Discussion of

Comparing Predictive Accuracy, Twenty Years Later: A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests

by F. X. Diebold

Discussion by Andrew J. Patton*

Duke University

21 April 2014

1 Introduction

This interesting article provides an insider's view of the early days of the "modern" forecast comparison literature, which might be dated as beginning with Diebold and Mariano (1995) and West (1996), papers which spawned numerous extensions and refinements over the subsequent years.[1] The article also reviews how forecast evaluation methods have been employed in the literature, offering some criticism of the use of split-sample forecast comparison techniques (like the Diebold-Mariano test) when applied to model comparison rather than forecast comparison. My discussion is a personal perspective on a personal perspective on the DM test, and it is perhaps a bold move to debate the DM test with "D," but below I will do so.

The key ingredient in a DM test is the loss differential:

$$d_t = L\left(Y_t, \hat{Y}_{1,t}\right) - L\left(Y_t, \hat{Y}_{2,t}\right) \tag{1}$$

*I thank Jia Li for helpful comments. Contact address: Department of Economics, Duke University, 213 Social Sciences Building, Box 90097, Durham NC 27708-0097. Email: andrew.patton@duke.edu.

[1] I denote this the "modern" forecast comparison literature, as the literature on forecast evaluation and comparison stretches all the way back to the very birth of econometrics, with an article by Cowles (1933).

where $L : \mathcal{Y} \times \mathcal{Y}^f \to \mathbb{R}_+$ is some loss function, mapping from the range of possible values for the target variable and the forecast to the non-negative real line.[2] With this variable defined, a test of equal predictive accuracy is easily seen to be equivalent to a simple test that the mean of this scalar time series is zero:

$$H_0 : \mathbb{E}\left[L\left(Y_t, \hat{Y}_{1,t}\right)\right] = \mathbb{E}\left[L\left(Y_t, \hat{Y}_{2,t}\right)\right] \Leftrightarrow H_0 : \mathbb{E}\left[d_t\right] = 0 \tag{2}$$

Diebold praises the virtue of the single assumption that is required to implement the DM test, namely that of covariance stationarity of d_t. With such an assumption in place, a central limit theorem for the sample mean, $\bar{d}_T$, can be invoked, and the properties of the DM test easily established. As noted in his footnote 2, the assumption of covariance stationarity is stronger than needed; what is actually desired is simply a set of assumptions that are sufficient for the DM test statistic to be asymptotically standard Normal under the null hypothesis of equal predictive accuracy. Thus, one might consider swapping Assumption DM (equation 1 in the article) with the following higher-level assumption:

$$\text{Assumption AN:} \quad \begin{cases} & \sqrt{T}\left(\bar{d}_T - \mu\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma^2\right) \\ \text{and} \quad & \exists\ \hat{\sigma}_T^2 \quad s.t. \ \ \hat{\sigma}_T^2 \xrightarrow{p} \sigma^2, \text{ as } T \to \infty \end{cases} \tag{3}$$

In this discussion I argue that neither covariance stationarity nor asymptotic Normality more generally are needed for a test to be a "DM-type" test.

2 Asymptotic Normality is nice but not necessary

In his paper, Diebold argues that there is an important distinction between comparing forecasts over some historical period and comparing models at their pseudo-true parameter values. I completely concur, and I further suggest that, fundamentally, the biggest distinction between the many tests that have been proposed in the last twenty years is not the shape of the limiting distribution of the test statistics, or the various terms that appear in the asymptotic variance, but rather the *null hypothesis* that the statistic is used to test. This hypothesis summarizes the economic

[2] The notation here is slightly more general than that used in Diebold's paper. This notation allows for loss functions that cannot be expressed in terms of the forecast error, such as those based on proportional errors, $y/\hat{y}$, which commonly appear in volatility forecasting applications, see Patton (2011) for example, as well as those based on sign forecast errors, $sgn\left(y\right) - sgn\left(\hat{y}\right)$, see Cumby and Modest (1987) for example.

question being posed, and differences can matter greatly for the economic conclusions drawn. This distinction is noted in Diebold and Mariano (1995) and West (1996), but the clearest discussion of this distinction is in Giacomini and White (2006). These latter authors focus explicitly on hypotheses involving estimated parameters, and one of the contributions of their paper is to provide primitive conditions under which such tests, like the DM test, can be applied.

$$WCM \quad H_0 \quad : \quad \mathbb{E}\left[L\left(Y_t, \hat{Y}_t(\beta_1^*)\ \right)\right] = \mathbb{E}\left[L\left(Y_t, \hat{Y}_t(\beta_2^*)\ \right)\right] \tag{4}$$

$$DM\text{-}GW \quad H_0 \quad : \quad \mathbb{E}\left[L\left(Y_t, \hat{Y}_t(\hat{\beta}_{1,t})\right)\right] = \mathbb{E}\left[L\left(Y_t, \hat{Y}_t(\hat{\beta}_{2,t})\right)\right] \tag{5}$$

Tests comparing models at their pseudo-true parameter values correspond to "WCM" tests,[3] while tests comparing forecasts (from estimated models, or from surveys, judgement forecasts, etc.) correspond to DM-GW tests.

I propose that the defining feature of a "DM-type" test is not the asymptotic Normality of the test statistic (and the simple use of a HAC estimator of the long-run variance of d_t), but rather its focus on tests of *forecasts*, rather than tests of models evaluated at their pseudo-true parameter values. If one accepts this as a key feature, then any test of a null like that in equation (5) might reasonably be called a "DM-type" test, including those that do *not* have asymptotically Normal test statistics.

3 Non-Normal "DM" tests

In a recent working paper, Li and Patton (2013) extend the applicability of the tests of Diebold and Mariano (1995), West (1996), White (2000), Giacomini and White (2006), McCracken (2007), and others, to accommodate *latent* target variables like those encountered in financial applications, for example, forecasting volatility, correlation, beta, or jump risk. The goal in that paper is to provide conditions under which tests based on a proxy constructed using high frequency data (e.g., "realized correlation") have the same properties as the corresponding test based on the true, latent, target variable (e.g., "integrated correlation"). As part of the evaluation of the proposed theory, the simple case where the target variable *is* observable is considered, which corresponds to a standard

[3] I follow Diebold in calling these tests "WCM" after West (1996) and Clark and McCracken (2001), although several other authors have focussed on tests of nulls involving population parameters, including White (2000), Corradi and Swanson (2002), Elliott, Komunjer and Timmermann (2005) and Rossi (2005).

DM-type test. One simulation design in Li and Patton (2013) focuses on correlation forecasting, using a bivariate stochastic volatility process from Barndorff-Nielsen and Shephard (2004) as the data generating process, and comparing forecasts from two DCC models (Engle, 2002) for daily correlations.[4] The null to be tested is:

$$H_0 : \mathbb{E}\left[\left(\rho_t^{True} - \rho_t^{DCC_1}\left(\hat{\beta}_{1,t}\right)\right)^2\right] = \mathbb{E}\left[\left(\rho_t^{True} - \rho_t^{DCC_2}\left(\hat{\beta}_{2,t}\right)\right)^2\right] \tag{6}$$

i.e., a "DM-type" null hypothesis. The usual heteroskedasticity and autocorrelation (HAC) robust t-statistics for this test were found to be far from Normal under the null hypothesis. After some investigation, it appeared that the need for many lags in the HAC variance estimation was distorting the behavior of the test statistic, prompting the consideration of the "fixed b" asymptotics of Kiefer and Vogelsang (2005). The Kiefer-Vogelsang approach also uses a "t test," with possibly many autocorrelations included in the HAC variance estimate, but the limiting distribution is not Normal; rather it follows a non-standard distribution, with critical values provide by Kiefer and Vogelsang. As Table 1 below reveals, the Kiefer-Vogelsang approach to obtaining a (non-Normal) limiting distribution for the DM t-statistic provides better size control than the familiar approach based on asymptotic Normality.

[INSERT TABLE 1 ABOUT HERE]

Other examples of non-Normal "DM-type" tests exist in the literature. Giacomini and Rossi (2010) propose a forecast comparison test that allows for the possibility of time variation in the relative performance of the competing forecasts over the sample. Importantly, the null hypothesis is a function of the forecasts evaluated using estimated parameters, not at the parameters' probability limits. The limiting distribution of the test statistic in this application is a functional of a Brownian motion. The "reality check" of White (2000) is presented as a test of a "WCM-type" null, focusing on forecasts generated using the pseudo-true values of the parameters, however a DM-GW version of the "reality check" is a natural extension (and indeed is hinted at in White's article). The test statistic in that case has a limiting distribution that is the maximum of a vector of correlated Normal random variables, and is thus also non-Normal.

[4] Details on the data generating process and forecast models are omitted in the interests of space; interested readers are referred to Li and Patton (2013).

4 Summary

One of the main arguments made in Diebold's interesting article is that if the objective of an analysis is to compare forecasts over some sample period, then the variables of the interest are the actual forecasts obtained in that sample period, not those that would be obtained using the pseudo-true parameters of the models (if any) that generated them. I agree, and following through on this point, it suggests that the key aspect of the "DM" approach to forecast comparison is the choice to state the null hypothesis as a function of estimated parameters rather than pseudo-true probability limits. Thus, while covariance stationarity of the loss differences ("Assumption DM" in the article), or, more generally, asymptotic normality of the sample mean of the loss differences, is convenient when applicable, it is not necessary for a "DM-type" test; in some applications a DM-type test will have a non-Normal limit distribution.

References

[1] Barndorff-Nielsen, O. E. and N. Shephard, 2004, Econometric Analysis of Realized Covariation: High Frequency Based Covariance, Regression, and Correlation in Financial Economics, *Econometrica* 72, 885–925.

[2] Clark, T. E. and M. W. McCracken, 2001, Tests of Equal Forecast Accuracy and Encompassing for Nested Models, *Journal of Econometrics* 105, 85–110.

[3] Corradi, V. and N. R. Swanson, 2002, A Consistent Test for Out of Sample Nonlinear Predictive Ability, *Journal of Econometrics* 110, 353–381.

[4] Cowles, Alfred 3rd, 1933, Can Stock Market Forecasters Forecast?, *Econometrica* 1, 309–324.

[5] Cumby, R. E. and D. M. Modest, 1987, Testing for Market Timing Ability: A Framework for Forecast Evaluation, *Journal of Financial Economics* 19, 169–189.

[6] Diebold, F. X. and R. S. Mariano, 1995, Comparing Predictive Accuracy, *Journal of Business and Economic Statistics* 13, 253–263.

[7] Elliott, G., I. Komunjer and A. Timmermann, 2005, Estimation and Testing of Forecast Rationality under Flexible Loss, *Review of Economic Studies* 72, 1107–1125.

[8] Engle, R. F., 2002, Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models, *Journal of Business and Economic Statistics* 20, 339–350.

[9] Giacomini, R. and B. Rossi, 2010, Forecast Comparisons in Unstable Environments, *Journal of Applied Econometrics* 25, 595–620.

[10] Giacomini, R. and H. White, 2006, Tests of Conditional Predictive Ability, *Econometrica* 74, 1545–1578.

[11] Kiefer, N. M. and T. J. Vogelsang, 2005, A New Asymptotic Theory for Heteroskedasticity-Autocorrelation Robust Tests, *Econometric Theory* 21, 1130–1164.

[12] Li, J. and A. J. Patton, 2013, Asymptotic Inference about Predictive Accuracy using High Frequency Data, working paper, Duke University.

[13] McCracken, M. W., 2007, Asymptotics for Out-of-Sample Tests of Granger Causality, *Journal of Econometrics* 140, 719–752.

[14] Newey, W. K. and K. D. West, 1987, A Simple, Positive Semidefinite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica* 55, 703–708.

[15] Patton, A. J., 2011, Volatility Forecast Comparison using Imperfect Volatility Proxies, *Journal of Econometrics* 160, 246–256.

[16] Rossi, B., 2005, Testing Long-Horizon Predictive Ability with High Persistence, and the Meese-Rogoff Puzzle, *International Economic Review* 46, 61–92.

[17] West, K. D., 1996, Asymptotic Inference about Predictive Ability, *Econometrica* 64, 1067–1084.

[18] White, H., 2000, A Reality Check for Data Snooping, *Econometrica* 68, 1097–1126.

Table 1: Rejection probabilities under the null

	Out-of-sample size					
	500	*1000*	*2000*	*500*	*1000*	*2000*
In-sample size	**Normal**			**Kiefer-Vogelsang**		
500	0.22	0.17	0.17	0.07	0.05	0.03
1000	0.24	0.22	0.20	0.13	0.08	0.05

Notes: This table presents the proportion of simulations in which the null hypothesis of equal predictive accuracy is rejected at the 0.05 level, in a design in which the null hypothesis is true. All tests are based on a standard Diebold-Mariano test statistic, with HAC standard errors constructed using Newey and West (1987). The "in-sample size" refers to the number of observations used to estimate the forecast models; the "out-of-sample size" (P) refers to the number of observations used to compare the forecast models. In the left panel the number of lags used in the HAC estimate is set to $3P^{1/3}$ and critical values are obtained from the standard Normal distribution. In the right panel the number of lags is set to $P/2$ and critical values are obtained from Kiefer and Vogelsang (2005). Further details are available in Li and Patton (2013).