NBER WORKING PAPER SERIES

COMPARING PREDICTIVE ACCURACY, TWENTY YEARS LATER: A PERSONAL PERSPECTIVE ON THE USE AND ABUSE OF DIEBOLD-MARIANO TESTS

Francis X. Diebold

Working Paper 18391
http://www.nber.org/papers/w18391

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2012

This paper is based on a lecture delivered in March 2012 at CREATES, University of Aarhus, Denmark. For valuable comments I am grateful to participants at the CREATES lecture, especially Niels Haldrup, Soren Johansen, Asger Lunde, and Timo Terasvirta. I am also grateful to participants at Penn's Econometrics Lunch Seminar, as well as Peter Hansen, Lutz Kilian, Barbara Rossi, Frank Schorfheide, Minchul Shin, Norman Swanson, Allan Timmermann, Ken Wolpin and Jonathan Wright. The usual disclaimer most definitely applies. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

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

Francis X. Diebold

NBER Working Paper No. 18391
September 2012
JEL No. C01,C52,C53

ABSTRACT

The Diebold-Mariano (DM) test was intended for comparing forecasts; it has been, and remains, useful in that regard. The DM test was not intended for comparing models. Unfortunately, however, much of the large subsequent literature uses DM-type tests for comparing models, in (pseudo-) out-of-sample environments. In that case, much simpler yet more compelling full-sample model comparison procedures exist; they have been, and should continue to be, widely used. The hunch that (pseudo-) out-of-sample analysis is somehow the "only," or "best," or even a "good" way to provide insurance against in-sample over-fitting in model comparisons proves largely false. On the other hand, (pseudo-) out-of-sample analysis may be useful for learning about comparative historical predictive performance.

Francis X. Diebold
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297
and NBER
fdiebold@sas.upenn.edu

1 Introduction

One routinely has competing forecasts of the same object and seeks to determine which is better. To take a concrete example, consider U.S. inflation forecasting. One might obtain survey-based forecasts from the Survey of Professional Forecasters (S), $\{\pi_t^S\}_{t=1}^T$, and simultaneously one might obtain market-based forecasts from inflation-indexed bonds (B), $\{\pi_t^B\}_{t=1}^T$. Suppose that loss is quadratic and that during $t = 1, ..., T$ the sample mean-squared errors are $\widehat{MSE}(\pi_t^S) = 1.80$ and $\widehat{MSE}(\pi_t^B) = 1.92$. Evidently "$S$ wins," and one is tempted to conclude that S provides better inflation forecasts than does B. The forecasting literature is filled with such horse races, with associated declarations of superiority based on outcomes.

Obviously, however, the fact that $\widehat{MSE}(\pi_t^S) < \widehat{MSE}(\pi_t^B)$ in a particular sample realization does not mean that S is necessarily truly better than B in population. That is, even if in population $MSE(\pi_t^S) = MSE(\pi_t^B)$, in any particular sample realization one or the other of S and B must "win," so the question arises in any particular sample as to whether S is truly superior or merely lucky. The Diebold and Mariano (1995) (DM) test is an econometric tool for answering that question, allowing one to assess the significance of apparent predictive superiority.[1] It provides a test of the hypothesis of equal expected loss (in our example, $MSE(\pi_t^S) = MSE(\pi_t^B)$), valid under quite general conditions including, for example, wide classes of loss functions.

2 *Forecast* Comparisons

DM is a test for comparing forecasts, not models. Here I elaborate on that basic point, which was missed by much of the ensuing literature.[2] In particular, I discuss its implications for construction and justification of the DM test statistic.

[1] The DM paper has a rather colorful history. It was written in summer 1991 when Diebold visited the Institute for Empirical Macroeconomics at the Federal Reserve Bank of Minneapolis; see Diebold and Mariano (1991) at `http://econpapers.repec.org/paper/fipfedmem/default1.htm`. Subsequently it was curtly rejected by *Econometrica* after a long refereeing delay, with a quarter-page "report" expressing bewilderment as to why anyone would care about the subject it addressed. I remain grateful to the *Journal of Business and Economic Statistics* for quickly recognizing the paper's contribution and eventually publishing it in 1995. Quite curiously, *Econometrica* published a Diebold-Mariano extension the next year. In 2002 the Diebold-Mariano paper appeared in the *JBES*'s Twentieth Anniversary Commemorative Issue (Ghysels and Hall (2002)), containing reprints of the ten most influential papers published in the *JBES*'s first twenty years. As of August 2012 it had more than 3000 Google Scholar citations.

[2] Alas, Diebold and Mariano are partly responsible. They were careful to focus exclusively on forecast comparison, as opposed to model comparison, throughout the bulk of their paper, but they nevertheless did speculate briefly on the possibility of DM-based model comparison in their concluding remarks.

2.1 The DM Perspective, Assumption DM, and the DM Statistic

The essence of the DM approach is to take forecast errors as primitives, intentionally. Those forecast errors need not come from models, and even if they do, the models need not be known to the econometrician. This makes for wide DM applicability, because in many important applications the models are either unknown, as for example with proprietary models provided by a third party, or there simply *are* no "models," as for example with forecasts based on surveys, forecasts extracted from financial markets, forecasts obtained from explicit prediction markets, and forecasts based on expert judgment (entirely or in part).

DM makes assumptions directly on the forecast errors, or more precisely, on the forecast error *loss differential*. Denote the loss associated with forecast error e_t by $L(e_t)$; hence, for example, time-t quadratic loss would be $L(e_t) = e_t^2$. The time-t loss differential between forecasts 1 and 2 is then $d_{12t} = L(e_{1t}) - L(e_{2t})$. DM requires only that the loss differential be covariance stationary. That is, DM assumes that:

$$Assumption\ DM: \quad \begin{cases} E(d_{12t}) = \mu, \ \forall t \\ cov(d_{12t}, d_{12(t-\tau)}) = \gamma(\tau), \ \forall t \\ 0 < var(d_{12t}) = \sigma^2 < \infty. \end{cases} \tag{1}$$

The key hypothesis of equal predictive accuracy (i.e., equal expected loss) corresponds to $E(d_{12t}) = 0$, in which case, under the maintained Assumption DM:

$$DM_{12} = \frac{\bar{d}_{12}}{\hat{\sigma}_{\bar{d}_{12}}} \rightarrow N(0,1), \tag{2}$$

where $\bar{d}_{12} = \frac{1}{T}\sum_{t=1}^{T} d_{12t}$ is the sample mean loss differential and $\hat{\sigma}_{\bar{d}_{12}}$ is a consistent estimate of the standard deviation of $\bar{d}_{12}$ (more on that shortly). That's it, there's nothing more to do, it really *is* that trivial: If Assumption DM holds, then the $N(0,1)$ limiting distribution of test statistic DM holds.

DM is obviously extremely simple, almost embarrassingly so. It is simply an asymptotic z-test of the hypothesis that the mean of an observed series (the loss differential) is zero. The only wrinkle is that forecast errors, and hence loss differentials, may be serially correlated for a variety of reasons, the most obvious being forecast sub-optimality. Hence the standard error in the denominator of the DM statistic (2) should be calculated robustly. Diebold and Mariano (1995) use $\hat{\sigma}_{\bar{d}} = \sqrt{\hat{g}(0)/T}$, where $\hat{g}(0)$ is a lag-window estimator of the loss

differential spectrum at frequency zero.

DM is also readily extensible. The key is to recognize that the DM statistic can be trivially calculated by regression of the loss differential on an intercept, using heterskedasticity and autocorrelation robust (HAC) standard errors. Immediately, then (and as noted in the original Diebold-Mariano paper), one can potentially extend the regression to condition on additional variables that may explain the loss differential, thereby moving from an unconditional to a conditional expected loss perspective.[3] For example, comparative predictive performance may differ over the business cycle, in which case one might include a business cycle chronology variable in the DM HAC regression.

2.2 Thoughts on Assumption DM

In the previous section I praised DM rather effusively, and its great simplicity and wide applicability certainly *are* virtues: There is just one Assumption DM, just one DM test statistic, and just one DM limiting distribution, always and everywhere (in sharp contrast to the situation where one uses DM-type tests to compare models as opposed to forecasts, as I discuss subsequently). But of course everything hinges on Assumption DM. Here I offer some perspectives on the validity of Assumption DM.

First, as George Box (1979) famously and correctly noted, "All models are false, but some are useful." Precisely the same is true of *assumptions*, as models are just sets of assumptions. Indeed all areas of economics benefit from assumptions that are surely false if taken literally, but that are nevertheless useful. So too with Assumption DM. Surely d_t is likely never *precisely* covariance stationary, just as surely *no* economic time series is likely precisely covariance stationary. But in many cases Assumption DM may be an accurate and useful approximation.[4]

Second, special forecasting considerations lend support to the validity of Assumption DM. Forecasters strive to achieve forecast optimality, which corresponds to unforecastable covariance-stationary errors (indeed white-noise errors in the canonical 1-step-ahead case), and hence unforecastable covariance-stationary loss differentials. Of course they may not achieve optimality, resulting in serially-correlated, and indeed forecastable, forecast errors.

[3] For an important more recent development from the conditional perspective, see Giacomini and White (2006).

[4] To take an example about which I will have much more to say throughout this paper, suppose that one uses DM to compare models, as opposed to forecasts, in a (pseudo-) out-of-sample environment. Although loss-differential nonstationarity is induced as model parameters converge to their pseudo-true values in expanding estimation samples, the induced nonstationarity may be small. Hence the loss differential may be approximately stationary, and the DM null distribution approximately valid.

But non-stationarity of forecast errors (e.g., $I(1)$ behavior) takes serial correlation to the extreme.

Third, even if nonstationary components *do* exist in forecast errors, there is reason to suspect that they may be shared. Information sets overlap across forecasters, so that forecast-error nonstationarities may vanish from the loss differential. For example, two loss series, each integrated of order one, may nevertheless be cointegrated with cointegrating vector $(1, -1)$. Suppose for example that

$$\begin{aligned} L(e_{1t}) &= x_t + \varepsilon_{1t} \\ L(e_{2t}) &= x_t + \varepsilon_{2t}, \end{aligned} \tag{3}$$

where x_t is a common nonstationary $I(1)$ loss component, and ε_{1t} and ε_{2t} are idiosyncratic stationary $I(0)$ loss components. Then $d_{12t} = L(e_{1t}) - L(e_{2t}) = \varepsilon_{1t} - \varepsilon_{2t}$ is $I(0)$, so that the loss differential series is covariance stationary despite the fact that neither individual loss series is covariance stationary.

Fourth, and most importantly, standard and powerful tools enable empirical assessment of Assumption DM. That is, the approximate validity of Assumption DM is ultimately an empirical matter, and a wealth of diagnostic procedures are available to help assess its validity. One can plot the loss differential series, examine its sample autocorrelations and spectrum, test it for unit roots and other nonstationarities including trend, structural evolution, and structural breaks.[5]

3 *Model* Comparisons

Now I consider the use of DM-type tests in model, as opposed to forecast, comparisons. Unavoidable and crucially-important issues arise, related both to finite-sample analysis vs. asymptotic analysis, and more importantly, to comparisons of two models vs. many models.

[5]Even with apparent nonstationarity due to apparent breaks in the loss differential series, Assumption DM may nevertheless hold if the breaks have a stationary rhythm, as for example in hidden-Markov processes in the tradition of Hamilton (1989).

3.1 Two Models

I have emphasized, and I will continue to emphasize, that DM compares *forecasts* via the null hypothesis of a zero expected loss differential,

$$H_0: \ E(d_{12t}) = E\left(L(e(F_{1t})) - L(e(F_{2t}))\right) = 0, \tag{4}$$

where the new and slightly more detailed notation ($e(F_t)$ rather than e_t) is designed to emphasize that the errors are driven by forecasts, not models. As I have also emphasized, in the DM framework the loss differential d_{12t} is the primitive, and one makes Assumption DM directly on d_{12t}.

Many researchers, however, have used DM and DM-type tests not for comparing forecasts, but rather for comparing *models*, via forecasts, in (pseudo-) "out-of-sample" situations. That approach traces to the work of West (1996) and Clark and McCracken (2001), *inter alia*, and in what follows I will use "WCM" in broad reference to it. WCM assumes that the forecasts are from fully-articulated econometric models, known to the researcher. It follows the DM approach and effectively tests a null hypothesis based on the loss differential,

$$H_0: \ E(d_{12t}) = E\left(L(e(F_{1t}(M_1(\theta_1)))) - L(e(F_{2t}(M_2(\theta_2))))\right) = 0, \tag{5}$$

where I now write $e(F_t(M(\theta)))$ to emphasize that the error e is ultimately driven by a model M, which in turn involves a vector of pseudo-true parameters θ.

Mechanically, WCM proceeds roughly as follows. First split the data into a (pseudo-) in-sample period $t = 1, ..., t^*$ and a (pseudo-) out-of-sample period $t = t^* + 1, ..., T$. Then recursively estimate the models with the last (pseudo-) in-sample observation starting at $t = t^*$ and ending at $T - 1$, at each t predicting $t + 1$. Finally, base a DM-style test on the sample mean quadratic loss differential,

$$\bar{d}_{12} = \frac{\sum_{t=t^*+1}^{T}(e^2_{1,t/t-1} - e^2_{2,t/t-1})}{T - t^*}, \tag{6}$$

where $e_{t/t-1}$ is a time-t 1-step-ahead (pseudo-) forecast error, or "recursive residual." There are of course many variations. For example, the (pseudo-) in-sample period could be fixed or rolling, as opposed to expanding, but (6) serves as something of a canonical benchmark.

A key observation is that in the WCM framework the ultimate primitives are not forecasts (or the loss differential), but rather *models*, so WCM proceeds by making assumptions

not about the loss differential, but about the models. Complications arise quickly, however, as one may entertain a wide variety of models and model assumptions. Indeed there is no single "Assumption WCM" analogous to Assumption DM; instead, one must tiptoe carefully across a minefield of assumptions depending on the situation.[6] Such assumptions include but are not limited to: (1) Nesting structure. Are the models nested, non-nested, or partially overlapping? (2) Functional form. Are the models linear or non-linear? (3) Model disturbance properties. Are the disturbances Gaussian? Martingale differences? Something else? (4) Estimation method. Are the models estimated by OLS? MLE? GMM? Something else? (5) Estimation sample(s). Is the (pseudo-) in-sample estimation period fixed? Recursively expanding? Rolling? (6) Asymptotics. What asymptotics are invoked as $T \to \infty$? $t^*/T \to 0$? $t^*/T \to \infty$? $t^*/T \to const$?

But in many respects I digress. The key issue involves not details of implementation of the (pseudo-) out-of-sample paradigm, but rather the paradigm itself. It is not only tedious (one must construct the (pseudo-) out-of-sample forecast errors and ascertain the correct limiting distribution), but also largely misunderstood and sup-optimal in certain important respects, as I argue throughout the rest of this paper. I begin by stepping back and extracting some basic principles of model comparison that emerge from the massive literature.

3.1.1 Optimal Finite-Sample Comparisons

I proceed by example, the first quite specialized and the second quite general. First consider the classical model comparison paradigm, and the very special and simple comparison of two nested Gaussian linear models, M_1 and M_2, where $M_1 \subset M_2$ and M_2 is assumed true. (Hence M_1 may or may not be true.) In that time-honored case, and at the risk of belaboring the obvious, one achieves exact finite-sample optimal inference using the F-test of linear restrictions,

$$F_{12} = \frac{(SSR_1 - SSR_2)/(k-1)}{SSR_2/(T-k)}, \tag{7}$$

where SSR denotes a sum of squared residuals, T is sample size, and k is the number of restrictions imposed under the null hypothesis. As is well-known, F is the uniformly most powerful test. Any other approach is sub-optimal. The key observation for our purposes is that the optimal model comparison procedure is based on full-sample residuals, not (pseudo-) out-of-sample forecast errors.

[6] Lengthy surveys of the WCM approach, and implicitly the many varieties of "Assumption WCM," include West (2006) and Clark and McCracken (2011).

Now maintain focus on exact finite-sample analysis but go in some sense to an opposite extreme, considering the Bayesian model comparison paradigm, and a more general two-model comparison (nested or non-nested, linear or non-linear, Gaussian or non-Gaussian). Like the classical F test above, the Bayesian paradigm produces exact finite-sample inference, but the perspective and mechanics are very different. Its essence, which is to say the essence of rational behavior (via the complete-class theorem), is to condition inference on the observed data y – *all* observed data. In the model comparison context, the prescription for doing so is simply to select the model favored by posterior odds,

$$\underbrace{\frac{p(M_1|y)}{p(M_2|y)}}_{posterior\ odds} = \underbrace{\frac{p(y|M_1)}{p(y|M_2)}}_{Bayes\ factor} \times \underbrace{\frac{p(M_1)}{p(M_2)}}_{prior\ odds} . \tag{8}$$

As equation (8) emphasizes, however, all data-based information in the posterior odds comes from the Bayes factor, which is the ratio of marginal likelihoods. Indeed if prior odds are 1:1, the Bayesian prescription is simply to select the model with higher marginal likelihood. The marginal likelihood is

$$p(y|M) = \int p(y|\theta, M)p(\theta|M)d\theta. \tag{9}$$

The key observation for our purposes is that the marginal likelihood is a full-sample construct, not a (pseudo-) out-of-sample predictive likelihood.

Thus from two-model classical hypothesis testing in very simple environments, to Bayesian two-model posterior comparisons in much more general environments, *optimal finite-sample model comparison is full-sample model comparison.* Indeed it's hard to imagine otherwise: If one discards data in finite samples, both intuition and mathematics suggest that surely one must pay a price relative to an efficient procedure that uses all data.

3.1.2 Asymptotically-Optimal Comparisons

Now consider asymptotic analysis. First consider again the classical nested model comparison paradigm, but now including non-linear and/or non-Gaussian environments. Little can generally be said analytically in finite-sample environments, but simulation studies find clear superiority of full-sample procedures (e.g., Kilian and Taylor (2003)). Moreover, powerful analytic results are available asymptotically, and they lead to the same conclusion. In particular, as has been known for many decades, each of the "trinity" of likelihood-ratio, Lagrange

multiplier and Wald tests achieves maximum asymptotic local power. The key observation for our purposes is that each member of that trinity is based on the full sample of available data.

Second, consider again the Bayesian marginal likelihood paradigm. Asymptotic analysis is in a certain sense ill-posed there, as the Bayesian perspective is fundamentally finite-sample, conditioning precisely and exclusively on the available sample information. From that perspective, once one determines the model with higher marginal likelihood there's nothing more to do, regardless of whether the sample size is small or huge. The Bayesian optimal finite-sample two-model comparison procedure (8) remains the Bayesian asymptotically-optimal two-model comparison procedure – nothing changes.

Nevertheless, one can ask interesting and important asymptotic questions related to Bayesian model selection. For example, because the marginal likelihood (8) can be very difficult to calculate, the question arises as to whether one can approximate it asymptotically with a simpler construct. Schwarz (1978) answers in the affirmative, showing that, under conditions including $T \to \infty$, the model with higher marginal likelihood is the model with smaller Schwarz information criterion (SIC), where

$$SIC = k\, lnT - 2\, lnL, \tag{10}$$

and k is the number of parameters estimated. Indeed SIC is often, and appropriately, called the Bayesian information criterion (BIC). The key observation for our purposes should by now be familiar: SIC is based on the full-sample likelihood, not a (pseudo-) out-of-sample predictive likelihood.

3.2 Many Models

But there's much more to consider. In reality we typically compare many models, nested and non-nested, one or *none* of which may be coincide with the true data-generating process (DGP).[7] Let us continue our asymptotic discussion from that much more compelling perspective.

SIC extends immediately to comparisons of many models; one simply selects the model

[7]Note that the explicit or implicit assumption thus far has been that at least one of the two models considered is true. The classical nested approach clearly assumes that at least the larger model is correctly specified, as mentioned earlier. Interestingly, the Bayesian (possibly non-nested) approach also implicitly assumes that one of the models is correctly specified, as first emphasized in Diebold (1991). Only recently has that assumption begun to be relaxed, as in Geweke (2010).

with smallest SIC. Closely-related and equally popular, but derived from a different (approximation-theoretic) perspective, is the Akaike (1974) information criterion (AIC). The AIC is

$$AIC = 2k - 2\,lnL, \tag{11}$$

and one selects the model with smallest AIC. SIC and AIC are central because they have the key properties of consistency (SIC) and efficiency (AIC) in model selection.

Consistency and efficiency have very different definitions in the model selection literature than in the traditional estimation literature. Model selection consistency refers to situations where the DGP is among a fixed set of models considered. Very roughly put, consistency of a selection procedure (sometimes called the "oracle property") means that it selects the true DGP asymptotically almost surely. Model selection efficiency refers to situations where the DGP is not among an expanding set of models considered. Again very roughly put, efficiency of a selection procedure means that it selects the Kullback-Leibler- (KLIC-) optimal sequence of approximations to the the DGP asymptotically.[8] The key observation for our purposes is that SIC and AIC – and their optimality properties – are based on full-sample likelihoods, not (pseudo-) out-of-sample predictive likelihoods.

It is illuminating from a model comparison perspective to specialize SIC and AIC to the Gaussian linear regression case, in which they can be written in terms of penalized in-sample mean-squared error (MSE),

$$SIC = T^{\left(\frac{k}{T}\right)}\ MSE \tag{12}$$

$$AIC = e^{\left(\frac{2k}{T}\right)}\ MSE, \tag{13}$$

where $MSE = \frac{\sum_{t=1}^{T} e_t^2}{T}$. SIC and AIC inflate in-sample MSE in just the right ways, relative to their respective optimality criteria, to offset the MSE deflation inherent in model fitting. This is an important lesson: good ways of estimating out-of-sample predictive MSE typically first estimate full-sample residual MSE (thereby using all data) and then transform it appropriately. In particular, AIC and SIC asymptotically guard against *in-sample overfitting* – the spurious appearance of good forecast performance when selecting over many models – not by moving to (pseudo-) out-of-sample analysis, but rather by using *all* data as embedded in full-sample sums of squared errors (SSE's), and simultaneously deflating those SSE's with degree-of-freedom penalties harsher than those associated with traditional F and

[8] It is important to note that of the two optimality properties consistency is the less compelling, as the DGP is surely *never* among the models considered.

related tests.[9]

In closing this section, it is useful to step back and note that although SIC and AIC were developed as pure model selection tools, not as hypothesis tests for model comparison, they can be readily adapted in that way, so that my basic points extend in that direction. The leading example is Vuong (1989) and Rivers and Vuong (2002), who develop inferential methods for AIC.[10] That is, the AIC measures KLIC divergence from the DGP, and they develop methods for testing the pairwise null hypothesis of equal population KLIC divergence. Hansen et al. (2011) go even farther by developing methods for controlling the family-wise error rate when performing many Vuong-type tests, allowing them to obtain a set of models containing the KLIC-optimal approximating model with controlled error rate, the so-called model confidence set.

4 Whither Out-of-Sample Model Comparisons?

Several questions remain. First I ask whether any (pseudo-) out-of-sample model comparison procedure can compete with the full-sample procedures discussed above. The answer turns out to be yes, but only if one takes an asymptotic perspective and invokes the less-compelling optimality concept of consistency, and even then there is a much simpler procedure with the same asymptotic justification and likely-superior finite-sample properties. In light of this, I then proceed to ask whether there is *any* role for (pseudo-) out-of-sample model comparisons. The answer is a cautious yes.

4.1 Can Out-of-Sample *Ever* Compete with Full-Sample?

I have considered a variety of model comparison situations: two-model and many-model, nested and non-nested, finite-sample and asymptotic. In every case, optimal procedures were full-sample procedures. I emphasized, moreover, that it is possible to perform model selection in ways that are asymptotically robust to data mining. But again, in every case,

[9] F and related tests were not designed for large-scale model selection, and they have poor properties (even asymptotically) when used in that way, as do the closely-related strategies of model selection by maximizing $\bar{R}^2$ or minimizing S^2. Indeed max $\bar{R}^2$ model selection is equivalent to min S^2 model selection, where $S^2 = \frac{\sum_{t=1}^{T} e_t^2}{T-k} = \frac{T}{T-k} \frac{\sum_{t=1}^{T} e_t^2}{T}$. Hence its form matches the "penalty $\times$ MSE" form of AIC and SIC, but with penalty $\frac{T}{T-k}$. Efficient model selection requires the harsher penalty factor $e^{\left(\frac{2k}{T}\right)}$ associated with AIC, and consistent model selection requires the even harsher penalty factor $T^{\left(\frac{k}{T}\right)}$ associated with SIC.

[10] See also Li (2009).

optimal procedures were full-sample procedures. Is there no role for (pseudo-) out-of-sample procedures?

It turns out that there is some role, at least from an asymptotic perspective. That is, despite the fact that they discard data, certain (pseudo-) out-of-sample procedures can be justified asymptotically, because the discarded data become asymptotically irrelevant. Rather than estimating out-of-sample MSE by inflating in-sample MSE, such out-of-sample procedures attempt to estimate out-of-sample MSE directly by mimicking real-time forecasting. The key example is "predictive least squares" (PLS). PLS should sound familiar, as it is precisely the foundation on which WCM-type procedures are built. First split the data into a (pseudo-) in-sample period $t = 1, ..., t^*$ and a (pseudo-) out-of-sample period $t = t^* + 1, ..., T$. Then recursively estimate the models over $t = t^* + 1, ..., T$, at each t predicting $t + 1$, and finally construct for each model

$$PLS = \frac{\sum_{t=t^*+1}^{T} e_{t/t-1}^2}{T - t^*}, \tag{14}$$

where $\hat{e}_{t/t-1}$ is the time-t 1-step-ahead (pseudo-) forecast error, or "recursive residual," and select the model with smallest PLS.

Wei (1992) establishes consistency of PLS, but not the more compelling property of efficiency, and it appears that a procedure cannot be both consistent and efficient, as discussed in Yang (2005).[11] So PLS has the less-compelling asymptotic optimality property of consistency, and it's more tedious to compute than SIC, which also has that property. Moreover, one would expect better finite-sample SIC performance, because SIC uses all data.

Hence, based on the considerations invoked thus far, it's hard to imagine why one would do PLS with WCM-type inference as opposed to, say, AIC with Vuong-type inference. Introducing additional considerations, moreover, often worsens matters for PLS/WCM. For example, (pseudo-) out-of-sample methods actually *expand* the scope for data mining in finite samples, as emphasized in Rossi and Inoue (2012) and Hansen and Timmermann (2011), because one can also mine over the sample split point t^*.[12] They develop methods robust to choice of split point, but only at the cost of (additional) power loss.

[11] See also Inoue and Kilian (2006).

[12] All procedures under consideration, even those that achieve robustness to data mining asymptotically, are subject to strategic data mining in finite samples. Achieving robustness to data mining in finite samples requires simulation methods, as in the "reality check" of White (2000) or the bootstrap model confidence set procedure of Hansen et al. (2011).

4.2 What Role for (Pseudo-) Out-of-Sample Model Comparisons?

Nevertheless, in my view there is still a potential role for (pseudo-) out-of-sample model comparisons. Importantly, however, it comes with a caveat.

Quite apart from testing models, (pseudo-) out-of-sample model comparisons may be useful for learning about comparative predictive performance during particular historical episodes. Suppose, for example, that using a full-sample Vuong test one finds that M_1 KLIC-approximates the DGP significantly better than M_2. It may nevertheless be of great interest to go farther, assessing (pseudo-) out-of-sample predictive performance period-by-period via recursive residuals, with particular attention (say) to performance over different business cycles. Such analyses may help to dig into the *reasons* – the "whens and whys and hows" – for M_1's predictive superiority. Rapach et al. (2010), for example, use out-of-sample predictive methods to argue that stock market returns can be forecast during recessions but not during expansions.

An important caveat arises, however. Accurate and informative real-time comparisons require using period-by-period "vintage" data, in contrast to simply using the most recent vintage as if had been available in real time. This is rarely done in the (pseudo-) out-of-sample model comparison literature. It is of course irrelevant for data not subject to revision, such as various financial series, but tremendously relevant for variables subject to revision, such as most macroeconomic series.[13] Moreover, incorporating vintage data causes (even more) complications if one wants to do inference, as emphasized by Clark and McCracken (2009).

5 Conclusion

The DM test was intended for comparing forecasts; it has been, and remains, useful in that regard. The DM test was *not* intended for comparing models. Unfortunately, however, much of the large subsequent literature uses DM-type tests for comparing models, in (pseudo-) out-of-sample environments. In that case, much simpler yet more compelling full-sample model comparison procedures exist; they have been, and should continue to be, widely used.[14] The hunch that (pseudo-) out-of-sample analysis is somehow the "only," or "best," or even a "good" way to provide insurance against in-sample over-fitting in model comparisons proves

[13] For an overview, see Croushore (2006).

[14] Many open and issues remain, of course, even for full-sample procedures. Recent work, for example, has begun to tackle the challenging problem of model comparison in the presence of possible structural change, as in Giacomini and Rossi (2010).

largely false.[15] On the other hand, (pseudo-) out-of-sample analysis may be useful for learning about comparative historical predictive performance.

[15]See Inoue and Kilian (2004) for complementary discussion and insights.

References

Akaike, H (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

Box, G.E.P. (1979), "Robustness in the Strategy of Scientific Model Building," In R.L. Launer and G.N. Wilkinson (eds.), *Robustness in Statistics: Proceedings of a Workshop*, Academic Press.

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

Clark, T.E. and M.W. McCracken (2009), "Tests of Equal Predictive Ability with Real-Time Data," *Journal of Business and Economic Statistics*, 27, 441–454.

Clark, T.E. and M.W. McCracken (2011), "Advances in Forecast Evaluation," In G. Elliott and A. Timmerman (eds.), *Handbook of Economic Forecasting, Volume 2*, Elsevier, in press.

Croushore, D. (2006), "Forecasting with Real-Time Macroeconomic Data," in "Handbook of Economic Forecasting," (edited by Elliot, G., C.W.J. Granger, and A. Timmermann), 961–1012, Amsterdam: North-Holland.

Diebold, F.X. (1991), "A Note on Bayesian Forecast Combination Procedures," In A. Westlund and P. Hackl (eds.) *Economic Structural Change: Analysis and Forecasting*, Springer-Verlag, 225-232.

Diebold, F.X and R.S. Mariano (1991), "Comparing Predictive Accuracy I: An Asymptotic Test," Discussion Paper 52, Institute for Empirical Macroeconomics, Federal Reserve Bank of Minneapolis.

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

Geweke, J. (2010), *Complete and Incomplete Econometric Models*, Princeton University Press.

Ghysels, E. and A. Hall (2002), "Twentieth Anniversary Commemorative Issue of the *JBES*," *Journal of Business and Economic Statistics*, 20, 1–144.

Giacomini, R. and B. Rossi (2010), "Forecast Comparisons in Unstable Environments," *Journal of Applied Econometrics*, 25, 595–620.

Giacomini, R. and H. White (2006), "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545–1578.

Hamilton, J.D. (1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57, 357–384.

Hansen, P.R., A. Lunde, and J.M. Nason (2011), "The Model Confidence Set," *Econometrica*, 79, 453–497.

Hansen, P.R. and A. Timmermann (2011), "Choice of Sample Split in Out-of-Sample Forecast Evaluation," Manuscript, Stanford and UCSD.

Inoue, A. and L. Kilian (2004), "In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?" *Econometric Reviews*, 23, 371–402.

Inoue, A. and L. Kilian (2006), "On the Selection of Forecasting Models," *Journal of Econometrics*, 130, 273–306.

Kilian, L. and M.P. Taylor (2003), "Why Is It so Difficult to Beat the Random Walk Forecast of Exchange Rates?" *Journal of International Economics*, 60, 85–107.

Li, T. (2009), "Simulation-Based Selection of Competing Structural Econometric Models," *Journal of Econometrics*, 148, 114–123.

Rapach, D.E., J.K. Strauss, and G. Zhou (2010), "Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy," *Review of Financial Studies*, 23, 821–862.

Rivers, D. and Q. Vuong (2002), "Selection Tests for Nonlinear Dynamic Models," *The Econometrics Journal*, 5, 1–39.

Rossi, B. and A. Inoue (2012), "Out-of-Sample Forecast Tests Robust to the Choice of Window Size," *Journal of Business and Economic Statistics*, 30, 432–453.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.

Vuong, Q. (1989), "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica*, 57, 307–333.

Wei, C.Z. (1992), "On Predictive Least Squares Principles," *Annals of Statistics*, 20, 1–42.

West, K.D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica*, 64, 1067–1084.

West, K.D. (2006), "Forecast Evaluation," In G. Elliott, C. Granger and A. Timmerman (eds.), *Handbook of Economic Forecasting, Volume 1*, Elsevier, 100-134.

White, H. (2000), "A Reality Check for Data Snooping," *Econometrica*, 68, 1097–1126.

Yang, Y. (2005), "Can the Strengths of AIC and BIC be Shared?" *Biometrika*, 92, 937–950.