8-2014

Jackknife Model Averaging for Quantile Regressions

Xun LU
Hong Kong University of Science and Technology

Liangjun SU
Singapore Management University, ljsu@smu.edu.sg

SMU ECONOMICS & STATISTICS
WORKING PAPER SERIES

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Jackknife Model Averaging for Quantile Regressions

Xun Lu and Liangjun Su

August 2014

Paper No. 11-2014

Jackknife Model Averaging for Quantile Regressions*

Xun Lu[a] and Liangjun Su[b]
[a]Department of Economics, Hong Kong University of Science & Technology
[b]School of Economics, Singapore Management University, Singapore

June 12, 2014

Abstract

In this paper we consider the problem of frequentist model averaging for quantile regression (QR) when all the M models under investigation are potentially misspecified and the number of parameters in some or all models is diverging with the sample size n. To allow for the dependence between the error terms and the regressors in the QR models, we propose a jackknife model averaging (JMA) estimator which selects the weights by minimizing a leave-one-out cross-validation criterion function and demonstrate that the jackknife selected weight vector is asymptotically optimal in terms of minimizing the out-of-sample final prediction error among the given set of weight vectors. We conduct Monte Carlo simulations to demonstrate the finite-sample performance of the proposed JMA QR estimator and compare it with other model selection and averaging methods. We find that in terms of out-of-sample forecasting, the JMA QR estimator can achieve significant efficiency gains over the other methods, especially for extreme quantiles. We apply our JMA method to forecast quantiles of excess stock returns and wages.

JEL Classification: C51, C52

Key Words: Final prediction error; High dimensionality; Model averaging; Model selection; Misspecification; Quantile regression

1 Introduction

In practice researchers are often confronted with a large number of candidate models and are not sure which model to use. Model selection helps to choose a single optimal model, ignores the information in other models, and often produces a rather unstable estimator in applications despite the fact that it

*The authors gratefully thank the Co-editor Han Hong, the associate editor, two anonymous referees for their many helpful comments. They are also indebted to Peter C. B. Phillips for his constructive comments on the paper and valuable discussions on the subject matter. Su gratefully acknowledges the Singapore Ministry of Education for Academic Research Fund under grant number MOE2012-T2-2-021. Address correspondence to: Liangjun Su, School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903; E-mail: ljsu@smu.edu.sg, Phone: +65 6828 0386.

has a long history and nice theoretical properties in both statistics and econometrics literature.[1] As an alternative to model selection, model averaging, on the other hand, seeks to obtain a combined estimator by taking the weighted average of the estimators obtained from all candidate models under investigation. It allows researchers to diversify, account for model uncertainty, and improve out-of-sample performance.

Model averaging can be classified as Bayesian model averaging (BMA) and frequentist model averaging (FMA). See Hoeting et al. (1999) for an overview on BMA and Moral-Benito (2013) for a recent overview on both BMA and FMA. FMA has a relatively shorter history than BMA. Buckland et al. (1997) and Burnham and Anderson (2002, ch.6) construct model averaging weights based on the values of AIC or BIC scores. Yang (2001) and Yuan and Yang (2005) propose a model averaging method known as adaptive regression by mixing (ARM). In a local asymptotic framework Hjort and Clasekens (2003) and Clasekens and Hjort (2008, ch.7) study the asymptotic properties of the FMA maximum likelihood estimator by studying perturbations around a given narrow model in certain directions. Other works on the asymptotic property of averaging estimators include Leung and Barron (2006), Pötscher (2006), Hansen (2009, 2010), and Liu (2012). In particular, Liu (2012) proposes a plug-in estimator of the optimal weights by minimizing the sample analog of the asymptotic mean squared error (MSE) for linear regression models. In a similar spirit, Liang et al. (2011) derive an exact unbiased estimator of the MSE of the model average estimator and propose selecting the weights that minimize the trace of the MSE estimate of focus parameters.

In a seminal article, Hansen (2007) proposes selecting the model weights in least squares model averaging by minimizing the Mallows' criterion over a set of discrete weights. The justification of this method lies in the fact that the Mallows' criterion is asymptotically equivalent to the squared error so that the Mallows model averaging (MMA) estimator is asymptotically optimal in terms of minimizing the MSE. Thus his approach marks a significant step toward the development of optimal weight choice in the FMA estimator. Hansen (2008, 2009, 2010) extends his MMA method to the forecast combination literature, to models with structural break, and to models with a near unit root, respectively. Note that Hansen (2007) only considers nested models and his MMA estimator does not allow for (conditional) heteroskedasticity. Wan et al. (2010) extend Hansen's MMA estimator to allow for non-nested model and selection of continuous weights in a unit simplex. Liu and Okui (2013) extends Hansen's MMA estimator to allow for heteroskedasticity and non-discrete weights. To allow for both non-nested models and heteroskedasticity, Hansen and Racine (2012) propose jackknife model averaging (JMA) for least squares regression when the weights are selected by minimizing a leave-one-out cross-validation criterion function. Zhang et al. (2013) extend JMA to models with dependent data. In the case of instrument uncertainty, Kuersteiner and Okui (2010) apply the MMA approach to the first stage of the 2SLS, LIML and FIML estimators. In contrast, Lee and Zhou (2011) take an average over the second stage estimators. Sueishi (2010) proposes a new simultaneous model and instrument selection method for IV models based on 2SLS estimation when the true model is of infinite dimension.

Almost all of the above papers on FMA focus on the least squares regression and MSE criterion. The only exceptions are Hjort and Clasekens (2003) and Clasekens and Hjort (2008) who concentrate

[1] It is well known that a small perturbation of the data can result in selecting a very different model. As a consequence, estimators of the regression function based on model selection often have larger variance than usual. See Yang (2001).

on the likelihood framework but with the MSE criterion too. The MSE criterion seems natural in the least squares regression framework because it balances the asymptotic bias and variance in a nice way. Nevertheless, it is also interesting to apply the idea of FMA to different contexts where MSE may not be the best criterion choice.

In this paper we extend the JMA of Hansen and Racine (2012) to the quantile regression (QR) framework. QR provides much more information about the conditional distribution of a response variable than the traditional conditional mean regression. Since the seminal paper of Koenker and Bassett (1978), QR has attracted huge attention in the literature. Just as in the least squares regression, model selection and model averaging can play an important role in the QR model building process. There is a growing literature on model selection for QR models or more generally, M-estimation. For example, Hurwich and Tsai (1990) develop a small sample criterion for the selection of LAD regression models; Machado (1993) and Burman and Nolan (1995) consider variants of the Schwarz information criterion (SIC) and Akaike information criterion (AIC), respectively, for M-estimation which includes the QR as a special case; Koenker et al. (1994) consider using SIC in QR models. More recently, Wu and Liu (2009) study variable selection in penalized QR (see Su and Zhang (2014) for an overview on this); Belloni and Chernozhukov (2011) consider l_1-penalized QR in high-dimensional sparse models. Nevertheless, to the best of our knowledge, there still is a lack of an FMA method in the QR framework. This work seeks to fill this gap. It is well known that quantile estimates tend to be unstable when the quantile index is very high or very low (say, close to 0.95 or 0.05). This implies that model averaging can certainly play an important role in this case.

To proceed, it is worth mentioning that the major motivation for model averaging is to address the problem of *model uncertainty* for forecasting. Kapetanios et al. (2008) provide compelling reasons for using model averaging for the purpose of forecasting. They consider two broad cases: one is when the model that generates the data belongs to the class of candidate models, and the other, which is perhaps more relevant in empirical applications, is when the true model does not belong to the class of models under consideration. In the first case, model averaging addresses the issue that the chosen model is not necessarily the true model, and by assigning probabilities to various models can yield an out-of-sample forecast that is robust to model uncertainty. In the second case, it is impossible that the chosen model could capture all the features of the true model, which makes the motivation for model averaging even stronger because it has been well documented in the forecasting literature that forecasts from different models can inform the overall forecast in different ways and tend to outperform individual forecasts significantly. Admittedly, forecasting a variable of interest and discovering the true model (or true structural/causal relation) can be quite different objectives in econometrics. As Ng (2013) puts it in her abstract, "*(i)rrespective of the model size, there is an unavoidable tension between prediction accuracy and consistent model determination.*" Consistent model selection of the true model, if existing, does not necessarily lead to a model that yields minimum forecast error. The main purpose of this paper is to provide a FMA method for the purpose of forecasting a variable of interest under the check loss function but not to discover the underlying true model because it is possible in practice that none of the models considered is the true model or even close to the truth.

Since we use the check loss function as a base for model averaging, we do not have the usual bias-

variance decomposition for the MSE-based evaluation criterion, and it is difficult to define a Mallows-type criterion for the QR model averaging as in Hansen (2007).[2] For this reason, we focus on the extension of Hansen and Racine's (2012) JMA to the QR framework. Such an extension is not trivial for several reasons. First, there is no closed form solution for QR, and the asymptotic properties of jackknife QR estimators are not well studied in the literature. Second, since we do not adopt the local asymptotic framework, it is possible that all the models under investigation are incorrectly specified even asymptotically. The literature on QR under misspecification is quite limited. Third, we allow the number of parameters in the QR models to diverge with the sample size n, which also complicates the analysis of QR estimators under model misspecification. We shall study the consistency and asymptotic normality of QR estimators for a single potentially misspecified QR model with a diverging number of parameters, and then study the uniform consistency of the leave-one-out QR estimators. These results are needed in order to establish the asymptotic optimality of our JMA estimator. Fourth, we also allow the number of the candidate models to increase with the sample size at a suitable polynomial rate.

We conduct Monte Carlo simulations to compare the finite sample performance of our JMA QR estimators with other model averaging and model selection methods, such as those based on AIC and BIC. We find that our JMA QR estimators clearly dominate other methods for the 0.05th conditional quantile regression. For the conditional median regression, there is no clearly dominating method, but JMA QR estimators perform well in most of the cases. We apply our new method to predict the conditional quantiles of excess stock returns and wages.

The rest of the paper is structured as follows. Section 2 proposes the quantile regression model averaging estimator. We study the asymptotic properties of the quantile regression estimators and the asymptotic optimality of our jackknife selected weight vector in Section 3. Section 4 reports the Monte Carlo simulation results. In Section 5 we apply the proposed method to predict conditional quantiles of excess stock returns and wages. Section 6 concludes. The proofs of the main results in Section 3 are relegated to Appendices A-C. Supplementary appendices D-E contain some additional theoretical and simulation results.

NOTATION. Throughout the paper we adopt the following notation. For an $m \times n$ real matrix A, we denote its transpose as A', its Frobenius norm as $\|A\|$ $(\equiv [\mathrm{tr}\,(AA')]^{1/2})$, and its Moore-Penrose generalized inverse as A^+. When A is symmetric, we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote its largest and smallest eigenvalues, respectively. I_l denotes an $l \times l$ identity matrix and "p.s.d." abbreviates "positive semidefinite". The operator $\xrightarrow{p}$ denotes convergence in probability, $\xrightarrow{d}$ convergence in distribution, and plim probability limit.

2 Quantile regression model averaging

In this section we present the quantile regression model averaging estimators.

[2] Alternatively, one can continue to adopt the MSE as an evaluation criterion for QR estimators. It remains unknown whether Hansen's MMA has a straightforward extension to QR.

2.1 Quantile regression model averaging

Let $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ be a random sample, where y_i is a scalar dependent variable and $\mathbf{x}_i = (x_{i1}, x_{i2}, ...)$ is of countably infinite dimension. Without loss of generality, we assume that $x_{i1} = 1$. As in Koenker and Bassett (1982), we consider the following data generating process (DGP)

$$y_i = \sum_{j=1}^{\infty} \beta_j x_{ij} + \left(\sum_{j=1}^{\infty} \alpha_j x_{ij} \right) \epsilon_i \tag{2.1}$$

where β_j and α_j are unknown parameters, $\alpha_1 = 1$, and ϵ_i are independent and identically distributed (IID) unobservable error terms and are independent of $\mathbf{x}_i$. Let $q_\epsilon(\tau)$ denote the τth quantile of ϵ_i for some $\tau \in (0,1)$. Under the condition that the conditional scale function $\sigma(\mathbf{x}_i) = \sum_{j=1}^{\infty} \alpha_j x_{ij}$ is nonnegative almost surely (a.s.), the τth conditional quantile of y_i given $\mathbf{x}_i$ is given by

$$Q_\tau(\mathbf{x}_i) = \sum_{j=1}^{\infty} \left[\beta_j + \alpha_j q_\epsilon(\tau) \right] x_{ij} = \sum_{j=1}^{\infty} \theta_j(\tau) x_{ij}, \tag{2.2}$$

where $\theta_j(\tau) \equiv \beta_j + \alpha_j q_\epsilon(\tau)$. It follows that we have the following linear QR model

$$y_i = \mu_i + \varepsilon_i = \sum_{j=1}^{\infty} \theta_j x_{ij} + \varepsilon_i, \tag{2.3}$$

where $\mu_i \equiv \mu_i(\tau) \equiv \sum_{j=1}^{\infty} \theta_j x_{ij}$, $\theta_j \equiv \theta_j(\tau)$, and $\varepsilon_i \equiv \varepsilon_i(\tau) \equiv y_i - Q_\tau(\mathbf{x}_i)$ satisfies the quantile restriction[3]

$$P(\varepsilon_i(\tau) \leq 0 | \mathbf{x}_i) = \tau. \tag{2.4}$$

We consider a sequence of approximating models $m = 1, 2, ..., M$, where the m'th model uses k_m regressors belonging to $\mathbf{x}_i$ and M may go to infinity with the sample size. We write the m'th approximating model as

$$y_i = \Theta'_{(m)} \mathbf{x}_{i(m)} + b_{i(m)} + \varepsilon_i = \sum_{j=1}^{k_m} \theta_{j(m)} x_{ij(m)} + b_{i(m)} + \varepsilon_i, \tag{2.5}$$

where $\Theta_{(m)} \equiv (\theta_{1(m)}, ..., \theta_{k_m(m)})'$, $\mathbf{x}_{i(m)} = (x_{i1(m)}, ..., x_{ik_m(m)})'$, $x_{ij(m)}$, $j = 1, ..., k_m$, are variables in $\mathbf{x}_i$ that appear as regressors in the m'th model, $\theta_{j(m)}$ are the corresponding coefficients, and $b_{i(m)} = \mu_i - \sum_{j=1}^{k_m} \theta_{j(m)} x_{ij(m)}$ signifies the approximation error in the m'th model. Although k_m and consequently $\mathbf{x}_{i(m)}$ and $\Theta_{(m)}$ may depend on n, we suppress their dependence on n for notational simplicity. In particular, k_m is permitted to diverge to infinity with n, which may be important in both practice and theory. First, in practice allowing k_m to diverge with n is a way of allowing the model to become more complicated as the sample size increases and, through the restrictions on the rate at which k_m can increase as $n \to \infty$, suggests restrictions on the complexity of the model for each finite n. Second, in the theory for nonparametric sieve estimation, the number of approximating terms which is k_m here has to diverge with n at a certain rate in order to achieve a desirable balance between the approximating bias and the asymptotic variance of the resulting sieve estimator.

[3] For notational simplicity, we frequently suppress the dependence of $\mu_i(\tau)$, $\theta_j(\tau)$, and $\varepsilon_i(\tau)$ on τ.

Let $\rho_\tau(e) \equiv e[\tau - \mathbf{1}\{e \leq 0\}]$ where $\mathbf{1}\{\cdot\}$ denotes the usual indicator function. The τth QR estimate of $\Theta_{(m)}$ is given by

$$\hat{\Theta}_{(m)} \equiv \arg\min_{\Theta_{(m)}} Q_{n(m)}\left(\Theta_{(m)}\right) = \arg\min_{\Theta_{(m)}} \sum_{i=1}^{n} \rho_\tau\left(y_i - \Theta'_{(m)}\mathbf{x}_{i(m)}\right). \tag{2.6}$$

Let $\hat{\varepsilon}_{i(m)} \equiv y_i - \hat{\Theta}'_{(m)}\mathbf{x}_{i(m)}$. Let $\mathbf{w} \equiv (w_1, ..., w_m)'$ be a weight vector in the unit simplex of $\mathbb{R}^M$ and $\mathcal{W} \equiv \left\{\mathbf{w} \in [0,1]^M : \sum_{m=1}^{M} w_m = 1\right\}$. For $i = 1, ..., n$, the model average estimator of μ_i is given by

$$\hat{\mu}_i(\mathbf{w}) = \sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\hat{\Theta}_{(m)}. \tag{2.7}$$

Ideally, one might consider choosing $\mathbf{w}$ to minimize the average quantile loss

$$L_n(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(\mu_i - \hat{\mu}_i(\mathbf{w})\right), \tag{2.8}$$

or its associated conditional risk

$$R_n(\mathbf{w}) = E\left[L_n(\mathbf{w})|\mathbf{X}\right], \tag{2.9}$$

where $\mathbf{X} \equiv \{\mathbf{x}_1, ..., \mathbf{x}_n\}$. Unlike the squared error loss function in Hansen (2007), it is not easy for us to study $L_n(\mathbf{w})$ or $R_n(\mathbf{w})$ directly in order to establish their connection with any known information criterion (e.g., AIC and BIC) in the quantile regression literature. Below we follow the lead of Hansen and Racine (2012) and propose jackknife selection of $\mathbf{w}$ (also known as leave-one-out cross-validation).

2.2 Jackknife weighting

Here we propose jackknife selection of $\mathbf{w}$. We shall show in the next section that the jackknife weight vector is optimal in terms of minimizing final prediction error (FPE) in the sense of Akaike (1970).

For $m = 1, ..., M$, let $\hat{\Theta}_{i(m)}$ denote the jackknife estimator of $\Theta_{(m)}$ in model m with the i'th observation deleted. Define the leave-one-out cross-validation criterion function as

$$CV_n(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(y_i - \sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\hat{\Theta}_{i(m)}\right). \tag{2.10}$$

The jackknife choice of weight vector $\hat{\mathbf{w}} = (\hat{w}_1, ..., \hat{w}_M)$ is obtained by choosing $\mathbf{w} \in \mathcal{W}$ to minimize the above criterion function, i.e.,

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{W}} CV_n(\mathbf{w}). \tag{2.11}$$

Given $\hat{\mathbf{w}}$, one can obtain the jackknife model averaging (JMA) estimator of μ_i by

$$\hat{\mu}_i(\hat{\mathbf{w}}) = \sum_{m=1}^{M} \hat{w}_m \mathbf{x}'_{i(m)}\hat{\Theta}_{(m)}. \tag{2.12}$$

Note that $CV_n(\mathbf{w})$ is convex in $\mathbf{w}$ and can be minimized by running the quantile regression of y_i on $\mathbf{x}'_{i(m)}\hat{\Theta}_{i(m)}$. But this procedure cannot guarantee the resulting solution lies in $\mathcal{W}$. Fortunately, we can write the constrained minimization problem in (2.11) as a linear programming problem:

$$\min_{\mathbf{w},\mathbf{u},\mathbf{v}} \left\{ \tau \mathbf{1}'_n \mathbf{u} + (1-\tau)\mathbf{1}'_n \mathbf{v} \left| \sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)} \hat{\Theta}_{i(m)} + u_i - v_i = y_i, i = 1, ...n \right. \right\}$$

$$st \quad : \quad 0 \leq u_i,\ 0 \leq v_i \text{ for } i = 1, ..., n,\ 0 \leq w_m \leq 1 \text{ for } m = 1, ..., M, \text{ and } \sum_{m=1}^{M} w_m = 1,$$

where $\mathbf{u} \equiv (u_1, u_2, ..., u_n)$ and $\mathbf{v} \equiv (v_1, v_2, ..., v_n)$ are the positive and negative slack variables and $\mathbf{1}_n$ is $n \times 1$ vector of ones. This linear programming can be implemented in standard software. For example, one can use the algorithm of *linprog* in Matlab.

Let $(y, \mathbf{x})$ be an independent copy of $(y_i, \mathbf{x}_i)$ and $\mathcal{D}_n \equiv \{(y_i, \mathbf{x}_i)\}_{i=1}^n$. Define the *out-of-sample* quantile prediction error (or *final prediction error*, FPE) as follows

$$FPE_n(\mathbf{w}) = E\left[\rho_\tau \left(y - \sum_{m=1}^{M} w_m \mathbf{x}'_{(m)} \hat{\Theta}_{(m)} \right) \middle| \mathcal{D}_n \right], \tag{2.13}$$

where $\mathbf{x}_{(m)} \equiv (x_{1(m)}, ..., x_{k_m(m)})'$ and $x_{j(m)}$, $j = 1, ..., k_m$, are variables in $\mathbf{x}$ that correspond to the k_m regressors in the m'th model. We will show that $\hat{\mathbf{w}}$ is asymptotically optimal in terms of minimizing $FPE_n(\mathbf{w})$.

3 Asymptotic Optimality

In this section we first study the asymptotic properties of $\hat{\Theta}_{(m)}$ and $\hat{\Theta}_{i(m)}$ for a fixed model, and then show that jackknife weight $\hat{\mathbf{w}}$ is asymptotically optimal in terms of minimizing $FPE_n(\mathbf{w})$.

3.1 Asymptotic properties of $\hat{\Theta}_{(m)}$ and $\hat{\Theta}_{i(m)}$

Since Koenker and Bassett (1978) a large literature on quantile regression (QR) has developed; see Koenker (2005) for an excellent exposition on this. While QR estimates are as easy to compute as OLS regression coefficients, most of the theoretical and applied work on QR postulates a correctly specified parametric (usually linear) model for conditional quantiles. Asymptotic theory for QR under misspecification is limited. Angrist, Chernozhukov, and Fernández-Val (2006, ACF hereafter) study QR under misspecification and show that the QR minimizes a weighted mean-squared error loss function for specification error and establish the asymptotic distributional result for QR estimators when the number of parameters is fixed.

For model averaging, all the models under investigation are potentially misspecified. So the classical distributional result for QR estimator in Koenker (2005) cannot be used. In addition, we allow *diverging* number of parameters in some or all QR models. This means that the results in ACF (2006) are not applicable either. Although there are some asymptotic results in the literature on M-estimation which allow diverging number of parameters (see, e.g., Portnoy (1984, 1985) for smooth influence functions

and Welsh (1989) for smooth or monotone influence functions), none of these allow the models to be misspecified. Therefore we need to study the asymptotic properties of $\hat{\Theta}_{(m)}$ and $\hat{\Theta}_{i(m)}$ when the underlying model is potentially misspecified and the number of parameters in the model may diverge to infinity with the sample size n.

To proceed, let $f(\cdot|\mathbf{x}_i)$ and $F(\cdot|\mathbf{x}_i)$ denote the conditional probability density function (PDF) and cumulative distribution function (CDF) of ε_i given $\mathbf{x}_i$, respectively. Let $f_{y|\mathbf{x}}(\cdot|\mathbf{x}_i)$ denote the conditional PDF of y_i given $\mathbf{x}_i$. Define the pseudo-true parameter

$$\Theta^*_{(m)} \equiv \arg\min_{\Theta_{(m)}} E\left[\rho_\tau\left(y_i - \mathbf{x}'_{i(m)}\Theta_{(m)}\right)\right]. \tag{3.1}$$

In view of the fact that the objective function in (3.1) is convex, one can readily show that $\Theta^*_{(m)}$ exists and is unique under Assumptions A.1-A.3 given below. For $m = 1, ..., M$, define

$$\begin{aligned} A_{(m)} &\equiv E\left[f\left(-u_{i(m)}|\mathbf{x}_i\right)\mathbf{x}_{i(m)}\mathbf{x}'_{i(m)}\right] = E\left[f_{y|\mathbf{x}}(\Theta^{*\prime}_{(m)}\mathbf{x}_{i(m)}|\mathbf{x}_i)\mathbf{x}_{i(m)}\mathbf{x}'_{i(m)}\right], \\ B_{(m)} &\equiv E\left[\psi_\tau\left(\varepsilon_i + u_{i(m)}\right)^2\mathbf{x}_{i(m)}\mathbf{x}'_{i(m)}\right], \\ V_{(m)} &\equiv A^{-1}_{(m)}B_{(m)}A^{-1}_{(m)}, \end{aligned} \tag{3.2}$$

where $u_{i(m)} \equiv \mu_i - \mathbf{x}'_{i(m)}\Theta^*_{(m)}$ indicates the approximation bias for the m'th QR model and $\psi_\tau(\varepsilon_i) \equiv \tau - \mathbf{1}\{\varepsilon_i \leq 0\}$. Let $\bar{k} \equiv \max_{1\leq m\leq M} k_m$.

We make the following assumptions.

Assumption A.1. (i) $(y_i, \mathbf{x}_i)$, $i = 1, ..., n$, are IID such that (2.3) holds.

(ii) $P(\varepsilon_i(\tau) \leq 0|\mathbf{x}_i) = \tau$ a.s.

(iii) $E\left(\mu_i^4\right) < \infty$ and $\sup_{j\geq 1} E(x_{ij}^8) \leq c_{\mathbf{x}}$ for some $c_{\mathbf{x}} < \infty$.

Assumption A.2 (i) $f_{y|\mathbf{x}}(\cdot|\mathbf{x}_i)$ is bounded above by a finite constant c_f and continuous over its support a.s.

(ii) There exist constants $\underline{c}_{A(m)}$ and $\bar{c}_{A(m)}$ that may depend on k_m such that $0 < \underline{c}_{A(m)} \leq \lambda_{\min}\left(A_{(m)}\right) \leq \lambda_{\max}\left(A_{(m)}\right) \leq c_f\lambda_{\max}(E[\mathbf{x}_{i(m)}\mathbf{x}'_{i(m)}]) \leq \bar{c}_{A(m)} < \infty$.

(iii) There exist constants $\underline{c}_{B(m)}$ and $\bar{c}_{B(m)}$ that may depend on k_m such that $0 < \underline{c}_{B(m)} \leq \lambda_{\min}\left(B_{(m)}\right) \leq \lambda_{\max}\left(B_{(m)}\right) \leq \bar{c}_{B(m)} < \infty$.

(iv) $(\bar{c}_{A(m)} + \bar{c}_{B(m)})/k_m = O(\underline{c}^2_{A(m)})$.

Assumption A.3 Let $\underline{c}_A \equiv \min_{1\leq m\leq M}\underline{c}_{A(m)}$, $\underline{c}_B \equiv \min_{1\leq m\leq M}\underline{c}_{B(m)}$, $\bar{c}_A \equiv \max_{1\leq m\leq M}\bar{c}_{A(m)}$, and $\bar{c}_B \equiv \max_{1\leq m\leq M}\bar{c}_{B(m)}$.

(i) As $n \to \infty$, $\bar{k}^4\bar{c}_A/(n\underline{c}_B) \to 0$, and $\bar{k}^4(\log n)^4/(n\underline{c}^2_B) \to 0$.

(ii) $nM\, n^{-0.5L^2\bar{k}\underline{c}^3_A/(\bar{c}_A\bar{c}_B)} = o(1)$ for a sufficiently large constant L.

Assumption A.1(i) specifies the data are IID. It is easy to see that the results in this paper continue to hold for weakly dependent time series data under some mixing conditions.[4] A.1(ii) specifies the

[4]For the weakly dependent data, our method does not take into account the dependence between the in-sample and out-sample data. In other words, we pretend that the in-sample and out-sample data are independent, following the same strategy taken by Hansen (2008) and Zhang et al. (2013).

quantile restriction. A.1(iii) implies that $E\left\|\mathbf{x}_{i(m)}\right\|^{2s} \leq c_{\mathbf{x}} k_m^s$ for $m = 1, ..., M$ and $s = 1, 2, 4$. These moment conditions are needed for the application of Boole's, Bernstein's, and Markov's inequalities to obtain the uniform probability orders of certain sample mean objects. The first requirement in A.1(iii) is implied by $\sum_{j=1}^{\infty} \left|\beta_j\right| < \infty$, and $\sum_{j=1}^{\infty} |\alpha_j| < \infty$, in conjunction with the last condition in A.1(iii). Our asymptotic study mainly requires the finiteness of $E(\mu_i^4)$ so that the absolute summability of these coefficients are not necessary.

A.2(i) is weak and it allows conditional heteroskedasticity in the QR model. A.2(ii)-(iii) are often assumed in typical QR models when the regressors and quantile error terms are not independent of each other. Note that we allow $\underline{c}_{A(m)}$, $\bar{c}_{A(m)}$, $\underline{c}_{B(m)}$, and $\bar{c}_{B(m)}$ to depend on the dimension (k_m) of the regressors in model m. In particular, it is possible that $\bar{c}_{A(m)}$ and $\bar{c}_{B(m)}$ diverge to infinity and $\underline{c}_{A(m)}$ and $\underline{c}_{B(m)}$ converge to zero, both at slow rates when $k_m \to \infty$. The rates are restricted in A.2(iv) so that the usual $\sqrt{k_m/n}$-consistency for the parameter estimate is not affected.

A.3(i)-(ii) impose restrictions on the largest dimension of the models $\left(\bar{k}\right)$, the potential number of models under investigation (M), and the constants $\underline{c}_A$, $\underline{c}_B$, $\bar{c}_A$,and $\bar{c}_B$. A.3(i) is comparable with the conditions in the literature on inference with diverging number of parameters. For example, to obtain the distributional result, Welsh (1989) requires $p^4 \left(\log n\right)^2 / n \to 0$ for M-estimation with discontinuous but monotone influence function by assuming that the regressors are nonrandom and uniformly bounded and the error terms are homoskedastic, and Fan and Peng (2004) and Lam and Fan (2008) require $p^5/n \to 0$ for their nonconcave penalized likelihood and profile-kernel likelihood estimation, respectively, where p is the number of parameters in their models. A.3(ii) suggests that we allow M to grow at a polynomial rate with n.

The following theorem studies the asymptotic property of $\hat{\Theta}_{(m)}$, which is of interest in its own.

Theorem 3.1 *Suppose Assumptions A.1-A.3 hold. Let $C_{(m)}$ denote an $l_m \times k_m$ matrix such that $C_0 \equiv \lim_{n\to\infty} C_{(m)} C'_{(m)}$ exists and is positive definite, where $l_m \in [1, k_m]$ is a fixed integer. Then*

(i) $\left\|\hat{\Theta}_{(m)} - \Theta^*_{(m)}\right\| = O_P\left(\sqrt{k_m/n}\right)$;

(ii) $\sqrt{n} C_{(m)} V_{(m)}^{-1/2} [\hat{\Theta}_{(m)} - \Theta^*_{(m)}] \xrightarrow{d} N\left(0, C_0\right)$.

The proof of the rate of convergence in Theorem 3.1(i) is standard and straightforward. But this is not the case for the proof of asymptotic normality in Theorem 3.1(ii). Intuitively, we allow the number of parameters k_m in the m'th QR model to diverge to infinity with n. For this reason, we cannot consider and derive the asymptotic normality of $\hat{\Theta}_{(m)}$ itself as in Pollard (1991), Knight (1998), or Koenker (2005, ch.4.2) by using the convexity lemma. Instead, we prove the asymptotic normality for any arbitrary linear combinations of elements of $\hat{\Theta}_{(m)}$ by relying on the stochastic equicontinuity of the gradient function as argued in Ruppert and Carroll (1980) and extending the usual Euclidean norm for a fixed dimensional vector to a weighted norm for a vector with possible diverging dimension. In the special case where k_m is fixed, we can simply take $C_{(m)} = I_{k_m}$ and obtain the usual asymptotic

normality result as in ACF (2006). In this case, we have the usual Bahadur representation:[5]

$$\sqrt{n}\left(\hat{\Theta}_{(m)}-\Theta^{*}_{(m)}\right)=A^{-1}_{(m)}n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_{i(m)}\left[\tau-\mathbf{1}\left\{y_i\leq\Theta^{*\prime}_{(m)}\mathbf{x}_{i(m)}\right\}\right]+o_p\left(1\right), \tag{3.3}$$

where the first order condition to the minimization problem in (3.1) yields[6]

$$E\left\{\left[\tau-\mathbf{1}\left\{y_i\leq\Theta^{*\prime}_{(m)}\mathbf{x}_{i(m)}\right\}\right]\mathbf{x}_{i(m)}\right\}=0. \tag{3.4}$$

When $k_m\to\infty$ as $n\to\infty$, the Bahadur representation for $\sqrt{n}[\hat{\Theta}_{(m)}-\Theta^{*}_{(m)}]$ is quite complicated and reported in (A.15) at the end of Appendix A.

Note that $\hat{\Theta}_{i(m)}$ is asymptotically equivalent to $\hat{\Theta}_{(m)}$ and its asymptotic normality follows from Theorem 3.1(ii). The following theorem studies the uniform convergence property of $\hat{\Theta}_{(m)}$ and $\hat{\Theta}_{i(m)}$.

Theorem 3.2 *Suppose Assumptions A.1-A.3 hold. Then*
(*i*) $\max_{1\leq i\leq n}\max_{1\leq m\leq M}\left\|\hat{\Theta}_{i(m)}-\Theta^{*}_{(m)}\right\|=O_P\left(\sqrt{n^{-1}\bar{k}\log n}\right)$;
(*ii*) $\max_{1\leq m\leq M}\left\|\hat{\Theta}_{(m)}-\Theta^{*}_{(m)}\right\|=O_P\left(\sqrt{n^{-1}\bar{k}\log n}\right)$.

The above theorem establishes the uniform convergence of $\hat{\Theta}_{i(m)}$ and $\hat{\Theta}_{(m)}$ to $\Theta^{*}_{(m)}$. Under our conditions, the uniform convergence rate depends only on the sample size n and the largest number of parameters $\bar{k}$ in all M models under investigation.

The proof of Theorem 3.2 is not technically trivial because there is no closed form expression for the QR estimator. Fortunately, Rice (1984) demonstrates that one can choose a bandwidth in nonparametric regression based on an unbiased estimate of the relevant mean squared error and prove the bandwidth thus chosen is asymptotically optimal. We extend Rice's proof strategy to prove the uniform convergence of our QR estimators by using Shibata's (1981, 1982) inequality for χ^2 distributions.

3.2 Asymptotic optimality of the jackknife model averaging

Following Li (1987), Andrews (1991) and Hansen (2007), Hansen and Racine (2012) demonstrate the asymptotic optimality of their jackknife selected weight vector in the sense of making the average squared error and its associated conditional risk as small as possible among all feasible weight vectors. In their case, the conditional risk is equivalent to the out-of-sample prediction mean squared error (MSE). So the optimally chosen weight vector also minimizes their out-of-sample prediction MSE.

Unfortunately, in our QR framework, we cannot demonstrate the asymptotic equivalence between the conditional risk in (2.9) and the out-of-sample quantile prediction error in (2.13) under general conditions. Nevertheless, we can show that our JMA selected weight vector $\hat{\mathbf{w}}$ is optimal in the sense of making the FPE as small as possible among all feasible weight vectors. Specifically, we prove the following theorem.

[5] A close examination of ACF (2006) indicates a negative sign is missing in their representation of the influence function.

[6] If the m'th QR model is correctly specified, then $E\left[\psi_\tau\left(\varepsilon_i\right)\ |\mathbf{x}_{i(m)}\right]=E\left[\tau-1\{y_i\leq\Theta^{*\prime}_{(m)}\mathbf{x}_{i(m)}\}|\mathbf{x}_{i(m)}\right]=0$ a.s.

Theorem 3.3 *Suppose Assumptions A.1-A.3 hold. Suppose that* $n^3/[\bar{k}^2 M(\log M)^4] \to \infty$ *as* $n \to \infty$. *Then* $\hat{\mathbf{w}}$ *is asymptotically optimal in the sense that*

$$\frac{FPE\left(\hat{\mathbf{w}}\right)}{\inf_{\mathbf{w}\in\mathcal{W}} FPE\left(\mathbf{w}\right)} = 1 + o_P\left(1\right). \tag{3.5}$$

The optimality statement in the above theorem specifies an oracle property for the JMA selected weight vector $\hat{\mathbf{w}}$ in the sense that $\hat{\mathbf{w}}$ is asymptotically equivalent to the infeasible best weight vector to minimize $FPE\left(\mathbf{w}\right)$. As in the case of least squares model averaging, the limitation of such an optimality property is obvious: it restricts attention to the estimators which are weighted average of the original estimators $\hat{\mu}_{i(m)} \equiv \mathbf{x}'_{i(m)}\hat{\Theta}_{(m)}$, $m = 1, ..., M$. If one changes the set of models under consideration, then $\hat{\mathbf{w}}$ will also change and any JMA estimator based on the given set of models may not outperform an estimator that is not considered by the given models. Similar remarks also hold for Hansen's (2007) MMA estimator and Hansen and Racine's (2012) and Zhang et al.'s (2013) JMA estimator.

3.3 Quantile regression information criterion

Despite the asymptotic optimality of the QR jackknife model averaging, it is computationally expensive. The speed of calculating the QR JMA estimator may slow down significantly if the number of models (M) and the number of observations (n) are both large. For this reason, it is worthwhile to propose a Mallows-type information criterion for QR model averaging.

Let $\hat{\varepsilon}_{i(m)} \equiv y_i - \mathbf{x}'_{i(m)}\hat{\Theta}_{(m)}$ and $\hat{\varepsilon}_i\left(\mathbf{w}\right) \equiv \sum_{m=1}^{M} w_m \hat{\varepsilon}_{i(m)}$. Define the *quantile regression information criterion* (QRIC) as

$$QRIC_n\left(\mathbf{w}\right) = nQ_n\left(\mathbf{w}\right) + \frac{\tau\left(1-\tau\right)}{f\left(F^{-1}\left(\tau\right)\right)} \sum_{m=1}^{M} w_m k_m, \tag{3.6}$$

where $Q_n\left(\mathbf{w}\right) \equiv n^{-1}\sum_{i=1}^{n} \rho_\tau\left(\hat{\varepsilon}_i\left(\mathbf{w}\right)\right)$ indicates the average *in-sample* QR prediction error, F and f denote the CDF and PDF of $\varepsilon_i\left(\tau\right)$, respectively, and $\sum_{m=1}^{M} w_m k_m$ signifies the number of effective parameters in the combined estimator. To extend AIC to the least absolute deviation (LAD) regression, Burman and Nolan (1995) consider the following information criterion for model m under the assumption of correct model specification:

$$IC_{n,m} = \sum_{i=1}^{n} \left| y_i - \mathbf{x}'_{i(m)}\hat{\Theta}_{(m)} \right| + \frac{k_m}{2f\left(0\right)}. \tag{3.7}$$

Apparently, $QRIC_n\left(\mathbf{w}\right)$ generalizes the above information criterion to the general QR-based model averaging information criterion. In particular, in the case of LAD regression ($\tau = 0.5$), if $F^{-1}\left(0.5\right) = 0$ (i.e., the median model is correctly specified under the assumption that the error terms are independent of the regressors) and one assigns weight 1 to model m and 0 to all other models, then $QRIC_n\left(\mathbf{w}\right)$ reduces to $IC_{n,m}/2$.

In the supplementary appendix (Appendix D) we motivate the derivation of the criterion function in (3.6) from the perspective of JMA, which is similar to the Mallows criterion in the least squares regression framework. Interestingly, we demonstrate that the second term on the right hand side of (3.6) signifies

the dominant term in the difference between $nCV_n(\mathbf{w})$ and $nQ_n(\mathbf{w})$ under certain conditions:

$$nCV_n(\mathbf{w}) = nQ_n(\mathbf{w}) + \frac{\tau(1-\tau)}{f(F^{-1}(\tau))}\sum_{m=1}^{M} w_m k_m + \text{ smaller order terms.} \tag{3.8}$$

One condition requires that the error term $\varepsilon_i(\tau)$ should be independent of the regressor $\mathbf{x}_i$. Another condition requires that all M models should be approximately correct in the sense the model approximation biases are asymptotically $o(1)$ almost surely. The latter condition can be met in the case of nonparametric sieve estimation where $k_m \to \infty$ as $n \to \infty$ for $m = 1, ..., M$. Alternatively, it is also automatically satisfied if one would like to consider the local asymptotic framework as in Hjort and Clasekens (2003), Leung and Barron (2006), Pötscher (2006), Clasekens and Hjort (2008), Hansen (2009, 2010), and Liu (2012) so that all models under consideration are asymptotically correctly specified. Note that these two conditions are not required for our JMA estimator.

To use the above QRIC to select the weight vector $\mathbf{w}$, one has to estimate $s(\tau) \equiv 1/f\left(F^{-1}(\tau)\right)$, the sparsity function of $\varepsilon_i(\tau)$. Following Koenker (2005, p. 77 and p. 139), we can estimate $s(\tau)$ by

$$\tilde{s}(\tau) = \left[\tilde{F}_n^{-1}(\tau + h_n) - \tilde{F}_n^{-1}(\tau - h_n)\right]/(2h_n)$$

where $\tilde{F}_n^{-1}$ is an estimate of the quantile function F^{-1} of $\varepsilon_i(\tau)$ based on the quantile residual obtained from the largest approximating model, $h_n = n^{-1/5}\{4.5\phi^4\left(\Phi^{-1}(\tau)\right)/[2\Phi^{-1}(\tau)^2 + 1]^2\}^{1/5}$, and ϕ and Φ are the standard normal PDF and CDF, respectively. Define

$$\tilde{\mathbf{w}} \equiv (\tilde{w}_1, ..., \tilde{w}_M) \equiv \arg\min_{\mathbf{w}\in\mathcal{W}} \widetilde{QRIC}_n(\mathbf{w}) = \arg\min_{\mathbf{w}\in\mathcal{W}} \left[Q_n(\mathbf{w}) + \tau(1-\tau)\tilde{s}(\tau)\sum_{m=1}^{M} w_m k_m\right], \tag{3.9}$$

the empirical QRIC selected weight vector. Obviously, there is no closed-form solution to (3.9) and one has to find the optimal weight vector by linear programming as in typical quantile regressions. Given $\tilde{\mathbf{w}}$, one can obtain the QRIC-based estimator of μ_i by

$$\tilde{\mu}_i(\tilde{\mathbf{w}}) = \sum_{m=1}^{M} \tilde{w}_m \mathbf{x}'_{i(m)} \hat{\Theta}_{(m)}. \tag{3.10}$$

We will examine the performance of this estimator with that of $\hat{\mu}_i(\hat{\mathbf{w}})$ in (2.12) through simulations.

4 Monte Carlo Simulations

In this section, we conduct a small set of Monte Carlo simulations to evaluate the finite sample performance of our proposed quantile regression model averaging estimators.

4.1 Data generating processes

The first DGP is similar to that in Hansen (2007):

$$\text{DGP 1: } y_i = \theta \sum_{j=1}^{1000} j^{-1} x_{ij} + \varepsilon_i,$$

where $x_{i1} = 1$ and x_{ij}, $j = 2, 3....$, are each IID $N(0,1)$ and mutually independent of each other. For ε_i, we consider two cases: (1) homoskedasticity where ε_i is $N(0,1)$ and independent of x_{ij}, $j = 2, 3, ...$; (2) heteroskedasticity where $\varepsilon_i = \sum_{j=2}^{6} x_{ij}^2 \epsilon_i$ and ϵ_i is $N(0,1)$ and independent of $x_{ij}, j = 2, 3, ...$ As in Hansen (2007), the population $R^2 = [var(y_i) - var(\varepsilon_i)]/var(y_i)$ is controlled by θ.[7] We consider different choices of θ such that $R^2 = 0.1, 0.2, ..., 0.9$. We consider the sample size $n = 50$, 100, and 150 and the number of models is given by $M = \lfloor 3n^{1/3} \rfloor$ where $\lfloor \cdot \rfloor$ denotes the integer part of $\cdot$. For simplicity, we consider nested models by specifying $x_{i(1)} = \{x_{i1}\}$, $x_{i(2)} = \{x_{i1}, x_{i2}\}$, etc.

The second DGP is

$$\text{DGP 2: } y_i = \theta \frac{\exp(x_i)}{1 + \exp(x_i)} + \varepsilon_i,$$

where x_i follows IID Weibull(1,1) distribution. For the homoskedastic case, ε_i is IID $N(0,1)$. For the heteroskedastic case, $\varepsilon_i = (0.01 + x_i)\epsilon_i$, where ϵ_i is IID $N(0,1)$. We choose different θ's to control for the population R^2 such that $R^2 = 0.1, 0.2, ..., 0.9$. We consider nonparametric sieve estimators of the τth conditional quantile function of y_i given x_i. Specifically, we use Hermite polynomials to approximate the unknown function $\theta \exp(x_i)/(1 + \exp(x_i))$. The jth term in the Hermite polynomial is:

$$x_{ij} = (x_i - \bar{x})^{j-1} \cdot \exp\left[\frac{-(x_i - \bar{x})^2}{2s_x^2}\right], \ j = 1, 2, ...,$$

where $\bar{x}$ and s_x are the sample mean and standard deviation of $\{x_i\}$, respectively. We consider $M = \lfloor 3n^{1/3} \rfloor$ nested models: $x_{i(1)} = \{x_{i1}\}$, $x_{i(2)} = \{x_{i1}, x_{i2}\}$, ..., for $n = 50$, 100, and 150.

The third DGP is similar to DGP 1, but we consider different distributions of x_{ij}'s and different heteroskedasticity structures, and fix the number of models for all sample sizes under investigation. Specifically,

$$\text{DGP 3: } y_i = \theta \sum_{j=1}^{30} j^{-1} x_{ij} + \varepsilon_i,$$

where $x_{i1} = 1$; x_{ij}, $j = 2, 3, ..., 50$, are each IID $\chi^2(1)$ and mutually independent of each other; and $\varepsilon_i = \epsilon_i$ and $\varepsilon_i = \sum_{j=2}^{30} j^{-1} x_{ij} \epsilon_i$ for the homoskedasticity and heteroskedasticity cases, respectively, where ϵ_i is normalized $\chi^2(3)$ with mean zero and variance one and independent of x_{ij}'s. Different θ's are chosen to control for the population R^2 such that $R^2 = 0.1, 0.2, ..., 0.9$. The number of models (M) is fixed at 20, which is relatively large for the sample sizes we consider here ($n = 50$, 100, and 150). Again, only nested models are considered: $x_{i(1)} = \{x_{i1}\}$, $x_{i(2)} = \{x_{i1}, x_{i2}\}$, ..., $x_{i(20)} = \{x_{i1}, x_{i2}, ..., x_{i20}\}$ for $m = 1, 2, ..., 20$, respectively.

The fourth DGP is nonlinear in each term:

$$\text{DGP 4: } y_i = \theta \left[x_{1i} + \sum_{j=2}^{25} j^{-1} \Phi(x_{ij}) \right] + \varepsilon_i,$$

where $x_{i1} = 1$, the remaining x_{ij}'s ($j = 2, ..., 25$) are each IID $N(0,1)$ and mutually independent of each other, and $\Phi(\cdot)$ is the standard normal CDF function. $\varepsilon_i = \epsilon_i$ and $\varepsilon_i = (0.01 + \sum_{j=2}^{11} x_{ij}^2)\epsilon_i$ for the

[7] For the ease of generating data in the simulation, here we use R^2 defined in the least square sense, as in Hansen (2007). Alternatively, one may use the R^2 defined for quantile regressions, see, e.g., Koenker and Machado (1999, eq.(7)).

homoskedasticity and heteroskedasticity cases respectively, where ϵ_i is IID $N(0,1)$ and independent of x_{ij}'s. Different θ's are chosen to ensure that the population $R^2 = 0.1, 0.2, ..., 0.9$. We consider $M = 20$ nested linear models: $x_{i(1)} = \{x_{i1}\}$, $x_{i(2)} = \{x_{i1}, x_{i2}\}, ..., x_{i(20)} = \{x_{i1}, x_{i2}, ..., x_{i20}\}$ for different sample sizes ($n = 50$, 100, and 150) and for $m = 1, 2, ..., 20$, respectively.

4.2 Implementation

We use the following four methods to choose the weights in the model averaging: (*i*) JMA or cross-validation (CV) averaging, (*ii*) AIC model averaging, (*iii*) BIC model averaging, and (*iv*) QRIC.[8] JMA and QRIC are defined above. AIC and BIC model averaging are alternative ways of implementing model averaging. They are often referred to as "smoothed" averaging (see, e.g., Buckland et al. (1997)), which uses exponential weights of the form $\exp(-I_m/2)/\sum_{j=1}^{M}\exp(-I_j/2)$, where I_m is an information criterion for model m. In the quantile regression context, following Machado (1993), for the mth model, the AIC and BIC are respectively defined as

$$\begin{aligned} AIC_m &= 2n\ln\left[\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i - \hat{\Theta}'_{(m)}\mathbf{x}_{i(m)}\right)\right] + 2k_m, \text{ and} \\ BIC_m &= 2n\ln\left[\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i - \hat{\Theta}'_{(m)}\mathbf{x}_{i(m)}\right)\right] + k_m\ln(n). \end{aligned}$$

Thus, the AIC and BIC weights for model m are respectively defined as

$$\hat{w}_m^{AIC} = \frac{\exp\left(-\frac{1}{2}AIC_m\right)}{\sum_{j=1}^{M}\left(\exp\left(-\frac{1}{2}AIC_j\right)\right)} \text{ and } \hat{w}_m^{BIC} = \frac{\exp\left(-\frac{1}{2}BIC_m\right)}{\sum_{j=1}^{M}\left(\exp\left(-\frac{1}{2}BIC_j\right)\right)}.$$

We evaluate each method using the out-of-sample quantile prediction error. For each replication, we generate $\{x_s, y_s\}_{s=1}^{100}$ as out-of-sample observations. For the rth replication, the final prediction error is calculated as

$$FPE(r) = \frac{1}{100}\sum_{s=1}^{100}\rho_\tau\left[y_s - \sum_{m=1}^{M}\hat{w}_m\cdot\hat{\Theta}'_{(m)}x_{s(m)}\right],$$

where $\hat{w}_m$ is chosen by one of the four methods. Then we average the out-of-sample prediction error over $R = 200$ replications: $FPE = \frac{1}{R}\sum_{r=1}^{R}FPE(r)$. The smaller FPE, the better the method in terms of the out-of-sample quantile prediction error.

4.3 Evaluations

We normalize the final quantile predication error by dividing by the prediction error of the infeasible optimal single model, as in Hansen (2007). To save space, we only report the results for the heteroskedasticity case. The results for the homoskedasticity case can be found in the supplementary appendix (see Figures S1-S4 in Appendix E).

[8] We also try the corresponding model selection criteria and find that the model selection is always dominated by the corresponding model averaging. For example, CV model selection is dominated by JMA model averaging. Thus we only show the results of model averaging.

Figure 1 shows the results for DGP 1 with $\tau = 0.5$ and $\tau = 0.05$. When $\tau = 0.5$, no method clearly denominates the others. The performances of QRIC are slightly better than JMA when the sample size is large. The performance of BIC seems to be the worst. When $\tau = 0.05$, it is clear that JMA dominates all the other three methods. BIC seems to be the second best.

Figure 2 shows the results for DGP 2. When $\tau = 0.5$, JMA is the best, followed by BIC. The performances of AIC and QRIC are worse than those of JMA and BIC. When $\tau = 0.05$, similar to DGP 1, JMA dominates all other three methods in all cases.

The results for DGPs 3 and 4 are reported in Figures 3 and 4, respectively. We find the same pattern that for $\tau = 0.05$, JMA is clearly the dominating method. When $\tau = 0.5$, there is no clear dominating method. But for most cases, both JMA and QRIC perform relatively well, especially when the sample size is not small.

In general, the performance of QRIC is poor when $\tau = 0.05$. One possible explanation is that QRIC requires the estimation of the sparsity function, which is difficult to estimate for $\tau = 0.05$. When $\tau = 0.5$, as a referee kindly points out, the performance of QRIC is also relatively poor for DGP 2. This could be due to the poor approximation of the sieve estimator when the number of terms is small. As discussed in Section 3.3, one condition for QRIC to work is that *all* models under consideration should be approximately correct. The relatively good performance of QRIC in DGPs 1, 3 and 4 could be due to the design of our DGPs. In all these three DGPs, the regressors (x_{ij} or $\Phi(x_{ij})$) have a coefficient j^{-1}, which means that the additional regressors become less and less important. Thus, ignoring the latter regressors may not lead to serious misspecification of the models.

5 Empirical Applications

5.1 Quantile forecast of excess stock returns

Forecasting quantiles of stock returns is widely used for VaR (Value-at-Risk) and essential to financial risk management. In this subsection, we apply our JMA and QRIC estimators to predict the quantiles of excess stock returns.[9]

The data is the same as in Campbell and Thompson (2008) and Jin et al. (2014). The data is monthly from January 1950 to December 2005 with total number of observations $T = 672$. The dependent variable y is the excess stock returns, which is defined as the monthly returns of S&P 500 index minus the risk-free rate. There are 12 regressors in the dataset as shown in Table 1.

The detailed explanations of these variables can be found in Jin et al. (2014). The order of the 12 regressors above is based on the absolute value of their correlation with the dependent variable. For example, x_1 has the largest absolute value of correlation with y. We construct 13 candidate nested models with regressors $\{1\}, \{1, x_1\}, ...\{1, x_1, x_2, ..., x_{12}\}$, respectively.

We construct one-period-ahead forecasts of the quantiles of excess stock returns using the fixed in-sample size (T_1) of 48, 60, 72, 96, 120, 144 and 180. We compare different forecast methods with

[9] In both empirical applications, we focus on quantile regression. The results for mean regressions are available upon request.

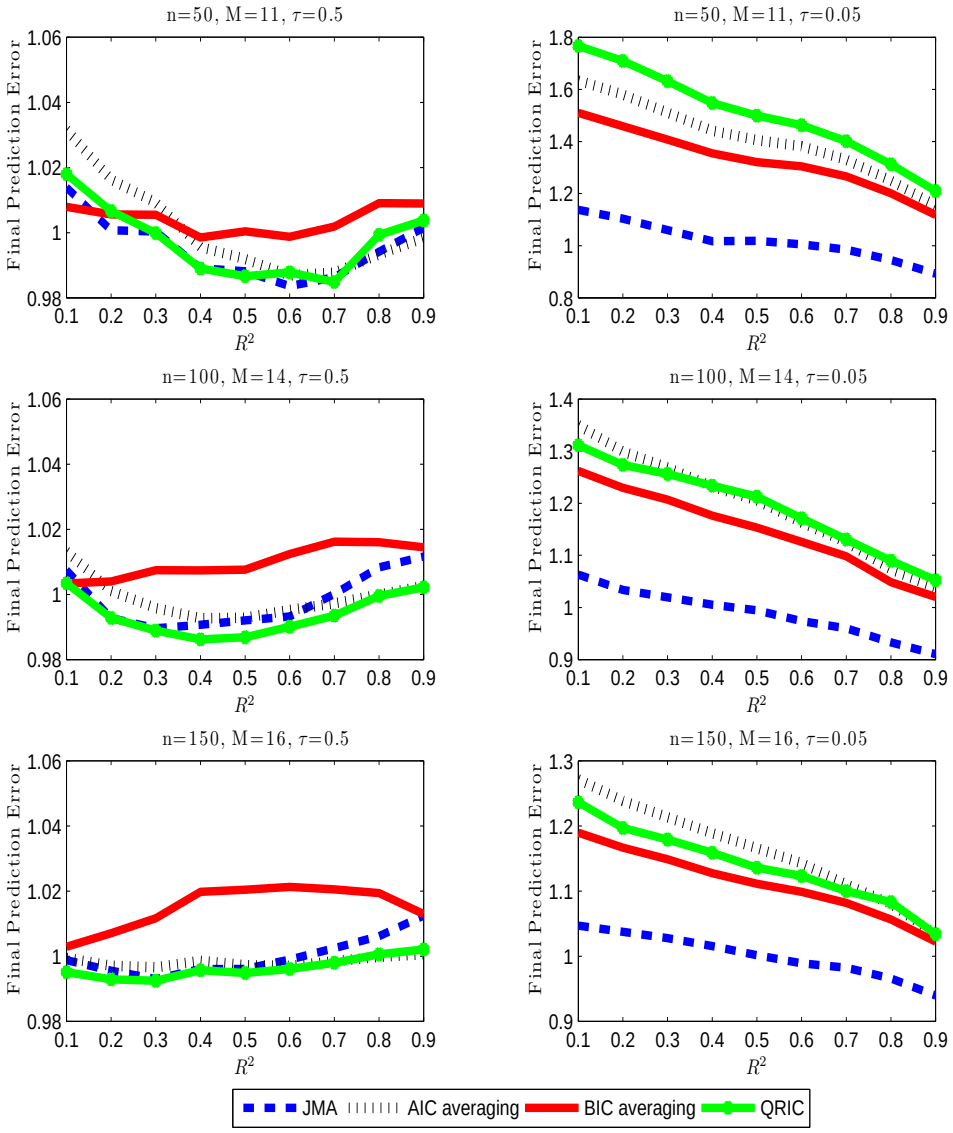

Figure 1: Out-of-sample performance: DGP 1, Heteroskedasticity

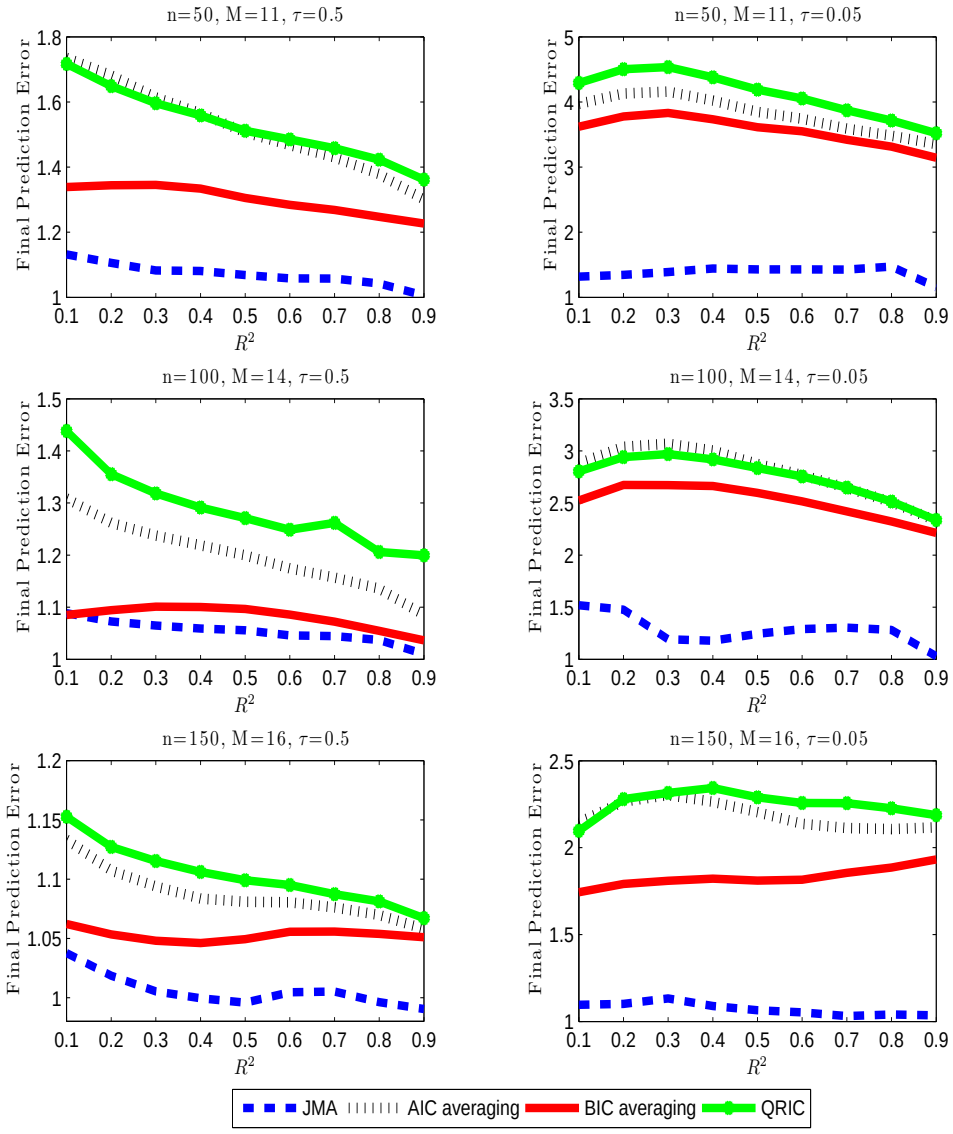

Figure 2: Out-of-sample performance: DGP 2, Heteroskedasticity

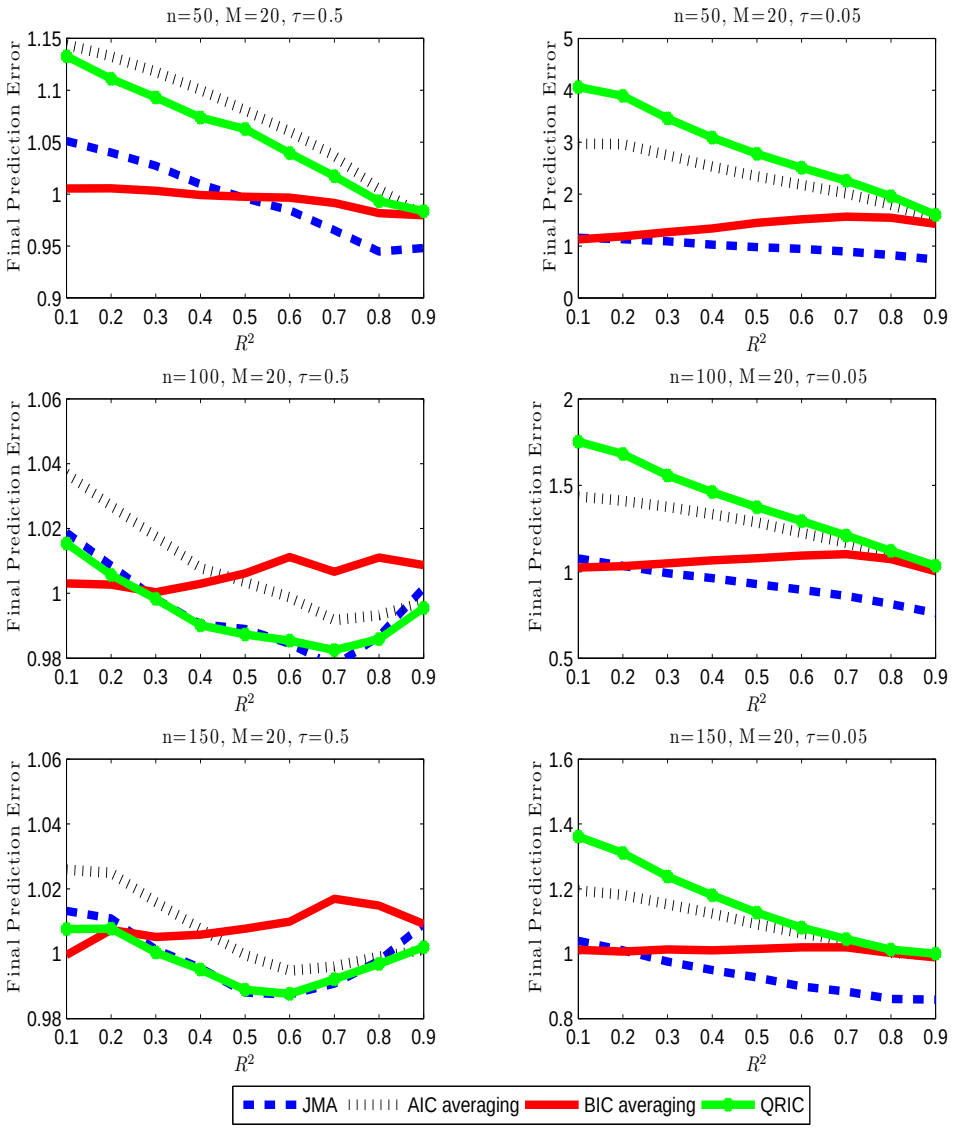

Figure 3: Out-of-sample performance: DGP 3, Heteroskedasticity

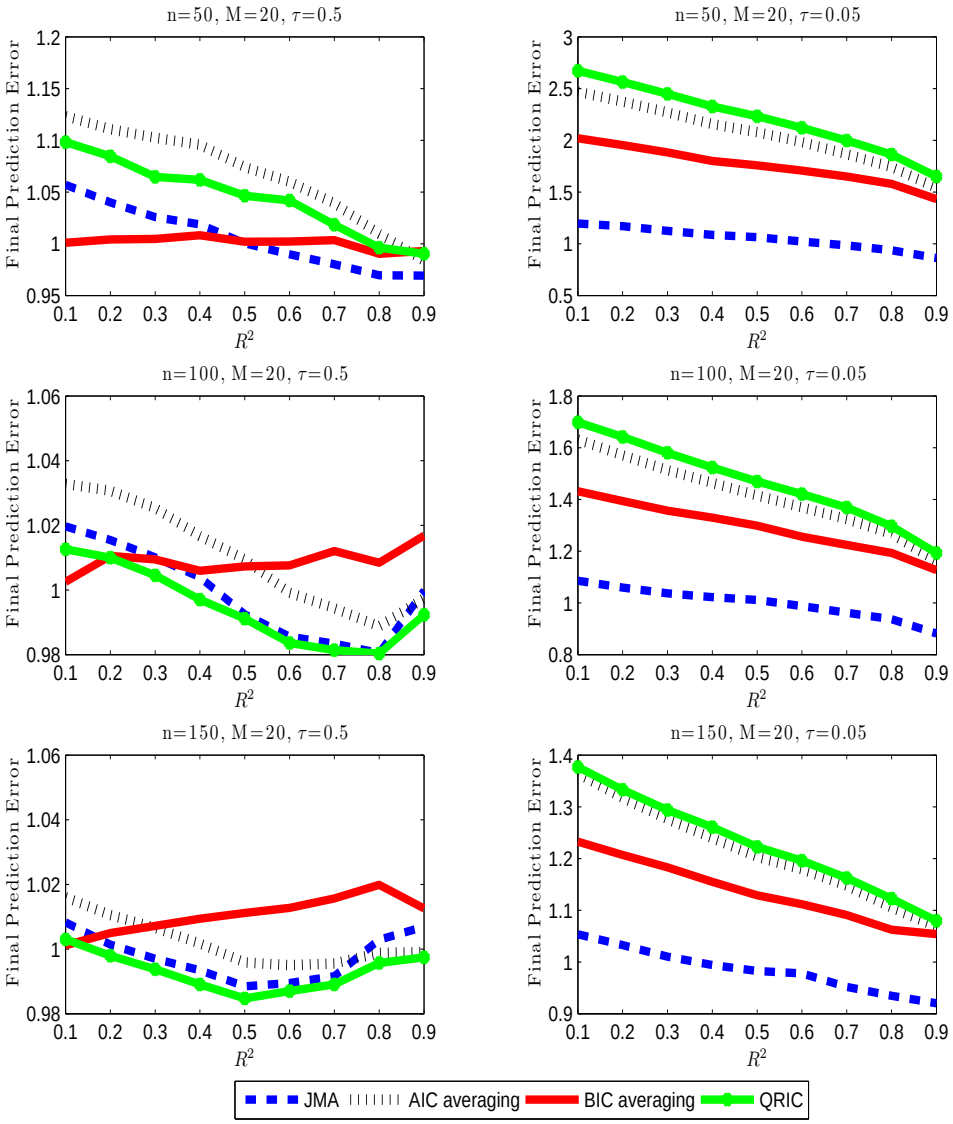

Figure 4: Out-of-sample performance: DGP 4, Heteroskedasticity

Table 1: Regressors for the stock returns data

Regressor	Names	Correlation with y
x_1	Default Yield Spread	0.075
x_2	Treasury Bill Rate	-0.063
x_3	Net Equity Expansion	0.056
x_4	Term Spread	0.053
x_5	Dividend Price Ratio	0.048
x_6	Earnings Price Ratio	0.043
x_7	Long Term Yield	0.042
x_8	Book-to-Market Ratio	-0.028
x_9	Inflation	0.019
x_{10}	Return on Equity	-0.015
x_{11}	Lagged dependent variable	0.014
x_{12}	Smoothed Earnings Price Ratio	-0.012

Table 2: Out-of-sample R^2 for the stock returns data

		Model Averaging				Model Selection		
T_1	τ	JMA	AIC	BIC	QRIC	CV	AIC	BIC
48	0.5	0.075	-0.009	0.007	0.031	-0.023	-0.012	-0.010
60	0.5	0.075	-0.014	0.012	0.038	-0.037	-0.027	-0.007
72	0.5	0.063	-0.026	-0.003	0.019	-0.043	-0.025	-0.017
96	0.5	0.058	-0.031	-0.018	0.011	-0.031	-0.034	-0.025
120	0.5	0.012	-0.044	-0.043	-0.011	-0.060	-0.044	-0.056
144	0.5	-0.003	-0.067	-0.060	-0.026	-0.068	-0.067	-0.074
180	0.5	0.032	-0.006	0.024	0.032	-0.014	-0.010	0.011
48	0.05	-0.135	-0.629	-0.602	-0.670	-0.527	-0.683	-0.653
60	0.05	-0.142	-0.680	-0.672	-0.695	-0.634	-0.694	-0.726
72	0.05	-0.007	-0.456	-0.450	-0.487	-0.390	-0.483	-0.460
96	0.05	-0.029	-0.251	-0.247	-0.210	-0.248	-0.253	-0.263
120	0.05	0.051	-0.135	-0.147	-0.136	-0.122	-0.140	-0.151
144	0.05	0.038	-0.107	-0.116	-0.100	-0.090	-0.107	-0.128
180	0.05	0.025	-0.043	-0.045	-0.034	-0.114	-0.041	-0.045

the simple historical unconditional quantile. Following Campbell and Thompson (2008), we define the out-of-sample R^2 as $R^2 = 1 - \frac{\sum_{t=T_1}^{T-1} \rho_\tau\left(y_{t+1} - \hat{y}_{t+1|t}\right)}{\sum_{t=T_1}^{T-1} \rho_\tau\left(y_{t+1} - \bar{y}_{t+1|t}\right)}$, where $\hat{y}_{t+1|t}$ is the one-period-ahead prediction of the τth quantile of the excess return at time t using data from the past T_1 periods (period t to period $t - T_1 + 1$); $\bar{y}_{t+1|t}$ is the simple historical τth unconditional quantile over the past T_1 periods and is used as the benchmark prediction. The results of the out-of-sample R^2 are presented in Table 2. It is clear that JMA dominates all other model averaging and model selection methods. As discussed in the literature, for predicting stock returns, the simple historical mean estimator cannot be easily beaten (see, e.g., Goyal and Welch, 2008). However, our JMA estimators can outperform the historical unconditional quantiles for many scenarios. For $\tau = 0.5$, JMA outperforms the benchmark for most of the cases. For $\tau = 0.05$, JMA outperforms the benchmark when the in-sample size is relatively large.

Since conditional quantiles can also be interpreted as the VaR, we also define the out-of-sample violation rate as the percentage that the out-of-sample realization is smaller than the prediction of its

Table 3: Out-of-sample violation rate

		Model Averaging				Model Selection		
T_1	τ	JMA	AIC	BIC	QRIC	CV	AIC	BIC
48	0.5	0.502	0.503	0.498	0.498	0.522	0.504	0.522
60	0.5	0.487	0.482	0.489	0.482	0.489	0.478	0.498
72	0.5	0.503	0.500	0.503	0.493	0.510	0.503	0.495
96	0.5	0.474	0.460	0.463	0.451	0.472	0.460	0.469
120	0.5	0.462	0.451	0.460	0.458	0.469	0.451	0.449
144	0.5	0.445	0.434	0.443	0.441	0.430	0.434	0.428
180	0.5	0.441	0.415	0.437	0.435	0.429	0.413	0.439
48	0.05	0.139	0.251	0.248	0.253	0.186	0.256	0.255
60	0.05	0.144	0.258	0.258	0.266	0.201	0.263	0.265
72	0.05	0.128	0.212	0.203	0.217	0.168	0.213	0.210
96	0.05	0.101	0.155	0.155	0.142	0.137	0.153	0.158
120	0.05	0.082	0.130	0.132	0.130	0.121	0.130	0.132
144	0.05	0.070	0.097	0.095	0.097	0.102	0.097	0.102
180	0.05	0.049	0.053	0.053	0.051	0.083	0.051	0.053

τth quantile (see, e.g., Kuester et al., 2006): $\hat{p} = \frac{1}{T-T_1}\sum_{t=T_1}^{T-1} \mathbf{1}\{y_{t+1} < \hat{y}_{t+1|t}\}$, where $\hat{y}_{t+1|t}$ is an estimator of the τth conditional quantile of y_{t+1} given x_t at time t. Thus, ideally, $\hat{p}$ should be close to τ. Table 3 presents the out-of-sample performance of various methods. The performances of all the methods are similar when $\tau = 0.5$. When $\tau = 0.05$, JMA clearly dominates all the other methods. Also in general, when the estimation sample size (T_1) increases, $\hat{p}$ becomes closer to 0.05 for all methods.

5.2 Quantile forecast of wages

In this subsection, our new averaging estimators are applied to predict the quantiles of wages. This is an important topic in labor economics, as quantiles are often used to characterize wage inequality (see, e.g., Angrist and Pischke, 2009, Chapter 7).

The data is a random sample of the US Current Population Survey for the year 1976 from Wooldridge (2003). It is a widely used dataset, see, e.g., Hansen and Racine (2012). The sample size is $n = 526$. The dependent variable is the logarithm of average hourly earnings. There are 20 regressors in the dataset. We order these regressors according to the absolute value of their correlation with the dependent variable and construct 11 candidate nested models using the first 10 regressors. These 10 regressors are shown in Table 4.

We randomly split the sample into an estimation sample of size n_1 and an evaluation sample of size $n_2 \equiv n - n_1$. We construct the out-of-sample R^2 : $R^2 = 1 - \frac{\sum_{s=1}^{n_2} \rho_\tau(y_s - \hat{y}_s)}{\sum_{s=1}^{n_2} \rho_\tau(y_s - \bar{y}_s)}$, where $\{y_s\}_{s=1}^{n_2}$ is the evaluation sample, $\{\hat{y}_s\}_{s=1}^{n_2}$ and $\{\bar{y}_s\}_{s=1}^{n_2}$ are the τth conditional quantile predictions and the unconditional quantile estimates using the estimation sample, respectively. We repeat this sample splitting exercise for 200 times and report the average of the out-of-sample R^2. We consider different estimation sample sizes $n_1 = 50$, 100, 150, and 200. Table 5 presents the out-of-sample R^2. For $\tau = 0.5$, the performances of JMA, AIC model averaging, and QRIC are similar and better than other methods. For $\tau = 0.05$, JMA clearly dominates all other methods. This again confirms the simulation results that JMA has the best

Table 4: Regressors for the wage data

Regressor	Names	Explanation	Correlation with y
x_1	Professional Occupation	=1 if in professional occupation	0.442
x_2	Education	years of education	0.406
x_3	Tenure	years with current employer	0.347
x_4	Female	=1 if female	-0.340
x_5	Service Occupation	=1 if in service occupation	-0.253
x_6	Married	=1 if married	0.229
x_7	Trade	=1 if in wholesale or retail	-0.190
x_8	SMSA	=1 if live in SMSA	0.178
x_9	Services	=1 if in services industry	-0.142
x_{10}	Clerk Occupation	=1 if in clerical occupation	-0.141

Table 5: Out-of-sample R^2 for the wage data

		Model Averaging				Model Selection		
n_1	τ	JMA	AIC	BIC	QRIC	CV	AIC	BIC
50	0.5	0.177	0.173	0.165	0.175	0.150	0.151	0.144
100	0.5	0.217	0.218	0.204	0.218	0.209	0.211	0.192
150	0.5	0.231	0.232	0.220	0.232	0.225	0.230	0.210
200	0.5	0.240	0.241	0.233	0.241	0.236	0.240	0.225
50	0.05	-0.071	-0.419	-0.255	-0.588	-0.139	-0.510	-0.342
100	0.05	0.036	-0.035	-0.015	-0.039	-0.036	-0.051	-0.044
150	0.05	0.064	0.033	0.032	0.036	0.012	0.026	0.019
200	0.05	0.074	0.061	0.060	0.063	0.038	0.057	0.054

performance for extreme quantiles.

6 Concluding Remarks

In this paper, we provide a new averaging quantile regression estimator, namely jackknife model averaging (JMA) quantile regression estimator. The estimator uses leave-one-out cross-validation to choose the weight. We show that the weight chosen by our method is asymptotically optimal in terms of minimizing the out-of-sample final prediction error. The numerical algorithm is also simple using a linear programming. Our simulations suggest that our new method outperforms the other QR model averaging and selection methods, especially for extreme quantiles. We apply our new JMA estimator to predict quantiles of excess stock returns and wages.

APPENDIX

In the following we will use Knight's (1998) identity repeatedly:

$$\rho_\tau(u+v) - \rho_\tau(u) = v\psi_\tau(u) + \int_0^{-v} [\mathbf{1}\{u \le s\} - \mathbf{1}\{u \le 0\}]\, ds,$$

where $\psi_\tau(u) = \tau - \mathbf{1}\{u < 0\}$. Recall that $Q_{n(m)}\left(\Theta_{(m)}\right) \equiv \sum_{i=1}^n \rho_\tau(y_i - \mathbf{x}'_{i(m)}\Theta_{(m)})$.

A Proof of Theorem 3.1

(i) Let $a_n = \sqrt{k_m/n}$. Let $\mathbf{v}_{(m)} \in \mathbb{R}^{k_m}$ such that $\left\|\mathbf{v}_{(m)}\right\| = C$ where C is a large enough constant. We want to show that for any given $\epsilon > 0$ there is a large constant C such that, for large n we have $P\left\{\inf_{\|\mathbf{v}_{(m)}\|=C} Q_{n(m)}\left(\Theta^*_{(m)} + \alpha_n \mathbf{v}_{(m)}\right) > Q_{n(m)}\left(\Theta^*_{(m)}\right)\right\} \geq 1-\epsilon$. This implies that with probability approaching 1 (w.p.a.1) there is a local minimum $\hat{\Theta}_{(m)}$ in the ball $\{\Theta^*_{(m)} + a_n \mathbf{v}_{(m)} : \left\|\mathbf{v}_{(m)}\right\| \leq C\}$ such that $||\hat{\Theta}_{(m)} - \Theta^*_{(m)}|| = O_P(a_n)$. It is also the global minimum by the convexity of $Q_{n(m)}$.

Let $u_{i(m)} \equiv \mu_i - \mathbf{x}'_{i(m)}\Theta^*_{(m)}$. Then by Knight's identity

$$\begin{aligned}
Z_{n(m)}\left(\mathbf{v}_{(m)}\right) &\equiv Q_{n(m)}\left(\Theta^*_{(m)} + \alpha_n \mathbf{v}_{(m)}\right) - Q_{n(m)}\left(\Theta^*_{(m)}\right) \\
&= \sum_{i=1}^{n}\left[\rho_\tau\left(\varepsilon_i + u_{i(m)} - a_n \mathbf{x}'_{i(m)}\mathbf{v}_{(m)}\right) - \rho_\tau\left(\varepsilon_i + u_{i(m)}\right)\right] \\
&= -a_n \sum_{i=1}^{n} \psi_\tau\left(\varepsilon_i + u_{i(m)}\right)\mathbf{x}'_{i(m)}\mathbf{v}_{(m)} + \sum_{i=1}^{n}\int_0^{a_n \mathbf{x}'_{i(m)}\mathbf{v}_{(m)}} \alpha_{i(m)}(s)\, ds \\
&= -a_n \sum_{i=1}^{n} \psi_\tau\left(\varepsilon_i + u_{i(m)}\right)\mathbf{x}'_{i(m)}\mathbf{v}_{(m)} + \sum_{i=1}^{n} E\left[\int_0^{a_n \mathbf{x}'_{i(m)}\mathbf{v}_{(m)}} \alpha_{i(m)}(s)\, ds \;|\mathbf{x}_i\right] \\
&\quad + \sum_{i=1}^{n}\left\{\int_0^{a_n \mathbf{x}'_{i(m)}\mathbf{v}_{(m)}} \alpha_{i(m)}(s)\, ds - E\left[\int_0^{a_n \mathbf{x}'_{i(m)}\mathbf{v}_{(m)}} \alpha_{i(m)}(s)\, ds \;|\mathbf{x}_i\right]\right\} \\
&\equiv Z_{n(m),1}\left(\mathbf{v}_{(m)}\right) + Z_{n(m),2}\left(\mathbf{v}_{(m)}\right) + Z_{n(m),3}\left(\mathbf{v}_{(m)}\right), \text{ say,}
\end{aligned} \tag{A.1}$$

where $\alpha_{i(m)}(s) = \mathbf{1}\left\{\varepsilon_i + u_{i(m)} \leq s\right\} - \mathbf{1}\left\{\varepsilon_i + u_{i(m)} \leq 0\right\}$.

The first order condition for the population minimization problem (3.1) implies that

$$E\left[\psi_\tau\left(\varepsilon_i + u_{i(m)}\right)\mathbf{x}_{i(m)}\right] = 0, \tag{A.2}$$

which is analogous to the last identity on page 545 of Angrist et al. (2006). It follows that by Assumption A.2(iii), $E\left|Z_{n(m),1}\left(\mathbf{v}_{(m)}\right)\right|^2 \leq a_n^2 \sum_{i=1}^n \mathbf{v}'_{(m)} E\{\left[\psi_\tau\left(\varepsilon_i + u_{i(m)}\right)\right]^2 \mathbf{x}_{i(m)}\mathbf{x}'_{i(m)}\}\mathbf{v}_{(m)} \leq \bar{c}_{B(m)} n a_n^2 \left\|\mathbf{v}_{(m)}\right\|^2$ and therefore by Chebyshev's inequality

$$Z_{n(m),1}\left(\mathbf{v}_{(m)}\right) = \bar{c}^{1/2}_{B(m)} O_P\left(a_n\sqrt{n}\right)\left\|\mathbf{v}_{(m)}\right\| = O_P\left(\bar{c}^{1/2}_{B(m)} a_n^2 n/\sqrt{k_m}\right)\left\|\mathbf{v}_{(m)}\right\|. \tag{A.3}$$

For $Z_{n(m),2}\left(\mathbf{v}_{(m)}\right)$, by the law of iterated expectations, Taylor expansion, Lebesgue dominated convergence theorem, and Assumption A.2(ii) we have w.p.a.1

$$\begin{aligned}
Z_{n(m),2}\left(\mathbf{v}_{(m)}\right) &= \sum_{i=1}^{n}\left[\int_0^{a_n \mathbf{x}'_{i(m)}\mathbf{v}_{(m)}} F\left(-u_{i(m)} + s|\mathbf{x}_i\right) - F\left(-u_{i(m)}|\mathbf{x}_i\right) ds\right] \\
&= \sum_{i=1}^{n}\left[\int_0^{a_n \mathbf{x}'_{i(m)}\mathbf{v}_{(m)}} f\left(-u_{i(m)}|\mathbf{x}_i\right) s ds\right]\{1 + o_P(1)\} \\
&= \frac{1}{2}a_n^2 \mathbf{v}'_{(m)}\left[\sum_{i=1}^{n} f\left(-u_{i(m)}|\mathbf{x}_i\right)\mathbf{x}_{i(m)}\mathbf{x}'_{i(m)}\right]\mathbf{v}_{(m)}\{1 + o_P(1)\} \\
&= \frac{1}{2}a_n^2 n \mathbf{v}'_{(m)} A_{(m)}\mathbf{v}_{(m)}\{1 + o_P(1)\} \geq \frac{\underline{c}_{A(m)}}{4} a_n^2 n \left\|\mathbf{v}_{(m)}\right\|^2.
\end{aligned} \tag{A.4}$$

Noting that $E\left(Z_{n(m),3}|\mathbf{X}\right)=0$ and by Assumption A.1

$$\begin{aligned}\mathrm{Var}\left(Z_{n(m),3}|\mathbf{X}\right) &\leq \sum_{i=1}^{n} E\left\{\left[\int_0^{a_n\mathbf{x}'_{i(m)}\mathbf{v}_{(m)}} \alpha_{i(m)}(s)\,ds\right]^2 |\mathbf{x}_i\right\} \leq \sum_{i=1}^{n}\left[a_n\mathbf{x}'_{i(m)}\mathbf{v}_{(m)}\right]^2 \\ &= na_n^2\mathbf{v}'_{(m)}E\left[\mathbf{x}_{i(m)}\mathbf{x}'_{i(m)}\right]\mathbf{v}_{(m)}\left\{1+o_P(1)\right\} \leq \frac{\bar{c}_{A(m)}}{c_f}na_n^2\left\|\mathbf{v}_{(m)}\right\|^2\left\{1+o_P(1)\right\},\end{aligned}$$

we have

$$Z_{n(m),3}=\bar{c}_{A(m)}^{1/2}O_P\left(a_n\sqrt{n}\right)\left\|\mathbf{v}_{(m)}\right\|=O_P\left(\bar{c}_{A(m)}^{1/2}a_n^2n/\sqrt{k_m}\right)\left\|\mathbf{v}_{(m)}\right\|. \tag{A.5}$$

Observe that $\left(\bar{c}_{A(m)}+\bar{c}_{B(m)}\right)/k_m=O(\underline{c}_{A(m)}^2)$ under Assumption A.2(iv). By (A.3)-(A.5) and allowing $\left\|\mathbf{v}_{(m)}\right\|$ to be sufficiently large, both $Z_{n(m),1}$ and $Z_{n(m),3}$ are dominated by $Z_{n(m),2}$, which is positive w.p.a.1. This, in conjunction with (A.1), implies that $Z_{n(m)}\left(\mathbf{v}_{(m)}\right)>0$ w.p.a.1. This proves (i).

(ii) Let $\hat{\Delta}_{(m)}\equiv\sqrt{n}(\hat{\Theta}_{(m)}-\Theta^*_{(m)})$ and $\Delta_{(m)}\equiv\sqrt{n}(\boldsymbol{\Theta}_{(m)}-\Theta^*_{(m)})$. It follows that $\hat{\Delta}_{(m)}=\underset{\Delta_{(m)}}{\arg\min}\sum_{i=1}^n \rho_\tau\left(y_i-\left[\Theta^*_{(m)}+n^{-1/2}\Delta_{(m)}\right]'\mathbf{x}_{i(m)}\right)$. Let $V_{(m)}(\Delta)\equiv n^{-1/2}\sum_{i=1}^n\psi_\tau\left(y_i-\left[\Theta^*_{(m)}+n^{-1/2}\Delta\right]'\mathbf{x}_{i(m)}\right)\mathbf{x}_{i(m)}$ and $\bar{V}_{(m)}(\Delta)\equiv E\left[V_{(m)}(\Delta)\right]$. Define the weighted norm $\|\cdot\|_{c_{(m)}}$ by

$$\|A\|_{c_{(m)}}=\left\|c_{(m)}A\right\|$$

where $c_{(m)}$ is an arbitrary $l_m\times k_m$ matrix with $\left\|c_{(m)}\right\|\leq\underline{c}_{B(m)}^{-1/2}L_c$ for a large constant $L_c<\infty$. We want to show that for any large constant $L<\infty$,

$$\sup_{\|\Delta\|\leq\sqrt{k_m}L}\left\|V_{(m)}(\Delta)-V_{(m)}(0)-\bar{V}_{(m)}(\Delta)+\bar{V}_{(m)}(0)\right\|_{c_{(m)}} = o_P(1), \tag{A.6}$$

$$\sup_{\|\Delta\|\leq\sqrt{k_m}L}\left\|\bar{V}_{(m)}(\Delta)-\bar{V}_{(m)}(0)+A_{(m)}\Delta\right\|_{c_{(m)}} = o_P(1), \tag{A.7}$$

$$||V_{(m)}(\hat{\Delta}_{(m)})||_{c_{(m)}} = o_P(1). \tag{A.8}$$

(A.6)-(A.7) and the result in part (i) imply that $\left\|V_{(m)}(\hat{\Delta}_{(m)})-V_{(m)}(0)+A_{(m)}\hat{\Delta}_{(m)}\right\|_{c_{(m)}}=o_P(1)$ and consequently we have by Assumptions A.2(ii)-(iii), $\hat{\Delta}_{(m)}=\sqrt{n}\left[\hat{\Theta}_{(m)}-\Theta^*_{(m)}\right]=A_{(m)}^{-1}V_{(m)}(0)$ $+A_{(m)}^{-1}V_{(m)}(\hat{\Delta}_{(m)})+A_{(m)}^{-1}R_{(m)}$, and

$$\begin{aligned}C_{(m)}V_{(m)}^{-1/2}\hat{\Delta}_{(m)} &= \sqrt{n}C_{(m)}V_{(m)}^{-1/2}\left[\hat{\Theta}_{(m)}-\Theta^*_{(m)}\right] \\ &= C_{(m)}V_{(m)}^{-1/2}A_{(m)}^{-1}V_{(m)}(0)+C_{(m)}V_{(m)}^{-1/2}A_{(m)}^{-1}V_{(m)}(\hat{\Delta}_{(m)})+C_{(m)}V_{(m)}^{-1/2}A_{(m)}^{-1}R_{(m)} \\ &\equiv T_{1(m)}+T_{2(m)}+T_{3(m)},\ \text{say},\end{aligned}$$

where $\left\|R_{(m)}\right\|_{c_{(m)}}=o_P(1)$ for any $c_{(m)}$ with $\left\|c_{(m)}\right\|\leq\underline{c}_{B(m)}^{-1/2}L_c$, $V_{(m)}^{1/2}$ denotes the symmetric square root of $V_{(m)}$ and $V_{(m)}^{-1/2}$ the inverse of $V_{(m)}^{1/2}$.

Let $\eta_{ni}\equiv n^{-1/2}C_{(m)}V_{(m)}^{-1/2}A_{(m)}^{-1}\psi_\tau\left(\varepsilon_i+u_{i(m)}\right)\mathbf{x}_{i(m)}$. Then $T_{1(m)}=\sum_{i=1}^n\eta_{ni}$. Noting that $E(\eta_{ni})=0$ by (A.2), we have

$$\begin{aligned}\mathrm{Var}\left(T_{1(m)}\right) &= \sum_{i=1}^{n}\mathrm{Var}(\eta_{ni})=n^{-1}\sum_{i=1}^{n}C_{(m)}V_{(m)}^{-1/2}A_{(m)}^{-1}E\left[\psi_\tau\left(\varepsilon_i+u_{i(m)}\right)^2\mathbf{x}_{i(m)}\mathbf{x}'_{i(m)}\right]A_{(m)}^{-1}V_{(m)}^{-1/2}C'_{(m)} \\ &= C_{(m)}V_{(m)}^{-1/2}A_{(m)}^{-1}B_{(m)}A_{(m)}^{-1}V_{(m)}^{-1/2}C'_{(m)}=C_{(m)}C'_{(m)}.\end{aligned}$$

By the fact that $\mathrm{tr}(AB) \leq \lambda_{\max}(A)\mathrm{tr}(B)$ for symmetric matrix A and p.s.d. matrix B (e.g., Bernstein (2005, Proposition 8.4.13)),

$$\begin{aligned}
E\|\eta_{ni}\|^4 &= n^{-2}E\left\{\left[\mathrm{tr}\left(C_{(m)}V_{(m)}^{-1/2}A_{(m)}^{-1}\psi_\tau\left(\varepsilon_i+u_{i(m)}\right)^2\mathbf{x}_{i(m)}\mathbf{x}_{i(m)}'A_{(m)}^{-1}V_{(m)}^{-1/2}C_{(m)}'\right)\right]^2\right\} \\
&\leq n^{-2}E\left\{\left[\mathrm{tr}\left(\mathbf{x}_{i(m)}\mathbf{x}_{i(m)}'A_{(m)}^{-1}V_{(m)}^{-1/2}C_{(m)}'C_{(m)}V_{(m)}^{-1/2}A_{(m)}^{-1}\right)\right]^2\right\} \\
&\leq n^{-2}E\left\{\left[\mathrm{tr}\left(\mathbf{x}_{i(m)}\mathbf{x}_{i(m)}'\right)\right]^2\left[\lambda_{\max}\left(A_{(m)}^{-1}V_{(m)}^{-1/2}C_{(m)}'C_{(m)}V_{(m)}^{-1/2}A_{(m)}^{-1}\right)\right]^2\right\} \\
&\leq n^{-2}E\left\|\mathbf{x}_{i(m)}\right\|^4\left[\lambda_{\max}\left(C_{(m)}'C_{(m)}\right)\right]^2\left[\lambda_{\max}\left(A_{(m)}^{-1}V_{(m)}^{-1}A_{(m)}^{-1}\right)\right]^2 \\
&= n^{-2}E\left\|\mathbf{x}_{i(m)}\right\|^4\left[\lambda_{\max}\left(C_{(m)}C_{(m)}'\right)\right]^2\left[\lambda_{\max}\left(B_{(m)}^{-1}\right)\right]^2 \\
&= O\left(n^{-2}k_m^2\underline{c}_{B(m)}^{-2}\right). \qquad \text{(A.9)}
\end{aligned}$$

Then for any $\epsilon > 0$, $\sum_{i=1}^n E\left[\|\eta_{ni}\|^2\mathbf{1}\{\|\eta_{ni}\| \geq \epsilon\}\right] = nE\left[\|\eta_{ni}\|^2\mathbf{1}\{\|\eta_{ni}\| \geq \epsilon\}\right] \leq n\left\{E\|\eta_{ni}\|^4\right\}^{1/2}$ $\times\{P(\|\eta_{ni}\| \geq \epsilon)\}^{1/2} \leq n\epsilon^{-2}E\|\eta_{ni}\|^4 = O(n^{-1}k_m^2\underline{c}_{B(m)}^{-2}) = o(1)$ by Assumption A.3(i). Thus $\{\eta_{ni}\}$ satisfies the conditions of the Lindeberg-Feller central limit theorem and we have

$$T_{1(m)} \xrightarrow{d} N(0, C_0). \qquad \text{(A.10)}$$

For $T_{2(m)}$ and $T_{3(m)}$, we take $c_{(m)} = C_{(m)}V_{(m)}^{-1/2}A_{(m)}^{-1}$. By the fact that $\mathrm{tr}(AB) \leq \lambda_{\max}(A)\mathrm{tr}(B)$ for symmetric matrix A and p.s.d. matrix B and that $\lambda_{\max}(A'A) = \lambda_{\max}(AA')$ for any matrix A, we have

$$\begin{aligned}
\left\|c_{(m)}\right\| &= \left\{\mathrm{tr}\left(V_{(m)}^{-1/2}A_{(m)}^{-1}A_{(m)}^{-1}V_{(m)}^{-1/2}C_{(m)}'C_{(m)}\right)\right\}^{1/2} \\
&\leq \left\|C_{(m)}\right\|\left\{\lambda_{\max}\left(V_{(m)}^{-1/2}A_{(m)}^{-1}A_{(m)}^{-1}V_{(m)}^{-1/2}\right)\right\}^{1/2} = \left\|C_{(m)}\right\|\left\{\lambda_{\max}\left(A_{(m)}^{-1}V_{(m)}^{-1}A_{(m)}^{-1}\right)\right\}^{1/2} \\
&= \left\|C_{(m)}\right\|\left\{\lambda_{\max}\left(B_{(m)}^{-1}\right)\right\}^{1/2} \leq \left\|C_{(m)}\right\|\underline{c}_{B(m)}^{-1/2} \leq L_c\underline{c}_{B(m)}^{-1/2}
\end{aligned}$$

for sufficiently large L_c. Then by (A.8)

$$\left\|T_{2(m)}\right\| = \left\|V_{(m)}(\hat{\Delta}_{(m)})\right\|_{C_{(m)}V_{(m)}^{-1/2}A_{(m)}^{-1}} = o_P(1) \qquad \text{(A.11)}$$

and

$$\left\|T_{3(m)}\right\| = \left\|R_{(m)}\right\|_{C_{(m)}V_{(m)}^{-1/2}A_{(m)}^{-1}} = o_P(1). \qquad \text{(A.12)}$$

Combining (A.10)-(A.12) yields the claim in part (ii).

Below we demonstrate (A.6)-(A.8) hold under Assumptions A.1-A.3. Since l_m is fixed, without loss of generality we assume that $l_m = 1$. First, we show (A.6). Write $a_{i(m)} \equiv c_{(m)}\mathbf{x}_{i(m)} = a_{i(m)}^+ - a_{i(m)}^-$ where $a_{i(m)}^+ = \max\{a_{i(m)}, 0\}$ and $a_{i(m)}^- = \max\{-a_{i(m)}, 0\}$. Then by Minkowski's inequality we have

$$\begin{aligned}
&\sup_{\|\Delta\| \leq \sqrt{k_m}L}\left\|V_{(m)}(\Delta) - V_{(m)}(0) - \bar{V}_{(m)}(\Delta) + \bar{V}_{(m)}(0)\right\|_{c_{(m)}} \\
\leq\ &\sup_{\|\Delta\| \leq \sqrt{k_m}L}\left|V_{(m)}^+(\Delta) - V_{(m)}^+(0) - \bar{V}_{(m)}^+(\Delta) + \bar{V}_{(m)}^+(0)\right| \\
&+ \sup_{\|\Delta\| \leq \sqrt{k_m}L}\left|V_{(m)}^-(\Delta) - V_{(m)}^-(0) - \bar{V}_{(m)}^-(\Delta) + \bar{V}_{(m)}^-(0)\right| \qquad \text{(A.13)}
\end{aligned}$$

where $V^+_{(m)}(\Delta) \equiv n^{-1/2}\sum_{i=1}^n \psi_\tau(y_i - [\Theta^*_{(m)} + n^{-1/2}\Delta]'\mathbf{x}_{i(m)})a^+_{i(m)}$, $\bar{V}^+_{(m)}(\Delta) \equiv E[V^+_{(m)}(\Delta)]$, and $V^-_{(m)}(\Delta)$ and $\bar{V}^-_{(m)}(\Delta)$ are analogously defined. It suffices to show that each term on the right hand side of (A.13) is $o_P(1)$. We only show the first term is $o_P(1)$ as the second term can be treated analogously.

Let $\mathbf{D} \equiv \{\Delta \in \mathbb{R}^{k_m} : \|\Delta\| \leq \sqrt{k_m}L\}$ for some $L < \infty$. Let $|t|_\infty$ denote the maximum of the absolute values of the coordinates of t. By selecting $N_1 = \left(2n^2\right)^{k_m}$ grid points, $\Delta_1, \cdots, \Delta_{N_1}$, we can cover $\mathbf{D}$ by cubes $\mathbf{D}_s = \left\{\Delta \in \mathbb{R}^{k_m} : |\Delta - \Delta_s|_\infty \leq \delta_n\right\}$ with sides of length δ_n where $\delta_n = Lk_m^{1/2}/n^2$. Let $\varepsilon_{i(m)} \equiv y_i - \mathbf{x}'_{i(m)}\Theta^*_{(m)}$. In view of the fact $\psi_\tau(\cdot)$ is monotone and by Minkowski's inequality, we can readily show that

$$
\begin{aligned}
&\sup_{\|\Delta\| \leq \sqrt{k_m}L} \left|V^+_{(m)}(\Delta) - V^+_{(m)}(0) - \bar{V}^+_{(m)}(\Delta) + \bar{V}^+_{(m)}(0)\right| \\
\leq\ & \max_{1\leq s\leq N_1} \left|V^+_{(m)}(\Delta_s) - V^+_{(m)}(0) - \bar{V}^+_{(m)}(\Delta_s) + \bar{V}^+_{(m)}(0)\right| \\
&+ \max_{1\leq s\leq N_1} \left|n^{-1/2}\sum_{i=1}^n E\left[\psi_{si(m)}(\delta_n)\, a^+_{i(m)}\right] - E\left[\psi_{si(m)}(-\delta_n)\, a^+_{i(m)}\right]\right| \\
&+ \max_{1\leq s\leq N_1} \left|n^{-1/2}\sum_{i=1}^n \left[\left[\psi_{si(m)}(\delta_n) - \psi_{si(m)}(0)\right] a^+_{i(m)} - E\left\{\left[\psi_{si(m)}(\delta_n) - \psi_{si(m)}(0)\right] a^+_{i(m)}\right\}\right]\right| \\
\equiv\ & I_{1(m)} + I_{2(m)} + I_{3(m)}, \text{ say,}
\end{aligned}
$$

where $\psi_{si(m)}(\delta) = \psi_\tau\left(\varepsilon_{i(m)} - n^{-1/2}\Delta'_s\mathbf{x}_{i(m)} + n^{-1/2}\delta\left\|\mathbf{x}_{i(m)}\right\|\right)$. For $I_{2(m)}$, we apply Taylor expansion, Assumption A.2(i), and the fact that $a^+_{i(m)} \leq \left|c_m\mathbf{x}_{i(m)}\right| \leq \|c_m\|\left\|\mathbf{x}_{i(m)}\right\|$ to obtain

$$
\begin{aligned}
I_{2(m)} &= \max_{1\leq s\leq N_1} \Big|n^{1/2}E\left\{F\left(-u_{i(m)} + n^{-1/2}\Delta'_s\mathbf{x}_{i(m)} + n^{-1/2}\delta_n||\mathbf{x}_{i(m)}||\ |\mathbf{x}_i\right) a^+_{i(m)}\right\} \\
&\quad - E\left\{F\left(-u_{i(m)} + n^{-1/2}\Delta'_s\mathbf{x}_{i(m)} - n^{-1/2}\delta_n||\mathbf{x}_{i(m)}||\ |\mathbf{x}_i\right) a^+_{i(m)}\right\}\Big| \\
&\leq 2c_f\delta_n E\left\{\left\|\mathbf{x}_{i(m)}\right\| a^+_{i(m)}\right\} \leq 2c_f\delta_n\|c_m\| E\left\|\mathbf{x}_{i(m)}\right\|^2 \\
&= O\left(\delta_n\underline{c}^{-1/2}_{B(m)}k_m\right) = O(\underline{c}^{-1/2}_{B(m)}k_m^{3/2}/n^2) = o(1).
\end{aligned}
$$

For $I_{1(m)}$, note that

$$
\begin{aligned}
&V^+_{(m)}(\Delta_s) - V^+_{(m)}(0) - \bar{V}^+_{(m)}(\Delta_s) + \bar{V}^+_{(m)}(0) \\
=\ & n^{-1}\sum_{i=1}^n \eta_{is(m)}\mathbf{1}\left\{a^+_{i(m)} \leq e_{1n}\right\} + n^{-1}\sum_{i=1}^n \eta_{is(m)}\mathbf{1}\left\{a^+_{i(m)} > e_{1n}\right\} \equiv D_{1s} + D_{2s}, \text{ say,}
\end{aligned}
$$

where $\eta_{is(m)} \equiv n^{1/2}[\eta_{is(m),0} - E(\eta_{is(m),0})]$, $\eta_{is(m),0} = [\psi_\tau\left(\varepsilon_{i(m)} - n^{-1/2}\Delta'_s\mathbf{x}_{i(m)}\right) - \psi_\tau\left(\varepsilon_{i(m)}\right)]a^+_{i(m)}$, and $e_{1n} = (nk_m^4\underline{c}^{-4}_{B(m)})^{1/8}$. It suffices to prove $I_{1(m)} = o_P(1)$ by showing that

$$
\max_{1\leq s\leq N_1} \|D_{ls}\| = o_P(1) \text{ for } l = 1 \text{ and } 2. \tag{A.14}
$$

Note that $\mathrm{Var}\left[\eta_{is(m)}\mathbf{1}\{a^+_{i(m)} \leq e_{1n}/n^{1/2}\}\right] \leq nE[\left|\psi_\tau\left(\varepsilon_{i(m)} - n^{-1/2}\Delta'_s\mathbf{x}_{i(m)}\right) - \psi_\tau\left(\varepsilon_{i(m)}\right)\right|\ (a^+_{i(m)})^2] \leq C_1\underline{c}^{-1}_{B(m)}n^{1/2}k_m$ for some $C_1 < \infty$. By Boole's and Bernstein's inequalities (e.g., Serfling (1980, p.95)),

we have

$$\begin{aligned} P\left(\max_{1\leq s\leq N_1}\|D_{1s}\|\geq\epsilon\right) &\leq N_1\max_{1\leq s\leq N_1}P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\eta_{is(m)}\mathbf{1}\{a_{i(m)}^{+}\leq e_{1n}\}\right\|\geq\epsilon\right)\\ &\leq 2N_1\exp\left(-\frac{n\epsilon^2}{2C_1\underline{c}_{B(m)}^{-1}n^{1/2}k_m+4\epsilon n^{1/2}e_{1n}/3}\right)\\ &\leq 2\exp\left(3k_m\log n\right)\times\exp\left(-4k_m\log n\right)=2\exp\left(-k_m\log n\right)=o\left(1\right), \end{aligned}$$

because $n/(\underline{c}_{B(m)}^{-1}n^{1/2}k_m)=n^{1/2}\underline{c}_{B(m)}/k_m\gg k_m\log n$ and $n/(n^{1/2}e_{1n})=n^{3/8}k_m^{-1/2}\underline{c}_{B(m)}^{1/2}\gg k_m\log n$ by Assumption A.3(i). Let $\bar{a}_{i(m)}=a_{i(m)}k_m^{-1/2}\underline{c}_{B(m)}^{1/2}$. Noting that $E[\left|c_{(m)}\mathbf{x}_{i(m)}\right|^8]=O(k_m^4\underline{c}_{B(m)}^{-4})$ by arguments as used to obtain (A.9), $E\left|\bar{a}_{i(m)}\right|^8=O\left(1\right)$. By Boole's and Markov's inequalities, Assumption A.1(iii), the fact that $nk_m^4\underline{c}_{B(m)}^{-4}/e_{1n}^8=O(1)$ by construction, and Lebesgue dominated convergence theorem

$$\begin{aligned} P\left(\max_{1\leq s\leq N_1}\|D_{2s}\|\geq\epsilon\right) &\leq P\left(\max_{1\leq i\leq n}a_{i(m)}^{+}>e_{1n}\right)\leq nP\left(\left|\bar{a}_{i(m)}\right|>k_m^{-1/2}\underline{c}_{B(m)}^{1/2}e_{1n}\right)\\ &\leq \frac{nk_m^4\underline{c}_{B(m)}^{-4}}{e_{1n}^8}E\left[\left|\bar{a}_{i(m)}\right|^8\mathbf{1}\left\{\left|\bar{a}_{i(m)}\right|>k_m^{-1/2}\underline{c}_{B(m)}^{1/2}e_{1n}\right\}\right]=o\left(1\right). \end{aligned}$$

Thus (A.14) follows and we have shown $I_{1(m)}=o_P\left(1\right)$. By the same token, we can show that $I_{3(m)}=o_P\left(1\right)$. Consequently (A.6) follows.

Next, we show (A.7). By Assumptions A.1 and A.2,

$$\begin{aligned} &\sup_{\|\Delta\|\leq\sqrt{k_m}L}\left\|\bar{V}_{(m)}\left(\Delta\right)-\bar{V}_{(m)}\left(0\right)+A_{(m)}\Delta\right\|_{c_{(m)}}\\ =&\sup_{\|\Delta\|\leq\sqrt{k_m}L}\left\|n^{-1/2}\sum_{i=1}^{n}E\left\{\left[F\left(-u_{i(m)}+n^{-1/2}\Delta'\mathbf{x}_{i(m)}|\mathbf{x}_i\right)-F\left(-u_{i(m)}|\mathbf{x}_i\right)\right]\mathbf{x}_{i(m)}\right\}-A_{(m)}\Delta\right\|_{c_{(m)}}\\ =&\sup_{\|\Delta\|\leq\sqrt{k_m}L}\left\|n^{-1}\sum_{i=1}^{n}E\left\{\int_0^1\left[f\left(\left(-u_{i(m)}+sn^{-1/2}\Delta'\mathbf{x}_{i(m)}\right)|\mathbf{x}_i\right)-f\left(-u_{i(m)}|\mathbf{x}_i\right)\right]ds\ \mathbf{x}_{i(m)}\mathbf{x}_{i(m)}'\Delta\right\}\right\|_{c_{(m)}}\\ \leq& C\sup_{\|\Delta\|\leq\sqrt{k_m}L}n^{-3/2}\sum_{i=1}^{n}E\left\|\Delta'\mathbf{x}_{i(m)}\mathbf{x}_{i(m)}\mathbf{x}_{i(m)}'\Delta\right\|_{c_{(m)}}\\ \leq& Cn^{-1/2}L_c\underline{c}_{B(m)}^{-1/2}\sup_{\|\Delta\|\leq\sqrt{k_m}L}\left\{E\left\|\Delta'\mathbf{x}_{i(m)}\right\|^2\right\}^{1/2}\left\{E\left\|\mathbf{x}_{i(m)}\mathbf{x}_{i(m)}'\Delta\right\|^2\right\}^{1/2}\\ =& n^{-1/2}\underline{c}_{B(m)}^{-1/2}O\left(\bar{c}_{A(m)}^{1/2}k_m^{1/2}\right)O\left(k_m^{3/2}\right)=O\left(\underline{c}_{B(m)}^{-1/2}\bar{c}_{A(m)}^{1/2}k_m^2/n^{1/2}\right)=o(1), \end{aligned}$$

where we use the fact that $E\left\|\Delta'\mathbf{x}_{i(m)}\right\|^2\leq\lambda_{\max}(E[\mathbf{x}_{i(m)}\mathbf{x}_{i(m)}'])\left\|\Delta\right\|^2=O\left(\bar{c}_{A(m)}k_m\right)$ by Assumptions A.2(i)-(ii).

Now we show (A.8). By the proof of Lemma A2 in Ruppert and Carroll (1980) (see, Welsh (1989,

p. 360)) and Assumptions A.1-A.3,

$$\begin{aligned}||V_{(m)}(\hat{\Delta}_{(m)})||_{c_{(m)}} &= \left\| n^{-1/2}\sum_{i=1}^{n}\psi_\tau(y_i-\hat{\Theta}'_{(m)}\mathbf{x}_{i(m)})\mathbf{x}_{i(m)}\right\|_{c_{(m)}} \\ &\leq n^{-1/2}\sum_{i=1}^{n}\mathbf{1}\left\{y_i-\hat{\Theta}'_{(m)}\mathbf{x}_{i(m)}=0\right\}\left|c_{(m)}\mathbf{x}_{i(m)}\right| \\ &\leq n^{-1/2}k_m\max_{1\leq i\leq n}\left|c_{(m)}\mathbf{x}_{i(m)}\right| = o_P(1)\end{aligned}$$

because by Boole's and Markov's inequalities $P\left(\max_{1\leq i\leq n}\left|c_{(m)}\mathbf{x}_{i(m)}\right|\geq n^{1/2}/k_m\right)\leq nP(\left|c_{(m)}\mathbf{x}_{i(m)}\right|\geq n^{1/2}/k_m)\leq k_m^8n^{-3}E[\left|c_{(m)}\mathbf{x}_{i(m)}\right|^8]=k_m^8n^{-3}O(k_m^4\underline{c}_{B(m)}^{-4})=O(n^{-3}k_m^{12}\underline{c}_{B(m)}^{-4})=o(1)$ by Assumption A.3(i). This completes the proof of part (ii). ■

Remark. The proof of (A.10) indicates $\sqrt{n}C_{(m)}V_{(m)}^{-1/2}[\hat{\Theta}_{(m)}-\Theta^*_{(m)}]=C_{(m)}V_{(m)}^{-1/2}A_{(m)}^{-1}\,n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_{i(m)}\times\left[\tau-\mathbf{1}\left\{y_i\leq\Theta^{*\prime}_{(m)}\mathbf{x}_{i(m)}\right\}\right]+o_P(1)$, from which we obtain the following Bahadur representation for $\sqrt{n}[\hat{\Theta}_{(m)}-\Theta^*_{(m)}]$:

$$\begin{aligned}\sqrt{n}[\hat{\Theta}_{(m)}-\Theta^*_{(m)}] &= \left[C_{(m)}V_{(m)}^{-1/2}\right]^+C_{(m)}V_{(m)}^{-1/2}A_{(m)}^{-1}\,n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_{i(m)}\left[\tau-\mathbf{1}\left\{y_i\leq\Theta^{*\prime}_{(m)}\mathbf{x}_{i(m)}\right\}\right]+\text{s.m.} \\ &= P_{V_{(m)}^{-1/2}C'_{(m)}}A_{(m)}^{-1}\,n^{-1/2}\sum_{i=1}^{n}\mathbf{x}_{i(m)}\left[\tau-\mathbf{1}\left\{y_i\leq\Theta^{*\prime}_{(m)}\mathbf{x}_{i(m)}\right\}\right]+\text{s.m.}, \qquad \text{(A.15)}\end{aligned}$$

where s.m. denotes smaller order terms and $P_A=A\left(A'A\right)^{-1}A'$.

B Proof of Theorem 3.2

We only prove (i) as the proof of (ii) is analogous. Let $\delta_n\equiv L\sqrt{n^{-1}\bar{k}\log n}$ for some large constant $L<\infty$. Let $\bar{Q}_{(m)}\left(\Theta_{(m)}\right)\equiv E[\rho_\tau(y_i-\mathbf{x}'_{i(m)}\Theta_{(m)})]$. Define

$$D\left(\delta_n\right)\equiv\inf_{1\leq m\leq M}\inf_{\left\|\Theta_{(m)}-\Theta^*_{(m)}\right\|>\delta_n}\left[\bar{Q}_{(m)}\left(\Theta_{(m)}\right)-\bar{Q}_{(m)}(\Theta^*_{(m)})\right], \qquad \text{(B.1)}$$

and $\mathcal{S}_m\left(\delta_n\right)\equiv\{\Theta_{(m)}:||\Theta_{(m)}-\Theta^*_{(m)}||>\delta_n,\,||\Theta_{(m)}-\Theta^*_{(m)}||=o\left(1\right)\}$. By Knight's identity, the definition of $u_{i(m)}(\equiv\mu_i-\mathbf{x}'_{i(m)}\Theta^*_{(m)})$ and Assumption A.2(ii) for any $\Theta_{(m)}\in\mathcal{S}_m\left(\delta_n\right)$ we have

$$\begin{aligned}\bar{Q}_{(m)}\left(\Theta_{(m)}\right)-\bar{Q}_{(m)}\left(\Theta^*_{(m)}\right) &= E\left[\rho_\tau\left(y_i-\mathbf{x}'_{i(m)}\Theta_{(m)}\right)-\rho_\tau\left(y_i-\mathbf{x}'_{i(m)}\Theta^*_{(m)}\right)\right] \\ &= E\left[\rho_\tau\left(\varepsilon_i+u_{i(m)}-\mathbf{x}'_{i(m)}\left[\Theta_{(m)}-\Theta^*_{(m)}\right]\right)-\rho_\tau\left(\varepsilon_i+u_{i(m)}\right)\right] \\ &= E\left\{\int_{\mathbf{0}}^{\mathbf{x}'_{i(m)}\left[\Theta_{(m)}-\Theta^*_{(m)}\right]}\left[\mathbf{1}\left\{\varepsilon_i+u_{i(m)}\leq s\right\}-\mathbf{1}\left\{\varepsilon_i+u_{i(m)}\leq 0\right\}\right]ds\right\} \\ &= E\left\{\int_{0}^{\mathbf{x}'_{i(m)}\left[\Theta_{(m)}-\Theta^*_{(m)}\right]}\left[F\left(-u_{i(m)}+s|\mathbf{x}_i\right)-F\left(-u_{i(m)}|\mathbf{x}_i\right)\right]ds\right\} \\ &\simeq \frac{1}{2}\left[\Theta_{(m)}-\Theta^*_{(m)}\right]'A_{(m)}\left[\Theta_{(m)}-\Theta^*_{(m)}\right]\geq\frac{\underline{c}_A\delta_n^2}{2}.\end{aligned}$$

Then by Boole's inequality, (B.1), and the fact that $\bar{Q}_{(m)}(\hat{\Theta}_{i(m)}) - \bar{Q}_{(m)}(\Theta^*_{(m)}) \simeq \frac{1}{2}[\hat{\Theta}_{i(m)} - \Theta^*_{(m)}]' A_{(m)} [\hat{\Theta}_{i(m)} - \Theta^*_{(m)}]$, we have

$$\begin{aligned} P\left(\max_{1\le i\le n}\max_{1\le m\le M}\left\|\hat{\Theta}_{i(m)} - \Theta^*_{(m)}\right\| \ge \delta_n\right) &\le nM \max_{1\le i\le n}\max_{1\le m\le M} P\left(\left\|\hat{\Theta}_{i(m)} - \Theta^*_{(m)}\right\| \ge \delta_n\right) \\ &\le nM \max_{1\le i\le n}\max_{1\le m\le M} P\left\{\bar{Q}_{(m)}\left(\hat{\Theta}_{i(m)}\right) - \bar{Q}_{(m)}\left(\Theta^*_{(m)}\right) \ge D\left(\delta_n\right)\right\} \\ &\approx nM \max_{1\le i\le n}\max_{1\le m\le M} P\left\{\mathbb{W}_{i(m)} \ge 2nD\left(\delta_n\right)\right\}, \end{aligned} \tag{B.2}$$

where $\mathbb{W}_{i(m)} \equiv n[\hat{\Theta}_{i(m)} - \Theta^*_{(m)}]' A_{(m)} [\hat{\Theta}_{i(m)} - \Theta^*_{(m)}]$. The key is to bound $P\left\{\mathbb{W}_{i(m)} \ge 2nD\left(\delta_n\right)\right\}$.

Following the proof of Theorem 3.1(ii), we can also show that $n^{1/2} C_{(m)} V_{(m)}^{-1/2} [\hat{\Theta}_{i(m)} - \Theta^*_{(m)}] \xrightarrow{d} N\left(0, C_0\right)$, which implies that

$$\tilde{\beta}_{i(m)} \equiv n^{1/2}\left[C_{(m)} C'_{(m)}\right]^{-1/2} C_{(m)} V_{(m)}^{-1/2} [\hat{\Theta}_{i(m)} - \Theta^*_{(m)}] \xrightarrow{d} N\left(0, I_{l_m}\right). \tag{B.3}$$

Rewriting $n^{1/2}[\hat{\Theta}_{i(m)} - \Theta^*_{(m)}]$ in terms of $\tilde{\beta}_{i(m)}$ yields $n^{1/2}[\hat{\Theta}_{i(m)} - \Theta^*_{(m)}] = \{[C_{(m)} C'_{(m)}]^{-1/2} C_{(m)} V_{(m)}^{-1/2}\}^{+} \tilde{\beta}_{i(m)}$. It follows that

$$\begin{aligned} \mathbb{W}_{i(m)} &= n\left[\hat{\Theta}_{i(m)} - \Theta^*_{(m)}\right]' A_{(m)} \left[\hat{\Theta}_{i(m)} - \Theta^*_{(m)}\right] \\ &= \tilde{\beta}'_{i(m)} \left[\left\{\left[C_{(m)} C'_{(m)}\right]^{-1/2} C_{(m)} V_{(m)}^{-1/2}\right\}^{+}\right]' A_{(m)} \left\{\left[C_{(m)} C'_{(m)}\right]^{-1/2} C_{(m)} V_{(m)}^{-1/2}\right\}^{+} \tilde{\beta}_{i(m)} \\ &\le \lambda_{\max}\left(A_{(m)}\right) \tilde{\beta}'_{i(m)} \left[\left\{\left[C_{(m)} C'_{(m)}\right]^{-1/2} C_{(m)} V_{(m)}^{-1/2}\right\}^{+}\right]' \left\{\left[C_{(m)} C'_{(m)}\right]^{-1/2} C_{(m)} V_{(m)}^{-1/2}\right\}^{+} \tilde{\beta}_{i(m)} \\ &= \lambda_{\max}\left(A_{(m)}\right) \tilde{\beta}'_{i(m)} \left\{\left[C_{(m)} C'_{(m)}\right]^{-1/2} C_{(m)} V_{(m)}^{-1} C'_{(m)} \left[C_{(m)} C'_{(m)}\right]^{-1/2}\right\}^{-1} \tilde{\beta}_{i(m)} \\ &= \lambda_{\max}\left(A_{(m)}\right) \tilde{\beta}'_{i(m)} \left\{\left[C_{(m)} C'_{(m)}\right]^{1/2} \left(C_{(m)} V_{(m)}^{-1} C'_{(m)}\right)^{-1} \left[C_{(m)} C'_{(m)}\right]^{1/2}\right\} \tilde{\beta}_{i(m)} \\ &\le \lambda_{\max}\left(A_{(m)}\right) \lambda_{\max}\left(V_{(m)}\right) \tilde{\beta}'_{i(m)} \tilde{\beta}_{i(m)} \le \left(\bar{c}_A \bar{c}_B / \underline{c}_A^2\right) \left\|\tilde{\beta}_{i(m)}\right\|^2 \text{ by Assumptions A.2 } (ii)\text{-}(iii) \end{aligned}$$

where we have used the fact that $A'BA \le \lambda_{\max}(B) A'A$ for any real symmetric matrix B and conformable matrix A and that $A^{+\prime} A^{+} = (AA')^{+}$ (see, e.g., Bernstein (2005, Proposition 6.1.6xvii)). Let $c_{AB} \equiv \bar{c}_A \bar{c}_B / \underline{c}_A^2$ and $\bar{l} = \max_{1\le m\le M} l_m$. Then by Lemma 2.1 of Shibata (1981)

$$\begin{aligned} & nM \max_{1\le i\le n}\max_{1\le m\le M} P\left\{\mathbb{W}_{i(m)} \ge 2nD\left(\delta_n\right)\right\} \\ \le\ & nM \max_{1\le i\le n}\max_{1\le m\le M} P\left\{\left\|\tilde{\beta}_{i(m)}\right\|^2 \ge 2nD\left(\delta_n\right)/c_{AB}\right\} \\ \le\ & \limsup_{n\to\infty} nM \max_{1\le m\le M} P\left\{\chi^2\left(l_m\right) \ge 2nD\left(\delta_n\right)/c_{AB}\right\} \\ \le\ & \limsup_{n\to\infty} nM\ P\left\{\chi^2\left(\bar{l}\right) \ge 2nD\left(\delta_n\right)/c_{AB}\right\} \\ \le\ & \limsup_{n\to\infty} nM\ P\left\{\chi^2\left(\bar{l}\right) \ge \bar{l} + [n\delta_n^2 \underline{c}_A / c_{AB} - \bar{l}]\right\} \\ =\ & \limsup_{n\to\infty} nM\ \exp\left(-\frac{[n\delta_n^2 \underline{c}_A / c_{AB} - \bar{l}]}{2}\left\{1 - \log\left(n\delta_n^2 \underline{c}_A / \bar{l} c_{AB}\right) / [n\delta_n^2 \underline{c}_A / \bar{l} c_{AB} - 1]\right\}\right) = 0 \end{aligned} \tag{B.4}$$

because $nM\exp\left(-0.5n\delta_n^2\underline{c}_A/c_{AB}\right)=nM\,n^{-0.5L^2\bar{k}\underline{c}_A^3/(\bar{c}_A\bar{c}_B)}=o(1)$ for sufficiently large L and $\log\left(n\delta_n^2\underline{c}_A/\bar{l}c_{AB}\right)/[n\delta_n^2\underline{c}_A/\bar{l}c_{AB}-1]=o(1)$ under our assumptions. Combining (B.2)-(B.4), we have shown that $\limsup_{n\to\infty} P\left(\max_{1\leq i\leq n}\max_{1\leq m\leq M}\left\|\hat{\Theta}_{i(m)}-\Theta^*_{(m)}\right\|\geq\delta_n\right)=0$ and thus (i) follows. ■

C Proof of Theorem 3.3

Following the proof of Theorem 2.1 in Li (1987), if we can show that the difference $CV_n(\mathbf{w})-FPE_n(\mathbf{w})$ is negligible compared with $FPE_n(\mathbf{w})$ uniformly for any $\mathbf{w}\in\mathcal{W}$, then the optimality property (3.5) is established for $\hat{\mathbf{w}}$. More precisely, it suffices to show that

$$\sup_{\mathbf{w}\in\mathcal{W}}\left|\frac{CV_n(\mathbf{w})-FPE_n(\mathbf{w})}{FPE_n(\mathbf{w})}\right|=o_P(1). \tag{C.1}$$

Let $E_{\mathbf{x}_i}(\cdot)$ denote expectation with respect to $\mathbf{x}_i$. By Knight's identity and the fact that

$$E\left\{\int_0^{\sum_{m=1}^M w_m\mathbf{x}'_{(m)}\hat{\Theta}_{(m)}-\mu}[F(s|\mathbf{x})-F(0|\mathbf{x})]ds|\mathcal{D}_n\right\}=E_{\mathbf{x}_i}\left\{\int_0^{\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\hat{\Theta}_{(m)}-\mu_i}[F(s|\mathbf{x}_i)-F(0|\mathbf{x}_i)]ds\right\},$$

we have

$$\begin{aligned}
& CV_n(\mathbf{w})-FPE_n(\mathbf{w})\\
= & \left\{\frac{1}{n}\sum_{i=1}^n\left[\rho_\tau\left(y_i-\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\hat{\Theta}_{i(m)}\right)-\rho_\tau(\varepsilon_i)\right]\right\}-\{FPE_n(\mathbf{w})-E[\rho_\tau(\varepsilon)]\}\\
& +\frac{1}{n}\sum_{i=1}^n\{\rho_\tau(\varepsilon_i)-E[\rho_\tau(\varepsilon)]\}\\
= & \frac{1}{n}\sum_{i=1}^n\left[\mu_i-\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\hat{\Theta}_{i(m)}\right]\psi_\tau(\varepsilon_i)+\frac{1}{n}\sum_{i=1}^n\int_0^{\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\hat{\Theta}_{i(m)}-\mu_i}[\mathbf{1}\{\varepsilon_i\leq s\}-\mathbf{1}\{\varepsilon_i\leq 0\}]\,ds\\
& -E\left[\int_0^{\sum_{m=1}^M w_m\mathbf{x}'_{(m)}\hat{\Theta}_{(m)}-\mu}[\mathbf{1}\{\varepsilon\leq s\}-\mathbf{1}\{\varepsilon\leq 0\}]\,ds\mid\mathcal{D}_n\right]+\frac{1}{n}\sum_{i=1}^n\{\rho_\tau(\varepsilon_i)-E[\rho_\tau(\varepsilon_i)]\}\\
= & CV_{1n}(\mathbf{w})+CV_{2n}(\mathbf{w})+CV_{3n}(\mathbf{w})+CV_{4n}(\mathbf{w})+CV_{5n},
\end{aligned}$$

where

$$\begin{aligned}
CV_{1n}(\mathbf{w}) &\equiv \frac{1}{n}\sum_{i=1}^n\left[\mu_i-\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\hat{\Theta}_{i(m)}\right]\psi_\tau(\varepsilon_i),\\
CV_{2n}(\mathbf{w}) &\equiv \frac{1}{n}\sum_{i=1}^n\int_0^{\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\hat{\Theta}_{i(m)}-\mu_i}[\mathbf{1}\{\varepsilon_i\leq s\}-\mathbf{1}\{\varepsilon_i\leq 0\}-F(s|\mathbf{x}_i)+F(0|\mathbf{x}_i)]\,ds,\\
CV_{3n}(\mathbf{w}) &\equiv \frac{1}{n}\sum_{i=1}^n\left\{\int_0^{\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\hat{\Theta}_{i(m)}-\mu_i}[F(s|\mathbf{x}_i)-F(0|\mathbf{x}_i)]\,ds\right.\\
&\qquad\left.-E_{\mathbf{x}_i}\left[\int_0^{\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\hat{\Theta}_{i(m)}-\mu_i}[F(s|\mathbf{x}_i)-F(0|\mathbf{x}_i)]\,ds\right]\right\},
\end{aligned}$$

$$
\begin{aligned}
CV_{4n}(\mathbf{w}) &\equiv \frac{1}{n}\sum_{i=1}^{n} E_{\mathbf{x}_i}\left\{\int_0^{\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\hat{\Theta}_{i(m)}-\mu_i} [F(s|\mathbf{x}_i)-F(0|\mathbf{x}_i)]\,ds\right. \\
&\quad \left. -E_{\mathbf{x}_i}\left[\int_0^{\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\hat{\Theta}_{(m)}-\mu_i} [F(s|\mathbf{x}_i)-F(0|\mathbf{x}_i)]\,ds\right]\right\}, \text{ and} \\
CV_{5n} &\equiv \frac{1}{n}\sum_{i=1}^{n}\{\rho_\tau(\varepsilon_i)-E[\rho_\tau(\varepsilon_i)]\}.
\end{aligned}
$$

We prove (C.1) by showing that (i) $\min_{\mathbf{w}\in\mathcal{W}} FPE_n(\mathbf{w}) \geq E[\rho_\tau(\varepsilon)] - o_P(1)$; (ii) $\sup_{\mathbf{w}\in\mathcal{W}}|CV_{1n}(\mathbf{w})| = o_P(1)$; (iii) $\sup_{\mathbf{w}\in\mathcal{W}}|CV_{2n}(\mathbf{w})| = o_P(1)$; (iv) $\sup_{\mathbf{w}\in\mathcal{W}}|CV_{3n}(\mathbf{w})| = o_P(1)$; (v) $\sup_{\mathbf{w}\in\mathcal{W}}|CV_{4n}(\mathbf{w})| = o_P(1)$; and (vi) $CV_{5n} = o_P(1)$. (vi) follows by the weak law of large numbers so we only show (i)-(v) below.

We first show (i). Let $u(\mathbf{w}) \equiv \mu - \sum_{m=1}^{M} w_m \mathbf{x}'_{(m)}\Theta^*_{(m)}$. Then by Knight's identity, (A.2), Taylor expansion, Jensen inequality, Assumption A2, and Theorem 3.2

$$
\begin{aligned}
& FPE_n(\mathbf{w}) - E[\rho_\tau(\varepsilon+u(\mathbf{w}))] \\
= \;& E\left[\rho_\tau\left(\varepsilon+u(\mathbf{w})-\sum_{m=1}^{M} w_m\mathbf{x}'_{(m)}\left(\hat{\Theta}_{(m)}-\Theta^*_{(m)}\right)\right)-\rho_\tau(\varepsilon+u(\mathbf{w}))\mid \mathcal{D}_n\right] \\
= \;& E\left[\int_0^{\sum_{m=1}^{M} w_m\mathbf{x}'_{(m)}\left(\hat{\Theta}_{(m)}-\Theta^*_{(m)}\right)} [\mathbf{1}\{\varepsilon+u(\mathbf{w})\leq s\}-\mathbf{1}\{\varepsilon+u(\mathbf{w})\leq 0\}]\,ds \mid \mathcal{D}_n\right] \\
= \;& E_{\mathbf{x}}\left[\int_0^{\sum_{m=1}^{M} w_m\mathbf{x}'_{(m)}\left(\hat{\Theta}_{(m)}-\Theta^*_{(m)}\right)} [F(s-u(\mathbf{w})|\mathbf{x})-F(-u(\mathbf{w})|\mathbf{x})]\,ds\right] \\
= \;& E_{\mathbf{x}}\left[\int_0^{\sum_{m=1}^{M} w_m\mathbf{x}'_{(m)}\left(\hat{\Theta}_{(m)}-\Theta^*_{(m)}\right)} f(-u(\mathbf{w})|\mathbf{x})\,s\,ds\right]+o_P(1) \\
= \;& \frac{1}{2}E_{\mathbf{x}}\left\{f(-u(\mathbf{w})|\mathbf{x})\left[\sum_{m=1}^{M} w_m\mathbf{x}'_{(m)}\left(\hat{\Theta}_{(m)}-\Theta^*_{(m)}\right)\right]^2\right\}+o_P(1) \\
\leq \;& \frac{1}{2}E_{\mathbf{x}}\left\{f(-u(\mathbf{w})|\mathbf{x})\sum_{m=1}^{M} w_m\left[\mathbf{x}'_{(m)}\left(\hat{\Theta}_{(m)}-\Theta^*_{(m)}\right)\right]^2\right\}+o_P(1) \\
= \;& \frac{1}{2}\left\{\sum_{m=1}^{M} w_m\left(\hat{\Theta}_{(m)}-\Theta^*_{(m)}\right)' E\left[f(-u(\mathbf{w})|\mathbf{x})\,\mathbf{x}_{(m)}\mathbf{x}'_{(m)}\right]\left(\hat{\Theta}_{(m)}-\Theta^*_{(m)}\right)\right\}+o_P(1) \\
\leq \;& \frac{\bar{c}_A}{2}\max_{1\leq m\leq M}\left\|\hat{\Theta}_{(m)}-\Theta^*_{(m)}\right\|^2+o_P(1)=o_P(1).
\end{aligned}
$$

Let $D(t) \equiv E[\rho_\tau(\varepsilon+t)-\rho_\tau(\varepsilon)]$ where $t\in\mathbb{R}$. It is well known that $D(t)$ has a global minimum at $t=0$. This implies that $\min_{\mathbf{w}\in\mathcal{W}} E[\rho_\tau(\varepsilon+u(\mathbf{w}))] \geq E[\rho_\tau(\varepsilon)]$. Consequently, we have

$$
\min_{\mathbf{w}\in\mathcal{W}} FPE_n(\mathbf{w}) = \min_{\mathbf{w}\in\mathcal{W}} E[\rho_\tau(\varepsilon+u(\mathbf{w}))] - o_P(1) \geq E[\rho_\tau(\varepsilon)] - o_P(1).
$$

(ii) We decompose $CV_{1n}(\mathbf{w})$ as follows

$$\begin{aligned} CV_{1n}(\mathbf{w}) &= \frac{1}{n}\sum_{i=1}^{n}\left[\mu_i - \sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\Theta^*_{(m)}\right]\psi_\tau(\varepsilon_i) - \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\hat{\Theta}_{i(m)} - \Theta^*_{(m)}\right]\right\}\psi_\tau(\varepsilon_i) \\ &\equiv CV_{1n,1}(\mathbf{w}) - CV_{1n,2}(\mathbf{w}). \end{aligned}$$

In view of that $E[CV_{1n,1}(\mathbf{w})] = 0$ and $\mathrm{Var}(CV_{1n,1}(\mathbf{w})) = O(\bar{k}/n)$, we have $CV_{1n,1}(\mathbf{w}) = o_P(1)$ for each $\mathbf{w} \in \mathcal{W}$. If both M and $\bar{k} \equiv \max_{1\le m\le M} k_m$ are finite, we argue that one can apply the Glivenko-Cantelli theorem (e.g., Theorem 2.4.1 in van der Vaart and Wellner (1996)) to conclude $\sup_{\mathbf{w}\in\mathcal{W}} |CV_{1n,1}(\mathbf{w})| = o_P(1)$. To see this, consider the class of functions

$$\mathcal{G} \equiv \{g(\cdot,\cdot;\mathbf{w}) : \mathbf{w} \in \mathcal{W}\}$$

where $g(\cdot,\cdot;\mathbf{w}) : \mathbb{R}\times\mathbb{R}^{d_\mathbf{x}} \to \mathbb{R}$ is defined by $g(\varepsilon_i, \mathbf{x}_i; \mathbf{w}) = [\mu_i - \sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\Theta^*_{(m)}]\psi_\tau(\varepsilon_i)$. Define the metric $|\cdot|_1$ on $\mathcal{W}$ where

$$|\mathbf{w} - \bar{\mathbf{w}}|_1 = \sum_{m=1}^{M} |w_m - \bar{w}_m|,$$

for any $\mathbf{w} = (w_1, ..., w_M) \in \mathcal{W}$ and $\bar{\mathbf{w}} = (\bar{w}_1, ..., \bar{w}_M) \in \mathcal{W}$. It is easy to see the ϵ-covering number of $\mathcal{W}$ with respect to $|\cdot|_1$ is given by $\mathcal{N}(\epsilon, \mathcal{W}, |\cdot|_1) = O(1/\epsilon^{M-1})$. By Theorem 2.7.11 in van der Vaart and Wellner (1996), this, together with the fact that

$$|g(\varepsilon_i, \mathbf{x}_i; \mathbf{w}) - g(\varepsilon_i, \mathbf{x}_i; \bar{\mathbf{w}})| = \left|\sum_{m=1}^{M} (w_m - \bar{w}_m)\mathbf{x}'_{i(m)}\Theta^*_{(m)}\psi_\tau(\varepsilon_i)\right| \le c_\Theta |\mathbf{w} - \bar{\mathbf{w}}|_1 \max_{1\le m\le M} \left\|\mathbf{x}_{i(m)}\right\|$$

where $c_\Theta \equiv \max_{1\le m\le M} ||\Theta^*_{(m)}|| = O\left(\bar{k}^{1/2}\right)$ and that $E\max_{1\le m\le M}\left\|\mathbf{x}_{i(m)}\right\| < \infty$ in the case of finite M and $\bar{k}$ implies that the ϵ-bracketing number of $\mathcal{G}$ with respect to the $L_1(P)$-norm is given by $\mathcal{N}_{[]}(\epsilon, \mathcal{G}, L_1(P)) \le C/\epsilon^{M-1}$ for some finite C. As a result, one can apply Theorem 2.4.1 in van der Vaart and Wellner (1996) to conclude that $\mathcal{G}$ is Glivenko-Cantelli.

The above argument breaks down when either $M \to \infty$ or $\bar{k} \to \infty$ as $n \to \infty$. To allow for diverging M or $\bar{k}$, let $h_n \equiv 1/(\bar{k}\log n)$. We create grids using regions of the form $W_j = \left\{\mathbf{w} : |\mathbf{w} - \mathbf{w}_j|_1 \le h_n\right\}$. By selecting $\mathbf{w}_j = (w_{j1}, ..., w_{jM})$ to lay on a grid, $\mathcal{W}$ can be covered with $N = O\left(1/h_n^{M-1}\right)$ regions W_j, $j = 1, ..., N$. Observe that

$$\begin{aligned} \sup_{\mathbf{w}\in W_j} |CV_{1n,1}(\mathbf{w}) - CV_{1n,1}(\mathbf{w}_j)| &= \sup_{\mathbf{w}\in W_j}\left|\sum_{m=1}^{M}(w_m - w_{jm})\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}'_{i(m)}\Theta^*_{(m)}\psi_\tau(\varepsilon_i)\right| \\ &\le c_\Theta \max_{1\le m\le M}\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_{i(m)}\right\| \sup_{\mathbf{w}\in W_j}\sum_{m=1}^{M}|w_m - w_{jm}| \\ &\le c_\Theta O_P\left(\bar{k}^{1/2}\right) h_n = o_P(1), \end{aligned}$$

where the result holds uniformly in j. Here we have used the fact that

$$\begin{aligned} \max_{1\le m\le M}\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_{i(m)}\right\| &\le \max_{1\le m\le M}\frac{1}{n}\sum_{i=1}^{n}E\left\|\mathbf{x}_{i(m)}\right\| + \max_{1\le m\le M}\left|\frac{1}{n}\sum_{i=1}^{n}\left[\left\|\mathbf{x}_{i(m)}\right\| - E\left\|\mathbf{x}_{i(m)}\right\|\right]\right| \\ &= O\left(\bar{k}^{1/2}\right) + o_P(1) = O_P\left(\bar{k}^{1/2}\right) \end{aligned}$$

by the analogous arguments as used in the study of $CV_{1n,1}(\mathbf{w}_j)$ below. Therefore

$$\begin{aligned}\sup_{\mathbf{w}\in\mathcal{W}}|CV_{1n,1}(\mathbf{w})| &= \max_{1\le j\le N}\sup_{\mathbf{w}\in W_j}|CV_{1n,1}(\mathbf{w})| \\ &\le \max_{1\le j\le N}|CV_{1n,1}(\mathbf{w}_j)| + \max_{1\le j\le N}\sup_{\mathbf{w}\in W_j}|CV_{1n,1}(\mathbf{w})-CV_{1n,1}(\mathbf{w}_j)| \\ &= \max_{1\le j\le N}|CV_{1n,1}(\mathbf{w}_j)| + o_P(1).\end{aligned}$$

Let $u_i(\mathbf{w}_j) \equiv \mu_i - \sum_{m=1}^M w_{jm}\mathbf{x}'_{i(m)}\Theta^*_{(m)}$ and $e_n = \left(Mn\bar{k}^2\right)^{1/4}$. Noting that $|u_i(\mathbf{w}_j)| = |\sum_{m=1}^M w_{jm}[\mu_i - \mathbf{x}'_{i(m)}\Theta^*_{(m)}]| \le \max_{1\le m\le M}\left|b_{i(m)}\right|$ where $b_{i(m)} \equiv \mu_i - \mathbf{x}'_{i(m)}\Theta^*_{(m)}$, we have for any $\epsilon > 0$,

$$\begin{aligned}&\Pr\left(\max_{1\le j\le N} CV_{1n,1}(\mathbf{w}_j) \ge 2\epsilon\right) \\ \le\ &\Pr\left(\max_{1\le j\le N}\frac{1}{n}\sum_{i=1}^n u_i(\mathbf{w}_j)\psi_\tau(\varepsilon_i) \ge 2\epsilon\right) \le \Pr\left(\max_{1\le m\le M}\frac{1}{n}\sum_{i=1}^n\left|b_{i(m)}\right| \ge 2\epsilon\right) \\ \le\ &\Pr\left(\max_{1\le m\le M}\frac{1}{n}\sum_{i=1}^n\left|b_{i(m)}\right|\cdot\mathbf{1}\left\{b_{i(m)}\le e_n\right\} \ge \epsilon\right) + \Pr\left(\max_{1\le m\le M}\frac{1}{n}\sum_{i=1}^n\left|b_{i(m)}\right|\cdot\mathbf{1}\left\{b_{i(m)}\ge e_n\right\} \ge \epsilon\right) \\ \equiv\ &T_{n1}+T_{n2},\ \text{say.}\end{aligned}$$

Noting that $\mathrm{Var}\left[\left|b_{i(m)}\right|\cdot\mathbf{1}\left\{b_{i(m)}\le e_n\right\}\right] \le 2E\left(\mu_i^2\right) + 2E[\mathbf{x}'_{i(m)}\Theta^*_{(m)}]^2 \le \bar{k}\bar{\sigma}^2$ for some $\bar{\sigma}^2 < \infty$, by Boole's and Bernstein's inequalities, we have

$$\begin{aligned}T_{n1} &\le M\max_{1\le j\le N}\Pr\left(\frac{1}{n}\sum_{i=1}^n\left|b_{i(m)}\right|\cdot\mathbf{1}\left\{b_{i(m)}\le e_n\right\}\ge\epsilon\right) \\ &\le 2M\exp\left(-\frac{n\epsilon^2}{2\bar{k}\bar{\sigma}^2+2\epsilon e_n/3}\right) \le 2\exp\left(-\frac{n\epsilon^2}{2\bar{k}\bar{\sigma}^2+2\epsilon\left(Mn\bar{k}^2\right)^{1/4}/3}+\log M\right) = o(1)\end{aligned}$$

where the last equality follows from Assumption A.3(i) and the condition that $n^3/[\bar{k}^2M(\log M)^4] \to \infty$ as $n\to\infty$. Similarly, by Boole's and Markov's inequalities, Assumption A.1(iii) and Lebesgue dominated convergence theorem,

$$\begin{aligned}T_{n2} &\le \Pr\left(\max_{1\le m\le M}\max_{1\le i\le n}\left|\mu_i - \mathbf{x}'_{i(m)}\Theta^*_{(m)}\right| > e_n\right) \le \sum_{m=1}^M\sum_{i=1}^n\Pr\left(\left|\mu_i-\mathbf{x}'_{i(m)}\Theta^*_{(m)}\right|>e_n\right) \\ &\le \frac{1}{e_n^4}\sum_{m=1}^M\sum_{i=1}^n E\left[\left|\mu_i-\mathbf{x}'_{i(m)}\Theta^*_{(m)}\right|^4\mathbf{1}\left(\left|\mu_i-\mathbf{x}'_{i(m)}\Theta^*_{(m)}\right|^4>e_n^4\right)\right] = o(1).\end{aligned}$$

It follows that $\max_{1\le j\le N} CV_{1n,1}(\mathbf{w}_j) = o_P(1)$ and thus $\sup_{\mathbf{w}\in\mathcal{W}}|CV_{1n,1}(\mathbf{w})| = o_P(1)$. By the triangle inequality

$$\begin{aligned}\sup_{\mathbf{w}\in\mathcal{W}}|CV_{1n,2}(\mathbf{w})| &\le \sup_{\mathbf{w}\in\mathcal{W}}\sum_{m=1}^M w_m\frac{1}{n}\sum_{i=1}^n\left|\mathbf{x}'_{i(m)}\left[\hat{\Theta}_{i(m)}-\Theta^*_{(m)}\right]\psi_\tau(\varepsilon_i)\right| \\ &\le \max_{1\le i\le n}\max_{1\le m\le M}\left\|\hat{\Theta}_{i(m)}-\Theta^*_{(m)}\right\|\max_{1\le m\le M}\sum_{i=1}^n\left\|\mathbf{x}_{i(m)}\right\| \\ &\le O_P\left(n^{-1/2}\bar{k}^{1/2}[\log n]^{1/2}\right)O_P\left(\bar{k}^{1/2}\right) = o_P(1).\end{aligned}$$

Consequently $\sup_{\mathbf{w}\in\mathcal{W}}|CV_{1n,2}(\mathbf{w})| = o_P(1)$.

(iii) Observe that $CV_{2n}(\mathbf{w}) = CV_{2n,1}(\mathbf{w}) + CV_{2n,2}(\mathbf{w})$ where

$$CV_{2n,1}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\Theta^*_{(m)}-\mu_i}[\mathbf{1}\{\varepsilon_i\le s\} - \mathbf{1}\{\varepsilon_i\le 0\} - F(s|\mathbf{x}_i) + F(0|\mathbf{x}_i)]\,ds$$

and

$$CV_{2n,2}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\int_{\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\Theta^*_{(m)}-\mu_i}^{\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\hat{\Theta}_{i(m)}-\mu_i}[\mathbf{1}\{\varepsilon_i\le s\} - \mathbf{1}\{\varepsilon_i\le 0\} - F(s|\mathbf{x}_i) + F(0|\mathbf{x}_i)]\,ds.$$

In view of the fact that $|\mathbf{1}\{\varepsilon_i\le s\} - \mathbf{1}\{\varepsilon_i\le 0\} - F(s|\mathbf{x}_i) + F(0|\mathbf{x}_i)| \le 2$, we have

$$\begin{aligned} CV_{2n,2}(\mathbf{w}) &\le \frac{2}{n}\sum_{i=1}^{n}\left|\sum_{m=1}^{M} w_m\mathbf{x}'_{i(m)}\left[\hat{\Theta}_{i(m)} - \Theta^*_{(m)}\right]\right| \\ &\le 2\max_{1\le i\le n}\max_{1\le m\le M}\left\|\hat{\Theta}_{i(m)} - \Theta^*_{(m)}\right\|\max_{1\le m\le M}\frac{1}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_{i(m)}\right\| \\ &= O_P\left(n^{-1/2}\bar{k}^{1/2}[\log n]^{1/2}\right)O_P\left(\bar{k}^{1/2}\right) = o_P(1). \end{aligned} \tag{C.2}$$

Observing that $E[CV_{2n,1}(\mathbf{w})] = 0$ and $\mathrm{Var}(CV_{2n,1}(\mathbf{w})) = O(\bar{k}/n)$, we have $CV_{2n,1}(\mathbf{w}) = O_P((\bar{k}/n)^{1/2})$ for each $\mathbf{w}\in\mathcal{W}$. Analogous to the proof of $CV_{1n,1}(\mathbf{w})$, we can show that $\sup_{\mathbf{w}\in\mathcal{W}}|CV_{2n,1}(\mathbf{w})| = o_P(1)$.

(iv) Observe that $CV_{3n}(\mathbf{w}) = CV_{3n,1}(\mathbf{w}) + CV_{3n,2}(\mathbf{w})$ where

$$\begin{aligned} CV_{3n,1}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{&\int_0^{\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\Theta^*_{(m)}-\mu_i}[F(s|\mathbf{x}_i) - F(0|\mathbf{x}_i)]\,ds \\ &-E\left[\int_0^{\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\Theta^*_{(m)}-\mu_i}[F(s|\mathbf{x}_i) - F(0|\mathbf{x}_i)]\,ds\right]\Bigg\}, \end{aligned}$$

and

$$\begin{aligned} CV_{3n,2}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{&\int_{\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\Theta^*_{(m)}-\mu_i}^{\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\hat{\Theta}_{i(m)}-\mu_i}[F(s|\mathbf{x}_i) - F(0|\mathbf{x}_i)]\,ds \\ &-E_{\mathbf{x}_i}\left[\int_{\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\Theta^*_{(m)}-\mu_i}^{\sum_{m=1}^M w_m\mathbf{x}'_{i(m)}\hat{\Theta}_{i(m)}-\mu_i}[F(s|\mathbf{x}_i) - F(0|\mathbf{x}_i)]\,ds\right]\Bigg\}. \end{aligned}$$

In view of that $|F(s|\mathbf{x}_i) - F(0|\mathbf{x}_i)| \le 1$, we have

$$\begin{aligned} |CV_{3n,2}(\mathbf{w})| &\le \frac{1}{n}\sum_{i=1}^{n}\left|\sum_{m=1}^{M} w_m\mathbf{x}'_{i(m)}\left[\hat{\Theta}_{i(m)} - \Theta^*_{(m)}\right]\right| + \frac{1}{n}\sum_{i=1}^{n}E_{\mathbf{x}_i}\left|\sum_{m=1}^{M} w_m\mathbf{x}'_{i(m)}\left[\hat{\Theta}_{i(m)} - \Theta^*_{(m)}\right]\right| \\ &\equiv CV_{3n,21}(\mathbf{w}) + CV_{3n,22}(\mathbf{w}). \end{aligned}$$

The first term is studied above in (C.2). For the second term, by the triangle and Cauchy-Schwarz

inequalities, the fact $A'BA \leq \lambda_{\max}(B) A'A$ for any real symmetric matrix B, and Theorem 3.2, we have

$$\begin{aligned}
\sup_{\mathbf{w}\in\mathcal{W}} CV_{3n,22}(\mathbf{w}) &\leq \sup_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_m E_{\mathbf{x}_i}\left|\mathbf{x}'_{i(m)}\left[\hat{\Theta}_{i(m)} - \Theta^*_{(m)}\right]\right| \\
&\leq \sup_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n}\sum_{m=1}^{M} w_m \left\{\left[\hat{\Theta}_{i(m)} - \hat{\Theta}_{(m)}\right]' E\left[\mathbf{x}_{i(m)}\mathbf{x}'_{i(m)}\right]\left[\hat{\Theta}_{i(m)} - \hat{\Theta}_{(m)}\right]\right\}^{1/2} \\
&\leq \max_{1\leq m\leq M}\left[\lambda_{\max}\left(E\left[\mathbf{x}_{i(m)}\mathbf{x}'_{i(m)}\right]\right)\right]^{1/2} \max_{1\leq i\leq n}\max_{1\leq m\leq M}\left\|\hat{\Theta}_{i(m)} - \hat{\Theta}_{(m)}\right\| \\
&= o_P(1).
\end{aligned}$$

Consequently $\sup_{\mathbf{w}\in\mathcal{W}} |CV_{3n,2}(\mathbf{w})| = o_P(1)$. The proof that $\sup_{\mathbf{w}\in\mathcal{W}} |CV_{3n,1}(\mathbf{w})| = o_P(1)$ is analogous to that of $\sup_{\mathbf{w}\in\mathcal{W}} |CV_{1n,1}(\mathbf{w})| = o_P(1)$ and thus omitted.

(v) For $CV_{4n}(\mathbf{w})$, noting that $|F(s|\mathbf{x}_i) - F(0|\mathbf{x}_i)| \leq 1$ and by the study of $CV_{3n,22}(\mathbf{w})$ we have

$$\sup_{\mathbf{w}\in\mathcal{W}} CV_{4n}(\mathbf{w}) \leq \sup_{\mathbf{w}\in\mathcal{W}} \frac{1}{n}\sum_{i=1}^{n} E_{\mathbf{x}_i}\left|\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\hat{\Theta}_{i(m)} - \hat{\Theta}_{(m)}\right]\right| = o_P(1).$$

This completes the proof of the theorem. ■

References

Akaike, H., 1970. Statistical predictor identification. Annals of the Institute of Statistical Mathematics 22, 203-217.

Andrews, D. W. K., 1991. Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. Journal of Econometrics 47, 359-377.

Angrist, J., Chernozhukov, V., Fernández-Val, I., 2006. Quantile regression under misspecification, with an application to the U.S. wage structure. Econometrica 74, 539-563.

Angrist, J., Pischke, J., 2009. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press, Princeton, NJ.

Belloni, A., Chernozhukov, V., 2011. l_1-penalized quantile regression in high-dimensional sparse models. Annals of Statistics 39, 82-130.

Bernstein, D. S., 2005. Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems Theory. Princeton University Press, Princeton.

Buckland, S. T., Burnham, K. P., Augustin, N. H., 1997. Model selection: an integral part of inference. Biometrics 53, 603-618.

Burnham, K. P., Anderson, D. R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach. Berlin: Springer-Verlag.

Burman, P., Nolan, D., 1995. A general Akaike-type criterion for model selection in robust regression. Biometrika 82, 877-886.

Campbell, J. Y., Thompson, S. B., 2008. Predicting the equity premium out of sample: can anything beat the historical average? Review of Financial Studies 21, 1509-1531.

Claeskens, G., Hjort, N. L., 2008. Model Selection and Model Averaging. Cambridge University Press, Cambridge.

Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. Annals of Statistics 32, 928-961.

Goyal, A., Welch, I., 2008. A comprehensive look at the empirical performance of equity premium prediction. Review of Financial Studies 21, 1455-1508.

Hansen, B. E., 2005. Challenges for econometric model selection. Econometric Theory 21, 60-68.

Hansen, B. E., 2007. Least squares model averaging. Econometrica 75, 1175-1189.

Hansen, B. E., 2008. Least-squares forecast averaging. Journal of Econometrics 146, 342-350.

Hansen, B. E., 2009. Averaging estimators for regressions with a possible structural break. Econometric Theory 25, 1489-1514.

Hansen, B. E., 2010. Averaging estimators for autoregressions with a near unit root. Journal of Econometrics 158, 142-155.

Hansen, B. E., Racine, J. S., 2012. Jackknife model averaging. Journal of Econometrics 167, 38-46.

Hjort, N. L., Claeskens, G., 2003. Frequentist model average estimators. Journal of the American Statistical Association 98, 879-899.

Hoeting, J. A., Madigan, D., Raftery, A. E., Volinsky, C. T., 1999. Bayesian Model Averaging: A Tutorial. Statistical Science 14, 382-417.

Hurvich, C. M., Tsai, C-L., 1990. Model selection for least absolute deviations regression in small samples. Statistics & Probability Letters 9, 259-265.

Jin, S., Su, L., Ullah, A., 2014. Robustify financial time series forecasting with bagging. Econometric Reviews 33, 575-605.

Kapetanios, G., Labhard, V., Price, S., 2008. Forecasting using Bayesian and information-theoretical model averaging: an application to U.K. inflation. Journal of Business & Economic Statistics 26, 33-41.

Knight, K., 1998. Limiting distributions for L_1 regression estimators under general conditions. Annals of Statistics 26, 755-770.

Koenker, R., 2005. Quantile Regression. Cambridge University Press, New York.

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46, 33-50.

Koenker, R., Bassett, G., 1982. Robust tests for heteroscedasticity based on regression quantiles. Econometrica 50, 43-61.

Koenker, R., Machado, J.A.F., 1999. Goodness of fit and related inference processes for quantile regression. Journal of the American Statistical Association 94, 1296-1310.

Koenker, R., Ng, P., Portnoy, S., 1994. Quantile smoothing splines. Biometrika 81, 673-680.

Kuersteiner, G., Okui, R., 2010. Constructing optimal instruments by first-stage prediction averaging. Econometrica 78, 697-718.

Kuester, K., Mittnik, S., Paolella, M., 2006. Value-at-risk prediction: a comparison of alternative strategies. Journal of Financial Econometrics 4, 53-89.

Lam, C., Fan, J., 2008. Profile-kernel likelihood inference with diverging number of parameters. Annals of Statistics 36, 2232-2260.

Lee, Y., Zhou, Y., 2011. Averaged instrumental variable estimators. Working paper, University of Michigan.

Leung, G., Barron, A. R., 2006. Information theory and mixing least-squares regressions. IEEE Transactions on Information Theory 52, 3396-3410.

Li, K-C., 1987. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. Annals of Statistics 15, 958-975.

Liang, H., Zou, G., Wan, A. T. K., Zhang, X., 2011. Optimal weight choice for frequentist model average estimators. Journal of the American Statistical Association 106, 1053-1066.

Liu, C-A., 2012. A plug-in averaging estimator for regression with heteroskedastic errors. Working paper, National University of Singapore.

Liu, Q., Okui, R., 2013. Heteroskedasticity-robust C_p model averaging. Econometrics Journal 16, 463-472.

Machado, J. A. F., 1993. Robust model selection and M-estimation. Econometric Theory 9, 478-493.

Moral-Benito, E., 2013. Model averaging in economics: an overview. Journal of Economic Surveys, forthcoming.

Ng., S., 2013. Variable selection in predictive regressions. In G. Elliott, A. Timmermann (eds.), Handbook of Economic Forecasting, Vol 2B, Chapter 14, pp. 752-789.

Pollard, D., 1991. Asymptotics for least absolute deviation regression estimators. Econometric Theory 7, 186-199.

Portnoy, S., 1984. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large: I. consistency. Annals of Statistics 12, 1298-1309.

Portnoy, S.,1985. Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large: II. normal approximation. Annals of Statistics 13, 1403-1417.

Portnoy, S., 1986. On the central limit theorem in R^p when $p \to \infty$. Probability Theory and Related Fields 73, 571-583.

Pötscher, B. M., 2006. The distribution of model averaging estimators and an impossibility result regarding its estimation. Time Series and Related Topics, IMS Lecture Notes-Monograph Series 52, 113-129.

Rice, J., 1984. Bandwidth choice for nonparametric regression. Annals of Statistics 12, 1215-1230.

Ruppert, D., Carroll, R. J., 1980. Trimmed least squares estimation in the linear model. Journal of the American Statistical Association 75, 828-838.

Serfling, R. J., 1980. Approximation Theorems of Mathematical Statistics. John Wiley & Sons, New York.

Shibata, R., 1981. An optimal selection of regression variables. Biometrika 68, 45-54.

Shibata, R., 1982. Correction for "An optimal selection of regression variables". Biometrika 69, 492.

Su, L., Zhang, Y., 2014. Variable selection in nonparametric and semiparametric regression models. In J. Racine, L. Su, A. Ullah, (eds), The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics, pp. 249-307. Oxford University Press, New York.

Sueishi, N., 2010. Model selection criterion for infinite dimensional instrumental variable models. Working paper, Kyoto University.

Van der Vaart, A., Wellner, J. A., 1996. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer, New York.

Wan, A. T. K., Zhang, X., Zhou, G., 2010. Least squares model averaging by Mallows criterion. Journal of Econometrics 156, 277-283.

Welsh, A. H., 1989. On M-processes and M-estimation (Corr: v18, p1500). Annals of Statistics 17, 337-361.

Wooldridge, J. M., 2003. Introductory Econometrics. Thompson South-Western.

Wu, Y., Liu, Y., 2009. Variable selection in quantile regression. Statistica Sinica 19, 801-817.

Yang, Y., 2001. Adaptive regression by mixing. Journal of the American Statistical Association 96, 574-586.

Yuan, Z., Yang, Y., 2005. Combining linear regression models: when and how? Journal of the American Statistical Association 100, 1202-1214.

Zhang, X., Wan, A. T. K., Zou, G., 2013. Model averaging by jackknife criterion in models with dependent data. Journal of Econometrics 174, 82-94.

Supplementary Material for "Jackknife Model Averaging for Quantile Regressions"

Xun Lu[a], Liangjun Su[b]
[a] *Department of Economics, HKUST*
[b] *School of Economics, Singapore Management University*

THIS SUPPLEMENTARY MATERIAL PROVIDES AN INFORMAL DERIVATION OF (3.8) IN THE TEXT AND SIMULATION RESULTS FOR THE HOMOSKEDASTICITY CASE.

D An Informal Derivation of (3.8)

By Knight's identity

$$\begin{aligned} nCV_n(\mathbf{w}) - nQ_n(\mathbf{w}) &= \sum_{i=1}^{n}\left\{\rho_\tau\left(\varepsilon_i + u_i(\mathbf{w}) + \sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\Theta^*_{(m)} - \hat{\Theta}_{i(m)}\right]\right) - \rho_\tau(\varepsilon_i + u_i(\mathbf{w}))\right. \\ &\quad \left. - \left[\rho_\tau\left(\varepsilon_i + u_i(\mathbf{w}) + \sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\Theta^*_{(m)} - \hat{\Theta}_{(m)}\right]\right) - \rho_\tau(\varepsilon_i + u_i(\mathbf{w}))\right]\right\} \\ &= \sum_{i=1}^{n}\sum_{m=1}^{M}\{w_m \mathbf{x}'_{i(m)}\left[\hat{\Theta}_{(m)} - \hat{\Theta}_{i(m)}\right]\psi_\tau(\varepsilon_i + u_i(\mathbf{w})) \\ &\quad + \int_{\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\hat{\Theta}_{(m)} - \Theta^*_{(m)}\right]}^{\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\hat{\Theta}_{i(m)} - \Theta^*_{(m)}\right]} [\mathbf{1}\{\varepsilon_i + u_i(\mathbf{w}) \le s\} - \mathbf{1}\{\varepsilon_i + u_i(\mathbf{w}) \le 0\}]\, ds\} \\ &\equiv A_{1n}(\mathbf{w}) + A_{2n}(\mathbf{w}), \text{ say.} \end{aligned}$$

It is straightforward to show that $\sup_{\mathbf{w}\in\mathcal{W}} A_{sn}(\mathbf{w})/n = o_P(1)$ for $s = 1, 2$, which means that $CV_n(\mathbf{w})$ and $Q_n(\mathbf{w})$ differ only in smaller order terms. To derive the Mallows-type QR information criterion in (3.6), we have to examine the smaller order terms. By (A.15),

$$\sqrt{n}\left(\hat{\Theta}_{(m)} - \Theta^*_{(m)}\right) = P_{V_{(m)}^{-1/2} C'_{(m)}} A_{(m)}^{-1}\, n^{-1/2} \sum_{i=1}^{n} \mathbf{x}_{i(m)}\left[\tau - \mathbf{1}\left\{y_i \le \Theta^{*\prime}_{(m)} \mathbf{x}_{i(m)}\right\}\right] + \text{s.m.}, \tag{D.1}$$

where s.m. denotes terms of smaller order than the preceding one. Similarly,

$$\sqrt{n}\left(\hat{\Theta}_{i(m)} - \Theta^*_{(m)}\right) = P_{V_{(m)}^{-1/2} C'_{(m)}} A_{(m)}^{-1} n^{-1/2} \sum_{j=1, j\neq i}^{n}\left[\tau - \mathbf{1}\left\{y_j \le \Theta^{*\prime}_{(m)} \mathbf{x}_{j(m)}\right\}\right] \mathbf{x}_{j(m)} + \text{s.m.} \tag{D.2}$$

It follows that

$$\sqrt{n}\left[\hat{\Theta}_{(m)} - \hat{\Theta}_{i(m)}\right] = n^{-1/2} P_{V_{(m)}^{-1/2} C'_{(m)}} A_{(m)}^{-1}\left[\tau - \mathbf{1}\left\{y_i \le \Theta^{*\prime}_{(m)} \mathbf{x}_{i(m)}\right\}\right] \mathbf{x}_{i(m)} + \text{s.m.} \tag{D.3}$$

where we conjecture that the difference in the two smaller terms in (D.1) and (D.2) is also of smaller order in (D.3). The former derivation of such a claim require one to replace the indictor function by a CDF-type smooth function as in smoothed quantile regressions (see, e.g., Kato and Galvao (2010) and

Su and White (2012)). Then

$$\begin{aligned} A_{1n}(\mathbf{w}) &= n^{-1}\sum_{i=1}^{n}\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)} P_{V_{(m)}^{-1/2}C'_{(m)}} A_{(m)}^{-1}\mathbf{x}_{i(m)}\left[\tau - \mathbf{1}\left\{y_i \le \Theta_{(m)}^{*\prime}\mathbf{x}_{i(m)}\right\}\right]\psi_\tau(\varepsilon_i + u_i(\mathbf{w})) + \text{ s.m.} \\ &= \sum_{m=1}^{M} w_m E\left[\mathbf{x}'_{i(m)} P_{V_{(m)}^{-1/2}C'_{(m)}} A_{(m)}^{-1}\mathbf{x}_{i(m)}\psi_\tau\left(\varepsilon_i + u_{i(m)}\right)\psi_\tau(\varepsilon_i + u_i(\mathbf{w}))\right] + \text{ s.m.} \end{aligned}$$

Without additional assumptions it is difficult to simplify the dominant term in the last expression.

Note that

$$\begin{aligned} A_{2n}(\mathbf{w}) &= \sum_{i=1}^{n}\sum_{m=1}^{M}\int_{\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\hat{\Theta}_{(m)} - \Theta_{(m)}^{*}\right]}^{\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\hat{\Theta}_{i(m)} - \Theta_{(m)}^{*}\right]} \left[F(-u_i(\mathbf{w}) + s|\mathbf{x}_i) - F(-u_i(\mathbf{w})|\mathbf{x}_i)\right] ds \\ &\quad + \sum_{i=1}^{n}\sum_{m=1}^{M}\int_{\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\hat{\Theta}_{(m)} - \Theta_{(m)}^{*}\right]}^{\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\hat{\Theta}_{i(m)} - \Theta_{(m)}^{*}\right]} \alpha_i(\mathbf{w}, s)\, ds \\ &\equiv A_{2n,1}(\mathbf{w}) + A_{2n,2}(\mathbf{w}) \end{aligned}$$

where $\alpha_i(\mathbf{w}, s) \equiv \mathbf{1}\{\varepsilon_i + u_i(\mathbf{w}) \le s\} - \mathbf{1}\{\varepsilon_i + u_i(\mathbf{w}) \le 0\} - F(-u_i(\mathbf{w}) + s|\mathbf{x}_i) + F(-u_i(\mathbf{w})|\mathbf{x}_i)$. By Taylor expansion, Theorem 3.2 and (D.3),

$$\begin{aligned} A_{2n,1}(\mathbf{w}) &= \sum_{i=1}^{n}\sum_{m=1}^{M}\int_{\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\hat{\Theta}_{(m)} - \Theta_{(m)}^{*}\right]}^{\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\hat{\Theta}_{i(m)} - \Theta_{(m)}^{*}\right]} f(-u_i(\mathbf{w})|\mathbf{x}_i)\, s ds + \text{s.m.} \\ &= \sum_{i=1}^{n}\sum_{m=1}^{M} f(-u_i(\mathbf{w})|\mathbf{x}_i)\left[\left\{\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\hat{\Theta}_{i(m)} - \Theta_{(m)}^{*}\right]\right\}^2 - \left\{\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\hat{\Theta}_{(m)} - \Theta_{(m)}^{*}\right]\right\}^2\right] \\ &\quad + \text{s.m.} \\ &= \sum_{i=1}^{n}\sum_{m=1}^{M} f(-u_i(\mathbf{w})|\mathbf{x}_i)\left\{\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\hat{\Theta}_{i(m)} - \hat{\Theta}_{(m)}\right]\right\}^2 \\ &\quad + 2\sum_{i=1}^{n}\sum_{m=1}^{M} f(-u_i(\mathbf{w})|\mathbf{x}_i)\left\{\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\hat{\Theta}_{i(m)} - \hat{\Theta}_{(m)}\right]\right\}\left\{\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\left[\hat{\Theta}_{(m)} - \Theta_{(m)}^{*}\right]\right\} \\ &\quad + \text{s.m.} \\ &= o_P(1). \end{aligned}$$

Let $\hat{\Delta}_{(m)} \equiv \sqrt{n}[\hat{\Theta}_{(m)} - \Theta_{(m)}^{*}]$, $\hat{\Delta}_{i(m)} \equiv \sqrt{n}[\hat{\Theta}_{i(m)} - \Theta_{(m)}^{*}]$, $\hat{\mathbf{\Delta}} \equiv \left(\hat{\Delta}_{(1)}, ..., \hat{\Delta}_{(M)}\right)$, and $\hat{\mathbf{\Delta}}_{-i} \equiv (\hat{\Delta}_{-i(1)}, ..., \hat{\Delta}_{-i(M)})$ for $i = 1, ..., n$. Let $\mathbf{\Delta} \equiv \left(\Delta_{(1)}, ..., \Delta_{(M)}\right)$ and $\mathbf{\Delta}_{-i} \equiv (\Delta_{-i(1)}, ..., \Delta_{-i(M)})$ for $i = 1, ..., n$. Define

$$\mathcal{A}_n(\mathbf{\Delta}, \mathbf{\Delta}_{-1}, ..., \mathbf{\Delta}_{-n}; \mathbf{w}) \equiv \sum_{i=1}^{n}\int_{n^{-1/2}\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\Delta_{(m)}}^{n^{-1/2}\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\Delta_{i(m)}} \alpha_i(\mathbf{w}, s)\, ds$$

Clearly, $A_{2n,2}(\mathbf{w}) = \sum_{i=1}^{n}\sum_{m=1}^{M}\int_{n^{-1/2}\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\hat{\Delta}_{(m)}}^{n^{-1/2}\sum_{m=1}^{M} w_m \mathbf{x}'_{i(m)}\hat{\Delta}_{i(m)}} \alpha_i(\mathbf{w}, s)\, ds = \mathcal{A}_n\left(\hat{\mathbf{\Delta}}, \hat{\mathbf{\Delta}}_{-1}, ..., \hat{\mathbf{\Delta}}_{-n}; \mathbf{w}\right)$. Let L be a large fixed constant. Define

$$\mathcal{S}_L \equiv \{(\mathbf{\Delta}, \mathbf{\Delta}_{-1}, ..., \mathbf{\Delta}_{-n}) : \ \left\|\Delta_{(m)}\right\| \le \bar{k}^{1/2}L, \left\|\Delta_{-i(m)}\right\| \le \bar{k}^{1/2}L, \left\|\Delta_{(m)} - \Delta_{-i(m)}\right\| \le \bar{k}^{1/2}Ln^{-1/2}\log n \text{ for } i = 1, ..., n, \ m = 1..., M\}.$$

We can prove $\sup_{\mathbf{w}\in\mathcal{W}}|A_{2n,2}(\mathbf{w})| = o_P(1)$ by showing that

$$\sup_{\mathbf{w}\in\mathcal{W}}\sup_{(\boldsymbol{\Delta},\boldsymbol{\Delta}_{-1},...,\boldsymbol{\Delta}_{-n})\in\mathcal{S}_L}|\mathcal{A}_n(\boldsymbol{\Delta},\boldsymbol{\Delta}_{-1},...,\boldsymbol{\Delta}_{-n};\mathbf{w})| = o_P(1).$$

We first show that $\mathcal{A}_n(\boldsymbol{\Delta},\boldsymbol{\Delta}_{-1},...,\boldsymbol{\Delta}_{-n};\mathbf{w}) = o_P(1)$ for each $(\boldsymbol{\Delta},\boldsymbol{\Delta}_{-1},...,\boldsymbol{\Delta}_{-n};\mathbf{w})$, and then show the above uniform result. By Jensen's inequality, Taylor expansions, and the fact that $\left\|\Delta_{-i(m)}-\Delta_{(m)}\right\| = O\left(\bar{k}^{1/2}n^{-1/2}\log n\right)$, we have

$$\begin{aligned}
&\mathrm{Var}\left(\mathcal{A}_n(\boldsymbol{\Delta},\boldsymbol{\Delta}_{-1},...,\boldsymbol{\Delta}_{-n};\mathbf{w})\right)\\
=&\sum_{i=1}^{n}\mathrm{Var}\left(\int_{n^{-1/2}\sum_{m=1}^{M}w_m\mathbf{x}'_{i(m)}\Delta_{(m)}}^{n^{-1/2}\sum_{m=1}^{M}w_m\mathbf{x}'_{i(m)}\Delta_{i(m)}}\alpha_i(\mathbf{w},s)\,ds\right)\\
\leq&\sum_{i=1}^{n}E\left[\left(\int_{n^{-1/2}\sum_{m=1}^{M}w_m\mathbf{x}'_{i(m)}\Delta_{(m)}}^{n^{-1/2}\sum_{m=1}^{M}w_m\mathbf{x}'_{i(m)}\Delta_{i(m)}}\left[\mathbf{1}\left\{\varepsilon_i+u_i(\mathbf{w})\leq s\right\}-\mathbf{1}\left\{\varepsilon_i+u_i(\mathbf{w})\leq 0\right\}\right]ds\right)^2\right]\\
=&\sum_{i=1}^{n}E\left[\int_{n^{-1/2}\sum_{m=1}^{M}w_m\mathbf{x}'_{i(m)}\Delta_{(m)}}^{n^{-1/2}\sum_{m=1}^{M}w_m\mathbf{x}'_{i(m)}\Delta_{i(m)}}\int_{n^{-1/2}\sum_{m=1}^{M}w_m\mathbf{x}'_{i(m)}\Delta_{(m)}}^{n^{-1/2}\sum_{m=1}^{M}w_m\mathbf{x}'_{i(m)}\Delta_{i(m)}}\right.\\
&\left.\times F\left((s\wedge t)-u_i(\mathbf{w})|\mathbf{x}_i\right)-F\left((s\wedge 0)-u_i(\mathbf{w})|\mathbf{x}_i\right)-F\left((0\wedge t)-u_i(\mathbf{w})|\mathbf{x}_i\right)+F\left(-u_i(\mathbf{w})|\mathbf{x}_i\right)\right]dsdt\\
=&\sum_{i=1}^{n}E\left[\int_{n^{-1/2}\sum_{m=1}^{M}w_m\mathbf{x}'_{i(m)}\Delta_{(m)}}^{n^{-1/2}\sum_{m=1}^{M}w_m\mathbf{x}'_{i(m)}\Delta_{i(m)}}\int_{n^{-1/2}\sum_{m=1}^{M}w_m\mathbf{x}'_{i(m)}\Delta_{(m)}}^{n^{-1/2}\sum_{m=1}^{M}w_m\mathbf{x}'_{i(m)}\Delta_{i(m)}}f\left(-u_i(\mathbf{w})|\mathbf{x}_i\right)\left[(s\wedge t)-(s\wedge 0)-(0\wedge t)\right]dsdt\right]\\
&+\text{s.m.}\\
=&O\left(\bar{k}^{1/2}n^{-1/2}\log n\right)=o(1).
\end{aligned}$$

Then $\mathcal{A}_n(\boldsymbol{\Delta},\boldsymbol{\Delta}_{-1},...,\boldsymbol{\Delta}_{-n};\mathbf{w}) = o_P(1)$ for each $(\boldsymbol{\Delta},\boldsymbol{\Delta}_{-1},...,\boldsymbol{\Delta}_{-n};\mathbf{w})$ by the Chebyshev inequality. As in the proof of Theorems 3.2 and 3.3, we can apply Bernstein's inequality to show this convergence also holds uniformly in $(\boldsymbol{\Delta},\boldsymbol{\Delta}_{-1},...,\boldsymbol{\Delta}_{-n};\mathbf{w})\in\mathcal{S}_L\times\mathcal{W}$. Consequently, we have

$$nCV_n(\mathbf{w}) = nQ_n(\mathbf{w})+\sum_{m=1}^{M}w_mE\left[\mathbf{x}'_{i(m)}P_{V_{(m)}^{-1/2}C'_{(m)}}A_{(m)}^{-1}\mathbf{x}_{i(m)}\psi_\tau\left(\varepsilon_i+u_{i(m)}\right)\psi_\tau\left(\varepsilon_i+u_i(\mathbf{w})\right)\right]+\text{ s.m.}$$

To simplify, we assume that (D.1) and (D.2) continue to hold by taking $C_m = I_{k_m}$ (say when $\bar{k}$ is fixed) which implies that $P_{V_{(m)}^{-1/2}C'_{(m)}} = I_{k_m}$. If in addition the approximation bias $u_{i(m)} = o_{a.s.}(1)$ for all $m = 1,...,M$ (say when all models under consideration are approximately correct), then by the dominated convergence theorem we have

$$\begin{aligned}
nCV_n(\mathbf{w}) &= nQ_n(\mathbf{w})+\sum_{m=1}^{M}w_m\mathrm{tr}\left\{A_{(m)}^{-1}E\left[\mathbf{x}_{i(m)}\mathbf{x}'_{i(m)}\psi_\tau\left(\varepsilon_i+u_{i(m)}\right)\psi_\tau\left(\varepsilon_i+u_i(\mathbf{w})\right)\right]\right\}+\text{s.m.}\\
&= nQ_n(\mathbf{w})+\sum_{m=1}^{M}w_m\mathrm{tr}\left\{A_{(m)}^{-1}E\left[\mathbf{x}_{i(m)}\mathbf{x}'_{i(m)}\psi_\tau(\varepsilon_i)^2\right]\right\}+\text{s.m.}\\
&= nQ_n(\mathbf{w})+\tau(1-\tau)\sum_{m=1}^{M}w_m\mathrm{tr}\left\{A_{(m)}^{-1}E\left[\mathbf{x}_{i(m)}\mathbf{x}'_{i(m)}\right]\right\}+\text{s.m.}
\end{aligned}$$

If ε_i and $\mathbf{x}_{i(m)}$ are independent for all $m = 1,2,...,M$, then $A_{(m)} = f(F^{-1}(\tau))E\left[\mathbf{x}_{i(m)}\mathbf{x}'_{i(m)}\right]$ and

$$nCV_n(\mathbf{w}) = nQ_n(\mathbf{w})+\frac{\tau(1-\tau)}{f(F^{-1}(\tau))}\sum_{m=1}^{M}w_mk_m+\text{s.m.}$$

E Simulation Results for the Homoskedasticity Case

This appendix contains some simulation results for DGPs 1-4 when the error terms are homoskedastic. The findings from Figures S1-S4 are largely consistent with those for DGPs 1-4 when the error terms are heteroskedastic. In particular, when $\tau = 0.05$, JMA clearly dominates all other model averaging estimators for all DGPs under investigation.

REFERENCES

Kato, K., Galvao, A. F., 2010. Smoothed quantile regression for panel data. Working paper, Dept. of Economics, University of Iowa.

Su, L., White, H., 2012. Conditional independence specification testing for dependent processes with local polynomial quantile regression. Advances in Econometrics 29, 355-434.

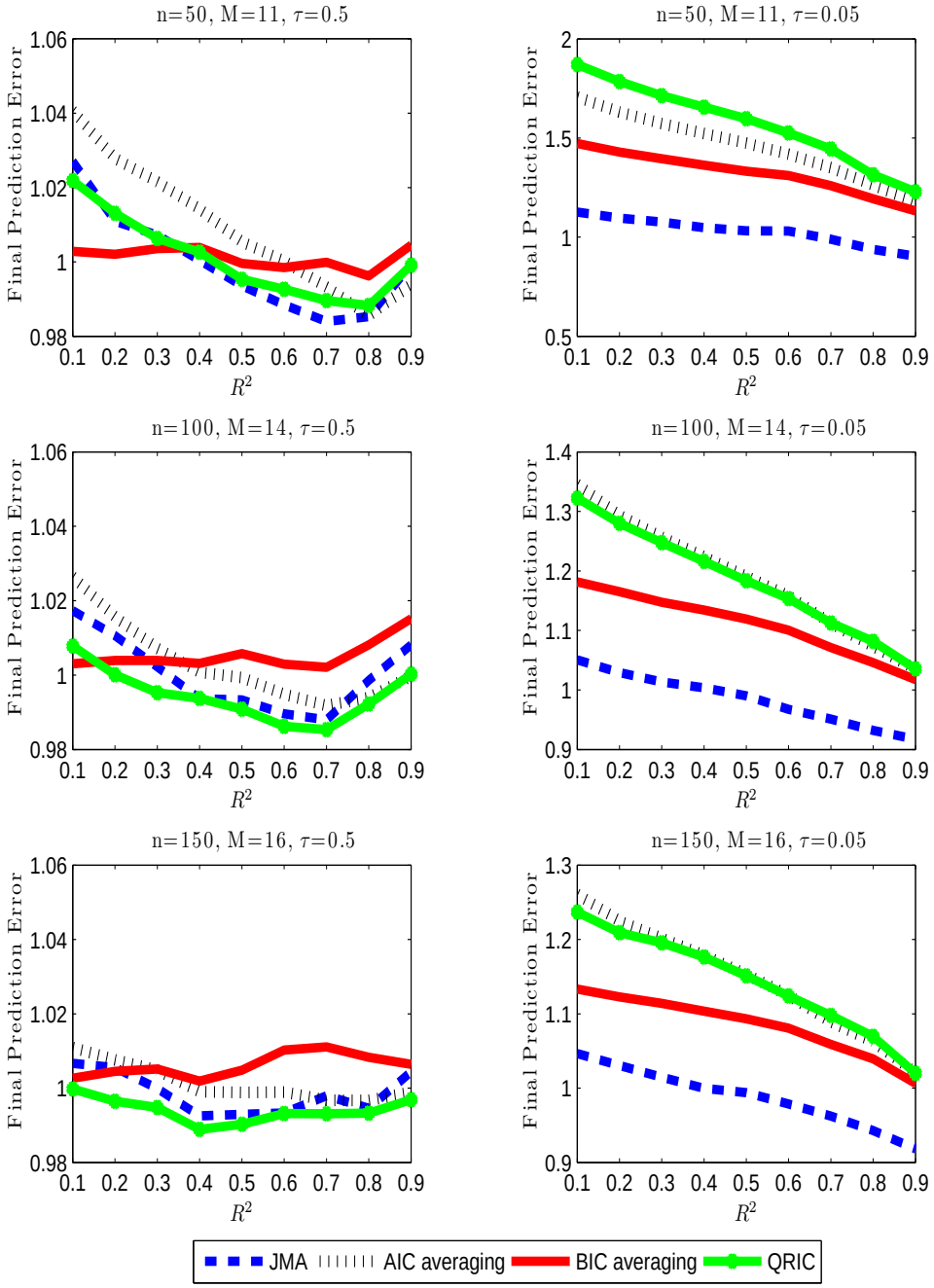

Figure S1: Out-of-sample performance: DGP 1, Homoskedasticity

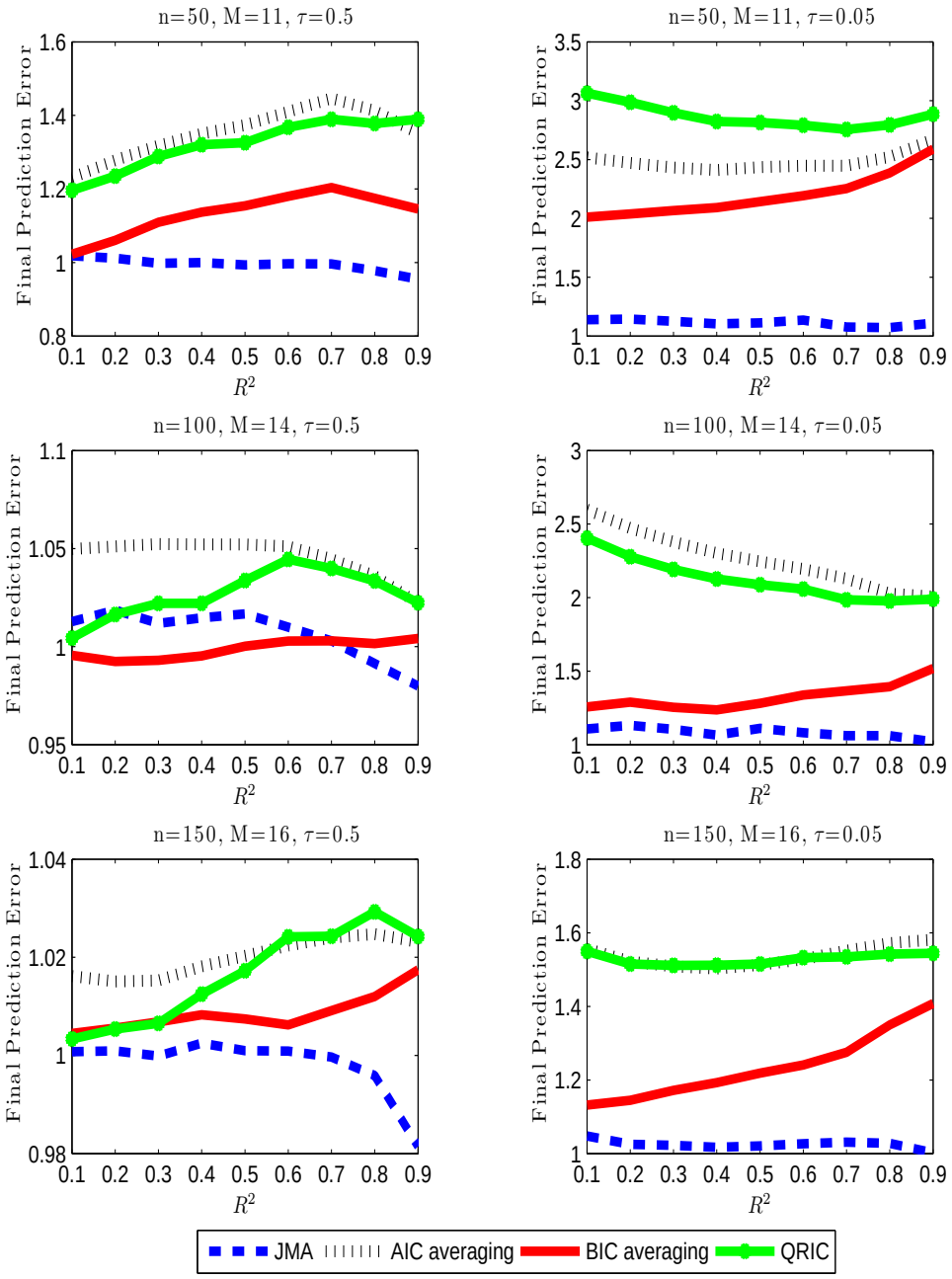

Figure S2: Out-of-sample performance: DGP 2, Homoskedasticity

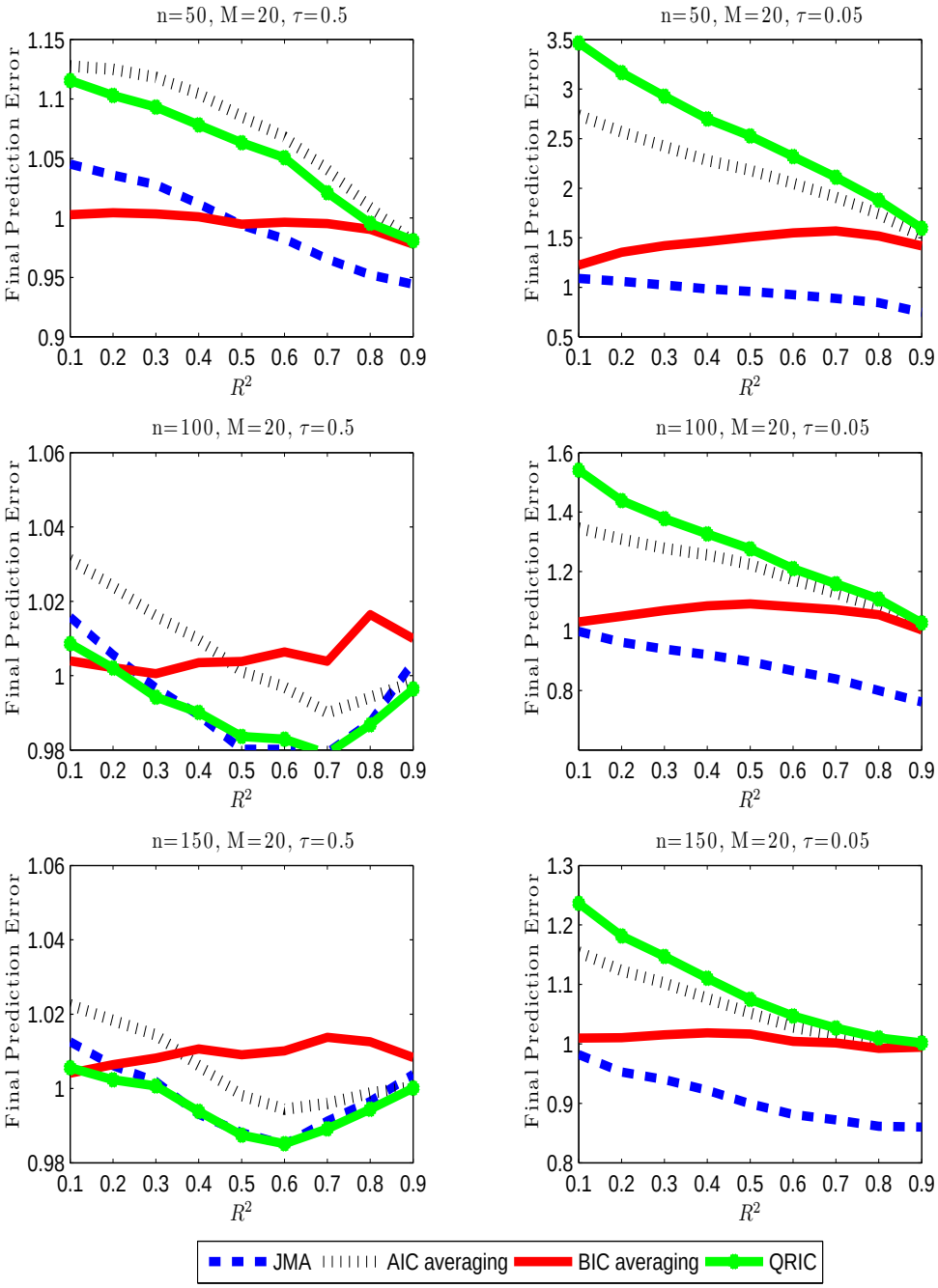

Figure S3: Out-of-sample performance: DGP 3, Homoskedasticity

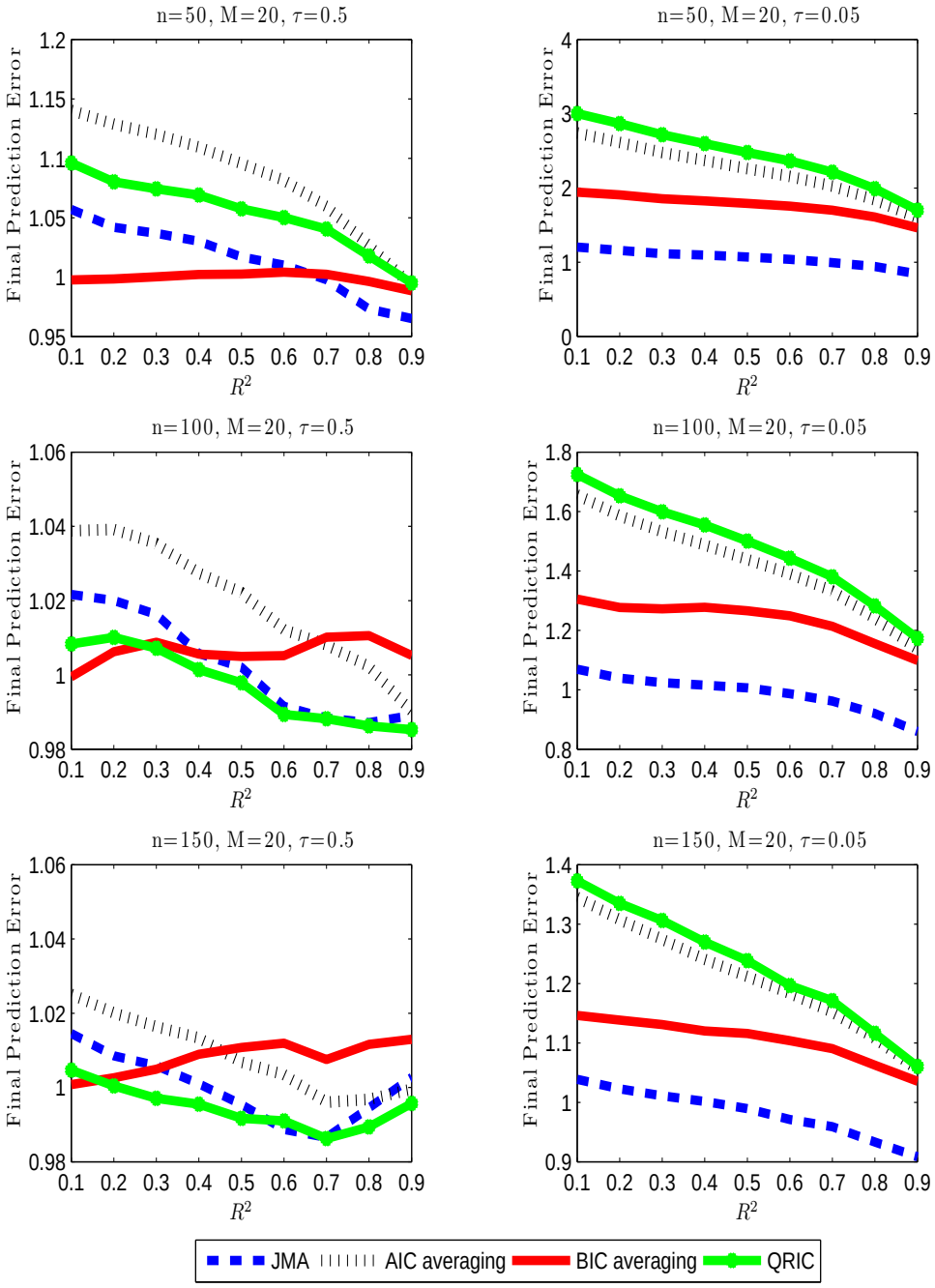

Figure S4: Out-of-sample performance: DGP 4, Homoskedasticity