

Comparing forecast accuracy: a Monte Carlo investigation

Fabio Busetti, Juri Marcucci* and Giovanni Veronese
Bank of Italy - Research Department

Abstract

The properties of several tests of equal Mean Square Prediction Error (MSPE) and tests of Forecast Encompassing (FE) are evaluated, using simulation methods, in the context of dynamic regressions. For nested models, larger differences in the behavior of the tests occur when the number of out-of-sample observations is relatively small compared to the size of the estimation sample. In this case the standard tests of equal MSPE and of FE retain good size properties but they pay a big price in terms of power; overall the FE test *ENC-F* of Clark and McCracken (2001), despite being slightly oversized, is clearly the most powerful. For longer spans of the prediction sample, the power advantage of *ENC-F* tends to become smaller, and thus a standard FE test, based on Gaussian critical values, may become relatively more attractive. The ranking among the tests does not change significantly for multi-step ahead predictions, as well as for cases where the estimated models are partly misspecified. A similar simulation setup is used to analyze the case of non-nested models. Again, we find that FE tests have a significantly better performance with respect to tests of equal MSPE, for discriminating between correct and misspecified models. An empirical example with models of prediction of euro-area and US inflation is provided.

Keywords: Forecast encompassing, Model evaluation, Nested models, Predictive accuracy.

JEL Classification: C12, C52, C53

*The views expressed are those of the authors and do not necessarily reflect those of the Bank of Italy. March 2008, *First draft*. Corresponding author: Juri Marcucci. Emails: fabio.busetti@bancaditalia.it (Fabio Busetti), juri@sssup.it (Juri Marcucci), giovanni.veronese@bancaditalia.it (Giovanni Veronese).

1 Introduction

Evaluating the out-of-sample performance 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 prediction accuracy measured in terms of some general loss function. In most practical applications, however, little attention is paid to the shape of the loss function and models are generally compared on the basis of their mean square prediction error (MSPE). An alternative approach looks at the out-of-sample correlation between prediction errors, which leads to tests of forecast encompassing (FE) or, in the terminology of Granger and Newbold (1986), of conditional forecast efficiency. 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), 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 it may enter 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 they are not asymptotically normally distributed. For nested models McCracken (2007) and Clark and McCracken (2001) derive the appropriate non Gaussian limit for tests of, respectively, equal MSPE and FE; the critical values are tabulated across two nuisance parameters (the ratio of the magnitudes of prediction sample to estimation sample and the number of additional regressors in the larger model) and they are, in general, valid only 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. A different approach is taken by Giacomini and White (2006) that focus on comparing forecasting methods as opposed to forecasting models: their test statistic of equal conditional predictive ability has a Gaussian null distribution as the prediction sample size tends to infinite for a finite length of the estimation sample. An excellent survey on asymptotic inference for predictive ability for nested and non-nested models is West (2006).

In this paper we evaluate the properties of several tests of equal MSPE and tests of FE, with the goal to provide practical guidance to forecasters needing to choose among a set of predictions from (a small number of)

competing models¹. We use Monte Carlo simulation methods to compute empirical size and empirical power functions in the context of dynamic regression models. One-step and multi-step ahead predictions are considered for correctly specified and misspecified regressions; we also investigate the properties of the tests across different values of the ratio between prediction and estimation sample sizes.

The tests under scrutiny are the followings: (i) the standard Diebold-Mariano test of equal MSPE; (ii) the $MSE - t$ and (iii) the $MSE - F$ modifications of McCracken (2007) for nested models; (iv) the forecast encompassing test of Harvey, Leybourne and Newbold (1998); (v) the $ENC - t$ and (vi) $ENC - F$ modifications of Clark and McCracken (2001) for nested models; (vii) the forecast encompassing test of Chao, Corradi and Swanson (2001) for nested models.²

As discussed in section 3, our results extend the more limited Monte Carlo comparison among a few of these tests carried out in Clark and McCracken (2001, 2005a). We confirm their findings that, overall, for nested models the $ENC-F$ test has the best properties, noticing however that its power advantage tends to become smaller as the dimension of the prediction sample increases. In particular, we find that the relative ranking among the different tests changes according to whether the number of out-of-sample observations is "small" or "large". For non-nested models we again obtain a significant advantage for the tests of forecast encompassing over those of equal MSPE in discriminating between a correct and a mis-specified model.

In summary, the paper proceeds as follows. Section 2 briefly reviews the test statistics under scrutiny. Section 3 and 4 contain the simulation results for nested and non-nested models respectively. In section 5 we provide an empirical example with prediction models for euro-area and US inflation. Section 6 concludes.

2 The setup and the tests under scrutiny

We consider a sample of T observations on a target series y_t and two k_i -dimensional vectors of (non mutually exclusive) predictors X_{it} , $i = 1, 2$. The

¹For issues arising on comparing a large number of models, see White (2000), Hansen (2005), Hubrich and West (2007).

²In the comparison we do not include the method of Giacomini and White (2006) because it relates to a different null hypothesis from the other tests. In their framework, under the null hypothesis the true model is the larger one, implying a bias-variance tradeoff in the forecasts. In our experiments for the nested case, the null hypothesis is the smaller, or restricted, model and there is no bias. See the discussion of Giacomini and White (2006, p.1559-1561)

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 generated by the linear models

$$\hat{y}_{it} = X'_{i,t-h} \hat{\beta}_{i,t-h}, \quad t = R + h, R + h + 1, \dots, T \quad (1)$$

where $\hat{\beta}_{i,t-h}$ is the least square 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 the recursive or the rolling scheme: the recursive least square 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 $e_{it} = y_t - \hat{y}_{it}$, $i = 1, 2$, for $t = R + h, R + h + 1, \dots, R + P$; for simplicity we have suppressed the dependency on h in the notation. The tests under scrutiny are briefly detailed below.

2.1 Tests of equal MSPE

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

$$DM = P^{\frac{1}{2}} \bar{d} / \hat{\sigma}_{DM}(m), \quad (2)$$

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

$$\hat{\sigma}_{DM}^2(m) = P^{-1} \sum_{t=R+h}^T (d_t - \bar{d})^2 + 2P^{-1} \sum_{j=1}^m w(j, m) \sum_{t=j+R+h}^T (d_t - \bar{d})(d_{t-j} - \bar{d}), \quad (3)$$

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 if e_{1t} , e_{2t} are h -step ahead forecast errors, then m should be at least equal to $h - 1$.³ The DM statistic in (2) tests the null hypothesis $H_0 : E d_t = 0$. If the models are *non nested*, the limiting null distribution of (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;

³A lag truncation parameter $m > 0$ should be used also for 1-step ahead predictions if the second moments of the data follow some GARCH or related process.

however the Gaussian critical values would still approximately hold if P/R is small (e.g. less than 0.1, see West, 2006).

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 critical values, will be called $MSE - t$. The following F -type statistic is also proposed

$$MSE - F = P\bar{d} / \hat{\sigma}_2^2 \quad (4)$$

where $\hat{\sigma}_2^2 = P^{-1} \sum_{t=R+1}^T e_{2t}^2$ 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 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 forecasts. The limiting distributions are different for the case of multi-step ahead predictions, but critical values can be obtained by bootstrap; see Clark and McCracken (2005a).

For the case of *nested* models the standard DM test turns out to be heavily undersized and with low power. Although the correct limiting distribution is not 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 size close to, but a little less than, the nominal one. Specifically, the Clark-West adjusted statistic is

$$DM - adj = P^{\frac{1}{2}} \bar{d}^* / \hat{\sigma}_{DM-adj}(m), \quad (5)$$

where $\bar{d}^* = \frac{1}{P} \sum_{t=R+h}^T d_t^*$, $d_t^* = e_{1t}^2 - e_{2t}^2 + (\hat{y}_{1t} - \hat{y}_{2t})^2$ and $\hat{\sigma}_{DM-adj}^2(m)$ is the non-parametric estimator of the long run variance of d_t^* , that parallels the definition in (3). Since $(\hat{y}_{1t} - \hat{y}_{2t})^2 = (e_{2t} - e_{1t})^2$ one can write $d_t^* = 2e_{1t}(e_{1t} - e_{2t})$. Thus, as noted in West (2006), the DM -adjusted statistic is based on the covariance between e_{1t} and $e_{1t} - e_{2t}$, and it corresponds to the test of forecast encompassing given in (6) below.

2.2 Tests of forecast encompassing

It is said that the forecast \hat{y}_{1t} encompasses \hat{y}_{2t} if there is no gain from combining them into a composite forecast $\hat{y}_{ct} = (1 - \lambda)\hat{y}_{1t} + \lambda\hat{y}_{2t}$, for some weight $\lambda > 0$; see *inter alia* Clements and Hendry (1993), Granger and Newbold (1986) where forecast encompassing is termed “conditional efficiency” and the early empirical work of Nelson (1972). As the combined forecast error e_{ct} satisfies the relation $e_{1t} = \lambda(e_{1t} - e_{2t}) + e_{ct}$, 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, Leybourne and

Newbold (1998) write the null hypothesis of forecast encompassing as $H_0 : Ef_t = 0$, with $f_t = e_{1t}(e_{1t} - e_{2t})$, and construct a t-test on $\bar{f} = \frac{1}{P} \sum_{t=R+1}^T f_t$; their statistic is

$$HLN = P^{\frac{1}{2}} \bar{f} / \hat{\sigma}_{HLN}(m), \quad (6)$$

where $\hat{\sigma}_{HLN}^2(m)$ is a non-parametric estimator of the long run variance of f_t , that parallels the definition in (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 when applied to nested models the *HLN* statistic is no longer asymptotically Gaussian and they obtain the correct null limiting distribution for one-step ahead forecasts: the test that uses their critical value will be called *ENC-t*. They also propose the F-type statistic

$$ENC - F = P\bar{f} / \hat{\sigma}_2^2, \quad (7)$$

where $\hat{\sigma}_2^2$ is the mean squared forecast error of the nesting model as in (4). The distributions of *ENC-t* and *ENC-F* depend on the ratio 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 forecasts. The extension for multi-step ahead forecasts is given in Clark and McCracken (2005a).

A different test of forecast encompassing for nested models has been proposed by Chao, Corradi and Swanson (2001): the null hypothesis is $H_0 : Ec_t = 0$, where $c_t = e_{1t}(Z_{2t} - \bar{Z}_2)$ and Z_{2t} are the additional $k_2 - k_1$ predictors in X_{2t} not included in X_{1t} .⁴ This is again a Wald-type test with statistic given by

$$CCS = P\bar{c}' (\hat{\Sigma}_{CCS}(m))^{-1} \bar{c}, \quad (8)$$

where $\bar{c} = \frac{1}{P} \sum_{t=R+1}^T c_t$ and $\hat{\Sigma}_{CCS}(m)$ is a non-parametric estimator of the long run variance-covariance matrix of c_t , that parallels the definition in (3). Under the null hypothesis of forecast encompassing *CCS* is asymptotically distributed as chi square with $k_2 - k_1$ degrees of freedom.⁵

⁴In the original formulation of Chao et al. (2001) the regressors Z_{2t} are not demeaned in the expression for c_t , as the statistic has been derived from a zero-mean data generating process. However, for non-zero mean data, demeaning of the regressors is required to achieve the correct limiting distribution.

⁵Chao et al. (2001) also propose a version of the test that takes into account estimation uncertainty, with $\hat{\Sigma}_{CCS}(m)$ replaced by a 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 turns out to be less powerful.

3 Monte Carlo evaluation for nested models

To evaluate the properties of the tests for nested models we start by considering the following VAR(1) data generating process (for $t = 1, 2, \dots, T$)

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

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

with Gaussian i.i.d. innovations

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

Note that, if $c \neq 0$, y_t can be represented as a Gaussian ARMA(2,1) process with 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 by 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.⁶ Given that the null hypothesis is $c = 0$, the tests can also be interpreted as out-of-sample tests of Granger causality.

We consider sample sizes of $T = R + P$ where $R = (100, 200)$ and $P = \pi R$ with $\pi = (.1, .25, .5, 1)$. The properties of the tests clearly depend on the number of out-of-sample observations P , with power expected to increase with π (for given 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 (9), and (10).

In the first subsection below we evaluate the properties of the tests under the case of 1-step ahead forecasts and correct specification (in the sense that the estimated unrestricted model is the same as the true data generating process). In the second subsection we study the effect of miss-specification and overparameterization, while the third one investigates the case of multi-step ahead predictions.

⁶For 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; see Clark and McCracken (2005a, p.376).

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 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 (2), (5), (6), and (8) for $m = 0$, i.e. with scaling provided by the sample variance instead of the long-run variance.

Table 1 provides the empirical sizes of the tests (case $c = 0$) run at 5% and 10% level of significance for $R = (100, 200)$. Results are shown for both recursive and rolling forecasts. We presents figures where the values of the parameters in the data generating process are set to $\phi_y = \phi_x = 0.8$, $q = 1$, $\rho_{\varepsilon u} = 0$. The size of the tests does not change in any significant way if different values of these parameters are considered.

Consider first the case $R = 100$ for recursive regressions with tests run at 10% level of significance. For $\pi = 0.1$ (10 out-of-sample observations) all tests except *DM* are oversized, in particular *MSE-t* (0.17), *ENC-t* (0.16) and *CCS* (0.15). As π increases 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% consistently with the arguments of Clark and West (2006, 2007) for the (equivalent) adjusted *DM* statistic. Doubling the sample ($R = 200$) yields more reliable sizes for all tests except, to some extent, *DM* and *HLN*. The figures for rolling and recursive regressions are nearly identical. Qualitatively similar arguments apply for the tests run at 5%.

Figure 1 shows the empirical power functions (with respect to the parameter c governing the distance from the null hypothesis) of tests⁷ run at 10% significance level for $R = 200$; results for $R = 100$ and for tests run at 5% significance are qualitatively similar and therefore will be not discussed. Power is affected by the parameter q that governs the variance of x_t (for given c , the higher q the more powerul the tests) and, to a lesser extent, by the value of the correlation $\rho_{\varepsilon u}$; however, as the relative ranking among the tests turns out to be unaffected by the values of q and $\rho_{\varepsilon u}$, to save space we only present results for $q = 1$ and $\rho_{\varepsilon u} = 0$. The four panels of figure 1 refer to different magnitudes of the prediction sample, $\pi = (0.1, 0.25, 0.5, 1)$; clearly, for fixed R , the larger π the more power.

For all values of π the *ENC - F* test of Clark and McCracken (2001) clearly turns out to be the most powerful one. The second ranked test depends on the value of π , the parameter governing the length of prediction sample relative to estimation sample. If the prediction sample is short

⁷The reported empirical power functions refer to the recursive case, but the results are very similar for the case of rolling regressions.

($\pi = 0.1$) then the $MSE - F$ test is preferable, otherwise $ENC - t$ is better. For large π the HLN test, that uses Gaussian critical values, behaves not so differently from $MSE - F$, and it is quite more powerful than $MSE - t$. The DM test has by far the lowest power, while the CCS test (that uses χ^2 critical values) has relatively good power only for large π , but it is always dominated by HLN .

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 latter has high rejection rates under the alternative hypothesis (second only to $ENC - F$). For higher π tests tend to behave more similarly: while $ENC - F$ clearly dominates, the HLN test may become attractive being based on Gaussian critical values. Detailed simulation results are available from the authors upon request.

To sum up, this sub-section extends the findings of Clark and McCracken (2001) by providing empirical power functions of the tests (while they only reported two specific values of the alternative) and by including HLN and CCS in the comparison. We confirm their findings that, overall, for nested models the $ENC - F$ test has the better properties, noticing however that its power advantage tends to become smaller as the dimension of the prediction sample increases. Furthermore, the relative ranking among the different tests changes according to whether the number of out-of-sample observations is "small" or "large".

3.2 Mis-specification of the nesting model

We consider three cases where the nesting model is somehow different from the true data generating process, so to understand to what extent the tests of equal MSPE and FE are still effective and whether some of them are more robust to misspecification.

(1) Autoregression. We take as 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 $ARMA(2, 1)$, it is plausible that an autoregressive model provides a reasonable approximation. The power loss from the misspecification, however, turns out to be very relevant; for example, if $c = 0.50$ and $\pi = 0.1$, the $ENC - F$ and CCS tests reject, respectively, 56% and 23% of the times, against 100% and 98% for the case of correct specification. The empirical power functions of DM , $MSE - F$, $ENC - F$, HLN , CCS are depicted in figure 2, for the case of recursive regression and $R = 200$. While the relative ranking among the tests remains mostly unaffected compared with figure 1, the power of the CCS test becomes now

very low (this could be related to the many nuisance parameters implicitly incorporated in the statistic, now distributed as χ_{p-1}^2 under the null, instead of χ_1^2)

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

$$w_t = x_t + u_{w,t}, \quad u_{w,t} \sim NIID(0, q_w^2 \sigma_x^2), \quad (11)$$

so that w_t and x_t are positively correlated with coefficient $\rho_{xw} = 1/(1 + q_w^2)$. Thus, for example, $\rho_{xw} = 0.5$ if $q_w^2 = 1$. All tests undergo a reduction of power with respect to the case of exact specification⁸, but again the relative ranking of the tests remains the same. Clearly, the higher ρ_{xw} , the smaller the power reduction.

(3) Over-parameterization. In the unrestricted model we take as predictors $(1, y_{t-1}, x_{t-1}, w_{t-1})'$ instead of $(1, y_{t-1}, x_{t-1})'$, where w_t is given by (11). Again the relative ranking is mostly unaffected. However, it turns out that, while $ENC - F$ is still the most powerful test, for HLN the power loss from over-parameterization is rather small; in fact now HLN becomes more attractive than $MSE - F$ for $\pi \geq 0.50$.

Detailed results for cases (2) and (3) are available upon request.

3.3 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 asymptotic approximation generally depends on many unknown 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 $ENC-t$ and $MSE-t$ asymptotic critical values 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. 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 computed as usual. The statistics have been calculated setting $m = 1.5h$ in the long run variance estimator (3). The bootstrap algorithm is that in Kilian (1998), as also implemented by Clark and McCracken (2005a). We

⁸For example, if $c = 0.10$ and $\pi = 1$, the $ENC - F$ and $MSE - F$ tests reject, respectively, 68% and 58% of the times, against 92% and 79% for the case of correct specification.

denote the bootstrap version of the tests by adding a * to the original name, e.g. $MSE - F^*$.

Table 2 provides the empirical sizes of the tests, run at 10% significance level, for the cases of $h = 2$ and 4 step ahead predictions, $R = (100, 200)$, and recursive regressions. For small π all tests (except CCS^*) appear to be oversized, with huge distortions for CCS , $MSE-t$ and $ENC-t$. Size generally improves as the number of out-of-sample observations increases, falling below the nominal significance level in the case of DM and HLN tests. If the number of out-of-sample observations is not too small, the good size properties of DM and HLN permit to carry out tests without bootstrapping; the question is again whether a price is paid in terms of power.⁹

Figure 3 provides the empirical power functions of the DM , $MSE-t^*$, $MSE-F^*$, HLN , CCS^* , $ENC-t^*$, $ENC-F^*$ tests for $\pi = (.25, .50)$ and $h = (2, 4)$. Again, $ENC - F^*$ turns out to be the most powerful test, with $MSE - F^*$ now unambiguously ranked second. Here CCS^* performs worst, followed by DM and $MSE - t^*$ which display similar rejection rates for $h = 4$. As for one step ahead predictions, we find larger differences in the behavior of the tests for smaller π . The HLN test appears to work relatively well, especially when $h = 4$. Thus, given the computational burden of bootstrapping the two best statistics, HLN appears to be a simple and good method for comparing forecast accuracy in multi-step ahead predictions.

4 Monte Carlo evaluation for non-nested models

The same data generating process of section 3 is used to evaluate the properties of the tests of equal mean square prediction error and of forecast encompassing for non-nested models. In particular, we consider the VAR(1) process (9), and (10) and the error-in-variables model (11). Let M_x denote the linear regression model of y_t on $(1, y_{t-1}, x_{t-1})'$ and M_w that of y_t on $(1, y_{t-1}, w_{t-1})'$. Then, if c and q_w are different from zero in the data generating process (9), (10), and (11), the two models M_x and M_w are non-nested. Of course, if $q_w = 0$ the two models and forecast errors are identical, while if $c = 0$ it is not possible to consistently discriminate between the correct model, M_x , and

⁹Contrary to the results of Clark and McCracken (2005a) we find that size distortions tend to vanish as the number of out-of-sample observations P increases. One difference, however, is that in their simulation experiment regressors 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.

the alternative regression, M_w since they both essentially boil down to the same.

The Monte Carlo simulations in this section aims at measuring the ability of the tests of MSPE and of FE towards rejecting the mis-specified model M_w in favor of M_x . The important parameter is the correlation $\rho_{xw} = 1/(1 + q_w^2)$ between the regressors x_t and w_t . If ρ_{xw} is nearly one (q_w small) the two models produce very similar forecasts; on the other hand, the smaller ρ_{xw} the better the tests are likely to discriminate between the models.

We consider the following tests: (i) the *DM* test of equal MSPE; (ii) the *HLN* test of FE (where the null hypothesis is that M_w encompasses M_x); (iii) the often used t-test of FE based on the regression of y_t on the forecasts from the two alternative models, which we denote as *FE-REG*; (iv) a *DM* test comparing M_w with a combined forecast that use predictions from both models, denoted as *DM-FC*.

For the *DM* test we present results for the test against the one-sided alternative that the model M_x is better (in the MSPE sense) than M_w , denoted as DM1, and also that against the two-sided alternative that either model is better, denoted as DM2. The main difference between *FE-REG* and *HLN* is that the former is a two-sided test. For the *DM-FC* we have used equal (50%) weights test for getting combined forecasts.

As in the previous section, we present results where the parameters of the data generating process are set as follows, $\phi_y = \phi_x = 0.8$, $q = 1$, $\rho_{\varepsilon u} = 0$ (and, without loss of generality, we set $\mu_y = \mu_x = 0$). For each ρ_{xw} , the rejection probabilities of the tests depend on the magnitude of c ; in the Monte Carlo simulations below (with $R = 200$) we have set $c = 0.2$, but qualitatively similar results would hold for other values of this parameter.

Figure 4 shows the rejection frequencies of the tests (against the correlation parameter ρ_{xw} taking decreasing values from 0.99 to 0) for $R = 200$ in-sample and $P = \pi R$ out-of-sample observations, where $\pi = (0.1, 0.25, 0.50, 1)$, for the case of one-step ahead predictions and tests run at 10% significance level. Consider the first panel of the figure, corresponding to the case of $\pi = 0.1$. For ρ_{xw} near 1 all tests display rejection frequencies very close to the nominal size, with the probability of rejection increasing as ρ_{xw} becomes smaller. Clearly, the one-sided tests DM1 and *HLN* are more powerful than the corresponding two-sided version, DM2 and *FE-REG*. Again, as for the case of nested models, the encompassing tests are significantly more powerful than the tests of equal MSPE: for example, for $\rho_{xw} = 0.5$ the simulated rejection probability of *HLN* is 49%, against 30% of DM1. It is interesting to see that the *DM-FC* test between the combined forecast and the mis-specified model M_w has properties similar to DM1.

Higher rejection probabilities characterize the remaining panels of figure

4, corresponding to a larger number of out-of-sample observations (50,100,200, respectively). While the ranking $HLN \succ FE - REG \succ DM1 \succ DM2$, applies for all π , for $\pi \geq 0.25$ (and ρ_{xw} not too big) $DM - FC$ seems preferable to $DM1$.¹⁰

Figure 5 shows the corresponding results for two and four steps ahead forecasts, $h = (2, 4)$, and $\pi = (0.25, 0.50)$, where the tests have been corrected by serial correlation by estimating long run variances with bandwidth parameter $m = 1.5h$, as done in the previous section. The ranking between the tests remains unchanged. Note that it is more difficult to reject when we move from $h = 2$ to $h = 4$ except when ρ_{xw} is very near zero (but this is probably related to the choice of the lag truncation parameter $m = 1.5h$, which may be too small for this case).

5 An empirical application

In this section we illustrate with two empirical applications the behavior of the tests discussed in the previous sections. In the first we look at the problem of forecasting euro area inflation, and compare the performance of a set of univariate and bivariate models, with the one of the nested simple *naive* model.

The forecasting model for euro area inflation has the following ARX structure:

$$\hat{\pi}_{t+1|t} = \hat{\alpha} + \hat{\beta}\pi_{t-1} + \hat{\gamma}x_t \quad (12)$$

where π_t denotes the monthly year-on-year inflation $\pi_t = 1200 * \log(P_t/P_{t-12})$, and P_t is the euro area Harmonized Consumer Price index.¹¹ The model forecasts inflation based on lagged inflation and another predictor variable x_t , defined as past deviations of inflation from a measure of core inflation, that is $x_t = (\pi_{t-12} - \pi_{t-12}^{core})$. Hence, as in Cogley (2005) we expect positive deviations (i.e., inflation above core inflation) to be followed by periods of decreasing inflation. We look at the forecasting performance of the model in (12) using different measures of core inflation, comparing their forecasts with the one from the *naive* benchmark. The latter forecast is obtained from (12) in the case $\alpha = 0$, $\gamma = 0$, and $\beta = 1$. We consider various so called exclusion measures of core inflation, which are computed as the inflation rate net of

¹⁰The advantage should however not be statistically significant. In fact, for $R = 200$, the DM test (run at 10% significance) of the combined forecast against the correct model M_x rejects the null hypothesis with probability around 0.11-0.14 for all values of ρ_{xw} .

¹¹The out of sample exercise runs from 1999.01 to 2007.12.

specific items in the CPI basket, as well as another core measure obtained as the 6-months growth rate in the seasonally adjusted price index (at an annual rate).

The MSFE of the various models compared to the *naive* benchmark are reported in panel A of Table 3. The results show that, as found for the US by Clark and West (2006), euro area inflation forecasts from the *naive* benchmark are hardly outperformed. In our case the relative MSFE is close but above unity for all but one of the measures of core inflation. Only the inflation rate excluding energy and food (labelled XEF in the table) achieves a relative MSFE of 0.784, and in this case all the tests considered reject the null of equal forecast accuracy with the benchmark *naive* model.

As to the other measures of core inflation no particular gain seems to be achieved when introducing an additional variable in the forecasting model (12). This outcome, in the light of the size results shown in the previous sections, may come as no surprise for the DM test, which is affected by strong under-sizing as first pointed out by Clark and McCracken (2001). Furthermore, those tests which are found to have greater power in the simulation section generally fail to reject the null of equal forecast accuracy. Only in the XSA model does the $ENC - F$ detect a rejection of the null hypothesis, albeit marginally significant (10%).

In the second example we compare non nested models, in the context of forecasting methods that exploit a large number of predictors. To this end, we replicate the forecasting exercise of a recent work by De Mol et al. (2008) on forecasting with a large number of predictors. In particular, they propose a new set of bayesian forecasting techniques, and compare their performance with that based on principal components regressions and other standard methods. The out-of-sample forecasting exercise assesses the forecasting accuracy for US inflation and industrial production of these alternative models and uses the full Stock and Watson (2005) monthly data set on US macroeconomic variables as the underlying information set. The forecasts of the target variable, $\hat{y}_{t+h|t}$, are calculated as:

$$\hat{y}_{t+h|t} = \hat{z}_{t+h|t} + y_t$$

where $\hat{y}_{t+h|t}$, is either the (log) IP or the the level of year-on-year inflation, and $\hat{z}_{t+h|t}$ is the forecast of the growth of the target variable over the forecast horizon h . We focus on the four methods described in De Mol et al. (2008):

- Principal components forecasts (PC): $\hat{z}_{t+h|t}$ is obtained using the first k normalized principal components of the data set.
- Ridge regression forecasts (RIDGE): $\hat{z}_{t+h|t}$ is obtained using a ridge-regression approach, where all the variables in the data set are weighted

according to a particular posterior mode estimator.

- Lasso forecasts (LASSO): $\hat{z}_{t+h|t}$ is obtained using a Lasso forecasting technique, which unlike PC and Ridge, gives zero weights to a large number of predictors, thereby favoring sparse regression coefficients. Furthermore, in the Lasso regression the selection and shrinkage of the variables in the panel depends on the choice of the target variable, while in PC and Ridge regressions that choice is independent of the series to be forecasted.
- Lars forecasts (LARS): $\hat{z}_{t+h|t}$ is obtained using the so called LARS (Least Angle Regression), an efficient alternative to the LASSO algorithm.

The accuracy of the predictions obtained by the various models is evaluated using the mean-square forecast error (MSFE) metric. We focus on the out of sample results from a *rolling scheme* exercise, where the sample for evaluation is 1970:01 to 2002:12, and the forecast horizon is $h = 12$. We compare the forecasting performance of RIDGE, LASSO and LARS to the one of PC forecasts. This is a non-nested model comparison, hence as in section (4) we report the results from the DM and HLN test, along with the ones from the FE-REG and DM-FC.

Panel B of Table 3 reports the MSFE relative to the PC forecasts, for industrial production and inflation. According to this metric, for industrial production the PC forecasts perform well compared to the other methods. Instead, in the case of inflation, PC forecasts are generally underperforming, but these differences are not statistically significant.

In the case of industrial production, while the DM cannot reject the null neither at the 10% nor at the 5% level, the HLN test indicates that PC forecasts encompass the Ridge ones. In contrast, when comparing the PC to the LARS and LASSO methods we find that while the DM test cannot reject, the HLN test indicates the potential improvement of using either LARS or LASSO methods along with the PC forecasts. In sum, PC forecasts do not encompass the forecasts from LARS or LASSO.

Overall, these results are consistent with those from our Monte Carlo simulations.

6 Concluding remarks

The performance of several tests for comparing out-of-sample forecasts between competing models has been evaluated. Overall, the tests of forecast

encompassing seem preferable to those of equal mean square prediction error. In particular, for nested models the best properties are displayed by the $ENC - F$ test of Clark and McCracken (2001); its power advantage however tends to become smaller for longer prediction samples, for which a standard forecast encompassing test, with Gaussian critical values, may become relatively more attractive. Moreover, as the issue surrounding the standard tests in nested comparisons is that of under sizing and low power, it is clear that if these reject the null hypothesis then the use of the $ENC - F$ or similar sophisticated techniques may become superfluous.

The simulation results presented, however, do not account either for structural breaks or for model uncertainty. In fact, Clark and McCracken (2004, 2005b) show that breaks significantly affect the properties of tests of predictive ability and thus they may render harder the task of discriminating between competing models, and they also argue that the choice of estimation window in rolling regressions becomes crucial given the bias variance tradeoff induced by parameter instability. Finally, as real world forecasts can never be generated by the underlying "true model", we believe that taking into account mis-specification and model uncertainty is an important direction for future research.

References

- [1] Chao, J., Corradi V., and Swanson N., (2001), “An out of sample test for Granger causality”, *Macroeconomic Dynamics*, 5, 598-620.
- [2] Chong, Y. Y., and Hendry D.F., (1986), “Econometric evaluation of linear macroeconomic models”, *Review of Economic Studies*, 53, 671-690.
- [3] Clark, T.E., and McCracken, M.W., (2001), “Tests of equal forecast accuracy and encompassing for nested models”, *Journal of Econometrics*, 105, 85-110.
- [4] Clark, T.E., McCracken, M.W. (2004), Forecast Accuracy and the Choice of Observation Window, *mimeo*.
- [5] Clark, T.E., and McCracken, M.W., (2005a), “Evaluating direct multistep forecasts”, *Econometric Reviews*, 24, 369-404, doi:10.1080/07474930500405683.
- [6] Clark, T.E., and McCracken, M.W., (2005b), “The power of tests of predictive ability in the presence of structural breaks”, *Journal of Econometrics*, 124, 1-31, doi:10.1016/j.jeconom.2003.12.011.
- [7] Clark, T.E., and West, K.D., (2006), “Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis”, *Journal of Econometrics*, 135, 155-186, doi:10.1016/j.jeconom.2005.07.014.
- [8] Clark, T.E., and West, K.D., (2007), “Approximately normal tests for equal predictive accuracy in nested models”, *Journal of Econometrics*, 138, 291-311, doi:10.1016/j.jeconom.2006.05.023.
- [9] Clements, M.P., and Hendry, D.F., (1993), “On the limitations of comparing mean square forecast errors” (with discussion), *Journal of Forecasting*, 12, 617-676.
- [10] Cogley, T., (2002). “A Simple Adaptive Measure of Core Inflation”, *Journal of Money, Credit and Banking*, 34, 94-113.
- [11] De Mol, C., Giannone D., and Reichlin, L., (2008), “Forecasting Using a Large Number of Predictors: Is Bayesian Regression a Valid Alternative to Principal Components?”, *forthcoming* in the *Journal of Econometrics*.

- [12] Diebold, F.X., and Mariano, R.S., (1995), “Comparing Predictive Accuracy”, *Journal of Business & Economic Statistics*, 13, 253-263.
- [13] 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.
- [14] Giacomini, R., and White, H., (2006), “Tests of conditional predictive ability”, *Econometrica*, 74, 1545-1578.
- [15] Granger, C.W.J., and Newbold, P., (1986), *Forecasting economic times series* (2nd ed.), New York: Academic Press.
- [16] Hansen, P.R., (2005), “A Test for superior predictive ability”, *Journal of Business & Economic Statistics*, 23, 365-380.
- [17] Harvey, D.I., Leybourne, S.J., and Newbold P., (1998), “Tests of forecast encompassing”, *Journal of Business & Economic Statistics*, 16, 254-259.
- [18] Hubrich K., and West K.D., (2007), “Forecast Evaluation of small nested model sets”, European Central Bank, *mimeo*.
- [19] Kilian, L. (1998), Small-sample confidence intervals for impulse response functions, *Review of Economics and Statistics*, 80, 218–230.
- [20] McCracken, M.W., (2007), “Asymptotics for out of sample tests of Granger causality”, *Journal of Econometrics*, 140, 719-752, doi:10.1016/j.jeconom.2006.07.020.
- [21] Nelson, C. R., (1972), “The Prediction Performance of the FRB-MIT-PENN Model of the U.S. Economy”, *American Economic Review*, 62, 902-917.
- [22] Newey, W.K., and West K.D., (1987), “A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix”, *Econometrica*, 55, 703-708.
- [23] Stock, J.H., and Watson, M.W., (2005), “Implications of Dynamic Factor Models for VAR Analysis”, NBER Working paper 11467, National Bureau of Economic Research, Inc.
- [24] West, K.D., (1996), “Asymptotic inference about predictive ability” *Econometrica*, 64, 1067-1084.

- [25] West, K.D., (2006), “Forecast Evaluation”, 100-134, in Handbook of Economic Forecasting, Vol. 1, G. Elliott, C.W.J. Granger and A. Timmerman (eds), Amsterdam: Elsevier.
- [26] White, H. (2000). “A reality check for data snooping”, *Econometrica*, 68, 1097-1126.

Table 1: Empirical size of the tests of equal forecast accuracy for one-step ahead forecasts run at nominal 5 and 10% (Nested case).

	$R = 100$				$R = 200$			
π	0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00
(A) Recursive 5%								
<i>DM</i>	0.06	0.03	0.02	0.01	0.04	0.02	0.01	0.01
<i>MSE - t</i>	0.11	0.09	0.07	0.06	0.08	0.08	0.06	0.06
<i>MSE - F</i>	0.07	0.07	0.06	0.06	0.06	0.06	0.05	0.05
<i>HLN</i>	0.07	0.05	0.04	0.03	0.05	0.04	0.04	0.03
<i>ENC - t</i>	0.10	0.08	0.07	0.06	0.08	0.07	0.06	0.06
<i>ENC - F</i>	0.08	0.07	0.07	0.06	0.07	0.06	0.06	0.06
<i>CCS</i>	0.09	0.07	0.07	0.06	0.07	0.06	0.06	0.06
(B) Rolling 5%								
<i>DM</i>	0.06	0.03	0.02	0.01	0.04	0.02	0.01	0.00
<i>MSE - t</i>	0.11	0.09	0.07	0.06	0.08	0.08	0.06	0.06
<i>MSE - F</i>	0.07	0.07	0.06	0.06	0.06	0.06	0.06	0.05
<i>HLN</i>	0.07	0.05	0.04	0.04	0.05	0.04	0.04	0.03
<i>ENC - t</i>	0.10	0.08	0.07	0.06	0.08	0.07	0.06	0.06
<i>ENC - F</i>	0.08	0.07	0.07	0.06	0.07	0.06	0.06	0.06
<i>CCS</i>	0.09	0.07	0.07	0.07	0.07	0.06	0.06	0.06
(C) Recursive 10%								
<i>DM</i>	0.10	0.06	0.04	0.02	0.08	0.05	0.03	0.02
<i>MSE - t</i>	0.17	0.14	0.12	0.11	0.14	0.13	0.11	0.11
<i>MSE - F</i>	0.12	0.11	0.11	0.10	0.11	0.10	0.10	0.10
<i>HLN</i>	0.12	0.09	0.08	0.07	0.10	0.08	0.07	0.06
<i>ENC - t</i>	0.16	0.14	0.12	0.11	0.14	0.13	0.11	0.11
<i>ENC - F</i>	0.13	0.12	0.12	0.11	0.12	0.11	0.11	0.11
<i>CCS</i>	0.15	0.13	0.12	0.12	0.13	0.11	0.11	0.11
(D) Rolling 10%								
<i>DM</i>	0.10	0.06	0.03	0.01	0.08	0.05	0.03	0.01
<i>MSE - t</i>	0.17	0.15	0.12	0.11	0.14	0.13	0.11	0.11
<i>MSE - F</i>	0.12	0.12	0.11	0.11	0.11	0.11	0.10	0.11
<i>HLN</i>	0.12	0.09	0.08	0.07	0.10	0.08	0.07	0.06
<i>ENC - t</i>	0.16	0.14	0.13	0.12	0.14	0.13	0.11	0.11
<i>ENC - F</i>	0.13	0.12	0.12	0.12	0.12	0.11	0.11	0.11
<i>CCS</i>	0.15	0.13	0.13	0.13	0.13	0.12	0.11	0.12

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

Table 2: Empirical size of the tests of equal forecast accuracy for multi-step ahead forecasts run at nominal 10% (Nested case).

	$R = 100$				$R = 200$			
π	0.10	0.25	0.50	1.00	0.10	0.25	0.50	1.00
(A) Recursive $h = 2, 10\%$								
<i>DM</i>	0.18	0.10	0.06	0.04	0.13	0.07	0.05	0.03
<i>MSE - t</i>	0.25	0.18	0.15	0.16	0.19	0.15	0.13	0.12
<i>MSE - t*</i>	0.11	0.10	0.09	0.10	0.11	0.10	0.09	0.10
<i>MSE - F*</i>	0.10	0.09	0.10	0.11	0.11	0.10	0.09	0.11
<i>HLN</i>	0.19	0.12	0.09	0.08	0.14	0.09	0.08	0.08
<i>ENC - t</i>	0.23	0.17	0.13	0.13	0.18	0.14	0.11	0.13
<i>ENC - t*</i>	0.11	0.10	0.10	0.10	0.11	0.10	0.10	0.10
<i>ENC - F*</i>	0.10	0.10	0.10	0.11	0.11	0.09	0.10	0.12
<i>CCS</i>	0.32	0.18	0.14	0.10	0.21	0.13	0.10	0.12
<i>CCS*</i>	0.10	0.09	0.10	0.10	0.10	0.09	0.09	0.09
(B) Recursive $h = 4, 10\%$								
<i>DM</i>	0.28	0.15	0.10	0.07	0.18	0.10	0.07	0.03
<i>MSE - t</i>	0.33	0.24	0.20	0.21	0.25	0.20	0.18	0.14
<i>MSE - t*</i>	0.10	0.11	0.10	0.10	0.11	0.10	0.10	0.11
<i>MSE - F*</i>	0.10	0.10	0.10	0.10	0.11	0.10	0.10	0.12
<i>HLN</i>	0.28	0.16	0.12	0.10	0.18	0.12	0.09	0.09
<i>ENC - t</i>	0.32	0.22	0.16	0.16	0.23	0.17	0.14	0.15
<i>ENC - t*</i>	0.10	0.11	0.10	0.10	0.10	0.10	0.10	0.11
<i>ENC - F*</i>	0.10	0.10	0.10	0.11	0.11	0.10	0.10	0.12
<i>CCS</i>	0.43	0.27	0.19	0.13	0.30	0.18	0.13	0.14
<i>CCS*</i>	0.10	0.10	0.10	0.10	0.09	0.10	0.10	0.10

Notes: Results from 5,000 Monte Carlo iterations. h -step ahead forecasts computed with the direct method with recursive and rolling schemes. Forecast horizons: $h = (2, 4)$. In-sample sizes: $R = (100, 200)$. Starred (*) tests are bootstrapped.

Table 3: Empirical application: forecasting inflation and industrial production (Nested and non-nested case).

Panel (A)

Forecasting Euro area inflation one-month ahead

Nested case: random walk benchmark

Relative MSFE	AR			ARX		
		AR+ XE	AR+ XES	AR+ XEU	AR+ XSA	AR+ XEF
	1.038	1.046	1.075	1.075	1.063	0.7837
DM	-	-	-	-	-	**
HLN	-	-	-	-	-	**
Enc-t	-	-	-	-	-	**
Enc-F	-	-	-	-	*	**
Mse-t	-	-	-	-	-	**
Mse-F	-	-	-	-	-	**
CCS	-	-	-	-	-	**

Panel (B)

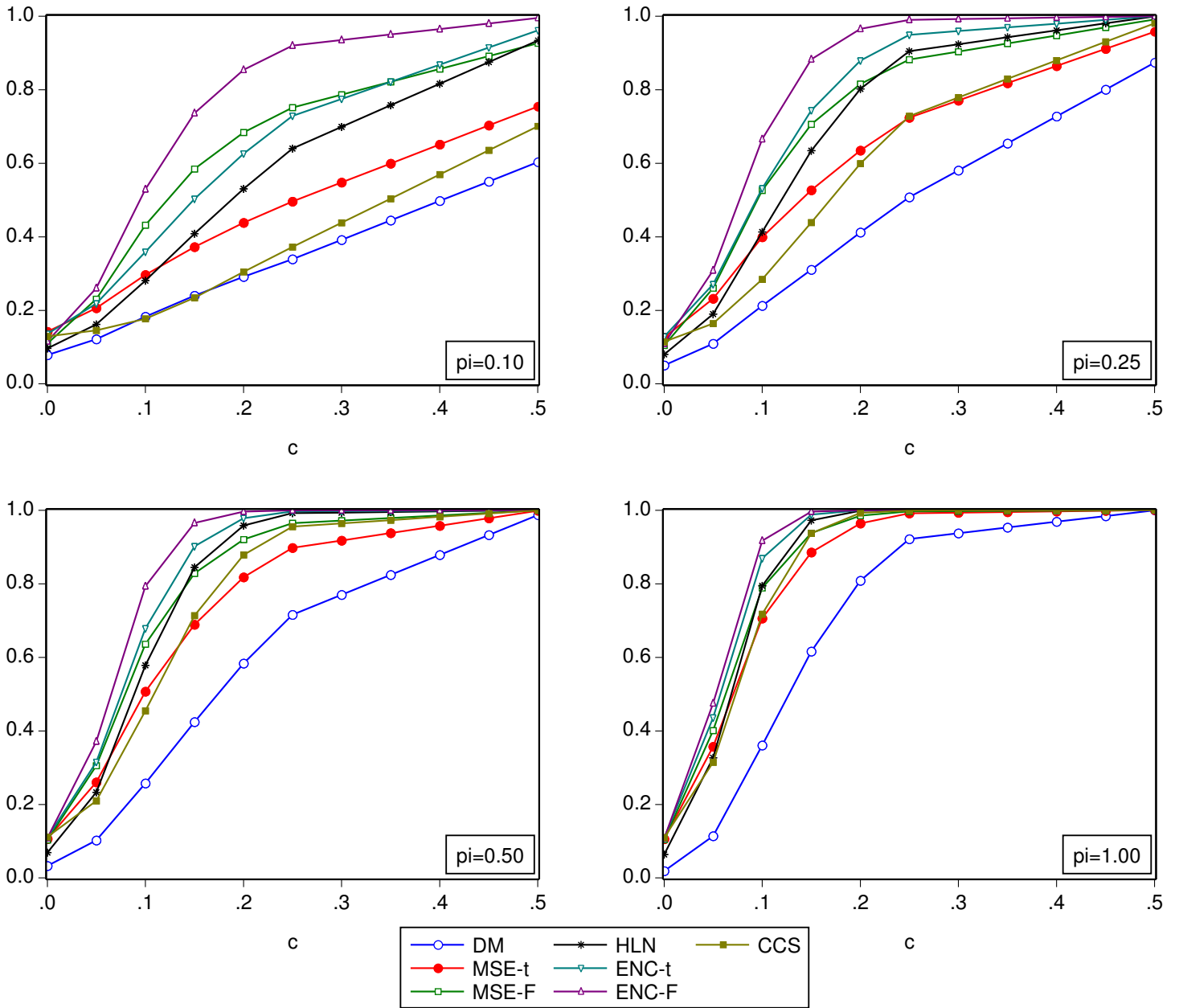
Forecasting US Industrial production and Inflation 12-months ahead with many predictors
(as in De Mol et al., 2008)

Non nested case: Principal Component benchmark

Relative MSFE	Ridge		Lars		Lasso	
	Ind.Prod.	Inflation	Ind.Prod.	Inflation	Ind.Prod.	Inflation
	1.031	0.888	1.106	0.855	1.081	0.912
DM	-	-	-	-	-	-
HLN	-	-	**	-	**	-
FE-REG	-	-	-	-	-	-
DM-FC	-	-	-	-	-	-

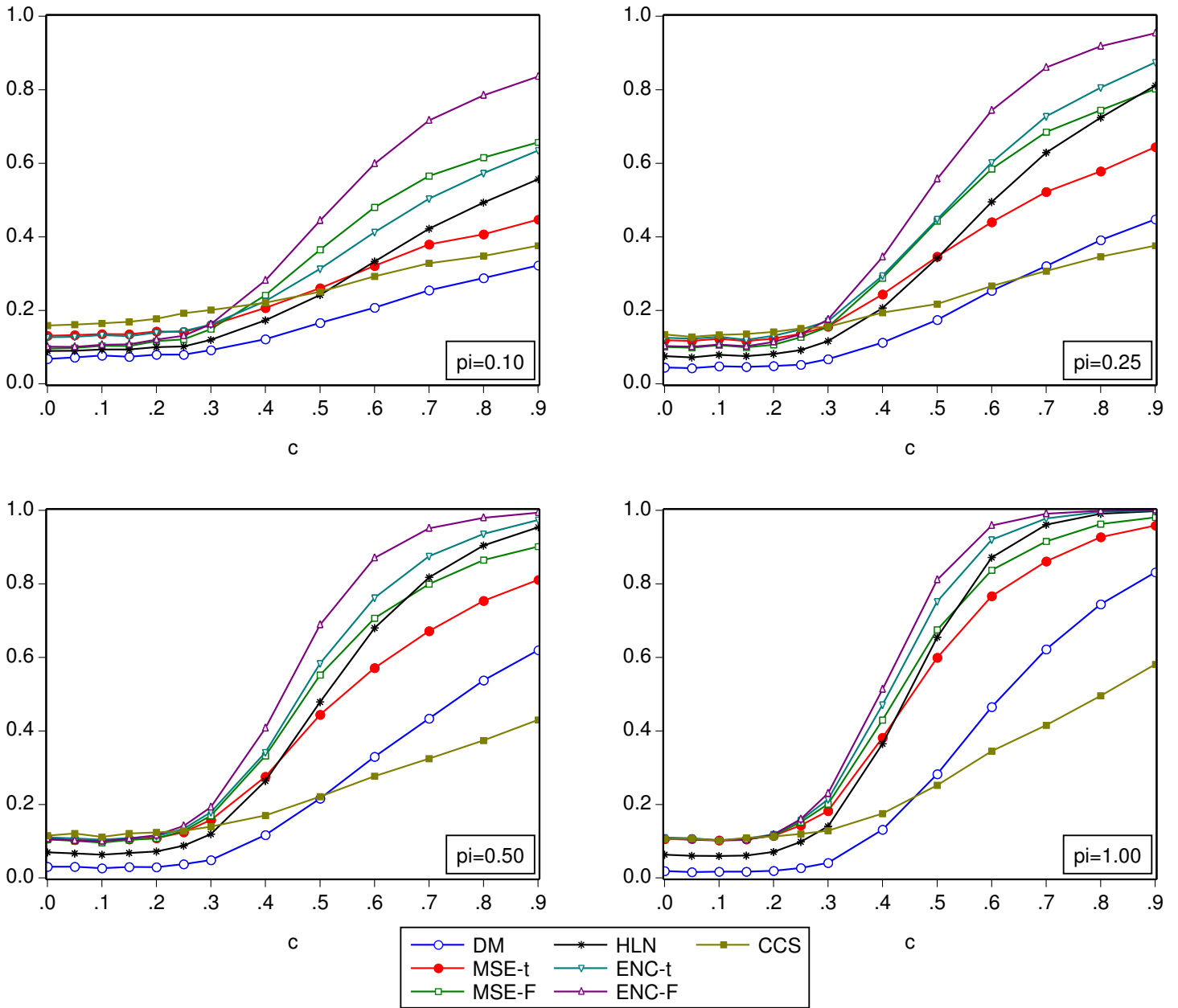
Notes: Panel (A) shows the results of different tests of equal forecast accuracy and FE at one-step ahead for the year-on-year (yoy) euro area inflation. The benchmark model is a random walk and the alternative models are AR(1) models with (or without) an adjustment factor. This factor is the past deviation of inflation from different measures of the core inflation. The core inflation measures are: excluding energy (XE), excluding energy and seasonal food (XES), excluding energy and unprocessed food (XEU), seasonally adjusted inflation (XSA) and excluding energy and food (XEF). Panel (B) reports the results of different tests of equal forecast accuracy and forecast encompassing at 12-steps ahead for US inflation and industrial production from the data set of Stock and Watson (2005) according to the estimates by De Mol et al. (2008). Here the model that uses principal components to predict is the benchmark. In both panels *, ** and - indicate rejection at 10%, at 5% and no rejection, respectively.

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



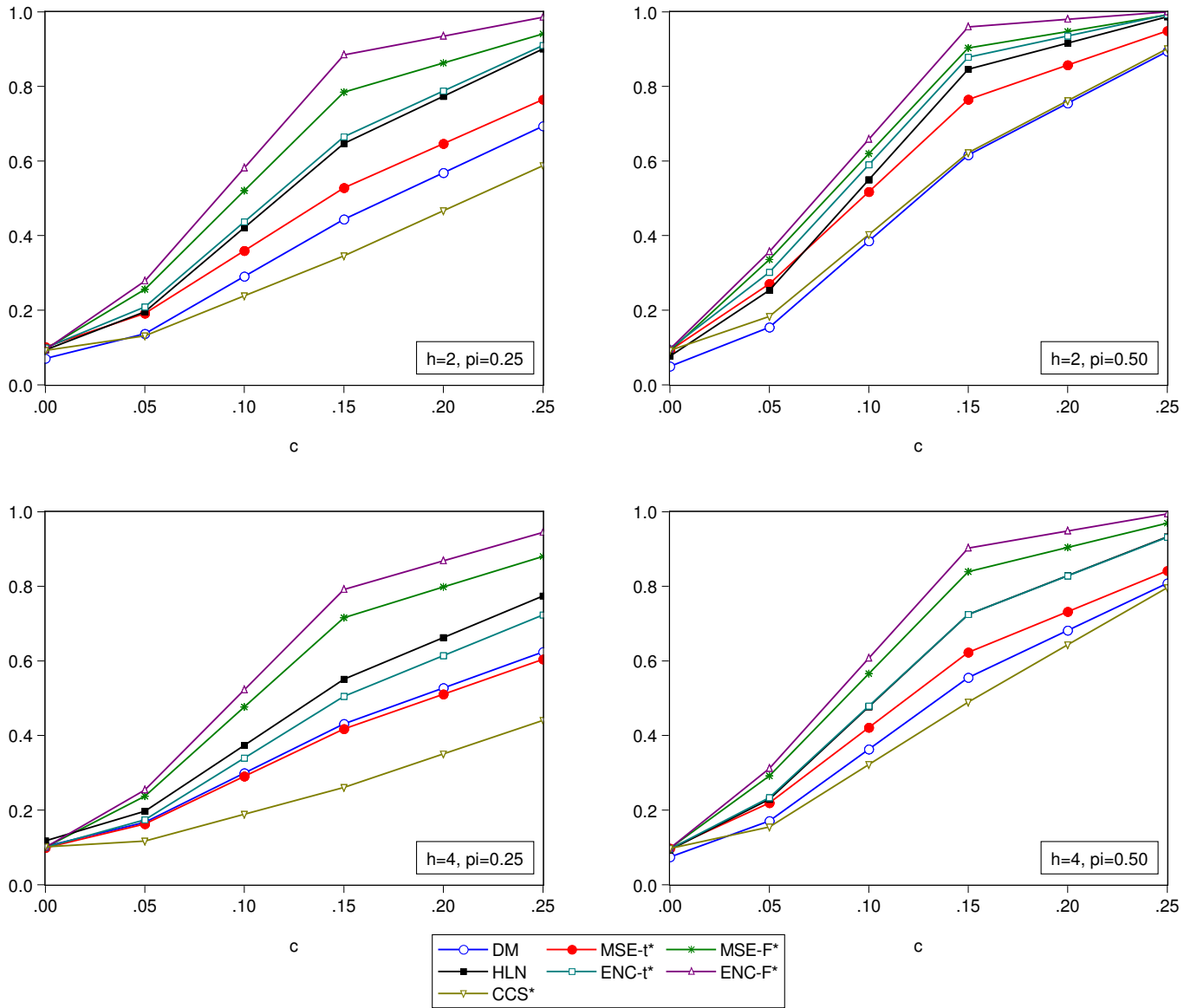
Notes: Results from 50,000 Monte Carlo simulations of one-step ahead forecasts. Recursive scheme with in-sample $R = 200$.

Figure 2: Empirical power functions for the case of one-step ahead forecasts under misspecification ($R=200$, recursive regressions)



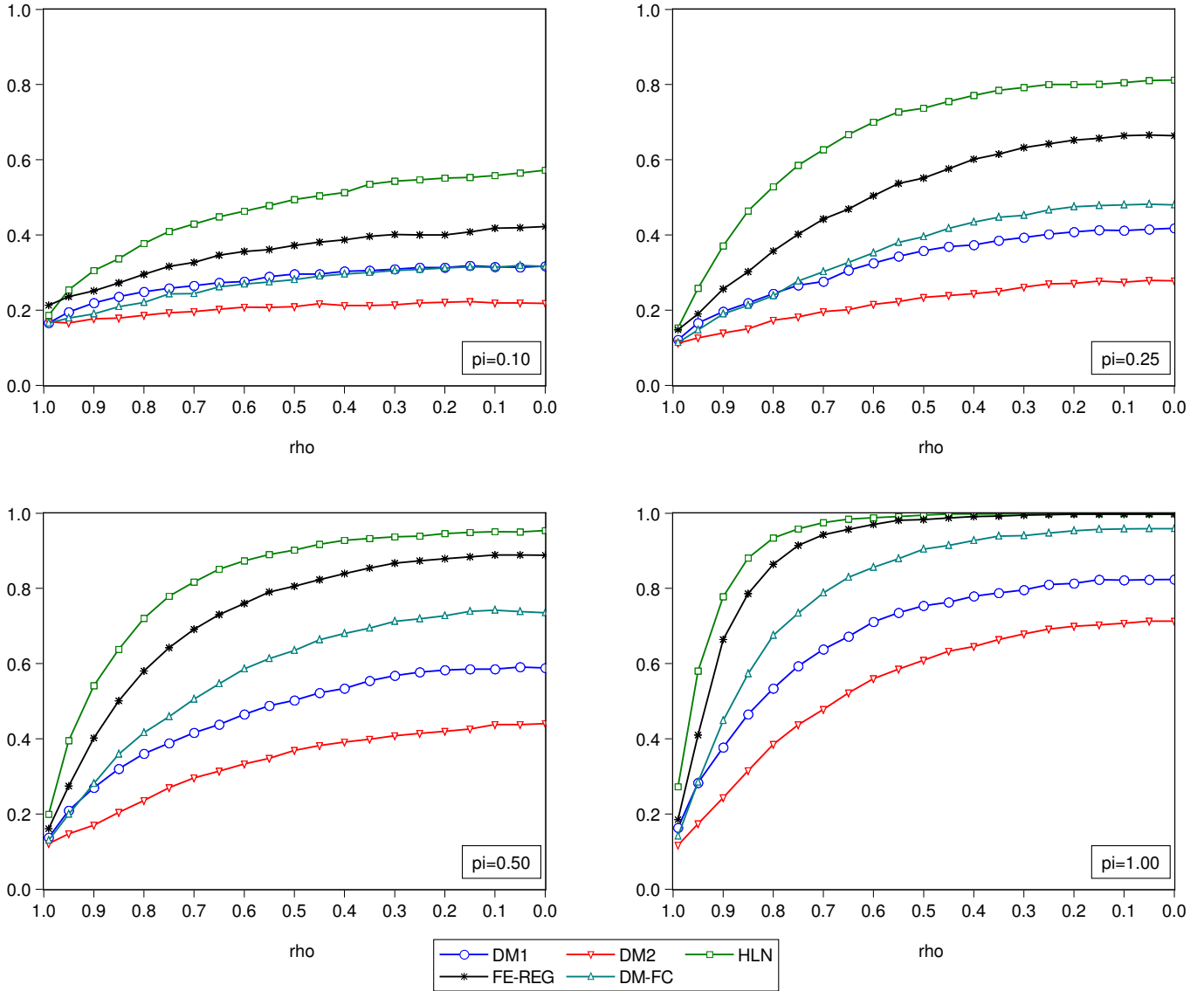
Notes: Results from 50,000 Monte Carlo simulations of one-step ahead forecasts. Recursive scheme with in-sample $R = 200$.

Figure 3: Empirical power functions for the case of multi-step ahead forecasts under correct specification ($R=200$, recursive regressions)



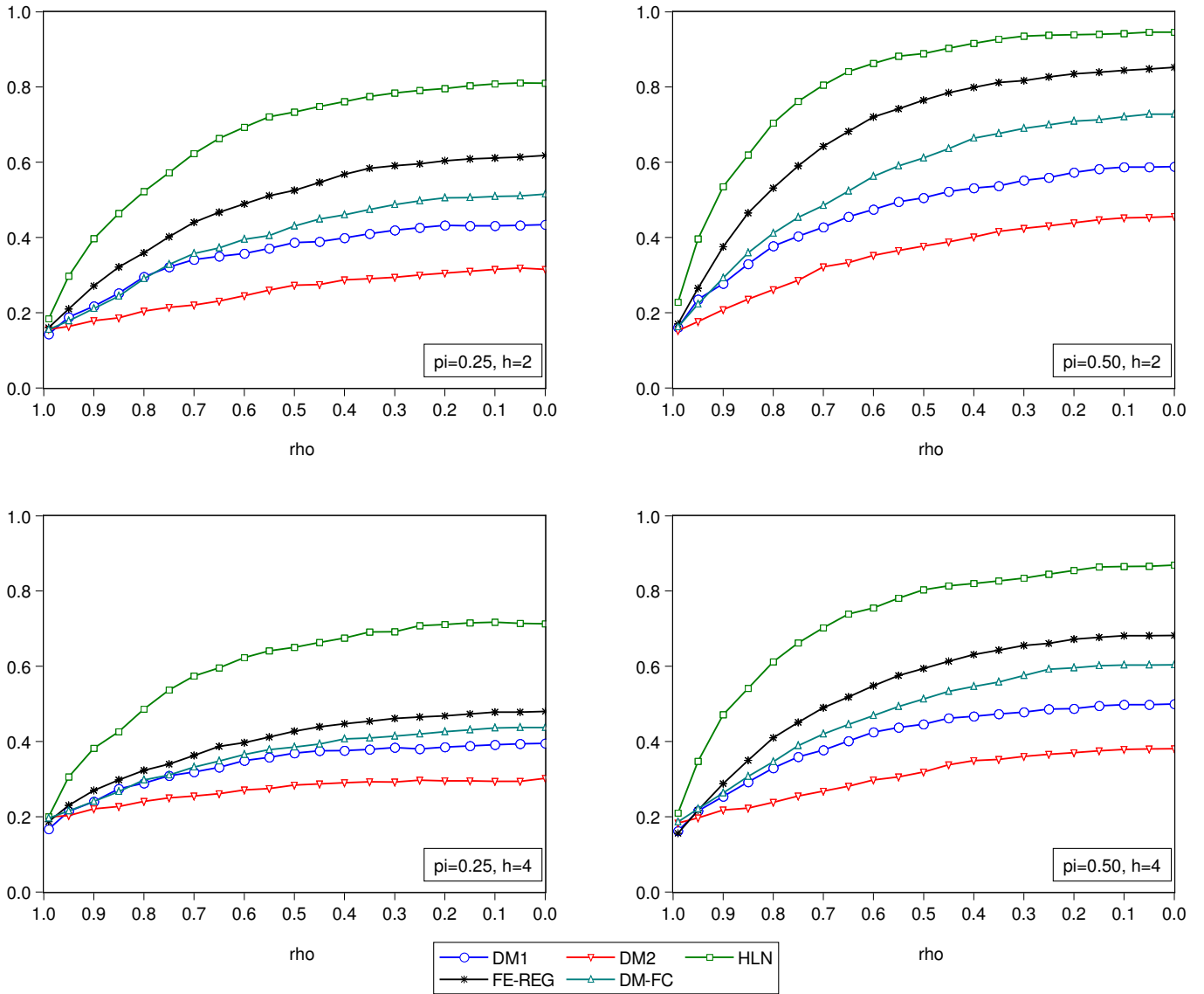
Notes: Results from 50,000 Monte Carlo simulations of multi-step ahead forecasts. Recursive scheme with in-sample $R = 200$.

Figure 4: Empirical power functions for the non-nested case of one-step ahead forecasts (R=200, recursive regressions)



Notes: Results from 10,000 Monte Carlo simulations of one-step ahead forecasts. Recursive scheme with in-sample $R = 200$. Non-nested case.

Figure 5: Empirical power functions for the non-nested case of multi-step ahead forecasts (R=200, recursive regressions)



Notes: Results from 10,000 Monte Carlo simulations of h -step ahead forecasts. Recursive scheme with in-sample $R = 200$ and $h = (2, 4)$. Non-nested case.