Finite sample properties of the Conditional Predictive Ability test

Jan Acedański[1]

Abstract

In the paper we analyze finite sample properties of the conditional predictive ability test proposed by Giacomini and White [9]. The test is designed for comparison of the out-of-sample forecast accuracy of two competing models. We simulate various two-dimensional series of forecast errors and calculate the empirical power and size of the test in both conditional and unconditional version for few different sample lengths. We find that the test has very appealing properties as far as the forecast errors of the two models are highly correlated. Otherwise for moderate sample lengths it has high power only when the forecast errors differ significantly in terms of the unconditional standard deviation.

Keywords: *predictive ability tests, forecasting, mean square error*

JEL Classification: C53, C12
AMS Classification: 62G10, 60G25

1. Introduction

Predictive ability tests are used for comparing ex post accuracy of series of forecasts generated by two competing models. In other words they are designed to answer the question whether the difference in forecast accuracy measured by some loss function observed in a sample can be attributed to a pure chance or it will likely occur also out of the sample. The first such test was proposed by Diebold and Mariano [8]. Now the Diebold-Mariano test and its few important modifications are commonly applied in literature for comparing model predictive abilities [see 6, 11].

In the paper we analyze the power and the size of the Conditional Predictive Ability (CPA) test which is an important generalization of the Diebold-Mariano test. We utilize the Monte Carlo approach and generate realizations of few two-dimensional stochastic processes representing forecast errors from two models that differ in terms of mean square errors. Then we count the number of cases where the hypothesis of equal predictive abilities was rejected. In the paper we focus on short series of errors since this is usually the case when one works with Polish data. The finite sample properties of the CPA test were examined only scarcely by Giacomini and White [9]. The properties of some other predictive ability tests were analyzed by Clark [3], Clark, McCracken [5] and Busetti, Marcucci [2] to name a few.

[1] University of Economics in Katowice, Department of Economics, 1-go Maja 50, 40-136 Katowice, Poland, jan.acedanski@ue.katowice.pl

The paper is organized as follow. In the first part we introduce the Diebold-Mariano and Giacomini-White predictive ability tests. Then we describe the simulation exercises. Finally we present the results.

2. Methodology - predictive ability tests

Let $\hat{Y}_{t+\tau}(\beta_1)$ and $\hat{Y}_{t+\tau}(\beta_2)$ denote forecasts for the horizon $t + \tau$ generated in period $t = 1, 2, \ldots, n$ from two models represented by parameter vectors β_1 and β_2. Moreover let $\hat{\beta}_{i,mt}$ stands for parameter estimates based on a sample of length m_t. Forecasts accuracy is measured by a loss function $L_{t+\tau}(Y_{t+\tau}, \hat{Y}_{t+\tau})$. By $\Delta L_{t+\tau,mt} = L_{t+\tau}(Y_{t+\tau}, \hat{Y}_{t+\tau}(\hat{\beta}_{1,mt})) - L_{t+\tau}(Y_{t+\tau}, \hat{Y}_{t+\tau}(\hat{\beta}_{2,mt}))$ we denote a difference between forecast loss functions for two models that parameters are estimated on a sample with m_t observations. The test proposed by Diebold and Mariano [8] has the null hypothesis of the form:

$$H_0 : E(\Delta L_{t+\tau,mt}) = 0, \quad t = 1, 2, \ldots, \tag{1}$$

and its alternative is:

$$H_1 : E(\Delta L_{t+\tau,mt}) \neq 0, \quad t = 1, 2, \ldots. \tag{2}$$

E stands for expectation operator. The null hypothesis states that on average there are no differences between predictive abilities of two models measured by a loss function. It should be noted that the hypothesis refers to the estimates $\hat{\beta}_{1,mt}$ and $\hat{\beta}_{2,mt}$ but not to the true values β_1 and β_2. Therefore this is so called finite sample level prediction ability test [6]. The test made no assumptions about both estimation method and samples length.

The test statistic has the standard zero-mean form:

$$Z = \frac{\Delta\bar{L}_{mt,\tau}}{\hat{\sigma}(\Delta\bar{L}_{mt,\tau})}\sqrt{n}, \tag{3}$$

where $\Delta\bar{L}_{mt,\tau} = \frac{1}{n}\sum_{t=1}^{n}\Delta\bar{L}_{t+\tau,mt}$, $\hat{\sigma}(\Delta\bar{L}_{mt,\tau})$ is an estimator of standard deviation of $\Delta\bar{L}_{mt,\tau}$ and *n* is a number of ex post forecasts. Since the series of $\Delta\bar{L}_{mt,\tau}$ may be autocorrelated the HAC-type estimators (see Newey, West [10], Andrews [1]) of $\hat{\sigma}(\Delta\bar{L}_{mt,\tau})$ are suggested. The limiting distribution of the test statistic is standard normal.

Generalization of the DM test was proposed by Giacomini and White [9]. The null hypothesis in that test has the following form:

$$H_0 : E(\Delta L_{t+\tau,m} \mid \Psi_t) = 0, \quad t = 1, 2, \ldots, \tag{4}$$

where Ψ_t represents a set of additional information available in period *t*. The null differs from its DM counterpart in two ways. First the expectation is conditional, so additional information can be taken into account for comparing forecasts. Therefore the test can answer the question whether differences in forecast ability depend on business cycle phase or other factors. And secondly it is assumed that models are estimated on samples of constant size *m* which need not be the same for both models. This assumption excludes models estimated on expanding window samples.

The test statistic has also the standard multivariate zero-mean form:

$$\chi_q^2 = nZ'_{m,n}\hat{\Omega}_Z^{-1}Z_{m,n}, \tag{5}$$

where $Z_{m,n} = \left[Z_{1,m,n}, Z_{2,m,n}, \ldots, Z_{q,m,n}\right]$, $Z_{i,t,m,n} = X_{i,t}\Delta L_{t+\tau,m}$, and $X_{i,t}$ denotes value of *i*-th instrumental variable in period *t*. Moreover $\hat{\Omega}_Z$ denotes covariance estimator of matrix $Z_{m,n}$.

If the horizon $\tau > 2$ the authors suggest using HAC estimators. Giacomini and White [9] showed that under some mild assumptions on data used for estimating the models the test statistic converges asymptotically to χ^2 distribution with *q* degrees of freedom. Therefore the test is right-sided. If there is no additional information the test statistic collapses to (3). In our analysis we used both unconditional and conditional version of the test.

We should also mention that there is another approach to testing predictive abilities based on a population level view. In this approach the hypotheses refer to the true values of parameters β_1 and β_2, which generally makes the testing procedure significantly more complicated. Such tests are considered for example by West [12] and Clark and McCracken [4, 7] among others.

3. Simulation study

The properties of the test were examined using simulated series of forecast errors from two models. We considered several data generating processes. In every variant we generated two series that differ in terms of unconditional standard deviation. The series consist of

n = 30, 60, 90 and 120 observations. Below we describe the data generating processes that we worked with.

- P1a: $u_{1t} = \varepsilon_{1t}$, $u_{2t} = \sigma_2 \varepsilon_{2t}$, $\varepsilon_{1t}, \varepsilon_{2t} \sim N(0,1)$, $\sigma_2 = 1, 1.05, \ldots, 3$;
- P1b: $u_{1t} = \frac{\rho_{12}}{\sigma_2} u_{2t} + \sqrt{1-\rho_{12}^2}\varepsilon_{1t}$, $u_{2t} = \sigma_2 \varepsilon_{2t}$, $\varepsilon_{1t}, \varepsilon_{2t} \sim N(0,1)$, $\rho_{12} = 0.95$;
- P2a: $u_{1t} = \varepsilon_{1t}$, $u_{2t} = \gamma u_{2t-1} + \varepsilon_{2t}$, $\varepsilon_{1t}, \varepsilon_{2t} \sim N(0,1)$, $\gamma = 0, 0.05, \ldots, 0.9$;
- P2b: $u_{1t} = \rho_{12}\sqrt{1-\gamma^2} u_{2t} + \sqrt{1-\rho_{12}^2}\varepsilon_{1t}$, $u_{2t} = \gamma u_{2t-1} + \varepsilon_{2t}$, $\varepsilon_{1t}, \varepsilon_{2t} \sim N(0,1)$, $\gamma = 0, 0.05, \ldots, 0.9$, $\rho_{12} = 0.95$.

As far as W1 processes are concerned both series are normal and serially uncorrelated. In version W1a the series are mutually independent whereas the version W1b assumes that the correlation coefficient between the series equals 0.95. The first process has always unitary unconditional standard deviation. The standard deviation of the second one takes values ranging from 1 to 3. The processes W2 differ from W1 by the fact that the second series has constant unitary conditional variance but are increasingly serially correlated.

For estimating the covariance matrices in the test the HAC estimators with Bartlett kernel were used, where the automatic procedure proposed by Andrews [1] were utilized for calculating the optimal window lag length. In every variant 10000 series were simulated for calculating the power and 100000 for calculating the size of the tests.

4. Results

The results of the power analysis were presented for two nominal significance levels α = 0.1 and α = 0.01. Figure 1 depictures relationship between the power of the unconditional version of the test and the unconditional standard deviation of the second variable s_2 for the process P1a and two nominal significance levels. For α = 0.1 and longer samples n = 90 or n = 120 the rejection probability is close to 1 as far as the unconditional standard deviation of the second error process is about 50% higher than for the first one. The results change only slightly if one considers lower significance level α = 0.01. However if shorter samples are taken into account this difference needs to be significantly higher. For example if the error series have only 30 observations then the test achieves high power if the standard deviation of the second model reaches about 100% (s_2 = 2) of the first standard deviation for α = 0.1 and about 200% (s_2 = 3) when α = 0.01.

The problem with the low power of the test does not occur when the two error series are strongly mutually dependent as it is presented on figure 2 for the process P1b. For the nominal

significance level $\alpha = 0.1$ and longer samples the test achieves the high power just for $s_2 = 1.1$. For $n = 30$ observations and = 1.25 the test properly rejects the null in almost all simulations if only s_2 exceeds 1.25. If $\alpha = 0.01$ these differences need to be about two times higher but they are still lower than for the uncorrelated series. We do not present the results for the conditional version of the test since for all discussed cases they are virtually the same.

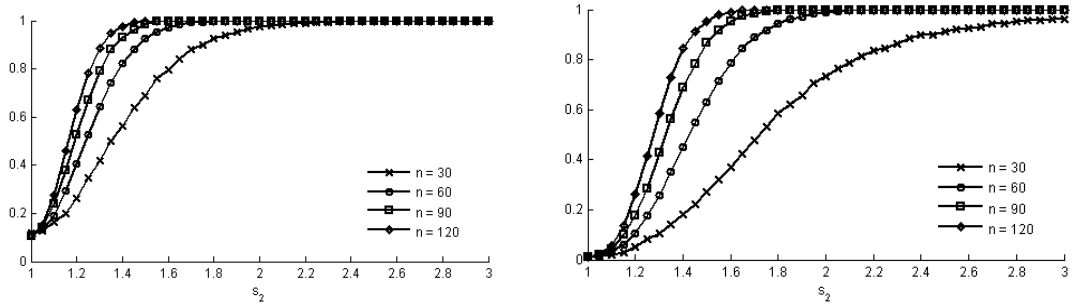

Fig. 1. Power of the unconditional test for the process P1a and nominal significance levels $\alpha = 0.1$ (left axis) and $\alpha = 0.01$ (right axis).

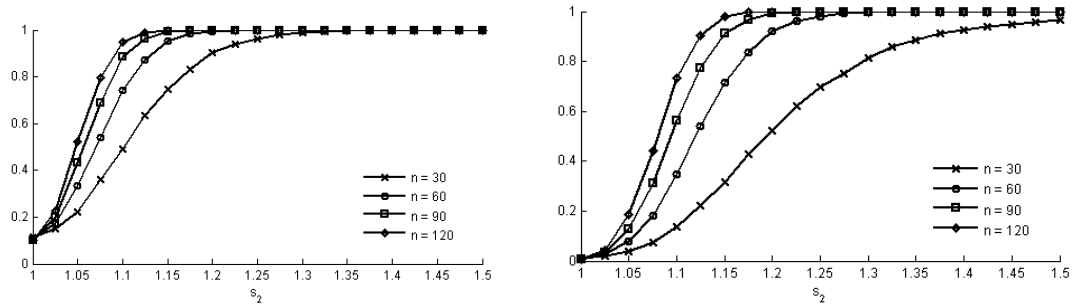

Fig. 2. Power of the unconditional test for the process P1b and nominal significance levels $\alpha = 0.1$ (left axis) and $\alpha = 0.01$ (right axis).

We also analyzed the size of the test for the process P1a assuming that the second errors have also unitary variance. The results are presented in table 1. It can be easily seen that regardless of the correlation level between the two series the test has correct sizes which are close to the nominal significance levels. In all cases they exceed the nominal counterparts only slightly even for the shortest sample. We do not report size of the conditional version of the test since it is almost the same as in table 1.

The power of the unconditional version of the test for the process P2a is presented on figure 3. In that case the errors from one model are serially correlated. Similarly to the previous figures the x-axis depicts unconditional standard deviation of the serially correlated errors that now is determined by the autocorrelation coefficient. We see that the power of the test is now lower than in case of no serial correlation. If we consider the higher significance

level $\alpha = 0.1$ the unconditional standard deviation s_2 needs to be approximately doubled compared to uncorrelated case to guarantee the high power of the test. To be more concrete, for the short sample with 30 observations even for large autocorrelation coefficient $\rho_{12} = 0.9$ that results in high unconditional standard deviation of more than 3.2 the rejection fraction is only about 70%. The results for $\alpha = 0.01$ are more extreme. In that case the test is simply unable to reject the null frequently enough even for huge differences in unconditional standard deviations. All considered samples are too short for correct inference.

		$\rho_{12} = 0$				$\rho_{12} = 0.95$			
		Sample length *n*				**Sample length *n***			
		30	**60**	**90**	**120**	**30**	**60**	**90**	**120**
Significance level α	0.1	0.117	0.108	0.105	0.103	0.117	0.109	0.107	0.106
	0.05	0.061	0.054	0.053	0.052	0.059	0.055	0.054	0.053
	0.01	0.013	0.011	0.011	0.011	0.013	0.011	0.011	0.010

Table 1 Size of the unconditional version of the test for the processes P1a and P1b.

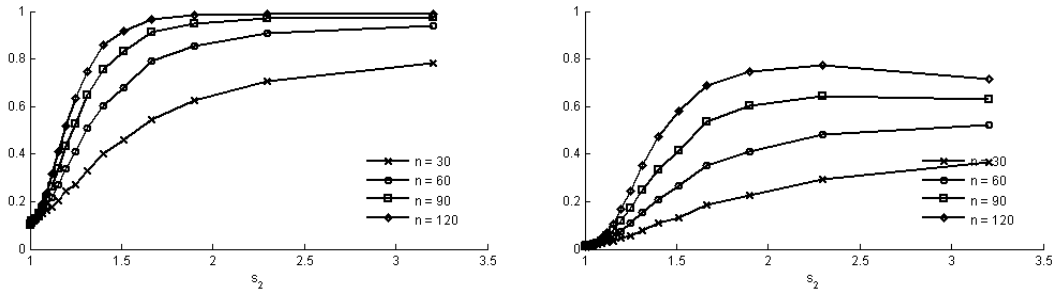

Fig. 3. Power of the unconditional test for the process P2a and nominal significance levels $\alpha = 0.1$ (left axis) and $\alpha = 0.01$ (right axis).

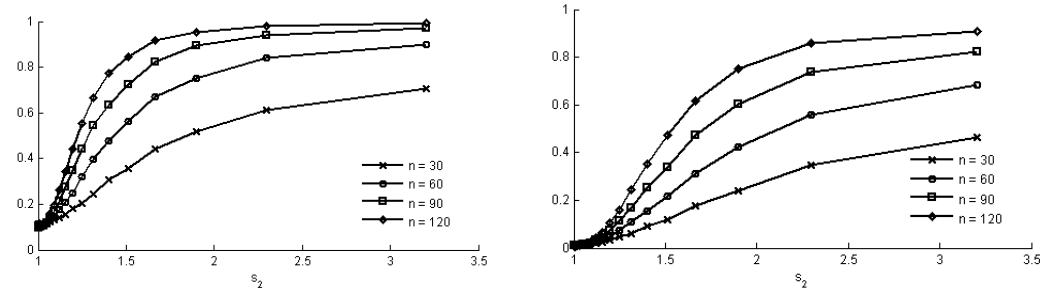

Fig. 4. Power of the conditional test for the process P2a and nominal significance levels $\alpha = 0.1$ (left axis) and $\alpha = 0.01$ (right axis).

This is however not the case as far as the conditional version of the test is concerned with lagged values of the serially correlated process as additional information. This situation is illustrated on figure 4. Now the power steadily rises as the standard deviation grows, although the increase is very slow. For $\alpha = 0.1$ the power of the conditional test is slightly lower than in its unconditional version. The problem almost completely disappears when the series are mutually correlated (process P2b). Then the power of the test is high even for small differences in unconditional standard deviations. We do not report exact results here but they are similar to that presented on figure 2.

		$\rho_{12} = 0$				$\rho_{12} = 0.95$			
		Sample length *n*				**Sample length *n***			
		30	**60**	**90**	**120**	**30**	**60**	**90**	**120**
Significance level α	0.1	0.402	0.349	0.320	0.298	0.178	0.167	0.163	0.160
	0.05	0.314	0.272	0.246	0.228	0.106	0.099	0.096	0.094
	0.01	0.174	0.160	0.146	0.135	0.033	0.030	0.029	0.027

Table 2 Size of the unconditional version of the test for the process P2a and nominal significance levels $\alpha = 0.1$ and $\alpha = 0.01$.

Finally we conduct the analysis of the size of the unconditional test for the serially correlated processes. The results are presented in table 2. The autocorrelation causes severe size distortions. If the series are mutually independent and the nominal level equals 0.1 the empirical size is usually above 0.3 and for $\alpha = 0.01$ it is at least of one order of magnitude higher. The distortions for mutually correlated series are somewhat smaller. However the empirical size still exceeds its nominal level two- or threefold.

5. Conclusions

The presented results lead to a few conclusions. First the power of the tests is significantly higher when one compares forecasts that are highly mutually correlated. For normally distributed and serially independent errors the test correctly rejects the null when forecast standard deviation from one model exceeds the second one by 10% provided a sample has at least 100 observations. For short samples the test achieves high power for differences of order 25%. The properties of the conditional test are very similar in that case. If the error series are mutually independent the power is considerably lower. The properties of the test worsen significantly when one of the error series is serially correlated. This can result in severe size

distortions as well as power loss. The conditional version of the test behaves slightly better but still suffers from the mentioned problems.

To sum up we can recommend using the CPA test for comparison of short samples forecasts at 0.1 significance level provided the error series do not exhibit strong serial dependence. It is also highly desirable to compare forecasts that are mutually dependent. One should be careful using the test in case of serially correlated series. The conditional version of the test should then be applied.

References

[1] Andrews, D.W., 1991. Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. Econometrica 59, 817-858.
[2] Busetti, F., Marcucci, J., 2013. Comparing Forecast Accuracy: A Monte Carlo Investigation. International Journal of Forecasting 29, 13-27.
[3] Clark, T.E., 1999. Finite-Sample Properties of Tests for Equal Forecast Accuracy. Journal of Forecasting 18, 489-04.
[4] Clark, T.E., McCracken, M.W., 2001. Tests of Equal Forecast Accuracy and Encompassing for Nested Models. Journal of Econometrics 105, 85-110.
[5] Clark, T.E., McCracken, M.W., 2005. The Power of Tests of Predictive Ability in the Presence of Structural Breaks. Journal of Econometrics 124, 1-31.
[6] Clark, T.E., McCracken, M.W., 2011. Advances in Forecast Evaluation. Federal Reserve Bank of Cleveland Working Paper 11-20.
[7] Clark, T.E., West, K.D., 2007. Approximately Normal Tests for Equal Predictive Accuracy in Nested Models. Journal of Econometrics 138, 291-311.
[8] Diebold, F.X., Mariano, R.S., 1995. Comparing Predictive Accuracy. Journal of Business and Economic Statistics 13, 253-263.
[9] Giacomini, R., White, H., 2006. Tests of Conditional Predictive Ability. Econometrica 74, 1545-1578.
[10] Newey, W.K., West, K.D., 1987. A Simple, Positive Semidefinite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica 55, 703-708.
[11] Rossi, B., 2013. Advances in Forecasting Under Instability. In: Elliott, G., Timmermann, A. (eds.), Handbook of Economic Forecasting, Volume 2. Amsterdam: Elsevier-North Holland Publications.
[12] West, K.D., 1996. Asymptotic Inference about Predictive Ability. Econometrica 64, 1067-1084.