*Econometric Reviews*, 33(5–6):575–605, 2014

ISSN: 0747-4938 print/1532-4168 online
DOI: 10.1080/07474938.2013.825142

Taylor & Francis
Taylor & Francis Group

# ROBUSTIFY FINANCIAL TIME SERIES FORECASTING WITH BAGGING

**Sainan Jin[1], Liangjun Su[1], and Aman Ullah[2]**

[1]*School of Economics, Singapore Management University, Singapore, Singapore*
[2]*Department of Economics, University of California, Riverside, California, USA*

□ *In this paper we propose a revised version of (bagging)* **b***ootstrap* **aggr***egat***ing** *as a forecast combination method for the out-of-sample forecasts in time series models. The revised version explicitly takes into account the dependence in time series data and can be used to justify the validity of bagging in the reduction of mean squared forecast error when compared with the unbagged forecasts. Monte Carlo simulations show that the new method works quite well and outperforms the traditional one-step-ahead linear forecast as well as the nonparametric forecast in general, especially when the in-sample estimation period is small. We also find that the bagging forecasts based on misspecified linear models may work as effectively as those based on nonparametric models, suggesting the robustification property of bagging method in terms of out-of-sample forecasts. We then reexamine forecasting powers of predictive variables suggested in the literature to forecast the excess returns or equity premium. We find that, consistent with Goyal and Welch (2008), the historical average excess stock return forecasts may beat other predictor variables in the literature when we apply traditional one-step linear forecast and the nonparametric forecasting methods. However, when using the bagging method or its revised version, which help to improve the mean squared forecast error for "unstable" predictors, the predictive variables have a better forecasting power than the historical mean.*

**Keywords** Bagging; Combined forecasts; Nonparametric models; Predictability; Time series.

**JEL Classification** C14; C53.

## 1. INTRODUCTION

Since the introduction of bagging by Breiman (1996a,b) to reduce the variance of a predictor, (bagging) **b**ootstrap **agg**regat**ing** has attracted a great deal of attention in both statistics and economics. See Bühlmann and Yu (2002), Buja and Stuetzle (2006), Lee and Yang (2006), Friedman and

Address correspondence to Aman Ullah, Department of Economics, University of California, Riverside, CA 92521, USA; E-mail: aman.ullah@ucr.edu

Hall (2007), Inoue and Kilian (2008), Lee et al. (2010), among others. Lee and Yang (2006) showed that bagging may improve the binary and quantile predictions in small samples using asymmetric loss functions. Inoue and Kilian (2008) studied how useful bagging is in forecasting economic time series by examining the inflation prediction in great details. Stock and Watson (2009) found that bagging is asymptotically a shrinkage forecast. Lee et al. (2010) considered nonparametric and semiparametric regression models with the use of bagging to impose economic constraints. It is worth mentioning that the statistical theory of bagging by Breiman (1996a,b) was only developed for independent and identically distributed (IID) data, and the extension of the application of bagging to economic time series data has largely ignored the dependence structure in the data despite the fact that the usual bootstrapping procedure in the bagging was typically replaced by the method of block bootstrapping. In particular, the formal justification of bagging on the reduction on mean squared forecast error (MSFE) was not well offered in the literature for time series data.

In this paper, we examine how to revise the current bagging method for the use in the time series framework, and propose some new combined forecasting methods based on the idea of bagging. The new methods combine forecasts by bagging the out-of-sample time series forecasts based on linear, local constant, and local linear regression models. Monte Carlo simulations show that the new methods work quite well and outperform the traditional one-step linear forecast as well as the nonparametric (local constant and local linear) forecasts in most of the cases under investigation, especially when the in-sample estimation period ($R$) is small. When for example, $R = 20$, the percentage gains of our proposed methods are usually above 5%.

In addition, we find that bagging forecasts based on different forecasting models all yield similar percentage of gains on the reduction of MSFE when compared with the simple linear forecast models despite the fact that the unbagged forecasts based on nonparametric regressions may significantly outperform the unbagged forecasts based on linear regressions when the underlying model is nonlinear. In other words, the gains by using the nonparametric model in conjunction with the bagging method tend to be marginal in comparison with the gains of bagging forecasts over unbagged forecasts. This suggests that bagging forecasts based on misspecified parametric models may work as effectively as those based on correctly specified nonparametric models. So the great advantage of bagging lies in its robustifying forecasts based on different approximating models.

It is a long tradition in finance and economics to attempt to predict stock market returns or the equity premium. Recently there is a vast amount of empirical studies on the predictability of stock market returns using various lagged financial or macro variables, such as dividend price

ratio, earnings yield and dividend-earnings (payout) ratio, various interest rates and spreads, inflation rates, book-to-market ratio, investment-capital ratio, consumption, wealth, and income ratio, and aggregate net or equity issuing activity. Goyal and Welch (2008) claimed that the historical average excess stock return forecasts beat other predictor variables in the literature and predictive regressions have performed poorly out-of-sample. Campbell and Thompson (2008) showed that many predictive regressions perform better than the historical average return if some restrictions are imposed on the signs of regression coefficients and return forecasts. Chen and Hong (2009) argued that the restriction proposed by Campbell and Thompson (2008) is a form of nonlinearity and found nonparametric predictive methods outperform the historical average in quarterly or annual frequency. Lee et al. (2010) considered nonparametric and semiparametric regression models with the use of bagging to impose economic constraints. They, like Chen and Hong (2009), found annual return prediction favors much of their proposed models over the historical average. However, monthly return prediction is beyond the reign of all models they considered with few exceptions. In view of this, here we only analyze the case of monthly return prediction.

In this paper, we reexamine the forecasting performances of predictive variables suggested in the literature. In addition, it is also interesting to consider forecasting models for stock returns by using lagged stock returns as an additional predictive variable. This is because there are quite a few works devoted to testing if the stock returns are autocorrelated or martingale difference sequences, or exhibit other types of dependent structures. (For an overview, see Campbell et al. (1997).) We find that, consistent with Goyal and Welch (2008), the monthly historical average excess stock returns forecasts beat other predictor variables in the literature when we apply traditional one-step-ahead linear forecast and the nonparametric forecasting methods.

When using our new combined forecast methods and the bagging methods, which help to improve the mean squared error for "unstable" predictors, we find that the monthly excess returns are actually predictable for a variety of predictor variables. When the in-sample estimation period is small, say, e.g., $R = 24$ (i.e., one uses only the last two years of observations to estimate the model), the percentage gains over the historical mean are above 3% in all the cases (with different subsamples) using our combined forecast methods and above 1% using the bagging methods. As Pesaran and Timmermann (2002, 2004) put it, simply using as many observations as possible is not a sensible choice for out-of-sample prediction in time series where model instability is of our concern or structural breaks may occur frequently. In this circumstance, small sample training period may be desired and various combined forecasts are particularly relevant and effective. When $R$ is large (e.g., $R = 100$), the percentage gains of our

method are small but it still outperforms the historical mean in most cases. As claimed in Campbell and Thompson (2008), the small out-of-sample predictive power is actually economically meaningful.

The rest of the paper is structured as follows. Section 2 motivates and proposes the new forecasting method in the time series framework. A small set of Monte Carlo simulation results are reported in Section 3. In section 4 we apply the proposed method to predict the U.S. excess stock returns. Final remarks are contained in Section 5.

## 2. MOTIVATION AND A REVISED VERSION OF BAGGING

In this section we first motivate the idea of bagging in time series framework and propose a revised version of bagging,

### 2.1. Motivation

In this paper we consider a simple data generating process

$$y_{t+1} = m(x_t) + \varepsilon_{t+1}, \quad t = 0, \ldots, T-1, \tag{2.1}$$

where $x_t$ is a $q \times 1$ vector of variables, $y_{t+1}$ is a scalar, $m(\cdot)$ is a smooth function, and $\varepsilon_{t+1}$ is an error term such that $E(\varepsilon_{t+1} \mid x_t) = 0$ almost surely (a.s.).

Define the training set at time $t = R, \ldots, T-1$

$$\mathscr{D}_t \equiv \{(y_j, x_{j-1})\}_{j=t-R+1}^{t}, \tag{2.2}$$

which consists of $R$ observations. Let $\varphi(x_t, \mathscr{D}_t)$ denote a forecast of $y_{t+1}$ using the training set $\mathscr{D}_t$ and input vector $x_t$.

Assuming that each training set $\mathscr{D}_t$ consists of $R$ observations drawn from a distribution $\mathbf{P}_{R,t}$, Lee and Yang (2006) tried to use the set $\mathscr{D}_t$ to obtain a better predictor than the single training set predictor $\varphi(x_t, \mathscr{D}_t)$. If $\mathbf{P}_{R,t}$ is known, one can easily update the predictor $\varphi(x_t, \mathscr{D}_t)$ by the *ensemble aggregating predictor*[1]

$$\varphi_A(x_t) \equiv E_{\mathscr{D}_t}[\varphi(x_t, \mathscr{D}_t)], \tag{2.3}$$

where $E_{\mathscr{D}_t}$ denotes expectation with respect to $\mathscr{D}_t$. With a simple application of the Jensen inequality, Breiman (1996a) showed that $\varphi_A(x_t)$ has no larger *mean squared forecast error* (MSFE) than $\varphi(x_t, \mathscr{D}_t)$ under the condition that $\{(y_{t+1}, x_t)\}$ forms an IID sequence. Lee and Yang (2006)

[1]In practice, one could draw the multiple training sets $\mathscr{D}_t^{(b)}$ $(b = 1, \ldots, B)$ from $\mathbf{P}$ and employ a weighted version of the ensemble aggregating predictor $\varphi_{A,w}(x_t) \equiv \sum_{b=1}^{B} w_{b,t}\varphi(x_t, \mathscr{D}_t^{(b)})$, where $w_{b,t}$ is the weight function with $\sum_{b=1}^{B} w_{b,t} = 1$.

tried to extend Breiman (1996a) theory from the IID case to time series case. But a close examination of the proof of their Proposition 1 suggests they are not proving the above claim. Instead, they proved

$$\mathbb{E}_{\mathcal{D}_t, y_{t+1}, x_t}[y_{t+1} - \varphi(x_t, \mathcal{D}_t)]^2 \geq E[y_{t+1} - \varphi_A(x_t)]^2 \tag{2.4}$$

where $\mathbb{E}_{\mathcal{D}_t, y_{t+1}, x_t}(\cdot) \equiv E_{x_t} E_{y_{t+1}|x_t}[E_{\mathcal{D}_t}(\cdot)]$ and $E_{y_{t+1}|x_t}[\cdot]$ denotes expectation taken over $y_{t+1}$ conditional on $x_t$. Clearly, the object on right hand side of (2.4) is the MSFE for the ensemble aggregate predictor $\varphi_A(x_t)$, but the object on the left hand side of (2.4) is distinct from $E[y_{t+1} - \varphi(x_t, \mathcal{D}_t)]^2$, the MSFE of the original predictor $\varphi(x_t, \mathcal{D}_t)$. In short, Lee and Yang (2006) did not make comparison between MSFE's per se for the general time series case, and (2.4) reduces to the usual MSFE comparison result only in some special cases, say, when $\{(y_{t+1}, x_t)\}$ forms an independent sequence.

To motivate our method, we apply the law of iterated expectations and Jensen's inequality to obtain

$$\begin{aligned}
&E[y_{t+1} - \varphi(x_t, \mathcal{D}_t)]^2 \\
&\quad = E[E\{[y_{t+1} - \varphi(x_t, \mathcal{D}_t)]^2 \mid x_t, y_{t+1}\}] \\
&\quad = E\{y_{t+1}^2 - 2y_{t+1}E[\varphi(x_t, \mathcal{D}_t) \mid x_t, y_{t+1}] + E[\varphi^2(x_t, \mathcal{D}_t) \mid x_t, y_{t+1}]\} \\
&\quad \geq E\{y_{t+1}^2 - 2y_{t+1}E[\varphi(x_t, \mathcal{D}_t) \mid x_t, y_{t+1}] + \{E[\varphi(x_t, \mathcal{D}_t) \mid x_t, y_{t+1}]\}^2\} \\
&\quad = E\{y_{t+1} - E[\varphi(x_t, \mathcal{D}_t) \mid x_t, y_{t+1}]\}^2.
\end{aligned}$$

Consequently,

$$E[y_{t+1} - \varphi(x_t, \mathcal{D}_t)]^2 \geq E\{y_{t+1} - E[\varphi(x_t, \mathcal{D}_t) \mid x_t, y_{t+1}]\}^2. \tag{2.5}$$

To proceed, it is worthwhile to make several remarks.

**Remark 1.** The equality holds in (2.5) if and only if $E[\varphi(x_t, \mathcal{D}_t) \mid x_t, y_{t+1}] = \varphi(x_t, \mathcal{D}_t)$ a.s. Obviously, such a condition is rarely met in any meaningful forecast situation because it indicates that observing the future value $y_{t+1}$ does not affect the forecast at time $t$. As a result, (2.5) implies that $E[\varphi(x_t, \mathcal{D}_t) \mid x_t, y_{t+1}]$, as a predictor for $y_{t+1}$, has a strictly smaller MSFE than $\varphi(x_t, \mathcal{D}_t)$.

**Remark 2.** Notice that at time $t$, $\varphi(x_t, \mathcal{D}_t)$ is observable but $E[\varphi(x_t, \mathcal{D}_t) \mid x_t, y_{t+1}]$ is generally not. So the latter is an infeasible predictor and cannot be employed in practice. Furthermore, at the first sight it seems impossible for us to estimate $E[\varphi(x_t, \mathcal{D}_t) \mid x_t, y_{t+1}]$ consistently at time $t$ because $y_{t+1}$ is not observed at time $t$. Nevertheless, we shall argue that such a conditional expectation can be approximated by $E[\varphi(x_t, \mathcal{D}_t) \mid x_t]$

under certain regularity conditions, and the latter can be estimated consistently so that we can improve over the original predictor $\varphi(x_t, \mathscr{D}_t)$.

To introduce the regularity condition, let $\mathscr{A}\backslash\mathscr{B}$ denote the complement of $\mathscr{B}$ with respect to $\mathscr{A}$ and $\#\mathscr{A}$ the cardinality of the set $\mathscr{A}$. We use $\mathscr{A}_1 \perp \mathscr{A}_2 \,|\, \mathscr{A}_3$ to denote that $\mathscr{A}_1$ and $\mathscr{A}_2$ are conditionally independent given $\mathscr{A}_3$. We make the following assumption.

**Assumption I.**

(i) *There exists a partition of the training set* $\mathscr{D}_t$ *such that* $\mathscr{D}_t = \widetilde{\mathscr{D}}_t \cup (\mathscr{D}_t \backslash \widetilde{\mathscr{D}}_t)$ *and* $\widetilde{\mathscr{D}}_t \perp y_{t+1} \,|\, x_t$.
(ii) $\#(\mathscr{D}_t \backslash \widetilde{\mathscr{D}}_t)$ *is finite.*

Heuristically speaking, the above assumption implies that we can partition the training set $\mathscr{D}_t$ into two subsets, one contains some information that is not independent of the forecasting object $y_{t+1}$ given the input vector $x_t$ at time $t$, and the other contains some information that is independent of $y_{t+1}$ given $x_t$. The requirement on the finite cardinality of the first subset can be relaxed at the cost of lengthy arguments. To be concrete, Assumption I(ii) implies that $\widetilde{D}_t$ is the set $D_t$ excluding some finite number of elements including $x_t, x_{t-j_1}, \ldots, x_{t-j_m}$, $y_t, y_{t-k_1}, \ldots, y_{t-k_l}$, where $j_1, j_2, \ldots, j_m$, $k_1, k_2, \ldots, k_l$ are integers and $m$ and $l$ are finite. Without loss of generality and for notational simplicity, we assume that $j_i = i$ for $i = 1, 2, \ldots, m$, and $k_i = i$ for $i = 1, 2, \ldots, l$.

The above assumption can be easily satisfied in some empirical applications. The leading example is when $y_t$ is a $p$th order Markov process satisfying

$$f_t(y_t \,|\, \mathscr{F}_{t-1}) = f_t(y_t \,|\, y_{t-1}, y_{t-2}, \ldots, y_{t-p}) \text{ a.s.} \tag{2.6}$$

and $x_t = (y_t, y_{t-1}, \ldots, y_{t-p+1})'$ in (2.1), where $f_t(\cdot \,|\, \mathscr{A})$ denotes the conditional probability density function of $y_t$ given $\mathscr{A}$, $\mathscr{F}_t$ is a filtration to which $y_t$ is adapted, and $p$ is finite. For many applications, the most relevant choice of $\mathscr{F}_t$ appears to be the natural filtration of $y_t$, in which case $\mathscr{F}_t$ for each $t \geq 1$ is defined to be the $\sigma$-field generated by $y_s$ for all $s \leq t$. Another example is when $\{(x_t, \varepsilon_{t+1})\}$ is a $p$th order Markov process in (2.1).

We will argue that under Assumption I and some standard conditions on the data generating process, $E[\varphi(x_t, \mathscr{D}_t) \,|\, x_t, y_{t+1}]$ can be estimated consistently, and its consistent estimate can outperform the original forecaster $\varphi(x_t, \mathscr{D}_t)$ in finite samples.

## 2.2. Forecasting Models

In this paper we consider both parametric and nonparametric forecasting models. For the parametric forecasting models we focus on the linear forecasting models for the easiness of illustration.

### 2.2.1. *Linear Forecasting Models*

We first consider the simple linear forecasting model

$$y_{t+1} = \beta' x_t + e_{t+1}, \tag{2.7}$$

where $x_t$ typically includes the constant term. Based on the training set $\mathscr{D}_t$ at time $t$, we obtain the ordinary least squares (OLS) estimate of $\beta$ given by

$$\hat{\beta}_t \equiv \hat{\beta}(\mathscr{D}_t) \equiv \left( \sum_{s=t-R}^{t-1} x_s x_s' \right)^{-1} \left( \sum_{s=t-R}^{t-1} x_s y_{s+1} \right),$$

where $A \equiv B$ denotes $A$ is defined as or by $B$. The traditional one-step-ahead forecast is given by

$$\hat{y}_{t+1\,|\,t} = \hat{\beta}_t' x_t \equiv \varphi_1(x_t, \mathscr{D}_t) \qquad \text{for } t = R, \ldots, T-1. \tag{2.8}$$

The new forecasting method we propose is based upon $E[\varphi_1(x_t, \mathscr{D}_t) \,|\, x_t, y_{t+1}]$, which is an infeasible predictor. To obtain an approximation for it, let $k = \max(l+1, m)$, and then we have

$$\begin{aligned} \varphi_1(x_t, \mathscr{D}_t) &= x_t' \left( \sum_{s=t-R}^{t-1} x_s x_s' \right)^{-1} \left( \sum_{s=t-R}^{t-1} x_s y_{s+1} \right) \\ &= x_t' \left( \sum_{s=t-R}^{t-k-1} x_s x_s' \right)^{-1} \left( \sum_{s=t-R}^{t-1} x_s y_{s+1} \right) \\ &\quad + x_t' \left[ \left( \sum_{s=t-R}^{t-1} x_s x_s' \right)^{-1} - \left( \sum_{s=t-R}^{t-k-1} x_s x_s' \right)^{-1} \right] \left( \sum_{s=t-R}^{t-1} x_s y_{s+1} \right). \end{aligned} \tag{2.9}$$

For the second term in (2.9), we have

$$x_t' \left[ \left( \sum_{s=t-R}^{t-1} x_s x_s' \right)^{-1} - \left( \sum_{s=t-R}^{t-k-1} x_s x_s' \right)^{-1} \right] \left( \sum_{s=t-R}^{t-1} x_s y_{s+1} \right)$$

$$
\begin{aligned}
&= x_t' \left( \frac{1}{R} \sum_{s=t-R}^{t-k-1} x_s x_s' \right)^{-1} \left[ \frac{1}{R} \sum_{s=t-R}^{t-k-1} x_s x_s' - \frac{1}{R} \sum_{s=t-R}^{t-1} x_s x_s' \right] \\
&\quad \times \left( \frac{1}{R} \sum_{s=t-R}^{t-1} x_s x_s' \right)^{-1} \left( \frac{1}{R} \sum_{s=t-R}^{t-1} x_s y_{s+1} \right) \\
&= x_t' \left( \frac{1}{R} \sum_{s=t-R}^{t-k-1} x_s x_s' \right)^{-1} \left( -\frac{1}{R} \sum_{s=t-k}^{t-1} x_s x_s' \right) \left( \frac{1}{R} \sum_{s=t-R}^{t-1} x_s x_s' \right)^{-1} \left( \frac{1}{R} \sum_{s=t-R}^{t-1} x_s y_{s+1} \right) \\
&= O_p\left( \frac{1}{R} \right) = o_p(1)
\end{aligned}
$$

as $R \to \infty$ and $k$ is finite, since by standard assumptions on heterogenous and weakly dependent data (see, e.g., White, 2001),

$$
\begin{aligned}
&\frac{1}{R} \sum_{s=t-R}^{t-k-1} x_s x_s' = O_p(1), \quad \frac{1}{R} \sum_{s=t-R}^{t-1} x_s x_s' = O_p(1), \\
&\frac{1}{R} \sum_{s=t-k}^{t-1} x_s x_s' = O_p\left( \frac{1}{R} \right), \quad \frac{1}{R} \sum_{s=t-R}^{t-1} x_s y_{s+1} = O_p(1).
\end{aligned}
$$

By the same token,

$$
x_t' \left( \sum_{s=t-R}^{t-k-1} x_s x_s' \right)^{-1} \left( \sum_{s=t-R}^{t-1} x_s y_{s+1} \right) = x_t' \left( \sum_{s=t-R}^{t-k-1} x_s x_s' \right)^{-1} \left( \sum_{s=t-R}^{t-k-1} x_s y_{s+1} \right) + O_p\left( \frac{1}{R} \right).
$$

It follows that

$$
\begin{aligned}
\varphi_1(x_t, \mathcal{D}_t) &= x_t' \left( \sum_{s=t-R}^{t-k-1} x_s x_s' \right)^{-1} \left( \sum_{s=t-R}^{t-k-1} x_s y_{s+1} \right) + O_p\left( \frac{1}{R} \right) \\
&\equiv \varphi_1(x_t, \widetilde{\mathcal{D}}_t) + O_p\left( \frac{1}{R} \right),
\end{aligned}
$$

where the conditional expectation of the $O_p(\frac{1}{R})$ term given $(x_t, y_{t+1})$ can also be shown to be $O_p(\frac{1}{R})$ under some regularity conditions.

By Assumption I(i) and arguments similar to those used above,

$$
\begin{aligned}
E[\varphi_1(x_t, \widetilde{\mathcal{D}}_t) \mid x_t, y_{t+1}] &= E[\varphi_1(x_t, \widetilde{\mathcal{D}}_t) \mid x_t] \\
&= E\left[ \left( x_t' \left( \sum_{s=t-R}^{t-k-1} x_s x_s' \right)^{-1} \left( \sum_{s=t-R}^{t-k-1} x_s y_{s+1} \right) \right) \middle| \, x_t \right]
\end{aligned}
$$

$$= E\left[\left(x_t'\left(\sum_{s=t-R}^{t-1} x_s x_s'\right)^{-1}\left(\sum_{s=t-R}^{t-1} x_s y_{s+1}\right)\right)\middle| x_t\right] + O_p\left(\frac{1}{R}\right)$$

$$= E[\hat{\beta}_t' x_t \mid x_t] + O_p\left(\frac{1}{R}\right) = E\left[\varphi_1(x_t, \mathcal{D}_t) \mid x_t\right] + O_p\left(\frac{1}{R}\right).$$

It follows that

$$E\left[\varphi_1(x_t, \mathcal{D}_t) \mid x_t, y_{t+1}\right] = E\left[\varphi_1(x_t, \mathcal{D}_t) \mid x_t\right] + O_p\left(\frac{1}{R}\right).$$

### 2.2.2. *Nonparametric Forecasting Models*

Now we consider the following nonparametric forecasting model:

$$y_{t+1} = m(x_t) + e_{t+1}, \tag{2.10}$$

where $x_t$ does not include the constant term, and the functional form of the smoothing function $m(\cdot)$ is unknown and has to been estimated from the data.

The nonparametric one-step-ahead local constant forecast is given by

$$\varphi_2(x_t, \mathcal{D}_t) \equiv \hat{m}(x_t) \equiv \frac{\sum_{s=t-R}^{t-1} y_{s+1} K_h(x_s - x_t)}{\sum_{s=t-R}^{t-1} K_h(x_s - x_t)} \quad \text{for } t = R, \ldots, T-1, \tag{2.11}$$

where $K_h(x_s - x_t) = \prod_{j=1}^{q} h_j^{-1} k(\frac{x_{sj} - x_{tj}}{h_j})$, $k(\cdot)$ is a univariate kernel function, and $h_1, \ldots, h_q$ are bandwidth sequences that converge to zero as $t \to \infty$. See Pagan and Ullah (1999) for details. Following a similar argument as above and using the results in Andrews (1995) and Kristensen (2009) for kernel estimators with heterogenous and weakly dependent data, we can show that as $h_1, \ldots, h_q \to 0$ and $Rh_1 \cdots h_q \to \infty$,

$$\varphi_2(x_t, \mathcal{D}_t) = \frac{\sum_{s=t-R+1}^{t-k-1} y_{s+1} K_h(x_s - x_t)}{\sum_{s=t-R}^{t-k-1} K_h(x_s - x_t)} + O_p\left(\frac{1}{Rh_1 \cdots h_q}\right)$$

$$\equiv \varphi_2(x_t, \tilde{\mathcal{D}}_t) + O_p\left(\frac{1}{Rh_1 \cdots h_q}\right)$$

and

$$E[\varphi_2(x_t, \tilde{\mathcal{D}}_t) \mid x_t, y_{t+1}] = E[\varphi_2(x_t, \tilde{\mathcal{D}}_t) \mid x_t]$$

$$= E[\varphi_2(x_t, \mathcal{D}_t) \mid x_t] + O_p\left(\frac{1}{Rh_1 \cdots h_q}\right)$$

by Assumption I(i).

Similarly, the nonparametric one-step-ahead local linear forecast is given by

$$\varphi_3(x_t, \mathscr{D}_t) \equiv \tilde{m}(x_t) \equiv e_1'(\mathbf{X}_t'\mathbf{K}_t\mathbf{X}_t)^{-1}\mathbf{X}_t'\mathbf{K}_t\mathbf{Y}_{t+1}, \tag{2.12}$$

where $e_1 = (1, 0, \ldots, 0)'$ is a $(q+1) \times 1$ vector, $\mathbf{X}_t = (X_{t-R,x_t}, \ldots, X_{t-1,x_t})$, $X_{s,x_t} = (1, (x_s - x_t)')'$, $\mathbf{K}_t = \operatorname{diag}(K_h(x_{t-R} - x_t), \ldots, K_h(x_{t-1} - x_t))$, and $\mathbf{Y}_t = (y_{t-R}, \ldots, y_{t-1})'$. We can show that under Assumption I

$$E[\varphi_3(x_t, \mathscr{D}_t) \mid x_t, y_{t+1}] = E[\varphi_3(x_t, \mathscr{D}_t) \mid x_t] + O_p\left(\frac{1}{Rh_1 \cdots h_q}\right).$$

### 2.3. Bagging and Its Revised Version

In practice, the probability distribution $\mathbf{P}_{R,t}$ of $\mathscr{D}_t$ is unknown and we have only a single training set $\mathscr{D}_t$ at time $t$. In this case, we can estimate $\mathbf{P}_{R,t}$ by the empirical distribution $\widehat{\mathbf{P}}_{R,t}$ of $\mathscr{D}_t$ and then draw bootstrap resamples $\{\mathscr{D}_t^{*(b)}\}_{b=1}^B$ with $\mathscr{D}_t^{*(b)} \equiv \{(y_j^{*(b)}, x_{j-1}^{*(b)})\}_{j=t-R+1}^t$, say by the method of block bootstrap from $\widehat{\mathbf{P}}_{R,t}$ to form multiple training sets. Then we obtain the following bagging predictor:

$$\varphi_{it}^* \equiv \varphi_i^*(x_t, \mathscr{D}_t) \equiv \sum_{b=1}^B w_{b,t}\varphi_i(x_t, \mathscr{D}_t^{*(b)}), \quad i = 1, 2, 3,$$

where for each $t$, $\{w_{b,t}\}_{b=1}^B$ are probability weights such that $w_{b,t} \geq 0$ for each $b$ and $\sum_{b=1}^B w_{b,t} = 1$.

We showed in Section 2.1 that for $i = 1, 2$, and 3, $E[\varphi_i(x_t, \mathscr{D}_t) \mid x_t, y_{t+1}]$ has a smaller MSFE than $\varphi_i(x_t, \mathscr{D}_t)$, and in Section 2.2 that the former can be approximated by $E[\varphi_i(x_t, \mathscr{D}_t) \mid x_t]$ under Assumption I and some standard conditions on kernel regressions. To obtain a feasible forecast, we propose to estimate $E[\varphi_i(x_t, \mathscr{D}_t) \mid x_t]$ by regressing $\{\varphi_i(x_\tau, \mathscr{D}_\tau)\}_{\tau=t-\overline{R}+1}^t$ on $x_\tau$ for $t = R + \overline{R}, \ldots, T-1$ and $i = 1, 2, 3$ nonparametrically, so that we could utilize historical information to improve forecasting. Let

$$\widehat{E}[\varphi_i(x_t, \mathscr{D}_t) \mid x_t] = \frac{\sum_{\tau=t-\overline{R}+1}^t K_h(x_\tau - x_t)\varphi_{i\tau}}{\sum_{\tau=t-\overline{R}+1}^t K_h(x_\tau - x_t)}$$
$$\text{for } t = R + \overline{R}, \ldots, T-1 \quad \text{and} \quad i = 1, 2, 3,$$

where $\varphi_{i\tau} \equiv \varphi_i(x_\tau, \mathscr{D}_\tau)$. Under the assumption that $R \to \infty$, $\overline{R} \to \infty$, $Rh_1 \cdots h_q \to \infty$, $\overline{R}h_1 \cdots h_q \to \infty$, and $\sum_{j=1}^q h_j^2 \to 0$, and standard conditions on the process $\{(y_s, x_{s-1})\}$, using the results in Andrews (1995)

and Kristensen (2009) for kernel estimators with heterogenous and weakly dependent data we can show that

$$\begin{aligned}
&\widehat{E}[\varphi_i(x_t, \mathscr{D}_t) \mid x_t = x] \\
&\quad = E[\varphi_i(x_t, \mathscr{D}_t) \mid x_t = x] + O_p\left(\frac{1}{\sqrt{\min(R, \overline{R})h_1 \cdots h_q}} + \sum_{j=1}^{q} h_j^2\right) \\
&\quad = E[\varphi_i(x_t, \mathscr{D}_t) \mid x_t = x] + o_p(1) \qquad \text{for } i = 1, 2,
\end{aligned}$$

for any $x$ on the interior of the support of $x_t$.

Noting that $E[\varphi_i(x_t, \mathscr{D}_t) \mid x_t]$ has smaller MSFE than $\varphi_i(x_t, \mathscr{D}_t)$, in principle one can use $\widehat{E}[\varphi_i(x_t, \mathscr{D}_t) \mid x_t]$ as an unbagged forecast for $y_{t+1}$. But it may be unstable and work poorly due to the small to moderate large value of $R$ or $\overline{R}$ as in typical financial forecast applications, and the model and forecast uncertainty using the single training sample $\mathscr{D}_t$. It is well known that bagging can serve as a device to improve the accuracy of unstable predictors. Therefore, we recommend to robustify $\widehat{E}[\varphi_i(x_t, \mathscr{D}_t) \mid x_t]$ by a bagging predictor:

$$\widehat{E}^* \varphi_{it} \equiv \sum_{b=1}^{B} w_{b,t} \widehat{E}[\varphi_i(x_t, \mathscr{D}_t^{*(b)}) \mid x_t]$$
$$\text{for } t = R + \overline{R}, \ldots, T - 1 \text{ and } i = 1, 2, 3.$$

For simplicity, we can apply the simple weights: $w_{b,t} = \frac{1}{B}$ for all $b$ and $t$. We could view this new forecasting method as an updated version of bagging method. It can also be viewed as a combined forecast which treats each bootstrap training set $\{\mathscr{D}_t^{*(b)}\}$ as equally important. Alternatively, one can compute the weights by using a Bayesian model averaging (BMA) approach; see the Appendix in Lee and Yang (2006) for details.

Thus, the procedure to obtain our combined forecast $\widehat{E}^* \varphi_{it}$, $i = 1, 2, 3$, is summarized as follows:

1. For $t = R + \overline{R}, \ldots, T - 1$, based on the empirical distribution of $\{(y_j, x_{j-1})\}_{j=t-R-\overline{R}+1}^{t}$, we construct the $b$th bootstrap sample $\{(y_j^{*(b)}, x_{j-1}^{*(b)})\}_{j=t-(R+\overline{R})+1}^{t}$ by the method of block bootstrap. Let $\mathscr{D}_t^{*(b)} \equiv \{(y_j^{*(b)}, x_{j-1}^{*(b)}\}_{j=t-R+1}^{t}$.
2. For each $b$, we obtain $\varphi_{it}^{*(b)} \equiv \varphi_i(x_t, \mathscr{D}_t^{*(b)})$, $i = 1, 2, 3$, following (2.8), (2.11), or (2.12). In particular, $\varphi_{1t}^{*(b)}$ is obtained by first calculating $\hat{\beta}_t^{*(b)}$

$$\hat{\beta}_t^{*(b)} \equiv \hat{\beta}(\mathscr{D}_t^{*(b)}) \equiv \left(\sum_{s=t-R}^{t-1} x_s^{*(b)} (x_s^{*(b)})'\right)^{-1} \left(\sum_{s=t-R}^{t-1} x_s^{*(b)} y_{s+1}^{*(b)}\right),$$

and then forming

$$\varphi_{1t}^{*(b)} = \varphi_1(x_t, \mathscr{D}_t^{*(b)}) = (\hat{\beta}_t^{*(b)})' x_t \qquad \text{for } t = R, \ldots, T-1. \tag{2.13}$$

Similarly, form $\varphi_{2t}^{*(b)}$ and $\varphi_{3t}^{*(b)}$ by using $\mathscr{D}_t^{*(b)}$ in place of $\mathscr{D}_t$ for $t = R, \ldots, T-1$. Note that the forecast at time $t$ is formed using the original predictor variable $x_t$ instead of $x_t^{*(b)}$ in each case.

3. For $i = 1, 2, 3$ and $t = R + \overline{R}, \ldots, T-1$, regress $\{\varphi_{i\tau}^{*(b)}\}_{\tau=t-\overline{R}}^{t-1}$ on $x_\tau$ nonparametrically to obtain

$$\widehat{E}\left[\varphi_i(x_t, \mathscr{D}_t^{*(b)}) \mid x_t\right] = \frac{\sum_{\tau=t-\overline{R}+1}^{t} K_h(x_\tau^{*(b)} - x_t)\varphi_{i\tau}^{*(b)}}{\sum_{\tau=t-\overline{R}+1}^{t} K_h(x_\tau^{*(b)} - x_t)}.$$

Note that the forecast at time $t$ is formed using the original predictor variables $x_t$ instead of $x_t^{*(b)}$.

4. Repeat the above procedure $B$ times, and obtain $\widehat{E}^*\varphi_{it} \equiv \sum_{b=1}^{B} w_{b,t}\widehat{E}[\varphi_i(x_t, \mathscr{D}_t^{*(b)}) \mid x_t]$ for $i = 1, 2, 3$ and $t = R + \overline{R}, \ldots, T-1$.

## 3. MONTE CARLO SIMULATIONS

In this section, we conduct a small set of Monte Carlo simulations to evaluate the finite sample performance of our proposed predictors and compare them with that of several existing predictors in the literature.

### 3.1. Data Generating Processes

We consider a small class of data generating processes (DGPs):

DGP 1: $y_{t+1} = 0.95y_t \exp(-y_t^2) + \varepsilon_{t+1}$;
DGP 2: $y_{t+1} = 2\varphi(y_t)y_t + \varepsilon_{t+1}$;
DGP 3: $y_{t+1} = \frac{0.5}{1+\exp(-y_t)} + \varepsilon_{t+1}$;
DGP 4: $y_{t+1} = 0.95x_t \exp(-x_t^2) + \varepsilon_{t+1}$;
DGP 5: $y_{t+1} = 2\varphi(x_t)x_t + \varepsilon_{t+1}$;
DGP 6: $y_{t+1} = \frac{0.5}{1+\exp(-x_t)} + \varepsilon_{t+1}$;

where $t = 0, 1, \ldots, T-1$, $\varphi(\cdot)$ is the standard normal density function, $x_t$ is an AR (1) process in DGPs 4–6: $x_t = \rho x_{t-1} + \epsilon_t$ with the coefficient $\rho$ set to be 0, 0.5, 0.95, respectively,[2] and $\epsilon_t$ is IID standard normal. In all DGPs,

[2]We also consider the cases of negative autocorrelation. We set $\rho = -0.5$ and $-0.95$. The results seem to be similar to the results when $\rho = 0.5$ and 0.95, respectively. To save space for the tables, the results are not reported here.

$\{\varepsilon_t\}$ is specified as the following GARCH $(1,1)$ process:

$$\begin{aligned} \varepsilon_t &= v_t \eta_t, \\ v_t^2 &= 1 + \alpha \varepsilon_{t-1}^2 + \beta v_{t-1}^2, \end{aligned} \tag{3.1}$$

where $\eta_t$ is IID standard normal, and we consider $(\alpha, \beta) = (0,0)$, $(0.3,0)$, $(0.9,0)$, $(0.3,0.4)$, $(0.7,0.2)$. In addition, one can apply the results in Masry and Tjøstheim (1995, 1997) and verify that the nonlinear AR(1) process $\{y_t\}$ is strictly stationary and strong mixing.

Clearly, when $(\alpha, \beta) = (0,0)$ in (3.1), $\{\varepsilon_t\}$ reduces to an IID process. If $\beta = 0$, DGPs 1–3 specify a first order Markov process for $\{y_t\}$, DGPs 4–6 specify a first order Markov process for $\{w_t \equiv (y_t, x_{t-1})'\}$, and Assumption I is satisfied. If $\beta \neq 0$, Assumption I is not satisfied so that we can also investigate how bagging is working in this case.

### 3.2. Forecasting Methods

We first study forecasts based on the linear forecasting model

$$y_{t+1} = \beta_0 + \beta_1 x_t + e_{t+1}, \tag{3.2}$$

where $x_t = y_t$ in DGPs 1–3, and is otherwise as specified in DGPs 4–6. We consider the following three forecasts:

1. The traditional one-step-ahead forecast is given by $\varphi_{1t} \equiv \varphi_1(x_t, \mathcal{D}_t) = \hat{\beta}_{0t} + \hat{\beta}_{1t} x_t$, where $(\hat{\beta}_{0t}, \hat{\beta}_{1t})'$ is the OLS estimate of $(\beta_0, \beta_1)'$ based on $\{(y_s, x_{s-1})\}_{s=t-R+1}^{t}$.
2. We ignore the time series structure of $\{y_t\}$ in DGPs 1–3 and $\{w_t\}$ in DGPs 4–6 by pretending $\{y_t\}$ or $\{w_t\}$ is an independent process, and then we use bagging method on $\varphi_1(x_t, \mathcal{D}_t)$ to obtain $\varphi_{1t}^* \equiv \varphi_1^*(x_t, \mathcal{D}_t)$ for $t = R, \ldots, T-1$.
3. We use the new forecasting method proposed in this paper to obtain $\widehat{E}^* \varphi_{1t} \equiv \sum_{b=1}^{B} w_{b,t} \widehat{E}(\varphi_1(x_t, \mathcal{D}_t^{*(b)}) \mid x_t)$, where

$$\widehat{E}[\varphi_1(x_t, \mathcal{D}_t^{*(b)}) \mid x_t] = \frac{\sum_{\tau=t-\overline{R}+1}^{t} K_h(x_\tau^{*(b)} - x_t) \varphi_{1\tau}^{*(b)}}{\sum_{\tau=t-\overline{R}+1}^{t} K_h(x_\tau^{*(b)} - x_t)}, \quad t = R + \overline{R}, \ldots, T-1.$$

We then study forecasts based on the nonparametric forecasting model

$$y_{t+1} = m(x_t) + e_{t+1}, \tag{3.3}$$

where $x_t = y_t$ in DGPs 1–3, and is otherwise as specified in DGPs 4–6. We consider the following six forecasts:

4. The one-step-ahead local constant predictor is given by $\varphi_{2t} \equiv \varphi_2(x_t, \mathcal{D}_t) = \hat{m}_t(x_t)$, where the local constant estimator $\hat{m}_t(x_t)$ of $m(x_t)$ is obtained by using $\{(y_s, x_{s-1})\}_{s=t-R+1}^{t}$.
5. We ignore the time series structure of $w_t = (y_t, x_{t-1})'$ by pretending $w_t$ is an independent process and then use bagging method on $\varphi_2(x_t, \mathcal{D}_t)$ to obtain $\varphi_{2t}^* \equiv \varphi_2^*(x_t, \mathcal{D}_t)$ for $t = R, \ldots, T-1$.
6. We use the new forecasting method proposed in this paper $\widehat{E}^*\varphi_{2t} \equiv \sum_{b=1}^{B} w_{b,t}\widehat{E}(\varphi_2(x_t, \mathcal{D}_t^{*(b)}) \mid x_t)$, where

$$\widehat{E}[\varphi_2(x_t, \mathcal{D}_t^{*(b)}) \mid x_t] = \frac{\sum_{\tau=t-\overline{R}+1}^{t} K_h(x_\tau^{*(b)} - x_t)\varphi_{2\tau}^{*(b)}}{\sum_{\tau=t-\overline{R}+1}^{t} K_h(x_\tau^{*(b)} - x_t)}, \quad t = R + \overline{R}, \ldots, T-1.$$

7. The one-step-ahead local linear predictor is given by $\varphi_{3t} \equiv \varphi_3(x_t, \mathcal{D}_t) = \tilde{m}_t(x_t)$, where the local linear estimator is obtained by using $\{(y_s, x_{s-1})\}_{s=t-R+1}^{t}$.
8. We ignore the time series structure of $w_t = (y_t, x_{t-1})'$ by pretending $w_t$ is an independent process and then use bagging method on $\varphi_3(x_t, D_t)$ to obtain $\varphi_{3t}^* \equiv \varphi_3^*(x_t, D_t)$ for $t = R, \ldots, T-1$.
9. We use the new forecasting method proposed in this paper $\widehat{E}^*\varphi_{3t} \equiv \sum_{b=1}^{B} w_{b,t}\widehat{E}(\varphi_3(x_t, D_t^{*(b)}) \mid x_t)$, where

$$\widehat{E}[\varphi_3(x_t, \mathcal{D}_t^{*(b)}) \mid x_t] = \frac{\sum_{\tau=t-\overline{R}+1}^{t} K_h(x_\tau^{*(b)} - x_t)\varphi_{3\tau}^{*(b)}}{\sum_{\tau=t-\overline{R}+1}^{t} K_h(x_\tau^{*(b)} - x_t)}, \quad t = R + \overline{R}, \ldots, T-1.$$

We consider four choices of $R$, namely, $R = 20, 50, 100, 200$, for the in-sample estimation. We try to be thrift on $\overline{R}$ and set $\overline{R} = 20$ for all cases. We set the out-of-sample period $P \equiv T - R - \overline{R}$ to be 50 in all cases.[3] $w_{b,t}$ is set to be $\frac{1}{B}$ in all cases for convenience.[4] For the nonparametric estimation, we need to choose both the kernel function $k(\cdot)$ and the bandwidth parameter $h$. We use the standard normal kernel function throughout the simulations and applications, i.e., $k(x) = (2\pi)^{-1/2}\exp(-x^2/2)$. We apply the least squares cross validation method to select the bandwidth and use

[3]We try different values of $\overline{R}$ by setting $\overline{R} = R/2$ when $R = 50, 100$, and 200. We also try different out-of-sample periods by setting $P$ to be 100, 200, and 500. The results are similar and not reported here for brevity.

[4]As suggested by one referee, The choice of equal weights is optimal in the iid case but not necessarily with dependent series like the ones considered in the paper. We also try to compute the weights by using Bayesian model averaging (BMA) technique as introduced in Lee and Yang (2006). The BMA gives a large weight to the $b$th bootstrap predictor at each period $t$ when it has forecasted well over the past $k$ periods and a small weight to the predictor at period $t$ when it forecasted poorly over the past $k$ periods. We have set $k = 1, 5$ and $R$. The results are similar to those based on equal weights and not reported here for brevity.

the same bandwidth for the bootstrap resamples.[5] For the out-of-sample evaluation, we first calculate the MSFE for each replication and each forecasting method, e.g.,

$$MSFE(\varphi_i) = \frac{1}{P}\sum_{t=R+\bar{R}}^{T-1}(\varphi_{it} - y_{t+1})^2, \quad i = 1, 2, 3,$$

and then obtain the final MSFE by averaging $MSFE(\varphi_i)$ across replications. Further, we compute the percentage reduction in the MSFE of other predictors relative to that of $\varphi_1$.

In each scenario, the number of replications in the Monte Carlo study is 200, and the number of bootstrap resamples is $B = 100$. The selection of optimal block length is based on Politis and White (2004).

### 3.3. Simulation Results

Tables 1–6 report the results of the experiments for DGPs 1–6, respectively. We summarize some important findings from these tables.

First, for all DGPs under our investigation, forecast methods 2–6 and 8–9 in Subsection 3.2 outperform the benchmark one-step-ahead linear forecast method in terms of out-of-sample MSFE in most cases. Since we do not restrict the conditioning variable to be compactedly supported, the one-step-ahead local linear forecast method may yield very odd forecasts and thus does not work as well as the benchmark method at all.

Second, both the bagging methods and our proposed methods outperform the traditional one-step-ahead linear, local constant, or local linear forecast methods significantly. In particular, when the in-sample training period is small (say, $R = 20$), the percentage gains of both methods are above 5% in all cases except for DGPs 2 and 5 when the errors are IID distributed. For DGPs 1–3 with $(\alpha, \beta) = (0.9, 0)$ in (3.1), the percentage gains of both methods are above 45%, and for DGPs 1–3

[5]As one referee remarks, the choice of $h$ affects the results and may not be optimal as chosen. We are dealing with dependent series and cross validation methods require blocking here too. We follow Hart and Vieu (1990) and set different leave-out sequences to take care of the dependence structure of the time series. We set $l_n = 0, 1, 2, 3, 4, 5$ as in Hart and Vieu (1990), where $l_n = 0$ corresponds to the ordinary leave-one-out cross validation, $l_n > 0$ corresponds to leave $2l_n + 1$ observations out, and the leave-out sequence is $\{X_j\}$ with $|j - t| \le l_n$. The results for $l_n > 0$ are similar to those based on the usual leave-one-out least squares cross validation and thus not reported here.

We also try to choose the bandwidth by the "rule of thumb": $h_l = c_0 s_l n^{-1/(4+q)}$, where $s_l$ stands for the sample standard deviations of $X_{it,l}$, the $l$th regressor in $X_{it}$. We set $c_0 = 0.5, 1$, and 2 to examine the sensitivity of our test to the choice of bandwidth. It turns out that the results of our proposed methods and bagging methods are robust to different bandwidth choice. On the other hand, the usual one-step-ahead local constant and local linear predictors are sensitive to the bandwidth choice.

**TABLE 1** Percentage gain of MSFE compared to one-step linear forecast method: DGP 1

| | | $(\alpha, \beta)$ in (3.1) | | | | |
|---|---|---|---|---|---|---|
| $R$ | Forecasts | (0,0) | (0.3,0) | (0.9,0) | (0.3,0.4) | (0.7,0.2) |
| 20 | $\varphi_1^*$ | 5.0636 | 7.7969 | 46.8776 | 9.0388 | 29.4740 |
| | $\widehat{E}^*\varphi_1$ | 5.0984 | 8.6463 | 48.1612 | 10.6490 | 31.5870 |
| | $\varphi_2$ | 3.2394 | 5.4276 | 41.7799 | 6.5571 | 15.6565 |
| | $\varphi_2^*$ | 5.4952 | 8.4527 | 47.7829 | 9.6084 | 30.4567 |
| | $\widehat{E}^*\varphi_2$ | 5.0038 | 8.6240 | 48.2825 | 10.7088 | 31.6435 |
| | $\varphi_3$ | −9.0487 | −46.5875 | −95.5993 | −28.9733 | −23.5670 |
| | $\varphi_3^*$ | 5.2620 | 7.1193 | 46.0044 | 7.5773 | −6.9829 |
| | $\widehat{E}^*\varphi_3$ | 5.0251 | 8.4680 | 47.9631 | 10.4431 | 8.0923 |
| 50 | $\varphi_1^*$ | 0.9994 | 2.4195 | 14.7864 | 2.7445 | 9.1133 |
| | $\widehat{E}^*\varphi_1$ | 0.2733 | 2.6082 | 15.9809 | 3.8076 | 11.0263 |
| | $\varphi_2$ | 0.2208 | 1.1133 | 4.3821 | 2.0653 | 3.7369 |
| | $\varphi_2^*$ | 1.8089 | 2.9164 | 14.9160 | 3.0902 | 10.0952 |
| | $\widehat{E}^*\varphi_2$ | 0.3273 | 2.6853 | 16.0379 | 3.8640 | 10.9503 |
| | $\varphi_3$ | −5.1195 | −18.8208 | −67.5567 | −14.6733 | −51.6003 |
| | $\varphi_3^*$ | 1.4807 | 2.4179 | 13.4518 | 2.2813 | 8.7441 |
| | $\widehat{E}^*\varphi_3$ | 0.4459 | 2.7631 | 15.8721 | 3.7451 | 10.7342 |
| 100 | $\varphi_1^*$ | −0.1656 | 0.4277 | 17.6806 | 1.0200 | 8.3792 |
| | $\widehat{E}^*\varphi_1$ | −2.0588 | −0.2985 | 13.2545 | 1.1636 | 6.0999 |
| | $\varphi_2$ | 0.1085 | −1.2894 | 2.3753 | 0.9066 | 1.6915 |
| | $\varphi_2^*$ | 0.9011 | 1.1263 | 9.7124 | 0.8754 | 6.2815 |
| | $\widehat{E}^*\varphi_2$ | −2.0898 | −0.3038 | 12.4770 | 1.1801 | 5.6702 |
| | $\varphi_3$ | −2.3898 | −12.3332 | −37.2586 | −8.8236 | −10.0113 |
| | $\varphi_3^*$ | 1.0515 | 0.3365 | 7.4648 | 0.4096 | 5.1681 |
| | $\widehat{E}^*\varphi_3$ | −2.0162 | −0.3560 | 12.2303 | 1.0191 | 5.4734 |
| 200 | $\varphi_1^*$ | −0.1490 | 0.3633 | 3.4777 | 0.7497 | 1.6493 |
| | $\widehat{E}^*\varphi_1$ | −2.4147 | −0.4804 | 4.2079 | 0.8863 | 2.2850 |
| | $\varphi_2$ | 1.5921 | −0.5383 | 1.4677 | 0.0608 | 0.9852 |
| | $\varphi_2^*$ | 1.9854 | 1.4136 | 4.2796 | 0.9027 | 1.8040 |
| | $\widehat{E}^*\varphi_2$ | −2.2262 | −0.3134 | 4.4460 | 0.9971 | 2.5250 |
| | $\varphi_3$ | 0.2857 | −10.3706 | −11.7520 | −8.2181 | −21.2924 |
| | $\varphi_3^*$ | 2.0352 | 0.4070 | 1.7330 | 0.2846 | −1.2830 |
| | $\widehat{E}^*\varphi_3$ | −2.1929 | −0.5060 | 4.1248 | 0.8522 | 2.2711 |

*Note:* In Tables 1–6, the results are based on the out-of-sample forecast MSE averaged over 200 repetitions. The in-sample period is $R = 20, 50, 100$, and 200, respectively; the out-of-sample period is $P = 50$. For the bagging methods, the number of bootstrap resamples is $B = 100$.

with $(\alpha, \beta) = (0.7, 0.2)$ in (3.1), the percentage gains of both methods are above 29%.

Third, the traditional bagging methods are not as good as our proposed methods for all the six DGPs when the errors exhibit strong conditional heteroskedasticity (especially when $(\alpha, \beta) = (0.9, 0)$, $(0.7, 0.2)$ in (3.1)). For DGPs 4–6, when the independent variable is autocorrelated (e.g., $\rho = 0.5$ and 0.95), our proposed methods tend to outperform the traditional bagging methods in most cases. For example, for DGP 4, when the independent variable is not autocorrelated (i.e., $\rho = 0$), the traditional bagging methods perform slightly better when the errors are

**TABLE 2** Percentage gain of MSFE compared to one-step linear forecast method: DGP 2

| | | $(\alpha, \beta)$ in (3.1) | | | | |
|---|---|---|---|---|---|---|
| $R$ | Forecasts | (0,0) | (0.3,0) | (0.9,0) | (0.3,0.4) | (0.7,0.2) |
| 20 | $\varphi_1^*$ | 3.6020 | 6.9709 | 47.0760 | 8.9088 | 29.4783 |
| | $\widehat{E}^*\varphi_1$ | 1.3974 | 6.4438 | 47.9435 | 10.1641 | 31.4088 |
| | $\varphi_2$ | 0.5824 | 3.1874 | 41.7083 | 4.7606 | 15.4826 |
| | $\varphi_2^*$ | 3.1500 | 7.1847 | 47.5911 | 9.7129 | 30.3673 |
| | $\widehat{E}^*\varphi_2$ | 1.1649 | 6.4202 | 48.0851 | 10.2023 | 31.4541 |
| | $\varphi_3$ | −12.1327 | −14.9638 | 20.3422 | −21.8105 | −39.0167 |
| | $\varphi_3^*$ | 3.5330 | 6.7643 | 45.3948 | 8.2039 | 27.6878 |
| | $\widehat{E}^*\varphi_3$ | 1.5131 | 6.5206 | 47.7107 | 9.9977 | 31.2595 |
| 50 | $\varphi_1^*$ | 0.1081 | 1.6648 | 14.7116 | 2.7400 | 9.0764 |
| | $\widehat{E}^*\varphi_1$ | −3.3138 | 0.5524 | 15.6238 | 3.4299 | 10.7244 |
| | $\varphi_2$ | −0.1844 | 0.0557 | 6.8603 | 2.1074 | 3.0693 |
| | $\varphi_2^*$ | 0.6787 | 1.8980 | 14.7759 | 2.9533 | 10.1112 |
| | $\widehat{E}^*\varphi_2$ | −3.2416 | 0.5931 | 15.8271 | 3.4426 | 10.7427 |
| | $\varphi_3$ | −20.0629 | −18.8279 | −20.3801 | −11.8201 | −45.6133 |
| | $\varphi_3^*$ | 0.5757 | 1.9487 | 14.6281 | 1.7704 | 8.3726 |
| | $\widehat{E}^*\varphi_3$ | −2.9570 | 0.7280 | 15.7878 | 3.4183 | 10.2564 |
| 100 | $\varphi_1^*$ | −0.8634 | 0.0148 | 17.2495 | 0.7395 | 8.1590 |
| | $\widehat{E}^*\varphi_1$ | −6.1687 | −2.8421 | 12.9646 | 0.4472 | 6.0955 |
| | $\varphi_2$ | 0.8386 | −0.9932 | 5.6705 | 0.0574 | 1.1922 |
| | $\varphi_2^*$ | 0.4596 | 1.0014 | 10.0248 | 1.1427 | 6.1554 |
| | $\widehat{E}^*\varphi_2$ | −6.1443 | −2.8373 | 12.2321 | 0.4638 | 5.6478 |
| | $\varphi_3$ | −3.0155 | −5.3967 | −18.8311 | −7.4191 | −9.4973 |
| | $\varphi_3^*$ | 0.6852 | 0.9734 | 11.0005 | 0.5912 | 4.9188 |
| | $\widehat{E}^*\varphi_3$ | −5.9863 | −2.8351 | 12.5015 | 0.3885 | 5.3408 |
| 200 | $\varphi_1^*$ | −0.7558 | −0.0916 | 3.4361 | 0.5814 | 1.7834 |
| | $\widehat{E}^*\varphi_1$ | −6.4397 | −2.9894 | 3.8493 | 0.2561 | 2.1137 |
| | $\varphi_2$ | 1.4469 | −0.6051 | 2.1881 | 0.1102 | 0.5990 |
| | $\varphi_2^*$ | 1.0940 | 0.9938 | 3.7153 | 1.0978 | 2.1834 |
| | $\widehat{E}^*\varphi_2$ | −6.2262 | −2.8597 | 4.0627 | 0.4108 | 2.2913 |
| | $\varphi_3$ | 0.4264 | −11.6600 | −21.3443 | −7.9908 | −27.8649 |
| | $\varphi_3^*$ | 1.1489 | 0.7962 | 0.9443 | 0.6603 | −1.5120 |
| | $\widehat{E}^*\varphi_3$ | −6.1946 | −2.9194 | 3.7087 | 0.3466 | 2.0430 |

IID ($(\alpha, \beta) = (0, 0)$ in (3.1)) or weakly conditional heteroskedastic ($(\alpha, \beta) = (0.3, 0)$ in (3.1)) for most choices of in-sample periods; on the other hand, our proposed methods perform better when the errors exhibit stronger conditional heteroskedasticity, e.g., $(\alpha, \beta) = (0.9, 0)$ in (3.1) for in-sample periods $R = 20, 50$, and 100, $(\alpha, \beta) = (0.3, 0.4)$ in (3.1) for $R = 20$ and 50, $(\alpha, \beta) = (0.7, 0.2)$ in (3.1) for all in-sample periods. When the independent variable is autocorrelated, say, $\rho = 0.5$, our proposed methods perform better for the cases where $(\alpha, \beta) = (0.3, 0)$ in (3.1) for $R = 20$, $(\alpha, \beta) = (0.9, 0)$, $(0.7, 0.2)$ in (3.1) for all in-sample periods, $(\alpha, \beta) = (0.3, 0.4)$ in (3.1) for $R = 20, 50$ and 100. When the independent variable is strongly autocorrelated ($\rho = 0.95$), our proposed methods beat the traditional bagging methods in all cases except when the errors are IID with large

**TABLE 3** Percentage gain of MSFE compared to one-step linear forecast method: DGP 3

| | | $(\alpha, \beta)$ in (3.1) | | | | |
|---|---|---|---|---|---|---|
| $R$ | Forecasts | (0,0) | (0.3,0) | (0.9,0) | (0.3,0.4) | (0.7,0.2) |
| 20 | $\varphi_1^*$ | 5.2233 | 7.8038 | 46.9761 | 8.9925 | 29.4590 |
| | $\widehat{E}^*\varphi_1$ | 6.3668 | 9.0364 | 48.4371 | 10.3755 | 31.5110 |
| | $\varphi_2$ | 3.2917 | 3.6964 | 34.9334 | 5.5565 | 8.9866 |
| | $\varphi_2^*$ | 5.7143 | 8.1639 | 47.8880 | 9.5750 | 30.0790 |
| | $\widehat{E}^*\varphi_2$ | 6.4099 | 9.0842 | 48.5770 | 10.4363 | 31.6936 |
| | $\varphi_3$ | −7.7736 | −11.5636 | 21.9036 | −23.9950 | −45.1837 |
| | $\varphi_3^*$ | 4.5557 | 6.3786 | 46.6719 | 6.1486 | 27.5119 |
| | $\widehat{E}^*\varphi_3$ | 6.2842 | 8.8654 | 48.4397 | 10.1058 | 31.5648 |
| 50 | $\varphi_1^*$ | 1.5386 | 2.5787 | 14.8946 | 2.7349 | 9.5699 |
| | $\widehat{E}^*\varphi_1$ | 1.8592 | 3.2077 | 15.9254 | 3.7367 | 10.9140 |
| | $\varphi_2$ | 0.8987 | 0.9688 | 4.8649 | 1.0031 | 3.1202 |
| | $\varphi_2^*$ | 1.5713 | 2.6934 | 14.8634 | 2.7878 | 9.7862 |
| | $\widehat{E}^*\varphi_2$ | 1.8439 | 3.2299 | 16.1256 | 3.7104 | 10.9297 |
| | $\varphi_3$ | −6.3813 | −39.0208 | −23.9676 | −10.2531 | −46.6232 |
| | $\varphi_3^*$ | 0.6394 | 1.4103 | 13.3343 | 1.7787 | 6.3712 |
| | $\widehat{E}^*\varphi_3$ | 1.8107 | 3.1846 | 15.9650 | 3.5480 | 10.4347 |
| 100 | $\varphi_1^*$ | 0.3312 | 0.6462 | 17.5590 | 0.7554 | 8.2044 |
| | $\widehat{E}^*\varphi_1$ | −0.0038 | 0.4033 | 13.5445 | 0.8048 | 6.1454 |
| | $\varphi_2$ | −1.3182 | 0.0882 | 2.9650 | 0.5407 | 1.0213 |
| | $\varphi_2^*$ | −0.0693 | 0.3093 | 9.7459 | 0.7241 | 6.1008 |
| | $\widehat{E}^*\varphi_2$ | −0.0750 | 0.3488 | 12.9012 | 0.7728 | 5.8667 |
| | $\varphi_3$ | −3.2023 | −17.1408 | −27.5696 | −7.7353 | −18.3625 |
| | $\varphi_3^*$ | −0.0815 | −0.0561 | 9.1729 | −0.0552 | 4.6040 |
| | $\widehat{E}^*\varphi_3$ | −0.1156 | 0.1880 | 12.8920 | 0.6416 | 5.4761 |
| 200 | $\varphi_1^*$ | 0.1482 | 0.4777 | 3.1238 | 0.6178 | 1.7745 |
| | $\widehat{E}^*\varphi_1$ | −0.4739 | 0.1037 | 3.9241 | 0.5213 | 2.0009 |
| | $\varphi_2$ | −0.5396 | 0.0367 | 2.7106 | 0.0258 | 0.1806 |
| | $\varphi_2^*$ | −0.0825 | 0.3609 | 3.8725 | 0.6692 | 2.3080 |
| | $\widehat{E}^*\varphi_2$ | −0.5048 | 0.1359 | 4.0887 | 0.5257 | 2.1299 |
| | $\varphi_3$ | −1.3725 | −7.4971 | −10.5715 | −8.6772 | −12.9924 |
| | $\varphi_3^*$ | 0.0455 | −0.3299 | −0.3065 | −3.7956 | −0.1829 |
| | $\widehat{E}^*\varphi_3$ | −0.5089 | 0.0525 | 3.7141 | 0.0472 | 1.9037 |

in-sample period ($R = 200$). This is as what we expected: the proposed methods outperform the traditional bagging methods in the case of strong correlation.

Fourth, in general the percentage gains of our proposed methods decrease when $R$ increases for all DGPs. When $R$ is large, the percentage gains of our proposed methods tend to be negative when the errors are IID ($(\alpha, \beta) = (0, 0)$ in (3.1)) or weakly conditional heteroskedastic ($(\alpha, \beta) = (0.3, 0)$ in (3.1)). But when the errors exhibit strong conditional heteroskedasticity, especially when $(\alpha, \beta) = (0.9, 0), (0.7, 0.2)$ in (3.1), the percentage gains of our proposed methods are positive for all in-sample periods. For DGPs 4–6, the higher degree of dependence in the predicting variable, the more positive gains of our proposed methods. Take DGP 4 as

an example. When $\rho = 0$, the percentage gains of our proposed methods are negative when $(\alpha, \beta) = (0, 0)$ and $(0.3, 0)$ in (3.1) for $R = 100, 200$, and $(\alpha, \beta) = (0.3, 0.4)$ in (3.1) for $R = 200$; when $\rho = 0.5$, the percentage gains of our proposed methods are negative when $(\alpha, \beta) = (0, 0)$ and $(0.3, 0)$ in (3.1) for $R = 100, 200$, and the percentage gains are positive for all other cases. When $\rho = 0.95$, the percentage gains are positive for all cases.

Fifth, the percentage gain in terms of MSFE reduction for forecasting methods 2–6 and 8–9 is larger when the in-sample training period $R$ is small (e.g., $R = 20, 50$) than the case when $R$ is large (e.g., $R = 100, 200$). But for DGPs 1–3 with $(\alpha, \beta) = (0.9, 0)$ in (3.1), the percentage gains for both bagging and our methods are above 10% even when $R = 100$; and for DGPs 1–3 with $(\alpha, \beta) = (0.7, 0.2)$ in (3.1), the percentage gains of our proposed methods are about 5% higher than those of the traditional one-step-ahead linear predictor when $R = 100$.

Sixth, the most prominent predictors to forecast the excess returns proposed in the literature include the dividend price ratio and dividend yield, the earnings price ratio and dividend-earnings (payout) ratio, various interest rates and spreads, the inflation rates, the book-to-market ratio, volatility, the investment-capital ratio, the consumption, wealth, and income ratio, and aggregate net or equity issuing activity. Many of the predictor variables are highly persistent. Thus we pay special attention to the cases of $\rho = 0.95$ for DGPs 4–6. It is apparent from Tables 4–6 that the higher degree of dependence in the predicting variable, the higher percentage gains of our proposed method. For example, for DGP 4 with IID errors ($(\alpha, \beta) = (0, 0)$ in (3.1)), when $\rho = 0$, the percentage gains of our proposed methods using linear, local constant and local linear approaches are 5.361, 5.269, 5.289, respectively. They increase to 6.861, 6.792, and 6.929, respectively, when $\rho = 0.5$, and rise to 11.006, 11.658, and 7.446, respectively, when $\rho = 0.95$.

## 4. EMPIRICAL APPLICATION: PREDICTING EXCESS STOCK RETURNS

Goyal and Welch (2008) claimed that the historical average excess stock return forecasts beat other predictor variables in the literature. Campbell and Thompson (2008) showed that many predictive regressions perform better than the historical average return if some restrictions are imposed on the signs of regression coefficients and return forecasts. Chen and Hong (2009) argued that the restriction proposed by Campbell and Thompson (2008) is a form of nonlinearity and found nonparametric predictive methods outperform the historical average in quarterly or annual frequency. Lee et al. (2010) considered nonparametric and semiparametric regression models with the use of bagging to impose

**TABLE 4** Percentage gain of MSFE compared to one-step linear forecast method: DGP 4 with different AR (1) coefficients

| | | $\rho = 0, (\alpha, \beta) =$ | | | | | $\rho = 0.5, (\alpha, \beta) =$ | | | | | $\rho = 0.95, (\alpha, \beta) =$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $R$ | Forecast | (0,0) | (.3,0) | (.9,0) | (.3,.4) | (.7,.2) | (0,0) | (.3,0) | (.9,0) | (.3,.4) | (.7,.2) | (0,0) | (.3,0) | (.9,0) | (.3,.4) | (.7,.2) |
| 20 | $\varphi_1^*$ | 5.727 | 6.146 | 6.118 | 6.512 | 6.830 | 6.955 | 6.923 | 6.189 | 7.180 | 6.951 | 9.564 | 9.370 | 7.929 | 9.571 | 7.954 |
| | $\widehat{E}^*\varphi_1$ | 5.361 | 5.994 | 7.647 | 7.174 | 8.021 | 6.861 | 7.249 | 7.974 | 8.063 | 8.443 | 11.006 | 10.860 | 10.550 | 10.983 | 10.817 |
| | $\varphi_2$ | 2.626 | 2.208 | 3.392 | 3.497 | 3.225 | 2.785 | 2.812 | 3.421 | 2.437 | 1.725 | 2.474 | 3.274 | 4.671 | 3.379 | 3.947 |
| | $\varphi_2^*$ | 5.921 | 6.209 | 6.451 | 6.640 | 6.799 | 7.308 | 7.551 | 7.027 | 7.779 | 7.363 | 11.048 | 10.654 | 9.662 | 10.760 | 9.858 |
| | $\widehat{E}^*\varphi_2$ | 5.269 | 5.962 | 7.730 | 7.100 | 8.019 | 6.792 | 7.268 | 8.042 | 8.101 | 8.494 | 11.658 | 11.653 | 10.963 | 11.587 | 11.145 |
| | $\varphi_3$ | −8.310 | −9.524 | −11.353 | −9.120 | −9.889 | −17.090 | −13.713 | −23.571 | −11.267 | −24.482 | −23.712 | −26.596 | −14.269 | −15.648 | −15.854 |
| | $\varphi_3^*$ | 5.271 | 5.027 | 4.557 | 5.560 | 5.611 | 6.826 | 6.886 | 4.016 | 6.569 | 6.277 | −2.581 | 4.119 | 5.492 | −0.613 | 5.907 |
| | $\widehat{E}^*\varphi_3$ | 5.289 | 6.017 | 7.595 | 7.036 | 7.955 | 6.929 | 7.249 | 7.616 | 7.940 | 8.457 | 7.446 | 10.939 | 10.093 | 10.175 | 10.484 |
| 50 | $\varphi_1^*$ | 1.209 | 1.484 | 1.609 | 1.803 | 1.674 | 1.794 | 2.075 | 1.797 | 2.048 | 1.911 | 3.062 | 3.107 | 2.776 | 3.084 | 3.012 |
| | $\widehat{E}^*\varphi_1$ | 0.408 | 1.069 | 2.076 | 1.955 | 2.091 | 1.243 | 1.750 | 2.489 | 2.433 | 2.589 | 3.603 | 3.645 | 3.151 | 3.824 | 3.438 |
| | $\varphi_2$ | 0.648 | 0.076 | 0.196 | 0.494 | 0.438 | 0.420 | 0.079 | 1.286 | 1.396 | 1.233 | 1.932 | 1.834 | 2.983 | 1.797 | 2.185 |
| | $\varphi_2^*$ | 1.844 | 1.914 | 1.737 | 1.971 | 1.651 | 2.336 | 2.453 | 1.840 | 2.171 | 2.065 | 3.404 | 3.330 | 2.694 | 3.315 | 3.126 |
| | $\widehat{E}^*\varphi_2$ | 0.410 | 1.059 | 2.063 | 1.998 | 2.107 | 1.252 | 1.743 | 2.492 | 2.412 | 2.609 | 3.640 | 3.691 | 3.465 | 3.822 | 3.848 |
| | $\varphi_3$ | −3.879 | −3.943 | −19.034 | −4.487 | −17.850 | −4.524 | −5.317 | −6.645 | −5.645 | −4.652 | −14.310 | −12.428 | −15.382 | −8.740 | −12.800 |
| | $\varphi_3^*$ | 1.456 | 1.659 | 0.585 | 1.512 | 0.859 | 2.112 | 1.975 | 0.357 | 1.378 | 1.191 | 2.255 | 2.072 | 1.969 | 2.339 | 2.099 |
| | $\widehat{E}^*\varphi_3$ | 0.442 | 1.045 | 2.031 | 1.981 | 2.071 | 1.159 | 1.644 | 2.423 | 2.292 | 2.540 | 3.366 | 3.434 | 3.222 | 3.687 | 3.578 |
| 100 | $\varphi_1^*$ | −0.182 | 0.174 | 1.225 | 0.754 | 1.024 | 0.132 | 0.449 | 1.287 | 0.764 | 0.961 | 1.358 | 1.322 | 2.469 | 1.373 | 1.995 |
| | $\widehat{E}^*\varphi_1$ | −1.527 | −0.609 | 1.755 | 0.696 | 1.559 | −0.595 | 0.027 | 1.869 | 0.898 | 1.614 | 1.407 | 1.375 | 2.930 | 1.438 | 2.384 |
| | $\varphi_2$ | 1.708 | 0.502 | 0.987 | 0.525 | 0.181 | 1.313 | 0.402 | 0.579 | 0.143 | 0.521 | 1.372 | 0.507 | 0.779 | 1.100 | 0.231 |
| | $\varphi_2^*$ | 0.976 | 0.929 | 1.008 | 0.932 | 0.970 | 1.049 | 1.042 | 0.789 | 0.937 | 0.665 | 1.728 | 1.519 | 2.694 | 1.385 | 2.113 |
| | $\widehat{E}^*\varphi_2$ | −1.396 | −0.525 | 1.679 | 0.735 | 1.477 | −0.623 | −0.034 | 1.837 | 0.819 | 1.598 | 1.519 | 1.470 | 3.139 | 1.571 | 2.612 |
| | $\varphi_3$ | −6.190 | −6.342 | −3.962 | −8.138 | −4.778 | −3.222 | −3.462 | −2.500 | −3.203 | −2.876 | −12.895 | −12.198 | −4.242 | −8.544 | −4.559 |
| | $\varphi_3^*$ | 1.089 | 0.915 | 0.797 | 0.865 | 0.447 | 0.907 | 0.974 | −0.095 | 0.666 | −0.259 | 1.210 | −0.800 | 1.954 | 0.783 | 1.128 |
| | $\widehat{E}^*\varphi_3$ | −1.370 | −0.528 | 1.645 | 0.778 | 1.445 | −0.689 | −0.102 | 1.923 | 0.740 | 1.619 | 1.371 | 1.232 | 2.860 | 1.353 | 2.347 |
| 200 | $\varphi_1^*$ | −0.964 | −0.378 | 0.343 | 0.088 | 0.342 | −0.640 | −0.221 | 0.307 | 0.220 | 0.425 | 0.438 | 0.380 | 0.328 | 0.318 | 0.365 |
| | $\widehat{E}^*\varphi_1$ | −2.551 | −1.588 | 0.211 | −0.305 | 0.327 | −1.610 | −0.914 | 0.415 | 0.003 | 0.514 | 0.412 | 0.401 | 0.462 | 0.418 | 0.480 |
| | $\varphi_2$ | 2.262 | 1.158 | 0.400 | 0.267 | 0.618 | 3.596 | 2.461 | 0.350 | 0.717 | 0.269 | 0.245 | 0.407 | 0.679 | 0.684 | 0.700 |
| | $\varphi_2^*$ | 0.536 | 0.589 | 0.314 | 0.389 | 0.175 | 1.327 | 1.378 | 0.769 | 0.875 | 0.805 | 0.883 | 0.667 | 0.372 | 0.309 | 0.379 |
| | $\widehat{E}^*\varphi_2$ | −2.506 | −1.580 | 0.179 | −0.359 | 0.302 | −1.478 | −0.806 | 0.499 | 0.071 | 0.594 | 0.540 | 0.461 | 0.573 | 0.378 | 0.554 |
| | $\varphi_3$ | 1.330 | 0.101 | −0.938 | −0.835 | −1.188 | 2.325 | 1.472 | −0.734 | 0.026 | −0.740 | −4.702 | −3.462 | −2.327 | −2.693 | −2.478 |
| | $\varphi_3^*$ | 0.584 | 0.697 | 0.360 | 0.484 | 0.195 | 1.280 | 1.391 | 0.521 | 0.845 | 0.567 | 0.323 | 0.128 | −0.708 | −0.223 | −0.628 |
| | $\widehat{E}^*\varphi_3$ | −2.480 | −1.535 | 0.284 | −0.293 | 0.362 | −1.468 | −0.773 | 0.516 | 0.086 | 0.574 | 0.433 | 0.368 | 0.342 | 0.292 | 0.326 |

**TABLE 5** Percentage gain of MSFE compared to one-step linear forecast method: DGP 5 with different AR (1) coefficients

| | | $\rho = 0, (\alpha, \beta) =$ | | | | | $\rho = 0.5, (\alpha, \beta) =$ | | | | | $\rho = 0.95, (\alpha, \beta) =$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $R$ | Forecast | (0,0) | (.3,0) | (.9,0) | (.3,.4) | (.7,.2) | (0,0) | (.3,0) | (.9,0) | (.3,.4) | (.7,.2) | (0,0) | (.3,0) | (.9,0) | (.3,.4) | (.7,.2) |
| 20 | $\varphi_1^*$ | 4.238 | 5.047 | 6.078 | 5.999 | 6.622 | 5.437 | 6.041 | 6.317 | 6.902 | 6.888 | 9.475 | 9.332 | 7.956 | 9.338 | 8.184 |
| | $\widehat{E}^*\varphi_1$ | 1.706 | 3.449 | 7.161 | 6.010 | 7.541 | 3.644 | 5.007 | 7.611 | 7.123 | 8.100 | 10.304 | 10.380 | 10.470 | 10.778 | 10.748 |
| | $\varphi_2$ | 1.099 | 2.446 | 3.839 | 2.157 | 3.692 | 1.886 | 2.234 | 2.193 | 2.049 | 3.498 | 3.109 | 3.539 | 4.441 | 3.343 | 3.682 |
| | $\varphi_2^*$ | 3.732 | 4.779 | 6.201 | 5.744 | 6.677 | 5.104 | 3.794 | 6.484 | 6.979 | 7.003 | 11.387 | 10.841 | 9.736 | 10.520 | 9.801 |
| | $\widehat{E}^*\varphi_2$ | 1.560 | 3.354 | 7.286 | 5.947 | 7.521 | 3.499 | 4.933 | 7.623 | 7.114 | 8.069 | 11.138 | 11.184 | 10.860 | 11.476 | 11.021 |
| | $\varphi_3$ | −8.713 | −10.084 | −11.611 | −9.475 | −8.679 | −8.960 | −10.996 | −15.049 | −11.442 | −18.555 | −15.140 | −20.674 | −14.032 | −18.102 | −17.100 |
| | $\varphi_3^*$ | 3.630 | 3.927 | 4.868 | 5.056 | 5.710 | −3.149 | 5.735 | 4.616 | 5.987 | −6.160 | −3.759 | −3.410 | 6.658 | 2.777 | 6.133 |
| | $\widehat{E}^*\varphi_3$ | 1.513 | 3.359 | 7.238 | 5.935 | 6.325 | 2.842 | 4.890 | 7.502 | 6.875 | 7.527 | 10.181 | 8.223 | 9.991 | 10.443 | 10.529 |
| 50 | $\varphi_1^*$ | −0.662 | 0.431 | 1.340 | 1.342 | 1.458 | 0.232 | 0.983 | 1.610 | 1.695 | 1.632 | 2.911 | 3.065 | 2.900 | 3.119 | 2.938 |
| | $\widehat{E}^*\varphi_1$ | −3.301 | −1.525 | 1.453 | 0.867 | 1.536 | −1.996 | −0.508 | 1.938 | 1.516 | 2.140 | 3.146 | 3.343 | 3.190 | 3.656 | 3.408 |
| | $\varphi_2$ | 0.664 | 0.031 | 0.495 | 0.863 | 0.631 | 1.170 | 0.316 | 1.305 | 0.483 | 1.332 | 0.269 | 0.399 | 2.601 | 0.856 | 2.051 |
| | $\varphi_2^*$ | −0.189 | 0.716 | 1.287 | 1.360 | 1.384 | 0.798 | 1.425 | 1.703 | 1.621 | 1.754 | 3.685 | 3.756 | 2.763 | 3.346 | 3.020 |
| | $\widehat{E}^*\varphi_2$ | −3.288 | −1.543 | 1.412 | 0.866 | 1.566 | −2.033 | −0.541 | 1.876 | 1.460 | 2.120 | 3.247 | 3.435 | 3.390 | 3.652 | 3.707 |
| | $\varphi_3$ | −15.605 | −29.754 | −19.362 | −5.284 | −17.658 | −4.550 | −5.160 | −6.523 | −5.480 | −4.806 | −21.681 | −16.699 | −20.772 | −11.666 | −17.966 |
| | $\varphi_3^*$ | −1.245 | −0.146 | 0.044 | 0.945 | 0.760 | 0.622 | 1.438 | 0.062 | 0.882 | 0.966 | 2.792 | 2.922 | 1.965 | 2.341 | 1.760 |
| | $\widehat{E}^*\varphi_3$ | −3.210 | −1.604 | 1.373 | 0.674 | 1.564 | −2.052 | −0.517 | 1.828 | 1.387 | 2.1033 | 2.963 | 3.239 | 3.136 | 3.494 | 3.476 |
| 100 | $\varphi_1^*$ | −2.274 | −1.157 | 1.048 | 0.272 | 0.940 | −1.930 | −0.754 | 1.092 | 0.276 | 0.786 | 1.284 | 1.327 | 2.370 | 1.331 | 1.964 |
| | $\widehat{E}^*\varphi_1$ | −5.705 | −3.603 | 1.250 | −0.600 | 1.071 | −4.264 | −2.555 | 1.294 | −0.261 | 1.0726 | 1.205 | 1.225 | 2.844 | 1.320 | 2.306 |
| | $\varphi_2$ | 1.729 | 0.500 | 0.968 | −0.693 | 1.213 | 1.779 | 0.672 | 0.692 | −0.310 | 0.769 | 0.190 | 0.475 | 0.915 | 0.827 | 0.395 |
| | $\varphi_2^*$ | −1.388 | −0.624 | 0.724 | 0.263 | 0.728 | −0.696 | −0.156 | 0.494 | 0.301 | 0.321 | 2.313 | 2.140 | 2.812 | 1.633 | 2.286 |
| | $\widehat{E}^*\varphi_2$ | −5.643 | −3.569 | 1.150 | −0.561 | 0.996 | −4.358 | −2.677 | 1.247 | −0.334 | 1.051 | 1.280 | 1.315 | 3.136 | 1.422 | 2.551 |
| | $\varphi_3$ | −7.498 | −7.627 | −4.517 | −8.837 | −5.689 | −0.847 | −2.154 | −2.313 | −2.482 | −2.814 | −11.824 | −11.216 | −3.465 | −8.077 | −3.949 |
| | $\varphi_3^*$ | −1.088 | −0.423 | 0.524 | 0.405 | 0.381 | −0.781 | −0.095 | −0.207 | 0.225 | −0.185 | 1.630 | −0.118 | 2.085 | 0.921 | 1.425 |
| | $\widehat{E}^*\varphi_3$ | −5.589 | −3.519 | 1.163 | −0.548 | 1.025 | −4.387 | −2.671 | 1.285 | −0.398 | 1.076 | 1.114 | 1.321 | 2.843 | 1.200 | 2.268 |
| 200 | $\varphi_1^*$ | −3.074 | −1.736 | 0.089 | −0.431 | 0.109 | −2.562 | −1.389 | 0.170 | −0.306 | 0.199 | 0.266 | 0.292 | 0.305 | 0.311 | 0.357 |
| | $\widehat{E}^*\varphi_1$ | −6.944 | −4.653 | −0.613 | −1.658 | −0.350 | −5.510 | −3.616 | −0.259 | −1.201 | −0.035 | 0.128 | 0.176 | 0.409 | 0.334 | 0.442 |
| | $\varphi_2$ | 2.394 | 1.203 | −0.354 | −0.152 | −0.623 | 4.196 | 2.778 | 0.392 | 0.862 | 0.285 | 2.193 | 1.013 | −0.537 | −0.174 | −0.627 |
| | $\varphi_2^*$ | −1.746 | −0.951 | 0.063 | −0.179 | −0.047 | −0.413 | 0.194 | 0.674 | 0.386 | 0.606 | 1.907 | 1.294 | 0.534 | 0.604 | 0.527 |
| | $\widehat{E}^*\varphi_2$ | −6.901 | −4.656 | −0.638 | −1.687 | −0.356 | −5.341 | −3.516 | −0.151 | −1.130 | 0.031 | 0.345 | 0.298 | 0.534 | 0.359 | 0.534 |
| | $\varphi_3$ | 1.626 | 0.452 | −0.843 | −0.844 | −1.091 | 3.295 | 1.986 | −0.622 | 0.267 | −0.640 | −1.988 | −2.277 | −2.378 | −2.036 | −2.261 |
| | $\varphi_3^*$ | −1.563 | −0.759 | 0.181 | 0.003 | 0.085 | −0.375 | 0.234 | 0.539 | 0.467 | 0.458 | 1.491 | 0.860 | −0.546 | 0.479 | −0.475 |
| | $\widehat{E}^*\varphi_3$ | −6.850 | −4.593 | −0.534 | −1.623 | −0.291 | −5.316 | −3.469 | −0.131 | −1.094 | 0.024 | 0.275 | 0.198 | 0.352 | 0.333 | 0.344 |

**TABLE 6** Percentage gain of MSFE compared to one-step linear forecast method: DGP 4 with different AR (1) coefficients

| | | $\rho = 0, (\alpha, \beta) =$ | | | | | $\rho = 0.5, (\alpha, \beta) =$ | | | | | $\rho = 0.95, (\alpha, \beta) =$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $R$ | Forecast | (0,0) | (.3,0) | (.9,0) | (.3,.4) | (.7,.2) | (0,0) | (.3,0) | (.9,0) | (.3,.4) | (.7,.2) | (0,0) | (.3,0) | (.9,0) | (.3,.4) | (.7,.2) |
| 20 | $\varphi_1^*$ | 6.043 | 6.324 | 6.070 | 6.601 | 6.846 | 6.853 | 6.885 | 6.102 | 7.110 | 6.772 | 8.587 | 8.615 | 7.699 | 8.988 | 8.187 |
| | $\widehat{E}^*\varphi_1$ | 6.803 | 7.095 | 7.855 | 7.621 | 8.185 | 7.398 | 7.642 | 8.028 | 8.196 | 8.484 | 9.163 | 9.563 | 10.249 | 10.489 | 10.542 |
| | $\varphi_2$ | 2.243 | 2.648 | 3.212 | 2.905 | 3.269 | 2.063 | 3.846 | 3.692 | 2.417 | 3.318 | 3.121 | 4.919 | 5.316 | 4.163 | 4.512 |
| | $\varphi_2^*$ | 5.937 | 6.343 | 6.687 | 6.843 | 7.090 | 6.998 | 7.186 | 7.239 | 7.549 | 7.217 | 9.260 | 9.480 | 9.443 | 10.090 | 9.698 |
| | $\widehat{E}^*\varphi_2$ | 6.770 | 7.072 | 7.942 | 7.584 | 8.205 | 7.362 | 7.674 | 8.186 | 8.221 | 8.547 | 9.379 | 9.865 | 10.644 | 10.884 | 10.798 |
| | $\varphi_3$ | −8.359 | −8.679 | −8.384 | −8.369 | −8.328 | −8.091 | −10.412 | −13.685 | −8.743 | −14.664 | −16.684 | −18.985 | −13.172 | −18.725 | −14.567 |
| | $\varphi_3^*$ | 4.777 | 5.244 | 5.215 | 5.804 | 5.959 | 6.053 | 5.943 | 5.518 | 6.673 | 4.544 | 2.388 | 5.614 | 5.182 | 3.311 | 5.235 |
| | $\widehat{E}^*\varphi_3$ | 6.638 | 6.985 | 7.891 | 7.505 | 8.161 | 7.189 | 7.712 | 8.093 | 8.071 | 8.222 | 8.801 | 8.446 | 9.852 | 9.806 | 10.072 |
| 50 | $\varphi_1^*$ | 1.798 | 1.867 | 1.692 | 1.966 | 1.764 | 1.904 | 2.172 | 1.929 | 2.106 | 1.968 | 2.145 | 2.597 | 2.600 | 2.845 | 2.886 |
| | $\widehat{E}^*\varphi_1$ | 1.859 | 2.063 | 2.352 | 2.316 | 2.310 | 1.963 | 2.288 | 2.674 | 2.635 | 2.697 | 1.974 | 2.627 | 2.926 | 3.432 | 3.308 |
| | $\varphi_2$ | 0.981 | 0.764 | 0.259 | 0.646 | 0.627 | 0.273 | 0.102 | 1.904 | 0.267 | 0.915 | 0.837 | 0.935 | 2.445 | 0.827 | 2.057 |
| | $\varphi_2^*$ | 1.690 | 1.897 | 1.791 | 1.921 | 1.687 | 1.807 | 2.103 | 2.038 | 2.059 | 2.102 | 1.758 | 2.217 | 2.572 | 2.928 | 2.977 |
| | $\widehat{E}^*\varphi_2$ | 1.857 | 2.040 | 2.361 | 2.354 | 2.368 | 1.904 | 2.226 | 2.632 | 2.609 | 2.682 | 1.908 | 2.295 | 3.150 | 3.260 | 3.647 |
| | $\varphi_3$ | −6.293 | −5.489 | −18.585 | −4.770 | −17.732 | −4.444 | −4.289 | −6.066 | −4.546 | −3.999 | −9.882 | −8.801 | −15.330 | −9.227 | −13.938 |
| | $\varphi_3^*$ | 1.103 | 1.323 | 0.528 | 1.190 | 0.793 | 1.201 | 1.687 | 0.560 | 1.411 | 1.438 | 1.137 | 1.565 | 1.750 | 2.047 | 1.910 |
| | $\widehat{E}^*\varphi_3$ | 1.910 | 1.989 | 2.366 | 2.414 | 2.324 | 1.927 | 2.165 | 2.494 | 2.513 | 2.608 | 1.987 | 2.427 | 3.040 | 3.168 | 3.556 |
| 100 | $\varphi_1^*$ | 0.726 | 0.854 | 1.332 | 1.042 | 1.188 | 0.426 | 0.768 | 1.250 | 0.933 | 1.012 | 0.232 | 0.541 | 2.266 | 0.932 | 1.777 |
| | $\widehat{E}^*\varphi_1$ | 0.562 | 0.855 | 1.910 | 1.350 | 1.704 | 0.684 | 0.832 | 2.199 | 1.230 | 1.899 | 0.394 | 0.595 | 2.650 | 1.008 | 2.166 |
| | $\varphi_2$ | −0.539 | −0.562 | −2.218 | −0.377 | −0.996 | −0.386 | −0.028 | −0.323 | 0.132 | −0.282 | −1.280 | −1.302 | 0.694 | −1.038 | 0.196 |
| | $\varphi_2^*$ | 0.550 | 0.736 | 1.061 | 1.004 | 1.031 | 0.367 | 0.733 | 0.803 | 0.986 | 0.777 | −0.246 | 0.130 | 2.375 | 0.736 | 1.766 |
| | $\widehat{E}^*\varphi_2$ | 0.553 | 0.864 | 1.817 | 1.352 | 1.655 | 0.436 | 0.804 | 2.084 | 1.168 | 1.839 | −0.223 | 0.172 | 2.766 | 0.958 | 2.280 |
| | $\varphi_3$ | −6.598 | −6.397 | −4.210 | −7.081 | −4.613 | −2.548 | −2.116 | −2.057 | −2.135 | −2.323 | −4.289 | −4.787 | −3.819 | −4.424 | −3.092 |
| | $\varphi_3^*$ | 0.503 | 0.402 | 0.628 | 0.366 | 0.422 | 0.297 | 0.500 | −0.002 | 0.485 | −0.065 | −0.366 | −0.151 | 1.783 | 0.206 | 1.166 |
| | $\widehat{E}^*\varphi_3$ | 0.539 | 0.816 | 1.785 | 1.324 | 1.636 | 0.307 | 0.648 | 2.139 | 1.041 | 1.852 | −0.332 | 0.144 | 2.544 | 0.902 | 2.062 |
| 200 | $\varphi_1^*$ | 0.108 | 0.273 | 0.446 | 0.400 | 0.426 | −0.149 | 0.168 | 0.416 | 0.318 | 0.420 | −1.647 | −0.473 | 0.252 | 0.043 | 0.243 |
| | $\widehat{E}^*\varphi_1$ | −0.421 | −0.058 | 0.582 | 0.362 | 0.673 | −0.668 | −0.161 | 0.609 | 0.309 | 0.677 | −1.695 | −0.988 | 0.259 | 0.079 | 0.298 |
| | $\varphi_2$ | −0.854 | −0.925 | 0.466 | 0.848 | 0.594 | −0.358 | −0.210 | 0.004 | 0.063 | 0.063 | −0.655 | −0.639 | −0.669 | −0.561 | −0.617 |
| | $\varphi_2^*$ | −0.110 | −0.018 | 0.265 | 0.272 | 0.246 | −0.279 | 0.042 | 0.500 | 0.294 | 0.465 | −1.316 | −0.759 | 0.104 | −0.191 | 0.061 |
| | $\widehat{E}^*\varphi_2$ | −0.451 | −0.096 | 0.549 | 0.307 | 0.650 | −0.687 | −0.196 | 0.639 | 0.282 | 0.677 | −1.350 | −1.146 | 0.231 | −0.190 | 0.299 |
| | $\varphi_3$ | −1.544 | −1.591 | −1.201 | −1.556 | −1.332 | −0.897 | −0.868 | −1.160 | −0.737 | −1.113 | −2.210 | −1.961 | −2.017 | −1.713 | −1.879 |
| | $\varphi_3^*$ | 0.071 | 0.149 | 0.146 | 0.285 | 0.131 | −0.189 | 0.126 | 0.188 | 0.252 | 0.194 | −1.213 | −0.687 | −0.260 | −0.271 | −0.384 |
| | $\widehat{E}^*\varphi_3$ | −0.411 | −0.044 | 0.610 | 0.364 | 0.723 | −0.687 | −0.175 | 0.659 | 0.273 | 0.683 | −1.316 | −0.997 | 0.181 | −0.265 | 0.265 |

economic constraints, and like Chen and Hong (2009), they found annually return prediction favors much of their proposed models over the historical average. However, monthly return prediction is beyond the reign of all models they considered with few exceptions.

Following the previous section, we consider applying both the linear forecasting model in (3.2) and the nonparametric forecasting model in (3.3) to forecast the monthly excess stock returns one month ahead by using some predictor variables. We consider the nine different forecasting methods detailed in Section 3.2.

### 4.1. Data

Our dependent variable ($y_{t+1}$) is always the excess returns defined by the monthly stock returns minus the corresponding risk-free rate. To be concrete, the monthly stock returns used here are the continuously compounded returns on the S&P 500 index, including dividends. The S&P 500 index returns from 1927M1 to 2005M12 are taken from Center for Research in Security Press (CRSP) month-end values. Monthly stock returns before that time are constructed by interpolation of lower-frequency data and may be not reliable. Therefore, we start the in-sample estimation and out-of -sample forecasts from 1927 to 2005. We use the Treasury-bill rate from the same period as the risk-free rate.

We construct predictors ($x_t$) based on Campbell and Shiller (1988, 1998), Goyal and Welch (2008), Campbell and Thompson (2008), and Lee et al. (2010). They are listed as follows:

**Dividend Price Ratio** ($d/p$): Dividends are 12-month moving sums of dividends paid on the S&P 500 index. The Dividend Price Ratio is the difference between the log of dividends and the log of prices.

**Earnings Price Ratio** ($e/p$): Earnings are 12-month moving sums of earnings on the S&P 500 index. The Earnings Price Ratio is the difference between the log of earnings and the log of prices.

**Smoothed Earnings Price Ratio** ($se/p$): The smoothed earnings price ratio is the ratio of a 10-year moving average of real earnings to current real prices.

**Book-to-Market Ratio** ($b/m$): The Book to Market Ratio is the ratio of book value to market value for the Dow Jones Industrial Average.

**Return on Equity** ($roe$): Return on equity is a fiscal year's net income (after preferred stock dividends but before common stock dividends) divided by total equity (excluding preferred shares), expressed as a percentage.

**Treasury Bill** (*tbl*): Treasury-bill rates from 1920 to 1933 are the U.S. yields on Short-Term United States Securities, Three-Six Month Treasury Notes and Certificates, Three Month Treasury series in the NBER Macrohistory data base. Treasury-bill rates from 1934 to 2005 are the 3-Month Treasury Bill from the economic research data base at the Federal Reserve Bank at St. Louis.

**Long Term Yield** (*lty*): The long term government bond yields from 1926 to 2005 are from Ibbotson's *Stocks, Bonds, Bills and Inflation Yearbook.*

**Term Spread** (*ts*): The Term Spread is the difference between the long term yield on government bonds and the Treasury-bill.

**Default Yield Spread** (*ds*): Default Yield Spread is the difference between BAA and AAA-rated corporate bond yields.

**Inflation** (*inf*): Inflation is the Consumer Price Index for all urban consumers from 1919 to 2005 from the Bureau of Labor Statistics.

**Net Equity Expansion** (*ntis*): Net Equity Expansion is the ratio of twelve-month moving sums of net issues by S&P listed stocks divided by the total end-of-year market capitalization of S&P stocks.

**Lagged dependent variable** (*lagy*): one-period lagged value of the excess stock returns.

Following the literature, for each forecasting model we only consider including one of the above defined predicting variables into the linear or nonparametric forecasting models.

### 4.2. Results

Following Lee et al. (2010), the in-sample estimation starts from 1950 M1. We keep a fixed in-sample size ($R$) of 24, 60, and 120 observations and roll the in-sample estimation window forward till the last available observation. We also leave $\overline{R} = 24$ observations for the estimation of $\widehat{E}^*\varphi_{1t}$, $\widehat{E}^*\varphi_{2t}$, and $\widehat{E}^*\varphi_{3t}$. Thus, the forecast begins in 1954 M1, 1957 M1, and 1962 M1 for $R = 24, 60$, and 120, respectively.

To evaluate the nine different forecast methods considered in Section 3.2, we take Campbell and Thompson (2008) out-of-sample $R^2$ defined as

$$R^2 = 1 - \frac{\sum_{t=R+\overline{R}}^{T-1}(r_{t+1} - \hat{r}_t)^2}{\sum_{t=R+\overline{R}}^{T-1}(r_{t+1} - \bar{r}_t)^2},$$

where $\hat{r}_t$ is the fitted value from a predictive regression at time $t$ (e.g., $\hat{r}_t = \varphi_{1t}$), and $\bar{r}_t$ is the historical average return estimated through period $t$. For

each forecast method, we compute $100R^2$ to measure the percentage gain in the MSE of corresponding forecast method over that of historical mean. The results are reported in Tables 7, 8, and 9 for $R = 24$, 60, and 120, respectively.

We summarize several important patterns that emerge from Tables 7–9.

First, the traditional linear predictive models without bagging tend to have higher MSE than the historical mean model. If one uses the traditional linear model to test the predictability of stock returns, one may conclude that the stock returns are not predictable.

Second, the local constant and local linear nonparametric models perform better than the linear model in some cases, but they are beaten by the historical mean models for most predictive variables.[6] This suggests that the nonparametric method can reveal certain predictability of stock returns by using some predictive variables, but not always.

Third, both the bagging method and the method proposed in this paper significantly outperform the historical mean models when $R$ is small and in majority cases our proposed methods perform better than the bagging methods. This is conformable with the findings in Tables 4–6 where the independent variable has strong serial correlation. When the in-sample number of observations is 24, the percentage gains of our proposed methods over the historical mean in terms of out-of-sample $R^2$ are all around 3.2–5.3%, while the percentage gains of the bagging method are around 0.45–4.6%. When the in-sample number of observations is 60, the gains of our proposed methods reduce to about 0.1–1.6% for most of the predicting variables under consideration, while the gains of the bagging methods are mostly negative, ranging from −2.8–0.7%. Even when the in-sample number of observations is set to be 120 (i.e., one uses all the last ten years of observations in estimating a model), we find that in some cases the percentage gains of our proposed methods and bagging methods in the out-of-sample $R^2$ are positive. Therefore, our proposed method does outperform the historical mean when $R$ is small and beat the bagging method in most cases. And as discussed in Campbell and Thompson (2008), a small value of gain in out-of-sample $R^2$, say 0.43%, is actually large enough and has its economic significance, compared to the squared monthly Sharpe ratio of stocks. The monthly Sharpe ratio of stocks is about 0.5 since 1950 according to Ibbotson's Stocks, Bonds, Bills, and Inflation Yearbook, and the squared monthly Sharpe ratio is about 2.1%. For example, when the percentage gain in the out-of-sample $R^2$ is 1.60% for the dividend price ratio ($d/p$) in the last row of Table 8, a mean-variance investor can use dividend price ratio to increase the average

[6]As we discussed in the previous section, the one-step-ahead local constant and local linear predictors are sensitive to the bandwidth choice. If we try to choose the bandwidth by the "rule of thumb": $h_l = c_0 s_l n^{-1/(4+q)}$, and set different values of $c_0 = 0.5, 1$, and 2, the results are quite different.

**TABLE 7** Percentage Gain in Out-of-Sample R-squared: Monthly Excess Returns Forecasts with R = 24

| Forecasts\Variables | *d/p* | *e/p* | *se/p* | *b/m* | *roe* | *tbl* | *lty* | *ts* | *ds* | *inf* | *ntis* | *lagy* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\varphi_1$ | −8.2240 | −14.5846 | −8.3316 | −10.1558 | −9.6488 | −7.2329 | −7.6159 | −7.6536 | −5.8778 | −12.0413 | −9.2246 | −6.8085 |
| $\varphi_1^*$ | 3.2555 | 2.6394 | 1.3576 | 2.1788 | 3.0017 | 3.3507 | 2.5648 | 3.6292 | 2.3966 | 2.5808 | 3.6174 | 1.5402 |
| $\widehat{E}^*\varphi_1$ | 4.6448 | 4.5151 | 3.6055 | 4.1706 | 4.3518 | 4.4549 | 3.9141 | 4.0061 | 3.2163 | 3.5051 | 3.7932 | 3.5870 |
| $\varphi_2$ | −6.6287 | −11.7913 | −6.5592 | −9.2151 | −6.0135 | −5.1911 | −6.1353 | −6.4879 | −4.8324 | −0.2333 | −9.3964 | −9.3710 |
| $\varphi_2^*$ | 4.5553 | 3.6299 | 3.8205 | 3.2128 | 1.9253 | 2.3522 | 0.4474 | 3.9360 | 1.9620 | 3.1199 | 3.3390 | 1.6544 |
| $\widehat{E}^*\varphi_2$ | 5.2044 | 4.5114 | 4.7119 | 5.0684 | 4.3239 | 4.0160 | 4.8039 | 4.0719 | 3.6111 | 3.8483 | 3.7299 | 3.7699 |
| $\varphi_3$ | −6.6207 | −11.8983 | −10.5276 | −11.7377 | −9.0100 | −18.6040 | −13.3777 | −20.9161 | −11.8367 | −1.5835 | −9.5950 | −10.3653 |
| $\varphi_3^*$ | 4.6067 | 3.6247 | 3.9234 | 2.8969 | 1.8063 | 2.5304 | 0.5670 | 3.8138 | 1.8100 | 3.0727 | 3.1610 | 1.3766 |
| $\widehat{E}^*\varphi_3$ | 5.2128 | 4.5170 | 4.7333 | 5.2715 | 4.3013 | 3.9427 | 4.7531 | 3.8373 | 3.5200 | 3.7584 | 3.7414 | 3.7828 |

*Note:* Sample begins:1950 M1, Forecast begin: 1954 M1, Forecast end: 2005 M12. The benchmark predictor: historical mean.

**TABLE 8** Percentage Gain in Out-of-Sample R-squared: Monthly Excess Returns Forecasts with R = 60

| Forecasts\Variables | *d/p* | *e/p* | *se/p* | *b/m* | *roe* | *tbl* | *lty* | *ts* | *ds* | *inf* | *ntis* | *lagy* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\varphi_1$ | −4.0532 | −7.0712 | −8.2401 | −5.1787 | −4.0299 | −2.4608 | −4.3541 | −0.7002 | −2.2242 | −5.9840 | −3.6829 | −4.0492 |
| $\varphi_1^*$ | −0.0307 | −0.4813 | −0.9729 | −1.0798 | −0.8034 | 0.4344 | −0.3992 | 0.6823 | 0.1024 | −0.2290 | −0.2141 | −0.7814 |
| $\widehat{E}^*\varphi_1$ | 0.6146 | 0.7553 | 0.6527 | 1.0218 | 0.3307 | 0.7477 | 0.5267 | 0.4659 | 0.2025 | −0.0396 | 0.1033 | 0.1179 |
| $\varphi_2$ | −5.9285 | −7.8908 | −9.6220 | −4.0823 | −1.8749 | −3.6750 | −2.2200 | −2.8847 | −1.1837 | −1.6589 | −7.4792 | −5.9525 |
| $\varphi_2^*$ | −0.0385 | −0.5007 | −1.7497 | −1.9121 | −0.4788 | −0.9332 | −2.7792 | −1.3844 | −1.0956 | −0.1836 | −0.1851 | −0.6824 |
| $\widehat{E}^*\varphi_2$ | 1.5864 | 0.4766 | 1.3569 | 1.5278 | 0.8122 | 0.5829 | 0.9399 | 0.6127 | −0.0622 | −0.0202 | 0.2965 | 0.0981 |
| $\varphi_3$ | −15.9111 | −8.3537 | −14.5788 | −16.1263 | −1.9405 | −31.6770 | −28.2073 | −30.5305 | −19.1754 | −2.5849 | −7.8460 | −8.2352 |
| $\varphi_3^*$ | 0.1422 | −0.5314 | −1.8670 | −2.1924 | −0.5044 | −1.1205 | −2.6848 | −1.4659 | −1.0849 | −0.2855 | −0.2246 | −0.8213 |
| $\widehat{E}^*\varphi_3$ | 1.5961 | 0.4885 | 1.3911 | 1.6055 | 0.8216 | 0.5440 | 0.9441 | 0.4372 | −0.1793 | −0.0532 | 0.2794 | 0.0761 |

*Note:* Sample begin: 1950 M1, Forecast begin: 1957 M1, Forecast end: 2005 M12. The benchmark predictor: historical mean.

**TABLE 9** Percentage Gain in Out-of-Sample R-squared: Monthly Excess Returns Forecasts with R = 120

| Forecasts\Variables | *d/p* | *e/p* | *se/p* | *b/m* | *roe* | *tbl* | *lty* | *ts* | *ds* | *inf* | *ntis* | *lagy* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\varphi_1$ | −3.3337 | −3.8498 | −5.2462 | −2.0959 | −3.2765 | −1.9689 | −3.5374 | −0.4953 | −1.1875 | −3.4901 | −1.6121 | −1.3427 |
| $\varphi_1^*$ | −0.1949 | −0.2160 | −0.5091 | −0.2117 | −0.4360 | 0.9993 | −0.2695 | 0.9175 | 0.0358 | −0.3458 | −0.2054 | −0.5406 |
| $\widehat{E}^*\varphi_1$ | 0.3926 | 0.0461 | −0.0868 | 0.4070 | −0.4087 | −0.0813 | −0.0175 | 0.0605 | −0.1073 | −0.5599 | −0.4589 | −0.2364 |
| $\varphi_2$ | −1.5064 | −6.1910 | −8.0379 | −6.4064 | −0.9793 | −1.4017 | −7.1698 | −1.0480 | −8.8851 | −2.4638 | −5.4001 | −5.0508 |
| $\varphi_2^*$ | −0.2580 | −0.6491 | −0.6470 | −0.9036 | −0.1043 | −0.3053 | −0.7836 | 0.4491 | −0.4641 | −0.3224 | −0.6356 | −0.3632 |
| $\widehat{E}^*\varphi_2$ | 0.1766 | −0.2390 | −0.0063 | 0.4759 | −0.0580 | −0.5751 | −0.3005 | 0.2871 | 0.0106 | −0.6132 | −0.2998 | −0.2354 |
| $\varphi_3$ | −1.6443 | −7.0657 | −8.2465 | −9.3440 | −1.2928 | −1.5296 | −8.0415 | −1.0399 | −8.6600 | −3.3177 | −6.0378 | −6.2492 |
| $\varphi_3^*$ | −0.2244 | −0.7170 | −0.7625 | −1.3501 | −0.1575 | −0.2575 | −0.7664 | 0.4714 | −0.5180 | −0.4212 | −0.7514 | −0.4546 |
| $\widehat{E}^*\varphi_3$ | 0.2420 | −0.1866 | 0.0325 | 0.5962 | −0.0617 | −0.5690 | −0.3144 | 0.1859 | −0.0076 | −0.6238 | −0.3368 | −0.2435 |

*Note:* Sample begin: 1950 M1, Forecast begin: 1962 M1, Forecast end: 2005 M12. The benchmark predictor: historical mean.

monthly portfolio return by a proportional factor of $1.60/2.1 = 76.19\%$. The absolute increase in portfolio return depends on risk aversion, but is about 160 basis points per month or 21.85% per year for an investor with unit risk aversion and about 7.28% per year for an investor with a risk aversion coefficient of three.[7]

In addition, we also try different in-sample estimation starting time. Since the stock return data start from 1927 M1, we use *lagy* (one month lagged value of the excess returns) to forecast the excess returns beginning from 1930 M1, 1940 M1, 1960 M1, 1970 M1, 1980 M1, and 1990 M1. And as the data for other predictive variables are available from 1936 M6, we use different predictive regression models to forecast excess returns beginning from 1940 M1, 1960 M1, 1970 M1, 1980 M1, and 1990 M1. The results are similar to what we have here for Tables 7, 8, and 9. To save space, we do not report them here.

## 5. CONCLUDING REMARKS

Combined forecasts have been suggested in the economics literature to improve forecasts over individual forecasting models. See Bates and Granger (1969), Granger et al. (1994), Granger and Jeon (2004), Stock and Watson (1999, 2009), Yang (2004), and Timmermann (2006) for details. On the one hand, combination can be formed over a set of different forecasting models with potentially different predictive variables, which includes the recent advance in forecast combination based on the method of Mallows model averaging (MMA) as advocated by Hansen (2008). On the other hand, combination can also be formed over a set of bootstrap-based training sets for a given model, which is the idea of bagging introduced by Breiman (1996a,b). In this paper, we construct combined forecasts by bagging the out-of-sample time series forecasts based on linear, local constant, and local linear regression models and consider a revised version of the traditional bagging method. We show that the bagging method and its revised version work quite well and outperform the traditional one-step-ahead linear, local constant and local linear forecasts in general and a large percentage of gains can be obtained especially when the in-sample estimation period ($R$) is small.

Interestingly, we find that bagging forecasts based on different forecasting models all yield similar percentage of gains in terms of MSFE reduction when compared with the simple linear forecast models

[7]The calculation is based on Eq. (13) in Campbell and Thompson (2008). For illustration, the paper considered an investor with a single-period horizon and mean-variance preference and calculated the expected excess return when the investor observes the predicting variable and the expected excess return when the investor does not observe the predicting variable. The difference between these two expected excess returns is $\frac{1}{\gamma}(\frac{R^2}{1-R^2})(1+S^2)$, where $\gamma$ is the coefficient of relative risk aversion, and $S$ is the unconditional Sharpe ratio of the risky asset.

despite the fact that the unbagged forecasts based on local constant regressions may significantly outperform the unbagged forecasts based on linear regressions when the underlying model is nonlinear. The gains by using the nonparametric model in conjunction with bagging tend to be incrementally marginal in comparison with the gains of bagging forecasts over unbagged forecasts. This suggests that bagging forecasts based on misspecified models may work as effectively as those based on correctly specified models (the nonparametric models here). From this point of view, bagging has the great advantage in robustifying forecasts based on different approximating models.

There is a long debate on the predictability of excess stock returns in economics and finance. In view of this, we reexamine the forecasting performances of predictive variables suggested in the literature and also consider lagged excess stock returns as an additional forecast. We find that, consistent with Goyal and Welch (2008), the monthly historical average excess stock return forecasts beat other predictor variables in the literature when we apply the traditional one-step-ahead linear forecast and the nonparametric forecasting methods. However, when using the bagging forecast methods, the monthly excess returns are actually predictable. When the in-sample estimation period $R$ is small, the percentage gains over the historical mean are very significant with different sub-sample periods. And when $R$ is large, the percentage gains are small but still outperform the historical mean in some cases. And as claimed in Campbell and Thompson (2008), the small out-of-sample predictive power is actually economically meaningful.

## ACKNOWLEDGMENTS

The authors are grateful to a referee for useful comments and suggestions. They are also thankful to Tae-Hwy Lee and Yundong Tu for discussions on the subject matter of this paper. The first author gratefully acknowledges financial support from the SMU research grant (Grant number: C244/MSS10E006).

## REFERENCES

Andrews, D. W. K. (1995). Nonparametric kernel estimation for semiparametric models. *Econometric Theory* 11:560–596.

Bates, J. M., Granger, C. W. J. (1969). The combination of forecasts. *Operations Research Quarterly* 20:451–468.

Breiman, L. (1996a). Bagging predictors. *Machine Learning* 36:105–139.

Breiman, L. (1996b). Heuristics of instability and stabilization in model selection. *Annals of Statistics* 24(6):2350–2383.

Bühlmann, P., Yu, B. (2002). Analyzing bagging. *The Annals of Statistics* 30:927–961.

Buja, A., Stuetzle, W. (2006). Observations on bagging. *Statistica Sinica* 16:323–351.

Campbell, J. Y., Lo, A. W., MacKinlay C. A. (1997). *The Econometrics of Financial Markets*. Princeton: Princeton University Press.

Campbell, J. Y., Shiller, R. J. (1988). Stock prices, earnings, and expected dividends. *Journal of Finance* 43(3):661–76.

Campbell, J. Y., Shiller, R. J. (1998). Valuation ratios and the long-run stock market outlook. *Journal of Portfolio Management* 24(2):11–26.

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

Chen, Q., Hong, Y. (2009). Predictability of equity returns over different time horizons: A nonparametric approach. Working paper, Cornell University.

Friedman, J. H., Hall, P. (2007). On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference* 137:669–683.

Goyal, A., Welch, I. 2008. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies* 21(4):1455–1508.

Granger, C. W. J., Deutsch, M., Teräsvirta, T. (1994). The combination of forecasts using changing weights. *International Journal of Forecasting* 10:47–57.

Granger, C. W. J., Jeon, Y. (2004). Thick modeling. *Economic Modeling* 21:323–343.

Hansen, B. E. (2008). Least-squares forecast averaging. *Journal of Econometrics* 146:342–350.

Hart, J., Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data. *The Annals of Statistics* 18(2):873–890.

Inoue, A., Kilian, L. (2008). How useful is bagging in forecasting economic time series? a case study of U.S. CPI inflation. *Journal of the American Statistical Association* 130:511–522.

Kristensen, D. (2009). Uniform convergence rates of kernel estimators with heterogeneous dependent data. *Econometric Theory* 25:1433–1445.

Lee, T., Yang, Y. (2006). Bagging binary and quantile predictors for time series. *Journal of Econometrics* 135:465–497.

Lee, T., Tu, Y., Ullah, A. (2010). Bagging nonparametric and semiparametric forecasts with constraints. Working paper, University of California, Riverside.

Masry, E., Tjøstheim, D. (1995). Nonparametric estimation and identification of non-linear ARCH time series. *Econometric Theory* 11:258–289.

Masry, E., Tjøstheim, D. (1997). Additive non-linear ARX time series and projection estimates. *Econometric Theory* 13:214–252.

Pagan, A., Ullah, A. (1999). *Nonparametric Econometrics*. Cambridge, UK: Cambridge University Press.

Politis, D. N., White, H. (2004). Automatic block-length selection for the dependent bootstrap. *Econometric Reviews* 23(1):53–70.

Pesaran, M. H., Timmermann, A. (2002). Model instability and choice of observation window. Working paper, Cambridge University.

Pesaran, M. H., Timmermann, A. (2004). How costly is it to ignore breaks when forecasting the direction of a time series? *International Journal of Forecasting* 20:411–424.

Stock, J. H., Watson, M. W. (1999). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In: Engle, R. F., White, H. eds. *Cointegration, Causality, and Forecasting, A Festschrift in Honor of C. W. J. Granger*. London: Oxford University Press, pp. 1–44.

Stock, J. H., Watson, M. W. (2009). Generalized shrinkage methods for forecasting using many predictors, Working paper, Harvard University.

Timmermann, A. (2006). Forecast combinations. In: Elliott, G., Granger, C. W. J., Timmermann, A. eds. *Handbook of Economic Forecasting*. Amsterdam, North Holland, pp. 135–196.

White, H. (2001). *Asymptotic Theory for Econometricians*. Bingley, UK: Emerald Group.

Yang, Y. (2004). Combining forecasting procedures: Some theoretical results. *Econometric Theory* 20:176–222.