

COMPLETE SUBSET AVERAGING FOR QUANTILE REGRESSIONS

Ji HYUNG LEE
University of Illinois

YOUNGKI SHIN
McMaster University

We propose a novel conditional quantile prediction method based on complete subset averaging (CSA) for quantile regressions. All models under consideration are potentially misspecified, and the dimension of regressors goes to infinity as the sample size increases. Since we average over the complete subsets, the number of models is much larger than the usual model averaging method which adopts sophisticated weighting schemes. We propose to use an equal weight but select the proper size of the complete subset based on the leave-one-out cross-validation method. Building upon the theory of Lu and Su (2015, *Journal of Econometrics* 188, 40–58), we investigate the large sample properties of CSA and show the asymptotic optimality in the sense of Li (1987, *Annals of Statistics* 15, 958–975) We check the finite sample performance via Monte Carlo simulations and empirical applications.

1. INTRODUCTION

Quantile regression (QR) has emerged as an essential tool since Koenker and Bassett (1978) (see, e.g., Koenker, 2005). QR estimates the response of conditional quantiles of outcome variables with respect to changes in the covariates. The entire response distribution of outcome variables in economic models provides a broader insight than the classical mean regression. Moreover, in many economic applications, tail quantiles have highly valuable information. See, for example, wage distribution in labor economic applications (Buchinsky, 1998) and stock return quantiles (Value-at-Risk) in financial market analysis (Duffie and Pan, 1997). Recently, policymakers have begun to pay attention to the left tail quantiles of GDP growth (Growth-at-Risk) as a measure of downside risks associated with tight financial conditions (Adrian, Boyarchenko, and Giannone, 2019). There has also been an increasing interest in climate change, in particular, more frequent and intense extreme weather conditions. A tail quantile is the main object of interest

We would like to thank the Editor, Peter Phillips, the Co-Editor, Arthur Lewbel, and three anonymous referees for helpful comments and suggestions, which have led to substantial improvements. We would also like to thank Xun Lu and Liangjun Su for helpful discussion and sharing their codes. Shin is grateful for partial support by the Social Sciences and Humanities Research Council of Canada (SSHRC-435-2018-0275). This work was made possible by the facilities of WestGrid (www.westgrid.ca) and Compute Canada (www.compute-canada.ca).

© The Author(s), 2021. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

in this analysis (Bhatia et al., 2019). Estimation, inference, and prediction of the conditional quantiles are thus important but require a careful econometric analysis due to their nonlinear structure and nonstandard limit theory.

In this paper, we propose a novel prediction method based on complete subset averaging (CSA) for QRs. Following Lu and Su (2015), we work on the framework such that all models under consideration are potentially misspecified and that the dimension of regressors goes to infinity as the sample size increases. The CSA method that we propose works as follows. First, pick the numbers of regressors k out of all regressors K available in the data. Then, there exist $K!/(k!(K-k)!)$ complete subsets of size k . Second, estimate all the QR models and save all the conditional quantile predictors from each model. Finally, the conditional quantile predictor is constructed as the average of all the quantile predictors estimated in Step 2. Since we average over the complete subsets, the number of models is much larger than the usual model averaging methods selecting the weight of each model. We propose to use an equal weight but select the optimal size of the complete subset k^* based on the leave-one-out cross-validation method.

The CSA approach has a couple of advantages over the existing model averaging method which adopts sophisticated weighting schemes. First, it may produce better forecasts in practice, because there is no sampling variance from the weight estimation. This result is already reported both in the forecasting and machine learning literature in the mean regression setup (see, e.g., Clemen, 1989; Breiman, 1996; Stock and Watson, 2004; Smith and Wallis, 2009; Elliott, Gargano, and Timmermann, 2013). Second, it does not ask a researcher to choose the initial set of models and the order of each model. In practice, the model averaging methods with different weights usually construct the set of models in an encompassing way, and the forecasting performance could depend on the researcher's discretion. Third, CSA averages over a larger number of submodels, and one could expect an additional noise reduction from it. However, CSA is possibly more demanding in computation, and we will discuss this issue in detail later.

The contribution of this paper is twofold. First, building upon the theory of Lu and Su (2015), we show that the complete subset quantile regression (CSQR) estimator converges the pseudo-true value and satisfies asymptotic normality under mild regularity conditions. The uniform convergence property of CSQR is also provided. Based on these pointwise and uniform limit theories, we prove the asymptotic optimality of \hat{k} in the sense of Li (1987). Second, we implement the CSA method and show that it performs quite well both in simulations and real datasets. Especially, we show that the performance is still satisfactory when we use a fixed number of subsets randomly drawn from the complete subsets when the time budget does not allow estimating the QRs of the whole subsets. We also provide regularity conditions on the choice of the fixed number of subsets. Finally, we provide a theory that compares the performance of equal weighting and optimal weighting in QR. This result justifies our intuition such that optimal weighting forecasts poorly when the number of models increases and extends the existing result in mean regression.

Finally, we summarize related literature. Lu and Su (2015) and Elliott et al. (2013) are closely related to this paper. The former proposes the jackknife model averaging (JMA) method for the quantile prediction problem and derives the nonstandard asymptotic properties of the estimator. Our approach is different from theirs in that we use complete subsets for models to be averaged and that we choose a scalar \widehat{k} from the cross-validation method instead of a weighting vector \widehat{w} . The latter proposes the CSA method in the mean prediction problem and shows by simulation studies that the CSA predictor outperforms alternative methods like bagging, ridge, lasso, and Bayesian model averaging. However, they do not show any optimality result of the estimator. Hansen (2007) and Hansen and Racine (2012) show the optimality of model averaging based on the Mallows criterion and that of the JMA, respectively. Ando and Li (2014) propose a model averaging method in a high-dimensional setting and show the optimality result. Komunjer (2013) provides a great review on the quantile prediction problem of time-series data. Meinshausen (2006) proposes a quantile prediction method based on random forest. Lee (2016) studies the inference problem of the predictive QR when the regressors are persistent. In the empirical finance literature, Meligkotsidou et al. (2019, 2021) apply CSQR to forecast realized volatility and the risk premium.

The rest of the paper is organized as follows. Section 2 introduces the model and the CSQR estimator. Section 3 presents the asymptotic properties of the CSQR estimator and the asymptotic optimality. The Monte Carlo simulation results are reported in Section 4. Section 5 investigates two empirical applications and illustrates the advantage of the proposed method. Section 6 concludes. All the proofs are deferred to the Appendix.

We use the following notation. For a matrix A , $\|\cdot\|$ represents its Frobenius norm $\|A\| = \sqrt{\text{tr}(AA')}$. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues of A . We use the notation $x_n \approx y_n$ to denote $x_n = y_n + o_p(1)$, and $a_n \ll b_n$ to denote $a_n = o(b_n)$.

2. MODEL AND ESTIMATOR

In this section, we lay out the model under study and propose the CSA quantile predictor. We also discuss the choice of the subset size based on the cross-validation method.

2.1. CSA Quantile Predictor

Consider a random sample $\{(y_i, x'_i)\}$, for $i = 1, \dots, n$, where the dimension of x_i can be countably infinite. Following Lu and Su (2015), we assume that $\{(y_i, x'_i)\}_{i=1}^n$ is generated from the following linear QR model: for $\tau \in (0, 1)$,

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

where $\mu_i = \mu_i(\tau) := \sum_{j=1}^{\infty} \theta_j x_{ij}$, $\theta_j = \theta_j(\tau)$, $\varepsilon_i = \varepsilon_i(\tau) := y_i - Q_y(\tau|x_i)$, and $Q_y(\tau|x)$ is the τ th conditional quantile function of y given x . Note that we drop τ from each expression for notational simplicity and that ε_i satisfies the quantile restriction $P(\varepsilon_i(\tau) \leq 0|x_i) = \tau$. Equivalently, we can also express $Q_y(\tau|x_i) := \sum_{j=1}^{\infty} \theta_j(\tau)x_{ij}$ as is often done in the QR literature.

We consider a sequence of covariates available, which approximate the above QR model:

$$y_i = \sum_{j=1}^{K_n} \theta_j(\tau)x_{ij} + b_i(\tau) + \varepsilon_i(\tau),$$

where $b_i = b_i(\tau) := \mu_i(\tau) - \sum_{j=1}^{K_n} \theta_j(\tau)x_{ij}$ is the approximation error and K_n is the total number of available regressors that may increase as the sample size n increases. Thus, we presume that all models are misspecified in a finite sample as in Hansen (2007).¹

Given K_n regressors, we consider a model composed of k regressors, where $k \in \{1, 2, \dots, K_n\}$. There are $\frac{K_n!}{k!(K_n-k)!}$ different ways to select k regressors out of K_n . Therefore, a subset of size k is composed of $M_{(K_n,k)} = \frac{K_n!}{k!(K_n-k)!}$ different elements, and a model is defined as a single element of them. We use index $m_{(K_n,k)} \in \{1, 2, \dots, M_{(K_n,k)}\}$ for each model. For example, consider that we have $K_n = 3$ regressors $\{x_{i1}, x_{i2}, x_{i3}\}$ and construct a subset of size $k = 2$. Then, we have $M_{(3,2)} = 3$ different ways to choose a model as follows: (x_{i1}, x_{i2}) , (x_{i1}, x_{i3}) , and (x_{i2}, x_{i3}) . Each model is indexed by $m_{(3,2)} \in \{1, 2, 3\}$. For succinct notation, we drop all subscripts from K_n , $M_{(K_n,k)}$, and $m_{(K_n,k)}$, and denote them as K , M , and m unless there is any confusion.

We now consider a QR model with regressors in a complete subset. Let model m with a size k be given. For observation i , let $x_{i(m,k)}$ be a k -dimensional vector of regressors corresponding to model m , i.e., $x_{i(2,2)} = (x_{i1}, x_{i3})$ in the above example. We can construct a linear QR model with regressors $x_{i(m,k)}$:

$$y_i = x'_{i(m,k)} \Theta_{(m,k)} + b_{i(m,k)} + \varepsilon_i, \tag{2}$$

where $b_{i(m,k)} := \mu_i - x'_{i(m,k)} \Theta_{(m,k)}$ is again the approximation error when we use only $x_{i(m,k)}$ regressors. The model in (2) is estimated by the standard method in linear QR:

$$\widehat{\Theta}_{(m,k)} = \arg \min_{\Theta_{(m,k)} \in \Theta} \sum_{i=1}^n \rho_{\tau} (y_i - x'_{i(m,k)} \Theta_{(m,k)}) \tag{3}$$

$$:= \arg \min_{\Theta_{(m,k)} \in \Theta} Q_n (\Theta_{(m,k)}), \tag{4}$$

¹Using quantile crossings, Phillips (2015) also shows that QR is always at the risk of model misspecification unless the parameters are local to constant over τ .

where Θ is a parameter space and $\rho_\tau(u) := u(\tau - 1\{u \leq 0\})$ is the check function. Note that the estimator $\widehat{\Theta}_{(m,k)}$ is defined for each subset size k and for each model m with k regressors. As noted above, we can think of M different models and corresponding estimators that have k regressors.

We have a few remarks here. First, we use the subscript (m, k) to denote a generic model with k regressors. However, the index set $\{1, \dots, M_{(K_n, k)}\}$ itself is defined in terms of k , which implies that m is also determined by k . Recall the original notation $m_{(K_n, k)}$ above. Therefore, model $m \in \{1, \dots, M_{(K_n, k)}\}$ has the same number of regressors k , and we cannot choose m and k in an arbitrary way. Second, we allow that the subset size k goes to infinity as n increases. In other words, there exists a sequence of subset sizes $\{k(n)\}$ that diverges. This setting is natural as the upper bound K_n goes to infinity as n increases. Note that the number of regressors in each model (k_m in their notation) is also allowed to diverge in Lu and Su (2015). Both approaches allow more complex models to be averaged as n grows, which is measured by k and k_m , respectively. However, Lu and Su (2015) require controlling the growth rates of M and $\max_m k_m$, separately. The proposed method constructs submodels based on the complete subsets, and M is tightly related to K and k . As a result, the regularity condition on the complexity of the models is expressed only in terms of K_n (see Assumption 3 in Section 3).

We finalize this subsection by defining the CSA quantile predictor. Let the size of the complete subset k be given. For each model, we estimate the parameter $\widehat{\Theta}_{(m,k)}$ by (3) and construct the linear index $x'_{(m,k)} \widehat{\Theta}_{(m,k)}$. The CSA quantile predictor of y given x is defined as a simple average of those indices over M different models:

$$\widehat{y}(k) = \frac{1}{M} \sum_{m=1}^M x'_{(m,k)} \widehat{\Theta}_{(m,k)}.$$

The CSA quantile predictor is different from the JMA quantile predictor of Lu and Su (2015) in two respects. First, we do not select the set of models to be averaged, since we average over the complete subsets of size k . Second, CSA does not estimate the weights over different models. The idea of averaging over the complete subsets was first introduced by Elliott et al. (2013) in the conditional mean prediction setup. Heuristically speaking, since the weights can be seen as additional parameters to be estimated in the model, the equal weight could perform better in a finite sample when the number of models (i.e., the dimension of a weight vector) is large.

2.2. Choice of Subset Size k

We propose to choose the subset size k using the leave-one-out cross-validation method. We will show in the next section that the subset size \widehat{k} chosen by this method is optimal in the sense that it is asymptotically equivalent to the infeasible optimal choice.

For $k = 1, \dots, K$, we define a cross-validation objective function as follows:

$$CV_n(k) = \frac{1}{n} \sum_{i=1}^n \rho_\tau \left(y_i - \frac{1}{M} \sum_{m=1}^M x'_{i(m,k)} \widehat{\Theta}_{i(m,k)} \right) \tag{5}$$

$$= \frac{1}{n} \sum_{i=1}^n \rho_\tau (y_i - \widehat{y}_i(k)), \tag{6}$$

where $\widehat{\Theta}_{i(m,k)}$ is the jackknife estimator for $\widehat{\Theta}_{(m,k)}$, which is estimated by (3) without using the i th observation (x_i, y_i) , and $\widehat{y}_i(k)$ is a corresponding jackknife CSA quantile predictor for the i th outcome variable y_i . The prediction error is measured by the check function $\rho_\tau(\cdot)$. Then, we can choose the complete subset size k that minimizes the cross-validation objective function as follows:

$$\widehat{k} = \arg \min_{1 \leq k \leq K} CV_n(k). \tag{7}$$

After choosing the complete subset size, the CSA quantile predictor is finally defined as

$$\widehat{y}(\widehat{k}) = \frac{1}{M} \sum_{m=1}^M x'_{(m,\widehat{k})} \widehat{\Theta}_{(m,\widehat{k})}, \tag{8}$$

where the plugged-in \widehat{k} is chosen by (7).

We finalize this subsection by adding some remarks on computation. First, we propose to use a fixed number M_{max} of random draws of models when M is too large to implement the method. Since $M = K! / (k!(K - k)!)$, it can be quite large when the model has large potential regressors. The simulation studies in Section 4 reveal that the CSA quantile predictor still performs well with a feasible size of submodels randomly drawn from the complete subsets. We also provide regularity conditions that assure the asymptotic equivalence between using M and M_{max} in Section 3. Second, the proposed jackknife method can be immediately extended to the b -fold cross-validation method, where b is the partition size of the sample. Algorithm 1 below summarizes the leave-one-out cross-validation method for choosing \widehat{k} .

3. ASYMPTOTIC THEORY

In this section, we investigate the asymptotic properties of the CSQR estimator. We first provide the pointwise and uniform convergence results of $\widehat{\Theta}_{(m,k)}$ and $\widehat{\Theta}_{i(m,k)}$, respectively. Then, we show the optimality of CSA in the sense of Li (1987), which implies that \widehat{k} is asymptotically equivalent to the infeasible optimal choice of the subset size.

In addition to the model described in Section 2, we define some notation for latter use. Let $f_{y|x}(\cdot|x)$ be a conditional probability density function for generic random variables x and y . Since all models are potentially misspecified in the

Algorithm 1: Cross-validation for CSA

Input: $\{(y_i, x_i) : i = 1, \dots, n\}, M_{max}$

Output: \hat{k}

```

1 Set  $K = \dim(x_i)$ ;
2 for  $k = 1$  to  $K$  do
3   Set  $\mathcal{X}_{i,k} = \{\text{all combinations with } k \text{ regressors out of } x_i\}$ ;
4   Set  $M = |\mathcal{X}_{i,k}|_0 = K! / (k!(K-k)!)$ ;
5   if  $M \leq M_{max}$  then
6     for  $m = 1$  to  $M$  do
7       Set  $x_{i(m,k)} = (\text{the } m\text{th element of } \mathcal{X}_{i,k})$  for  $i = 1, \dots, n$ ;
8       for  $i = 1$  to  $n$  do
9         Estimate the jackknife estimator  $\hat{\Theta}_{i(m,k)}$ :
          
$$\hat{\Theta}_{i(m,k)} = \arg \min_{\Theta_{(m,k)} \in \Theta} \sum_{j=1, j \neq i}^n \rho_\tau(y_j - x'_{j(m,k)} \Theta_{(m,k)}) \tag{9}$$

10        end
11      end
12      Set  $\hat{y}_i(k) = \frac{1}{M} \sum_{m=1}^M x'_{i(m,k)} \hat{\Theta}_{i(m,k)}$ ;
13      Set  $CV_n(k) = \frac{1}{n} \sum_{i=1}^n \rho_\tau(y_i - \hat{y}_i(k))$ ;
14    end
15    if  $M > M_{max}$  then
16      for  $m = 1$  to  $M_{max}$  do
17        Set  $x_{i(m,k)} = (\text{a random element of } \mathcal{X}_{i,k})$  for  $i = 1, \dots, n$ ;
18        for  $i = 1$  to  $n$  do
19          Estimate the jackknife estimator  $\hat{\Theta}_{i(m,k)}$  using (9);
20        end
21      end
22      Set  $\hat{y}_i(k) = \frac{1}{M} \sum_{m=1}^M x'_{i(m,k)} \hat{\Theta}_{i(m,k)}$ ;
23      Set  $CV_n(k) = \frac{1}{n} \sum_{i=1}^n \rho_\tau(y_i - \hat{y}_i(k))$ ;
24    end
25  end
26 Set  $\hat{k} = \arg \min_k CV_n(k)$ ;

```

model averaging literature, we define the pseudo-true parameter value for any given (m, k) :

$$\Theta_{(m,k)}^* := \arg \min_{\Theta_{(m,k)} \in \Theta} E[\rho_\tau(y_i - x'_{i(m,k)} \Theta_{(m,k)})].$$

Let $\psi_\tau(c) := \tau - 1\{c \leq 0\}$. For any (m, k) such that $m = 1, \dots, M$ and $k = 1, \dots, K$, we define

$$A_{(m,k)} := E \left[f_{y|x} \left(\Theta'_{(m,k)} x_{i(m,k)} \right) x_{i(m,k)} x'_{i(m,k)} \right],$$

$$B_{(m,k)} := E \left[\psi_{\tau} \left(y_i - \Theta'_{(m,k)} x_{i(m,k)} \right)^2 x_{i(m,k)} x'_{i(m,k)} \right],$$

and

$$V_{(m,k)} := A_{(m,k)}^{-1} B_{(m,k)} A_{(m,k)}^{-1}.$$

We need the following regularity conditions.

Assumption 1.

- (i) (y_i, x_i) is i.i.d. generated by (1).
- (ii) $P(\varepsilon_i(\tau) \leq 0 | x_i) = \tau$ a.s.
- (iii) $E[\mu_i^4] < \infty$ and $\sup_{j \geq 1} E[x_{ij}^8] < c_x$, for some $c_x < \infty$.

Assumption 2.

- (i) $f_{y|x}(\cdot | x_i)$ is bounded above by $c_f < \infty$ and continuous over its support a.s.
- (ii) There exist constants $\underline{c}_{A(m,k)}$ and $\bar{c}_{A(m,k)}$ such that $0 < \underline{c}_{A(m,k)} \leq \lambda_{\min}(A_{(m,k)}) \leq \lambda_{\max}(A_{(m,k)}) \leq c_f \lambda_{\max}(E[x_{i(m,k)} x'_{i(m,k)}]) \leq \bar{c}_{A(m,k)} < \infty$.
- (iii) There exist constants $\underline{c}_{B(m,k)}$ and $\bar{c}_{B(m,k)}$ such that $0 < \underline{c}_{B(m,k)} \leq \lambda_{\min}(B_{(m,k)}) \leq \lambda_{\max}(B_{(m,k)}) \leq \bar{c}_{B(m,k)} < \infty$.
- (iv) $(\bar{c}_{A(m,k)} + \bar{c}_{B(m,k)})/k = O(\underline{c}_{A(m,k)}^2)$.

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

- (i) $\frac{K^4 \bar{c}_A}{n \underline{c}_B} = o(1)$ and $\frac{K^4 (\log n)^4}{n \underline{c}_B^2} = o(1)$.
- (ii) $\frac{K}{\log n} = O(1)$ and $(\log n)^{K+1} n^{-K \bar{c}_A^3 / (\bar{c}_A \bar{c}_B)} = o(1)$.

Conditions (i) and (ii) in Assumption 1 are the standard *i.i.d.* and the quantile restrictions. Assumption 1(iii) requires some finite moment restrictions to achieve the probability bounds of various sample mean objects in the proof. Assumption 2 allows conditional heteroskedasticity. Note that the eigenvalues of $A_{(m,k)}$ and $B_{(m,k)}$ are bounded and bounded away from zero for a given (m, k) . However, these bounds $(\underline{c}_{A(m,k)}, \underline{c}_{B(m,k)}, \bar{c}_{A(m,k)}, \bar{c}_{B(m,k)})$ can converge to zero or diverge to infinity as n increases. The speed of convergence is restricted by Assumption 2(iv). These bounded eigenvalue restrictions are commonly imposed in the literature that studies the increasing dimension of parameters (see, e.g., Portnoy, 1984, 1985). Assumptions 1 and 2 are standard and similar to those in Lu and Su (2015). See the additional remarks therein. Assumption 3 imposes some regularity conditions on the number of potential regressors K_n and the sequence of the uniform bounds $(\underline{c}_A, \underline{c}_B, \bar{c}_A, \bar{c}_B)$. Different from the regularity condition of JMA in Lu and Su (2015), we need not restrict the growth rate of potential models M directly, since $M_{(K_n, k)}$ is determined by K_n . However, $M_{(K_n, k)}$ increases very quickly at a factorial rate of K_n , and we need a stronger restriction on K_n . As noted in Assumption 3(ii), K_n can increase at most the logarithmic rate of n . In the case of JMA, the number

of regressors can increase at the polynomial rate if we set $\bar{k} = k_M = M$ in their notation. This is a trade-off in proving the uniform convergence results over a larger index set than that of JMA. We discuss this point in detail below in Theorem 2. The second part of Assumption 3(ii) holds if $\underline{c}_A^3/\bar{c}_A\bar{c}_B$ is bounded away from zero or converges to zero at the slower rate than $\log(\log n)/\log n$ when K increases at the rate of $\log n$.

First, we prove the convergence rate and the asymptotic normality of $\widehat{\Theta}_{(m,k)}$ when the dimension of parameter k increases.

THEOREM 1. *Suppose that Assumptions 1, 2, and 3(i) hold. Let $C_{(m,k)}$ denote an $l_{(m,k)} \times k$ matrix such that $C_0 := \lim_{n \rightarrow \infty} C_{(m,k)}C'_{(m,k)}$ exists and is positive definite, where $l_{(m,k)} \in [1, k]$ is a fixed integer. Then,*

- (i) $\left\| \widehat{\Theta}_{(m,k)} - \Theta_{(m,k)}^* \right\| = O_p\left(\sqrt{\frac{k}{n}}\right),$
- (ii) $\sqrt{n}C_{(m,k)}V_{(m,k)}^{-1/2}\left[\widehat{\Theta}_{(m,k)} - \Theta_{(m,k)}^*\right] \xrightarrow{d} N(0, C_0).$

This theorem provides an asymptotic theory for the QR estimator when the model is misspecified and the number of parameters diverges to infinity as similarly seen in Lu and Su (2015). The convergence rate in (i) is a standard result when k diverges as n increases. To show the asymptotic normality with a diverging number of parameters, we also consider an arbitrary linear combination of $\widehat{\Theta}_{(m,k)}$ represented by $C_{(m,k)}$. The difference between two estimators, CSA and JMA, originates from the fact that CSA chooses the total number of the regressors K_n first and the number of complete subset models $M_{(K_n,k)}$ follows automatically for each $k = 1, \dots, K_n$, whereas JMA selects the set of models M_n (in their notation) in advance. Then, the size of regressors k_m in case of JMA is determined by the sequence of models $m = 1, \dots, M_n$ chosen by a researcher. Although there are slight differences in the definition of $c_{A(m,k)}$ and $c_{B(m,k)}$ and their bounds from those in Lu and Su (2015), the proof of Theorem 1 is identical to theirs, so is omitted.

We next turn our attention to the uniform convergence results of $\widehat{\Theta}_{i(m,k)}$ and $\widehat{\Theta}_{(m,k)}$. In addition to its own interest, the uniform convergence rates in the next theorem are required to prove to the asymptotic optimality of \widehat{k} .

THEOREM 2. *Suppose that Assumptions 1, 2, and 3(ii) hold. Then,*

- (i) $\max_{1 \leq i \leq n} \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \left\| \widehat{\Theta}_{i(m,k)} - \Theta_{(m,k)}^* \right\| = O_p\left(\sqrt{n^{-1}K \log n}\right),$
- (ii) $\max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \left\| \widehat{\Theta}_{(m,k)} - \Theta_{(m,k)}^* \right\| = O_p\left(\sqrt{n^{-1}K \log n}\right).$

Since CSA is defined on the index sets of m and k , the uniform convergence rates are defined over those sets, $m \in \{1, \dots, M\}$ and $k \in \{1, \dots, K\}$. In case of $\widehat{\Theta}_{i(m,k)}$, we need additional uniformity over $i \in \{1, \dots, n\}$. As a result, the regularity conditions that control the growth rates of K_n and $M_{(K_n,k)}$ are different from those of JMA in Assumption 3(ii). As discussed before, since the number of complete

subsets increases at the factorial rate of K_n , we need a restriction on K_n slightly stronger than that of JMA. We follow the proof strategy in Lu and Su (2015) which extends the results of Rice (1984) by using the inequality in Shibata (1981, 1982). To handle the different growth rates, we provide new technical lemmas. The proof of Theorem 2 as well as these lemmas are provided in the Appendix. Finally, the uniform convergence rates are expressed in terms of the sample size n and the total number of regressors K that goes to infinity as n increases.

We next prove the prediction equivalence when we replace M with M_{max} . Let \mathcal{M}_{max} be a subset of $\{1, \dots, M\}$ such that M_{max} elements are randomly drawn. Define $\tilde{y}(k)$ to be the CSA quantile predictor using only M_{max} models:

$$\tilde{y}(k) := \frac{1}{M_{max}} \sum_{m' \in \mathcal{M}_{max}} x'_{(m',k)} \hat{\Theta}_{(m',k)}.$$

Let $y_k^* := \lim_{M \rightarrow \infty} M^{-1} \sum_{m=1}^M E[x'_{(m,k)} \Theta_{(m,k)}^*] < \infty$. We show the validity of M_{max} in the following theorem.

THEOREM 3. *Suppose that Assumptions 1–3 hold. Let $M_{max} \rightarrow \infty$ and $K/M_{max} \rightarrow 0$ as $n \rightarrow \infty$. Furthermore, we assume that*

$$\max_{1 \leq k \leq K} M^{-1} \sum_{m=1}^M x'_{(m,k)} \Theta_{(m,k)}^* - y_k^* = o_p(1),$$

$$\max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \|x_{(m,k)}\| = O_p(1).$$

Then, we have

$$\max_{1 \leq k \leq K} |\hat{y}(k) - \tilde{y}(k)| = o_p(1).$$

The rate requirement for M_{max} is mild, and $M_{max} = O(n^{1/2})$ would work given $K = O(\log n)$. The uniform boundedness assumption on $\|x_{(m,k)}\|$ is weak and holds easily in most applications. We have some remarks on the uniform convergence assumption of the model average with the pseudo-true parameter $\Theta_{(m,k)}^*$. Let $z_{(m,k)} = x'_{(m,k)} \Theta_{(m,k)}^* - E[x'_{(m,k)} \Theta_{(m,k)}^*]$. Note that k is discrete and the functional class size over k is small. Thus, it depends on the dependent structure of $z_{(m,k)}$ to hold the uniform law of large numbers. For example, consider the following maximal inequality: for $\delta > 0$,

$$\begin{aligned} P\left(\max_{1 \leq k \leq K} \left| M^{-1} \sum_{m=1}^M z_{(m,k)} \right| > \delta\right) &\leq K \max_{1 \leq k \leq K} P\left(\left| M^{-1} \sum_{m=1}^M z_{(m,k)} \right| > \delta\right) \\ &\leq \frac{K}{M} \max_{1 \leq k \leq K} \frac{E[\sum_{m=1}^M z_{m,k}]^2}{M\delta^2}, \end{aligned}$$

where the second line holds from the Markov inequality. Since $K/M = o(1)$, a sufficient condition for the uniform convergence is $\max_{1 \leq k \leq K} E[\sum_{m=1}^M z_{m,k}]^2/M = O(1)$. If $z_{(m,k)}$ is covariance stationary over m for all k , then the sufficient condition

becomes the absolute summability condition $\max_{1 \leq k \leq K} \sum_{j=0}^{\infty} |E[z_{(m,k)} z_{(m+j,k)}]| < \infty$. See, e.g., Fazekas and Klesov (2001) for more general conditions on the partial sums in a different dependent structure.

We now prove the asymptotic optimality of \widehat{k} in the sense of Li (1987). Following Lu and Su (2015), we use the final prediction error (FPE, or the out-of-sample quantile prediction error) as a criterion to evaluate the prediction performance:

$$FPE_n(k) := E \left[\rho_{\tau} \left(y - \frac{1}{M} \sum_{m=1}^M X'_{(m,k)} \widehat{\Theta}_{(m,k)} \right) | \mathcal{D}_n \right],$$

where $\mathcal{D}_n := \{(y_i, x_i) : i = 1, \dots, n\}$ is a sample. The next theorem shows that \widehat{k} is asymptotically equivalent to the infeasible best subset size choice that is defined as a minimizer of $FPE(k)$.

THEOREM 4. *Suppose that Assumptions 1–3 hold. Then,*

$$\frac{FPE(\widehat{k})}{\inf_{k \in \mathcal{K}} FPE(k)} \xrightarrow{p} 1,$$

where $\mathcal{K} := \{1, \dots, K_n\}$.

A similar optimality concept has been adopted in the context of the weighted average estimator (e.g., Hansen, 2007; Hansen and Racine, 2012; Lu and Su, 2015) and in the context of the instrumental variable (IV) estimator (e.g., Donald and Newey, 2001; Kuersteiner and Okui, 2010; Lee and Shin, 2020). Different from JMA, CSA considers the complete subsets given (K_n, k) and does not require the preselection of models to be considered nor the order of models. Thus, the optimality result is also independent of the initial model selection/ordering issue once the total number of regressors is given. The index set \mathcal{K} of CSA is discrete while that of JMA or the JMA in Hansen and Racine (2012) is compact. All require the finite moment condition similar to Assumption (A.1) in Li (1987), which is assured by Assumption 1(iii) above. The idea of CSA has been adopted in the forecasting literature (e.g., Rapach, Strauss, and Zhou, 2010; Elliott et al., 2013, 2015). This is the first formal result to show the optimality of the subset size selection.

Finally, we compare the performance of the nonstochastic equal weight with that of the optimal weight. In the mean prediction context, it has been observed that a simple arithmetic mean, i.e., the equal weight, outperforms the *estimated* optimal weight. This empirical phenomenon is known as the “forecast combination puzzle,” and some formal explanations under the mean squared error are provided by Smith and Wallis (2009), Elliott (2011), and Claeskens et al. (2016), to name a few. Heuristically speaking, it happens when the estimation error of the optimal weight is large enough to dominate the efficiency loss caused by the equal weight. We extend this result to the class of smooth expected loss functions. This is crucial

in our analysis, since the check function $\rho_\tau(\cdot)$ does not give a closed-form solution, which is different from the mean squared error used in the existing literature.

We consider the following simplified framework to focus on the main idea. Let be $\hat{y}_1, \dots, \hat{y}_M$ be predictors for y based on M different models. For example, we can think of $\hat{y}_m = X'_{(m,k)} \hat{\Theta}_{(m,k)}$ for any given k . Let w be an M -dimensional weight vector combining the M predictors. We consider only positive weights with $1'_M w = 1$, where 1_M is an M -dimensional unit vector. Let $\hat{y} := (\hat{y}_1, \dots, \hat{y}_M)'$ and $e_m := y - \hat{y}_m$ be the prediction error of \hat{y}_m and $e := (e_1, \dots, e_M)'$ be a vector of these prediction errors. We define the prediction error of the combined predictor as $e_c(w) := y - w'\hat{y} = w'(1_M \cdot y - \hat{y}) = w'e$. Then, we can define an optimal weight w^* as

$$w^* = \arg \min_{w \in \Delta^{M-1}} F(w),$$

where Δ^{M-1} is the standard $(M - 1)$ -simplex and $F(w) := E[L(w; e_c)]$ is an expected loss function. For example, the mean squared error in Elliott (2011) can be written in terms of the quadratic loss function: $F(w) = E[e_c^2] = E[w'ee'w] = w'\Sigma w$, where $\Sigma = E[ee']$. The quantile prediction error adopted in this paper can be written in terms of the check function: $F(w) = E[\rho_\tau(e_c)] = E[\rho_\tau(y - w'\hat{y})]$. Let $\bar{w} := M^{-1}1_M$ be an equal-weight vector and \hat{w} be an estimator for w^* with $\hat{\eta} := \hat{w} - w^*$. To illustrate our main point, we further impose that $E[\hat{\eta}] = 0$ and $\max_m \text{Var}(\hat{\eta}_m) = \bar{\sigma}_\eta^2 > 0$.

THEOREM 5. *Suppose that $F(w)$ is twice differentiable on Δ^{M-1} and that $\sup_{w \in \Delta^{M-1}} \|\nabla_2 F(w)\| \leq C < \infty$ uniformly in M . Let $\bar{\lambda}_{max} := \limsup_M \sup_w \lambda_{max}(\nabla_2 F(w))$.*

- (i) $|F(\bar{w}) - F(w^*)| \leq 2^{-1} \bar{\lambda}_{max} (1 + 3M^{-1})$.
- (ii) $E|F(\hat{w}) - F(w^*)| \leq 2^{-1} \bar{\lambda}_{max} M \bar{\sigma}_\eta^2$.

We have some remarks. First, it shows that the equal weight \bar{w} may work better than the estimated optimal weight \hat{w} when we average many models, i.e., when M is large. Compared to the optimal prediction error $F(w^*)$, the efficiency loss by \bar{w} is bounded by $2^{-1} \bar{\lambda}_{max} (1 + 3M^{-1})$, which converges to $2^{-1} \bar{\lambda}_{max}$ for large enough M . On the contrary, the upper bound of the mean efficiency loss by \hat{w} diverges as M increases. We admit that these upper bounds only reflect the worst case scenario. However, it confirms the intuition formally that the equal weight can outperform the *estimated* optimal weight under the class of smooth expected loss functions. Second, the prediction error of \bar{w} under a quadratic loss function converges to the optimal prediction error as M increases. The same result is also proved in Proposition 1 in Elliott (2011). Different from his result, it does not require decomposing the prediction error into the common component and the idiosyncratic component. This result is summarized in Corollary 6 below. Third, to achieve the optimality, the estimation errors of the weight \hat{w} , $\{\hat{\eta}_m\}$, should

vanish fast enough. Let $\bar{\sigma}_\eta^2 = O(c_n)$. A sufficient condition for the optimality is $c_n = o(M_n^{-1})$. For example, if c_n is a parametric rate, $n^{-1/2}$, then M_n should be bounded by $o(n^{1/2})$. When $M_n = O(\log n)$, for example, this condition is satisfied. However, if M_n increases too fast, then $M_n \bar{\sigma}_\eta^2$ will diverge and \hat{w} may work worse than \bar{w} .² Fourth, $\|\nabla_2 F(w)\| = (\sum_{m=1}^M \lambda_m^2)^{1/2}$, where $\{\lambda_m\}$ are eigenvalues of $\nabla_2 F(w)$, since $\nabla_2 F(w)$ is symmetric. Thus, the uniform bound C exists if $\{\lambda_m\}$ is absolutely summable, $\sum_{m=1}^\infty |\lambda_m| < \infty$. Fifth, if we restrict our attention to the expected check function adopted in this paper, $F(w)$ is twice differentiable if the conditional density $f(y|\hat{y})$ is smooth for all \hat{y} . From Theorem 1 in Angrist, Chernozhukov, and Fernández-Val (2006), we have

$$F(w) = E[\bar{\omega}_\tau(\hat{y}, w)(w'\hat{y} - Q_\tau(y|\hat{y}))^2], \tag{10}$$

where $Q_\tau(y|\hat{y})$ is the conditional quantile function of y given \hat{y} and $\bar{\omega}_\tau(\hat{y}, w) := \int_0^1 (1-u) \cdot f(u \cdot w'\hat{y} + (1-u) \cdot Q_\tau(y|\hat{y})|\hat{y}) du$. Thus, the smoothness of $F(w)$ is implied by the twice differentiability of $f(y|\hat{y})$. Finally, equation (10) shows that CSA would not work well if we include many irrelevant models. Similar to the QR specification error in Angrist et al. (2006), we call $(w'\hat{y} - Q_\tau(y|\hat{y}))$ the quantile prediction specification error. If there are many irrelevant models, the optimal weight ω^* would be sparse, i.e., many elements of ω^* would be zeros. In such a case, CSA with $\bar{\omega} = M^{-1}1_M$ results in a larger quantile prediction specification error given M and n . For example, if there is only one relevant regressor and all other coefficients $\theta_j(\tau)$ equal zero besides one, the complete subsets will be composed of many irrelevant models. As we will see in the simulations studies in the next section, CSA does not perform well under this situation. Therefore, a prescreening process is desirable to achieve a satisfactory result of CSA.

COROLLARY 6. *Suppose that we have a quadratic loss function, $L(w; e_c) = (e_c(w))^2$, and that $\lambda_{\max}(\Sigma) < \infty$ uniformly in M , where $\Sigma := E[ee']$. Then, we have*

$$|F(\bar{w}) - F(w^*)| \rightarrow 0 \text{ as } M \rightarrow \infty.$$

4. MONTE CARLO SIMULATIONS

In this section, we investigate the finite sample performance of the proposed estimator in simple Monte Carlo experiments. We consider two categories of the simulation designs: (i) all candidate models are misspecified, and (ii) candidate models include the true model.

First, we adopt the following data generating process (DGP):

$$y_i = \theta \sum_{j=1}^{1000} j^{-1} x_{ij} + \varepsilon_i, \tag{11}$$

²We thank an anonymous referee and the Co-Editor for pointing out this intuition. Furthermore, note that it is one sufficient condition. It is still possible that there exists a different set of conditions that guarantee the optimality.

where $x_{i1} = 1$ and $(x_{i2}, \dots, x_{i1000})$ follows a multivariate normal distribution, $N(0, \Sigma)$ with $\Sigma_{jk} = \rho_x$ if $j \neq k$ and 1 if $j = k$. Therefore, the regressors are possibly dependent on each other, which is a more general feature of the design than the existing literature (see, e.g., Hansen, 2007; Lu and Su, 2015). The term ε_i follows $N(0, 1)$ independent of x_{ij} . The sample is *i.i.d.* over i . The population $R^2 := (\text{Var}(y_i) - \text{Var}(\varepsilon_i))/\text{Var}(y_i)$ is controlled by θ . We consider two sample sizes, $n = 50, 150$. The number of potential regressors is set to $K = 4 \log(n)$, which is 15 and 20, respectively. Note that all candidate models are misspecified, since there remain many missing regressors in the sample. We consider various DGPs by combining different $R^2 = \{0.1, \dots, 0.9\}$, $\tau = \{0.1, \dots, 0.9\}$, and $\rho_x = \{0.0, 0.1, 0.2, \dots, 0.9\}$. We consider 38 different DGPs in total and estimate 74 different quantile models.

We compare the performance of the proposed CSA estimator with the JMA estimator in Lu and Su (2015), the ℓ_1 -penalized quantile regression (L1QR) in Belloni and Chernozhukov (2011), the bootstrap aggregating (BAG) methods in Breiman (1996), and ℓ_2 -penalized quantile regression (L2QR). L1QR and L2QR are also called the lasso and the ridge regression in the mean regression setup. The set of models used for JMA is constructed in an encompassing way, e.g., $\{x_{i1}\}, \{x_{i1}, x_{i2}\}, \dots, \{x_{i1}, \dots, x_{i20}\}$. For CSA, we set the maximum submodels to $M_{max} = 100$. Thus, we draw 100 models randomly from the complete subsets of size k if $M = K!/(k!(K - k!))$ is bigger than 100. Furthermore, we reduce some computational burden by applying 10-fold cross-validation when $n = 150$. The tuning parameter of L1QR is chosen by equation (2.7) in Belloni and Chernozhukov (2011). The bootstrap size of BAG is set to be 1,000. The tuning parameter of L2QR is chosen by 10-fold cross-validation over the set $\{0.01, 0.05, 0.1, 0.5, 1.0\}$, which is constructed after some presimulation studies.

To compare the performance, we first compute $FPE(r)$ for each replication $r = 1, \dots, R$ as follows. After estimating the model with n in-sample observations, we generate additional 100 out-of-sample observations. Then, $FPE(r)$ is calculated by

$$FPE(r) := \frac{1}{100} \sum_{s=1}^{100} \rho_\tau(y_s - \hat{y}_s),$$

where \hat{y}_x is a predicted value by each method. Then, we construct the following three comparison measures:

$$\text{Average FPE}_A := R^{-1} \sum_{r=1}^R FPE(r)_A,$$

$$\text{Winning Ratio}_A := R^{-1} \sum_{r=1}^R 1\{FPE(r)_A < FPE(r)_B, \dots, FPE(r)_A < FPE(r)_E\},$$

$$\text{Loss to CSA}_A := R^{-1} \sum_{r=1}^R 1\{FPE(r)_{CSA} < FPE(r)_A\},$$

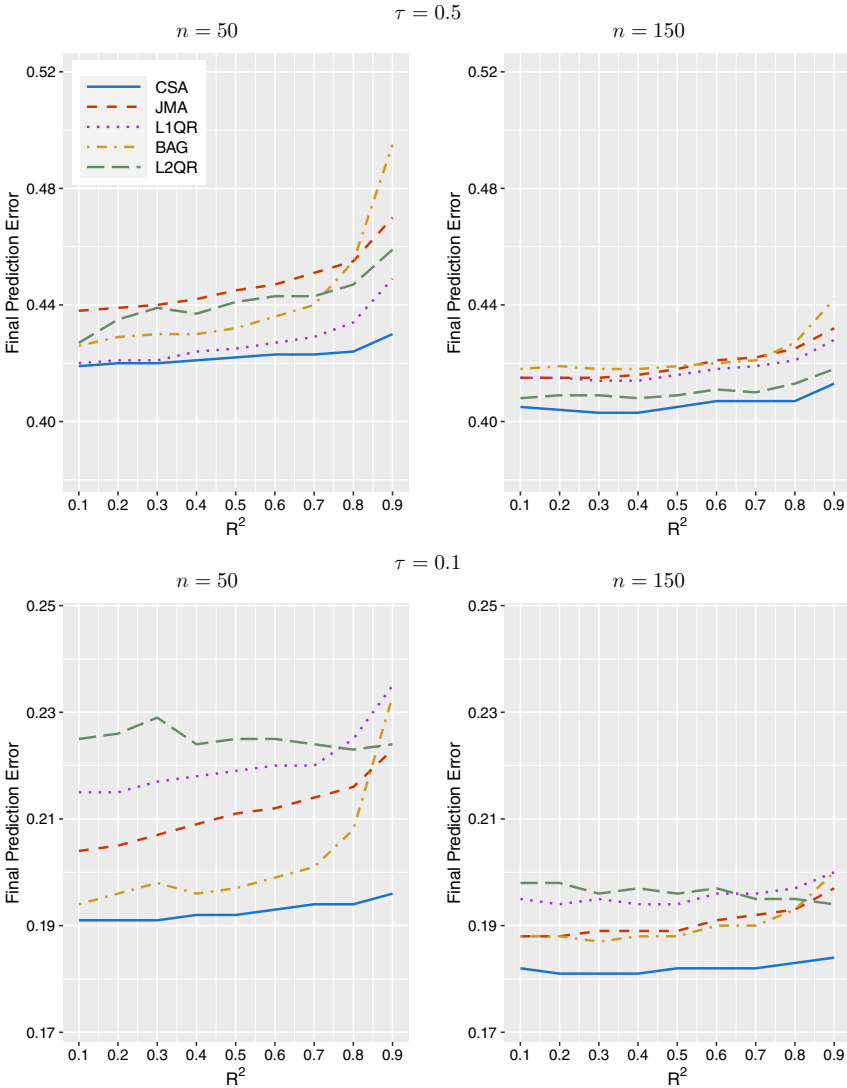


FIGURE 1. Prediction errors over R^2 .

where each subscript denotes generic notation for a forecasting method. Note that the loss to CSA ratio provides more direct binary comparison of each method to CSA. We set the total number of replications $R = 1,000$.

Figures 1–3 and Tables 1–4 summarize the simulation results over all designs. Overall, the performance of CSA compared to the alternative is quite satisfactory. We first direct our attention to Figure 1 and Tables 1 and 2. In these simulation

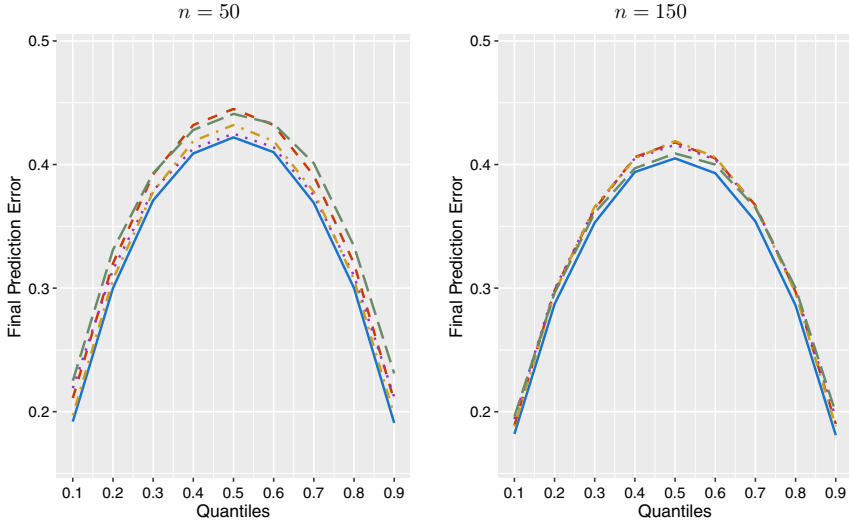


FIGURE 2. Prediction errors over various quantiles.

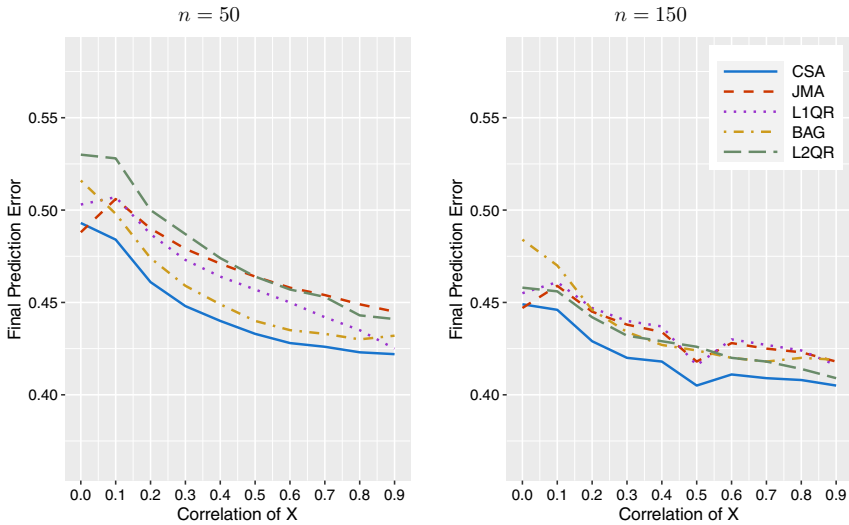


FIGURE 3. Prediction errors over various ρ_x .

designs, we vary R^2 over $\{0.1, 0.2, \dots, 0.9\}$ while setting $\rho_x = 0.9$. We consider two quantiles, $\tau = 0.1$ and 0.5 , respectively. From the four graphs in Figure 1, we confirm that CSA outperforms the alternative uniformly over R^2 's in terms of FPE in both quantiles. The prediction performance of CSA is better when the sample

TABLE 1. Simulation results over various R^2 : $\tau = 0.5$

R^2	$n = 50$					$n = 150$				
	CSA	JMA	L1QR	BAG	L2QR	CSA	JMA	L1QR	BAG	L2QR
	<u>Average FPE</u>									
0.1	0.419 (0.043)	0.438 (0.047)	0.420 (0.037)	0.426 (0.035)	0.427 (0.048)	0.405 (0.033)	0.415 (0.035)	0.415 (0.033)	0.418 (0.033)	0.408 (0.033)
0.2	0.420 (0.042)	0.439 (0.046)	0.421 (0.036)	0.429 (0.035)	0.435 (0.051)	0.404 (0.032)	0.415 (0.034)	0.415 (0.033)	0.419 (0.033)	0.409 (0.033)
0.3	0.420 (0.041)	0.440 (0.045)	0.421 (0.036)	0.430 (0.035)	0.439 (0.053)	0.403 (0.032)	0.415 (0.033)	0.414 (0.033)	0.418 (0.032)	0.409 (0.032)
0.4	0.421 (0.042)	0.442 (0.047)	0.424 (0.035)	0.430 (0.035)	0.437 (0.049)	0.403 (0.032)	0.416 (0.034)	0.414 (0.033)	0.418 (0.032)	0.408 (0.031)
0.5	0.422 (0.042)	0.445 (0.046)	0.425 (0.035)	0.432 (0.035)	0.441 (0.053)	0.405 (0.032)	0.418 (0.034)	0.416 (0.033)	0.419 (0.032)	0.409 (0.032)
0.6	0.423 (0.041)	0.447 (0.045)	0.427 (0.035)	0.436 (0.036)	0.443 (0.055)	0.407 (0.032)	0.421 (0.034)	0.418 (0.033)	0.420 (0.032)	0.411 (0.033)
0.7	0.423 (0.043)	0.451 (0.046)	0.429 (0.036)	0.440 (0.036)	0.443 (0.052)	0.407 (0.031)	0.422 (0.033)	0.419 (0.033)	0.421 (0.032)	0.410 (0.032)
0.8	0.424 (0.042)	0.455 (0.046)	0.434 (0.036)	0.455 (0.038)	0.447 (0.051)	0.407 (0.031)	0.425 (0.033)	0.421 (0.033)	0.427 (0.032)	0.413 (0.032)
0.9	0.430 (0.043)	0.470 (0.048)	0.449 (0.038)	0.495 (0.049)	0.459 (0.055)	0.413 (0.033)	0.432 (0.035)	0.428 (0.034)	0.442 (0.035)	0.418 (0.033)
	<u>Winning Ratio</u>									
0.1	27.8%	8.6%	14.9%	19.5%	29.2%	34.2%	10.1%	6.8%	11.3%	37.6%
0.2	29.6%	7.7%	17.2%	19.7%	25.8%	35.9%	10.1%	7.7%	11.8%	34.5%
0.3	33.6%	6.9%	16.2%	20.9%	22.4%	39.4%	8.9%	7.9%	10.9%	32.9%
0.4	34.3%	6.4%	14.4%	21.0%	23.9%	38.5%	7.4%	7.7%	12.5%	33.9%
0.5	35.8%	5.9%	16.3%	18.0%	24.0%	38.4%	8.8%	7.2%	13.4%	32.2%
0.6	36.6%	7.1%	15.6%	17.7%	23.0%	40.1%	7.4%	7.5%	13.0%	32.0%
0.7	38.8%	4.4%	15.5%	16.0%	25.3%	40.4%	6.3%	7.1%	12.3%	33.9%
0.8	47.2%	4.4%	12.7%	11.1%	24.6%	43.2%	6.7%	6.2%	10.0%	33.9%
0.9	54.1%	4.2%	9.9%	3.6%	28.2%	43.0%	5.3%	6.1%	6.7%	38.9%
	<u>Loss to CSA</u>									
0.1	NA	77.3%	59.6%	57.0%	55.6%	NA	77.8%	81.8%	65.4%	52.5%
0.2	NA	77.8%	59.4%	58.4%	59.8%	NA	79.1%	82.6%	64.2%	55.2%
0.3	NA	79.8%	62.2%	61.1%	61.5%	NA	80.4%	82.8%	66.3%	57.3%
0.4	NA	79.3%	65.1%	60.0%	62.3%	NA	82.8%	83.2%	65.2%	54.2%
0.5	NA	80.7%	65.6%	61.6%	62.1%	NA	81.9%	81.8%	65.5%	55.5%
0.6	NA	82.4%	67.2%	64.3%	64.7%	NA	82.7%	82.4%	66.1%	57.1%
0.7	NA	84.7%	69.8%	64.5%	62.3%	NA	85.8%	82.9%	65.2%	54.8%
0.8	NA	86.4%	75.5%	73.2%	65.4%	NA	85.6%	85.4%	69.5%	56.5%
0.9	NA	89.2%	82.6%	86.0%	68.0%	NA	87.2%	84.6%	74.5%	54.3%

Note: The standard error of the average FPE is denoted inside the parentheses.

TABLE 2. Simulation results over various R^2 : $\tau = 0.1$

R^2	$n = 50$					$n = 150$				
	CSA	JMA	L1QR	BAG	L2QR	CSA	JMA	L1QR	BAG	L2QR
<u>Average FPE</u>										
0.1	0.191 (0.028)	0.204 (0.036)	0.215 (0.040)	0.194 (0.023)	0.225 (0.046)	0.182 (0.019)	0.188 (0.021)	0.195 (0.025)	0.188 (0.020)	0.198 (0.026)
0.2	0.191 (0.027)	0.205 (0.035)	0.215 (0.039)	0.196 (0.024)	0.226 (0.050)	0.181 (0.019)	0.188 (0.021)	0.194 (0.024)	0.188 (0.021)	0.198 (0.027)
0.3	0.191 (0.029)	0.207 (0.036)	0.217 (0.040)	0.198 (0.025)	0.229 (0.049)	0.181 (0.019)	0.189 (0.021)	0.195 (0.025)	0.187 (0.020)	0.196 (0.025)
0.4	0.192 (0.028)	0.209 (0.037)	0.218 (0.039)	0.196 (0.023)	0.224 (0.047)	0.181 (0.019)	0.189 (0.021)	0.194 (0.024)	0.188 (0.021)	0.197 (0.027)
0.5	0.192 (0.029)	0.211 (0.036)	0.219 (0.040)	0.197 (0.023)	0.225 (0.046)	0.182 (0.019)	0.189 (0.021)	0.194 (0.025)	0.188 (0.020)	0.196 (0.026)
0.6	0.193 (0.029)	0.212 (0.037)	0.220 (0.042)	0.199 (0.024)	0.225 (0.047)	0.182 (0.018)	0.191 (0.021)	0.196 (0.024)	0.190 (0.020)	0.197 (0.026)
0.7	0.194 (0.032)	0.214 (0.038)	0.220 (0.040)	0.201 (0.025)	0.224 (0.046)	0.182 (0.018)	0.192 (0.021)	0.196 (0.024)	0.190 (0.021)	0.195 (0.025)
0.8	0.194 (0.028)	0.216 (0.035)	0.225 (0.041)	0.208 (0.027)	0.223 (0.046)	0.183 (0.018)	0.193 (0.022)	0.197 (0.025)	0.193 (0.020)	0.195 (0.024)
0.9	0.196 (0.027)	0.223 (0.035)	0.235 (0.041)	0.233 (0.032)	0.224 (0.046)	0.184 (0.018)	0.197 (0.022)	0.200 (0.025)	0.200 (0.021)	0.194 (0.023)
<u>Winning Ratio</u>										
0.1	38.8%	10.0%	8.0%	34.2%	9.0%	38.4%	12.8%	8.3%	26.7%	13.8%
0.2	41.2%	12.5%	8.5%	31.2%	6.6%	37.0%	12.8%	8.8%	28.0%	13.4%
0.3	43.0%	9.5%	8.1%	32.5%	6.9%	37.5%	11.1%	8.8%	27.4%	15.2%
0.4	42.0%	7.9%	7.8%	32.3%	10.0%	41.2%	11.2%	8.2%	26.0%	13.4%
0.5	43.7%	7.7%	8.9%	31.5%	8.2%	38.6%	10.7%	9.2%	26.7%	14.8%
0.6	45.1%	7.8%	7.9%	27.7%	11.5%	40.5%	11.3%	8.7%	23.9%	15.6%
0.7	44.9%	6.8%	9.4%	27.6%	11.3%	41.6%	11.0%	6.7%	22.6%	18.1%
0.8	47.1%	8.5%	8.7%	18.7%	17.0%	42.4%	9.2%	8.4%	19.4%	20.6%
0.9	57.1%	5.4%	7.3%	7.1%	23.1%	44.9%	8.4%	8.9%	11.2%	26.6%
<u>Loss to CSA</u>										
0.1	NA	74.4%	82.9%	55.8%	75.9%	NA	73.8%	84.4%	61.2%	70.0%
0.2	NA	75.5%	83.2%	59.4%	79.0%	NA	73.1%	83.3%	60.4%	70.8%
0.3	NA	79.7%	84.0%	59.8%	80.3%	NA	76.3%	83.8%	58.7%	69.9%
0.4	NA	80.8%	84.2%	57.2%	76.3%	NA	76.2%	83.9%	62.5%	69.7%
0.5	NA	82.6%	84.6%	58.9%	77.4%	NA	75.0%	82.6%	60.7%	68.8%
0.6	NA	81.8%	84.6%	60.7%	74.6%	NA	78.9%	83.5%	63.5%	70.2%
0.7	NA	83.3%	83.6%	62.4%	75.8%	NA	80.5%	86.9%	62.4%	66.9%
0.8	NA	83.1%	84.8%	67.9%	74.5%	NA	81.5%	83.9%	65.8%	66.8%
0.9	NA	88.3%	88.5%	83.7%	72.0%	NA	84.1%	85.0%	76.3%	64.8%

TABLE 3. Simulation results over various quantiles

τ	$n = 50$					$n = 150$				
	CSA	JMA	L1QR	BAG	L2QR	CSA	JMA	L1QR	BAG	L2QR
	<u>Average FPE</u>									
0.1	0.192 (0.029)	0.211 (0.036)	0.219 (0.040)	0.197 (0.023)	0.225 (0.046)	0.182 (0.019)	0.189 (0.021)	0.194 (0.025)	0.188 (0.020)	0.196 (0.026)
0.2	0.300 (0.038)	0.320 (0.044)	0.315 (0.039)	0.307 (0.029)	0.331 (0.053)	0.287 (0.024)	0.298 (0.027)	0.299 (0.028)	0.296 (0.026)	0.296 (0.030)
0.3	0.371 (0.038)	0.392 (0.043)	0.379 (0.037)	0.378 (0.033)	0.393 (0.051)	0.353 (0.029)	0.364 (0.032)	0.365 (0.031)	0.366 (0.030)	0.361 (0.032)
0.4	0.409 (0.040)	0.432 (0.044)	0.413 (0.036)	0.419 (0.035)	0.428 (0.053)	0.394 (0.031)	0.406 (0.034)	0.405 (0.033)	0.405 (0.032)	0.397 (0.032)
0.5	0.422 (0.042)	0.445 (0.046)	0.425 (0.035)	0.432 (0.035)	0.441 (0.053)	0.405 (0.032)	0.418 (0.034)	0.416 (0.033)	0.419 (0.032)	0.409 (0.032)
0.6	0.410 (0.042)	0.432 (0.045)	0.414 (0.037)	0.419 (0.035)	0.433 (0.053)	0.393 (0.032)	0.405 (0.033)	0.404 (0.033)	0.406 (0.031)	0.400 (0.032)
0.7	0.369 (0.041)	0.391 (0.045)	0.375 (0.036)	0.378 (0.033)	0.401 (0.053)	0.354 (0.029)	0.367 (0.031)	0.366 (0.031)	0.366 (0.029)	0.365 (0.031)
0.8	0.300 (0.036)	0.320 (0.043)	0.310 (0.035)	0.307 (0.030)	0.334 (0.052)	0.286 (0.025)	0.297 (0.027)	0.298 (0.027)	0.296 (0.026)	0.300 (0.029)
0.9	0.191 (0.029)	0.209 (0.034)	0.212 (0.036)	0.197 (0.024)	0.231 (0.047)	0.181 (0.019)	0.190 (0.022)	0.195 (0.025)	0.188 (0.020)	0.199 (0.027)
	<u>Winning Ratio</u>									
0.1	43.7%	7.7%	8.9%	31.5%	8.2%	38.6%	10.7%	9.2%	26.7%	14.8%
0.2	41.9%	7.1%	11.9%	25.0%	14.1%	38.5%	10.2%	7.8%	18.7%	24.8%
0.3	37.0%	6.0%	13.9%	23.3%	19.8%	37.0%	10.3%	8.7%	15.2%	28.8%
0.4	34.0%	5.0%	15.7%	22.0%	23.3%	35.9%	7.8%	8.3%	12.3%	35.7%
0.5	35.7%	5.9%	16.3%	18.1%	24.0%	38.3%	8.8%	7.2%	13.5%	32.2%
0.6	36.5%	6.6%	15.2%	23.9%	17.8%	38.3%	8.4%	8.0%	15.6%	29.7%
0.7	38.1%	8.1%	14.7%	25.9%	13.2%	42.3%	8.5%	7.3%	18.5%	23.4%
0.8	41.0%	7.2%	11.6%	28.1%	12.1%	40.5%	9.7%	8.0%	23.2%	18.6%
0.9	43.3%	8.6%	9.6%	30.7%	7.8%	41.5%	11.6%	7.6%	26.5%	12.8%
	<u>Loss to CSA</u>									
0.1	NA	82.6%	84.6%	58.9%	77.4%	NA	75.0%	82.6%	60.7%	68.8%
0.2	NA	82.1%	78.0%	63.3%	71.6%	NA	78.3%	83.2%	62.5%	61.6%
0.3	NA	81.6%	70.8%	59.4%	64.0%	NA	79.9%	83.9%	63.2%	58.4%
0.4	NA	81.4%	65.7%	58.2%	59.4%	NA	81.9%	80.0%	62.9%	53.0%
0.5	NA	80.7%	65.6%	61.1%	62.1%	NA	81.9%	81.8%	65.6%	55.3%
0.6	NA	80.1%	67.9%	59.8%	66.9%	NA	80.2%	80.5%	63.1%	56.9%
0.7	NA	78.9%	71.4%	60.9%	71.1%	NA	82.5%	84.1%	64.4%	62.3%
0.8	NA	82.6%	74.3%	60.1%	72.5%	NA	81.5%	82.1%	63.4%	65.3%
0.9	NA	80.1%	81.6%	61.2%	81.0%	NA	77.1%	85.2%	62.2%	73.1%

TABLE 4. Simulation results over various ρ_x

ρ_x	$n = 50$					$n = 150$				
	CSA	JMA	L1QR	BAG	L2QR	CSA	JMA	L1QR	BAG	L2QR
	<u>Average FPE</u>									
0.0	0.493	0.488	0.503	0.516	0.530	0.449	0.447	0.455	0.484	0.458
	(0.047)	(0.050)	(0.052)	(0.042)	(0.065)	(0.037)	(0.036)	(0.037)	(0.038)	(0.038)
0.1	0.484	0.506	0.507	0.498	0.528	0.446	0.459	0.461	0.470	0.456
	(0.046)	(0.049)	(0.049)	(0.040)	(0.060)	(0.036)	(0.038)	(0.037)	(0.037)	(0.038)
0.2	0.461	0.490	0.487	0.474	0.500	0.429	0.445	0.447	0.446	0.442
	(0.045)	(0.049)	(0.045)	(0.039)	(0.053)	(0.034)	(0.036)	(0.037)	(0.035)	(0.036)
0.3	0.448	0.479	0.473	0.459	0.487	0.420	0.438	0.440	0.434	0.432
	(0.044)	(0.049)	(0.042)	(0.037)	(0.054)	(0.034)	(0.035)	(0.036)	(0.032)	(0.035)
0.4	0.440	0.471	0.464	0.449	0.474	0.418	0.434	0.437	0.427	0.429
	(0.042)	(0.046)	(0.042)	(0.036)	(0.052)	(0.032)	(0.034)	(0.034)	(0.032)	(0.034)
0.5	0.433	0.464	0.457	0.440	0.464	0.405	0.418	0.416	0.424	0.426
	(0.043)	(0.047)	(0.042)	(0.035)	(0.053)	(0.032)	(0.034)	(0.033)	(0.033)	(0.035)
0.6	0.428	0.458	0.450	0.435	0.457	0.411	0.428	0.430	0.420	0.420
	(0.042)	(0.047)	(0.041)	(0.033)	(0.053)	(0.032)	(0.034)	(0.034)	(0.032)	(0.034)
0.7	0.426	0.454	0.442	0.433	0.453	0.409	0.425	0.427	0.418	0.418
	(0.041)	(0.046)	(0.038)	(0.034)	(0.053)	(0.033)	(0.033)	(0.034)	(0.032)	(0.033)
0.8	0.423	0.449	0.435	0.430	0.443	0.408	0.423	0.424	0.420	0.414
	(0.040)	(0.046)	(0.038)	(0.034)	(0.052)	(0.031)	(0.034)	(0.033)	(0.033)	(0.034)
0.9	0.422	0.445	0.425	0.432	0.441	0.405	0.418	0.416	0.419	0.409
	(0.042)	(0.046)	(0.035)	(0.035)	(0.053)	(0.032)	(0.034)	(0.033)	(0.032)	(0.032)
	<u>Winning Ratio</u>									
0.0	20.2%	33.9%	14.3%	15.6%	16.0%	23.0%	31.1%	8.7%	6.7%	30.5%
0.1	38.4%	11.9%	10.5%	26.6%	12.6%	40.0%	11.5%	6.4%	11.6%	30.5%
0.2	44.3%	7.7%	7.9%	28.4%	11.7%	43.4%	9.1%	6.9%	16.2%	24.4%
0.3	45.4%	6.4%	6.5%	29.4%	12.3%	44.5%	7.1%	5.4%	19.1%	23.9%
0.4	46.0%	6.6%	6.2%	28.6%	12.6%	43.4%	8.1%	5.4%	21.2%	21.9%
0.5	42.5%	5.3%	7.1%	31.7%	13.4%	46.6%	10.2%	8.6%	18.3%	16.3%
0.6	44.4%	4.8%	6.3%	27.1%	17.4%	43.7%	5.8%	4.2%	21.4%	24.9%
0.7	41.1%	6.8%	7.7%	27.0%	17.4%	44.5%	5.7%	3.3%	20.1%	26.4%
0.8	38.7%	6.7%	10.3%	22.3%	22.0%	38.7%	7.9%	4.3%	17.5%	31.6%
0.9	35.8%	5.9%	16.3%	18.0%	24.0%	38.5%	8.7%	7.2%	13.5%	32.1%
	<u>Loss to CSA</u>									
0.0	NA	40.4%	63.0%	63.9%	68.3%	NA	44.0%	67.7%	73.5%	58.0%
0.1	NA	73.7%	77.0%	61.5%	73.7%	NA	76.6%	82.7%	70.1%	59.0%
0.2	NA	79.1%	81.0%	61.3%	73.9%	NA	81.4%	86.5%	65.7%	62.9%
0.3	NA	83.8%	82.9%	58.8%	73.2%	NA	84.9%	88.7%	63.3%	61.7%
0.4	NA	83.9%	82.6%	59.5%	72.1%	NA	82.6%	87.9%	61.4%	61.7%
0.5	NA	87.1%	82.5%	55.8%	70.2%	NA	81.9%	81.8%	67.3%	70.2%
0.6	NA	85.7%	82.7%	58.4%	68.6%	NA	85.3%	88.6%	59.3%	59.1%
0.7	NA	84.3%	79.1%	57.4%	67.7%	NA	85.2%	88.8%	58.8%	58.9%
0.8	NA	81.3%	74.5%	59.3%	63.1%	NA	83.3%	87.8%	61.7%	53.9%
0.9	NA	80.7%	65.6%	61.6%	62.1%	NA	81.9%	81.8%	65.6%	55.4%

size is small, $n = 50$, and the gap decreases as the sample size increases to $n = 150$. At $\tau = 0.5$, L1QR performs the second when $n = 50$, but L2QR does the second when $n = 150$. Thus, the performance order next to CSA is not stable. At $\tau = 0.1$, BAG performs the second overall, but it is deteriorated when R^2 is very high, e.g., $R^2 = 0.9$. We also note that the performance of CSA is relatively stable over R^2 while that of the alternative increases steeply for larger R^2 when $n = 50$. The same results are confirmed in Tables 1 and 2. CSA shows the highest winning ratios over all designs except $\tau = 0.5$ and $R^2 = 0.1$, where that of L2QR is slightly higher. When we conduct the binary comparison (loss to CSA), all methods lose more than 50% to CSA over all designs and more than 80% in some designs. Therefore, we conclude that both the winning ratio and the loss to CSA are more favorable to CSA in this set of simulation designs.

In the next simulation, we study the performance over a wider range of quantiles. We vary the quantile $\tau = \{0.1, 0.2, \dots, 0.9\}$ while setting $R^2 = 0.5$ and $\rho_x = 0.9$. The results are summarized in Figure 2 and Table 3. In Figure 2, CSA outperforms the alternative uniformly over all quantiles in both sample sizes followed by BAG and L2QR. Again, the gap decreases as the sample size increases. It is also interesting that all estimators predict better at the tail distributions and they show the largest prediction errors at the median. The winning ratio and the loss to CSA in Table 3 are also satisfactory.

Third, we check the performance over different levels of dependency among the predictors. We vary $\rho_x = \{0, 0.1, 0.2, \dots, 0.9\}$ while setting $R^2 = 0.5$ and $\tau = 0.5$. Since $(x_{i2}, \dots, x_{i1000})$ are generated from the multivariate normal distribution, they are independent when $\rho_x = 0$. Figure 3 reveals an interesting point. CSA performs better than the alternative when there exists any correlation between the predictors, i.e., $\rho_x > 0$. Recall that most simulation studies in the literature consider independent predictors. As we can see from the empirical applications in the next section, however, the predictors are usually correlated with each other. Therefore, it is promising that CSA performs better when there is any correlation among predictors. Elliott et al. (2013) also report in the conditional mean prediction settings that the CSA approach performs better when predictors are correlated with each other. In Table 4, both the winning ratio and the loss to CSA statistics improve dramatically when ρ_x is away from zero, where JMA performs the best.

We next consider the second category of simulation designs, where the candidate models include the true DGP. The new simulations are based on the following model:

$$y_i = \theta \sum_{j=1}^K \beta_j x_{ij} + \varepsilon_i,$$

where we observe all K predictors in the sample. We consider $K = 5, 15$ when $n = 50$ and $K = 10, 20$ when $n = 150$. Similar to the previous simulations, the

TABLE 5. Correct specification: Decreasing signal

		<i>n</i> = 50				
		CSA	JMA	L1QR	BAG	L2QR
<u>Average FPE</u>						
<i>K</i> = 5		0.415 (0.033)	0.424 (0.036)	0.419 (0.034)	0.430 (0.036)	0.419 (0.035)
<i>K</i> = 15		0.420 (0.039)	0.443 (0.044)	0.424 (0.034)	0.426 (0.035)	0.427 (0.047)
<u>Winning Ratio</u>						
<i>K</i> = 5		28.5%	11.1%	13.2%	12.5%	34.7%
<i>K</i> = 15		32.7%	5.4%	11.1%	21.7%	29.1%
<u>Loss to CSA</u>						
<i>K</i> = 5		NA	73.6%	66.2%	62.7%	53.7%
<i>K</i> = 15		NA	81.8%	70.1%	56.7%	54.4%
		<i>n</i> = 150				
		CSA	JMA	L1QR	BAG	L2QR
<u>Average FPE</u>						
<i>K</i> = 10		0.406 (0.032)	0.412 (0.033)	0.411 (0.032)	0.422 (0.034)	0.406 (0.031)
<i>K</i> = 20		0.407 (0.032)	0.419 (0.034)	0.419 (0.033)	0.417 (0.032)	0.408 (0.032)
<u>Winning Ratio</u>						
<i>K</i> = 10		30.6%	10.8%	8.6%	8.7%	41.3%
<i>K</i> = 20		35.2%	9.2%	6.5%	13.0%	36.1%
<u>Loss to CSA</u>						
<i>K</i> = 10		NA	71.6%	72.7%	65.0%	49.0%
<i>K</i> = 20		NA	77.9%	82.9%	61.6%	50.9%

population R^2 is controlled by θ . We set $R^2 = 0.5$, $\tau = 0.5$, and $\rho_X = 0.9$. Instead of varying R^2 , τ , and ρ_X , we consider three signal structures in this simulation:

Decreasing signal : $\beta_j = j^{-1}$,

Constant signal : $\beta_j = 1$, for all j ,

$$\text{Sparse signal : } \beta_j = \begin{cases} 1 & \text{if } j = 1, 2 \\ 0 & \text{if } j > 2 \end{cases} .$$

Therefore, we consider 12 new DGPs in total.

Tables 5–7 summarize the simulation results. First of all, we take a look at the loss to CSA ratio in the second column (JMA) in these tables. Note that the loss ratio increases as K increases over all different designs, which is expected by the

TABLE 6. Correct specification: Constant signal

		<i>n</i> = 50				
		CSA	JMA	L1QR	BAG	L2QR
<u>Average FPE</u>						
<i>K</i> = 5		0.414 (0.033)	0.426 (0.036)	0.419 (0.034)	0.429 (0.035)	0.415 (0.034)
<i>K</i> = 15		0.418 (0.040)	0.443 (0.046)	0.422 (0.036)	0.429 (0.034)	0.429 (0.049)
<u>Winning Ratio</u>						
<i>K</i> = 5		30.3%	7.0%	12.6%	12.4%	37.7%
<i>K</i> = 15		33.1%	4.7%	14.5%	16.9%	30.8%
<u>Loss to CSA</u>						
<i>K</i> = 5		NA	80.6%	68.0%	61.6%	50.6%
<i>K</i> = 15		NA	85.0%	65.8%	60.2%	56.0%
		<i>n</i> = 150				
		CSA	JMA	L1QR	BAG	L2QR
<u>Average FPE</u>						
<i>K</i> = 10		0.406 (0.031)	0.414 (0.032)	0.411 (0.032)	0.422 (0.032)	0.406 (0.032)
<i>K</i> = 20		0.406 (0.033)	0.419 (0.034)	0.416 (0.034)	0.420 (0.033)	0.410 (0.032)
<u>Winning Ratio</u>						
<i>K</i> = 10		31.2%	6.2%	11.4%	8.4%	42.8%
<i>K</i> = 20		40.3%	6.6%	8.8%	11.9%	32.4%
<u>Loss to CSA</u>						
<i>K</i> = 10		NA	78.3%	73.7%	67.4%	47.8%
<i>K</i> = 20		NA	82.2%	81.8%	65.3%	56.3%

theoretical results in Theorem 5. Second, CSA performs worse in the sparse signal models compared to the other two designs. As discussed under equation (10), this is expected from the theory in Section 3, since the sparse design generates many subsets with totally irrelevant predictors. Third, it is interesting that JMA does not particularly outperform in this setup, where the candidate models include the true one. Furthermore, note that L1QR does not particularly outperform in the sparse signal model. In fact, L2QR performs well over all three signal designs. Given that L2QR is understudied in the literature, this would be an interesting topic for future research.

In sum, we confirm that CSA shows satisfactory finite sample properties via Monte Carlo simulation studies. Related to the forecast combination puzzle, we observe a similar phenomenon in quantile forecasting and confirm some theoretical predictions developed in Section 3.

TABLE 7. Correct specification: Sparse signal

		<i>n</i> = 50				
		CSA	JMA	L1QR	BAG	L2QR
<u>Average FPE</u>						
<i>K</i> = 5		0.422 (0.034)	0.422 (0.036)	0.420 (0.035)	0.430 (0.036)	0.421 (0.035)
<i>K</i> = 15		0.437 (0.042)	0.442 (0.046)	0.431 (0.038)	0.430 (0.035)	0.439 (0.049)
<u>Winning Ratio</u>						
<i>K</i> = 5		13.5%	20.9%	20.3%	12.7%	32.6%
<i>K</i> = 15		14.7%	16.9%	18.3%	25.6%	24.5%
<u>Loss to CSA</u>						
<i>K</i> = 5		NA	47.2%	41.9%	57.2%	50.4%
<i>K</i> = 15		NA	52.3%	44.9%	45.5%	48.8%
		<i>n</i> = 150				
		CSA	JMA	L1QR	BAG	L2QR
<u>Average FPE</u>						
<i>K</i> = 10		0.414 (0.032)	0.410 (0.032)	0.411 (0.032)	0.423 (0.032)	0.411 (0.031)
<i>K</i> = 20		0.418 (0.033)	0.416 (0.034)	0.419 (0.034)	0.422 (0.031)	0.415 (0.032)
<u>Winning Ratio</u>						
<i>K</i> = 10		13.3%	22.8%	15.5%	12.3%	36.1%
<i>K</i> = 20		13.9%	24.9%	12.6%	17.4%	31.2%
<u>Loss to CSA</u>						
<i>K</i> = 10		NA	38.2%	39.5%	58.0%	45.2%
<i>K</i> = 20		NA	41.0%	51.0%	52.3%	45.9%

5. EMPIRICAL ILLUSTRATION

In this section, we investigate the performance of the proposed method with real datasets. Specifically, we revisit two empirical applications in Lu and Su (2015): (i) quantile forecast of excess stock returns; and (ii) quantile forecast of wages. Following the simulation studies in Section 4, we compare the performance of the CSA method to the JMA, the L1QR, the BAG method, and the L2QR.

5.1. Stock Return

The same dataset is composed of monthly observations of the US stock market from January 1950 to December 2005 ($T = 672$). The dependent variable is the excess stock return. We use the following 12 regressors: default yield spread,

TABLE 8. Out-of-sample R^2 for the excess stock return data

τ	T_1	CSA	JMA	L1QR	BAG	L2QR	$E[\widehat{k}]$	$Med[\widehat{k}]$
0.05	48	-0.071 (2)	-0.117 (4)	-0.088 (3)	0.031 (1)	-4.331 (5)	7.4	8
	60	-0.063 (3)	-0.125 (4)	-0.038 (2)	0.009 (1)	-4.395 (5)	8.1	8
	72	-0.001 (2)	-0.023 (4)	-0.005 (3)	0.020 (1)	-3.955 (5)	8.2	9
	96	0.055 (1)	-0.020 (4)	-0.012 (3)	0.027 (2)	-4.010 (5)	8.1	9
	120	0.104 (1)	0.053 (2)	0.028 (4)	0.033 (3)	-3.655 (5)	8.7	9
	144	0.082 (1)	0.045 (2)	0.012 (4)	0.019 (3)	-3.735 (5)	9.1	9
	180	0.039 (1)	0.033 (2)	0.023 (3)	-0.011 (4)	-2.311 (5)	9.6	10
0.5	48	0.103 (1)	0.076 (2)	-0.040 (4)	-0.016 (3)	-2.341 (5)	9.8	10
	60	0.089 (1)	0.079 (2)	-0.036 (4)	-0.013 (3)	-2.078 (5)	9.9	10
	72	0.057 (2)	0.067 (1)	-0.003 (3)	-0.009 (4)	-1.953 (5)	10.0	10
	96	0.049 (2)	0.053 (1)	-0.013 (3)	-0.014 (4)	-2.206 (5)	10.3	11
	120	0.003 (2)	0.013 (1)	0.003 (3)	-0.011 (4)	-1.882 (5)	10.5	11
	144	-0.012 (3)	-0.002 (1)	-0.006 (2)	-0.022 (4)	-1.648 (5)	10.6	11
	180	0.032 (2)	0.034 (1)	0.018 (3)	-0.012 (4)	-1.031 (5)	10.5	11

Note: The number in the parentheses denotes the performance ranking among the five different methods.

treasury bill rate, net equity expansion, term spread, dividend price ratio, earnings price ratio, long-term yield, book-to-market ratio, inflation, return on equity, lagged dependent variable, and smoothed earnings price ratio. See Campbell and Thompson (2007) and Lu and Su (2015) for the details of the dataset. Note that JMA needs to select the order of important regressors, but we do not need such a selection for CSA, BAG, and L2QR. L1QR would select important regressors automatically by the ℓ_1 -penalty.

We forecast the one-period-ahead excess stock returns at 0.5 and 0.05 quantiles using various fixed in-sample sizes, $T_1 = 48, 60, 72, 96, 120, 144,$ and 180 . The forecast performance is measured by the out-of-sample R^2 defined as

$$R^2 = 1 - \frac{\sum_{t=T_1}^{T-1} \rho_\tau (y_{t+1} - \widehat{y}_{t+1|t})}{\sum_{t=T_1}^{t-1} \rho_\tau (y_{t+1} - \bar{y}_{t+1|t})},$$

where $\widehat{y}_{t+1|t}$ the one-period-ahead τ -quantile prediction at time t using the data from the past T_1 periods, and $\bar{y}_{t+1|t}$ is the unconditional τ -quantile for the same T_1 periods. The out-of-sample R^2 measures the relative performance of a forecast method compared to the unconditional historical quantile. The higher values of R^2 imply better forecasting performance.

Table 8 summarizes the forecasting results. In addition to R^2 , we report the ranking of each forecasting method, the mean of \widehat{k} , and the median of \widehat{k} . The upper panel of Table 8 reports the results when $\tau = 0.05$. The R^2 of CSA is better than that of JMA *uniformly* over different sample sizes (T_1). The gap between two

R^2 's is substantial except $T = 180$. BAG performs well when T_1 is small. The performance of L2QR is not satisfactory over all in-sample sizes. We next turn our attention to the lower panel when $\tau = 0.5$. Again, CSA performs the best or second best except when $T_1 = 144$. CSA performs better when T_1 is small, while JMA does better when T_1 is larger. Overall, the gap between R^2 's is small when $\tau = 0.5$. As we have observed from the simulation studies, the performance of the two estimators becomes similar as the sample size increases in both panels. It is also noticeable that the selected \hat{k} of CSA increases as the sample size increases and that CSA selects relatively large \hat{k} across all T_1 and τ . Different from $\tau = 0.05$, BAG performs poorly when $\tau = 0.5$. L2QR also shows poor performance.

In sum, the performance of CSA is satisfactory in this forecasting exercise. It is quite stable over different in-sample sizes (T_1) and different quantiles in terms of the performance ranking. Among the alternative, BAG and JMA perform well in certain quantiles (0.05 and 0.5, respectively), but they do poorly when we apply them in different quantiles.

5.2. Wage

In this subsection, we conduct the quantile forecast exercises using the Current Population Survey data in 1975. The same dataset is also used by Lu and Su (2015) and Hansen and Racine (2012) for quantile and mean forecast exercises, respectively. The sample size is $n = 526$, and we use the logarithm of the average hourly wage as the dependent variable. We use the following 10 regressors: professional occupation, years of education, years with current employer, female, service occupation, married, trade, standard metropolitan statistical area (SMSA), services, and clerk occupation.

We split the sample into the estimation sample randomly drawn n_1 observations and the evaluation sample of $n - n_1$ observations. The estimation sample size varies $n_1 = 50, 100, 150,$ and 200 , and the random splitting is repeated 200 times for each n_1 . The out-of-sample R^2 is defined as

$$R^2 = 1 - \frac{\sum_{s=1}^{n_2} \rho_\tau(y_s - \hat{y}_s)}{\sum_{s=1}^{n_2} \rho_\tau(y_s - \bar{y}_s)},$$

where \hat{y}_s is the τ th conditional quantile predictor and \bar{y}_s is the unconditional τ -quantile estimate from the estimation sample. Again, R^2 measures the prediction performance relative to the unconditional quantile estimate.

Table 9 summarizes the exercise results.³ We confirm that CSA shows good and stable quantile prediction performance. In this application, BAG shows quite a similar performance to CSA. Similar to the stock return application, CSA performs better than BAG when $\tau = 0.5$, and BAG does when $\tau = 0.05$. The prediction results of JMA, L1QR, and L2QR are worse than CSA and BAG. The performance

³ R^2 's of JMA are different from the numbers reported in Table 5 in Lu and Su (2015), because they implemented the level of wage as a dependent variable which is supposed to be $\log(wage)$. We use $\log(wage)$ in this empirical illustration.

TABLE 9. Out-of-sample R^2 for the wage data

τ	n_1	CSA	JMA	L1QR	BAG	L2QR	$E[\hat{k}]$	$Med[\hat{k}]$
0.05	50	0.066 (2)	-0.034 (3)	-0.035 (4)	0.104 (1)	-0.139 (5)	3.9	4
	100	0.122 (2)	0.073 (4)	0.078 (3)	0.133 (1)	0.020 (5)	5.5	6
	150	0.138 (2)	0.112 (4)	0.113 (3)	0.144 (1)	0.076 (5)	6.1	6
	200	0.158 (1)	0.125 (4)	0.132 (3)	0.154 (2)	0.111 (5)	6.6	7
0.5	50	0.252 (1)	0.233 (3)	0.198 (5)	0.248 (2)	0.212 (4)	6.5	6
	100	0.287 (1)	0.276 (3)	0.233 (5)	0.285 (2)	0.260 (4)	7.7	8
	150	0.302 (1)	0.293 (3)	0.253 (5)	0.301 (2)	0.290 (4)	8.2	8
	200	0.307 (2)	0.302 (3)	0.267 (5)	0.312 (1)	0.302 (4)	8.4	9

Note: The number in the parentheses denotes the performance ranking among the five different methods.

gaps are larger when the sample size (n_1) is small and they narrow as n_1 increases. As predicted by the theory and also confirmed in the stock return application, the selected \hat{k} increases as n_1 increases.

6. CONCLUSION

In this paper, we propose a novel conditional quantile prediction method based on CSA of QRs. We show the asymptotic properties of the estimator when the dimension of regressors diverges to infinity as the sample size increases. The size of the complete subset is chosen by the leave-one-out cross-validation method. We prove that the subset size chosen by this method is optimal in the sense that it is asymptotically equivalent to the infeasible optimal size minimizing the FPE. The prediction performance in the simulation studies and empirical applications is satisfactory.

We conclude with two potential extensions of the proposed method. First, we can think of a different approach in choosing the complete subset size. Recently, Hirano and Wright (2021) have proposed a Laplace cross-validation method, where the tuning parameter of interest is chosen by the pseudo-Bayesian posterior mean, and showed that it works better than the standard cross-validation method when the risk function is asymmetric. It would be interesting to check how it performs in the CSA quantile prediction. Second, it will be useful if one can extend the results into the time-series data possibly including persistent regressors (e.g., Fan and Lee, 2019). We leave them for future research.

APPENDIX

In this appendix, we provide all necessary lemmas and technical proofs of the main text.

LEMMA 1. *Let $e_n := (nMK^2)^{1/4}$. Suppose that $K/\log(n) = O(1)$. Then, we can show the following rate conditions:*

- (i) $M = O(2^K)$,
- (ii) $K \log M/n = o(1)$,
- (iii) $(e_n \log M)/n = o(1)$.

Proof of Lemma 1. (i) Recall the dependency of M on K and k . By construction, $M_{K,k} = \binom{K}{k}$. Then, the result follows from $2^K = \sum_{k=0}^K \binom{K}{k}$. (ii) Note that

$$\begin{aligned} \frac{K \log M}{n} &= O\left(\frac{\log n(\log 2^{\log n})}{n}\right) \\ &= O\left(\log 2 \cdot \frac{(\log n)^2}{n}\right) = o(1). \end{aligned}$$

(iii) Note that

$$\begin{aligned} \frac{e_n \log M}{n} &= \frac{(nMK^2)^{1/4} \log M}{n} \\ &= \left(\frac{MK^2 (\log M)^4}{n^3}\right)^{1/4} \\ &= O\left(\left(\frac{2^{\log n}}{n}\right)^{1/4}\right) O\left(\left(\frac{(\log n)^2}{n}\right)^{1/4}\right) O\left(\left(\frac{(\log 2^{\log n})^2}{n}\right)^{1/4}\right). \end{aligned}$$

It is enough to show that $2^{\log n}/n = o(1)$. Let $c_{1n} = 2^{\log n}/n$. Then,

$$\log c_{1n} = \log n(\log 2 - 1) \rightarrow -\infty.$$

Therefore, $c_{1n} = o(1)$, and the desired result is established. □

LEMMA 2. Suppose that (i) $\sup_{j \geq 1} E[x_{ij}^2] < c_x$ with $c_x < \infty$, and (ii) $E[\mu_i^2] < \infty$. Then,

$$\max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \frac{1}{n} \sum_{i=1}^n \|x_{i(m,k)}\| = O_p(K^{1/2}).$$

Proof of Lemma 2. The triangle inequality implies that

$$\begin{aligned} &\max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \frac{1}{n} \sum_{i=1}^n \|x_{i(m,k)}\| \\ &\leq \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \frac{1}{n} \sum_{i=1}^n E\|x_{i(m,k)}\| + \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \left| \frac{1}{n} \sum_{i=1}^n (\|x_{i(m,k)}\| - E\|x_{i(m,k)}\|) \right| \\ &\equiv A_1 + A_2. \end{aligned}$$

We first investigate A_1 :

$$\begin{aligned}
 A_1 &= \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \frac{1}{n} \sum_{i=1}^n E \left[x'_{i(m,k)} x_{i(m,k)} \right]^{1/2} \\
 &\leq \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \frac{1}{n} \sum_{i=1}^n \left(E \left[x'_{i(m,k)} x_{i(m,k)} \right] \right)^{1/2} \\
 &\leq \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \frac{1}{n} \sum_{i=1}^n (kc_x)^{1/2} \\
 &\leq \max_{1 \leq k \leq K} k^{1/2} c_x^{1/2} \\
 &\leq K^{1/2} c_x^{1/2} = O(K^{1/2}).
 \end{aligned}$$

We next turn our attention to A_2 . Let $v_{i(m,k)} := \|x_{i(m,k)}\| - E \|x_{i(m,k)}\|$. Note that $\text{Var}(v_{i(m,k)}) \leq CK$ for some generic constant $C > 0$. Let $e_n := (nMK^2)^{1/4}$. We have

$$\begin{aligned}
 P(A_2 \geq 2\varepsilon) &= P \left(\max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \left| \frac{1}{n} \sum_{i=1}^n v_{i(m,k)} \right| \geq 2\varepsilon \right) \\
 &\leq P \left(\max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \frac{1}{n} \sum_{i=1}^n |v_{i(m,k)}| \geq 2\varepsilon \right) \\
 &\leq P \left(\max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \frac{1}{n} \sum_{i=1}^n |v_{i(m,k)}| \mathbb{1}(|v_{i(m,k)}| \leq e_n) \geq \varepsilon \right) \\
 &\quad + P \left(\max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \frac{1}{n} \sum_{i=1}^n |v_{i(m,k)}| \mathbb{1}(|v_{i(m,k)}| > e_n) \geq \varepsilon \right) \\
 &\equiv A_{21} + A_{22}.
 \end{aligned}$$

Boole’s and Bernstein inequalities imply that

$$\begin{aligned}
 A_{21} &\leq KM \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} P \left(\frac{1}{n} \sum_{i=1}^n |v_{i(m,k)}| \mathbb{1}(|v_{i(m,k)}| \leq e_n) \geq \varepsilon \right) \\
 &\leq 2KM \exp \left(-\frac{n\varepsilon^2}{2CK + 2\varepsilon e_n/3} \right) \\
 &= 2 \exp \left(-\frac{n\varepsilon^2}{2CK + 2\varepsilon e_n/3} + \log M + \log K \right) \\
 &= 2 \exp \left(\frac{-n\varepsilon^2}{2CK + 2\varepsilon e_n/3} \left(1 - \frac{2CK(\log M + \log K) + (2/3)\varepsilon e_n(\log M + \log K)}{n\varepsilon^2} \right) \right) = o(1).
 \end{aligned}$$

The convergence result follows from $K = o(M)$ by Lemma 1(i), $(K \log M)/n = o(1)$ by Lemma 1(ii), and $e_n \log M/n = o(1)$ by Lemma 1(iii).

Finally, we show that $A_{22} = o(1)$.

$$\begin{aligned}
 A_{22} &= P\left(\max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \frac{1}{n} \sum_{i=1}^n |v_{i(m,k)}| \mathbb{1}(|v_{i(m,k)}| > e_n) \geq \varepsilon\right) \\
 &\leq P\left(\max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \max_{1 \leq i \leq n} |v_{i(m,k)}| \mathbb{1}(|v_{i(m,k)}| > e_n) \geq \varepsilon\right) \\
 &\leq P\left(\max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \max_{1 \leq i \leq n} \mathbb{1}(|v_{i(m,k)}| > e_n)\right) \\
 &\leq \sum_{k=1}^K \sum_{m=1}^M \sum_{i=1}^n P(|v_{i(m,k)}| > e_n) \\
 &\leq \frac{1}{e_n^4} \sum_{k=1}^K \sum_{m=1}^M \sum_{i=1}^n E(|v_{i(m,k)}|^4) = \frac{O(1)}{K} = o(1).
 \end{aligned}$$

□

Proof of Theorem 2. The proof is similar to Lu and Su (2015, Theorem 3.2) except the last part that shows the convergence of the maximal inequality bound. Let $\delta_n := L\sqrt{n^{-1}K \log n}$ for some large constant $L < \infty$. Let $\bar{Q}_{(m,k)}(\Theta_{(m,k)}) := E[\rho_\tau(y_i - x'_{i(m,k)} \Theta_{(m,k)})]$. We also define

$$\begin{aligned}
 D(\delta_n) &:= \inf_{1 \leq m \leq M} \inf_{\|\Theta_{(m,k)} - \Theta^*_{(m,k)}\| > \delta_n} \left[\bar{Q}_{(m,k)}(\Theta_{(m,k)}) - \bar{Q}_{(m,k)}(\Theta^*_{(m,k)}) \right], \\
 \mathcal{S}_{(m,k)}(\delta_n) &:= \{\Theta_{(m,k)} : \|\Theta_{(m,k)} - \Theta^*_{(m,k)}\| > \delta_n, \|\Theta_{(m,k)} - \Theta^*_{(m,k)}\| = o(1)\}.
 \end{aligned}$$

The same arguments in Lu and Su (2015) imply that, for any $\Theta_{(m,k)} \in \mathcal{S}_{(m,k)}(\delta_n)$,

$$\begin{aligned}
 &\bar{Q}_{(m,k)}(\Theta_{(m,k)}) - \bar{Q}_{(m,k)}(\Theta^*_{(m,k)}) \\
 &= E[\rho_\tau(y_i - x'_{i(m,k)} \Theta_{(m,k)}) - \rho_\tau(y_i - x'_{i(m,k)} \Theta^*_{(m,k)})] \\
 &= E[\rho_\tau(\varepsilon_i + u_{i(m,k)} - x'_{i(m,k)} [\Theta_{(m,k)} - \Theta^*_{(m,k)}]) - \rho_\tau(\varepsilon_i - u_{i(m,k)})] \\
 &= E\left\{ \int_0^{x'_{i(m,k)} [\Theta_{(m,k)} - \Theta^*_{(m,k)}]} [1\{\varepsilon_i + u_{i(m,k)} \leq s\} - 1\{\varepsilon_i + u_{i(m,k)} \leq 0\}] ds \right\} \\
 &= E\left\{ \int_0^{x'_{i(m,k)} [\Theta_{(m,k)} - \Theta^*_{(m,k)}]} [F(-u_{i(m,k)} + s|x_i) - F(-u_{i(m,k)}|x_i)] ds \right\} \\
 &\approx \frac{1}{2} [\Theta_{(m,k)} - \Theta^*_{(m,k)}]' A_{(m,k)} [\Theta_{(m,k)} - \Theta^*_{(m,k)}] \geq \frac{c_A \delta_n^2}{2}.
 \end{aligned}$$

The claim in (i) is established by showing that the following maximal inequality converges to zero:

$$P\left(\max_{1 \leq i \leq n} \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \|\hat{\Theta}_{i(m,k)} - \Theta^*_{(m,k)}\| \geq \delta_n\right) = o_p(1).$$

We first derive the upper bound of it:

$$\begin{aligned}
 &P\left(\max_{1 \leq i \leq n} \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \|\widehat{\Theta}_{i(m,k)} - \Theta^*_{(m,k)}\| \geq \delta_n\right) \\
 &\leq nKM \max_{1 \leq i \leq n} \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} P\left(\|\widehat{\Theta}_{i(m,k)} - \Theta^*_{(m,k)}\| \geq \delta_n\right) \\
 &\leq nKM \max_{1 \leq i \leq n} \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} P\left(\widehat{Q}_{i(m,k)}(\Theta_{(m,k)}) - \bar{Q}_{i(m,k)}(\Theta^*_{(m,k)}) \geq D(\delta_n)\right) \\
 &\approx nKM \max_{1 \leq i \leq n} \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} P\left(\mathbb{W}_{i(m,k)} \geq 2nD(\delta_n)\right),
 \end{aligned}$$

where $\mathbb{W}_{i(m,k)} := n[\Theta_{(m,k)} - \Theta^*_{(m,k)}]' A_{(m,k)} [\Theta_{(m,k)} - \Theta^*_{(m,k)}]$. We apply similar arguments in the proof of Theorem 3.2 of Lu and Su (2015) to show that

$$\mathbb{W}_{i(m,k)} \leq \left(\bar{c}_A \bar{c}_B / \underline{c}_A^2\right) \|\tilde{\beta}_{i(m,k)}\|^2,$$

where $\tilde{\beta}_{i(m,k)} := \sqrt{n} \left[C_{(m,k)} C'_{(m,k)} \right]^{-1/2} C_{(m,k)} V_{(m,k)}^{-1/2} [\widehat{\Theta}_{i(m,k)} - \Theta^*_{(m,k)}] \stackrel{d}{\rightarrow} N(0, I_{l(m,k)})$. Let $c_{AB} := \bar{c}_A \bar{c}_B / \underline{c}_A^2$ and $\bar{l} := \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} l(m,k)$. Then, the above inequality for $\mathbb{W}_{i(m,k)}$ and the corrected version of Lemma 2.1 of Shibata (1981, 1982) imply that

$$\begin{aligned}
 &nKM \max_{1 \leq i \leq n} \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} P\left(\mathbb{W}_{i(m,k)} \geq 2nD(\delta_n)\right) \\
 &\leq nKM \max_{1 \leq i \leq n} \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} P\left(\|\tilde{\beta}_{i(m,k)}\|^2 \geq 2nD(\delta_n)/c_{AB}\right) \\
 &\leq \limsup_{n \rightarrow \infty} nKM \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} P\left(\chi^2(l(m,k)) \geq 2nD(\delta_n)/c_{AB}\right) \\
 &\leq \limsup_{n \rightarrow \infty} nKMP\left(\chi^2(\bar{l}) \geq 2nD(\delta_n)/c_{AB}\right) \\
 &\leq \limsup_{n \rightarrow \infty} nKMP\left(\chi^2(\bar{l}) \geq \bar{l} + \left(n\delta_n^2 \underline{c}_A / c_{AB} - \bar{l}\right)\right) \\
 &\leq \limsup_{n \rightarrow \infty} nKM \exp\left(-0.5\left(n\delta_n^2 \underline{c}_A / c_{AB} - \bar{l}\right)\left(1 - \log\left(n\delta_n^2 \underline{c}_A / (\bar{l}c_{AB})\right) / \left(n\delta_n^2 \underline{c}_A / (\bar{l}c_{AB}) - 1\right)\right)\right) \\
 &= o(1).
 \end{aligned}$$

For the last equality, note first that $\log(n\delta_n^2 \underline{c}_A / (\bar{l}c_{AB})) / (n\delta_n^2 \underline{c}_A / (\bar{l}c_{AB}) - 1) = o(1)$ by Assumption 3(ii). The leading term becomes

$$\begin{aligned}
 nKM \exp\left(-0.5\left(n\delta_n^2 \underline{c}_A / c_{AB}\right)\right) &= nKM n^{-0.5(L^2 K \underline{c}_A^3 / (\bar{c}_A \bar{c}_B))} \\
 &\ll KK! n^{1-0.5(L^2 K \underline{c}_A^3 / (\bar{c}_A \bar{c}_B))} \\
 &\ll KK^K n^{1-0.5(L^2 K \underline{c}_A^3 / (\bar{c}_A \bar{c}_B))} \\
 &= K^{K+1} n^{1-0.5(L^2 K \underline{c}_A^3 / (\bar{c}_A \bar{c}_B))} \\
 &\leq C(\log n)^{K+1} n^{1-0.5(L^2 K \underline{c}_A^3 / (\bar{c}_A \bar{c}_B))} \\
 &= o(1)
 \end{aligned}$$

where $C < \infty$ is a generic constant. The second line holds by the definition of M , the third line holds by the fact that $K! \ll K^K$, the fifth line holds by Assumption 3(ii), and the last convergence result holds by Assumption 3(ii) and by taking some large L .

Therefore, we establish the result in (i). Analogously, we can prove the result in (ii). \square

Proof of Theorem 3. Using the definition of $\widehat{y}(k)$ and $\widetilde{y}(k)$ and the triangular inequality, we have

$$\begin{aligned} & \max_{1 \leq k \leq K} |\widehat{y}(k) - \widetilde{y}(k)| \\ &= \max_{1 \leq k \leq K} \left| \frac{1}{M} \sum_{m=1}^M x'_{(m,k)} \widehat{\Theta}_{(m,k)} - \frac{1}{M_{max}} \sum_{m' \in \mathcal{M}_{max}} x'_{(m',k)} \widehat{\Theta}_{(m',k)} \right| \\ &= \max_{1 \leq k \leq K} \left| \frac{1}{M} \sum_{m=1}^M (x'_{(m,k)} \widehat{\Theta}_{(m,k)} - y^*) - \frac{1}{M_{max}} \sum_{m' \in \mathcal{M}_{max}} (x'_{(m',k)} \widehat{\Theta}_{(m',k)} - y^*) \right| \\ &\leq \max_{1 \leq k \leq K} \left| \frac{1}{M} \sum_{m=1}^M (x'_{(m,k)} \widehat{\Theta}_{(m,k)} - y^*) \right| \\ &\quad + \max_{1 \leq k \leq K} \left| \frac{1}{M_{max}} \sum_{m' \in \mathcal{M}_{max}} (x'_{(m',k)} \widehat{\Theta}_{(m',k)} - y^*) \right| \\ &\equiv \max_{1 \leq k \leq K} EQ_1 + \max_{1 \leq k \leq K} EQ_2. \end{aligned}$$

We first investigate EQ_1 :

$$\begin{aligned} \max_{1 \leq k \leq K} EQ_1 &\leq \max_{1 \leq k \leq K} \left| \frac{1}{M} \sum_{m=1}^M (x'_{(m,k)} \Theta_{(m,k)}^* - y_k^*) \right| \\ &\quad + \max_{1 \leq k \leq K} \left| \frac{1}{M} \sum_{m=1}^M x'_{(m,k)} (\widehat{\Theta}_{(m,k)} - \Theta_{(m,k)}^*) \right| \\ &\leq \max_{1 \leq k \leq K} \left| \frac{1}{M} \sum_{m=1}^M (x'_{(m,k)} \Theta_{(m,k)}^* - y_k^*) \right| \\ &\quad + \max_{1 \leq k \leq K} \frac{1}{M} \sum_{m=1}^M \|x_{(m,k)}\| \|\widehat{\Theta}_{(m,k)} - \Theta_{(m,k)}^*\| \\ &\leq \max_{1 \leq k \leq K} \left| \frac{1}{M} \sum_{m=1}^M (x'_{(m,k)} \Theta_{(m,k)}^* - y_k^*) \right| \\ &\quad + \left(\max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \|x_{(m,k)}\| \right) \left(\max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \|\widehat{\Theta}_{(m,k)} - \Theta_{(m,k)}^*\| \right) \\ &= o_p(1) + O_p(1) \cdot o_p(1) \\ &= o_p(1). \end{aligned}$$

The first inequality holds from the triangular inequality and the second one from the Cauchy–Schwartz inequality. The final inequality holds from the uniform convergence and boundedness assumptions, and Theorem 2(ii) above. Similarly, we can show $EQ_2 = o_p(1)$. \square

Proof of Theorem 4. It suffices to show that, with $\mathcal{K} := \{1, \dots, K_n\}$,

$$\sup_{k \in \mathcal{K}} \left| \frac{CV_n(k) - FPE_n(k)}{FPE_n(k)} \right| = o_p(1). \tag{12}$$

We first expand the numerator by applying Knight’s identity repeatedly.

$$\begin{aligned} & CV_n(k) - FPE_n(k) \\ &= \left\{ \frac{1}{n} \sum_{i=1}^n \left[\rho_\tau \left(y_i - M^{-1} \sum_{m=1}^M x'_{i(m,k)} \widehat{\Theta}_{i(m,k)} \right) - \rho_\tau(\varepsilon_i) \right] \right\} \\ & \quad - \{FPE_n(k) - E[\rho_\tau(\varepsilon)]\} + \frac{1}{n} \sum_{i=1}^n \{\rho_\tau(\varepsilon_i) - E[\rho_\tau(\varepsilon)]\} \\ &= \frac{1}{n} \sum_{i=1}^n \left[\mu_i - M^{-1} \sum_{m=1}^M x'_{i(m,k)} \widehat{\Theta}_{i(m,k)} \right] \psi_\tau(\varepsilon_i) \\ & \quad + \frac{1}{n} \sum_{i=1}^n \int_0^{M^{-1} \sum_{m=1}^M x'_{i(m,k)} \widehat{\Theta}_{i(m,k)} - \mu_i} [1\{\varepsilon_i \leq s\} - 1\{\varepsilon_i \leq 0\}] ds \\ & \quad - E \left[\int_0^{M^{-1} \sum_{m=1}^M x'_{(m,k)} \widehat{\Theta}_{(m,k)} - \mu} [1\{\varepsilon \leq s\} - 1\{\varepsilon \leq 0\}] ds | \mathcal{D}_n \right] \\ & \quad + \frac{1}{n} \sum_{i=1}^n \{\rho_\tau(\varepsilon_i) - E[\rho_\tau(\varepsilon)]\} \\ &= \frac{1}{n} \sum_{i=1}^n \left[\mu_i - M^{-1} \sum_{m=1}^M x'_{i(m,k)} \widehat{\Theta}_{i(m,k)} \right] \psi_\tau(\varepsilon_i) \\ & \quad + \frac{1}{n} \sum_{i=1}^n \int_0^{M^{-1} \sum_{m=1}^M x'_{i(m,k)} \widehat{\Theta}_{i(m,k)} - \mu_i} [1\{\varepsilon_i \leq s\} - 1\{\varepsilon_i \leq 0\} - F(s|x_i) + F(0|x_i)] ds \\ & \quad + \frac{1}{n} \sum_{i=1}^n \left\{ \int_0^{M^{-1} \sum_{m=1}^M x'_{i(m,k)} \widehat{\Theta}_{i(m,k)} - \mu_i} [F(s|x_i) - F(0|x_i)] ds \right. \\ & \quad \quad \left. - E_{x_i} \left[\int_0^{M^{-1} \sum_{m=1}^M x'_{i(m,k)} \widehat{\Theta}_{i(m,k)} - \mu_i} [F(s|x_i) - F(0|x_i)] ds \right] \right\} \\ & \quad + \frac{1}{n} \sum_{i=1}^n \left\{ E_x \left[\int_0^{M^{-1} \sum_{m=1}^M x'_{(m,k)} \widehat{\Theta}_{i(m,k)} - \mu} [F(s|x) - F(0|x)] ds \right] \right\} \end{aligned}$$

$$\begin{aligned}
 & - E_x \left[\int_0^{M^{-1} \sum_{m=1}^M x'_{(m,k)} \hat{\Theta}_{(m,k)}^{-\mu}} [F(s|x) - F(0|x)] ds \right] \Big\} \\
 & + \frac{1}{n} \sum_{i=1}^n \{ \rho_\tau(\varepsilon_i) - E[\rho_\tau(\varepsilon)] \} \\
 & \equiv CV_{1n} + CV_{2n} + CV_{3n} + CV_{4n} + CV_{5n}.
 \end{aligned}$$

It is straightforward to derive all terms except CV_{4n} . We need the following two results to get CV_{4n} . Let E_x be an expectation with respect to a random variable x .

$$\begin{aligned}
 & \frac{1}{n} \sum_{i=1}^n E_{x_i} \left[\int_0^{M^{-1} \sum_{m=1}^M x'_{i(m,k)} \hat{\Theta}_{i(m,k)}^{-\mu} } [F(s|x_i) - F(0|x_i)] ds \right] \\
 & = \frac{1}{n} \sum_{i=1}^n E_x \left[\int_0^{M^{-1} \sum_{m=1}^M w_m x'_{(m,k)} \hat{\Theta}_{i(m,k)}^{-\mu}} [F(s|x) - F(0|x)] ds \right], \tag{13}
 \end{aligned}$$

$$\begin{aligned}
 & E \left[\int_0^{M^{-1} \sum_{m=1}^M x'_{(m,k)} \hat{\Theta}_{(m,k)}^{-\mu}} [1\{\varepsilon \leq s\} - 1\{\varepsilon \leq 0\}] ds | D_n \right] \\
 & = E_x \left[\int_0^{M^{-1} \sum_{m=1}^M x'_{(m,k)} \hat{\Theta}_{(m,k)}^{-\mu}} [F(s|x) - F(0|x)] ds \right]. \tag{14}
 \end{aligned}$$

The identity (13) follows from the fact that $\hat{\Theta}_{i(m)}$ does not depend on the i th observation. The second result (14) comes from

$$\begin{aligned}
 & E \left[\int_0^{M^{-1} \sum_{m=1}^M x'_{(m,k)} \hat{\Theta}_{(m,k)}^{-\mu}} [1\{\varepsilon \leq s\} - 1\{\varepsilon \leq 0\}] ds | D_n \right] \\
 & = \int_{(x, \varepsilon)} \int_0^{M^{-1} \sum_{m=1}^M x'_{(m,k)} \hat{\Theta}_{(m,k)}^{-\mu}} [1\{\varepsilon \leq s\} - 1\{\varepsilon \leq 0\}] ds f(x, \varepsilon | D_n) dx d\varepsilon \\
 & = \int_{(x, \varepsilon)} \int_0^{M^{-1} \sum_{m=1}^M x'_{(m,k)} \hat{\Theta}_{(m,k)}^{-\mu}} [1\{\varepsilon \leq s\} - 1\{\varepsilon \leq 0\}] ds f(x, \varepsilon) dx d\varepsilon \\
 & = \int_x \int_0^{M^{-1} \sum_{m=1}^M x'_{(m,k)} \hat{\Theta}_{(m,k)}^{-\mu}} \int_\varepsilon [1\{\varepsilon \leq s\} - 1\{\varepsilon \leq 0\}] f(\varepsilon|x) f(x) d\varepsilon dx ds \\
 & = \int_x \int_0^{M^{-1} \sum_{m=1}^M x'_{(m,k)} \hat{\Theta}_{(m,k)}^{-\mu}} [F(s|x) - F(0|x)] f(x) dx ds \\
 & = E_x \left[\int_0^{M^{-1} \sum_{m=1}^M x'_{(m,k)} \hat{\Theta}_{(m,k)}^{-\mu}} [F(s|x) - F(0|x)] ds \right].
 \end{aligned}$$

The second equality holds by the independence of the sample $\{x_i, \varepsilon_i\}$ and the generic random variable (x, ε) .

We are now ready to prove (12). We first show that the denominator of (12) is uniformly bounded above zero and show the uniform convergence of CV_{1n}, \dots, CV_{5n} .

Claim 1: $\min_{k \in K} FPE_n(k) \geq E[\rho_\tau(\varepsilon)] - o_p(1)$. This result shows that the denominator of the LHS in (12) is bounded above zero. Let $u_k := \mu - M^{-1} \sum_{m=1}^M x'_{(m,k)} \Theta_{(m,k)}^*$.

$$\begin{aligned}
 & FPE_n(k) - E[\rho_\tau(\varepsilon + u_k)] \\
 &= E \left[\rho_\tau \left(\varepsilon + u_k - M^{-1} \sum_{m=1}^M x'_{(m,k)} \left(\widehat{\Theta}_{(m,k)} - \Theta_{(m,k)}^* \right) \right) - \rho_\tau(\varepsilon + u_k) \mid \mathcal{D}_n \right] \\
 &= E \left[\int_0^{M^{-1} \sum_{m=1}^M x'_{(m,k)} \left(\widehat{\Theta}_{(m,k)} - \Theta_{(m,k)}^* \right)} [1\{\varepsilon + u_k \leq s\} - 1\{\varepsilon + u_k \leq 0\}] ds \mid \mathcal{D}_n \right] \\
 &= E_x \left[\int_0^{M^{-1} \sum_{m=1}^M x'_{(m,k)} \left(\widehat{\Theta}_{(m,k)} - \Theta_{(m,k)}^* \right)} [F(s - u_k \mid x) - F(-u_k \mid x)] ds \right] \\
 &= E_x \left[\int_0^{M^{-1} \sum_{m=1}^M x'_{(m,k)} \left(\widehat{\Theta}_{(m,k)} - \Theta_{(m,k)}^* \right)} f(-u_k \mid x) s ds \right] + o_p(1) \\
 &= 2^{-1} E_x \left[f(-u_k \mid x) \left[M^{-1} \sum_{m=1}^M x'_{(m,k)} \left(\widehat{\Theta}_{(m,k)} - \Theta_{(m,k)}^* \right) \right]^2 \right] + o_p(1) \\
 &\leq 2^{-1} E_x \left[f(-u_k \mid x) M^{-1} \sum_{m=1}^M \left[x'_{(m,k)} \left(\widehat{\Theta}_{(m,k)} - \Theta_{(m,k)}^* \right) \right]^2 \right] + o_p(1) \\
 &= 2^{-1} \left\{ M^{-1} \sum_{m=1}^M \left(\widehat{\Theta}_{(m,k)} - \Theta_{(m,k)}^* \right)' E_x [f(-u_k \mid x) x_{(m,k)} x'_{(m,k)}] \left(\widehat{\Theta}_{(m,k)} - \Theta_{(m,k)}^* \right) \right\} \\
 &\quad + o_p(1) \\
 &\leq \frac{\bar{c}_A}{2} \max_{1 \leq k \leq K} \max_{1 \leq m \leq M} \|\widehat{\Theta}_{(m,k)} - \Theta_{(m,k)}^*\|^2 + o_p(1) = o_p(1).
 \end{aligned}$$

Claim 2: $\sup_{k \in \mathcal{K}} |CV_{1n}(k)| = o_p(1)$.

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

We first show that $\sup_{k \in \mathcal{K}} CV_{1n,1} = o_p(1)$. Let $b_{i(m,k)} = \mu_i - x'_{i(m,k)} \Theta_{i(m,k)}^*$ and $e_n = (MnK^2)^{1/4}$. Note that

$$\begin{aligned}
 &P\left(\max_{1 \leq k \leq K} |CV_{1n,1}| \geq 2\varepsilon\right) \\
 &\leq P\left(\max_{1 \leq k \leq K} \frac{1}{nM} \sum_{i=1}^n \sum_{m=1}^M |b_{i(m,k)}| \geq 2\varepsilon\right) \\
 &\leq P\left(\max_{1 \leq k \leq K} \frac{1}{nM} \sum_{i=1}^n \sum_{m=1}^M |b_{i(m,k)}| 1(|b_{i(m,k)}| \leq e_n) \geq \varepsilon\right) \\
 &\quad + P\left(\max_{1 \leq k \leq K} \frac{1}{nM} \sum_{i=1}^n \sum_{m=1}^M \left|\mu_i - x'_{i(m,k)} \Theta_{i(m,k)}^*\right| 1(|b_{i(m,k)}| > e_n) \geq \varepsilon\right) \\
 &\equiv CV_{1n,11} + CV_{1n,12}.
 \end{aligned}$$

We next show that $CV_{1n,11} = o(1)$ and $CV_{1n,12} = o(1)$, respectively.

$$\begin{aligned}
 CV_{1n,11} &\leq K \max_{1 \leq k \leq K} P\left(\frac{1}{nM} \sum_{i=1}^n \sum_{m=1}^M |b_{i(m,k)}| 1(|b_{i(m,k)}| \leq e_n) \geq \varepsilon\right) \\
 &\leq 2K \exp\left(-\frac{nM\varepsilon^2}{2KC + 2\varepsilon e_n/3}\right) \\
 &\leq 2 \exp\left(-\frac{nM\varepsilon^2}{2KC + 2\varepsilon e_n/3} + \log K\right) \\
 &= 2 \exp\left(-\frac{nM\varepsilon^2}{2KC + 2\varepsilon e_n/3} \left(1 - \frac{(2KC + 2\varepsilon e_n/3) \log K}{nM\varepsilon^2}\right)\right) = o(1).
 \end{aligned}$$

The last convergence result follows from the order conditions in Assumption 3.

$$\begin{aligned}
 CV_{1n,12} &\leq P\left(\max_{1 \leq k \leq K} \max_{1 \leq i \leq n} \max_{1 \leq m \leq M} |b_{i(m,k)}| > e_n\right) \\
 &\leq \sum_{k=1}^K \sum_{i=1}^n \sum_{m=1}^M P(|b_{i(m,k)}| > e_n) \\
 &\leq \frac{1}{e_n^4} \sum_{k=1}^K \sum_{i=1}^n \sum_{m=1}^M E\left[|b_{i(m,k)}|^4 1(|b_{i(m,k)}| > e_n^4)\right] = o(1).
 \end{aligned}$$

We next turn our attention to $CV_{1n,2}$:

$$\begin{aligned}
 \sup_{k \in \mathcal{K}} |CV_{1n,2}| &\leq \sup_{k \in \mathcal{K}} \frac{1}{nM} \sum_{i=1}^n \sum_{m=1}^M \left| x'_{i(m,k)} \left(\widehat{\Theta}_{i(m,k)} - \Theta_{i(m,k)}^* \right) \psi_{\tau}(\varepsilon_i) \right| \\
 &\leq \sup_{k \in \mathcal{K}} \frac{1}{nM} \sum_{i=1}^n \sum_{m=1}^M \left| x'_{i(m,k)} \left(\widehat{\Theta}_{i(m,k)} - \Theta_{i(m,k)}^* \right) \right| \\
 &\leq \sup_{k \in \mathcal{K}} \frac{1}{nM} \sum_{i=1}^n \sum_{m=1}^M \|x_{i(m,k)}\| \left\| \widehat{\Theta}_{i(m,k)} - \Theta_{i(m,k)}^* \right\|
 \end{aligned}$$

$$\begin{aligned}
 &= \sup_{k \in \mathcal{K}} \frac{1}{M} \sum_{m=1}^M \left\{ \frac{1}{n} \sum_{i=1}^n \|x_{i(m,k)}\| \left\| \widehat{\Theta}_{i(m,k)} - \Theta_{i(m,k)}^* \right\| \right\} \\
 &\leq \sup_{k \in \mathcal{K}} \frac{1}{M} \sum_{m=1}^M \left\{ \frac{1}{n} \sum_{i=1}^n \|x_{i(m,k)}\| \left(\max_{1 \leq i \leq n} \left\| \widehat{\Theta}_{i(m,k)} - \Theta_{i(m,k)}^* \right\| \right) \right\} \\
 &= \sup_{k \in \mathcal{K}} \frac{1}{M} \sum_{m=1}^M \left\{ \left(\max_{1 \leq i \leq n} \left\| \widehat{\Theta}_{i(m,k)} - \Theta_{i(m,k)}^* \right\| \right) \frac{1}{n} \sum_{i=1}^n \|x_{i(m,k)}\| \right\} \\
 &\leq \sup_{k \in \mathcal{K}} \frac{1}{M} \sum_{m=1}^M \max_{1 \leq m \leq M} \left\{ \left(\max_{1 \leq i \leq n} \left\| \widehat{\Theta}_{i(m,k)} - \Theta_{i(m,k)}^* \right\| \right) \frac{1}{n} \sum_{i=1}^n \|x_{i(m,k)}\| \right\} \\
 &= \sup_{k \in \mathcal{K}} \max_{1 \leq m \leq M} \left\{ \left(\max_{1 \leq i \leq n} \left\| \widehat{\Theta}_{i(m,k)} - \Theta_{i(m,k)}^* \right\| \right) \frac{1}{n} \sum_{i=1}^n \|x_{i(m,k)}\| \right\} \frac{1}{M} \sum_{m=1}^M 1 \\
 &= \sup_{k \in \mathcal{K}} \max_{1 \leq m \leq M} \left\{ \left(\max_{1 \leq i \leq n} \left\| \widehat{\Theta}_{i(m,k)} - \Theta_{i(m,k)}^* \right\| \right) \frac{1}{n} \sum_{i=1}^n \|x_{i(m,k)}\| \right\} \\
 &= \left\{ \sup_{k \in \mathcal{K}} \max_{1 \leq i \leq n} \max_{1 \leq m \leq M} \left\| \widehat{\Theta}_{i(m,k)} - \Theta_{i(m,k)}^* \right\| \right\} \left\{ \sup_{k \in \mathcal{K}} \max_{1 \leq m \leq M} \frac{1}{n} \sum_{i=1}^n \|x_{i(m,k)}\| \right\} \\
 &= O_p \left(\sqrt{n^{-1} K \log n} \right) O_p \left(\sqrt{K} \right) = o_p(1).
 \end{aligned}$$

The convergence results follow from Theorem 2, Lemma 2, and Assumption 3.

Claim 3: $\sup_{k \in \mathcal{K}} |CV_{2n}(k)| = o_p(1)$.

$$|CV_{2n}(k)| \leq |CV_{2n,1}(k)| + |CV_{2n,2}(k)|,$$

where

$$CV_{2n,1}(k) = \frac{1}{n} \sum_{i=1}^n \int_0^{M^{-1} \sum_{m=1}^M x'_{i(m,k)} \Theta_{(m,k)}^* - \mu_i} [1\{\varepsilon_i \leq s\} - 1\{\varepsilon_i \leq 0\} - F(s|x_i) + F(0|x_i)] ds$$

and

$$CV_{2n,2}(k) = \frac{1}{n} \sum_{i=1}^n \int_{M^{-1} \sum_{m=1}^M x'_{i(m,k)} \Theta_{(m,k)}^* - \mu_i}^{M^{-1} \sum_{m=1}^M x'_{i(m,k)} \widehat{\Theta}_{i(m,k)} - \mu_i} [1\{\varepsilon_i \leq s\} - 1\{\varepsilon_i \leq 0\} - F(s|x_i) + F(0|x_i)] ds.$$

Since $1\{\varepsilon_i \leq s\} - 1\{\varepsilon_i \leq 0\} - F(s|x_i) + F(0|x_i) \leq 2$, we have

$$|CV_{2n,1}(k)| \leq \frac{2}{nM} \sum_{i=1}^n \sum_{m=1}^M \left| x'_{i(m,k)} \Theta_{(m,k)}^* - \mu_i \right|.$$

Thus, $\sup_{k \in \mathcal{K}} |CV_{2n,1}(k)| = o_p(1)$ follows from the same arguments used for $CV_{1n,1}$ above.

We next investigate $CV_{2n,2}$. We have

$$\begin{aligned} \sup_{k \in \mathcal{K}} |CV_{2n,2}(k)| &\leq \sup_{k \in \mathcal{K}} \frac{2}{nM} \sum_{i=1}^n \sum_{m=1}^M \left| x'_{i,(m,k)} (\widehat{\Theta}_{i(m,k)} - \Theta_{(m,k)}^*) \right| \\ &\leq \sup_{k \in \mathcal{K}} \frac{2}{nM} \sum_{i=1}^n \sum_{m=1}^M \|x_{i,(m,k)}\| \left\| \widehat{\Theta}_{i(m,k)} - \Theta_{(m,k)}^* \right\| \\ &\leq 2 \sup_{k \in \mathcal{K}} \max_{1 \leq i \leq n} \max_{1 \leq m \leq M} \left\| \widehat{\Theta}_{i(m,k)} - \Theta_{(m,k)}^* \right\| \sup_{k \in \mathcal{K}} \max_{1 \leq m \leq M} \frac{1}{n} \sum_{i=1}^n \|x_{i,(m,k)}\| \\ &= O_p \left(\sqrt{n^{-1} K \log n} \right) \cdot O_p \left(\sqrt{K} \right) = o_p(1). \end{aligned}$$

Claim 4: $\sup_{k \in \mathcal{K}} |CV_{3n}(k)| = o_p(1)$.

$$|CV_{3n}(k)| \leq |CV_{3n,1}(k)| + |CV_{3n,2}(k)|,$$

where

$$\begin{aligned} CV_{3n,1}(k) &= \frac{1}{n} \sum_{i=1}^n \left\{ \int_0^{M^{-1} \sum_{m=1}^M x'_{i(m,k)} \Theta_{(m,k)}^* - \mu_i} [F(s|x_i) - F(0|x_i)] ds \right. \\ &\quad \left. - E_{x_i} \left[\int_0^{M^{-1} \sum_{m=1}^M x'_{i(m,k)} \Theta_{(m,k)}^* - \mu_i} [F(s|x_i) - F(0|x_i)] ds \right] \right\} \end{aligned}$$

and

$$\begin{aligned} CV_{3n,2}(k) &= \frac{1}{n} \sum_{i=1}^n \left\{ \int_{M^{-1} \sum_{m=1}^M x'_{i(m,k)} \Theta_{(m,k)}^* - \mu_i}^{M^{-1} \sum_{m=1}^M x'_{i(m,k)} \widehat{\Theta}_{i(m,k)} - \mu_i} [F(s|x_i) - F(0|x_i)] ds \right. \\ &\quad \left. - E_{x_i} \left[\int_{M^{-1} \sum_{m=1}^M x'_{i(m,k)} \Theta_{(m,k)}^* - \mu_i}^{M^{-1} \sum_{m=1}^M x'_{i(m,k)} \widehat{\Theta}_{i(m,k)} - \mu_i} [F(s|x_i) - F(0|x_i)] ds \right] \right\}. \end{aligned}$$

The proof of $\sup_{k \in \mathcal{K}} |CV_{3n,1}| = o_p(1)$ is similar to that of $\sup_{k \in \mathcal{K}} |CV_{1n,1}| = o_p(1)$ in Claim 2 and is omitted. Note that

$$\begin{aligned} |CV_{3n,2}(k)| &\leq \frac{1}{n} \sum_{i=1}^n \left| M^{-1} \sum_{m=1}^M x'_{i(m,k)} \left(\widehat{\Theta}_{i(m,k)} - \Theta_{(m,k)}^* \right) \right| \\ &\quad \left| \frac{1}{n} \sum_{i=1}^n E_{x_i} \left| M^{-1} \sum_{m=1}^M x'_{i(m,k)} \left(\widehat{\Theta}_{i(m,k)} - \Theta_{(m,k)}^* \right) \right| \right| \\ &\equiv CV_{3n,21}(k) + CV_{3n,22}(k). \end{aligned}$$

The proof of $\sup_{k \in \mathcal{K}} |CV_{3n, 21}| = o_p(1)$ is similar to that of $\sup_{k \in \mathcal{K}} |CV_{2n, 2}| = o_p(1)$ in Claim 3 and is also omitted. It remains to show that $\sup_{k \in \mathcal{K}} |CV_{3n, 22}| = o_p(1)$. Note that

$$\begin{aligned} & \sup_{k \in \mathcal{K}} CV_{3n, 22}(k) \\ & \leq \sup_{k \in \mathcal{K}} \frac{1}{nM} \sum_{i=1}^n \sum_{m=1}^M E_{x_i} \left| x'_{i(m,k)} \left(\widehat{\Theta}_{i(m,k)} - \Theta_{(m,k)}^* \right) \right| \\ & = \sup_{k \in \mathcal{K}} \frac{1}{nM} \sum_{i=1}^n \sum_{m=1}^M E_{x_i} \left[\left(\widehat{\Theta}_{i(m,k)} - \Theta_{(m,k)}^* \right)' x_{i(m,k)} x'_{i(m,k)} \left(\widehat{\Theta}_{i(m,k)} - \Theta_{(m,k)}^* \right) \right]^{1/2} \\ & \leq \sup_{k \in \mathcal{K}} \frac{1}{nM} \sum_{i=1}^n \sum_{m=1}^M \left[\left(\widehat{\Theta}_{i(m,k)} - \Theta_{(m,k)}^* \right)' E_{x_i} \left[x_{i(m,k)} x'_{i(m,k)} \right] \left(\widehat{\Theta}_{i(m,k)} - \Theta_{(m,k)}^* \right) \right]^{1/2} \\ & \leq \sup_{k \in \mathcal{K}} \frac{1}{nM} \sum_{i=1}^n \sum_{m=1}^M \left[\lambda_{\max} \left(E_{x_i} \left[x_{i(m,k)} x'_{i(m,k)} \right] \right) \right]^{1/2} \left\| \widehat{\Theta}_{i(m,k)} - \Theta_{(m,k)}^* \right\| \\ & \leq \left(\max_{k \in \mathcal{K}} \max_{1 \leq m \leq M} \left[\lambda_{\max} \left(E_{x_i} \left[x_{i(m,k)} x'_{i(m,k)} \right] \right) \right]^{1/2} \right) \\ & \quad \times \left(\max_{1 \leq i \leq n} \max_{k \in \mathcal{K}} \max_{1 \leq m \leq M} \left\| \widehat{\Theta}_{i(m,k)} - \Theta_{(m,k)}^* \right\| \right) \\ & = o_p(1) \end{aligned}$$

by the triangle inequality, $|x| = (x^2)^{1/2}$, Jensen's inequality, $A'BA \leq \lambda_{\max}(B)A'A$ for any real symmetric matrix B .

Claim 5: $\sup_{k \in \mathcal{K}} |CV_{4n}(k)| = o_p(1)$.

$$\begin{aligned} |CV_{4n}(k)| & = \left| \frac{1}{n} \sum_{i=1}^n E_x \left[\int_{M^{-1} \sum_{m=1}^M x'_{(m,k)} \widehat{\Theta}_{(m,k)} - \mu}^{M^{-1} \sum_{m=1}^M x'_{(m,k)} \widehat{\Theta}_{i(m,k)} - \mu} [F(s|x) - F(0|x)] ds \right] \right| \\ & \leq \frac{1}{n} \sum_{i=1}^n E_x \left| \int_{M^{-1} \sum_{m=1}^M x'_{(m,k)} \widehat{\Theta}_{(m,k)} - \mu}^{M^{-1} \sum_{m=1}^M x'_{(m,k)} \widehat{\Theta}_{i(m,k)} - \mu} [F(s|x) - F(0|x)] ds \right| \\ & \leq \frac{1}{n} \sum_{i=1}^n E_x \left| \left(M^{-1} \sum_{m=1}^M x'_{(m,k)} \widehat{\Theta}_{i(m,k)} - \mu \right) - \left(M^{-1} \sum_{m=1}^M x'_{(m,k)} \widehat{\Theta}_{(m,k)} - \mu \right) \right| \\ & = \frac{1}{n} \sum_{i=1}^n E_x \left| M^{-1} \sum_{m=1}^M x'_{(m,k)} \left(\widehat{\Theta}_{i(m,k)} - \widehat{\Theta}_{(m,k)} \right) \right| \\ & = o_p(1) \end{aligned}$$

by the triangle inequality, $F(s|x) - F(0|x) \leq 1$, and the similar arguments in the proof of $\sup_{k \in \mathcal{K}} |CV_{3n, 22}(k)| = o_p(1)$.

Claim 6: $CV_{5n} = o_p(1)$. Since CV_{5n} does not depend on k , this result follows from the weak law of large numbers. □

Proof of Theorem 5. (i) There exists \tilde{w} between \bar{w} and w^* such that

$$\begin{aligned} F(\bar{w}) &= F(w^*) + \nabla_1 F(w^*)'(\bar{w} - w^*) + \frac{1}{2}(\bar{w} - w^*)'\nabla_2 F(\tilde{w})(\bar{w} - w^*) \\ &= F(w^*) - \tilde{\lambda} \cdot 1'_M(\bar{w} - w^*) + \frac{1}{2}(\bar{w} - w^*)'\nabla_2 F(\tilde{w})(\bar{w} - w^*) \\ &= F(w^*) + \frac{1}{2}(\bar{w} - w^*)'\nabla_2 F(\tilde{w})(\bar{w} - w^*), \end{aligned}$$

where $\tilde{\lambda}$ is a Lagrange multiplier from the constraint optimization problem:

$$w^* = \arg \max_{w \in \mathbb{R}^M} F(w) + \tilde{\lambda} \cdot (1'_M w - 1). \tag{15}$$

Note that the second equality above comes from the first-order condition for w^* and that the third equality holds by the normalization, $1'_M w = 1$, for any weight w . We investigate the upper bound of the quadratic term:

$$\begin{aligned} \frac{1}{2}(\bar{w} - w^*)'\nabla_2 F(\tilde{w})(\bar{w} - w^*) &= 2^{-1}(\tilde{w}'\nabla_2 F(\tilde{w})\bar{w} - 2\tilde{w}'\nabla_2 F(\tilde{w})w^* + w^{*'}\nabla_2 F(\tilde{w})w^*) \\ &\equiv 2^{-1}(I + II + III). \end{aligned}$$

Since $\nabla_2 F(\tilde{w})$ is an $M \times M$ symmetric matrix, we can factorize it as $S\Lambda S'$, where Λ is a diagonal matrix composed of the eigenvalues $\{\lambda_m\}$ and S is composed of the corresponding orthonormal eigenvectors $\{s_m\}$. Note that

$$\begin{aligned} I &= M^{-2}1'_M \nabla_2 F(\tilde{w})1_M \\ &= M^{-2}1'_M \left(\sum_{m=1}^M \lambda_m s_m s'_m \right) 1_M \\ &\leq M^{-2} \tilde{\lambda}_{max} \|1_M\|^2 \\ &= M^{-1} \tilde{\lambda}_{max}, \\ II &\leq 2|\tilde{w}'\nabla_2 F(\tilde{w})w^*| \\ &= 2M^{-1} \left| 1'_M \left(\sum_{m=1}^M \lambda_m s_m s'_m \right) w^* \right| \\ &\leq 2M^{-1} \tilde{\lambda}_{max} |1'_M w^*| \\ &= 2M^{-1} \tilde{\lambda}_{max}, \\ III &= w^{*'} \left(\sum_{m=1}^M \lambda_m s_m s'_m \right) w^* \\ &\leq \tilde{\lambda}_{max} \|w^*\|^2 \\ &\leq \tilde{\lambda}_{max} \|w^*\|_1^2 \\ &= \tilde{\lambda}_{max}, \end{aligned}$$

where $\|\cdot\|_1$ denotes ℓ_1 -norm. Therefore, we have

$$\frac{1}{2}(\tilde{w} - w^*)' \nabla_2 F(\tilde{w})(\tilde{w} - w^*) \leq 2^{-1} \bar{\lambda}_{max} \left(1 + 3M^{-1}\right),$$

which establishes the desired result.

(ii) Using the similar arguments above, we have

$$\begin{aligned} F(\hat{w}) &= F(w^*) + 2^{-1} \hat{\eta}' \nabla_2 F(\tilde{w}) \hat{\eta} \\ &= F(w^*) + 2^{-1} \hat{\eta}' \left(\sum_{m=1}^M \tilde{\lambda}_m \tilde{s}_m \tilde{s}_m' \right) \hat{\eta} \\ &\leq F(w^*) + 2^{-1} \bar{\lambda}_{max} \|\hat{\eta}\|^2, \end{aligned}$$

where $\hat{\eta}$, \tilde{w} , $\tilde{\lambda}_m$, and \tilde{s}_m are all random objects in the second equality. The third inequality holds almost surely by the definition of $\bar{\lambda}_{max}$. We establish the desired result by noting that $E \|\hat{\eta}\|^2 \leq M \bar{\sigma}_{\eta}^2$. □

Proof of Corollary 6. Following similar arguments in Theorem 5, we only need to show that $w^{*'} \Sigma w^* \rightarrow 0$ as $M \rightarrow \infty$. Note that we have a closed-form solution $w^* = (1_M' \Sigma^{-1} 1_M)^{-1} \Sigma^{-1} 1_M$ for the optimization problem in (15). Then, we have

$$w^{*'} \Sigma w^* = (1_M' \Sigma^{-1} 1_M)^{-1}.$$

Abusing notation on eigenvalues/eigenvectors, we have

$$\begin{aligned} 1_M' \Sigma^{-1} 1_M &= 1_M' \left(\sum_{m=1}^M \lambda_m^{-1} s_m s_m' \right) 1_M \\ &\geq \bar{\lambda}_{max}^{-1} \|1_M\|^2 \\ &= \bar{\lambda}_{max}^{-1} M, \end{aligned}$$

which diverges to infinity as M increases. Therefore, $(1_M' \Sigma^{-1} 1_M)^{-1} \leq \bar{\lambda}_{max} M^{-1} \rightarrow 0$ as $M \rightarrow \infty$. □

REFERENCES

Adrian, T., N. Boyarchenko, & D. Giannone (2019) Vulnerable growth. *American Economic Review* 109(4), 1263–1289.

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

Angrist, J., V. Chernozhukov, & I. Fernández-Val (2006) Quantile regression under misspecification, with an application to the US wage structure. *Econometrica* 74(2), 539–563.

Belloni, A. & V. Chernozhukov (2011) ℓ_1 -penalized quantile regression in high-dimensional sparse models. *Annals of Statistics* 39(1), 82–130.

Bhatia, K.T., G.A. Vecchi, T.R. Knutson, H. Murakami, J. Kossin, K.W. Dixon, & C.E. Whitlock (2019) Recent increases in tropical cyclone intensification rates. *Nature Communications* 10(1), 1–9.

Breiman, L. (1996) Bagging predictors. *Machine Learning* 24(2), 123–140.

Buchinsky, M. (1998) The dynamics of changes in the female wage distribution in the USA: A quantile regression approach. *Journal of Applied Econometrics* 13(1), 1–30.

- Campbell, J.Y. & S.B. Thompson (2007) Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies* 21(4), 1509–1531.
- Claeskens, G., J.R. Magnus, A.L. Vasnev, & W. Wang (2016) The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting* 32(3), 754–762.
- Clemen, R.T. (1989) Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting* 5(4), 559–583.
- Donald, S.G. & W.K. Newey (2001) Choosing the number of instruments. *Econometrica* 69(5), 1161–1191.
- Duffie, D. & J. Pan (1997) An overview of value at risk. *Journal of Derivatives* 4(3), 7–49.
- Elliott, G. (2011) Averaging and the optimal combination of forecasts. Manuscript, Department of Economics, University of California, San Diego.
- Elliott, G., A. Gargano, & A. Timmermann (2013) Complete subset regressions. *Journal of Econometrics* 177(2), 357–373.
- Elliott, G., A. Gargano, & A. Timmermann (2015) Complete subset regressions with large-dimensional sets of predictors. *Journal of Economic Dynamics and Control* 54, 86–110.
- Fan, R. & J.H. Lee (2019) Predictive quantile regressions under persistence and conditional heteroskedasticity. *Journal of Econometrics* 213(1), 261–280.
- Fazekas, I. & O. Klesov (2001) A general approach to the strong law of large numbers. *Theory of Probability and Its Applications* 45(3), 436–449.
- Hansen, B.E. (2007) Least squares model averaging. *Econometrica* 75(4), 1175–1189.
- Hansen, B.E. & J.S. Racine (2012) Jackknife model averaging. *Journal of Econometrics* 167(1), 38–46.
- Hirano, K. & J.H. Wright (2021) Analyzing cross-validation for forecasting with structural instability. *Journal of Econometrics*, forthcoming.
- Koenker, R. (2005) Quantile Regression. *Econometric Society Monographs*, vol. 38. Cambridge University Press.
- Koenker, R. & G. Bassett (1978) Regression quantiles. *Econometrica* 46(1), 33–50.
- Komunjer, I. (2013). Quantile prediction. In G. Elliot and A. Timmerman (eds.), *Handbook of Economic Forecasting*, vol. 2, pp. 961–994. Elsevier.
- Kuersteiner, G. & R. Okui (2010) Constructing optimal instruments by first-stage prediction averaging. *Econometrica* 78(2), 697–718.
- Lee, J.H. (2016) Predictive quantile regression with persistent covariates: IVX-QR approach. *Journal of Econometrics* 192(1), 105–118.
- Lee, S. and Y. Shin (2021) Complete subset averaging with many instruments. *The Econometrics Journal*, 24(2), 290–314.
- Li, K.-C. (1987) Asymptotic optimality for C_p , C_L , cross-validation and generalized cross-validation: Discrete index set. *Annals of Statistics* 15(3), 958–975.
- Lu, X. & L. Su (2015) Jackknife model averaging for quantile regressions. *Journal of Econometrics* 188(1), 40–58.
- Meinshausen, N. (2006) Quantile regression forests. *Journal of Machine Learning Research* 7(35), 983–999.
- Meligkotsidou, L., E. Panopoulou, I.D. Vrontos, & S.D. Vrontos (2019) Quantile forecast combinations in realised volatility prediction. *Journal of the Operational Research Society* 70(10), 1720–1733.
- Meligkotsidou, L., E. Panopoulou, I.D. Vrontos, & S.D. Vrontos (2021) Out-of-sample equity premium prediction: A complete subset quantile regression approach. *The European Journal of Finance* 27(1–2), 110–135.
- Phillips, P.C. (2015) Halbert White Jr. memorial JFEC lecture: Pitfalls and possibilities in predictive regression. *Journal of Financial Econometrics* 13(3), 521–555.
- Portnoy, S. (1984) Asymptotic behavior of m -estimators of p regression parameters when p^2/n is large. I. Consistency. *Annals of Statistics* 12(4), 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(4), 1403–1417.

- Rapach, D.E., J.K. Strauss, & G. Zhou (2010) Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies* 23(2), 821–862.
- Rice, J. (1984) Bandwidth choice for nonparametric regression. *Annals of Statistics* 12(4), 1215–1230.
- Shibata, R. (1981) An optimal selection of regression variables. *Biometrika* 68(1), 45–54.
- Shibata, R. (1982) Amendments and corrections: An optimal selection of regression variables. *Biometrika* 69(2), 492–492.
- Smith, J. & K.F. Wallis (2009) A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics* 71(3), 331–355.
- Stock, J.H. & M.W. Watson (2004) Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting* 23(6), 405–430.