Persistence in Forecasting Performance and Conditional Combination Strategies

Marco Aiolfi
Bocconi University

Allan Timmermann
UCSD

November 17, 2004

Abstract

This paper considers various measures of persistence in the (relative) forecasting performance of linear and nonlinear time-series models applied to a large cross-section of economic variables in the G7 countries. We find strong evidence of persistence among top and bottom forecasting models, but also systematic evidence of 'crossings' - where a previously good (poor) forecasting model delivers poor (good) future forecasting performance - among the linear models. Persistence in forecasting performance is related to the possibility of improving performance through forecast combinations. We propose a new four-stage conditional model combination method that first sorts models into clusters based on their past performance, then pools forecasts within each cluster, followed by estimation of the optimal forecast combination weights for these clusters and shrinkage towards equal weights. These methods are shown to work well empirically in out-of-sample forecasting experiments.

Keywords: Forecast combination, shrinkage, clustering, persistence in forecasting performance

JEL Classifications: C22, C53.

1 Introduction

Forecasts are of considerable importance to decision makers throughout economics and finance and are routinely used by private enterprises, government institutions and professional economists. It is therefore not surprising that much effort has gone into developing forecasting models ranging from simple, autoregressive specifications to complicated non-linear models and models with time-varying parameters. A multitude of forecasting models is typically considered because the true data generating process underlying a particular series of interest is unknown. Even the most complicated model is likely to be misspecified and can, at best, provide a reasonable 'local' approximation to the process driving the target variable.[1]

Model instability is a source of misspecification that is likely to be particularly relevant in practice, c.f. Stock and Watson (1996). In its presence, it is highly unlikely that a single model will dominate uniformly across time and the identity of the best local approximation is likely to change over time. If the identity of the best local model is time-varying, it is implausible that a forecasting strategy that, at each point in time, attempts to select the best current model will work well. Most obviously, if (ex-ante) the identity of the best model varies in a purely random way from period to period, it will not be possible to identify this model by considering past forecasting performance across models. Similarly, if a single best model exists but it only outperforms other models by a margin that is small relative to random sampling variation, it becomes difficult to identify this model by means of statistical methods based on past performance. Even if the single best model could be identified in this situation, it is likely that diversification gains from combining across a set of forecasting models with similar performance will dominate the strategy of only using a single forecasting model.

In practice, the factors that give rise to long-lasting changes in the ranking of different forecasting models - e.g., major oil price shocks, policy changes, institutional shifts or market participants' learning behavior - can either take the form of discrete shocks or gradually

[1] Conditions under which the true model is selected asymptotically are quite strict, c.f. White (1990) and Sin and White (1996), and are unlikely to be empirically relevant in situations characterized by a large cross-section of forecasting models and a short time-series dimension.

evolving shifts and one may expect the relative performance of forecasting models to display moderate degrees of persistence. How much persistence is a question of great practical relevance. Indeed, the popular strategy of assigning equal weights to the individual forecasting models (e.g., Clemen (1989)) becomes an optimal strategy if there is no ex-ante indication of the individual models' prospective out-of-sample forecasting performance, either because the models are of similar quality or because their (relative) performance is unstable over time.

Unfortunately, little is known about persistence in forecasting performance, so the first part of our paper considers this question, establishing 'stylized facts' by studying empirically a large cross-section of economic variables and forecasting methods.[2] We find systematic evidence of persistence among both top and bottom forecasting models, but also find evidence of 'crossings' - where a previously good (poor) forecasting model delivers poor (good) forecasting performance out-of-sample - among linear models.

In the presence of model misspecification of unknown form and moderate degrees of persistence in the relative performance of different forecasting models, no single econometric model can be expected to outperform all others and an attractive option is to combine forecasts from several models. In their seminal paper on forecast combinations, Bates and Granger (1969) already pointed to the importance of changes in models' relative performance over time as a determinant of the scope for combining forecasts. Key questions that arise when forecast combinations are considered is how wide a set of models to include (or, similarly, how many models to exclude), whether to estimate the combination weights, use a simple combination scheme such as equal-weighting or apply shrinkage methods. The answer to such questions depends on the distribution of (relative) forecasting performance across models and the degree of persistence and is hence closely linked to the first part of our analysis. We address these issues in the second part of the paper by comparing a wide range of combination schemes that differ along these dimensions, including a new set of conditional combination strategies.

[2]Stock and Watson (1999) consider combination methods based on expanding and rolling window estimators, two approaches that are usually associated with a stable and unstable data generating process, respectively.

The contributions of our paper are three-fold. First, we analyze the persistence in the relative forecasting performance of a range of linear and nonlinear models using a large international data set. Second, we propose a new four-stage approach for model combination that (i) sorts models into clusters based on their past performance; (ii) pools forecasts within each cluster; (iii) estimates the optimal forecast combination weights for these clusters; and (iv) shrinks the least squares combination weights towards equal weights. Third, we investigate empirically the out-of-sample forecasting performance of this new combination method and compares it with existing ones. We find that our approach improves upon existing combination methods using a range of economic variables in the G7 countries.

The paper is organized as follows. Section 2 studies the persistence of forecasting performance across a range of linear and nonlinear time-series models. Section 3 introduces the forecast combination problem and studies the out-of-sample forecasting performance of a range of standard combination methods proposed in the literature as well as our new four-stage combination method. Section 4 concludes.

2 Persistence in Forecasting Performance

2.1 Data Set

The seven-country data set that we use is the same as that used in Stock and Watson (2003). It consists of up to 43 quarterly time series for each of the G7 economies (Canada, France, Germany, Italy, Japan, UK, and the US) over the period 1959.I – 1999.IV, although some series are available only for a shorter period. The 43 series comprise a range of asset prices (including returns, interest rates and spreads); measures of real economic activity; wages and prices; and various measures of the money stock.[3] In many cases we use more than

[3] Following Stock and Watson (2004) the variables were subject to the following transformations. First, in a few cases the series contained a large outlier−such as spikes associated with strikes−and these outliers were replaced by interpolated values. Second, series that showed significant seasonal variation were seasonally adjusted using a linear approximation to X11 in order to avoid problems with non-linearities, c.f. Ghysels, Granger and Siklos (1996). Third, data series available on a monthly basis were aggregated to get quarterly observations.

one transformation of a given series. For example, interest rates are used both in levels and in first differences. Counting all the constructed variables (such as spreads) and different transformations of the same variable, the maximum number of time series per country is 75. Because the full data set contains some series that are available for short subsamples, for each country we select a balanced panel subset of the full data set that includes between 46 and 71 series per country.

2.2 Forecasting Models and Methods

h-step ahead forecasts of the conditional mean of the target variable, Y, are generated by time-series models of the form

$$y_{t+h} = f_i\left(\mathbf{x}_t; \boldsymbol{\theta}_{i,h}\right) + \varepsilon_{t+h,t,i}. \tag{1}$$

Here i is an index for the forecasting model, $\boldsymbol{\theta}_{i,h}$ is a vector of unknown parameters, $\varepsilon_{t+h,t,i}$ is an $h-$step error term and $\mathbf{x}_t$ is a vector of predictor variables that are known at time t and may include y_t. In general, individual forecasting models only use a subset of the elements of $\mathbf{x}_t$. All forecasts are computed recursively out-of-sample, so the forecast of y_{t+h} by the ith model is computed as $f_i\left(\mathbf{x}_t; \hat{\boldsymbol{\theta}}_{i,h,t}\right)$, where $\hat{\boldsymbol{\theta}}_{i,h,t}$ is the estimate of $\boldsymbol{\theta}_{i,h}$ given period-t information.

Following the analysis of Stock and Watson (1999), we consider both linear and non-linear forecasting models. The class of linear models comprises simple autoregressions with lag lengths selected recursively using the Bayes information criterion (BIC), including up to four lags:

$$y_{t+h} = c + A\left(L\right) y_t + \epsilon_{t+h}. \tag{2}$$

We also consider bivariate autoregressive models that include a single additional regressor, x_t, which is an element of $\mathbf{x}_t$:

$$y_{t+h} = c + A\left(L\right) y_t + B\left(L\right) x_t + \epsilon_{t+h}. \tag{3}$$

Lag lengths are again selected recursively using the BIC with between 1 and 4 lags of x_t and between 0 and 4 lags of y_t. The average number of linear model specifications varies

across series and countries. For example, it ranges from 36 for France, 38 for Italy, 43 for the UK, 44 for Canada and Germany, 51 for Japan to 67 for the US.

The class of non-linear forecasting models includes many of the models considered in Terasvirta, Tjostheim and Granger (1994). It includes 18 Artificial Neural Network (ANN) models with one and two hidden layers and different numbers of lags, p. Single layer feed-forward neural network models take the form

$$\begin{aligned} y_{t+h} &= \boldsymbol{\beta}_0'\boldsymbol{\zeta}_t + \sum_{i=1}^{n_1} \gamma_{1i} g\left(\boldsymbol{\beta}_{1i}'\boldsymbol{\zeta}_t\right) + \epsilon_{t+h}, \\ \boldsymbol{\zeta}_t &= \left(1, y_t, y_{t-1}, ..., y_{t-p-1}\right), \quad p = 1, 2, 3, \end{aligned} \tag{4}$$

where $g(z) = \frac{1}{1+e^z}$ is the logistic function. Neural network models with two hidden layers take the form

$$y_{t+h} = \boldsymbol{\beta}_0'\boldsymbol{\zeta}_t + \sum_{j=1}^{n_2} \gamma_{2j} g\left[\sum_{i=1}^{n_1} \beta_{2ji} g\left(\boldsymbol{\beta}_{1i}'\boldsymbol{\zeta}_t\right)\right] + \epsilon_{t+h}. \tag{5}$$

Our choice of design parameters for the single hidden layer ANN models are $n_1 = 1, 2, 3$ and $p = 1, 2, 3$, giving a total of nine basic models. Our choice for the ANN models with two hidden layers are $n_1 = 2, n_2 = 1, 2, 3$ and $p = 1, 2, 3$, producing nine basic models. These choices cover many of the basic neural net designs, c.f. Swanson and White (1995, 1997).[4]

We also consider 15 Logistic Smooth Transition Autoregression *(LSTAR)* models:

$$\begin{aligned} y_{t+h} &= \boldsymbol{\alpha}_0'\boldsymbol{\zeta}_t + d_t\boldsymbol{\beta}'\boldsymbol{\zeta}_t + \epsilon_{t+h} \\ d_t &= \frac{1}{1+\exp\left(\gamma_0 + \gamma_1\xi_t\right)}, \\ \boldsymbol{\zeta}_t &= \left(1, y_t, y_{t-1}, ..., y_{t-p-1}\right) \quad p = 1, 2, 3 \\ \xi_t &\in \left\{y_{t-1}, y_{t-2}, y_{t-3}, \Delta y_{t-1}, \Delta^2 y_{t-1}\right\}, \end{aligned} \tag{6}$$

where the scalar ξ_t is selected from the set in (6). LSTAR models differ by the variable used to define the transition and by the lag length p, c.f. Granger and Terasvirta (1993).

Finally, we consider time-varying autoregressions *(TVARs)* whose parameters are allowed

[4]For all ANN models, coefficients were estimated by recursive non-linear least squares, minimizing the objective function by an ad-hoc algorithm developed by Stock and Watson.

to evolve according to a multivariate random walk:

$$\begin{aligned} y_{t+h} &= \boldsymbol{\theta}'_{t,h}\boldsymbol{\xi}_t + \varepsilon_{t+h} \\ \boldsymbol{\theta}_{t,h} &= \boldsymbol{\theta}_{t-1,h} + \mathbf{u}_{t,h}, \\ \mathbf{u}_{t,h} &\sim iid\left(0, \lambda^2\sigma^2\mathbf{Q}\right). \end{aligned} \tag{7}$$

Here $\sigma^2\mathbf{Q}$ is the variance of $\mathbf{u}_{t,h}$. We consider seven different values of λ in the set $\{0.00,$ $0.0025,$ $0.005,$ $0.0075,$ $0.010,$ $0.015,$ $0.020\}$ and up to three lags for a total of 21 TVAR models, all of which are estimated by the Kalman filter.

To avoid extreme forecasts−a problem often associated with highly non-linear models−we implement the following trimming scheme. Forecasts exceeding four recursive standard deviations of the target variable are replaced by a recursive estimate of the unconditional mean of the dependent variable computed at the time of the forecast.

2.3 Sorting Windows

We implement an automatic procedure to control for missing values and outliers to produce a balanced panel of forecasts. Let T_0 be the point at which the first forecast is computed and let T be the final period. For each variable and forecast horizon, h, we produce a $((T-h-T_0+1)\times N_j)$ panel of forecasts

$$\widehat{\mathbf{Y}}_{T_0+h:T} = \begin{bmatrix} \widehat{y}^{(1)}_{T,T-h} & \widehat{y}^{(2)}_{T,T-h} & \cdots & \widehat{y}^{(N_j)}_{T,T-h} \\ \widehat{y}^{(1)}_{T-1,T-h-1} & \widehat{y}^{(2)}_{T-1,T-h-1} & \cdots & \widehat{y}^{(N_j)}_{T-1,T-h-1} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{y}^{(1)}_{T_0+h,T_0} & \widehat{y}^{(2)}_{T_0+h,T_0} & \cdots & \widehat{y}^{(N_j)}_{T_0+h,T_0} \end{bmatrix},$$

where $\widehat{y}^{(i)}_{t,t+h}$ is the h-step ahead forecast computed under the ith model at time t. The superscript, i, tracks the model, $i=1,...,N_j$, and N_j is the number of models for country or forecasting method j.[5]

The $h-$period performance of the ith forecasting model at time t is measured through the loss function, $L^{(i)}_{t+h,t} \equiv L\left(y_{t+h}, \widehat{y}^{(i)}_{t+h,t}\right)$. In line with common practice, we assume mean

[5]For each country some series are available only for shorter subsamples so instead of dropping these series from the panel we trade off the time-series and cross-sectional dimension. Given a $T\times N$ panel we minimize

squared forecast error (MSFE) loss:

$$L(y_{t+h}, \widehat{y}_{t+h,t,i}) = \left(e^{(i)}_{t+h,t}\right)^2, \tag{8}$$

where $e^{(i)}_{t+h,t} = y_{t+h} - \widehat{y}^{(i)}_{t+h,t}$ is the $h-$step forecast error associated with the ith model's prediction at time t. We track the historical forecasting performance over a window of the last win periods by computing $S^{(i)}_t = (1/win) \sum_{\tau=t-win+1}^{t} L^{(i)}_{\tau,\tau-h}$. To study how persistent forecasting performance is through time, we consider three different tracking or 'sorting' windows used to rank the forecasting models based on their historical performance:

1. Short window: $win = 1 : S^{(i)}_t = \left(e^{(i)}_{t,t-h}\right)^2$;
2. Rolling window: $win = 20 : S^{(i)}_t = (1/win) \sum_{\tau=t-win+1}^{t} \left(e^{(i)}_{\tau,\tau-h}\right)^2$;
3. Expanding window: $S^{(i)}_t = (1/(t-h-T_0+1)) \sum_{\tau=T_0+h}^{t} \left(e^{(i)}_{\tau,\tau-h}\right)^2$.

Future out-of-sample performance at time t is based on the h-period loss, $L^{(i)}_{t+h,t}$.

For each model, i, we record its rank at time t, $\mathcal{R}_{it} = f(S^{(1)}_t, ..., S^{(N_j)}_t)$. The model with the best MSFE forecasting performance, gets a rank of 1, the second best a rank of 2 and so on. Using these rank orders, we sort the models into quartiles and use 4×4 contingency tables to cross-tabulate the forecasting models' sorting-period performance against their out-of-sample performance.

2.4 Empirical Evidence

Tables 1-4 report empirical evidence on persistence in forecasting performance for the three sorting windows using linear (Tables 1 and 2) or nonlinear (Tables 3 and 4) models and forecast horizons of $h = 1, 2, 4, 8$ quarters. Transition probability estimates, $\hat{P}_{ij}$, in these

---

the following loss function with respect to α

$$\alpha^* = \arg\min_{\alpha} L(\alpha) = (T - T(\alpha))^2 + (N - N(\alpha))^2.$$

If a time series has more than αT missing values we drop it. This gives us a $T(\alpha^*) \times N(\alpha^*)$ panel of forecasts.

tables give the probability of moving from quartile i (based on historical performance up to time t) to quartile j (based on future performance, $e_{t+h,t}$, $t = T_0, ..., T-h$). We show only the top corners of the tables (i.e., $\hat{P}_{11}, \hat{P}_{14}, \hat{P}_{41}, \hat{P}_{44}$) since these effectively convey information about persistence or 'anti-persistence' in forecasting performance.

Under the null of no forecasting persistence, we have $P_{ij} = 0.25$ for all i, j since the probability of good (or bad) future performance should be unaffected by past performance. Persistent forecasting performance would lead to estimates of P_{11} and P_{44} above 0.25, while P_{14} and P_{41} (the probability that a historically good model becomes a poor future model or vice versa) should be well below 0.25. Conversely, anti-persistence corresponds to small values of P_{11} and P_{44} and large values of P_{14} and P_{41}. A chi-squared test statistic can easily be constructed for the estimated transition probabilities when $h = 1$. However, at longer horizons ($h \geq 2$) the data is overlapping so the performance statistics are serially correlated. To assess the statistical significance in this situation, we therefore use a bootstrap procedure to construct confidence intervals for the transition probability estimates, $\hat{P}_{ij}$. The proportion of transition probability estimates exceeding 0.25 with a p-value below 5% is reported in Tables 2 and 4 for linear and nonlinear models, respectively.

Several interesting results emerge from the tables. First, there is robust evidence of persistence among the linear forecasting models (Table 1). Across all sorting windows, forecast horizons and countries the average estimate of P_{11} is 0.30 with 76% of the estimates exceeding 25% at a statistically significant margin. Similar numbers are obtained for the worst performing models where the average estimate of P_{44} - averaged across countries, forecast horizons and sorting windows - is 0.30 with 73% of the estimates exceeding 0.25 at the 5% significance level.

The average estimate of P_{14} is 0.29 with 75% of the estimates being significantly greater than 0.25, suggesting that there is also a high chance that the historically best performing models become the future worst models. There is clearly a smaller chance of the reverse happening - i.e., that the historically worst models become the best future models - as the average estimate of P_{41} is 0.27 and only 52% of these estimates exceed 0.25 at the 5% critical level.

There is also considerable variation in the results across sorting windows. Persistence is systematically weaker the longer the sorting window, consistent with what one would expect under model instability. Going from an expanding via a rolling to a short sorting window, the average estimate of P_{11} rises from 0.29 to 0.30 and 0.31 with 60%, 76% and 91% of these estimates being significantly greater than 0.25. A similar pattern is observed in the average estimate of P_{44} which rises from 0.28 to 0.30 and 0.33 (with 49%, 72% and 98% of these estimates being significantly greater than 0.25 at the 5% level) and in the estimates of P_{14} and P_{41} with the former rising from 0.29 to 0.31 and the latter rising from 0.26 to 0.29 as the sorting window is shortened.

Results are largely invariant with respect to the forecast horizons (h) where both $\hat{P}_{11}$ and $\hat{P}_{44}$ are close to 0.30 irrespective of the value of h.[6] Between 73% and 78% of the $\hat{P}_{11}-$values and between 69% and 76% of the $\hat{P}_{44}-$values are significant at the 5% critical level.

Disaggregating the results by country, many interesting variations are observed in our data. The mean estimate of P_{11} (averaged across series, sorting windows and forecast horizons) is 0.30 for all countries, with the estimates for Japan and the US taking the smallest values. Among the worst models, the smallest persistence, measured by the average value of $\hat{P}_{44}$, is 0.28 for the US while the largest value is 0.32, recorded for Japan. This suggests that the weakest persistence is generally found in US time series. Large variations across sorting windows are also observed. For example, for the US forecasts at the shortest horizon ($h = 1$) the proportion of estimates of P_{44} that is significant at the 5% critical level increases from 8% to 68% and 97% as we move from the expanding via the rolling to the short sorting window.

Turning to the nonlinear models (Tables 3 and 4), there is generally a lower probability of 'crossings' and fewer of the off-diagonal transition probability estimates exceed 0.25, the average value of $\hat{P}_{14}$ and $\hat{P}_{41}$ being 0.22 and 0.25, respectively. Only 15% and 35% of these estimates are significant at the 5% critical level. Overall, persistence among top models is reduced to 0.26 with only 39% (compared with 76% in the case of the linear models) of the $\hat{P}_{11}-$values being significantly greater than 0.25. There is stronger persistence among the

[6] To save space, disaggregated results are not reported here, but these results are available on request from the authors.

worst nonlinear models than was found for the linear models, however: the average estimate of P_{44} is 0.33 and 85% of the estimates of P_{44} from the nonlinear models exceed 0.25 at the 5% critical level (compared to 73% for the linear models).

To verify the robustness of our results we also partitioned the forecasting models into groups of three based on their previous forecasting performance. We found very similar results with persistence in the top and bottom models' forecasting performance, stronger persistence among the worst linear forecasting models the shorter the sorting window and stronger persistence among the worst models for the nonlinear forecasts than for the linear forecasts.

We conclude the following from these findings. First, there is systematic evidence of persistence in forecasting performance both at the top end and at the bottom end of the rankings. Second, there is unfortunately also a strong tendency for the previous best linear models to become future underperformers. Third, there is in general stronger persistence in the forecasting performance of the worst non-linear models and a much lower probability of crossings in these models' forecasting performance than was observed for the linear models.

3 Forecast Combinations

The empirical evidence reported in Section 2 suggests that there is systematic persistence in the relative forecasting performance of standard time-series models. The extent to which this evidence can be translated into improved out-of-sample forecasting performance is still an open question, however. The moderate (albeit statistically significant) degree of persistence observed among top and bottom models suggests that a strategy of using a single 'top' model is unlikely to work well and that averaging across models could improve forecasting performance. As argued earlier, persistence in forecasting performance is likely to be a key determinant of the optimal degree of averaging across models, with less averaging being required the more persistent the performance is since this makes it easier to identify the best models from their historical track record.

Under mean squared error (MSE) loss the general forecast combination problem can be posed as that of choosing a mapping $\mathcal{R}^N \rightarrow \mathcal{R}$ from a vector of N predictions $\hat{\mathbf{y}}_{t+h,t} =$

$(\hat{y}^{(1)}_{t+h,t}, \hat{y}^{(2)}_{t+h,t}, ..., \hat{y}^{(N)}_{t+h,t})'$ to the real line, that best approximates the conditional expectation, $E[y_{t+h}|\hat{\mathbf{y}}_{t+h,t}]$. This general class of combination schemes comprises non-linear and time-varying combination methods, but it is far more common to limit the analysis by assuming a linear combination and choosing weights, $\boldsymbol{\omega}_{t,h} = (\omega^{(1)}_{t,h}, ..., \omega^{(N)}_{t,h})'$ to produce a combined forecast, $\hat{y}^{c}_{t+h,t} = \boldsymbol{\omega}'_{t,h}\hat{\mathbf{y}}_{t+h,t}$, resulting in the forecast error $e^{c}_{t+h,t} = y_{t+h} - \hat{y}^{c}_{t+h,t}$.

Assuming again that the forecaster's loss function $L(.)$ only depends on the forecast error, $e^{c}_{t+h,t}$, the optimal combination weights, ω^{*}_{th}, solve the problem

$$\boldsymbol{\omega}^{*}_{th} = \arg\min_{\boldsymbol{\omega}_{th}} E\left[L\left(e^{c}_{t+h,t}\right)|\hat{\mathbf{y}}_{t+h,t}\right]. \tag{9}$$

Under MSE loss, $L(e) = e^2$, the combination weights are easy to characterize in population and only depend on the first two conditional moments of the joint distribution of y_{t+h} and $\hat{\mathbf{y}}_{t+h,t}$,

$$\begin{pmatrix} y_{t+h} \\ \hat{\mathbf{y}}_{t+h,t} \end{pmatrix} \sim \left(\begin{pmatrix} \mu_{yth} \\ \boldsymbol{\mu}_{th} \end{pmatrix} \begin{pmatrix} \sigma^2_{yth} & \boldsymbol{\sigma}'_{y\hat{\mathbf{y}}th} \\ \boldsymbol{\sigma}_{y\hat{\mathbf{y}}th} & \boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th} \end{pmatrix} \right).$$

Assuming that $\boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th}$ is invertible, the solution to equation (9) is

$$\boldsymbol{\omega}^{*}_{th} = (\boldsymbol{\mu}_{th}\boldsymbol{\mu}'_{th} + \boldsymbol{\Sigma}^{-1}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th})(\boldsymbol{\mu}_{th}\mu_{yth} + \boldsymbol{\sigma}_{y\hat{\mathbf{y}}th}). \tag{10}$$

If y_{t+h} is projected on a constant as well as on the forecasts, $\hat{\mathbf{y}}_{t+h,t}$,[7] the optimal (population) values of the constant and the combination weights, ω^{c*}_{0th} and $\boldsymbol{\omega}^{*}_{th}$, are

$$\begin{aligned} \omega^{c*}_{0th} &= \mu_{yth} - \boldsymbol{\omega}'_{th}\boldsymbol{\mu}_{th} \\ \boldsymbol{\omega}^{*}_{th} &= \boldsymbol{\Sigma}^{-1}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th}\boldsymbol{\sigma}_{y\hat{\mathbf{y}}th}. \end{aligned} \tag{11}$$

These weights depend on the full conditional covariance matrix of forecasts, $\boldsymbol{\Sigma}_{\hat{\mathbf{y}}\hat{\mathbf{y}}th}$. However, given a large number of forecasting models (N) relative to the number of time-series

[7] Including a constant to capture bias effects is a strategy recommended (under MSE loss) by Granger and Ramanathan (1984) and, for a variety of loss functions, by Elliott and Timmermann (2004). Ruling out that the covariance matrix, $\Sigma_{\hat{\mathbf{y}}\hat{\mathbf{y}}th}$, is singular is innocuous here since one can always drop superfluous forecasts from the combination. One could alternatively consider non-linear combination schemes that do not impose this restriction and allow individual forecasts to be perfectly linearly correlated as long as non-linear transformations of the forecasts are not perfectly correlated.

observations (T), it is generally not feasible or desirable to estimate optimal combination weights at the level of the individual forecasts.

A special case of (11) arises when one model - e.g. the ith model - has a much smaller forecast error than the other models. In this case, to an approximation, only a single forecast gets selected:

$$\boldsymbol{\omega}_{th}^{*} \approx \boldsymbol{\vartheta}_{i}, \tag{12}$$

where $\boldsymbol{\vartheta}_{i}$ is an $N-$vector with zeros everywhere except for unity in the ith place.

The opposite case arises when the forecasting errors are all (roughly) of the same size with similar correlations, in which case

$$\boldsymbol{\omega}_{t}^{*} \approx \boldsymbol{\iota}_{N}/N, \tag{13}$$

where $\boldsymbol{\iota}_{N}$ is an $N-$vector of ones. It is often found in the empirical literature that estimated "optimal" combination weights based on (11) lead to worse forecasting performance than such simple equal-weighted averages, (13), c.f. Clemen (1989).

3.1 Conditional Forecast Combination Strategies

Standard model selection schemes such as (12) and forecast combination schemes such as (11) or (13) suffer from a number of problems. With N large relative to T, estimation of the "optimal" combination weights (11) is either not feasible or is surrounded by considerable sampling error. While the forecasting methods in (12) and (13) do not suffer from this problem, they ignore correlation structure across different forecasts and do not efficiently use all information in the joint distribution of the forecast errors. For this reason, we propose a range of new (conditional) combination strategies that in a first stage sort the forecasting models into groups based on their recent historical forecasting performance, then pool forecasts within groups and finally combine the pooled forecasts for selected groups of models using least squares estimates of the combination weights followed by shrinkage towards equal weights.

3.1.1 Combination of Forecasts from Pre-selected Quartiles

The first set of combination methods operates at the level of quartile-sorted forecasts and uses information on the estimated transition probabilities to select which quartiles to include in the combination. Forecasting models are initially assigned to quartiles based on their historical forecasting performance up to the point of the prediction, t. For each quartile, a pooled (average) forecast is then computed. If the transition probability estimates (using information up to time t) suggest that a particular quartile of models produced better than average forecasts, then the pooled forecast from models in this quartile is included in the combination.

Pooling by quartile reduces the number of forecasts to between one and four. This is a number that is small enough to let us consider estimating optimal combination weights by least squares. We also consider shrinking the least-squares estimates of the combination weights towards equal-weights, c.f. Diebold and Pauly (1990):

$$\widehat{\boldsymbol{\omega}}_t(\hat{\psi}_t) = \hat{\psi}_t \widehat{\boldsymbol{\omega}}_t^{OLS} + (1 - \hat{\psi}_t)\overline{\boldsymbol{\omega}}, \tag{14}$$

where $\hat{\psi}_t$, the parameter governing the amount of shrinkage, is a function of the data. This estimator shrinks the least squares estimate of the combination weights, $\widehat{\boldsymbol{\omega}}_t^{OLS}$, towards equal weights, $\overline{\boldsymbol{\omega}}$. As an extreme case, this includes simply using equal weights. Shrinkage estimators can often improve the small sample performance of forecast combinations.[8]

To set out the combination strategy, let $\hat{\mathbf{y}}_{t+h,t}^{q}$ be the $N_q \times 1$ vector containing the forecasts belonging to quartile q, where N_q is the number of models in quartile q. We use the persistence information contained in the estimated transition probabilities at time t, $\hat{P}_{ijt}$, to select quartiles as follows:

If $\hat{P}_{11t} > \hat{P}_{14t}$: include the pooled forecast from models in the top quartile.

If $\hat{P}_{21t} + \hat{P}_{22t} > \hat{P}_{23t} + \hat{P}_{24t}$: include the pooled forecast from models in the second quartile.

If $\hat{P}_{31t} + \hat{P}_{32t} > \hat{P}_{33t} + \hat{P}_{34t}$: include the pooled forecasts from models in the third quartile.

[8] Elliott (2002) establishes conditions under which the expected loss from averaging gets closer to the expected loss from using the optimal weights as the number of forecasts (N) increases.

If $\hat{P}_{41t} > \hat{P}_{44t}$: include the pooled forecasts from models in the fourth quartile.

Let $\mathcal{I}_i$ be an indicator variable taking the value 1 if the ith quartile is included and otherwise zero, while $\boldsymbol{\iota}_{N_q}$ is an $N_q \times 1$ vector of ones. Then we consider four types of combination weights applied to the forecasts pooled into quartiles, namely previous best (PB), equal-weighted (EW), optimally weighted (OW) and shrinkage-weighted (SW) combinations:

1. $PB : \hat{y}^c_{t+h,t} = (\boldsymbol{\iota}'_{N_1}/N_1)\hat{\mathbf{y}}^1_{t+h,t}$.
2. $EW : \hat{y}^c_{t+h,t} = \left(\sum_{q=1}^4 \mathcal{I}_q\right)^{-1} \sum_{q=1}^4 \mathcal{I}_q(\boldsymbol{\iota}'_{N_q}/N_q)\hat{\mathbf{y}}^q_{t+h,t}$.
3. $OW : \hat{y}^c_{t+h,t} = \sum_{q=1}^4 \mathcal{I}_q\hat{\omega}_{qt}\left[(\boldsymbol{\iota}'_{N_q}/N_q)\hat{\mathbf{y}}^q_{t+h,t}\right]$, where $\hat{\omega}_{qt}$ are least squares estimates of the optimal combination weights for the included quartiles.
4. $SW : \hat{y}^c_{t+h,t} = \sum_{q=1}^4 \mathcal{I}_q\hat{s}_{qt}\left[(\boldsymbol{\iota}'_{N_q}/N_q)\hat{\mathbf{y}}^q_{t+h,t}\right]$, where $\hat{s}_{qt}$ are shrinkage weights applied to the selected quartiles, computed as $\hat{s}_{qt} = \psi_t\hat{\omega}_{qt} + (1-\psi_t)\left(\sum_{q=1}^4 \mathcal{I}_q\right)^{-1}$, $\quad \psi_t = \max\left\{0, 1-\kappa\left(\frac{\sum_{q=1}^4 \mathcal{I}_q}{t-h-T_0-\sum_{q=1}^4 \mathcal{I}_q}\right)\right\}$.

Quartile-sorted combinations are referred to as $Q(W,Z)$ where $W \in \{PB, EW, OW, SW\}$ and $Z \in \{L, M, H\}$ captures the degree of shrinkage. As κ goes up, ψ_t declines and the degree of shrinkage increases so the choices of $\kappa = 2.5, 5, 7.5$ represent low, medium and strong shrinkage. These are denoted by $Q(SW,L)$, $Q(SW,M)$ and $Q(SW,H)$, respectively.[9] If none of the quartiles passes the test in the first step, we set $Q(EW) = Q(OW) = Q(SW,.) = \frac{1}{N}\sum_{i=1}^N \hat{y}^{(i)}_{t+h,t}$ and average across all forecasting models.

Application of these methods requires that part of the out-of-sample period is used to establish an initial ranking of the models. We use the first 20 out-of-sample observations as our initial sorting period.

[9] These values are higher than the values (0.25, 0.5 and 1) considered by Stock and Watson (2004). This is because we apply shrinkage to the grouped (quartile) forecasts whereas Stock and Watson apply shrinkage to the original set of models (N) so the ratio of the number of forecasts to the effective sample size is much larger in their application than in ours. Hence we need larger values of κ to accomplish a similar degree of shrinkage.

3.1.2 Clustering by K-mean algorithm

The approach of sorting models into quartiles can be criticized for using arbitrary cut-off points. Two models with very similar in-sample forecasting performance may get assigned to different quartiles with different weights. To deal with this problem, we propose to use a $K-$mean clustering algorithm that divides the models into a finite number of clusters based on their past forecasting performance.

To motivate this approach, Figure 1 plots the in-sample MSFE performance for output growth (up to period T_0) against the out-of-sample forecasting performance across linear forecasting models while Figure 2 does the same for the non-linear models. In general there is not much support for a simple monotonic or linear relationship between past and future forecasting performance. However, there are indications of performance clusters in some countries, notably France, Germany and Japan. There is also some evidence - notably for Italy and Japan - that the models with the very worst in-sample forecasting performance tend to generate the highest out-of-sample MSFE-values. This suggests trimming the worst models prior to computing forecasts. Trimming is particularly appealing if many models underperform the unconditional mean forecast but may also work more generally if there is a large and persistent spread in the forecasting performance across models.[10]

Suppose we identify K clusters and let $\hat{\mathbf{y}}^k_{t+h,t}$ be the $N_k \times 1$ vector containing the subset of forecasts belonging to cluster $k \in \{1, ..., K\}$ where the first cluster contains the models with the lowest historical MSFE values. We consider the following conditional combination strategies:

1. $PB : \hat{y}^c_{t+h,t} = (\boldsymbol{\iota}'_{N_1}/N_1)\hat{\mathbf{y}}^1_{t+h,t}$ select the cluster with the lowest in-sample MSFE-values and use the simple mean of the forecasts in this cluster.

2. $EW : \hat{y}^c_{t+h,t} = \frac{1}{K-1}\sum_{k=1}^{K-1}(\boldsymbol{\iota}'_{N_k}/N_k)\hat{\mathbf{y}}^k_{t+h,t}$ exclude the worst cluster and apply equal-weights to the forecasts from the top $K-1$ clusters.

3. $OW : \hat{y}^c_{t+h,t} = \sum_{k=1}^{K}\hat{\omega}_{kt}\left[(\boldsymbol{\iota}'_{N_k}/N_k)\hat{\mathbf{y}}^k_{t+h,t}\right]$, where $\hat{\omega}_{kt}$ are least-squares estimates of the

[10] See also Aiolfi and Favero (2003) and Granger and Jeon (2004) who argue in favor of trimming the worst models followed by computation of a simple equal-weighted average of the remaining forecasts.

optimal combination weights for the K clusters.

4. SW : $\hat{y}^c_{t+h,t} = \sum_{k=1}^{K} \hat{s}_{kt} \left[(\boldsymbol{\iota}'_{N_k}/N_k) \hat{\mathbf{y}}^k_{t+h,t} \right]$, where $\hat{s}_{kt}$ are the shrinkage weights for the K clusters, computed as: $\hat{s}_{kt} = \psi_t \hat{\omega}_{kt} + (1 - \psi_t)\frac{1}{K}$, $\psi_t = \max\left\{0, 1 - \kappa\left(\frac{K}{t-h-T_0-K}\right)\right\}$, $\kappa = 2.5, 5, 7.5$.

Cluster-sorted combinations are referred to as $C(K, W, Z)$ where K is the number of clusters, $W \in \{PB, EW, OW, SW\}$ and $Z \in \{L, M, H\}$ measures the degree of shrinkage. Hence we use the notation $C(K, SW, L)$, $C(K, SW, M)$ and $C(K, SW, H)$ for the cluster combination based on K clusters with low, medium and high shrinkage weights, respectively. We set $K = 2, 3$ and use either two or three clusters.

3.2 Empirical Results

Results from a set of standard forecasting strategies (previous best single model (PB), equal-weighted average ($Q(EW)$) and top quartile ($Q(PB)$)) as well as from the four-step conditional combination strategies are presented in Table 5 (linear models) and Table 6 (non-linear models) which summarize the distribution of out-of-sample MSFE performance across countries and horizons.[11] Performance is reported relative to the out-of-sample MSFE performance of the previous best (PB) single model selected using an expanding sorting window.

First consider the results for the linear forecasting models (Table 5). Consistent with earlier studies we find that the equal-weighted combination ("mean" or $Q(EW)$ forecast) produces good forecasts that dominate the forecasts from the previous best (PB) model. Interestingly, however, the better conditional combination strategies outperform the equal-weighted forecasts overall.

Comparing the overall forecasting performance across combination strategies, the best methods appear to be either least squares estimates of the combination weights for the selected quartiles followed by relatively strong shrinkage towards equal weights ($Q(SW, M)$

[11] Our notation implies that PB is the forecast from the previous best single model while $Q(PB)$ is the average forecast from the quartile of previous best models and $Q(EW)$ is the average forecast computed across all models.

and $Q(SW, H)$) or a simple average of forecasts in the top cluster ($C(2, PB)$ or $C(3, PB)$). Shrinkage towards equal weights systematically improves the forecasting performance.[12]

Turning to the results for the non-linear models shown in Table 6, the methods involving pooling within quartile-ranked forecasts followed by estimation of optimal combination weights and shrinkage towards equal weights ($Q(SW, M)$ and $Q(SW, H)$) or pooling within the top cluster of models continue to perform better on average than any of the other methods, including using the mean forecast or the average forecast from the top quartile of models.[13] Once again, the combination schemes with the strongest degree of shrinkage lead to the best overall forecasting performance.

The robustness of our results across linear and non-linear forecasting models is reassuring and suggests that two mechanisms lead to better forecasting performance. First, even though it is difficult to identify the top model among forecasting models with similar performance, it is possible to identify clusters of good and bad models. Second, and related to this point, provided that the models are pooled into groups based on their past performance, least squares estimation of the combination weights (which accounts for the correlation structure between forecasts) is a useful step. However, the estimated combination weights are surrounded by sufficiently large sampling errors that shrinkage towards equal weights generally improves on the forecasting performance.

4 Conclusion

This paper investigated the extent of persistence in forecasting performance across a large set of linear and nonlinear models. Much of the paper was exploratory since there is not, to our knowledge, any previous research on this question. We found significant evidence of persistence in forecasting performance. Models that were in the top and bottom quartiles

[12] We do not report formal tests for significance of relative performance since the model choice is data driven. Hence the model under the null is sometimes nesting, while at other times does not nest, the models under the alternative. See also Stock and Watson (2004) for a discussion of this point.

[13] For the nonlinear models, in most cases only two clusters were clearly identified so we restrict the non-linear results to two clusters.

when ranked by their recent historical performance have a higher than average chance of remaining in the top and bottom quartiles, respectively, in future periods. However, we also found systematic evidence of 'crossings'−where the previous best models become the future worst models or vice versa−among the linear forecasting models. The ranking of the worst forecasts tended to be more persistent for non-linear models than for linear models, possibly due to the fact that some of the nonlinear models are grossly misspecified−and more strongly affected by parameter estimation error−while the performance of the linear models tends to be more robust in this regard.

We next linked this evidence to the possibility of producing improved forecasts, arguing that it is likely that conditional combination strategies (which use information on past forecasting performance) can be designed under the persistence in forecasting performance documented in our paper. We proposed a set of new combination strategies that first sort models into either quartiles or clusters on the basis of the distribution of past forecasting performance across models, pool forecasts within each cluster and then estimate optimal combination weights and shrink these towards equal weights. This combination scheme makes use of many of the techniques proposed in the literature for improving forecast combinations such as trimming, pooling, optimal weighting and shrinkage estimation. We find evidence in our data that these conditional combination strategies lead to better overall forecasting performance than simpler strategies in common use such as using the previous best model or simply averaging across all forecasting models or a small subset of these.

Acknowledgements

We received many helpful comments from an anonymous referee, Eric Ghysels, seminar participants at Bocconi University and at the January 2004 San Diego conference in honor of Clive Granger.

References

Aiolfi, M. and C. A. Favero, 2003, Model Uncertainty, Thick Modeling and the Predictability of Stock Returns. Forthcoming in Journal of Forecasting.

Bates, J.M. and C.W.J. Granger, 1969, The Combination of Forecasts, Operations Research Quarterly 20, 451-468.

Clemen, R.T., 1989, Combining Forecasts: A Review and Annotated Bibliography, International Journal of Forecasting 5, 559-581.

Diebold, F.X. and P. Pauly, 1990, The Use of Prior Information in Forecast Combination, International Journal of Forecasting 6, 503-508.

Elliott, G., 2002, Forecast Combination with Many Forecasts, Mimeo, UCSD.

Elliott, G. and A. Timmermann, 2004, Optimal Forecast Combinations under General Loss Functions and Forecast Error Distributions, Journal of Econometrics 122, 47-79.

Ghysels, E., C.W.J. Granger and P.L. Siklos, 1996, Is Seasonal Adjustment a Linear of Nonlinear Data Filtering Process?, Journal of Business and Economic Statistics 14, 374-86.

Granger, C.W.J. and Y. Jeon, 2004, Thick Modeling, Economic Modelling 21, 323-343.

Granger, C.W.J. and R. Ramanathan, 1984, Improved Methods of Combining Forecasts. Journal of Forecasting 3, 197-204.

Granger, C.W.J. and T. Terasvirta, 1993, Modeling Nonlinear Economic Relationships (Oxford University Press, Oxford).

Sin, C.Y. and H. White, 1996, Information Criterion for Selecting Possibly Mis-specified Parametric Models, Journal of Econometrics 71, 201-225.

Stock, J.H. and M.W. Watson, 1996, Evidence on structural instability in macroeconomic time series relations, Journal of Business and Economic Statistics 14, 11-30.

Stock, J.H. and M.W. Watson, 1999, A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series, in: R.F. Engle and H. White, eds, Festschrift in Honour of Clive Granger, 1-44.

Stock, J.H. and M.W. Watson, 2004, Combination Forecasts of Output Growth in a Seven-Country Data Set, Journal of Forecasting 23, 405-430.

Swanson, N.R. 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, 265-75.

Swanson, N.R. and H. White, 1997, A Model Selection Approach to Real-Time Macroeco-

nomic Forecasting Using Linear Models and Artificial Neural Networks, Review of Economics and Statistics 79, 540-550.

Terasvirta, T., D. Tjostheim, and C.W.J. Granger, 1994, Aspects of Modelling Nonlinear Time Series, in: R. Engle and D. McFadden, eds, Handbook of Econometrics, Vol. 4, (Elsevier: Amsterdam), 2919-60.

White, H., 1990, A Consistent Model Selection, in: C.W.J. Granger, ed., Modeling Economic Series, (Oxford University Press, Oxford)

Table 1: Transition probabilities estimated for the linear models. Each cell reports the corner probabilities of the 4x4 contingency table tracking the forecasting models' initial and subsequent h-period rankings. Transition probabilities P_{ij} give the probability of moving from quartile i (based on historical performance, $e^2_{t,t-h}$) to quartile j (based on future performance, $e^2_{t+h,t}$). All estimates are averaged across variables within a particular country.

Expanding Sorting Window								
	h=1		h=2		h=4		h=8	
USA	0.30	0.31	0.29	0.30	0.29	0.29	0.29	0.30
	0.23	0.23	0.24	0.25	0.26	0.26	0.26	0.26
UK	0.29	0.29	0.28	0.28	0.29	0.28	0.29	0.31
	0.27	0.29	0.27	0.30	0.27	0.30	0.28	0.30
France	0.28	0.28	0.28	0.27	0.28	0.28	0.28	0.28
	0.26	0.28	0.27	0.29	0.27	0.29	0.29	0.29
Germany	0.29	0.30	0.29	0.29	0.28	0.29	0.29	0.30
	0.26	0.26	0.26	0.27	0.27	0.28	0.27	0.29
Japan	0.29	0.27	0.28	0.26	0.29	0.26	0.27	0.27
	0.26	0.29	0.26	0.30	0.26	0.32	0.28	0.30
Canada	0.30	0.31	0.30	0.31	0.30	0.30	0.28	0.28
	0.24	0.25	0.24	0.25	0.26	0.27	0.27	0.27
Italy	0.28	0.28	0.28	0.27	0.27	0.27	0.29	0.28
	0.27	0.28	0.26	0.28	0.27	0.28	0.28	0.29
Rolling Sorting Window								
	h=1		h=2		h=4		h=8	
USA	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.31
	0.25	0.27	0.26	0.29	0.27	0.28	0.27	0.28
UK	0.29	0.29	0.30	0.29	0.30	0.30	0.30	0.31
	0.28	0.31	0.28	0.32	0.28	0.31	0.28	0.31
France	0.30	0.28	0.30	0.28	0.31	0.29	0.31	0.31
	0.28	0.31	0.27	0.31	0.27	0.31	0.28	0.30
Germany	0.31	0.30	0.30	0.31	0.30	0.32	0.31	0.32
	0.27	0.29	0.27	0.29	0.28	0.29	0.28	0.29
Japan	0.30	0.27	0.29	0.27	0.30	0.27	0.30	0.30
	0.27	0.31	0.27	0.32	0.26	0.33	0.29	0.31
Canada	0.31	0.31	0.31	0.32	0.31	0.31	0.31	0.32
	0.26	0.27	0.26	0.28	0.26	0.29	0.27	0.28
Italy	0.29	0.28	0.30	0.29	0.30	0.29	0.31	0.30
	0.28	0.31	0.27	0.32	0.28	0.31	0.28	0.31
Short Sorting Window								
	h=1		h=2		h=4		h=8	
USA	0.30	0.30	0.31	0.31	0.31	0.30	0.30	0.30
	0.30	0.33	0.29	0.33	0.29	0.32	0.28	0.30
UK	0.30	0.32	0.31	0.32	0.32	0.30	0.32	0.30
	0.31	0.34	0.30	0.35	0.28	0.36	0.29	0.33
France	0.32	0.30	0.33	0.30	0.32	0.30	0.34	0.30
	0.29	0.35	0.28	0.35	0.29	0.33	0.27	0.34
Germany	0.32	0.31	0.31	0.32	0.32	0.31	0.31	0.31
	0.30	0.34	0.30	0.33	0.30	0.33	0.29	0.31
Japan	0.30	0.31	0.31	0.30	0.32	0.29	0.32	0.30
	0.30	0.34	0.29	0.34	0.27	0.36	0.28	0.34
Canada	0.30	0.32	0.31	0.32	0.31	0.31	0.31	0.31
	0.30	0.32	0.29	0.32	0.29	0.32	0.28	0.31
Italy	0.32	0.30	0.32	0.30	0.33	0.28	0.33	0.30
	0.30	0.35	0.29	0.35	0.27	0.36	0.28	0.35

Table 2: Significance of transition probabilities estimated for the linear models. Each cell reports the percentage of corner probabilities in Table 1 that is greater than 0.25 at the 5% significance level.

Expanding Sorting Window								
	h=1		h=2		h=4		h=8	
USA	0.88	0.89	0.72	0.80	0.66	0.78	0.68	0.81
	0.05	0.08	0.21	0.23	0.32	0.36	0.30	0.42
UK	0.68	0.66	0.54	0.61	0.56	0.65	0.65	0.69
	0.46	0.52	0.39	0.66	0.44	0.67	0.53	0.67
France	0.57	0.57	0.60	0.45	0.47	0.49	0.36	0.45
	0.32	0.49	0.36	0.49	0.47	0.60	0.55	0.64
Germany	0.62	0.66	0.62	0.66	0.53	0.67	0.52	0.70
	0.30	0.40	0.36	0.40	0.57	0.63	0.58	0.64
Japan	0.65	0.47	0.50	0.39	0.55	0.35	0.48	0.53
	0.31	0.60	0.34	0.69	0.33	0.78	0.52	0.64
Canada	0.75	0.82	0.67	0.82	0.65	0.67	0.46	0.52
	0.18	0.18	0.14	0.18	0.31	0.45	0.43	0.44
Italy	0.43	0.61	0.53	0.53	0.43	0.43	0.49	0.53
	0.41	0.47	0.41	0.57	0.43	0.51	0.49	0.62
Rolling Sorting Window								
	h=1		h=2		h=4		h=8	
USA	0.95	0.92	0.93	0.89	0.93	0.93	0.77	0.90
	0.23	0.68	0.36	0.76	0.53	0.70	0.45	0.62
UK	0.68	0.57	0.70	0.66	0.67	0.76	0.71	0.90
	0.57	0.79	0.57	0.82	0.56	0.76	0.47	0.78
France	0.72	0.49	0.70	0.62	0.71	0.69	0.64	0.80
	0.51	0.85	0.45	0.81	0.47	0.78	0.45	0.70
Germany	0.79	0.79	0.74	0.92	0.82	0.88	0.72	0.88
	0.51	0.74	0.53	0.72	0.71	0.65	0.46	0.64
Japan	0.71	0.50	0.61	0.44	0.70	0.48	0.66	0.79
	0.40	0.81	0.50	0.85	0.47	0.85	0.69	0.76
Canada	0.88	0.98	0.93	0.96	0.84	0.93	0.85	0.93
	0.25	0.42	0.32	0.44	0.29	0.73	0.37	0.57
Italy	0.69	0.71	0.69	0.59	0.68	0.60	0.70	0.70
	0.47	0.71	0.41	0.84	0.43	0.74	0.55	0.74
Short Sorting Window								
	h=1		h=2		h=4		h=8	
USA	0.84	0.83	0.99	0.93	0.99	0.93	0.93	0.97
	0.83	0.97	0.81	1.00	0.84	1.00	0.81	0.92
UK	0.88	0.93	0.84	0.91	0.87	0.69	0.84	0.73
	0.93	0.98	0.80	0.96	0.59	1.00	0.61	0.96
France	0.91	0.85	0.96	0.74	0.82	0.80	0.93	0.82
	0.66	1.00	0.64	1.00	0.60	0.96	0.45	0.93
Germany	0.98	0.94	0.92	0.94	0.98	0.94	0.92	0.98
	0.89	1.00	0.85	0.96	0.84	0.98	0.70	0.90
Japan	0.73	0.84	0.87	0.69	0.90	0.62	0.91	0.81
	0.82	1.00	0.68	0.98	0.55	1.00	0.62	0.97
Canada	0.88	0.91	0.96	1.00	0.95	0.89	0.96	0.96
	0.88	0.96	0.89	1.00	0.80	0.96	0.65	0.96
Italy	0.92	0.78	0.92	0.88	0.89	0.68	0.89	0.77
	0.78	1.00	0.71	1.00	0.40	1.00	0.57	1.00

Table 3: Transition probabilities estimated for the non linear models. Each cell reports the corner probabilities of the 4x4 contingency table tracking the forecasting models' initial and subsequent h-period rankings. Transition probabilities P_{ij} give the probability of moving from quartile i (based on historical performance, $e^2_{t,t-h}$) to quartile j (based on future performance, $e^2_{t+h,t}$). All estimates are averaged across variables within a particular country.

Expanding Sorting Window								
	h=1		h=2		h=4		h=8	
USA	0.25	0.20	0.26	0.19	0.26	0.22	0.26	0.23
	0.25	0.34	0.25	0.33	0.26	0.31	0.25	0.30
UK	0.25	0.19	0.25	0.19	0.26	0.21	0.25	0.21
	0.25	0.35	0.25	0.33	0.26	0.33	0.25	0.33
France	0.25	0.19	0.25	0.21	0.25	0.22	0.26	0.23
	0.25	0.34	0.25	0.31	0.26	0.31	0.26	0.29
Germany	0.26	0.19	0.25	0.19	0.26	0.19	0.27	0.20
	0.24	0.35	0.25	0.33	0.24	0.34	0.23	0.35
Japan	0.25	0.18	0.25	0.19	0.26	0.20	0.27	0.23
	0.25	0.35	0.26	0.32	0.25	0.32	0.25	0.30
Canada	0.25	0.20	0.26	0.20	0.26	0.22	0.26	0.21
	0.25	0.35	0.25	0.33	0.25	0.32	0.25	0.33
Italy	0.25	0.20	0.25	0.20	0.26	0.22	0.25	0.22
	0.25	0.33	0.25	0.32	0.25	0.30	0.25	0.29
Rolling Sorting Window								
	h=1		h=2		h=4		h=8	
USA	0.26	0.20	0.26	0.20	0.26	0.21	0.26	0.22
	0.26	0.36	0.26	0.34	0.26	0.32	0.26	0.31
UK	0.26	0.19	0.26	0.20	0.27	0.20	0.27	0.20
	0.25	0.36	0.25	0.34	0.25	0.34	0.25	0.33
France	0.26	0.20	0.26	0.21	0.26	0.22	0.27	0.23
	0.25	0.35	0.25	0.33	0.26	0.32	0.26	0.30
Germany	0.26	0.19	0.26	0.20	0.27	0.20	0.28	0.20
	0.24	0.35	0.25	0.33	0.24	0.34	0.23	0.34
Japan	0.27	0.18	0.26	0.19	0.27	0.20	0.28	0.22
	0.25	0.36	0.25	0.34	0.25	0.33	0.25	0.31
Canada	0.25	0.20	0.26	0.21	0.26	0.22	0.26	0.22
	0.25	0.35	0.26	0.33	0.25	0.33	0.25	0.33
Italy	0.26	0.20	0.26	0.20	0.27	0.21	0.27	0.22
	0.26	0.34	0.26	0.33	0.24	0.32	0.25	0.31
Short Sorting Window								
	h=1		h=2		h=4		h=8	
USA	0.27	0.25	0.27	0.24	0.27	0.25	0.27	0.25
	0.26	0.34	0.26	0.33	0.27	0.31	0.25	0.30
UK	0.27	0.25	0.27	0.24	0.27	0.24	0.27	0.24
	0.26	0.35	0.26	0.34	0.25	0.32	0.26	0.32
France	0.28	0.24	0.27	0.24	0.28	0.24	0.27	0.26
	0.26	0.35	0.26	0.33	0.26	0.31	0.26	0.30
Germany	0.27	0.23	0.26	0.24	0.27	0.23	0.27	0.24
	0.26	0.35	0.25	0.33	0.25	0.32	0.25	0.30
Japan	0.27	0.24	0.28	0.24	0.28	0.24	0.29	0.24
	0.26	0.35	0.25	0.34	0.25	0.32	0.24	0.31
Canada	0.26	0.25	0.26	0.24	0.26	0.25	0.27	0.25
	0.26	0.34	0.26	0.33	0.26	0.31	0.25	0.32
Italy	0.27	0.24	0.27	0.24	0.28	0.23	0.27	0.24
	0.26	0.35	0.26	0.33	0.25	0.33	0.26	0.31

Table 4: Significance of transition probabilities estimated for the non linear models. Each cell reports the percentage of corner probabilities in Table 3 that is greater than 0.25 at the 5% significance level.

Expanding Sorting Window								
	h=1		h=2		h=4		h=8	
USA	0.35	0.13	0.36	0.08	0.48	0.22	0.36	0.29
	0.43	0.88	0.40	0.88	0.40	0.75	0.36	0.66
UK	0.21	0.03	0.26	0.02	0.41	0.09	0.38	0.20
	0.36	0.93	0.36	0.91	0.48	0.77	0.42	0.76
France	0.22	0.02	0.18	0.04	0.30	0.11	0.22	0.22
	0.24	0.90	0.24	0.78	0.34	0.83	0.33	0.62
Germany	0.40	0.07	0.33	0.11	0.47	0.07	0.46	0.22
	0.26	0.93	0.32	0.84	0.25	0.80	0.24	0.74
Japan	0.25	0.05	0.30	0.05	0.30	0.07	0.39	0.25
	0.33	0.94	0.44	0.87	0.30	0.77	0.36	0.55
Canada	0.32	0.14	0.32	0.08	0.35	0.19	0.34	0.21
	0.32	0.93	0.29	0.83	0.30	0.79	0.38	0.68
Italy	0.22	0.06	0.34	0.09	0.43	0.20	0.30	0.30
	0.43	0.94	0.43	0.86	0.32	0.66	0.30	0.63
Rolling Sorting Window								
	h=1		h=2		h=4		h=8	
USA	0.31	0.01	0.31	0.11	0.37	0.12	0.40	0.26
	0.44	0.99	0.45	0.95	0.49	0.88	0.42	0.66
UK	0.36	0.02	0.34	0.00	0.46	0.05	0.48	0.12
	0.41	0.98	0.36	0.91	0.36	0.88	0.40	0.74
France	0.31	0.00	0.31	0.06	0.28	0.04	0.38	0.22
	0.29	1.00	0.31	0.88	0.34	0.83	0.47	0.64
Germany	0.37	0.00	0.32	0.07	0.44	0.07	0.57	0.11
	0.30	0.96	0.32	0.86	0.27	0.84	0.28	0.81
Japan	0.40	0.02	0.37	0.06	0.41	0.07	0.45	0.21
	0.25	0.97	0.35	0.89	0.36	0.82	0.41	0.73
Canada	0.25	0.07	0.34	0.07	0.26	0.07	0.46	0.14
	0.32	0.95	0.39	0.95	0.35	0.86	0.39	0.68
Italy	0.41	0.04	0.50	0.02	0.50	0.14	0.53	0.20
	0.41	0.98	0.36	0.98	0.34	0.86	0.28	0.85
Short Sorting Window								
	h=1		h=2		h=4		h=8	
USA	0.41	0.28	0.45	0.23	0.48	0.32	0.42	0.26
	0.41	0.92	0.37	0.95	0.44	0.86	0.36	0.67
UK	0.36	0.26	0.48	0.21	0.48	0.20	0.46	0.32
	0.45	0.95	0.28	0.93	0.27	0.84	0.38	0.80
France	0.55	0.10	0.43	0.16	0.47	0.23	0.40	0.36
	0.35	0.96	0.35	0.90	0.30	0.79	0.33	0.67
Germany	0.47	0.12	0.42	0.23	0.40	0.20	0.43	0.26
	0.30	0.95	0.30	0.93	0.35	0.95	0.22	0.74
Japan	0.32	0.27	0.37	0.29	0.44	0.20	0.59	0.16
	0.40	0.95	0.30	0.90	0.25	0.89	0.21	0.75
Canada	0.34	0.20	0.39	0.20	0.42	0.25	0.46	0.23
	0.34	0.93	0.34	0.95	0.40	0.75	0.21	0.79
Italy	0.41	0.22	0.45	0.16	0.52	0.20	0.35	0.30
	0.41	0.98	0.41	0.98	0.27	0.91	0.38	0.78

Table 5: Out-of-sample forecasting performance of combination schemes applied to linear models. Each panel reports the distribution of out-of-sample MSFE - relative to that of the previous best model using an expanding window - averaged across variables, countries and forecast horizons (1,095 forecasts) for different combination strategies. Standard combination strategies include PB, the previous best model, the average across the models in the top quartile $Q(PB)$, and across all models $Q(EW)$. Quartile and cluster-sorted conditional combination strategies are referred to as $Q(W, Z)$, and $C(K, W, Z)$ respectively, where $W \in \{EW, OW, PB, SW\}$ (equal weighted, optimally weighted, previous best, and shrinkage weighted), $Z \in \{L, M, H\}$ is the degree of shrinkage (low, medium, high), and K is the number of clusters.

	Min	**10%**	**25%**	**Median**	**75%**	**90%**	**Max**	**Mean**
Expanding Sorting Window								
PB	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$Q(PB)$	0.223	0.762	0.866	0.939	0.991	1.036	2.360	0.923
$Q(EW)$	0.204	0.755	0.860	0.940	0.996	1.060	2.694	0.928
$Q(OW)$	0.152	0.705	0.820	0.919	0.994	1.073	3.518	0.912
$Q(SW, L)$	0.251	0.701	0.821	0.917	0.985	1.056	2.820	0.904
$Q(SW, M)$	0.304	0.705	0.828	0.916	0.981	1.047	2.695	0.900
$Q(SW, H)$	0.328	0.709	0.834	0.916	0.979	1.040	2.688	0.901
$C(3, PB)$	0.145	0.762	0.870	0.944	0.993	1.038	1.951	0.924
$C(3, EW)$	0.204	0.755	0.860	0.940	0.996	1.060	2.694	0.928
$C(3, OW)$	0.271	0.808	0.934	1.045	1.240	1.552	13.748	1.187
$C(3, SW, L)$	0.292	0.780	0.897	1.000	1.120	1.313	9.226	1.058
$C(3, SW, M)$	0.203	0.765	0.872	0.969	1.051	1.171	5.966	0.980
$C(3, SW, H)$	0.211	0.760	0.866	0.953	1.019	1.097	3.818	0.945
$C(2, PB)$	0.173	0.756	0.862	0.941	0.994	1.046	2.490	0.925
$C(2, OW)$	0.198	0.741	0.864	0.976	1.110	1.389	8.287	1.057
$C(2, SW, L)$	0.237	0.734	0.855	0.958	1.070	1.264	6.009	1.006
$C(2, SW, M)$	0.270	0.731	0.851	0.947	1.037	1.182	4.228	0.968
$C(2, SW, H)$	0.213	0.737	0.847	0.938	1.020	1.129	2.943	0.943
Rolling Sorting Window								
PB	0.510	0.874	0.948	1.005	1.060	1.133	1.967	1.010
$Q(PB)$	0.163	0.754	0.863	0.940	0.995	1.049	2.408	0.924
$Q(EW)$	0.204	0.755	0.860	0.940	0.996	1.060	2.694	0.928
$Q(OW)$	0.152	0.697	0.822	0.919	0.992	1.079	3.541	0.911
$Q(SW, L)$	0.253	0.694	0.822	0.916	0.986	1.055	2.824	0.903
$Q(SW, M)$	0.304	0.697	0.827	0.914	0.983	1.047	2.776	0.900
$Q(SW, H)$	0.328	0.704	0.829	0.915	0.981	1.042	2.760	0.900
$C(3, PB)$	0.172	0.757	0.864	0.942	0.995	1.048	1.922	0.926
$C(3, EW)$	0.204	0.755	0.860	0.940	0.996	1.060	2.694	0.928
$C(3, OW)$	0.222	0.845	0.971	1.114	1.331	1.662	16.616	1.263
$C(3, SW, L)$	0.235	0.802	0.924	1.035	1.186	1.389	11.452	1.113
$C(3, SW, M)$	0.272	0.778	0.892	0.992	1.093	1.229	7.642	1.017
$C(3, SW, H)$	0.304	0.763	0.877	0.969	1.044	1.149	5.023	0.970
$C(2, PB)$	0.178	0.752	0.860	0.939	0.995	1.044	2.259	0.923
$C(2, OW)$	0.245	0.763	0.895	1.019	1.189	1.464	10.246	1.131
$C(2, SW, L)$	0.251	0.747	0.882	0.994	1.135	1.332	8.357	1.064
$C(2, SW, M)$	0.283	0.742	0.872	0.977	1.091	1.238	6.710	1.013
$C(2, SW, H)$	0.295	0.740	0.863	0.963	1.058	1.171	5.303	0.977
Short Sorting Window								
PB	0.291	0.825	0.924	1.016	1.098	1.207	3.034	1.023
$Q(PB)$	0.160	0.755	0.865	0.942	0.999	1.062	2.585	0.929
$Q(EW)$	0.204	0.755	0.860	0.940	0.996	1.060	2.694	0.928
$Q(OW)$	0.269	0.695	0.821	0.921	1.004	1.098	3.756	0.920
$Q(SW, L)$	0.280	0.697	0.819	0.920	0.997	1.078	3.014	0.910
$Q(SW, M)$	0.296	0.701	0.820	0.917	0.990	1.063	2.575	0.905
$Q(SW, H)$	0.281	0.708	0.827	0.917	0.985	1.054	2.590	0.904
$C(3, PB)$	0.175	0.749	0.855	0.938	0.998	1.063	2.449	0.924
$C(3, EW)$	0.204	0.755	0.860	0.940	0.996	1.060	2.694	0.928
$C(3, OW)$	0.313	0.868	1.005	1.172	1.464	2.022	15.199	1.404
$C(3, SW, L)$	0.330	0.819	0.948	1.068	1.253	1.570	9.402	1.190
$C(3, SW, M)$	0.275	0.783	0.902	1.008	1.127	1.317	5.183	1.055
$C(3, SW, H)$	0.282	0.768	0.882	0.977	1.064	1.194	3.074	0.989
$C(2, PB)$	0.189	0.749	0.856	0.934	0.992	1.058	2.625	0.923
$C(2, OW)$	0.254	0.797	0.921	1.055	1.262	1.671	9.782	1.223
$C(2, SW, L)$	0.258	0.783	0.900	1.021	1.180	1.479	7.376	1.127
$C(2, SW, M)$	0.275	0.766	0.881	0.989	1.116	1.336	5.425	1.053
$C(2, SW, H)$	0.321	0.755	0.865	0.968	1.079	1.233	3.929	1.002

Table 6: Out-of-sample forecasting performance of combination schemes applied to non linear models. Each panel reports the distribution of out-of-sample MSFE - relative to that of the previous best model using an expanding window - averaged across variables, countries and forecast horizons (1,095 forecasts) for different combination strategies. Standard combination strategies include PB, the previous best model, the average across the models in the top quartile $Q(PB)$, and across all models $Q(EW)$. Quartile and cluster-sorted conditional combination strategies are referred to as $Q(W, Z)$, and $C(K, W, Z)$ respectively, where $W \in \{EW, OW, PB, SW\}$ (equal weighted, optimally weighted, previous best, and shrinkage weighted), $Z \in \{L, M, H\}$ is the degree of shrinkage (low, medium, high), and K is the number of clusters.

	Min	**10%**	**25%**	**Median**	**75%**	**90%**	**Max**	**Mean**
Expanding Sorting Window								
PB	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$Q(PB)$	0.369	0.748	0.833	0.907	0.970	1.018	1.630	0.896
$Q(EW)$	0.355	0.708	0.814	0.906	0.988	1.065	1.651	0.900
$Q(OW)$	0.089	0.641	0.746	0.858	0.968	1.062	2.925	0.862
$Q(SW, L)$	0.153	0.647	0.747	0.855	0.960	1.036	2.259	0.853
$Q(SW, M)$	0.207	0.652	0.747	0.854	0.951	1.024	1.727	0.849
$Q(SW, H)$	0.232	0.661	0.755	0.857	0.949	1.016	1.555	0.850
$C(2, PB)$	0.356	0.709	0.813	0.893	0.963	1.018	1.339	0.881
$C(2, OW)$	0.154	0.679	0.807	0.938	1.101	1.438	7.130	1.026
$C(2, SW, L)$	0.164	0.671	0.798	0.928	1.069	1.339	5.659	0.990
$C(2, SW, M)$	0.167	0.672	0.798	0.916	1.041	1.257	4.423	0.962
$C(2, SW, H)$	0.183	0.676	0.797	0.913	1.026	1.194	3.789	0.944
Rolling Sorting Window								
PB	0.520	0.870	0.948	1.017	1.097	1.187	3.710	1.034
$Q(PB)$	0.381	0.749	0.830	0.912	0.982	1.041	1.825	0.904
$Q(EW)$	0.355	0.708	0.814	0.906	0.988	1.065	1.651	0.900
$Q(OW)$	0.089	0.639	0.752	0.869	0.977	1.073	2.894	0.867
$Q(SW, L)$	0.135	0.643	0.750	0.865	0.963	1.049	2.246	0.858
$Q(SW, M)$	0.143	0.649	0.752	0.860	0.957	1.033	1.808	0.853
$Q(SW, H)$	0.181	0.660	0.757	0.864	0.954	1.023	1.633	0.853
$C(2, PB)$	0.357	0.714	0.810	0.897	0.970	1.029	1.467	0.885
$C(2, OW)$	0.176	0.687	0.813	0.944	1.117	1.448	5.869	1.041
$C(2, SW, L)$	0.158	0.679	0.807	0.932	1.076	1.359	4.906	1.001
$C(2, SW, M)$	0.157	0.682	0.803	0.919	1.053	1.285	4.320	0.970
$C(2, SW, H)$	0.170	0.685	0.802	0.914	1.034	1.218	3.787	0.949
Short Sorting Window								
PB	0.392	0.850	0.960	1.091	1.242	1.461	3.869	1.130
$Q(PB)$	0.385	0.751	0.844	0.943	1.047	1.179	2.221	0.958
$Q(EW)$	0.355	0.708	0.814	0.906	0.988	1.065	1.651	0.900
$Q(OW)$	0.088	0.642	0.758	0.880	0.996	1.118	2.682	0.886
$Q(SW, L)$	0.144	0.648	0.759	0.874	0.983	1.086	2.245	0.876
$Q(SW, M)$	0.166	0.649	0.762	0.874	0.976	1.070	1.880	0.871
$Q(SW, H)$	0.227	0.659	0.768	0.878	0.970	1.055	1.621	0.871
$C(2, PB)$	0.371	0.722	0.819	0.915	0.991	1.073	1.765	0.907
$C(2, OW)$	0.158	0.684	0.813	0.939	1.093	1.373	4.455	1.003
$C(2, SW, L)$	0.150	0.682	0.803	0.926	1.059	1.282	3.947	0.968
$C(2, SW, M)$	0.157	0.678	0.799	0.916	1.037	1.213	3.483	0.944
$C(2, SW, H)$	0.180	0.677	0.801	0.911	1.022	1.170	3.060	0.930

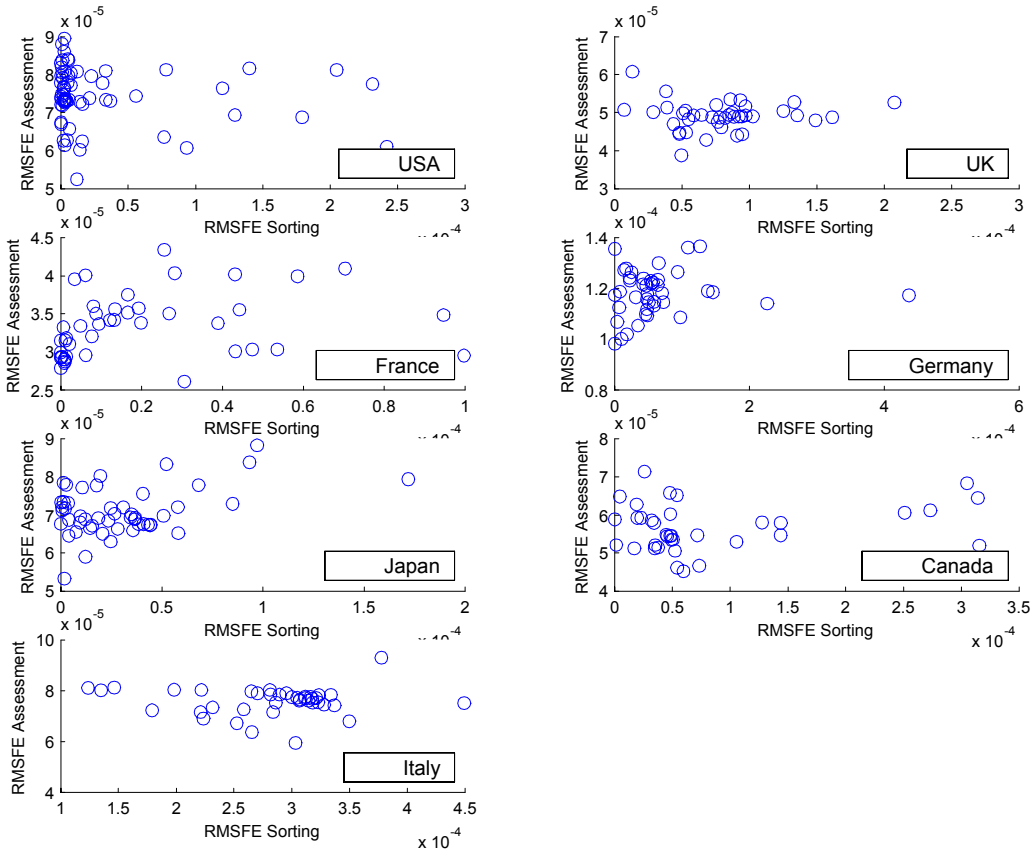

Fig. 1. Scatter plot of the average in-sample versus out-of-sample MSFE values generated by linear forecasting models estimated for output growth.

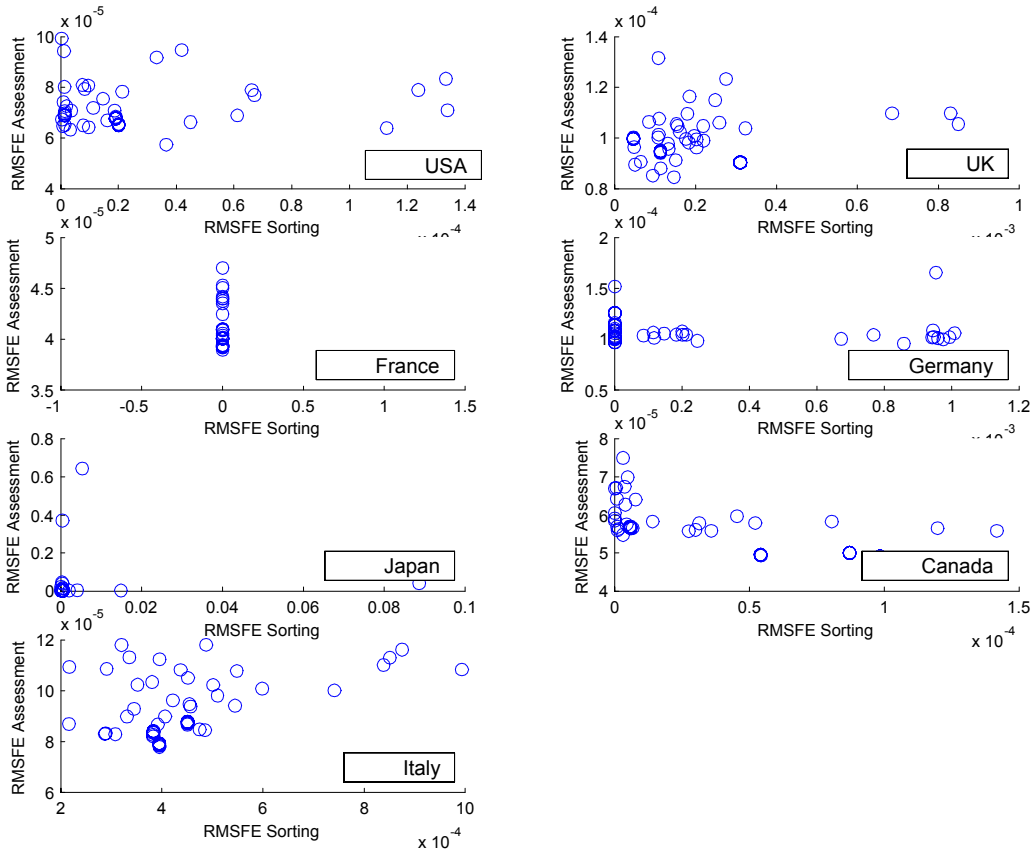

Fig. 2. Scatter plot of the average in-sample versus out-of-sample MSFE values generated by non-linear forecasting models estimated for output growth.