Bagging Binary and Quantile Predictors for Time Series*

Tae-Hwy Lee†
Department of Economics
University of California, Riverside
Riverside, CA 92521
tae.lee@ucr.edu

Yang Yang
Department of Economics
University of California, Riverside
Riverside, CA 92521
yang.yang@email.ucr.edu

First Version: February 29, 2004
This Version: April 25, 2005

Abstract

Bootstrap aggregating or Bagging, introduced by Breiman (1996a), has been proved to be effective to improve on unstable forecast. Theoretical and empirical works using classification, regression trees, variable selection in linear and non-linear regression have shown that bagging can generate substantial prediction gain. However, most of the existing literature on bagging has been limited to the cross sectional circumstances with symmetric cost functions. In this paper, we extend the application of bagging to time series settings with asymmetric cost functions, particularly for predicting signs and quantiles. We use quantile predictions to construct a binary predictor and the majority-voted bagging binary prediction. We show that bagging may improve the binary prediction in small sample, but it does not improve in large sample. Various bagging forecast combination weights are used such as equal weighted and Bayesian model averaging (BMA) weighted combinations. For demonstration, we present results from Monte Carlo experiments and from empirical applications using monthly S&P500 and NASDAQ stock index returns.

Key Words: Asymmetric cost function, Bagging, Binary prediction, BMA, Forecast combination, Majority voting, Quantile prediction, Time Series.

JEL Classification: C3, C5, G0.

*We would like to thank an anonymous referee, Graham Elliott, Chuan Goh, Clive Granger, Huiyu Huang, Yongmiao Hong, Atsushi Inoue, Lutz Kilian, Sinich Sakata, as well as seminar participants at the conference in honor of Professor Clive Granger, Far Eastern ES2004, Canadian CESG2004, USC, UBC, and Victoria, for their very useful comments. All remaining errors are our own.

†Corresponding author. Phone: +1 (951) 827-1509. Fax: +1 (951) 827-5685.

1 Introduction

To improve forecasts over individual forecasting models, combining forecasts has been suggested. Bates and Granger (1969), Granger, Deutsch and Teräsvirta (1994), Granger and Jeon (2004), Stock and Watson (1999, 2005), Yang (2004), and Timmermann (2005) show that forecast combinations can improve forecast accuracy over a single model. Combining forecasts diversifies risk of forecast errors, analogous to investing on portfolios rather than individual securities. On the other hand, to improve forecasts of a given model, combination can also be formed over a set of training sets. While usually we have a single training set, it can be replicated via bootstrap. Combining forecasts trained over the bootstrap-replicated training sets is the idea of *b*ootstrap *agg*regat*ing* (or bagging), introduced by Breiman (1996a).

In this paper we examine how bagging may improve binary predictions and quantile predictions. It is well known that, while financial returns $\{Y_t\}$ may not be predictable, their variance, sign, and quantiles may be predictable. Christofferson and Diebold (2003) show that binary variable $G_{t+1} \equiv \mathbf{1}(Y_{t+1} > 0)$ is predictable when some conditional moments are time varying, where $\mathbf{1}(\cdot)$ takes the value of 1 if the statement in the parenthesis is true, and 0 otherwise. Hong and Lee (2003), Hong and Chung (2003), Linton and Whang (2004), Pesaran and Timmermann (2002a), among many others, find some evidence that the directions of stock returns and foreign exchange rate changes are predictable. While there remains much work to be done in the literature, many theoretical and numerical works have shown that bagging is effective to improve forecasts.[1]

Bagging is a device to improve the accuracy of unstable predictors. A predictor is said to be unstable if perturbing the training sample can cause significant change in the predictor (Breiman 1996b). Bagging smooths instabilities by averaging over bootstrap predictors and thus lowering predictors' sensitivity to training samples. It has been shown that bagging is effective thanks to the variance reduction stemming from the averaging and so it is most effective if the volatility of the predictor is very high, as is the case for highly nonlinear models (Friedman and Hall 2000, Buja and Stuetzle 2002, Bühlmann and Yu 2002).

However, while classifier prediction (binary prediction as a special case) has been empirically shown to be very successful in the machine learning literature, it has not been analytically explained. As noticed by some researchers bagging may outperform unbagged predictor, but this may not be always the case. The aim of this paper is to show how and why bagging binary prediction may work or fail to work. To demonstrate our analytical results, Monte Carlo experiments and empirical applications are also presented. We construct binary predictors based on quantile predictors so that binary and quantile predictions are linked. Various

[1] Most bagging research has been in statistics and engineering. Recent bagging research in econometrics includes Kitamua (2001), Inoue and Kilian (2005), and Stock and Watson (2005).

bagging forecast combinations with equal weights and weights based on Bayesian model averaging (BMA) are considered.

There are some limitations in the already substantial bagging literature. One is that most of the existing bagging literature is dealing with independent data. With the obvious dynamic structures in most economic and financial variables, it is important to see how bagging works for time series. One concern of applying bagging to time series is whether a bootstrap can provide a sound simulation sample for dependent data. Fortunately, it has been shown that the block bootstrap can provide consistent densities for moment estimators and quantile estimators (Fitzenberger 1997). Another limitation of current bagging research is the use of symmetric cost functions (such as mean squared forecast error) as prediction evaluation criteria. However, it is widely accepted that asymmetric cost functions are more relevant (Granger 1969, 1999a, 1999b, 2002; Granger and Pesaran 2000; Elliott *et al.* 2003a, 2003b). That is, our utility changes differently with positive and negative forecast error, or with false-alert and failure-to-alert. In this paper, we analyze bagging binary predictors formed via majority voting, for weakly dependent time series, under asymmetric cost functions.

The plan of this paper is as follows. Section 2 explains ensemble aggregating predictor and bootstrap aggregating predictor. We show how ensemble aggregating predictor can improve predictive ability of a conditional mean model for *time series* under a symmetric L_2-cost function. In Section 3 we set up a binary prediction problem based on utility maximization behavior of an economic agent. We introduce a way to form a binary predictor through a quantile predictor. In Section 4, extending the results in Section 2, we show that the ensemble aggregating predictors can improve predictive ability of the conditional quantile model and conditional binary model for time series under asymmetric L_1-cost functions. In Section 5 we show how bagging for binary predictions works as the training sample size grows. In Section 6 we examine bagging for quantile predictions. Section 7 presents Monte Carlo experiments. Section 8 reports empirical results on binary and quantile predictions using the monthly returns in S&P500 and NASDAQ stock indexes. Section 9 concludes.

2 Aggregating Predictor

2.1 Ensemble Aggregating Predictor $\varphi_A(\mathbf{X}_t)$

To improve forecasts of a model, a combination can be formed over a set of training sets. Let

$$\mathcal{D}_t \equiv \{(Y_s, \mathbf{X}_{s-1})\}_{s=t-R+1}^{t} \qquad (t = R, \ldots, T),$$

be a training set at time t, and $\varphi(\mathbf{X}_t, \mathcal{D}_t)$ be a forecast of Y_{t+1} using this training set $\mathcal{D}_t$ and input vector $\mathbf{X}_t$. Suppose each training set $\mathcal{D}_t$ consists of R observations drawn from *strictly stationary* probability

distribution $\mathbf{P}$. Our mission is to use the $\mathcal{D}_t$ to get a better predictor than the single training set predictor $\varphi(\mathbf{X}_t, \mathcal{D}_t)$. Ideally, if $\mathbf{P}$ is known and multiple training sets $\mathcal{D}_t^{(j)}$ $(j = 1, \ldots, J)$ may be drawn from $\mathbf{P}$, an obvious procedure is to replace $\varphi(\mathbf{X}_t, \mathcal{D}_t)$ by the weighted average of $\varphi(\mathbf{X}_t, \mathcal{D}_t^{(j)})$ over j, i.e.,

$$\varphi_A(\mathbf{X}_t) \equiv \mathbb{E}_{\mathcal{D}_t}\varphi(\mathbf{X}_t,\mathcal{D}_t) \equiv \sum_{j=1}^{J} w_{j,t}\varphi(\mathbf{X}_t, \mathcal{D}_t^{(j)}), \tag{1}$$

where $\mathbb{E}_{\mathcal{D}_t}(\cdot)$ denotes the expectation over $\mathbf{P}$, $w_{j,t}$ is the weight function with $\sum_{j=1}^{J} w_{j,t} = 1$, and the subscript A in φ_A denotes "aggregation". We call $\varphi_A(\mathbf{X}_t)$ the ensemble aggregating predictor.

From the first sight, it may be hard to understand the meaning of multiple training set $\mathcal{D}_t^{(j)}$ in the time series circumstances since time is not repeatable. However, considering an example of the estimation and forecast procedure with panel data may be helpful. Suppose we want to forecast consumption of a household in next period. When the historical observations of the interested household is very limited, our parameters estimated and the predictors will have rather large variances, especially for non-linear regression models. If we can find some other households that have similar consumption patterns (similar underlying probability distribution $\mathbf{P}$), it would be better to use historical observations from all similar households than just from this interested household in the estimation process, though we only use data of this interested households to do forecast. Therefore, the ensemble aggregating predictor is just like to find similar households.

We now show that $\varphi_A(\mathbf{X}_t)$ has no larger mean squared forecast error than $\varphi(\mathbf{X}_t, \mathcal{D}_t)$, extending Breiman's (1996, p. 129) proof for the time series case:

Proposition 1. *The ensemble aggregating predictor $\varphi_A(\mathbf{X}_t)$ has no larger symmetric L_2-cost of the forecast error than the original predictor $\varphi(\mathbf{X}_t, \mathcal{D}_t)$, i.e.,*

$$\mathbb{E}_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t}(Y_{t+1} - \varphi(\mathbf{X}_t, \mathcal{D}_t))^2 \geq \mathbb{E}_{Y_{t+1}, \mathbf{X}_t}[(Y_{t+1} - \varphi_A(\mathbf{X}_t))^2], \tag{2}$$

where $E_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t}(\cdot) \equiv E_{\mathbf{X}_t}[E_{Y_{t+1}|\mathbf{X}_t}\{E_{\mathcal{D}_t}(\cdot)\,|X_t\}]$ denotes the expectations taken over the training set $\mathcal{D}_t$ first conditioning on Y_{t+1} and $\mathbf{X}_t$, then taking an expectation of Y_{t+1} conditioning on $\mathbf{X}_t$, and finally taking an expectation of $\mathbf{X}_t$, and similarly $\mathbb{E}_{Y_{t+1}, \mathbf{X}_t}(\cdot) \equiv \mathbb{E}_{\mathbf{X}_t}[\mathbb{E}_{Y_{t+1}|\mathbf{X}_t}(\cdot)\,|\mathbf{X}_t]$.

Proof: Let $e_{t+1} = Y_{t+1} - \varphi(\mathbf{X}_t, \mathcal{D}_t)$. Taking expectations on the squared forecast error $c(e_{t+1}) = e_{t+1}^2$ gives

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t}(Y_{t+1} - \varphi(\mathbf{X}_t, \mathcal{D}_t))^2 &= \mathbb{E}_{Y_{t+1}, \mathbf{X}_t}\left[Y_{t+1}^2 - 2Y_{t+1}\mathbb{E}_{\mathcal{D}_t}\varphi(\mathbf{X}_t, \mathcal{D}_t) + \mathbb{E}_{\mathcal{D}_t}\varphi^2(\mathbf{X}_t, \mathcal{D}_t)\right] \\
&\geq \mathbb{E}_{Y_{t+1}, \mathbf{X}_t}\left[Y_{t+1}^2 - 2Y_{t+1}\mathbb{E}_{\mathcal{D}_t}\varphi(\mathbf{X}_t, \mathcal{D}_t) + (\mathbb{E}_{\mathcal{D}_t}\varphi(\mathbf{X}_t, \mathcal{D}_t))^2\right] \\
&= \mathbb{E}_{Y_{t+1}, \mathbf{X}_t}\left[(Y_{t+1} - \varphi_A(\mathbf{X}_t))^2\right],
\end{aligned} \tag{3}$$

where we recall $\varphi_A(\mathbf{X}_t) \equiv \mathbb{E}_{\mathcal{D}_t}\varphi(\mathbf{X}_t, \mathcal{D}_t)$, and the inequality comes from (conditional on the values of $\mathbf{X}_t$)

$$\mathbb{E}_{\mathcal{D}_t}\varphi^2(\mathbf{X}_t, \mathcal{D}_t) \geq (\mathbb{E}_{\mathcal{D}_t}\varphi(\mathbf{X}_t, \mathcal{D}_t))^2. \tag{4}$$

■

We see that aggregating over training sets will lower the expected cost. How much this aggregating predictor can improve depends on the variance of the prediction function

$$\mathbb{V}_{\mathcal{D}_t}[\varphi(\mathbf{X}_t, \mathcal{D}_t)] \equiv \mathbb{E}_{\mathcal{D}_t}\left[\varphi(\mathbf{X}_t, \mathcal{D}_t) - \mathbb{E}_{\mathcal{D}_t}\varphi(\mathbf{X}_t, \mathcal{D}_t)\right]^2 = \mathbb{E}_{\mathcal{D}_t}\varphi^2(\mathbf{X}_t, \mathcal{D}_t) - (\mathbb{E}_{\mathcal{D}_t}\varphi(\mathbf{X}_t, \mathcal{D}_t))^2. \tag{5}$$

Therefore, the effect of instability is clear. A predictor is said to be unstable if perturbing the training sample can cause significant changes in the predictor (Breiman 1996b), i.e., $\mathbb{V}_{\mathcal{D}_t}\left[\varphi(\mathbf{X}_t, \mathcal{D}_t)\right]$ is large.

2.2 Bootstrap Aggregating Predictor $\varphi_B(\mathbf{X}_t|\mathcal{D}_t)$

In reality, $\mathbf{P}$ is unknown, and we only have a single training set. In this case, $\mathbf{P}$ may be estimated by its empirical distribution of a given $\mathcal{D}_t$, from which multiple training sets may be drawn by bootstrap method. A combined forecast can then be formed using the bootstrap-replicated training sets. Take bootstrap samples $\{\mathcal{D}_t^{*(j)}\}_{j=1}^J$ from the empirical distribution $\hat{\mathbf{P}}(\mathcal{D}_t)$ of $\mathcal{D}_t$, and form $\{\varphi(\mathbf{X}_t, \mathcal{D}_t^{*(j)})\}$. Therefore, the ensemble aggregating predictor $\varphi_A(\mathbf{X}_t)$ can be approximated by

$$\varphi_B(\mathbf{X}_t|\mathcal{D}_t) \equiv \mathbb{E}_{\mathcal{D}_t^*}\varphi(\mathbf{X}_t, \mathcal{D}_t^*) \equiv \sum_{j=1}^{J} w_{j,t}\varphi(\mathbf{X}_t, \mathcal{D}_t^{*(j)}). \tag{6}$$

We call $\varphi_B(\mathbf{X}_t|\mathcal{D}_t)$ the bootstrap aggregating or bagging predictor. Note that $\varphi_A(\mathbf{X}_t)$ does not depend on the training set $\mathcal{D}_t$ because an expectation has been taken over $\mathbf{P}$, but $\varphi_B(\mathbf{X}_t|\mathcal{D}_t)$ still depends on the training set $\mathcal{D}_t$ since all the bootstrap training samples are drawn from $\hat{\mathbf{P}}(\mathcal{D}_t)$.

Bagging is a device to improve the accuracy of unstable predictors. The unstableness of predictors may come from many sources – discrete choice regressions (e.g. decision trees and classification problem), non-smooth target functions (e.g. median or quantile function involving indicator function), small sample size, outliers or extreme values, etc. Under these circumstances the model uncertainty in the prediction function φ and/or parameter estimation uncertainty using the training sample $\mathcal{D}_t$ may render an unstable prediction.

Why does bagging work? Bagging is effective thanks to the variance reduction stemming from averaging predictors and so it is most effective if the volatility of the predictors is very high. Friedman and Hall (2000) and Buja and Stuetzle (2002) decompose a predictor or an estimator into linear and higher order parts by Taylor-expansion. They show that bagging reduces the variance for the higher order nonlinear component by replacing it with an estimate of its expected value, while leaving the linear part unaffected. Therefore bagging works most successfully with highly nonlinear estimators such as decision trees and neural network.

Bühlmann and Yu (2002) consider the cases that bagging can transform a hard-thresholding function into a soft-thresholding function and thus decrease the instabilities of predictors. However, the relevance of this argument depends on how the bagging predictor is formed. As discussed in Breiman (1996a), bagging predictors can be formed via voting instead of averaging. If the target variable of interest is continuous (as for quantile prediction in this paper), we can form a bagging predictor by an average over bootstrap predictors. However, if the target variable takes only discrete values, then a voting scheme over the discrete values is to be called for from the bootstrap predictors (as for binary prediction in this paper). Both averaging and voting can be weighted. Bagging transforms a hard-thresholding function into a soft-thresholding function only for averaged-bagging, but not for the voted-bagging because a voted-bagging predictor will remain as a binary hard-thresholding function.

Even so, the ability of the voted-bagging to stabilize classifier prediction has been proved to be very successful in the machine learning literature (Bauer and Kohavi 1999, Kuncheva and Whitaker 2003, and Evgeniou et al. 2004). The method for voting classification or voted-bagging is now an established research area known under different names in the literature – combining classifiers, classifier ensembles, committees of learners, a team of classifiers, consensus of learners, mixture of experts, etc. In the next three sections, we attempt to understand how voted-bagging (bagging binary prediction) may work and when it fails.

3 Binary Prediction

3.1 Cost function

Granger (2002) notes that a conditional risk measure of financial return Y_{t+1} that has a conditional predictive distribution $P_{Y_{t+1}}(y|\mathbf{X}_t) = \Pr(Y_{t+1} < y|\mathbf{X}_t)$ may be written as

$$R(\mathbf{X}_t) = A_1 \int_0^{\infty} |y-m|^p dP_{Y_{t+1}}(y|\mathbf{X}_t) + A_2 \int_{-\infty}^{0} |y-m|^p dP_{Y_{t+1}}(y|\mathbf{X}_t), \tag{7}$$

with A_1, A_2 both > 0 and some $p > 0$, where m is the predictor of y that minimizes the risk. One problem raised here is how to choose optimal L_p-norm in empirical works. Granger (2002) considers the absolute return ($p = 1$) as a preferable measure given that returns from the stock market are known to come from a distribution with particularly long tails. In particular, Granger (2002) refers to a trio of papers (Nyguist 1983, Money *et al.* 1982, Harter 1977) who find that the optimal $p = 1$ from Laplace and Cauchy distribution, $p = 2$ for Gaussian and $p = \infty$ (min/max estimator) for a rectangular distribution. Granger (2002) also notes that in terms of the kurtosis κ, Harter (1977) suggests to use $p = 1$ for $\kappa > 3.8$; $p = 2$ for $2.2 \leq \kappa \leq 3.8$; and $p = 3$ for $\kappa < 2.2$. In finance, the kurtosis of returns can be thought of as being well over 4 and so

$p = 1$ is preferred. In this paper, we follow Granger (2002) to consider binary and quantile predictions with $A_1 \neq A_2$ and $p = 1$.

We consider the asymmetric risk function to discuss a binary prediction. Let $G_{t+1} \equiv \mathbf{1}(Y_{t+1} > 0)$. To define the asymmetric risk with $A_1 \neq A_2$ and $p = 1$, we consider binary decision problem of Granger and Pesaran (2000) with the following 2×2 payoff or utility matrix:

Utility	$G_{t+1} = 1$	$G_{t+1} = 0$
$G_{t,1}(\mathbf{X}_t) = 1$	u_{11}	u_{01}
$G_{t,1}(\mathbf{X}_t) = 0$	u_{10}	u_{00}

(8)

where u_{ij} is the utility when $G_{t,1}(\mathbf{X}_t) = j$ is predicted and $G_{t+1} = i$ is realized $(i, j = 1, 2)$. Assume $u_{11} > u_{10}$ and $u_{00} > u_{01}$, and u_{ij} are constant over time. $(u_{11} - u_{10}) > 0$ is the utility gain from taking correct forecast when $G_{t,1}(\mathbf{X}_t) = 1$, and $(u_{00} - u_{01}) > 0$ is the utility gain from taking correct forecast when $G_{t,1}(\mathbf{X}_t) = 0$. Denote

$$\pi(\mathbf{X}_t) \equiv \mathbb{E}_{Y_{t+1}}(G_{t+1}|\mathbf{X}_t) = \Pr(G_{t+1} = 1|\mathbf{X}_t). \tag{9}$$

The expected utility of $G_{t,1}(\mathbf{X}_t) = 1$ is $u_{11}\pi(\mathbf{X}_t) + u_{01}(1 - \pi(\mathbf{X}_t))$, and the expected utility of $G_{t,1}(\mathbf{X}_t) = 0$ is $u_{10}\pi(\mathbf{X}_t) + u_{00}(1 - \pi(\mathbf{X}_t))$. Hence, to maximize utility, conditional on the values of $\mathbf{X}_t$, the prediction $G_{t,1}(\mathbf{X}_t) = 1$ will be made if

$$u_{11}\pi(\mathbf{X}_t) + u_{01}(1 - \pi(\mathbf{X}_t)) > u_{10}\pi(\mathbf{X}_t) + u_{00}(1 - \pi(\mathbf{X}_t)),$$

or

$$\pi(\mathbf{X}_t) > \frac{(u_{00} - u_{01})}{(u_{11} - u_{10}) + (u_{00} - u_{01})} \equiv 1 - \alpha.$$

Proposition 2 (Granger and Pesaran, 2000). *The optimal predictor that can maximize expected utility is:*

$$G^{\dagger}_{t,1}(\mathbf{X}_t) = \mathbf{1}(\pi(\mathbf{X}_t) > 1 - \alpha). \tag{10}$$

■

By making correct prediction, our net utility gain is $(u_{00} - u_{01})$ when $G_{t+1} = 0$, and $(u_{11} - u_{10})$ when $G_{t+1} = 1$. We can put it in another way, our opportunity cost (in the sense that you lose the gain) of wrong prediction is $(u_{00} - u_{01})$ when $G_{t+1} = 0$ and $(u_{11} - u_{10})$ when $G_{t+1} = 1$. Since a multiple of a utility function represents the same preference, $(1 - \alpha)$ can be viewed as the utility-gain from correct prediction when $G_{t+1} = 0$, or the opportunity cost of a false-alert. Similarly,

$$\alpha \equiv \frac{(u_{11} - u_{10})}{(u_{11} - u_{10}) + (u_{00} - u_{01})} \tag{11}$$

can be treated as the utility-gain from correct prediction when $G_{t+1} = 1$ is realized, or the opportunity cost of a failure-to-alert. We thus can define a cost function $c(e_{t+1})$ with $e_{t+1} = G_{t+1} - G_{t,1}(\mathbf{X}_t)$:

Cost	$G_{t+1} = 1$	$G_{t+1} = 0$
$G_{t,1}(\mathbf{X}_t) = 1$	0	$1 - \alpha$
$G_{t,1}(\mathbf{X}_t) = 0$	α	0

That is

$$c(e_{t+1}) = \begin{cases} \alpha & \text{if} \quad e_{t+1} = \quad 1 \\ 1-\alpha & \text{if} \quad e_{t+1} = -1 \\ 0 & \text{if} \quad e_{t+1} = \quad 0 \end{cases},$$

which can be equivalently written as $c(e_{t+1}) = \rho_\alpha(e_{t+1})$, where

$$\rho_\alpha(z) \equiv [\alpha - \mathbf{1}(z < 0)]z \tag{12}$$

is the *check function* of Koenker and Basset (1978). Hence we have obtained the following result.

Proposition 3. *The optimal binary predictor $G^\dagger_{t,1}(X_t) = 1(\pi(\mathbf{X}_t) > 1 - \alpha)$ maximizing the expected utility (obtained in Proposition 2) minimizes the expected cost $\mathbb{E}_{Y_{t+1}}(\rho_\alpha(e_{t+1})|\mathbf{X}_t)$.*

Proof: From Proposition 2 the optimal predictor that can maximize expected utility is $G^\dagger_{t,1}(X_t) = 1(\pi(X_t) > 1 - \alpha)$. We need to show that this minimizes the expected cost. The cost function may be written as

$$c(e_{t+1}) = \alpha G_{t+1} + (1 - \alpha - G_{t+1})G_{t,1}(\mathbf{X}_t). \tag{13}$$

and the expected cost (or the conditional risk) is

$$\mathbb{E}_{Y_{t+1}}(c(e_{t+1})|\mathbf{X}_t) = \alpha\pi(\mathbf{X}_t) + [1 - \alpha - \pi(\mathbf{X}_t)]G_{t,1}(\mathbf{X}_t). \tag{14}$$

When $\pi(\mathbf{X}_t) > 1 - \alpha$, the minimizer of $\mathbb{E}_{Y_{t+1}}(c(e_{t+1})|\mathbf{X}_t)$ is $G^\dagger_{t,1}(\mathbf{X}_t) = 1$. When $\pi(\mathbf{X}_t) < 1 - \alpha$, the minimizer of $\mathbb{E}_{Y_{t+1}}(c(e_{t+1})|\mathbf{X}_t)$ is $G^\dagger_{t,1}(\mathbf{X}_t) = 0$. Hence, $G^\dagger_{t,1}(X_t) = 1(\pi(X_t) > 1 - \alpha)$ is the minimizer of the expected cost. ■

In other words, the optimal binary prediction that minimizes $\mathbb{E}_{Y_{t+1}}(\rho_\alpha(e_{t+1})|\mathbf{X}_t)$ is the conditional α-quantile of G_{t+1}, denoted as

$$G^\dagger_{t,1}(\mathbf{X}_t) = Q^\dagger_\alpha(G_{t+1}|\mathbf{X}_t) = \arg\min_{G_{t,1}(\mathbf{X}_t)} \mathbb{E}_{Y_{t+1}}(\rho_\alpha(G_{t+1} - G_{t,1}(\mathbf{X}_t))|\mathbf{X}_t). \tag{15}$$

This is the maximum score problem of Manski (1975, 1985), Manski and Thompson (1989), and Kordas (2005).

Also, as noted by Powell (1986), using the fact that for any monotonic function $h(\cdot)$, $Q_\alpha(h(Y_{t+1})|\mathbf{X}_t) = h(Q_\alpha(Y_{t+1}|\mathbf{X}_t))$, which follows immediately from observing that $\Pr(Y_{t+1} < y|\mathbf{X}_t) = \Pr[h(Y_{t+1}) < h(y)|\mathbf{X}_t]$,

and noting that the indicator function is monotonic, $Q_\alpha(G_{t+1}|\mathbf{X}_t) = Q_\alpha(\mathbf{1}(Y_{t+1} > 0)|\mathbf{X}_t) = \mathbf{1}(Q_\alpha(Y_{t+1}|\mathbf{X}_t) > 0)$. Hence,

$$G^\dagger_{t,1}(\mathbf{X}_t) = \mathbf{1}(Q^\dagger_\alpha(Y_{t+1}|\mathbf{X}_t) > 0). \tag{16}$$

where $Q_\alpha(Y_{t+1}|\mathbf{X}_t)$ is the α-quantile function of Y_{t+1} conditional on $\mathbf{X}_t$. Note that $Q^\dagger_\alpha(G_{t+1}|\mathbf{X}_t) = \arg\min \mathbb{E}_{Y_{t+1}}(\rho_\alpha(e_{t+1})|\mathbf{X}_t)$ with $e_{t+1} \equiv G_{t+1} - Q_\alpha(G_{t+1}|\mathbf{X}_t)$, and $Q^\dagger_\alpha(Y_{t+1}|\mathbf{X}_t) = \arg\min \mathbb{E}_{Y_{t+1}}(\rho_\alpha(u_{t+1})|\mathbf{X}_t)$ with $u_{t+1} \equiv Y_{t+1} - Q_\alpha(Y_{t+1}|\mathbf{X}_t)$.

Proposition 3 indicates that the optimal binary prediction can be made from binary quantile regression for G_{t+1} as shown in equation (15). Binary prediction can also be made from a binary function of the α-quantile for Y_{t+1} in equation (16). In the next section where we consider bagging binary predictors, we choose to use equation (16) instead of equation (15) due to the following reasons.

First, Manski (1975, 1985) and Kim and Pollard (1990) show that parameter estimators obtained from minimizing $\sum \rho_\alpha(e_{t+1})$ has a slower $n^{1/3}$-convergence rate and has a non-normal limiting distribution. This is unattractive for our subsequent analysis (Proposition 5) in Section 5, where we need asymptotic normality. Instead, minimizing $\sum \rho_\alpha(u_{t+1})$ gives $n^{1/2}$-consistency and asymptotic normality. See Kim and White (2003), Chernozhukov and Umantsev (2001), and Komunjer (2005). Second, quantile-based binary regression models allow more structure than the maximum score regression models. We may have much more information than just an indicator. In the maximum score literature, Y_{t+1} is usually latent and unobservable. In our case, however, Y_{t+1} is not latent but observable. Information could be lost when one reduces Y_{t+1} to binary data. If the sufficient statistics are functions of binary variable G_{t+1} then there would be no information loss. If the sufficient statistics are functions that cannot be written as binary data there is information loss if binary variables are used and hence using the more informative variable Y_{t+1} in quantile regression may give better prediction than just using binary variable G_{t+1}. Finally, quantile itself is a very interesting topic. Quantile prediction is not only used in generating binary prediction, but also quantile itself is often the objective of interests in finance and economics, e.g., Value-at-Risk (VaR). For this reason, in addition to bagging binary predictors, we also consider bagging quantile predictors in Section 6.

3.2 Training binary predictor

We consider a simple univariate polynomial quantile regression function of Chernozhukov and Umantsev (2001):

$$Q_\alpha(Y_{t+1}|\mathbf{X}_t) = \tilde{\mathbf{X}}'_t\boldsymbol{\beta}_\alpha, \tag{17}$$

with $\mathbf{X}_t = Y_t$, $\tilde{\mathbf{X}}_t = (1\ Y_t\ Y_t^2)'$, and $\boldsymbol{\beta}_\alpha = [\beta_{\alpha,1}\ \beta_{\alpha,2}\ \beta_{\alpha,3}]'$. Denote the optimal binary predictor as

$$G^\dagger_{t,1}(\mathbf{X}_t) \equiv \mathbf{1}(Q^\dagger_\alpha(Y_{t+1}|\mathbf{X}_t) > 0) = \mathbf{1}(\tilde{\mathbf{X}}_t'\boldsymbol{\beta}^\dagger_\alpha > 0),$$

where $\boldsymbol{\beta}^\dagger_\alpha \equiv \arg\min_{\boldsymbol{\beta}_\alpha} \mathbb{E}_{Y_{t+1}}(\rho_\alpha(u_{t+1})|\mathbf{X}_t)$ is the pseudo-true value of the parameter β_α that minimizes the expected out-of-sample cost.

We estimate $\boldsymbol{\beta}^\dagger_\alpha$ recursively using the "rolling" in-sample of size R. Suppose there are $T+1$ ($\equiv R+P$) observations in total. We use the most recent R observations available at time t, $R \le t < T+1$, as a training sample $\mathcal{D}_t \equiv \{(Y_s, \mathbf{X}_{s-1})\}_{s=t-R+1}^t$. We then generate P ($= T+1-R$) one-step-ahead forecasts for the remaining validation sample. For each time t in the P prediction periods, we use a rolling training sample $\mathcal{D}_t$ of size R to estimate model parameters

$$\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t) \equiv \arg\min_{\boldsymbol{\beta}_\alpha} R^{-1} \sum_{s=t-R+1}^{t} \rho_\alpha(u_s), \qquad t = R, \ldots, T, \tag{18}$$

where $u_s \equiv Y_s - Q_\alpha(Y_s|\mathbf{X}_{s-1}) = Y_s - \tilde{\mathbf{X}}_{s-1}'\boldsymbol{\beta}_\alpha$.

We can then generate a sequence of one-step-ahead forecast $\hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t) = \tilde{\mathbf{X}}_t'\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t)$ and

$$\hat{G}_{t,1}(\mathbf{X}_t, \mathcal{D}_t) \equiv \mathbf{1}(\hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t) > 0) = \mathbf{1}(\tilde{\mathbf{X}}_t'\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t) > 0), \qquad t = R, \ldots, T. \tag{19}$$

Note that the notation, $\hat{G}_{t,1}(\mathbf{X}_t, \mathcal{D}_t)$, is to indicate the parameter estimation uncertainty in $\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t)$ due to the training of the unknown parameter $\boldsymbol{\beta}_\alpha$ using the training sample $\mathcal{D}_t$.

4 Ensemble Aggregating Predictors for Binary and Quantile Variables

We now extend Proposition 1 and show how the ensemble aggregating predictor can improve the predictive ability of the conditional quantile model and conditional binary model in time series under asymmetric L_1-cost functions:

Proposition 4. *(a) The ensemble aggregating binary predictor $\varphi_A(\mathbf{X}_t) = \mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}(\mathbf{X}_t, \mathcal{D}_t)$ has no larger asymmetric L_1-cost than the original predictor $\varphi(\mathbf{X}_t, \mathcal{D}_t) = \hat{G}_{t,1}(\mathbf{X}_t, \mathcal{D}_t)$, i.e.,*

$$\mathbb{E}_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t}\left[c(G_{t+1} - \hat{G}_{t,1}(\mathbf{X}_t, \mathcal{D}_t))\right] \ge \mathbb{E}_{Y_{t+1}, \mathbf{X}_t}\left[c(G_{t+1} - \mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}(\mathbf{X}_t, \mathcal{D}_t))\right].$$

(b) The ensemble aggregating quantile predictor $\varphi_A(\mathbf{X}_t) = \mathbb{E}_{\mathcal{D}_t}\hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t)$ has no larger asymmetric L_1-cost than the original predictor $\varphi(\mathbf{X}_t, \mathcal{D}_t) = \hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t)$, i.e.,

$$\mathbb{E}_{\mathcal{D}_t, Y_{t+1}, \mathbf{X}_t}\left[c(Y_{t+1} - \hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t))\right] \ge \mathbb{E}_{Y_{t+1}, \mathbf{X}_t}\left[c(Y_{t+1} - \mathbb{E}_{\mathcal{D}_t}\hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t))\right].$$

Proof: See Proposition 1 for the notation of $\mathbb{E}_{\mathcal{D}_t,Y_{t+1},\mathbf{X}_t}(\cdot)$ and $\mathbb{E}_{Y_{t+1},\mathbf{X}_t}(\cdot)$. Since both $\hat{G}_{t,1}(\mathbf{X}_t,\mathcal{D}_t)$ and $\hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t,\mathcal{D}_t)$ are quantile predictors, we prove just for (b). The proof of (a) is similar. Using (12), the cost function for the original predictor $\varphi(\mathbf{X}_t,\mathcal{D}_t)=\hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t,\mathcal{D}_t)$ is

$$c(Y_{t+1}-\varphi(\mathbf{X}_t,\mathcal{D}_t))=[\alpha\mathbf{1}(Y_{t+1}\geq\varphi(\mathbf{X}_t,\mathcal{D}_t))+(1-\alpha)\mathbf{1}(Y_{t+1}<\varphi(\mathbf{X}_t,\mathcal{D}_t))]\cdot|Y_{t+1}-\varphi(\mathbf{X}_t,\mathcal{D}_t)|,$$

and the cost function for the ensemble aggregating predictor $\varphi_A(\mathbf{X}_t)=\mathbb{E}_{\mathcal{D}_t}\hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t,\mathcal{D}_t)$ is

$$c(Y_{t+1}-\varphi_A(\mathbf{X}_t))=[\alpha\mathbf{1}(Y_{t+1}\geq\varphi_A(\mathbf{X}_t))+(1-\alpha)\mathbf{1}(Y_{t+1}<\varphi_A(\mathbf{X}_t))]\cdot|Y_{t+1}-\varphi_A(\mathbf{X}_t)|.$$

With some algebra, it can be shown that

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_t}c(Y_{t+1}-\varphi(\mathbf{X}_t,\mathcal{D}_t)) &= \mathbb{E}_{\mathcal{D}_t}c(Y_{t+1}-\varphi_A(\mathbf{X}_t))+\mathbb{E}_{\mathcal{D}_t}\left(Y_{t+1}-\varphi(\mathbf{X}_t,\mathcal{D}_t)\right)\mathbf{1}\left(Y_{t+1}>\varphi(\mathbf{X}_t,\mathcal{D}_t)\right)\\ &\quad+\mathbb{E}_{\mathcal{D}_t}\left(\varphi(\mathbf{X}_t,\mathcal{D}_t)-Y_{t+1}\right)\mathbf{1}\left(Y_{t+1}<\varphi(\mathbf{X}_t,\mathcal{D}_t)\right),\end{aligned}$$

where the first term $\mathbb{E}_{\mathcal{D}_t}c(Y_{t+1}-\varphi_A(\mathbf{X}_t))=c(Y_{t+1}-\varphi_A(\mathbf{X}_t))$, and the second and the third terms are non-negative. Therefore, $\mathbb{E}_{\mathcal{D}_t}c(Y_{t+1}-\varphi(\mathbf{X}_t,\mathcal{D}_t))\geq c(Y_{t+1}-\varphi_A(\mathbf{X}_t))$. Taking expectations with respect to Y_{t+1} and $\mathbf{X}_t$ gives the desired result. ■

Actually, for any convex cost function $c(\cdot)$, we will have

$$\mathbb{E}_{\mathcal{D}_t,Y_{t+1},\mathbf{X}_t}c(\hat{z}_{t+1})\geq\mathbb{E}_{Y_{t+1},\mathbf{X}_t}c(\mathbb{E}_{\mathcal{D}_t}(\hat{z}_{t+1}))$$

where $\mathbb{E}_{\mathcal{D}_t}(\hat{z}_{t+1})$ is the aggregating forecast error, and $\hat{z}_{t+1}$ is either $\hat{e}_{t+1}=G_{t+1}-\hat{G}_{t,1}(\mathbf{X}_t,\mathcal{D}_t)$ or $\hat{u}_{t+1}=Y_{t+1}-\hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t,\mathcal{D}_t)$. Therefore, the aggregating predictor will always have no larger expected cost than the original predictor for convex cost functions.

5 Bootstrap Aggregating Predictor for Binary Variable

Proposition 4 (for the ensemble aggregating predictors) is derived based on the assumption that we can infinitely draw random samples from the true data generating process $\mathbf{P}$. In practice, we do not have such luxury. Given a training set $\mathcal{D}_t$ at time t, the predictor can be formed from the bootstrap training sets drawn from the empirical distribution $\hat{\mathbf{P}}(\mathcal{D}_t)$ of $\mathcal{D}_t$. If the target variable is continuous as for the stock returns Y_{t+1} or for its quantiles $Q_\alpha(Y_{t+1}|\mathbf{X}_t)$, bagging procedure is to take an average of the forecasts from the bootstrap training samples. If the target variable is a class as for binary variable G_{t+1}, then a method of aggregating the bootstrap-trained forecasts is a voting.

One of the most representative unstable predictor studied in bagging literature is the classifier predictor. What we focus here is a basic case (two classes) of classification problem – binary prediction. Since all multi-classification problems can be decomposed into many binary predictions, our analysis on binary prediction can be easily extended to multi-classifier problems. In this section we first discuss how to form the voted-bagging binary predictor $\hat{G}^B_{t,1}(\mathbf{X}_t|\mathcal{D}_t)$ (defined below) and then we compare it with the original unbagged predictor $\hat{G}_{t,1}(\mathbf{X}_t, \mathcal{D}_t)$.

5.1 How to form bagging binary predictor

The procedure of bagging for binary predictors can be conducted in the following steps:

1. Given a training set of data at time t, $\mathcal{D}_t \equiv \{(Y_s, \mathbf{X}_{s-1})\}_{s=t-R+1}^{t}$, construct the jth bootstrap sample $\mathcal{D}_t^{*(j)} \equiv \{(Y_s^{*(j)}, \mathbf{X}_{s-1}^{*(j)})\}_{s=t-R+1}^{t}$, $j = 1, \ldots, J$, according to the empirical distribution of $\hat{\mathbf{P}}(\mathcal{D}_t)$ of $\mathcal{D}_t$.

2. For each j, estimate $\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)}) \equiv \arg\min_{\boldsymbol{\beta}_\alpha} R^{-1} \sum_{s=t-R+1}^{t} \rho_\alpha(Y_s^{*(j)} - \tilde{\mathbf{X}}_{s-1}^{*(j)\prime}\boldsymbol{\beta}_\alpha)$, $t = R, \ldots, T$.

3. For each j, compute the bootstrap binary predictor

$$\hat{G}_{t,1}^{*(j)}(\mathbf{X}_t, \mathcal{D}_t^{*(j)}) \equiv \mathbf{1}(\hat{Q}_\alpha^{*(j)}(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t^{*(j)}) > 0) = \mathbf{1}(\tilde{\mathbf{X}}_t'\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)}) > 0). \tag{20}$$

   Note that here we use $\tilde{\mathbf{X}}_t$ instead of $\tilde{\mathbf{X}}_t^{*(j)}$ so that only the parameter $\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})$ is trained on the bootstrap samples $\mathcal{D}_t^{*(j)}$, but the forecast is formed using the original predictor variables $\tilde{\mathbf{X}}_t$.

4. Construct an average over the J bootstrap predictors

$$\mathbb{E}_{\mathcal{D}_t^*}\hat{G}_{t,1}^*(\mathbf{X}_t, \mathcal{D}_t^*) = \sum_{j=1}^{J} \hat{w}_{j,t}\hat{G}_{t,1}^{*(j)}(\mathbf{X}_t, \mathcal{D}_t^{*(j)})$$

   with $\sum_{j=1}^{J} \hat{w}_{j,t} = 1$. We will discuss how to decide $\hat{w}_{j,t}$ in calculating both binary bagging predictors and quantile bagging predictors in Appendix. $\mathbb{E}_{\mathcal{D}_t^*}(\cdot)$ denotes the expectation over $\hat{\mathbf{P}}(\mathcal{D}_t)$, that is the average over the bootstrap training samples $\mathcal{D}_t^*$. This is a key step in bagging to smooth out the instability of the predictor due to the parameter estimation (training or learning). The weights $\hat{w}_{j,t}$ is typically set equal to J^{-1}, but can be computed via a Bayesian approach (see Appendix). J is often chosen in the range of 50, depending on sample size and on the computational cost to evaluate the predictor. See Breiman (1996a, Section 6.2). Our Monte Carlo results and empirical results reported in Sections 7 and 8 suggest $J = 50$ is more than sufficient, and even $J = 20$ is often good enough.

5. Take a majority voting over the J bootstrap predictors, i.e.,

$$\hat{G}_{t,1}^B(\mathbf{X}_t|\mathcal{D}_t) \equiv \mathbf{1}(\mathbb{E}_{\mathcal{D}_t^*}\hat{G}_{t,1}^*(\mathbf{X}_t, \mathcal{D}_t^*) > 1/2).$$

5.2 Evaluating bagging binary predictor

We compare the expected cost of bagging binary predictor and the original unbagged binary predictor. We need to find out the expected cost of these predictors.

First, the minimum possible expected cost is that of the optimal predictor $G^{\dagger}_{t,1}(\mathbf{X}_t)$. See Proposition 3. Plugging $G^{\dagger}_{t,1}(\mathbf{X}_t)$ into the conditional risk function in (14), we have

$$\mathbb{E}_{Y_{t+1}}(c(e^{\dagger}_{t+1})|\mathbf{X}_t) = \alpha\pi(\mathbf{X}_t) + [1-\alpha-\pi(\mathbf{X}_t)]G^{\dagger}_{t,1}(\mathbf{X}_t), \tag{21}$$

where $e^{\dagger}_{t+1} \equiv G_{t+1} - G^{\dagger}_{t,1}(\mathbf{X}_t)$. As defined in equation (7), the conditional expectation $\mathbb{E}_{Y_{t+1}}(\cdot|\mathbf{X}_t)$ is taken over $P_{Y_{t+1}}(y|\mathbf{X}_t)$, the conditional distribution of Y_{t+1} given the values of the predictor $\mathbf{X}_t$. Recall that $G_{t+1} \equiv \mathbf{1}(Y_{t+1} > 0)$ is a transform of Y_{t+1}.

Once the unknown parameters are trained, the conditional risks of unbagged predictor and bagging predictors (for a given value of $\mathbf{X}_t$) can be written as

$$\mathbb{E}_{\mathcal{D}_t,Y_{t+1}}(c(\hat{e}_{t+1})|\mathbf{X}_t) = \alpha\pi(\mathbf{X}_t) + [1-\alpha-\pi(\mathbf{X}_t)]\mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}(\mathbf{X}_t,\mathcal{D}_t) \tag{22}$$

$$\mathbb{E}_{\mathcal{D}_t,Y_{t+1}}(c(\hat{e}^{B}_{t+1})|\mathbf{X}_t) = \alpha\pi(\mathbf{X}_t) + [1-\alpha-\pi(\mathbf{X}_t)]\mathbb{E}_{\mathcal{D}_t}\hat{G}^{B}_{t,1}(\mathbf{X}_t|\mathcal{D}_t) \tag{23}$$

where $\hat{e}_{t+1} \equiv G_{t+1} - \hat{G}_{t,1}(\mathbf{X}_t,\mathcal{D}_t)$ and $\hat{e}^{B}_{t+1} \equiv G_{t+1} - \hat{G}^{B}_{t,1}(\mathbf{X}_t|\mathcal{D}_t)$.

From (21), (22), and (23), it is easy to see the following result. Comparison of the predictive ability of unbagged predictor and bagging predictor can be done by comparing the conditional risks in (22) and (23). Conditional on the values of $\mathbf{X}_t$, if $\mathbb{E}_{\mathcal{D}_t,Y_{t+1}}(c(\hat{e}_{t+1})|\mathbf{X}_t) > \mathbb{E}_{\mathcal{D}_t,Y_{t+1}}(c(\hat{e}^{B}_{t+1})|\mathbf{X}_t)$, we say bagging "works" for binary predictor. Hence, when $1-\alpha-\pi(\mathbf{X}_t) > 0$, (i.e., $G^{\dagger}_{t,1}(\mathbf{X}_t) = 0$), bagging works if $\mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}(\mathbf{X}_t,\mathcal{D}_t) > \mathbb{E}_{\mathcal{D}_t}\hat{G}^{B}_{t,1}(\mathbf{X}_t|\mathcal{D}_t) \geq G^{\dagger}_{t,1}(\mathbf{X}_t) = 0$. When $1-\alpha-\pi(\mathbf{X}_t) < 0$, (i.e., $G^{\dagger}_{t,1}(\mathbf{X}_t) = 1$), bagging works if $\mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}(\mathbf{X}_t,\mathcal{D}_t) < \mathbb{E}_{\mathcal{D}_t}\hat{G}^{B}_{t,1}(\mathbf{X}_t|\mathcal{D}_t) \leq G^{\dagger}_{t,1}(\mathbf{X}_t) = 1$.

Depending on this condition, bootstrap aggregating predictor may not always be better than the original unbagged predictor. We note that this is different from the ensemble aggregating predictor, which is always no worse than the original predictor as shown in Proposition 4. We also note that if the conditional risk of unbagged predictor (22) is very close to the minimum possible conditional risk in (21), that occurs when $\mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}(\mathbf{X}_t,\mathcal{D}_t)$ is very close to $G^{\dagger}_{t,1}(\mathbf{X}_t)$, then there is little room for improvement by bagging.

5.3 Asymptotic behavior of bagging binary predictor

Given $\alpha \in (0,1)$, let $u_t \equiv Y_t - Q_\alpha(Y_t|\mathbf{X}_{t-1}) = Y_t - \tilde{\mathbf{X}}'_{t-1}\boldsymbol{\beta}_\alpha$. Let $\boldsymbol{\beta}^{\dagger}_\alpha$ be the pseudo-true value of the parameter $\boldsymbol{\beta}_\alpha$ of interest, in the sense that $\boldsymbol{\beta}^{\dagger}_\alpha \equiv \arg\min_{\boldsymbol{\beta}_\alpha\in\Theta}\mathbb{E}_{Y_{t+1}}(\rho_\alpha(u_{t+1})|\mathbf{X}_t)$. Let $\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t) \equiv \arg\min_{\boldsymbol{\beta}_\alpha\in\Theta} R^{-1}\sum_{s=t-R+1}^{t}\rho_\alpha(u_s)$. Let $\hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t,\mathcal{D}_t) = \tilde{\mathbf{X}}'_t\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t)$ and $Q^{\dagger}_\alpha(Y_{t+1}|\mathbf{X}_t) = \tilde{\mathbf{X}}'_t\boldsymbol{\beta}^{\dagger}_\alpha$.

We now compare $\mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}(\mathbf{X}_t, \mathcal{D}_t)$ and $\mathbb{E}_{\mathcal{D}_t}\hat{G}^B_{t,1}(\mathbf{X}_t|\mathcal{D}_t)$ when R is large to examine the large sample behavior of bagging binary predictor. We make the following assumptions, *sufficient* for the asymptotic normality of quantile estimator $\hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t)$, and the consistency of the block bootstrap for the distribution of $\hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t)$.

Assumption 1. The conditional α-quantile model of Y_t given $\mathbf{X}_{t-1}$ is $Y_t = \tilde{\mathbf{X}}'_{t-1}\boldsymbol{\beta}_\alpha + u_t$, where $\boldsymbol{\beta}_\alpha$ is identified on Θ, a compact subset of $\mathbb{R}^k$.

Assumption 2. $\boldsymbol{\beta}^\dagger_\alpha$ is an interior point of Θ.

Assumption 3. The sequence $\{Y_t, \tilde{\mathbf{X}}_{t-1}\}$ is strong mixing with size $-r/(r-2)$ for $r > 2$.

Assumption 4. The distribution of $\{u_t\}$ is absolutely continuous and has a conditional density $f_t(u_t|\mathbf{X}_{t-1})$, where $f_t(u_t|\mathbf{X}_{t-1})$ is Lipschitz *a.s.*, $f_t(u_t|\mathbf{X}_{t-1})$ is bounded *a.s.*, and $f_t(0|\mathbf{X}_{t-1}) > 0$ *a.s.*, for all t.

Assumption 5. (i) For some $\delta > 0$, $\mathbb{E}_{\mathcal{D}_t}\left|\tilde{\mathbf{X}}_{ti}\right|^{2r+\delta} < \infty$ *a.s.* for all t and $i = 1, \ldots, k$. (ii) $M_R = \mathbb{E}_{\mathcal{D}_t}(R^{-1}\sum_{s=t-R+1}^{t}\tilde{\mathbf{X}}_{s-1}\tilde{\mathbf{X}}'_{s-1})$ and $L_R = \mathbb{E}_{\mathcal{D}_t}(R^{-1}\sum_{s=t-R+1}^{t} f_s(0|\mathbf{X}_{s-1})\tilde{\mathbf{X}}_{s-1}\tilde{\mathbf{X}}'_{s-1})$ are uniformly positive definite *a.s.* in R.

Assumption 6. Let $u^\dagger_t \equiv Y_t - \tilde{\mathbf{X}}'_{t-1}\boldsymbol{\beta}^\dagger_\alpha$. (i) $\mathbb{E}_{\mathcal{D}_t}[\alpha - 1(u^\dagger_t < 0)]\tilde{\mathbf{X}}_{t-1} = 0$ *a.s.* for every t. (ii) $J_R = \mathbb{V}_{\mathcal{D}_t}\left(R^{-1/2}\sum_{s=t-R+1}^{t}[\alpha - 1(u^\dagger_t < 0)]\tilde{\mathbf{X}}_{s-1}\right)$ is uniformly positive definite *a.s.* in R, where $\mathbb{V}_{\mathcal{D}_t}(\cdot)$ is the variance as defined in equation (5). (iii) The sequence $\{[\alpha - 1(u^\dagger_t < 0)]\tilde{\mathbf{X}}_{t-1}\}$ satisfies the conditions of a central limit theorem (e.g., White 1994, Theorem A.3.7; Newey and McFadden 1994, Theorem 7.2).

Our result on the asymptotic behavior of bagging binary predictor in Proposition 5 is based on the following two results.

Asymptotic normality (Fitzenberger 1997, Theorem 2.2; Komunjer 2005, Theorem 4). Let Assumptions 1-6 hold. Then, (i) $D_R^{-1/2}R^{1/2}(\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t) - \boldsymbol{\beta}^\dagger_\alpha) \to^d N(0, \mathbf{I}_k)$ as $R \to \infty$, where $D_R = L_R^{-1}J_R L_R^{-1}$ and $\mathbf{I}_k$ is the $k \times k$ identity matrix. (ii) Conditioning on the values of $\mathbf{X}_t$,

$$\hat{Z}_R \mid \mathbf{X}_t \to^d N(0, 1), \tag{24}$$

as $R \to \infty$, where $\hat{Z}_R \equiv R^{1/2}(\tilde{\mathbf{X}}'_t\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t) - \tilde{\mathbf{X}}'_t\boldsymbol{\beta}^\dagger_\alpha)/\sigma_R(\mathbf{X}_t)$ and $\sigma^2_R(\mathbf{X}_t) = \mathbb{V}_{\mathcal{D}_t}\left(R^{1/2}(\tilde{\mathbf{X}}'_t\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t) - \tilde{\mathbf{X}}'_t\boldsymbol{\beta}^\dagger_\alpha)\right) = \tilde{\mathbf{X}}'_t D_R \tilde{\mathbf{X}}_t$. ■

Given a training set of data at time t, $\mathcal{D}_t \equiv \{(Y_s, \mathbf{X}_{s-1})\}_{s=t-R+1}^t$, construct a bootstrap sample $\mathcal{D}^*_t \equiv \{(Y^*_s, \mathbf{X}^*_{s-1})\}_{s=t-R+1}^t$ according to the empirical distribution of $\hat{\mathbf{P}}(\mathcal{D}_t)$ of $\mathcal{D}_t$. Estimate $\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}^*_t) \equiv \arg\min_{\boldsymbol{\beta}_\alpha}$

$R^{-1}\sum_{s=t-R+1}^{t}\rho_\alpha(Y_s^* - \tilde{\mathbf{X}}_{s-1}^{*\prime}\boldsymbol{\beta}_\alpha)$, $t = R,\ldots,T$. The bootstrap binary predictor is $\hat{Q}_\alpha^*(Y_{t+1}|\mathbf{X}_t,\mathcal{D}_t^*) = \tilde{\mathbf{X}}_t'\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^*)$.

Bootstrap consistency (Fitzenberger 1997, Theorem 3.3). Let Assumptions 1-6 hold. Then, conditioning on the values of $\mathbf{X}_t$,

$$Z_R^* \mid \mathbf{X}_t \to^d N(0,1),$$

as $R\to\infty$, where $Z_R^* \equiv R^{1/2}(\tilde{\mathbf{X}}_t'\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^*) - \tilde{\mathbf{X}}_t'\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t))/\sigma_R(\mathbf{X}_t)$. That is, if we let $\Phi_R^*(z)\equiv \Pr(Z_R^* < z)$ and $\Phi(z)$ be the $N(0,1)$ distribution function, then

$$\sup_{z\in\mathbb{R}}||\Phi_R^*(z) - \Phi(z)|| = O_p(R^{-1}), \tag{25}$$

as $\hat{Z}_R$ is asymptotically pivotal (Hall 1992). ■

Under the asymptotic normality and the bootstrap consistency, we show in Proposition 5 that the difference between $\mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}(\mathbf{X}_t,\mathcal{D}_t)$ and $\mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}^B(\mathbf{X}_t|\mathcal{D}_t)$, and thus the difference between the expected costs of unbagged predictor and bagging predictor, $\mathbb{E}_{\mathcal{D}_t,Y_{t+1}}(c(\hat{e}_{t+1})|\mathbf{X}_t)$ and $\mathbb{E}_{\mathcal{D}_t,Y_{t+1}}(c(\hat{e}_{t+1}^B)|\mathbf{X}_t)$, vanish *a.s.* as $R\to\infty$.

Proposition 5. *Under Assumptions 1-6,* $\mathbb{E}_{\mathcal{D}_t,Y_{t+1}}(c\left(\hat{e}_{t+1}\right)|\mathbf{X}_t) - \mathbb{E}_{\mathcal{D}_t,Y_{t+1}}(c(\hat{e}_{t+1}^B)|\mathbf{X}_t) = O_p(R^{-1})$.

Proof: From (22) and (23),

$$\mathbb{E}_{\mathcal{D}_t,Y_{t+1}}(c\left(\hat{e}_{t+1}\right)|\mathbf{X}_t) - \mathbb{E}_{\mathcal{D}_t,Y_{t+1}}(c(\hat{e}_{t+1}^B)|\mathbf{X}_t) = [1-\alpha-\pi(\mathbf{X}_t)]\left[\mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}(\mathbf{X}_t,\mathcal{D}_t) - \mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}^B(\mathbf{X}_t|\mathcal{D}_t)\right].$$

Thus, it suffices to show $\mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}(\mathbf{X}_t,\mathcal{D}_t) - \mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}^B(\mathbf{X}_t|\mathcal{D}_t) = O_p(R^{-1})$. Let us set up some notation: $d_R^\dagger = -\sqrt{R}Q_\alpha^\dagger(Y_{t+1}|\mathbf{X}_t)/\sigma_R(\mathbf{X}_t)$, $d_R^* \equiv -\sqrt{R}\hat{Q}_\alpha^*(Y_{t+1}|\mathbf{X}_t,\mathcal{D}_t^*)/\sigma_R(\mathbf{X}_t)$, and $\hat{d}_R \equiv -\sqrt{R}\hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t,\mathcal{D}_t)/\sigma_R(\mathbf{X}_t)$. Note that $\hat{Z}_R \equiv d_R^\dagger - \hat{d}_R$ and $Z_R^* \equiv \hat{d}_R - d_R^*$.

First, for the original binary predictor, conditioning on the values of $\mathbf{X}_t$,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}(\mathbf{X}_t,\mathcal{D}_t) &= \Pr(\hat{G}_{t,1}(\mathbf{X}_t,\mathcal{D}_t)=1)\\
&= \Pr(\hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t,\mathcal{D}_t)>0)\\
&= \Pr(\hat{d}_R<0)\\
&= \Pr(\hat{Z}_R > d_R^\dagger)\\
&= 1-\Pr(\hat{Z}_R < d_R^\dagger) \qquad (26)\\
&\to 1-\Phi(d_R^\dagger) \ \ a.s. \ \text{ as } R\to\infty, \qquad (27)
\end{aligned}$$

where the third equality follows from the definition of $\hat{d}_R$, the fourth equality follows from the definition of $\hat{Z}_R$, and the last line follows from the asymptotic normality $\hat{Z}_R$ as shown in (24).

Next, we examine the bootstrap binary predictor. For each bootstrap forecast,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_t^*}\hat{G}_{t,1}^*(\mathbf{X}_t, \mathcal{D}_t^*) &= \Pr(\hat{G}_{t,1}^*(\mathbf{X}_t, \mathcal{D}_t^*) = 1) \\
&= \Pr(\hat{Q}_\alpha^*(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t^*) > 0) \\
&= \Pr(d_R^* < 0) \\
&= \Pr(Z_R^* > \hat{d}_R) \\
&= 1 - \Phi_R^*(\hat{d}_R) \\
&= 1 - \Phi(\hat{d}_R) - Bias_R(\hat{d}_R), \qquad (28)
\end{aligned}$$

where $Bias_R(z) \equiv \Phi_R^*(z) - \Phi(z)$. Therefore, for the voted-bagging predictor, conditioning on the values of $\mathbf{X}_t$,

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}^B(\mathbf{X}_t|\mathcal{D}_t) &= \Pr(\hat{G}_{t,1}^B(\mathbf{X}_t, \mathcal{D}_t^*) = 1) \\
&= \Pr(\mathbb{E}_{\mathcal{D}_t^*}\hat{G}_{t,1}^*(\mathbf{X}_t, \mathcal{D}_t^*) > \tfrac{1}{2}) \\
&= \Pr\left(1 - \Phi(\hat{d}_R) - Bias_R(\hat{d}_R) > \tfrac{1}{2}\right) \\
&= \Pr\left(\hat{d}_R < \Phi^{-1}(\tfrac{1}{2} - Bias_R(\hat{d}_R))\right) \\
&= \Pr\left(d_R^\dagger - \hat{Z}_R < \Phi^{-1}(\tfrac{1}{2} - Bias_R(\hat{d}_R))\right) \\
&= 1 - \Pr\left(\hat{Z}_R < d_R^\dagger - \Phi^{-1}(\tfrac{1}{2} - Bias_R(\hat{d}_R))\right) \qquad (29) \\
&\to 1 - \Phi(d_R^\dagger) \;\; a.s. \;\; \text{as } R \to \infty.
\end{aligned}$$

Comparing (26) and (29), we get $\mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}(\mathbf{X}_t, \mathcal{D}_t) - \mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}^B(\mathbf{X}_t|\mathcal{D}_t) = O_p(R^{-1})$, as $Bias_R(\hat{d}_R) = O_p(R^{-1})$ from (25). ■

From the above analysis, we see that bagging can push a predictor to its optimal value and also see that ability of bagging to improve predictability is limited. If there is no predictability at all, if unbagged predictor is already very close to the optimal predictor, or if we have a large enough sample, then there will be no room for improvement by bagging. Therefore, we can not expect bagging to still improve unbagged predictor when sample size is very large. We also note that bagging can be worse than unbagged predictors. Therefore, we cannot arbitrarily select a prediction model and expect bagging to work like magic and to give a better prediction automatically. In case of classification (as in our binary predictors), bagging predictor is

formed via voting, for which case Breiman (1996a, Section 4.2) also shows that it is possible that a bagging predictor could be worse than an unbagged predictor. As we will see from the Monte Carlo experiment in Section 7 and the empirical applications in Section 8, bagging could be worse for binary predictors and quantile predictions, although in general it works well particularly with a small R.

Remark 1. As noted before, if $\mathbb{E}_{\mathcal{D}_t}\hat{G}_{t,1}(\mathbf{X}_t, \mathcal{D}_t)$ is very close to $G^{\dagger}_{t,1}(\mathbf{X}_t)$, there is no room for improvement by bagging. According to (27), when $d^{\dagger}_R$ is close to 0, $\Phi(d^{\dagger}_R)$ is most away from 1 or 0. That means when the pseudo-true conditional α-quantile $Q^{\dagger}_{\alpha}(Y_{t+1}|\mathbf{X}_t)$ is close to 0, there is largest room for bagging to work. This is because when $Q^{\dagger}_{\alpha}(Y_{t+1}|\mathbf{X}_t) = 0$ unbagged predictor $\hat{G}_{t,1}(\mathbf{X}_t, \mathcal{D}_t) = \mathbf{1}(\hat{Q}_{\alpha}(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t) > 0)$ will be most "unstable" (This is when the sign of the estimated quantile can come out either way due to sample variation). Notice that $Q^{\dagger}_{\alpha}(Y_{t+1}|\mathbf{X}_t)$ is a function of $\mathbf{X}_t$, we can expect that the instability of binary predictor will also depend on how much mass the distribution of $\mathbf{X}_t$ puts in the regions where $Q^{\dagger}_{\alpha}(Y_{t+1}|\mathbf{X}_t)$ is close to zero. If the distribution of $\mathbf{X}_t$ is dense where $Q^{\dagger}_{\alpha}(Y_{t+1}|\mathbf{X}_t)$ is close to zero, $\hat{Q}_{\alpha}(Y_{t+1}|\mathbf{X}_t)$ will be less volatile, so unbagged predictor will perform relatively well. Otherwise, $\hat{Q}_{\alpha}(Y_{t+1}|\mathbf{X}_t)$ will be more volatile, so unbagged predictor will be very unstable. By adding more observations to that point through bootstrap, bagging will generate a more stable predictor around $Q^{\dagger}_{\alpha}(Y_{t+1}|\mathbf{X}_t)$. Our Monte Carlo results with skewed error distribution show this property clearly. Hence, for economic agents or financial investors with the cost function parameter α that gives $Q^{\dagger}_{\alpha}(Y_{t+1}|\mathbf{X}_t) = 0$, bagging binary prediction could work most effectively.

Remark 2. The "voted-bagging" does not transform the hard-thresholding decision into a soft-thresholding decision. As Bühlmann and Yu (2002) have shown, "averaged-bagging" transforms a hard-thresholding function (e.g., indicator) into a soft-thresholding function (smooth function) and thus decreases the instabilities of predictors.[2] However, in the majority voting $\hat{G}^{B}_{t,1}(\mathbf{X}_t|\mathcal{D}_t) = \mathbf{1}(\mathbb{E}_{\mathcal{D}^*_t}\hat{G}^{*}_{t,1}(\mathbf{X}_t, \mathcal{D}^*_t) > 1/2)$, bagging predictor is still a hard-thresholding decision. Bagging binary prediction remains unstable (particularly around the thresholding value). Hence, the explanation of Bühlmann and Yu (2002) does not apply to the voted-bagging binary predictor.

[2]If the bagging predictor is not via majority voting but via averaging, then the bagging can be effective. Note that

$$\mathbb{E}_{\mathcal{D}^*_t}\hat{G}^{*}_{t,1}(\mathbf{X}_t, \mathcal{D}^*_t) = 1 - \Phi^*(\hat{d}_R) = 1 - \Phi^*(d^{\dagger}_R - \hat{Z}_R).$$

In particular, consider the case when $Q^{\dagger}_{\alpha}(Y_{t+1}|\mathbf{X}_t) = 0$ or $d^{\dagger}_R = 0$, so that the binary predictor $\hat{G}_{t,1} = \mathbf{1}(\hat{Q}_{\alpha}(Y_{t+1}|\mathbf{X}_t) > 0)$ is most unstable. Then

$$\mathbb{E}_{\mathcal{D}^*_t}\hat{G}^{*}_{t,1}(\mathbf{X}_t, \mathcal{D}^*_t) = 1 - \Phi^*(-\hat{Z}_R) \sim U[0,1],$$

so that its mean is $\frac{1}{2}$ and its variance $\frac{1}{12}$. This is clearly an improvement over the unbagged *binary* predictor $\hat{G}_{t,1}(\mathbf{X}_t, \mathcal{D}_t)$ that may have mean $\frac{1}{2}$ and variance $\frac{1}{2}(1-\frac{1}{2}) = \frac{1}{4}$.

6 Bootstrap Aggregating Predictor for Quantile Variable

Another unstable predictor used in bagging literature is non-linear predictors. Quantile predictor is the minimizer of cost function $\rho_\alpha(\cdot)$ as defined in (12), and is a non-linear function of sample moments. According to Friedman and Hall (2000) and Buja and Stuetzle (2002), bagging can increase the prediction accuracy for non-linear predictors. Knight and Bassett (2002) show that under i.i.d. and some other assumptions, bagged non-parametric quantile estimators and linear regression quantile estimators can outperform their corresponding standard unbagged predictors. Therefore, we will examine the effect of bagging on quantile predictions, by comparing unbagged predictor $\hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t)$ and bagging predictor $\hat{Q}^B_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t)$ (defined below).

The procedure of bagging for quantile predictors can be conducted in the following steps:

1. Construct the jth bootstrap sample $\mathcal{D}^{*(j)}_t$, $j = 1, \ldots, J$, the bootstrap samples, according to the empirical distribution of $\mathcal{D}_t$.

2. Estimate $\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}^{*(j)}_t) = \arg\min_{\boldsymbol{\beta}_\alpha \in \Theta} R^{-1} \sum_{s=t-R+1}^{t} \rho_\alpha(Y^{*(j)}_s - \tilde{\mathbf{X}}^{*(j)\prime}_{s-1}\boldsymbol{\beta}_\alpha)$, $t = R, \ldots, T$.

3. Compute the bootstrap quantile predictor from the jth bootstrapped sample, that is,

$$\hat{Q}^{*(j)}_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}^{*(j)}_t) \equiv \tilde{\mathbf{X}}'_t \hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}^{*(j)}_t).$$

4. Finally, bagging predictor $\hat{Q}^B_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t)$ can be constructed by averaging over the J bootstrap predictors, i.e.,

$$\hat{Q}^B_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t) \equiv \mathbb{E}_{\mathcal{D}^*_t} \hat{Q}^*_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}^*_t),$$

where $\mathbb{E}_{\mathcal{D}^*_t} \hat{Q}^*_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}^*_t) = \sum_{j=1}^{J} \hat{w}_{j,t} \hat{Q}^{*(j)}_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}^{*(j)}_t)$.

Why does bagging work for quantile prediction? There have been several papers that are useful to answer this question. It is generally explained in two ways. The first explanation why bagging works for quantile prediction is different than for the classification problem as we discussed in the previous section. Friedman and Hall (2000), Buja and Stuetzle (2002), and Knight and Bassett (2002) all use a certain kind of Taylor-expansion to rewrite interested estimators, and show that bagging predictors are first order equivalent to the standard unbagged predictors, but will lower the non-linearities of the predictor. Bagging can drive an estimator towards its linear approximation, which usually has a lower variance.[3] The second explanation

[3]Friedman and Hall (2000) use a Taylor-expansion to decompose statistical estimators into linear and higher order parts, or decompose the objective function that they optimize into quadratic and higher order terms. To apply a Taylor-expansion, the estimator must be a smooth function of sample moments and the cost function is differentiable. Therefore this does not help

why bagging works for quantile prediction has to do with the unstableness of quantile prediction. One source of unstableness is the non-differentiable feature in the objective function (12). The sample objective function used in regression as an estimator of the objective function will be even worse behaved because of the limitation of sample size. There may be several equivalent minimizers for sample objective function, and the numerical searching method may stop at any of these minimizers, or even a local minimizer depending on the beginning point of the search. So quantile estimator may have a very high volatility. Bagging can smooth the sample objective function, so that bagging predictor will converge to the global minimizer of the sample objective function.

For the above reasons, we conjecture that bagging will also improve quantile prediction with time series data. However, we leave more rigorous analytical work for our future research. For now, we show the performance of bagging for quantile predictions via simulation and empirical analysis in the next two sections. These results show that bagging would be very useful in improving quantile prediction (e.g., VaR forecasts in financial risk management and the fan-chart of the Bank of England).

7 Monte Carlo

In this section, we use a set of Monte Carlo simulations to gain further insights of conditions under which bagging works. From both binary and quantile predictions, we can obtain the out-of-sample average costs for unbagged predictors (S_1) and bagging predictors (S_2). We consider the asymmetric cost parameter $\alpha = 0.1, 0.3, 0.5, 0.7,$ and 0.9. It will be said that bagging "works" if $S_1 > S_2$. To rule out the chance of pure luck by a certain criterion, we compute the following four summary performance statistics from 100 Monte Carlo replications ($r = 1, \ldots, 100$): $T_1 \equiv \frac{1}{100}\sum_{r=1}^{100} S_a^r$, $T_2 \equiv \left(\frac{1}{100}\sum_{r=1}^{100}(S_a^r - T_1)\right)^{1/2}$, $T_3 \equiv \frac{1}{100}\sum_{r=1}^{100} \mathbf{1}(S_1^r > S_2^r)$, and $T_4 \equiv \frac{1}{100}\sum_{r=1}^{100} \mathbf{1}(S_1^r = S_2^r)$, where $a = 1$ for the non-bagged predictor and $a = 2$ for bagging predictor. T_1 measures the Monte Carlo mean of the out-of-sample cost, T_2 measures the Monte Carlo standard deviation of the out-of-sample cost, T_3 measures the Monte Carlo frequency that bagging works, and $(T_3 + T_4)$ measures the Monte Carlo frequency that bagging is no worse than unbagged predictors. (T_4 is usually zero for quantile prediction, but usually non-zero for binary prediction.)

analyzing bagging quantile predictor. Buja and Stuetzle (2002) extend the application of bagging regression to more general circumstance by expressing predictors as statistical functionals (especially those can be written as U-statistics). They show how quantile estimators can be written as U-statistics. Since a Taylor expansion is no longer applicable now, they use the von Mises expansion technique to prove that the leading effects of bagging on variance, squared bias, and MSE are of order R^{-2}, where R is the estimation sample size. So, bagging may works for non-smooth target functions, such as median predictors and quantile predictors. Knight and Bassett (2002) illustrate that if quantile estimator is asymptotically normal, we can decompose quantile estimator into linear and non-linear part using the Bahadur-Kiefer representation. They show that bagging quantile estimators are first-order equivalent to the standard unbagged quantile predictor, however, bagging can reduce the non-linearity of the sample quantile.

We generate the data from

$$\begin{aligned} Y_t &= \rho Y_{t-1} + \varepsilon_t, \\ \varepsilon_t &= z_t \left[(1-\theta) + \theta \varepsilon_{t-1}^2\right]^{1/2} \\ z_t &\sim \text{i.i.d. } MW_i \end{aligned}$$

where the i.i.d. innovation z_t is generated from the first eight mixture normal distributions of Marron and Wand (1992, p. 717), each of which will be denoted as MW_i $(i = 1, \ldots, 8)$.[4] In Table 1, we consider the data generating processes for ARCH-MW$_1$ with $\theta = 0.5$ (and $\rho = 0$), while in Tables 2-5, we consider the data generating processes for AR-MW$_i$ $(i = 1, \ldots, 4)$ with $\rho = 0.6$ (and $\theta = 0$). Therefore, our data generating processes fall into two categories: the (mean-unpredictable) martingale-difference ARCH(1) processes without AR structure and the mean-predictable AR(1) processes without ARCH structure.

For each series, 100 extra series is generated and then discarded to alleviate the effect of the starting values in random number generation. We consider one fixed out-of-sample size $P = 100$ and a range of estimation sample sizes $R = 20$, 50, 100, 200. Our bagging predictors are generated by voting or by averaging over $J = 50$ bootstrap predictors (the results with $J = 20$, not reported here, basically tell the same story).

Our binary predictors are based on quantile predictors as suggested before: $\hat{G}_{t,1}(\mathbf{X}_t, \mathcal{D}_t) \equiv \mathbf{1}(\hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t) > 0)$, and we are using univariate quantile regression model as discussed in Section 3. Quantile model is estimated using the interior-point algorithm used by Portnoy and Koenker (1997).

To generate bootstrap samples, we use the block bootstrap for both Monte Carlo experiments in this section and the empirical application in the next section. We choose the block size that minimizes the in-sample average cost recursively and therefore we use a different block size at each forecasting time and for the cost function with different α's.

The Monte Carlo results are reported in Tables 1-5. For both binary predictions and quantile predictions, bagging works well when the sample size is small. The improvement of bagging predictors over unbagged predictors becomes less significant when the sample size increases. This is true in terms of all three criteria, T_1, T_2, and T_3. When $R = 20$, bagging gives the largest reduction in the mean cost, the largest reduction in the variance of the cost, and the highest frequency of out-performance.

In the previous sections, we introduced several weights $\hat{w}_{j,t}$ to form bagging predictors: equal weight and the BMA-weights with lags $k = 1, 5$, and R, defined in (31) in Appendix. In Tables 1-5, EW and W_k denote

[4] MW_1 is Gaussian, MW_2 is Skewed unimodal, MW_3 Strongly skewed, MW_4 Kurtoic unimodal, MW_5 Outlier, MW_6 Bimodal, MW_7 Separated bimodal, and MW_8 is Skewed bimodal. See Marron and Wand (1992, p. 717). To save space we report only for MW_i $(i = 1, \ldots, 4)$. The other four results for $i = 5, \ldots, 8$ are basically similar in the pattern how the bagging works, and are available upon request.

the weighted-bagging with equal weights and with BMA-weights using k-most recent in-sample observations. The BMA is to give a large weight to the jth bootstrap predictor at each period t when it has forecast well over the past k periods, and gives a small weight to the jth bootstrap predictor at period t when it forecasted poorly over the past k periods. With smaller k, we intend to focus on the most recent performance of each bootstrap predictor. According to our simulation results, the BMA-weight with $k = 1$ performs the best for our DGP's, although all weights with different k often work quite similarly.

Let us examine Table 1 in some details. The mean cost reduction (in terms of T_1) for both binary prediction and quantile prediction can be as much as about 20% for all α's; the variance of cost (in terms of T_2) can be reduced by as much as about 20% for binary prediction and about 50% for quantile prediction; and T_3 is as high as 95% for binary prediction and 100% for quantile prediction. This result shows that bagging can significantly mitigate the problem of the sample shortage.

This function of bagging can also be observed by the performance of bagging on tails. The scarce of observations on tails usually will lead to the poor predictions, however, the degree of improvement of bagging predictors are very significant for $\alpha = 0.1$ and 0.9 according to our Monte Carlo results. We observe this fact using Tables 2 to 5 in case of $R = 20$. For binary prediction, the average cost reduction (in terms of T_1) for $\alpha = 0.1$ and 0.9 is about 15%, however, the average cost reduction for $\alpha = 0.5$ is only about 5%. The average cost variance (in terms of T_2) reduction for $\alpha = 0.1$ and 0.9 is about 20%, however, the average cost variance reduction for $\alpha = 0.5$ is not significant. The frequency that bagging works (in terms of T_3) for $\alpha = 0.1$ and 0.9 is about 80%, however, the frequency that bagging works for $\alpha = 0.5$ is only about 60%. For quantile prediction, the average cost reduction (in terms of T_1) for $\alpha = 0.1$ and 0.9 is about 20%, however, the average cost reduction for $\alpha = 0.5$ is only about 5%. The average cost variance reduction (in terms of T_2) for $\alpha = 0.1$ and 0.9 is about 40%, however, the average cost variance reduction for $\alpha = 0.5$ is about 20%. The frequency that bagging works (in terms of T_3) for $\alpha = 0.1$ and 0.9 and $\alpha = 0.5$ are similarly around 90%.

The advantage of bagging for quantile predictions (in terms of T_1, T_2, and T_3) decreases gradually as the sample size increases, but still exists for $R = 200$ though the advantage is not very significant. However, for binary prediction, this advantage disappears much faster, and will almost disappear when the sample size exceeds 100 for most data generating processes (DGP). With a large sample size, the average cost of bagging predictor converges to that of unbagged predictor, confirming the analytical result in Proposition 5.

Another interesting phenomenon we observe is that bagging works better (in terms of T_1, T_2, and T_3) for both binary prediction and quantile when there is ARCH structure in the DGP. For the AR processes,

the mean is predictable, so binary predictors perform pretty well even without bagging. Therefore, there is not much room left for bagging to improve on. However, for the ARCH processes without the AR term, the conditional mean is unpredictable. Although there is some binary predictability through the time-varying conditional higher moments, the predictability is harder to explore than in the AR models. As we can see from the tables, the mean cost (T_1) of unbagged binary predictor in Tables 2-5 is much lower than the mean cost of unbagged binary predictors in Table 1. At the same time, the non-linear structure in the ARCH models lead to the difficulty in the parameter estimation via numerical optimization, which also leaves more room for bagging to work. Therefore, we can see that bagging improves more (in all of T_1, T_2, and T_3) in Table 1 than in Tables 2-5 for both binary predictions and quantile predictions.

One more observation from the Monte Carlo simulation is that bagging works asymmetrically (in terms of T_1, T_2, and T_3) for asymmetric data. For MW distribution with asymmetric distributions like in Table 3 (MW_2) and 4 (MW_3), bagging works better if $\hat{Q}_\alpha(Y_{t+1}|\mathbf{X}_t)$ lies on the flatter tail for both binary and quantile prediction. For example, MW_2 has flatter left tail, therefore bagging works better for a smaller α in Tables 3; while MW_3 has flatter right tail, therefore bagging works better for a larger α in Table 4.

8 Empirical Application

Christofferson and Diebold (2003) find, among other things, that the sign dependence is highly nonlinear and is not likely to be found in high-frequency (e.g., daily) or low-frequency (e.g., annual) returns. Instead, it is more likely to appear at intermediate return horizons of two or three months. Thus our empirical application of binary prediction is to time the market for the S&P500 and NASDAQ indexes in monthly frequency. Let Y_t represent the log difference of a stock index at month t, and suppose we are interested in predicting whether the stock index will rise or not in the next month.

The S&P 500 series, retrieved from *finance.yahoo.com*, is monthly data from October 1982 to February 2004 ($T + 1 = 257$). The NASDAQ series is also retrieved from *finance.yahoo.com,* monthly from October 1984 to February 2004 ($T + 1 = 233$). We split the series into two parts: one for in-sample estimation with the size R = 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140 and 150, and another for out-of-sample forecast validation with sample size $P = 100$ (fixed for all R's). We choose the most recent $P = 100$ months in the sample as a out-of-sample validation sample. We use a rolling-sample scheme, that is, the first forecast is based on observations $T - P - R + 1$ through $T - P$, the second forecast is based on observations

$T-P-R+2$ through $T-P+1$, and so on. The two series are summarized as follows:

	In-sample period	Out-of-sample period	$T+1$	P
S&P 500	October 1982 $\sim$ October 1995	November 1995 $\sim$ February 2004	257	100
Nasdaq	October 1984 $\sim$ October 1995	November 1995 $\sim$ February 2004	233	100

We consider the asymmetric cost parameter $\alpha = 0.1, 0.2, \ldots, 0.9$ to represent different possible preferences. With the cost function given in (13), if we predict the market to go down when the market will go up, the cost is α; and if we predict the market to go up when the market will go down, the cost is $1-\alpha$. Therefore, if a person only buys and holds the stock, she tends to have a value of α smaller than 0.5 because missing an earning opportunity will not be as bad as losing money. However, if a person wants to sell short, she tends to have a value of α larger than 0.5 because of the leverage effect. Since most investors belong to the first category, the more predictability may be exploited for small α's.

Figures 1-2 present the graphs of the out-of-sample average costs (S_1 or S_2) in vertical axis against the training sample size R in the horizontal axis, for the nine values of α. There are lines for each α – the dark solid line is for the cost S_1 of unbagged predictor, and the other four lines are for the cost S_2 of bagging predictor with different weights (equal weight or the three BMA weights with $k = 1, 5, R$, as discussed in Appendix). bagging works similarly for S&P500 and NASDAQ. It works better with the smaller R. Bagging predictors with different weighting schemes seem to work similarly. For all α's and for both binary and quantile predictions, the costs of bagging predictors converge to the optimal costs much faster than those of unbagged predictors. We can see that bagging predictors converge to the stable level of the cost for R as small as 20 in most cases for both binary predictions and quantile predictions. However, a larger R is needed for unbagged predictors converge to that level of the cost. When the sample size is small, bagging can lower the cost to larger extent, and bagging predictors almost never get worse than unbagged predictors. When $R = 20$, bagging can lower the cost as much as or even more than 50% for both binary predictions and quantile predictions! When R grows larger, unbagged predictors and bagging predictors will converge to the same stable cost level as expected from Proposition 5.

9 Conclusions

We have examined how bagging works for binary and quantile prediction with an asymmetric cost function for time series. We construct binary predictor from quantile predictor. Bagging binary predictors are constructed via majority voting on binary predictors trained on the bootstrapped training samples. We have shown the conditions under which bagging works for binary prediction. Based on the asymmetric quantile check functions, by treating it as a quasi likelihood, we have also derived the various BMA-weights to form the

weighted bagging both for binary and quantile predictors. The simulation results and the empirical results using two U.S. stock index monthly returns, not only confirm but also clearly demonstrate our analytical results – the main finding of the paper is that bagging works when the size of the training sample is small and the predictor is unstable. We prove that bagging does not work for binary prediction when the training sample size is very large. Hence, the potential advantage of bagging lies in areas where small sample is common. Bagging will be particularly relevant and useful in practice when structural breaks are frequent so that simply using as many observations as possible is not a wise choice for out-of-sample prediction, as emphasized in Pesaran and Timmermann (2002b, 2004) and Paye and Timmermann (2003).

References

Avramov, K. (2002), "Stock Return Predictability and Model Uncertainty", *Journal of Financial Economics*, 64, 423-458.

Bates, J.M. and C.W.J. Granger (1969), "The Combination of Forecasts", *Operations Research Quarterly,* 20, 451-468.

Bauer, E. and R. Kohavi (1999), "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants", *Machine Learning,* 36, 105-139.

Breiman, L. (1996a), "Bagging Predictors", *Machine Learning,* 24, 123-140.

Breiman, L. (1996b), "Heuristics of Instability and Stabilization in Model Selection", *Annals of Statistics*, 24(6), 2350–2383.

Bühlmann, P. and B. Yu (2002), "Analyzing Bagging", *Annals of Statistics*, 30(4), 927-961.

Buja, A. and W. Stuetzle (2002), "Observations on Bagging", University of Pennsylvania and University of Washington, Seattle.

Christofferson, P.F. and F.X. Diebold (2003), "Financial Asset Returns, Direction-of-Change Forecasting, and Volatility Dynamics", McGill University and University of Pennsylvania.

Chernozhukov, V. and Len Umantsev (2001), "Conditional Value-at-Risk: Aspects of Modeling and Estimation", *Empirical Economics*, 26, 271-292.

Elliott, G., I. Komunjer, and A. Timmermann (2003a), "Biases in Macroeconomic Forecasts: Irrationality or Asymmetric Loss?", UCSD and Caltech.

Elliott, G., I. Komunjer, and A. Timmermann (2003b), "Estimating Loss Function Parameters", UCSD and Caltech.

Evgeniou, T., M. Pontil, and A. Elisseeff (2004), "Leave One Out Error, Stability, and Generalization of Voting Combinations of Classifiers", *Machine Learning*, 55(1), 71-97.

Fitzenberger, B. (1997), "The Moving Blocks Bootstrap and Robust Inference for Linear Least Squares and Quantile Regressions", *Journal of Econometrics*, 82, 235-287.

Friedman, J.H. and P. Hall (2000), "On Bagging and Nonlinear Estimation", Stanford University and Australian National University.

Garratt, A., K. Lee, H.M. Pesaran, and Y. Shin (2003), "Forecast Uncertainties in Macroeconometric Modelling: An Application to the UK Economy", *Journal of American Statistical Association,* 98, 829-838.

Granger, C.W.J. (1969), "Prediction with a Generalized Cost of Error Function", *Operational Research Quarterly,* 20, 199-207.

Granger, C.W.J. (1999a), *Empirical Modeling in Economics: Specification and Evaluation,* Cambridge University Press: London.

Granger, C.W.J. (1999b), "Outline of Forecast Theory Using Generalized Cost Functions", *Spanish Economic Review* 1, 161-173.

Granger, C.W.J. (2002), "Some Comments on Risk", *Journal of Applied Econometrics*, 17, 447-456.

Granger, C.W.J., M. Deutsch, and T. Teräsvirta (1994), "The Combination of Forecasts Using Changing Weights", *International Journal of Forecasting*, 10, 47-57.

Granger, C.W.J. and Y. Jeon (2004), "Thick Modeling", *Economic Modeling*, 21, 323-343.

Granger, C.W.J. and M.H. Pesaran (2000), "Economic and Statistical Measures of Forecast Accuracy", *Journal of Forecasting,* 19, 537-560.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer.

Harter, H.L. (1977), "Nonuniqueness of Least Absolute Values Regression", *Communications in Statistics - Theory and Methods,* A6, 829-838.

Hong, Y. and J. Chung (2003), "Are the Directions of Stock Price Changes Predictable? Statistical Theory and Evidence", Cornell University.

Hong, Y. and T.-H. Lee (2003), "Inference on Predictability of Foreign Exchange Rates via Generalized Spectrum and Nonlinear Time Series Models", *Review of Economics and Statistics*, 85(4), November, in press.

Inoue, A. and L. Kilian (2005), "How Useful is Bagging in Forecasting Economic Time Series? A Case Study of U.S. CPI Inflation", NCSU and University of Michigan.

Kim, J. and D. Pollard (1990), "Cube Root Asymptotics", *Annals of Statistics,* 18, 191-219.

Kim, T.-H. and H. White (2003), "Estimation, Inference, and Specification Testing for Possibly Misspecified Quantile Regressions", in T. Fomby and R.C. Hill, eds., *Maximum Likelihood Estimation of Misspecified Models: Twenty Years Later*. New York: Elsevier, 107-132.

Kitamura, Y. (2001), "Predictive Inference and the Bootstrap", Yale University.

Knight, K. and G.W. Bassett Jr. (2002), "Second Order Improvements of Sample Quantiles Using Subsamples", University of Toronto and University of Illinois, Chicago.

Koenker, R and G. Basset (1978), "Asymptotic Theory of Least Absolute Error Regression", *Journal of the American Statistical Association*, 73, 618-622.

Komunjer, I. (2005), "Quasi-Maximum Likelihood Estimation for Conditional Quantiles", *Journal of Econometrics*, forthcoming.

Kordas, G. (2005), "Smoothed Binary Regression Quantiles", *Journal of Applied Econometrics*, forthcoming.

Kuncheva, L.I. and C.J. Whitaker (2003), "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy", *Machine Learning,* 51, 181-207.

Linton, O. and Y.-J. Whang (2004), "A Quantilogram Approach to Evaluating Directional Predictability", Cowles Foundation Discussion Paper No. 1454.

Manski, C.F. (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice", *Journal of Econometrics*, 3(3), 205-228.

Manski, C.F. (1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator", *Journal of Econometrics*, 27(3), 313-333.

Manski, C.F., and T. S. Thompson (1989), "Estimation of Best Predictors of Binary Response", *Journal of Econometrics,* 40, 97-123.

Marron, J.S. and M.P. Wand (1992), "Exact Mean Integrated Squared Error", *Annals of Statistics*, 20, 712–736.

Money, A.H., J.F. Affleck-Graves, M.L. Hart, and G.D.I. Barr (1982), "The Linear Regression Model and the Choice of p", *Communications in Statistics - Simulations and Computations,* 11(1), 89-109.

Newey, W.K. and D.L. McFadden (1994), "Large Sample Estimation and Hypothesis Testing", in R.F. Engle and D.L. McFadden (eds.), *Handbook of Econometrics,* 4, 2113-2247, Elsevier Science.

Nyguist, H. (1983), "The Optimal L_p-norm Estimation in Linear Regression Models", *Communications in Statistics - Theory and Methods,* 12, 2511-2524.

Paye, B.S. and A. Timmermann (2003), "Instability of Return Predictability Models", UCSD.

Pesaran, M.H. and A. Timmermann (2002a), "Market Timing and Return Prediction Under Model Instability", *Journal of Empirical Finance*, 9, 495-510.

Pesaran, M.H. and A. Timmermann (2002b), "Model Instability and Choice of Observation Window", UCSD and Cambridge.

Pesaran, M.H. and A. Timmermann (2004), "How Costly Is It To Ignore Breaks When Forecasting the Direction of a Time Series?" *International Journal of Forecasting*, 20, 411-424.

Powell, J.L. (1986), "Censored Regression Quantiles", *Journal of Econometrics,* 32, 143-155.

Portnoy, S. and R. Koenker (1997), "The Gaussian Hare and the Laplacean Tortoise: Computability of l_1 vs l_2 Regression Estimators", *Statistical Science*, 12, 279-300.

Stock, J.H. and M.W. Watson (1999), "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series", in *Cointegration, Causality, and Forecasting, A Festschrift in Honor of C.W.J. Granger*, (eds.) R.F. Engle and H. White, Oxford University Press: London, pp. 1-44.

Stock, J.H. and M.W. Watson (2005), "An Empirical Comparison of Methods for Forecasting Using Many Predictors", Harvard University and Princeton University.

Timmermann, A. (2005), "Forecast Combinations", in G. Elliott, C.W.J. Granger, and A. Timmermann (eds.), *Handbook of Economic Forecasting.* Amsterdam: North-Holland, forthcoming.

White, H. (1994), *Estimation, Inference, and Specification Analysis,* Cambridge University Press.

Yang, Y. (2004), "Combining Forecasting Procedures: Some Theoretical Results", *Econometric Theory*, 20, 176-222.

Zou, H. and Y. Yang (2004), "Combining Time Series Models for Forecasting", *International Journal of Forecasting*, 20, 69-84.

Appendix

In this appendix, we discuss the Bayesian model averaging (BMA) technique to find a proper weight function $\{\hat{w}_{j,t}\}$ in forming bagging binary predictor via majority voting and bagging quantile predictor by averaging as discussed in Sections 5 and 6.[5] Usually, bagging predictor uses the equally weighting ($\hat{w}_{j,t} = J^{-1}$, $j = 1, \ldots, J$.) over the bootstrapped predictions. However, $\hat{w}_{j,t}$ can be estimated depending on the performance of each bootstrapped predictor. There are several candidates that we can borrow from estimated weighting forecast combination. One method is to estimate the weight by regression (minimizing the cost function) as initially suggested by Granger *et al.* (1994). This weight scheme assumes that predictors to be combined are independent and the number of predictors are small compared to the sample size, so the regular regression method can be applied. However, the number of our bootstrapped predictors are large and the predictors are closely related, the regression-based weights are not applicable here. Another method for forecast combination is via the Bayesian averaging, which is what we use in this paper for Monte Carlo experiments and empirical work.

Bagging prediction can be decomposed into two parts – the prediction based on estimated model and the parameter estimation given the data, therefore, we compute the BMA-weighted bootstrap average $\mathbb{E}_{\mathcal{D}_t^*}\hat{G}_{t,1}^*(\mathbf{X}_t, \mathcal{D}_t^*)$ as follows:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_t^*}\hat{G}_{t,1}^*(\mathbf{X}_t, \mathcal{D}_t^*) &= \sum_{j=1}^{J} \hat{G}_{t,1}^{*(j)}(\mathbf{X}_t, \mathcal{D}_t^{*(j)}) \Pr\left[\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})|\mathcal{D}_t^{*(j)}, \mathcal{D}_t\right] \Pr(\mathcal{D}_t^{*(j)}|\mathcal{D}_t) \\ &\equiv \sum_{j=1}^{J} \hat{w}_{j,t}\hat{G}_{t,1}^{*(j)}(\mathbf{X}_t, \mathcal{D}_t^{*(j)}),\end{aligned}$$

where the last equality follows from $\Pr\left[\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})|\mathcal{D}_t^{*(j)}, \mathcal{D}_t\right] \Pr(\mathcal{D}_t^{*(j)}|\mathcal{D}_t) = \Pr\left[\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})|\mathcal{D}_t\right]$ and by setting $\hat{w}_{j,t} \equiv \Pr\left[\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})|\mathcal{D}_t\right]$.

The posterior probability of $\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})$ given $\mathcal{D}_t$ through the jth bootstrap data set $\mathcal{D}_t^{*(j)}$ is calculated by Bayes' rule:

$$\hat{w}_{j,t} = \Pr\left[\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})|\mathcal{D}_t\right] = \frac{\Pr\left[\mathcal{D}_t|\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})\right] \Pr(\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)}))}{\sum_{j=1}^{J} \Pr\left[\mathcal{D}_t|\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})\right] \Pr(\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)}))},$$

and now the problem is to estimate $\Pr\left[\mathcal{D}_t|\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})\right]$ and $\Pr(\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)}))$. When we do not have any information for the prior $\Pr(\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)}))$, we may just use some non-informative prior, $\Pr(\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)}))$ is the same

[5] BMA has been proved to be useful in forecasting financial returns (Avramov 2002) and in macroeconomic forecasting for inflation and output growth (Garratt *et al.* 2003), to deal with the parameter and model uncertainties. The out-of-sample forecast performance using BMA is often shown to be superior to that using model selection criteria.

for all j, so that

$$\hat{w}_{j,t} = \Pr\left[\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})|\mathcal{D}_t\right] = \frac{\Pr\left[\mathcal{D}_t|\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})\right]}{\sum_{j=1}^{J}\Pr\left[\mathcal{D}_t|\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})\right]}. \tag{30}$$

where $\Pr\left[\mathcal{D}_t|\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})\right]$ is usually estimated by a likelihood function. According to Komunjer (2005), the exponential of quantile cost function can be treated as quasi-likelihood function, so $\Pr\left[\mathcal{D}_t|\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})\right]$ can be calculated by

$$\Pr\left[\mathcal{D}_t|\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})\right] \propto \exp\left(-k^{-1}\sum_{s=t-k+1}^{t} c(\hat{u}_s^{*(j)})\right), \tag{31}$$

with using the k-most recent in-sample observations, where $\hat{u}_s^{*(j)} = Y_s^{*(j)} - \tilde{\mathbf{X}}_{s-1}^{*\prime}\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_{s-1}^{*(j)})$. We select $k = 1, 5$, and R in the simulations (Section 7) and in the empirical experiments (Section 8). Intuitively, $\hat{w}_{j,t}$ gives a large weight to the jth bootstrap predictor at period t when it has forecasted well over the past k periods, and gives a small weight to the jth bootstrap predictor at period t when it forecasted poorly over the past k periods.[6] In Tables 1-5 (discussed in Section 7), EW and W_k denote the weighted-bagging with equal weights and with BMA-weights using k-most recent in-sample observations.

We can derive the BMA-weight bagging quantile prediction similarly as follows:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_t^*}\hat{Q}_\alpha^*(Y|\mathbf{X}_t, \mathcal{D}_t^*) &= \sum_{j=1}^{J}\hat{Q}_\alpha^{*(j)}(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t^{*(j)})\Pr\left[\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})|\mathcal{D}_t^{*(j)}, \mathcal{D}_t\right]\Pr(\mathcal{D}_t^{*(j)}|\mathcal{D}_t) \\
&\equiv \sum_{j=1}^{J}\hat{w}_{j,t}\hat{Q}_\alpha^{*(j)}(Y_{t+1}|\mathbf{X}_t, \mathcal{D}_t^{*(j)}),
\end{aligned}$$

where $\hat{w}_{j,t} \equiv \Pr\left[\hat{\boldsymbol{\beta}}_\alpha(\mathcal{D}_t^{*(j)})|\mathcal{D}_t\right]$.

[6]The weighting scheme proposed in Yang (2004) may be regarded as a special case of the above BMA-weighted forecasts. Zou and Yang (2004) have applied this method to choose ARIMA models and find that it has a clear stability advantage in forecasting over some existing popular model selection criteria.

Table 1: $\rho = 0, \theta = 0.5,$ and MW_1 (Gaussian)

		R=20				R =50				R =100				R =200			
		J=1	J =50			J=1	J =50			J=1	J =50			J=1	J =50		
Binary			*EW*	W_1	W_R		*EW*	W_1	W_R		*EW*	W_1	W_R		*EW*	W_1	W_R
α=.1	T_1	6.89	5.71	5.56	5.63	5.41	5.11	5.08	5.09	5.10	5.05	5.05	5.05	5.04	5.04	5.04	5.04
	T_2	1.28	1.09	0.98	1.05	0.80	0.54	0.52	0.53	0.55	0.51	0.51	0.51	0.54	0.54	0.54	0.54
	T_3		0.80	0.80	0.81		0.35	0.36	0.34		0.08	0.08	0.08		0.00	0.00	0.00
	T_4		0.08	0.09	0.07		0.55	0.56	0.56		0.90	0.90	0.90		1.00	1.00	1.00
α=.3	T_1	18.42	15.50	15.20	15.48	16.36	15.28	15.27	15.31	15.56	15.17	15.18	15.16	15.25	15.14	15.13	15.13
	T_2	2.44	2.28	2.13	2.19	1.94	1.69	1.62	1.71	1.74	1.50	1.50	1.51	1.66	1.65	1.64	1.65
	T_3		0.89	0.95	0.91		0.80	0.83	0.79		0.39	0.37	0.40		0.16	0.16	0.16
	T_4		0.01	0.00	0.02		0.07	0.08	0.08		0.51	0.53	0.51		0.82	0.81	0.82
α=.5	T_1	25.56	21.26	21.24	21.48	25.17	22.63	22.67	22.69	25.04	23.22	23.22	23.27	24.93	23.23	23.33	23.24
	T_2	2.29	3.38	3.11	3.32	2.20	2.49	2.36	2.38	2.60	2.35	2.48	2.31	2.56	2.78	2.54	2.77
	T_3		0.93	0.97	0.93		0.81	0.82	0.82		0.80	0.80	0.80		0.70	0.72	0.65
	T_4		0.04	0.00	0.03		0.09	0.06	0.09		0.07	0.03	0.06		0.07	0.03	0.14
α=.7	T_1	17.80	15.57	15.41	15.60	15.82	14.93	14.88	14.95	15.15	14.88	14.87	14.87	14.98	14.87	14.86	14.88
	T_2	2.19	2.08	2.09	2.11	1.96	1.63	1.61	1.63	1.68	1.60	1.58	1.60	1.70	1.63	1.63	1.62
	T_3		0.88	0.91	0.86		0.73	0.75	0.73		0.37	0.37	0.37		0.16	0.17	0.14
	T_4		0.00	0.01	0.00		0.11	0.11	0.11		0.54	0.57	0.54		0.80	0.79	0.81
α=.9	T_1	6.94	5.89	5.83	5.92	5.22	4.96	4.97	4.96	5.03	4.95	4.95	4.96	4.96	4.96	4.96	4.96
	T_2	1.49	1.45	1.39	1.47	0.69	0.50	0.52	0.50	0.62	0.51	0.51	0.52	0.54	0.54	0.54	0.54
	T_3		0.77	0.78	0.77		0.29	0.28	0.29		0.09	0.09	0.09		0.00	0.00	0.00
	T_4		0.07	0.09	0.07		0.62	0.64	0.62		0.89	0.89	0.89		0.99	0.99	0.99
Quantile																	
α=.1	T_1	25.46	19.20	18.30	18.93	19.27	17.13	17.07	17.13	17.83	16.77	16.75	16.78	16.73	16.28	16.29	16.28
	T_2	6.75	3.91	3.47	3.64	3.02	2.47	2.43	2.47	3.27	2.57	2.57	2.57	2.73	2.47	2.49	2.47
	T_3		0.96	0.98	0.97		0.97	0.97	0.97		0.93	0.92	0.93		0.88	0.88	0.88
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.3	T_1	40.75	33.69	33.57	33.67	35.74	32.83	32.91	32.86	34.28	32.86	32.84	32.87	32.86	32.21	32.21	32.21
	T_2	7.03	5.22	5.19	5.10	4.83	4.02	4.09	4.03	5.12	4.70	4.64	4.70	4.51	4.37	4.38	4.37
	T_3		0.97	0.99	0.99		0.99	0.99	0.99		0.96	0.97	0.96		0.90	0.89	0.90
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.5	T_1	45.31	38.12	38.25	38.15	40.62	37.32	37.56	37.35	38.87	37.23	37.29	37.24	37.53	36.80	36.81	36.80
	T_2	7.12	6.14	6.50	6.08	5.43	4.50	4.73	4.51	6.13	5.04	5.18	5.04	4.90	4.85	4.84	4.85
	T_3		1.00	1.00	1.00		0.99	0.98	0.99		0.99	0.99	0.99		0.93	0.92	0.93
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.7	T_1	40.10	33.98	33.87	33.92	35.56	32.39	32.52	32.41	33.86	32.44	32.48	32.45	32.70	32.09	32.13	32.10
	T_2	6.32	5.93	6.10	5.75	5.12	3.76	3.91	3.77	5.53	4.74	4.82	4.75	4.42	4.26	4.31	4.26
	T_3		0.97	0.98	0.98		0.98	0.99	0.98		0.96	0.96	0.96		0.90	0.90	0.90
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.9	T_1	24.64	19.88	18.96	19.45	19.19	16.74	16.72	16.74	17.58	16.58	16.57	16.58	16.86	16.34	16.36	16.34
	T_2	5.18	5.06	4.15	4.38	3.96	2.44	2.50	2.45	3.13	2.72	2.72	2.72	2.58	2.40	2.43	2.40
	T_3		0.91	0.97	0.96		0.98	0.98	0.98		0.92	0.91	0.92		0.88	0.87	0.88
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00

Table 2: $\rho = 0.6$, $\theta = 0$, and MW_1 (Gaussian)

		R=20				R =50				R =100				R =200			
		J=1	J =50			J=1	J =50			J=1	J =50			J=1	J =50		
Binary			EW	W_1	W_R		EW	W_1	W_R		EW	W_1	W_R		EW	W_1	W_R
α=.1	T_1	6.52	5.66	5.36	5.55	4.97	4.78	4.78	4.78	5.09	4.95	4.93	4.94	4.91	4.94	4.94	4.94
	T_2	1.45	1.54	1.28	1.44	1.08	0.86	0.86	0.86	0.98	0.90	0.88	0.89	0.92	0.84	0.84	0.84
	T_3		0.72	0.78	0.76		0.43	0.43	0.43		0.31	0.34	0.32		0.18	0.19	0.18
	T_4		0.01	0.03	0.02		0.15	0.15	0.15		0.28	0.24	0.28		0.21	0.21	0.21
α=.3	T_1	14.46	13.35	13.11	13.33	12.94	12.63	12.64	12.64	12.91	12.87	12.85	12.86	12.06	12.35	12.35	12.35
	T_2	2.41	2.24	2.14	2.22	2.57	2.24	2.28	2.25	1.90	2.00	1.97	1.99	2.28	2.21	2.22	2.20
	T_3		0.76	0.82	0.79		0.55	0.56	0.53		0.39	0.39	0.41		0.33	0.32	0.33
	T_4		0.01	0.00	0.00		0.00	0.01	0.00		0.08	0.08	0.08		0.05	0.06	0.04
α=.5	T_1	17.44	16.26	15.91	16.32	15.32	15.24	15.05	15.22	15.79	15.80	15.71	15.80	14.76	14.75	14.76	14.73
	T_2	2.73	2.68	2.64	2.76	2.80	2.78	2.70	2.78	2.32	2.21	2.15	2.18	2.50	2.39	2.42	2.42
	T_3		0.69	0.73	0.66		0.46	0.57	0.51		0.36	0.40	0.35		0.36	0.30	0.36
	T_4		0.10	0.13	0.13		0.18	0.16	0.16		0.27	0.25	0.29		0.33	0.38	0.36
α=.7	T_1	14.08	13.12	12.94	13.21	12.50	12.36	12.26	12.35	12.79	12.89	12.81	12.85	12.33	12.47	12.50	12.46
	T_2	2.46	2.40	2.25	2.45	2.38	2.14	2.24	2.26	2.47	2.36	2.28	2.31	2.05	2.06	2.10	2.05
	T_3		0.69	0.73	0.68		0.55	0.57	0.54		0.45	0.43	0.43		0.41	0.39	0.41
	T_4		0.02	0.04	0.02	1.00	0.03	0.04	0.05		0.03	0.07	0.03		0.06	0.06	0.06
α=.9	T_1	6.22	5.18	5.00	5.21	5.12	4.92	4.90	4.92	4.96	4.87	4.88	4.88	5.01	4.91	4.90	4.91
	T_2	1.53	1.36	1.13	1.37	1.01	0.91	0.89	0.92	1.00	0.83	0.84	0.84	0.96	0.84	0.84	0.85
	T_3		0.73	0.77	0.73		0.45	0.48	0.44		0.36	0.34	0.34		0.32	0.32	0.32
	T_4		0.02	0.03	0.02		0.14	0.12	0.15		0.11	0.13	0.12		0.20	0.22	0.22
Quantile																	
α=.1	T_1	26.34	22.24	19.98	21.29	20.01	18.74	18.50	18.70	18.52	18.02	17.94	18.02	17.90	17.84	17.83	17.84
	T_2	4.89	4.54	3.05	3.77	2.18	1.85	1.69	1.82	2.11	1.95	1.89	1.95	1.76	1.72	1.71	1.72
	T_3		0.86	0.99	0.92		0.86	0.93	0.87		0.75	0.78	0.75		0.50	0.50	0.50
	T_4		0.00	0.00	0.00	1.00	0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.3	T_1	42.30	38.84	37.17	38.31	37.52	36.62	36.29	36.58	35.85	35.60	35.49	35.60	35.29	35.26	35.21	35.26
	T_2	4.41	4.19	3.59	3.87	3.22	3.11	2.97	3.10	3.36	3.28	3.22	3.28	2.92	3.00	2.97	3.00
	T_3		0.90	0.99	0.94		0.82	0.85	0.83		0.60	0.67	0.60		0.57	0.57	0.57
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.5	T_1	47.65	43.98	42.47	43.60	42.35	41.33	41.02	41.29	40.76	40.48	40.34	40.48	40.23	40.18	40.13	40.17
	T_2	4.56	4.54	3.82	4.22	3.27	3.23	3.06	3.21	3.52	3.48	3.45	3.47	3.11	3.15	3.13	3.15
	T_3		0.87	0.99	0.89		0.82	0.91	0.85		0.63	0.73	0.63		0.58	0.62	0.58
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.7	T_1	42.57	39.18	37.46	38.68	36.74	35.99	35.68	35.96	35.43	35.26	35.15	35.25	34.93	34.92	34.88	34.92
	T_2	4.10	4.48	3.79	4.12	2.85	2.76	2.63	2.74	3.24	3.11	3.08	3.11	2.65	2.63	2.62	2.63
	T_3		0.84	0.98	0.90		0.68	0.77	0.71		0.68	0.71	0.68		0.55	0.58	0.55
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.9	T_1	25.92	21.52	19.70	20.74	19.61	18.36	18.16	18.33	18.30	17.87	17.83	17.86	17.80	17.71	17.69	17.70
	T_2	4.29	4.07	2.83	3.42	2.26	1.89	1.77	1.87	2.11	1.88	1.88	1.88	1.60	1.43	1.42	1.43
	T_3		0.85	0.98	0.93		0.85	0.89	0.86		0.74	0.76	0.74		0.55	0.56	0.55
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00

Table 3: $\rho = 0.6$, $\theta = 0$, **and** MW_2 **(Skewed unimodal)**

		R=20				R =50				R =100				R =200			
		J=1	J =50			J=1	J =50			J=1	J =50			J=1	J =50		
Binary			EW	W_1	W_R		EW	W_1	W_R		EW	W_1	W_R		EW	W_1	W_R
α=.1	T_1	7.36	6.00	5.71	5.88	5.57	5.14	5.13	5.14	5.51	5.29	5.29	5.29	5.28	5.27	5.27	5.27
	T_2	1.77	1.69	1.27	1.53	1.12	0.84	0.83	0.85	1.12	0.97	0.97	0.97	0.95	0.88	0.88	0.88
	T_3		0.78	0.83	0.81		0.60	0.60	0.60		0.42	0.42	0.41		0.24	0.24	0.24
	T_4		0.04	0.04	0.03		0.09	0.08	0.09		0.25	0.23	0.24		0.33	0.32	0.32
α=.3	T_1	15.14	14.50	14.05	14.50	13.68	13.48	13.43	13.46	13.16	13.17	13.10	13.19	12.88	12.98	12.95	12.98
	T_2	2.57	2.71	2.53	2.70	2.41	2.20	2.25	2.18	2.34	2.25	2.30	2.24	2.34	2.46	2.47	2.47
	T_3		0.66	0.68	0.66		0.52	0.55	0.52		0.51	0.54	0.48		0.48	0.49	0.48
	T_4		0.01	0.02	0.01		0.03	0.05	0.03		0.03	0.03	0.03		0.02	0.03	0.02
α=.5	T_1	16.61	16.01	15.50	15.98	15.33	15.17	15.07	15.21	14.62	14.71	14.65	14.70	14.24	14.23	14.22	14.22
	T_2	2.84	2.55	2.36	2.48	3.07	2.83	2.80	2.83	3.00	3.16	3.06	3.21	2.46	2.41	2.41	2.43
	T_3		0.56	0.65	0.60		0.48	0.51	0.44		0.43	0.41	0.38		0.31	0.30	0.32
	T_4		0.08	0.10	0.04		0.15	0.19	0.20		0.17	0.22	0.19		0.40	0.44	0.40
α=.7	T_1	12.80	12.04	11.85	12.09	11.41	11.40	11.40	11.38	10.83	10.85	10.83	10.83	11.20	11.31	11.27	11.29
	T_2	2.43	2.24	2.17	2.27	2.21	2.14	2.18	2.16	2.33	2.11	2.08	2.09	2.20	2.23	2.22	2.21
	T_3		0.64	0.71	0.63		0.50	0.46	0.50		0.50	0.52	0.49		0.46	0.45	0.44
	T_4		0.01	0.00	0.01		0.04	0.07	0.05		0.05	0.06	0.06		0.04	0.06	0.05
α=.9	T_1	5.46	4.79	4.69	4.85	4.50	4.48	4.47	4.48	4.29	4.41	4.40	4.40	4.35	4.46	4.43	4.44
	T_2	1.62	1.46	1.33	1.48	1.11	0.94	0.96	0.96	0.94	0.92	0.92	0.93	0.93	0.79	0.80	0.80
	T_3		0.64	0.70	0.63		0.39	0.37	0.37		0.26	0.25	0.25		0.23	0.25	0.24
	T_4		0.05	0.02	0.04		0.05	0.07	0.07		0.07	0.07	0.07		0.09	0.08	0.09
Quantile																	
α=.1	T_1	30.66	24.13	21.78	23.25	23.46	21.41	21.13	21.39	21.57	20.80	20.74	20.80	20.87	20.59	20.57	20.59
	T_2	5.98	5.05	3.56	4.26	3.20	2.64	2.55	2.62	3.09	2.71	2.66	2.71	2.32	2.23	2.22	2.23
	T_3		0.87	1.00	0.95		0.90	0.97	0.91		0.77	0.79	0.77		0.67	0.68	0.67
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.3	T_1	43.96	40.60	38.92	39.95	38.52	37.42	37.11	37.40	36.49	36.21	36.07	36.20	36.01	35.92	35.87	35.91
	T_2	5.35	5.34	4.66	4.79	4.21	3.85	3.92	3.86	4.26	4.13	4.06	4.13	3.23	3.22	3.22	3.22
	T_3		0.82	0.97	0.89		0.79	0.89	0.79		0.62	0.67	0.62		0.64	0.67	0.64
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.5	T_1	45.58	42.49	41.09	42.09	40.89	40.09	39.73	40.05	39.02	39.04	38.90	39.03	38.36	38.31	38.27	38.31
	T_2	5.07	4.85	4.22	4.55	3.86	3.56	3.55	3.55	4.01	3.98	3.91	3.97	3.31	3.23	3.23	3.23
	T_3		0.79	0.93	0.87		0.75	0.91	0.75		0.53	0.65	0.54		0.52	0.60	0.54
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.7	T_1	38.32	35.59	34.23	35.17	33.77	33.11	32.77	33.08	32.35	32.25	32.16	32.25	31.69	31.74	31.70	31.74
	T_2	4.55	4.14	3.60	3.84	2.72	2.65	2.54	2.65	2.97	2.91	2.89	2.91	2.58	2.52	2.52	2.52
	T_3		0.82	0.96	0.84		0.70	0.83	0.70		0.56	0.62	0.56		0.50	0.51	0.50
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.9	T_1	22.62	19.42	17.79	18.62	16.93	16.17	15.93	16.14	15.78	15.66	15.63	15.66	15.32	15.46	15.45	15.46
	T_2	4.49	3.66	2.74	3.12	1.96	1.62	1.46	1.60	1.54	1.53	1.52	1.53	1.35	1.27	1.26	1.27
	T_3		0.77	0.93	0.87		0.71	0.82	0.71		0.54	0.54	0.54		0.45	0.45	0.45
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00

Table 4: $\rho = 0.6,\ \theta = 0,$ **and** MW_3 **(Strongly skewed)**

		R=20				R =50				R =100				R =200			
		J=1	J =50			J=1	J =50			J=1	J =50			J=1	J =50		
Binary			*EW*	W_1	W_R		*EW*	W_1	W_R		*EW*	W_1	W_R		*EW*	W_1	W_R
α=.1	T_1	3.32	3.38	3.30	3.33	2.97	3.26	3.28	3.26	3.12	3.43	3.45	3.43	3.00	3.23	3.25	3.23
	T_2	0.81	0.71	0.61	0.64	0.63	0.56	0.56	0.56	0.68	0.72	0.72	0.72	0.66	0.70	0.70	0.70
	T_3		0.34	0.36	0.37		0.17	0.18	0.17		0.13	0.13	0.13		0.06	0.06	0.06
	T_4		0.06	0.06	0.07		0.08	0.05	0.08		0.08	0.06	0.08		0.11	0.10	0.11
α=.3	T_1	9.46	9.34	9.04	9.28	8.35	8.30	8.29	8.30	8.62	8.74	8.74	8.74	8.57	8.60	8.60	8.61
	T_2	2.03	2.11	1.83	1.99	1.55	1.58	1.56	1.58	1.59	1.60	1.61	1.61	1.70	1.78	1.78	1.79
	T_3		0.52	0.60	0.56		0.32	0.31	0.32		0.30	0.30	0.31		0.34	0.34	0.34
	T_4		0.06	0.07	0.04		0.34	0.34	0.34		0.23	0.23	0.22		0.33	0.33	0.31
α=.5	T_1	14.59	14.09	13.64	14.00	12.85	12.72	12.61	12.70	12.99	13.06	13.02	13.02	13.13	13.01	13.04	13.00
	T_2	2.68	3.03	2.82	2.96	2.38	2.28	2.25	2.27	2.36	2.38	2.30	2.33	2.61	2.55	2.54	2.56
	T_3		0.57	0.67	0.57		0.45	0.52	0.45		0.32	0.32	0.31		0.39	0.41	0.39
	T_4		0.12	0.14	0.14		0.26	0.17	0.26		0.34	0.38	0.37		0.38	0.35	0.37
α=.7	T_1	16.12	14.44	14.16	14.52	14.83	14.58	14.37	14.46	14.25	14.47	14.32	14.43	14.12	14.18	14.11	14.18
	T_2	2.62	2.61	2.48	2.56	2.65	2.37	2.28	2.36	2.32	2.28	2.22	2.21	2.38	2.43	2.35	2.42
	T_3		0.76	0.82	0.76		0.56	0.62	0.61		0.42	0.49	0.43		0.40	0.42	0.37
	T_4		0.00	0.00	0.01		0.00	0.03	0.02		0.03	0.03	0.04		0.07	0.08	0.08
α=.9	T_1	8.80	6.88	6.86	6.99	6.45	5.98	5.98	5.99	5.90	5.72	5.72	5.72	5.83	5.79	5.78	5.78
	T_2	1.94	1.85	1.69	1.87	1.18	0.89	0.91	0.91	1.05	0.90	0.90	0.90	1.13	0.99	0.98	0.98
	T_3		0.85	0.84	0.84		0.51	0.52	0.52		0.30	0.29	0.30		0.14	0.14	0.14
	T_4		0.01	0.04	0.02		0.16	0.18	0.17		0.40	0.40	0.40		0.66	0.66	0.66
Quantile																	
α=.1	T_1	12.96	13.34	11.83	12.30	9.52	10.31	10.23	10.28	9.39	10.16	10.12	10.15	9.19	9.74	9.72	9.74
	T_2	2.70	2.96	1.66	1.79	1.01	1.14	1.12	1.13	1.07	1.35	1.33	1.34	0.98	1.18	1.18	1.18
	T_3		0.46	0.66	0.61		0.05	0.08	0.06		0.07	0.08	0.08		0.03	0.03	0.03
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.3	T_1	30.27	29.96	28.27	29.10	25.85	25.89	25.71	25.86	25.50	25.65	25.58	25.64	25.07	25.16	25.14	25.15
	T_2	4.04	4.34	3.46	3.54	2.47	2.54	2.47	2.53	2.90	2.94	2.91	2.93	2.68	2.65	2.65	2.65
	T_3		0.61	0.76	0.69		0.49	0.56	0.51		0.41	0.45	0.42		0.41	0.46	0.42
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.5	T_1	43.05	40.51	39.27	40.01	37.70	37.19	36.97	37.15	37.07	36.96	36.88	36.96	36.25	36.29	36.25	36.29
	T_2	5.17	5.09	4.67	4.84	3.82	3.74	3.65	3.73	4.23	4.05	4.05	4.05	3.90	3.72	3.72	3.72
	T_3		0.79	0.89	0.82		0.64	0.76	0.65		0.56	0.64	0.57		0.49	0.51	0.49
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.7	T_1	47.22	42.05	40.95	41.61	40.95	39.29	39.18	39.27	39.69	39.29	39.15	39.29	38.56	38.37	38.34	38.37
	T_2	5.69	5.70	5.34	5.44	4.65	4.52	4.47	4.51	4.72	4.62	4.55	4.62	3.99	3.77	3.78	3.77
	T_3		0.91	0.98	0.95		0.91	0.94	0.92		0.68	0.75	0.68		0.63	0.63	0.64
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.9	T_1	35.58	27.37	25.20	26.31	26.28	23.70	23.47	23.70	24.36	23.09	22.98	23.08	23.36	22.87	22.84	22.87
	T_2	6.64	5.74	4.33	4.47	3.92	3.23	3.09	3.21	3.10	2.84	2.75	2.83	2.44	2.18	2.17	2.18
	T_3		0.91	0.98	0.96		0.94	0.99	0.96		0.89	0.94	0.89		0.80	0.81	0.80
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00

Table 5: $\rho = 0.6$, $\theta = 0$, **and** MW_4 **(Kurtotic unimodal)**

		R=20				R =50				R =100				R =200			
		J=1	J =50			J=1	J =50			J=1	J =50			J=1	J =50		
Binary			EW	W_1	W_R		EW	W_1	W_R		EW	W_1	W_R		EW	W_1	W_R
α=.1	T_1	6.51	5.46	5.12	5.31	5.26	4.74	4.77	4.75	4.98	4.83	4.84	4.84	4.90	4.95	4.95	4.95
	T_2	1.58	1.41	1.14	1.26	1.37	0.97	0.99	0.97	1.10	0.92	0.90	0.92	1.08	0.96	0.96	0.96
	T_3		0.72	0.79	0.77		0.58	0.57	0.58		0.38	0.36	0.37		0.22	0.22	0.22
	T_4		0.04	0.04	0.03		0.11	0.11	0.11		0.18	0.18	0.18		0.16	0.16	0.16
α=.3	T_1	13.02	12.27	12.09	12.32	11.55	11.53	11.48	11.55	11.50	11.58	11.52	11.59	11.34	11.31	11.31	11.30
	T_2	2.23	2.11	1.92	2.13	2.23	2.25	2.21	2.25	2.15	2.06	2.06	2.06	2.13	2.22	2.22	2.21
	T_3		0.65	0.71	0.65		0.49	0.46	0.46		0.47	0.45	0.46		0.48	0.48	0.47
	T_4		0.01	0.01	0.00		0.00	0.03	0.00		0.02	0.02	0.02		0.04	0.04	0.04
α=.5	T_1	14.51	14.33	13.97	14.32	12.61	12.78	12.58	12.74	12.64	12.69	12.66	12.70	12.30	12.27	12.28	12.28
	T_2	2.51	2.64	2.55	2.66	2.07	2.18	2.06	2.17	2.36	2.37	2.33	2.37	2.45	2.46	2.48	2.48
	T_3		0.49	0.60	0.49		0.29	0.38	0.32		0.30	0.26	0.27		0.25	0.24	0.25
	T_4		0.08	0.09	0.07		0.30	0.24	0.25		0.36	0.47	0.42		0.61	0.58	0.59
α=.7	T_1	12.91	11.97	11.68	11.96	11.54	11.64	11.54	11.56	11.47	11.56	11.55	11.56	11.22	11.22	11.24	11.22
	T_2	2.37	2.38	2.18	2.42	2.14	1.98	1.95	1.97	2.17	2.09	2.12	2.11	2.33	2.25	2.25	2.28
	T_3		0.69	0.73	0.69		0.50	0.49	0.56		0.49	0.48	0.46		0.43	0.40	0.41
	T_4		0.01	0.00	0.00		0.00	0.01	0.00		0.02	0.03	0.02		0.04	0.04	0.06
α=.9	T_1	6.55	5.24	5.24	5.33	5.21	4.89	4.89	4.89	5.00	4.94	4.94	4.94	4.83	4.89	4.88	4.89
	T_2	1.65	1.42	1.43	1.47	1.20	1.01	1.03	1.02	1.09	0.89	0.90	0.90	1.01	0.97	0.97	0.97
	T_3		0.81	0.79	0.80		0.55	0.55	0.54		0.33	0.31	0.31		0.21	0.22	0.22
	T_4		0.02	0.02	0.02		0.04	0.06	0.04		0.11	0.15	0.15		0.22	0.21	0.21
Quantile																	
α=.1	T_1	29.60	23.20	20.80	22.14	22.02	19.98	19.79	19.97	20.06	19.35	19.26	19.34	19.26	19.14	19.13	19.14
	T_2	6.16	5.30	3.57	4.28	2.91	2.50	2.35	2.47	2.60	2.36	2.22	2.35	2.37	2.19	2.18	2.19
	T_3		0.85	0.99	0.95		0.90	0.93	0.90		0.82	0.83	0.83		0.59	0.56	0.59
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.3	T_1	39.25	37.30	35.50	36.58	34.56	33.95	33.74	33.92	32.98	33.09	33.01	33.08	32.53	32.75	32.72	32.74
	T_2	4.76	5.00	4.43	4.63	4.46	4.11	4.04	4.11	3.93	3.68	3.66	3.68	3.80	3.74	3.72	3.74
	T_3		0.72	0.92	0.76		0.61	0.67	0.62		0.51	0.51	0.51		0.39	0.40	0.39
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.5	T_1	40.39	40.21	38.38	39.50	35.15	35.75	35.39	35.69	34.41	34.70	34.62	34.69	34.14	34.29	34.28	34.29
	T_2	4.58	5.14	4.57	4.88	3.93	4.08	3.91	4.05	3.87	3.93	3.91	3.93	3.66	3.72	3.71	3.72
	T_3		0.52	0.79	0.68		0.27	0.43	0.29		0.32	0.36	0.32		0.37	0.40	0.37
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.7	T_1	39.79	36.77	35.32	36.27	33.55	33.39	33.13	33.35	32.94	32.92	32.82	32.91	32.55	32.74	32.72	32.74
	T_2	5.66	5.01	4.58	4.82	4.16	4.15	4.04	4.13	4.14	3.91	3.87	3.91	3.56	3.42	3.42	3.42
	T_3		0.84	0.93	0.90		0.51	0.57	0.51		0.50	0.53	0.50		0.39	0.39	0.39
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00
α=.9	T_1	29.40	23.19	21.07	22.21	21.69	19.62	19.47	19.61	19.97	19.22	19.16	19.22	19.31	19.11	19.08	19.11
	T_2	5.55	5.68	3.84	4.56	3.45	2.69	2.60	2.67	2.63	2.51	2.43	2.51	2.29	2.05	2.01	2.05
	T_3		0.92	0.99	0.95		0.92	0.93	0.92		0.77	0.81	0.77		0.58	0.60	0.58
	T_4		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00		0.00	0.00	0.00

Figure 1(a): SP500 Binary Prediction

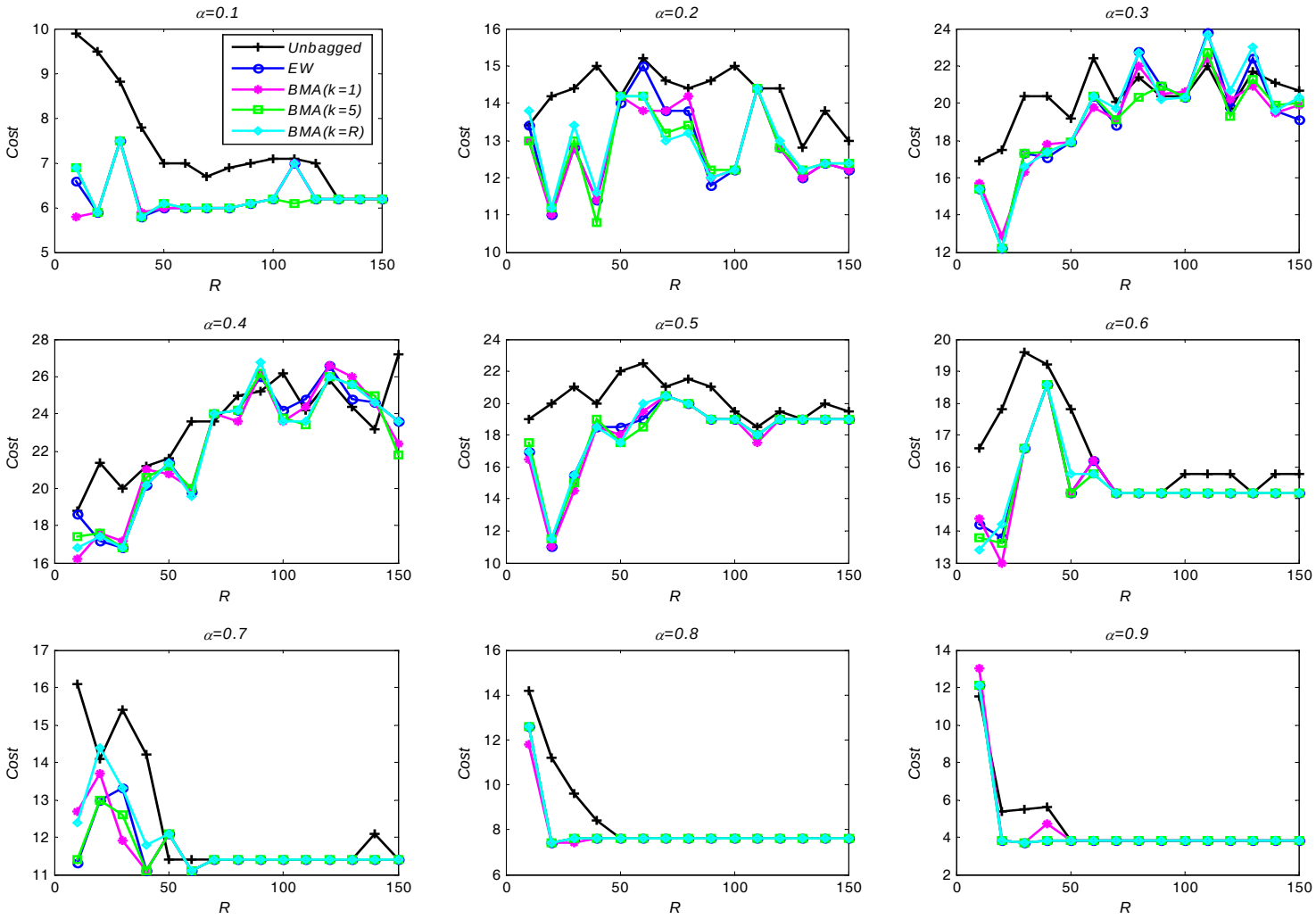

Figure 1(b): SP500 Quantile Prediction

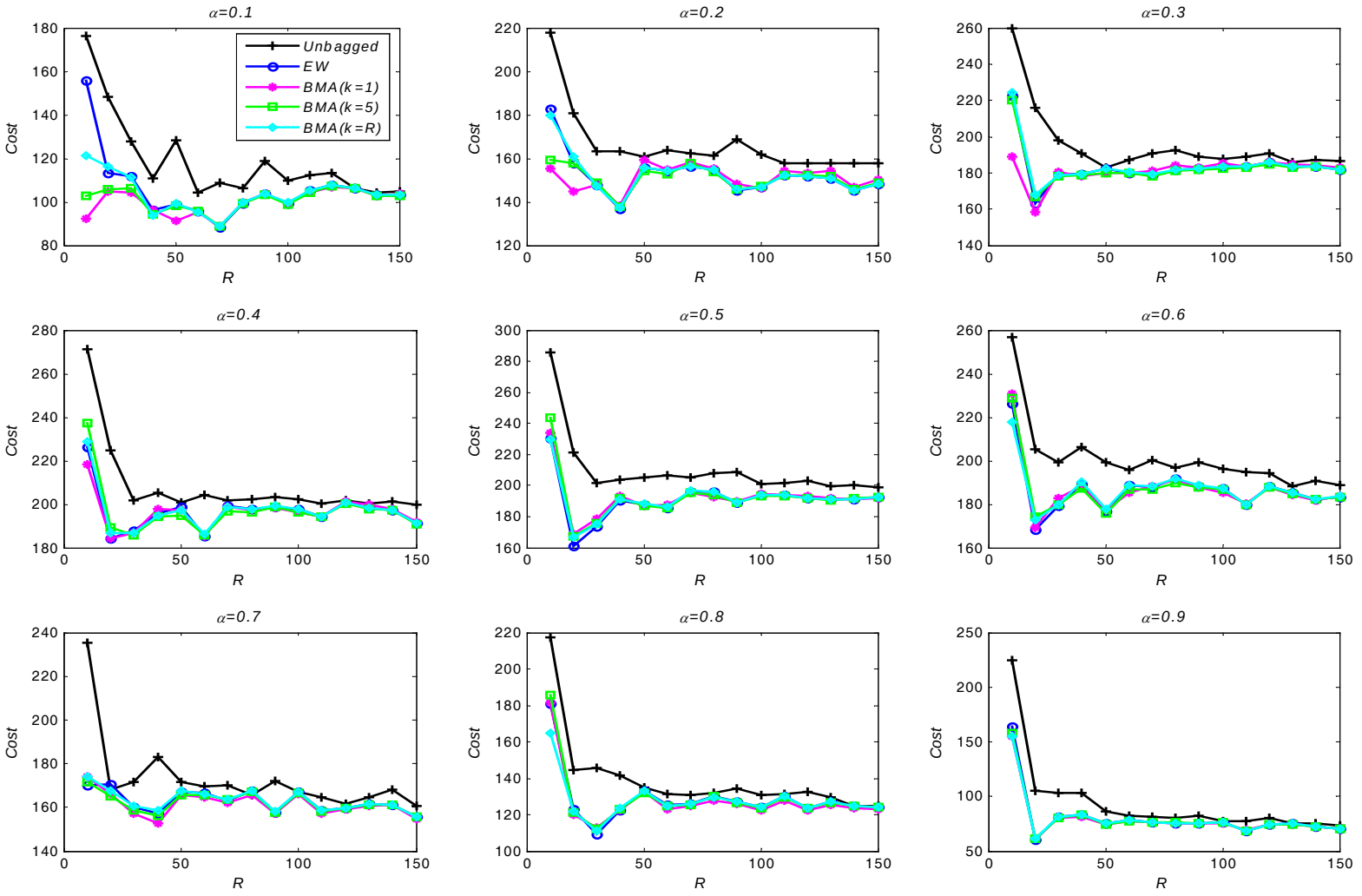

Figure 2(a): NASDAQ Binary Prediction

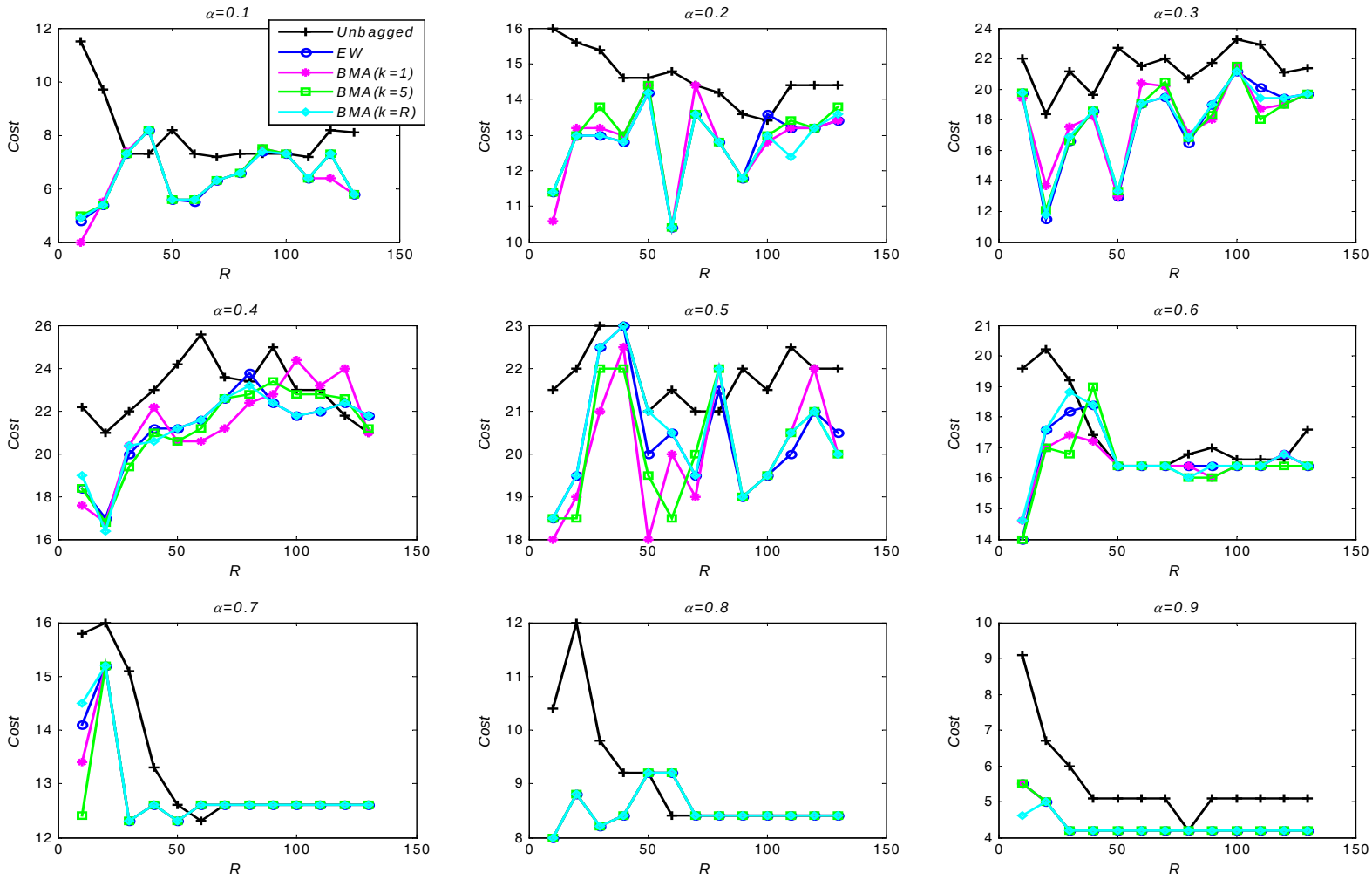

Figure 2(b): NASDAQ Quantile Prediction