International Journal of Forecasting 29 (2013) 13–27

Contents lists available at SciVerse ScienceDirect

International Journal of Forecasting

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

Comparing forecast accuracy: A Monte Carlo investigation

Fabio Busetti, Juri Marcucci *

Bank of Italy, Research Department, Italy

ARTICLE INFO

Keywords:
Forecast encompassing
Model evaluation
Nested models
Equal predictive ability
Forecast evaluation

ABSTRACT

The size and power properties of several tests of equal Mean Square Prediction Errors (MSPE) and of Forecast Encompassing (FE) are evaluated, using Monte Carlo simulations, in the context of nested dynamic regression models. The highest size-adjusted power is achieved by the F-type test of forecast encompassing proposed by Clark and McCracken (2001); however, the test tends to be slightly oversized when the number of out-of sample observations is 'small' and in cases of (partial) misspecification. The relative performances of the various tests remain broadly unaltered for one- and multi-step-ahead predictions and when the predictive models are partially misspecified. Interestingly, the presence of highly persistent regressors leads to a loss of power of the tests, but their size properties remain nearly unaffected. An empirical example compares the performances of models for short term predictions of Italian GDP.

1. Introduction

Evaluating the out-of-sample performances of competing models is an important aspect of economic forecasting and model selection. Diebold and Mariano (1995) have proposed a simple test for the null hypothesis of equal predictive accuracy *in population*, measured in terms of a general loss function. However, in most applications, little attention is paid to the shape of the loss function, and models are generally compared on the basis of their mean square prediction errors (MSPE). An alternative approach looks at the out-of-sample correlation between prediction errors, which leads to tests of forecast encompassing (FE). A preferred forecast is said to encompass some competing alternative if the latter contains no additional useful information for prediction; see, inter alia, Chong and Hendry (1986), Clements and Hendry (1993), Granger and Newbold (1986), and Harvey, Leybourne, and Newbold (1998).

The recent literature on out-of-sample prediction has highlighted two important issues that may render invalid the standard large sample inference *à la* Diebold and Mariano (1995). First, West (1996) has showed that parameter estimation error may not be asymptotically irrelevant, and therefore may affect the limiting distribution of the test statistics. Second, if models are nested, the statistics based on average comparisons of prediction errors have a degenerate limiting variance under the null hypothesis, and are not asymptotically normally distributed. For nested models, McCracken (2007) and Clark and McCracken (2001) derive the appropriate non-Gaussian limit for tests of equal MSPE and FE, respectively; the critical values are tabulated across two nuisance parameters (the ratio of the magnitude of the prediction sample to that of the estimation sample and the number of additional regressors in the larger model), and are, in general, only valid for one-step-ahead predictions. The test of forecast encompassing for nested models proposed by Chao, Corradi, and Swanson (2001) does not suffer from this degeneracy: its limiting distribution is a chi-square under the null hypothesis. Giacomini and White (2006) take a different approach, focusing on comparing forecasting methods, as opposed to forecasting models; their test statistic of equal conditional predictive ability has a chi-square null distribution, as the prediction sample size tends to infinity for a finite length of the estimation sample. Comprehensive surveys of the evaluation of predictive ability for nested and non-nested

* Corresponding author.
E-mail addresses: fabio.busetti@bancaditalia.it (F. Busetti), juri@sssup.it, juri.marcucci@bancaditalia.it (J. Marcucci).

0169-2070/$ – see front matter
doi:10.1016/j.ijforecast.2012.04.011

models include those of Clark and McCracken (2011) and West (2006).

In this paper we evaluate the finite sample properties of several tests of equal MSPE and tests of FE, with the aim of providing practical guidance for forecasters who need to choose among a set of predictions from (a small number of) competing models.[1] We focus on nested model comparisons, for which several modifications of the standard MSPE and FE tests have been suggested.[2] Monte Carlo simulation methods are used to compute the empirical size and empirical power functions in the context of dynamic regression models. One- and multi-step-ahead predictions are considered for both correctly specified and misspecified regressions. The properties of the tests across different values of the ratio between prediction and estimation sample sizes and for various degrees of persistence of the data generating process are also investigated.

The tests under scrutiny are the following: (i) the standard Diebold–Mariano test of equal MSPE; (ii) the *MSE-t* and (iii) *MSE-F* modifications of McCracken (2007) for nested models; (iv) the forecast encompassing test of Harvey et al. (1998); (v) the *ENC-t* and (vi) *ENC-F* modifications of Clark and McCracken (2001) for nested models; and (vii) the forecast encompassing test of Chao et al. (2001) for nested models.[3]

Our results extend previous analyses (which have mostly been concerned with the size properties of the tests) by providing empirical power functions in a variety of settings, including misspecification of the regression models and high persistence in the data generating process. We confirm the findings of Clark and McCracken (2001, 2005a) that the *ENC-F* test achieves the highest (size-adjusted) power, noting however that it tends to be somewhat oversized when the prediction sample is short, and for cases of model misspecification. In fact, the relative ranking among the different tests changes based on whether the number of out-of-sample observations is "small" or "large". Interestingly, the presence of highly persistent regressors leads to a loss of power, but the size of the tests is broadly unaffected.

In summary, the paper proceeds as follows. Section 2 briefly reviews the test statistics under scrutiny. Sections 3 and 4 contain the simulation results for one-step-ahead and multi-step-ahead forecasts, respectively. The size and power properties of the tests under different degrees of persistence of the predictors are evaluated in Section 5. A short empirical application to short term predictions of Italian GDP is presented in Section 6, and Section 7 concludes.

[1] For issues arising when comparing a large number of models, see Hansen (2005) and White (2000), while for issues arising when comparing a small number of models, see Hubrich and West (2010).

[2] Results for non-nested models are contained in an earlier draft of this paper (Busetti, Marcucci, & Veronese, 2009).

[3] We do not include the method of Giacomini and White (2006) in the comparison, because it relates to a different null hypothesis from the other tests. We also do not consider the test of Corradi and Swanson (2002), which is consistent against generic nonlinear alternatives, because we adopt a linear setup.

2. The setup and the tests under scrutiny

We consider a sample of T observations of a target series y_t, and two k_i-dimensional vectors of (non-mutually exclusive) predictors X_{it}, $i = 1, 2$. The sample is divided into R in-sample and P out-of-sample observations, with $T = R + P$.

We want to compare two sets of h-step-ahead forecasts, $h \geq 1$, generated by the linear models

$$\widehat{y}_{it} = X'_{i,t-h}\widehat{\beta}_{i,t-h}, \quad t = R + h, R + h + 1, \ldots, T, \tag{1}$$

where $\widehat{\beta}_{i,t-h}$ is the least squares estimate for model i constructed using observations up to time $t - h$, and the predictors $X_{i,t-h}$ may include lags of the dependent variable y_{t-j} for $j \geq h$. The models are estimated under either the recursive or the rolling scheme: the recursive least squares estimates are constructed using observations indexed from 1 to $t - h$, while the rolling coefficients are estimated using the R observations indexed from $t - R - h + 1$ to $t - h$.

The forecasting performance of the models is evaluated using the two sets of h-step-ahead forecast errors $e_{it} = y_t - \widehat{y}_{it}$, $i = 1, 2$, for $t = R + h, R + h + 1, \ldots, R + P$; for the sake of simplicity, we suppress the dependency on h in the notation. The tests under scrutiny are detailed briefly below. Table 1 provides a concise summary of the sources of the tests and of the corresponding critical values.

2.1. Tests of equal MSPE

The test of equal mean squared prediction error of Diebold and Mariano (1995) is based on the following t-type statistic

$$DM = \widehat{P}^{\frac{1}{2}}\overline{d}/\widehat{\sigma}_{DM}(m), \tag{2}$$

where $\overline{d} = \widehat{P}^{-1}\sum_{t=R+h}^{T} d_t$, $d_t = e_{1t}^2 - e_{2t}^2$, $\widehat{P} = P - h + 1$, and $\widehat{\sigma}^2_{DM}(m)$ is the non-parametric estimator of the long run variance of d_t:

$$\widehat{\sigma}^2_{DM}(m) = \widehat{P}^{-1}\sum_{t=R+h}^{T}\left(d_t - \overline{d}\right)^2 + 2\widehat{P}^{-1}\sum_{j=1}^{m} w(j, m) \times \sum_{t=j+R+h}^{T}\left(d_t - \overline{d}\right)\left(d_{t-j} - \overline{d}\right), \tag{3}$$

where $w(j, m)$ is a weight function truncated at $m \ll T$; e.g., $w(j, m) = 1 - j/(m+1)$, as in Newey and West (1987); note that, in large samples, P can replace $\widehat{P}$ in the definition of Eq. (2). The *DM* statistic tests the null hypothesis of equal forecast accuracy $H_0 : E\, d_t^* = 0$, where d_t^*, is the population version of d_t, i.e., excluding parameter estimation error. If the models are *non-nested*, the limiting null distribution of Eq. (2) is a standard Gaussian. By contrast, if the models are *nested*, the denominator converges to zero under the null, and the limiting distribution of the *DM* statistic is non-Gaussian.[4]

[4] However, it is argued that the Gaussian critical values would still hold approximately if P/R is small (e.g., less than 0.1, see West, 2006); mathematically, the limiting distribution is Gaussian if $P/R \to 0$.