On the usefulness of the Diebold-Mariano test in the selection of prediction models: Some Monte Carlo evidence

Costantini, Mauro
University of Vienna, Department of Economics
Brunner Strasse 72
Vienna 1210, Austria
E-mail: mauro.costantini@univie.ac.at

Kunst, Robert M.
Institute for Advanced Studies, Department of Economics and Finance
Stumpergasse 56
Vienna 1060, Austria
E-mail: kunst@ihs.ac.at

In their search for the best forecasting model or procedure for their data, researchers routinely reserve a portion of their samples for out-of-sample prediction experiments. Instinctively, they feel that a model or procedure that has shown its advantages for a training sample will also be a good choice for predicting the unknown future. Following the seminal publication by Diebold and Mariano (1995, DM) who introduced the Diebold-Mariano (DM) test, it has become customary and often required to add an evaluation of significance to forecast comparisons. This may have led to widespread doubts on the recommendation by the primary comparisons, if differences among rivals cannot be shown to be statistically significant. Typically, one of the procedures is chosen as the 'simple' or 'benchmark' procedure and significance is assigned to the increase in accuracy achieved by a more sophisticated rival. The impression conveyed by this practice is that the sophisticated procedure is recommended only if it is 'significantly' better than the benchmark, not just if it has better accuracy statistics.

Two arguments can be raised against this practice. First, the null hypothesis of the DM test, i.e. the exact equality of population values or expectations of statistics from two comparatively simple forecasting models or other procedures, is unlikely *a priori*. Except in artificial designs, the true data-generating process will be more complex than all rival prediction models. Classical hypothesis testing, however, requires a plausible null. An implausible null implies a sizeable small-sample bias in its favor. Here, this means that the benchmark model implicitly obtains a strong prior.

Second, the original forecast comparison, if based on a true out-of-sample experiment, is a strong model-selection tool on its own grounds. Minimizing prediction errors over a training sample can be asymptotically equivalent to traditional information criteria (see Wei, 1992, Inoue and Kilian, 2006, Ing, 2007). Conducting a test 'on top' of the information criterion decision is tantamount to increasing the penalty imposed in these criteria and may lead to an unwanted bias in favor of simplicity. Even if a bias in favor of simplicity corresponds to the forecaster's preferences, the same effect can be obtained by an information criterion with a stronger penalty without any additional statistical testing.

Within this paper, we restrict attention to binary comparisons between a comparatively simple time-series model and a more sophisticated rival.

The theoretical background

Typically, the DM test is performed on accuracy measures such as MSE (mean squared errors) following an out-of-sample forecasting experiment, in which a portion of size S from a sample of size T is predicted. In a notation close to DM, the null hypothesis of such tests is $\mathrm{E}g(e_1) = \mathrm{E}g(e_2)$, where $e_j, j = 1, 2$ denote the prediction errors for the two rival forecasts, $g(.)$ is some function—for example,

$g(x) = x^2$ for the MSE—and E denotes the expectation operator. The out-of-sample prediction experiment (SOOS for simulated out-of-sample according to Inoue and Kilian, 2006) is, however, in itself comparable to an information criterion. The asymptotic properties of this SOOS criterion critically depend on the large-sample assumptions for S/T.

If S/T converges to a constant in the open interval $(0, 1)$, Inoue and Kilian (2006) show that the implied SOOS criterion is comparable to traditional 'efficient' criteria such as AIC. The wording 'efficient' is due to McQuarrie and Tsai (1998) and relates to the property of optimizing predictive performance at the cost of some inconsistency. If $S/T \to 1$, Wei (1992) shows that the implied SOOS criterion is consistent in the sense that it selects the true model with probability one as $T \to \infty$.

If a consistent model-selection procedure is flanked by a further hypothesis test that has the traditional test-consistency property, in the sense that it achieves its nominal significance level on its null and rejection with probability one on its alternative, this does not affect the asymptotic property of selection consistency, unless there is a strong and negative dependence between the test statistic and the information criterion. Whereas, in the issue of concern this dependence is more likely to be positive, we briefly consider the case of independence as a benchmark.

Proposition 1 *Suppose there exists a consistent information criterion τ_1 and an independent test-consistent significance test τ_2 at a given significance level α_2. Then, the joint decision from rejecting H_0 if both criteria prefer the alternative is a consistent model selection procedure.*

While this result appears to imply that flanking a consistent criterion with a hypothesis test is innocuous, note that this joint test does not preserve the original significance level.

Proposition 2 *Suppose there exists an information criterion τ_1 with implicit significance level $\alpha_1(T)$ at T, and an independent test-consistent significance test τ_2 at level α_2. Then, the joint test has critical level $\alpha_1(T)\alpha_2$.*

For the inconsistent AIC, the asymptotic implicit significance level is around 0.14. Flanking it with a 5% test implies a level of 0.007. Thus, even if the asymptotic decision is correct, the procedure entails a strong preference for the null model.

Clearly, the DM statistic and a typical consistent information criterion, whether SOOS or BIC, will not be independent, which mitigates this strong *a priori* null preference. With exact dependence, the implicit level $\alpha_1(T)$ is attained as it is usually lower than the specified level α_2. In this case, the DM test decision is ignored. In any other case, the preference for the null will be stronger than that implied by the information criterion. This fact promises a bleak prospect for flanking the IC decision: either flanking is not activated or it generates a bias toward the null. The strength of this bias will be the subject of our simulation experiments.

In particular, we find it useful to study the situation given by the following proposition:

Proposition 3 *Suppose there exists a consistent information criterion τ such that between two models M_1 and M_2 the event $\tau > 0$ indicates a preference for M_2, while $\tau \leq 0$ prefers M_1. Assume the user instead bases her decision on $\tau > \tau_0$ with $\tau_0 > 0$. This decision will be inconsistent in the sense that, as $T \to \infty$, the probability of preferring M_1 although M_2 is true will not converge to 0.*

Depending on the nature of the true data-generation mechanism, particularly on whether the models are nested or not, flanking the consistent SOOS criterion with a DM statistic may lead to situations close to the one being described by Proposition 3. In typical applications of significance tests, the criterion statistic τ can be properly scaled to $(\tau - \tau_0)/f(T)$, such that it converges to 0 for M_1 and to ∞ for M_2. Then, it will not hit a non-zero interval $(0, \tau_0]$ for large T, and consistency is

unaffected. If the significance level for the DM test, however, is gradually reduced as $T \to \infty$, as it is often recommended in order to obtain a fully consistent test, the inconsistency may be relevant.

This offers an even bleaker prospect for the practice of testing on top of the training-sample comparison. However, we are less interested here in asymptotic properties than in finite-sample effects. These can only be reliably studied by means of Monte Carlo with realistic assumptions on the data-generating process (DGP) and on entertained prediction models. For the second and third simulation designs, we particularly assume that the DGP is more complex than all entertained rivals.

Simulations: Nested design

The original DM test is known to suffer from severe distortions for nested model situations, see Clark and McCracken (2001). Nevertheless, it has been used repeatedly by empirical forecasters, and we see this simple nested design as a benchmark case with some practical relevance.

Our basic design does not allow for mis-specification in the sense that at least one of the forecasting models corresponds to the data-generating process. We simulate ARMA(1,1) series of length N according to $X_t = \phi X_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1}$ for $t = 1, \ldots, N$, with Gaussian N(0,1) noise (ε_t). The autoregressive coefficient ϕ is varied over the set $\{0, 0.3, 0.5, 0.7\}$, such that all models are stationary and no model touches upon the sensitive non-stationarity boundary. The moving-average coefficient θ is varied over the set $\{-0.9, -0.7, -0.5, -0.3, 0,\ 0.3, 0.5, 0.7, 0.9\}$. A burn-in of 100 observations shields the experiments against a potential dependence on starting values. 1000 replications of each constellation are generated.

As forecasting models, we consider the autoregressive AR(1) model $X_t = \phi X_{t-1} + \varepsilon_t$ and the ARMA(1,1) model. The AR(1) candidate is correctly specified on $\Theta_R = \{(\phi, \theta) | \theta = 0 \,\text{or}\, \theta = \phi\}$. Otherwise, the AR(1) model is theoretically misspecified. It is to be expected that a reasonable selection procedure chooses the AR(1) on its 'home ground' Θ_R, and the ARMA(1,1) model for stronger deviations from Θ_R. It is also expected that in small samples AR(1) will outperform ARMA(1,1) even for cases outside Θ_R and will be selected accordingly.

Our expectations are met by the simulation results for $N = 100$. Observations $t = 52, \ldots, 99$ are used as a training sample in the sense that models are estimated from samples $t = 1, \ldots, T$ and the mean squared error of one-step out-of-sample forecasts for observations X_{T+1} is evaluated by averaging over $T = 51, \ldots, 98$. The AR(1) forecast is clearly superior on Θ_R and appears to dominate for some other cases. In fact, the AR(1) model yields a smaller MSE for two thirds of all replications for $(\phi, \theta) = (0.3, 0.5)$, while this quota falls to 3 out of 1000 for $(\phi, \theta) = (0.3, -0.9)$.

The simpler AR forecast dominates slightly on the two branches of the set Θ_R and is markedly worse as the parameter values move away from the set. This picture is surprisingly similar for $N = 100$ and $N = 200$, excepting a slight gain for ARMA forecasting in larger samples. While the 'true' model should clearly dominate for larger N, the ratios summarize expanding windows over a wider range of N values and thus do not correspond to expectations. This situation changes for the second step of the prediction experiment, as observations at positions $t = 100$ and $t = 200$ are then evaluated.

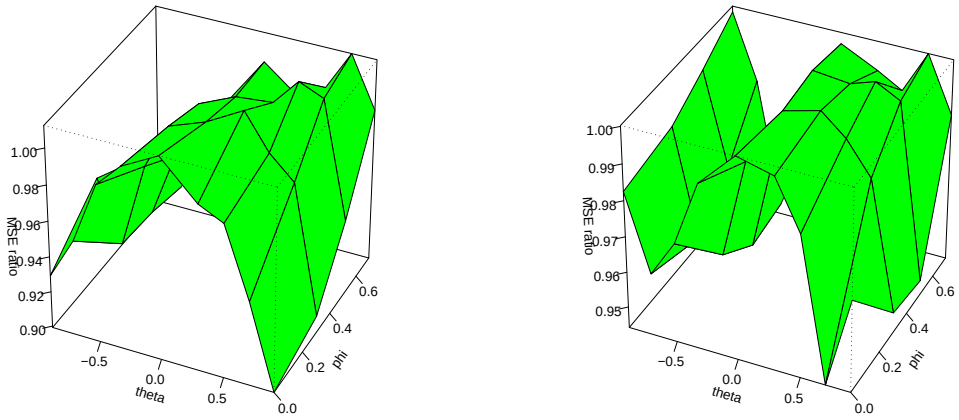
A virtual forecaster who is interested in forecasting observation X_N may use this comparison to choose the better forecasting model, thus extrapolating the observed relative performance. We were surprised at the quality of this procedure. It appears that even the 'incorrect' choice of an AR(1) model at larger distance from Θ_R can benefit forecasting accuracy. Some trajectories are infested by short sequences of large errors, for example, which may create poor estimates for the ARMA(1,1) parameters. The more 'robust' AR(1) estimation at $t < N$ often continues its dominance for $t = N$.

If this procedure is modified by conducting a DM test and sticking to the AR model unless the dominance of the ARMA scheme is significant at 5%, the MSE increases over almost the whole parameter space. Only in some cases with $\theta = 0$ does the MSE decrease, as the DM test enhances the

support for the pure AR model that is beneficial for such values. In other words, the bias in favor of the null has small benefits if the null is true but causes a sizeable deterioration if it is false.

With $N = 100$, the forecast based on the selection dictated by an MSE evaluation over the training sample strictly dominates the forecast that used an additional DM test, excepting a part of the Θ_R set. With $N = 200$, the race between the two selection strategies becomes closer. Particularly for the cases with $\theta = -0.9$ and $\theta = 0.9$, close to non-invertibility, a relative gain for the procedure using the flanking DM test becomes obvious, though even there the procedure without that test still dominates. Performance becomes trimodal, with near-equivalence between approaches for nearly non-invertible cases and for pure AR models, and more palpable advantages for skipping the DM-test step for intermediate values of θ.

Generally, dominance or at least equal performance of the DM-guided model selection is mainly restricted to the case $\theta = 0$, i.e. the pure AR model. For most other cases, the additional DM step yields a deterioration in forecasting accuracy.



MSE ratio DM selected AR or ARMA model divided by selected model without DM testing. Left graph for $N = 100$***, right graph for*** $N = 200$***.***

Simulations: Non-nested design

In this second experiment, data are generated from ARMA(2,2) processes. There are twelve pairs of AR coefficients. Eight pairs yield complex conjugates in the roots of the characteristic AR polynomial and hence cyclical behavior in the generated processes. Three pairs imply real roots, and one case is the origin to include the case of a pure MA structure. These autoregressive designs are combined with several moving-average specifications: a benchmark case without MA component, a first-order MA model, an MA(2) model with $\theta_1 = 0$, and a full specification with $\theta_1 = \theta_2$.

The prevailing impression from these simulations is that the AR(2) model dominates at most parameter values. This dominance is partly caused by the comparatively simpler MA part of the generating processes, but it may also indicate greater robustness in the estimation of autoregressive models as compared to mixed models. The relative performance of the two rival models, measured by the ratio of MSE(AR) and MSE(ARMA), remains almost constant as N increases from 100 to 200, which indicates that the large-sample ratios may already have been attained. The absolute performance, however, improves perceptibly as the sample size increases.

For the pure AR(2) model, there are mostly gains for imposing the DM step. The null model of the test is the true model, and the extra step helps in supporting it. For strong MA effects, the DM step tends to incur some deterioration.

In summary, the DM procedure is beneficial for prediction performance in 34 out of 48 designs

for $N = 100$, but this dominance decreases to 25 cases for $N = 200$. The training procedure without the DM step wins 10 (12) cases for $N = 100(200)$, the remainder are ties at three digits. A rough explanation is that the AR(2) model usually forecasts better than the ARMA(1,1) model, often simply due to a better fit to the generating ARMA(2,2) by the asymptotic pseudo-model or due to better estimation properties of the autoregressive estimator, which uses simple and straightforward conditional least squares. The DM step enhances the preference for the AR(2) model and thus improves predictive accuracy, though this effect becomes less pronounced as the sample size increases.

Simulations: Nonlinear generation mechanism

In this experiment, the data are generated from a nonlinear time-series process. Tiao and Tsay (1994) considered a self-exciting threshold autoregressive (SETAR) model for the growth rate of U.S. gross national product (GNP). Four regimes correspond to economic recessions and expansions with accelerating or decelerating tendencies:

$$y_t = \begin{cases} -0.015 - 1.076y_{t-1} + \varepsilon_{1,t}, & y_{t-1} \leq y_{t-2} \leq 0, \\ -0.006 + 0.630y_{t-1} - 0.756y_{t-2} + \varepsilon_{2,t}, & y_{t-1} > y_{t-2},\ y_{t-2} \leq 0, \\ 0.006 + 0.438y_{t-1} + \varepsilon_{3,t}, & y_{t-1} \leq y_{t-2},\ y_{t-2} > 0, \\ 0.004 + 0.443y_{t-1} + \varepsilon_{4,t}, & y_{t-1} > y_{t-2} > 0. \end{cases}$$

The standard deviations of the errors $\sigma_j = \sqrt{\mathrm{E}\varepsilon_{j,t}^2}$ are $\sigma_1 = 0.0062$, $\sigma_2 = 0.0132$, $\sigma_3 = 0.0094$, and $\sigma_4 = 0.0082$. In contrast with linear models, threshold models may behave differently in qualitative terms if the relative scales of the error processes change. For a recent summary of results on statistical properties of such models, see Fan and Yao (2005). Within regime 1, which corresponds to a deepening economic recession, the model is 'locally unstable', as the coefficient is less than -1. Nevertheless, the model is 'globally stable'.

For our prediction experiment, we use samples drawn from the SETAR process with $N = 100$ and $N = 200$ observations. Burn-in samples of 1000 observations are generated and discarded. 1000 replications are performed. The hypothetical forecaster is supposed to be unaware of the nonlinear nature of the DGP, and she fits AR(p) and ARMA(p, p) models to the time series. In analogy to the other experimental designs, the models deliver out-of-sample forecasts for the latter half of the observation range, excepting the very last time point. This latter half is viewed as a training sample. Either the better one of the two models or the one that is 'significantly' better according to a DM test, is used to forecast this last time point. We also compare the accuracy of these two strategies with the forecasts that always use the autoregressive or the ARMA model.

In contrast to the other two experiments, the lag order p has not been fixed *a priori*, but it is rather determined by minimizing AIC over the range $1, \ldots, p^*$. The maximum lag orders p^* are set at $\sqrt{N}$ for the AR and at $0.5\sqrt{N}$ for the ARMA model, for smaller samples, and at $2\sqrt{N}/3$ and $\sqrt{N}/3$ for larger samples. This choice is not very influential, as AIC minimization often implies low lag orders, most frequently $p = 1$.

Table 1 summarizes the results. For $N = 100$, the pure AR appears to approximate better than the ARMA model. Choosing the better model on the basis of a pure comparison of performance over the training sample yields an MSE comparable to the pure AR model. This average hides some specific features in single replications. The AR model is preferred on the basis of the training sample in 697 out of 1000 replications, while in the remaining cases the ARMA model can be substantially better. Applying the DM test in order to revise the comparison incurs an improvement in accuracy. Whereas this evidence is turned on its head once the ARMA model is defined as the simple model and the AR model as the complex one, this may not be the natural choice. At $N = 200$, the effect in favor of DM testing weakens, which is in keeping with our second experiment.

Table 1: Results of the SETAR experiment.

	MSE$\times 10^{-4}$		frequency $\succ$	
	$N = 100$	$N = 200$	$N = 100$	$N = 200$
AR	1.115	1.037	0.518	0.479
ARMA	1.133	1.044	0.482	0.521
50% training				
lower MSE	1.113	1.041	0.123	0.118
DM–based	1.112	1.038	0.122	0.106
25% training				
lower MSE	1.106	1.042	0.144	0.144
DM–based	1.114	1.035	0.127	0.137

Note: 'frequency $\succ$' gives the empirical frequency of the model yielding the better prediction for the observation at $t = N$.

For distributions with high variance, MSE may not be the most reliable evaluation criterion. When the cases of improvement among the replications are counted, the slight advantage for test-based selection is turned on its head. At $N = 200$, in 118 cases is the pure training-sample comparison better, while there are only 106 cases with the opposite ranking. By construction, the forecasts are identical for the remaining 776 cases. At $N = 100$, wins and losses are fairly identical. Application of the DM test helps as much as tossing a coin.

REFERENCES

Clark, T.E., and M.W. McCracken (2001) 'Tests of equal forecast accuracy and encompassing for nested models,' *Journal of Econometrics* **105**, 85–110.

Diebold, F.X., and R.S. Mariano (1995) 'Comparing Predictive Accuracy,' *Journal of Business and Economic Statistics* **13**, 253–263.

Fan, J., and Q. Yao (2005) *Nonlinear Time Series,* Springer.

Ing, C.K. (2007) 'Accumulated prediction errors, information criteria and optimal forecasting for autoregressive time series,' *Annals of Statistics* **35**, 1238–1277.

Inoue, A., and L. Kilian (2006) 'On the selection of forecasting models,' *Journal of Econometrics* **130**, 273–306.

McQuarrie, A.D.R., and C.-L. Tsai (1998) *Regression and Time Series Model Selection,* World Scientific.

Tiao, G.C., and R.S. Tsay (1994) 'Some Advances in Non Linear and Adaptive Modelling in Time Series,' *Journal of Forecasting* **13**, 109–131.

Wei, C.Z. (1992) 'On predictive least squares principles,' *Annals of Statistics* **20**, 1–42.

ABSTRACT

In evaluating prediction models, many researchers flank comparative ex-ante prediction experiments by significance tests on accuracy improvement, such as the Diebold-Mariano test. We argue that basing the choice of prediction models on such significance tests is problematic, as this practice may favor the null model, usually a simple benchmark. We explore the validity of this argument by extensive Monte Carlo simulations with linear (ARMA) and nonlinear (SETAR) generating processes. For many parameter constellations, we find that utilization of additional significance tests in selecting the forecasting model fails to improve predictive accuracy.