Elusive Return Predictability*

Allan Timmermann
University of California San Diego

June 29, 2007

Abstract

Investors' search for successful forecasting models leads the data generating process for financial returns to change over time which means that individual return forecasting models can at best hope to uncover evidence of 'local' predictability. We illustrate this point on a suite of forecasting models used to predict US stock returns and propose an adaptive forecast combination approach. Most of the time the forecasting models perform rather poorly, but there is evidence of relatively short-lived periods with modest return predictability. The short duration of the episodes where return predictability appears to be present and the relatively weak degree of predictability even during such periods makes predicting returns an extraordinarily challenging task.

Keywords: Out-of-sample forecasting performance, predictability of stock returns, creative self-destruction, adaptive forecast combination

1 Introduction

The possibility of predicting stock market returns has fascinated professional investors, laymen and academics for decades. The fact that the quest is still continuing indicates just how difficult it is to predict returns. Adding further to this challenge, forecasters of stock returns face a moving target that is constantly changing over time. Just when a forecaster may think that he has figured out how to predict returns, the dynamics of market prices will, in all likelihood, have moved on−possibly as a consequence of the forecaster's own efforts.

Stock prices are formed as a result of the complex interaction between heterogenous groups of investors. At one extreme is highly informed, technically sophisticated professional investors with access to substantial capital reserves and financial leverage. At the other extreme is essentially uninformed individuals whose trades may reflect liquidity needs or Keynesian animal spirits. Returns arise as the change in prices between adjacent dates and hence reflect revisions

[*] I wish to thank participants at the IIF conference in New York, June 2007, for many helpful comments.

in investors' beliefs caused by the arrival and interpretation of new information, changes in liquidity needs and the resulting interaction among traders.

Even the most sophisticated investors have to deal with the basic premise that they are attempting to price assets and predict returns using forecasting models that at best can be viewed as local approximations to a complicated and evolving market situation. Faced with such challenges, investors are led to constantly search across several competing forecasting approaches and investment strategies.

At any given point in time, investors explore a variety of approaches to forecast returns and so different forecasting methods effectively compete against each other. Moreover, the more successful a particular forecasting approach has been in recent times, the more likely it is to have been detected and adopted by a wider group of investors. Once enough investors adopt a particular forecasting approach−or a set of closely correlated approaches−and put substantial money behind it, we would expect their forecasts to start having a price impact: When the approach predicts an asset to have unusually high future returns, this will lead investors to acquire the underlying asset, thus pushing up its price during the current period so that most, if not all, of the predicted future return effectively gets incorporated in the current price.[1]

We would expect competition between a multitude of forecasting methods to cause instability both in the parameter estimates associated with particular forecasting models and also in their (relative) forecasting performance.[2] Indeed, the performance of individual forecasting methods may follow a life cycle pattern. Before a particular forecasting approach is widely discovered and adopted, it may perform quite well. Then, given a suitably long historical track record indicating good performance, the forecasting method will become more broadly adopted. Finally, as this learning and adoption process gets more complete and the information in the forecasts gets incorporated into prices, the method will cease to predict future return movements.[3]

How long this process of "creative self-destruction" involving model identification and adoption by market participants takes depends on several factors. Clearly the strength of the underlying prediction signal matters. If this is weak, it will be difficult for investors to identify profit-making opportunities. In addition the duration of the profit-making opportunity−which itself is endogenous−is important. If this is relatively short, again it speaks against the likelihood that investors can successfully implement strategies that exploit any

[1] The dynamics described here is consistent with the adaptive markets hypothesis in Lo (2004), but does not necessarily require that investors are boundedly rational.

[2] We intentionally refer to forecasting methods rather than models, the former comprising aspects of the forecasting cycle such as the model specification, estimation technique and choice of estimation window.

[3] The life cycle process does not rule out that a particular forecasting approach or state variable is useful on more than one occasion. It is quite possible that an approach works more than once, e.g. if a similar macroeconomic state repeats itself such as in the case of the oil price shocks that occurred around 1974, 1979 or more recently in 2005. The length of the life cycle might get shortened somewhat, however, if investors can pool recent data with data from similar historical events which will speed up the learning process.

predictable patterns in returns.

The process whereby investors search for the best forecasting methods is constantly perturbed due to the regular occurrence of outside shocks to the financial markets and the economy, reflecting changes in institutions (e.g. fiscal or monetary policy), large macroeconomic shocks, introduction of new financial instruments, changes in trading practices and the arrival of new types of investors (e.g. private equity and hedge funds).

Identifying intermittent predictive components in stock returns is difficult even by means of the best modern forecasting tools and is subject to an important trade-off. Highly adaptive methods that are capable of rapidly identifying prediction opportunities are also likely to be sensitive to outliers and hence will be subject to the 'false alarm' problem associated with type I errors in statistical inference - in this case wrongly identifying predictability during periods where it is genuinely absent. 'Spurious predictability' becomes a real concern in this situation. Conversely, less adaptive methods may miss short-lived episodes of predictability altogether.

This somewhat stylized picture ignores a number of complicating factors. First, our discussion assumes that the predicted return represents a profit making opportunity and not simply compensation for holding systematic risk and thus a fair risk premium. Put differently, our discussion is intended to apply to predictable return components that are not simply time-varying risk premia. Distinguishing between the two can be difficult in practice and requires modeling and identifying the source of variations in risk premia. Indeed, if the methods worked over long spells of time, investors were either not paying attention to them (which is unlikely) or the predicted component must represent a risk premium.

Second, given the substantial uncertainty surrounding which forecasting method to use at any point in time, investors are likely to adopt more than one approach. This may take the form of forecast combination strategies or Bayesian model averaging. Our discussion applies to this situation, although the more general issue then becomes how large a weight investors assign to different forecasting methods and how much this varies over time.

Important implications for return predictability follow from this discussion. Given the intense competition among informed investors acting in highly liquid financial markets, we would expect to find that most forecasting methods have a poor out-of-sample forecasting performance on "average" if estimated over long spans of time.

However, a poor "average" track record need not hold uniformly through time. There may be periods of time where one or more approaches work well, but these spells are likely to be fairly short-lived. Indeed, the identity of the "best" forecasting method can be expected to vary over time and there are likely to be periods of "model breakdown" where no approach seems to work.

In what follows, Section 2 briefly reviews the existing evidence of predictability of stock returns and instability in return forecasting models. Section 3 conducts an empirical analysis of US stock market returns that we use to illustrate the points discussed so far. Section 4 presents an alternative adaptive forecast combination approach, while Section 5 concludes.

2 Evidence of Return Predictability

Beginning with a series of academic studies published in the eighties, evidence emerged that stock returns were predictable, see e.g. Fama and Schwert (1981), Campbell (1987), Campbell and Shiller (1988) and Fama and French (1988, 1989). This early literature was mostly concerned with the presence of *ex-post* or in-sample return predictability by means of linear time-series models using predictor variables such as the dividend yield, the price-earnings ratio, interest rates, default premia or macroeconomic variables such as inflation.

Many of the predictor variables proposed in the early literature were subsequently noticed not to produce good predictions during the bull market that characterized a large part of the 1990s. Indeed, as noted by Lettau and Ludvigsson (2001), Schwert (2002) and others, return prediction models based on valuation ratios such as the dividend yield seemed to break down. During this period where stock prices soared, the dividend yield systematically drifted downwards, thus generating a negative sample correlation between returns and the dividend yield, in stark contrast with their positive historical association.[4]

Studies of *ex-ante* (out-of-sample) return predictability have found either that return predictability is confined to particular sub-samples (Pesaran and Timmermann (1995)) or that it is largely absent (Bossaerts and Hillion (1999)). Goyal and Welch (2003, 2006) have recently gone as far as arguing that none of the conventional predictor variables proposed in the literature on stock return predictability seems capable of systematically predicting stock returns out-of-sample, i.e. after accounting for parameter estimation error.

This conclusion has been disputed by, inter alia, Campbell and Thompson (2006) and Cochrane (2006) and the debate is currently not settled. One point is certain, however: Caught between the twin challenges of low predictive power and unstable regression coefficients, standard forecasting models find it difficult to consistently predict stock returns over long sample periods.

2.1 Predictability and Investment Horizon

Evidence of return predictability is particularly important, economically speaking, the shorter the forecast horizon. Intuitively this is easy to see: the shorter the time interval, the more times a trading strategy can be implemented to take advantage of any return predictability and so the greater the potential for high annualized returns. On economic grounds we would therefore expect the strength of return predictability to weaken, the shorter the forecast horizon.

Indeed, the evidence of return predictability at short horizons such as one day is generally very weak. There is some evidence of negative serial correlation

[4] Paye and Timmermann (2006) and Rapach and Wohar (2006) report broad evidence of breaks in a wide range of models used to predict stock returns.

in high frequency returns. However, this is likely a reflection of market microstructure effects such as the bid-ask bounce. Consistent with the notion that such effects account for the mean reversion observed in prices at high frequencies, Avramov, Chordia and Goyal (2006) find that mean reversion at the daily interval is strongest for the least liquid stocks. Their evidence indicates that return predictability at high frequencies such as one day is not strong enough to be exploited by regular traders.

The evidence of return predictability and mean reversion at horizons such as one year or longer is not clear-cut. However, given the lack of opportunities for repeating a particular investment strategy, it is clearly very risky to base investments on strategies tracking return behavior at very long horizons, should the payoffs fall short of expectations.

2.2 Instability of Return Forecasting Models

There are many reasons to expect the relation between predictor variables and asset returns to vary over time. First, incomplete learning effects are likely to play a role. If financial markets are not in a steady state but constantly get perturbed, investors' learning process will never converge. When investors act on evidence of predictability, we would expect a successful model adopted by one investor (e.g. a mutual fund, hedge fund or some other investment vehicle) fairly quickly to be adopted by other investors through diffusion of information. Depending on the functional form mapping investors' forecasts to asset prices, this can induce serial correlation and volatility clustering in returns.[5]

Second, structural changes in the underlying data generating process for returns may ensue as a reflection of increased participation in stock markets, availability of new low-cost investment vehicles such as index funds, exchange traded funds (ETFs), extended trading opportunities (e.g. after-hours trading through electronic communication networks) and lower transaction costs in many markets, including derivatives markets (e.g. options).

These considerations appear to be important in practice. In many cases, previous 'anomalies' have disappeared after their existence became publicized. For example, the evidence of a small-firm effect, originally documented in the early eighties appears to have weakened in subsequent sample periods (Schwert (2002)) as has the evidence of predictability of stock market returns by means of variables such as the inflation rate (Fama and Schwert (1981)) or the dividend yield (Fama and French (1988)).

As a further example, Sullivan, Timmermann and White (1999) find that the apparently superior performance of a range of technical trading rules reported by Brock, Lakonishok and LeBaron (1992) disappeared in the period after their publication. This may be a coincidence, but it may also reflect that the Brock et al study was published following a period where such rules performed well. After this became public knowledge, the rules gained more widespread use and ceased to continue to have predictive power over future returns.[6]

[5] See Guidolin and Timmermann (2007) in the context of a simple binomial model.

[6] The obvious alternative to this explanation, which is the subject of the study by Sullivan

As a final example of how predictability evolves over time and how this can be caused by exogenous events, some studies identified oil prices as having predictive power over stock returns during the period surrounding the oil price shocks in 1973 and 1974. Prior to these events it would have been difficult for investors to identify oil prices as a predictor variable of stock returns since oil prices were fluctuating less freely. While the oil price became an important macroeconomic state variable during the seventies, its significance has vanished in subsequent periods, see Pesaran and Timmermann (2000).

3 Empirical Application to US Stock Returns

To illustrate the earlier points we next turn to an empirical application. We are interested in forecasting monthly returns in the US stock market. To this end we use returns data going back to 1959:12 and ending in 2005:12. Three return series are considered, namely the return on value- and equal-weighted portfolios of US stocks and the return differential between small and big shares (as measured by their market capitalization), i.e. the SMB spread portfolio studied by, inter alia, Fama and French (1992).

Our data source for stock returns is the Center for Research in Security Prices (CRSP) at the University of Chicago. Returns on the SMB portfolio are obtained from Ken French's web site.

We use data from 1959:12 to 1969:12 as the initial estimation sample and retain the period from 1970:01 to 2005:12 as an out-of-sample evaluation period. Two estimation approaches are considered. The first "expanding window" approach uses recursive estimation starting with data from 1959:12 up to the time of the forecast to generate a series of one-step-ahead forecasts. Thus, the first forecast is generated for 1970:01, using data from 1959:12 to 1969:12. The following month (January 1970), the data window is expanded to also include 1970:01, the parameters of the forecasting models are re-estimated and then used to predict stock returns for 1970:02 and so forth up to the end of the sample.

The second, rolling window, approach uses a fixed-length window of the most recent ten years of data (120 monthly observations) to estimate the parameters of the forecasting models and then predicts returns next period conditional on those parameter estimates.

3.1 Forecasting Models

We consider a suite of eleven forecasting models.[7]

The first model simply uses the prevailing mean, i.e.

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

et al. (1999) is that the apparent success of the technical trading rules was the outcome of data-snooping.

[7] These were selected as a subset of the models considered by Elliott and Timmermann (2007). I am grateful to Gray Calhoun for research assistance with the empirical analysis.

Here and elsewhere the error term, ε_{t+1}, is treated as white noise whose mean cannot be predicted. The return forecast for period $t+1$ produced at time t is thus given by $t^{-1}\sum_{\tau=1}^{t} r_\tau$ under the recursive approach or by $\tau_0^{-1}\sum_{\tau=t-\tau_0+1}^{t} r_\tau$ under the rolling window approach, where $\tau_0 = 120$ is the window length. The recursively estimated prevailing mean is the benchmark model considered by Goyal and Welch (2003, 2006). When estimated using an expanding window, this model essentially assumes 'no predictability' (i.e. a constant mean), while under the rolling estimation window it incorporates a slowly changing mean.

The second forecasting model is an autoregressive (AR) specification

$$r_{t+1} = \beta_0 + \sum_{j=1}^{k} \beta_j r_{t+1-j} + \varepsilon_{t+1}, \tag{2}$$

where k is selected to minimize the Bayes Information Criterion.

The third model is a factor-augmented AR specification which in addition to autoregressive terms considers the inclusion of a set of common factors:

$$r_{t+1} = \beta_0 + \sum_{j=1}^{k} \alpha_j r_{t+1-j} + \sum_{j=1}^{q} \beta_j \psi_{j,t} + \varepsilon_{t+1}, \tag{3}$$

where $\psi_{j,t}$ is the jth factor and k and q are again selected to minimize the BIC. Factors are obtained by adopting the principal components approach of Stock and Watson (2002) to a cross-section of 131 macroeconomic time series which begin in 1960. The factors are extracted in real time using either a recursive or a rolling 10-year estimation window. Since these macroeconomic data only go as far as 2003:12, our forecasts from this model stop at this date.

In view of the long literature that models stock returns as having a slowly moving highly persistent component (e.g. Fama and French (1989)), it is natural to consider simple adaptive approaches so two smoothing methods constitute models four and five. Under the exponential smoothing approach the forecast, f_{t+1}, is generated by the recursion

$$f_{t+1} = \alpha f_t + (1-\alpha) r_t, \tag{4}$$

subject to the initial condition that $f_1 = r_1$. We also consider double exponential (Holt) smoothing:

$$\begin{aligned} f_{t+1} &= \alpha(f_t + \lambda_{t-1}) + (1-\alpha) r_t \\ \lambda_t &= \beta(f_{t+1} - f_t) + (1-\beta)\lambda_{t-1}, \end{aligned} \tag{5}$$

where $f_1 = 0, f_2 = r_2$ and $\lambda_2 = (r_2 - r_1)$. For these cases, α and α and β, respectively, are selected to minimize the sum of squared forecast errors in real time.

The next class of forecasting models comprises a set of non-linear specifications including two logistic STAR models of the form

$$r_{t+1} = \theta_0' \eta_t + d_t \theta_1' \eta_t + \varepsilon_{t+1} \tag{6}$$

where $\eta_t = (1, r_t)'$. Under the STAR1 model,

$$d_t = 1/(1 + exp(\gamma_0 + \gamma_1 r_{t-3})), \tag{7}$$

while, under the STAR2 model,

$$d_t = 1/(1 + exp(\gamma_0 + \gamma_1(r_t - r_{t-6}))). \tag{8}$$

We also consider more flexible nonlinear forecasting models in the form of a single-layer neural net model with two hidden units ($n = 2$)

$$r_{t+1} = \theta_0' \eta_t + \sum_{i=1}^{n} \theta_i g(\beta_i' \eta_t) + \varepsilon_{t+1}, \tag{9}$$

as well as a two-layer neural net model

$$r_{t+1} = \theta_0' \eta_t + \sum_{i=1}^{n_2} \theta_i g\left(\sum_{j=1}^{n_1} \beta_j g(\alpha_j' \eta_t)\right) + \varepsilon_{t+1}, \tag{10}$$

with two hidden units in the first layer ($n_1 = 2$) and one hidden layer in the second layer ($n_2 = 1$). For both neural net models, $g(.)$ is the logistic function and $\eta_t = (1, r_t, r_{t-1}, r_{t-2})$.

Nonlinear forecasting models are known to sometimes generate extreme forecasts. To deal with this problem we adopt an 'insanity filter' that constrains such forecasts. More specifically, if the predicted change in the underlying variable is greater than any of the historical changes up to a given point in time, the forecast is replaced with a 'no change' forecast.

Note also that parameter estimates need no longer be consistent, nor is it clear which properties models selected by information criteria such as the BIC will have under regularly occurring breaks to the data generating process or when a rolling window estimator is used.

We finally consider two approaches that build on the initial nine forecasting methods. The 'previous best' approach selects that forecasting model which, at a given point in time, has produced the best historical forecasting record, using historical root mean squared error (RMSE) as the criterion. In the case of the rolling window approach this works as follows:

$$\begin{aligned} f_{t+1} &= f_{j^*,t+1}, \\ j^* &= \arg \min_{j=1,\dots,N} \tau_0^{-1} \sum_{\tau=0}^{\tau_0 - 1} (y_{t-\tau} - f_{j,t-\tau}), \end{aligned} \tag{11}$$

where $f_{j,t}$ is the forecast of returns for period t produced by the jth model and N is the number of underlying models.

Conversely, the average approach uses an equal-weighted average of the forecasts to predict future returns:

$$f_{t+1} = N^{-1} \sum_{j=1}^{N} f_{j,t+1}. \tag{12}$$

3.2 Results

To illustrate the out-of-sample forecasts generated by two of the models, Figure 1 plots these for the factor-augmented autoregressive model and for the two-layer neural net, both estimated by means of a rolling window. The two sets of forecasts are quite different most of the time. In particular, the factor-augmented forecasts are smoother and more persistent than those generated by the neural net which are also subject to occasional spikes. Even so, compared to the variation in the actual or realized returns, the range of values taken by the forecasts is very small.

Tables 1-3 present results in the form of annualized RMSE-values for the full out-of-sample period (1970:01-2005:12) in addition to the three subsamples spanning the 1970s, 1980s and 1990s. Table 1 reports results for the value-weighted portfolio under both the recursive (expanding) and rolling estimation window. Under the recursive approach the best performance is produced by the equal-weighted average which is marginally better than the forecasts generated by the autoregressive, prevailing mean, exponential smoothing, two-layer neural net and previous best models.[8] At the other end of the performance spectrum, the Holt smoothing model generates the worst out-of-sample forecasts.

Similar results are obtained under the 10-year rolling estimation window. Here the best forecasting performance is delivered by the prevailing mean followed by the autoregressive, equal-weighted average and previous best forecasts. Again, the worst performance is produced by the Holt smoothing and STAR2 models.

Under the recursive estimation approach, the factor-augmented AR model is able to outperform the prevailing mean but only during the 1980s. In contrast, under the rolling window approach the prevailing mean produces the lowest RMSE-values in all subsamples.

While the value-weighted return series is largely serially uncorrelated, the equal-weighted returns display stronger serial correlation reflecting the less active trading in the smaller stocks that dominate this portfolio. Table 2 shows that some of the forecasting models are capable of identifying serial persistence in returns. Indeed the neural net models are now best followed by the simple average and autoregressive forecasts, while the worst is again the Holt smoothing model.

Whereas the good performance of the autoregressive model is independent of the choice of estimation window, the absolute and relative forecasting performance of the two neural nets deteriorate under the shorter rolling estimation window. This is to be expected given the difficulty in precisely estimating the parameters of these models.

Turning to the returns on the SMB portfolio that captures the differential performance of small and big stocks, Table 3 shows that under the recursive

[8] The simple autoregressive model performs well when the lag order is selected by the BIC. This method for selecting lag order often reduces to excluding all past returns (particularly for the value-weighted portfolio) which explains the similarity between its performance and that of the prevailing mean.

estimation approach the average forecast does best overall, closely followed by the AR model, the previous best forecast and the prevailing mean. Under the rolling window the prevailing mean is best by some margin.

In conclusion, these results suggest that, as far as out-of-sample RMSE performance is concerned, the prevailing mean is a top contender for the best overall approach. While some forecasting models work well for certain return series and certain subsamples, none of them appears to outperform the simple prevailing mean on a consistent basis.

3.3 Time-varying Predictability

Our earlier discussion suggested that we should not expect a particular model's ability to predict stock returns to remain constant over time. In the absence of a formal model for capturing changes in predictive accuracy, it is difficult to tell how best to detect or monitor such time-variations. We simply resort to present rolling window estimates of the out-of-sample R^2-value using a 36-month rolling estimation window and our recursive forecasts. Our estimates are computed relative to the sum of squared errors associated with the prevailing mean benchmark:

$$\hat{R}^2_{i,t-m+1:t} = 1 - \frac{e'_{i,t-m+1:t} e_{i,t-m+1:t}}{\bar{e}'_{t-m+1:t} \bar{e}_{t-m+1:t}}, \tag{13}$$

where $\bar{e}_{t-m+1:t}$ is the $m-$vector of out-of-sample forecast errors associated with the prevailing mean, $e_{i,t-m+1:t}$ is the $m-$vector of out-of-sample forecast errors from the ith model measured between period $t-m+1$ and period t and m is the length of the estimation window. Notice that whenever the sum of squared forecast errors for a particular model exceeds that of the prevailing mean, the R^2-value will be negative. Hence there is a one-to-one correspondence between out-of-sample RMSE and this R^2-measure.

Figure 2 shows the sequence of $\hat{R}-$estimates associated with the factor-augmented autoregressive model and the 2-layer neural net model, assuming $m = 36$. On average there is little evidence that the factor-augmented model can predict returns in a mean squared error sense: The out-of-sample RMSE-values hover around zero and are slightly negative most of the time. However, there are two periods where return predictability seems to have been present, namely during 1974-76, following the oil price shocks, and during 1983-1987.

Turning to the graph for the two-layer neural net, again most of the time this model is unable to predict return variations. However, there is mild evidence of predictability during 1982-85 and perhaps during a briefer period in 2000. Despite their differences, both models identify the increasing R^2 from 1980-85, the subsequent decline until around 1988 followed by another decline in 1995 and the small increase thereafter.

3.4 Evaluation of Relative Forecasting Performance

To get an indication of the relative forecasting performance of the approaches under consideration, we next computed the test statistic proposed by Giacomini and White (2006). This facilitates pair-wise comparisons of out-of-sample forecasting performance. We compare models estimated under rolling windows which ensures that one forecasting model is not asymptotically nested by the other.

Tables 4-6 report the results. A word of caution is necessary when interpreting these results. An implication of our earlier discussion is that out-of-sample tests of return predictability are not necessarily appropriate diagnostics for corroborating in-sample (historical) predictability. The distributional properties of most statistical tests assume a stationary setting (at least asymptotically) which is unlikely to be a valid assumption here.

Bearing this in mind, while the worst forecasting approach, namely the Holt smoothing method, generally is rejected against the better forecasting models, it is also clear that the data is not very informative when it comes to distinguishing between the best approaches. For example, the difference in the forecasting performance of the prevailing mean versus the autoregressive, factor-augmented or two-layer neural net models is generally not significant.

These findings suggest that it is difficult to distinguish between the average out-of-sample mean squared error performance (computed over a fairly long data sample) of some relatively sophisticated forecasting models and the alternative of simply using the prevailing mean. It thus adds further evidence to the existing literature that, in the context of linear forecasting models, has found it very difficult to identify predictor variables with reliable out-of-sample forecasting power.

3.5 An Alternative Criterion for Measuring Forecasting Performance

Mean squared error performance and out-of-sample R^2 are standard statistical measures of forecast precision. However, they overlook that, ultimately, the economic value of return forecasts hinges upon their use in investors portfolio decisions. Other criteria for forecasting performance may well be more closely related to the possibility of exploiting return forecasts in trading strategies that aim to generate abnormal (risk-adjusted) profits.

Forecasts of returns in financial markets are of interest predominantly because of the possibility of exploiting such forecasts in portfolio selection rules that enhance risk-adjusted investment performance. To be valuable, forecast signals hence need to be implemented in profitable trading rules. Predictability is not 'per se' precluded by the efficient market hypothesis, but economic theory suggests that predictability should not offer easy ways to enhance the expected return versus risk trade-off established by more passive, low-cost investment strategies.

One criterion that has been adopted to this end is the sign criterion, see e.g.

Henriksson and Merton (1981) and Pesaran and Timmermann (1992). This looks at the ability of forecasts to correctly predict the direction of the market, i.e. the sign of the return measured in excess of a benchmark such as the risk-free T-bill rate. This measure has been shown to be more closely related to the possibility of converting forecast signals into profits than standard statistical measures such as mean squared error (Leitch and Tanner (1991)).

Table 7 shows the outcome of applying the sign test proposed by Pesaran and Timmermann (1992) to our forecasts and realized returns, both calculated net of the risk-free rate in the case of the value-weighted and equal-weighted stock portfolios. A 'zero' indicates that the predicted excess return always has the same sign. This represents a 'broken clock' forecast and hence conveys no information. According to the sign criterion the factor-augmented AR, average, prevailing mean and previous best forecasts (ranked in that order) appear able to predict the sign of value-weighted excess returns over the full sample. Closer inspection of the results reveals that this occurs despite poor forecasting performance during the 1990s.

Turning to the equal-weighted excess returns, several of the forecasting approaches perform quite well and generate a test statistic above two. This holds for the AR, one- and two-layer neural nets, the STAR2 model, the previous best and the average forecast. For this portfolio there is less evidence of a deterioration in the forecasting performance during the 1990s although the 1980s is the period with the strongest evidence of sign predictability.

Finally, the exponential smoothing and one- and two-layer neural net models produce evidence of sign predictability for the SMB portfolio returns. Moreover, all forecasting models with exception of the prevailing mean produce evidence of sign predictability during the 1970s. Hence there is no evidence that the sign of SMB portfolio returns could be predicted after the publication of the small firm effect in 1981.

4 An Adaptive Forecast Combination Approach

Our empirical results suggest that no single forecasting model consistently outperforms the simple prevailing mean over long periods of time and that any return predictability is, at best, short-lived and likely to deteriorate fairly quickly. Forecasting approaches that always use the same model are therefore unlikely to be successful. Rather than sticking to a single forecasting model, we will therefore consider a forecast combination approach.

Before introducing our adaptive combination approach, we briefly review some alternative methods for dealing with model instability in a forecasting context.

4.1 Methods Dealing with Instability in Forecasting Models

Several approaches have been proposed to deal with the instability found in the parameter estimates and performance of many forecasting models. Clements and Hendry (1999, 2006) provide a taxonomy and analysis of this area.[9]

One approach is to identify specific historical breaks (e.g., Pesaran and Timmermann (2002), Lettau and van Nieuwerburgh (2006) and Pesaran, Pettenuzzi and Timmermann (2006)). Common to variants of this approach is that they seek to test for discrete breaks in real time and estimate the parameters of the forecasting models either on the post-break data alone (if the break is large or the most recent break happened a long time ago) or some combination that weights post-break data more than pre-break data (Pesaran and Timmermann (2007)).

A second approach which does not require identifying the dates and size of the breaks was proposed by Clements and Hendry (1999) and reviewed, more recently, by Hendry (2005). This uses intercept corrections or, in the context of vector equilibrium correction models, differencing of the forecasting model. This approach has the potential to account for sizeable forecast errors following a structural break and thus to catch up with the data generating process following a levels shift.

A third approach tracks gradual shifts in coefficients which are subject to small changes each period, e.g. by using a time-varying parameter model. One example is provided by Mamaysky, Spiegel and Zhang (2006) who use the Kalman filter to track time-variations in the ability of fund managers to outperform their benchmark on a risk-adjusted basis as measured by the funds' alphas. This approach only attempts to identify a predictable return component indirectly through the managers' ability to select assets and outperform their benchmark on a persistent basis (see Brown and Goetzmann (1995) for a discussion of performance persistence).

A fourth approach uses regime switching models. These assume that 'history repeats' in the sense that the parameters only shift between a small set of possible states. In particular, the underlying data generating process is subject to discrete shifts governed either by some latent variable (as in Hamilton (1989)) or by means of observable threshold variables (see Terasvirta (2006) for a survey). Ang and Bekaert (2002), Perez-Quiros and Timmermann (2000) and Guidolin and Timmermann (2005) are examples of these models applied to predict stock market returns.

A fifth approach uses exponentially weighted moving averages of the data to estimate model parameters, putting greater weight on more recent observations and reducing the effect of past data, more so the further back in time this occur.

A sixth approach is to use rolling estimation windows. The idea is to use only the most recent τ_0 observations to estimate the parameters of the best forecasting model and exclude data further back in time than this estimation

[9] Rapach and Wohar (2006) and Paye and Timmermann (2006) provide empirical evidence of model instability in the context of stock returns.

window. The approach is easy to implement and is very popular in practice. One problem with this approach is that there is no theory for how to select the length of the rolling window, nor is it clear that one can come up with a data generating process for which this is the optimal strategy to use.

Finally, it has been argued that forecast combination methods provide a way to hedge against model instability (Hendry and Clements (2002) and Timmermann (2006)). Breaks are likely to affect the individual forecasting models to different degrees and so a combined forecast may well provide more robust performance in the presence of breaks.

We next describe our simple adaptive forecast combination approach which uses rolling windows both to estimate model parameters and select the forecasting model.

4.2 A Simple Adaptive Combination Approach

To be successful, an adaptive forecasting approach must search for and monitor local predictability patterns in returns and, if these can be identified, attempt to exploit them before they disappear. In as far as possible, the approach must also be robust to the effects of parameter estimation error and uncertainty about choosing among individual forecasting models. There are many ways this can be done; we just consider a simple approach here. Because identification of the single best individual model is surrounded by considerable noise, we restrict our attention to using either the combined (equal-weighted) forecast or the prevailing mean forecast. Individual forecasting models are used as "testers" that can detect short-lived periods with return predictability.

First, we use a rolling window of m observations to calculate a backward-looking estimate of the out-of-sample R^2 over the previous m periods. We then check if any of the individual recursively estimated forecasting models produces an estimate, $\hat{R}^2$, above a certain threshold, $R^2_{\min}$. If this condition fails, the forecast is set equal to the prevailing mean which effectively represents the benchmark.[10] If the condition is satisfied and the combined forecast has a positive out-of-sample R^2-value (computed using the same rolling window), we use the combined forecast.

We consider different lengths of the selection window $m = 18$, 36, 60 and 120 months, while the threshold $R^2_{\min}-$value is set to $\{0, 0.01, 0.02, 0.05, 0.10\}$.

Before interpreting the empirical results, note the trade-off in selecting the hurdle value, $R^2_{\min}$. If $R^2_{\min}$ is set too low (i.e. at zero), it is more likely that periods with spurious predictability are identified. On the other hand, periods with genuine predictability are less likely to be ignored. Setting the threshold too high has the reverse effect of increasing the probability of selecting

[10] The idea of letting the forecasting model change over time is of course not novel. Pesaran and Timmermann (1995, 2000) and Swanson and White (1995) use model selection methods recursively over time to select the forecasting model that optimizes a penalized likelihood function. What appears to be new is to introduce a threshold for the R^2-value and only produce predictions other than the prevailing mean during periods where this threshold is exceeded.

periods with genuine predictability, but also reduces the proportion of times when predictability is deemed to be present.

A similar trade-off holds with regard to the choice of the length of the evaluation window, m. The shorter this is, the noisier estimates of R^2 will be and hence the higher the risk of wrongly identifying non-existent predictability. On the other hand, the shorter the window, the better the approach will be at identifying temporary, fleeting predictability patterns. These may be missed by a long window which, conversely, offers more stable estimates of the out-of-sample R^2.

Results from the adaptive forecast combination approach are reported in Table 8. The first column shows the hurdle value, R^2_{min}, used in selecting the forecast. If this is not exceeded by any model, the forecast is simply set equal to the prevailing mean. The second column shows the proportion of periods where the threshold was exceeded. Naturally this proportion declines as the threshold value is increased. Moreover, for the smallest threshold value, $R^2_{\text{min}} = 0$, the exceedance rate increases as a function of the window length while the reverse pattern is observed for the higher hurdle values $R^2_{\text{min}} = 0.05$ or 0.10. In fact, when the threshold is set as high as 0.10, the prevailing mean is adopted 100% of the time under the 120-month window. This makes sense since it is difficult to generate this high an R^2-value over lengthy periods of time.

The third column reports root mean squared errors under the prevailing mean. These vary across different lengths of the evaluation window because the beginning of the out-of-sample period is 1970 plus the length of the initial evaluation window used to compute the R^2-value, e.g. mid-1971 for an 18-month window and 1980 for a 120-month window. Finally, the fourth column reports the corresponding RMSE-values under the adaptive forecast combination approach.

In the majority of cases (i.e. across different selection windows and R^2-thresholds), the adaptive combination approach is capable of producing forecasts with lower RMSE-values than the prevailing mean. The best results are generated under either the short windows of 18 or 36 months or the longest window of 120 months.

Figure 3 shows the forecasts assuming a selection window of 36 months and an $R^2_{\text{min}}-$value of 0.02. During a three-year period from 1978-1981 and a five-year in the mid-nineties there was not sufficiently strong evidence of return predictability and so the adaptive combination forecasts fall back on the prevailing mean. At other times, however, the adaptive combination forecasts can differ substantially from the prevailing mean value.

Changing the selection window to 120 months, as we do in Figure 4, has a large effect on the forecasts. Reflecting the ‘model breakdown’ during the 1990s, the forecasts always fall back on the prevailing mean after 1993. In other words, no single model managed to produce forecasts that, over a 10-year period, had an out-of-sample R^2-value above 0.02.

Figures 5-7 illustrate the selection of the combined forecast for different lengths of the window used to compute the out-of-sample R^2-value, again using a cutoff value $R^2_{\text{min}} = 0.02$. When this window is very short (18 months), Figure 5 shows that periods where the approach identifies return predictability sometimes are very short-lived. Increasing the selection window from 18 to 36 months, Figure 6 shows that the adaptive approach tends to identify fewer but longer blocks of time with return predictability. Moreover, under the longest 120-month window Figure 7 shows that, with few exceptions, the combined forecast gets selected most of the time between the early 1980s and 1995. Consistent with the earlier comments on weaker evidence of return predictability in the 1990s, the prevailing mean dominates after this period.

These findings indicate the potential for adaptive combination approaches based on the principle that return predictability is likely to be short-lived. Different approaches could well be used, so more evidence is required to see how general this finding is.

4.3 Statistical Significance

The empirical evidence should be interpreted cautiously in the absence of a theory for the small-sample distribution of the rolling window estimates of the out-of-sample R^2-values. Complicating interpretation of these statistics is, first, that the underlying forecasts are generated using recursively estimated parameter estimates. Moreover, the rolling window estimates of R^2 are, by construction, serially dependent and so any inference would have to account for this feature. Finally, the alternative hypothesis of intermittent predictability is non-standard, so power is also an issue.

There is clearly a need to develop a distributional theory, e.g. by means of the probability distribution for the out-of-sample R^2-estimate measured over relatively short sample periods. To get a first idea of the distribution of the RMSE-values reported in Table 8, we use a simple parametric bootstrap approach which first estimates an AR(1) model to the individual forecasts (with persistence parameter $\hat{\phi}_j$ for the jth forecast, $f_{j,t+1}$), saves the residuals and generates pseudo-random forecasts

$$f^b_{j,t+1} = \hat{\phi}_j f^b_{j,t} + \hat{\varepsilon}^b_{j,\tau},$$

where the b superscript refers to the bootstrap number and $\hat{\varepsilon}^b_{j\tau}$ is the randomly selected residual for period $\tau \in [1, T]$, T being the sample size.[11] This setup breaks the link between the forecasts and actual returns while it preserves the basic persistence and volatility properties of the data. We use 1,000 bootstrap simulations to generate test statistics for the RMSE-value of the adaptive combination method. Finally, we order the bootstrapped values in descending order to obtain the 95% critical values.

Column 5 of Table 8 presents the results from this analysis. The bootstrapped 95% critical values for the RMSE-statistic, exceed the RMSE-value of the prevailing mean for the lowest hurdle value, $R^2_{\min} = 0$, but generally fall as the hurdle rate is increased. This is explained by the less frequent use of

[11] Since a separate value of τ is drawn for each value of $t+1$, the notation $\tau(t+1)$ is more precise.

the combined forecast and the resulting narrowing of the sample distribution of RMSE-values.

Using these critical values, all or all but one of the windows generate significantly smaller RMSE-values than can be attributed to randomness when the hurdle value is set at or below $R^2_{\text{min}} = 0.05$, while no window does so for the highest hurdle rate of $R^2_{\text{min}} = 0.10$.

Notice the contrast between these findings and the earlier results based on the Giacomini-White test which considered average out-of-sample forecasting performance and failed to find evidence that the individual forecasts outperformed the prevailing mean. It is easy to explain these differences since our adaptive combination approach does not consider the average forecasting performance of individual models but instead looks for temporary predictability as identified by a multitude of forecasting approaches.

5 Conclusion

What makes predictability of financial returns so difficult is that it is influenced by market participants' own attempts to identify and exploit any purported predictability and so constantly evolves over time. Just as Heisenberg's uncertainty principle implies that a scientist's attempts to increase the accuracy with which he can measure the position of an object leads to a corresponding decline in the precision with which the object's momentum can be measured, investors' efforts to exploit predictability patterns lead to their self-destruction. As a consequence, there is a sense in which the stronger (i.e. easier to detect) the evidence of past return predictability, the greater the expected decline in future predictability, as predictability patterns get more rapidly incorporated into current market prices.

How can investors then utilize forecasts from return prediction models that (i) only generate weak evidence of predictability and (ii) do not work most of the time but on some occasions produce valuable signals? A crucial component in any answer to this question is to get some indication of when different models produce valuable forecasts and when they fail to do so, e.g. in the form of a real-time monitoring system tracking how reliable the forecasts have been over a recent period.

Our empirical findings suggest that most of the time stock returns are not predictable, but there appear to be pockets in time with modest evidence of local predictability. Interestingly, none of the forecasting models appears able to predict returns during the sustained bull market during the second half of the 1990s. The adaptive forecast combination approach responds to this dearth of predictability by not attempting to identify time-varying components in stock returns during this period and hence avoids making mistakes that could have proved costly if implemented in an investment strategy.

References

[1] Ang A., and G., Bekaert, 2002, International Asset Allocation with Regime Shifts. Review of Financial Studies, 15, 1137-1187.

[2] Avramov, D. , T. Chordia and A. Goyal, 2006, Liquidity and Autocorrelation in Individual Stock Returns. Journal of Finance 61(5), 2365-2394.

[3] Brock, W., J. Lakonishok, and B. LeBaron, 1992, Simple technical trading rules and the stochastic properties of stock returns. Journal of Finance 47, 1731-1764.

[4] Brown, S. and W. Goetzmann, 1995, Performance Persistence. Journal of Finance 50, 679-698.

[5] Campbell, J.Y., 1987, Stock Returns and the Term Structure. Journal of Financial Economics 18(2), 373-399.

[6] Campbell, J.Y., and R., Shiller, 1988, The Dividend-Price Ratio, Expectations of Future Dividends and Discount Factors. Review of Financial Studies, 1, 195-227.

[7] Campbell, J.Y. and S. Thompson, 2007, Predicting the Equity Premium Out of Sample: Can Anything Beat the Historical Average? Forthcoming in Review of Financial Studies.

[8] Clements, M.P. and D.F. Hendry, 1999, Forecasting Non-stationary Economic Time Series. Cambridge, Mass.: MIT Press.

[9] Clements, M.P. and D.F. Hendry, 2006, Forecasting with Breaks in Data Processes. Pages 605-658 in Handbook of Economic Forecasting, edited by C.W.J. Granger, A. Timmermann and G. Elliott, North Holland.

[10] Cochrane, J., 2006, The Dog that Didn't Bark. Mimeo, GSB University of Chicago.

[11] Elliott, G., and A. Timmermann, 2007, Economic Forecasting. Mimeo, UCSD.

[12] Fama, E.F. and K.R. French, 1988, Dividend Yields and Expected Stock Returns. Journal of Financial Economics 22(1), 3-25.

[13] Fama, E.F. and K.R. French, 1989, Business Conditions and Expected Returns on Stocks and Bonds. Journal of Financial Economics 25(1), 23-49.

[14] Fama, E.F. and K.R. French, 1992, The Cross-Section of Expected Stock Returns. Journal of Finance 47, 427-465.

[15] Fama, E.F. and G.W. Schwert, 1981, Stock Returns, Real Activity, Inflation and Money. American Economic Review 71, 545-565.

[16] Giacomini, R., and H., White, 2006, Tests of Conditional Predictive Ability. Econometrica 74, 6, 1545-1578.

[17] Goyal, A. and I. Welch, 2003, Predicting the Equity Premium with Dividend Ratios. Management Science 49(5), 639-654.

[18] Goyal, A. and I. Welch, 2006, A Comprehensive Look at the Empirical Performance of Equity Premium Prediction. Forthcoming in Review of Financial Studies.

[19] Guidolin, M. and A. Timmermann, 2005, Economic Implications of Bull and Bear Regimes in UK Stock Returns. Economic Journal, 111-143.

[20] Guidolin, M. and A. Timmermann, 2007, Properties of Equilibrium Asset Prices under Alternative Learning Schemes. Journal of Economic Dynamics and Control 31(1), 161-217.

[21] Hamilton, J., 1989, A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle, Econometrica, 57, 357-384.

[22] Hendry, D.F. 2005, Unpredictability and the Foundations of Economic Forecasting. Mimeo, Nuffield College.

[23] Hendry, D.F. and M.P. Clements, 2002, Pooling of Forecasts. Econometrics Journal 5, 1-26.

[24] Henriksson, R.D. and R.C. Merton, 1981, On Market Timing and Investment Performance. II. Statistical Procedures for Evaluating Forecasting Skills. Journal of Business 54, 513-533.

[25] Leitch, G. and J.E. Tanner, 1991, Economic Forecast Evaluation: Profits Versus the Conventional Error Measures, American Economic Review 81, 580-90.

[26] Lettau, M. and S. Ludvigsson, 2001, Consumption, aggregate wealth, and expected stock returns. Journal of Finance 56, 815-850.

[27] Lettau, M. and S. van Nieuwerburgh, 2006, Reconciling the Return Predictability Evidence. Forthcoming in Review of Financial Studies.

[28] Lo, A.W., 2004, The Adaptive Markets Hypothesis. Market Efficiency from an Evolutionary Perspective. Journal of Portfolio Management 30, 15-29.

[29] Mamaysky, H., M. Spiegel and H. Zhang, 2006, Improved Forecasting of Mutual Fund Alphas and Betas. Forthcoming in Review of Finance.

[30] Paye, B. and A. Timmermann, 2006, Instability of Return Prediction Models. Journal of Empirical Finance 13 (3), 274-315.

[31] Perez-Quiros, G. and A. Timmermann, 2000, Firm Size and Cyclical Variations in Stock Returns. Journal of Finance, 1229-1262.

[32] Pesaran, M.H., D. Pettenuzzo and A. Timmermann, 2006, Forecasting Time Series Subject to Multiple Structural Breaks. Review of Economic Studies 73, 1057-1084.

[33] Pesaran, M.H. and A. Timmermann, 1992, A Simple Nonparametric Test of Predictive Performance. Journal of Business and Economic Statistics 10, 461-465.

[34] Pesaran, M.H. and A. Timmermann, 1995, Predictability of Stock Returns: Robustness and Economic Significance. Journal of Finance 50, 1201-1228.

[35] Pesaran, M.H. and A. Timmermann, 2000, A Recursive Modelling Approach to Predicting UK Stock Returns. Economic Journal 159-191.

[36] Pesaran, M.H. and A. Timmermann, 2002, Market Timing and Return Prediction under Model Instability. Journal of Empirical Finance 9, 495-510.

[37] Pesaran, M.H. and A. Timmermann, 2007, Selection of Estimation Window in the Presence of Breaks. Journal of Econometrics 137(1), 134-161.

[38] Rapach, D. and M. Wohar, 2006, Structural Breaks and Predictive Regression Models of Aggregate US Stock Returns. Journal of Financial Econometrics 4(2), 238-274.

[39] Schwert, G.W., 2002, Anomalies and market efficiency. In G.M. Constantinides, M. Harris, R.M. Stulz (eds) Handbook of the Economics of Finance. North Holland: Amsterdam.

[40] Stock, J.H., and M. W. Watson, 2002, Macroeconomic Forecasting Using Diffusion Indexes. Journal of Business and Economic Statistics 20:147-162.

[41] Sullivan, R., A. Timmermann and H. White, 1999, Data-Snooping, Technical Trading Rules and the Bootstrap. Journal of Finance 54, 1647-1692.

[42] Swanson, N. and H. White, 1995, A Model Selection Approach to Assessing the Information in the Term Structure using Linear Models and Artificial Neural Networks. Journal of Business and Economic Statistics 13, 265-276.

[43] Terasvirta, T., 2006. Forecasting Economic Variables with Nonlinear Models. Pages 423-458 in G. Elliott, C.W.J. Granger, A. Timmermann, eds. Handbook of Economic Forecasting. North-Holland: Amsterdam.

[44] Timmermann, A., 2006, Forecast Combinations. Pages in 135-196 in G. Elliott, C.W.J. Granger, A. Timmermann, eds. Handbook of Economic Forecasting. North-Holland: Amsterdam.

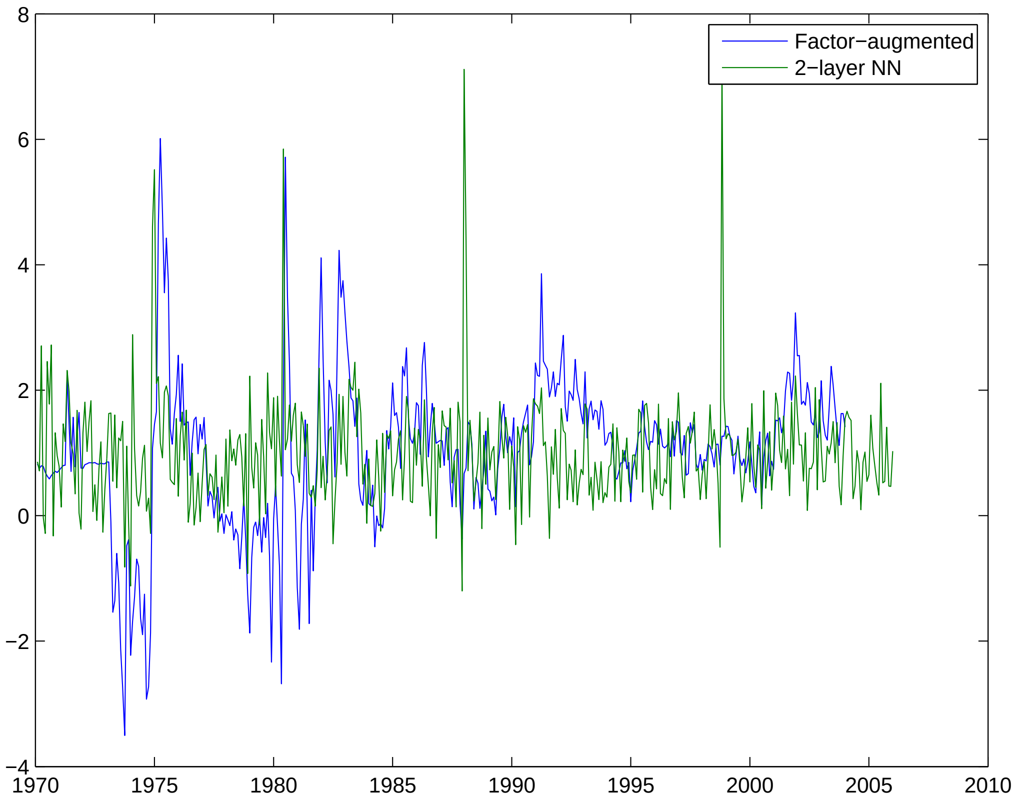

Figure 1: Out-of-sample forecasts of value-weighted returns, 1970-2005.

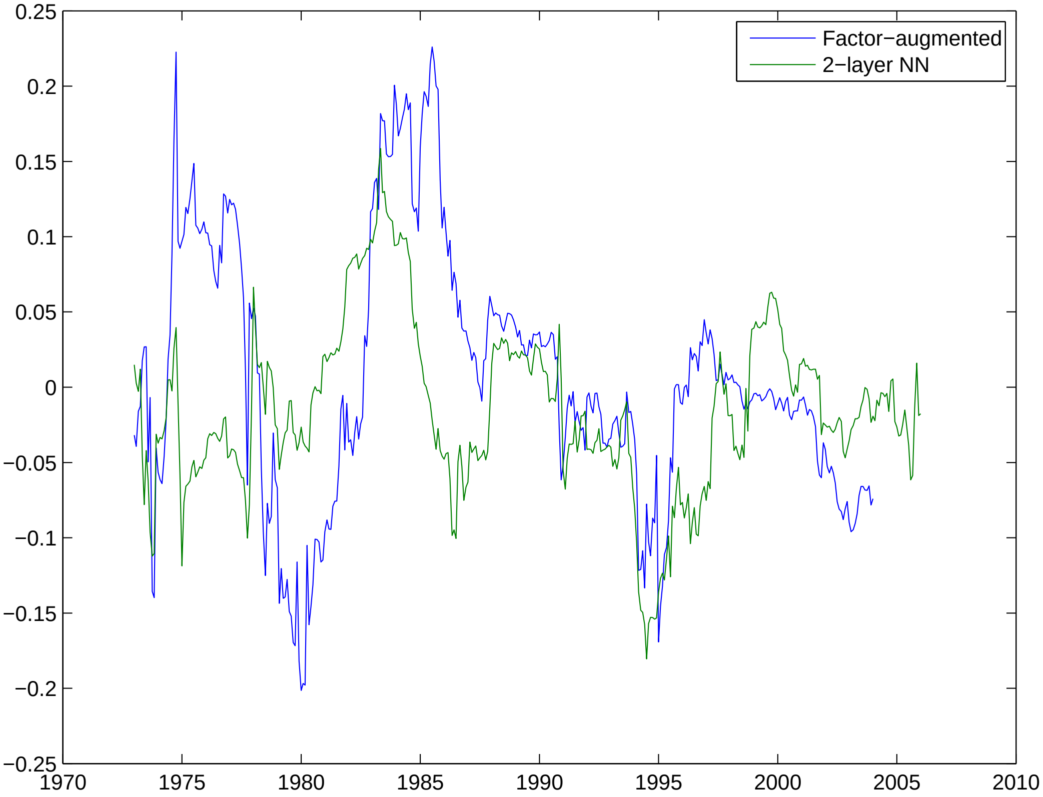

Figure 2: 36-month rolling window estimates of out-of-sample R^2, 1970-2005.

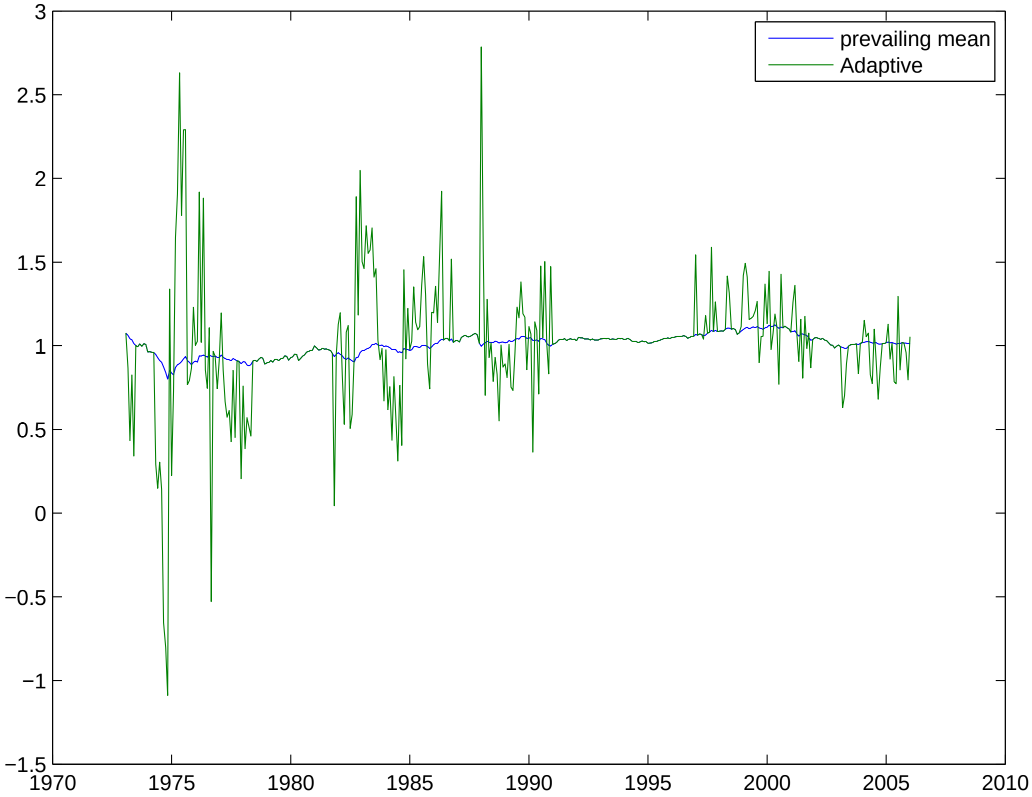

Figure 3: Out-of-sample return forecasts from the adaptive combination approach, assuming a 36-month window.

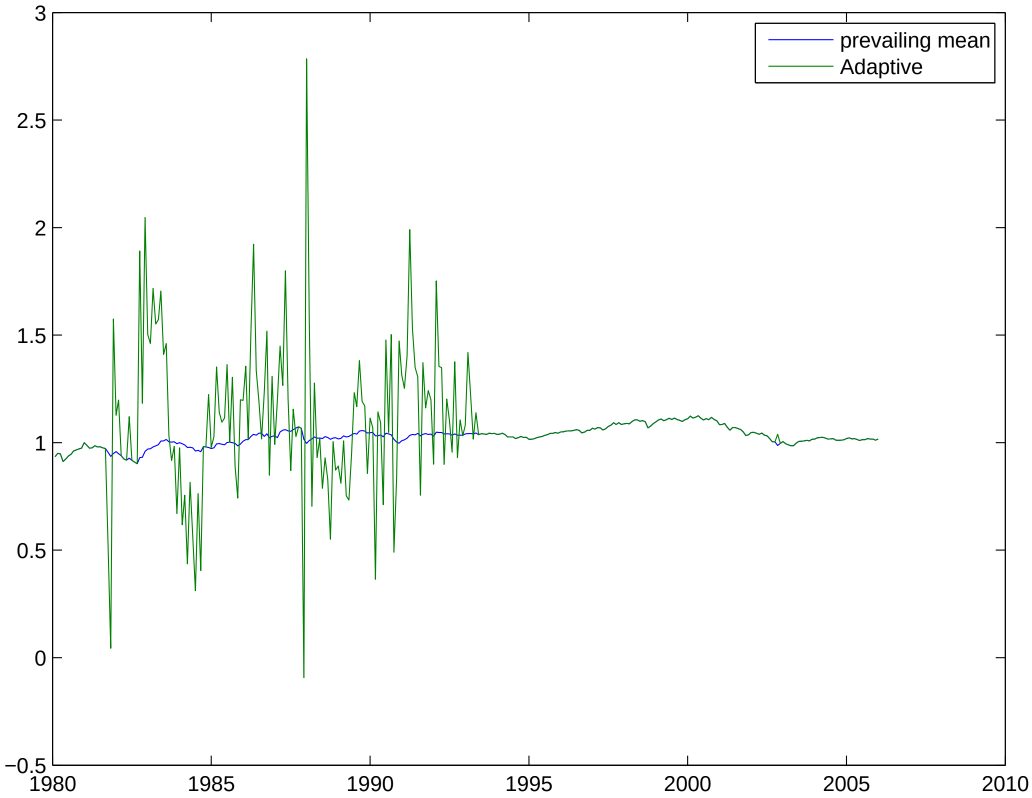

Figure 4: Out-of-sample return forecasts from the adaptive combination approach, assuming a 120-month window.

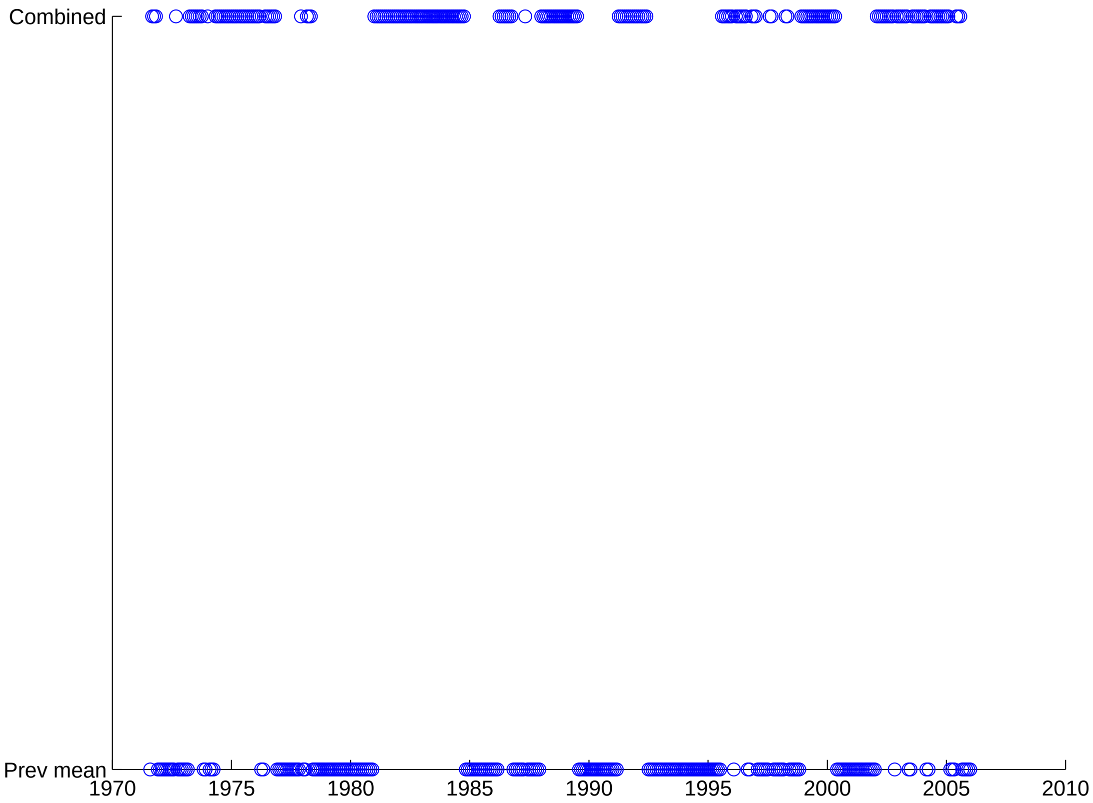

Figure 5: Periods with return predictability identified by the adaptive forecast combination approach (18-month window).

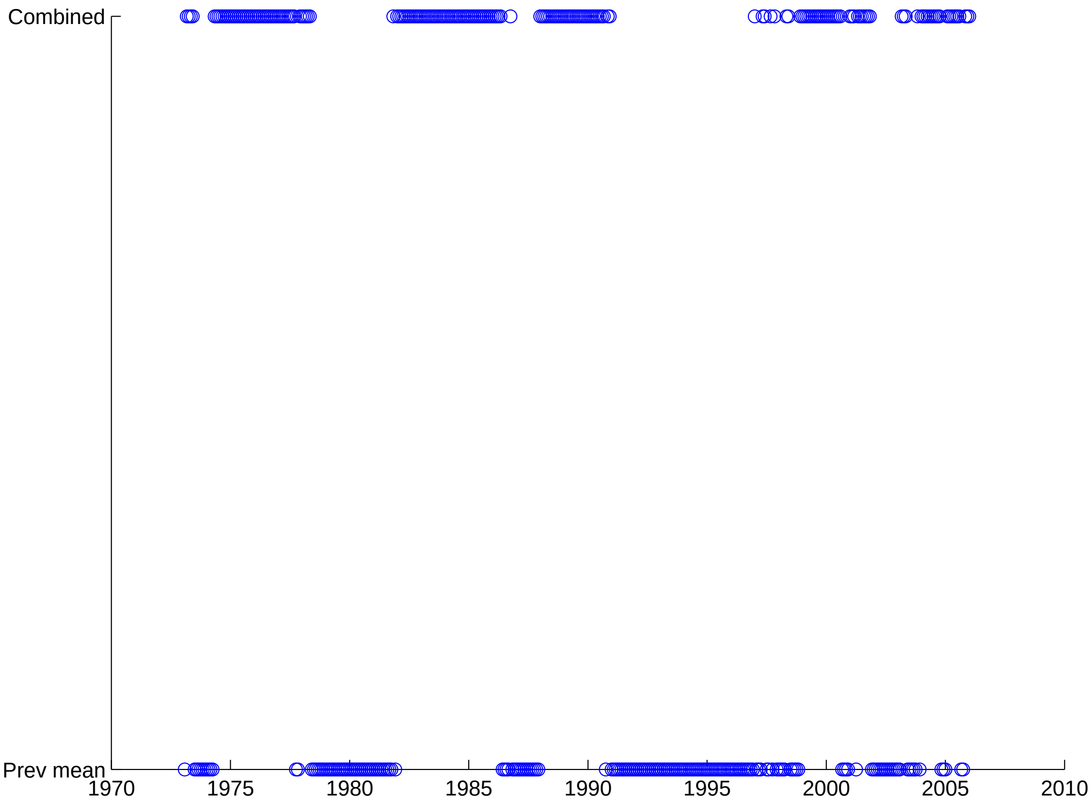

Figure 6: Periods with return predictability identified by the adaptive forecast combination approach (36-month window).

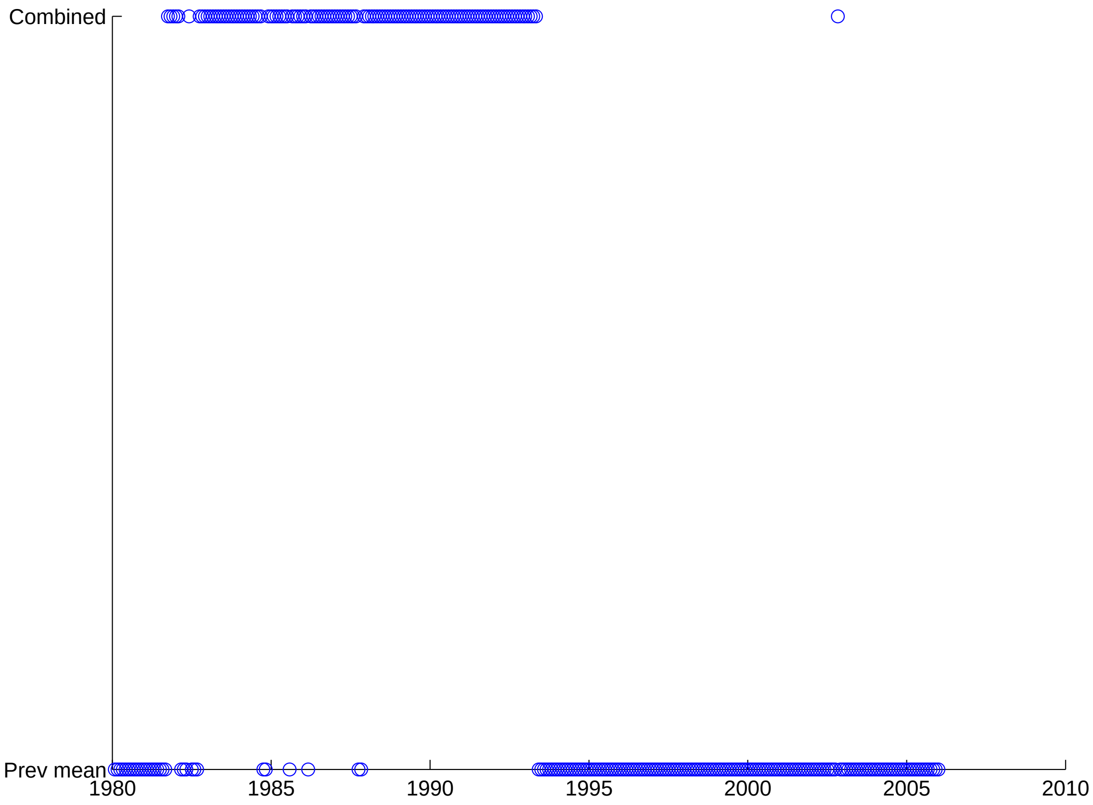

Figure 7: Periods with return predictability identified by the adaptive forecast combination approach (120-month window).

Table 1: Root Mean Squared Error Performance (value-weighted portfolio)

Window Strategy	**Model**	**1970-2005**	**1970s**	**1980s**	**1990s**
Recursive	Prevailing Mean	15.87	16.91	16.69	13.60
	Autoregressive	15.92	17.08	16.69	13.60
	Factor-augmented AR	16.11	16.90	16.04	13.71
	Exp. Smoothing	15.94	17.03	16.74	13.59
	Holt Smoothing	16.41	17.60	17.31	13.91
	STAR_1	16.22	17.26	17.23	13.93
	STAR_2	16.55	18.25	17.38	13.94
	One Layer NN	16.00	17.19	16.70	13.62
	Two Layer NN	15.92	17.13	16.44	13.63
	Previous Best Model	15.88	17.22	16.04	13.71
	Equal-weighted Avg.	15.80	16.87	16.46	13.59
Ten Year Rolling	Prevailing Mean	15.93	16.93	16.73	13.62
	Autoregressive	16.01	17.17	16.77	13.63
	Factor-augmented AR	16.28	16.96	16.77	13.62
	Exp. Smoothing	16.31	17.34	17.14	14.08
	Holt Smoothing	18.36	19.12	19.36	16.63
	STAR_1	16.91	17.85	18.25	13.66
	STAR_2	17.39	19.62	17.12	14.82
	One Layer NN	16.50	17.57	17.17	13.86
	Two Layer NN	16.32	17.36	16.93	13.66
	Previous Best Model	16.02	17.08	16.77	13.62
	Equal-weighted Avg.	15.98	16.95	16.66	13.76

Note: This table reports out-of-sample RMSE-values generated by a set of return forecasting models estimated using either recursive (expanding window) estimation or rolling window estimation. Returns are measured monthly and computed for a value-weighted portfolio of US stocks.

Table 2: Root Mean Squared Error Performance (equal-weighted portfolio)

Window Strategy	**Model**	**1970-2005**	**1970s**	**1980s**	**1990s**
Recursive	Prevailing Mean	20.03	23.54	18.74	16.03
	Autoregressive	19.64	23.32	18.14	15.40
	Factor-augmented AR	20.11	23.94	17.72	15.51
	Exp. Smoothing	20.29	23.94	18.93	16.21
	Holt Smoothing	21.74	25.67	20.10	17.32
	STAR_1	20.14	24.39	18.41	15.51
	STAR_2	19.79	23.72	18.26	15.62
	One Layer NN	19.48	23.67	17.57	15.67
	Two Layer NN	19.26	22.95	17.58	15.67
	Previous Best Model	19.80	23.81	18.14	15.40
	Equal-weighted Avg.	19.57	23.33	17.93	15.56
Ten Year Rolling	Prevailing Mean	20.18	23.72	18.89	16.14
	Autoregressive	20.03	23.71	18.91	15.55
	Factor-augmented AR	20.71	24.67	18.91	15.55
	Exp. Smoothing	20.57	24.12	19.09	16.42
	Holt Smoothing	22.86	26.39	20.66	18.27
	STAR_1	20.99	23.94	21.85	15.52
	STAR_2	20.51	24.32	18.36	16.37
	One Layer NN	20.24	23.48	18.14	16.09
	Two Layer NN	20.64	23.44	18.16	16.35
	Previous Best Model	20.76	24.13	20.05	15.94
	Equal-weighted Avg.	19.88	23.46	18.25	15.69

Note: This table reports out-of-sample RMSE-values generated by a set of return forecasting models estimated using either recursive (expanding window) estimation or rolling window estimation. Returns are measured monthly and computed for an equal-weighted portfolio of US stocks.

Table 3: Root Mean Squared Error Performance (SMB portfolio)

Window Strategy	**Model**	**1970-2005**	**1970s**	**1980s**	**1990s**
Recursive	Prevailing Mean	11.44	11.81	8.15	10.10
	Autoregressive	11.35	11.73	8.14	9.97
	Factor-augmented AR	11.77	11.77	8.18	9.98
	Exp. Smoothing	11.53	11.88	8.26	10.30
	Holt Smoothing	13.08	13.44	9.16	11.18
	STAR_1	11.72	12.91	8.15	9.99
	STAR_2	11.77	12.04	8.20	10.07
	One Layer NN	11.63	12.11	8.42	10.05
	Two Layer NN	11.53	12.03	8.19	10.17
	Previous Best Model	11.37	11.79	8.14	9.97
	Equal-weighted Avg.	11.27	11.75	8.10	9.99
Ten Year Rolling	Prevailing Mean	11.53	11.90	8.34	10.11
	Autoregressive	11.86	11.95	8.35	10.16
	Factor-augmented AR	12.25	12.55	8.38	10.16
	Exp. Smoothing	11.92	11.84	8.28	10.27
	Holt Smoothing	13.29	12.75	9.15	11.49
	STAR_1	12.57	12.47	8.39	10.28
	STAR_2	12.92	12.54	8.37	10.13
	One Layer NN	12.56	12.46	8.56	10.57
	Two Layer NN	12.23	12.21	8.20	10.24
	Previous Best Model	12.70	12.66	8.36	10.45
	Equal-weighted Avg.	11.92	11.86	8.16	10.11

Note: This table reports out-of-sample RMSE-values generated by a set of return forecasting models estimated using either recursive (expanding window) estimation or rolling window estimation. Returns are measured monthly and computed for the Small minus Big (SMB) portfolio tracking return differentials between small and big stocks.

Table 4: Pair-wise comparison of out-of-sample Mean Squared Error Performance (value-weighted portfolio)

	AR	Factor-augm. AR	Exp Smoothing	Holt Smoothing	STAR1	STAR2	One Layer NN	Two Layer NN	Previous Best	Equal. Wht Average
Prev Mean	0.246	0.894	0.061	0.000	0.001	0.002	0.004	0.032	0.560	0.722
AR		0.594	0.077	0.000	0.001	0.002	0.011	0.082	0.969	0.776
Factor			0.062	0.000	0.001	0.002	0.010	0.060	0.498	0.776
Exp Smoothing				0.000	0.058	0.020	0.466	0.986	0.166	0.022
Holt Smoothing					0.015	0.139	0.002	0.001	0.000	0.000
STAR_1						0.346	0.153	0.035	0.001	0.001
STAR_2							0.059	0.022	0.004	0.002
One Layer NN								0.026	0.022	0.005
Two Layer NN									0.114	0.039
Previous Best										0.750

Note: This table reports p-values associated with pair-wise comparisons of out-of-sample MSE forecasting performance using the Giacomini-White (2006) test statistic. Small values indicate that the null of equal forecasting performance is rejected. Results are computed over the period from 1970-2005.

Table 5: Pair-wise comparison of out-of-sample Mean Squared Error Performance (equal-weighted portfolio)

	AR	Factor-	Exp	Holt	STAR1	STAR2	One Layer	Two Layer	Previous	Equal. Wht
		augm. AR	Smoothing	Smoothing			NN	NN	Best	Average
Prev Mean	0.500	0.625	0.188	0.001	0.292	0.456	0.909	0.509	0.154	0.267
AR		0.226	0.046	0.000	0.199	0.200	0.741	0.371	0.037	0.360
Factor			0.549	0.000	0.395	0.757	0.609	0.685	0.295	0.093
Exp Smoothing				0.000	0.581	0.875	0.305	0.910	0.621	0.014
Holt Smoothing					0.045	0.000	0.000	0.015	0.001	0.000
STAR_1						0.490	0.252	0.700	0.722	0.107
STAR_2							0.296	0.814	0.489	0.033
One Layer NN								0.376	0.121	0.270
Two Layer NN									0.867	0.214
Previous Best										0.003

Note: This table reports p-values associated with pair-wise comparisons of out-of-sample MSE forecasting performance using the Giacomini-White (2006) test statistic. Small values indicate that the null of equal forecasting performance is rejected. Results are computed over the period from 1970-2005.

Table 6: Pair-wise comparison of out-of-sample Mean Squared Error Performance (Small minus Big (SMB) portfolio)

	AR	Factor-	Exp	Holt	STAR1	STAR2	One Layer	Two Layer	Previous	Equal. Wht
		augm. AR	Smoothing	Smoothing			NN	NN	Best	Average
Prev Mean	0.292	0.142	0.373	0.036	0.258	0.194	0.050	0.141	0.191	0.296
AR		0.168	0.761	0.014	0.452	0.172	0.016	0.105	0.359	0.647
Factor			0.447	0.031	0.573	0.256	0.083	0.444	0.465	0.455
Exp Smoothing				0.006	0.498	0.145	0.017	0.148	0.402	0.951
Holt Smoothing					0.548	0.484	0.152	0.026	0.632	0.010
STAR_1						0.797	0.993	0.735	0.607	0.454
STAR_2							0.572	0.292	0.875	0.175
One Layer NN								0.124	0.866	0.003
Two Layer NN									0.630	0.129
Previous Best										0.357

Note: This table reports p-values associated with pair-wise comparisons of out-of-sample MSE forecasting performance using the Giacomini-White (2006) test statistic. Small values indicate that the null of equal forecasting performance is rejected. Results are computed over the period from 1970-2005.

Table 7: Sign tests for return forecasts

Model	**Full Sample**	**1970s**	**1980s**	**1990s**
Value-weighted returns				
Prevailing Mean	2.04	-0.62	2.56	0.00
Autoregressive	1.64	-1.06	2.32	0.00
Factor-augmented AR	2.49	1.68	2.57	-1.01
Exp. Smoothing	1.82	-0.52	2.12	0.00
Holt Smoothing	0.63	0.03	-0.28	-0.94
STAR_1	1.93	0.65	1.63	-0.35
STAR_2	1.58	0.14	1.75	-0.28
One Layer NN	0.80	0.85	1.75	-2.56
Two Layer NN	0.76	1.04	1.93	-2.31
Previous Best Model	2.02	0.05	2.57	-1.01
Equal-weighted Avg.	2.45	0.68	1.91	0.51
Equal-weighted returns				
Prevailing Mean	0.81	1.48	-0.94	0.00
Autoregressive	3.45	1.66	1.89	2.06
Factor-augmented AR	1.96	-0.38	2.49	1.58
Exp. Smoothing	1.34	1.23	-0.94	0.17
Holt Smoothing	1.93	1.23	1.35	0.96
STAR_1	0.80	-0.75	0.34	1.38
STAR_2	2.06	1.83	0.83	-0.93
One Layer NN	2.70	0.06	2.66	1.20
Two Layer NN	3.00	0.84	2.66	1.20
Previous Best Model	3.11	1.04	1.89	2.06
Equal-weighted Avg.	2.29	0.84	1.07	0.76
SMB portfolio returns				
Prevailing Mean	-1.39	-1.34	0.00	0.00
Autoregressive	1.82	3.04	-0.64	0.45
Factor-augmented AR	1.70	3.23	-0.64	-0.27
Exp. Smoothing	2.68	4.17	0.97	-0.24
Holt Smoothing	1.72	1.55	0.88	1.06
STAR_1	1.49	3.23	-0.64	0.06
STAR_2	1.19	3.05	-1.24	0.05
One Layer NN	2.09	3.07	0.37	0.20
Two Layer NN	2.57	3.98	0.20	0.26
Previous Best Model	1.72	2.85	-0.64	0.45
Equal-weighted Avg.	1.69	2.85	-1.23	0.90

Note: This table reports the value of the Pesaran-Timmermann (1992) test statistic for sign predictability which is asymptotically normally distributed.

Table 8: Out-of-sample root mean squared error (RMSE) performance of the adaptive forecast combination approach versus the prevailing mean.

window length	hurdle value	hurdle exceedance	RMSE prevailing mean	RMSE adaptive model	Bootstrapped 95% value
18 months	0.00	0.52	15.759	15.709	15.896
	0.01	0.52	--	15.709	15.882
	0.02	0.50	--	15.716	15.872
	0.05	0.39	--	15.728	15.814
	0.10	0.20	--	15.725	15.716
36 months	0.00	0.57	15.945	15.901	15.959
	0.01	0.57	--	15.904	15.928
	0.02	0.51	--	15.918	15.919
	0.05	0.28	--	15.876	15.904
	0.10	0.16	--	15.917	15.891
60 months	0.00	0.69	15.447	15.453	15.552
	0.01	0.60	--	15.456	15.539
	0.02	0.50	--	15.472	15.533
	0.05	0.31	--	15.423	15.484
	0.10	0.09	--	15.458	15.442
120 months	0.00	0.86	15.454	15.419	15.462
	0.01	0.54	--	15.420	15.454
	0.02	0.41	--	15.412	15.454
	0.05	0.18	--	15.463	15.454
	0.10	0.00	--	15.454	15.454

Note: The hurdle value is the minimum value of the rolling-window estimate of R-squared used to identify periods with return predictability. Hurdle exceedance is the percentage of periods where at least one forecasting model produces an R-squared higher than the hurdle value. Bootstrapped 95% values report the critical values for the RMSE-statistic which the RMSE must fall below in order to indicate return predictability.