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 $P/R \to 0$.

Table 1
Tests under scrutiny: sources and critical values.

Test	H_0	Critical values		Source
		one-step	multi-step	
DM	Equal MSPE	$N(0, 1)$	$N(0, 1)$	Diebold and Mariano (1995)
MSE-t	Equal MSPE	cf. source	Bootstrap	McCracken (2007)
MSE-F	Equal MSPE	cf. source	Bootstrap	McCracken (2007)
HLN	FE	$N(0, 1)$	$N(0, 1)$	Harvey et al. (1998)
ENC-t	FE	cf. source	Bootstrap	Clark and McCracken (2001)
ENC-F	FE	cf. source	Bootstrap	Clark and McCracken (2001)
CCS	FE	χ^2	χ^2	Chao et al. (2001)

McCracken (2007) obtains the correct null limiting distribution of the *DM* statistic for the case of one-step-ahead forecasts between *nested* models; the test based on McCracken's critical values will be called *MSE-t*. The following *F*-type statistic is also proposed:

$$MSE\text{-}F = \widehat{P}\overline{d}/\widehat{\sigma}_2^2, \tag{4}$$

where $\widehat{\sigma}_2^2 = \widehat{P}^{-1}\sum_{t=R+h}^{T} e_{2t}^2$ is the estimate of the second moment of the forecast errors of the nesting model. The distributions of *MSE-t* and *MSE-F* depend on the ratio $\pi = P/R$ and on the number $k_2 - k_1$ of excess parameters in the nesting model; critical values are tabulated for recursive and rolling one-step-ahead forecasts. The limiting distributions change for the case of multi-step-ahead predictions, but critical values can be obtained by bootstrapping; see Clark and McCracken (2005a).

For the case of *nested* models, the standard *DM* test turns out to be heavily undersized and to have low power. Although the correct limiting distribution is non-Gaussian, Clark and West (2006, 2007) argue that most of the bias can be corrected by a simple adjustment in the statistic; this leads to a test with Gaussian critical values that has a size close to, but a little less than, the nominal one.[5] Specifically, the Clark–West adjusted statistic is

$$DM_{AD} = \widehat{P}^{\frac{1}{2}}\overline{d}_{AD}/\widehat{\sigma}_{AD}(m), \tag{5}$$

where $\overline{d}_{AD} = \widehat{P}^{-1}\sum_{t=R+h}^{T} d_{AD,t}$, $d_{AD,t} = e_{1t}^2 - e_{2t}^2 + (\widehat{y}_{1t} - \widehat{y}_{2t})^2$, and $\widehat{\sigma}_{AD}^2(m)$ is the non-parametric estimator of the long run variance of $d_{AD,t}$, which parallels the definition in Eq. (3). Since $(\widehat{y}_{1t} - \widehat{y}_{2t})^2 = (e_{2t} - e_{1t})^2$, one can write $d_{AD,t} = 2e_{1t}(e_{1t} - e_{2t})$. Thus, as was noted by West (2006), the DM_{AD} statistic is based on the covariance between e_{1t} and $e_{1t} - e_{2t}$, and corresponds to the test of forecast encompassing given in Eq. (6).

2.2. Tests of forecast encompassing

It is said that the forecast $\widehat{y}_{1t}$ encompasses $\widehat{y}_{2t}$ if there is no gain from combining them into a composite forecast $\widehat{y}_{ct} = (1-\lambda)\widehat{y}_{1t} + \lambda\widehat{y}_{2t}$, for some weight $\lambda \neq 0$; see, inter alia, Chong and Hendry (1986), Clements and Hendry (1993), and Granger and Newbold (1986), and the early empirical work of Nelson (1972). As the combined forecast error e_{ct} satisfies the relationship $e_{1t} = \lambda(e_{1t} - e_{2t}) + e_{ct}$, Ericsson (1992) tests the null hypothesis of forecast encompassing, $H_0 : \lambda = 0$, by a t-test on λ in the regression of e_{1t} on $e_{1t} - e_{2t}$. In a similar way, Harvey et al. (1998) write the null hypothesis of forecast encompassing as $H_0 : Ef_t^* = 0$, where f_t^* is the population version of $f_t = e_{1t}(e_{1t} - e_{2t})$, and they construct a t-test on $\overline{f} = \widehat{P}^{-1}\sum_{t=R+h}^{T} f_t$; specifically, their statistic is

$$HLN = \widehat{P}^{\frac{1}{2}}\overline{f}/\widehat{\sigma}_{HLN}(m), \tag{6}$$

where $\widehat{\sigma}_{HLN}^2(m)$ is a non-parametric estimator of the long run variance of f_t, which parallels the definition in Eq. (3). If the models are *non-nested*, the limiting null distribution of the *HLN* statistic is a standard Gaussian.

Clark and McCracken (2001) show that the *HLN* statistic is no longer asymptotically Gaussian when applied to nested models, and obtain the correct null limiting distribution for one-step-ahead forecasts; the test that uses their critical values will be called *ENC-t*. They also propose the *F*-type statistic

$$ENC\text{-}F = \widehat{P}\overline{f}/\widehat{\sigma}_2^2, \tag{7}$$

where $\widehat{\sigma}_2^2$ is the mean squared forecast error of the nesting model, as in Eq. (4). The distributions of *ENC-t* and *ENC-F* depend on both the ratio P/R and the number $k_2 - k_1$ of excess parameters in the nesting model; critical values are tabulated for recursive and rolling one-step-ahead forecasts. The extension to multi-step-ahead forecasts is given by Clark and McCracken (2005a).

A different test of forecast encompassing for nested models has been proposed by Chao et al. (2001): the null hypothesis is $H_0 : Ec_t^* = 0$, where c_t^* is the population version of $c_t = e_{1t}\left(Z_{2t} - \overline{Z}_2\right)$, and Z_{2t} are the additional $k_2 - k_1$ predictors in X_{2t} which are not included in X_{1t}.[6] Again, this is a Wald-type test, with the statistic being given by

$$CCS = \widehat{P}\overline{c}\,'\left(\widehat{\Sigma}_{CCS}(m)\right)^{-1}\overline{c}, \tag{8}$$

where $\overline{c} = \widehat{P}^{-1}\sum_{t=R+h}^{T} c_t$, and $\widehat{\Sigma}_{CCS}(m)$ is a non-parametric estimator of the long run variance–covariance matrix of c_t, which parallels the definition in Eq. (3). Under the null hypothesis of forecast encompassing, *CCS* is asymptotically distributed as chi squared with $k_2 - k_1$ degrees of freedom.[7]

[5] The limiting distribution of the Clark–West statistic is exactly Gaussian if the null model is a martingale difference sequence and the rolling scheme is used.

[6] While the regressors Z_{2t} are not demeaned in the expression for c_t in the original formulation of Chao et al. (2001), we find slightly better properties of the test after demeaning (in real time).

[7] Chao et al. (2001) also propose a version of the test that takes into account estimation uncertainty, with $\widehat{\Sigma}_{CCS}(m)$ replaced by a

3. Properties of the tests: one-step-ahead forecasts

The properties of the tests are evaluated by means of Monte Carlo simulations in the context of nested linear dynamic regression models. We start by considering the following VAR(1) data generating process (for $t = 1, 2, \ldots, T$)

$$y_t = \mu_y + \phi_y y_{t-1} + c x_{t-1} + \varepsilon_t, \tag{9}$$

$$x_t = \mu_x + \phi_x x_{t-1} + u_t, \tag{10}$$

with Gaussian i.i.d. innovations

$$\begin{pmatrix} \varepsilon_t \\ u_t \end{pmatrix} \sim NIID\left(0, \begin{pmatrix} 1 & \rho_{\varepsilon u} q \\ \rho_{\varepsilon u} q & q^2 \end{pmatrix}\right). \tag{11}$$

Note that, if $c \neq 0$, y_t can be represented as a Gaussian ARMA(2, 1) process with the degree of persistence, as measured by the sum of the autoregressive roots, equal to $\phi_x + \phi_y - \phi_x \phi_y$. If $c = 0$, then y_t is not Granger-caused by x_t.

The object is to forecast y_t using a dynamic univariate regression. We compare two sets of out-of-sample forecasts: the first one is obtained by an autoregression of order 1 (the restricted model), the other by including additional predictors (the unrestricted or nesting model). The case $c = 0$ measures the size of the tests of equal MSPE and FE, while $c \neq 0$ provides the power. All tests are one-sided, in the sense that the alternative hypothesis is that the nesting model yields better forecasts.[8] Given that the null hypothesis is $c = 0$, the tests can also be interpreted as out-of-sample tests of Granger causality.

We provide results for sample sizes of $T = R + P$, where $R = (100, 200)$ and $P = \pi R$, with $\pi = (0.1, 0.25, 0.5, 1)$. The properties of the tests clearly depend on the number of out-of-sample observations P, with the power expected to increase with π (for a given value of R). Since a constant term will always be included in the set of predictors, without loss of generality we set $\mu_y = \mu_x = 0$ in the data generating process (Eqs. (9)–(11)).

In this section, the size and power properties of the tests are evaluated for one-step-ahead forecasts, using 50,000 Monte Carlo replications. The first subsection contains the results for correctly specified models under Gaussianity (i.e., the estimated unrestricted model is the same as the true data generating process); the effects of misspecification and overparameterization form the subject of the second subsection; and fat-tailed distributions are considered in the third subsection.

3.1. The nesting model is correctly specified

The restricted model is the regression of y_t on $X_{1t} = (1, y_{t-1})'$; in the unrestricted or nesting model, the predictors are given by $X_{2t} = (1, y_{t-1}, x_{t-1})'$. Since there is no additional temporal dependence to be taken into account, we calculate the statistics in Eqs. (2), (5), (6) and (8) for $m = 0$, i.e., with scaling provided by the sample variance instead of the long-run variance.[9]

Table 2 provides the empirical sizes of the tests ($c = 0$) run at the 5% and 10% levels of significance for $R = (100, 200)$ for the case of recursive forecasts.[10] We present figures where the values of the parameters in the data generating process are set to $\phi_y = \phi_x = 0.8$, $q = 1$, and $\rho_{\varepsilon u} = 0$.

Consider first the case of $R = 100$, with tests run at the 10% level of significance. For $\pi = 0.1$ (10 out-of-sample observations), all tests except *DM* are oversized, particularly *MSE-t* (0.17), *ENC-t* (0.16) and *CCS* (0.15). As π increases, the size improves for all tests except *DM* and *HLN*; however, while *DM* is deeply undersized for $\pi \geq 0.5$, the rejection frequencies for *HLN* do not fall below 7%, consistent with the arguments of Clark and West (2006, 2007) for the (equivalent) adjusted *DM* statistic in Eq. (5). Doubling the sample ($R = 200$) yields more reliable sizes for all tests, except, to some extent, *DM* and *HLN*. Qualitatively similar arguments apply for the tests run at the 5% level.

Fig. 1 shows the empirical power functions (with respect to the parameter c governing the distance from the null hypothesis) of tests run at the 10% significance level for $R = 200$; the results for $R = 100$ and for tests run at the 5% significance level are qualitatively similar, and therefore will be not discussed. The power is affected by the parameter q that controls the variance of x_t (for a given value of c, the higher q is, the more powerful the tests are), as well as, to a lesser extent, by the value of the correlation $\rho_{\varepsilon u}$; however, as the relative ranking among the tests turns out not to be affected by the values of q and $\rho_{\varepsilon u}$, to save space here, we only present the results for $q = 1$ and $\rho_{\varepsilon u} = 0$. The four panels of Fig. 1 refer to different magnitudes of the prediction sample, $\pi = (0.1, 0.25, 0.5, 1)$; clearly, for a fixed R, the larger the value of π, the greater the power.

For all values of π, the *ENC-F* test of Clark and McCracken (2001) presents the highest rejection frequencies; although it appears to be slightly oversized, the computation of the size-adjusted power shown in Fig. 2 confirms that the *ENC-F* is indeed the most powerful test.[11] The second ranked test depends on the length of the prediction sample relative to the estimation sample. If the prediction sample is short ($\pi = 0.1$), then the *MSE-F* test is preferable, otherwise the *ENC-t* is better. For large values of π, the *HLN* test, which uses Gaussian critical values, behaves similarly to *MSE-F*, and is considerably more powerful than *MSE-t*. The *DM* test has the lowest power, while the *CCS* test (which uses χ^2 critical values) only has a relatively good power for large values of π, and is always dominated by *HLN*.

more complicated expression which depends on the sampling scheme. However, they also argue that the modified test does not provide a clear advantage in terms of size, and it actually turns out to be less powerful.

[8] For the *HLN* test, it can be shown that if x_{t-1} has predictive power for y_t, then the covariance between e_{1t} and $e_{1t} - e_{2t}$ is positive. Thus, the test is one-sided in the right tail; see Clark and McCracken (2005a, p. 376).

[9] However, note that if the data were characterized by time variation in the second moments (GARCH effects), then a long-run variance correction ($m > 0$) should be used as well, in the context of one-step-ahead predictions.

[10] The results for rolling forecasts are very similar, and therefore are not presented.

[11] Clearly, in terms of size-adjusted power, *DM* and *HLN* are the same as *ENC-t* and *ENC-F*, respectively.

Table 2
Empirical size of the tests of equal forecast accuracy for one-step-ahead forecasts run at the nominal 5% and 10% levels.

π	$R = 100$				$R = 200$			
	0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00
(A) Recursive 5%								
DM	0.06	0.03	0.02	0.01	0.04	0.02	0.01	0.01
MSE-t	0.11	0.08	0.07	0.06	0.08	0.07	0.06	0.06
MSE-F	0.07	0.07	0.06	0.06	0.06	0.06	0.05	0.05
HLN	0.07	0.05	0.04	0.03	0.05	0.04	0.04	0.03
ENC-t	0.10	0.08	0.07	0.06	0.08	0.07	0.06	0.06
ENC-F	0.08	0.07	0.07	0.06	0.07	0.06	0.06	0.06
CCS	0.09	0.07	0.07	0.06	0.07	0.06	0.06	0.06
(B) Recursive 10%								
DM	0.10	0.06	0.04	0.02	0.08	0.05	0.03	0.02
MSE-t	0.17	0.13	0.12	0.11	0.14	0.12	0.11	0.11
MSE-F	0.12	0.12	0.11	0.10	0.11	0.11	0.10	0.10
HLN	0.12	0.09	0.08	0.07	0.10	0.08	0.07	0.06
ENC-t	0.16	0.14	0.12	0.11	0.14	0.13	0.11	0.11
ENC-F	0.13	0.12	0.12	0.11	0.12	0.11	0.11	0.11
CCS	0.15	0.13	0.12	0.12	0.13	0.11	0.11	0.11

Notes: Results from 50,000 Monte Carlo iterations. One-step-ahead forecasts with the recursive scheme. In-sample sizes: $R = (100, 200)$.

Larger differences in the behavior of the tests occur when the number of out-of-sample observations is small. In particular, when $\pi = 0.1$, the better sized tests are *DM*, *HLN* and *MSE-F*, but only the last has high rejection rates under the alternative hypothesis (being second only to *ENC-F*). For higher values of π, the tests tend to behave more similarly; while *ENC-F* clearly dominates, the *HLN* test may become attractive, being based on Gaussian critical values.

3.2. Misspecification and overparameterization

We now compare the properties of the tests of equal MSPE and FE for models that are misspecified[12] or overparameterized. We consider the following cases.

(1) *Error-in-variables*. In the unrestricted model, we take $(1, y_{t-1}, w_{t-1})'$ as predictors instead of $(1, y_{t-1}, x_{t-1})'$, where

$$w_t = x_t + u_{w,t}, \quad u_{w,t} \sim NIID\left(0, q_w^2 \sigma_x^2\right), \tag{12}$$

so that w_t and x_t are positively correlated with coefficient $\rho_{xw} = 1/\left(1 + q_w^2\right)$. Fig. 3 reports the empirical power functions for a correlation parameter $\rho_{xw} = 0.5$ (i.e., $q_w = 1$) and $R = 200$. All tests undergo some reduction of power with respect to the case of correct specification,[13] but, interestingly, the relative ranking among them remains the same. Clearly, higher values of ρ_{xw} (i.e., lower values of q_w) generate smaller losses of power from misspecification.[14]

(2) *Autoregression*. We take as our unrestricted model an autoregression of order p, where $2 \leq p \leq 8$ is chosen according to the *BIC* method. As the true data generating process is an *ARMA*(2, 1), an autoregressive model should provide a reasonable approximation. The empirical power functions of the tests are depicted in Fig. 4, for the case of $R = 200$. Here, the loss of power due to misspecification turns out to be very relevant; for example, if $c = 0.50$ and $\pi = 0.25$, the *ENC-F* and *CCS* tests reject, respectively, 56% and 22% of the times, compared to 100% and 98% for the case of correct specification. The power loss is most extreme for the *CCS* test, probably in connection with the fact that several nuisance parameters are now embedded in the statistic (now distributed as χ^2_{p-1} under the null, instead of χ^2_1).

(3) *Over-parameterization*. In the unrestricted model, we take $(1, y_{t-1}, x_{t-1}, w_{t-1})'$ as predictors instead of $(1, y_{t-1}, x_{t-1})'$, where w_t is given by Eq. (12). When measured in terms of the size-adjusted power, the relative ranking of the tests remains unaltered once again, as expected. However, our simulation results document considerable finite sample size distortions for *F*-type tests of equal MSPE and FE, compared with the case of correctly specified models. Table 3 reports our findings for $R = 200$. The empirical size of *ENC-F* is now between 0.17 and 0.19 at the 10% level of significance; the *MSE-F* is also significantly oversized when $\pi \leq 0.5$. On the other hand, the size properties of the other tests are nearly unchanged. On balance, in the case of overparameterization, the use of *ENC-t* or *HLN* seems to be probably the best compromise between size and power properties.

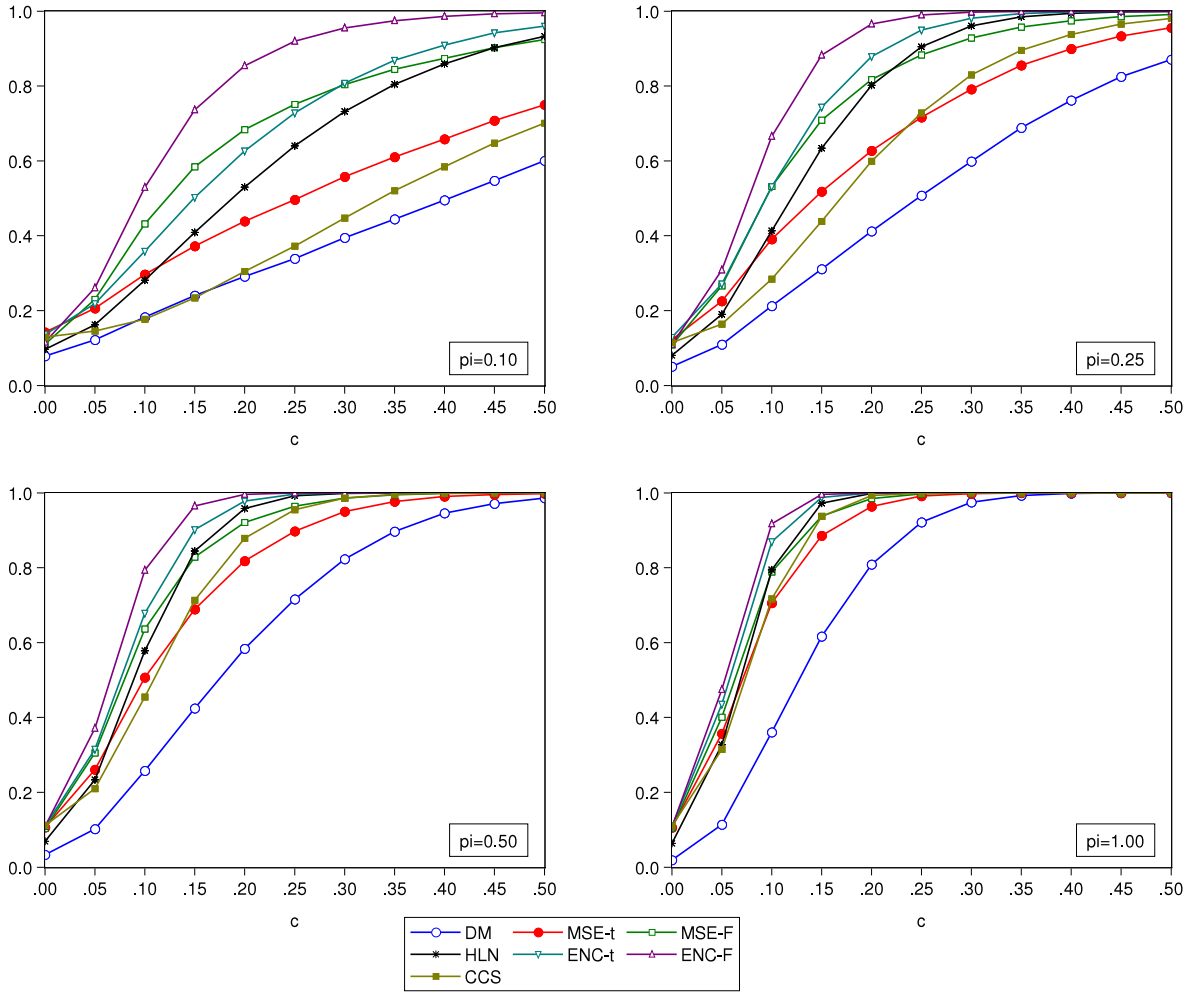
3.3. Fat-tailed distributions

This section considers the properties of the tests for data characterized by fat-tailed distributions. The disturbance terms in the data generating process in Eqs. (9)–(10) are no longer Gaussian, but follow a Student-*t* distribution with three degrees of freedom, implying the existence of

[12] Armah and Swanson (2008) and Corradi and Swanson (2007) also consider tests of predictive accuracy under (dynamic) model misspecification.

[13] When $\rho_{xw} = 0.5$, $\pi = 0.25$ and $R = 200$, the power loss for all tests is roughly 30%–35% with respect to the case of correct specification (at $c = 0.10$).

[14] When $\rho_{xw} = 0.8$, $\pi = 0.25$ and $R = 200$, the power loss for all tests is roughly 10%–15% with respect to the case of correct specification (at $c = 0.10$).



Notes: Results from 50,000 Monte Carlo simulations of one-step-ahead forecasts. Recursive scheme with in-sample $R = 200$. Nominal size = 10%.

Fig. 1. Empirical power functions for the case of one-step-ahead forecasts under correct specification ($R = 200$, recursive regressions).

finite moments of order at most two; this case is therefore not covered by the regularity assumptions used to derive the asymptotic critical values of the tests in the papers referenced in Table 1.

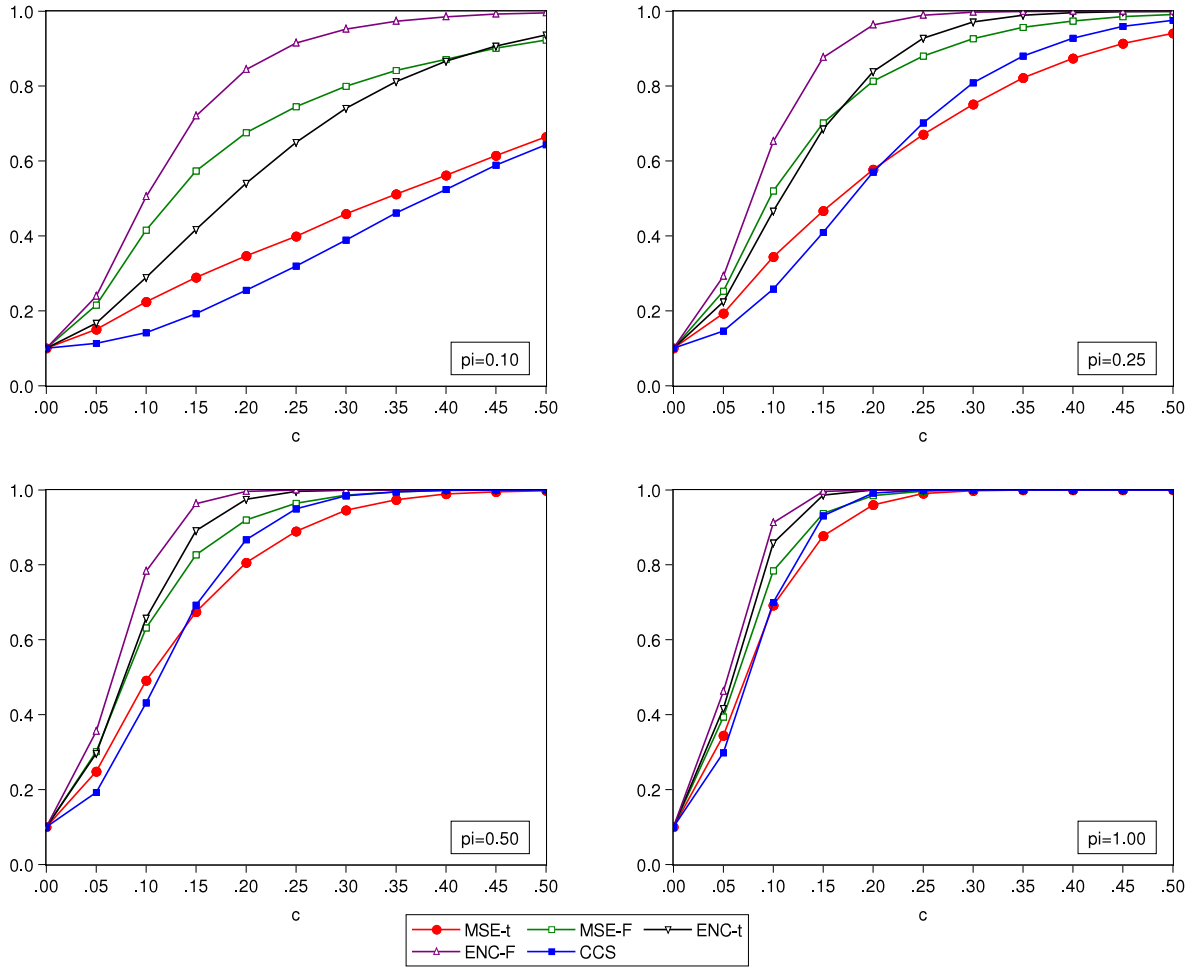
Table 4 contains the empirical size of the tests run at the nominal 5% and 10% levels, for $R = 100$ and $R = 200$. Interestingly, the sizes of the tests are not very different from the case of Gaussian observations, as reported in Table 2; however, there appears to be some stronger oversizing for the *MSE-t*, *ENC-t* and *ENC-F*, particularly when the fraction π of out-of-sample observations is low.

4. Properties of the tests: multi-step-ahead forecasts

Clark and McCracken (2005a) argue that for multi-step-ahead predictions, the critical values of the *ENC-t*, *MSE-t*, *ENC-F* and *MSE-F* tests should be obtained by bootstrap or simulation methods, as the limiting approximation generally depends on several nuisance parameters, which makes it infeasible to tabulate. However, in the case of a single additional regressor in the unrestricted model (as in the simulation experiment of this section), the asymptotic critical values for *ENC-t* and *MSE-t* coincide with those tabulated for the case of one-step-ahead forecasts.

Here, we consider multi-step-ahead predictions for correctly specified models, as in Section 3.1. We consider a "direct" forecasting scheme at horizon $h > 1$; thus, the prediction errors follow a MA($h-1$) process, and the test statistics should be computed using an estimate of the long-run variance. More precisely, the simulated data generating process is given by Eqs. (9)–(11), the restricted model is the regression of y_t on $(1, y_{t-h})$, and the predictors that characterize the nesting model are $(1, y_{t-h}, x_{t-h})$.

For the *ENC-F* and *MSE-F* tests, we provide results using bootstrap critical values; for *ENC-t*, *MSE-t* and *CCS*, we consider both asymptotic and bootstrap critical values; the *DM* and *HLN* tests are calculated as usual using the Gaussian critical values. The bootstrap algorithm is that of Kilian (1998), and is also implemented by Clark and McCracken (2005a). We denote the bootstrap



Notes: Results from 50,000 Monte Carlo simulations of one-step-ahead forecasts. Recursive scheme with in-sample $R = 200$. Nominal size = 10%.

Fig. 2. Empirical power functions (size-adjusted) for the case of one-step-ahead forecasts under correct specification ($R = 200$, recursive regressions).

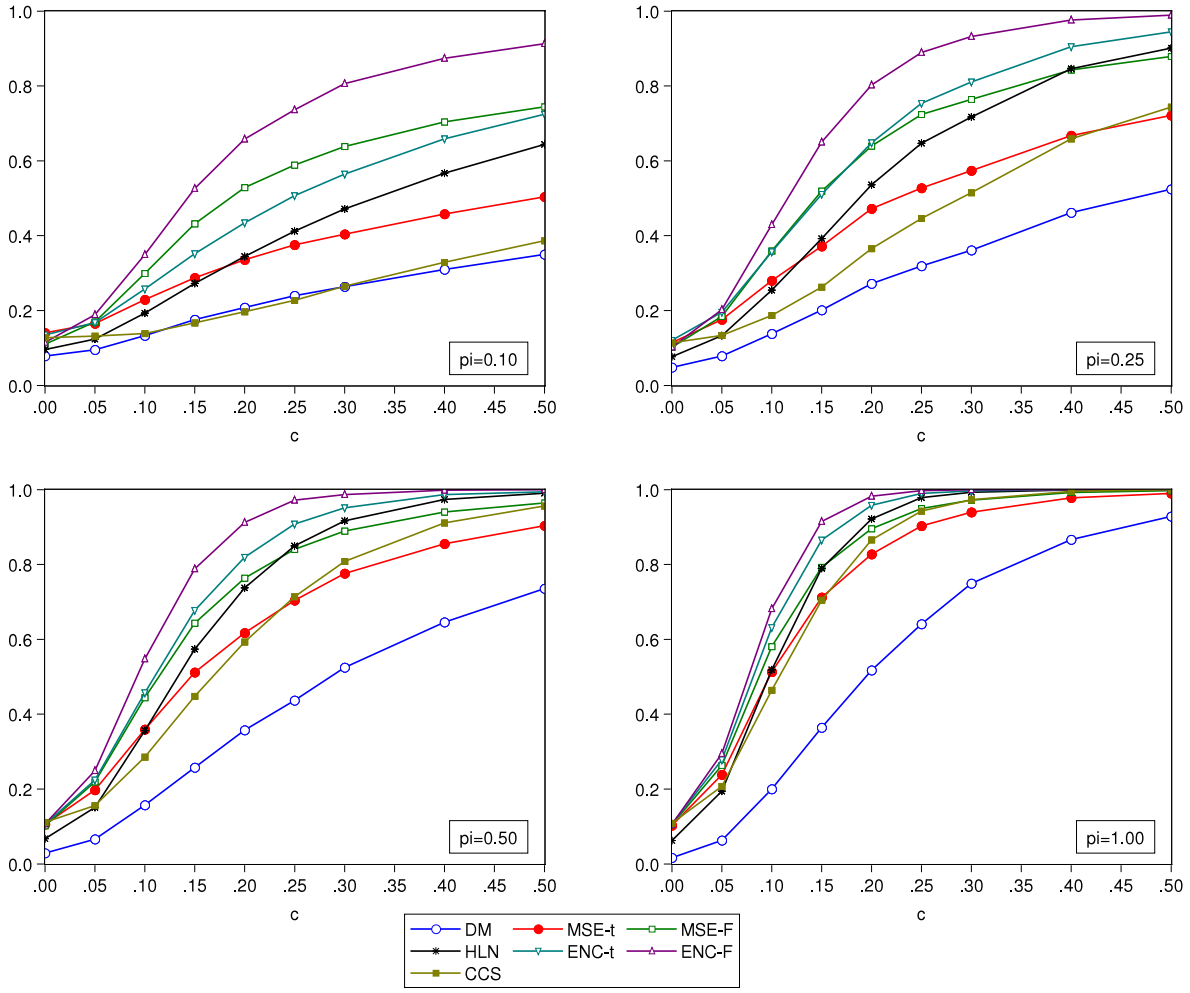
version of the tests by adding an asterisk to the name of the statistics, e.g. *MSE-F**.[15] For the tests compared with asymptotic critical values, we compute the statistics using alternative long run variance estimators, in order to check the sensitivity of the results to the correction for serial correlation. We consider four possibilities: (a) the Newey–West estimator with $m = 1.5h$, as suggested by Clark and McCracken (2005a); (b) the rectangular kernel with $m = h - 1$, as with Hansen (1982); (c) the rectangular kernel with finite sample correction, from Harvey, Leybourne, and Newbold (1997) (HLN); and (d) the quadratic spectral kernel of Andrews and Monahan (1992). An appropriate subscript is appended to the statistics for each of these cases (*NW*, *Rec*, *HLN*, and *QS*, respectively).

Tables 5 and 6 report the size of the tests, run at the 10% significance level, for the cases of $h = 2$- and 4-step-ahead predictions, $R = (100, 200)$, and recursive regressions.

For $\pi = 0.10$ and $h = 4$, the tests not based on bootstrap critical values are grossly oversized (except when the statistics are computed with the QS kernel), with huge distortions affecting *CCS*, *MSE-t* and *ENC-t*; on the other hand, the bootstrap allows us to control the size in all cases.[16] When $R = 200$ and $\pi \geq 0.25$, the HLN_{NW} test has reasonably good size properties for both the 2- and 4-step-ahead projections, while the *DM* test displays a strong tendency toward under-rejection, as for the case of one-step-ahead predictions. In general,

[15] Given the fact that Monte Carlo simulations are highly time consuming for multi-step-ahead forecasts (a double bootstrap is run at each iteration), we chose 5000 iterations, which took several days for each DGP.

[16] Contrary to the results of Clark and McCracken (2005a), we find that these size distortions tend to vanish as the number of out-of-sample observations *P* increases. However, one difference is that the regressors in their simulation experiment are chosen according to information criteria; in finite samples, this may be an important source of additional noise and mis-specification of the restricted model.



Notes: Results from 50,000 Monte Carlo simulations of one-step-ahead forecasts. Recursive scheme with in-sample $R = 200$, and correlation parameter $\rho_{xw} = 0.5$ in equation (12). Nominal size = 10%.

Fig. 3. Empirical power functions for the case of one-step-ahead forecasts under error-in-variable mis-specification ($R = 200$, recursive regressions).

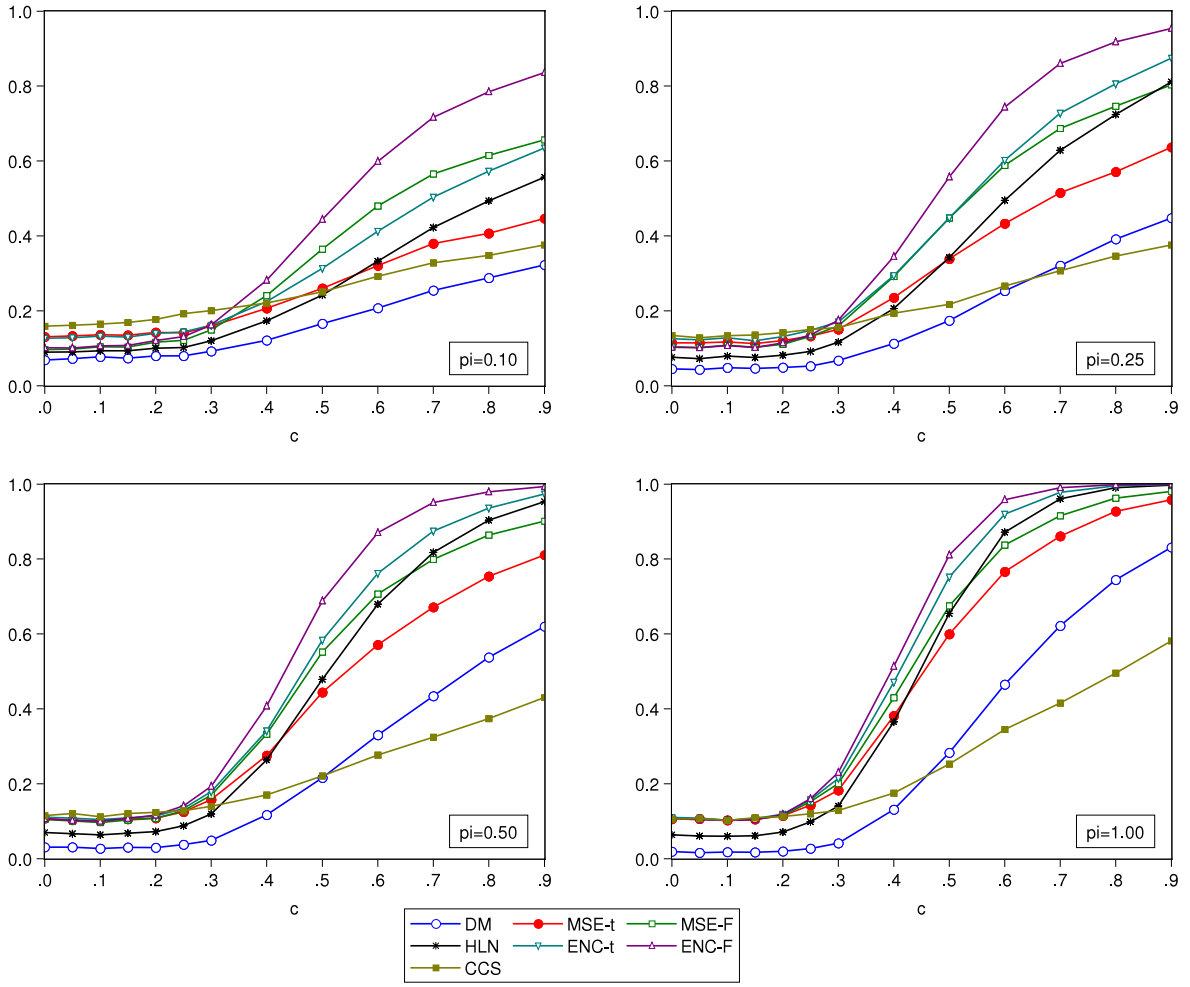
we confirm the finding of Clark and McCracken (2011) that the QS and HLN types of correction for serial correlation provide distinctly better results than NW. In our simulations, QS tends to 'overcorrect', leading to conservative tests. Based on our results, we would recommend using QS if the number of out-of-sample observations is 'small' ($\pi = 0.10$), and the rectangular kernel with the small-sample correction otherwise.

Fig. 5 provides the empirical power functions of the DM_{NW}, *MSE-t**, *MSE-F**, HLN_{NW}, *CCS**, *ENC-t**, and *ENC-F** tests for $\pi = (0.25, 0.50)$, $h = (2, 4)$, $R = 200$. The bootstrap version of the *F*-type test of Clark and McCracken (2001), *ENC-F**, displays the highest power, as for one-step-ahead predictions. The properties of the *HLN* test are now even better than in the one-step-ahead case: it is broadly equivalent to *MSE-F** and *ENC-t** for $h = 2$, and is more powerful for $h = 4$. Thus, given the computational burden of bootstrapping *ENC-F*, the use of *HLN*, based on Gaussian critical values, is a simple and reliable way of comparing the forecast accuracy for multi-step-ahead predictions.

5. Additional evidence for highly persistent data

The asymptotic distributions of the tests of equal MSPE and FE have been derived under the assumption of weakly dependent data. In large samples, the size and power of the tests should remain unaffected by the degree of persistence of x_t, as long as the data remain weakly dependent. However, there might be non-negligible finite sample effects when the series gets closer to the region of non-stationarity.

In our data generating process in Eqs. (9)–(11), increasing the persistence parameter ϕ_x yields a higher variance of the regressor x_t, and thus higher rejection rates for all tests, for a given value of q and $c > 0$. The finite sample effects of varying ϕ_x therefore need to be computed while holding constant the variance of x_t, $\sigma_x^2 = q^2/(1 - \phi_x^2)$, by



Notes: Results from 50,000 Monte Carlo simulations of one-step-ahead forecasts. Recursive scheme with in-sample $R = 200$. Nominal size = 10%.

Fig. 4. Empirical power functions for the case of one-step-ahead forecasts under mis-specification ($AR(p)$ model selected by BIC, $R = 200$, recursive regressions).

changing the parameter q correspondingly. Thus, here we compare the properties of the tests described in Sections 3 and 4 (where $\phi_x = 0.8$) with cases of lower and higher persistence levels, namely $\phi'_x = (0.5, 0.95, 0.99)$, where q is selected such that $1/(1 - 0.8^2) = q^2/(1 - \phi'_x)$.

Consider first Table 7, which reports the empirical sizes of tests run at the 10% significance level in the case of one-step-ahead errors, with $R = (100, 200)$ in-sample observations. The rejection frequencies are very similar across all values of ϕ_x for all tests, although there is a slightly bigger oversizing for the *MSE-t*, *MSE-F*, *ENC-t*, *ENC-F* and *CCS* tests for $\pi = 0.1$ when ϕ_x gets near to 1. However, the effects of high persistence in the data are very evident in terms of power. Fig. 6 compares the empirical power functions in terms of ϕ_x for the *MSE-F*, *ENC-F*, *HLN*, *CCS* tests (one for each quadrant), with $\pi = 0.5$ and $R = 200$ in the case of one-step-ahead errors; the results for the other tests, and for $\pi \neq 0.50$, are qualitatively similar, and therefore are not discussed. It is seen that the power of the tests is reduced considerably as ϕ_x tends to 1; e.g., when $\phi_x = 0.99$ for *ENC-F*, the rejection rates at $c = 0.10$ are about 30% lower than in the baseline case of $\phi_x = 0.80$. The *CCS* test is the most heavily affected, losing most of its power when there is a near-unit root. Except for *CCS*, the relative performances of the tests remain broadly the same as in the previous section.[17]

In summary, our results reveal a loss of power in the presence of highly persistent regressors, but the effects are negligible in terms of size. This is an important finding, as forecasting models in many empirical applications are built on highly persistent predictors; a prime example is the ongoing debate on the predictability of excess stock returns, see for example Welch and Goyal (2008).[18] Finally,

[17] Clearly, for each test and for a given distance c from the null hypothesis, the power loss is lower, the larger the value of R.

[18] Following a referee's suggestion, we have also considered a 'bias correction' (as did Stambaugh, 1999) for the coefficient of a highly

Table 3
Empirical size of the tests of equal forecast accuracy for one-step-ahead forecasts run at the nominal 5% and 10% levels under over-parameterization.

π	$q_w^2 = 0.25$				$q_w^2 = 1.0$			
	0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00
(A) Recursive 5%								
DM	0.04	0.02	0.01	0.00	0.04	0.02	0.01	0.00
MSE-t	0.08	0.06	0.05	0.05	0.08	0.06	0.05	0.05
MSE-F	0.10	0.09	0.08	0.06	0.10	0.09	0.08	0.06
HLN	0.05	0.04	0.04	0.04	0.05	0.04	0.04	0.04
ENC-t	0.08	0.08	0.07	0.07	0.08	0.08	0.07	0.07
ENC-F	0.11	0.10	0.10	0.10	0.11	0.10	0.10	0.10
CCS	0.07	0.06	0.06	0.06	0.07	0.06	0.06	0.06
(B) Recursive 10%								
DM	0.07	0.04	0.02	0.01	0.07	0.04	0.02	0.01
MSE-t	0.13	0.11	0.10	0.09	0.13	0.11	0.10	0.09
MSE-F	0.16	0.14	0.13	0.11	0.16	0.14	0.13	0.11
HLN	0.10	0.08	0.08	0.07	0.10	0.08	0.08	0.07
ENC-t	0.14	0.14	0.12	0.12	0.14	0.14	0.12	0.12
ENC-F	0.19	0.17	0.17	0.17	0.19	0.17	0.17	0.17
CCS	0.13	0.12	0.11	0.11	0.13	0.12	0.11	0.11

Notes: Results from 50,000 Monte Carlo iterations. One-step-ahead forecasts with the recursive scheme and an in-sample size $R = 200$. Misspecification due to over-parameterization, i.e., in the unrestricted model, we take $(1, y_{t-1}, x_{t-1}, w_{t-1})'$ as predictors instead of $(1, y_{t-1}, x_{t-1})'$, where w_t is given by Eq. (12).

we have also considered increasing only the coefficient, ϕ_y, attached to the lagged dependent variable. We find very small effects on the properties of the tests (even if $\phi_y = 1$), which remain nearly identical to those described in the previous sections (detailed results are available upon request).

6. An empirical application to forecasting Italian GDP growth

We investigate the forecast accuracy of prediction models for quarterly GDP growth based on monthly indicators for the Italian economy; these are sometimes called *bridge models*, see for example Parigi and Golinelli (2007). As in Section 3, the restricted model is an AR(1) process. In addition, the nesting model also considers one of the following indicators: industrial production (IP), an error correction term between industrial production and GDP (ECM), net exports (IE), car registrations (CAR), and business climate in the construction sector (CC). We also estimate a nesting model with all of these indicators (denoted 'bridge'). The prediction sample runs from 1999Q1 to 2007Q4, while the estimation sample recursively expands one quarter at each step, from 1987Q1 until 1998Q4.

The results are presented in Table 8. The last row of the table reports the mean squared prediction errors (MSPE) of the different models, which are compared with that of the AR(1) model (0.3868); the test results (in terms of the significance of each test) are contained in the previous rows.

According to all of the tests, the use of industrial production (either alone or combined with the other indicators) leads to significantly more accurate GDP forecasts than those obtained by a simple AR(1) model; among the other indicators, only the confidence in the construction sector seems to be a useful one, as is shown by the rejection of the *ENC-F* and *ENC-t* tests. We have also compared the industrial production model (IP) with the model that considers all indicators (Bridge). Despite the fact that the out-of-sample MSPEs for the two regressions are very similar (see the last row of the table), we found that the *ENC-t* and *ENC-F* tests reject the null hypothesis of equal forecast accuracy at the 1% significance level, while *CCS* and *HLN* (based on χ^2 and Gaussian critical values, respectively) reject at only the 5% level. The *DM* test does not reject at all, even at the 10% significance level, confirming the low power of this test for nested model comparisons.

7. Concluding remarks

The performances of several tests for comparing out-of-sample forecasts between competing nested models have been evaluated. Overall, the tests of forecast encompassing seem to be preferable to those of equal mean square prediction error. In particular, the highest (size-adjusted) power is achieved by the *ENC-F* test of Clark and McCracken (2001); however, it appears to be slightly oversized when the number of out-of-sample observations is 'small', and for cases of misspecified models. For multi-step-ahead predictions, a standard forecast encompassing test (based on Gaussian critical values) becomes relatively attractive, since it maintains good properties without the

persistent regressor that is negatively correlated with the disturbance of the predictive regression, i.e. $E(\hat{c} - c) = \frac{\sigma_{\varepsilon u}}{\sigma_u^2} E(\hat{\phi}_x - \phi_x) = -\frac{\sigma_{\varepsilon u}}{\sigma_u^2}(\frac{1+3\phi_x}{R})$, where R is the in-sample size. We compared tests run with and without this correction for bias. We found that the size of the tests was basically lower and closer to the nominal one using the correction, while the power was marginally to appreciably higher—depending on both the test and the ratio between the prediction and estimation sample sizes—if the bias correction was used. However, the relative 'ranking' of the tests was unaffected.

Table 4
Empirical size of the tests of equal forecast accuracy for one-step-ahead forecasts run at the nominal 5% and 10% levels under fat-tailed distributions (Student-t(3)).

π	$R = 100$				$R = 200$			
	0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00
(A) Recursive 5%								
DM	0.06	0.03	0.02	0.01	0.04	0.03	0.01	0.01
MSE-t	0.11	0.09	0.07	0.06	0.09	0.08	0.06	0.06
MSE-F	0.07	0.07	0.06	0.05	0.06	0.06	0.05	0.05
HLN	0.07	0.06	0.05	0.04	0.05	0.05	0.04	0.04
ENC-t	0.10	0.09	0.08	0.07	0.08	0.08	0.07	0.07
ENC-F	0.09	0.07	0.07	0.06	0.07	0.06	0.06	0.06
CCS	0.08	0.06	0.06	0.06	0.06	0.06	0.05	0.05
(B) Recursive 10%								
DM	0.10	0.06	0.04	0.02	0.08	0.05	0.03	0.02
MSE-t	0.18	0.14	0.12	0.11	0.15	0.13	0.11	0.11
MSE-F	0.12	0.11	0.10	0.10	0.11	0.10	0.10	0.10
HLN	0.13	0.10	0.09	0.08	0.10	0.09	0.08	0.07
ENC-t	0.17	0.16	0.14	0.13	0.15	0.15	0.12	0.12
ENC-F	0.14	0.12	0.12	0.11	0.12	0.11	0.11	0.11
CCS	0.14	0.13	0.12	0.12	0.12	0.12	0.11	0.11

Notes: Results from 50,000 Monte Carlo iterations. One-step-ahead forecasts with the recursive scheme. In-sample sizes: $R = (100, 200)$. Fat-tailed distributions (Student-t with 3 degrees of freedom).

Table 5
Empirical size of the tests of equal forecast accuracy for two-step-ahead forecasts run at the nominal 10% level.

π	$R = 100$				$R = 200$			
	0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00
Recursive 10%, $h = 2$								
DM_{NW}	0.18	0.09	0.06	0.03	0.13	0.07	0.04	0.02
DM_{Rec}	0.17	0.09	0.03	0.02	0.12	0.06	0.03	0.02
DM_{HLN}	0.14	0.07	0.03	0.02	0.10	0.06	0.03	0.02
DM_{QS}	0.04	0.05	0.02	0.01	0.05	0.04	0.02	0.01
$MSE\text{-}t_{NW}$	0.25	0.17	0.15	0.14	0.22	0.14	0.12	0.10
$MSE\text{-}t_{Rec}$	0.24	0.16	0.12	0.12	0.20	0.13	0.12	0.10
$MSE\text{-}t_{HLN}$	0.21	0.15	0.12	0.12	0.18	0.13	0.12	0.10
$MSE\text{-}t_{QS}$	0.16	0.13	0.10	0.10	0.14	0.11	0.09	0.09
$MSE\text{-}t$[a]	0.11	0.10	0.11	0.11	0.11	0.10	0.10	0.09
$MSE\text{-}F$[a]	0.11	0.10	0.10	0.11	0.11	0.10	0.10	0.10
HLN_{NW}	0.20	0.13	0.11	0.10	0.17	0.11	0.09	0.07
HLN_{Rec}	0.18	0.11	0.09	0.07	0.14	0.10	0.07	0.06
HLN_{HLN}	0.15	0.10	0.08	0.07	0.12	0.09	0.07	0.06
HLN_{QS}	0.04	0.07	0.05	0.04	0.06	0.06	0.04	0.04
$ENC\text{-}t_{NW}$	0.25	0.19	0.16	0.15	0.23	0.16	0.13	0.12
$ENC\text{-}t_{Rec}$	0.23	0.17	0.13	0.12	0.19	0.15	0.12	0.11
$ENC\text{-}t_{HLN}$	0.19	0.16	0.13	0.12	0.18	0.15	0.12	0.11
$ENC\text{-}t_{QS}$	0.10	0.12	0.10	0.09	0.11	0.11	0.08	0.08
$ENC\text{-}t$[a]	0.11	0.09	0.11	0.11	0.11	0.10	0.09	0.09
$ENC\text{-}F$[a]	0.10	0.10	0.10	0.11	0.11	0.10	0.09	0.10
CCS_{NW}	0.29	0.21	0.16	0.14	0.20	0.15	0.15	0.12
CCS_{Rec}	0.27	0.14	0.14	0.13	0.16	0.13	0.13	0.10
CCS_{HLN}	0.19	0.12	0.13	0.12	0.13	0.12	0.12	0.10
CCS_{QS}	0.03	0.03	0.05	0.06	0.03	0.05	0.05	0.05
CCS[a]	0.10	0.11	0.10	0.10	0.10	0.11	0.11	0.09

Notes: Results from 5000 Monte Carlo iterations. Two-step-ahead forecasts computed by the direct method with the recursive scheme. Different kernels: Newey–West (NW), Rectangular (Rec), Rectangular with HLN correction (HLN) and Quadratic Spectral (QS). In-sample sizes: $R = (100, 200)$.
[a] Tests are bootstrapped.

complications involved in bootstrapping the critical values of the *ENC-F* test.

However, the simulation results presented do not account for either structural breaks or model uncertainty. In fact, Clark and McCracken (2005b, 2009) show that breaks have a significant effect on the properties of tests of predictive ability, and thus they may render the task of discriminating between competing models harder; they also argue that the choice of estimation window in rolling regressions is crucial, given the bias-variance tradeoff induced by parameter instability. As real world forecasts can never be generated by the underlying "true model", we believe that taking misspecification and model uncertainty into account is an important direction for research.

Table 6
Empirical size of the tests of equal forecast accuracy for four-step-ahead forecasts run at the nominal 10% level.

π	$R = 100$				$R = 200$			
	0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00
Recursive 10%, $h = 4$								
DM_{NW}	0.27	0.13	0.09	0.05	0.20	0.08	0.06	0.03
DM_{Rec}	0.26	0.12	0.06	0.03	0.19	0.07	0.05	0.03
DM_{HLN}	0.15	0.10	0.05	0.02	0.14	0.06	0.05	0.03
DM_{QS}	0.04	0.04	0.03	0.02	0.05	0.03	0.03	0.02
$MSE\text{-}t_{NW}$	0.33	0.21	0.16	0.15	0.27	0.16	0.15	0.12
$MSE\text{-}t_{Rec}$	0.32	0.20	0.14	0.12	0.27	0.15	0.14	0.11
$MSE\text{-}t_{HLN}$	0.25	0.19	0.14	0.11	0.22	0.13	0.13	0.11
$MSE\text{-}t_{QS}$	0.14	0.14	0.13	0.11	0.15	0.11	0.12	0.10
$MSE\text{-}t$[a]	0.11	0.10	0.11	0.11	0.12	0.09	0.10	0.09
$MSE\text{-}F$[a]	0.12	0.12	0.12	0.13	0.11	0.10	0.09	0.10
HLN_{NW}	0.30	0.18	0.13	0.12	0.23	0.12	0.11	0.07
HLN_{Rec}	0.29	0.17	0.11	0.09	0.21	0.11	0.09	0.06
HLN_{HLN}	0.17	0.13	0.10	0.08	0.17	0.09	0.08	0.06
HLN_{QS}	0.03	0.06	0.06	0.05	0.06	0.04	0.06	0.05
$ENC\text{-}t_{NW}$	0.34	0.24	0.18	0.17	0.27	0.18	0.15	0.12
$ENC\text{-}t_{Rec}$	0.33	0.23	0.16	0.13	0.27	0.16	0.14	0.11
$ENC\text{-}t_{HLN}$	0.23	0.19	0.14	0.13	0.21	0.15	0.14	0.10
$ENC\text{-}t_{QS}$	0.08	0.12	0.12	0.11	0.11	0.11	0.10	0.08
$ENC\text{-}t$[a]	0.10	0.11	0.11	0.12	0.12	0.08	0.10	0.08
$ENC\text{-}F$[a]	0.11	0.12	0.12	0.14	0.10	0.10	0.11	0.10
CCS_{NW}	0.40	0.26	0.20	0.15	0.27	0.20	0.17	0.15
CCS_{Rec}	0.38	0.24	0.16	0.12	0.24	0.17	0.13	0.12
CCS_{HLN}	0.14	0.18	0.13	0.12	0.17	0.14	0.11	0.11
CCS_{QS}	0.02	0.04	0.05	0.06	0.03	0.05	0.05	0.06
CCS[a]	0.11	0.11	0.10	0.12	0.09	0.10	0.10	0.11

Notes: Results from 5000 Monte Carlo iterations. Four-step-ahead forecasts computed by the direct method with the recursive scheme. Different kernels: Newey–West (NW), Rectangular (Rec), Rectangular with HLN correction (HLN) and Quadratic Spectral (QS). In-sample sizes: $R = (100, 200)$.
[a] Tests are bootstrapped.

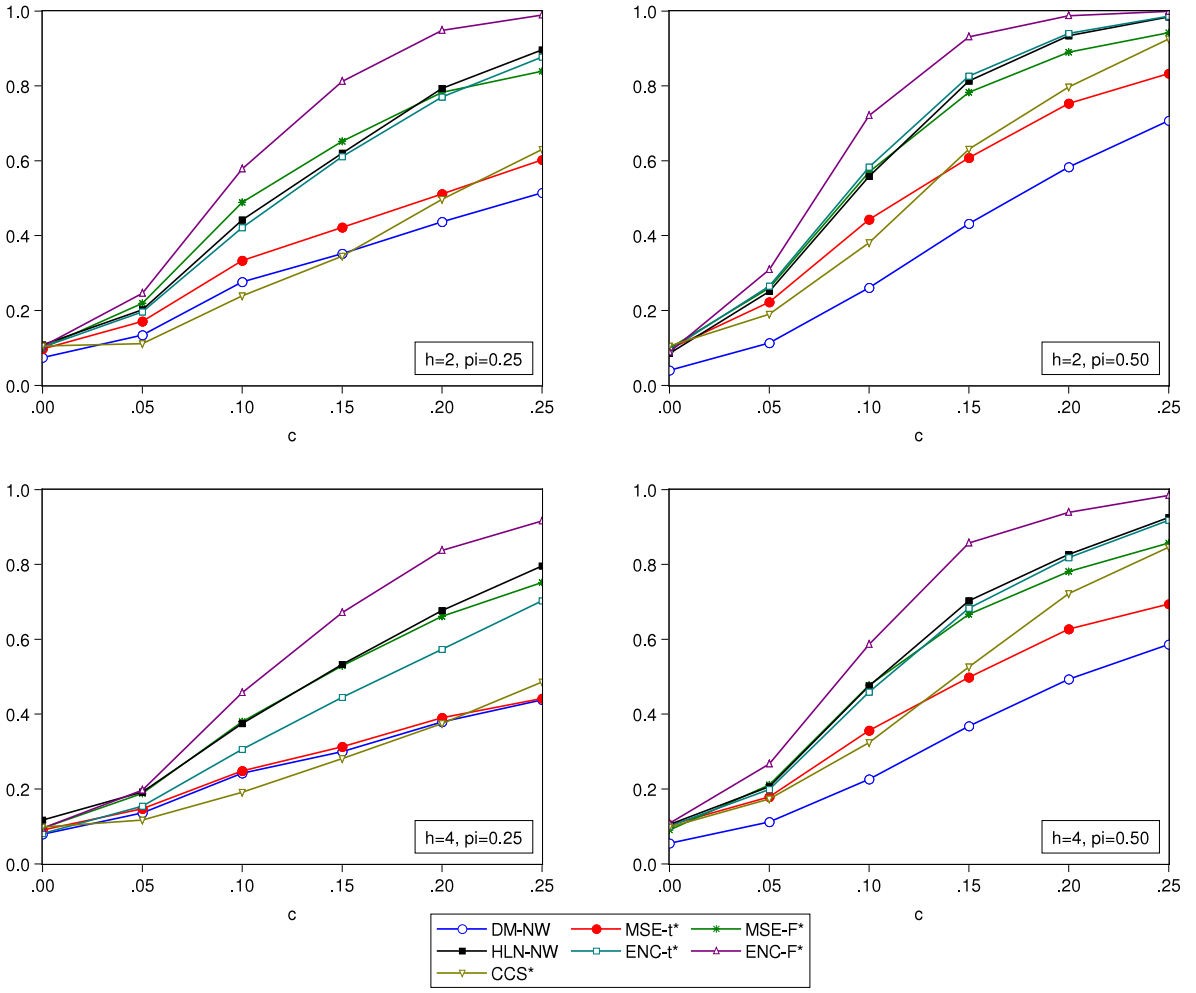
Table 7
Empirical size of the tests of equal forecast accuracy for one-step-ahead forecasts run at the nominal 10% level. Different degrees of persistence in x.

π	$\phi_x = 0.5$				$\phi_x = 0.8$				$\phi_x = 0.95$				$\phi_x = 0.99$			
	0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00
(A) Recursive 10%, $R = 100$																
DM	0.10	0.06	0.03	0.02	0.10	0.06	0.04	0.02	0.11	0.06	0.04	0.02	0.11	0.06	0.04	0.02
MSE-t	0.16	0.13	0.11	0.11	0.17	0.13	0.12	0.11	0.17	0.14	0.12	0.10	0.17	0.14	0.11	0.10
MSE-F	0.12	0.11	0.10	0.10	0.12	0.12	0.11	0.10	0.13	0.13	0.11	0.11	0.14	0.13	0.12	0.10
HLN	0.11	0.09	0.07	0.06	0.12	0.09	0.08	0.07	0.12	0.10	0.09	0.07	0.13	0.10	0.09	0.08
ENC-t	0.15	0.14	0.11	0.11	0.16	0.14	0.12	0.11	0.17	0.15	0.13	0.12	0.17	0.15	0.13	0.12
ENC-F	0.12	0.11	0.11	0.11	0.13	0.12	0.12	0.11	0.14	0.14	0.14	0.13	0.16	0.15	0.15	0.15
CCS	0.15	0.12	0.11	0.11	0.15	0.13	0.12	0.12	0.16	0.14	0.13	0.12	0.16	0.14	0.13	0.12
(B) Recursive 10%, $R = 200$																
DM	0.07	0.05	0.03	0.02	0.08	0.05	0.03	0.02	0.08	0.06	0.03	0.02	0.09	0.05	0.03	0.01
MSE-t	0.14	0.12	0.11	0.10	0.14	0.12	0.11	0.11	0.15	0.13	0.11	0.10	0.15	0.12	0.10	0.08
MSE-F	0.11	0.11	0.10	0.10	0.11	0.11	0.10	0.10	0.11	0.11	0.10	0.10	0.12	0.11	0.10	0.09
HLN	0.09	0.08	0.07	0.06	0.10	0.08	0.07	0.06	0.10	0.09	0.07	0.06	0.11	0.09	0.07	0.06
ENC-t	0.13	0.13	0.11	0.10	0.14	0.13	0.11	0.11	0.14	0.14	0.11	0.10	0.15	0.14	0.11	0.11
ENC-F	0.11	0.11	0.11	0.10	0.12	0.11	0.11	0.11	0.12	0.12	0.12	0.11	0.13	0.13	0.13	0.13
CCS	0.13	0.11	0.11	0.10	0.13	0.11	0.11	0.11	0.13	0.12	0.11	0.11	0.13	0.12	0.11	0.10

Notes: Results from 50,000 Monte Carlo iterations. One-step-ahead forecasts with the recursive scheme. Different degrees of persistence in x. In-sample sizes: $R = (100, 200)$.

Acknowledgments

We would like to thank Michael Clements, the editor, and two anonymous referees for many helpful suggestions which helped to improve the paper considerably. We also wish to acknowledge the contribution of Giovanni Veronese to the previous version of this paper, and the help of Roberto Stok for the implementation of the computer-intensive simulations. Furthermore, we thank Joerg Breitung, Graham Elliott, Raffaella Giacomini, Lutz Kilian, Michael McCracken and Ken West for useful comments and suggestions. Finally, we would like to thank participants at the 2008 workshop on "Factor models, high frequency data and short term



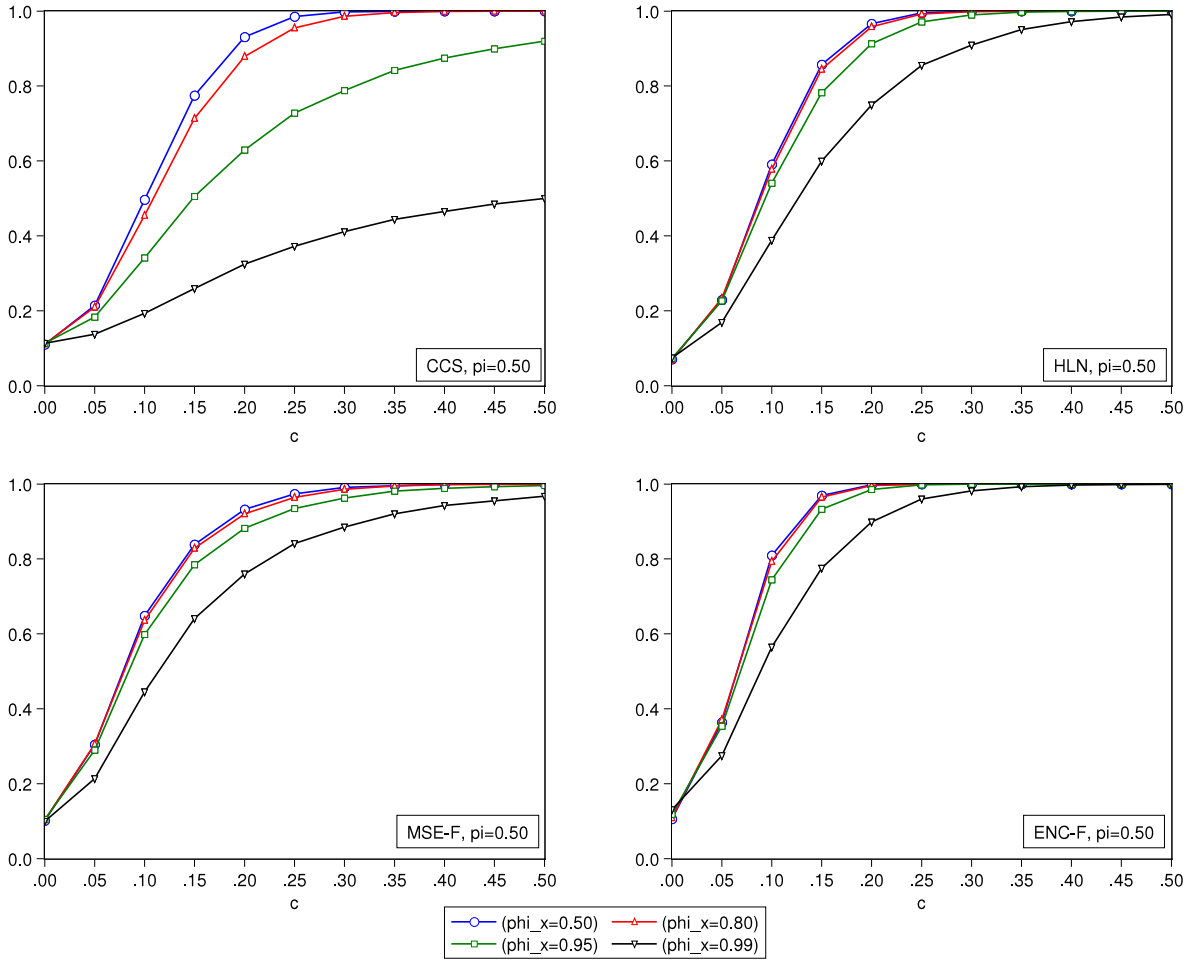
Notes: Results from 5,000 Monte Carlo simulations of multi-step-ahead forecasts. Recursive scheme with in-sample $R = 200$, $h = (2, 4)$, $\pi = (0.25, 0.50)$. Nominal size = 10%.

Fig. 5. Empirical power functions for the case of multi-step-ahead forecasts under correct specification ($R = 200$, recursive regressions).

Table 8
Empirical application: forecasting the Italian GDP one quarter ahead.

	ARX_{car}	ARX_{CC}	ARX_{ECM}	ARX_{IE}	ARX_{IP}	Bridge
DM	–	–	–	–	***	***
MSE-t	–	–	–	–	***	***
MSE-F	–	–	–	–	***	***
HLN	–	–	–	–	***	***
ENC-t	–	*	–	–	***	***
ENC-F	*	***	–	–	***	***
CCS	–	–	–	–	***	***
MSE	0.3926	0.3861	0.3943	0.3905	0.2527	0.2409

Notes: AR(1) benchmark vs. the model in the column. The table shows the results of different tests of equal forecast accuracy and FE for the quarterly Italian GDP. The benchmark model is an AR(1) and the alternative models are ARX(1) models and a Bridge model. The exogenous variables in the ARX(1) models are the new car sales ARX_{car}, the construction confidence indicator ARX_{CC}, an error correction term ARX_{ECM}, the growth rate of net exports ARX_{IE}, and the growth rate of industrial production ARX_{IP}. The Bridge model combines the lagged dependent variable with all of the exogenous indicators.
'–' indicates no rejection of the null hypothesis at the 10% level.
* indicates rejection at the 10% level.
*** indicates rejection at the 1% level.



Notes: Results from 50,000 Monte Carlo simulations of one-step-ahead forecasts. Recursive scheme with in-sample $R = 200$, $\pi = 0.50$, $\phi_x = (0.5, 0.8, 0.95, 0.99)$. Nominal size = 10%.

Fig. 6. Empirical power functions for the case of one-step-ahead forecasts under different degrees of persistence ($R = 200, \pi = 0.50$, recursive regressions).

forecasting", the 3rd ICEEE Conference, the 1st IFO-INSEE-ISAE Macroeconomic Forecasting Conference, the XVII SNDE Conference, and seminar participants at the Federal Reserve Bank of Boston. The views expressed are those of the authors and do not necessarily reflect those of the Bank of Italy.

References

Andrews, D. W. K., & Monahan, J. C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, *60*, 953–966.

Armah, N. A., & Swanson, N. (2008). Predictive inference under model misspecification with an application to assessing the marginal predictive content of money for output. In M. Wohar (Ed.), *Forecasting in the presence of structural breaks and model uncertainty* (pp. 195–230). Bingley, UK: Emerald.

Busetti, F., Marcucci, J., & Veronese, G. (2009). *Comparing forecast accuracy: a Monte Carlo investigation*. Bank of Italy Discussion Papers, No. 723.

Chao, J., Corradi, V., & Swanson, N. (2001). An out-of-sample test for Granger causality. *Macroeconomic Dynamics*, *5*, 598–620.

Chong, Y. Y., & Hendry, D. F. (1986). Econometric evaluation of linear macroeconomic models. *Review of Economic Studies*, *53*, 671–690.

Clark, T. E., & McCracken, M. W. (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics*, *105*, 85–110.

Clark, T. E., & McCracken, M. W. (2005a). Evaluating direct multistep forecasts. *Econometric Reviews*, *24*, 369–404.

Clark, T. E., & McCracken, M. W. (2005b). The power of tests of predictive ability in the presence of structural breaks. *Journal of Econometrics*, *124*, 1–31.

Clark, T. E., & McCracken, M. W. (2009). Improving forecast accuracy by combining recursive and rolling forecasts. *International Economic Review*, *50*, 363–395.

Clark, T. E., & McCracken, M. W. (2011). *Advances in forecast evaluation*. Mimeo.

Clark, T. E., & West, K. D. (2006). Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics*, *135*, 155–186.

Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, *138*, 291–311.

Clements, M. P., & Hendry, D. F. (1993). On the limitations of comparing mean square forecast errors (with discussion). *Journal of Forecasting*, *12*, 617–676.

Corradi, V., & Swanson, N. (2002). A consistent test for nonlinear out-of-sample predictive accuracy. *Journal of Econometrics*, *110*, 353–381.

Corradi, V., & Swanson, N. (2007). Nonparametric bootstrap procedures for predictive inference based on recursive estimation schemes. *International Economic Review*, *48*, 67–109.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *13*, 253–263.
Ericsson, N. R. (1992). Parameter constancy, mean square forecast errors, and measuring forecast performance: an exposition, extensions, and illustration. *Journal of Policy Modeling*, *14*, 465–495.
Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*, 1545–1578.
Granger, C. W. J., & Newbold, P. (1986). *Forecasting economic time series* (2nd ed.). New York: Academic Press.
Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, *50*, 1029–1054.
Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business and Economic Statistics*, *23*, 365–380.
Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared error. *International Journal of Forecasting*, *13*, 281–291.
Harvey, D. I., Leybourne, S. J., & Newbold, P. (1998). Tests of forecast encompassing. *Journal of Business and Economic Statistics*, *16*, 254–259.
Hubrich, K., & West, K. D. (2010). Forecast evaluation of small nested model sets. *Journal of Applied Econometrics*, *25*, 574–594.
Kilian, L. (1998). Small-sample confidence intervals for impulse response functions. *Review of Economics and Statistics*, *80*, 218–230.
McCracken, M. W. (2007). Asymptotics for out-of-sample tests of Granger causality. *Journal of Econometrics*, *140*, 719–752.
Nelson, C. R. (1972). The prediction performance of the FRB-MIT-PENN model of the US economy. *American Economic Review*, *62*, 902–917.
Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*, 703–708.
Parigi, G., & Golinelli, R. (2007). The use of monthly indicators to forecast quarterly GDP in the short run: an application to the G7 countries. *Journal of Forecasting*, *26*, 77–94.
Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics*, *54*, 375–421.
Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*, 1455–1508.
West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, *64*, 1067–1084.
West, K. D. (2006). Forecast evaluation. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of economic forecasting*: *Vol. 1* (pp. 100–134). Amsterdam: Elsevier.
White, H. (2000). A reality check for data snooping. *Econometrica*, *68*, 1097–1126.

Fabio Busetti is an economist in the Research Department of the Bank of Italy. His research interests are in nonstationary time series, unobserved component models, structural change, seasonality, cointegration, time-varying distributions and copulas, convergence and forecasting. He is a fellow of the Granger Center for Time Series Econometrics, the Euro Area Business Cycle Network and the Working Group of Forecasting of the ESCB. He has been a visiting scholar at the Department of Economics at the University of California San Diego. He has also been a lecturer at the University of Rome Tor Vergata. His work has appeared in many international journals: *Journal of Econometrics, Econometric Theory, Journal of Business and Economic Statistics, Journal of Time Series Analysis, Journal of Applied Econometrics,* and *Journal of Forecasting*. He obtained his Ph.D. in Economics from the London Schoool of Economics.

Juri Marcucci is an economist in the Research Department of the Bank of Italy. His research interests are in volatility modeling and forecasting, pure variance common features models, forecast evaluation, forecasting with nonlinear time series, predictability of returns, nonlinear panel data models, credit risk cyclicality. He has been a lecturer at the University of Bologna and the University of Rome Tor Vergata. His work has appeared in the *Journal of Econometrics, Studies in Nonlinear Dynamics and Econometrics, Journal of Banking and Finance, Journal of Economics and Business, Journal of International Financial Markets, Institution and Money*, and *International Review of Financial Analysis*. He obtained his Ph.D. in economics from the University of California San Diego.