Are financial returns really predictable out-of-sample?: Evidence from a new bootstrap test[☆]

Li Liu[a], Ruijun Bu[b], Zhiyuan Pan[c,d,*], Yuhua Xu[a]

[a] *School of Finance, Nanjing Audit University, China*
[b] *University of Liverpool, UK*
[c] *Institute of Chinese Financial Studies, Southwestern University of Finance and Economics, China*
[d] *Collaborative Innovation Center of Financial Security, China*

Abstract

Testing the out-of-sample return predictability is of great interest among academics. A wide range of studies have shown the predictability of stock returns, but fail to test the statistical significance of economic gains from the predictability. In this paper, we develop a new statistical test for the directional accuracy of stock returns. Monte Carlo experiments reveal that our bootstrap-based tests have more correct size and better power than the existing tests. We use the forecast combinations and find that the stock return predictability is statistically significant in terms of reduction of mean squared predictive error relative to the benchmark of historical average forecasts. However, the results from our tests show that the predictability is not economically significant. We conclude that there will be still a long way to go for forecasting stock returns for market participants.

Keywords: Mean predictability, Block bootstrap, Stock returns, S&P 500 index

[☆]We would like to show our sincere gratitude to the handling editor (Prof. Sushanta Mallick) and the anonymous referee whose comments and suggestions greatly improved the quality of the manuscript. This work was supported by the Chinese National Science Foundation through grant numbers 71401077 and 71771124 (Li Liu), 71601161 (Zhiyuan Pan), 61673221 (Yuhua Xu), the National Social Science Fund of China through grant number 18VSJ073 (Zhiyuan Pan), the State Scholarship Fund organized by the China Scholarship Council (CSC) through grant number 201806985006 (Zhiyuan Pan) and Qinglan project in Jiangsu province (Li Liu and Yuhua Xu).

[*]Corresponding author. Tel: +86-15021412552. Address: 555, Liutai Avenue, Wenjiang District, Chengdu 611130, China

Email address: panzhiyuancd@126.com (Zhiyuan Pan)

1. Introduction

The out-of-sample predictability of financial returns has important implications for many areas such as asset pricing and portfolio allocation. As an influential paper, Goyal and Welch (2008) document that the financial or economic models cannot outperform the benchmark of historical average model under the criterion of mean squared predictive error (MSPE). In comparison with the statistical predictability, market investors are more concerned about the profitability of making investment decisions using return forecasts. Therefore, the evaluation of forecasting performance should stand at the view point of profit maximization rather than loss function minimization, i.e., the economic predictability of stock returns.

Recently, quite a large number of studies propose the state-of-the-art methods to reveal the return predictability from both statistical and economic perspectives. These methods include the forecast combinations (Rapach, Strauss, and Zhou, 2010), time-varying parameter models (Dangl and Halling, 2012; Zhu and Zhu, 2013), diffusion index (Neely, Rapach, Tu, and Zhou, 2014; Huang, Jiang, Tu, and Zhou, 2015), economic constraints (Campbell and Thompson, 2008; Pettenuzzo, Timmermann, and Valkanov, 2014) and machine learning approaches (Zhang, Zeng, Ma, and Shi, 2018). They evaluate the significance of statistical predictability using the popular methods such as Clark and West (2007) and Diebold and Mariano (1995) tests. However, existing studies do not show the significance of economic predictability explicitly. This is important because the model of interest will be useless in practice if its improvement of economic gains over the benchmark model is insignificant. In this paper, we fill this gap by developing a new statistical test for the significance of the excess profit.

Our test is actually an extension of Anatolyev and Gerko (2005) (AG thereafter) test. The AG test is constructed relying on a trading strategy which issues a buy signal if a forecast of next period return is positive and a sell signal otherwise. The average return of this trading strategy is compared to the benchmark strategy that issues buy/sell signal at random to test for the significance of return predictability. AG show that the power of their excess profit test is higher than that of the directional accuracy test of Pesaran and

Timmermann (1992) (PT thereafter).

In this paper, we improve upon both AG and PT tests by proposing bootstrapped methods to test for the return predictability. The first order validity of block bootstrap for mean predictability under the serial dependence is demonstrated. Our methods display two advantages over AG and PT methods. First, we allow for dependence in financial asset returns and their forecasts, while the statistics of AG and PT methods are strictly based on the hypothesis of independent and identically distributed series. Therefore, our test is expected to be less suffered from size distortion. Second, with the increase of sample size T, bootstrapped distribution theoretically converges to the true distribution at the rate of T, faster than the convergence rate of asymptotic distribution of PT. In this sense, bootstrapped statistics are expected to have better power property than the corresponding statistics based on asymptotic distribution, especially when the sample size is small. This property is of more importance for financial forecast evaluation because macroeconomic data we use is often monthly or quarterly and the sample size is not large. Our simulating results further confirm these two advantages of our test, the more corrected size and better power property than the AG and PT tests.

Empirically, we also contribute to the literature by differentiating two major sources affecting forecasting performance. The literature agrees that parameter instability (i.e., time-variation in coefficients) is one of the potential challenges and might influence many of the forecasting results. For example, Dangl and Halling (2012) show that predictive regressions with time-varying coefficients dominate regressions with constant coefficients in predicting S&P 500 index returns. More studies also document that model uncertainty (i.e., the choice of predictors) is another major problem in forecasting returns. For example, Avramov (2002) uses a Bayesian model averaging approach to deal with model uncertainty and find significant predictability of stock returns. Rapach, Strauss, and Zhou (2010) show that the combined forecasts of different models are significantly more accurate than the historical average forecasts. Our empirical analysis aims to compare the effects of parameter instability and model uncertainty on forecasting performances. This issue has not been considered in the literature.

To capture the parameter instability, we employ three types of predictive regressions which differ depending on the time-variation in coefficients, constant coefficient (CC) model, Markov regime switching (MRS) model and time-varying parameter (TVP) model. Four forecast combination strategies are employed to handle the problem of model uncertainty. A total of 12 economic and financial variables are employed to predict stock excess return over the period from January 1967 through December 2012. We first use the popular out-of-sample R^2 (R^2_{OoS}) to evaluate forecasting performance. This index measures the percent reduction of MSPE of the model of interest relative to the historical average benchmark (see, e.g., Campbell and Thompson, 2008). The Clark and West (2007) (CW) statistic is applied to test for the null hypothesis that the MSPE of the model of interest is higher than or equal to the MSPE of the benchmark model. Our statistical evaluation results show that individual models cannot significantly beat the benchmark model. Models with time-varying coefficients do not perform better than models with constant coefficients. The combination strategies can generate forecasts with positive R^2_{OoS} and CW test results also indicate that the improvement of predictability is significant. This result suggests that model uncertainty, rather than parameter instability plays the major role in affecting the return forecasting performance in the statistical sense.

We further evaluate the economic significance of return predictability using the proposed test. We find that most of macroeconomic variables cannot predict stock returns individually in the economic sense. The only exceptions are that the buy/sell trading rules based on forecasts from long-term government bond returns (LTR) and the difference between Moodys BAA- and AAA-rated bond yields (DFY) can cause the significant profit when CC models, not MRS or TVP models are used. The combined forecasts cannot provide significant profits. Overall, we find little evidence about economically significant predictability of stock returns and the unpredictability cannot be explained by parameter instability or model uncertainty. Therefore, forecasting stock returns is still an unsolved problem from economic perspective.

The remainder of this paper is organized as follows: In Section 2, we propose the methodology of bootstrap-based tests for mean predictability and discuss the size and power property of our tests. Section 3 gives an application of our tests. Section 4 concludes the paper.

2. Setup and statistics of interest

According to Anatolyev and Gerko (2005)'s statement, the size of their statistics may be distorted due to serial dependence and parameter uncertainty. In this section, we propose a modified statistic based on bootstrap. Bootstrap distribution converges to the true distribution at the rate of T, which is higher than the asymptotic distribution ($\sqrt{T}$). Therefore, bootstrap statistics are more accurate than the asymptotic statistics theoretically. In recent years, bootstrap methods have been widely used to modify test statistics. For example, Corradi and Swanson (2006) use block bootstrap method to extend the CK test of Andrews (1997) and the DGT test of Diebold et al. (1998) under dynamic misspecification and parameter estimation error. Dovonon et al. (2013) propose a bootstrap statistic for high frequency returns and show that the finite sample performance of the bootstrap is superior to the existing statistics. Su and Qu (2015) adopt a wild bootstrap method to mimic the distribution of the original statistics for testing spatial autoregressive models. We modify the statistic of Anatolyev and Gerko (2005) using the bootstrap method, and expect that inference based on bootstrap critical value is more accurate than that based on asymptotically normal critical values.

Prior to constructing the bootstrap versions of the test statistics, it is worthwhile to recall two statistics of our interest. To be readable, we adopt the same notations as Anatolyev and Gerko (2005). Let y_t denote the returns of financial assets such as stocks and foreign currencies, and let $\hat{y}_t$ be the forecasting value of y_t, the null hypothesis is given as

$$\mathbb{H}_0 : E[y_t|I_{t-1}] = c, \tag{1}$$

where c means the constant variable, I_t denotes all available predictive information at time t. Obviously, equation (1) implies that the past information set cannot help to improve the accuracy of prediction for y_t.

2.1. The excess profit test of Anatolyev and Gerko (2005)

Anatolyev and Gerko (2005) form their statistic based on a trading strategy. They assume an investor goes long if the forecast $\hat{y}_t$ is nonnegative and goes short otherwise.

Thus the return of the trading rule can be expressed as

$$r_t = \text{sign}(\hat{y}_t) y_t, \tag{2}$$

where $\text{sign}(w) = 1$ when $w \geq 0$ and $\text{sign}(w) = -1$ otherwise.

The idea of their statistics is as follows:

$$\begin{aligned} E[\text{sign}(\hat{y}_t)]\mathbb{E}[y_t] &= E\big[\text{sign}(\hat{y}_t)E[y_t]\big] \\ &\overset{H_0}{=} E\big[\text{sign}(\hat{y}_t)E[y_t|I_{t-1}]\big] \\ &= E\big[E[\text{sign}(\hat{y}_t)y_t|I_{t-1}]\big] \\ &= E[\text{sign}(\hat{y}_t)y_t] = E[r_t]. \end{aligned} \tag{3}$$

Under the null hypothesis, the average return of buy/sell trading rules formed by sign of forecasts should statistically equal to a benchmark strategy that issues buy/sell signals at random with probabilities corresponding to the proportion of "buys" and "sells" implied ex post by trading strategy (Anatolyev and Gerko, 2005). To design a feasible test statistic, we should estimate the expectation given above. Using the sample data, the right-hand side (RHS) estimator can be easily calculated as

$$A_T = \frac{1}{T}\sum_{t=1}^{T} r_t, \tag{4}$$

and the left-hand side (LHS) estimator is

$$B_T = \big(\frac{1}{T}\sum_{t=1}^{T}\text{sign}(\hat{y}_t)\big)\big(\frac{1}{T}\sum_{t=1}^{T} y_t\big). \tag{5}$$

Under the null hypothesis, the variance of $A_T - B_T$ is

$$Var[A_T - B_T] = \frac{4(T-1)}{T^2} p_{\hat{y}}(1-p_{\hat{y}}) Var[y_t] \tag{6}$$

where

$$p_{\hat{y}} = Pr(\text{sign}(\hat{y}_t) = 1). \tag{7}$$

The estimator for (6) is

$$\hat{V}_{AG} = \frac{4}{T^2}\hat{p}_{\hat{y}}(1-\hat{p}_{\hat{y}})\sum_{t=1}^{T}(y_t - \bar{y})^2 \tag{8}$$

and the estimator for (7) is

$$\hat{p}_{\hat{y}} = \frac{1}{2}\big(1 + \frac{1}{T}\sum_{t=1}^{T}\text{sign}(\hat{y}_t)\big). \tag{9}$$

According to Hausman (1978)'s suggestion, AG test statistics can be defined and can has asymptotic normal distribution as

$$AG \equiv \frac{A_T - B_T}{\sqrt{\hat{V}_{AG}}} \xrightarrow{d} \mathcal{N}(0,1). \tag{10}$$

2.2. The directional accuracy test of Pesaran and Timmermann (1992)

Besides, we can use the Pesaran and Timmermann (1992)'s directional accuracy statistic to test the null hypothesis (1). Firstly, we let $\tilde{A}$ and $\tilde{B}$ be, respectively,

$$\tilde{A}_T = \frac{1}{T}\sum_{t=1}^{T}\text{sign}(\hat{y}_t)\text{sign}(y_t), \tag{11}$$

and

$$\tilde{B}_T = \big(\frac{1}{T}\sum_{t=1}^{T}\text{sign}(\hat{y}_t)\big)\big(\frac{1}{T}\sum_{t=1}^{T}\text{sign}(y_t)\big) \tag{12}$$

Then, the PT statistic can be expressed similarly as

$$PT \equiv \frac{\tilde{A}_T - \tilde{B}_T}{\sqrt{\hat{V}_{PT}}} \xrightarrow{d} \mathcal{N}(0,1), \tag{13}$$

where

$$\hat{V}_{PT} = \frac{16(T-1)}{T^2}\hat{p}_{\hat{y}}(1-\hat{p}_{\hat{y}})\hat{p}_y(1-\hat{p}_y), \tag{14}$$

and

$$\hat{p}_y = \frac{1}{2}(1 + \frac{1}{T}\sum_{t=1}^{T}\text{sign}(y_t)). \tag{15}$$

2.3. Validity of the block bootstrap

From the simulation experiment of Anatolyev and Gerko (2005), we can see that both statistics still have low power even as the sample size goes to 1000, and the size performance is not clear at finite sample case (see Table 1 in AG paper for more details). These shortcomings may be caused by the presence of serial dependence. Block bootstrap theory has been

established and shown to be valid for serial dependence (e.g., Gonçalves and White, 2004; Corradi and Swanson, 2006; Dovonon et al., 2013).

We use the bootstrap statistics AG^* (PT^*) defined below to obtain the critical values for the original statistics AG (PT), instead of their asymptotic critical values. The bootstrap procedures can be implemented as follows.

1. Let $w_t = (\hat{y}_t, y_t)$. Given a sample $\{w_1, w_2, \cdots, w_T\}$, draw the 1st block of length l from the sample, $\{w^*_{s_0+1}, w^*_{s_0+2}, \cdots, w^*_{s_0+l}\}$, where s_i means the $(i+1)$th draw on $[0, 1, \cdots, T-l]$ with the independent and identical uniform distribution. Thus, each draw has the probability of $1/(T-l+1)$.

2. As the replacement, draw the 2nd block from the same sample to obtain $\{w^*_{s_1+1}, w^*_{s_1+2}, \cdots, w^*_{s_1+l}\}$.

3. Let $T = bl$. Until the bth block, we have a sample $\{w^*_{s_0+1}, \cdots, w^*_{s_1+1}, \cdots, w^*_{s_{b-1}+1}, \cdots\}$ and calculate the AG^* (PT^*) statistics, denoted by $AG^*_{(1)}$ ($PT^*_{(1)}$).

4. Replicate steps 1-3 B times, we collect two sets, $\{AG^*_{(1)}, AG^*_{(2)}, \cdots, AG^*_{(B)}\}$ and $\{PT^*_{(1)}, PT^*_{(2)}, \cdots, PT^*_{(B)}\}$. Then the bootstrapped p value is

$$p^*(\tau) = \frac{1}{B}\sum_{i=1}^{B} I(\tau^*_{(i)} > \tau)$$

where $\tau \in \{AG, PT\}$ and $I(E)$ is an indicator function, which takes the value of one when the condition E is satisfied and 0 otherwise.

The proposed versions of AG and PT statistics used in the 3rd of the bootstrap procedure are given as follows:

$$AG^* \equiv \frac{A^*_T - B^*_T - (A_T - B_T)}{\sqrt{\hat{V}^*_{AG}}}, \tag{16}$$

$$PT^* = \frac{\tilde{A}^*_T - \tilde{B}^*_T - (\tilde{A}_T - \tilde{B}_T)}{\sqrt{\hat{V}^*_{PT}}}, \tag{17}$$

where notation $(^*)$ denotes the corresponding bootstrap value. For example, $A^*_T = \sum_{t=1}^{T}$ sign$(\hat{y}^*_t)y^*_t$ and $\hat{y}^*_t, y^*_t$ are bootstrap samples. And the corresponding bootstrapped variances

of these two statistics are given by,

$$\begin{aligned}\hat{V}^*_{AG} &= Var^*[A^*_T - B^*_T] = Var^*[A^*_T] - Var^*[B^*_T] \\ &= \frac{4}{T^2}\hat{p}^*_{\hat{y}^*}(1-\hat{p}^*_{\hat{y}^*})\sum_{t=1}^{T}(y^*_t - \bar{y}^*)^2 \qquad (18)\\ \hat{V}^*_{PT} &= \frac{16(T-1)}{T^2}\hat{p}^*_{\hat{y}^*}(1-\hat{p}^*_{\hat{y}^*})\hat{p}^*_{y^*}(1-\hat{p^*}_{y^*}) \qquad (19)\end{aligned}$$

The following theorems show that our bootstrap method is valid.

Theorem 1. *Under some regularity conditions and* $\mathbb{H}_0$ *holds, and let* $T = bl$*, such that as* $T \to \infty, l/T \to 0$*. Then,* $\sup_{x\in R}|P^*(AG^* \le x) - P(AG \le x)| \xrightarrow{p} 0$*, where* P^* *means conditional probability given* $\omega_1, \cdots, \omega_T$*.*

Theorem 2. *Under some regularity conditions and* $\mathbb{H}_0$ *holds, and let* $T = bl$*, such that as* $T \to \infty, l/T \to 0$*. Then,* $\sup_{x\in R}|P^*(PT^* \le x) - P(PT \le x)| \xrightarrow{p} 0$*, where* P^* *means conditional probability given* $\omega_1, \cdots, \omega_T$*.*

Remark: Our conditions are the same with Hausman (1978) since our statistics belong to Hausman-type statistics, and is therefore omitted. As Paparoditis and Politis (2005) claim, there exist two ways to construct a bootstrap test of the hypothesis by imposing the null hypothesis or not. Imposing the null hypothesis will certainly cause the loss of power but the statistics become simpler. Our simulating analysis will show that even in our case of imposing the null hypothesis, the power is still much higher than the asymptotic statistics of AG and PT.

2.4. Finite sample performance

In this subsection, we investigate the finite sample performance of our bootstrap statistics. In detail, we check if the test statistics (10) and (13) have a reasonable size and good power. This procedure is necessary because the test statistic may lead to a misleading inference if the size or power has poor performance. For comparison, we simulate data using the same models as Anatolyev and Gerko (2005) to examine the size and power. Constant model and GARCH are used to evaluate size performance and AR and SETAR models

are employed to assess power performance. The detailed model specifications are given as follows:

Constant model:

$$\begin{aligned} y_t &= 0.001526 + \epsilon_t, \\ \epsilon_t &\overset{i.i.d}{\sim} \mathcal{N}(0, 0.000025) \end{aligned} \tag{20}$$

GARCH model:

$$\begin{aligned} y_t &= 0.002483 + \epsilon_t \\ \epsilon_t &= \sqrt{h_t}\eta_t \\ h_t &= 0.0000223 + 0.1773\epsilon_{t-1}^2 + 0.7397h_{t-1} \\ \eta_t &\overset{i.i.d}{\sim} \mathcal{N}(0, 1), \end{aligned} \tag{21}$$

AR model:

$$\begin{aligned} y_t &= 0.1256y_{t-1} + \epsilon_t \\ \epsilon_t &\overset{i.i.d}{\sim} \mathcal{N}(0, 0.000249), \end{aligned} \tag{22}$$

SETAR model:

$$\begin{aligned} y_t &= \begin{cases} 0.000844 + 0.2453y_{t-1} + \epsilon_t, & \text{if } |y_{t-1}| \leq 0.1848 \\ 0.002679 + 0.0664y_{t-1} + \epsilon_t, & \text{if } |y_{t-1}| < 0.1848 \end{cases} \\ \epsilon_t &\overset{i.i.d}{\sim} \mathcal{N}(0, 0.000245). \end{aligned} \tag{23}$$

To forecast the $\hat{y}_t$, we run the simple linear model with a constant term and the first lag of y_t as regressors, i.e., $y_t = \alpha + \beta y_{t-1} + \epsilon_t$. Therefore, the optimal forecast $\hat{y}_t = \hat{\alpha} + \hat{\beta}y_{t-1}$. For each sample size T, we estimate the parameters using rolling window of 100 observations, then the predict sample size $T - 100$ is used to compute statistics. This approach is also adopted by Anatolyev and Gerko (2005) and used quite intensively in empirical work (Pesaran and Timmermann, 1995).

We apply the sample sizes 250, 500, 750 and 1000 in our simulation. To obtain the block bootstrap critical value, we set the bootstrap resample $B = 300$ times. At each

experiment, we repeat 1000 times and record the frequency that p value is smaller than the given significant level.

Table 1 reports the empirical sizes of the original statistics (10), (13) at 1%, 5% and 10% significance levels under the constant model and GARCH model. We can see that both the sizes of AG and PT statistics are distorted under the constant model at small sample case. For example, when sample size is equal to 250 ($T = 250$), the sizes of AG and PT statistics are 0.050 and 0.038 at 10% significance level, respectively. Notably, AG statistic is over-rejected when GARCH process is used.

Insert Table 1 here

Tables 2-3 display the size performances of our statistics based on the block bootstrap when constant model and GARCH model are employed to simulate series, respectively. Following Fitzenberger (1998) and Corradi and Swanson (2006), the block length (l) is chosen as $1, 5, 10$ and 15. As expected, empirical significance levels are in general closer to nominal levels even when the sample is 250, indicating that our proposed statistics are quit more reliable for small sample sizes.

Insert Tables 2-3 here

Turning to the empirical power shown in the tables 4-6, the power increases when the sample size become large, as expected. We find that the power of AG statistic is large than that of PT statistic, and the percent increase of power is about 10%(see Table 4), which is consistent with Anatolyev and Gerko (2005)'s result. When using block bootstrap method, both AG^* statistics and PT^* statistics appear to have higher power no matter what the length (l) is used comparing with the corresponding results in table 4. Our bootstrap methods can improve about 10%-20% of the power (see Tables 5-6), suggesting that bootstrap statistics have better finite sample performances.

Insert Tables 4-6 here

3. Empirical application

3.1. Forecast methodology

For the application of our test, we turn to the problem of forecasting stock excess returns, one of the hottest issues in the area of financial economics. In this section, we will check the performances of stock trading rules constructed on a wide range of individual and combined macroeconomic signals. In the literature, it has been well documented that the benchmark of historical average forecasts is very difficult to be beaten (Goyal and Welch, 2008). Nevertheless, some recent studies still reveal the predictability by handling with two important problems in predictive regressions, parameter instability (Dangl and Halling, 2012) and model uncertainty (Avramov, 2002; Rapach et al., 2010). We try to find which factor has greater impacts on forecasting performances. This topic has not been considered in existing studies.

To investigate the effect of parameter instability, we employ three types of predictive regressions which differ depending on the degree of parameter variation. The first is the constant coefficient (CC) model which assumes that the predictive relationship does not change over time. The specification of CC model is given by,

$$r_{t+1} = \alpha_i + \beta_i x_{i,t} + \epsilon_{t+1}, \tag{24}$$

where r_{t+1} is the return on a stock market index in excess of the risk-free rate, $x_{i,t}$ is a predictive variable of interest and ϵ_{t+1} is a disturbance term.

The second type of predictive regression is time-varying parameter (TVP) model. In this specification, the regression parameter is assumed to follow the process of random walk without drift. Therefore, the predictive relationship implied by this model changes at each point of time. Recently, Dangl and Halling (2012) show that TVP predictive regression dominates the CC one in forecasting stock returns. The specification of a TVP model is given by,

$$\begin{aligned} r_{t+1} &= X_{i,t}\theta_{i,t} + \epsilon_{t+1} \\ \theta_{i,t} &= \theta_{i,t-1} + \eta_{i,t} \end{aligned} \tag{25}$$

where $X_{i,t} = [1, x_{i,t}], \theta_{i,t} = [\alpha_{i,t}, \beta_{i,t}]'$ and $\eta_{i,t} \overset{i.i.d}{\sim} \mathcal{N}(0, Q_i)$.

It may be argued that TVP specification over-estimates the change of predictive relationship. For this consideration, we employ a Markov regime switching (MRS) model in which the predictive relationship is allowed to change between different regimes. In this sense, MRS regression can be considered as a midpoint between CC and TVP regressions. The combination of MRS models is found to deliver consistent out-of-sample forecasting gains relative to the historical average (Zhu and Zhu, 2013). The specification of an MRS model is given by,

$$r_{t+1} = X_{i,t}\theta_{i,s_t} + \epsilon_{t+1}, \quad s_t = 0, 1, \tag{26}$$

where $\theta_{i,s_t} = [\alpha_{i,s_t}, \beta_{i,s_t}]', \epsilon_t \overset{i.i.d}{\sim} \mathcal{N}(0, H_{s_t})$, and $s_t = 0$ or $s_t = 1$ represents one of the two regimes that follow a first-order Markov process with a constant transition probability $Pr(s_t = j|s_{t-1} = i) = p_{ij}, i, j \in (0, 1)$. Under the MRS framework, we obtain the regime dependent forecasts. The final forecast used is the weighted average of forecast in each of the two regimes, where the weight equals to the transition probability. The parameter of TVP and MRS models are obtained via the maximum likelihood estimation (MLE) method. Following the majority of the literature, we generate out-of-sample forecasts of stock return using a recursive (expanding) estimation window.

Although parameter instability can be well addressed, the predictive ability of an individual model is still rather unstable due to the problem of model uncertainty. Model uncertainty recognizes that forecasters do not know which variables should enter the predictive regression and the inclusion of irrelevant predictive variable can cause overfitting, a situation that in-sample performance is improved but out-of-sample performance becomes worse. We follow Rapach et al. (2010) by combining forecasts from different individual models to address model uncertainty.

We consider four standard forecast combination methods used in Rapach et al. (2010). The first method is the mean forecast combination (MFC) which takes the simple equal-weighted average of forecasts from individual models of interest. MFC is imposed over CC, MRS and TVP models. We denote these methods by MFC-CC, MFC-MRS and MFC-TVP,

respectively.

The second method is the trimmed mean combination (TMC) which uses the equal-weighted average of forecasts of individual models after trimming the one with the worst past performance (i.e., the highest mean squared predictive error). We denote the TMC for CC, MRS and TVP forecasts by TMC-CC, TMC-MRS and TMC-TVP, respectively.

The last two methods refer to the discounted mean squared predictive error (DMSPE). DMSPE uses the weighted average of individual forecasts and the weights of each model are given by:

$$\omega_{i,t} = \frac{\phi_{i,t-1}^{-1}}{\sum_{j=1}^{N} \phi_{j,t-1}^{-1}}, \tag{27}$$

where N is the number of individual models, $\phi_{i,t} = \sum_{j=1}^{t} \delta^{t-j}(r_t - \hat{r}_{i,t})^2$ and $\hat{r}_{i,t}$ is the forecast of model i. We follow Rapach et al. (2010) and Zhu and Zhu (2013) by using the discounting factor $\delta = 1$ and $\delta = 0.9$. We denote these DMSPE strategies for CC, MRS and TVP models for DMSPE(δ)-CC, DMSPE(δ)-MRS, and DMSPE(δ)-TVP, respectively.

3.2. Forecast evaluation

We use two criteria to evaluate forecast accuracy. The first is the statistical measure, the out-of-sample R-square (see, e.g., Campbell and Thompson, 2008; Rapach et al., 2010; Neely et al., 2014). This measure computes percent reduction of mean squared predictive error (MSPE) of the model of interest relative ($MSPE_{model}$) to the MSPE of the benchmark of historical average ($MSPE_{bench}$), defined as:

$$R_{oos}^2 = 1 - \frac{MSPE_{model}}{MSPE_{bench}}, \tag{28}$$

where $MSPE_i = \frac{1}{M}\sum_{t=1}^{M}(r_t - \hat{r}_{t,i})^2$, $(i = model, bench)$; r_t and $\hat{r}_t$ are, respectively, the true value and forecast of stock return. A positive R_{oos}^2 implies that the model forecasts of interest display lower MSPE than the benchmark ones, implying the greater accuracy. To examine whether the differences in forecasting accuracy of the two different models are significant, we use the MSPE-adjusted statistic of Clark and West (2007) (CW). We can obtain a p-value for a one-sided (upper-tail) test with the standard normal distribution.

The null hypothesis of CW test is that the historical average MSPE is not greater than the predictive model MSPE (corresponding to $H_0 : R^2_{oos} \leq 0$ against $H_A : R^2_{oos} > 0$).

We also use our proposed profit test to find whether the trading rules based on return forecasts can make significant profit. In existing studies, to evaluate the economic significance of return predictability, a utility function is pre-specified and return and volatility forecasts are taken as the key inputs to pre-determine the optimal weights of stock index in a portfolio during the next period that can maximize the investor utility. The return forecast which can form the portfolio with greater performance is considered to have higher economic significance. However, this methodology for evaluating return forecasts is sensitive to the volatility forecasts and the choice of utility function. An advantage of our test is that it relies on the return forecasts uniquely.

3.3. Data and variables

We use the monthly excess returns of the S&P 500 Index from January 1947 to December 2012 to conduct empirical analysis. The start of the sample is the same as that of Rapach et al. (2010). We calculate the excess returns (or risk premium) as the aggregate returns of the S&P 500 Index (including dividends) minus the short-term interest rate proxied by a risk-free bill rate. The stock return and predictor data taken from the literature enable us to compare results. We use 12 variables to predict stock returns. A representative study by Goyal and Welch (2008)[1] offers a detailed explanation of these predictors. For the sake of brevity, we describe the following predictive variables briefly :

- Dividend yield (DY): the log of dividends minus the log of lagged stock prices (S&P 500 Index). The dividends used here are the 12-month moving sums of dividends paid on the S&P 500 Index.

- Earning price ratio (EP): the log of earnings minus the log of stock prices. The earnings used here are the 12-month moving sums on the S&P 500 Index.

[1]We particularly thank Amit Goyal for providing these data at his homepage (http://www.hec .unil.ch/agoyal/).

- Dividend payout ratio (DP): the log of dividends minus the log of earnings.
- Stock return volatility (SVOL): the sum of squared daily stock returns on the S&P 500 Index in each month.
- Book-to-market ratio (BM): the ratio of book value to market value for the Dow Jones Industrial Average.
- Net equity expansion (NTIS): the ratio of 12-month moving sums of net issues by New York Stock Exchange (NYSE) listed stocks to the total market capitalization of NYSE stocks.
- Treasury bill rate (TBL): the interest rate on a three-month Treasury bill (secondary market).
- Long-term yield (LTY): long-term government bond yield.
- Long-term return (LTR): long-term government bond returns.
- Default yield spread (DFY): the difference between Moodys BAA- and AAA-rated bond yields.
- Default return spread (DFR): the difference between long-term corporate and government bond returns.
- Inflation rate (INFL): the inflation rate calculated from the Consumer Price Index for all urban consumers. As the inflation rate information is reported with a delay, we follow the literature and use one-month lagged inflation rates (e.g., Dangl and Halling, 2012; Rapach et al., 2010; Neely et al., 2014)

Based on the dataset used by Goyal and Welch (2008) and Neely et al. (2014), we exclude the DP variable because it is almost perfectly correlated with DY. We do not use the term spread to avoid the problem of collinearity, as it is the difference between ITY and TBL.

3.4. Forecasting results

Table 7 reports the forecasting results of individual models in which only a predictive variable is included based on the criterion of R^2_{oos}. We can see that almost none of the individual variables can significantly beat the benchmark of historical average, evidenced by the negative values of R^2_{oos}. The only two exceptions are that DY and SVOL forecasts are slightly more accurate than historical average forecasts, with the R^2_{oos} of 0.033 and 0.078 percent, respectively. This result is generally consistent with the finding in the main stream literature that a single model is difficult to outperform the historical average benchmark (Goyal and Welch, 2008). More importantly, the R^2_{oos} values of forecasts of MRS and TVP models do not higher than the simple CC models. This result is contrary to the finding in the paper by Dangl and Halling (2012) who point out that predictive models with time-varying coefficients dominate models with constant coefficients. The reason is that Dangl and Halling (2012) use a combination of several TVP models with various degrees of parameter variation, while we use single predictive models only. It is likely that the predictive relationships change at different extents during different periods of time. The combined strategy in Dangl and Halling (2012) paper actually accounts for both parameter instability and model uncertainty. Our evidence suggests that parameter instability is not the main source about the inferior forecasting performances of individual models.

Insert Table 7 here

Turning to the performances of forecast combinations reported in Table 8, we can see that each combination over CC or MRS models results in significantly more accurate forecasts than the historical average forecasts, with the R^2_{oos} values of 0.7-1 percent. The R^2_{oos} values of combination strategies are also higher than all individual models, suggesting the benefit of improvement of predictive ability. Consistently, combinations over MRS and TVP models do not perform better than combinations over CC models in forecasting stock returns. Therefore, the return predictability is available by addressing the problem of model uncertainty, rather than parameter instability.

Insert Table 8 here

We have found predictability of stock returns in the statistical sense using forecast combinations. To examine whether the predictability is economically significant, we consider a trading rule determined by the sign of forecasts of excess returns. Specifically, the investor holds a long (short) position of stock index if the predicted stock return is positive (negative) (see Anatolyev and Gerko, 2005). We use our bootstrap based AG method to test the significance of the returns of this trading strategy[2]. The null hypothesis is that the average return resulting from use of the trading strategy is lower than or equal to the average return of a benchmark strategy that issues buy/sell signals at random with probabilities corresponding to the proportion of "buys" and "sells" implied *ex post* by the trading strategy. Table 9 shows the average annualized returns of buy/sell strategy based on the sign of return forecasts from individual models. We also report p-values via 1000 block bootstraps and use the block lengths of 5 and 15. Generally speaking, the average returns of buy/sell strategy depend heavily on the use of predictive variable. Most of individual models cannot generate significant profit with only a few exceptions. In detail, LTR and DFR forecasts can on average result in annualized returns of 670 and 500 bps when CC specifications are employed, respectively. The obtained returns are also significant according to the bootstrapped p-values of AG test. Consistently, predictive models with time variation in coefficients cannot perform better than models with constant coefficients.

Insert Table 9 here

Table 10 reports the annualized returns of trading rules formed by combined forecasts, as well as bootstrapped p-values. Unlike the forecasting performances in the statistical sense, we cannot find any combination strategy yielding significant returns by looking at the p-values. The combination strategy does not necessarily perform better than individual models under the economic evaluation. Therefore, although we have deal with the problems of model uncertainty and parameter instability carefully, we do not obtain economically significant predictability of stock returns. None of these two factors can explain the failure

[2]We do not use the PT test because its power is demonstrated to be lower than AG test.

in finding return predictability. In this economic sense, forecasting stock returns will be still a long way to go.

Insert Table 10 here

4. Conclusions

Testing for the predictability of financial asset returns has been of great interest for academicians. In this paper, we give the first order validity of the block bootstrap for predictability under the serial dependence. We construct new bootstrap tests for mean predictability based on a simple buy/sell trading strategy. Two test statistics are considered; one is based on excess profit test of AG, and the other is in the spirit of direction accuracy test of PT. The simulating results show that our bootstrapped method displays more correct size and higher power than the asymptotic AG and PT statistics, especially for smaller sample size.

Our tests are also applied to examine the stock return predictability. We use 12 macroeconomic variables to predict excess stock return of S&P 500 index. CC, MRS and TVP models are employed to deal with the problem of parameter instability. Four forecast combinations are used to account for the effect of model uncertainty. We find significant predictability in terms of MSPE using combination strategies. Model uncertainty plays the major role in affecting forecasting performances in the statistical sense. However, when using our test for the significance of profit, we find little evidence of return predictability. Parameter instability or model uncertainty cannot explain the unpredictability of returns in economic sense. Revealing stock return predictability is still a difficult task for academicians.

References

Anatolyev, S., Gerko, A., 2005. A trading approach to testing for predictability. Journal of Business & Economic Statistics 23 (4), 455–461.

Andrews, D. W., 1997. A conditional kolmogorov test. Econometrica: Journal of the Econometric Society, 1097–1128.

Avramov, D., 2002. Stock return predictability and model uncertainty. Journal of Financial Economics 64 (3), 423–458.

Campbell, J. Y., Thompson, S. B., 2008. Predicting excess stock returns out of sample: Can anything beat the historical average? Review of Financial Studies 21 (4), 1509–1531.

Clark, T. E., West, K. D., 2006. Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. Journal of Econometrics 135 (1), 155–186.

Clark, T. E., West, K. D., 2007. Approximately normal tests for equal predictive accuracy in nested models. Journal of econometrics 138 (1), 291–311.

Corradi, V., Swanson, N. R., 2006. Bootstrap conditional distribution tests in the presence of dynamic misspecification. Journal of Econometrics 133 (2), 779–806.

Dangl, T., Halling, M., 2012. Predictive regressions with time-varying coefficients. Journal of Financial Economics 106 (1), 157–181.

Diebold, F., Mariano, R., 1995. Comparing predictive accuracy. Journal of Business & Economic Statistics 13 (3), 253–63.

Diebold, F. X., Gunther, T. A., Tay, A. S., 1998. Evaluating density forecasts with applications to financial risk management. International Economic Review, 863–883.

Dovonon, P., Goncalves, S., Meddahi, N., 2013. Bootstrapping realized multivariate volatility measures. Journal of Econometrics 172 (1), 49–65.

Fitzenberger, B., 1998. The moving blocks bootstrap and robust inference for linear least squares and quantile regressions. Journal of Econometrics 82 (2), 235–287.

Gonçalves, S., White, H., 2004. Maximum likelihood and the bootstrap for nonlinear dynamic models. Journal of Econometrics 119 (1), 199–219.

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

Hausman, J. A., 1978. Specification tests in econometrics. Econometrica: Journal of the Econometric Society, 1251–1271.

Huang, D., Jiang, F., Tu, J., Zhou, G., 2015. Investor sentiment aligned: A powerful predictor of stock returns. The Review of Financial Studies 28 (3), 791–837.

Neely, C. J., Rapach, D. E., Tu, J., Zhou, G., 2014. Forecasting the equity risk premium: the role of technical indicators. Management Science 60 (7), 1772–1791.

Paparoditis, E., Politis, D. N., 2005. Bootstrap hypothesis testing in regression models. Statistics & probability letters 74 (4), 356–365.

Pesaran, M. H., Timmermann, A., 1992. A simple nonparametric test of predictive performance. Journal of Business & Economic Statistics 10 (4), 461–465.

Pesaran, M. H., Timmermann, A., 1995. Predictability of stock returns: Robustness and economic significance. The Journal of Finance 50 (4), 1201–1228.

Pettenuzzo, D., Timmermann, A., Valkanov, R., 2014. Forecasting stock returns under economic constraints. Journal of Financial Economics 114 (3), 517–553.

Rapach, D. E., Strauss, J. K., Zhou, G., 2010. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. Review of Financial Studies 23 (2), 821–862.

Su, L., Qu, X., 2015. Specification test for spatial autoregressive models. Journal of Business & Economic Statistics (just-accepted).

Zhang, Y., Zeng, Q., Ma, F., Shi, B., 2018. Forecasting stock returns: Do less powerful predictors help? Economic Modelling.

Zhu, X., Zhu, J., 2013. Predicting stock returns: A regime-switching combination approach and economic links. Journal of Banking & Finance 37 (11), 4120–4133.

Table 1: Empirical proportion of rejections (size).

T	AG stat			PT stat		
	1% level	5% level	10% level	1% level	5% level	10% level
			DGP: Constant model			
250	0.007	0.027	0.050	0.002	0.012	0.038
500	0.010	0.039	0.079	0.003	0.025	0.060
750	0.007	0.044	0.084	0.003	0.025	0.070
1000	0.013	0.057	0.119	0.008	0.048	0.098
			DGP: GARCH model			
250	0.023	0.083	0.144	0.006	0.034	0.082
500	0.023	0.088	0.139	0.013	0.053	0.100
750	0.024	0.088	0.141	0.009	0.054	0.095
1000	0.019	0.076	0.127	0.008	0.054	0.113

Notes: The table shows the size of AG statistics (10) and PT statistics (13) based on asymptotic critical value, respectively. The data are generated from constant model (20) and GARCH model (21). All results are based on 1000 Monte Carlo simulation.

Table 2: Empirical proportion of rejections (size)

T	AG* stat			PT* stat		
	1% level	5% level	10% level	1% level	5% level	10% level
			$l = 1$			
250	0.005	0.030	0.093	0.003	0.023	0.075
500	0.005	0.040	0.115	0.003	0.027	0.078
750	0.009	0.047	0.108	0.005	0.033	0.076
1000	0.010	0.045	0.089	0.007	0.041	0.079
			$l = 5$			
250	0.006	0.034	0.102	0.008	0.031	0.088
500	0.003	0.037	0.105	0.006	0.039	0.092
750	0.008	0.055	0.095	0.009	0.036	0.082
1000	0.009	0.046	0.099	0.008	0.042	0.087
			$l = 10$			
250	0.005	0.039	0.101	0.007	0.031	0.072
500	0.007	0.041	0.096	0.004	0.030	0.071
750	0.008	0.043	0.106	0.008	0.040	0.077
1000	0.008	0.052	0.111	0.012	0.045	0.096
			$l = 15$			
250	0.017	0.056	0.119	0.011	0.047	0.087
500	0.005	0.041	0.110	0.009	0.035	0.080
750	0.009	0.050	0.111	0.008	0.029	0.076
1000	0.008	0.041	0.107	0.007	0.026	0.082

Notes: The table shows the size of AG^* statistics (16) and PT^* statistics (17) based on bootstrap critical value, respectively. The data are generated from constant model (20). All results are based on 1000 Monte Carlo simulation and $B = 300$ bootstrap replications.

Table 3: Empirical proportion of rejections (size).

T	AG* stat			PT* stat		
	1% level	5% level	10% level	1% level	5% level	10% level
			$l = 1$			
250	0.009	0.038	0.080	0.004	0.037	0.078
500	0.013	0.039	0.072	0.007	0.034	0.071
750	0.008	0.036	0.082	0.006	0.050	0.104
1000	0.008	0.034	0.080	0.014	0.046	0.085
			$l = 5$			
250	0.009	0.051	0.103	0.011	0.049	0.097
500	0.007	0.036	0.071	0.005	0.040	0.080
750	0.007	0.036	0.081	0.011	0.043	0.097
1000	0.007	0.032	0.069	0.003	0.042	0.096
			$l = 10$			
250	0.012	0.046	0.081	0.013	0.046	0.084
500	0.014	0.046	0.081	0.022	0.048	0.091
750	0.009	0.038	0.088	0.008	0.054	0.113
1000	0.008	0.044	0.095	0.010	0.055	0.095
			$l = 15$			
250	0.019	0.052	0.110	0.020	0.051	0.100
500	0.009	0.038	0.074	0.012	0.047	0.099
750	0.008	0.043	0.093	0.014	0.049	0.092
1000	0.003	0.039	0.071	0.014	0.039	0.088

Notes: The table shows the size of AG^* statistics (16) and PT^* statistics (17) based on bootstrap critical value, respectively. The data are generated from GARCH model (21). All results are based on 1000 Monte Carlo simulation and $B = 300$ bootstrap replications.

Table 4: Empirical proportion of rejections (power).

T	AG stat			PT stat		
	1% level	5% level	10% level	1% level	5% level	10% level
			DGP: AR model			
250	0.024	0.093	0.177	0.015	0.071	0.134
500	0.102	0.206	0.303	0.049	0.161	0.235
750	0.151	0.314	0.417	0.099	0.215	0.299
1000	0.212	0.398	0.520	0.116	0.293	0.389
			DGP: SETAR model			
250	0.036	0.118	0.186	0.027	0.089	0.161
500	0.074	0.195	0.295	0.048	0.144	0.218
750	0.171	0.337	0.413	0.099	0.242	0.332
1000	0.220	0.377	0.480	0.126	0.283	0.383

Notes: The table shows the power of AG statistics (10) and PT statistics (13) based on asymptotic critical value, respectively. The data are generated from AR model (22) and SETAR model (23). All results are based on 1000 Monte Carlo simulation.

Table 5: Empirical proportion of rejections (power).

T	AG* stat			PT* stat		
	1% level	5% level	10% level	1% level	5% level	10% level
			$l = 1$			
250	0.061	0.145	0.240	0.043	0.118	0.188
500	0.118	0.298	0.424	0.097	0.227	0.336
750	0.213	0.416	0.547	0.151	0.320	0.442
1000	0.269	0.495	0.604	0.193	0.377	0.513
			$l = 5$			
250	0.083	0.180	0.261	0.055	0.148	0.213
500	0.134	0.319	0.420	0.101	0.247	0.349
750	0.205	0.401	0.541	0.129	0.312	0.450
1000	0.272	0.528	0.663	0.187	0.398	0.513
			$l = 10$			
250	0.079	0.183	0.244	0.057	0.156	0.226
500	0.136	0.307	0.438	0.107	0.245	0.352
750	0.210	0.422	0.556	0.150	0.317	0.439
1000	0.276	0.503	0.632	0.180	0.397	0.524
			$l = 15$			
250	0.105	0.200	0.270	0.073	0.164	0.236
500	0.139	0.300	0.395	0.101	0.233	0.338
750	0.214	0.399	0.521	0.153	0.321	0.448
1000	0.292	0.519	0.656	0.203	0.400	0.514

Notes: The table shows the power of AG^* statistics (16) and PT^* statistics (17) based on bootstrap critical value, respectively. The data are generated from AR model (22). All results are based on 1000 Monte Carlo simulation and $B = 300$ bootstrap replications.

Table 6: Empirical proportion of rejections (power).

T	AG* stat			PT* stat		
	1% level	5% level	10% level	1% level	5% level	10% level
			$l = 1$			
250	0.067	0.150	0.242	0.045	0.114	0.192
500	0.128	0.274	0.363	0.072	0.203	0.306
750	0.225	0.413	0.511	0.139	0.317	0.436
1000	0.284	0.494	0.629	0.178	0.384	0.512
			$l = 5$			
250	0.084	0.173	0.254	0.046	0.126	0.208
500	0.113	0.253	0.380	0.068	0.194	0.298
750	0.211	0.393	0.515	0.134	0.291	0.381
1000	0.276	0.484	0.623	0.189	0.371	0.480
			$l = 10$			
250	0.083	0.187	0.256	0.055	0.128	0.202
500	0.148	0.318	0.405	0.096	0.237	0.351
750	0.214	0.405	0.498	0.138	0.294	0.391
1000	0.296	0.525	0.656	0.212	0.422	0.551
			$l = 15$			
250	0.108	0.205	0.284	0.070	0.152	0.226
500	0.165	0.329	0.421	0.114	0.235	0.341
750	0.201	0.385	0.513	0.133	0.289	0.423
1000	0.268	0.455	0.583	0.175	0.369	0.492

Notes: The table shows the power of AG^* statistics (16) and PT^* statistics (17) based on bootstrap critical value, respectively. The data are generated from SETAR model (23). All results are based on 1000 Monte Carlo simulation and $B = 300$ bootstrap replications.

Table 7: Forecasting performances of individual models evaluated by out-of-sample R-square.

	CC		MRS		TVP	
	R^2_{oos}	p-value	R^2_{oos}	p-value	R^2_{oos}	p-value
DY	0.033	0.070	-0.421	0.135	-1.295	0.452
EP	-0.424	0.342	-0.658	0.359	-2.825	0.613
DP	-0.643	0.386	0.025	0.177	-2.486	0.456
SVOL	0.078	0.098	-0.093	0.057	-2.272	0.598
BM	-1.130	0.792	-1.992	0.776	-4.036	0.691
NTIS	-0.701	0.374	-0.453	0.283	-2.472	0.443
TBL	-0.956	0.043	-1.451	0.032	-2.603	0.189
LTY	-0.978	0.084	-1.517	0.087	-4.933	0.466
LTR	-0.020	0.032	-3.051	0.329	-1.249	0.124
DFY	-0.823	0.739	-0.643	0.379	-3.560	0.325
DFR	-0.605	0.574	-1.433	0.594	-3.963	0.752
INFL	-0.354	0.419	-0.434	0.466	-1.490	0.501

Notes: This table provides the R^2_{oos} of stock return forecasts from individual models which takes a macroeconomic variable as predictor. The values of R^2_{oos} are calculated as the percent reduction of mean squared predictive error (MSPE) of model of interest relative to the MSPE of the benchmark model of historical average. p-values are for Clark and West (2006) test for the null hypothesis that the MSPE of the model of interest is higher than or equal to the benchmark model.

Table 8: Forecasting performances of combination strategies evaluated by out-of-sample R-square.

	CC		MRS		TVP	
	R^2_{oos}	p-value	R^2_{oos}	p-value	R^2_{oos}	p-value
MFC	0.821	0.005	0.745	0.088	-0.988	0.441
TMC	0.967	0.004	0.690	0.102	-0.904	0.423
DMSPE(1)	0.802	0.007	0.716	0.095	-1.014	0.450
DMSPE(0.9)	0.817	0.012	0.708	0.102	-1.026	0.443

Notes: This table provides the R^2_{oos} of stock return forecasts from combination strategies over 12 individual models. The values of R^2_{oos} are calculated as the percent reduction of mean squared predictive error (MSPE) of the combination strategy of interest relative to the MSPE of the benchmark model of historical average. p-values are for Clark and West (2006) test for the null hypothesis that the MSPE of the strategy of interest is higher than or equal to the benchmark model.

Table 9: Forecasting performances of individual models evaluated by economic value.

	CC			MRS			TVP		
	return	p-value $l=5$	p-value $l=15$	return	p-value $l=5$	p-value $l=15$	return	p-value $l=5$	p-value $l=15$
DY	0.016	0.707	0.657	0.067	0.540	0.557	-2.840	0.963	0.947
EP	3.625	0.417	0.357	3.666	0.350	0.347	1.933	0.670	0.630
DP	2.198	0.617	0.630	3.089	0.413	0.510	2.710	0.393	0.377
SVOL	3.696	0.453	0.467	1.677	0.887	0.917	2.525	0.353	0.380
BM	3.930	0.500	0.500	4.148	0.377	0.400	-2.968	0.957	0.983
NTIS	3.266	0.933	0.957	2.621	0.873	0.930	2.894	0.273	0.287
TBL	2.862	0.250	0.230	2.624	0.257	0.243	4.230	0.157	0.137
LTY	3.108	0.153	0.147	2.088	0.277	0.270	1.709	0.473	0.447
LTR	6.710	0.023	0.003	5.018	0.097	0.080	2.459	0.383	0.387
DFY	3.558	0.910	0.930	2.709	0.657	0.797	0.263	0.710	0.710
DFR	5.017	0.062	0.050	3.347	0.483	0.497	1.332	0.697	0.703
INFL	3.684	0.357	0.367	3.909	0.320	0.323	0.410	0.807	0.833

Notes: This table reports the annualized average returns of the buy/sell trading rules constructed based on the sign of return forecasts obtained from individual predictive regressions. We use the newly bootstrapped version of AG statistic to test for the null hypothesis of no return predictability. p-values are obtained based on 1000 block bootstraps with the block lengths of 5 and 15 ($l = 5, 15$).

Table 10: Forecasting performances of combination strategies evaluated by economic value.

	CC			MRS			TVP		
	return	p-value $l=5$	p-value $l=15$	return	p-value $l=5$	p-value $l=15$	return	p-value $l=5$	p-value $l=15$
MFC	3.616	0.970	0.907	3.710	0.520	0.473	3.216	0.350	0.360
TMC	3.544	0.830	0.897	3.488	0.547	0.557	3.186	0.343	0.427
DMSPE(1)	3.616	0.973	0.943	3.710	0.467	0.397	3.145	0.410	0.370
DMSPE(0.9)	3.544	0.803	0.877	3.790	0.447	0.457	2.588	0.487	0.503

Notes: This table reports the annualized average returns of the buy/sell trading rules constructed based on the sign of return forecasts obtained from combination strategies. We use the newly bootstrapped version of AG statistic to test for the null hypothesis of no return predictability. p-values are obtained based on 1000 block bootstraps with the block lengths of 5 and 15 ($l = 5, 15$).

Appendix A. Proof

This appendix gives some lemmas for the proof of 1 and 2.

Lemma 1. *Draw independently each block on* $[0, 1, \cdots, T-l]$ *with probability* $1/(T-l+1)$, *and let* $T = bl$ *where* b *is the number of blocks,* l *denotes the length of each block, then*

$$E^*[A_T^*] = \frac{1}{T}\sum_{t=1}^{T} sign(\hat{y}_t)y_t + O_{p^*}(l/T) \tag{A.1}$$

$$E^*[B_T^*] = \big(\frac{1}{T}\sum_{t=1}^{T} sign(\hat{y}_t)\big)\big(\frac{1}{T}\sum_{t}^{T} y_t\big) + O_{p^*}(l/T) \tag{A.2}$$

$$E^*[\tilde{A}_T^*] = \frac{1}{T}\sum_{t=1}^{T} sign(\hat{y}_t)sign(y_t) + O_{p^*}(l/T) \tag{A.3}$$

$$E^*[\tilde{B}_T^*] = \big(\frac{1}{T}\sum_{t=1}^{T} sign(\hat{y}_t)\big)\big(\frac{1}{T}\sum_{t=1}^{T} sign(y_t)\big) + O_{p^*}(l/T) \tag{A.4}$$

where $E^*[\cdot]$ *means conditional expectation given* $\omega_1, \cdots, \omega_T$.

Firstly, we present the proof of (A.1).

$$\begin{aligned} E^*[A_T^*] &= E^*[\frac{1}{T}\sum_{t=1}^{T}\text{sign}(\hat{y}_t^*)y_t^*] \\ &= E^*[\frac{1}{bl}\sum_{i=1}^{b}\sum_{j=1}^{l}\text{sign}(\hat{y}_{s_i+j})y_{s_i+j}] \\ &= E^*[\frac{1}{l}\sum_{j=1}^{l}\text{sign}(\hat{y}_{s_1+j})y_{s_1+j}] \end{aligned} \tag{A.5}$$

since $s_i, i = 1, \cdots, b$ are independent uniform, they are *i.i.d.* uniform. Thus (A.5) can be rewritten as

$$\begin{aligned} & \frac{1}{l}(\text{sign}(\hat{y}_1)y_1 + \text{sign}(\hat{y}_2)y_2 + \cdots + \text{sign}(\hat{y}_l)y_l)P(s_1 = 0) \\ + \ & \frac{1}{l}(\text{sign}(\hat{y}_2)y_2 + \text{sign}(\hat{y}_3)y_3 + \cdots + \text{sign}(\hat{y}_{l+1})y_{l+1})P(s_1 = 1) \\ + \ & \vdots \\ + \ & \frac{1}{l}(\text{sign}(\hat{y}_{bl-l+1})y_{bl-l+1} + \text{sign}(\hat{y}_{bl-l+2})y_{bl-l+2} + \cdots + \text{sign}(\hat{y}_{bl})y_{bl})P(s_1 = T-l), \end{aligned} \tag{A.6}$$

and $P(s_1 = 0) = P(s_1 = 1) = \cdots = P(s_1 = T - l) = \frac{1}{T-l+1}$.

Note that for $l + 1 \leq t \leq T - l$ we obtain $l \times \text{sign}(\hat{y}_t)y_t$ summands, while we have only 1 $\text{sign}(\hat{y}_t)y_t$ and $\text{sign}(\hat{y}_{bl})y_{bl}$, 2 $\text{sign}(\hat{y}_2)y_2$ and $\text{sign}(\hat{y}_{bl-1})y_{bl-1}$ and so on. To sum up the terms in (A.6)

$$\begin{aligned} & \frac{1}{T-l+1}\sum_{t=l+1}^{T-l}\text{sign}(\hat{y}_t)y_t + O_{p^*}(l/T) \qquad \text{(A.7)} \\ = \ & \frac{1}{T}\sum_{t=1}^{T}\text{sign}(\hat{y}_t)y_t + O_{p^*}(l/T) \qquad \blacksquare \end{aligned}$$

Similarly, we also show A.2-A.4 hold, and the proof is thus omitted.

Lemma 2. *Under the null hypthesis $\mathbb{H}_0$ hold, then*

$$|Var^*[A_T^*] - Var[A_T]| \xrightarrow{p^*} 0 \qquad \text{(A.8)}$$

$$|Var^*[B_T^*] - Var[B_T]| \xrightarrow{p^*} 0 \qquad \text{(A.9)}$$

$$|Cov^*[A_T^*, B_T^*] - Cov[A_T, B_T]| \xrightarrow{p^*} 0 \qquad \text{(A.10)}$$

$$|Var^*[\tilde{A}_T^*] - Var[\tilde{A}_T]| \xrightarrow{p^*} 0 \qquad \text{(A.11)}$$

$$|Var^*[\tilde{B}_T^*] - Var[\tilde{B}_T]| \xrightarrow{p^*} 0 \qquad \text{(A.12)}$$

$$|Cov^*[\tilde{A}_T^*, \tilde{B}_T^*] - Cov[\tilde{A}_T, \tilde{B}_T]| \xrightarrow{p^*} 0 \qquad \text{(A.13)}$$

where $Var^*[\cdot]$ and $Cov^*[\cdot]$ are the conditional variance and covariance given $\omega_1, \cdots, \omega_T$, respectively.

The proof of (A.8):

$$\begin{aligned} Var^*[A_T^*] &= Var^*[\frac{1}{T}\sum_{t=1}^{T}\text{sign}(\hat{y}_t^*)y_t^*] \\ &= E^*[(\frac{1}{T}\sum_{t=1}^{T}\text{sign}(\hat{y}_t^*)y_t^*)^2] - (E^*[\frac{1}{T}\sum_{t=1}^{T}\text{sign}(\hat{y}_t^*)y_t^*])^2 \\ &= E^*[(\frac{1}{bl}\sum_{i=1}^{b}\sum_{j=1}^{l}\text{sign}(\hat{y}_{s_i+j})y_{s_i+j})^2] - (E^*[\frac{1}{bl}\sum_{i=1}^{b}\sum_{j=1}^{l}\text{sign}(\hat{y}_{s_i+j})y_{s_i+j}])^2 \\ &\overset{H_0}{=} \frac{1}{T}\overline{\text{sign}(\hat{y})^2y^2} - \frac{1}{T}\overline{[\text{sign}(\hat{y})]}^2\overline{y}^2 + o_{p^*}(1) \\ &= Var[A_T] + o_{p^*}(1). \qquad \blacksquare \qquad \text{(A.14)} \end{aligned}$$

Thus, $|Var^*[A_T^*] - Var[A_T]| \xrightarrow{p^*} 0$ as desired.

$$\begin{aligned}
Var^*[B_T^*] &= E^*[(\frac{1}{T}\sum_{t=1}^{T}\text{sign}(\hat{y}_t^*))^2(\frac{1}{T}\sum_{t=1}^{T}y_t^*)^2] - (E^*[(\frac{1}{T}\sum_{t=1}^{T}\text{sign}(\hat{y}_t^*))(\frac{1}{T}\sum_{t=1}^{T}y_t^*)])^2 \\
&\overset{H_0}{=} E^*[(\frac{1}{bl}\sum_{i=1}^{b}\sum_{j=1}^{l}\text{sign}(\hat{y}_{s_i+j}))^2]E^*[(\frac{1}{bl}\sum_{i=1}^{b}\sum_{j=1}^{l}y_{s_i+j})^2] \\
&\quad -(E^*[\frac{1}{bl}\sum_{i=1}^{b}\sum_{j=1}^{l}\text{sign}(\hat{y}_{s_i+j})])^2(E^*[\frac{1}{bl}\sum_{i=1}^{b}\sum_{j=1}^{l}y_{s_i+j}])^2 \\
&= [\frac{1}{T}\overline{\text{sign}(\hat{y})^2} + \frac{T-1}{T}\overline{\text{sign}(\hat{y})}^2][\frac{1}{T}\overline{y^2} + \frac{T-1}{T}\overline{y}^2] - (\overline{\text{sign}(\hat{y})})^2\overline{y}^2 + o_{p^*}(1) \\
&= Var[B_T] + o_{p^*}(1). \qquad \blacksquare
\end{aligned} \tag{A.15}$$

Therefore, $|Var^*[B_T^*] - Var[B_T]| \xrightarrow{p^*} 0$ as desired. And

$$\begin{aligned}
Cov^*[A_T^*, B_T^*] &= E^*[(\frac{1}{T}\sum_{t=1}^{T}\text{sign}(\hat{y}_t^*)y_t^*)(\frac{1}{T}\sum_{t=1}^{T}\text{sign}(\hat{y}_t^*))(\frac{1}{T}\sum_{t=1}^{T}y_t^*)] \\
&\quad -E^*[(\frac{1}{T}\sum_{t=1}^{T}\text{sign}(\hat{y}_t^*)y_t^*)]E^*[(\frac{1}{T}\sum_{t=1}^{T}\text{sign}(\hat{y}_t^*))(\frac{1}{T}\sum_{t=1}^{T}y_t^*)] \\
&= E^*[(\frac{1}{bl}\sum_{i=1}^{b}\sum_{j=1}^{l}\text{sign}(\hat{y}_{s_i+j})y_{s_i+j})(\frac{1}{bl}\sum_{i=1}^{b}\sum_{j=1}^{l}\text{sign}(\hat{y}_{s_i+j}))(\frac{1}{bl}\sum_{i=1}^{b}\sum_{j=1}^{l}y_{s_i+j})] \\
&\quad -E^*[(\frac{1}{bl})\sum_{i=1}^{b}\sum_{j=1}^{l}\text{sign}(\hat{y}_{s_i+j})y_{s_i+j}]E^*[(\frac{1}{bl}\sum_{i=1}^{b}\sum_{j=1}^{l}\text{sign}(y)_{s_i+j})(\frac{1}{bl}\sum_{i=1}^{b}\sum_{j=1}^{l}y_{s_i+j})] \\
&\overset{H_0}{=} [\frac{1}{T}\overline{\text{sign}(\hat{y})^2} + \frac{T-1}{T}\overline{\text{sign}(\hat{y})}^2][\frac{1}{T}\overline{y^2} + \frac{T-1}{T}\overline{y}^2] - (\overline{\text{sign}(\hat{y})})^2\overline{y}^2 + o_{p^*}(1) \\
&= Cov[A_T, B_T] + o_{p^*}(1). \qquad \blacksquare
\end{aligned} \tag{A.16}$$

Then, the proof of equation (A.10) is finished. Similarly, (A.11)-(A.13) can be proven and is omitted here.

Corollary 1. *Under the conditions of Lemma 2, then*

$$|\hat{V}_{AG}^* - \hat{V}_{AG}| \xrightarrow{p^*} 0 \tag{A.17}$$

$$|\hat{V}_{PT}^* - \hat{V}_{PT}| \xrightarrow{p^*} 0 \tag{A.18}$$

where $\hat{V}_{AG}^*$ and $\hat{V}_{PT}^*$ are defined in (18) and (19), respectively.

Finally, using the Lemma 1 and Corollary 1, we can construct the Hausman-type statistics as

$$AG^* \equiv \frac{A_T^* - B_T^* - (A_T - B_T)}{\sqrt{\hat{V}_{AG}^*}} \to \mathcal{N}(0,1), \tag{A.19}$$

$$PT^* = \frac{\tilde{A}_T^* - \tilde{B}_T^* - (\tilde{A}_T - \tilde{B}_T)}{\sqrt{\hat{V}_{PT}^*}} \to \mathcal{N}(0,1). \quad \blacksquare \tag{A.20}$$

So, the proof of Theorem 1 and 2 is accomplished.