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 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.

The practice of reserving sample portions for out-of-sample prediction experiments presupposes that a procedure that has shown advantages for a training sample will also be a good choice for predicting the unknown future. Following the introduction of the DM (Diebold and Mariano, 1995) 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 represents the 'benchmark', and significance is assigned to the increase in accuracy achieved by a more sophisticated rival. The impression conveyed by this practice is that the rival is recommended only if it 'significantly' bests 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 expectations of statistics from two comparatively simple forecasting models or other procedures, is unlikely *a priori*. True data-generating processes will be more complex than all rival prediction models. Classical hypothesis testing, however, requires a plausible null. An implausible null implies a bias in its favor. Here, 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 (Wei, 1992, 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.

REFERENCES

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

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

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