Comparing Predictive Accuracy, Twenty Years Later:

A Personal Perspective on the Use and Abuse of Diebold-Mariano Tests*

Francis X. Diebold
University of Pennsylvania

December 5, 2013

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. Much of the large ensuing literature, however, uses DM-type tests for comparing models, in pseudo-out-of-sample environments. In that case, 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 necessarily 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 remains useful for certain tasks, perhaps most notably for providing information about comparative predictive performance during particular historical episodes.

Key words: Forecasting, model comparison, model selection, out-of-sample tests

JEL codes: C01, C53

Contact Info: F.X. Diebold, fdiebold@sas.upenn.edu

*This paper is prepared for *Journal of Business and Economic Statistics* Invited Lecture, Allied Social Science Association Meetings, Philadelphia, January 2014. It grew from a March 2012 lecture delivered 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 Todd Clark, Lutz Kilian, Mike McCracken, Andrew Patton, Barbara Rossi, Frank Schorfheide, Norman Swanson, Allan Timmermann, Ken Wolpin, and especially Peter Hansen, Minchul Shin and Jonathan Wright. The usual disclaimer most definitely applies.

1 Introduction

I was excited about the Diebold-Mariano paper (Diebold and Mariano (1995), DM) when it first circulated in 1991, almost 25 years ago. But I tend to be excited about all of my papers, so it's fascinating to watch which ones resonate in the intellectual marketplace and which ones don't. Certainly the warm reception accorded to DM was most gratifying.

The need for formal tests for comparing predictive accuracy is surely obvious. We've all seen hundreds of predictive horse races, with one or the other declared the "winner" (usually the new horse in the stable), but with no consideration given to the statistical significance of the victory. Such predictive comparisons are incomplete and hence unsatisfying. That is, in any particular realization, one or the other horse must emerge victorious, but one wants to know whether the victory is statistically *significant*. That is, one want to know whether a victory "in sample" was merely good luck, or truly indicative of a difference "in population."

If the need for predictive accuracy tests seems obvious ex post, it was not at all obvious to empirical econometricians circa 1991. Bobby Mariano and I simply noticed the defective situation, called attention to it, and proposed a rather general yet trivially-simple approach to the testing problem. And then – *boom!* – use of predictive accuracy tests exploded.

It's been a long, strange trip. In this paper I offer some perspectives on where we've been, where we are, and where we're going.

2 Comparing Model-Free Forecasts

Consider a model-free forecasting environment, as for example with forecasts based on surveys, forecasts extracted from financial markets, forecasts obtained from prediction markets, or forecasts based on expert judgment. One routinely has competing model-free forecasts of the same object, gleaned for example from surveys or financial markets, 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 $t = 1, ..., T$ 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. Diebold and Mariano (1995) propose a test 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 and forecast-error serial correlation of unknown form.

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, and to make assumptions directly on those forecast errors. (In a model-free environment there are obviously no models about which to make assumptions.) More precisely, DM relies on assumptions made directly 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.[2] That is, DM assumes that:

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

[1] The Diebold-Mariano 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 following year. In 2002 the paper was reprinted in the *JBES* Twentieth Anniversary Commemorative Issue (Ghysels and Hall (2002)).

[2] Actually covariance stationarity is sufficient but may not be strictly necessary, as less-restrictive types of mixing conditions could presumably be invoked.

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:

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

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 *must* hold.

DM is obviously extremely simple, almost embarrassingly so. It is simply an asymptotic z-test of the hypothesis that the mean of a constructed but 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 consistent 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 heteroskedasticity 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 by stage of the business cycle, in which case one might include a 0-1 NBER business cycle chronology variable (say) in the DM HAC regression.

2.2 Thoughts on Assumption *DM*

Thus far I have 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. 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

[3] Important subsequent work takes the conditional perspective farther; see Giacomini and White (2006).

are useful." Precisely the same is true 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 a useful approximation.

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 forecasters may not achieve optimality, resulting in serially-correlated, and indeed forecastable, forecast errors. But $I(1)$ non-stationarity of forecast errors takes serial correlation to the extreme.[4]

Third, even in the extreme case where nonstationary components somehow *do* exist in forecast errors, there is reason to suspect that they may be shared. In particular, 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 $L(e_{1t}) = x_t + \varepsilon_{1t}$ and $L(e_{2t}) = x_t + \varepsilon_{2t}$, 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 breaks or evolution, and so on.

[4]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 with hidden-Markov processes in the tradition of Hamilton (1989).

3 Comparing Models via Pseudo-Out-of-Sample Forecasts

DM emphasized comparing forecasts in model-free environments, but in their concluding remarks they also mentioned the possibility of DM-based model comparison. They envisioned applying the DM test in standard fashion to pseudo-out-of-sample forecast errors under Assumption DM (subject, of course, to its empirically-assessed approximate validity). The subsequent literature took a very different course, emphasizing the non-stationarity induced in loss differentials from estimated models as estimated parameters converge to their pseudo-true values. That literature eschews Assumption DM on loss differentials and replaces it with assumptions on underlying models. Here we explore the costs and benefits of the different approaches.

3.1 Unknown Models: Assumption DM is the Only Game in Town

Consider first the case of model-based forecasts, but where the models are not known to the econometrician, as for example with forecasts purchased from a vendor who uses a proprietary model. In this case of unknown models, one has only the forecast errors, so one has no choice but to proceed as in the model-free case of section 2 above. Moreover, that's not necessarily a problem. As long as Assumption DM is approximately true, the DM test is approximately valid. Of course parameter estimation may induce non-stationarity, but one might naturally conjecture that under a variety of sampling schemes of relevance in practice, parameter estimation uncertainty would be small, so that the induced nonstationarity would be small, in which case the loss differential would be approximately stationary and the DM $N(0,1)$ null distribution would be approximately valid. And of course, as also emphasized above, the importance of any non-stationarity is ultimately an empirical matter, easily checked.

3.2 Known Models I: Proceeding under Assumption DM

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

$$H_0: \; E(d_{12t}) = E\left(L(e_{1t}(F_{1t})) - L(e_{2t}(F_{2t}))\right) = 0, \tag{3}$$

where the new and slightly more detailed notation ($e_t(F_t)$ 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 subsequent researchers, in contrast, use DM and DM-type tests not for comparing forecasts, but rather for comparing fully-articulated econometric *models* (known to the researcher, who presumably specified and estimated them), 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.

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 $t = T - 1$, at each t predicting $t + 1$. Finally, base a DM-style test on the sample mean quadratic loss differential,

$$\bar{d}_{12} = \frac{\sum_{t=t^*+1}^{T} (e_{1,t/t-1}^2 - e_{2,t/t-1}^2)}{T - t^*}, \tag{4}$$

where $e_{t/t-1}$ is a time-t pseudo-out-of-sample 1-step-ahead forecast error, or "recursive residual." There are of course many variations. For example, the in-sample period could be fixed or rolling, as opposed to expanding, but (4) serves as something of a canonical benchmark.

Convergence of model parameters to their pseudo-true values as sample size grows introduces nonstationarities into forecast errors and hence into the loss differential, so it would seem that Assumption DM is violated in the model-based environment. But just as in the "unknown model" case of section 3.1, parameter estimation uncertainty might be small, so that the induced forecast-error nonstationarity would be small, in which case the loss differential would be approximately stationary and the DM $N(0, 1)$ null distribution would be approximately valid. Presumably that's why so many researchers have continued to do model comparisons using DM with $N(0, 1)$ critical values, despite the fact that they are precisely correct only under Assumption DM.

3.3 Known Models II: "Old-School" WCM

What I will call "Old-school" WCM follows the DM approach and effectively tests a null hypothesis based on the loss differential,

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

where I now write $e_{it}(F_{it}(M_i(\theta_i)))$ to emphasize that the error e_{it} is ultimately driven by a model M_i, which in turn involves a vector of pseudo-true parameters θ_i.

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 rather about the models.

Complications arise quickly in the WCM framework, 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 carefully tiptoe across a minefield of assumptions depending on the situation.[5] 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 sample(s). Is the pseudo-in-sample estimation period fixed? Recursively expanding? Rolling? Something else?

5. Estimation method. Are the models estimated by OLS? MLE? GMM? Something else? And crucially: Does the loss function embedded in the estimation method match the loss function used for pseudo-out-of-sample forecast accuracy comparison?

6. Asymptotics. What asymptotics are invoked as $T \to \infty$? $t^*/T \to 0$? $t^*/T \to \infty$? $t^*/T \to const$? Something else?

Unfortunately but not surprisingly, the relevant limiting distribution generally depends on the assumptions. Perhaps the most important WCM-style finding is that, in models with

[5] Lengthy surveys of the WCM approach, and implicitly the many varieties of "Assumption WCM," include West (2006) and Clark and McCracken (2013).

estimated parameters, the validity of the DM asymptotic $N(0,1)$ null distribution depends on the nesting structure of the models.[6] Asymptotic normality holds, for example, when non-nested models are compared, but not when nested models are compared.

3.4 Known Models III: "New-School" WCM

Old-school WCM makes assumptions on models, takes them literally, and wades through substantial mathematics to conclude that the appropriate limiting distribution depends on the model assumptions. Ironically, however, the subsequent WCM literature has recently changed course, coming full circle and implicitly arriving at the view that Assumption DM may often be credibly invoked even when comparing estimated models.

More precisely, an emerging "New-School" WCM takes a more nuanced approach using more compelling asymptotics, and, interestingly, winds up steering toward DM $N(0,1)$ critical values, as in Clark and McCracken (2011), who take a model-based approach but move away from the the sharp hypothesis of $E(d_{12t}) = 0$.[7] They consider instead the local-to-zero hypothesis $E(d_{12t}) = k/\sqrt{T}$, and they show that standard normal critical values often approximate the exact null distribution very well. Indeed, from both finite-sample and asymptotic perspectives the basic DM test (2) (the "$MSE-t$" test in their parlance), *using standard normal critical values*, features prominently. Given the ease and validity of such an approach, it's hard to imagine doing the more tedious bootstrap procedures that they also discuss.

3.5 Summation

For comparing forecasts, DM is the only game in town. There's really nothing more to say. For comparing models, with estimated parameters, the situation is more nuanced. DM-style tests are still indisputably relevant, but the issue arises as to appropriate critical values. The most obvious approach is to continue to use DM critical values, subject to empirical assessment of Assumption DM. Old-school WCM, in contrast, argues that comparison of models with estimated parameters requires a fundamentally new approach, making assumptions on models rather than forecast errors, and winds up suspicious of the DM $N(0,1)$ critical values. New-school WCM also considers comparison of models with estimated parameters, but it takes a different perspective and finds that asymptotic normality of DM is likely a trust-

[6] See Clark and McCracken (2001).

[7] See also Clark and McCracken (2013).

worthy approximation. So after twenty years, as the smoke finally clears, we see that DM with Gaussian large-sample critical values appears appropriate for forecast comparisons (as emphasized by DM), for pseudo-out-of-sample model comparisons subject to approximate empirical validity of Assumption DM as suggested by DM, *and* for pseudo-out-of-sample model comparisons subject to model assumptions as suggested by WCM.

4 Optimal Model Comparisons are Full-Sample Comparisons

Thus far I have argued that DM tests with Gaussian critical values are appropriate from both the DM and WCM perspectives. But the key issue involves not details of implementation of the WCM pseudo-out-of-sample model-comparison paradigm, but rather the paradigm itself. It is not only tedious (one must construct the pseudo-out-of-sample forecast errors), but also largely misunderstood and sup-optimal in certain important respects.

The key question is simple: Why would one *ever* want to do pseudo-out-of-sample model comparisons, as they waste data by splitting samples? For example, in comparing nested models, why not do the standard full-sample LR, LM or $Wald$ tests, with their unbeatable asymptotic optimality properties?

In this section I consider that question. The answer turns out to be that pseudo-out-of-sample model comparisons generally *are* wasteful and hence *are* dominated by full sample procedures. There are, however, many nuances.

4.1 Two Models

I begin by stepping back and extracting some basic principles of model comparison that emerge from the massive literature.

4.1.1 Optimal Finite-Sample Comparisons

I proceed by example, the first quite specialized and the second quite general. First consider the frequentist 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,

$$F_{12} = \frac{(SSR_1 - SSR_2)/(k-1)}{SSR_2/(T-k)}, \tag{6}$$

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_{12} is the uniformly most powerful test, so any other approach is sub-optimal. The key observation for our purposes is that the optimal frequentist model comparison procedure is *based on full-sample analysis, not pseudo-out-of-sample analysis.*

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 is to condition inference on the observed data $\tilde{y}_t = \{y_1, y_2, ..., y_t\}$. The Bayesian prescription for doing so is simply to select the model favored by posterior odds,

$$\underbrace{\frac{p(M_1|\tilde{y}_t)}{p(M_2|\tilde{y}_t)}}_{posterior\ odds} = \underbrace{\frac{p(\tilde{y}_t|M_1)}{p(\tilde{y}_t|M_2)}}_{Bayes\ factor} \times \underbrace{\frac{p(M_1)}{p(M_2)}}_{prior\ odds}. \tag{7}$$

As equation (7) emphasizes, however, all data-based information in the posterior odds comes from the Bayes factor, which is the ratio of marginal likelihoods. Hence if prior odds are 1:1, the Bayesian prescription is simply to select the model with higher marginal likelihood,

$$p(\tilde{y}_t|M_i) = \int p(\tilde{y}_t|\theta, M_i)p(\theta|M_i)d\theta. \tag{8}$$

A key observation for our purposes is that the optimal Bayesian model comparison procedure, like the optimal frequentist procedure, is based on full-sample analysis, not pseudo-out-of-sample analysis. An inescapable principle of Bayesian analysis is to condition inference on *all* observed data.

Thus from frequentist 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.

A second key observation, to which we will return in section 4.2, is that model selection

based on marginal likelihood, at which we have arrived, is intimately related to model selection based on predictive likelihood. By Bayes theorem the predictive likelihood is a ratio of marginal likelihoods,

$$p(y_{t+1}|\tilde{y}_t, M_i) = \frac{p(\tilde{y}_{t+1}|M_i)}{p(\tilde{y}_t|M_i)},$$

so that

$$\prod_{t=1}^{T-1} p(y_{t+1}|\tilde{y}_t, M_i) = \frac{p(\tilde{y}_T|M_i)}{p(\tilde{y}_1|M_i)}.$$

Hence Bayesian model selection is not only based on the (full-sample) marginal likelihood but also based on the (full-sample) predictive performance record.

4.1.2 Optimal Asymptotic Comparisons

Now consider asymptotic analysis, and let us stay with our consideration of 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 (7) 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 (7) can be 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

$$SIC = k\, lnT - 2\, lnL, \tag{9}$$

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: the SIC as used routinely is based on the full-sample likelihood, not a pseudo-out-of-sample predictive likelihood (and in Gaussian contexts it is based on full-sample residuals, not pseudo-out-of-sample forecast errors).

4.2 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).[8] Let us continue our asymptotic discussion from that more-compelling perspective.

SIC extends immediately to comparisons of many models; one simply selects the model with smallest SIC. SIC is of central importance because it has the key property of consistency in model selection, sometimes called the "oracle property." Consistency in model selection involves selection over a fixed set of models, and it refers to a procedure that asymptotically selects the true DGP with probability 1 (if the true DGP is among the models considered) and otherwise asymptotically selects the KLIC-optimal approximation to the true DGP with probability 1 (if the true DGP is not among the models considered, and if a unique KLIC-optimal approximation is among the models considered). The key observation for our purposes is that SIC – and its consistency property – is based on the full-sample likelihood, not on a pseudo-out-of-sample predictive likelihood.

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

$$SIC = T^{\left(\frac{k}{T}\right)} \; MSE, \tag{10}$$

where $MSE = \frac{\sum_{t=1}^{T} e_t^2}{T}$ and the e_t's are regression residuals. SIC is an estimate of out-of-sample mean squared forecast error; the key point for our purposes is that it achieves the oracle property by taking *in-sample* MSE and inflating it in just the right way to offset the in-sample MSE deflation inherent in model fitting. This is an important lesson: optimal estimators of *out-of-sample predictive* MSE are typically based on *in-sample residual* MSE (using the full sample of available data), inflated appropriately, as emphasized by Efron and Tibshirani (1993) and many others.

[8] 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 traditional Bayesian (possibly non-nested) approach also implicitly assumes that one of the models is correctly specified, as emphasized in Diebold (1991). Only recently has that assumption begun to be relaxed, as in Geweke (2010), Geweke and Amisano (2011) and Durham and Geweke (2013).

4.3 Selection vs. Testing

In closing this section, it is useful to step back and note that although criteria like 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.[9] 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."

5 More Problems with Pseudo-Out-of-Sample Model Comparisons

Several key questions remain. One is whether any pseudo-out-of-sample model comparison procedure can compete in terms of consistency with the full-sample procedures discussed above. The answer turns out to be yes, but simultaneously there are simpler procedures with the same asymptotic justification and likely-superior finite-sample properties (SIC being one example).

A second key question – perhaps *the* key question that drives the pseudo-out-of-sample model comparison literature – is whether pseudo-out-of-sample procedures can help insure against in-sample overfitting, or "data mining," *in finite samples*, quite apart from consistency in large samples. The answer is no (and the result is not new).

5.1 Pseudo-Out-of-Sample Model Comparisons are Justified Asymptotically but Have Reduced Power in Finite Samples

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, power considerations pointed to the desirability of full-sample procedures. I emphasized, moreover, that it is possible to perform model selection in ways that are asymptotically robust to data mining.

[9] See also Li (2009).

But again, in every case, 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

$$PLS = \frac{\sum_{t=t^*+1}^{T} e_{t/t-1}^2}{T - t^*}, \tag{11}$$

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 efficiency, and it appears that a procedure cannot be both consistent and efficient, as discussed in Yang (2005). So PLS has the asymptotic optimality property of consistency, but 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. This is important. In a sense PLS invokes asymptotic games. Cutting a sample in half is of no consequence asymptotically, because half of an infinite sample is still infinite, but one would nevertheless expect a clear loss in finite samples.

Based on the finite-sample power considerations invoked thus far, it's hard to imagine why one would do PLS with WCM-type inference as opposed to, say, SIC or AIC with Vuong-type inference. Hansen and Timmermann (2013) make this point very clearly in the nested case, showing precisely how the leading recursive pseudo-out-of-sample procedure of McCracken (2007) produces costly power reduction relative to the full-sample Wald statistic while simultaneously producing no offsetting benefits.

5.2 Pseudo-Out-of-Sample Model Comparisons Don't Provide Finite-Sample Insurance Against Overfitting

SIC and PLS *asymptotically* guard against in-sample overfitting – by which I mean tailoring fitted models to in-sample idiosyncrasies, effectively fitting noise rather than signal, and hence producing spuriously well-fitting models that then fail in out-of-sample forecasting – deflating in-sample MSE's with degree-of-freedom penalties harsher than those associated with traditional F and related tests.[10]

It is crucial to note, however, that all procedures under consideration, even those that achieve robustness to data mining asymptotically, are subject to strategic data mining in finite samples. This point was first made and elaborated upon by Inoue and Kilian (2004) and Inoue and Kilian (2006). Their message was true then and is true now. There is little more to say.

Introducing additional considerations, moreover, generally worsens matters for PLS/WCM. Consider, for example, and endogenous sample split point, t^*. Then 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 then also mine over t^*.

Achieving robustness to data mining in finite samples requires simulation methods that model the data mining, as in the "reality check" of White (2000). Note that one can model mining on split point as well, in addition to mining variables for a given split point, as in the bootstrap model confidence set procedure of Hansen et al. (2011). They develop methods robust to choice of in/out split point, but only at the cost of (additional) power loss. Finally, in any event there is always the issue of specifying the universe of models over which mining takes place; problems arise if it is too large or too small.

6 Whither Pseudo-Out-of-Sample Model Comparison?

The discussion thus far has cast pseudo-out-of-sample model comparisons in a bad light. Hence I now proceed to ask whether there is *any* role for pseudo-out-of-sample model com-

[10] 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 SIC in equation (10), but with penalty $\frac{T}{T-k}$. Consistency requires the much harsher penalty factor $T^{\left(\frac{k}{T}\right)}$ associated with SIC.

parison. The answer is at least a cautious yes – or maybe even an emphatic yes – for several reasons.

6.1 Interaction With Structural Break Testing

Quite apart from model comparison tests, pseudo-out-of-sample methods are a key tool for flagging structural change, as for example with $CUSUM$ and related procedures. The reason is simple: structural change produces real-time forecast breakdown. That will always remain true, so there will always be a role for recursive methods in diagnosing structural change.

Many open 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) and Giacomini and Rossi (2013). It is not yet clear where that work will lead, but there would seem to be a larger role for pseudo-out-of-sample procedures.

6.2 Pseudo-Out-of-Sample Model Comparison with Enforced "Honesty"

Let us not forget that *true* out-of-sample forecast comparisons remain an invaluable gold standard for finite-sample model comparison. The problem of course is that in time series, where the data through time T are generally public information at time T, one can strategically over-fit on any pseudo-out of sample period $t = t^* + 1, ..., T$. Hence truly-honest time-series model comparisons must use as-yet-unrealized data, and waiting for ten or twenty years is hardly appealing or realistic (although it does emphasize the desirability of performing ongoing reliability and replication studies).

In cross sections, however, the prognosis may be better. First, at least in principle, one can gather new cross-section observations at any time. Second, even when new observations are not gathered, there may be significant scope for the data-collection group truly to "hold out" a subsample to be used for subsequent predictive comparisons, as for example with Kaggle and related competitions.[11] Schorfheide and Wolpin (2013) explore such possibilities in the context of estimating treatment effects using competing structural models.

[11] See `http://www.kaggle.com/`.

6.3 Non-Standard Loss Functions

I have emphasized that even "large" samples the question remains as to why one would want to implement comparatively-tedious split-sample procedures when simpler procedures like information criteria are available. One reason is that split-sample procedures are typically more easily adapted to compare models under non-standard loss functions, such as asymmetric and multi-step. Indeed the original DM paper worked throughout with very general loss functions. Of course it emphasized forecast rather than model comparisons, but the point is that if one wants to use DM for model comparison, the set of admissible loss functions is very large. Most full-sample model-comparison procedures, in contrast, are tied closely to 1-step-ahead quadratic loss.

6.4 Comparative Historical Predictive Performance

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 methods, 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.

"Regular" full-sample model residuals may also inform about particular episodes, but they do so in a less-compelling way. Pseudo-out-of-sample residuals (recursive, rolling, split-sample, etc.) are in much closer touch with questions along the lines of "what was believed and projected, at what time, using information available at the time," which are often important in macroeconomics and financial economics.

For example, in assessing financial market efficiency it is of interest to know whether agents could have used publicly-available information in real time to out-perform a benchmark asset pricing model. The issue is naturally framed in terms of out-of-sample forecasting: Could investors have used the information and predictive technologies available in real time to make investment decisions that produce excess risk-adjusted returns?[12]

[12]See, among many others, Granger and Timmermann (2004).

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 it 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).

7 Conclusions and Directions for Future Research

Here I conclude and point to a promising direction for future work.

7.1 How Best to do, and why to do, Pseudo-out-of-Sample Model Comparisons?

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, 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 necessarily a "good" way to provide insurance against in-sample over-fitting in time-series model comparisons proves largely false. On the other hand, pseudo-out-of-sample analysis remains useful from several perspectives, including, and perhaps most importantly, providing direct information about comparative historical predictive performance.

The basic conclusion, as I see it, is two-fold:

1. If you insist on performing pseudo-out-of-sample model comparisons, then proceed if you must, but recognize that traditional DM tests, with traditional DM $N(0,1)$ critical values, are likely fine (subject to Assumption DM, of course, but as I've stressed, Assumption DM is empirically verifiable and typically reasonable.)

2. *But*: Think hard about why you're performing pseudo-out-of-sample model comparisons. They're generally costly in terms of power loss, and their benefits are unclear.

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

In particular, they don't guarantee protection against data-mining in finite samples. In general, full-sample model comparison procedures appear preferable.

7.2 Might Pseudo-out-of-Sample Procedures Have Better, if Still Imperfect, Finite-Sample Performance?

An overarching theme of this paper has been that pseudo-out-of-sample model comparisons are wasteful insofar as they come at a potentially high cost (reduced power) with no compensating benefit (all known procedures, including pseudo-out-of-sample procedures, can be "tricked" by data mining in finite samples).

The finite-sample possibility arises, however, that it may be harder, if certainly not impossible, for data mining to trick pseudo-out-of-sample procedures than to trick various popular full-sample procedures. Consider, for example, information criteria in linear regression environments, whose penalties are functions of k, the number of included regressors. For any fixed k, the penalty is irrelevant because it is identical across all k-variable models, so *such criteria will always select the most heavily-mined model*. Pseudo-out-of-sample procedures, in contrast, have the potential to be much more discriminating. In fascinating unpublished work, Hansen (2010) rigorously examines and verifies that intuition in some leading environments. Additional work along Hansen's lines may prove highly valuable.

References

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), "Nested Forecast Model Comparisons: A New Approach to Testing Equal Accuracy," Manuscript, Federal Reserve Banks of Cleveland and St. Louis.

Clark, T.E. and M.W. McCracken (2013), "Advances in Forecast Evaluation," In G. Elliott and A. Timmerman (eds.), *Handbook of Economic Forecasting, Volume 2*, Elsevier, 1107-1201.

Croushore, D. (2006), "Forecasting with Real-Time Macroeconomic Data," In G. Elliot, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting*, 961-1012.

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.

Durham, G. and J. Geweke (2013), "Improving Asset Price Prediction When all Models are False," *Journal of Financial Econometrics*, 11.

Efron, B. and R.J. Tibshirani (1993), *An Introduction to the Bootstrap*, New York: Chapman and Hall.

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

Geweke, J. and G. Amisano (2011), "Optimal Prediction Pools," *Journal of Econometrics*, 164, 130–141.

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 B. Rossi (2013), "Model Comparisons in Unstable Environments," Revised Manuscript, UCL and UPF.

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

Granger, C.W.J. and A. Timmermann (2004), "Efficient Market Hypothesis and Forecasting," *International Journal of Forecasting*, 20, 15–27.

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. (2010), "A Winners Curse for Econometric Models: On the Joint Distribution of In-Sample Fit and Out-of-Sample Fit and its Implications for Model Selection," Manuscript, Department of Economics, Stanford University.

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.

Hansen, P.R. and A. Timmermann (2013), "Equivalence Between Out-of-Sample Forecast Comparisons and Wald Statistics," Manuscript, EUI, UCSD and CREATES.

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.

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

McCracken, M.W. (2007), "Asymptotics for Out-of-Sample Tests of Granger Causality," *Journal of Econometrics*, 140, 719–752.

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.

Schorfheide, F. and K.I. Wolpin (2013), "To Hold Out or Not to Hold Out," Manuscript, Department of Economics, University of Pennsylvania.

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.