Ensemble Machine Learning and Stock Return Predictability*

Ben Jacobsen[1], Fuwei Jiang[2], and Hongwei Zhang [1,2]

[1]Tilburg University – TIAS School for Business and Society

[2]Central University of Finance and Economics – School of Finance

First Draft: March 2018

This Draft: September 2019

*We acknowledge the very helpful comments and suggestions from David Rapach, Guofu Zhou, Allen Timmerman, Dacheng Xiu, Michael Halling, Jan R. Magnus, Frans de Roon, and the conference and seminar participants at Tilburg University, Central University of Finance and Economics, Tsinghua university, Peking University, Renmin University, Xiamen University, Zhejiang University, University of Nottingham Ningbo China, 2018 Asian Financial Association Meetings, 2018 SoFiE Summer School Machine Learning and Finance, 2018 Conference on Financial Predictability and Data Science. This article is supported by the National Natural Science Foundation of China (No. 71602198, 71572052), Beijing Natural Science Foundation (No. 9174045), and the Program for Innovation Research in Central University of Finance and Economics. Send correspondence to Hongwei Zhang, TIAS School for Business and Society, Tilburg University; e-mail: zhanghongwei@tsinghua.org.cn.

Ensemble Machine Learning and Stock Return Predictability

Abstract

Many, even sophisticated, models cannot beat a simple mean combination of univariate stock market return forecasts. We introduce an ensemble machine learning method, which averages forecasts from sophisticated models (like BMA, WALS and LASSO) based on random subsamples and which learns from its mistakes by adaptively changing sampling distributions. Empirically, our novel method improves the simple mean forecast with statistically significant monthly out-of-sample R^2_{OS} of around 2-3% and annual utility gains around 3%. Our approach benefits from predicting well in volatile periods and especially from extreme market drops. The forecasting gains of our new method stem from improved diversity among individual forecasts. We obtain similar gains in forecasting accuracy when we use our method to predict macro economic variables.

JEL classifications: G17, G12, G02, C58

Keywords: Equity Premium Prediction, Machine Learning, Forecast Combination, Parameter Uncertainty, Diversification

1 Introduction

Out-of-sample predictability of equity premia remains a challenge both to academics in asset pricing and investment practitioners (see, Rapach and Zhou, 2013, for a recent survey). There are now many sophisticated forecast models that can be applied in stock return forecasting, such as the popular Bayesian model averaging (BMA), the famous LASSO introduced by Tibshirani (1996) and the weighted-average least squares (WALS) technique newly proposed by Magnus et al. (2010). However, to our knowledge, there is no empirical evidence that these sophisticated models can outperform a simple mean combination of univariate variables proposed by Rapach et al. (2010) . This is surprising, since one might expect model uncertainty[1] to be high in this environment and these sophisticated models could therefore perform well. One possible explanation is that, as is the case with any highly parameterized model, they may suffer from overfitting and parameter uncertainty (estimation error) when applying to stock return forecasting where the number of variables N is usually large relative to the sample size T. This suggests that the stock return forecasting problem seems a natural candidate for the application of bootstrap aggregation (bagging) methods that have already achieved great success in many real-world tasks (see, Breiman, 1996; Zhou, 2012).

Bagging is a machine learning ensemble meta-method designed to reduce the out of-sample mean squared forecast error (MSFE) of forecast models suffering from overfitting by reducing variance. It is an equal-weighted average of forecast models estimated using subsamples randomly drawn with replacement from the current available observations of the original forecasting problem. Although a number of recent studies suggest that bagging is a promising tool for improving forecast accuracy of economic variables such as inflation and employment growth (e.g., Inoue et al., 2008; Rapach and Strauss, 2010), to date few attempts have been made to apply bagging in stock return forecasting. The theoretic explanation is that, just as Inoue et al. (2008) pointed out, bagging cannot work well when the degree of predictability is very low, as is the case in forecasting stock returns.

We empirically confirm that standard bagging has little use in the context of stock return forecasting but introduce a variant of bagging - which we call adaptive bagging (AdaBagging) - and demonstrate that it can substantially improve forecast models we investigate. Like bagging, AdaBagging also combines a set of models estimated using random subsamples drawn from the current samples to a more accurate one. However, AdaBagging is adaptive in the sense that it

[1]Model uncertainty refers to the situation in which a forecaster knows neither the best model specification nor its corresponding parameter values.

adaptively changes the distribution of the samples based on the performance of previous models by using the sequential sampling mechanism borrowed from AdaBoost (Freund and Schapire, 1997).(AdaBoost is another representative method of ensemble learning, but is mainly applied in classification problems instead of regressions.)[2] The intuition behind our AdaBagging method is straightforward. AdaBagging learns from its initial mistakes by systematically assigning more weight in the re-estimation procedure on the error reduction of its mistakes. We show how this adaption works well in stock return forecasting using the standard data sets from previous studies and helps improve the forecast of the sophisticated models beyond the simple mean forecast. We also show its robustness by forecasting numerous factor portfolios. And to illustrate the general applicability of AdaBagging we show how it improves forecast of a number of economic variables.

Why does AdaBagging work so well? Based on the bias-variance-covariance decomposition (Ueda and Nakano, 1996), which is an extension of the bias-variance decomposition and can be dated back to Markowitz (1952)'s Modern Portfolio Theory (Brown et al., 2005), we explore how AdaBagging works and show that it achieves a better bias-variance trade-off by generating diversity successfully. We find that it is surprisingly resistant to overfitting in terms of the MSFE. Moreover, our results indicate that out-of-sample gains for the forecasts of sophisticated models with AdaBagging are mainly concentrated in extreme periods of stock returns, especially during extreme market downturns.

Our first contribution is to provide the novel ensemble learning method AdaBagging to stock return forecasting, and to study the behavior of sophisticated models built on our method. The combination of AdaBagging and these sophisticated models is designed to deal with parameter estimation error and model uncertainty simultaneously. We present empirical evidence regarding the usefulness of AdaBagging for out-of-sample forecasting of stock returns. Since parameter estimation error is a common challenge of financial econometrics, our ideas not only work for economic and financial forecasting, or forecasting in general, but can be applied to many other finance areas, such as portfolio optimization. The paper probably most closely related to ours is Rossi (2018). This paper also employs ensemble learning methods such as boosting and bagging, to forecast stock returns. However, Rossi (2018) experiments are based on regression trees, which are nonlinear models. We focus on linear models.

Specifically, to test the effectiveness of AdaBagging on improving stock return forecasts, we consider three representative sophisticated models. The first one is LASSO, which performs both variable selection and shrinkage in order to enhance the prediction accuracy and interpretability

[2]Our experiments also confirm that, the standard AdaBoost method cannot work well in the context of stock return forecasting. For brevity, these results are not reported but they are available upon request.

of the model it produces.[3] The other two are forecast combination models: BMA and WALS.[4] WALS has turned out to be an effective approach for dealing with model uncertainty. It has two advantages over BMA: It requires no prior and the computational time is linear in the number of dependent variables rather than exponential. This last point is particularly relevant in forecasting stock markets, where many variables can be in play.

Based on AdaBagging, we comprehensively compare the performance of BMA, LASSO and WALS with the simple mean combination forecast model and the standard multivariate predictive regression model called Kitchen sink. Dating back to the work of Bates and Granger (1969), it is generally believed that a simple equal-weighted combination forecast, is hard to beat, which is also known as the forecast combination puzzle (e.g., Smith and Wallis, 2009; Claeskens et al., 2016). We compare these techniques in terms of forecast accuracy and economic gains using returns on the S&P 500 index. To make our results comparable with the literature, following Rapach et al. (2010), we use updated monthly data from Welch and Goyal (2008) over the period from 1926:12 to 2016:12. This data set also includes the usual suspects, such as the dividend–price ratio, the dividend yield, the Treasury bill rate, inflation, and the term spread.

When estimated traditionally, without taking parameter uncertainty into account, we can basically confirm the results in the current literature that the simple mean combination forecast is superior. None of Kitchen sink, LASSO, BMA and WALS can outperform the historical average, indicating they seriously suffer from overfitting and parameter uncertainty. And there is not sufficient evidence to prefer WALS over LASSO and BMA in terms of the MSFE. The results change we use the scheme that incorporate the AdaBagging method with estimation windows given by the user ex-ante or determined by the data. All of the sophisticated models we investigate including Kitchen sink can significantly beat the simple mean combination forecast in terms of both the MSFE and utility gains. For example, the AdaBagging–WALS (AdaBagging–Kitchen sink) with 360-month rolling windows can generate statistically significant monthly out-of-sample R^2_{OS} of 2.87% (2.80%) and annual utility gain of 3.47% (3.58%), when the learning round number of AdaBagging are determined by data. Our results show that, contrary to the finding of Rapach et al. (2010), the predictability of forecast models with AdaBagging remains strong in both NBER-dated business-cycle recessions (bad times) and expansions (good times). Furthermore, our results show that AdaBagging is surprisingly resistant to overfitting in terms of the MSFE, even when the

[3]Efron et al. (2004) show there is a very close mathematical relationship between LASSO and boosting. But this issue is far beyond the scope of this paper.

[4]Forecast combination is a popular method for dealing with model uncertainty. The idea underlying forecast combination is that, instead of selecting the true or best possible model, we construct a combination forecast based on a weighted average over all candidate models.

number of learning rounds becomes very large (such as 500 or 1000) in our experiments. However, we also find that, in terms of utility gains, AdaBagging eventually overfits with the increase of the number of learning rounds. Meanwhile, we confirm Inoue et al. (2008)'s asymptotic analysis, which is that bagging cannot work well in stock return forecasting.

The second contribution is that, on the basis of bias-variance-covariance decomposition, we explore the work mechanism of AdaBagging and provide statistical explanations on how AdaBagging works in the context of forecasting stock returns. Our results show that the key to the success of AdaBagging in improving stock return forecasts is its excellent ability of generating diversity. Using sequential sampling with repeatedly reweighted data distribution like AdaBoost, instead of bootstrap sampling like bagging, AdaBagging achieves diversity successfully. Our work sheds light on the role of diversification in improving forecast accuracy. The extensive research results of diversity generating techniques in ensemble learning methods may help provide further insights into research on stock return forecasting.

Last but not least, we show our methodology is robust and can be applied in many cases. For instance, we provide evidence that, with the same set of economic variables, AdaBagging-WALS can significantly forecast many well-known macroeconomic variables, such as the Chicago Fed National Activity Index, Smoothed U.S. Recession Probabilities, Industrial Production Growth, the Output Gap and the Civilian Unemployment Rate.

2 The Equity Premium Prediction Problem and Ensemble Learning

2.1 Traditional Predictive Regression Models

The problem of forecasting stock returns is an instance of what is known as supervised learning in machine learning. The standard predictive regression model: Kitchen sink can be expressed as

$$r_{t+1} = \alpha + X_t\beta + e_{t+1}, \quad t = 0, 1, ..., T-1 \tag{1}$$

where r_{t+1} is the stock return in excess of the risk-free rate from the end of period t to the end of period $t+1$, α is a constant term, X_t is a $1 \times N$ vector of predictors available at the end of t, the $N \times 1$ vector β is the corresponding parameter vector, e_{t+1} is a zero-mean disturbance term, N is the number of predictors and T is the sample size.

The standard model is simple; however, it depends on two questionable crucial assumptions: (i) The predictor variables are given and fixed, meaning the underlying "best model" is known and

(ii) the model parameters do not change over time. There is abundant empirical evidence showing that both assumptions are invalid (e.g., Goyal and Welch, 2003; Welch and Goyal, 2008).

There are several ways to extend the standard model that allow us to drop either of the two assumptions. We consider three representative sophisticated models, as follows

(1) **Least Absolute Shrinkage and Selection Operator (LASSO)**: a popular regularization method for performing shrinkage in regressions, introduced by Tibshirani (1996). It can improve the prediction accuracy and interpretability of regression models by altering the model fitting process to select only a subset of the provided variables for use in the final model rather than using all of them.

(2) **Bayesian Model Averaging (BMA)**: a popular forecast combination method for dealing with model uncertainty. The idea underlying BMA is that, instead of selecting the true or best possible model, we construct a combination forecast based on a weighted average over all candidate models, where the weights of candidate models are determined by their posterior probabilities.

(3) **Weighted-Average Least Squares (WALS)**: a new forecast combination method, introduced by Magnus et al. (2010), which is a Bayesian combination of frequentist estimators and possesses both computational and theoretical advantages over frequentist and Bayesian methods. The computational advantage is that its computing time is linear in the number of predictor variables rather than exponential, as in Bayesian model averaging techniques. The theoretical advantage is that, in contrast to standard BMA, which is based on normal priors leading to unbounded risk (prediction variance), WALS is based on reflected Weibull, Subbotin, or Laplace priors, which imply a coherent treatment of ignorance and can generate bounded risk.

For comparison, we also consider the simple mean combination model introduced by Rapach et al. (2010). The detail description of these models are given in Appendix: Forecast Models.

2.2 AdaBagging Forecasts

Ensemble methods in machine learning, such as bagging, boosting and their variants, are meta methods to improve on weak learners. The idea underlying bagging and boosting is that a combination forecast is constructed by averaging forecasts estimated using different subsamples

randomly drawn from the current available observations. The main difference between those methods is how the subsamples are determined.

Bagging is appealing for its simplicity and elegance. The standard bagging procedure, includes two key steps: bootstrapping and aggregation. For the first step (bootstrapping), bootstrap sampling is applied to obtain subsamples from the available observations. In the second step (aggregation), the regression models are first estimated based on subsamples generated in the first step and then these submodels are averaged to obtain an aggregation model. The forecast of this aggregation model is the final prediction given by bagging.

Algorithm 1 AdaBagging Forecasts

Input:

Data set $D = (D_0, D_1, ..., D_{T-1}) = \{r_{t+1}, X_t\}_{t=0}^{T-1}$ of size T
The forecast model $r_{t+1} = F(X_t)$, which can be LASSO, WALS or any other forecast model
Number of learning rounds B

Process:

Initialize the data distribution $P^{(1)}(D_t) = 1/T$, $t = 0, ..., T-1$
for i = 1,...,B
  Resample a data set $D^{(i)}$ of size T with replacement under distribution $P^{(i)}(D)$
  Estimate the forecast model $F^{(i)}$ using data set $D^{(i)}$
  Obtain the forecast value $\hat{r}_{t+1}^{(i)} = F^{(i)}(X_t)$, $t = 0, ..., T-1$
  Compute the forecast error $e_i = \sum_{t=0}^{T-1} P^{(i)}(D_t)(r_{t+1} - \hat{r}_{t+1}^{(i)})^2$
  Compute the model weight $\lambda_i = \frac{1}{2} ln(\frac{1-e_i}{e_i})$
  Update the distribution $P^{(i+1)}(D_t) = P^{(i)}(D_t) exp(-\lambda_i r_{t+1} \hat{r}_{t+1}^{(i)})/Z_i$, $t = 0, ..., T-1$, where Z_i is a normalization factor that enables $P^{(i+1)}(D)$ to be a distribution
end

Output:

$$\hat{r}_{T+1}^{AdaBaging}(D) = \frac{1}{B} \sum_{i=1}^{B} \hat{r}_{T+1}^{(i)} \tag{2}$$

We provide a variant of bagging, called adaptive bagging, based on the sequential sampling mechanism borrowed from AdaBoost. AdaBoost is the representative method in boosting method family, which usually is used for cross-sectional classification. Similar to bagging, AdaBoost also combines a set of weak learners to create a strong learner that obtains better performance than a single one. However, there are two key differences between AdaBoost and bagging. One is that

AdaBoost adaptively changes the distribution of the samples based on the performance of previous models but bagging does not. The other is that bagging uses equal weights to combine models but AdaBoost does not.

The implementation of AdaBagging is described in Algorithm 1. It works by estimating a set of base forecast models sequentially based on the repeatedly reweighted data. The predictions from all these base forecast models are then combined through a weighted average to produce the final prediction. Initially, the weights of the data samples are all set to $1/N$ (step 1 in Algorithm 1), so that the first iteration simply estimates a base forecast using the original data (steps 3 and 4 in Algorithm 1). For each successive iteration, the data distribution weights are individually modified (step 8 in Algorithm 1) based on the performance of the previous forecast model (steps 5 to 7 in Algorithm 1) and the learning algorithm is reapplied to the reweighted data (steps 3 and 4 in Algorithm 1). At a given iteration, those samples that were incorrectly predicted by the base forecast model induced in the previous iteration have their weights increased, whereas the weights are decreased for those that were predicted correctly. As the iterations proceed, samples that are difficult to predict obtain ever-increasing influence. The later forecast models are thereby forced to focus more on the mistakes of the earlier forecast models. In recursive forecasting practice, the input dataset D of AdaBagging can be rolling windows specified by the user or optimal windows determined by the data.[5]

2.3 Forecast Evaluation Measures

Following Rapach et al. (2010), we use the out-of-sample R^2 statistic, R^2_{OS}, suggested by Campbell and Thompson (2008) to evaluate the forecast accuracy of different (combination) models. The out-of-sample R^2_{OS} is a convenient statistic for comparing MSFEs and measures the proportional reduction in the MSFE for model $\mathcal{M}$ relative to the benchmark model.

The MSFE is the most popular metric for evaluating forecast accuracy and is computed as

$$MSFE^M = \frac{1}{p}\sum_{t=T+1}^{T+p}(r_t - \hat{r}_t^M)^2 \tag{3}$$

where $[T+1, T+p]$ is the out-of-sample evaluation period and $\hat{r}_t^M$ comes from any forecast model we investigate, such as the Kitchen sink, WALS, and BMA.

[5]Note that we must ensure that the forecast error $e_i < 0.5$ (accordingly $\lambda_i > 0$), which depends on the data set $\{r_{t+1}\}_{t=T_0}^{T-1}$ and the forecast model $r = F(X)$. Otherwise, AdaBagging may not work properly. The data set we will test in Section 3 satisfies this condition; otherwise we can normalize the data to satisfy this condition.

Then, the R^2_{OS} statistic is defined as

$$R^2_{OS} = 1 - \frac{MSFE^M}{MSFE^{bmk}} = 1 - \frac{\frac{1}{p}\sum_{t=T+1}^{T+p}(r_t - \hat{r}^M_t)^2}{\frac{1}{p}\sum_{t=T+1}^{T+p}(r_t - \bar{r}_t)^2} \tag{4}$$

where $MSFE^{bmk}$ is the MSFE of the benchmark model. Following Rapach et al. (2010), we set the benchmark model to the historical average model, which is computed as

$$\bar{r}_{T+1} = \frac{1}{T}\sum_{t=1}^{T} r_t \tag{5}$$

When $R^2_{OS} > 0$, the forecast model $\mathcal{M}$ is more accurate than the benchmark model in terms of the MSFE. The associated p-value is based on the work of Clark and West (2007) to test the null hypothesis that $R^2_{OS} \leq 0$.

Following Campbell and Thompson (2008), Welch and Goyal (2008) and Rapach et al. (2010), among others, we also analyze stock return forecasts with utility gains, which is a profit-based metric and provides more direct measures of the value of forecasts to a mean–variance investor who is more interested in the economic value of a forecast model than its precision.

Assume that a mean–variance investor with relative risk aversion parameter γ will decide at the end of period T to allocate the following share of a portfolio to equities in period $T+1$:

$$w^M_T = \frac{1}{\gamma}\left(\frac{\hat{r}^M_{T+1}}{\hat{\sigma}^2_{T+1}}\right) \tag{6}$$

where $\hat{r}^M_{T+1}$ is the forecast of model $\mathcal{M}$, and following Campbell and Thompson (2008) $\hat{\sigma}^2_{T+1}$ is the five-year rolling-window estimate of the variance of stock returns[6]. In addition, following Rapach et al. (2010) and Campbell and Thompson (2008), we constrain the portfolio weight on stocks to lie between 0% and 150% (inclusive), and set the relative risk aversion parameter γ to three.

The investor then allocates $1 - w^M_T$ of the portfolio to risk-free bills, and the $T+1$ realized portfolio return is,

$$R^M_p = w^M_T \hat{r}^M_{T+1} + r^f_{T+1} \tag{7}$$

where r^f_{T+1} is the risk-free rate. Then, the investor realizes a certainty equivalent return (CER) of

[6]The results are qualitatively similar for a ten-year moving window suggested by Rapach et al. (2016).

the portfolio formed using the model M,

$$CER_M = \hat{u}_M - \frac{1}{2}\gamma(\hat{\sigma}_M^2) \tag{8}$$

where $\hat{u}_M$ and $\hat{\sigma}_M^2$ are the sample mean and variance, respectively, for the investor's portfolio over the evaluation period. Finally, in the case of monthly data, we obtain the utility (CER) gain in annualized percentage return as

$$\triangle(ann\%) = 1200 * (CER_M - CER_{bmk}) \tag{9}$$

where CER_{bmk} is the CER of the portfolio formed using the benchmark model.

As pointed out by Rapach et al. (2010), the utility gain can be interpreted as the portfolio management fee that an investor would be willing to pay to have access to the additional information available in a forecast model relative to the information in the benchmark model alone.

In addition, we also calculate the monthly Sharpe ratio of the portfolio, which is the mean portfolio return in excess of the risk-free rate divided by the standard deviation of the excess portfolio return. To examine the adverse effect of transaction costs, we also consider the case of 50bps transaction costs, which is generally considered as a relatively high number.

3 Empirical Results

3.1 Data

Following Rapach et al. (2010) and to make our results comparable with previous studies, we use updated monthly data from Welch and Goyal (2008) over the period 1926:12 to 2016:12.[7] Following Elliott et al. (2013) and to avoid multicollinearity when estimating some of the forecast combination models such as WALS, we consider only 12 of the 14 popular economic variables,[8] which can generate a total of $2^{12} = 4096$ candidate models. These 12 variables are as follows.

(1) Log dividend–price ratio [log(DP)]: log of a 12-month moving sum of dividends paid on the S&P 500 index minus the log of stock prices (S&P 500 index).

[7]The data are available at http://www.hec.unil.ch/agoyal/. Variable definitions and data sources are described in more detail by Welch and Goyal (2008).

[8]Following Elliott et al. (2013), we exclude the log dividend–earnings ratio and the long-term yield, because the log dividend earnings ratio is equal to the difference between the log dividend price ratio and the log earnings–price ratio, while the long- term yield is equal to the sum of the term spread and the Treasury bill rate.

(2) Log dividend yield [log(DY)]: log of a 12-month moving sum of dividends minus the log of lagged stock prices.

(3) Log earnings–price ratio [log(EP)]: log of a 12-month moving sum of earnings on the S&P 500 index minus the log of stock prices.

(4) Stock variance (SVAR): monthly sum of squared daily returns on the S&P 500 index.

(5) Book-to-market ratio (BM): book-to-market value ratio for the Dow Jones Industrial Average.

(6) Net equity expansion (NTIS): ratio of a 12-month moving sum of net equity issues by New York Stock Exchange-listed stocks to the total end-of-year market capitalization of the New York Stock Exchange stocks.

(7) Treasury bill rate (TBL): interest rate on a three-month Treasury bill (secondary market).

(8) Long-term return (LTR): return on long-term government bonds.

(9) Term spread (TMS): long-term yield minus the Treasury bill rate.

(10) Default yield spread (DFY): difference between BAA- and AAA-rated corporate bond yields.

(11) Default return spread (DFR): long-term corporate bond return minus the long-term government bond return.

(12) Inflation (INFL): calculated from the Consumer Price Index (all urban consumers).

The stock returns are measured as the difference between the log return on the S&P 500 (including dividends) and the log return on the risk-free Treasury bill. We follow Rapach and Zhou (2013) and divide the total sample into an in-sample period (1926:12–1956:12) and an out-of-sample evaluation period (1957:01–2016:12), including 720 observations. As the authors point out, the 1957:01–2016:12 forecast evaluation period covers most of the postwar era, including the oil price shocks of the 1970s; the deep recession associated with the Volcker disinflation in the early 1980s; the stock price shock in 1987; the long expansions of the 1960s, 1980s, and 1990s; and the recent global financial crisis in 2008 and the concomitant Great Recession. Following Rapach and Zhou (2013), we present not only the results of the full 1957:01–2016:12 forecast evaluation period, but also the results computed separately during National Bureau of Economic Research-dated business cycle expansions and recessions.

Table A.1 in Internet Appendix reports R^2_{OS} statistics for each of the individual predictive regression models relative to the historical average benchmark model. It reinforces the findings of Welch and Goyal (2008) that the individual predictive regression models can not generate reliable out-of-sample forecasts of the equity premium.

3.2 Out-of-sample Forecasting Results under Conventional Estimation Schemes

In this section, we investigate the performance of Kitchen sink, BMA, LASSO and WALS compared to Rapach et al. (2010)'s simple mean combination under traditional estimation schemes.

Table 1 provides the out-of-sample forecasting results of these models estimated using 360-month rolling windows. It shows that, none of the Kitchen sink, BMA, LASSO and WALS model can outperform the historical average model in terms of the MSFE, although they can outperform the Kitchen sink model. Hence, our results confirm the common finding in the literature on the forecast combination, which is, that the simple mean combination forecast is superior.[9] There is not sufficient evidence to prefer WALS over BMA and LASSO in terms of the MSFE. It is worthy pointing out, however, that WALS overwhelmingly beat BMA and LASSO in terms of the computational times. Especially, compared to BMA, WALS can reduce the computation time by 90% in our experiments using 12 dependent variables.[10] Furthermore, as pointed out by Magnus et al. (2010), the computational time of WALS is linear in the number of dependent variables rather than exponential, as in BMA. In stock return forecasting practice where many variables can be in play, this point is particularly relevant.

[Place Table 1 about here]

3.3 Out-of-sample Forecasting Results with AdaBagging

One possible explanation for the relatively poor results under the conventional estimation schemes is that these sophisticated models suffer from overfitting and parameter uncertainty. Now, we further investigate the performance of forecast models based on AdaBagging.

[9] With recursively expanding windows, we also draw a similar conclusion. For brevity, we do not report these results here but they are available upon request.

[10] We use a window 10 based PC with Core(TM) i7-6700 3.40Ghz processors and 32 G of DRAM.

3.3.1 The R^2_{OS} Values with Different Learning Rounds

Table 2 reports the out-of-sample performance of forecast models estimated using AdaBagging with 360-month rolling windows, given the number of learning rounds $B = 100, 200, 300$. It shows that, whatever the number of learning rounds is, all of these sophisticated models including the Kitchen sink model improve their R^2_{OS} dramatically while the simple mean combination only improves slightly. Impressively, for example, the R^2_{OS} value of WALS (Kitchen sink) is 2.26% (2.33%) when $B = 300$, which is significantly larger than that of the simple mean combination. It further shows that, surprisingly, these sophisticated models can obtain positive R^2_{OS} values in the period of NBER-dated business-cycle expansions. Although the simple mean combination is stable, its space for improving seems seriously limited. It does not benefit substantially from AdaBagging, suggesting it suffers from underfitting.

Table 2 also indicates that, compared to the case of $B = 100$, the R^2_{OS} values of all sophisticated models (BMA, LASSO and WALS) do increase substantially and then decrease slowly with the increase of B. However, the simple mean combination forecast suffers from the increase of the number of learning rounds of AdaBagging.

[Place Table 2 about here]

Figure 1 plots the performance (in terms of the R^2_{OS} values) of Kitchen sink, BMA, LASSO, WALS and the simple mean combination forecast model based on AdaBagging with 360-month rolling windows, against the number of learning rounds. It confirms that, in terms of the R^2_{OS} values, WALS achieves improvement substantially and rapidly with the increase of the number of learning rounds when $B \leq 200$, after that R^2_{OS} converges to a stable value of about $2.1\% \sim 2.3\%$. Our untabulated results indicate that, after $B > 600$, WALS begins overfit very slowly, but even given a large number of learning rounds such as $B = 1000$, the R^2_{OS} of WALS can still reach the value of 1.46, higher than that of $B = 100$. The figure also shows that while LASSO improves more rapidly than WALS, it also tends to overfit more rapidly than WALS. Meanwhile, the standard Kitchen sink model improves most slowly, but can achieve highest R^2_{OS} values than all other models we investigate when the number of learning rounds becomes large enough. Figure 1 also shows, as noted before, that the simple mean combination forecast suffers from the increase of B, confirming that it shrinks too much and suffers from underfitting. Last but not least, Figure 1 suggests a preference for WALS (and Kitchen sink) over BMA and LASSO because of its resistance to overfitting.

[Place Figure 1 about here]

3.3.2 What is the Optimal Number of Learning Rounds? Results with Early Stopping

We next investigate how to determine the optimal number of learning rounds via early stopping, which is a technique used to avoid overfitting and quicken the learning speed when training a model with an iterative method. Following Prechelt (2012), to dynamically determine the optimal learning rounds in recursive forecasting practice, we use a holdout-based early stopping strategy as follows

(1) Split the current available data sample into a training set of size, for example, 360 months and a validation set of size 120 months if possible, otherwise set the optimal learning rounds as 100.

(2) Train on the training set and evaluate the average error on the validation set once in a while after every fifty learning rounds.

(3) Stop training as soon as the error on the validation set is higher than it was the last time it was checked.

(4) Use the number of learning rounds of the previous step as the optimal number of learning rounds.

Table 3 reports the out-of-sample performance of Kitchen sink, BMA, LASSO, WALS and the simple mean combination forecast model estimated using AdaBagging with early stopping. It shows that, in terms of R^2_{OS}, all the models including the simple mean combination achieve some improvement compared to the case of the number of learning rounds $B = 100$, conforming the effectiveness of early stopping on avoiding overfitting. Specifically, the R^2_{OS} values of WALS and Kitchen sink are 2.87% and 2.80% respectively, which are surprisingly high. Furthermore, both WALS and Kitchen sink can always achieve a quite high positive R^2_{OS} value no matter what macroeconomic conditions are, indicating the predictability of stock returns in both good times and bad times.

[Place Table 3 about here]

Overall, our results indicate that AdaBagging achieves a good results and seems resistant to overfitting in terms of the R^2_{OS} values. It seems that AdaBagging benefits from the sequential sampling mechanism borrowed from AdaBoost that is famous for its resistance to overfitting (Zhou, 2012).

For comparison, Table 4 reports the results of bagging with 360-month rolling windows, given the number of learning rounds $B = 100, 200, 300$. It shows that, compared to the rolling windows scheme, bagging alone cannot substantially improve, and may even slightly deteriorate, the performance of forecast models we investigate. Our results of bagging confirm the asymptotic analysis of Inoue et al. (2008) that bagging cannot improve on forecasts when the degree of predictability is low, as tends to be the case in stock return forecasting.

[Place Table 4 about here]

3.4 Asset Allocation Results with AdaBagging

We further examine the economic value of the stock return predictability of different forecast models from an asset allocation perspective. Following Campbell and Thompson (2008) among others, we compute the certainty equivalent return (CER) gain and Sharpe ratio for a mean-variance investor who optimally allocates across equities and the risk-free asset using the out-of-sample predictive regression forecasts.

Table 5 reports the asset allocation results of AdaBagging with 360 months rolling windows, for different numbers of learning rounds $B = 100, 200, 300$. It indicates that the CER values of all forecast models we investigate decrease monotonically with an increase in the number of learning rounds B. However, it also shows that, in terms of both the CER gains and Sharpe ratio, Kitchen sink, BMA, LASSO and WALS substantially outperform the simple mean combination. For example, given the number of learning rounds $B = 100$, the CER gains of WALS equals 3.73% which is about 3 or 6 times higher than that of the simple mean combination when assuming no transaction costs. Even in the case of high transaction costs of 50bps, WALS still achieves a positive CER gain of 2.15%, while the CER gain of the simple mean combination forecast becomes negative.

[Place Table 5 about here]

Figure 2 confirms that the CER values of all forecast models we investigate decrease monotonically with an increase in the number of learning rounds. When B is larger than 100, the CER values of Kitchen sink, BMA, LASSO and WALS tend to decrease rapidly. Still, the CER gain of WALS when $B = 500$ reaches 2.67%, four times larger than the CER gain of the simple mean combination model. Figure 2 also provides evidence which indicates that WALS (and Kitchen sink) is preferred over BMA and LASSO.

[Place Figure 2 about here]

Table 6 reports the asset allocation results of AdaBagging with optimal learning rounds determined by our holdout-based early stopping rule. Table 6 shows that all forecast models we investigate, even in the case of 50bps transaction costs, achieve quite high positive CER values, regardless whether the risk aversion coefficient is 3 or 5. This confirms the ability of early stopping to avoid overfitting.

[Place Table 6 about here]

To summarize, our results show that, based on AdaBagging learning method, sophisticated models (such as Kitchen sink, BMA, LASSO and WALS) strongly beat the historical average benchmark and the simple mean combination in terms of both the MSFE and CER gains. Their predictability remains strong in both good times and bad times. One possible explanation for their excellent performance is that the AdaBagging machine learning method successfully handles the problems of parameter uncertainty and overfitting which seriously impair these sophisticated models return predictability.

3.5 Robust analysis of AdaBagging

To verify the robustness of AdaBagging, we further investigate the performance of AdaBagging with estimation windows determined by data, instead of specified by the user ex-ante.[11] We use the cross-validation method, developed by Pesaran and Timmermann (2007), to select optimal estimation windows.

Assuming forecast model is subject to one or more breaks with unknown break dates, the optimal estimation window size W^* is determined as follows:

$$W^* = \arg \min_{W=\underline{W},\ldots,(T-W_{cv})} \frac{\sum_{\tau=T-W_{cv}+1}^{T}(y_\tau - \hat{r}_\tau(W))^2}{W_{cv}}$$

where $\underline{W}$ is the minimum estimation window size, W_{cv} is the cross-validation window size, $\hat{r}_\tau(W)$ is the forecast value of a given model, estimated using the window of size W, at time τ (Pesaran and Timmermann, 2007, P11-12). Following Pesaran and Timmermann (2007), we set the cross-validation window size W_{cv} at 25% of T. And we set the minimum window size $\underline{W} = 240$.

[11] We thank Michael Halling for suggesting discussing this issue.

Figure 3 plots the performance (in terms of the R^2_{OS} values) of forecast models estimated using AdaBagging with estimation windows determined by data, against the number of learning rounds. It shows that, with the increase of the number of learning rounds, all sophisticated models we investigate including Kitchen sink can achieve improvement substantially and converge to a stable R^2_{OS} value of above 1.0%. However, it also indicates the so-called optimal windows are not reliable enough. A more sophisticated window selection method seems to be needed.

[Place Figure 3 about here]

4 How AdaBagging Helps? Accuracy-Diversity Trade-off

The fact that AdaBagging, which averages the same models but estimated over subsamples randomly drawn from the available observations, can improve performance might seem counterintuitive at first. It is however easy to see how it works. Via forecast encompassing tests, we illustrate that AdaBagging incorporates useful forecasting information from increasing learning rounds. Based on bias-variance-covariance decomposition (Ueda and Nakano, 1996), we demonstrate that, by introducing randomness, AdaBagging reduces the average covariance of individual forecasts and stabilizes the variance of the combination forecast, thereby improving forecasting performance.

4.1 Forecast Encompassing Test Results

To further assess the information content of AdaBagging relative to alternative schemes such as the traditional rolling windows or AdaBagging with different numbers of learning rounds, following Rapach et al. (2010) among others, we conduct a forecast encompassing test. Harvey et al. (1998) develop the MHLN statistic for testing the null hypothesis that a given forecast contains all of the relevant information found in a competing forecast (i.e., the given forecast encompasses the competitor) against the alternative that the competing forecast contains relevant information beyond that in the given forecast.

Table 7 reports p-values of the Harvey et al. (1998)'s MHLN statistic for Kitchen sink, BMA, LASSO, WALS and the simple mean combination forecast model estimated using 360 months rolling windows, and AdaBagging with 360 months rolling windows, given the number of learning rounds $B = 50, 100, 200$. Each entry in the table corresponds to a one-sided (upper-tail) test of the null hypothesis that the forecast under the scheme given in the column heading encompasses the

forecast under the scheme given in the row heading against the alternative hypothesis that the forecast under the scheme given in the column heading does not encompass the forecast under the scheme given in the row heading.

The first column of Panel A of Table 7 shows, not surprisingly, the forecast of BMA based on rolling windows fails to encompass the forecasts of AdaBagging–BMA with different numbers of learning rounds. The second column shows that AdaBagging–BMA with the number of learning rounds $B = 50$ significantly encompasses the traditional rolling windows scheme, indicating that AdaBagging contains incremental forecasting information beyond the traditional rolling windows scheme. The three and four columns further indicate that AdaBagging–BMA with a larger B encompasses AdaBagging–BMA with a smaller B, suggesting potential gains from increasing the number of learning rounds to fully make use of the relevant information. From Panel B, C and D, we can also draw a similar conclusion for AdaBagging–LASSO, AdaBagging–WALS and AdaBagging–Kitchen sink.

Panel E shows that AdaBagging–SMC can always encompass the simple mean combination forecast estimated using rolling windows. However, it cannot benefit obviously from the increase of learning rounds, indicating that the simple mean combination forecast model suffers from underfitting, which limits its potential in out-of-sample forecasting.

[Place Table 7 about here]

4.2 Explanations based on Bias-Variance-Covariance Decomposition

AdaBagging per se is a forecast combination. Intuitively, the method is analogous to including additional assets in a portfolio to reduce the portfolios variance (see, e.g. Dunis et al., 2001; Timmermann, 2006; Rapach et al., 2010, among others). With AdaBagging combining individual forecasts helps to reduce forecast variability when they are diverse.

Generating diverse individual forecasts, however, is not easy. The major obstacle lies in the fact that these individual forecasts are estimated for the same task from the same data based on the same type of models, and thus they are usually highly correlated. Another challenge of generating diverse individual forecasts is that the individual forecasts must not be very poor, otherwise their combination would not improve and could even worsen performance. So, it is desired that the individual forecasts should be accurate and diverse. The key to success of forecast combination lies in achieving a good trade-off between individual performance and diversity. It is known as the accuracy-diversity trade-off in the machine learning literature on ensemble methods (for more

details, see e.g. Zhou, 2012). One way to understand the accuracy-diversity trade-off is through the bias-variance-covariance decomposition (Ueda and Nakano, 1996)[12].

Looking back to the forecasts of AdaBagging, $\hat{r}^{AdaBagging} = \frac{1}{B}\sum_{i=1}^{B}\hat{r}^{(i)}$, which is an equal-weighted combination of the individual forecasts from B learning rounds. Following Ueda and Nakano (1996) and Brown et al. (2005), define the averaged bias, averaged variance, and averaged covariance of the individual forecasts of AdaBagging as, respectively,

$$\begin{aligned}
\overline{Bias} &= \frac{1}{B}\sum_{i=1}^{B}(E[\tilde{r}^{(i)}] - r), \\
\overline{Variance} &= \frac{1}{B}\sum_{i=1}^{B}(E[\tilde{r}^{(i)}] - \tilde{r}^{(i)})^2, \\
\overline{Covariance} &= \frac{1}{B(B-1)}\sum_{i=1}^{B}\sum_{j=1,j\neq i}^{B}(E[\tilde{r}^{(i)}] - \tilde{r}^{(i)})(E[\tilde{r}^{(j)}] - \tilde{r}^{(j)}).
\end{aligned} \tag{10}$$

Then, the bias-variance-covariance decomposition of the MSFE of $\hat{r}^{AdaBagging}$ is

$$MSFE = \overline{Bias}^2 + \frac{1}{B}\overline{Variance} + (1 - \frac{1}{B})\overline{Covariance}, \tag{11}$$

which divides the MSFE into three components: averaged squared bias, averaged variance and averaged covariance. It indicates that the MSFE depends heavily on the covariance term, which can be negative while the bias and variance terms are constrained to be positive, especially when B becomes a large number. It is obvious that the more diverse the individual forecasts are, the better.

Figure 4 plots the bias-variance-covariance decomposition of the MSFE of AdaBagging–WALS with 360 months rolling windows, against the number of learning rounds. It demonstrates that, the MSFE decreases dramatically with the increase of the number of learning rounds especially when the number of learning rounds $B \leq 30$. After that, the MSFE (and averaged variance, covariance) gradually decreases. It also clearly shows that the MSFE is mainly determined by the averaged covariance when B becomes larger such as $B \geq 30$. It provides insight to how AdaBagging gains from combination of the individual forecasts estimated over subsamples randomly drawn from the available observations, indicating that AdaBagging achieves diversity

[12]It is interesting to note that this is the first appearance of the decomposition only for the machine learning literature. In fact, just as Brown et al. (2005) pointed out, an equivalent decomposition can be found in Markowitz (1952), which is instrumental for Modern Portfolio Theory.

successfully by introducing randomness based on the repeatedly reweighted data.

However, the introduction of randomness may lead to a reduction in the accuracy of the individual forecasts. Figure 4 clearly shows that, with the increase of the number of learning rounds, the averaged squared bias will increase gradually. But the positive effect of diversity outweighs the negative effects of the lost accuracy of the individual forecasts because, relative to the averaged variance and covariance, the averaged squared bias is so small that it is negligible.[13]

We can also draw a similar conclusion from the bias-variance-covariance decomposition of AdaBagging–BMA, AdaBagging–LASSO and AdaBagging–Kitchen sink. See Panels A, B and C of Figure A.1 in Internet Appendix. Meanwhile, Panel D of Figure A.1 reinforces our argument that the simple mean combination forecast model suffers from underfitting and cannot benefit obviously from the increase of learning rounds.

[Place Figure 4 about here]

5 Forecasting Macro Conditions with the Same Set of 12 Economic Variables

In order to explore the economic sources of equity premium predictability, we investigate whether the stock return predictors also have predictive power for future business conditions. As pointed out by Cochrane (2008), if the return predictor shows predictive power for business cycle, then the predictable return variations are more plausibly related to macroeconomic risk.

In particular, we run the following regressions with AdaBagging–WALS,

$$Y_{t+1} = \alpha + X_t\beta + e_{t+1} \tag{12}$$

$$Y_{t+12} = \alpha + X_t\beta + e_{t+1} \tag{13}$$

$$\Delta Y_{t+1} = \alpha + X_t\beta + e_{t+1} \tag{14}$$

$$\Delta Y_{t+12} = \alpha + X_t\beta + e_{t+1} \tag{15}$$

where X_t is the same 12 economic predictors, Y_{t+1} is the macroeconomic condition for next month, Y_{t+12} is the macroeconomic condition for next year, $\Delta Y_{t+1} = Y_{t+1} - Y_{(t+1)-12}$ is the year change

[13] Our untabulated results discussed in the above subsection indicates that, when the learning rounds becomes large such as $B \geq 600$, the negative effects of the lost accuracy of the individual forecasts will be not negligible. Of course, the performance deterioration in this case may also result from decreased diversity due to the number of the learning rounds becoming large.

in macroeconomic condition for the next month, and $\Delta Y_{t+12} = Y_{t+12} - Y_t$ is the year change in macroeconomic condition for the next year.

We focus on the following macroeconomic conditions Y_t

(1) Chicago Fed National Activity Index (CFNAI). The CFNAI is a monthly index designed to capture economic activty and inflationary presure. The data are download from the Federal Reserve Bank of Chicago. Data spans from 1967:03 to 2017:12, monthly.

(2) Smoothed U.S. Recession Probabilities (SRP). Smoothed recession probabilities for the United States are obtained from a dynamic-factor markov-switching model applied to four monthly coincident variables: non-farm payroll employment, the index of industrial production, real personal income excluding transfer payments, and real manufacturing and trade sales. The data are download from the Federal Reserve Bank of St. Louis. Data spans from 1967:06 to 2017:12, monthly.

(3) Industrial Production Growth (IPG). The production growth rate data are also obtained from the Federal Reserve Bank of St. Louis. Data spans from 1919:01 to 2017:12, monthly.

(4) Macroeconomic Uncertainty Index (MU). MU is the macroeconomic uncertainty index proposed in Jurado et al. (2015), which is constructed as a common component of the unpredictable variation of macroeconomic variables. It can be download from Prof. Ludvigson's website. Data spans from 1960:07 to 2017:12, monthly.

(5) Output Gap (Gap): The output gap is the difference between actual GDP or actual output and potential GDP. Both of them can be obtained from Federal Reserve Bank of St. Louis. Data spans from 1968:01 to 2017:12, monthly.

(6) Cay: Cay, introduced by Lettau and Ludvigson (2001a,b), is a cointegrating residual between log consumption, log asset (nonhuman) wealth, and log labor income. It is proved to have striking forecasting power for excess returns on aggregate stock market indexes. It can be download from Prof. Ludvigson's website. Data spans from 1952:Q1 to 2017:Q3, quartly. We tranform it to monthly data spanning from 1952:01 to 2017:12.

(7) Civilian Unemployment Rate (UNRATE). The UNRATE is also obtained from Federal Reserve Bank of St. Louis. Data spans from 1948:01 to 2017:12, monthly.

Table 8 reports the R^2_{OS} values of forecasts of AdaBagging–WALS with early stopping for the above macroeconomic condition variables. It shows that, not surprisingly, all the out-of-sample

R^2_{OS} values are significantly positive, indicating that the 12 economic predictors also have strong predictive power for future business conditions.

[Place Table 8 about here]

6 Forecasting Extreme Market Movements

Fama and French (1989) and Campbell and Cochrane (1999); Cochrane (2007) argue that heightened risk aversion during economic downturns demands a higher risk premium, thereby generating equity premium predictability. Following Rapach et al. (2010), we examine the R^2_{OS} statistics computed separately during NBER-dated business-cycle recessions (bad times) and expansions (good times). Our results in Table 3 show that, however, the predictive ability of sophisticated models with AdaBagging remains strong in both bad times and good times.

We first investigate when and by how much AdaBagging helps these sophisticated models to outperform the simple mean combination. Following Goyal and Welch (2003) and Welch and Goyal (2008), Figure 5 depicts the relative performance of different sophisticated models over time in terms of the DCSFE that is defined as the difference in the cumulative squared forecast error of the historical average benchmark model and the given model. The larger the DCSFE value, the better the model's performance. These models are estimated using 360-month rolling windows or AdaBagging with the number of learning rounds as $B = 100, 200, 300$.

Panel A of Figure 5 illustrates that, when estimated using 360-month rolling windows, the Kitchen sink model almost always underperform the simple mean combination. However, when estimated using AdaBagging with the number of learning rounds $B = 100$, it begins to outperform the simple mean combination during the subperiod 1975:01–2008:01, but still fail to outperform it during the global financial crisis in 2008. With the increase of the number of learning rounds, such as when $B = 200$ or 300, it begins to consistently outperform the simple mean combination over the entire evaluation period since 1970s, demonstrating the success of AdaBagging–Kitchen sink in forecasting the extreme market movements during the global financial crisis in 2008. Panel B of Figure 5 shows that, when estimated using AdaBagging, Kitchen sink achieves a substantial performance leap during the October 1987 stock market crash period, further implying the predictive ability of AdaBagging–Kitchen sink to forecast extreme market movements. We can draw a similar conclusion for all other models we investigate, see Panels C and D of Figure 5 and Figure A.2 in Internet Appendix.

[Place Figure 5 about here]

In order to confirm the economic sources of equity premium predictability, we further investigate forecasting gains of different models, estimated using rolling windows or AdaBagging with early-stopping, during extreme (downturn) periods of normalized monthly returns r_t. For brevity, we only report the results of sophisticated models (BMA, LASSO and WALS), compared to the simple mean combination forecast model. A similar conclusion can be drawn from the results of the standard Kitchen sink model.

Panel A of Table 9 reports the results for four extreme periods, defined as $|r_t| \geq 0.5, 1.0, 1.5, 2.0$ respectively. It shows that, whichever estimation scheme we use, the R^2_{OS} statistics of all sophisticated models are always higher during extreme periods than during less extreme periods. For example, when estimated using AdaBagging, the R^2_{OS} statistics are approximately three times higher during the extreme period when $|r_t| \geq 2.0$ than during the extreme period when $|r_t| \geq 0.5$. Panel A also indicates that sophisticated models always beat the simple mean combination with much higher R^2_{OS} values during extreme periods.

Panel B reports the results for four extreme downturn periods, defined as $r_t \leq -0.5$, -1.0, -1.5, -2.0 respectively. It shows that the R^2_{OS} statistics of all models are always much higher during extreme downturn periods than during the corresponding extreme periods that also include extreme upturn periods. For example, the R^2_{OS} statistics of sophisticated models with AdaBagging are approximately three times higher during the extreme downturn period when $r_t \leq -0.5$ than during the extreme period when $|r_t| \geq 0.5$. In fact, our untabulated results further indicate that out-of-sample gains for forecasts of all models we investigate are mainly concentrated in downturn periods defined as $r_t < 0$, especially extreme downturn periods. It also shows that the differences of the R^2_{OS} statistics of all sophisticated models and the simple mean combination become larger during extreme downturn periods than during extreme periods, demonstrating the superior of sophisticated models during extreme downturn periods.

Table 9 shows that, compared to conventional rolling window schemes, AdaBagging substantially improve forecasts of all models we investigate during extreme (downturn) periods except a few extreme outliers.[14] It seems that forecasts of sophisticated models estimated using AdaBagging fluctuate less than when estimated traditionally. Overall, Table 9 points to enhanced forecasting gains for sophisticated models relative to the simple mean combination during extreme periods, especially extreme downturn periods, in agreement with the Fama and French (1989)

[14] In fact, our untabulated results indicate that, compared to conventional estimation schemes, AdaBagging can also substantially improve forecasts during normal periods.

and Campbell and Cochrane (1999); Cochrane (2007) view that heightened risk aversion during economic downturns requires a higher risk premium.

[Place Table 9 about here]

7 Conclusions

We provide a novel ensemble machine learning method AdaBagging, which is a variant of bagging, to improve on forecast models suffering from overfitting and parameter uncertainty. Based on AdaBagging, we compare several representative sophisticated models that accommodate model uncertainty, such as the recently proposed WALS, the popular BMA and the famous LASSO, with the standard Kitchen sink model and the simple mean combination in the context of stock return forecasting.

Our empirical results confirm, when estimated using conventional rolling window schemes, the common finding in the literature that the simple mean combination forecast is superior. None of the forecast models we investigate can outperform the historical average and the simple mean combination forecast, indicating they seriously suffer from overfitting and parameter uncertainty. However, under the schemes that incorporate the AdaBagging learning method with estimation windows given by the user ex-ante or determined by the data, all of the forecast models we investigate can significantly outperform the simple mean combination forecast in terms of both the MSFE and utility gains. For example, the AdaBagging–WALS (AdaBagging–Kitchen sink) model with early stopping generates statistically significant monthly out-of-sample R^2_{OS} of 2.87% (2.80%) and annual utility gain of 3.47% (3.58%).

Furthermore, our results show that, with the same set of economic variables, AdaBagging-WALS can significantly forecast many well-known macroeconomic variables, such as Chicago Fed National Activity Index, Smoothed U.S. Recession Probabilities, Output Gap, Civilian Unemployment Rate. Our results also indicate that out-of-sample gains for the forecasts of sophisticated models with AdaBagging are mainly concentrated in extreme periods, especially during extreme market downturns.

On the basis of bias-variance-covariance decomposition, we further investigate how and why AdaBagging works in the context of forecasting stock returns. Our results show that the key to the success of AdaBagging is its excellent ability of generating diversity. Through diversification, without demanding more data, AdaBagging help sophisticated models reduce estimation error

dramatically. However, the simple mean combination forecast cannot benefit obviously from AdaBagging, because it shrinks too much and suffers from underfitting. Furthermore, our results show that AdaBagging is surprisingly resistant to overfitting in terms of the MSFE. However, we also find that, in terms of utility gains, AdaBagging eventually overfits with the increase of the number of learning rounds. Because the bias-variance-covariance decomposition can be dated back to Modern Portfolio Theory, the benefit of the AdaBagging learning method can be viewed as theory-based, stemming from diversification.

Our work sheds light on the role of diversification in improving forecast accuracy. The extensive research results of diversity generating techniques in ensemble learning methods may provide further insights into stock return forecasting.

Appendix: Forecast Models

Simple Mean Combination

Rapach et al. (2010)'s simple mean combination forecast can be expressed as,

$$\hat{r}_{T+1}^{SMC} = \frac{1}{N}\sum_{i=1}^{N}\hat{r}_{T+1}^{(i)} \tag{A.1}$$

where $\hat{r}_{T+1}^{(i)}$ is the forecast result of the univariate regression model that only includes the i-th variable of X and the constant term.

Bayesian Model Averaging

Let $\lambda = (\lambda_1, ..., \lambda_m)$ be a vector of model weights with $0 \leq \lambda_i \leq 1$ and $\sum_{i=1}^{m}\lambda_i = 1$ where $m = 2^N$ is the number of candidate models and λ_i is the weight of model $\mathcal{M}_i$. A general forecast combination can be expressed as

$$\hat{r}_{T+1}^{C} = \sum_{i=1}^{m}\lambda_i\hat{r}_{T+1}^{(i)} \tag{A.2}$$

where $\hat{r}_{T+1}^{(i)}$ is the forecast result of model $\mathcal{M}_i$. A natural framework for model combination that has proved useful is Bayesian model averaging, where the weights of candidate models are determined by their posterior probabilities.

Let $P(\mathcal{M}_i)$ denote the prior probability that $\mathcal{M}_i$ is a true model and $P(R|\mathcal{M}_i)$ denote the marginal likelihood of R in model $\mathcal{M}_i$. Then, the BMA weight for $\mathcal{M}_i$ is given by

$$\lambda_i^{BMA} = P(\mathcal{M}_i|R) = \frac{P(\mathcal{M}_i)P(R|\mathcal{M}_i)}{\sum_{j=1}^{2^N}P(\mathcal{M}_j)P(R|\mathcal{M}_j)}, \quad i = 1, 2, ..., 2^N \tag{A.3}$$

Under the assumption of equal model priors and diffuse model priors on parameters, Buckland et al. (1997) provide a good approximation for the BMA weights, Equation (A.3), expressed as

$$\lambda_i^{BMA} = exp(-\frac{1}{2}BIC_i)/\sum_{j=1}^{2^N}exp(-\frac{1}{2}BIC_j), \quad i = 1, 2, ..., 2^N \tag{A.4}$$

where $BIC_i = Tlog(\hat{\sigma}_i^2) + log(T)(N_i)$ is the Bayesian information criterion (BIC) for model $\mathcal{M}_i$,

$\hat{\sigma}_i^2$ is the estimate of σ^2 in $\mathcal{M}_i$ and N_i is the actual number of predictors of $\mathcal{M}_i$.

Then, based on Equations (A.2) and (A.4), the BMA forecast model can be expressed as

$$\hat{r}_{T+1}^{BMA} = \sum_{i=1}^{2^N} \lambda_i^{BMA} \hat{r}_{T+1}^{(i)} \tag{A.5}$$

The BMA technique is straightforward and has proved useful (see, e.g. Avramov, 2002; Hoogerheide et al., 2010). However, it has several problems in practice. The most important is that the computation time will increase exponentially with the increase of N. The second is that, since the priors are based on the normal distribution, they can lead to unbounded prediction variance.

Weighted-Average Least Squares

Weighted-average least squares, introduced by Magnus et al. (2010), is a Bayesian combination of frequentist estimators and possesses both computational and theoretical advantages over frequentist and Bayesian methods. The computational advantage is that its computing time is linear in the number of predictor variables rather than exponential, as in Bayesian model averaging techniques. The theoretical advantage is that, in contrast to standard BMA, which is based on normal priors leading to unbounded risk, WALS is based on reflected Weibull, Subbotin, or Laplace priors, which imply a coherent treatment of ignorance and can generate bounded risk.

Rewrite the standard predictive regression model, Equation (1), in matrix form as

$$R = X_1\beta_1 + X_2\beta_2 + e \tag{A.6}$$

where X_1 is a $T \times (N_1 + 1)$ vector of "focus" predictors that must be included in the model based on theoretical or other grounds and always includes a constant term as the first variable, X_2 is a $T \times N_2$ vector of "auxiliary" predictors that may or may not be included in the model.

Let P be an orthogonal matrix and Λ a diagonal matrix with positive diagonal elements such that $P'X_2'M_1X_2P = \Lambda$, where

$$M_1 = I_T - X_1(X_1'X_1)^{-1}X_1' \tag{A.7}$$

and I_T represents a $T \times T$ identity matrix.

Define the transformed auxiliary variables and parameter as

$$X_2^* = X_2 P \Lambda^{-1/2}, \qquad \beta_2^* = \Lambda^{1/2} P' \beta_2 \tag{A.8}$$

such that $X_2^* \beta_2^* = X_2 \beta_2$.

Then we can rewrite Equation (A.6) equivalently as

$$R = X_1 \beta_1 + X_2^* \beta_2^* + \mathbf{e} \tag{A.9}$$

This transformation is called semi-orthogonal because the new matrix $(X_1 : X_2^*)$ is semi-orthogonal in the sense that $X_2^* M_1 X_2^{*'} = I_{N_2}$.

The equivalence theorem (Magnus et al., 2010, 2016) tells us that the WALS estimators of β_1 and $\beta_2^* = (\beta_{2,1}^*, \beta_{2,2}^*, ..., \beta_{2,N_2}^*)$ in Equation (A.9) can be expressed as, respectively,

$$\hat{\beta}_1 = \hat{\beta}_{1r} - Q^* \hat{\beta}_2^* \tag{A.10}$$

$$\hat{\beta}_{2,h}^* = \hat{\beta}_{2u,h}^* - \hat{\sigma}_h \frac{A_1(\frac{\hat{\beta}_{2u,h}^*}{\hat{\sigma}_h})}{A_0(\frac{\hat{\beta}_{2u,h}^*}{\hat{\sigma}_h})}, \quad (h = 1, ..., N_2) \tag{A.11}$$

where

$$\begin{aligned}
\hat{\beta}_{1r} &= (X_1' X_1)^{-1} X_1' R \\
Q^* &= (X_1' X_1)^{-1} X_1' X_2^* \\
\hat{\beta}_{2u}^* &= X_2^{*'} M_1 R \\
A_j(x) &= \int_{-\infty}^{\infty} (x - \gamma)^j \phi(x - \gamma) \pi(\gamma) d\gamma, \quad (j = 0, 1)
\end{aligned}$$

, ϕ denotes the standard normal density, and $\pi(\gamma)$ is the reflected Weibbull prior[15], and $\hat{\sigma}_h$ is the estimate of the standard deviation of $\beta_{2,h}^*$.

Based on Equations (A.9), (A.10) and (A.11), we can easily obtain the WALS forecast, as

[15]Following Magnus and Luca (2016), the reflected Weibull prior is preferred, which is defined as

$$\pi(\gamma) = \frac{qc}{2} |\gamma|^{-(1-q)} exp^{-c|\gamma|^q}$$

, where $q = 0.8876$, $c = log2$.

follows:

$$\hat{r}_{T+1}^{WALS} = X_{1,T}\hat{\beta}_1 + X_{2,T}^{*}\hat{\beta}_2^{*} \tag{A.12}$$

The WALS theory is appealing and has turned out to be an effective approach for dealing with model uncertainty. For example, Magnus et al. (2016) compare it with four competing predictors (unrestricted model, pretesting, ridge regression, MMA) in a wide range of simulation experiments. They find that the WALS predictor generally produces the lowest mean squared error. They also find that the estimated variance of the WALS predictor is typically larger than the variance of the pretesting and ridge predictor but more accurate in terms of the root mean squared error. Finally, when model uncertainty increases, the dominance of WALS becomes more pronounced.

LASSO

Least absolute shrinkage and selection operator (LASSO), introduced by Tibshirani (1996), is a popular regularization method for performing shrinkage in regressions. It can improve the prediction accuracy and interpretability of regression models by altering the model fitting process to select only a subset of the provided variables for use in the final model rather than using all of them. For the standard OLS model, Equation (1), the objective of LASSO can be expressed as

$$\underset{\alpha\in\mathbb{R},\beta\in\mathbb{R}^N}{\operatorname{argmin}} \left(\frac{1}{T}\sum_{t=0}^{T-1}(r_{t+1}-\alpha-X_t\beta)^2 + \lambda\sum_{i=1}^{N}|\beta_i|\right) \tag{A.13}$$

where λ is a regularization parameter. We use a 5-fold cross-validation to choose the optimal λ with minimum mean square error. The first component in parentheses is the familiar sum of squared residuals, so that the objective function reduces to that for OLS when $\lambda = 0$. The second component is an ι_1 penalty term that shrinks the slope coefficient estimates to prevent overfitting.

References

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

Bates, J. M., Granger, C. W. J., 1969. The combination of forecasts. Journal of the Operational Research Society 20 (4), 451–468.

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

Brown, G., Wyatt, J. L., Tio, P., 2005. Managing diversity in regression ensembles. Journal of Machine Learning Research 6 (1), 1621–1650.

Buckland, S. T., Burnham, K. P., Augustin, N. H., 1997. Model selection: An integral part of inference. Biometrics 53 (2), 603–618.

Campbell, J. Y., Cochrane, J. H., 1999. By force of habit: A consumptionbased explanation of aggregate stock market behavior. Journal of Political Economy 107 (2), 205–251.

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

Claeskens, G., Magnus, J. R., Vasnev, A. L., Wang, W., 2016. The forecast combination puzzle: A simple theoretical explanation. International Journal of Forecasting 32 (3), 754–762.

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

Cochrane, J. H., 2007. Financial markets and the real economy. in R. Mehra (ed.), Handbook of the Equity Premium. Amsterdam: Elsevier.

Cochrane, J. H., 2008. Chapter 7 financial markets and the real economy. Handbook of the Equity Risk Premium (January), 237–325.

Dunis, C., Moody, J., Timmermann, A., 2001. Developments in forecast combination and portfolio choice. International Journal of Forecasting 18 (3), 462–463.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32 (2), 407–499.

Elliott, G., Gargano, A., Timmermann, A., 2013. Complete subset regressions. Journal of Econometrics 177 (2), 357–373.

Fama, E. F., French, K. R., 1989. Business conditions and expected returns on stocks and bonds. Journal of Financial Economics 25 (1), 23–49.

Freund, Y., Schapire, R. E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55 (1), 119–139.

Goyal, A., Welch, I., 2003. Predicting the equity premium with dividend ratios. Management Science 49 (5), 639–654.

Harvey, D. I., Leybourne, S. J., Newbold, P., 1998. Tests for forecast encompassing. Journal of Business & Economic Statistics 16 (2), 254–259.

Hoogerheide, L., Kleijn, R., Ravazzolo, F., Dijk, H. K. v., Verbeek, M., 2010. Forecast accuracy and economic gains from Bayesian model averaging using time-varying weights. Journal of Forecasting 29 (1-2), 251–269.

Inoue, Atsushi, Kilian, Lutz, 2008. How useful is bagging in forecasting economic time series? A case study of U.S. consumer price inflation. Journal of the American Statistical Association 103 (482), 511–522.

Jurado, K., Ludvigson, S. C., Ng, S., March 2015. Measuring uncertainty. American Economic Review 105 (3), 1177–1216.

Lettau, M., Ludvigson, S., 2001a. Consumption, aggregate wealth, and expected stock returns. Journal of Finance 56 (3), 815–849.

Lettau, M., Ludvigson, S., 2001b. Resurrecting the (c)capm: a cross-sectional test when risk premia are time-varying. Journal of Political Economy 109 (6), 1238–1287.

Magnus, J. R., Luca, G. D., 2016. Weighted-average least squares (WALS): A survey. Journal of Economic Surveys 30 (1), 117–148.

Magnus, J. R., Powell, O., Prfer, P., 2010. A comparison of two model averaging techniques with an application to growth empirics. Journal of Econometrics 154 (2), 139–153.

Magnus, J. R., Wang, W., Zhang, X., 2016. Weighted-average least squares prediction. Econometric Reviews 35 (6), 1040–1074.

Markowitz, H., 1952. Portfolio selection. Journal of Finance 7 (1), 77–91.

Pesaran, M. H., Timmermann, A., 2007. Selection of estimation window in the presence of breaks. Journal of Econometrics 137 (1), 134–161.

Prechelt, L., 2012. Early Stopping — But When? Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 53–67.

Rapach, D., Strauss, J., 2010. Bagging or combining (or both)? An analysis based on forecasting U.S. employment growth. Econometric Reviews 29 (5-6), 511–533.

Rapach, D., Zhou, G., 2013. Forecasting stock returns. In: Handbook of Economic Forecasting. Elsevier B.V., Ch. 6, pp. 328–383.

Rapach, D. E., Ringgenberg, M. C., Zhou, G., 2016. Short interest and aggregate stock returns. Journal of Financial Economics 121 (1), 46–65.

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.

Rossi, A. G., 2018. Predicting stock market returns with machine learning. University of Maryland, Working Paper.

Smith, J., Wallis, K. F., 2009. A simple explanation of the forecast combination puzzle. Oxford Bulletin of Economics and Statistics 71 (3), 331–355.

Tibshirani, R. J., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society.Series B (methodological). Wiley 58, 267–288.

Timmermann, A., 2006. Chapter 4 forecast combinations. Vol. 1 of Handbook of Economic Forecasting. Elsevier, pp. 135 – 196.

Ueda, N., Nakano, R., 1996. Analysis of improvement effect for generalization error of ensemble estimators. In Proceedings of International Conference on Neural Networks,, 90–95.

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

Zhou, Z.-H., 2012. Ensemble Methods: Foundations and Algorithms, 1st Edition. Chapman & Hall/CRC.

Table 1
Performance comparison of sophisticated models estimated conventionally, 1957:01–2016:12

This table provides the out-of-sample performance comparison of forecast models estimated using 360 months rolling windows. These models are the standard multivariate predictive regression model: Kitchen sink, BMA with diffuse prior, LASSO with 5-fold cross-validation, Magnus et al. (2010)'s WALS, and Rapach et al. (2010)'s simple mean combination (SMC). R^2_{OS} measures the percent reduction in mean squared forecast error (MSFE) for the forecast model given in the first column relative to the historical average benchmark forecast. CW-test is the Clark and West (2007) MSFE-adjusted statistic. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively. $R^2_{OS,exp}$ ($R^2_{OS,rec}$) statistics are calculated over NBER-dated business-cycle expansions (recessions). The out-of-sample evaluation period is 1957:01–2016:12.

Model	$R^2_{OS}(\%)$	CW-test	$R^2_{OS,exp}(\%)$	$R^2_{OS,rec}(\%)$	Time consumed (s)
BMA	-2.50	1.09	-2.57	-2.35	1,492
LASSO	-1.34	1.72**	-2.92	2.23	248
WALS	-1.00	2.31**	-1.49	0.12	141
SMC	0.25	1.30*	-0.28	1.47	11
Kitchen sink	-5.69	2.30**	-5.88	-5.25	7

Table 2
Out-of-sample performance with AdaBagging, 1957:01–2016:12

This table reports the out-of-sample performance of various measures of forecast models estimated using AdaBagging with 360 months rolling windows, given the number of learning rounds of AdaBagging $B = 100, 200, 300$. These models are the standard multivariate predictive regression model: Kitchen sink, BMA with diffuse prior, LASSO with 5-fold cross-validation, Magnus et al. (2010)'s WALS, and Rapach et al. (2010)'s simple mean combination (SMC). R^2_{OS} measures the percent reduction in mean squared forecast error (MSFE) for the given forecast model relative to the historical average benchmark forecast. CW-test is the Clark and West (2007) MSFE-adjusted statistic. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively. $R^2_{OS,exp}$ ($R^2_{OS,rec}$) statistics are calculated over NBER-dated business-cycle expansions (recessions). The out-of-sample evaluation period is 1957:01–2016:12.

Learning Rounds	$R^2_{OS}(\%)$	CW-test	$R^2_{OS,exp}(\%)$	$R^2_{OS,rec}(\%)$
	Panel A: BMA			
B=100	1.38	2.34***	0.27	3.88
B=200	1.72	2.38***	0.51	4.44
B=300	1.82	2.47***	0.40	5.03
	Panel B: LASSO			
B=100	1.84	2.48***	0.80	4.19
B=200	2.10	2.46***	0.94	4.71
B=300	2.09	2.44***	0.80	4.99
	Panel C: WALS			
B=100	1.28	2.43***	0.51	3.05
B=200	2.10	2.56***	1.01	4.57
B=300	2.26	2.59***	1.04	5.04
	Panel D: Kitchen sink			
B=100	0.16	2.49***	-0.50	1.66
B=200	1.94	2.65***	0.63	4.90
B=300	2.33	2.66***	0.96	5.41
	Panel E: SMC			
B=100	0.31	1.45*	-0.50	1.66
B=200	0.32	1.48*	-0.66	2.52
B=300	0.29	1.49*	-0.83	2.83

Table 3
Out-of-sample forecasting results of AdaBagging with early stopping, 1957:01–2016:12

This table reports the out-of-sample performance of various measures of different forecast models, including the standard multivariate predictive regression model: Kitchen sink, BMA with diffuse prior, LASSO with 5-fold cross-validation, Magnus et al. (2010)'s WALS, and Rapach et al. (2010)'s simple mean combination (SMC). These models are estimated using AdaBagging with optimal learning rounds determined by holdout-based early stopping (Prechelt, 2012). In the implementation of early stopping, we set the size of training set and validation set as 360-month and 120-month respectively, and stop training as soon as the average error, which is evaluated on the validation set once in a while after every fifty epoch, is higher than it was the last time it was checked. R^2_{OS} measures the percent reduction in mean squared forecast error (MSFE) for the given forecast model relative to the historical average benchmark forecast. CW-test is the Clark and West (2007) MSFE-adjusted statistic. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively. $R^2_{OS,exp}$ ($R^2_{OS,rec}$) statistics are calculated over NBER-dated business-cycle expansions (recessions). $\bar{B}$ given in column six, is the average of optimal learning rounds. The out-of-sample evaluation period is 1957:01–2016:12.

	$R^2_{OS}(\%)$	CW-test	$R^2_{OS,exp}(\%)$	$R^2_{OS,rec}(\%)$	$\bar{B}$
BMA	1.85	2.62***	0.11	5.79	324
LASSO	2.18	2.87***	0.59	5.79	349
WALS	2.87	2.99***	1.65	5.61	361
Kitchen sink	2.80	3.00***	1.11	6.61	352
SMC	0.40	1.57*	-0.63	2.72	199

Table 4
Out-of-sample forecasting results with the standard bagging method, 1957:01–2016:12

This table reports the out-of-sample performance of various measures of forecast models estimated using bagging with 360 months rolling windows, given the number of learning rounds of bagging $B = 100, 200, 300$. These models are the standard multivariate predictive regression model: Kitchen sink, BMA with diffuse prior, LASSO with 5-fold cross-validation, Magnus et al. (2010)'s WALS, and Rapach et al. (2010)'s simple mean combination (SMC). R^2_{OS} measures the percent reduction in mean squared forecast error (MSFE) for the given forecast model relative to the historical average benchmark forecast. CW-test is the Clark and West (2007) MSFE-adjusted statistic. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively. $R^2_{OS,exp}$ ($R^2_{OS,rec}$) statistics are calculated over NBER-dated business-cycle expansions (recessions). The out-of-sample evaluation period is 1957:01–2016:12.

Learning Rounds	$R^2_{OS}(\%)$	CW-test	$R^2_{OS,exp}(\%)$	$R^2_{OS,rec}(\%)$
	Panel A: BMA			
B=100	-1.19	1.80**	-1.16	-1.25
B=200	-1.18	1.77**	-1.32	-0.87
B=300	-1.12	1.80**	-1.24	-0.85
	Panel B: LASSO			
B=100	-1.92	2.00**	-1.25	-3.41
B=200	-1.83	2.01**	-1.41	-2.78
B=300	-1.62	2.05**	-1.29	-2.35
	Panel C: WALS			
B=100	-2.25	2.16**	-1.81	-3.24
B=200	-2.01	2.20**	-1.68	-2.75
B=300	-1.99	2.18**	-1.62	-2.82
	Panel D: Kitchen sink			
B=100	-5.32	2.23**	-4.65	-6.81
B=200	-5.46	2.23**	-5.07	-6.35
B=300	-5.38	2.25**	-4.94	-6.39
	Panel E: SMC			
B=100	0.21	1.21	-0.30	1.35
B=200	0.21	1.23	-0.30	1.38
B=300	0.21	1.22	-0.32	1.41

Table 5
Asset allocation results with AdaBagging, 1957:01–2016:12

This table reports the portfolio performance measures for a mean-variance investor who allocates capital monthly between equities and risk-free bills using the monthly out-of-sample forecast results of the U.S. equity premium based on forecast models estimated using AdaBagging with 360 months rolling windows, given different number of learning rounds. These models are the standard multivariate predictive regression model: Kitchen sink, BMA with diffuse prior, LASSO with 5-fold cross-validation, Magnus et al. (2010)'s WALS, and Rapach et al. (2010)'s simple mean combination (SMC). The utility gain $\triangle(ann\%)$ is the annualized certainty equivalent return gain for the investor with risk aversion coefficient of three. The monthly Sharpe ratio is the mean portfolio return in excess of the risk-free rate divided by its standard deviation. The out-of-sample evaluation period is 1957:01–2016:12.

Learning Rounds	No transaction cost		50bps transaction cost	
	$\triangle(ann\%)$	Sharpe ratio	$\triangle(ann\%)$	Sharpe ratio
	Panel A: BMA			
B=100	3.34	0.17	1.77	0.13
B=200	2.86	0.16	1.46	0.12
B=300	2.33	0.15	1.06	0.12
	Panel B: LASSO			
B=100	3.76	0.17	2.18	0.14
B=200	3.39	0.17	1.89	0.13
B=300	2.96	0.17	1.63	0.13
	Panel C: WALS			
B=100	3.73	0.17	2.15	0.14
B=200	3.52	0.17	2.02	0.14
B=300	3.34	0.17	1.93	0.13
	Panel D: Kitchen sink			
B=100	3.27	0.16	1.64	0.13
B=200	3.52	0.17	2.02	0.14
B=300	3.30	0.17	1.90	0.13
	Panel E: SMC			
B=100	0.55	0.11	-0.12	0.09
B=200	0.47	0.11	-0.17	0.08
B=300	0.38	0.11	-0.20	0.08

Table 6
Asset allocation results of AdaBagging with early stopping, 1957:01–2016:12

This table reports the portfolio performance measures for a mean-variance investor who allocates capital monthly between equites and risk-free bills using the monthly out-of-sample forecast results of the U.S. equity premium based on forecast models estimated using AdaBagging with optimal learning rounds determined by holdout-based early stopping (Prechelt, 2012). In the implementation of early stopping, we set the size of training set and validation set as 360-month and 120-month respectively, and stop training as soon as the average error, which is evaluated on the validation set once in a while after every fifty epoch, is higher than it was the last time it was checked. These models are the standard multivariate predictive regression model: Kitchen sink, BMA with diffuse prior, LASSO with 5-fold cross-validation, Magnus et al. (2010)'s WALS, and Rapach et al. (2010)'s simple mean combination (SMC). The utility gain $\triangle(ann\%)$ is the annualized certainty equivalent return gain for the investor with risk aversion coefficient $\gamma = 3,5$. The monthly Sharpe ratio is the mean portfolio return in excess of the risk-free rate divided by its standard deviation. The out-of-sample evaluation period is 1957:01–2016:12.

Model	Risk Aversion	No transaction cost		50bps transaction cost	
		$\triangle(ann\%)$	Sharpe ratio	$\triangle(ann\%)$	Sharpe ratio
BMA	$\gamma = 3$	2.69	0.16	1.33	0.12
BMA	$\gamma = 5$	2.70	0.15	1.68	0.10
LASSO	$\gamma = 3$	2.93	0.16	1.54	0.13
LASSO	$\gamma = 5$	3.00	0.16	1.89	0.11
WALS	$\gamma = 3$	3.47	0.18	2.05	0.14
WALS	$\gamma = 5$	3.40	0.17	2.20	0.12
Kitchen sink	$\gamma = 3$	3.58	0.18	2.13	0.14
Kitchen sink	$\gamma = 5$	3.40	0.16	2.13	0.12
SMC	$\gamma = 3$	0.92	0.12	0.25	0.10
SMC	$\gamma = 5$	1.52	0.10	1.11	0.08

Table 7
Forecast encompassing test for AdaBagging, MHLN statistic p-values

The table reports p-values of the Harvey et al. (1998) MHLN statistic for combination models estimated using 360-month rolling windows and AdaBagging with 360-month rolling windows, given the number of learning rounds $B = 50, 100, 200$. These models are the standard multivariate predictive regression model: Kitchen sink, BMA with diffuse prior, LASSO with 5-fold cross-validation, Magnus et al. (2010)'s WALS, and Rapach et al. (2010)'s simple mean combination (SMC). The MHLN statistic corresponds to a one-sided (upper-tail) test of the null hypothesis that the forecast under the scheme given in the column heading encompasses the forecast under the scheme given in the row heading against the alternative hypothesis that the forecast under the scheme given in the column heading does not encompass the forecast under the scheme given in the row heading. 0.00 indicates less than 0.005. The MHLN statistic is computed for the entire 1957:01–2016:12 forecast evaluation period.

	Rolling Windows	AdaBagging, B=50	AdaBagging, B=100	AdaBagging, B=200
		Panel A: BMA		
Rolling Windows		0.99	0.99	0.97
AdaBagging, B=50	0.00		0.77	0.65
AdaBagging, B=100	0.00	0.07		0.72
AdaBagging, B=200	0.00	0.05	0.10	
		Panel B: LASSO		
Rolling Windows		0.77	0.93	0.88
AdaBagging, B=50	0.00		0.93	0.77
AdaBagging, B=100	0.00	0.01		0.68
AdaBagging, B=200	0.00	0.01	0.08	
		Panel C: WALS		
Rolling Windows		0.74	0.87	0.75
AdaBagging, B=50	0.02		0.99	0.92
AdaBagging, B=100	0.00	0.00		0.85
AdaBagging, B=200	0.00	0.00	0.03	
		Panel D: Kitchen sink		
Rolling Windows		0.97	0.97	0.92
AdaBoost, B=50	0.00		0.99	0.97
AdaBoost, B=100	0.00	0.00		0.97
AdaBoost, B=200	0.00	0.00	0.00	
		Panel E: SMC		
Rolling Windows		0.76	0.55	0.33
AdaBagging, B=50	0.12		0.35	0.25
AdaBagging, B=100	0.19	0.49		0.30
AdaBagging, B=200	0.21	0.42	0.50	

Table 8
Forecasting macro variables with AdaBagging–WALS

This table reports the R^2_{OS} values of forecasts of AdaBagging–WALS with early stopping and 360-month rolling windows, for a list of macroeconomic condition variables. Y_{t+1} is the macroeconomic condition for next month, Y_{t+12} is the macroeconomic condition for next year, $\Delta Y_{t+1} = Y_{t+1} - Y_{(t+1)-12}$ is the year change in macroeconomic condition for the next month, and $\Delta Y_{t+12} = Y_{t+12} - Y_t$ is the year change in macroeconomic condition for the next year. These macroeconomic condition variables considered include Chicago Fed National Activity Index (CFNAI), Smoothed U.S. Recession Probabilities (SRP), Industrial Production Growth (IPG), Jurado et al. (2015)'s Macroeconomic Uncertainty Index (MU), Output Gap (Gap), Lettau and Ludvigson (2001a,b)'s Cay that is a cointegrating residual between log consumption, log asset (nonhuman) wealth, and log labor income, Civilian Unemployment Rate (UNRATE). R^2_{OS} measures the percent reduction in mean squared forecast error (MSFE) for the given forecast model relative to the historical average benchmark forecast. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively, based on p-values for the Clark and West (2007) MSFE-adjusted statistic for testing the null hypothesis that the historical average MSFE is less than or equal to the predictive regression MSFE. The out-of-sample evaluation period given in column two depends on the data sample available.

	Evaluation Period	Y_{t+1}	$\triangle Y_{t+1}$	Y_{t+12}	$\triangle Y_{t+12}$
CFNAI	2000:01-2016:12	51.72***	17.44***	11.15***	21.36***
SRP	2000:01-2016:12	44.00***	8.88***	8.78***	18.96***
IPG	1957:01-2016:12	94.19***	64.13***	94.61***	25.61***
MU	1995:01-2016:12	18.79***	18.83***	33.91***	15.00***
GAP	1980:01-2016:12	72.53***	35.72***	42.33***	16.41***
Cay	1985:01-2016:12	24.95***	4.43***	9.33***	5.75***
UNRATE	1980:01-2016:12	70.60***	51.04***	55.68***	21.95***

Table 9
Forecasting extreme market movements, 1957:01–2016:12

This table reports the R^2_{OS} values of forecasts of different models, estimated using 360-month rolling windows or AdaBagging with 360-month rolling windows, for extreme (downturn) periods of normalized monthly returns r_t. These models are BMA with diffuse prior, LASSO with 5-fold cross-validation, Magnus et al. (2010)'s WALS, and Rapach et al. (2010)'s simple mean combination (SMC). In the implementation of AdaBagging with holdout-based early stopping (Prechelt, 2012), we set the size of training set and validation set as 360-month and 120-month respectively, and stop training as soon as the average error, which is evaluated on the validation set once in a while after every fifty epoch, is higher than it was the last time it was checked. R^2_{OS} measures the percent reduction in mean squared forecast error (MSFE) for the given forecast model relative to the historical average benchmark forecast. *, **, *** indicate significance at the 10%, 5% and 1% levels, respectively, based on p-values for the Clark and West (2007) MSFE-adjusted statistic for testing the null hypothesis that the historical average MSFE is less than or equal to the predictive regression MSFE. The out-of-sample evaluation period is 1957:01–2016:12.

Periods	Observations	R^2_{OS} (%) of models with AdaBagging				R^2_{OS} (%) of models estimated traditionally			
		BMA	LASSO	WALS	SMC	BMA	LASSO	WALS	SMC
Overall	720 / 720	1.85^{**}	2.18^{***}	2.87^{***}	0.40^{*}	-2.50	-1.99^{**}	-1.00^{**}	0.25^{*}
Panel A: extreme periods of normalized returns r_t									
$\lvert r_t\rvert \geq 0.5$	336 / 720	3.26^{***}	3.72^{***}	4.53^{***}	1.09^{***}	-0.07^{*}	1.38^{**}	2.95^{***}	0.75^{**}
$\lvert r_t\rvert \geq 1.0$	119 / 720	5.30^{***}	5.61^{***}	6.59^{***}	1.83^{***}	0.87	4.04^{**}	4.69^{**}	1.09^{***}
$\lvert r_t\rvert \geq 1.5$	42 / 720	7.30^{***}	7.45^{***}	9.24^{***}	1.75^{**}	1.62	6.10^{**}	7.24^{**}	1.13^{**}
$\lvert r_t\rvert \geq 2.0$	12 / 720	9.95^{***}	10.34^{***}	12.61^{***}	2.47^{**}	9.09^{**}	10.46^{**}	18.87^{***}	2.02^{***}
Panel B: extreme downturn periods of normalized returns r_t									
$r_t \leq -0.5$	167 / 720	11.90^{***}	11.64^{***}	12.06^{***}	6.61^{***}	9.74^{***}	9.85^{***}	9.53^{***}	2.45^{***}
$r_t \leq -1.0$	71 / 720	10.82^{***}	10.94^{***}	11.40^{***}	5.39^{***}	8.40^{***}	8.82^{***}	9.94^{***}	2.11^{***}
$r_t \leq -1.5$	26 / 720	11.80^{***}	11.93^{***}	12.85^{***}	4.35^{***}	7.50^{**}	8.83^{**}	12.52^{**}	1.71^{***}
$r_t \leq -2.0$	8 / 720	13.46^{***}	13.86^{***}	15.76^{***}	4.27^{***}	15.40^{***}	14.67^{***}	24.30^{***}	2.35^{***}

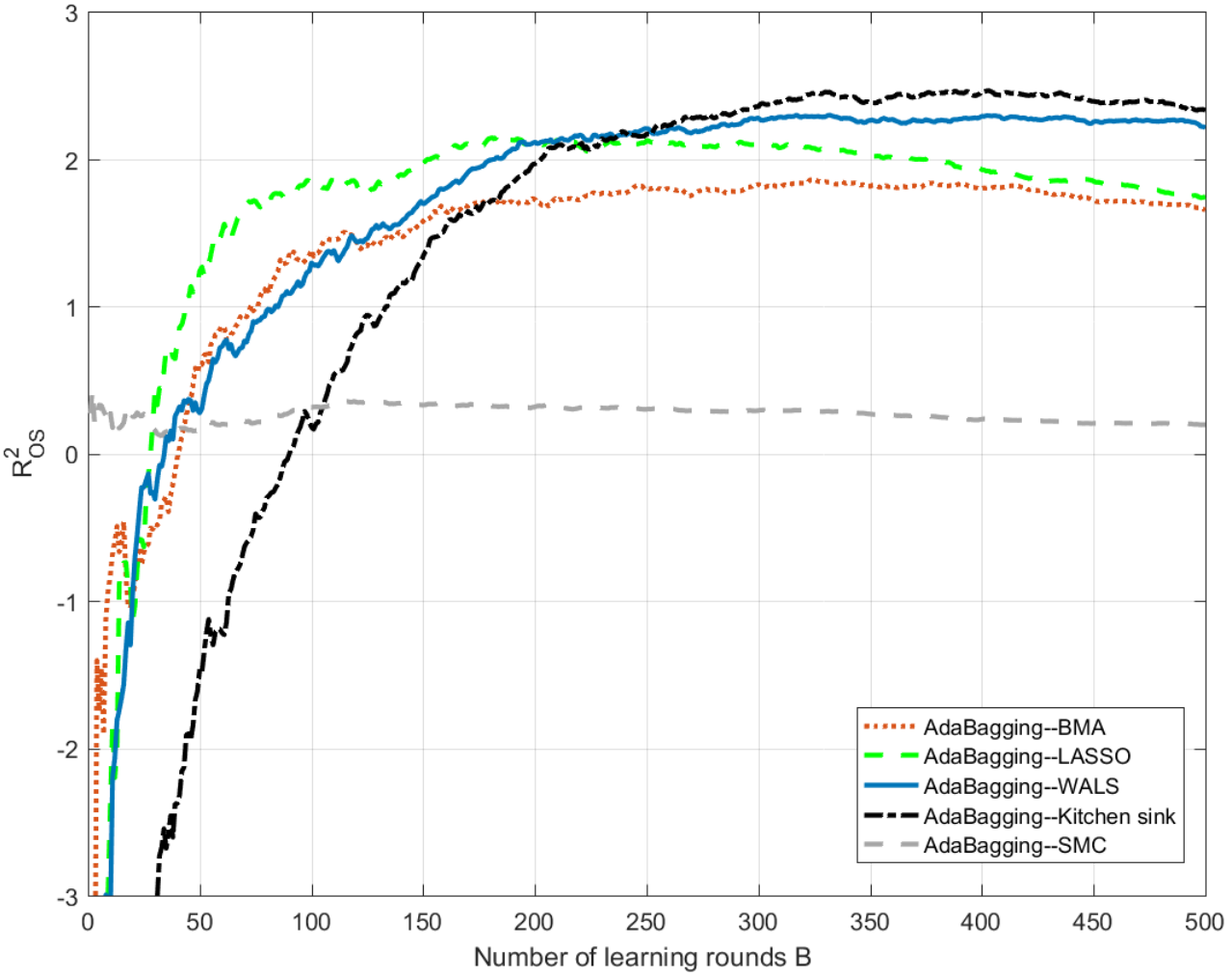


Fig. 1. Effects of the number of learning rounds of AdaBagging on R^2_{OS}

This figure depicts how the out-of-sample R^2_{OS} statistics changes with the number of learning rounds B of the AdaBagging method applied to different forecast models. These models are the standard multivariate predictive regression model: Kitchen sink, BMA with diffuse prior, LASSO with 5-fold cross-validation, Magnus et al. (2010)'s WALS, and Rapach et al. (2010)'s simple mean combination (SMC). R^2_{OS} measures the percentage reduction in mean squared forecast error (MSFE) for the AdaBagging method relative to the historical average benchmark forecast. The out-of-sample forecast evaluation period is 1957:01–2016:12.

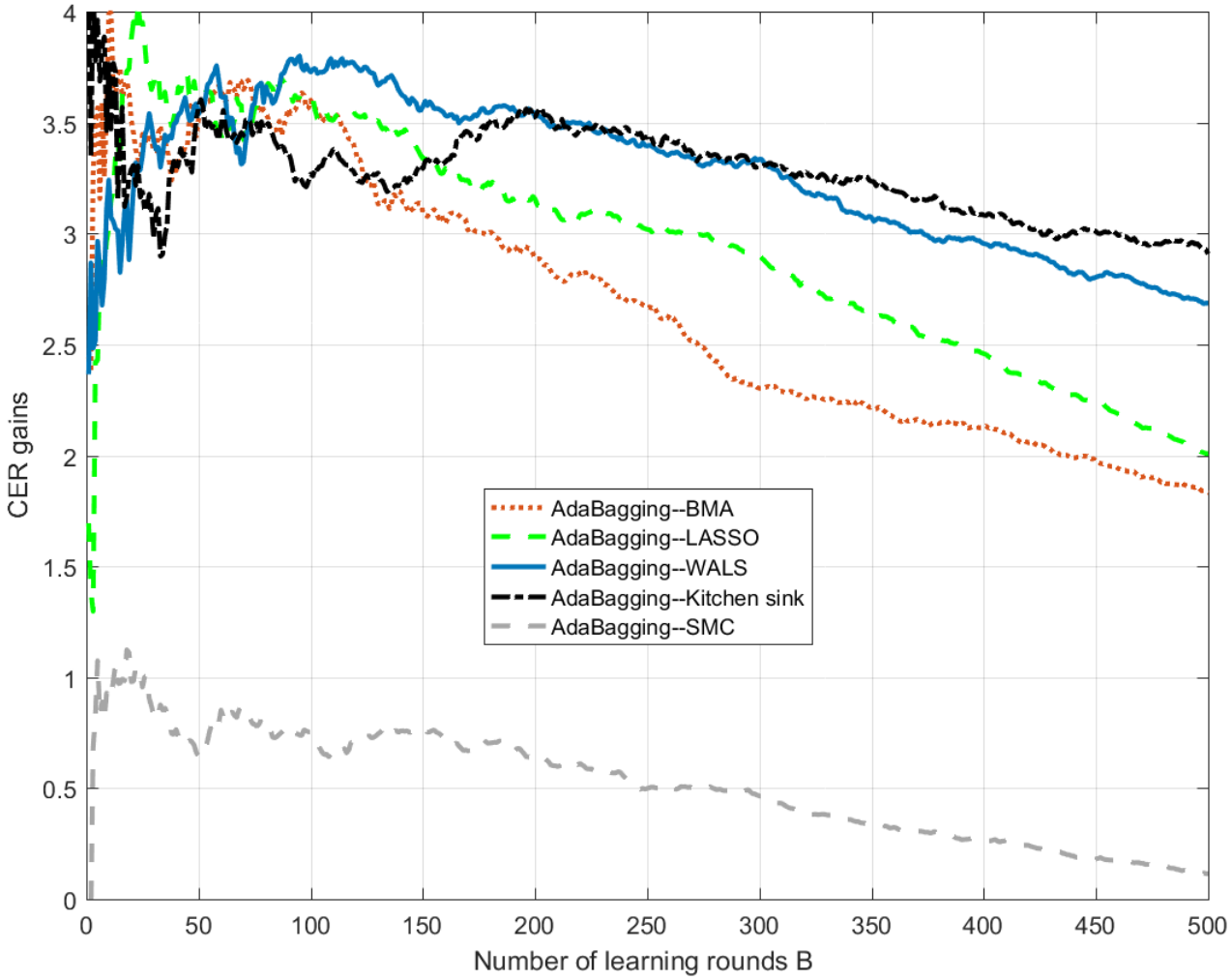


Fig. 2. Effects of the number of learning rounds of AdaBagging on CER gains

This figure depicts how the CER gains change with the number of learning rounds B of the AdaBagging method applied to different forecast models. These models are the standard multivariate predictive regression model: Kitchen sink, BMA with diffuse prior, LASSO with 5-fold cross-validation, Magnus et al. (2010)'s WALS, and Rapach et al. (2010)'s simple mean combination (SMC). The CER gain measures the annualized certainty equivalent return (CER) gain for a mean-variance investor with risk aversion coefficient of three for the AdaBagging method. The out-of-sample forecast evaluation period is 1957:01–2016:12.

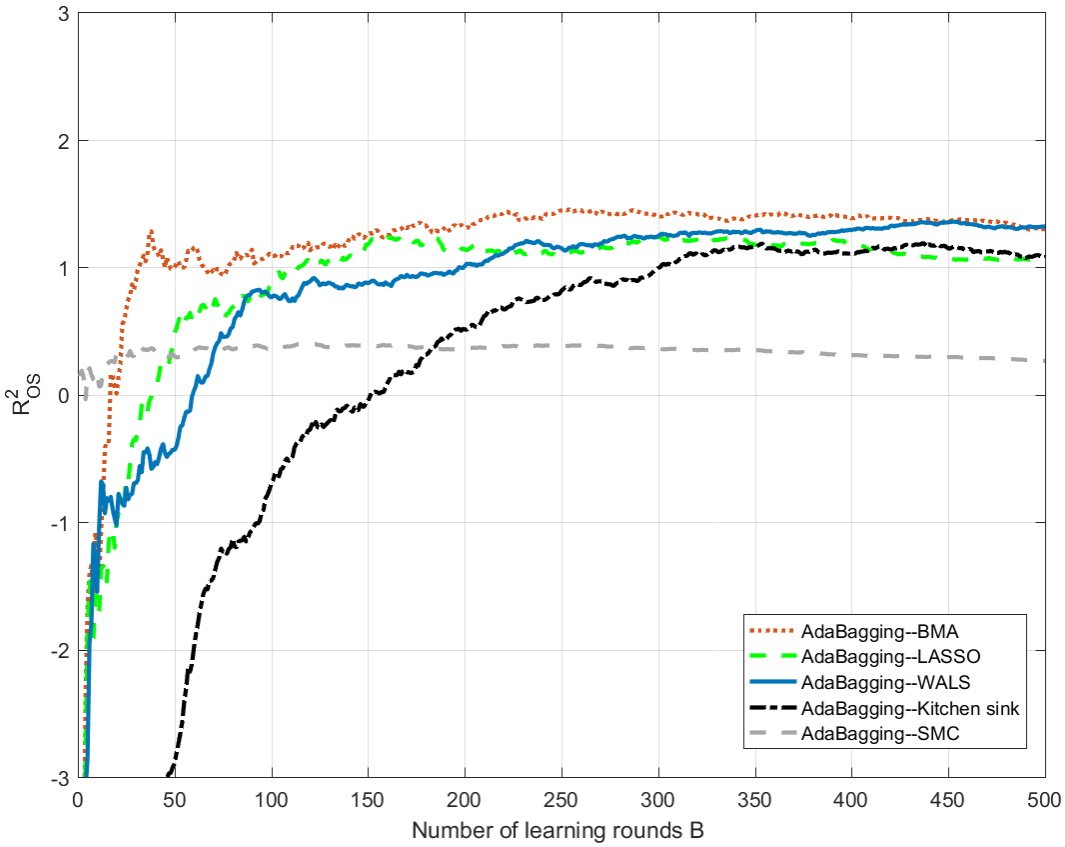


Fig. 3. Performance of AdaBagging with estimation windows determined by data

This figure depicts how the out-of-sample R^2_{OS} statistics of different forecast models, estimated using AdaBagging with optimal estimation windows determined by data, change with the number of learning rounds B of the AdaBagging method. These forecast models are the standard multivariate predictive regression model: Kitchen sink, BMA with diffuse prior, LASSO with 5-fold cross-validation, Magnus et al. (2010)'s WALS, and Rapach et al. (2010)'s simple mean combination (SMC). The optimal estimation windows are determined by the Pesaran and Timmermann (2007)'s cross-validation method with unknown break dates, given the minimum window size of 240. Following Pesaran and Timmermann (2007), we set the cross-validation window size W_{cv} of their cross-validation window selection method at 25% of the current observation sample. R^2_{OS} measures the percent reduction in mean squared forecast error (MSFE) for the given forecast model relative to the historical average benchmark forecast. The out-of-sample evaluation period is 1957:01–2016:12.

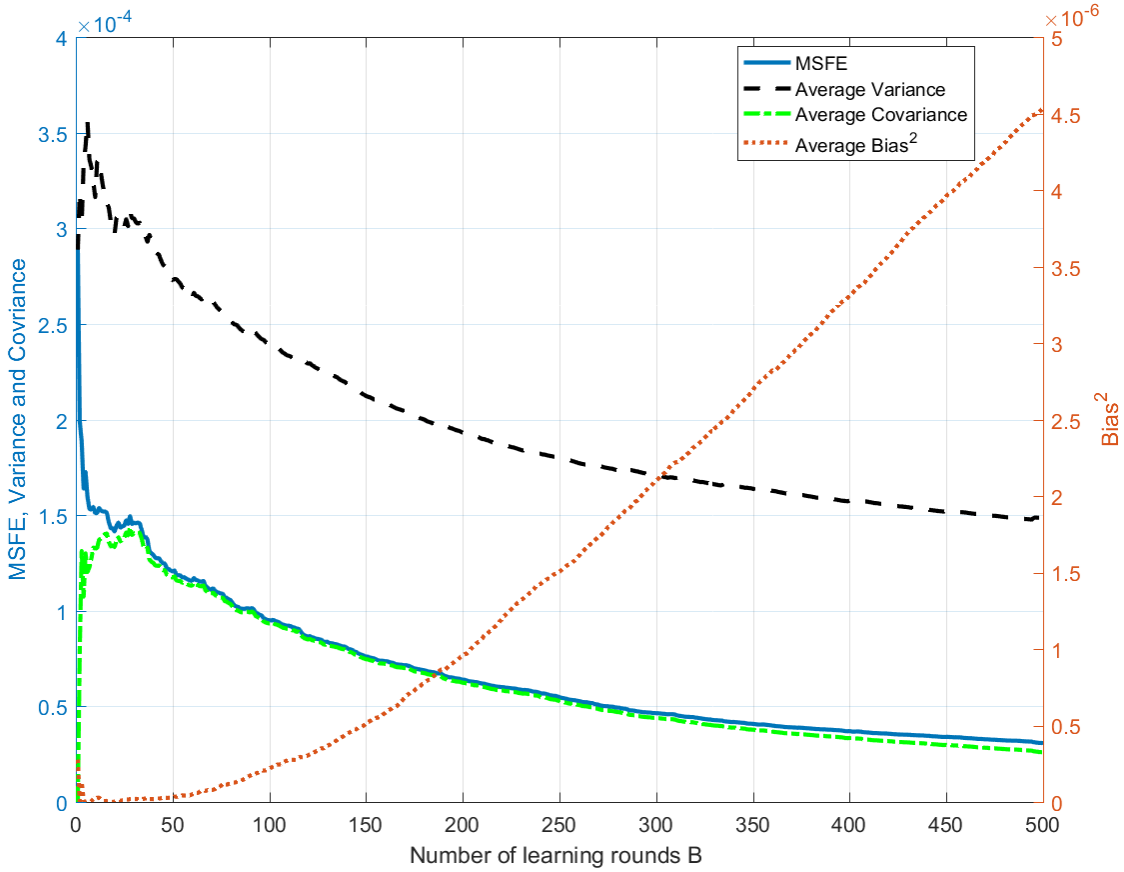
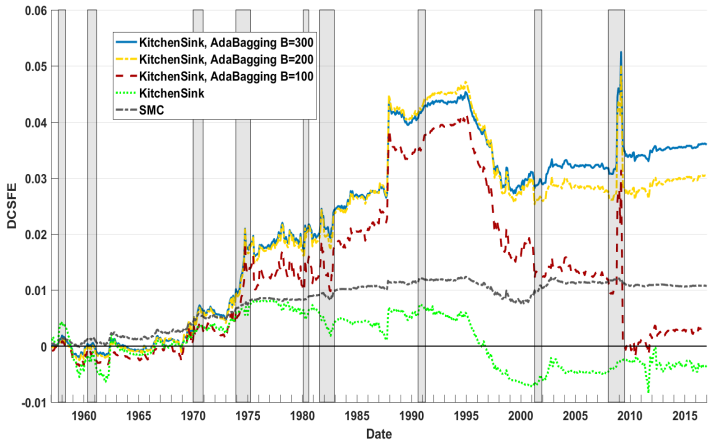


Fig. 4. Accuracy-diversity Trade-off of AdaBagging method

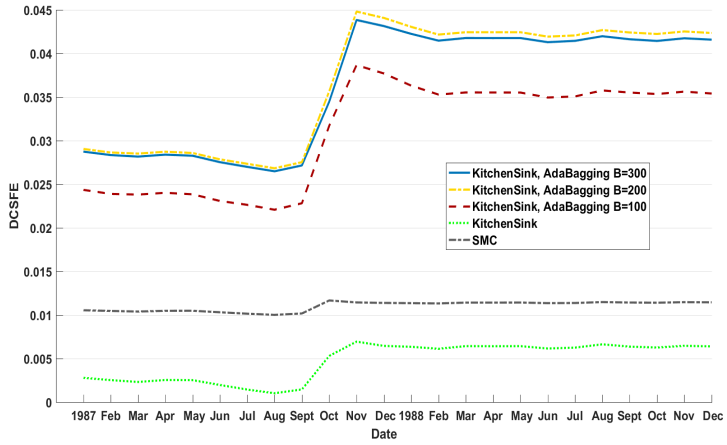
This figure depicts the bias-variance-covariance decomposition of the mean squared forecast errors (MSFE, solid line) for the AdaBagging–WALS method, with the number of learning rounds B changing from 1 to 500. The bias-variance-covariance decomposition of MSFE is defined as

$$MSFE = \overline{Bias}^2 + \frac{1}{B}\overline{Variance} + (1 - \frac{1}{B})\overline{Covariance}$$

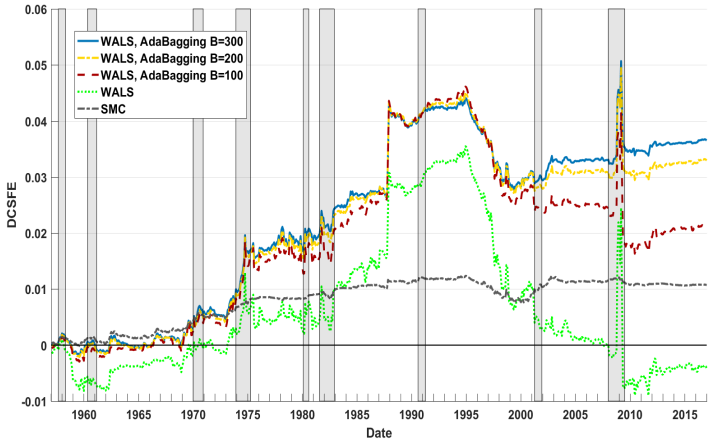
where $\overline{Bias}$, $\overline{Variance}$ and $\overline{Covariance}$ are the averaged squared bias (dotted line), averaged variance (dashed line), and averaged covariance (dash-dot line) of the individual forecasts, respectively, and B is the number of learning rounds. The out-of-sample forecast evaluation period is 1957:01–2016:12.



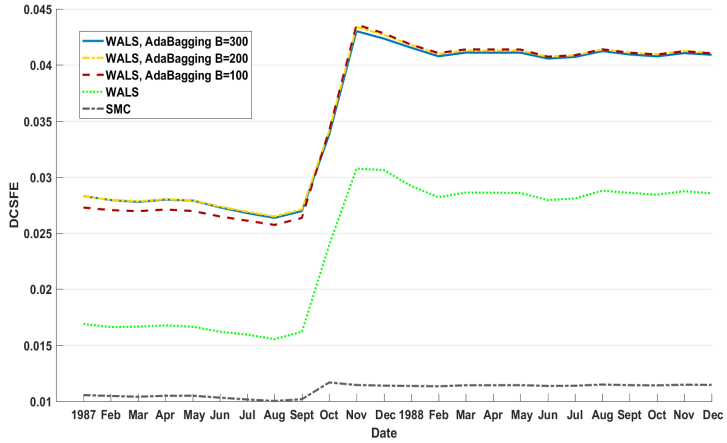
(a) Kitchen sink, 1957:01–2016:12



(b) Kitchen sink, 1987:01–1988:12



(c) WALS, 1957:01–2016:12



(d) WALS, 1987:01–1988:12

Fig. 5. Forecast accuracy over time

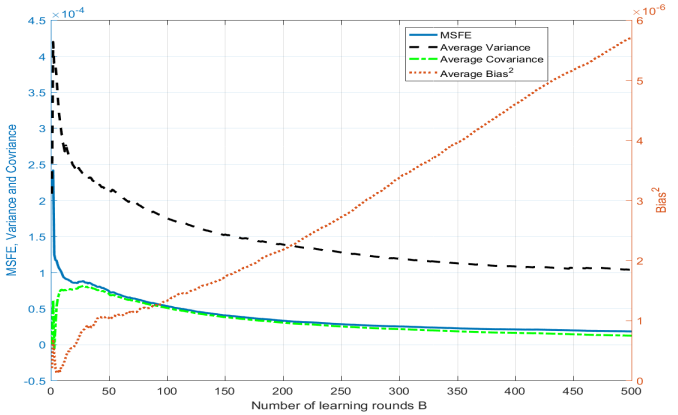
This figure shows the relative performance of the Kitchen sink model and Magnus et al. (2010)'s WALS in terms of the DCSFE, which is defined as the difference in the cumulative squared forecast error of the historical average benchmark model and the given model. The larger the DCSFE value, the better the model's performance. These models are estimated using 360-month rolling windows or AdaBagging with the number of learning rounds $B = 100, 200, 300$. For comparison, the DCSFE of the simple mean combination (SMC) from Rapach et al. (2010) is also depicted in gray dashed line. The vertical bars correspond to the NBER-dated recessions. The out-of-sample evaluation period is 1957:01–2016:12.

Internet Appendix

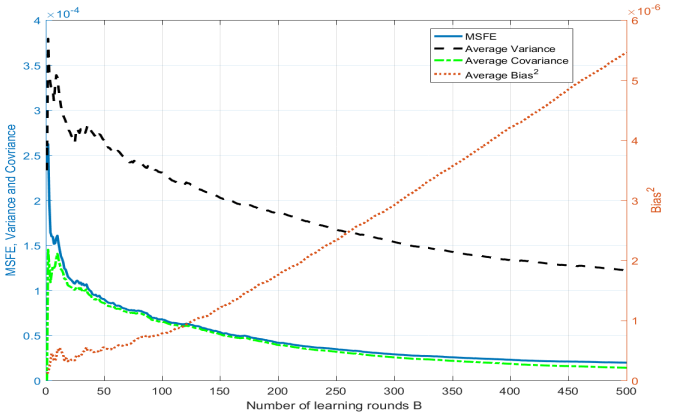
Table A.1
Predictive regression forecasts

This table reports the R^2_{OS} values of forecast models based on individual economic variables. These models are estimated using recursively expanding windows and rolling windows of size 360 months. R^2_{OS} measures the percent reduction in mean squared forecast error (MSFE) for the forecast model given in the first column relative to the historical average benchmark forecast. Brackets report p-values for the Clark and West (2007) MSFE-adjusted statistic for testing the null hypothesis that the historical average MSFE is less than or equal to the predictive regression MSFE against the alternative that the historical average MSFE is greater than the predictive regression MSFE (corresponding to $H_0 : R^2_{OS} \leq 0$ against $H_A : R^2_{OS} > 0$). CW-test is the Clark and West (2007) MSFE-adjusted statistic. The out-of-sample evaluation period is 1957:01–2016:12.

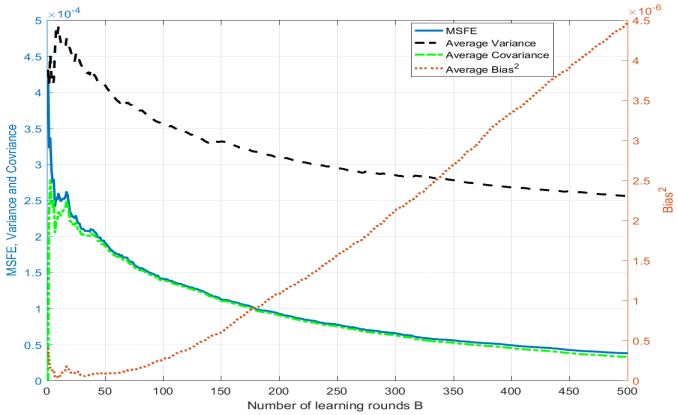
Economic variable	Rolling windows		Expanding windows	
	$R^2_{OS}(\%)$	CW-test	$R^2_{OS}(\%)$	CW-test
log(DP)	-0.71 [0.29]	0.54	-0.05 [0.10]	1.31
log(DY)	-0.75 [0.22]	0.78	-0.37 [0.07]	1.49
log(EP)	-1.80 [0.81]	-0.89	-1.88 [0.28]	0.57
SVAR	-0.69 [0.28]	0.58	0.32 [0.17]	0.97
BM	-1.99 [0.73]	-0.61	-1.74 [0.31]	0.50
NTIS	-1.00 [0.32]	0.47	-0.91 [0.41]	0.22
TBL	-1.30 [0.09]	1.33	-0.01 [0.09]	1.34
LTR	-0.69 [0.06]	1.57	-0.08 [0.20]	1.01
TMS	-0.31 [0.06]	1.57	0.06 [0.16]	0.84
DFY	-1.71 [0.83]	-0.95	-0.04 [0.59]	-0.24
DFR	-1.67 [0.64]	-0.37	-0.01 [0.38]	0.31
INFL	-0.48 [0.49]	0.03	-0.09 [0.50]	0.01



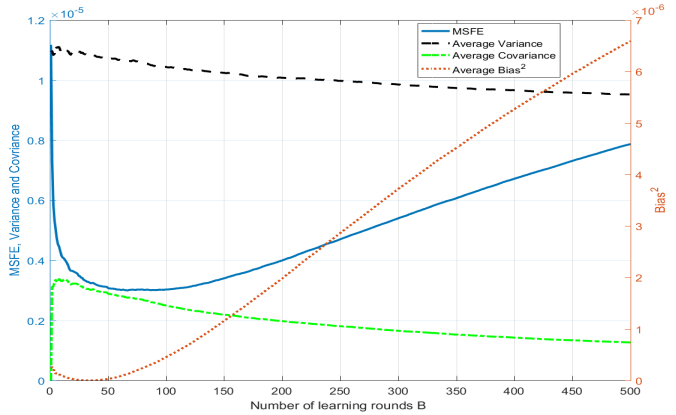
(a) AdaBagging–BMA



(b) AdaBagging–LASSO



(c) AdaBagging–Kitchen sink



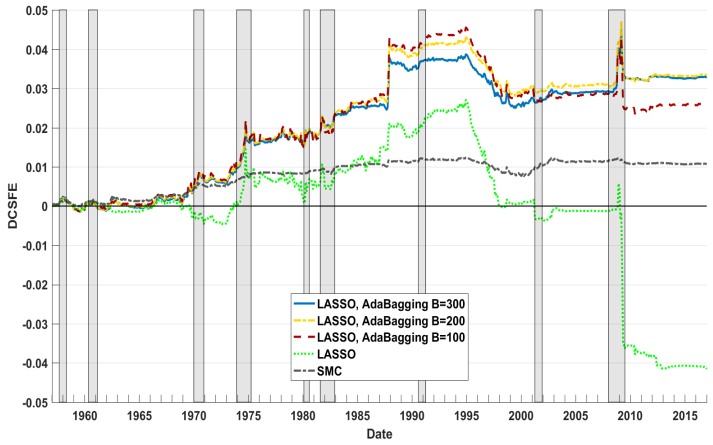
(d) AdaBagging–SMC

Fig. A.1. Accuracy-diversity Trade-off of AdaBagging method

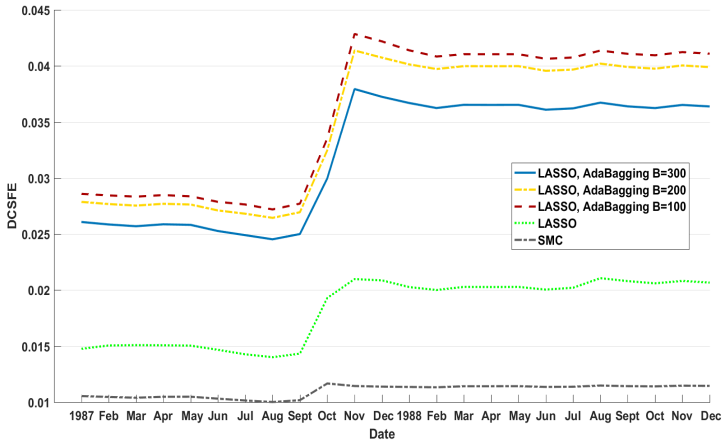
This figure depicts the bias-variance-covariance decomposition of the mean squared forecast errors (MSFE, solid line) for the AdaBagging–BMA, AdaBagging–LASSO, AdaBagging–Kithcen sink and AdaBagging-SMC method, with the number of learning rounds B changing from 1 to 500. The bias-variance-covariance decomposition of MSFE is defined as

$$MSFE = \overline{Bias}^2 + \frac{1}{B}\overline{Variance} + (1 - \frac{1}{B})\overline{Covariance}$$

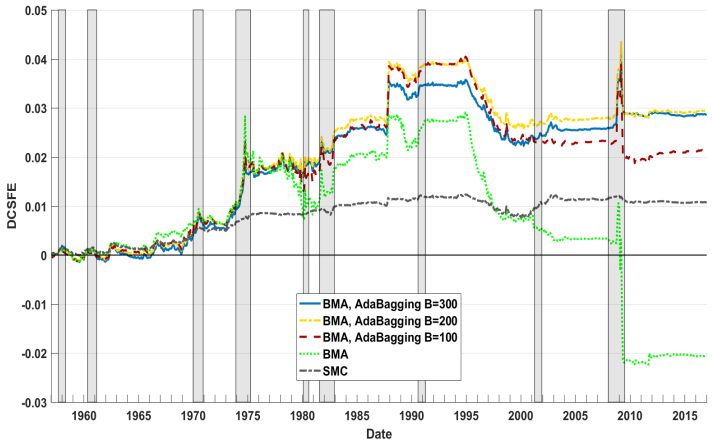
where $\overline{Bias}$, $\overline{Variance}$ and $\overline{Covariance}$ are the averaged squared bias (dotted line), averaged variance (dashed line), and averaged covariance (dash-dot line) of the individual forecasts, respectively, and B is the number of learning rounds. The out-of-sample forecast evaluation period is 1957:01–2016:12.



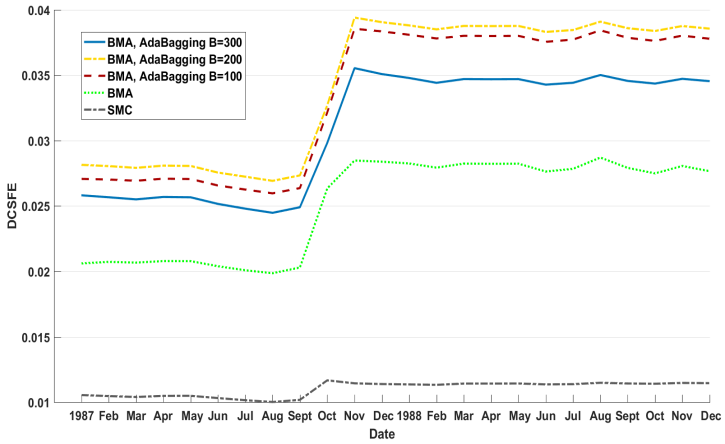
(a) LASSO, 1957:01–2016:12



(b) LASSO, 1987:01–1988:12



(c) BMA, 1957:01–2016:12



(d) BMA, 1987:01–1988:12

Fig. A.2. Forecast accuracy over time

This figure shows the relative performance of LASSO and BMA in terms of the DCSFE, which is defined as the difference in the cumulative squared forecast error of the historical average benchmark model and the given model. The larger the DCSFE value, the better the model's performance. These models are estimated using 360-month rolling windows or AdaBagging with the number of learning rounds $B = 100, 200, 300$. For comparison, the DCSFE of the simple mean combination (SMC) from Rapach et al. (2010) is also depicted in gray dashed line. The vertical bars correspond to the NBER-dated recessions. The out-of-sample evaluation period is 1957:01–2016:12.