Research article

Equity premium prediction: keep it sophisticatedly simple

Anwen Yin*

A.R. Sanchez, Jr. School of Business, Texas A&M International University, 5201 University Blvd, Laredo, TX 78041, USA

* Correspondence: Email: anwen.yin@tamiu.edu; Tel: +19563262513; Fax: +19563262479.

Abstract: Following the keep-it-sophisticatedly-simple principle, KISS, we propose using the averaging window approach to forecast the market equity premium in unstable environments. First, the estimation methodology of averaging window is a theoretically justified method robust to uncertainties on structural breaks and estimation window sizes. Second, the averaging window method has the obvious advantages of being understandable to forecast users and simple to implement, thus encouraging engagement and criticism. Our empirical results demonstrate the superior performance of the averaging window when forecasting the U.S. market equity premium, exceeding a wide range of methods which have been shown effective, such as shrinkage estimators and technical indicators.

Keywords: equity premium; forecast combination; structural break; averaging window

JEL Codes: C53, C58, G11, G17

---

1. Introduction

Accurate forecasts of the market equity premium play a vital role in empirical finance, as the aggregate equity predictions are often vital inputs into portfolio management and investment decisions. However, the predictability of the market equity premium has been the subject of contentious debate in academic research. Historically, studies such as Campbell (1987) and Fama and French (1988) have documented in-sample evidence supporting the predictability of the aggregate equity premium by means of exogenous variables such as the dividend-price ratio, dividend-yield, and various interest rate measures. However, this view is to a great extent challenged in Welch and Goyal (2008) in which the authors show that the previously documented in-sample evidence of predictability cannot translate into meaningful out-of-sample predictive gains on a consistent basis.

In light of the puzzle posed in Welch and Goyal (2008) regarding the disparity between in-sample and out-of-sample predictability, a number of studies in empirical finance have been published in the last decade proving that the equity premium can be meaningfully predicted out-of-sample. Generally speaking, the subsequent research on forecasting stock returns can be cast into two categories. The first group involves searching for new predictors, either new variables better reflecting the economic fundamentals or new composite indexes to a greater degree approximating market sentiment. For example, Li et al. (2013) show that the implied cost of capital can be used to forecast the excess equity returns, while Jiang et al. (2019) argue that the manager sentiment index possesses valuable information for forecasting stock returns over and above those contained in typical market sentiment indexes. The second category focuses on applying new estimation methods to address the econometric issues overlooked in Welch and Goyal (2008) which may have led to inferior forecasting performance. To illustrate, Campbell and Thompson (2008) argue that imposing constraints suggested by economic theory can largely restore the predictive content of many variables for the equity premium. Other studies such as Rapach et al. (2010) and Dangl and Halling (2012) center on using methods such as forecast combination and time-varying coefficients to account for the presence of structural breaks when generating equity premium forecasts.

Against the backdrop outlined above, in this paper, we do not attempt to make theoretical advances, such as searching for new predictive variables for the equity premium. Instead, we are interested in investigating if one can uncover the possible predictive content embedded in the set of predictors considered in Welch and Goyal (2008) by a sophisticatedly simple method, leading to superior predictive gains on a consistent basis. In the context of forecasting in business research, Green and Armstrong (2015) argue that in addition to being useful, a forecasting model or method should be simple, intuitive, and understandable to forecast users. The aim of sophisticatedly simple forecasting is to improve understanding, reduce mistakes and reveal bias. As such, Green and Armstrong (2015) use the phrase "keep-it-sophisticatedly-simple" (KISS) to describe the principle advocated in their article. Given the empirical challenges such as instability and tuning parameter selection in equity premium prediction, a sophisticatedly simple method or model should be capable of consistently and robustly delivering meaningful predictive gains while maintaining some degree of simplicity for forecast users to understand how predictions are made and why they work well. Furthermore, the sophisticatedly simple method should be flexible enough to accommodate its usage with diverse estimation methods and with new predictors to be discovered in future research.

While numerous studies have proposed complex methods or models for forecasting stock returns in the presence of instability, in this paper we contribute to the literature by showing that a sophisticatedly simple method such as the averaging window can achieve the same forecasting objective without the need for excess complexity, thereby encouraging engagement and criticism for further improvement among researchers, financial analysts and industry experts.

The estimation methodology of averaging window, originally proposed and analyzed in Pesaran and Pick (2011), is theoretically justified for being robust to a variety of uncertainties over the estimation window size and model instability. In addition, it is simple to construct, and can be used in conjunction with other forecasting methods such as model averaging or forecast combination, further extending its applicability to a plethora of settings in practice beyond forecasting the aggregate equity premium. The simplicity, usefulness, applicability and robustness ideally align the averaging window approach with the "keep-it-sophisticatedly-simple" principle advocated in Green and Armstrong (2015) when deciding between complex and simple methods in forecasting.

In empirical applications forecasting financial and economic time series, the estimation methodology of averaging window possesses the following advantages relative to many competing alternatives: (1) the construction of the averaging window estimator is theoretically justified in Pesaran and Pick (2011) regarding its usefulness dealing with the uncertainties over estimation window size and model instability; (2) the averaging window estimator is simple to construct without requiring the estimation of additional parameters, and is understandable to forecast users, thus facilitating engagement and criticism; (3) the averaging window estimator is robust to a variety of structural break types, including but not limited to types such as infrequent but identifiable large discrete breaks in coefficients, intermittent moderate breaks, break clusters, and smoothing transition to new parameter regimes; (4) the averaging window estimator is robust to the break distance, that is, the distance between the forecast origin and the latest structural break date if a break has occurred, thus eliminating the need to trim data when constructing the estimator; (5) the averaging window estimator is also robust to spurious structural breaks caused by unusual levels of volatility; (6) generating averaging window forecasts does not explicitly require the assumption of instability in the data generating process, hence validating the use of a number of forecast evaluation tests such as those introduced in Clark and McCracken (2013) when conducting inference in forecast evaluation; and (7) the averaging window estimator can be used in conjunction with either the rolling or recursive estimation window when generating forecasts for a single predictive model, or with a variety of forecast combination methods averaging forecasts from diverse sources.

Using an updated dataset maintained by Amit Goyal, we begin by showing that the averaging window forecasts from univariate predictive models uncover the predictive content for several exogenous variables such as the dividend-price ratio and the stock market variance, which were shown ineffective in Welch and Goyal (2008). Moreover, the gains delivered by the averaging window forecasts are broadly larger than those obtained from the restricted forecasts suggested in Campbell and Thompson (2008). Next, we conflate the averaging window estimator with forecast combination. The combined forecasts from the averaging window outperform many popular alternative methods which have been shown effective in the literature, such as the simple combination of Rapach et al. (2010), the elastic-net forecasts of Li and Tsiakas (2017), and the technical trading rules considered in Neely et al. (2014). The superior statistical performance of the averaging window carries over to its economic value to investors.

Why is the averaging window approach effective in forecasting the equity premium despite its relative simplicity compared with other complex methods such as Bayesian model averaging and time-varying coefficients? We explore the possible sources of the predictive gains accruing to the averaging window estimator via two channels. In the first channel investigating gains from univariate models, the weak performance documented in Welch and Goyal (2008) may be ascribed to the presence of structural breaks, either in the form of a one-time large break or smooth transitions. Many studies in the related literature have also argued that the failure of the univariate models originally considered in Welch and Goyal (2008) may be due to structural breaks, such as Rapach et al. (2010) and Dangl and Halling (2012), however, they do not explicitly provide empirical evidence of parameter instability associated with the data used in Welch and Goyal (2008). In contrast, we rigorously test for the presence of structural breaks using a variety of test statistics for all univariate models. Our results suggest that the significant presence of parameter instability in the divided-price ratio model may have caused its failure in forecasting stock returns under standard OLS estimation. However, with the averaging window estimator accounting for

instability, the divided-price ratio model outperforms the random walk benchmark, challenging the empirical conclusion drawn in Welch and Goyal (2008). In the second channel examining gains from forecast combinations, we show that the averaging window forecasts from various univariate models are largely less correlated with each other than their counterparts from simple OLS regressions, leading to superior combined forecasts following the standard arguments made for the benefits of forecast combination in Timmermann (2006).

Here we would like to emphasize that the aim of this paper is not about horse-racing: searching for the variable or model which reports the largest value of the predictive R-square, thus beating the empirical results documented in closely related articles published the past. Rather, we are interested in finding a sophisticatedly simple method which can uncover the possible predictive content embedded in a wealth of variables in unstable environments. Not only should this method be capable of delivering superior predictive gains consistently and robustly while being understandable to its users, but also it needs to be flexible enough to extend its applicability. For example, in the empirical results we show that the averaging window can be used in conjunction with the forecast combination to possibly further improve predictive accuracy. While we do not consider the following analysis for brevity, the averaging window approach also permits its usage with other estimation methods such as the elastic-net, or new predictive variables beyond those considered in Welch and Goyal (2008) such as the variance risk premium considered in Pyun (2019).

The remainder of this paper is structured as follows. Section two describes data and outlines the methods used in subsequent analysis. Section three describes metrics evaluating statistical and economic performances of forecasts. Section four presents empirical results. Section five explores the sources of predictive gains accruing to the averaging window. Section six concludes.

2. Data and econometric methodology

First, we describe data used in subsequent empirical analysis. Next, we outline the construction of the averaging window estimator in the context of univariate predictions and forecast combinations. Finally, we briefly discuss alternative predictive models and methods used when comparing forecasting performance.

2.1. Data source

We use monthly data from January 1927 to December 2017 on the aggregate U.S. equity premium along with a set of 14 predictive variables, for a total of 1092 observations. The data are hosted and periodically updated at Amit Goyal's website. The earlier part of data up to 2006 were used in the analysis of Welch and Goyal (2008). The equity premium (e.ret) is computed from the S&P 500 index including dividends minus the 3-month Treasury bill rate. The set of predictors comprises: the dividend-price ratio (dp); the dividend-yield (dy); earnings-price ratio (ep); dividend-payout ratio (de); the stock market variance (svar); book-to-market ratio (bm); net equity expansion (ntis); Treasury bill rate (tbl); long-term yield (lty); long-term return (ltr); term spread (tms); default yield spread (dfy); default return spread (dfr); inflation (infl). For brevity, we refer the interested readers to Welch and Goyal (2008) for details regarding the identity and construction of these financial variables.

2.2. Baseline predictive model

Following studies in this strand of research on forecasting stock returns such as Welch and Goyal (2008) and Pettenuzzo et al. (2014), we begin by presenting the baseline univariate predictive model used for forecasting the equity premium:

$$y_{t+1} = \beta_0 + \beta_1 x_{k,t} + \varepsilon_t \tag{1}$$

where y_{t+1} is the market equity premium at period t + 1, $x_{k,t}$ is the predictor k used at time t, and ε_t is the error term. The univariate model shown in Equation (1) is often named by the predictor that it contains. For example, if the predictor x_k used in Equation (1) is the dividend-yield (dy), then it is often called the dy model. In Welch and Goyal (2008), the baseline model is estimated by the ordinary least squares (OLS) before making predictions. In the following subsection, we argue that the baseline model should be estimated by the averaging window estimator in order to address some econometric issues overlooked in Welch and Goyal (2008), such as window size selection and model instability. For example, Rapach et al. (2010) argue that the presence of structural breaks which is not taken into account in the predictive regressions of Welch and Goyal (2008) may have led to their failure in forecasting the market equity premium. Using a nonparametric test statistic, Chen and Hong (2012) confirm the presence of a general breaking process in most univariate predictive regressions considered in Welch and Goyal (2008), thus supporting the argument made in Rapach et al. (2010).

2.3. Averaging window estimator

In practice, forecasters often need to make decisions regarding issues such as how to split the full sample before conducting out-of-sample analysis, and how to trim observations at both ends of the full sample before testing for structural breaks. Generally speaking, there is no clear guidance on how to optimally set these tuning parameters. For example, when spiting samples in out-of-sample forecasting, a relatively small training sample size could help reduce bias, but it may increase the variance component in the mean squared forecast error (MSFE), leading to imprecise predictions. While some studies such as Pesaran and Timmermann (2007) have attempted to address the optimality issue of sample selection, in practice, applying such optimality rules often requires estimating additional parameters, further complicating analysis.

Another type of uncertainty frequently encountered in the equity premium prediction literature is structural breaks or model instability. Studies such as Rapach and Wohar (2006) and Paye and Timmermann (2006) have provided empirical evidence of the statistically significant presence of structural breaks in the context of forecasting stock returns. However, the rejection of the null hypothesis of stability does not tell us what forms the break process follows. Furthermore, recent theoretical advances in econometrics examining the relationship between parameter instability and predictive gains have found that structural break only matters for forecasting if its size reaches a critical threshold, see Boot and Pick (2020) for a detailed discussion.

Against the backdrop described above regarding the uncertainties forecasters often face, here we outline the procedure of constructing out-of-sample forecasts with the averaging window estimator. The idea underlying the averaging window is fairly simple and intuitive: rather than choosing a fixed training window size to estimate predictive model parameters, one shall consider a series of nested estimation window sizes. Small windows close to the forecast origin are more likely to generate

unbiased forecasts, while large windows comprising more data in the distant past could help reduce forecast variance. Furthermore, small windows ignore the irrelevant breaks which have occurred in the distant past while large windows contain persistent breaks which may prove valuable to accurate predictions. Despite its simplicity, the robustness of the averaging window to uncertainties such as estimation window size and structural break type is solidly supported by the theoretical justifications provided in Pesaran and Pick (2011).

We implement the averaging window estimator in the framework of out-of-sample forecasting. Following the convention in the economic forecasting literature such as Clark and McCracken (2001), for a time series with T observations, we use R to indicate the initial training sample size used to estimate the predictive model parameters before making forecasts, $1 < R < T$. The remaining $P = T - R$ observations are reserved for forecast evaluation. In our empirical analysis, we set $R = 480$, resulting in an evaluation sample size of $P = 612$ which covers the period from January 1967 to December 2017. Our results are not sensitive to the value of R in preliminary analysis as long as it is not too small.

To make the first one-step-ahead out-of-sample point forecast at time period R for the equity premium, a sequence of nested m windows in the form of fractions within the first R observations is constructed as follows:

$$w_1 = R - T_{min}/R \tag{2}$$

$$w_2 = R - T_{min} - j(1)/R \tag{3}$$

$$w_3 = R - T_{min} - j(2)/R \tag{4}$$

$$\vdots \tag{5}$$

$$w_m = R - T_{min} - j(m-1)/R \tag{6}$$

where $w_1 = w_{min}$ indicates the minimum window size starting from the forecast origin moving backward, T_{min} denotes the number of observations included in the minimum window, and j denotes the number of observations included in windows beyond w_{min}. Note that $w_m = 1$, suggesting that with the largest window, all observations from 1 to R would be used to estimate model parameters. The window sizes can be compactly written as:

$$w_i = w_{min} + (i - 1/m - 1)(1 - w_{min}) \tag{7}$$

and $w_{min} \leq w_i \leq 1$, $I = 1, 2, \ldots, m$.

For each window, w_i, parameters are then estimated via OLS. The associated one-step ahead forecast for period $R + 1$ is:

$$\hat{y}_{R+1}^k(w_i) = \hat{\beta}_{0,k}(w_i) + \hat{\beta}_{1,k}(w_i)x_{k,R+1} \tag{8}$$

Then the one-step-ahead averaging window forecast (AveW) for period $R + 1$ is:

$$\hat{y}_{m,R+1}^k = \frac{1}{m}\sum_{i=1}^{m} \hat{y}_{R+1}^k(w_i), \tag{9}$$

where the item in the average is the period R + 1 forecast made with predictor x_k and window size w_i. In practice, Pesaran and Pick (2011) suggest the minimum window size w_{min} to be 15% of the training sample, and the number of windows m = 10. We report empirical results under the suggested parameter values of Pesaran and Pick (2011) for AveW forecasts. In our preliminary analysis, we have considered other values of w_{min} and m which are centered on the suggested values. The empirical results under alternative parameter configurations are qualitatively similar to the main results, thus we do not report them here for brevity.

The averaging window approach can be used in conjunction with either the recursive or the rolling estimation window. Under the recursive window, the first R + 1 observations would be used as the updated training sample to re-estimate model parameters for period R + 2 prediction. Under the rolling window, observations from 2 to R + 1 would be used as the updated training sample to re-estimate the predictive model for time R + 2 prediction, with the training sample size always fixed at R. Regardless of the estimation window choice, a sequence of P out-of-sample forecasts can be made, together with a series of P forecast errors.

2.4. Forecast combination

In the previous subsection we have outlined how to construct forecasts following the averaging window approach for a single predictive model with predictor k. However, pooling forecasts from different predictive models can help further improve performance in terms of statistical or economic gains, as shown in studies such as Pettenuzzo et al. (2014). Therefore, we construct a combined forecast for the equity premium by averaging across AveW forecasts made from Equation (1) with different predictors via equal weights. The combined forecast made for period t + 1 is:

$$\hat{y}_{m,t+1} = \frac{1}{K} \sum_{k=1}^{K} \hat{y}_{m,R+1}^{k} \tag{10}$$

where the item in the average is the AveW forecast made by Equation (1) with predictor k, and K is the total number of predictors or baseline models available.

Although alternative weighting schemes such as the discounted mean squared forecast error (DMSFE) weights and the approximate Bayesian model averaging (ABMA) weights are available to create the combined forecast, they do not improve upon the simple combination via equal weights in our preliminary analysis, so we do not consider other forecast combination weights in the empirical results section of this paper for succinctness.

2.5. Alternative models and methods

In light of the common practice in the literature of equity premium prediction such as Welch and Goyal (2008) and Pettenuzzo et al. (2014), we choose the simple yet empirically difficult-to-beat random walk model as the benchmark. The efficient market hypothesis inspired random walk model, also called the historical average or prevailing mean model in empirical finance, takes the following form:

$$y_{t+1} = \beta_0 + \varepsilon_t \tag{11}$$

Intuitively, the random walk benchmark assumes that the expected equity premium remains constant over time.

In addition to the historical average benchmark, we also consider a wide range of alternative models and methods with which we compare the AveW forecasts. For brevity, we refer interested readers to the articles cited below for details regarding these models and methods.

First, to ascertain how the averaging window forecasts improve upon the OLS forecasts considered in Welch and Goyal (2008) for univariate models, we make forecasts from Equation (1) which is estimated via OLS under the recursive estimation window (GW.REC) and the rolling estimation window (GW.ROLL). Analogously, to compare results in the context of forecast combination, following Rapach et al. (2010) we use equal weights to combine OLS forecasts obtained in the previous step. These combined forecasts are labeled RSZ.REC and RSZ.ROLL for recursive and rolling windows, respectively.

Second, we consider imposing restrictions on the univariate model forecasts following the suggestions made in Campbell and Thompson (2008). Specifically, we consider the forecast sign and slope sign restrictions on Equation (1), which are labeled CT.F and CT.S in subsequent empirical results, respectively. Moreover, we also consider imposing both restrictions, resulting in the forecasts named CT.B. The efficacy of imposing similar constraints on forecasts from Bayesian and dimension-reduction methods is investigated in studies such as Pettenuzzo et al. (2014) and Li and Tsiakas (2017).

Shrinkage estimators have been receiving growing attention in the economic forecasting literature. Studies such as Li and Tsiakas (2017) have demonstrated the superior performance of the shrinkage estimators when used for estimating the large kitchen-sink model comprising all available predictors for the stock returns. In the framework of shrinkage estimators, variables with weak past performance are permitted to exert limited influences while those exhibiting better performance are assigned greater weights for future forecasts. Two classic shrinkage methods, LASSO and Ridge, as well as the balanced elastic-net (ENET05) estimator which equally weights the LASSO and Ridge, are considered in our empirical applications. The usefulness of the elastic-net in forecasting stock returns has been shown in Li and Tsiakas (2017), making it a worthy competitor with which to compare AveW forecasts. Another dimension-reduction method we consider is the principle components regression (PCR), which has been used in the literature such as Neely et al. (2014).

Finally, we consider forecasts obtained from combining a wide range of technical indicator predictions. Specifically, following Neely et al. (2014) and Baetje and Menkhoff (2016), we generate equity premium forecasts using strategies such as moving average and momentum under various parameter configurations, then aggregate them to produce a combined forecast using the mean, the median, and the trimmed mean.

3. Forecast evaluation

We describe various metrics frequently used in the forecast evaluation literature assessing predictive gains. Both statistical and economic measures are presented.

3.1. Statistical evaluation

The most commonly used statistical measure evaluating forecasts is the OOS-R^2 proposed in Campbell and Thompson (2008), which compares the forecasts from the benchmark with those from a competing model under examination. Specifically, the OOS-R^2 can be computed according to the following:

$$OOS - R^2 = 100 \times \left(1 - \frac{\sum_{t=1}^{T-1}(y_{t+1} - \hat{y}_{t+1})^2}{\sum_{t=1}^{T-1}(y_{t+1} - \bar{y}_{t+1})^2}\right) \qquad (12)$$

Intuitively, the OOS-R^2 measures the percentage reduction in terms of the mean squared forecast error (MSFE) for the predictive model relative to the benchmark. The higher the OOS-R^2 value is, the better the gains would be for the predictive model.

Since OOS-R^2 is a point estimate of relative predictive accuracy, we assess its statistical significance via the MSFE-adjusted t-statistic (MSFE-t) proposed in Clark and West (2007). The MSFE-t tests the null hypothesis of equal forecasting accuracy against the one-sided alternative that the predictive model exceeds the benchmark. Although the asymptotic distribution of MSFE-t is not standard, Clark and West (2007) show that the standard normal distribution provides a good approximation.

Despite being widely used in empirical finance, the OOS-R^2 merely reports the predictive gains on average over the entire out-of-sample. Given the possible presence of instability in the underlying data generating process, it is likely that a model which forecasts well in the distant past may produce inferior contemporary predictions. Therefore, to gain a dynamic perspective on how each model performs over the entire evaluation sample, following the strategy suggested in Welch and Goyal (2008), we construct a new time series called the cumulative difference in squared forecast errors between the benchmark and the predictive model (CDSFE), then plot this series as a graphical device to evaluate forecasts.

At any time period t, if CDSFE is greater than zero, it implies that the predictive model exceeds the benchmark. The time series plot of the CDSFE can be employed to determine if the predictive model has a lower MSFE than the benchmark for any time window by simply comparing the heights of the curve at the beginning and end points of the segment corresponding to the period of evaluation. What matters is the slope of the CDSFE curve. A model which exceeds the benchmark would have a slope that is positive everywhere in its CDSFE plot. The closer the actual plot is to this ideal, the better the predictive gains are.

3.2. Economic evaluation

Cenesizoglu and Timmermann (2012) argue that statistical measures of forecasting performance may not necessarily be closely aligned with economic measures of predictive outcomes. A possible explanation is: the disagreement between statistical and economic measures may be caused by the fact that large predictive errors can be penalized more heavily by convex loss functions in statistical measures such as the MSFE relative to economic loss functions. As a result, we are interested in investigating if the AveW forecasts can generate meaningful economic value to investors who use them to guide investment decisions on a consistent basis.

Specifically, we consider a mean-variance investor who allocates capital between equities and risk-free instruments. At the end of each period t, the investor assigns an optimal portfolio share w_t of funds to equities for subsequent period according to the following rule:

$$w_t = \frac{1}{\gamma} \frac{\hat{y}_{t+1}}{\hat{\sigma}_{t+1}^2} \qquad (13)$$

where γ is the coefficient of relative risk aversion, y is the one-step ahead forecast of the equity premium, and σ^2 is the predicted variance of the equity premium. Following Rapach et al. (2016), we estimate the equity premium variance with a 10-year rolling window of historical data. In addition, we

impose the restriction that the optimal equity share w_t falls into the interval $[-0.5, 1.5]$, which allows for realistic short selling and leveraging as suggested in Rapach et al. (2016).

The investor who optimally allocates investment funds according to Equation (13) then realizes an average certainty equivalent return (CER) of

$$CER = \bar{R}_p - \frac{\gamma}{2}\sigma_p^2 \qquad (14)$$

where R_p and σ^2 in the above equation are the mean and variance of the optimal portfolio returns computed over the entire evaluation sample, respectively.

The CER can be viewed as the risk-free return that a mean-variance investor with a risk-aversion coefficient of γ would consider equivalent to investing according to the risky strategy. Similarly, we compute the CER value for the investor using benchmark forecasts. The CER gain (ΔCER) then is the difference between the CER from the predictive model and that from the benchmark. We report the annualized CER gain in percentage, thus it can be interpreted as the annual portfolio management fee in percentage that an investor would be willing to pay to gain access to the forecasts from predictive models in lieu of basing investment decisions on the benchmark.

In addition, we employ the Sharpe ratio (SR) to gauge the economic value of equity premium forecasts. The Sharpe ratio is the mean portfolio return in excess of the risk-free rate divided by the standard deviation of the portfolio return. Both the mean and the standard deviation of the portfolio returns are estimated over the entire evaluation sample. In keeping with the CER results, we report annualized Sharpe ratio in percentage in the empirical results.

4. Empirical results

In this section we present results evaluating forecasts in the context of univariate prediction and forecast combination. Both statistical and economic gains of forecasts are assessed.

4.1. Single model forecasting performance

We begin by providing a matrix plot of out-of-sample forecasts over 1967–2017 for all baseline univariate models as in Equation (1) estimated by the averaging window in Figure 1. The title of each panel in Figure 1 indicates the predictor used in the baseline predictive model, with the exception of the two panels in the lower-right corner, which are reserved for benchmark forecasts and realized excess returns, respectively. All forecasts are made with the rolling estimation window of 20 years of monthly data. An interesting pattern evinced in Figure 1 is that models such as dy and lty tend to generate smooth forecasts while models such as ltr and dfr produce volatile predictions.
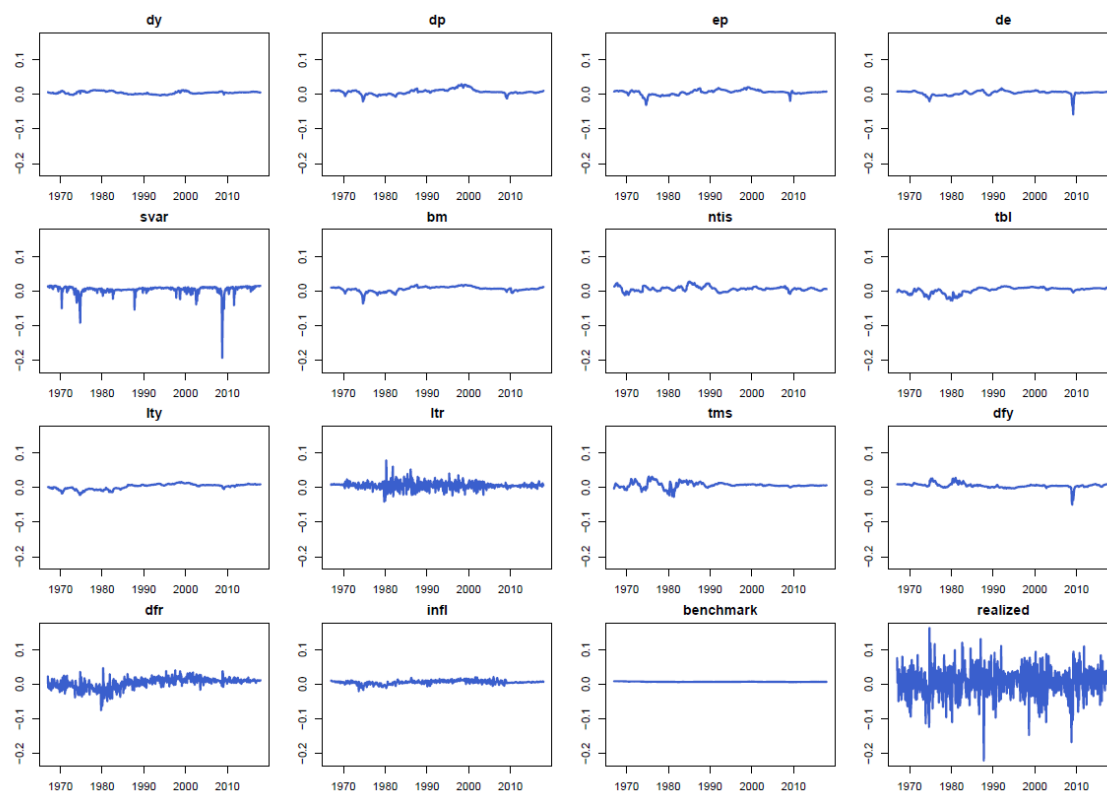
Figure 1. Averaging window forecasts from univariate models. This figure displays a matrix plot comprising the averaging window forecasts from all univariate predictive models, as well as the benchmark forecasts and the realized equity premium.

Table 1 reports the out-of-sample OOS-R^2 values in percentage across baseline univariate regression models estimated by various methods. The first column in Table 1 shows the name of the predictor used in Equation (1). The second and third columns report results for models estimated by the AveW approach with a rolling estimation window (AveW.ROLL), and the AveW method via a recursive estimation window (AveW.REC), respectively. The fourth and the fifth columns report results obtained via ordinary least squares (OLS) estimation as considered in Welch and Goyal (2008) with rolling (GW.ROLL) and recursive (GW.REC) windows, respectively. The remaining columns report results for OLS-estimated baseline models under various restrictions proposed in Campbell and Thompson (2008), namely, forecast sign (CF.F), slope sign (CT.S), and both sign (CT.B) constraints. ***, ** and * designate statistical significance via the MSFE-t statistic of Clark and West (2007) at levels of 1%, 5% and 10%, respectively. We make several observations from the results reflected in Table 1. First, a vast majority of univariate models estimated by OLS, regardless of estimation window, reports weak and insignificant performance against the historical mean benchmark, largely in keeping with the primary results shown in Welch and Goyal (2008) using a similar dataset ending in the year of 2006. Second, imposing the restrictions proposed in Campbell and Thompson (2008) indeed improves forecasting performance. Finally, the averaging window approach proposed in Pesaran and Pick (2011) can further improve upon the restricted forecasts, bringing the number of models being significant at least at 5% up to five under the rolling estimation window.

While the OOS-R^2 is useful in evaluating forecasts, it merely reports the predictive performance on average over the entire evaluation sample. To gain a dynamic perspective on how each AveW model fares over time, Figure 2 plots the time series of the difference of the cumulative sum of squared forecast errors between the benchmark and various predictive models (CDSFE) over 1967–2017. Recall that a positive slope of the CDSFE curve indicates better performance of the predictive model against the benchmark over a particular time window under examination. Overall, the results reflected in Figure 2 imply that variables such as dp, ltr and dfr to a great extent possess predictive content for the equity premium. Furthermore, Figure 2 suggests that the predictive performance of some variables seems elusive, such as de and ntis, corroborating the observation made in Timmermann (2008).

Table 1. Univariate model forecasting performance.

	AveW.ROLL	AveW.REC	GW.ROLL	GW.REC	CT.F	CT.S	CT.B
dy	−0.989	−0.570	−0.119	−0.964	−0.031	−0.535	−0.031
dp	0.282**	0.089	−0.023	−0.509	−2.246	−149.179	−1.130
ep	0.339**	0.425*	−0.508	−1.562	−0.720	−71.306	−1.130
de	−1.194	−0.563	−1.129	−0.688	−0.638	0.566**	0.566**
svar	5.241***	7.544***	−0.381	−0.577	1.702***	−1.583	−1.583
bm	0.166	0.093	−1.507	−3.443	−0.267	−10.964	−10.964
ntis	−0.446	−0.649	−0.849	−1.066	−1.001	−3.142	−3.142
tbl	−0.428	−0.134	−0.320	0.022*	0.241*	−1.937	−1.937
lty	−1.088	−0.750	−0.808	−0.717	0.316**	−5.137	−5.137
ltr	3.872***	1.690***	−0.225	0.235*	0.283*	0.765***	0.283**
tms	−0.038	−0.055	−0.042	0.120	0.046	0.214*	0.046
dfy	−0.391	−0.229	−0.836	0.172	−0.217	−0.267	−0.267
dfr	4.669***	4.215***	0.826***	1.196***	2.278***	1.713***	2.278***
infl	0.120*	0.249	0.140	0.372*	0.375**	−0.222	−0.222

Note: This table reports the OOS-R^2 values in percentage evaluating forecasts. ***, ** and * indicate statistical significance at levels of 1%, 5% and 10%, respectively. The first column displays the names of all univariate predictive models, while the first row shows the estimation methods used in estimating predictive models.

In sum, our results demonstrate that many variables with previously documented in-sample evidence of predictability, such as the dividend-price ratio and the term spread, indeed possess significant out-of-sample predictive content for the market equity premium as long as an appropriate estimation strategy is employed.

4.2. Regime-dependent evaluation

With the empirical evidence shown in Table 1 demonstrating the superior performance of averaging window forecasts, in this subsection, we are interested in investigating how AveW forecasts fare under different market conditions, highlighting the importance of regime-dependent evaluation for the equity risk premium advocated in Baltas and Karyampas (2018). Tables 2 and 3 report the OOS-R^2 values in percentage under various market conditions for the rolling averaging window forecasts and the recursive averaging window forecasts, respectively. In keeping with Table 1, the first column shows the predictor used in each baseline model. For the remaining columns in Tables 2 and 3: columns 2 and 3 report

separate results according to the expansion-recession business cycles defined by the NBER; columns 4 and 5 show
 7 report separate results according
to volatility, with periods above the sample average of the stock market variance (svar) labeled high while below the sample average of the svar labeled low. Again, ***, ** and * indicate statistical significance via the MSFE-t statistic of Clark and West (2007) at levels of 1%, 5% and 10%, respectively.
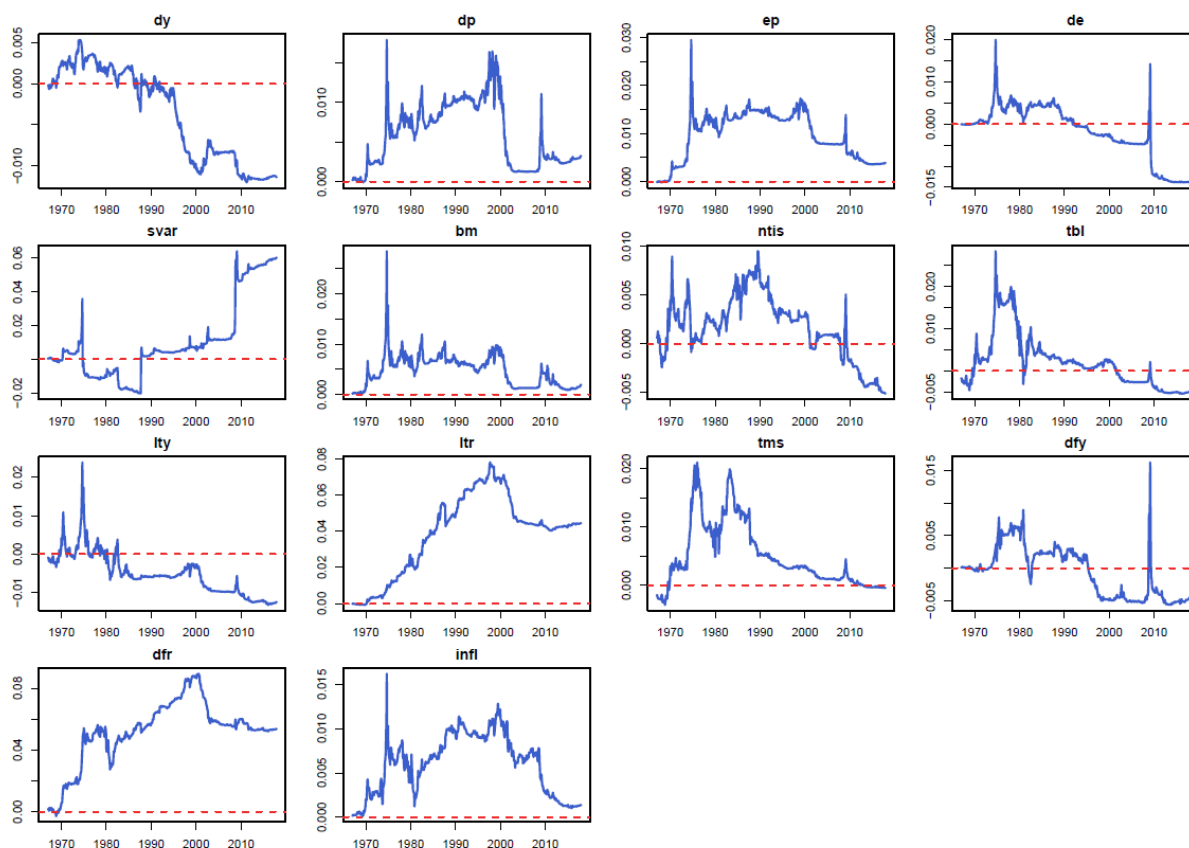
Figure 2. Univariate Predictive Model CDSFE plots. This figure presents the cumulative sum of the squared forecast error differences (CDSFE) for all univariate predictive models. A positive slope of the CDSFE curve indicates better forecasting performance for the predictive model relative to the random walk benchmark. Each panel is titled by the name of the predictive model.

Several interesting observations can be made after a thorough examination of Tables 2 and 3. First, the patterns shown in both tables largely support the conclusion reached in the literature of forecasting stock returns: the predictability of the aggregate equity premium tends to concentrate during recessions, volatile periods, or bearish regimes. For example, Rapach et al. (2010), Pettenuzzo et al. (2014) and Li and Tsiakas (2017) drew the same conclusion in separate but related studies. Second, the pattern favoring down markets is particularly discernible with the regimes separated by sentiments. To illustrate, in Table 2, only one model reports significant performance against the benchmark in bullish markets while all models become significantly better than the benchmark in bearish markets. Finally, the preference for down markets is more visible with forecasts implemented via a rolling estimation window. For instance, two models estimated by rolling windows report significant results during

expansions as opposed to six during recessions. However, about five models display significant results under a recursive estimation window, regardless of business cycles.

Table 2. Regime-dependent rolling average window forecast evaluation.

	Expansion	Recession	Bull	Bear	High	Low
dy	−0.729	−1.615	−8.307	4.473***	0.517	−2.906
dp	−0.721	2.687	−1.388	1.529**	−1.334	2.339***
ep	−0.210	1.656	−7.106	5.897***	−0.573	1.500***
de	−1.475	−0.519	−15.917	9.797***	−1.083	−1.335
svar	5.518***	4.577*	−24.766	27.639***	8.964***	0.503**
bm	−0.660	2.147	−8.417	6.573***	−0.754	1.337***
ntis	0.071	−1.687	−2.816	1.322**	−0.028	−0.979
tbl	−2.644	4.880**	−22.509	16.054***	1.573*	−2.974
lty	−2.076	1.280	−21.085	13.839***	−0.202	−2.214
ltr	3.038***	5.870***	1.748***	5.458***	1.397*	7.022***
tms	−3.333	7.855***	−0.702	0.458*	2.489**	−3.254
dfy	−1.671	2.674*	−7.437	4.868***	1.586*	−2.908
dfr	−0.553	17.180***	−5.474	12.240***	4.393***	5.020***
infl	−0.030	0.480	−7.293	5.654***	−0.021	0.300**

Note: This table reports the OOS-R^2 values in percentage evaluating the average window forecasts in various regimes. All forecasts are generated under the rolling estimation window. ***, ** and * indicate statistical significance at levels of 1%, 5% and 10%, respectively. The first column displays the names of all univariate predictive models, while the first row shows the names of various regimes.

Table 3. Regime-dependent recursive average window forecast evaluation.

	Expansion	Recession	Bull	Bear	High	Low
dy	−0.998	0.454	−11.823	7.830***	2.446***	−4.407
dp	−0.480	1.450	2.982***	−2.071	−1.640	2.289***
ep	0.042	1.341	−2.847	2.867**	−0.792	1.973***
de	−0.951	0.366	−11.425	7.544***	−0.129	−1.116
svar	7.967***	6.531***	−20.802	28.703***	12.991***	0.614**
bm	−0.267	0.956	−2.748	2.214**	−1.578	2.220***
ntis	0.756***	−4.015	2.682***	−3.136	−1.271	0.142**
tbl	−1.848	3.973**	−19.792	14.540***	1.305	−1.965
lty	−1.558	1.188	−22.063	15.160***	0.608	−2.477
ltr	0.262**	5.110***	2.343***	1.202*	−0.932	5.026***
tms	−2.823	6.576***	2.737***	−2.139	1.825*	−2.448
dfy	−0.826	1.203	−1.383	0.634	0.791	−1.526
dfr	−1.242	17.288***	−11.278	15.779***	5.432***	2.667***
infl	0.435**	−0.199	−4.798	4.015***	−0.399	1.073**

Note: This table reports the OOS-R^2 values in percentage evaluating the average window forecasts in various regimes. All forecasts are generated under the recursive estimation window. ***, ** and * indicate statistical significance at levels of 1%, 5% and 10%, respectively. The first column displays the names of all univariate predictive models, while the first row shows the names of various regimes.

To summarize, our regime-dependent evaluation broadly supports the conclusion drawn in the previous subsection: the sophisticatedly simple method of averaging window can uncover the valuable predictive content of many variables in difficult forecasting environments.

4.3. Forecast combination results

After examining the performance of averaging window forecasts from various univariate models, we are interested in investigating if further predictive gains can be achieved by combining all available forecasts. We apply the simple forecast combination method to combining all forecasts made from univariate models. That is, each baseline model forecast receives a constant weight value of 1/14 in the combined forecast. We construct combined forecasts for the following estimation methods: rolling averaging window (AveW.ROLL), recursive averaging window (AveW.REC), rolling window OLS (RSZ.ROLL), recursive window OLS (RSZ.REC), forecast sign restriction (CT.F), slope sign restriction (CT.S), and both sign restrictions (CT.B). Note that RSZ.ROLL and RSZ.REC forecasts correspond to the simple forecast combination results shown in Rapach et al. (2010) under rolling and recursive windows, respectively.

**Table 4.** Forecast combination performance.

	OOS-R^2	MSFE-t	p-value
AveW.ROLL	3.824	3.574	0.000
AveW.REC	3.772	4.047	0.000
RSZ.ROLL	1.010	2.613	0.004
RSZ.REC	0.884	2.656	0.004
CT.F	1.383	4.124	0.000
CT.S	1.388	2.957	0.002
CT.B	0.532	1.640	0.050
PCR	−0.525	−2.873	0.997
LASSO	−0.634	−0.646	0.741
RIDGE	2.289	2.813	0.002
ENET05	1.664	2.221	0.013
Mean	−0.214	−0.430	0.667
Median	−0.160	−0.175	0.569
Trim	−0.189	−0.321	0.626

Note: The first column shows the names of various combined forecasts and predictions from dimension-reduction methods. The second column reports OOS-R^2 values in percentage evaluating forecasts. The third column displays the MSFE-t statistic assessing the statistical significance of the OOS-R^2, with associated p-values shown in the last column.

In addition to combined forecasts, we also consider a number of dimension-reduction methods or shrinkage estimators which have been shown effective for forecasting stock returns in the literature, such as Neely et al. (2014) and Li and Tsiakas (2017). Specifically, we consider four shrinkage estimators: principles component regressions (PCR), Lasso estimator (LASSO), Ridge estimator (RIDGE), and an elastic-net estimator which equally weights Lasso and Ridge regressions (ENET05). Moreover, with the recently documented evidence supporting the use of tech indicators to forecast stock returns, such as Neely et al. (2014) and Ma et al. (2019), we also consider three schemes

averaging across a large number of forecasts made by technical trading rules. Specifically, we consider using the arithmetic mean (Mean), the median (Median) and the trimmed arithmetic mean after deleting the maximum and minimum (Trim) to combine forecasts from various moving average and momentum strategies proposed in Neely et al. (2014).

Table 4 reports forecast combination results. The first row displays the names of various forecast combination strategies and shrinkage estimators. The second row reports the out-of-sample OOS-R^2 values in percentage, with the third and fourth rows showing the associated MSFE-t statistics and p-values, respectively. Several patterns emerge from the results reflected in Table 4. First, the averaging window forecasts clearly dominate other methods by a relatively large margin, with the rolling AveW forecasts leading the recursive AveW predictions. Both AveW OOS-R^2 values are highly significant at the 1% level. Second, all of the remaining forecast combination methods report modest gains against the benchmark, largely in support of the findings documented in Rapach et al. (2010). Third, among shrinkage estimators, Ridge and the elastic net report relatively sizable gains, albeit weaker than those from the AveW, confirming the primary message conveyed in Li and Tsiakas (2017) that the elastic net method is effective for forecasting stock returns. Finally, turning to the combined forecasts from tech indicators, contrary to the evidence provided in Neely et al. (2014), they are uniformly inferior to the benchmark forecasts.
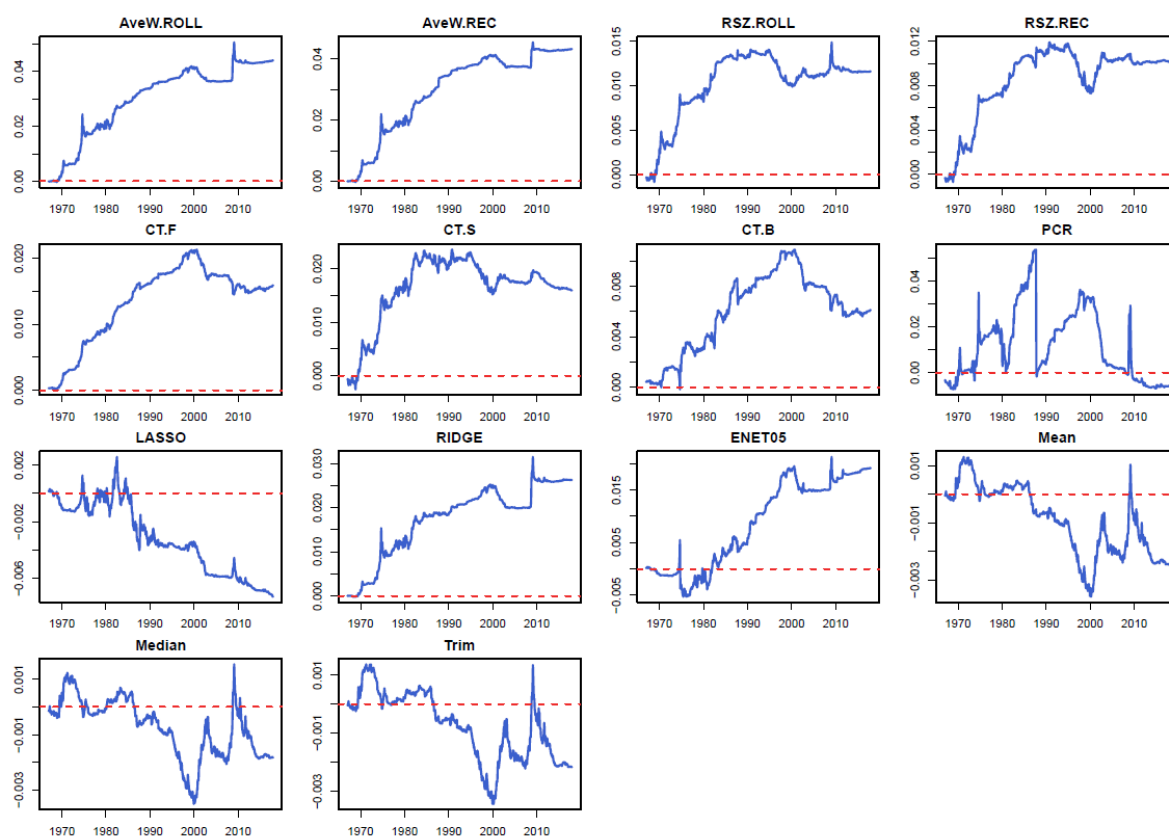
Figure 3. Forecast combination CDSFE plots. This figure presents the cumulative sum of the squared forecast error differences (CDSFE) for forecast combinations and shrinkage estimators. Each panel is titled by the name of the predictive model.

In keeping with the analysis performed for univariate models, to dynamically examine the forecasting performance of various combination methods, in Figure 3 we plot the CDSFE curves for all methods considered in Table 4. Overall, Figure 3 broadly demonstrates the superior performance of the averaging window forecasts relative to competing methods which have been shown effective for forecasting stock returns.

4.4. Economic value of forecasts

In the last subsection of empirical results, we are interested in examining the economic value of the averaging window forecasts delivered to investors who dynamically base optimal investment decisions on equity premium predictions. Following studies such as Pettenuzzo et al. (2014), we report the annualized certainty equivalent return (CER) gains and Sharpe ratio (SR) in percentage over the random walk benchmark. We assess the economic value of all forecast combination methods considered in Table 4, with two commonly used coefficient of relative risk-aversion (CRRA) values, $\gamma = 3$ and $\gamma = 5$. Table 5 reports results. Overall, the pattern evinced in Table 5 largely supports our conclusion drawn from statistical evaluation that the averaging window forecasts outperform competing methods in delivering gains to investors on a consistent basis.

Table 5. Economic value of forecasts.

	$\gamma = 3$		$\gamma = 5$	
	ΔCER	SR	ΔCER	SR
AveW.ROLL	3.571	1.199	2.201	1.127
AveW.REC	3.683	1.331	2.263	1.274
RSZ.ROLL	0.505	0.266	0.298	0.253
RSZ.REC	0.336	0.187	0.198	0.177
CT.F	1.601	0.330	0.957	0.328
CT.S	0.267	0.576	0.156	0.522
CT.B	0.871	0.056	0.520	0.057
PCR	3.889	0.736	2.484	0.641
LASSO	−0.978	−0.447	−0.588	−0.448
RIDGE	2.060	0.713	1.211	0.649
ENET05	2.168	0.641	1.237	0.567
Mean	−0.439	−0.232	−0.265	−0.233
Median	−0.350	−0.194	−0.212	−0.196
Trim	−0.408	−0.218	−0.246	−0.220

Note: This table reports the certainty equivalent return (CER) and Sharpe Ratio (SR) gains in percentage for various predictive models. We consider two commonly used coefficient of relative risk-aversion (CRRA) values, namely, $\gamma = 3$ and $\gamma = 5$.

To compare economic performance from a dynamic perspective, in Figure 4, we plot the log cumulative wealth for 14 portfolios named by the combination methods used when constructing forecasts. Without loss of generality, we assume that the investor starts with $1 and reinvests all proceeds over 1967–2017. For ease of comparison, two averaging window portfolios are marked in solid line while the others are denoted in dashed lines in various colors. Figure 4 reveals that the

superior predictive accuracy of the AveW forecasts can be translated into sizable economic gains, as the AveW portfolios clearly lead the rest in generating cumulative wealth.
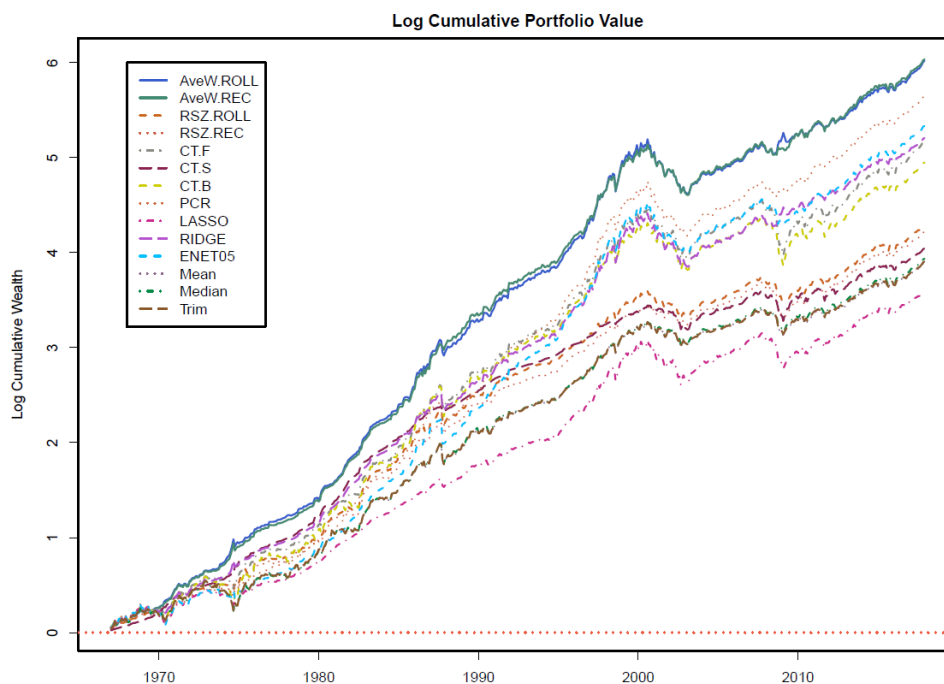
Figure 4. Log cumulative wealth growth. This figure delineates the log cumulative wealth growth for a mean-variance investor with relative risk coefficient of three, assuming that he or she starts with $1 and reinvests all proceeds over 1967–2017. The investor considers 14 portfolio strategies named by the method used in forecast construction. To highlight results, we exclusively use the solid line for the averaging window portfolio while others are represented by dashed lines in various colors.

5. Sources of predictive gains

In this section we investigate the possible sources of predictive gains from using the averaging window in univariate models and forecast combinations.

5.1. Source of gains with baseline predictive model

The estimation methodology of averaging windows is primarily motivated by the empirical evidence documenting the prevalence of instability among financial and economic regression models. There are a large number of tests designed to test the presence of structural break or model instability under different conditions. In the early stage of the structural break test literature, the focus was on designing optimal tests for large, abrupt and rarely occurring breaks, for example, see Bai and Perron (1998). Over the past twenty years, this strand of research has shifted to detecting a general breaking process encompassing cases such as sudden breaks and smooth transitions, for example, see Elliott and Muller (2006) and Chen and Hong (2012).

To ascertain if the predictive gains of the averaging window forecasts can be ascribed to instability, here we conduct various tests testing for the presence of structural breaks. Specifically, we consider four test statistics: SupF, AveF and ExpF statistics proposed in Andrews (1993) and Andrews and Ploberger (1994), and the qLL statistic proposed in Elliott and Muller (2006). The first three statistics test the null hypothesis of stability against the presence of a one-time large break, while the last one tests the null of no breaks against a general breaking process such as clustered breaks and smooth changes in coefficients.

Table 6. Structural break tests.

|	SupF	AveF	ExpF	qLL
dy	8.557	3.865*	2.365	−6.130
dp	30.907***	12.773***	11.779***	−8.754**
ep	36.579***	6.022**	12.465***	−5.529
de	5.072	1.459	1.002	−5.593
svar	17.943***	10.014***	6.643***	−4.883
bm	5.123	2.917	1.566	−2.320
ntis	8.432	4.719**	2.619*	−3.321
tbl	5.561	0.982	0.633	−4.285
lty	8.347	1.477	0.926	−3.746
ltr	37.675***	10.650***	15.029***	−27.486***
tms	4.829	1.602	0.910	−3.065
dfy	3.553	1.466	0.843	−1.081
dfr	29.408***	12.478***	11.902***	−12.049***
infl	3.876	1.046	0.609	−4.129

Note: This table reports the values of various test statistics testing for the presence of structural breaks, with superscripts ***, ** and * indicating statistical significance at levels of 1%, 5% and 10%, respectively. The first column displays the names of all univariate predictive models, while the first row shows the names of test statistics. For all tests, the null hypothesis is the absence of parameter instability in predictive regressions.

Table 6 reports structural break test results. In this exercise, we test for breaks for all baseline regression models, with model names indicated by the predictor it contains in the first column of Table 6. Columns 2–5 report values for SupF, AveF, ExpF and qLL tests, respectively. ***, ** and * indicate statistical significance at levels of 1%, 5% and 10%, respectively. We make several observations from Table 6. First, all tests provide evidence of instability for dp, ltr, and dfr models, suggesting that the predictive gains accruing to their AveW forecasts shown in Table 1 may arise from the fact that the averaging window approach account for the presence of structural breaks. Second, the SupF, AveF, ExpF tests report more cases of breaks than the qLL test, suggesting that models such as ep and svar are susceptible to abrupt and large breaks. Finally, only the AveF test reports weak evidence of break for the dy model, echoing the view expressed in studies such as Pesaran et al. (2013) that break dates and sizes are difficult to estimate precisely in practice.

5.2. Source of gains with forecast combination

Our previous results suggest that the superior performance of averaging window forecasts from baseline models may be attributed to the presence of structural breaks or instability. Here we are interested in investigating why the averaging window maintains its lead in forecast combinations.

A well-known result in forecast combination is that the combined forecasts are more likely to deliver superior gains if the underlying individual forecasts are correlated to a lesser degree, for example, see Timmermann (2006). Accordingly, in Figures 5 and 6, we plot the correlation matrix of out-of-sample forecasts for baseline models estimated via OLS and those estimated by the averaging window approach, respectively. All forecasts are made with the rolling window over 1967–2017. Statistically significant sample correlations are colored, with positive correlation in blue while negative in burned-orange. Each model is named by the predictor it contains.
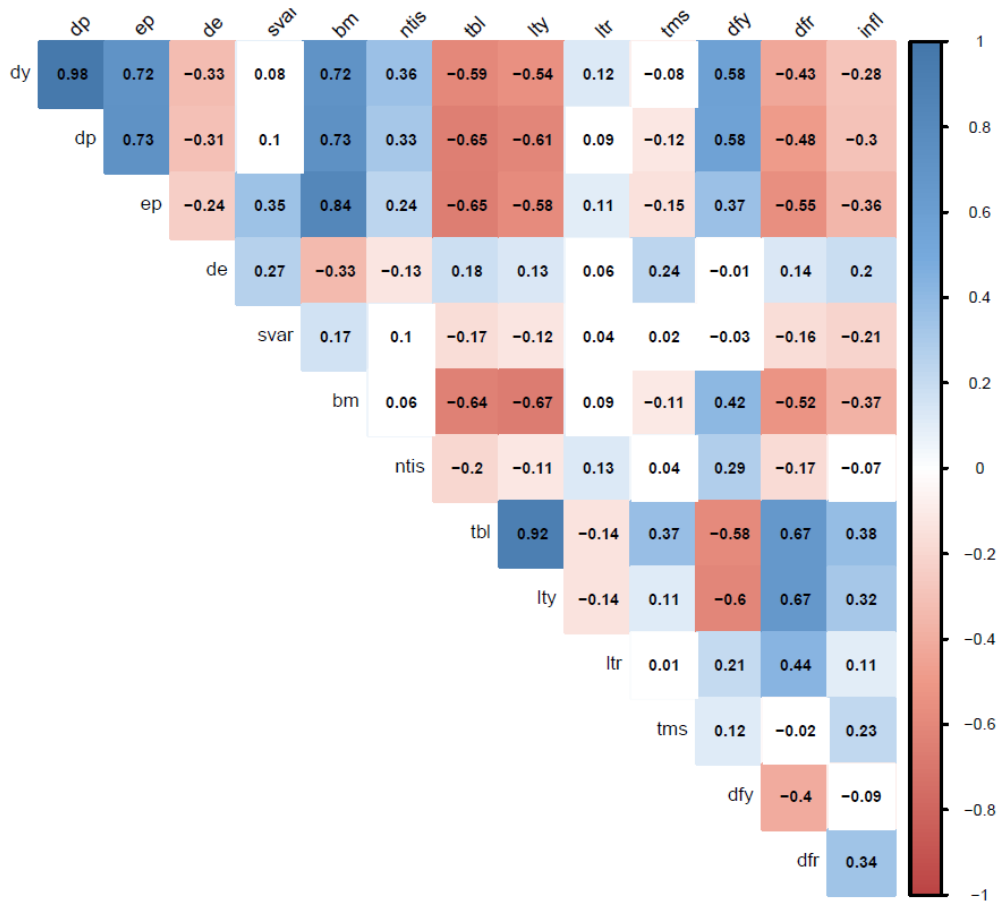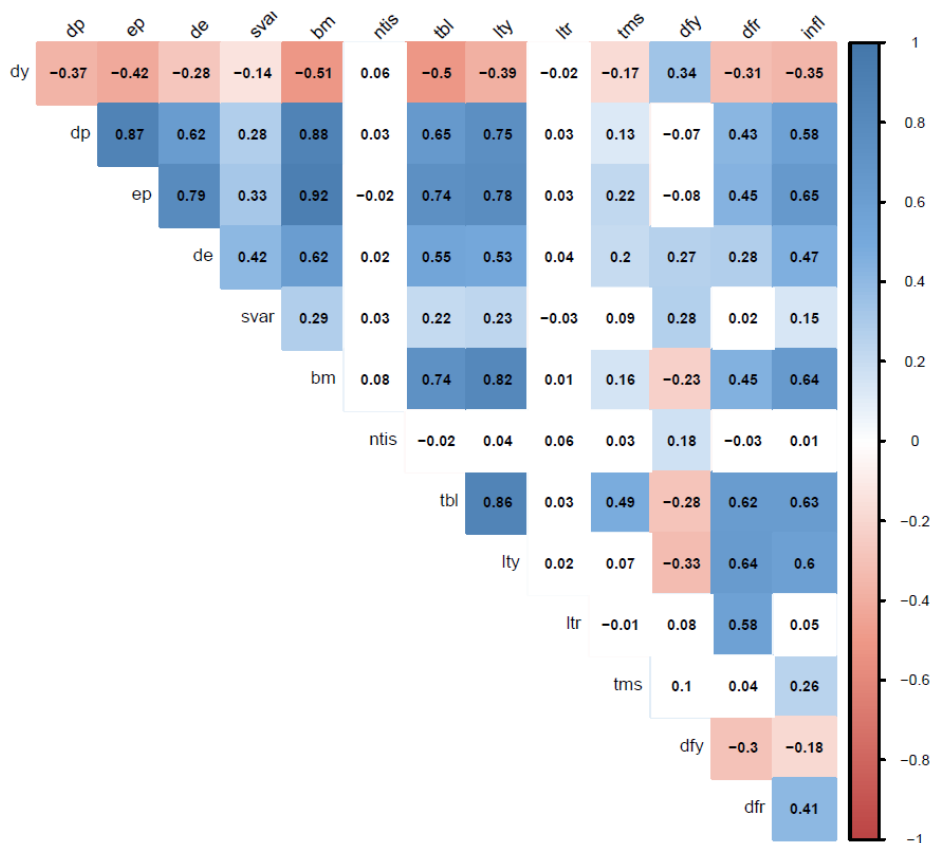

Figure 5. OLS forecast correlation matrix. This figure presents sample forecast correlations for each pair of univariate predictive models estimated via OLS. Statistically significant correlations are highlighted in colors.

Comparing Figure 5 with Figure 6, we observe that the number of insignificant pair-wise sample forecast correlations has increased from 15 to 30 when switching from OLS to the averaging window estimator. This observation has the implication that the AveW forecasts contain valuable information which

is not captured by the OLS forecasts, resulting in superior predictive gains when combined to form an average forecast. For example, in Table 4, the AveW.ROLL model constructed by combining AveW forecasts reports a much larger OOS-R^2 value of 3.824 than that of the RSZ.ROLL model built by combining OLS forecasts.

6. Conclusions

In practice, financial economists and professional forecasters may have access to a variety of models and methods to forecast the equity premium in unstable environments. Some models are certainly more complicated in terms of structure and the underlying theory supporting their use than others. However, as shown in empirical applications across diverse fields in finance, economics and business, complexity does not necessarily lead to predictive gains relative to simple, or even seemingly naive methods. Summarizing various empirical findings, Green and Armstrong (2015) advocate the principle of "keep-it-sophisticatedly-simple" (KISS) when deciding between complex and simple methods in forecasting, because simplicity has the obvious advantages of encouraging engagement and criticism by aiding in detecting mistakes, important omissions, spurious relationships and unsupported conclusions.

Figure 6. Averaging window forecast correlation matrix. This figure presents sample forecast correlations for each pair of univariate predictive models estimated via the rolling averaging window. Statistically significant correlations are highlighted in colors.

Echoing the views expressed in Green and Armstrong (2015), in this article we propose using the methodology of averaging window to forecast the market equity premium, and demonstrate its superior

performance relative to a variety of competing methods, such as restricted forecasts, simple combinations, shrinkage estimators, and tech indicators. The averaging window estimator, originally proposed and analyzed in Pesaran and Pick (2011), is theoretically justified for being robust to forecasting uncertainties such as window size and parameter instability. In addition, it is conceptually intuitive, understandable to forecast users, simple to implement, and can be used in conjunction with other predictive methods such as forecast combination, aligning ideally with the KISS principle advocated in Green and Armstrong (2015). Our empirical results show that a sophisticatedly simple method, such as the averaging window of Pesaran and Pick (2011), can achieve superior forecasting performance in unstable environments without the need for excessive complexity.

Acknowledgments

Conflict of interest

The author declares no conflict of interest in this paper.

References

Andrews D (1993) Tests for parameter instability and structural change with unknown change point. *Econometrica* 61: 821–856.

Andrews D, Ploberger W (1994) Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica* 62: 1383–1414.

Baetje F, Menkhoff L (2016) Equity premium prediction: Are economic and technical indicators unstable? *Int J Forecasting* 32: 1193–1207.

Bai J, Perron P (1998) Estimating and testing linear models with multiple structural changes. *Econometrica* 66: 47–78.

Baltas N, Karyampas D (2018) Forecasting the equity risk premium: The importance of regime-dependent evaluation. *J Financ Mark* 32: 83–102.

Boot T, Pick A (2010) Does modeling a structural break improve forecast accuracy? *J Econometrics* 215: 35–59.

Campbell J (1987) Stock returns and the term structure. *J Financ Econ* 18: 373–399.

Campbell J, Thompson S (2008) Predicting excess stock returns out of sample: can anything beat the historical average? *Rev Financ Stud* 21: 1509–1531.

Cenesizoglu T, Timmermann A (2012) Do return prediction models add economic value? *J Bank Financ* 36: 2974–2987.

Chen B, Hong Y (2012) Testing smooth structural changes in time series models via nonparametric regression. *Econometrica* 80: 1157–1183.

Clark T, McCracken M (2001) Tests of equal forecast accuracy and encompassing for nested models. *J Econometrics* 105: 85–110.

Clark T, McCracken M (2013) Advances in forecast evaluation. *Handb Econ Forecasting* 2: 1107–1201.

Clark T, West K (2007) Approximately normal tests for equal predictive accuracy in nested models. *J Econometrics* 138: 291–311.

Dangl T, Halling M (2012) Predictive regressions with time-varying coefficients. *J Financ Econ* 106: 157–181.

Elliott G, Muller U (2006) Efficient tests for general persistent time variation in regression coefficients. *Rev Econ Stud* 73: 907–940.

Fama E, French K (1988) Dividend yields and expected stock returns. *J Financ Econ* 22: 3–25.

Green K, Armstrong J (2015) Simple versus complex forecasting: the evidence. *J Bus Res* 68: 1678–1685.

Jiang F, Lee J, Martin X, et al. (2019) Manager sentiment and stock returns. *J Financ Econ* 132: 126–149.

Li J, Tsiakas I (2017) Equity premium prediction: the role of economic and statistical constraints. *J Financ Mark* 36: 56–75.

Li Y, Ng D, Swaminathan B (2013) Predicting market returns using aggregate implied cost of capital. *J Financ Econ* 110: 419–436.

Ma C, Wen D, Wang G, et al. (2019) Further mining the predictability of moving averages: evidence from the US stock market. *Int Rev Financ* 19: 413–433.

Neely C, Rapach D, Tu J, et al. (2014) Forecasting the equity risk premium: the role of technical indicators. *Manage Sci* 60: 1772–1791.

Paye B, Timmermann A (2006) Instability of return prediction models. *J Empir Financ* 13: 274–315.

Pesaran M, Pick A (2011) Forecast combination across estimation windows. *J Bus Econ Stat* 29: 307–318.

Pesaran M, Pick A, Pranovich M (2011) Optimal forecasts in the presence of structural breaks. *J Econometrics* 177: 134–152.

Pesaran M, Timmermann A (2007) Selection of estimation window in the presence of breaks. *J Econometrics* 137: 134–161.

Pettenuzzo A, Timmermann A, Rossen V (2014) Forecasting stock returns under economic constraints. *J Financ Econ* 114: 517–553.

Pyun S (2019) Variance risk in aggregate stock returns and time-varying return predictability. *J Financ Econ* 132: 150–174.

Rapach D, Strauss J, Zhou G (2010) Out-of-sample equity premium prediction: combination forecasts and links to the real economy. *Rev Financ Stud* 23: 821–862.

Rapach D, Wohar M (2006) Structural breaks and predictive regression models of aggregate US stock returns. *J Financ Econometrics* 4: 238–274.

Rapach D, Ringgenberg M, Zhou G (2016) Short interest and aggregate stock returns. *J Financ Econ* 121: 46–65.

Timmermann A (2006) Forecast combinations. *Handb Econ Forecasting* 1: 135–196.

Timmermann A (2008) Elusive return predictability. *Int J Forecasting* 24: 1–18.

Welch I, Goyal A (2008) A comprehensive look at the empirical performance of equity premium prediction. *Rev Financ Stud* 21: 1455–1508.