows-type information criterion (QRIC) as in Lu and Su (2015), smoothed AIC (SAIC) and smoothed BIC (SBIC). The data generation process (DGP) is given as follows.

DGP 1: First, we generate the data based on the following process:

$$y_t = \sum_{j=1}^{1000} \theta_j x_{tj} + \varepsilon_t, \quad t = 1, \ldots, n,$$

where $\theta_j = cj^{-1}$ and c is varied such that the signal-noise ratio of the population $R^2 = [\text{var}(y_i) - \text{var}(\varepsilon_i)]/\text{var}(y_i) = 0.1, 0.2, ..., 0.9$. Following Sun et al. (2023), we set $x_{i1} = 1$, $x_{tj} = 0.5x_{(t-1)j} + \epsilon_{tj}$ $(j = 2, ..., 1000)$ being an AR(1) process, where $\epsilon_{tj} \sim N(0, 1)$, i.i.d. over

t, with $\text{corr}(\epsilon_{ti}, \epsilon_{tj}) = 0.75$.[2] For robustness check, we consider two settings for the error term. In Setting I, the error is supposed to be homoscedastic and follows $N(0,1)$, which is independent of x_{tj}. Setting II considers a heteroscedastic process, where $\varepsilon_t = \sum_{j=2}^{6} x_{tj}^2 \epsilon_t$ and ϵ_t is $N(0,1)$ and independent of x_{tj}.

To mimic the situation that the number of candidate models M is diverging, we consider a nested class, and M is given by $[4n^{1/4}]$, where $[x]$ denotes the largest integer not larger than x. We set the training sample size $n = 50$, 100, and 150, and thus the corresponding numbers of candidate models are 10, 12 and 13.

DGP 2: We consider the following sparse setup:

$$y_i = \sum_{j=1}^{9} \theta_j x_{ij} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $\{x_{ij}, j = 1, ..., 9\}$ follows normal distribution $N(0,1)$ and the correlation between different x_{ij} is 0.75. We set the parameter $\boldsymbol{\theta} = (\theta_1, ..., \theta_9) = c \cdot (3, 0, 0, 1.5, 0, 0, 7, 0, 0)'$, where c is varied such that $R^2 = 0.1, 0.2, ..., 0.9$. We also consider two settings for $\{\varepsilon_i\}$ to check the robustness. In Setting I, for the homoscedasticity, ε_i is $N(0,1)$ and independent of x_{ij}. Setting II considers $\varepsilon_i = \sum_{j=1}^{9} j^{-1} x_{ij} \epsilon_i$ where ϵ_i is $N(0,1)$ and independent of x_{ij}. This data generating process is adopted from Zhang et al. (2016) and Zou (2006). In addition, we consider a nested framework with $M = 9$ candidate models, namely $\{\mathbf{x}_1\}, \{\mathbf{x}_1, \mathbf{x}_2\}, ..., \{\mathbf{x}_1, ..., \mathbf{x}_9\}$, and other settings are the same as those for DGP 1.

We consider three different quantiles with $\tau = 0.05, 0.5, 0.95$, corresponding to lower quartile, median, and upper quartiles. To examine forecast accuracy, we follow Lu and Su (2015) and Wang et al. (2023) to define the regression quantile prediction error of the rth

[2]Following Lu and Su (2015), we also consider the case that $\mathbf{x}$ is generated by an i.i.d. process, namely $x_{i1} = 1$ and $x_{ij} \sim N(0,1)$, $j = 2, ..., 1000$. Other settings are the same as those for DGP 1. The performances of QMA, JMA, SAIC, SBIC and QRIC under this process are shown in Figures B1 for Setting I and B2 for Setting II in Appendix B and the conclusions about Figures B1 and B2 are presented after each figure, respectively.

replication as follows:

$$\text{FPE}(r) = \frac{1}{n_f}\sum_{s=1}^{n_f}\rho_\tau\left(y_i - \sum_{m=1}^{M}\widehat{w}_m \mathbf{x}'_{i(m)}\widehat{\boldsymbol{\theta}}_\tau^{(m)}\right).$$

The forecast sample size n_f is set to be 100, and each experiment is repeated $n_{\text{sim}} = 1000$ times. Then, we average the out-of-sample prediction error over $n_{\text{sim}} = 1000$ replications: $\text{FPE} = \frac{1}{n_{\text{sim}}}\sum_{r=1}^{n_{\text{sim}}}\text{FPE}(r)$, and we normalize the quantile prediction error by dividing the prediction error of the infeasible optimal single candidate model.

We employ four methods as follows to compare with our model: jackknife model averaging, Mallows-type information criterion as in Hansen (2007) for quantile regression model averaging (QRIC), smoothed Akaike information criterion (SAIC), smoothed Bayesian information criterion (SBIC). The leave-one-out cross-validation criterion for JMA defined in Lu and Su (2015) is

$$\text{CV}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - \sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\widehat{\boldsymbol{\theta}}_\tau^{(m)}).$$

The weight of QRIC can be obtained from $\widehat{\mathbf{w}} = \min_{\mathbf{w}\in\mathcal{W}}\text{QRIC}(\mathbf{w})$, where

$$\text{QRIC}(\mathbf{w}) = \sum_{i=1}^{n}\rho_\tau(\widehat{\varepsilon}_i(\mathbf{w})) + \frac{\tau(1-\tau)}{f(F^{-1}(\tau))}\sum_{m=1}^{M} w_m k_m$$

with $\widehat{\varepsilon}_i(\mathbf{w}) = \sum_{m=1}^{M} w_m(y_i - \mathbf{x}'_{i(m)}\widehat{\boldsymbol{\theta}}_\tau^{(m)})$, and f and F denoting the cdf and pdf of ε_i. And the SAIC and SBIC weights for quantile regression model averaging are defined as

$$\widehat{w}_m^{AIC} = \frac{\exp(-\frac{1}{2}\text{AIC}_m)}{\sum_{m=1}^{M}(\exp(-\frac{1}{2}\text{AIC}_m))} \quad \text{and} \quad \widehat{w}_m^{BIC} = \frac{\exp(-\frac{1}{2}\text{BIC}_m)}{\sum_{m=1}^{M}(\exp(-\frac{1}{2}\text{BIC}_m))},$$

where AIC_m and BIC_m are information criteria for the mth candidate model, i.e., $\text{AIC}_m = A_n + 2k_m$ and $\text{BIC}_m = A_n + \ln(n)k_m$, where $A_n = 2n\ln\left[\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - \mathbf{x}'_{i(m)}\widehat{\boldsymbol{\theta}}_\tau^{(m)})\right]$.

Figures 1 - 2 report forecast results under DGP 1 with settings I and II. When modeling the homoscedastic process, several observations can obtained from Figure 1. First, QMA outperforms the other competing methods including JMA for nearly all T, τ and R^2. For

example, when $n = 50$, $\tau = 0.05$ and $R^2 = 0.50$, QMA provides a prediction error of 1.0975, which is 10.93%, 37.18%, 27.98% and 42.07% lower than the errors of JMA, SAIC, SBIC and QRIC. Second, when $\tau = 0.05$ and 0.95, QMA attains a slightly lower prediction error than JMA in most cases. For $\tau = 0.50$, we find evidence of a much larger difference in prediction error between QMA and JMA. Third, JMA always produces smaller prediction errors than SAIC, SBIC and QRIC with $\tau = 0.05$ and 0.95, while it can no longer dominate other methods with $\tau = 0.50$. This highlights the merits of QMA over other competing methods. Finally, the advantages of QMA are more pronounced when the sample size is small. One possible explanation is that QMA with two penalties provides a more parsimonious mega model than other model averaging methods and is suitable for selecting important predictors. Furthermore, when we consider a heteroscedastic error process, QMA estimator still frequently yields the most accurate estimates in most cases.

Figures 3 - 4 present the out-of-sample performance for DGP 2 with sparse settings when $\tau = 0.05$, 0.50 and 0.95. For the homoscedastic error, it can be observed that QMA enjoys the smallest FPEs for all τ, T and R^2. But JMA is no longer able to dominate SAIC, SBIC and QRIC. In fact, no method other than QMA can dominate the rest methods in all cases. And SAIC yields the worst outcomes when $\tau = 0.05$. With $\tau = 0.50$ and $n = 100$ and 150, SBIC is inferior to the other methods. One possible explanation for the strong showing of our method is that the penalized process can capture the sparsity of the model and produce more accurate predictions. In addition, it can be observed from Figure 4 that our method also outperforms the others in the DGP with heteroskedasticity.

6 An Empirical Example

In this section, we apply the proposed estimation procedure to predict the quantiles of excess stock returns. We consider the monthly excess stock returns of the U.S. S&P 500 Index as the dependent variable, and the dataset is from January 1994 to August 2021 with total number of observations $n = 332$. The excess stock returns is defined as the

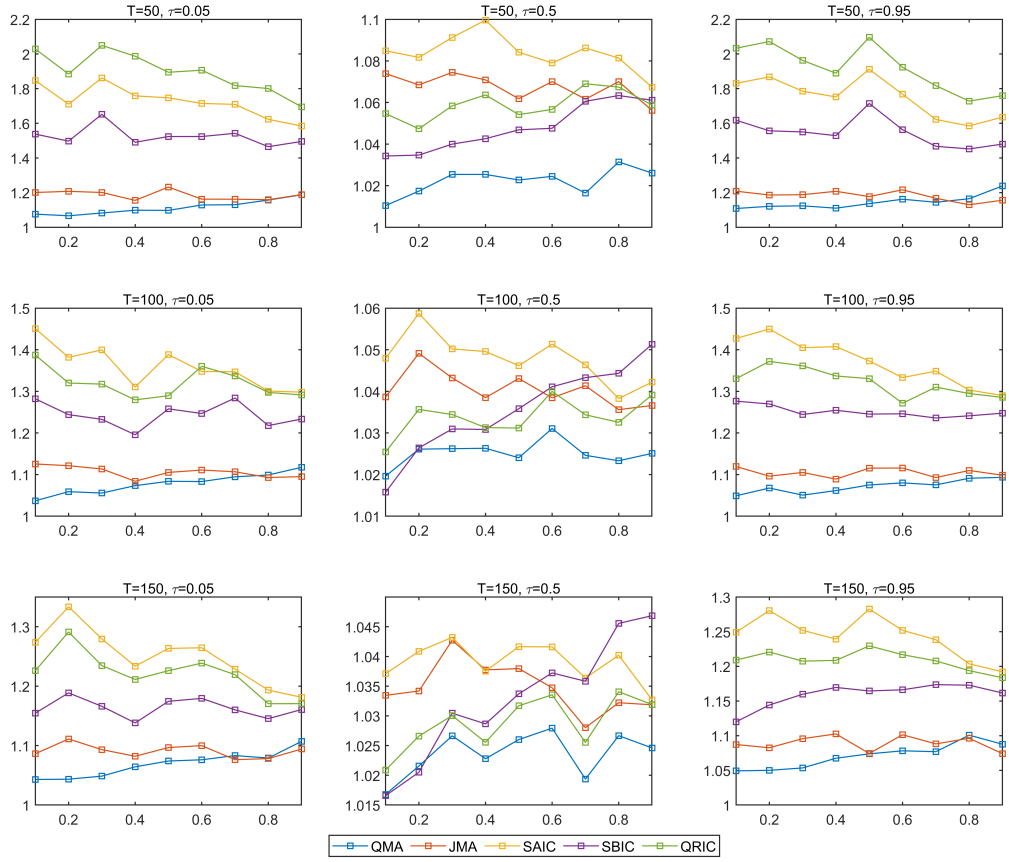

Figure 1: Out-of-sample performance of DGP 1 with homoscedasticity (Setting I).

monthly stock returns minus the treasury-bill rate, a proxy for the risk-free rate. The series of 12 exogenous predictors consist of stock market, bond market, monetary market and macroeconomic activity; see Table 1 for details. Most of these predictors are used in Campbell and Thompson (2007), Jin et al. (2014) and Lu and Su (2015).

We summarize some basic statistics of the exogenous variables in Table 2, including mean, median, standard deviation and data transformation. Note that an augmented Dickey-Fuller (ADF) test is applied to all exogenous variables. We find that all these variables are nonstationary. Thus, we transform these series to stationary time series by a logarithmic differentiation process.

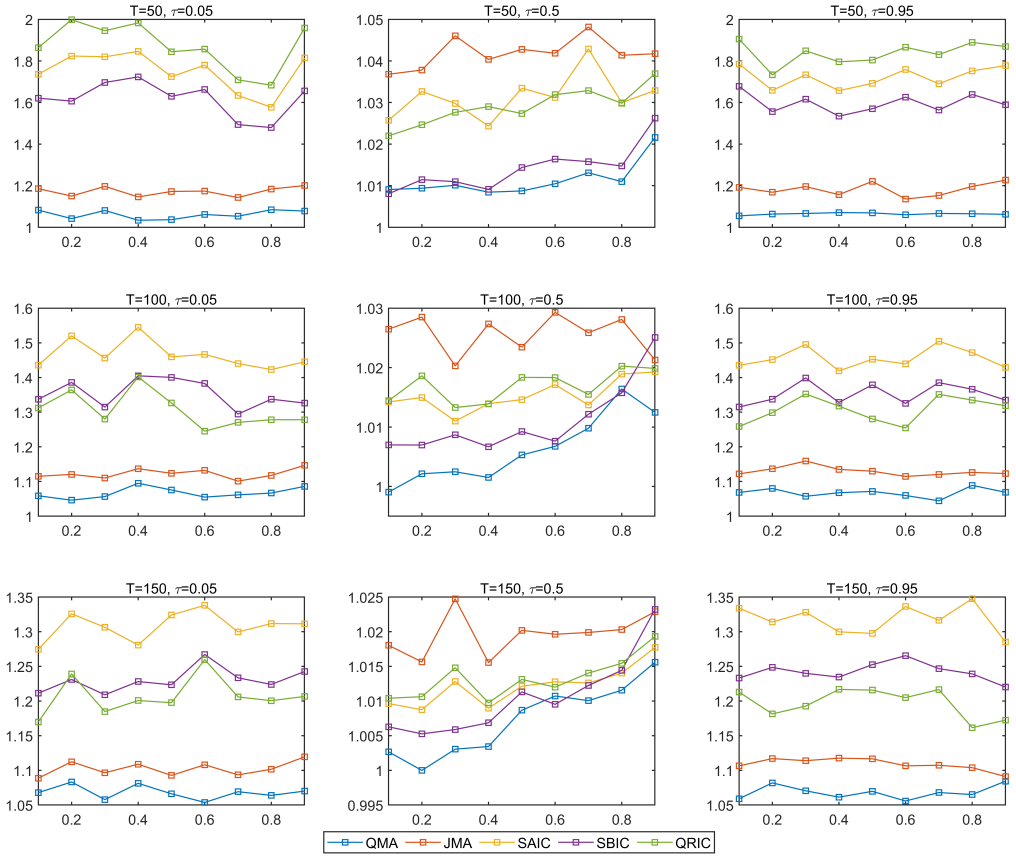

Figure 2: Out-of-sample performance of DGP 1 with heteroskedasticity (Setting II).

We follow Lu and Su (2015) to construct nested candidate models. As the correlation between the exogenous predictor and the dependent variable may perform differently at different quantiles, we follow He et al. (2013) and Wang et al. (2023) to order the predictors by the marginal quantile utility (MQU). The MQU of the j-th predictor $\mathbf{x}_j$ is defined as follows:

$$\text{MQU}(\mathbf{x}_j) = \frac{1}{n}\sum_{t=1}^{n}\left|x_{tj}\widehat{\beta}_j - q(\tau)\right|,$$

where $\widehat{\beta}_j = \text{argmin}_\beta \sum_{t=1}^{n}\rho_\tau(y_t - x_{tj}\beta)$ and $q(\tau)$ is the unconditional quantile of the sample $\{y_t\}$. Suppose the predictors $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{12}\}$ have been ranked in descending order of the

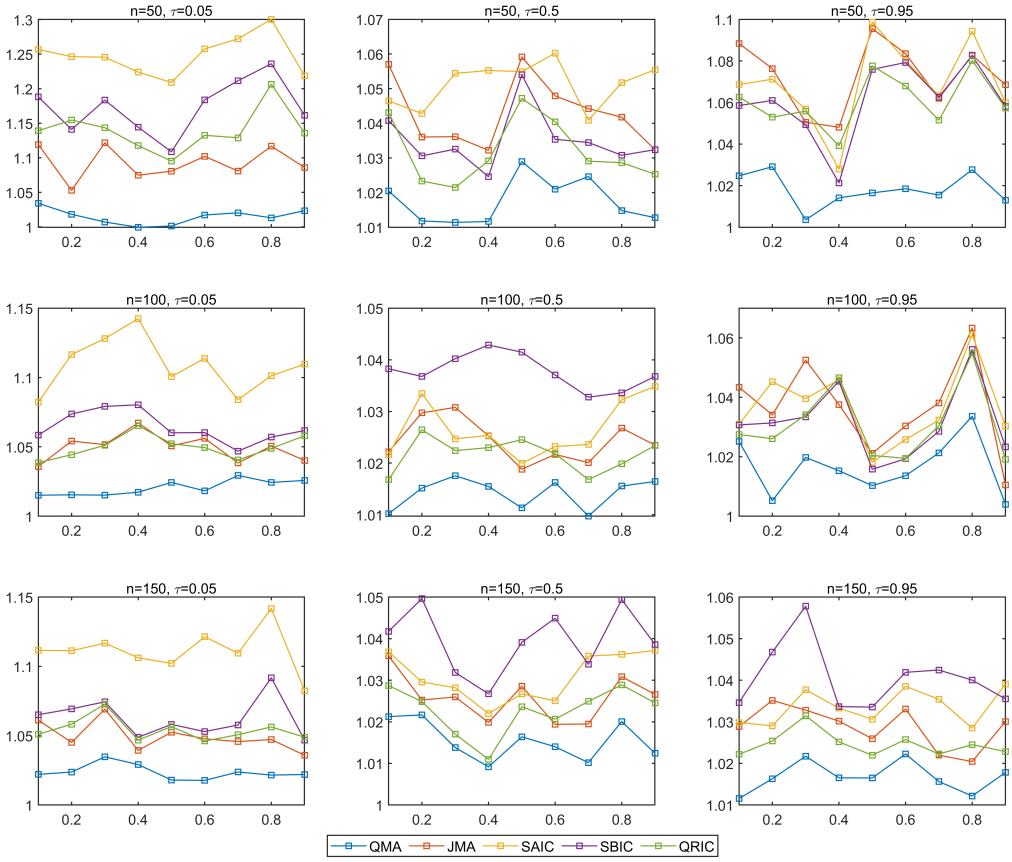

Figure 3: Out-of-sample performance of DGP 2 with homoscedasticity (Setting I).

marginal quantile utility. Then, the 12 nested candidate models are $\{1, \mathbf{x}_1\}$, $\{1, \mathbf{x}_1, \mathbf{x}_2\}$, ..., $\{1, \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{12}\}$, respectively.

We use a rolling window estimation to study the out-of-sample performance of our method. Each time, we forecast the one-period-ahead quantiles of excess stock returns. To verify the robustness of our model to the sample size, we take 7 different rolling window sizes of 48, 60, 72, 84, 96, 108, and 120. For example, to construct the first forecast time at January 1998 with a 48-observation fixed rolling window, we use the available observations from January 1994 to December 1997 to estimate parameters for each candidate model and model averaging weights, and then construct a combined one-step-ahead forecast. When a new observation is available, we add it to the estimation sample and delete the earliest

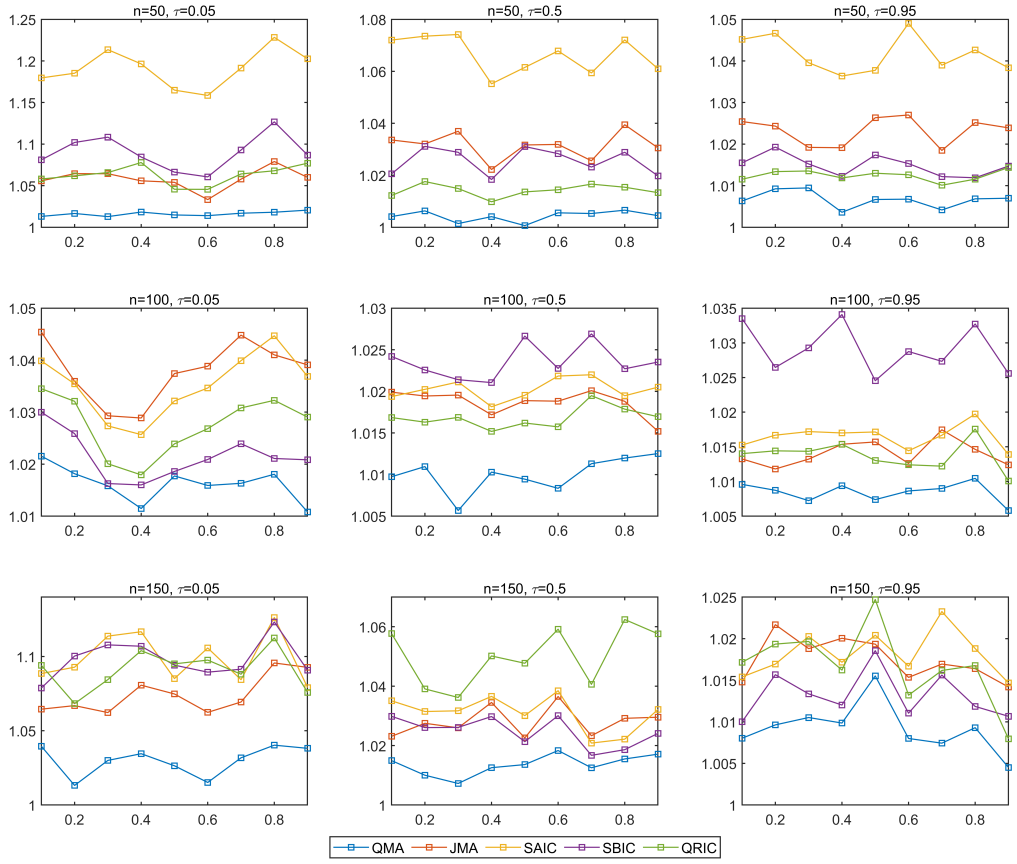

Figure 4: Out-of-sample performance of DGP 2 with heteroskedasticity (Setting II).

one, and recalculate a new set of model averaging weights with a 48-observation fixed rolling window and obtain a combined forecast at February 1998. The one-step-ahead out-of-sample forecast period is spanning from January 1998 to August 2021.

Following Campbell and Thompson (2007) and Lu and Su (2015), we define the out-of-sample R_o^2 of quantile estimations to evaluate the performance of our method,

$$R_o^2 = 1 - \frac{\sum_{t=n_1+1}^{n} \rho_\tau(y_t - \widehat{q}_t(\tau))}{\sum_{t=n_1+1}^{n} \rho_\tau(y_t - q_t(\tau))},$$

where y_t is the true excess stock return, $\widehat{q}_t(\tau)$ is the estimated conditional τ-th quantile of the employed methods, $q_t(\tau)$ is the unconditional τth quantile over the past n_1 samples,

Table 1: Explanatory variables and their explanations

Classification	Variables	Explanations
Stock Market	Dividend Price Ratio	The Dividend Price Ratio of S&P 500 is defined as the difference between the log of dividends and the log of prices, where dividends are 12-month moving sums of dividends.
	Earnings Price Ratio	The Earnings Price Ratio of S&P 500 is defined as the difference between the log of earnings and the log of prices, where earnings are 12-month moving sums of earnings.
	Book-to-market Ratio	The book-to-market ratio is the ratio of the book value to the market value of the Dow Jones Industrial Average.
	Lagged dependent variable	The lagged dependent variable is the one-step lagged value of excess stock returns of S&P 500 index.
Bond Market	Default Yield Spread	Default Yield Spread is the difference between BAA and AAA-rated corporate bond yields, which can be downloaded from Federal Reserve Economic Data.
	Term Spread	The Term Spread is the difference between the long-term yields on government bonds and Treasury bills.
	Treasury-bill Rate	The 3-Month Treasury Bill: Secondary Market Rate.
	Long Term Yield	The U.S. 10-year Treasury yields.
Monetary Market	Exchange Rate	The US Dollar Index.
	Money Supply	The U.S. M2 money supply.
Macroeconomic	Inflation	The U.S. Consumer Price Index (CPI).
	Labor	The U.S. unemployment rate.

and n_1 is the rolling window size. It is obvious that a larger R_o^2 implies a better out-of-sample forecast. Following Kuester et al. (2005), we also employ the violation rate (%Viol) to examine the model performance in terms of VaR violations. The violation rate can be explained as the percentage of the actual values that are smaller than their predicted τ-th quantiles, namely

$$\%\text{Viol} = \frac{1}{n - n_1} \sum_{t=n_1+1}^{n} \mathbf{1}(y_t < \widehat{q}_t(\tau)),$$

which implies that the closer %Viol is to τ, the better the model performs. In addition,

Table 2: Descriptive statistics and ADF test for the explanatory variables

Variables	Mean	Median	std	p-value	Transformation
Dividend Price Ratio	1.9066	1.8962	0.4391	0.1937	$\Delta \ln$
Earnings Price Ratio	20.6769	19.6086	4.7605	0.5608	$\Delta \ln$
Book-to-market Ratio	3.3351	3.2896	1.0964	0.8669	$\Delta \ln$
Default Yield Spread	0.9596	0.8800	0.4011	0.2847	$\Delta \ln$
Term Spread	1.6488	1.5990	1.0698	0.1427	$\Delta \ln$
Treasury-bill Rate	2.2486	1.6531	2.0895	0.2948	$\Delta \ln$
Long Term Yield	3.8752	3.8550	1.7568	0.1713	$\Delta \ln$
Exchange Rate	91.3774	90.5902	10.4205	0.5452	$\Delta \ln$
Money Supply	8.5557	7.4272	4.2679	0.9990	$\Delta \ln$
Inflation	205.7448	209.4860	35.1852	0.9990	$\Delta \ln$
Labor	5.7970	5.4000	1.7875	0.2668	$\Delta \ln$

Note: "Mean", "Median" and "std" denote the sample mean, median and standard deviation of the variables from January 1994 to August 2021. "p-value" is the p-value of ADF test for each variables. $\Delta \ln : X_t = \ln S_t - \ln S_{t-1}$, where $\{S_t\}$ is the original series.

JMA, SAIC, SBIC, and QRIC are employed here to work as the competing methods, and we estimate the conditional quantiles with $\tau = 0.05$, 0.50, and 0.95.

Panel A of Table 3 reports the out-of-sample R_o^2 of our model and the benchmark models. Several observations can be obtained from Panel A. First, it is obvious that our method dominates other competing methods for various rolling window sizes n_1 and τ. For example, when $n_1 = 120$ and $\tau = 0.05$, the out-of-sample R_o^2 of our method is 0.0976, which is the only positive value among all the results provided by the five employed methods. Second, as the sample size increases, the out-of-sample R_o^2 of all methods show an increasing trend. For instance, the out-of-sample R_o^2 increase from -0.0086 to 0.0976 when n_1 is ranged from 48 to 120. Third, JMA performs the second best among all the methods; that is, JMA provides a smaller out-of-sample R_o^2 than SAIC, SBIC, and QRIC. This observation is consistent with the conclusions in Lu and Su (2015). For example, for $n_1 = 48$ and $\tau = 0.05$, the out-of-sample R_o^2 of JMA is -0.1080, which is larger than -0.8387, -0.8002, and -0.8648.

The out-of-sample violation rates are presented in Panel B of Table 3. First, it is observed that QMA is frequently ranked on the top of all competing methods. For example, QMA provides the violation rates closer to τ in almost all cases (except for $\tau = 0.95$ and

Table 3: Evaluation of the excess stock returns

		Panel A: Out-of-sample R_o^2					*Panel B:* Out-of-sample violation rate				
n_1	τ	QMA	JMA	SAIC	SBIC	QRIC	QMA	JMA	SAIC	SBIC	QRIC
48		**-0.0086**	-0.1080	-0.8387	-0.8002	-0.8648	**0.0739**	0.1690	0.3415	0.3275	0.3310
60		**-0.0090**	-0.3520	-1.4105	-1.3919	-1.4250	**0.0846**	0.2243	0.3235	0.3199	0.3456
72		**0.0840**	-0.0341	-1.1869	-1.1975	-1.2993	**0.0654**	0.1731	0.2962	0.3000	0.2923
84	0.05	**0.0712**	0.0119	-1.0052	-0.9454	-1.0063	**0.0565**	0.1452	0.2944	0.2782	0.2823
96		**0.0409**	-0.0205	-0.7433	-0.7747	-0.6862	**0.0508**	0.1483	0.2839	0.2797	0.2627
108		**0.0757**	-0.0467	-0.5859	-0.4928	-0.5832	**0.0446**	0.1205	0.2991	0.2634	0.2723
120		**0.0976**	-0.0678	-0.5803	-0.5771	-0.5635	**0.0613**	0.1415	0.2689	0.2736	0.2689
48		**0.0463**	-0.0083	-0.1587	-0.0816	-0.0934	**0.4930**	0.5246	0.5211	0.4894	0.5246
60		**0.0708**	0.0056	-0.1022	-0.0556	-0.0154	**0.5037**	0.5551	0.5221	0.4926	0.5368
72		**0.0695**	0.0216	-0.1237	-0.0575	0.0042	**0.4923**	0.5269	0.5346	0.4885	0.5423
84	0.50	**0.0977**	0.0481	-0.0965	-0.0142	-0.0197	**0.5000**	0.5363	0.5605	0.5121	0.5524
96		**0.0933**	0.0264	-0.0975	-0.0102	-0.0205	**0.4873**	0.5551	0.5932	0.5424	0.5763
108		**0.1220**	0.0160	-0.1041	0.0104	-0.0041	**0.5000**	0.5357	0.6027	0.5089	0.5670
120		**0.0987**	0.0414	-0.0968	0.0255	0.0023	**0.4906**	0.5472	0.6038	0.5142	0.5802
48		**0.0111**	-0.1855	-1.7511	-1.5417	-1.9565	**0.9120**	0.8662	0.7077	0.7465	0.6796
60		**0.0745**	-0.1133	-0.8708	-0.7666	-0.9093	**0.9412**	0.8676	0.7574	0.7721	0.7426
72		**0.0510**	0.0302	-0.7008	-0.4850	-0.8267	**0.9308**	0.9192	0.7962	0.8154	0.7885
84	0.95	**0.0749**	-0.0004	-0.4844	-0.2645	-0.5124	**0.9395**	0.9073	0.8347	0.8669	0.8468
96		**0.0436**	0.0353	-0.1767	-0.0733	-0.2172	**0.9322**	0.8856	0.8602	0.8602	0.8729
108		**0.1995**	0.1918	-0.0112	0.0913	0.0304	0.9196	0.9152	0.8839	0.9107	**0.9286**
120		**0.1990**	0.1380	-0.0243	-0.0832	-0.0530	**0.9434**	0.9104	0.8726	0.8585	0.8774

Note: Our proposed method is denoted as QMA and the best result for each n_1 and τ is shown in bold.

$n_1 = 108$). Second, no methods except QMA can always outperform other methods in the sense of violation rates. For example, we observe that JMA outperforms SAIC, SBIC and QRIC for $\tau = 0.05, 0.95$ but can not dominate these three methods for $\tau = 0.50$. This highlights the importance of using penalties in model averaging for quantile regression when the time series may have a sparse representation. Furthermore, it is documented that the QMA forecasts perform quite well when $\tau = 0.50$. Especially, the estimated out-of-sample violation rates exactly equal to 0.50 when $n_1 = 84$ and 108 and $\tau = 0.50$.

Overall, the average improvements of forecasts by QMA can be clearly evidenced over other competing methods in terms of all evaluation criteria. Intuitively, the conventional model averaging method is potentially equivalent to a mega model, and some covariates and candidate models of this mega model may play a poor role in predicting conditional quantiles. However, the QMA-based mega model may be parsimonious and yield spareness from various potential predictors thanks to the use of penalties for both weights and predictors, which can eliminate the redundant regressors. Thus, it is highly desirable to consider penalized

model averaging schemes for quantile regression.

7 Conclusion

This paper proposes a novel parsimonious and general model averaging method for high-dimensional quantile regression to reduce model uncertainty and improve forecast accuracy. Both the number of candidate models and the dimension of predictors are allowed to be diverging. The proposed KL based weight choice criterion with penalties selects the optimal combination weights and important predictors simultaneously. We establish the asymptotic optimality and asymptotic consistency of the proposed model averaging estimator. We also develop a model screening process before model averaging for the ultra-high dimensional data. Simulation studies and empirical application to stock returns forecasting illustrate that the proposed method is promising.

The model averaging strategy proposed in this paper could be extend to other contexts. First, it would be interesting to extend our method to the time-varying model averaging for quantile regression with structural changes, which could capture the evolutionary changes of economic structure and improve the time-varying predictive abilities of the quantile forecasting models. Besides, estimating the conditional quantile with complex data, like missing data or censored data has been an important field in quantile regression and attracted much attention; see, for instance, the papers by Wang and Xiao (2022), Chen and Wang (2023), Chernozhukov et al. (2015) and references therein. How to develop a model averaging for quantile regression with complex data is challenging and deserves future research. Finally, we would like to mention that it is very interesting to investigate the quantile model averaging for the case that some regressors might be nonstationary, as addressed in Cai et al. (2023).

Acknowledgments

The authors thank seminar participants in the quantitative economics webinar at Lingnan College Sun Yat-sen University, 2022. All remaining errors are solely ours. Authors thank a partial support from Natural Sciences Foundation of China grants (72073126, 72091212, 71973116, 71631004, 72033008, 72133002) and Young Elite Scientists Sponsorship Program by CAST[YESS20200072].

References

Ando, T. and Li, K.-C. (2014). A model-averaging approach for high-dimensional regression. *Journal of the American Statistical Association*, 109(505):254–265.

Ando, T. and Li, K.-C. (2017). A weight-relaxed model averaging approach for high-dimensional generalized linear models. *Annals of Statistics*, 45(6):2654–2679.

Bertsimas, D. and Mazumder, R. (2014). Least quantile regression via modern optimization. *Annals of Statistics*, 42(6):2494–2525.

Briollais, L. and Durrieu, G. (2014). Application of quantile regression to recent genetic and-omic studies. *Human Genetics*, 133(8):951–966.

Brown, P. J., Vannucci, M., and Fearn, T. (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society: Series B*, 64(3):519–536.

Cade, B. S. and Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1(8):412–420.

Cai, Z. (2002). Regression quantiles for time series. *Econometric Theory*, 18(1):169192.

Cai, Z., Chen, H., and Liao, X. (2023). A new robust inference for predictive quantile regression. *Journal of Econometrics*, 234(1):227–250.

Cai, Z. and Xiao, Z. (2012). Semiparametric quantile regression estimation in dynamic models with partially varying coefficients. *Journal of Econometrics*, 167(2):413–425.

Cai, Z. and Xu, X. (2009). Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association*, 104(485):371–383.

Campbell, J. Y. and Thompson, S. B. (2007). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4):1509–1531.

Chen, J., Li, D., Linton, O., and Lu, Z. (2018). Semiparametric ultra-high dimensional model averaging of nonlinear dynamic time series. *Journal of the American Statistical Association*, 113(522):919–932.

Chen, S. and Wang, Q. (2023). Quantile regression with censoring and sample selection. *Journal of Econometrics*, 234(1).

Chernozhukov, V., Fernndez-Val, I., and Kowalski, A. E. (2015). Quantile regression with censoring and endogeneity. *Journal of Econometrics*, 186(1):201–221.

Engle, R. F. and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles. *Journal of Business & Economic Statistics*, 22(4):367–381.

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

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32(3):928–961.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, 75(4):1175–1189.

Hansen, B. E. and Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, 167(1):38–46.

He, X., Wang, L., and Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Annals of Statistics*, 41(1):342–369.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with comments by M. Clyde, D. Draper and E. I. George, and a rejoinder by the authors). *Statistical Science*, 14(4):382–417.

Honda, T. (2013). Nonparametric quantile regression with heavy-tailed and strongly dependent errors. *Annals of the Institute of Statistical Mathematics*, 65(1):23–47.

Jiang, X., Jiang, J., and Song, X. (2012). Oracle model selection for nonlinear models based on weighted composite quantile regression. *Statistica Sinica*, 22(4):1479–1506.

Jin, S., Su, L., and Ullah, A. (2014). Robustify financial time series forecasting with bagging. *Econometric Reviews*, 33(5-6):575–605.

Koenker, R. (2011). Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25(3):239–262.

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

Koenker, R. and Xiao, Z. (2004). Unit root quantile autoregression inference. *Journal of the American Statistical Association*, 99(467):775–787.

Koenker, R. and Xiao, Z. (2006). Quantile autoregression. *Journal of the American Statistical Association*, 101(475):980–990.

Komunjer, I. (2005). Quasi-maximum likelihood estimation for conditional quantiles. *Journal of Econometrics*, 128(1):137–164.

Kuester, K., Mittnik, S., and Paolella, M. S. (2005). Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, 4(1):53–89.

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

Li, J., Lv, J., Wan, A. T. K., and Liao, J. (2022). Adaboost semiparametric model averaging prediction for multiple categories. *Journal of the American Statistical Association*, 117(537):495–509.

Liang, H., Zou, G., Wan, A. T. K., and Zhang, X. (2011). Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association*, 106(495):1053–1066.

Liao, J., Zong, X., Zhang, X., and Zou, G. (2019). Model averaging based on leave-subject-out cross-validation for vector autoregressions. *Journal of Econometrics*, 209(1):35–60.

Linton, O. and Xiao, Z. (2017). Quantile regression applications in finance. In *Handbook of Quantile Regression*, pages 381–407. Chapman and Hall/CRC.

Lu, X. and Su, L. (2015). Jackknife model averaging for quantile regressions. *Journal of Econometrics*, 188(1):40–58.

Meier, L., Van De Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B*, 70(1):53–71.

Nguyen, L. H., Chevapatrakul, T., and Yao, K. (2020). Investigating tail-risk dependence in the cryptocurrency markets: A LASSO quantile regression approach. *Journal of Empirical Finance*, 58:333–355.

Sun, Y., Hong, Y., Lee, T.-H., Wang, S., and Zhang, X. (2021). Time-varying model averaging. *Journal of Econometrics*, 222(2):974–992.

Sun, Y., Hong, Y., Wang, S., and Zhang, X. (2023). Penalized time-varying model averaging. *Journal of Econometrics*, 235(2):1355–1377.

Sun, Y., Zhang, X., Wan, A. T., and Wang, S. (2022). Model averaging for interval-valued data. *European Journal of Operational Research*, 301(2):772–784.

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

Wan, A. T., Zhang, X., and Zou, G. (2010). Least squares model averaging by mallows criterion. *Journal of Econometrics*, 156(2):277–283.

Wang, M., Zhang, X., Wan, A. T. K., You, K., and Zou, G. (2023). Jackknife model averaging for high-dimensional quantile regression. *Biometrics*, 79(1):178–189.

Wang, Y. and Xiao, Z. (2022). Estimation and inference about tail features with tail censored data. *Journal of Econometrics*, 230(2):363–387.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.

White, H. (1984). *Asymptotic Theory for Econometricians*. Academic Press.

Xiao, Z. and Koenker, R. (2009). Conditional quantile estimation for generalized autoregressive conditional heteroscedasticity models. *Journal of the American Statistical Association*, 104(488):1696–1712.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.

Yuan, Z. and Yang, Y. (2005). Combining linear regression models. *Journal of the American Statistical Association*, 100(472):1202–1214.

Zhang, X., Lu, Z., and Zou, G. (2013). Adaptively combined forecasting for discrete response time series. *Journal of Econometrics*, 176(1):80–91.

Zhang, X., Ma, Y., and Carroll, R. J. (2019). Malmem: model averaging in linear measurement error models. *Journal of the Royal Statistical Society: Series B*, 81(4):763–779.

Zhang, X., Yu, D., Zou, G., and Liang, H. (2016). Optimal model averaging estimation for generalized linear models and generalized linear mixed-effects models. *Journal of the American Statistical Association*, 111(516):1775–1790.

Zhang, X., Zou, G., Liang, H., and Carroll, R. J. (2020). Parsimonious model averaging with a diverging number of parameters. *Journal of the American Statistical Association*, 115(530):972–984.

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

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Annals of Statistics*, 36(3):1108–1126.

Appendix

Appendix A

Following Fan and Peng (2004), to obtain the consistency of $\widehat{\boldsymbol{\theta}}\tau^{(m)}$ with diverging dimensions, Appendix A contains the additional conditions.

Condition (A.1). *f belongs to the tick-exponential family. And $\ln f(y_i|\boldsymbol{\theta}_\tau^{(m)}, \mathbf{x}_i)$ is continuously differentiable with probability 1.*

Condition (A.2). *Θ is compact, and $\boldsymbol{\theta}_\tau^{(m)*}$ is interior points of Θ.*

Condition (A.3). *For the mth candidate model, k_m is allowed to be diverging and satisfies $k_m^4/n \to 0$.*

Condition (A.4). *The log likelihood function satisfies the following conditions with probability 1: The Fisher information matrix*

$$I(\boldsymbol{\theta}_\tau^{(m)}) = \mathbb{E}\left[\frac{\partial \ln f(y_i|\boldsymbol{\theta}_\tau^{(m)}, \mathbf{x}_i)}{\partial \boldsymbol{\theta}_\tau^{(m)}} \frac{\partial \ln f(y_i|\boldsymbol{\theta}_\tau^{(m)}, \mathbf{x}_i)}{\partial \boldsymbol{\theta}_\tau^{(m)\prime}}\right]$$

satisfies $0 < C_1 < \lambda_m in\{I(\boldsymbol{\theta}_\tau^{(m)})\} \leq \lambda_m ax\{I(\boldsymbol{\theta}_\tau^{(m)})\} < C_2 < \infty$. And for $j, k = 1, ..., k_m$,

$$\mathbb{E}\left[\frac{\partial \ln f(y_i|\boldsymbol{\theta}_\tau^{(m)}, \mathbf{x}_i)}{\partial \theta_{\tau,j}^{(m)}} \frac{\partial \ln f(y_i|\boldsymbol{\theta}_\tau^{(m)}, \mathbf{x}_i)}{\partial \theta_{\tau,k}^{(m)}}\right]^2 < C_3 < \infty,$$

and

$$\mathbb{E}\left[\frac{\partial^2 \ln f(y_i|\boldsymbol{\theta}_\tau^{(m)}, \mathbf{x}_i)}{\partial \theta_{\tau,j}^{(m)} \partial \theta_{\tau,k}^{(m)}}\right].$$

Condition (A.5). *There is a large enough open subset $\Omega \in \Theta$, which contains $\boldsymbol{\theta}_\tau^{(m)*}$, such that for almost all $(y_i, \mathbf{x}_i)$ $\partial \ln f(y_i|\boldsymbol{\theta}_\tau^{(m)}, \mathbf{x}_i)/\partial \theta_{\tau,j}^{(m)} \partial \theta_{\tau,k}^{(m)} \partial \theta_{\tau,l}^{(m)}$ exists for all $\boldsymbol{\theta}_\tau^{(m)} \in \Omega$, with probability 1. And there are functions $M_{jkl}(\mathbf{x}_i)$ sunch that*

$$\left|\partial \ln f(y_i|\boldsymbol{\theta}_\tau^{(m)}, \mathbf{x}_i)/\partial \theta_{\tau,j}^{(m)} \partial \theta_{\tau,k}^{(m)} \partial \theta_{\tau,l}^{(m)}\right| \leq M_{jkl}(\mathbf{x}_i)$$

for all $\boldsymbol{\theta}_\tau^{(m)} \in \Omega$, and $\mathbb{E}[M_{jkl}^2(\mathbf{x}_i)] < C_5 < \infty$, for all k_m, n and j, k, l.

Appendix B

Appendix B contains the additional results for simulations for a comparison. Note that the conclusions are made at the end of each figure.

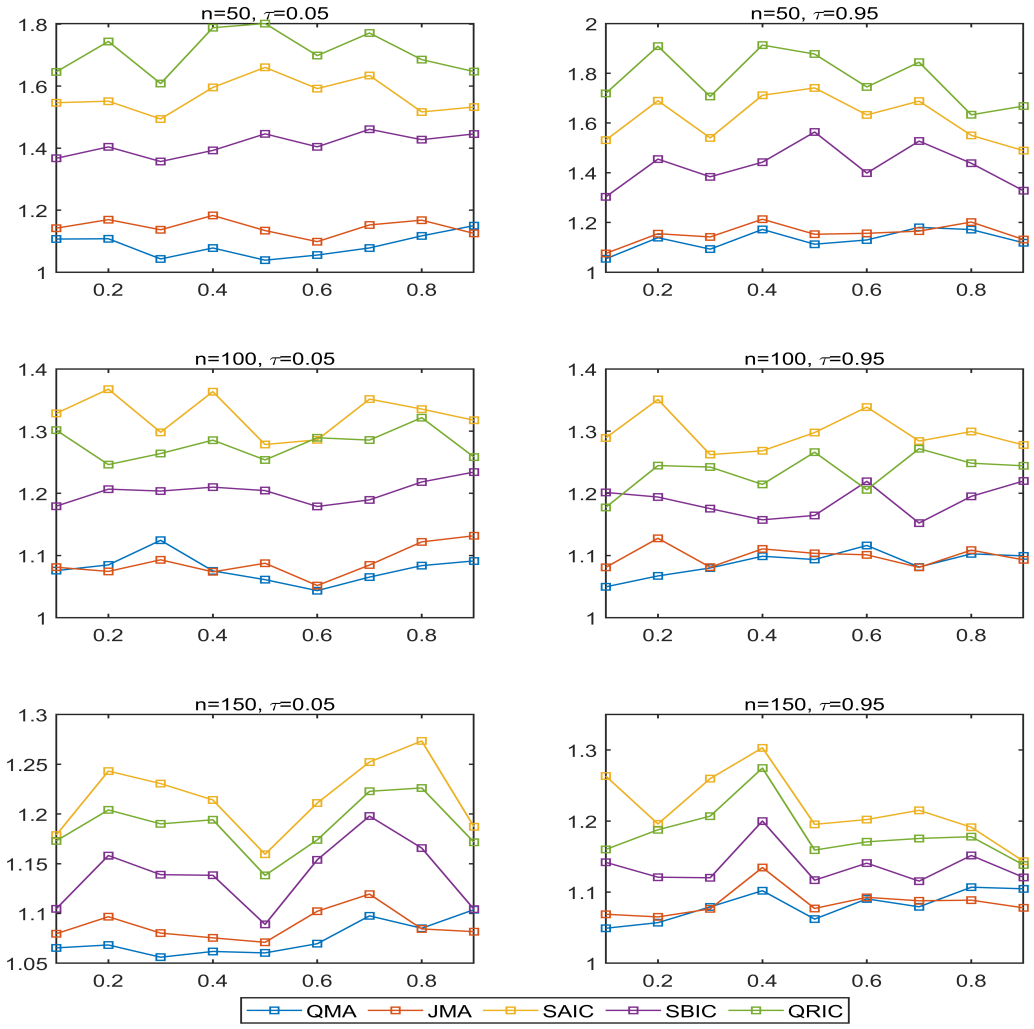

Figure B1: Out-of-sample performance with homoscedasticity (Setting I).

Note: This DGP is the same as DGP 1 in Lu and Su (2015). The result with $\tau = 0.50$ is not shown because Lu and Su (2015) has stated that no method clearly dominated the others when $\tau = 0.50$. It can be observed that JMA, SAIC, SBIC and QRIC provide similar results as in Lu and Su (2015); namely JMA clearly dominate other existing methods. Also, note that our method performs slightly better than JMA in most cases.

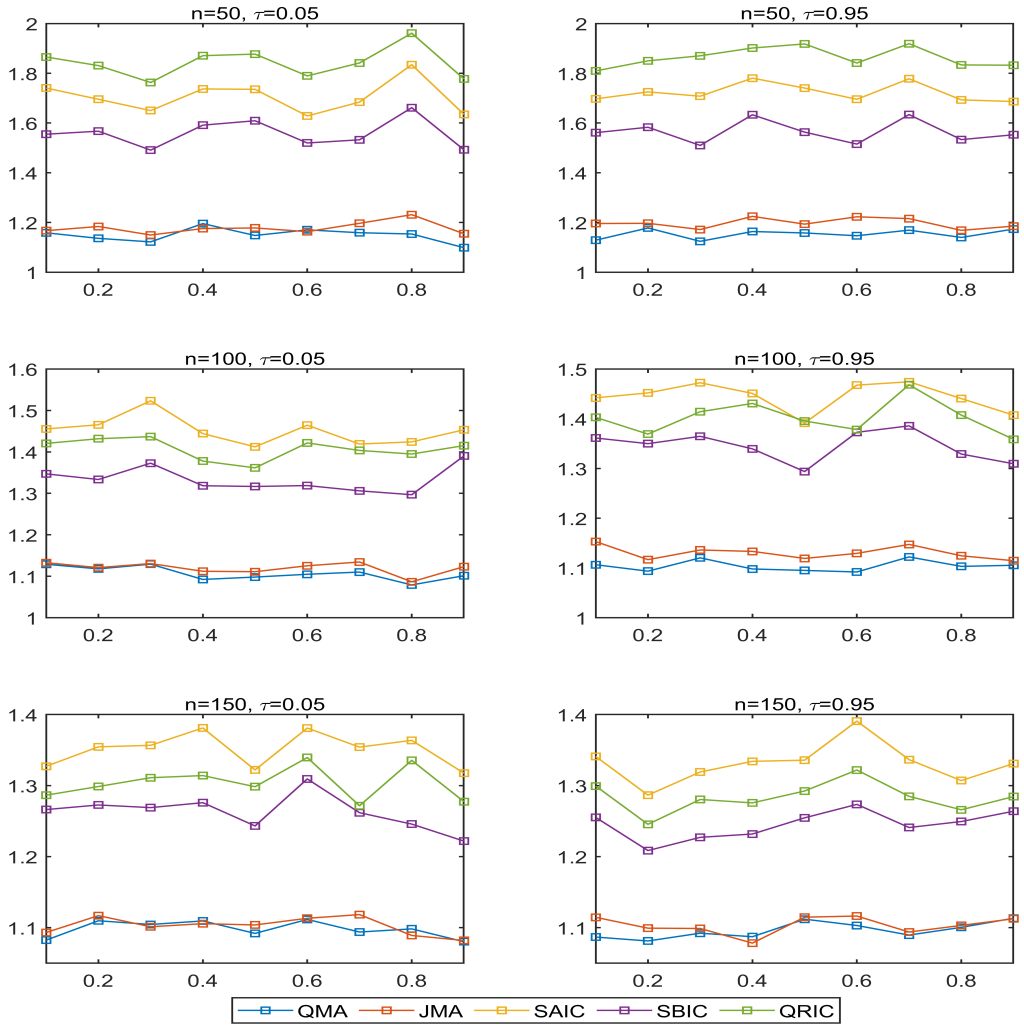

Figure B2: Out of sample performance with heteroscedasticity (Setting II).

Note: This DGP is the same as DGP 1 in Lu and Su (2015). The result with $\tau = 0.50$ is not shown because Lu and Su (2015) has stated that no method clearly dominated the others when $\tau = 0.50$. It can be observed that JMA, SAIC, SBIC and QRIC provide similar results as in Lu and Su (2015); namely JMA clearly dominates other existing methods. Also, note that our method performs slightly better than JMA in most cases.

Appendix C

In this section, we present the detailed proofs of Lemma 1 and Theorems 1 - 4.

Proof of Lemma 1. Let $\alpha_n = \sqrt{k_m/n}$ and set $\|\mathbf{u}\| = C$, where C is a large enough constant. To prove (8), it is equivalent to prove for any given ϵ there is a large C and n such that

$$P\{\sup_{\|\mathbf{u}\|=C} L_n(\boldsymbol{\theta}_\tau^{(m)*} + \alpha_n \mathbf{u}) < L_n(\boldsymbol{\theta}_\tau^{(m)*})\} > 1 - \epsilon. \tag{C.1}$$

Notice that the conditions about the differentiability of $\ln f(y_i|\boldsymbol{\theta}_\tau^{(m)}, \mathbf{x}_i)$ are satisfied with probability 1 and the set of non-differentiability has no effect on (C.1). More details can be found in Remark 5.

Now, we have

$$\begin{aligned} D_n(\mathbf{u}) &= L_n(\boldsymbol{\theta}_\tau^{(m)*} + \alpha_n \mathbf{u}) - L_n(\boldsymbol{\theta}_\tau^{(m)*}) \\ &= \alpha_n \nabla' L_n(\boldsymbol{\theta}_\tau^{(m)*})\mathbf{u} + \frac{1}{2}\mathbf{u}'\nabla^2 L_n(\boldsymbol{\theta}_\tau^{(m)*})\mathbf{u}\alpha_n^2 + \frac{1}{6}\nabla'[\mathbf{u}'\nabla^2 L_n(\tilde{\boldsymbol{\theta}}_\tau^{(m)})\mathbf{u}]\mathbf{u}\alpha_n^3 \\ &= \Psi_1 + \Psi_2 + \Psi_3, \end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_\tau^{(m)}$ is between $\boldsymbol{\theta}_\tau^{(m)*}$ and $\boldsymbol{\theta}_\tau^{(m)*} + \alpha_n \mathbf{u}$. Following the spirit of Fan and Peng (2004), we have

$$|\Psi_1| = |\alpha_n \nabla' L_n(\boldsymbol{\theta}_\tau^{(m)*})\mathbf{u}| \le \alpha_n \|\nabla' L_n(\boldsymbol{\theta}_\tau^{(m)*})\| \cdot \|\mathbf{u}\| = O_p(\alpha_n \sqrt{n k_m})\|\mathbf{u}\|.$$

For Ψ_2, by Chebyshev inequality and Lemma 8 of Fan and Peng (2004), it can be observed that

$$\begin{aligned} \Psi_2 &= \frac{1}{2}\mathbf{u}'[\frac{1}{n}(\nabla^2 L_n(\boldsymbol{\theta}_\tau^{(m)*}) - \mathbb{E}[\nabla^2 L_n(\boldsymbol{\theta}_\tau^{(m)*})])]\mathbf{u}n\alpha_n^2 - \frac{1}{2}\mathbf{u}' I(\boldsymbol{\theta}_\tau^{(m)*})\mathbf{u}n\alpha_n^2 \\ &= -\frac{1}{2}n\alpha_n^2 \mathbf{u}' I(\boldsymbol{\theta}_\tau^{(m)*})\mathbf{u} + o_p(1)n\alpha_n^2\|\mathbf{u}\|^2. \end{aligned}$$

Then, for Ψ_3, with Condition (A.5), we have

$$|\Psi_3| = \left|\frac{1}{6}\sum_{j,k,l=1}^{k_m}\frac{\partial L_n(\tilde{\boldsymbol{\theta}}_\tau^{(m)})}{\partial\theta_{\tau,j}^{(m)}\partial\theta_{\tau,k}^{(m)}\partial\theta_{\tau,l}^{(m)}}u_j u_k u_l \alpha_n^3\right| \leq \frac{1}{6}\sum_{t=1}^{n}\left[\sum M_{jkl}^2(\mathbf{x}_t)\right]^{1/2}\|\mathbf{u}\|^3\alpha_n^3 = O_p(nk_m^{2/3}\alpha_n^3)\|\mathbf{u}\|^3.$$

With Conditions (A.1)-(A.5) and allowing $\|\mathbf{u}\|$ to be large enough, it can be obtained that all the terms of Ψ_1, Ψ_2 and Ψ_3 are dominated by Ψ_2, which is negative. This implies that (C.1) is true. □

Proof of Theorem 1. Let $\widetilde{G}_n = G_n + \sum_{i=1}^n \mathbb{E}_0[\ln f_0(y_i|\mathbf{x}_i)]$. It is obvious that

$$\widehat{\mathbf{w}} = \text{argmin}_{\mathbf{w}\in\mathcal{W}}\widetilde{G}_n(\mathbf{w}). \tag{C.2}$$

From Zhang et al. (2016), Theorem 1 is equivalent to prove the following problems:

$$\sup_{w\in\mathcal{W}}\left[\frac{|\text{KL}(\mathbf{w}) - \text{KL}^*(\mathbf{w})|}{\text{KL}^*(\mathbf{w})}\right] = o_p(1), \tag{C.3}$$

and

$$\sup_{w\in\mathcal{W}}\left[\frac{|\widetilde{G}_n(\mathbf{w}) - \text{KL}^*(\mathbf{w})|}{\text{KL}^*(\mathbf{w})}\right] = o_p(1). \tag{C.4}$$

With conditions of this theorem, we have

$$\begin{aligned}\|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}) - \boldsymbol{\theta}^*(\mathbf{w})\| &= \left\|\sum_{m=1}^{M} w_m(\widehat{\boldsymbol{\theta}}_\tau^{(m)} - \boldsymbol{\theta}_\tau^{(m)*})\right\| \leq \sum_{m=1}^{M} w_m\|\widehat{\boldsymbol{\theta}}_\tau^{(m)} - \boldsymbol{\theta}_\tau^{(m)*}\| \\ \leq & M^{1/2}\|\mathbf{w}\|\cdot\|\widehat{\boldsymbol{\theta}}_\tau^{(m)} - \boldsymbol{\theta}_\tau^{(m)*}\| = O_p(M)\cdot O_p(p^{1/2}n^{-1/2}) = O_p(Mp^{1/2}n^{-1/2}).\end{aligned} \tag{C.5}$$

Assumption 2 implies that $\mathbb{E}_0\left\|\frac{\partial \ln f(y_i|\boldsymbol{\theta},\mathbf{x}_i,\tau)}{\partial\boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}(\mathbf{w})}\right\| = O_p(1)$. Then, we have,

$$\sum_{i=1}^{n}\int f_0(y_i|\mathbf{x}_i)\left\|\frac{\partial \ln f(y_i|\boldsymbol{\theta},\mathbf{x}_i,\tau)}{\partial\boldsymbol{\theta}'}|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}(\mathbf{w})}\right\|dy = O_p(n). \tag{C.6}$$

With Taylor formula and non-negativity of density function f_0, the numerator of (C.3) can

be expressed as

$$
\begin{aligned}
&|\mathrm{KL}(\mathbf{w}) - \mathrm{KL}^*(\mathbf{w})| \\
=&\left| \sum_{i=1}^{n} \int [f_0(y_i|\mathbf{x}_i) \ln f(y_i|\boldsymbol{\theta}^*(\mathbf{w}), \mathbf{x}_i, \tau) - f_0(y_i|\mathbf{x}_i) \ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)] dy \right| \\
\leq& \sum_{i=1}^{n} \int f_0(y_i|\mathbf{x}_i) \left\| \frac{\partial \ln f(y_i|\boldsymbol{\theta}, \mathbf{x}_i, \tau)}{\partial \boldsymbol{\theta}'} |_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}(\mathbf{w})} \right\| \cdot \left\| \widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}) - \boldsymbol{\theta}^*(\mathbf{w}) \right\| dy \\
=& \left\| \widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}) - \boldsymbol{\theta}^*(\mathbf{w}) \right\| \cdot \sum_{i=1}^{n} \int f_0(y_i|\mathbf{x}_i) \left\| \frac{\partial \ln f(y_i|\boldsymbol{\theta}, \mathbf{x}_i, \tau)}{\partial \boldsymbol{\theta}'} |_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}(\mathbf{w})} \right\| dy. \qquad \text{(C.7)}
\end{aligned}
$$

By (C.5) and (C.6), we obtain that $|\mathrm{KL}(\mathbf{w}) - \mathrm{KL}^*(\mathbf{w})| = O_p(M\sqrt{np})$. With Assumption 1, applying central limit theorem to $\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)$ leads to

$$
\left| \frac{1}{n} \sum_{i=1}^{n} \ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau) - \mathbb{E}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)] \right| = O_p(1/\sqrt{n}), \qquad \text{(C.8)}
$$

where

$$
\mathbb{E}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)] = \iint \ln f(y|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}, \tau) f_0(y|\mathbf{x}_i) \mathbf{g}_0(\mathbf{x}) d\mathbf{x} dy. \qquad \text{(C.9)}
$$

Similarly, for $\mathbb{E}_{y|\mathbf{x}}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)|\mathbf{x}_i] \equiv \int \ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau) \ln f_0(y|\mathbf{x}_i) dy$, we have

$$
\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{y|\mathbf{x}}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)|\mathbf{x}_i] - \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{y|\mathbf{x}}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)|\mathbf{x}_i]] \right| = O_p(1/\sqrt{n}), \qquad \text{(C.10)}
$$

where

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{x}}[\mathbb{E}_{y|\mathbf{x}}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)|\mathbf{x}_i]] = \int \mathbf{g}_0(\mathbf{x}) \mathbb{E}_{y|\mathbf{x}}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)|\mathbf{x}_i] d\mathbf{x} \\
&= \int \mathbf{g}_0(\mathbf{x}) \int f_0(y|\mathbf{x}_i) \ln f(y|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}, \tau) dy d\mathbf{x} = \iint \ln f(y|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}, \tau) f_0(y|\mathbf{x}_i) \mathbf{g}_0(\mathbf{x}) d\mathbf{x} dy. \qquad \text{(C.11)}
\end{aligned}
$$

A combination of (C.9) and (C.11) gives that

$$
\mathbb{E}_{\mathbf{x}}[\mathbb{E}_{y|\mathbf{x}}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)|\mathbf{x}_i]] = \mathbb{E}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)],
$$

so that the numerator of (C.4) can be expressed as

$$
\begin{aligned}
&|\widetilde{G}_n(\mathbf{w}) - \mathrm{KL}^*(\mathbf{w})| \\
=&\Big| -\sum_{i=1}^{n} \ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau) + \sum_{i=1}^{n} \int f_0(y_i|\mathbf{x}_i) \ln f(y_i|\boldsymbol{\theta}^*(\mathbf{w}), \mathbf{x}_i, \tau) dy \\
&+ \lambda_{n,1}\mathbf{w}'\mathbf{k} + \sum_{j=1}^{p} p_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w})|)\Big| \\
\leq&\Big| -\sum_{i=1}^{n} \ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau) + \sum_{i=1}^{n} \mathbb{E}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)]\Big| \\
&+ \Big| \sum_{i=1}^{n} \mathbb{E}_{y|\mathbf{x}}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)|\mathbf{x}_i] - \sum_{i=1}^{n} \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{y|\mathbf{x}}[\ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau)|\mathbf{x}_i]]\Big| \\
&+ \Big| -\sum_{i=1}^{n} \int f_0(y_i|\mathbf{x}_i) \ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau) dy + \sum_{i=1}^{n} \int f_0(y_i|\mathbf{x}_i) \ln f(y_i|\boldsymbol{\theta}^*(\mathbf{w}), \mathbf{x}_i, \tau) dy\Big| \\
&+ |\lambda_{n,1}\mathbf{w}'\mathbf{k}| + \sum_{j=1}^{p} p_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w})|). \qquad \text{(C.12)}
\end{aligned}
$$

Again, it can be obtained from Assumption 4 that

$$
\lambda_{n,1}\mathbf{w}'\mathbf{k} = \lambda_{n,1} \sum_{m=1}^{M} w_m k_m \leq \lambda_{n,1} p \sum_{m=1}^{M} w_m = O_p(M\sqrt{np}), \qquad \text{(C.13)}
$$

and

$$
\begin{aligned}
\sum_{j=1}^{p} p_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w})|) &= \sum_{j=1}^{p} p_{\lambda_{n,2}}(\mathbf{0}) + \sum_{j=1}^{p} \sum_{m=1}^{M} p'_{\lambda_{n,2}}(\mathbf{0})|\theta_{\tau,j}^{(m)}|w_m(1+o(1)) \\
&= O(p) + pMO_p(n^{1/2}p^{-1/2})(1+o(1)) \leq O_p(M\sqrt{np}). \qquad \text{(C.14)}
\end{aligned}
$$

Clearly, (C.8) and (C.10) conclude that the convergence rate of the first two absolute value is $O_p(\sqrt{n})$, the third part of (C.12) has been proved to be $O_p(M\sqrt{np})$ in (C.7), and (C.13) and (C.14) insure that $\lambda_{n,1}\mathbf{w}'\mathbf{k} = O_p(M\sqrt{np})$ and $\sum_{j=1}^{p} p_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w})|) \leq O_p(M\sqrt{np})$. Therefore, $|\widetilde{G}_n(\mathbf{w}) - \mathrm{KL}^*(\mathbf{w})| = O_p(M\sqrt{np})$. Finally, with Assumption 3, i.e., $M^2np\zeta_n^{-2} = o_p(1)$, (C.3) and (C.4) can be obtained easily from (C.7) and (C.12). This completes the proof of Theorem 1. □

Proof of Theorem 2. Let $\alpha_n = O_p(\xi_n^{1/2} n^{-1/2+\delta/2})$ and set $\|\mathbf{u}\| = C$, where C is a large enough constant. Following Fan and Peng (2004) and Chen et al. (2018), to prove the theorem, our aim is to show that for any given ϵ there is a large enough constant C such that, for large n we have

$$P\left\{\sup_{\|u\|=C, \mathbf{w}^*+\alpha_n\mathbf{u}\in\mathcal{W}} G_n(\mathbf{w}^*+\alpha_n\mathbf{u}) \geq G_n(\mathbf{w}^*)\right\} \geq 1-\epsilon,$$

where $G_n(\cdot)$ is the weight choice criterion defined as in (5), which implies that with probability tending to 1 there exists a minimum $\widehat{\mathbf{w}}_n$ in the ball $\{\mathbf{w}^*+\alpha_n\mathbf{u} : \|\mathbf{u}\| \geq C\}$, such that $\|\widehat{\mathbf{w}}_n - \mathbf{w}^*\| = O_p(\alpha_n)$. In the proof of lemma 1, note that the set of non-differentiability has no effect on our proof process. For the sake of notes, we introduce these notations in the following proof: $L_n(\mathbf{w}^*) = \sum_{i=1}^n \ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}^*), \mathbf{x}_i)$, $f_{i*} = f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}^*), \mathbf{x}_i)$ being the estimated conditional density function of y, $f_{i0} = f_0(y_i|\mathbf{x}_i)$ being the true conditional density function of y, and $\mathbb{E}[\cdot]$ being the expectation with true conditional density function f_{i0}.

By Taylor expansion of $L_n(\mathbf{w}^*+\alpha_n\mathbf{u})$ and $p_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w}^*+\alpha_n\mathbf{u})|)$ at $\mathbf{w}^*$, we have

$$\begin{aligned}
D_n(\mathbf{u}) =& G_n(\mathbf{w}^*+\alpha_n\mathbf{u}) - G_n(\mathbf{w}^*) \\
=& -\left(L_n(\mathbf{w}^*+\alpha_n\mathbf{u}) - L_n(\mathbf{w}^*)\right) + \lambda_{n,1}(\mathbf{w}^*+\alpha_n\mathbf{u})'k - \lambda_{n,1}\mathbf{w}^{*'}\mathbf{k} \\
&+ \sum_{j=1}^p p_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w}^*+\alpha_n\mathbf{u})|) - \sum_{j=1}^p p_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w}^*)|) \\
=& -\left(L_n(\mathbf{w}^*+\alpha_n\mathbf{u}) - L_n(\mathbf{w}^*)\right) + \lambda_{n,1}\alpha_n\mathbf{u}'\mathbf{k} \\
&+ \sum_{j=1}^p \left[p_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w}^*+\alpha_n\mathbf{u})|) - p_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w}^*)|)\right] \\
=& -\alpha_n\nabla' L_n(\mathbf{w}^*)\mathbf{u} - \frac{1}{2}\alpha_n^2\mathbf{u}'\nabla^2 L_n(\mathbf{w}^*)\mathbf{u} - \frac{1}{6}\alpha_n^3\nabla'[\mathbf{u}'\nabla^2 L_n(\tilde{\mathbf{w}})\mathbf{u}]\mathbf{u} + \lambda_{n,1}\alpha_n\mathbf{u}'\mathbf{k} \\
&+ \sum_{j=1}^p \left[\sum_{m=1}^M p'_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w}^*)|)\widehat{\theta}_{\tau,j}^{(m)}\mathrm{sgn}(\widehat{\theta}_{\tau,j}^{(m)})\alpha_n u_m + \sum_{m=1}^M p''_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w}^*)|)\widehat{\theta}_{\tau,j}^{(m)^2}\alpha_n^2 u_m^2\{1+o(1)\}\right] \\
=& -\Xi_{n1} - \Xi_{n2} + Xi_{n3} + \Xi_{n4},
\end{aligned}$$

where $\tilde{\mathbf{w}}$ lies between $\mathbf{w}^*$ and $\mathbf{w}^*+\alpha_n\mathbf{u}$. Next, we consider each term Ξ_{nj} for $1 \leq j \leq 4$.

First, we consider Ξ_{n1}. By Assumption 5, we have

$$\begin{aligned}\Xi_{n1} &= \alpha_n \nabla' L_n(\mathbf{w}^*)\mathbf{u} = \alpha_n n[\frac{1}{n}(\nabla' L_n(\mathbf{w}^*) - \mathbb{E}[\nabla' L_n(\mathbf{w}^*)])]\mathbf{u} + \alpha_n \mathbb{E}[\nabla' L_n(\mathbf{w}^*)]\mathbf{u} \\ &= \Xi_{n11} + \Xi_{n12}.\end{aligned}$$

Similar to the proof of Lemma 8 in Fan and Peng (2004), by applying Chebyshev inequality to Ξ_{n11}, for any $\epsilon > 0$, we have

$$\begin{aligned}&P\big(\|\frac{1}{n}(\nabla' L_n(\mathbf{w}^*) - \mathbb{E}[\nabla' L_n(\mathbf{w}^*)])\| \geq \epsilon n^{-1/2+\delta/2}\big|\mathbf{x}\big) \\ \leq &\frac{1}{n^2\epsilon^2 n^{-1+\delta}}\mathbb{E}\bigg[\sum_{m=1}^{M}\bigg(\frac{\partial L_n(\mathbf{w}^*)}{\partial w_m} - \mathbb{E}\bigg[\frac{\partial L_n(\mathbf{w}^*)}{\partial w_m}\bigg]\bigg)^2\bigg|\mathbf{x}\bigg] = O\left(\frac{M}{\epsilon^2 n^\delta}\right) = o(1),\end{aligned}$$

because the last equality holds because of Assumption 7, i.e., $M^{1/2}n^{-\delta/2} \to 0$, which implies that

$$\left\|\frac{1}{n}(\nabla' L_n(\mathbf{w}^*) - \mathbb{E}[\nabla' L_n(\mathbf{w}^*)])\right\| = o_p(n^{-1/2+\delta/2}).$$

Then,

$$|\Xi_{n11}| \leq o_p(n^{-1/2+\delta/2})n\alpha_n\|\mathbf{u}\| \leq o_p(n^{1/2+\delta/2}\alpha_n)\|\mathbf{u}\|.$$

For Ξ_{n12}, the fact that $\int \frac{\partial \ln f_{i*}}{\partial \mathbf{w}} f_{i*} dy = \mathbf{0}$, which comes from f_{i*} being the density function i.e., $\int f_{i*} dy = 1$, concludes that

$$\begin{aligned}\|\mathbb{E}[\nabla' L_n(\mathbf{w}^*)]\| &= \left\|\sum_{i=1}^{n}\int \frac{\partial \ln f_{i*}}{\partial \mathbf{w}} f_{i0} dy\right\| \\ &= \left\|\sum_{i=1}^{n}\left(\int \frac{\partial \ln f_{i*}}{\partial \mathbf{w}} f_{i0} dy - \int \frac{\partial \ln f_{i*}}{\partial \mathbf{w}} f_{i*} dy\right) + \sum_{i=1}^{n}\int \frac{\partial \ln f_{i*}}{\partial \mathbf{w}} f_{i*} dy\right\| \\ &= \left\|\sum_{i=1}^{n}\int \frac{\partial \ln f_{i*}}{\partial \mathbf{w}}(f_{i0} - f_{i*}) dy\right\| = \left\|\sum_{i=1}^{n}\int \frac{\partial \ln f_{i*}}{\partial \mathbf{w}}\sqrt{f_{i0}} \cdot \frac{f_{i0} - f_{i*}}{\sqrt{f_{i0}}} dy\right\| \\ &\leq \sum_{i=1}^{n}\left\{\sum_{m=1}^{M}\left[\int\left(\frac{\partial \ln f_{i*}}{\partial w_m}\right)^2 f_{i0} dy\right]\left[\int\left(\frac{f_{i0} - f_{i*}}{f_{i0}}\right)^2 f_{i0} dy\right]\right\}^{1/2} \\ &= \sum_{i=1}^{n}\left\{\sum_{m=1}^{M}\left[\int\left(\frac{\partial \ln f_{i*}}{\partial w_m}\right)^2 f_{i0} dy\right]\left[\int\left(1 - \frac{f_{i*}}{f_{i0}}\right)^2 f_{i0} dy\right]\right\}^{1/2}\end{aligned}$$

$$= \sum_{i=1}^{n} \Big\{ \sum_{m=1}^{M} \mathbb{E}\Big[\Big(\frac{\partial \ln f_{i*}}{\partial w_m} \Big)^2 \Big] \Big(\mathbb{E}\Big[\Big(\frac{f_{i*}}{f_{i0}} - 1 \Big)^2 \Big] \Big) \Big\}^{1/2}$$
$$\leq \sqrt{MC_3^{1/2}} \sqrt{n} \Big\{ \sum_{i=1}^{n} \mathbb{E}\Big[\Big(\frac{f_{i*}}{f_{i0}} - 1 \Big)^2 \Big] \Big\}^{1/2} = O_p(\sqrt{Mn\xi_n}).$$

The first inequality can be obtained from Cauthy-Schwarz inequality or Hölder inequality. This implies that $\Xi_{n12} = O_p(\sqrt{Mn\xi_n}\alpha_n)\|\mathbf{u}\|$. As a result, we have

$$|\Xi_{n1}| = |\Xi_{n11} + \Xi_{n12}| \leq |\Xi_{n11}| + |\Xi_{n12}| \leq o_p(n^{1/2+\delta/2}\alpha_n)\|\mathbf{u}\| + O_p(\sqrt{Mn\xi_n}\alpha_n)\|\mathbf{u}\|. \quad \text{(C.15)}$$

Now, for Ξ_{n2}, note that

$$\begin{aligned}
\Xi_{n2} =& \frac{1}{2}\alpha_n^2 \mathbf{u}' \nabla^2 L_n(\mathbf{w}^*)\mathbf{u} \\
=& \frac{1}{2} n\alpha_n^2 \mathbf{u}' \{ \frac{1}{n} (\nabla^2 L_n(\mathbf{w}^*) - \mathbb{E}[\nabla^2 L_n(\mathbf{w}^*)]) \} \mathbf{u} + \frac{1}{2}\alpha_n^2 \mathbf{u}' \mathbb{E}[\nabla^2 L_n(\mathbf{w}^*)]\mathbf{u} \\
=& \Xi_{n21} + \frac{1}{2}\alpha_n^2 \mathbf{u}' \mathbb{E}[\nabla^2 L_n(\mathbf{w}^*)]\mathbf{u} \\
=& \Xi_{n21} + \frac{1}{2}\alpha_n^2 \mathbf{u}' \sum_{i=1}^{n} \Big\{ \int \frac{\partial^2 \ln f_{i*}}{\partial \mathbf{w} \partial \mathbf{w}'} f_{i0} dy - \int \frac{\partial^2 \ln f_{i*}}{\partial \mathbf{w} \partial \mathbf{w}'} f_{i*} dy \\
& + \int \frac{\partial^2 \ln f_{i*}}{\partial \mathbf{w} \partial \mathbf{w}'} f_{i*} dy + \int \frac{\partial \ln f_{i*}}{\partial \mathbf{w}} \frac{\partial \ln f_{i*}}{\partial \mathbf{w}'} f_{i0} dy - \int \frac{\partial \ln f_{i*}}{\partial \mathbf{w}} \frac{\partial \ln f_{i*}}{\partial \mathbf{w}'} f_{i0} dy \Big\} \mathbf{u} \\
=& \Xi_{n21} + \frac{1}{2}\alpha_n^2 \mathbf{u}' \sum_{i=1}^{n} \Big\{ \int \frac{\partial^2 \ln f_{i*}}{\partial \mathbf{w} \partial \mathbf{w}'} (f_{i0} - f_{i*}) dy + \int \frac{\partial \ln f_{i*}}{\partial \mathbf{w}} \frac{\partial \ln f_{i*}}{\partial \mathbf{w}'} (f_{i0} - f_{i*}) dy \Big\} \mathbf{u} \\
& - \frac{1}{2}\alpha_n^2 \mathbf{u}' \Big(\sum_{i=1}^{n} \int \frac{\partial \ln f_{i*}}{\partial \mathbf{w}} \frac{\partial \ln f_{i*}}{\partial \mathbf{w}'} f_{i0} dy \Big) \mathbf{u} \\
=& \Xi_{n21} + \Xi_{n22} - \Xi_{n23}.
\end{aligned}$$

The penultimate equality holds because

$$\int \frac{\partial^2 \ln f_{i*}}{\partial \mathbf{w} \partial \mathbf{w}'} f_{i*} dy = - \int \frac{\partial \ln f_{i*}}{\partial \mathbf{w}} \frac{\partial \ln f_{i*}}{\partial \mathbf{w}'} f_{i*} dy.$$

Then, we obtain that

$$\Big\| \frac{1}{n} (\nabla^2 L_n(\mathbf{w}^*) - \mathbb{E}[\nabla^2 L_n(\mathbf{w}^*)]) \Big\| = o_p(Mn^{-1/2+\delta/2}). \quad \text{(C.16)}$$

Specifically, by Chebyshev inequality, we have

$$\begin{aligned}
&P(\|\frac{1}{n}(\nabla^2 L_n(\mathbf{w}^*) - \mathbb{E}[\nabla^2 L_n(\mathbf{w}^*)])\| \geq \epsilon M n^{-1/2+\delta/2}|\mathbf{x}) \\
\leq &\frac{1}{n^2\epsilon^2 M^2 n^{-1+\delta}}\mathbb{E}\Bigg\{\sum_{m,s=1}^{M}\left(\frac{\partial^2 L_n(\mathbf{w}^*)}{\partial w_m \partial w_s} - \mathbb{E}\left[\frac{\partial^2 L_n(\mathbf{w}^*)}{\partial w_m \partial w_s}\right]\right)^2\Bigg|\mathbf{x}\Bigg\} = O\left(\frac{M^2}{\epsilon^2 M^2 n^{\delta}}\right) = o(1).
\end{aligned}$$

Hence, it follows from (C.16) that

$$\Xi_{n21} \leq \frac{1}{2}n\alpha_n^2\|\mathbf{u}\|^2 o_p(Mn^{-1/2+\delta/2}) \leq O_p(Mn^{1/2+\delta/2}\alpha_n^2)\|\mathbf{u}\|^2. \tag{C.17}$$

Next, for any $1 \leq m, s \leq M$, by Cauthy-Schwarz inequality or Hölder inequality again, we have

$$\begin{aligned}
\int \frac{\partial^2 \ln f_{i*}}{\partial w_m \partial w_s}(f_{i0} - f_{i*})dy &= \int \frac{\partial^2 \ln f_{i*}}{\partial w_m \partial w_s}\sqrt{f_{i0}} \cdot \frac{f_{i0} - f_{i*}}{\sqrt{f_{i0}}}dy \\
&\leq \left\{\int \left(\frac{\partial^2 \ln f_{i*}}{\partial w_m \partial w_s}\right)^2 f_{i0}dy\right\}^{1/2} \cdot \left\{\int \left(\frac{f_{i0} - f_{i*}}{f_{i0}}\right)^2 f_{i0}dy\right\}^{1/2} \\
&= \left\{\mathbb{E}\left[\left(\frac{\partial^2 \ln f_{i*}}{\partial w_m \partial w_s}\right)^2\right]\right\}^{1/2} \cdot \left\{\mathbb{E}\left[\left(\frac{f_{i*}}{f_{i0}} - 1\right)^2\right]\right\}^{1/2}
\end{aligned}$$

and

$$\begin{aligned}
\int \frac{\partial \ln f_{i*}}{\partial w_m}\frac{\partial \ln f_{i*}}{\partial w_s}(f_{i0} - f_{i*})dy &= \int \frac{\partial \ln f_{i*}}{\partial w_m}\frac{\partial \ln f_{i*}}{\partial w_s}\sqrt{f_{i0}} \cdot \frac{f_{i0} - f_{i*}}{\sqrt{f_{i0}}}dy \\
&\leq \left\{\int \left(\frac{\partial \ln f_{i*}}{\partial w_m}\frac{\partial \ln f_{i*}}{\partial w_s}\right)^2 f_{i0}dy\right\}^{1/2} \cdot \left\{\int \left(\frac{f_{i0} - f_{i*}}{f_{i0}}\right)^2 f_{i0}dy\right\}^{1/2} \\
&= \left\{\mathbb{E}\left[\left(\frac{\partial \ln f_{i*}}{\partial w_m}\frac{\partial \ln f_{i*}}{\partial w_s}\right)^2\right]\right\}^{1/2} \cdot \left\{\mathbb{E}\left[\left(\frac{f_{i*}}{f_{i0}} - 1\right)^2\right]\right\}^{1/2}.
\end{aligned}$$

Thus, for Ξ_{n22}, by Assumption 5, we have

$$\begin{aligned}
\Xi_{n22} \leq &\frac{1}{2}\left\|\sum_{i=1}^{n}\int \frac{\partial^2 \ln f_{i*}}{\partial \mathbf{w}\partial \mathbf{w}'}(f_{i0} - f_{i*})dy\right\|\alpha_n^2\|\mathbf{u}\|^2 \\
&+ \frac{1}{2}\left\|\sum_{i=1}^{n}\int \frac{\partial \ln f_{i*}}{\partial \mathbf{w}}\frac{\partial \ln f_{i*}}{\partial \mathbf{w}'}(f_{i0} - f_{i*})dy\right\|\alpha_n^2\|\mathbf{u}\|^2 \\
\leq &\frac{1}{2}\sum_{i=1}^{n}\left\{\sum_{m,s}^{M}\mathbb{E}\left[\left(\frac{\partial^2 \ln f_{i*}}{\partial w_m \partial w_s}\right)^2\right]\mathbb{E}\left[\left(\frac{f_{i*}}{f_{i0}} - 1\right)^2\right]\right\}^{1/2}\alpha_n^2\|\mathbf{u}\|^2
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{2}\sum_{i=1}^{n}\left\{\sum_{m,s}^{M}\mathbb{E}\left[\left(\frac{\partial \ln f_{i*}}{\partial w_m}\frac{\partial \ln f_{i*}}{\partial w_s}\right)^2\right]\mathbb{E}\left[\left(\frac{f_{i*}}{f_{i0}}-1\right)^2\right]\right\}^{1/2}\alpha_n^2\|\mathbf{u}\|^2\\
\leq&\frac{1}{2}M^{1/2}C_3^{1/2}n^{1/2}\xi_n^{1/2}\alpha_n^2\|\mathbf{u}\|^2+\frac{1}{2}M^{1/2}C_4^{1/2}n^{1/2}\xi_n^{1/2}\alpha_n^2\|\mathbf{u}\|^2=O_p(\sqrt{Mn\xi_n}\alpha_n^2)\|\mathbf{u}\|^2.
\end{aligned}
\tag{C.18}
$$

The last inequality holds because of the arithmetic and geometric means inequality. For Ξ_{n23}, it can be obtained from Assumption 5 that

$$
\Xi_{n23}=\frac{1}{2}\alpha_n^2\mathbf{u}'\left(\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial \ln f_{i*}}{\partial \mathbf{w}}\frac{\partial \ln f_{i*}}{\partial \mathbf{w}'}\right]\right)\mathbf{u}=\frac{1}{2}n\alpha_n^2\mathbf{u}'I(\mathbf{w}^*)\mathbf{u}=O_p(n\alpha_n^2)\|\mathbf{u}\|^2. \tag{C.19}
$$

From (C.17), (C.18) and (C.19), we can get

$$
\Xi_{n2}=O_p(Mn^{1/2+\delta/2}\alpha_n^2)\|\mathbf{u}\|^2+O_p(\sqrt{Mn\xi_n}\alpha_n^2)\|\mathbf{u}\|^2-O_p(n\alpha_n^2)\|\mathbf{u}\|^2. \tag{C.20}
$$

For Ξ_{n3}, by triangular inequality, Cauthy-Schwarz inequality and Assumption 6, we have

$$
\begin{aligned}
|\Xi_{n3}|=&\left|\frac{1}{6}\alpha_n^3\nabla'[\mathbf{u}'\nabla^2L_n(\tilde{\mathbf{w}})\mathbf{u}]\mathbf{u}\right|=\frac{1}{6}\left|\sum_{l,m,s=1}^{M}\frac{\partial L_n(\tilde{\mathbf{w}})}{\partial w_l\partial w_m\partial w_s}u_lu_mu_s\alpha_n^3\right|\\
\leq&\frac{1}{6}\sum_{i=1}^{n}\left[\sum_{l,m,s=1}^{M}M_{lms}^2(y_i|\mathbf{x}_i)\right]^{1/2}\|\mathbf{u}\|^3\alpha_n^3=O_p(M^{3/2}n\alpha_n^3)\|\mathbf{u}\|^3.
\end{aligned}
\tag{C.21}
$$

Finally, to deal with Ξ_{n4}, by Assumption 4, we have

$$
\begin{aligned}
\Xi_{n4}=&\lambda_{n,1}\alpha_n\mathbf{u}'\mathbf{k}\\
&+\sum_{j=1}^{p}\left[\sum_{m=1}^{M}p'_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w}^*)|)\widehat{\theta}_{\tau,j}^{(m)}\mathrm{sgn}(\widehat{\theta}_{\tau,j}^{(m)})\alpha_nu_m\right.\\
&\left.+\sum_{m=1}^{M}p''_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w}^*)|)\widehat{\theta}_{\tau,j}^{(m)^2}\alpha_n^2u_m^2\{1+o(1)\}\right]\\
=&\Xi_{n41}+\Xi_{n42}.
\end{aligned}
$$

It can be obtained from Assumption 4 that

$$
\Xi_{n41}=\lambda_{n,1}\alpha_n\mathbf{u}'\mathbf{k}\leq\lambda_{n,1}\alpha_n\|\mathbf{k}\|\|\mathbf{u}\|=\lambda_{n,1}\alpha_nO(p\sqrt{M})\|\mathbf{u}\|=O_p(n^{1/2}p^{1/2}M^{1/2}\alpha_n)\|\mathbf{u}\|.
$$

Then, by Assumption 4, Ξ_{n42} can be expressed as

$$\begin{aligned}\Xi_{n42} =& \sum_{j=1}^{p}\sum_{m=1}^{M} p'_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w}^*)|)\widehat{\theta}^{(m)}_{\tau,j}\mathrm{sgn}(\widehat{\theta}^{(m)}_{\tau,j})\alpha_n u_m \\ &+ \sum_{j=1}^{p}\sum_{m=1}^{M} p''_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w}^*)|)\widehat{\theta}^{(m)^2}_{\tau,j}\alpha_n^2 u_m^2\{1+o(1)\} \\ \leq& O_p(n^{1/2}p^{-1/2})\cdot O_p(p)\alpha_n M^{1/2}\|\mathbf{u}\| + o_p(np^{-1})\cdot O_p(p)\alpha_n^2\|\mathbf{u}\|^2\{1+o(1)\} \\ =& O_p(n^{1/2}p^{1/2}M^{1/2}\alpha_n)\|\mathbf{u}\| + o_p(n\alpha_n^2)\|\mathbf{u}\|^2.\end{aligned}$$

Thus,

$$\Xi_{n4} \leq O_p(n^{1/2}p^{1/2}M^{1/2}\alpha_n)\|\mathbf{u}\| + o_p(n\alpha_n^2)\|\mathbf{u}\|^2. \tag{C.22}$$

Therefore, it can be obtained from the above proof that

$$D_n(\mathbf{u}) = \Xi_{n23} - \Xi_{n11} - \Xi_{n12} - \Xi_{n21} - \Xi_{n22} - \Xi_{n3} + \Xi_{n4} \tag{C.23}$$

From the definition $\alpha_n = O_p(\xi_n^{1/2}n^{-1/2+\delta/2})$, together with Assumption 7, (C.15), (C.20), (C.21) and (C.22), it is easy to see that all terms of (C.23) are dominated by its first term Ξ_{n23}, and the nonnegative property of Ξ_{n23} implies that $D_n(\mathbf{u})$ is asymptotically nonnegative, which implies that $\|\widehat{\mathbf{w}}_n - \mathbf{w}^*\| = O_p(\xi_n^{1/2}n^{-1/2+\delta/2})$. The proof of Theorem 2 is complete. □

Proof of Theorem 3. It can be obtained from Lemma 1 that

$$\|\widehat{\boldsymbol{\theta}}^{(m)}_{\tau} - \boldsymbol{\theta}^{*(m)}_{\tau}\| = O_p(k_m^{1/2}n^{-1/2}) \leq O_p(p^{1/2}n^{-1/2}),$$

where p is the number of all regressors in the candidate models. Then, with the conditions, we have

$$\begin{aligned}\|\widehat{\boldsymbol{\theta}}_{\tau}(\widehat{\mathbf{w}}) - \boldsymbol{\theta}^*_{\tau}(\mathbf{w}^*)\| &= \|(\widehat{\boldsymbol{\theta}}_{\tau}(\widehat{\mathbf{w}}) - \widehat{\boldsymbol{\theta}}_{\tau}(\mathbf{w}^*)) + (\widehat{\boldsymbol{\theta}}_{\tau}(\mathbf{w}^*) - \boldsymbol{\theta}^*_{\tau}(\mathbf{w}^*))\| \\ &\leq \|\widehat{\boldsymbol{\theta}}_{\tau}(\widehat{\mathbf{w}}) - \widehat{\boldsymbol{\theta}}_{\tau}(\mathbf{w}^*)\| + \|\widehat{\boldsymbol{\theta}}_{\tau}(\mathbf{w}^*) - \boldsymbol{\theta}^*_{\tau}(\mathbf{w}^*)\| \\ &= \|\widehat{\boldsymbol{\theta}}_{\tau}(\widehat{\mathbf{w}} - \mathbf{w}^*)\| + \|(\widehat{\boldsymbol{\theta}}_{\tau} - \boldsymbol{\theta}^*_{\tau})\mathbf{w}^*\| \\ &\leq \|\widehat{\boldsymbol{\theta}}_{\tau}\|_2 \cdot \|\widehat{\mathbf{w}} - \mathbf{w}^*\| + \|\widehat{\boldsymbol{\theta}}_{\tau} - \boldsymbol{\theta}^*_{\tau}\|_2 \cdot \|\mathbf{w}^*\|\end{aligned}$$

$$
\begin{aligned}
&= \sigma_{max}(\widehat{\boldsymbol{\theta}}_\tau) \cdot \|\widehat{\mathbf{w}} - \mathbf{w}^*\| + \sigma_{max}(\widehat{\boldsymbol{\theta}}_\tau - \boldsymbol{\theta}^*_\tau) \cdot \|\mathbf{w}^*\| \\
&\leq O_p(\sqrt{pM}) \cdot O_p(\xi_n^{1/2} n^{-1/2+\delta/2}) + O_p(\sqrt{pMn^{-1}}) \cdot O_p(M^{1/2}) \\
&= O_p(p^{1/2} M^{1/2} \xi_n^{1/2} n^{-1/2+\delta/2} + M p^{1/2} n^{-1/2}) = o_p(1),
\end{aligned}
$$

where $\|\widehat{\boldsymbol{\theta}}_\tau\|_2$ is the 2-norm of matrix $\widehat{\boldsymbol{\theta}}_\tau$ and σ_{max} is the largest singular value of the matrix. Therefore, Theorem 3 is established. □

Proof of Theorem 4. It has been defined in the proof of Theorem 1 that

$$
\begin{aligned}
\widetilde{G}_n &= G_n + \sum_{i=1}^{n} \mathbb{E}_0[\ln f_0(y_i|\mathbf{x}_i)] \\
&= -\sum_{i=1}^{n} \ln f(y_i|\widehat{\boldsymbol{\theta}}_\tau(\mathbf{w}), \mathbf{x}_i, \tau) + \sum_{i=1}^{n} \mathbb{E}_0[\ln f_0(y_i|\mathbf{x}_i)] + \lambda_{n,1}\mathbf{w}'\mathbf{k} + \sum_{j=1}^{p} p_{\lambda_{n,2}}(|\widehat{\boldsymbol{\theta}}_{\tau,j}(\mathbf{w})|).
\end{aligned}
$$

To prove the asymptotic optimality of model averaging weight in the screened space $\mathbf{W}^*$, for any $\gamma > 0$, we have

$$
\begin{aligned}
&P\left\{\left|\frac{\inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w})}{\mathrm{KL}(\widehat{\mathbf{w}}^*)} - 1\right| > \gamma\right\} = P\left\{\left|\frac{\inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w}) - \mathrm{KL}(\widehat{\mathbf{w}}^*)}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| > \gamma\right\} \\
=&P\left\{\left|\frac{\inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w}) + \widetilde{G}_n(\widehat{\mathbf{w}}^*) - \mathrm{KL}(\widehat{\mathbf{w}}^*) - \widetilde{G}_n(\widehat{\mathbf{w}}^*)}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| > \gamma\right\} = B_n.
\end{aligned}
$$

From (C.2), it follows that $\widetilde{G}_n(\widehat{\mathbf{w}}^*) = \inf_{\mathbf{w}\in\mathcal{W}^*} \widetilde{G}_n(\mathbf{w})$. Let $a_n(\mathbf{w}) = \widetilde{G}_n(\mathbf{w}) - \mathrm{KL}(\mathbf{w})$, we have

$$
\begin{aligned}
B_n =&P\left\{\left|\frac{\inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w}) + a_n(\widehat{\mathbf{w}}^*) - \inf_{\mathbf{w}\in\mathcal{W}^*} \widetilde{G}_n(\mathbf{w})}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| > \gamma\right\} \\
=&P\left\{\left|\frac{\inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w}) + a_n(\widehat{\mathbf{w}}^*) - \inf_{\mathbf{w}\in\mathcal{W}^*}(\mathrm{KL}(\mathbf{w}) + a_n(\mathbf{w}))}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| > \gamma\right\} \\
=&P\left\{\left|\frac{\inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w}) + a_n(\widehat{\mathbf{w}}^*) - \inf_{\mathbf{w}\in\mathcal{W}^*}(\mathrm{KL}(\mathbf{w}) + a_n(\mathbf{w}))}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| > \gamma, \mathbf{w}_n \in \mathcal{W}^*\right\} \\
&+ P\left\{\left|\frac{\inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w}) + a_n(\widehat{\mathbf{w}}^*) - \inf_{\mathbf{w}\in\mathcal{W}^*}(\mathrm{KL}(\mathbf{w}) + a_n(\mathbf{w}))}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right| > \gamma, \mathbf{w}_n \notin \mathcal{W}^*\right\}.
\end{aligned}
$$

For $\inf_{\mathbf{w}\in\mathcal{W}} \mathrm{KL}(\mathbf{w}) - \mathrm{KL}(\widehat{\mathbf{w}}^*) \leq 0$ and $\inf_{\mathbf{w}\in\mathcal{W}^*}(\mathrm{KL}(\mathbf{w}) + a_n(\mathbf{w})) \leq \mathrm{KL}(\mathbf{w}_n) + a_n(\mathbf{w}_n)$, we

have

$$
\begin{aligned}
B_n \leq & P\left\{\left|\frac{\inf_{\mathbf{w}\in\mathcal{W}}\mathrm{KL}(\mathbf{w})+a_n(\widehat{\mathbf{w}}^*)-(\mathrm{KL}(\mathbf{w}_n)+a_n(\mathbf{w}_n))}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right|>\gamma|\mathbf{w}_n\in\mathcal{W}^*\right\}P\{\mathbf{w}_n\in\mathcal{W}^*\} \\
& + P\{\mathbf{w}_n\notin\mathcal{W}^*\} \\
= & P\left\{\left|\frac{\mathrm{KL}(\mathbf{w}_n)-\nu_n+a_n(\widehat{\mathbf{w}}^*)-(\mathrm{KL}(\mathbf{w}_n)+a_n(\mathbf{w}_n))}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right|>\gamma|\mathbf{w}_n\in\mathcal{W}^*\right\}P\{\mathbf{w}_n\in\mathcal{W}^*\} \\
& + P\{\mathbf{w}_n\notin\mathcal{W}^*\} \\
\leq & P\left\{\left|\frac{\nu_n}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right|+\left|\frac{a_n(\widehat{\mathbf{w}}^*)}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right|+\left|\frac{a_n(\mathbf{w}_n)}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}\right|>\gamma|\mathbf{w}_n\in\mathcal{W}^*\right\}P\{\mathbf{w}_n\in\mathcal{W}^*\}+P\{\mathbf{w}_n\notin\mathcal{W}^*\} \\
\leq & P\left\{\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\nu_n}{\mathrm{KL}^*(\widehat{\mathbf{w}})}\right|\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\widehat{\mathbf{w}})}{\mathrm{KL}(\mathbf{w})}\right|+\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{a_n(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right|\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}(\mathbf{w})}\right|\right. \\
& \left.+\left|\frac{a_n(\mathbf{w}_n)}{\inf_{\mathbf{w}\in\mathcal{W}}\mathrm{KL}(\mathbf{w})}\right|>\gamma|\mathbf{w}_n\in\mathcal{W}^*\right\}P\{\mathbf{w}_n\in\mathcal{W}^*\}+P\{\mathbf{w}_n\notin\mathcal{W}^*\} \\
\leq & P\left\{\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\nu_n}{\mathrm{KL}^*(\widehat{\mathbf{w}})}\right|\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\widehat{\mathbf{w}})}{\mathrm{KL}(\mathbf{w})}\right|+\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{a_n(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right|\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}(\mathbf{w})}\right|\right. \\
& \left.+\left|\frac{a_n(\mathbf{w}_n)}{\mathrm{KL}(\mathbf{w}_n)-\nu_n}\right|>\gamma|\mathbf{w}_n\in\mathcal{W}^*\right\}P\{\mathbf{w}_n\in\mathcal{W}^*\}+P\{\mathbf{w}_n\notin\mathcal{W}^*\} \\
\leq & P\left\{\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\nu_n}{\mathrm{KL}^*(\widehat{\mathbf{w}})}\right|\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\widehat{\mathbf{w}})}{\mathrm{KL}(\mathbf{w})}\right|+\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{a_n(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right|\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}(\mathbf{w})}\right|\right. \\
& \left.+\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{a_n(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right|\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}(\mathbf{w})-\nu_n}\right|>\gamma|\mathbf{w}_n\in\mathcal{W}^*\right\}P\{\mathbf{w}_n\in\mathcal{W}^*\}+P\{\mathbf{w}_n\notin\mathcal{W}^*\}
\end{aligned}
$$

Assumption 8 implies that $P\{\mathbf{w}_n\notin\mathcal{W}^*\}\to 0$. Then, we have

$$
\begin{aligned}
B_n = & P\left\{\left|\frac{\inf_{\mathbf{w}\in\mathcal{W}}\mathrm{KL}(\mathbf{w})}{\mathrm{KL}(\widehat{\mathbf{w}}^*)}-1\right|>\gamma\right\} \\
\leq & P\left\{\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\nu_n}{\mathrm{KL}^*(\widehat{\mathbf{w}})}\right|\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\widehat{\mathbf{w}})}{\mathrm{KL}(\mathbf{w})}\right|+\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{a_n(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right|\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}(\mathbf{w})}\right|\right. \\
& \left.+\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{a_n(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right|\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}(\mathbf{w}_n)-\nu_n}\right|>\gamma|\mathbf{w}_n\in\mathcal{W}^*\right\}. \qquad \text{(C.24)}
\end{aligned}
$$

For each term of (C.24), it follows from Assumption 8 that $\nu_n/\inf_{\mathbf{w}\in\mathcal{W}}\mathrm{KL}^*(\mathbf{w})\to 0$, which implies

$$
\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\nu_n}{\mathrm{KL}^*(\mathbf{w})}\right|=o_p(1), \qquad \text{(C.25)}
$$

and by (C.3) and (C.4) in the proof of Theorem 1, we have

$$\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\widetilde{G}_n-\mathrm{KL}(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right| = \sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\widetilde{G}_n-\mathrm{KL}^*+\mathrm{KL}^*-\mathrm{KL}(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right|$$
$$\leq \sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\widetilde{G}_n-\mathrm{KL}^*}{\mathrm{KL}^*(\mathbf{w})}\right| + \sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*-\mathrm{KL}(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right| = o_p(1). \tag{C.26}$$

Then, note that

$$\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}(\mathbf{w})}\right| = \left(\inf_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right|\right)^{-1} = \left(\inf_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}(\mathbf{w})-\mathrm{KL}^*(\mathbf{w})+\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right|\right)^{-1}$$
$$= \left(\inf_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}(\mathbf{w})-\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}+1\right|\right)^{-1}$$
$$\leq \left(\inf_{\mathbf{w}\in\mathcal{W}}\left\{1-\left|\frac{\mathrm{KL}(\mathbf{w})-\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right|\right\}\right)^{-1}$$
$$\leq \left(1-\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}(\mathbf{w})-\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right|\right)^{-1} = (1-o_p(1))^{-1}\to 1. \tag{C.27}$$

In addition, we can also obtain that

$$\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}^*(\mathbf{w})}{\mathrm{KL}(\mathbf{w})-\nu_n}\right| = \left(\inf_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}(\mathbf{w})-\nu_n}{\mathrm{KL}^*(\mathbf{w})}\right|\right)^{-1} = \left(\inf_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}-\frac{\nu_n}{\mathrm{KL}^*(\mathbf{w})}\right|\right)^{-1}$$
$$= \left(\left|\inf_{\mathbf{w}\in\mathcal{W}}\left\{\left|\frac{\nu_n}{\mathrm{KL}^*(\mathbf{w})}\right|-\left|\frac{\mathrm{KL}(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right|\right\}\right|\right)^{-1}$$
$$\leq \left(\left|\inf_{\mathbf{w}\in\mathcal{W}}\left|\frac{\nu_n}{\mathrm{KL}^*(\mathbf{w})}\right|-\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{\mathrm{KL}(\mathbf{w})}{\mathrm{KL}^*(\mathbf{w})}\right|\right|\right)^{-1} = |o_p(1)-1|^{-1}\to 1,$$

which, together with (C.24)-(C.27), yields that

$$\frac{\mathrm{KL}(\widehat{\mathbf{w}}^*)}{\inf_{\mathbf{w}\in\mathcal{W}}\mathrm{KL}(\mathbf{w})}\to 1.$$

This concludes of Theorem 4. □