International Journal of Economics and Finance; Vol. 11, No. 12; 2019
ISSN 1916-971X E-ISSN 1916-9728
Published by Canadian Center of Science and Education

Equity Premium Prediction with Structural Breaks: A Two-Stage Forecast Combination Approach

Anwen Yin[1]

[1] A.R. Sanchez, Jr. School of Business, Texas A&M International University, Laredo, Texas, USA

Correspondence: Anwen Yin, WHT 217B, A.R. Sanchez, Jr. School of Business, Texas A&M International University, Laredo, TX., 78041, USA. Tel: 1-956-326-2513. E-mail: anwen.yin@tamiu.edu

Received: October 24, 2019 Accepted: November 15, 2019 Online Published: November 20, 2019

doi:10.5539/ijef.v11n12p50 URL: https://doi.org/10.5539/ijef.v11n12p50

Abstract

This paper introduces a two-stage out-of-sample predictive model averaging approach to forecasting the U.S. market equity premium. In the first stage, we combine the break and stable specifications for each candidate model utilizing schemes such as Mallows weights to account for the presence of structural breaks. Next, we combine all previously averaged models by equal weights to address the issue of model uncertainty. Our empirical results show that the double-averaged model can deliver superior statistical and economic gains relative to not only the historical average but also the simple forecast combination when forecasting the equity premium. Moreover, our approach provides an explicit theory-based linkage between forecast combination and structural breaks which distinguishes this study from other closely related works.

Keywords: equity premium, forecast combination, model averaging, structural break

1. Introduction

Forecasting the equity premium is of great importance to diverse areas such as portfolio allocation and performance evaluation of fund managers (e.g. Campbell, 1987; Campbell & Shiller, 1988; Fama & French, 1988; Fama & French, 1989; Ait-Sahali & Brandt, 2001; Avramov & Wermers, 2006). However, there is a long-standing debate regarding whether the equity premium can be meaningfully predicted out-of-sample by taking advantage of the information contained in various economic variables (e.g. Campbell & Thompson, 2008; Dangl & Halling, 2012; Pettenuzzo, Timmermann, & Rossen, 2014). By undertaking a comprehensive analysis investigating the aforementioned issue, Goyal and Welch (2008) provide empirical evidence showing that these variables perform poorly forecasting the equity premium out-of-sample relative to the historical mean which assumes a constant expected premium, and their predictive content seems episodic and unstable over time. For instance, in their article, most predictive gains come from the period following the oil shock in the 1970s, and they seem to disappear if those data are excluded. Goyal and Welch (2008) suggest that, in addition to the uncertainty on model selection, the unsatisfactory performance of many predictive variables could be attributed to issues related to structural break or parameter instability in the underlying data generating process.

In response to the findings reported in Goyal and Welch (2008), recent developments in the literature of forecasting equity returns show that the predictive power of various predictors can be uncovered or restored once an appropriate estimation methodology other than OLS is employed. Based on the general framework considered in Goyal and Welch (2008), Rapach, Strauss, and Zhou (2010) demonstrate that forecast combination could consistently improve upon the historical mean benchmark over time in terms of both statistical and economic gains. Additionally, they argue that, due to the uncertainty regarding model selection and parameter instability, the benefits of forecast combination come from taking advantage of all available information and its linkage to the real economy. However, it is not clear how the weighting methods combining models considered in Rapach et al. (2010) are explicitly linked with structural breaks. Specifically, despite the empirical evidence of instability documented in works such as Rapach and Wohar (2006), Rapach et al. (2010) do not consider any candidate model which allows for breaks when combining models, resulting in difficulty in interpreting the empirical results linking the success of forecast combination with structural breaks.

Our main contribution to the literature is to introduce a two-stage forecast combination methodology which improves upon the simple model averaging and can be easily implemented in empirical works. Our approach explicitly accounts for the possible presence of structural breaks beside the uncertainty on model selection,

leading to superior performance in terms of both statistical and economic gains when forecasting the equity premium out-of-sample relative to not only the historical mean, but also other competing models.

We consider a set of 14 bivariate predictive models, along with the historical average as those investigated in closely related studies such as Goyal and Welch (2008). However, our framework differs from them in that we also consider the structural break specification of those 14 bivariate models, i.e., all model coefficients are subject to a discrete break at an unknown date, thus, resulting in 28 candidate predictive models in total.

In the first step, for each model, we construct an averaged model by combining its break and stable specifications utilizing a number of weighting schemes, namely, equal weights, Schwarz information criterion weights (SIC), discounted mean squared forecast error weights (DMSFE), and Mallows weights. The first three are well documented and widely used in empirical analysis forecasting economic and financial variables (e.g. Stock & Watson, 2004). The last method, Mallows weights in the presence of possible breaks developed in Hansen (2009), is relatively new in the literature of forecast combination. This step is intent on eliminating the uncertainty surrounding parameter instability for each model. At the end of this stage, we have reduced the number of candidate models from 28 to 14. In the second stage, we simply average all 14 combined models constructed in the previous step based on equal weights, i.e., each previously averaged model receives a weight value of 1/14. Hence, we have completed constructing the double-averaged model. This stage of averaging aims to eliminate the uncertainty regarding model selection. The reason for selecting equal weights in the second stage is due to the fact that equal weights tend to outperform many estimation-based optimal weights when averaging over a large number of forecasting models in empirical analysis.

To empirically evaluate the statistical performance of our two-stage model averaging approach, we adopt the out-of-sample R_{OS}^2 statistic of Campbell and Thompson (2008), along with a graphical device based on the differences of the cumulative sum of the squared forecast errors between the historical mean benchmark and the double-averaged model. Our empirical results demonstrate that the double-averaged model significantly outperforms the simple forecast combination method considered in Rapach et al. (2010) and others in terms of the R_{OS}^2 statistic. For example, when predicting monthly returns, the simple forecast combination reports R_{OS}^2 value of 0.027%. However, all four weighting schemes within the two-stage forecast combination framework deliver at least a R_{OS}^2 value of 3.266%, suggesting that our approach can achieve about 3% more reduction in the mean squared forecast error than the simple forecast combination over the same predictive sample. Moreover, the statistical performance of the two-stage combination approach is consistent over time and robust to the choice of subsamples. For instance, when evaluating monthly forecasts for the smallest sample following the 2008-2009 financial crisis, the double-averaged model based on Mallows weights reports a R_{OS}^2 value of 2.563%, while that from the simple forecast combination is merely 0.008%. In addition, we find that the double-averaged model forecasts particularly well during economic recessions. Our empirical results are qualitatively the same across all forecast horizons.

In addition to the metric assessing statistical performance, the quality of returns prediction is often assessed based on the financial gains generated by the underlying models. Therefore, we evaluate and compare the economic gains measured according to the relative annualized certainty equivalent return (CER) and Sharpe ratio following related studies such as Ferreira and Santa-Clara (2011) and Li and Tsiakas (2017). In our empirical results, the two-stage forecast combination always leads to superior economic gains relative to the historical average across all time horizons and subsamples. For example, using the Mallows weight for monthly forecasts over the largest evaluation sample, we obtain an economic gain of 0.639% per year, the greatest among the four proposed weighting schemes. Turning to the Sharpe ratio, our empirical results are qualitatively the same as in the CER case: the two-stage forecast combination always results in superior Shape ratio gains relative to the historical average across all time horizons and subsamples.

The remainder of this paper is organized as follows. Section 2 outlines our two-stage model averaging methodology. Section 3 introduces the data and presents empirical results for the two-stage forecast combination approach and other competing models and methods. Section 4 examines the economic value of the two-stage forecast combination. Section 5 concludes.

2. Methodology

We begin by providing an overview of baseline linear predictive models for the market equity premium. Next, we discuss in detail the two-stage forecast combination when constructing out-of-sample forecasts. Finally, common statistical measures evaluating forecasts are discussed. We focus on the one-step ahead point forecast of the market equity premium across all available data frequencies.

2.1 Baseline Predictive Model

First, following related studies in the literature, we present the linear bivariate regression model forecasting one-period ahead equity premium where all coefficients are assumed constant over time:

$$r_{t+1} = \beta_0^i + \beta_1^i x_{i,t} + e_t, \tag{1}$$

where r_{t+1} is the one-step ahead excess returns, $x_{i,t}$ is the variable *i* available in time *t* to predict next period returns, stemming from a broad set of economic variables. e_t denotes the corresponding innovation. This specification is also referred to as the stable model subsequently as both β_0^i and β_1^i remain constant.

Turning to the construction of forecasts, we construct a series of forecasts of the market equity premium employing a recursive or expanding estimation window. Specifically, we divide the full sample of *T* observations into two non-overlapping segments: an estimation sample of size *R*, and an evaluation sample of size *P*, where *R* + *P* = *T*. Under the recursive estimation scheme, in each period, the previously estimated model coefficients are updated by including one more recently available observation, beginning with the first *R* observations. For instance, the first one-step ahead out-of-sample forecast utilizing predictor $x_{i,t}$ is

$$\hat{r}_{i,R+1} = \hat{\beta}_{0,R}^i + \hat{\beta}_{1,R}^i x_{i,R}, \tag{2}$$

where $\hat{\beta}_{0,R}^i$ and $\hat{\beta}_{1,R}^i$ are the OLS estimates of β_0^i and β_1^i, respectively, in Eq. (1) based on the first *R* observations. The second forecast can be generated by

$$\hat{r}_{i,R+2} = \hat{\beta}_{0,R+1}^i + \hat{\beta}_{1,R+1}^i x_{i,R+1}, \tag{3}$$

where $\hat{\beta}_{0,R+1}^i$ and $\hat{\beta}_{1,R+1}^i$ are the OLS estimates of β_0^i and β_1^i based on the first *R*+1 observations. Proceeding in this manner until the end of the entire sample, we have thus recursively generated a series of forecasts of size *P*, $\{\hat{r}_{i,s}\}_{s=R+1}^{T}$, using predictor x_i . We apply the procedure outlined above to all predictive models denoted by the predictor they include, x_i, where *i* = 1, ...,*M*, and *M* is the number of predictors available.

In practice, it is likely that there is no prior information suggesting that model (1) is the correct specification, given the well documented empirical evidence of structural breaks (see Paye & Timmermann, 2006; and Rapach & Wohar, 2006). Therefore, a seemingly natural competing alternative to the stable model (1) is a model allowing for instability in its coefficients:

$$r_{t+1} = \beta_{0,t}^i + \beta_{1,t}^i x_{i,t} + e_t. \tag{4}$$

Comparing with model (1), in Eq. (4), all coefficients become $\beta_{0,t}^i$ and $\beta_{1,t}^i$, indicating that they are time-dependent. The predictive model represented by Eq. (4) is called the break model subsequently.

Empirically, it is difficult to accurately estimate break sizes and locations, especially when the sample size is relatively small. Hence, increasing the number of discrete breaks for a predictive regression does not necessarily result in superior forecasting performance, as this involves a trade-off between predictability and complexity.

To boost predictive accuracy while reducing model complexity, here we only consider a discrete parameter break occurring at an unknown date τ in the break model (4). Hence, Eq. (4) can be rewritten as

$$r_{t+1} = \begin{cases} \beta_{0,1}^i + \beta_{1,1}^i x_{i,t} + e_t, & t < \tau, \\ \beta_{0,2}^i + \beta_{1,2}^i x_{i,t} + e_t, & t \geq \tau, \end{cases} \tag{5}$$

where τ denotes the period when break occurs. However, for identification, the break date τ is restricted to the closed interval $[\tau_1, \tau_2]$ which is bounded away from the ends on both sides of the estimation sample, i.e., $1 < \tau_1 < \tau_2 < R$.

The unknown break date τ can be estimated by concentration. Specifically, given a particular value of τ, we can estimate model (5) by OLS separately for each regime, then compute the corresponding sum of squared errors, $SSE(\tau) = \sum_{t=1}^{s} \hat{e}_t(\tau)$. We apply this approach to all possible values of τ specified in $[\tau_1, \tau_2]$, hence, generating a sequence of values of the sum of squared errors, $\{SSE(\tau)\}_{\tau=\tau_1}^{\tau_2}$. Our break date estimate, $\hat{\tau}$, would be the value of τ which is the global minimizer of $\{SSE(\tau)\}_{\tau=\tau_1}^{\tau_2}$. After dating the break, we employ the post-break window to estimate model coefficients in order to construct forecasts.

2.2 Benchmark Model

In the literature of forecasting stock returns, the efficient-market hypothesis inspired, historical mean model has

been proven to be a simple yet difficult to beat benchmark. Following related studies, we continue to use the historical average as benchmark. Specifically, it can be specified as

$$r_{t+1} = \beta_0 + e_t. \tag{6}$$

Again, we apply the recursive estimation window to construct a series of forecasts of size *P* based on model (6), and denote this series of forecasts as $\{\bar{r}_s\}_{s=R+1}^{T}$.

2.3 First Stage Forecast Combination

For a given bivariate model based on variable x_i, we have two naturally competing candidates to generate forecasts of the equity premium as shown in Eq. (1) and Eq. (5). In lieu of selecting a single best model among the two, here we combine Eq. (1) and Eq. (5) to form an averaged predictive model based on variable x_i, taking into account the uncertainty regarding parameter instability.

Specifically, we attach weight w to the break specification (5), and $1-w$ to the stable specification (1), where $w \in [0,1]$. Hence, the combined bivariate predictive model based on x_i is:

$$r_{t+1} = w\{\beta_{0,t}^i + \beta_{1,t}^i x_{i,t}\} + (1-w)\{\beta_0^i + \beta_1^i x_{i,t}\} + e_t. \tag{7}$$

Next, we will propose and discuss several approaches regarding how to assign weights, w and $1-w$, to the break and stable specifications.

2.3.1 Equal Weights

Forecast combination could be at a disadvantage over relying on a single model because it introduces additional estimation errors in situations where the weights need to be estimated. If so, the predictive gains from averaging could be wiped out by the noises introduced from estimating errors. This could help explain that, the seemingly suboptimal weighting schemes, such as equal weights, have widely been found to dominate complex methods which would be optimal in the absence of parameter estimation errors. Therefore, the first method is to use equal weights to combine the stable and break specification.

Specifically, the averaged model (7) based on equal weights becomes:

$$r_{t+1} = \frac{1}{2}\{\beta_{0,t}^i + \beta_{1,t}^i x_{i,t}\} + \frac{1}{2}\{\beta_0^i + \beta_1^i x_{i,t}\} + e_t. \tag{8}$$

2.3.2 DMSFE Weights

Stock and Watson (2003) propose a combination method based on the discounted mean squared forecast error (DMSFE) which computes weights according to the past performance of individual models over a holdout period. Specifically, for a predictive model *j* at time *t*, its weight is

$$w_{j,t}^d = \frac{\varphi_{j,t}^{-1}}{\sum_{s=1}^{M} \varphi_{s,t}^{-1}}, \tag{9}$$

where

$$\varphi_{j,t} = \sum_{l=1}^{t} \theta^{t-l} (r_{l+1} - \hat{r}_{l+1})^2, \tag{10}$$

and θ is a discount factor and *M* is the number of candidate models available. The DMSFE scheme assigns higher weight to a model with smaller historical error rate. When $\theta = 1$, there is no discounting, hence all historical observations are taken equally when calculating MSFE over the holdout sample. If $\theta < 1$, DMSFE allows for higher weights on the more recent observations. With DMSFE weights, the averaged model (7) becomes:

$$r_{t+1} = w_t^d\{\beta_{0,t}^i + \beta_{1,t}^i x_{i,t}\} + (1-w_t^d)\{\beta_0^i + \beta_1^i x_{i,t}\} + e_t. \tag{11}$$

2.3.3 Schwarz Information Criterion Weights

Bayesian model averaging (BMA) and its application in forecasting financial and economic variables have been receiving growing attention. In practice, the difficulty in constructing BMA estimates and forecasts lies in the complexity of obtaining the objects required to construct the weighted average. In situations where the predictive models' marginal likelihoods are difficult to compute, we can use a simple approximation based on the Bayesian information criterion. The Schwarz information criterion (SIC) takes the form

$$SIC_i = \frac{-2logLik_i}{T} + \frac{k_i ln(T)}{T}, \tag{12}$$

where $logLik_i$ is the log-likelihood of predictive model *i*, k_i is the number of parameters in model *i*, and *T* is the sample size. The SIC provides an asymptotic approximation to the marginal likelihood needed to compute BMA weights. Therefore, at time period *t*, if the SIC value of the break specification (5) is $SIC^b(t)$, and that of

the stable model (1) is $SIC^s(t)$, then the BMA weights can be approximated by the SIC weights. Specifically, the weight for the break model, w_t^s, is:

$$w_t^s = \frac{exp(-0.5SIC^b(t))}{exp(-0.5SIC^b(t))+exp(-0.5SIC^s(t))}. \quad (13)$$

Hence, the averaged model (7) under SIC weights becomes:

$$r_{t+1} = w_t^s\{\beta_{0,t}^i + \beta_{1,t}^i x_{i,t}\} + (1 - w_t^s)\{\beta_0^i + \beta_1^i x_{i,t}\} + e_t. \quad (14)$$

2.3.4 Mallows Weights

Hansen (2009) proposes a novel averaging estimator combining the structural break model and its stable specification by weights selected through minimizing a Mallows information criterion, whose penalty term is adjusted for the possible presence of non-stationarity.

In practice, the Mallows weight for the structural break model takes the form of a simple function of the SupF test statistic as presented in Andrews (1993), and a penalty parameter whose values are tabulated in Hansen (2009) for various combinations of trimming parameter and model size to meet regular empirical needs. Specifically, in period *t*, the Mallows weight for the break model (5), w_t^m, is

$$w_t^m = \begin{cases} 0, \ if \ F_t < \bar{p}, \\ 1 - \frac{\bar{p}}{F_t}, \ if \ F_t \geq \bar{p}, \end{cases} \quad (15)$$

where F_t is the SupF statistic in Andrews (1993), and $\bar{p}$ is the penalty parameter whose value relies on the asymptotic distribution of the SupF statistic.

Under Mallows weights, the averaged model (7) becomes:

$$r_{t+1} = w_t^m\{\beta_{0,t}^i + \beta_{1,t}^i x_{i,t}\} + (1 - w_t^m)\{\beta_0^i + \beta_1^i x_{i,t}\} + e_t. \quad (16)$$

2.4 Second Stage Forecast Combination

In our framework, we begin with *M* linear predictive models assuming constant coefficients as in Eq. (1) with each model differing from others by the unique predictor it includes. Since we are concerning with the presence of structural breaks, we extend the total number of candidate models to 2*M* by adding break specifications of the original *M* models.

In the first step, we use one of the four weighting methods outlined in previous sections to combine the break and stable specifications for each candidate model. In the end, we have reduced the number of models from 2*M* to *M*. This stage intends for accommodating the uncertainty regarding parameter instability for individual models. However, after the completion of averaging break and stable specifications, the uncertainty on which combined model out of the *M* averaged candidates best predict the equity premium remains. Therefore, in the second stage, we average the remaining *M* combined models to form a double-averaged model. Nonetheless, in the second step, we simply use equal weights to convert *M* candidates into a single forecasting model. The reason for our weighting choice in this step is that pooling a large number of models by equal weights tend to outperform other complex weighting schemes in empirical works. When the number of models is large, the possible gains from complex weights may be dominated by the extra noises from estimating additional parameters.

After the first stage, we have combined the break and stable specifications for each bivariate model *i* based on predictor x_i, using one of the four proposed weighting schemes to obtain $MODEL_i^j$, where $i = 1, \dots, M$, and $j \in \{e, d, s, m\}$ denotes the combination methodology used: *e*, equal weights; *d*, DMSFE weights; *s*, SIC weights; *m*, Mallows weights. Next, we assign equal weights to all *M* averaged models constructed previously. Therefore, our double-averaged predictive model is:

$$r_{t+1} = \frac{1}{M}\sum_{i=1}^{M}\{w_{i,t}^j[\beta_{0,t}^i + \beta_{1,t}^i x_{i,t}] + (1 - w_{i,t}^j)[\beta_0^i + \beta_1^i x_{i,t}]\} + e_t. \quad (17)$$

To summarize, in the first stage, to address the possibility of structural breaks, we combine the break and stable specifications for each model *i* by weighting methods such as Mallows weights. Then, to circumvent the danger of forcing extra noises into the construction of forecasts from complex weighting schemes, we simply use equal weights to pool all averaged models obtained from the first stage to form the double-averaged model.

2.5 Forecast Evaluation

Conventionally, the statistical measure assessing the quality of out-of-sample forecasts is the mean squared forecast error (MSFE), which is defined as:

$$MSFE = \frac{1}{P}\sum_{t=R}^{T-1}(r_{t+1} - \hat{r}_{t+1})^2, \tag{18}$$

where r_{t+1} is the realized value of the market equity premium in period $t + 1$, $\hat{r}_{t+1}$ is the forecast generated at period t, R is the training sample size, P is the evaluation sample size and $T = P + R$ is the total sample size.

For the purposes of comparing the forecasting performances with those reported in closely related studies, we adopt the MSFE-based out-of-sample R^2_{OS} statistic considered in Campbell and Thompson (2008) to measure statistical gains. Specifically,

$$R^2_{OS} = 100 \times \left(1 - \frac{MSFE^i}{MSFE^0}\right), \tag{19}$$

where i indexes the model under examination, and the superscript 0 represents the historical mean. The R^2_{OS} statistic measures the percentage reduction in terms of the mean squared forecast error for a model under examination relative to the historical average. Thus, intuitively, a positive value of the R^2_{OS} implies better predictive performance for model i than the historical mean. The higher the R^2_{OS} value, the better the out-of-sample predictive gains.

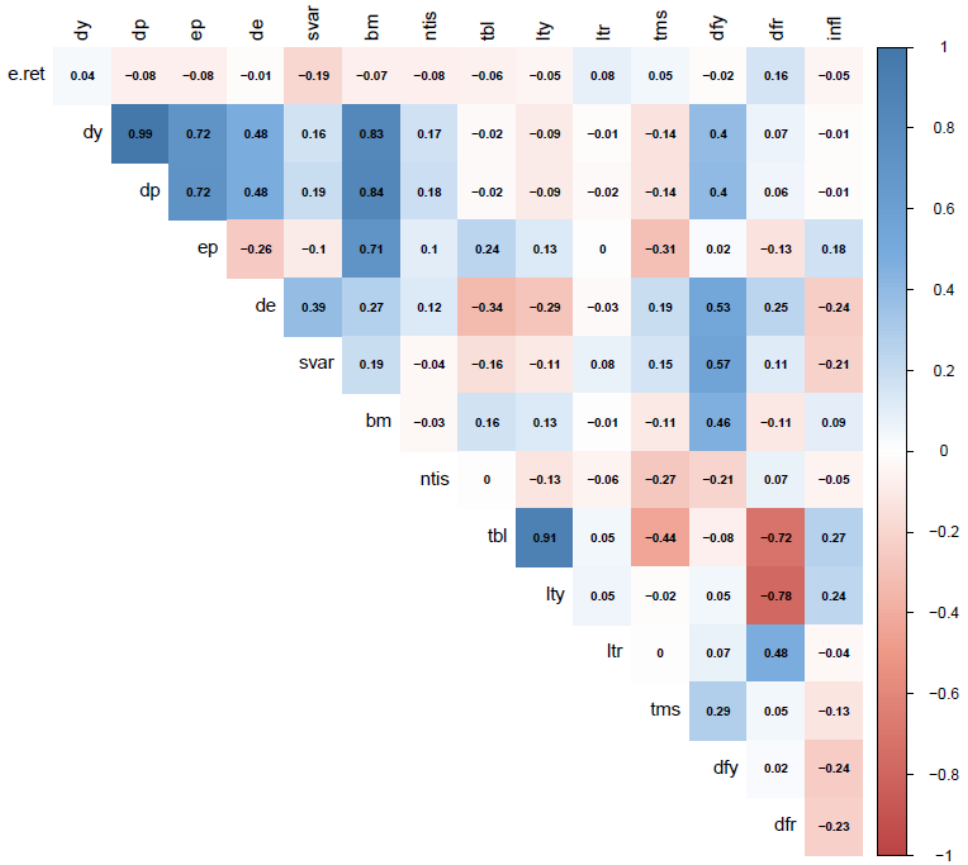

Figure 1. Monthly data sample correlation matrix plot

Note. Data sample runs from 1927:1 to 2016:12. Full descriptions of the variables are given in the data description section of this paper.

In addition, to better visually inspect the predictive gains over the entire evaluation sample, following the empirical methodology considered in Rapach et al. (2010), we use the cumulative differences in squared forecast errors (CDSFE) between the benchmark and the double-averaged model to construct a graphical device evaluating forecasts. Specifically,

$$CDSFE_t = \sum_{s=1}^{t}(r_{s+1} - \bar{r}_{s+1})^2 - \sum_{s=1}^{t}(r_{s+1} - \hat{r}_{s+1})^2, \tag{20}$$

where $\bar{r}_{s+1}$ is the one-step ahead prediction from model (6), and $\hat{r}_{s+1}$ is the one-period ahead point forecast from either model (1) or model (17). A positive value of $CDSFE_t$ indicates that the forecasting model under examination outperforms the benchmark by reporting a smaller value of MSFE as of time t.

Unlike related studies such as Rapach et al. (2010) and Dangl and Halling (2012), we do not base the significance of R^2_{OS} on the family of the MSFE-based test statistics proposed in Clark and West (2007). These tests are proposed under the assumption that forecasts are produced from a stationary environment, whereas here we explicitly allow for the possibility of structural breaks when constructing the double-averaged model. Hence, the notion of stationarity does not apply in our settings. Consequently, the critical values provided in Clark and West (2007) for testing the null hypothesis of equal predictive accuracy would not be statistically valid for our empirical results.

Table 1. Summary statistics

	Panel A: Monthly Data		Panel B: Quarterly Data		Panel C: Annual Data	
	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
ERET	0.006	0.055	0.011	0.112	0.042	0.194
DY	-3.362	0.458	-3.351	0.457	-3.315	0.447
DP	-3.367	0.460	-3.366	0.463	-3.372	0.468
EP	-2.733	0.417	-2.733	0.423	-2.738	0.410
DE	-0.634	0.331	-0.633	0.336	-0.634	0.321
SVAR	0.003	0.006	0.009	0.015	0.035	0.048
BM	0.572	0.265	0.577	0.267	0.565	0.262
NTIS	0.018	0.026	0.017	0.025	0.018	0.027
TBL	0.034	0.031	0.034	0.031	0.035	0.031
LTY	0.052	0.028	0.051	0.028	0.051	0.027
LTR	0.005	0.024	0.015	0.047	0.059	0.100
TMS	0.017	0.013	0.017	0.013	0.016	0.014
DFY	0.011	0.007	0.011	0.007	0.012	0.008
DFR	-0.046	0.034	-0.036	0.049	0.012	0.083
INFL	0.002	0.005	0.007	0.013	0.03	0.041

Note. This table reports summary statistics for the equity premium which is computed as returns on the Standard & Poor 500 portfolio minus the risk free rate, along with the predictive variables. The full sample runs from 1927 to 2016. The set of predictors examined are: ERET, equity premium or excess return; DY, dividend-yield; DP, dividend-price ratio; EP, earnings-price ratio; DE, dividend-payout ratio; SVAR, stock market variance; BM, book-to-market ratio; NTIS, net-equity expansion; TBL, Treasury bill rate; LTY, long-term yield; LTR, long-term return; TMS, term spread; DFY, default yield spread; DFR, default return spread; INFL, inflation.

3. Empirical Results

In this section, we present data and empirical results regarding statistical gains when forecasting the market equity premium with two-stage forecast combination.

3.1 Data Description

We use data on aggregate stock returns along with a set of 14 predictive variables. Our data come from an updated database maintained by Amit Goyal. Since the raw data samples vary substantially across individual regressors, to better compare results, we adopt the largest common sample from 1927 to 2016 in subsequent empirical analysis. All available data frequencies are analyzed.

Stock returns are measured as continuously compounded returns on the Standard and Poor's (S&P) value-weighted 500 index including dividends. To construct the series of equity premium, a 3-month Treasury bill rate is subtracted from stock returns. Turning to the set of predictive variables, it includes: 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 the sake of brevity, we refer the interested readers to Goyal and Welch (2008) for details regarding the identity and construction of these predictive variables.

Descriptive statistics for the equity premium along with 14 predictive variables across all available time horizons are reported in Table 1. On average, the U.S. market equity premium shows return rates of 0.6%, 1.1% and 4.2%, for monthly, quarterly and yearly data, respectively.

In light of the weak results documented in Goyal and Welch (2008), in Figure 1 we present a correlation matrix for the equity premium along with 14 predictors at the monthly horizon. Figure 1 shows that the dependent

variable, the U.S. market equity premium, is only weakly correlated with predictive variables. This could help explain the primary message conveyed in Goyal and Welch (2008) that forecasts from the simple bivariate models cannot reliably and consistently beat those from the prevailing mean. Moreover, some predictors, such as the dividend-price ratio (dp) and dividend-yield ratio (dy), are highly correlated with each other. For example, the correlation between the dividend-price ratio and the dividend-yield ratio is 0.99. While the correlation matrix is a simple statistical device, the insights it provides are profound. For instance, the poor performance of the OLS-estimated, "kitchen-sink" multivariate regression model comprising all available predictors in Goyal and Welch (2008), may be attributed to the fact that many regressors the model contains are highly correlated with each other as shown in Figure 1 for monthly data.

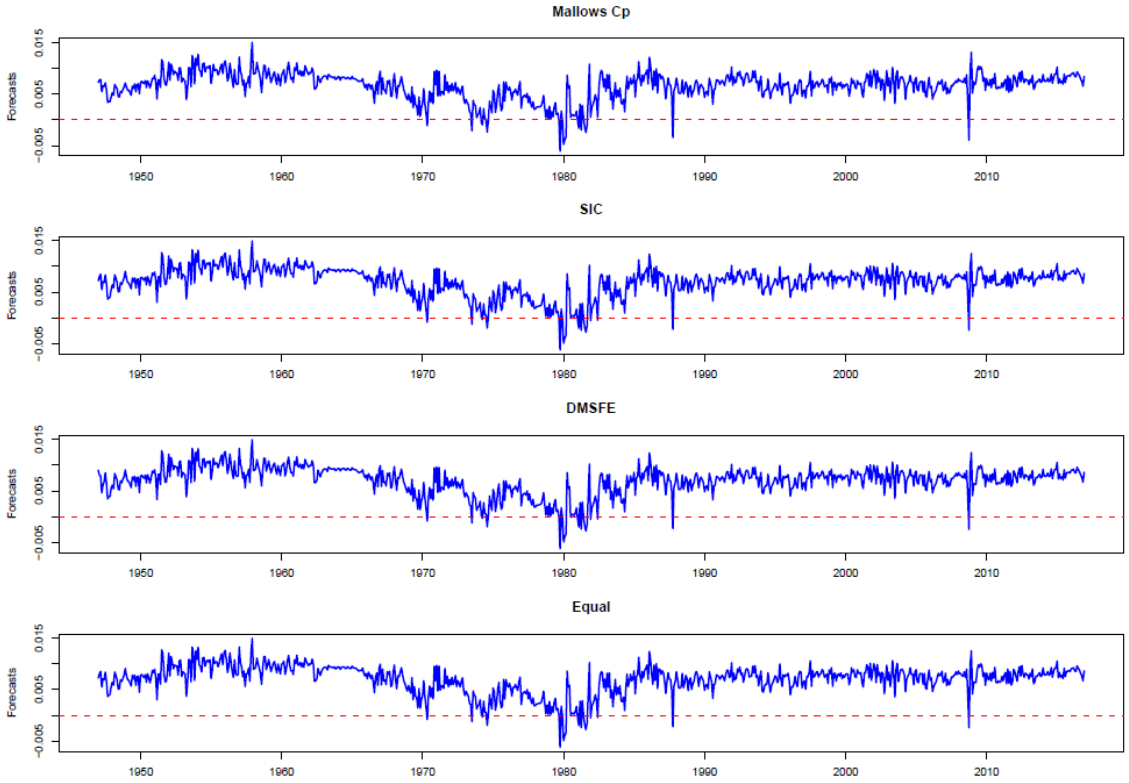

Figure 2. Equity premium forecasts with two-stage forecast combination

Note. Monthly forecasts generated by the recursive estimation window from January 1947 to December 2016. Each plot displays forecasts for a particular double-averaged model named by the weighting scheme used in the first stage of model construction.

3.2 Equity Premium Forecasts and Forecast Evaluation

3.2.1 Graphical Analysis

We start by providing a visual impression of the monthly out-of-sample forecasts. Figure 2 presents time-series plots of the monthly forecasts of the equity premium from January 1947 to December 2016 for the four two-stage forecast combination weights, namely, the Mallows (Cp), Schwarz information criterion (SIC), discounted mean squared forecast error (DMSFE) and equal weights. Overall, the patterns of all forecasts look quite similar to each other. Comparing with the results from bivariate regression models provided in Rapach et al. (2010), we find that forecasts from the two-stage model averaging are smoother and less volatile over the entire evaluation period. In addition, a closer examination of Figure 2 suggests that most volatilities of the equity premium forecasts concentrate on a few time periods, for instance, the oil shock of the mid-1970s, the Great Moderation of the mid-1980s, and the 2008-2009 global financial crisis. All of the aforementioned empirical features associated with the double-averaged model, namely, forecast stability and clustered volatility, are not reflected in the forecasts from bivariate models originally analyzed in Goyal and Welch (2008).

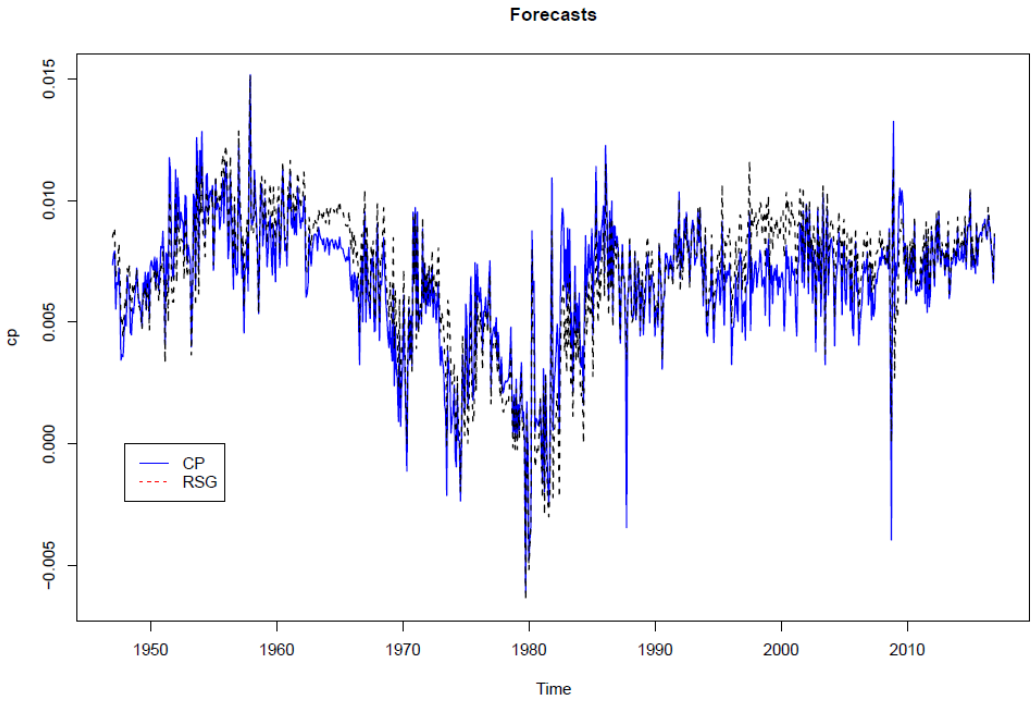

Figure 3. Comparison of equity premium forecasts

Note. Monthly forecasts from the double-averaged model with Mallows weights are represented by the solid line named CP, while those from the simple forecast combination are denoted by the dashed line titled RSG. All forecasts are generated by the recursive estimation window with data from January 1947 to December 2016.

Comparing with predictions from the simple forecast combination in Rapach et al. (2010), in Figure 3, we present time-series plots of the forecasts generated by the Mallows-weights based double-averaged model (solid line denoted by CP) and those from the simple one-step forecast combination (dashed line denoted by RSG) proposed in Rapach et al. (2010) for monthly data from 1947 to 2016. Figure 3 shows that forecasts from the double-averaged model are more volatile than those from the simple forecast combination during periods of financial crisis. For example, the forecasts from the double-averaged model are more volatile than the simple combination forecasts during the 2008-2009 financial crisis. This may be attributed to the fact that the two-stage forecast combination explicitly takes into account the possible presence of structural breaks which are more likely to occur during periods of economic crisis.

Next, we present the time-series plots of the cumulative differences between the squared forecast error of the historical mean and that from the two-stage forecast combination in Figures 4, 5 and 6, for monthly, quarterly and annual data, respectively, from 1947 to 2016.

Regarding monthly results, all panels in Figure 4 display CDSFE time series curves which are positively sloped for the most part of the evaluation window. All four weighting schemes demonstrate similar results. Two points are worth emphasizing here. First, all curves are strongly positively sloped roughly between 1970 and 1987, implying remarkably superior statistical performance for the double-averaged model relative to the benchmark. Note that some events, such as the oil shock and great moderation, occurred during this period of time. Second, the quality of forecasts from double-averaged models somehow deteriorates roughly between 2000 and 2008, however, they regain dominance over the benchmark after 2008. This phenomenon may be explained by the linkage between equity premium prediction and the real economy. Historically, Fama and French (1989) and Cochrane (1999) view that rising risk aversion during economic recessions requires a higher risk premium. Rapach et al. (2010) further argue that out-of-sample gains accruing to successful returns forecasts are often clustered in relatively extreme periods of economic expansion. Our approach reinforces their arguments by showing extraordinary gains relative to the benchmark during the 2008-2009 financial crisis, as reflected by a sharp jump in the curves right before the year 2010 in Figure 4.

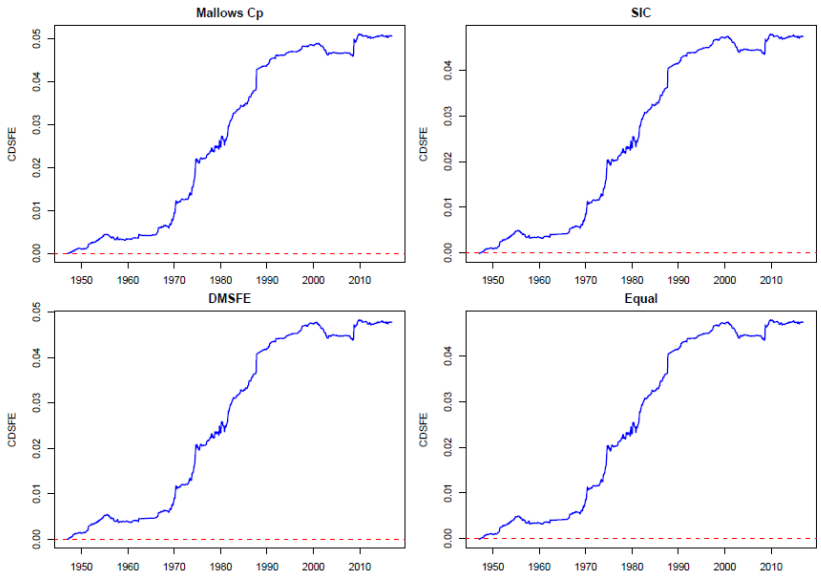

Figure 4. Monthly Cumulative Difference in Squared Forecast Error (CDSFE)

Note. At any time period, if the CDSFE curve moves up, it implies that the forecasting model outperforms the benchmark by having a smaller prediction error rate. Each plot displays CDSFE for a particular double-averaged model named by the weighting scheme used in the first stage of model construction. All forecasts are generated by the recursive window with data from 1947 to 2016.

Quarterly and annual results shown in Figures 5 and 6 are qualitatively similar to those revealed in monthly data: the two-stage forecast combination can lead to significant statistical gains during the 1970s, 1980s, and the 2008-2009 global financial crisis. In stark contrast to the results reported in Goyal and Welch (2008) and Rapach et al. (2010), generally, all CDSFE curves presented here are predominantly positively sloped over the entire evaluation period, especially for the monthly forecasts, suggesting that the two-stage forecast combination could deliver predictive gains on a considerably more consistent basis over time than the historical mean and bivariate predictive regressions.

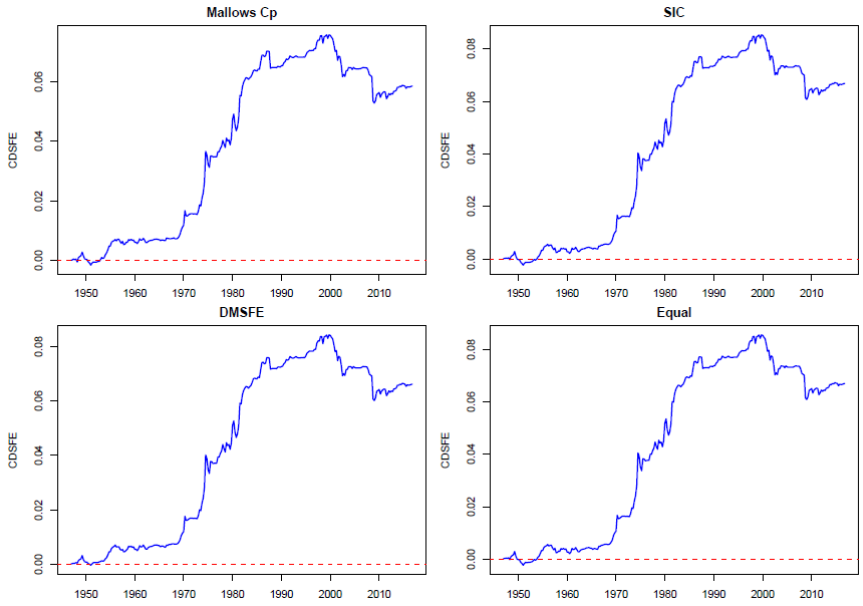

Figure 5. Quarterly Cumulative Difference in Squared Forecast Error (CDSFE)

Note. At any time period, if the CDSFE curve moves up, it implies that the forecasting model outperforms the benchmark by having a smaller prediction error rate. Each plot displays CDSFE for a particular double-averaged model named by the weighting scheme used in the first stage of model construction. All forecasts are generated by the recursive window with data from 1947 to 2016.

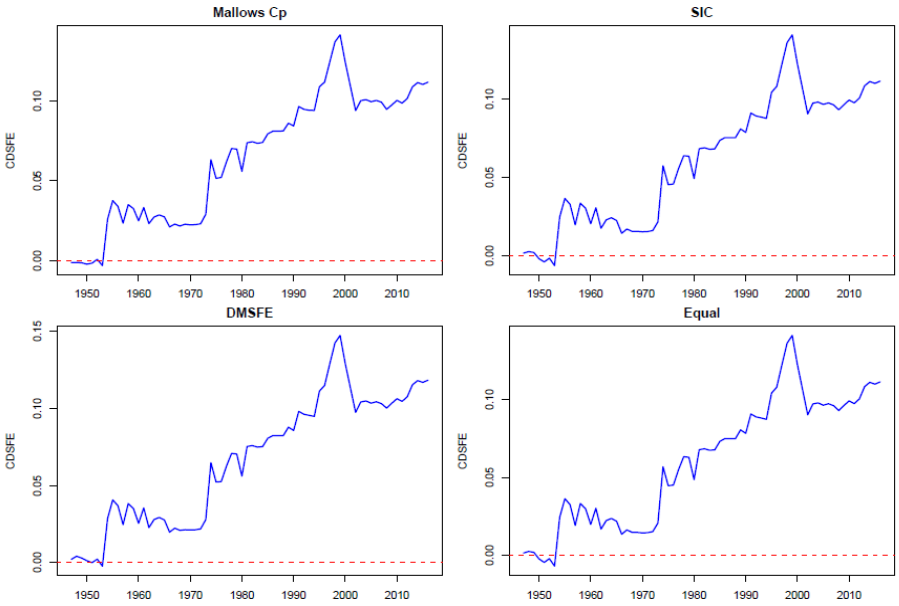

Figure 6. Yearly Cumulative Difference in Squared Forecast Error (CDSFE)

Note. At any time period, if the CDSFE curve moves up, it implies that the forecasting model outperforms the benchmark by having a smaller prediction error rate. Each plot displays CDSFE for a particular double-averaged model named by the weighting scheme used in the first stage of model construction. All forecasts are generated by the recursive window with data from 1947 to 2016.

3.2.2 Statistical Performance

In this section, we evaluate and compare the statistical gains based on the out-of-sample R_{OS}^2 statistic. We begin forecast 20 years after the first available observation, so the evaluation period runs from 1947 to 2016 for all data horizons. Furthermore, to check the robustness of all predictive models to the choice of out-of-sample split, we consider three subsamples beginning 40 years, 60 years and 80 years after the first available observation. All results are reported in Table 2.

In Table 2, in each panel of data frequency, we present empirical results for four evaluation periods. Within each panel, the first four columns present results from the two-stage forecast combination: Mallows weights, SIC weights, DMSFE weights and equal weights, denoted by CP, SIC, DMSFE and EQUAL, respectively. The fifth column, titled RSZ, reports out-of-sample performance of the simple forecast combination proposed in Rapach et al. (2010), that is, the forecasts constructed by simply averaging all 14 stable linear bivariate models in Eq. (1). For the remaining columns, they are out-of-sample results from several simple linear bivariate models originally considered in Goyal and Welch (2008), with each column titled by the name of the unique variable it contains.

We begin by examining the performance of all the bivariate models. Overall, the vast majority of the R_{OS}^2 is negative for all models across evaluation samples and forecast horizons. Only a few models report positive values of R_{OS}^2, but they are weak and inconsistent. This implies that the empirical performance of the simple bivariate models considered in Goyal and Welch (2008) is still weak and unreliable when predicting the equity premium even with extended data.

Next, we turn attention to the empirical performance of the simple combination proposed in Rapach et al. (2010). Our results show that the simple averaged forecast, which is the arithmetic mean of the forecasts of all 14 stable linear bivariate models, indeed improves upon bivariate predictions, as its R_{OS}^2 values are positive across all cases, with the exception of two instances with quarterly data. However, the improvement delivered by the simple combination is somewhat weak relative to the benchmark, as the gains in terms of the MSFE reduction is less than 0.1% for all cases.

Finally, in Table 2, our results show that the two-stage forecast combination significantly improves predictive performance when forecasting the equity premium, relative to not only the benchmark but also the simple forecast combination. Overall, all four weighting schemes have achieved more than 1% reduction in MSFE relative to the prevailing mean across all evaluation samples and horizons. For example, the maximum MSFE reduction, 10.661%, comes from the DMSFE method with annual data from 1967 to 2016, while the minimum MSFE reduction, 1.585%, is reported for the DMSFE method with quarterly data in the 1987-2016 subsample.

Turning to the comparison among the four weighting schemes within the two-stage model averaging framework, for monthly and quarterly forecasts, all SIC, DMSFE and EQUAL weights demonstrate similar performance over all subsamples. However, their performances become weaker when assessed at the end of the evaluation sample than those from the full sample. For example, the SIC method reports R^2_{OS} value of 3.267% when evaluated over the full monthly sample from 1947 to 2016, nevertheless, its R^2_{OS} value drops to 1.724% when evaluated over the smallest subsample period, 2007-2016. Surprisingly, Mallows weights demonstrate consistent gains across all subsamples, and seemingly are more robust to the subsample choice than the other three methods. As to yearly data, all four schemes exhibit even stronger forecasting performance than their counterparts in other horizons, as they achieve more than 4% reduction in MSFE across all evaluation samples.

Overall, our empirical results in Table 2 suggest that, when facing the uncertainty of structural breaks, the two-stage forecast combination could offer a robust way to significantly improve out-of-sample performance relative to many competing alternatives. Furthermore, our approach provides a clear linkage between structural break and model averaging, as break models are explicitly included in the construction of the double-averaged model.

Table 2. Out-of-sample forecast performance

Panel A: Monthly	CP	SIC	DMSFE	EQUAL	RSZ	DY	DP	EP	DE
1947-2016	3.491	3.267	3.302	3.266	0.027	-0.002	-0.053	-0.031	-0.008
1967-2016	3.742	3.756	3.766	3.758	0.031	-0.004	-0.023	-0.007	-0.008
1987-2016	1.814	1.787	1.927	1.789	0.012	-0.021	-0.030	-0.021	-0.004
2007-2016	2.563	1.724	1.942	1.722	0.008	-0.006	0.001	-0.005	-0.009
Panel B: Quarterly	CP	SIC	DMSFE	EQUAL	RSZ	DY	DP	EP	DE
1947-2016	3.550	4.056	4.010	4.069	0.052	-0.064	-0.195	-0.167	-0.016
1967-2016	6.170	7.338	7.527	7.374	0.061	-0.046	-0.085	-0.056	-0.014
1987-2016	3.371	2.000	1.585	1.996	-0.001	-0.121	-0.119	-0.138	-0.017
2007-2016	4.517	2.746	3.756	2.746	-0.013	-0.030	0.010	-0.147	-0.030
Panel C: Annual	CP	SIC	DMSFE	EQUAL	RSZ	DY	DP	EP	DE
1947-2016	5.747	5.744	6.091	5.720	0.073	-0.069	-0.442	-0.269	-0.060
1967-2016	6.581	9.506	10.661	9.595	0.085	-0.091	-0.217	-0.148	-0.015
1987-2016	4.951	4.391	4.716	4.378	0.061	-0.254	-0.312	-0.384	0.158
2007-2016	5.021	5.231	8.057	5.230	0.060	-0.051	0.103	-0.517	0.353

Note. This table reports the out-of-sample R^2_{OS} for the monthly, quarterly and annual equity premium forecasts. The R^2_{OS} of Campbell and Thompson (2008) measures the percent reduction in mean squared forecast error for a particular predictive model with name given in the first row of each panel relative to the benchmark. For each data frequency, we consider four forecast evaluation periods. In each panel, the first five columns report results for the forecast combinations while the rest are for simple linear bivariate models. CP: Mallows weights; SIC: Schwarz Information Criterion weights; DMSFE: discounted mean squared forecast error weights with discount factor $\theta = 1$; EQUAL: equal weights; RSZ: simple forecast combination considered in Rapach et al. (2010); DY, dividend-yield; DP, dividend-price ratio; EP, earnings-price ratio; DE, dividend-payout ratio.

3.2.3 Discussion

Rapach et al. (2010) conclude that forecast combination appears useful in equity prediction because it can sizably reduce predictive variance while including information from all available economic and financial predictors. They also suggest that the efficacy of forecast combination ultimately comes from the highly complex and constantly evolving environment generating the equity returns. These authors claim that combining forecasts by using schemes such as equal weights could consistently outperform the empirically hard-to-beat historical average.

Our main contribution is to show that, the two-stage forecast combination can substantially further improve the out-of-sample forecasting performance in terms of statistical gains consistently compared with the simple forecast combination. Not only can we obtain more statistical gains than the empirically difficult-to-beat historical mean, but also our approach is shown to be empirically superior to the simple, one-step forecast combination based on arithmetic mean originally analyzed in Rapach et al. (2010).

In addition, our empirical results are robust to the evaluation sample choice. The gains could come from the fact that the two-stage forecast combination explicitly accounts for the possible presence of parameter instabilities in predictive models. Comparing with the results reported in Rapach et al. (2010) and Goyal and Welch (2008), our approach explicitly addresses the uncertainty regarding both model selection and parameter instability.

Comparing our results with those in related studies such as Rapach et al. (2010) and Pettenuzzo et al. (2014), first we have explicitly provided the linkage between forecast combination and parameter instability as break specifications are included in the double-averaged model, thus better explaining the efficacy of model averaging in a non-stationary environment on solid theoretical grounds. Second, our methodology of two-stage forecast combination can be easily implemented in empirical applications following conventional econometric procedure in contrast to complex methods such as Bayesian model averaging which requires imposing priors on both candidate models and parameters.

3.3 Expansion and Recession Analysis

Following related literature such as Rapach et al. (2010) and Pettenuzzo et al. (2014), we are interested in examining the forecasting performance of the two-stage combination by investigating separately for the expansion and recession periods as defined by the National Bureau of Economic Research (NBER) indicator for monthly and quarterly data. For the sake of brevity, we only compare and evaluate the performance of the four weighting schemes within two-stage forecast combination. The results are reported in Table 3.

Consistent with similar findings documented in related studies such as Pettenuzzo et al. (2014) and Li and Tsiakas (2017), all four weighting schemes perform better during recessions than during expansions in terms of the R_{OS}^2 statistic. For monthly forecasts, all methods achieve about 2.6% in MSFE reduction relative to the prevailing mean during expansions. However, they obtain more than 5% gains in terms of MSFE reduction during recessions, with the Mallows weights achieving the largest reduction of about 5.5%.

Table 3. Recession-expansion analysis

Panel A: Monthly Data	CP	SIC	DMSFE	EQUAL
Recession	5.507	5.079	5.091	5.076
Expansion	2.763	2.613	2.657	2.612
Panel B: Quarterly Data	CP	SIC	DMSFE	EQUAL
Recession	5.225	5.767	5.598	5.786
Expansion	2.684	3.171	3.189	3.181

Note. This table reports the out-of-sample R_{OS}^2 for four double-averaged models during recession and expansion periods defined by the National Bureau of Economic Research indicator for monthly and quarterly data. The R_{OS}^2 measures the percent reduction in mean squared forecast error for a particular predictive model with name given in the first row of each panel relative to the historical average benchmark. CP: Mallows weights; SIC: Schwarz Information Criterion weights; DMSFE: discounted mean squared forecast error weights with discount factor $\theta = 1$; EQUAL: equal weights. The evaluation period runs from 1947 to 2016.

For quarterly forecasts, almost all weighting methods uniformly improve upon their counterparts in monthly prediction by achieving more statistical gains in terms of MSFE reduction. Nevertheless, the empirical fact that all weighting schemes perform better during recessions than during expansions remains.

4. Economic Value Evaluation

The economic value of equity-related forecasts is frequently evaluated based on the portfolio returns generated by predictive regressions. Following related studies such as Baetje and Menkhoff (2016), in this section we assess the economic value of the excess returns forecasts in the context of the optimal portfolio decision of a mean-variance investor. In each period, the investor rebalances a portfolio comprising a risky asset, the S&P 500 index approximating the aggregate equity returns, and a risk-free asset, the 3-month Treasury bill.

4.1 Framework

In each period *t*, the investor solves the problem of optimal portfolio decision,

$$\max_{w_t} E[U(w_t, r_{t+1})|I_t], \tag{21}$$

where $U(w_t, r_{t+1})$ is the mean-variance utility function depending on weight w_t for the equity. I_t is the information set available at time *t*. The solution to the optimal portfolio decision problem is:

$$w_t^* = \frac{r_{t+1} - rf_{t+1}}{\gamma \sigma_t^2}, \tag{22}$$

where rf_{t+1} is the risk-free rate for period $t + 1$ and is known in time *t*, γ is the coefficient of relative risk aversion, r_{t+1} is the risky asset returns in period $t + 1$, and σ_t^2 is the variance of the risky asset returns at time *t*. Empirically, we use one-step ahead forecast $\hat{r}_{t+1}$ from predictive models to replace r_{t+1}. Turning to the unknown variance of risky asset, σ_t^2, we can estimate it based on all the available data up to time *t* to obtain $\hat{\sigma}_t^2$ following the conventional statistical procedure.

Next, we compute the realized portfolio returns for each period in the evaluation sample according to

$$p_{t+1} = w_t^* r_{t+1} + (1 - w_t^*) r f_{t+1}. \quad (23)$$

When evaluating economic significance, we use the certainty equivalent return (CER) to measure the economic value of various forecast combination methods. Specifically,

$$CER = \bar{p} - \frac{\gamma}{2}\bar{\sigma}_p^2, \quad (24)$$

where $\bar{p}$ and $\bar{\sigma}_p^2$ are the sample average and sample variance of the portfolio returns, respectively.

The relative CER can be taken as a portfolio management fee that a mean-variance investor with a relative risk-aversion coefficient γ, is willing to pay to gain access to the information embedded in the averaged model rather than that in the historical average. In what follows, all reported values of CER are annualized, so that they can be understood as an annualized percentage portfolio management fee.

In addition, we also assess the economic performance by Sharpe ratio, which is a widely used measure to evaluate portfolio performance, and is defined as the average equity premium of a portfolio divided by its standard deviation.

Table 4. Annualized certainty equivalent returns

Panel A: Monthly Data	CP	SIC	DMSFE	EQUAL
1947-2016	0.639	0.617	0.621	0.617
1967-2016	0.803	0.821	0.824	0.822
1987-2016	0.423	0.410	0.431	0.410
2007-2016	0.679	0.499	0.540	0.499
Panel B: Quarterly Data	CP	SIC	DMSFE	EQUAL
1947-2016	0.294	0.344	0.337	0.345
1967-2016	0.634	0.736	0.746	0.740
1987-2016	0.324	0.195	0.159	0.194
2007-2016	0.559	0.345	0.451	0.345
Panel C: Annual Data	CP	SIC	DMSFE	EQUAL
1947-2016	0.426	0.428	0.449	0.427
1967-2016	0.545	0.812	0.884	0.820
1987-2016	0.217	0.253	0.258	0.254
2007-2016	0.612	0.641	0.921	0.641

Note. This table reports certainty equivalent return (CER) values in percentage for portfolios based on recursive out-of-sample forecasts of the equity premium using the two-stage forecast combination approach. CER values are annualized and measured relative to the prevailing mean benchmark, and can be interpreted as an annualized percentage portfolio management fee. For each data frequency, we consider four forecast evaluation periods. The coefficient of the relative risk aversion is set at five. Each model is named by the weighting scheme used in the first stage of constructing forecasts. CP: Mallows weights; SIC: Schwarz Information Criterion weights; DMSFE: discounted mean squared forecast error weights with discount factor θ=1; EQUAL: equal weights.

4.2 Economic Performance

Table 4 reports the annualized CER gains in percentage for all double-averaged models relative to investing based on the historical average. In all cases, the relative risk aversion coefficient is $\gamma = 5$. The full sample spans from 1947 to 2016. Moreover, we also consider three subsamples starting in 1967, 1987 and 2007, respectively. We investigate the economic performance for all four weighting schemes when combining the stable and break models within the framework of two-stage forecast combination. Empirical results are provided for all data horizons.

For monthly results shown in Panel A of Table 4, all four weighting methods have achieved more than 0.6% CER gains relative to the prevailing mean over the entire evaluation sample. The CP reports the largest gains of 0.639%. The economic gains also carry over to three subsamples. Moreover, there is no empirical evidence suggesting that the economic benefits obtained from forecast combination deteriorate over time, particularly for the model based on Mallows weights. For example, the smallest subsample beginning in 2007 reports a higher relative CER value of 0.679% than that from the full sample. Furthermore, Mallows weights deliver better economic gains than other methods when forecasting monthly returns following the 2008-2009 financial crisis. At the quarterly horizon shown in Panel B of Table 4, all forecast combination methods deliver superior economic benefits relative to the benchmark across all samples. However, for each method at a given period, the economic gains are uniformly smaller than their counterparts in Panel A with monthly data. Finally, at the annual

horizon, like previous results, all combination methods report positive economic gains consistently over time. The DMSFE model reports the largest CER gains of 0.921% when forecasting the last 10 years of annual returns.

Table 5 reports the values of annualized Sharpe ratio for all weighting schemes within the two-stage model averaging framework, across different horizons of data and evaluation samples. Overall, the results in Table 5 qualitatively confirm our previous findings: (1) the two-stage model averaging can consistently deliver superior economic gains relative to the historical mean; (2) longer-horizon returns seem harder to predict than shorter-horizon returns; and (3) the DMSFE weights seem to forecast long-term returns better than other schemes.

Table 5. Annualized sharpe ratio

Panel A: Monthly Data	CP	SIC	DMSFE	EQUAL
1947-2016	0.432	0.406	0.409	0.406
1967-2016	0.486	0.494	0.496	0.494
1987-2016	0.238	0.235	0.244	0.235
2007-2016	0.348	0.283	0.302	0.282
Panel B: Quarterly Data	CP	SIC	DMSFE	EQUAL
1947-2016	0.224	0.251	0.248	0.251
1967-2016	0.444	0.478	0.480	0.479
1987-2016	0.203	0.124	0.103	0.124
2007-2016	0.395	0.265	0.372	0.266
Panel C: Annual Data	CP	SIC	DMSFE	EQUAL
1947-2016	0.143	0.147	0.150	0.147
1967-2016	0.166	0.223	0.239	0.225
1987-2016	0.067	0.071	0.072	0.071
2007-2016	0.157	0.234	0.405	0.236

Note. This table reports the annualized Sharpe ratio values for portfolios based on recursive out-of-sample forecasts of the equity premium using the two-stage forecast combination approach. The numbers are Sharpe ratio gains for trading strategies timing the market relative to the historical average benchmark. For each data frequency, we consider four forecast evaluation periods. Each model is named by the weighting scheme used in the first stage of constructing forecasts. CP: Mallows weights; SIC: Schwarz Information Criterion weights; DMSFE: discounted mean squared forecast error weights with discount factor θ=1; EQUAL: equal weights.

5. Conclusion

This paper has extended the methodology of forecast combination when predicting the equity premium out-of-sample in a non-stationary environment. To address the uncertainty on both model selection and parameter instability, we propose a two-stage forecast combination approach: first, we combine the break and stable specifications of each baseline model using one of the four proposed weighting schemes, namely, equal weights, DMSFE weights, SIC weights and Mallows weights; next, we combine all averaged models obtained from the previous step to form a double-averaged model using equal weights. Empirically we apply the two-stage forecast combination to forecasting the aggregate equity premium in the context of our-of-sample analysis. Our main finding is that the double-averaged model, particularly the one based on Mallows weights, could potentially deliver significant statistical and economic gains relative to the simple forecast combination considered in Rapach et al. (2010), and the historical average.

Our methodology is restricted to the context of combining linear models, thus it would be greatly desirable to broaden the scope by considering more diverse model types. Another unexplored issue is statistical inference. As of the writing of this paper, to the best of our knowledge it is unclear how to rigorously test the significance of the R^2_{OS} for averaged models in a non-stationary environment, as the relaxation of stationarity complicates the derivation of asymptotic results. This is a challenging yet important topic for future investigation.

Acknowledgments

The author is grateful to the editor, two anonymous referees, Helle Bunzel, Gray Calhoun, and all seminar participants at Texas A&M International University for helpful comments and suggestions. All errors are mine.

References

Ait-Sahali, Y., & Brandt, M. W. (2001). Variable selection for portfolio choice. *The Journal of Finance, 56*, 1297-1351. https://doi.org/10.1111/0022-1082.00369

Andrews, D. W. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica, 61*(04), 821-856. https://doi.org/10.2307/2951764

Avramov, D., & Wermers, R. (2006). Investing in mutual funds when returns are predictable. *Journal of Financial Economics, 81*, 339-377. https://doi.org/10.1016/j.jfineco.2005.05.010

Baetje, F., & Menkhoff, L. (2016). Equity premium prediction: Are economic and technical indicators unstable? *International Journal of Forecasting, 32*, 1193-1207. https://doi.org/10.1016/j.ijforecast.2016.02.006

Campbell, J. Y. (1987). Stock returns and the term structure. *Journal of Financial Economics, 18*, 373-399. https://doi.org/10.1016/0304-405X(87)90045-6

Campbell, J. Y., & Shiller, R. J. (1988). The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies, 1*, 195-228. https://doi.org/10.1093/rfs/1.3.195

Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: can anything beat the historical average? *Review of Financial Studies, 21*(04), 1509-1531. https://doi.org/10.1093/rfs/hhm055

Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics, 138*, 291-311. https://doi.org/10.1016/j.jeconom.2006.05.023

Cochrane, J. H. (1999). New facts in fiance. *Economic Perspectives, 23*, 36-58.

Dangl, T., & Halling, M. (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics, 106*, 157-181. https://doi.org/10.1016/j.jfineco.2012.04.003

Fama, E. F., & French, K. R. (1988). Dividend yields and expected stock returns. *Journal of Financial Economics, 22*, 3-14. https://doi.org/10.1016/0304-405X(88)90020-7

Fama, E. F., & French, K. R. (1989). Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics, 25*, 23-49. https://doi.org/10.1016/0304-405X(89)90095-0

Ferreira, M. A., & Santa-Clara, P. (2011). Forecasting stock market returns: the sum of the parts is more than the whole. *Journal of Financial Economics, 100*, 514-537. https://doi.org/10.1016/j.jfineco.2011.02.003

Goyal, A., & Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies, 21*(04), 1455-1508. https://doi.org/10.1093/rfs/hhm014

Hansen, B. E. (2009). Averaging estimators for regressions with a possible structural break. *Econometric Theory, 25*(06), 1498-1514. https://doi.org/10.1017/S0266466609990235

Li, J., & Tsiakas, I. (2017). Equity premium prediction: the role of economic and statistical constraints. *Journal of Financial Markets, 36*, 56-75. https://doi.org/10.1016/j.finmar.2016.09.001

Paye, B., & Timmermann, A. (2006). Instability of return prediction models. *Journal of Empirical Finance, 13*(03), 274-315. https://doi.org/10.1016/j.jempfin.2005.11.001

Pettenuzzo, D., Timmermann, A., & Rossen, V. (2014). Forecasting stock returns under economic constraints. *Journal of Financial Economics, 114*, 517-553. https://doi.org/10.1016/j.jfineco.2014.07.015

Rapach, D. E., & Wohar, M. E. (2006). Structural breaks and predictive regression models of aggregate us stock returns. *Journal of Financial Econometrics, 4*, 238-274. https://doi.org/10.1093/jjfinec/nbj008

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*, 821-862. https://doi.org/10.1093/rfs/hhp063

Stock, J. H., & Watson, M. W. (2003). Forecasting output and inflation: the role of asset prices. *Journal of Economic Literature, 41*, 788-829. https://doi.org/10.1257/002205103322436197

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting, 23*, 405-430. https://doi.org/10.1002/for.928