International Journal of Forecasting 32 (2016) 458–474

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Finite sample weighting of recursive forecast errors

Chris Brooks [a,*], Simon P. Burke [a], Silvia Stanescu [b]

[a] University of Reading, UK

[b] University of Kent, UK

ARTICLE INFO

Keywords:
Forecast evaluation
Forecast comparison
Recursive model estimation
Mean squared error
Forecast weighting scheme

ABSTRACT

This paper proposes and tests a new framework for weighting recursive out-of-sample prediction errors according to their corresponding levels of in-sample estimation uncertainty. In essence, we show how to use the maximum possible amount of information from the sample in the evaluation of the prediction accuracy, by commencing the forecasts at the earliest opportunity and weighting the prediction errors. Via a Monte Carlo study, we demonstrate that the proposed framework selects the correct model from a set of candidate models considerably more often than the existing standard approach when only a small sample is available. We also show that the proposed weighting approaches result in tests of equal predictive accuracy that have much better sizes than the standard approach. An application to an exchange rate dataset highlights relevant differences in the results of tests of predictive accuracy based on the standard approach versus the framework proposed in this paper.

1. Introduction

The issue of forecast evaluation is key for many assessments of model adequacy, and has received a considerable amount of attention as a consequence. As Diebold (2013) notes, in practice, forecast evaluation is rarely the object of interest in its own right; instead, it is more often conducted as a way of evaluating the relative accuracy levels of competing models. In such circumstances, one of two approaches may be adopted: either focussing entirely on in-sample model estimation over all observations available, or splitting the data into an in-sample estimation part and a separate out-of-sample forecast portion, with the evaluation then taking place based entirely on the latter.

In the out-of-sample forecasting literature, three alternative frameworks have been employed for conducting performance tests: constant coefficients, rolling windows of a constant size, and recursive forecasting. Sometimes, the choice between these approaches is motivated by the particular forecasting application being considered, but more often it is entirely arbitrary. Examples of studies using the different approaches include those by Ashley, Granger, and Schmalensee (1980) (constant coefficients), Cheung, Chinn, and Pascual (2003) (rolling), and Faust, Rogers, and Wright (2004) (recursive). West (2006) argues that the constant coefficient approach is preferred in cases where it is impossible to take the (re-)estimation process into account, while the rolling window alternative may be preferable when the series includes structural breaks or regime shifts. One could argue intuitively that recursive forecasting would be preferable in many situations where the sample size is small, since it makes use of all of the information available to the forecaster at that point in time. Faust et al. (2004) compare the mean squared errors (MSEs) obtained using a constant coefficients forecasting scheme with those resulting from a recursive framework, and find that the recursive one almost always produces lower MSEs, with the differences between the

* Corresponding author.
E-mail address: C.Brooks@icmacentre.rdg.ac.uk (C. Brooks).

http://dx.doi.org/10.1016/j.ijforecast.2015.05.003

two forecasting methodologies being statistically significant in some of their samples.

In applications involving point predictions, forecasters often consider a set of candidate models with the aim of finding the one that minimises the value of a previously determined loss function, conditional on the information available at the time when the forecast is made. Such prediction comparison studies have been termed "forecasting horse races". Almost without exception, the key innovations in this literature over the past two decades have concerned either the models employed, which have become more sophisticated over time,[1] or the loss functions adopted, which have increasingly tended towards economically-relevant measures. The forecast evaluation framework itself has scarcely warranted a mention, and the vast majority of studies still base their evaluations on an ad hoc rule of thumb whereby a fixed proportion of the dataset is used for in-sample model estimation and the remainder is retained for the out-of-sample evaluation where the predictions are compared with actual values.

Given a fixed total quantity of data, it would be possible to have a short in-sample period with a long evaluation period, two samples of roughly equal length, or a long in-sample period and only a short hold-out sample. Clearly, there is a trade-off involved here. If the in-sample estimation period is too short, parameter uncertainty will be high, leading to forecast imprecision; on the other hand, if the out-of-sample estimation period is too short, even if the models are estimated reasonably accurately, there will be too few forecasts to be compared with the actual values, and hence, the forecast evaluation metrics (e.g., the out-of-sample MSE) will be noisy and unreliable.[2] There are examples of studies that use low, medium and high proportions of the data out-of-sample, and in the vast majority of cases it is not at all clear how this choice has been made: it appears to be subjective, capricious, and tilted towards a long estimation period with a consequent short evaluation part.[3] West (1996) shows that, asymptotically, parameter estimation will not affect the outcomes of tests of the equivalence of mean squared errors from non-nested models with conditionally homoscedastic errors, which leads intuitively to a preference for the use of relatively long in-sample estimation periods. When the in-sample period is relatively large (say, 90% of the available data or more), the impact of parameter estimation error can probably be ignored (West, 2006); however, it is debatable whether this asymptotic irrelevance will still hold in the context of nested models or when the number of observations is small.[4]

An interesting puzzle in the empirical economic forecasting literature is the fact that numerous studies have shown a particular variable or set of variables to possess in-sample predictability which cannot be maintained in out-of-sample tests. The conventional explanation has been that the in-sample forecasting ability was illusory and a consequence of data mining. An alternative explanation, which Inoue and Kilian (2005) support through asymptotic theory, is simply that, in many instances, out-of-sample tests lack sufficient power to detect the predictability that is actually present in the data – an inevitable consequence of the loss of data because of the splitting of the sample. As a result, they advocate the sole use of in-sample t- or F-tests for comparing nested models (with no out-of-sample analysis).[5]

When the model-building and forecasting exercise occurs in the context of small samples of data with non-standard features or in the presence of model misspecification, however, the asymptotical results concerning the desirability of in-sample testing may not apply any more. Inoue and Kilian (2006) note two circumstances in which out-of-sample testing may be favoured over in-sample model selection based on the SIC: when comparing non-nested models in the context of autocorrelated data, and when comparing nested models when the true model is the larger one. In the latter case, the smaller (incorrect) model will suffer less from parameter estimation error, a reduction in variance that will more than compensate for the additional forecast error bias arising from the use of the wrong model. Thus, in-sample analysis may be preferable in general, but this conclusion could vary considerably depending on the context and the precise nature of the data. A further reason to prefer out-of-sample testing to a pure in-sample evaluation is that a researcher might be interested in how a model performs for prediction at a particular point in time, rather than on average.[6]

While it may be difficult to generalise, it is common for around two thirds of the sample or more to be used for initial in-sample model estimation, leaving the

[1] As Diebold (2013) notes, it is usually the new horse in the stable that wins the race, perhaps not surprisingly.

[2] Ashley (2003) shows that demonstrating one forecasting model to be statistically significantly better than another would typically require more than 100 out-of-sample observations. However, fewer out-of-sample data points than this are commonly available when quarterly or even monthly data are employed with in-sample estimation windows of a conventional length.

[3] For example, West's (2006, p. 106) review paper explicitly takes the split point "as given", with no further discussion.

[4] A recent study that addresses the issue of the position of the split into in-sample and out-of-sample periods directly is that of Hansen and Timmerman (2012). Building on earlier work by Clark and McCracken (2001, 2005a) and McCracken (2007), Hansen and Timmerman develop an approach that modifies the p-values of tests of the null hypothesis of no predictability so that they become robust to sample-split-induced data-mining. A conceptually similar modification, albeit different in detail, is proposed by Rossi and Inoue (2012), and further comparisons of the powers of tests for the differences between out-of-sample forecasts are provided by Busetti and Marcucci (2013).

[5] On a related note, Clark and McCracken (2005b) demonstrate that an in-sample F-test of predictive ability is likely to be more powerful than out-of-sample tests. Schwarz's information criterion (SIC) could be used instead to penalise additional parameters within the model evaluated in-sample, and will deliver the correct model asymptotically with probability one when it is in the choice set. Moreover, the SIC will still select the best approximating model consistently even when all of the models in the choice set are misspecified (see Inoue & Kilian, 2006). While the penalty term in the SIC may help to weed out spurious predictability arising from data-mining in such contexts, as Diebold (2013) notes, it will not help if the researcher wishes to compare two (non-nested) models that contain the same numbers of parameters, one of which has arisen as a result of data-mining.

[6] Giacomini and Rossi (2010) develop a method for testing precisely this 'local' forecasting power.