Econometrica, Vol. 74, No. 6 (November, 2006), 1545–1578

TESTS OF CONDITIONAL PREDICTIVE ABILITY

BY RAFFAELLA GIACOMINI AND HALBERT WHITE[1]

We propose a framework for out-of-sample predictive ability testing and forecast selection designed for use in the realistic situation in which the forecasting model is possibly misspecified, due to unmodeled dynamics, unmodeled heterogeneity, incorrect functional form, or any combination of these. Relative to the existing literature (Diebold and Mariano (1995) and West (1996)), we introduce two main innovations: (i) We derive our tests in an environment where the finite sample properties of the estimators on which the forecasts may depend are preserved asymptotically. (ii) We accommodate *conditional* evaluation objectives (can we predict which forecast will be more accurate at a future date?), which nest *unconditional* objectives (which forecast was more accurate on average?), that have been the sole focus of previous literature. As a result of (i), our tests have several advantages: they capture the effect of estimation uncertainty on relative forecast performance, they can handle forecasts based on both nested and nonnested models, they allow the forecasts to be produced by general estimation methods, and they are easy to compute. Although both unconditional and conditional approaches are informative, conditioning can help fine-tune the forecast selection to current economic conditions. To this end, we propose a two-step decision rule that uses current information to select the best forecast for the future date of interest. We illustrate the usefulness of our approach by comparing forecasts from leading parameter-reduction methods for macroeconomic forecasting using a large number of predictors.

KEYWORDS: Forecast evaluation, out-of-sample, hypothesis test.

1. INTRODUCTION

FORECASTING IS CENTRAL to economic decision-making. Government institutions and regulatory authorities often base policy decisions on forecasts of major economic variables, and firms rely on forecasting for inventory management and production planning decisions. A problem economic forecasters often face is how to evaluate the relative merit of two or more forecast alternatives. One answer to this problem is to develop out-of-sample tests to compare the predictive ability of competing forecasts, given a general loss function. This literature was initiated by Diebold and Mariano (1995) and further formalized by West (1996), McCracken (2000), Clark and McCracken (2001), Corradi, Swanson, and Olivetti (2001), and Chao, Corradi, and Swanson (2001), among

[1]Discussions with Clive Granger, Graham Elliott, and Andrew Patton were essential to the paper. Useful comments from a co-editor and three anonymous referees led to a considerably improved version of the paper. We also thank Lutz Kilian for insightful suggestions and Farshid Vahid, Matteo Iacoviello, Mike McCracken, and seminar participants at UCSD, Nuffield College, LSE, University of Exeter, University of Warwick, University of Manchester, Cass Business School, North Carolina State University, Boston College, Texas A&M, University of Chicago GSB, the International Finance Division of the Federal Reserve Board, University of Houston, UCLA, Harvard/MIT, and the 2002 EC^2 conference in Bologna, Italy for helpful comments. We thank Vince Crawford for the use of the UCSD Experimental and Computational Lab.

others. This work represents a generalization of previous evaluation techniques that restricted attention to a particular loss function (e.g., Granger and Newbold (1977), Leitch and Tanner (1991), West, Edison, and Cho (1993), Harvey, Leybourne, and Newbold (1997)).

In this paper, we develop a framework for out-of-sample predictive ability testing and forecast selection designed for use when the forecasting model may be misspecified. It applies to multistep point, interval, probability, or density forecast evaluation for a general loss function. Our tests are a complement to the existing approach to predictive ability testing (which in the remainder of the paper we consider to be represented by Diebold and Mariano (1995) and West (1996), henceforth referenced as DMW), and at the same time they can be viewed as extending the DMW tests because they apply in all cases in which those tests are applicable and in many more besides.

We introduce two main methodological innovations: (i) motivated by the consequences of misspecification, we consider forecasts based on limited memory estimators, whose finite sample properties are preserved asymptotically; and (ii) we formulate the problem of forecast evaluation as a problem of inference about *conditional* expectations of forecasts and forecast errors that nests the *unconditional* expectations that are the sole focus of the existing literature. We accordingly propose two tests: a general test of equal *conditional predictive ability* of two competing forecasts and, as a special case, a test of equal *unconditional predictive ability*. Although the latter coincides with the test proposed by Diebold and Mariano (1995), we provide primitive conditions that ensure its validity and extend it to an environment that permits parameter estimation.

Regardless of whether we take a conditional or an unconditional perspective, preserving the finite sample behavior of the estimators in our evaluation procedure has a number of consequences that give our tests some appealing properties. First, they directly reflect the effect of estimation uncertainty on relative forecast performance, whereas the DMW tests do not, for example, take into account differing model complexities unless they are explicitly incorporated into the loss function (e.g., Akaike information criterion and Bayesian information criterion (BIC)).[2] As a result, our object of evaluation is not simply the forecasting model as in the DMW approach, but what we call the *forecasting method*. This includes the forecasting model along with a number of choices that must be made by the forecaster at the time of the prediction and that can affect future forecast performance, such as which estimation procedure to choose and what data to use for estimation. A second advantage is that our framework permits a unified treatment of nested and nonnested models, whereas the tests of West (1996) are not applicable to nested models. The comparison between nested models is important because it is often of interest to test whether forecasts from a given model can outperform those from

[2]A recent paper by Clark and West (2005) suggests an alternative way to overcome this problem in the context of testing the martingale difference hypothesis.

a nested benchmark model. Third, we can accommodate general estimation procedures in the derivation of the forecasts, including Bayesian and semi- and nonparametric estimation methods that are excluded from the DMW framework. A final, practical advantage of our tests is that they are easily computed using standard regression software, whereas the existing tests can be difficult to compute or have limiting distributions that are context-specific (e.g., the nested test of Clark and McCracken (2001)).

Concerning our second innovation, we emphasize that we are not recommending the conditional over the unconditional approach. Rather we provide a framework in which both make sense, and it is up to the researcher to decide which is more appropriate given her objectives. The unconditional approach asks which forecast was more accurate, on average, in the past; it may thus be appropriate for making recommendations about which forecast may be better for an unspecified future date. The conditional approach asks instead whether we can use available information—above and beyond past *average* behavior—to predict which forecast will be more accurate for a specific future date. A simple analogy may be helpful in understanding this distinction. Viewing the difference in forecast performance (e.g., squared prediction error) as the dependent variable in a regression that contains only a constant, the unconditional approach is like a test for whether the regression intercept is zero, whereas the conditional approach is like a test for serial correlation in the regression errors.

In applications, one rarely has sufficient knowledge to guarantee correct specification of one's forecasting model. Instead, misspecification is common, as a result of inadequately modeled dynamics, inadequately modeled heterogeneity, incorrect functional form, or any combination of these. To accommodate each of these possible sources of misspecification, we permit but do not require the underlying data-generating process (DGP) to be heterogeneous. For example, the DGP can have structural shifts at unknown dates. When the forecasting model is misspecified, it is often the case that forecasts based on estimators using an expanding data window can be less reliable than forecasts based on estimators with a limited memory. For example, when there is inadequately modeled heterogeneity, observations from the more distant past may lose their predictive relevance. Alternatively, when dynamics are inadequately modeled, a limited memory estimator can better track a series of interest. To illustrate this last point, consider predicting an AR(1) process using a regression model that is misspecified by omitting the lagged dependent variable and including only a constant. The expanding window forecast is just the sample mean. A simple equal-weight finite moving average (MA) of the preceding data values provides a limited memory forecasting method based on the same model. Both forecasts are unbiased, but the MA predictor can often track the target of interest well, whereas the sample mean performs essentially no tracking. The simplicity of this example is not crucial. Any dynamic misspecification yields the same essential features.

Similarly, when the prediction model functional form is misspecified and the target series exhibits memory (e.g., is autocorrelated), a limited memory estimator can provide more reliable forecasts than an expanding window-based forecast, because the limited memory estimator can provide a local approximation to the prediction relationship (and therefore potentially more accurate prediction, given the memory of the target series), whereas the expanding memory estimator provides a global approximation (and therefore potentially less accurate prediction).

Accordingly, we focus on limited memory forecasting methods. One class of well known and widely applied limited memory forecasts is "rolling window" forecasts, such as those introduced by Fama and MacBeth (1973) and Gonedes (1973). Not only are these methods familiar, but they also afford considerable analytical convenience. For these reasons, our primary focus will be on rolling window methods. As straightforward as rolling window methods are, they nevertheless permit a rich variety of possibilities. For example, two rolling window methods can have different estimation windows and apply different weighting schemes within those windows. The choice of estimation window can even be data driven, as in the procedure suggested by Pesaran and Timmermann (2006), so that two competing forecasting methods can use different data-driven window choice procedures. In any given application, it is an empirical matter as to whether a limited memory or an expanding memory method provides better forecasts. Our methods can provide direct evidence on this point, as demonstrated in our empirical example in Section 7, where we see numerous examples of limited memory predictors outperforming expanding window predictors.

Although our main focus is on rolling window methods, our results are also valid for a "fixed estimation sample" forecasting scheme, which involves estimating the models' parameters only once over the in-sample data and using these to produce all out-of-sample forecasts.

A final, important implication of our approach is that it provides a basis to make forecast selection decisions in cases where equal (conditional) predictive ability is rejected. As an example, we propose a simple decision rule for forecast selection based on the idea that, because rejection means that the relative performance of the competing forecasts is predictable, we should exploit current information to predict which forecast will be more accurate in the future.

To illustrate the usefulness of our approach, we consider, from both the conditional and the unconditional perspectives, the problem of macroeconomic forecasting using a large number of predictors and compare multistep forecasts of four macroeconomic variables (two measures of real activity and two price indexes) obtained by leading methods for parameter reduction: a simplified version of the general-to-specific model selection approach of Hoover and Perez (1999), the "diffusion indexes" approach of Stock and Watson (2002), and the use of Bayesian shrinkage estimators (Litterman (1986)). These forecasts cannot be compared using any previous method. We conclude that for

the price indexes, these methods are no better than a simple autoregression, whereas for the real variables, Bayesian shrinkage is the best performing method. The simplified general-to-specific method is characterized by an overall poor performance.

2. A NEW APPROACH TO OUT-OF-SAMPLE PREDICTIVE ABILITY TESTING

In this section, we set forth our approach and discuss the main differences between our approach and previous approaches to out-of-sample predictive ability testing.

2.1. *Null Hypothesis and Asymptotic Framework*

Suppose one wants to compare the accuracy of competing forecasts $f_t(\beta_1)$ and $g_t(\beta_2)$ for the τ-steps-ahead variable $Y_{t+\tau}$, using a loss function $L_{t+\tau}(\cdot)$. The DMW approach tests

$$H_0 : E\left[L_{t+\tau}(Y_{t+\tau}, f_t(\beta_1^*)) - L_{t+\tau}(Y_{t+\tau}, g_t(\beta_2^*))\right] = 0, \tag{1}$$

where β_1^* and β_2^* are population values (i.e., probability limits of the parameter estimates). This makes (1) a statement about the *forecasting models*: H_0 says that the models are equally accurate on average. A key feature of West's (1996) test of H_0 is the recognition and accommodation of the fact that, although H_0 concerns population values, the actual forecasts that appear in the test statistic depend on estimated parameters.

Our central idea is to test a null hypothesis that differs from the DMW null in two respects: (i) the losses depend on estimates $\hat{\beta}_{1t}$ and $\hat{\beta}_{2t}$, rather than on their probability limits; and (ii) the expectation is conditional on some information set $\mathcal{G}_t$:

$$H_0 : E\left[L_{t+\tau}(Y_{t+\tau}, f_t(\hat{\beta}_{1t})) - L_{t+\tau}(Y_{t+\tau}, g_t(\hat{\beta}_{2t}))|\mathcal{G}_t\right] = 0. \tag{2}$$

The focus on parameter estimates makes (2) a statement about the *forecasting methods*, which include the models as well as the estimation procedures and the possible choices of estimation window (note that the two forecasts may use different estimation windows). Our null says that one cannot predict which forecasting method will be more accurate at the forecast target date $t + \tau$ using the information in $\mathcal{G}_t$.

Regardless of the choice of $\mathcal{G}_t$, expressing the null in terms of parameter estimates is useful because it allows us to capture the impact of estimation uncertainty on relative forecast performance. For example, by comparing expected *estimated* mean squared forecast errors (MSE), rather than their population counterparts, we accommodate the possibility of a bias–variance trade-off such that forecasts from a small, misspecified model (biased with low variance)

are as accurate as forecasts from a large, correctly specified model (unbiased with high variance). Because of its focus on the forecasting model rather than the forecasting method, the DMW approach cannot accommodate such a trade-off. This emphasizes the distinction between evaluation of a forecasting method, which is a practical matter, and evaluation of a forecasting model, which may be appropriate for obtaining economic insight, but is less informative for prediction purposes.

An implication of testing different null hypotheses is that the tests of (1) and (2) are analyzed in different out-of-sample asymptotic environments. Whereas the test of West (1996) is analyzed in an environment where parameter estimates converge to their population values, we operate in an environment with asymptotically nonvanishing estimation uncertainty. This ensures that our tests capture the impact of estimation uncertainty on forecast performance. Furthermore, as we discuss in detail in Section 3.2, this has the important advantage that our tests can handle nested and nonnested models in a unified framework.

We achieve nonvanishing estimator uncertainty by considering estimators with limited memory, in particular, rolling window estimators, a method popular among practitioners ever since its influential use by Fama and MacBeth (1973) and Gonedes (1973). Limited memory estimators are especially appropriate in the misspecified predictor environments considered here, because they discount or exclude older data that may either no longer be informative about the predictive relationships of current interest or prevent a dynamically misspecified model from tracking well. Other relevant limited memory estimators are recursive estimators of the exponential smoothing type or, as suggested by a referee, expanding window weighted least squares estimators with weights that more heavily discount less recent observations. We work explicitly with rolling window estimators, not only because of their popularity among practitioners, but also for two further reasons: first, this approach affords significant generality, because it imposes no restrictions on the estimators other than finite memory, whereas the alternatives are comparatively specific; second, the analysis required for this approach is straightforward, whereas that for the alternatives is more involved, but has no compensating increase in insight.

Regarding the choice of the conditioning set $\mathcal{G}_t$, a leading case of interest is $\mathcal{G}_t = \mathcal{F}_t$, the time-$t$ information set. Another possibility is $\mathcal{G}_t = \{\emptyset, \Omega\}$, the trivial σ-field, which yields a test of equal *unconditional* predictive ability. The choice of the relevant conditioning set will depend on the objectives of the evaluator. Letting $\mathcal{G}_t = \{\emptyset, \Omega\}$ seems appropriate if the goal is to provide a forecast for an unspecified date in the future, in which case it makes sense to base recommendations on which forecast may be better on average. If, on the other hand, the goal is to produce a forecast for a specific date τ periods in the future, choosing $\mathcal{G}_t = \mathcal{F}_t$ may be more appropriate, because it allows us to ask whether there is additional current information that can help predict which forecast will be more accurate for that date. Conditioning (i.e., letting

$\mathcal{G}_t \neq \{\emptyset, \Omega\}$) when testing relative forecast performance is important, because it is plausible with misspecification to expect some predictability in future loss differences. For example, the relative performance may be characterized by persistence, so that if a forecast outperforms its competitor today, it may be likely to do so tomorrow. In this case, past loss differences may predict future loss differences. We may also expect the performance of certain models to depend on the state of the economy, so that a business cycle indicator may tell us which forecast is preferable for a future date, given current economic conditions.

Even though our framework nests both conditional and unconditional objectives, for succinctness we refer to a test of (2) as a test of equal *conditional predictive ability*.

2.2. *Data Assumptions*

One of the conclusions of Clements and Hendry (1998, 1999) is that the main explanation for systematic forecast failure in economics is the use of models that are inadequate to handle the nonconstant data-generating processes that govern real-world economic data. Specific sources of heterogeneity in economic series are several, including changes in the measurement process, changes in laws, and changes in technology. Any failure to model this heterogeneity will result in misspecification. As previously remarked, incorrect functional form or omission of lags (either own lags or lags of predictively relevant variables) also yields a misspecified prediction model. To accommodate each of these possible sources of misspecification, we operate in a data environment that permits but does not require data heterogeneity.[3]

3. THEORY

3.1. *Description of the Environment*

Consider a stochastic process $W \equiv \{W_t : \Omega \to \mathbb{R}^{s+1}, s \in \mathbb{N}, t = 1, 2, \dots\}$ defined on a complete probability space $(\Omega, \mathcal{F}, P)$. We partition the observed vector W_t as $W_t \equiv (Y_t, X_t')'$, where $Y_t : \Omega \to \mathbb{R}$ is the variable of interest and $X_t : \Omega \to \mathbb{R}^s$ is a vector of predictor variables, and we define $\mathcal{F}_t = \sigma(W_1', \dots, W_t')'$ (cf., White (1994, p. 96)).

We focus for simplicity on univariate forecasts. Suppose two alternative models are used to forecast the variable of interest τ steps ahead, $Y_{t+\tau}$. The (point, interval, probability, or density) forecasts formulated at time t are based on the information set $\mathcal{F}_t$ and are denoted by $\hat{f}_{t,m_f} \equiv f(W_t, W_{t-1}, \dots, W_{t-m_f+1})$;

[3]The type of nonstationarity we consider here is that induced by distributions that change over time. We also assume short memory, thus ruling out nonstationarity due to the presence of unit roots.

$\hat{\beta}_{t,m}$) and $\hat{g}_{t,m_g} \equiv g(W_t, W_{t-1}, \ldots, W_{t-m_g+1}; \hat{\beta}_{t,m})$, where f and g are measurable functions. The subscripts indicate that the time-t forecasts are measurable functions of a sample of size m_f for f and of size m_g for g. If the forecasts are based on parametric models, the parameter estimates from the two models are collected in the $k \times 1$ vector $\hat{\beta}_{t,m}$, where $m \equiv \max(m_f, m_g)$. Otherwise, $\hat{\beta}_{t,m}$ represents whatever semiparametric or nonparametric estimators are used to construct the forecasts. We allow general estimation procedures. We only require that the estimation window size is bounded.

We view m_f and m_g as either method-specific constants or as possibly time-dependent random integers determined by the forecasting method. For technical convenience, we require that $m \leq \bar{m}$, a finite constant (this can be relaxed, but at the cost of an explosion of technicality). For example, data-driven choices for m_f and m_g are given by the procedure suggested by Pesaran and Timmermann (2006). The requirement that $\bar{m}$ be finite rules out an expanding window forecasting scheme. In principle, however, our framework can also handle expanding estimation window procedures with observation weights that suitably discount older observations, such as exponential smoothers, with smoothing parameter bounded away from zero.

We produce the forecasts using a rolling window estimation scheme. Let T be the total sample size and let m_1 be the maximum size of the first estimation window.[4] We formulate the first τ-step-ahead forecasts at time m_1 using data indexed $1, \ldots, m_1$ and compare these forecasts to the realization $y_{m_1+\tau}$. At time $m_1 + 1$, we formulate the second set of forecasts using the previous m_2 observations (m_2 can be different from m_1) and compare them to the realization $y_{m_1+1+\tau}$. Iterating this procedure yields $n \equiv T - \tau - m_1 + 1$ out-of-sample forecasts and relative forecast errors.

Note that the requirement that $\bar{m}$ be finite is also compatible with a fixed estimation sample forecasting scheme, where the parameters are estimated only once on the first m_1 observations and used to produce all n out-of-sample forecasts (in which case $\hat{\beta}_{t,m} = \hat{\beta}_{m_1,m_1}$, $m_1 \leq t \leq T - 1$).

The preceding elements—the model, the estimation procedure, the size of estimation window, and any applied observation weights—are part of each forecasting method under evaluation.

We evaluate the sequence of out-of-sample forecasts by a loss function $L_{t+\tau}(Y_{t+\tau}, \hat{f}_{t,m_f})$ that is either an economically meaningful criterion, such as utility or profits (e.g., Leitch and Tanner (1991), West, Edison, and Cho (1993)), or a statistical measure of accuracy. Examples of loss functions for point forecasts considered in the literature and covered by our theory are squared error loss, absolute error loss, lin–lin loss, linex loss, direction-of-change loss, and predictive log-likelihood. Loss functions for quantile, prob-

[4]If m_f and m_g are time dependent, say $m_f = \{m_{ft}\}$ and $m_g = \{m_{gt}\}$, then $m_1 = \max(m_{f1}, m_{g1})$, $m_2 = \max(m_{f2}, m_{g2})$, etc., and we write $m = \max(m_1, m_2, \ldots)$.

ability, and density forecasts are discussed, e.g., in Diebold and Lopez (1996), Giacomini and Komunjer (2005), and Amisano and Giacomini (2006).

For a given loss function and σ-field $\mathcal{G}_t$, we write the null hypothesis of equal conditional predictive ability of forecasts f and g for the target date $t+\tau$ as

$$H_0 : E[L_{t+\tau}(Y_{t+\tau}, \hat{f}_{t,m_f}) - L_{t+\tau}(Y_{t+\tau}, \hat{g}_{t,m_g})|\mathcal{G}_t] \equiv E[\Delta L_{m,t+\tau}|\mathcal{G}_t] = 0 \quad \text{almost surely} \quad t = 1, 2, \ldots. \tag{3}$$

In writing (3), we are adopting the convention that $\hat{f}_{t,m_f}$ and $\hat{g}_{t,m_g}$ are measurable-$\mathcal{F}_t$. Note that we do not require $\mathcal{G}_t = \mathcal{F}_t$, although this is a leading case of interest that we analyze in the next two sections. We separately address the case $\mathcal{G}_t = \{\emptyset, \Omega\}$ in a subsequent section.

3.2. *One-Step Conditional Predictive Ability Test*

When $\tau = 1$ and $\mathcal{G}_t = \mathcal{F}_t$, the null hypothesis (3) claims that $\{\Delta L_{m,t}, \mathcal{F}_t\}$ is a martingale difference sequence (MDS). In this case, the conditional moment restriction (3) is equivalent to stating that $E[\tilde{h}_t \Delta L_{m,t+1}] = 0$ for all $\mathcal{F}_t$-measurable functions $\tilde{h}_t$. We restrict attention to a given subset of such functions, which we denote by the $q \times 1$ $\mathcal{F}_t$-measurable vector h_t and follow Stinchcombe and White (1998) by referring to this as the *test function*. For a given choice of test function h_t, we construct a test that exploits the consequence of the MDS property that $H_{0,h} : E[h_t \Delta L_{m,t+1}] = 0$.

Standard asymptotic normality arguments suggest using a Wald-type test statistic of the form

$$T^h_{m,n} = n\left(n^{-1}\sum_{t=m}^{T-1} h_t\,\Delta L_{m,t+1}\right)' \hat{\Omega}_n^{-1}\left(n^{-1}\sum_{t=m}^{T-1} h_t\,\Delta L_{m,t+1}\right) = n\bar{Z}'_{m,n}\hat{\Omega}_n^{-1}\bar{Z}_{m,n}, \tag{4}$$

where $\bar{Z}_{m,n} \equiv n^{-1}\sum_{t=m}^{T-1} Z_{m,t+1}$, $Z_{m,t+1} \equiv h_t\,\Delta L_{m,t+1}$, and $\hat{\Omega}_n \equiv n^{-1}\sum_{t=m}^{T-1} Z_{m,t+1} \times Z'_{m,t+1}$ is a $q \times q$ matrix that consistently estimates the variance of $Z_{m,t+1}$.

A level α test can be conducted by rejecting the null hypothesis of equal conditional predictive ability whenever $T^h_{m,n} > \chi^2_{q,1-\alpha}$, where $\chi^2_{q,1-\alpha}$ is the $(1-\alpha)$ quantile of a χ^2_q distribution. The asymptotic justification for the test is provided in the following theorem, which characterizes the behavior of the test statistic (4) under the null hypothesis.

THEOREM 1—One-Step Conditional Predictive Ability Test: *For forecast horizon* $\tau = 1$, (*maximum*) *estimation window size* $m \le \bar{m} < \infty$, *and* $q \times 1$ *test function sequence* $\{h_t\}$, *suppose* (i) $\{W_t\}$ *and* $\{h_t\}$ *are mixing with* ϕ *of size* $-r/(2r-1)$, $r \ge 1$, *or* α *of size* $-r/(r-1)$, $r > 1$; (ii) $E|Z_{m,t+1,i}|^{2(r+\delta)} < \infty$ *for*

some $\delta > 0$, $i = 1, \ldots, q$ *and for all* t; (iii) $\Omega_n \equiv n^{-1} \sum_{t=m}^{T-1} E[Z_{m,t+1} Z'_{m,t+1}]$ *is uniformly positive definite. Then, under* H_0 *in* (3), $T^h_{m,n} \xrightarrow{d} \chi^2_q$ *as* $n \to \infty$.

COMMENT 1: Assumption (i) is mild, allowing the data to be characterized by considerable heterogeneity as well as dependence. This is in contrast to the existing literature, which typically assumes stationarity of the loss differences. In particular, we allow the data to be characterized by arbitrary structural changes at unknown dates.

COMMENT 2: The asymptotic distribution is obtained for the number of out-of-sample observations n going to infinity, whereas the maximum estimation sample size m is finite. This leads to asymptotically nonvanishing estimation uncertainty. In contrast, in the framework of West (1996), both the in-sample and the out-of-sample sizes grow, causing estimation uncertainty to vanish asymptotically. As a result, in the DMW framework the choice of how to split the sample into in-sample and out-of-sample portions is arbitrary, whereas here the choice of estimation window is part of each forecasting method under evaluation.

COMMENT 3: Expanding window forecasting schemes are ruled out by assumption.

COMMENT 4: Assumption (iii), which imposes positive definiteness of the asymptotic variance of the test statistic, is related to a similar requirement made in the existing predictive ability testing literature (e.g., West (1996), McCracken (2000)), but it differs in a fundamental way. There, the asymptotic variance is computed at the probability limits of the parameters, which may cause singularity when the forecasts are based on nested models. Here, the nonvanishing estimation uncertainty prevents such singularity and thus makes our tests applicable to both nested and nonnested models.

COMMENT 5: In the construction of the test statistic, we exploit the simplifying feature that the null hypothesis imposes the time dependence structure of a MDS, which implies that the asymptotic variance can be consistently estimated by the sample variance. As suggested by a referee, one could instead use a heteroscedasticity and autocorrelation consistent (HAC) estimator (e.g., Andrews (1991)) in the construction of the test. This leaves the asymptotic distribution of the test statistic under the null hypothesis unchanged and results in a test with correct size. We prefer to exploit the MDS structure, however, because it not only yields a simpler test, but it may also increase power. The reason for this is that the asymptotic power depends on the asymptotic variance; the smaller is the variance, the more powerful is the test. If, as is often plausible under the alternative, there is positive autocorrelation in the loss differences that the HAC estimator accounts for, then the HAC estimator will be larger and the asymptotic power will be correspondingly lower.

COMMENT 6: As pointed out by a referee, the same theory outlined in Theorem 1 can be applied to testing for conditional bias, efficiency, and encompassing, provided the assumptions of the theorem are satisfied. One simply replaces $\Delta L_{m,t+1}$ with a suitable function of Y_{t+1} and the forecasts. Conditional encompassing for quantile forecasting is explored by Giacomini and Komunjer (2005).

COMMENT 7: It is easy to show that the test statistic $T_{m,n}^h$ can be alternatively computed as nR^2, where R^2 is the uncentered squared multiple correlation coefficient for the artificial regression of the constant unity on $(h_t \Delta L_{m,t+1})'$. Under the additional assumption of conditional homoscedasticity of $\Delta L_{m,t+1}$, the test can be based on the test statistic nR^2, where R^2 is the uncentered squared multiple correlation coefficient for the artificial regression of $\Delta L_{m,t+1}$ on h_t'.

3.2.1. *Alternative hypothesis*

We now analyze the behavior of the test statistic $T_{m,n}^h$ under a form of global alternative to H_0. Because we do not require identical distributions, we must exercise care in specifying the global alternative in this context. In fact, our test is consistent against

$$H_{A,h}: E[\bar{Z}'_{m,n}]E[\bar{Z}_{m,n}] \geq \delta > 0 \quad \text{for all } n \text{ sufficiently large.} \tag{5}$$

The following theorem characterizes the behavior of $T_{m,n}^h$ under the global alternative $H_{A,h}$.

THEOREM 2: *Given assumptions* (i), (ii), *and* (iii) *of Theorem* 1, *under* $H_{A,h}$ *in* (5) *and for any constant* $c \in \mathbb{R}$, $P[T_{m,n}^h > c] \to 1$ *as* $n \to \infty$.

Note that H_0 and $H_{A,h}$ are exhaustive under stationarity, but are not necessarily exhaustive under heterogeneity. For a given choice of $\{h_t\}$, with heterogeneity it may happen that $E[\bar{Z}'_{m,n'}]E[\bar{Z}_{m,n'}] = 0$ for some sequence $\{n'\}$, without $\{\Delta L_{m,t+1}\}$ being a MDS and thus the test may have no power against alternatives for which $\Delta L_{m,t+1}$ is correlated with some element of $\mathcal{F}_t$ that is not contained in h_t. This is not an issue with stationarity. The flexibility in the choice of test function is both a shortcoming and an advantage of our testing framework. On the one hand, for a given h_t the test may have no power against possibly important alternatives. On the other, one can choose which h_t is more relevant in any situation and thus focus power in that direction.

In practice, h_t is chosen by the researcher to include variables that are thought to help distinguish between the forecast performance of the two methods. Some examples are indicators of past relative performance (lagged loss differences or moving averages of past loss differences) or business cycle indicators that may capture possible asymmetries in relative performance during

booms and recessions. When choosing the number of elements for h_t, keep in mind that the properties of the test will be altered if either too few or too many elements are included. If h_t leaves out elements of the information set $\mathcal{F}_t$ that are correlated with $\Delta L_{m,t+1}$, the test may incorrectly "accept" a false null hypothesis. On the other hand, the inclusion of a number of elements that are either uncorrelated or weakly correlated with $\Delta L_{m,t+1}$ will in some sense dilute the significance of the important elements and thus erode the power of the test. A possible way to confront this difficulty is to apply the approaches advocated by Bierens (1990) or Stinchcombe and White (1998) that deliver consistent tests.

3.3. *Multistep Conditional Predictive Ability Test*

For a forecast horizon $\tau > 1$ and with $\mathcal{G}_t = \mathcal{F}_t$, the null hypothesis (3) implies that for all $\mathcal{F}_t$-measurable test functions h_t, the sequence $\{h_t \Delta L_{m,t+\tau}\}$ is "finitely correlated," so that $\operatorname{cov}(h_t \Delta L_{m,t+\tau}, h_{t-j} \Delta L_{t+\tau-j}) = 0$ for all $j \geq \tau$. Similarly to the previous section, we exploit this simplifying feature in the construction of the test statistic. Using reasoning that mirrors the development of the test for the one-step horizon, we consider the test statistic

$$
\begin{aligned}
(6) \qquad T^h_{m,n,\tau} &= n \left(n^{-1} \sum_{t=m}^{T-\tau} h_t \Delta L_{m,t+\tau} \right)' \tilde{\Omega}_n^{-1} \left(n^{-1} \sum_{t=m}^{T-\tau} h_t \Delta L_{m,t+\tau} \right) \\
&= n \bar{Z}'_{m,n} \tilde{\Omega}_n^{-1} \bar{Z}_{m,n},
\end{aligned}
$$

where h_t is a $q \times 1$ $\mathcal{F}_t$-measurable test function, $\bar{Z}_{m,n} \equiv n^{-1} \sum_{t=m}^{T-\tau} Z_{m,t+\tau}$, $Z_{m,t+\tau} \equiv h_t \Delta L_{m,t+\tau}$, and $\tilde{\Omega}_n \equiv n^{-1} \sum_{t=m}^{T-\tau} Z_{m,t+\tau} Z'_{m,t+\tau} + n^{-1} \sum_{j=1}^{\tau-1} w_{n,j} \times \sum_{t=m+j}^{T-\tau} [Z_{m,t+\tau} Z'_{m,t+\tau-j} + Z_{m,t+\tau-j} Z'_{m,t+\tau}]$, where $w_{n,j}$ is a weight function such that $w_{n,j} \to 1$ as $n \to \infty$ for each $j = 1, \ldots, \tau - 1$ (e.g., Newey and West (1987) and Andrews (1991)).

A level α test rejects the null hypothesis of equal conditional predictive ability whenever $T^h_{m,n,\tau} > \chi^2_{q,1-\alpha}$, where $\chi^2_{q,1-\alpha}$ is the $(1-\alpha)$ quantile of a χ^2_q distribution. The following result is the equivalent of Theorems 1 and 2 for the multistep forecast horizon case.

THEOREM 3—Multistep Conditional Predictive Ability Test: *For given forecast horizon $\tau > 1$, (maximum) estimation window size $m \leq \bar{m} < \infty$, and a $q \times 1$ test function sequence $\{h_t\}$, suppose* (i) *$\{W_t\}$ and $\{h_t\}$ are mixing with ϕ of size $-r/(2r-2)$, $r \geq 2$, or α of size $-r/(r-2)$, $r > 2$;* (ii) *$E|Z_{m,t+\tau,i}|^{r+\delta} < \infty$ for some $\delta > 0$, $i = 1, \ldots, q$ and for all t;* (iii) *$\Omega_n \equiv n^{-1} \sum_{t=m}^{T-\tau} E[Z_{m,t+\tau} Z'_{m,t+\tau}] + n^{-1} \sum_{j=1}^{\tau-1} \sum_{t=m+j}^{T-\tau} (E[Z_{m,t+\tau} Z'_{m,t+\tau-j}] + E[Z_{m,t+\tau-j} Z'_{m,t+\tau}])$ is uniformly positive definite. Then* (a) *under H_0 in (3), $T^h_{m,n,\tau} \xrightarrow{d} \chi^2_q$ as $n \to \infty$ and* (b) *under $H_{A,h}$ in (5), for any constant $c \in \mathbb{R}$, $P[T^h_{m,n,\tau} > c] \to 1$ as $n \to \infty$.*

3.4. *Multistep Unconditional Predictive Ability Test*

When $\mathcal{G}_t$ is the trivial σ-field $\mathcal{G}_t = \{\emptyset, \Omega\}$ and for forecast horizon $\tau \geq 1$, the null hypothesis (3) can be viewed as a test of equal unconditional predictive ability of forecasting methods f and g, $H_0 : E[\Delta L_{m,t+\tau}] = 0$, $t = 1, 2, \ldots$, against the alternative

$$H_A : |E[\Delta \bar{L}_{m,n}]| \geq \delta > 0 \quad \text{for all } n \text{ sufficiently large,} \tag{7}$$

where $\Delta \bar{L}_{m,n} \equiv n^{-1} \sum_{t=m}^{T-\tau} \Delta L_{m,t+\tau}$. The test is based on the statistic

$$t_{m,n,\tau} = \frac{\Delta \bar{L}_{m,n}}{\hat{\sigma}_n / \sqrt{n}}, \tag{8}$$

where $\hat{\sigma}_n^2$ is a suitable HAC estimator of the asymptotic variance $\sigma_n^2 = \operatorname{var}[\sqrt{n} \Delta \bar{L}_{m,n}]$, for example, $\hat{\sigma}_n^2 \equiv n^{-1} \sum_{t=m}^{T-\tau} \Delta L_{m,t+\tau}^2 + 2[n^{-1} \sum_{j=1}^{p_n} w_{n,j} \times \sum_{t=m+j}^{T-\tau} \Delta L_{m,t+\tau} \Delta L_{m,t+\tau-j}]$, with $\{p_n\}$ a sequence of integers such that $p_n \to \infty$ as $n \to \infty$, $p_n = o(n)$, and $\{w_{n,j} : n = 1, 2, \ldots; j = 1, \ldots, p_n\}$ a triangular array such that $|w_{n,j}| < \infty$, $n = 1, 2, \ldots$, $j = 1, \ldots, p_n$, and $w_{n,j} \to 1$ as $n \to \infty$ for each $j = 1, \ldots, p_n$ (cf. Andrews (1991)).

A level α test rejects the null hypothesis of equal unconditional predictive ability whenever $|t_{m,n,\tau}| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the $(1 - \alpha/2)$ quantile of a standard normal distribution. The test statistic $t_{m,n,\tau}$ coincides with that proposed by Diebold and Mariano (1995).

THEOREM 4 —Unconditional Predictive Ability Test: *For given forecast horizon* $\tau \geq 1$ *and* (*maximum*) *estimation window size* $m \leq \bar{m} < \infty$, *suppose* (i) $\{W_t\}$ *is mixing with* ϕ *of size* $-r/(2r-2)$, $r \geq 2$, *or* α *of size* $-r/(r-2)$, $r > 2$; (ii) $E|\Delta L_{m,t+\tau}|^{2r} < \infty$ *for all* t; (iii) $\sigma_n^2 \equiv \operatorname{var}[\sqrt{n} \Delta \bar{L}_{m,n}] > 0$ *for all* n *sufficiently large*. *Then* (a) *under* H_0 *in* (3), $t_{m,n,\tau} \xrightarrow{d} N(0,1)$ *as* $n \to \infty$ *and* (b) *under* H_A *in* (7), *for any constant* $c \in \mathbb{R}$, $P[|t_{m,n,\tau}| > c] \to 1$ *as* $n \to \infty$.

Note that, whereas for the conditional test the truncation lag for the HAC estimator is $p_n = \tau - 1$, for the unconditional test we require $p_n \to \infty$ as $n \to \infty$; thus in practice this must be selected by the user. The reason is that the unconditional null hypothesis, unlike the conditional null hypothesis, does not impose any particular dependence structure on the loss differences. Because the loss differences are mixing variables, a HAC estimator with $p_n \to \infty$ is needed for consistency. Nevertheless, in practical applications it is often the case that short truncation lags improve the finite-sample properties of the Diebold and

Mariano (1995) test (see, e.g., Clark (1999)).[5] Our simulations in Section 5 provide additional evidence on this point.

4. A DECISION RULE FOR FORECAST SELECTION

In this section, we consider the implications of rejecting equal conditional predictive ability and describe a method for adaptively selecting at time T a forecasting method for $T+\tau$. The basic idea is that rejection occurs because the test functions $\{h_t\}$ can predict the loss differences $\{\Delta L_{m,t+\tau}\}$ out of sample, which suggests using h_T to predict which method will yield lower loss at $T+\tau$. We propose the following two-step procedure:

STEP 1: Regress $\Delta L_{m,t+\tau} = L_{t+\tau}(Y_{t+\tau}, \hat{f}_{t,m_f}) - L_{t+\tau}(Y_{t+\tau}, \hat{g}_{t,m_g})$ on h_t over the out-of-sample period $t = m, \ldots, T-\tau$ and let $\hat{\delta}_n$ denote the regression coefficient. Apply one of the tests from Section 3 and, in case of rejection, proceed to Step 2.

STEP 2: The approximation $\hat{\delta}_n' h_T \approx E[\Delta L_{m,t+\tau}|\mathcal{F}_T]$ motivates the decision rule: use g if $\hat{\delta}_n' h_T > c$ and use f if $\hat{\delta}_n' h_T < c$, with c a user-specified threshold (e.g., $c = 0$).

This procedure is a simple example of how our tests can be used in forecast selection. More sophisticated approaches immediately suggest themselves, but the subject of forecast selection is a significant topic that deserves extensive attention beyond that possible in the space available here.

In general, the plot of out-of-sample period predicted loss differences $\{\hat{\delta}_n' h_t\}_{t=m}^{T-\tau}$ is useful for assessing the relative performance of f and g at different times. One can further summarize relative out-of-sample performance by computing the proportion of times the foregoing decision rule chooses g, i.e., $I_{n,c} = n^{-1} \sum_{t=m}^{T-\tau} \mathbb{1}\{\hat{\delta}_n' h_t > c\}$, where $\mathbb{1}\{A\}$ equals 1 if A is true and 0 otherwise. We report these proportions for our empirical application in Section 6.

5. MONTE CARLO EVIDENCE

We investigate the size and power properties of the tests of conditional and unconditional predictive ability in finite samples of the sizes typically available in macroeconomic forecasting applications.

[5]Diebold and Mariano (1995) also acknowledge that τ-step-ahead errors may not be $(\tau-1)$ dependent, but find that the assumption of $(\tau-1)$ dependence works well in practical applications and suggest using it as a benchmark. In the remainder of the paper, we adopt this approach.

5.1. *Size Properties*

The goal of our first Monte Carlo experiment is twofold: first, to consider a situation where our null hypothesis of equal forecasting *method* accuracy is satisfied when comparing nested models and, second, to contrast our test with tests for equal forecasting *model* accuracy previously available (McCracken (1999) and Clark and McCracken (2001)). We highlight the flexibility of our approach by presenting results for both a quadratic and a linex loss function. For comparability, we restrict attention in this subsection to the unconditional test and to the one-step forecast horizon. Solely for simplicity and brevity, we take $m = m_f = m_g$.

The idea is to consider a situation where the trade-off between misspecification and parameter estimation uncertainty is such that forecasts from a small, misspecified model are as accurate as those from a larger, correctly specified model. Thus, let the data-generating process be

$$Y_t = c + \mathrm{CPI}_t + \varepsilon_t, \quad \varepsilon_t \sim \text{i.i.d.}\, N(0, \sigma^2), \tag{9}$$

where CPI_t is the second log difference of the monthly U.S. consumer price index over the period 1959:1–1998:12. We use an actual time series to create data that exhibit realistic behavior. The two competing forecasting models are $M1: Y_t = \beta \mathrm{CPI}_t + u_{1t}$ and $M2: Y_t = \delta + \gamma \mathrm{CPI}_t + u_{2t}$. Note that $M1$ is misspecified in that it omits the intercept. The one-step-ahead forecasts of Y_{t+1} implied by the two models are, respectively,

$$\hat{f}^{(1)}_{t,m} = \hat{\beta}_{t,m} \mathrm{CPI}_{t+1} \quad \text{and} \quad \hat{f}^{(2)}_{t,m} = \hat{\delta}_{t,m} + \hat{\gamma}_{t,m} \mathrm{CPI}_{t+1}, \tag{10}$$

estimated by ordinary least squares (OLS) over a sample of size m. Here and in the following text, we treat CPI as known (i.e., CPI_{t+1} belongs to $\mathcal{F}_t$).

For each m and n pair in the range $(25, 75, 125, 150)$, we find values of c in (9) such that the two forecasting methods have equal expected MSE, using the following result:

PROPOSITION 5: *Let* $X_t \equiv \mathrm{CPI}_t$; $\bar{X} \equiv \frac{1}{m} \sum_{j=t-m+1}^{t} X_j$, $S_{xx} \equiv \sum_{j=t-m+1}^{t} X_j^2 - m\bar{X}^2$, $\sum_t \equiv \sum_{t=m}^{T-1}$, *and* $\sum_j \equiv \sum_{j=t-m+1}^{t}$. *If*

$$\begin{aligned} c = \sigma \Bigg(\Bigg\{ \sum_t \Bigg(\Bigg(\sum_j X_j^2 / m S_{xx} \Bigg) + X_{t+1}^2 / S_{xx} \\ - 2(\bar{X}/S_{xx}) X_{t+1} - X_{t+1}^2 \Big/ \sum_j X_j^2 \Bigg) \Bigg\} \\ \times \Bigg\{ \sum_t \Bigg(1 - \Bigg(\sum_j X_j \Big/ \sum_j X_j^2 \Bigg) X_{t+1} \Bigg)^2 \Bigg\}^{-1} \Bigg)^{1/2}, \end{aligned} \tag{11}$$

then $E[\frac{1}{n}\sum_t L(Y_{t+1}, \hat{f}^{(1)}_{t,m})] = E[\frac{1}{n}\sum_t L(Y_{t+1}, \hat{f}^{(2)}_{t,m})]$ *for* $L(Y_{t+1}, f) = (Y_{t+1} - f)^2$.

Using c from Proposition 5, $\sigma = 0.1$, and the last $T = m + n$ CPI observations, we generate 5,000 Monte Carlo replications of Y_t from (9) and compute rolling window forecasts as in (10). Note that we obtain a different c for each (m, n) pair. Also note that in this design the null of equal predictive ability only holds on average over $t = m, \ldots, T$.

To examine the robustness of the size properties of our test to the choice of loss function and to illustrate the flexibility of our method, we further consider a linex loss function. We generate 5,000 replications of Y_t from (9) as previously described, using values of c such that the two forecasting methods have equal expected average linex loss, obtained as follows:

PROPOSITION 6: *Using the notation of Proposition 5, if* c *solves* $F(c) = 0$, *where*

$$F(c) \equiv \sum_t \left\{ \exp\left[c\left(1 - \frac{\sum_j X_j}{\sum_j X_j^2} X_{t+1}\right) + \frac{\sigma^2}{2}\left(1 + \frac{X_{t+1}^2}{\sum_j X_j^2}\right)\right] - c\left(1 - \frac{\sum_j X_j}{\sum_j X_j^2} X_{t+1}\right) - \exp\left[\frac{\sigma^2}{2}\left(1 + \frac{\sum_j X_j^2}{m S_{xx}} + \frac{X_{t+1}^2}{S_{xx}} - 2\frac{\bar{X}}{S_{xx}} X_{t+1}\right)\right]\right\}, \tag{12}$$

then $E[\frac{1}{n}\sum_t L(Y_{t+1}, \hat{f}^{(1)}_{t,m})] = E[\frac{1}{n}\sum_t L(Y_{t+1}, \hat{f}^{(2)}_{t,m})]$ *for* $L(Y_{t+1}, f) = e^{Y_{t+1}-f} - (Y_{t+1} - f) - 1$.

We find values of c that solve the equation in Proposition 6 by numerical techniques. Table I reports the rejection frequencies of the hypotheses of equal forecasting method accuracy using quadratic and linex loss for a 5% nominal level using the test of Theorem 6. The truncation lag for the HAC estimator is $p_n = 0$.[6] For the quadratic loss, the table also shows the rejection frequencies for the test of equal forecasting model accuracy of McCracken (1999) and Clark and McCracken (2001) (henceforth the CM test), which relies on the same test statistic but uses critical values obtained by simulation from a nonstandard asymptotic distribution. For linex loss, the CM test cannot be applied because it requires the same loss function for estimation and evaluation, whereas we estimate by OLS and not by linex maximum likelihood.

[6]We also considered selecting p_n using either the data-dependent method of Andrews (1991) or the popular simple alternative $p_n = 0.75n^{1/3}$, which satisfies Andrews' (1991) optimal rate condition. The results, available upon request, suggest these alternative choices lead to slightly worse size properties, even though in the majority of cases Andrews' method selected $p_n = 0$ as the optimal bandwidth.

TABLE I

REJECTION FREQUENCIES OF UNCONDITIONAL PREDICTIVE ABILITY AND MCCRACKEN'S (1999) TESTS[a]

	A. Quadratic Loss								B. Linex Loss			
	Uncond. Pred. Ability				McCracken (1999)				Uncond. Pred. Ability			
	n				n				n			
m	25	75	125	150	25	75	125	150	25	75	125	150
25	0.053	0.037	0.035	0.024	0.087	0.360	0.481	0.525	0.060	0.055	0.046	0.046
75	0.062	0.048	0.040	0.037	0.147	0.070	0.256	0.279	0.069	0.065	0.064	0.058
125	0.073	0.054	0.044	0.042	0.120	0.146	0.063	0.199	0.072	0.070	0.075	0.077
150	0.061	0.056	0.048	0.046	0.091	0.134	0.204	0.058	0.075	0.075	0.077	0.073

[a]Rejection frequencies of the test of Theorem 6 and of McCracken's (1999) test in the Monte Carlo experiment described in Section 5.1, for nominal size 0.05: m is the estimation window size and n is the out-of-sample size.

The table reveals that our test is generally well sized, particularly when the estimation window m is small relative to the out-of-sample size n (for given m, the size tends to improve as n increases). This is true for both quadratic and linex loss functions, although for the linex loss the test is slightly oversized. Before discussing the rejection frequencies of the CM test, we emphasize that these do not represent the empirical size of the CM test, because this tests a different null hypothesis: for CM the losses are functions of population values of the parameters rather than parameter estimates, so the CM test is focused on the forecasting model rather than the forecasting method. Table I shows that in our scenario the CM test rejects the hypothesis that the forecasting models are equally accurate in favor of the larger model[7] more often than our test rejects its null hypothesis. In other words, by rejecting its null hypothesis relatively more frequently, the CM test signals that the larger forecasting model is superior in cases where the forecasting *method* based on the larger model is *not* superior. The disparity of conclusions between the two tests is greater when m is small relative to n (our test rejects 5% of the time, whereas the CM test rejects up to 50% of the time). Interestingly, the two tests have comparable rejection frequencies when m is equal to n.

5.2. *Power Properties*

We next investigate the power of our unconditional and conditional tests in two directions: (i) against serially correlated loss differences; and (ii) against different performance in different states of the economy. Again, solely for simplicity and brevity, we take $m = m_f = m_g$.

[7]The alternative hypothesis for the CM test is that the larger model is more accurate.

5.2.1. *Power against serial correlation in relative performance*

Here we consider the alternative that the loss differences $\Delta L_{m,t+1}$ follow an AR(1) process:

$$\Delta L_{m,t+1} = \mu(1-\rho) + \rho \Delta L_{m,t} + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim \text{i.i.d.}\, N(0,1). \tag{13}$$

For each of 5,000 Monte Carlo replications, we use (13) to generate a sequence of loss differences of length $n = 150$ starting from an initial value $\Delta L_{m,m}$ that equals the difference in squared errors for forecasts of $\text{CPI}_{1998:12}$ implied by (i) a white noise and (ii) an AR(1) model for CPI estimated over a window of size $m = 150$ using data up to 1998:11. We consider two scenarios: (I) the loss differences are not serially correlated ($\rho = 0$) but have nonzero unconditional mean; (II) the loss differences have zero unconditional mean ($\mu = 0$) but are serially correlated (and thus the unconditional null hypothesis is still satisfied). The corresponding parameterizations are (i) $\rho = 0$, $\mu = (0, 0.05, \ldots, 1)$ and (ii) $\mu = 0$, $\rho = (0, 0.05, \ldots, 0.9)$.

Figure 1 shows the power curves of the tests of Theorems 1 (conditional) and 6 (unconditional) in scenarios (I) and (II) computed as the proportion of rejections of the null hypotheses $H_{0,\text{cond}}$ and $H_{0,\text{unc}}$ at the 5% nominal level. In all cases, we let $h_t = (1, \Delta L_{m,t})'$ for the conditional test and $p_n = 0$ for the unconditional test.

The left panel of Figure 1 reveals that using the conditional rather than the unconditional test, even though there is no serial correlation in the loss differences, involves only a small loss of power. From the right panel of Figure 1, on the other hand, we see that the conditional test has appealing power properties but that the unconditional test suffers severe size distortions as the loss

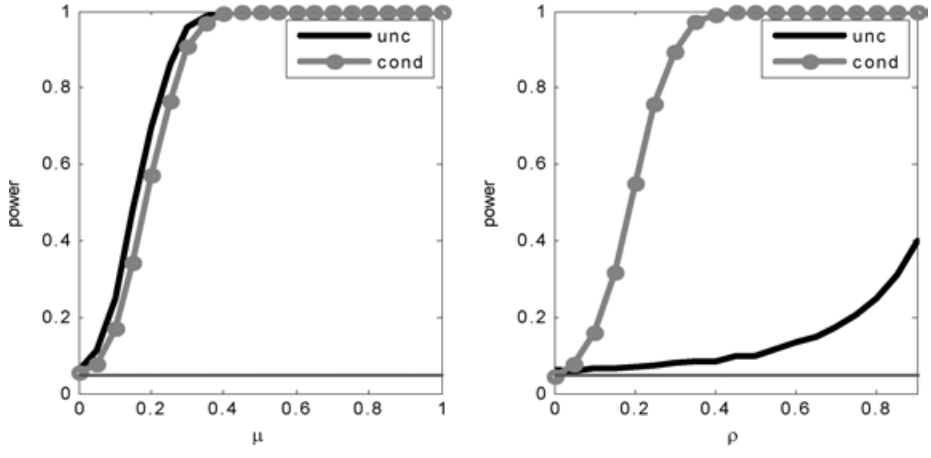

FIGURE 1.—Power curves for the conditional test of Theorem 1 and the unconditional test of Theorem 6. The DGP in the left panel is such that $E[\Delta L_{m,t+1}|F_t] = \mu$ and the DGP in the right panel is such that $E[\Delta L_{m,t+1}] = 0$, but $E[\Delta L_{m,t+1}|F_t] = (\Delta L_{m,t})$.

differences become more serially correlated (the power curve is upward sloping, whereas it should be flat because $H_{0,\text{unc}}$ is satisfied), a possible consequence of not using a more involved method for choosing p_n.

5.2.2. *Power against different performance in different states*

We next consider a situation where the two forecasts have equal predictive ability unconditionally, but each forecast is more accurate in a given state of the economy. For each of 5,000 Monte Carlo replications, we generate a sequence of loss differences of length $n = 150$ as

$$\Delta L_{m,t+1} = \frac{\mu}{p(1-p)}(S_t - p) + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim \text{i.i.d.}\, N(0,1),$$

where $S_t = 1$ with probability p and $S_t = 0$ with probability $1-p$. We thus have $E[\Delta L_{m,t+1}] = 0$, but

$$E[\Delta L_{m,t+1}|S_t] = \begin{cases} \mu/p, & \text{if } S_t = 1 \\ -\mu/(1-p), & \text{if } S_t = 0, \end{cases}$$

so that the second forecast is more accurate in the first state and the first forecast is more accurate in the second state. Figure 2 shows the rejection frequencies of the null hypotheses $H_{0,\text{cond}}$ and $H_{0,\text{unc}}$ at the 5% nominal level using the tests of Theorems 1 and 6. The power curves are obtained for $p = 0.5$ and $d \equiv \frac{\mu}{p(1-p)} = (0, 0.1, \ldots, 1)$ (d represents the difference in expected loss between the two states). We let $h_t = (1, S_t)'$ for the conditional test and let $p_n = 0$ for the unconditional test.

As expected, the conditional test has power to detect different performance in the different states, whereas the rejection frequencies for the unconditional

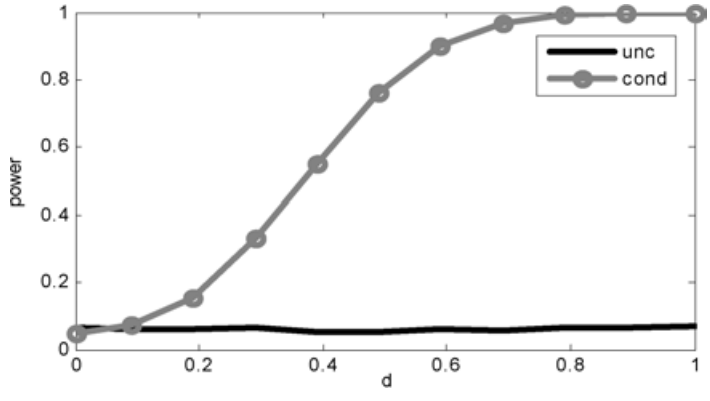

FIGURE 2.—Power curves for the conditional test of Theorem 1 and the unconditional test of Theorem 6. The DGP is such that $E[\Delta L_{m,t+1}] = 0$, but $E[\Delta L_{m,t+1}|F_t] = d(S_t - p)$, where $S_t = 1$ with probability p and is 0 otherwise.

test remain constant at the empirical size. Unlike the previous case, the unconditional test does not suffer size distortion.

6. APPLICATION: COMPARING PARAMETER-REDUCTION METHODS

A problem that often arises in macroeconomic forecasting is how to select a manageable subset of predictors from a large number of potentially useful variables. In this situation, one key determinant of the resulting forecast performance is the trade-off between the information content of each series and the estimation uncertainty that is introduced. The goal of our application is to analyze and compare the forecast performance, both conditionally and unconditionally, of three leading methods for parameter reduction: a sequential model-selection approach based on a simplified general-to-specific modeling strategy (Hoover and Perez (1999)), the "diffusion indexes" approach of Stock and Watson (2002), and the use of Bayesian shrinkage estimation (Litterman (1986)). We also compare each method to autoregressive and random walk benchmark forecasts. The DMW testing framework cannot be used here because some of the comparisons are between nested models and, furthermore, that framework does not easily accommodate Bayesian estimation or the presence of estimated regressors. In contrast, our approach is well suited for comparing methods based on nested models or on different modeling and estimation techniques.

We consider the "balanced panel" subset of the data set of Stock and Watson (2002) (henceforth SW), including 146 monthly economic time series measured over the period 1959:1–1998:12, and apply the same transformations as those documented in Appendix B of SW. We use the different parameter-reduction methods to construct multistep forecasts for four[8] U.S. macroeconomic variables: two real variables (industrial production and real personal income less transfers) and two price indexes (consumer price index and producer price index).

6.1. *Parameter-Reduction Methods*

All forecasting models project the τ-step-ahead variable $Y_{t+\tau}^{\tau}$ onto time-t predictors X_t and lags of the variable of interest $Y_t, Y_{t-1}, \ldots$. We consider the following forecasting methods.

The sequential model selection method (denoted Seq.) considers the model

$$Y_{t+\tau}^{\tau} = \alpha + \beta' X_t + \gamma_1 Y_t + \cdots + \gamma_6 Y_{t-5} + \varepsilon_{t+\tau}, \tag{14}$$

where X_t contains the 145 predictors, and applies a simplified version of the algorithm described by Hoover and Perez (1999, p. 175), which reduces the

[8]Results for additional series are available at http://www.econ.ucla.edu/giacomin/CPAappendix.pdf.

number of regressors by performing a sequence of stability tests, residual autocorrelation tests, and t- and F-tests of significance.[9] The significance level for all tests is $\alpha = 0.01$. A complete algorithm description is available upon request.

The diffusion indexes method (denoted DI) first uses principal component analysis to estimate k factors $\hat{F}_t$ from the predictors X_t ($1 \le k \le 12$) and then considers the model $Y_{t+\tau}^{\tau} = \alpha + \beta' \hat{F}_t + \gamma_1 Y_t + \cdots + \gamma_p Y_{t-p+1} + \varepsilon_{t+\tau}$, where both k and p are selected by BIC.

The Bayesian shrinkage method (denoted Bay) considers the full model (14) and applies Bayesian estimation of its coefficients using the Litterman (1986) prior. For variables in differences, the variance V for the prior distribution of $\theta \equiv (\alpha, \beta', \gamma')'$ is diagonal, with $\alpha \sim N(0, 10^8)$, $\beta_i \sim N(0, (w \cdot \lambda \cdot \hat{\sigma}_y / \hat{\sigma}_{x_i})^2)$, $i = 1, \ldots, 145$, and $\gamma_j \sim N(0, (\lambda / j))^2)$, $j = 1, \ldots, 6$. As in Litterman (1986), we set $w = 0.2$ and $\lambda = 0.2$, but the results were robust to a number of different choices for w and λ. The Bayesian estimate of θ is $\theta^B = (X'X + \hat{\sigma}^2 V^{-1})^{-1}(X'Y^{\tau})$, where X is $m \times 152$ (m is the size of the estimation sample) with rows $(1, X_t', Y_t, Y_{t-1}, \ldots, Y_{t-5})$, Y^{τ} is $m \times 1$ with elements $Y_{t+\tau}^{\tau}$, and $\hat{\sigma}$ is the estimated standard error of the residuals in a univariate autoregression for $Y_{t+\tau}^{\tau}$.

The benchmark methods are an autoregressive (denoted AR) model $Y_{t+\tau}^{\tau} = \alpha + \gamma_1 Y_t + \cdots + \gamma_p Y_{t-p+1} + \varepsilon_{t+\tau}$, where p is selected by BIC with $0 \le p \le 6$, and a random walk (denoted RW) in levels, corresponding to the forecasting model in differences $Y_{t+\tau}^{\tau} = \alpha + \varepsilon_{t+\tau}$.

6.2. *Real-Time Forecasting Experiment*

We use the preceding five methods to produce sequences of τ-step-ahead forecasts for $\tau = 1, 6, 12$ using a rolling window estimation procedure with $m = m_f = m_g = 150 + \tau$. The first estimation sample is from 1960:1–1972:6 + τ (the first 12 data were used as initial observations), the total sample has size $T = 468$, and the out-of-sample size is $n = 318 - \tau$.

At the outset, we described how limited memory estimators can have advantages relative to expanding memory procedures, especially in the presence of inadequately modeled heterogeneity, inadequately modeled dynamics, or incorrect functional form. To gain quantitative insight, one can compare the estimated loss from using a limited memory estimator (e.g., a rolling window estimator) to that of an expanding data window procedure. We do not provide a formal test based on this comparison here. Instead, however, we examine the relative performance of these different approaches by comparing the performance of forecasts of industrial production and consumer price index for

[9] We overcome multicollinearity in X_t by replacing the groups of variables whose correlation is greater than 0.98 with their average. The new X_t contains 130 regressors.

TABLE II

RELATIVE MSE OF ROLLING AND EXPANDING WINDOW FORECASTS[a]

	Industrial Production					Consumer Price Index				
τ	Seq.	DI	Bay	AR	RW	Seq.	DI	Bay	AR	RW
1 month	3.38	0.79	0.75	1.02	0.84	0.15	1.02	0.96	1.04	1.00
6 months	0.02	0.85	0.53	1.01	0.41	0.02	1.04	0.15	1.03	1.00
12 months	0.03	0.66	0.12	1.07	0.26	0.01	1.00	0.11	1.04	1.01

[a]Ratios of MSEs of τ-steps-ahead forecasts for the methods in the column estimated over either a rolling window of size $m = 150$ or an expanding window with the same initial size.

all models, and comparing forecast horizons based on rolling window methods to forecasts based on an expanding window of data from 1960:1 onward. Table II reports the relative MSEs of the rolling window and expanding window forecasts.

The table shows that MSEs for rolling window forecasts are often much smaller than those for expanding window forecasts (ratios are as small as 0.01). In the remaining cases, the MSEs for the two procedures are virtually identical (with one exception, ratios are no greater than 1.07).[10] We see that the rolling window procedure can result in substantial forecast accuracy gains relative to an expanding window for important economic time series.

6.3. *Results of Predictive Ability Tests*

For each forecast series we conduct pairwise tests of equal conditional predictive ability of the five forecasting methods using a squared error loss (results for absolute error loss are available on request). For $\tau = 1, 6$, and 12, we test $H_0 : E[(Y_{t+\tau} - \hat{f}_{t,m_f})^2 - (Y_{t+\tau} - \hat{g}_{t,m_g})^2 | \mathcal{G}_t] \equiv E[\Delta L_{t+\tau} | \mathcal{G}_t] = 0$ for $\mathcal{G}_t = \mathcal{F}_t$ (conditional test) and $\mathcal{G}_t = \{\emptyset, \Omega\}$ (unconditional test).

For the case $\mathcal{G}_t = \mathcal{F}_t$, we use the test function $h_t = (1, \Delta L_t)'$. Table III shows the results of conditional predictive ability tests for real variables and price indexes. Table IV shows the results for the unconditional case. The entries in the tables are the p-values of pairwise tests of equal conditional and unconditional predictive ability, using the tests of Theorems 5 and 6. In Table III, the numbers within parentheses below each entry are the indicators $I_{n,c}$ discussed in Section 4, for $c = 0$. A plus (minus) sign indicates rejection of the null hypothesis at the 10% level and signals that the method in the column would have been chosen more (less) often than the method in the row, as suggested by an entry $I_{n,c}$ greater (less) than 0.5. In Table IV, the numbers within parentheses are the ratios of MSEs for the method in the column relative to the method in

[10]Note that these results are for a fixed choice of estimation window. Optimizing the estimation window size could produce even greater improvements.

TABLE III

CONDITIONAL PREDICTIVE ABILITY TESTS[a]

	Industrial Production					Personal Income					CPI					Producer Price Index				
Bench	Seq.	DI	Bay	AR	RW	Seq.	DI	Bay	AR	RW	Seq.	DI	Bay	AR	RW	Seq.	DI	Bay	AR	RW
									A. Horizon = 1 month											
Bound	0.043	0.080	0.004	0.004	0.016	0.016	0.008	0.012	0.054	0.008	0.193	0.496	0.584	0.496	0.452	0.003	0.004	0.134	0.004	0.327
DI	0.026^{-} (0.00)				[0.010]	0.004^{-} (0.02)				[0.020]	0.175 (0.00)				[1.00]	0.001^{-} (0.02)				[0.009]
Bay	0.020^{-} (0.00)	0.080^{-} (0.02)				0.006^{-} (0.02)	0.731 (0.90)				0.146 (0.01)	0.644 (0.99)				0.046^{-} (0.01)	0.067^{+} (0.99)			
AR	0.034^{-} (0.00)	0.040^{+} (0.91)	0.001^{+} (0.93)			0.027^{-} (0.03)	0.018^{+} (0.97)	0.199 (0.93)			0.192 (0.00)	0.124 (0.74)	0.721 (0.18)			0.001^{-} (0.02)	0.161 (0.22)	0.047^{-} (0.01)		
RW	0.043^{-} (0.00)	0.049^{+} (0.81)	0.004^{+} (0.84)	0.163 (0.78)		0.108 (0.07)	0.002^{+} (0.86)	0.003^{+} (0.83)	0.023^{+} (0.80)		0.193 (0.00)	0.385 (0.84)	0.389 (0.78)	0.452 (0.80)		0.240 (0.01)	0.109 (1.00)	0.578 (0.76)	0.098^{+} (0.98)	
									B. Horizon = 6 months											
Bound	0.012	0.040	0.012	0.228	0.152	0.035	0.072	0.037	0.080	0.096	0.364	0.004	0.008	0.004	0.003	0.003	0.003	0.003	0.003	0.009
DI	0.010^{-} (0.05)				[0.030]	0.018^{-} (0.00)				[0.140]	0.091^{-} (0.00)				[0.009]	0.001^{-} (0.00)				[0.007]
Bay	0.003^{-} (0.01)	0.432 (0.00)				0.014^{-} (0.00)	0.037^{-} (0.00)				0.261 (0.00)	0.002^{+} (0.99)				0.493 (0.51)	0.001^{+} (0.99)			
AR	0.885 (0.01)	0.167 (0.98)	0.057^{+} (0.98)			0.031^{-} (0.00)	0.098^{+} (0.93)	0.020^{+} (1.00)			0.146 (0.00)	0.082^{+} (0.82)	0.055^{-} (0.02)			0.001^{-} (0.00)	0.809 (0.82)	0.001^{-} (0.01)		
RW	0.654 (0.30)	0.154 (0.98)	0.038^{+} (0.99)	0.193 (0.97)		0.035^{-} (0.00)	0.124 (1.00)	0.024^{+} (0.99)	0.591 (0.90)		0.935 (0.00)	0.001^{+} (1.00)	0.004^{+} (0.96)	0.001^{+} (1.00)		0.554 (0.95)	0.003^{+} (1.00)	0.404 (0.97)	0.003^{+} (1.00)	
									C. Horizon = 12 months											
Bound	0.004	0.012	0.004	0.116	0.124	0.012	0.024	0.012	0.112	0.164	0.200	0.003	0.004	0.004	0.003	0.003	0.002	0.003	0.002	0.003
DI	0.003^{-} (0.00)				[0.010]	0.006^{-} (0.00)				[0.030]	0.050^{-} (0.00)				[0.008]	0.001^{-} (0.00)				[0.005]
Bay	0.001^{-} (0.00)	0.201 (0.00)				0.003^{-} (0.00)	0.044^{-} (0.01)				0.271 (0.00)	0.001^{+} (0.98)				0.367 (0.00)	0.001^{+} (1.00)			
AR	0.029^{-} (0.02)	0.202 (0.96)	0.088^{+} (0.99)			0.028^{-} (0.00)	0.314 (0.94)	0.095^{+} (1.00)			0.224 (0.00)	0.059^{+} (0.98)	0.634 (0.06)			0.001^{-} (0.00)	0.152 (0.83)	0.001^{-} (0.01)		
RW	0.031^{-} (0.05)	0.174 (1.00)	0.073^{+} (1.00)	0.873 (0.01)		0.041^{-} (0.00)	0.227 (0.92)	0.113 (0.99)	0.096^{+} (0.85)		0.557 (0.03)	0.001^{+} (1.00)	0.005^{+} (0.99)	0.001^{+} (0.99)		0.484 (0.08)	0.001^{+} (1.00)	0.187 (0.96)	0.001^{+} (1.00)	

[a]Results of pairwise tests of equal conditional predictive ability for the forecast methods described in Section 6.1. The entries are the p-values of the test of equal conditional predictive ability of Theorem 5 for the forecast methods in the corresponding row and column. The loss is quadratic and the test function is $h_t = (1, \Delta L_{m,t})'$. The numbers within parentheses are the proportion of times the method in the column outperforms the method in the row over the out-of-sample period, according to the decision rule described in Section 4. A plus (minus) sign indicates that the test rejects equal conditional predictive ability at the 10% level and that the method in the column outperforms (is outperformed by) the method in the row more than 50% of the time. For example, for industrial production at the 1-month horizon, equal conditional predictive ability of the Bayesian shrinkage and the AR methods is rejected with a p-value of 0.001 and the Bayesian shrinkage method outperforms the AR method 93% of the time. The rows labeled "Bound" report the Hochberg–Bonferroni (HB) multiple hypothesis p-value bound for the method in the column relative to all other methods. The square brackets [ ] contain the HB p-value bound for the hypothesis that all pairwise comparisons are zero for that panel.

TABLE IV

UNCONDITIONAL PREDICTIVE ABILITY TESTS[a]

	Industrial Production					Personal Income					CPI					Producer Price Index				
Bench	Seq.	DI	Bay	AR	RW	Seq.	DI	Bay	AR	RW	Seq.	DI	Bay	AR	RW	Seq.	DI	Bay	AR	RW
									A. Horizon = 1 month											
Bound	0.065	0.072	0.008	0.008	0.116	0.160	0.040	0.174	0.040	0.156	0.229	0.472	0.635	0.635	0.362	0.012	0.016	0.063	0.016	0.120
DI	0.036^{-} (3.82)				[0.020]	0.055^{-} (1.83)				[0.100]	0.181 (3.45)				[1.00]	0.004^{-} (1.71)				[0.036]
Bay	0.029^{-} (4.22)	0.028^{-} (1.10)				0.054^{-} (1.79)	0.554 (0.98)				0.196 (3.19)	0.472 (0.93)				0.054^{-} (1.31)	0.019^{+} (0.77)			
AR	0.046^{-} (3.37)	0.027^{+} (0.88)	0.002^{+} (0.80)			0.099 (1.64)	0.010^{-} (0.89)	0.091^{+} (0.91)			0.186 (3.36)	0.287 (0.97)	0.635 (1.05)			0.004^{-} (1.75)	0.462 (1.03)	0.021^{-} (1.34)		
RW	0.065^{-} (2.97)	0.083^{+} (0.78)	0.029^{+} (0.70)	0.227 (0.88)		0.160 (1.50)	0.039^{+} (0.82)	0.058^{+} (0.84)	0.184 (0.92)		0.229 (2.81)	0.301 (0.81)	0.261 (0.88)	0.362 (0.84)		0.030^{-} (1.29)	0.108 (0.75)	0.823 (0.98)	0.102 (0.73)	
									B. Horizon = 6 months											
Bound	0.004	0.044	0.004	0.232	0.240	0.039	0.040	0.040	0.100	0.156	0.392	0.003	0.003	0.004	0.002	0.003	0.003	0.003	0.003	0.012
DI	0.011^{-} (1.67)				[0.010]	0.026^{-} (6.23)				[0.100]	0.098^{-} (2.67)				[0.007]	0.001^{-} (2.30)				[0.007]
Bay	0.001^{-} (1.83)	0.294 (1.10)				0.022^{-} (7.37)	0.010^{-} (1.18)				0.306 (1.65)	0.001^{+} (0.62)				0.988 (1.00)	0.001^{+} (0.43)			
AR	0.680 (1.11)	0.175 (0.67)	0.058^{+} (0.61)			0.037^{-} (4.78)	0.149 (0.77)	0.025^{+} (0.65)			0.145 (2.27)	0.143 (0.85)	0.003^{-} (1.37)			0.001^{-} (2.24)	0.801 (0.98)	0.001^{-} (2.25)		
RW	0.921 (1.03)	0.156 (0.62)	0.060^{+} (0.56)	0.145 (0.93)		0.039^{-} (4.62)	0.196 (0.74)	0.057^{+} (0.63)	0.655 (0.97)		0.742 (1.15)	0.001^{+} (0.43)	0.001^{+} (0.70)	0.001^{+} (0.51)		0.476 (0.84)	0.004^{+} (0.37)	0.188 (0.84)	0.003^{+} (0.37)	
									C. Horizon = 12 months											
Bound	0.003	0.004	0.004	0.105	0.108	0.003	0.004	0.004	0.068	0.102	0.180	0.003	0.003	0.004	0.002	0.003	0.003	0.003	0.003	0.012
DI	0.001^{-} (3.77)				[0.009]	0.001^{-} (2.97)				[0.009]	0.045^{-} (3.63)				[0.007]	0.001^{-} (2.76)				[0.007]
Bay	0.001^{-} (4.03)	0.442 (1.07)				0.001^{-} (3.42)	0.009^{-} (1.15)				0.095^{-} (2.48)	0.001^{+} (0.68)				0.339 (1.22)	0.001^{+} (0.44)			
AR	0.035^{-} (1.80)	0.064^{+} (0.48)	0.034^{+} (0.45)			0.017^{-} (2.08)	0.076^{+} (0.70)	0.025^{+} (0.61)			0.068^{-} (2.79)	0.110 (0.77)	0.389 (1.12)			0.001^{-} (2.52)	0.516 (0.91)	0.001^{-} (2.07)		
RW	0.036^{-} (1.83)	0.063^{+} (0.49)	0.032^{+} (0.46)	0.772 (1.02)		0.034^{-} (1.95)	0.083^{+} (0.65)	0.033^{+} (0.57)	0.212 (0.93)		0.344 (1.52)	0.001^{+} (0.42)	0.001^{+} (0.61)	0.001^{+} (0.54)		0.642 (0.88)	0.004^{+} (0.32)	0.075^{+} (0.72)	0.003^{+} (0.35)	

[a]Results of pairwise tests of equal unconditional predictive ability for the forecast methods described in Section 6.1. The entries are the p-values of the test of equal unconditional predictive ability of Theorem 6 for the forecast methods in the corresponding row and column. The loss is quadratic and the truncation lag for the HAC estimator is $\tau - 1$, where τ is the forecast horizon. The numbers within parentheses are the ratios of MSEs for the method in the column relative to the method in the row. A plus (minus) sign indicates that the test rejects equal unconditional predictive ability at the 10% level and that the method in the column has smaller (larger) MSE than the method in the row. For example, for industrial production at the 1-month horizon, equal unconditional predictive ability of the Bayesian shrinkage and the AR methods is rejected with a p-value of 0.002 and the Bayesian shrinkage method outperforms the AR method with a MSE ratio of 0.8. The rows labeled "Bound" report the Hochberg–Bonferroni (HB) multiple hypothesis p-value bound for the method in the column relative to all other methods. The square brackets [ ] contain the HB p-value bound for the hypothesis that all pairwise comparisons are zero for that panel.

the row and a plus (minus) sign indicates that the method in the column outperforms (underperforms) the method in the row at the 10% significance level, as evidenced by a relative MSE less (greater) than 1. The rows labeled "Bound" contain Hochberg's (1988) modified Bonferroni p-value bounds for testing the multiple hypothesis that all pairwise comparisons are zero for a given column reference method.[11] The square brackets contain Hochberg's (1988) p-value bound for the hypothesis that all pairwise comparisons are zero for that panel.

A sharp result that emerges from the tables is that the sequential model-selection method is characterized by the worst performance, likely due to its tendency to select overparameterized models (cases with 40 or more predictors in the final model were not uncommon). A second observation is that the predictors seem less useful for forecasting price indexes than real variables. For price indexes, the parameter-reduction methods do not generally outperform the AR benchmark. For real variables, both Bayesian shrinkage and the diffusion indexes methods mostly outperform the benchmarks. Bayesian shrinkage, however, often outperforms the diffusion indexes, thus emerging as the best forecasting method for real variables. Note that the use of Hochberg–Bonferroni modified p-values does not, in general, change the conclusions that emerge from the pairwise tests.

Finally, we draw two conclusions from the comparison of the results for the conditional and the unconditional tests. First, in some of the comparisons there is evidence of superior conditional performance even though we cannot reject equal unconditional performance (e.g., diffusion indexes versus AR forecasts of CPI). This suggests that in those cases, even though the two methods performed on average equally well, their relative performance could have been predicted by lagged relative performance. A second conclusion is that even though rejection of the unconditional hypothesis should imply rejection of the conditional hypothesis, in some cases the unconditional tests reject equal performance while the conditional tests fail to do so. This could be due either to the unconditional test being oversized or to the conditional test having low power. Our Monte Carlo simulations suggest that the more plausible explanations are the mild size distortions of the unconditional test and the test's sensitivity to lag length selection for the HAC estimator.

6.4. *Decision Rule Assessment*

To assess the effectiveness of the decision rule proposed in Section 4, we evaluate the performance of the "hybrid" forecast obtained by recursively applying the decision rule to select the best forecast for the next period. We consider the sequence of quadratic out-of-sample losses for 1-, 6-, and

[11] Hochberg's (1988) method involves ordering the p-values from testing r hypotheses as $p_{(1)}, \ldots, p_{(r)}$ and computing the bound as $\text{Bound} = \min_{j=1,\ldots,r}(r - j + 1)\, p_{(j)}$.

TABLE V

DECISION RULE ASSESSMENT: PERFORMANCE OF THE "HYBRID" FORECAST OF INDUSTRIAL PRODUCTION[a]

	Horizon = 1 month				Horizon = 6 months				Horizon = 12 months			
Bench	Seq.	DI	Bay	AR	Seq.	DI	Bay	AR	Seq.	DI	Bay	AR
DI	1				0				1			
Bayes	1	1			1	1			1	1		
AR	1	1	0		1	1	1		1	1	1	
RW	1	1	0	1	0	1	1	1	1	1	1	1

[a]Entries equal 1 if the MSE of the hybrid forecast (see Section 6.4) is less than or equal to the MSEs of both the method in the row and the method in the column, and they equal 0 otherwise.

12-months-ahead forecasts of industrial production obtained by the five forecasting methods, as described in Section 6.2. For each pair of forecasting methods and for each forecast horizon, we derive the hybrid forecast sequence by applying the two-step decision rule (using $h_t = (1, \Delta L_t)'$) on a rolling window of size 200, except that we proceed to Step 2 regardless of the test outcome. We evaluate the performance of the hybrid forecast and contrast it to that of the forecasts in the pair by (i) comparing the MSE of the hybrid forecast to the MSE of the individual forecasts and (ii) testing the optimality of each forecast for quadratic loss. The entries in Table V equal 1 if the MSE of the switching forecast is less than or equal to both the MSEs of the individual forecasts. We see that in 26 of 30 cases, the switching forecast is at least as accurate.

Overall, we observe that our simple decision rule behaves reasonably and adds useful information, suggesting that the model-selection implications of our testing approach may be a promising direction for future research.

7. CONCLUSION

We propose a general framework for out-of-sample predictive ability testing and forecast selection designed for use when the forecasting model may be misspecified. Our method can be applied to evaluation of point, interval, probability, and density forecasts for a general loss function.

We depart from the approach to predictive ability testing of Diebold and Mariano (1995) and West (1996) by evaluating the accuracy of a particular forecasting method, rather than the accuracy of the forecasting model. Because we consider forecasts based on estimators whose estimation uncertainty does not vanish asymptotically, our tests have a number of appealing properties: they directly capture the effect of estimation uncertainty on relative forecast performance, they can handle comparison of forecasts based on both nested and nonnested models, and they allow the forecasts to be produced by general parametric, semiparametric, and nonparametric estimation techniques.

Our framework can accommodate both unconditional objectives (which forecasting method was more accurate on average?), which have been the sole focus of the literature up to this point, as well as conditional objectives (can we predict which forecasting method will be more accurate at a specific future date?), which can help fine-tune the forecast selection decision to current economic conditions. We accordingly propose two tests: a test of equal conditional predictive ability and a test of equal unconditional predictive ability, which is the Diebold and Mariano (1995) test extended to an environment that permits parameter estimation.

Our Monte Carlo simulations suggest that our conditional tests have good finite-sample size and power properties. For the unconditional test, we show that when we compare nested models, our test correctly recognizes that forecasts from a misspecified but parsimonious model may be as accurate as forecasts from a correctly specified but less parsimonious model. Previously available tests (McCracken (1999) and Clark and McCracken (2001)) instead focus on the model rather than the forecasting method, and thus tend to favor the less parsimonious model. The disparity between the two approaches is greater the smaller is the ratio of in-sample to out-of-sample sizes. A drawback of the unconditional test implemented here is that it tends to falsely reject equal performance when the loss differences have zero mean but are highly serially correlated. This may be possible to remedy by more careful selection of HAC covariance estimators. On the other hand, the conditional tests emerge as useful tools for detecting persistence in the relative performance of the forecasts, as well as cases where the relative performance may depend on the state of the economy.

We explore the model-selection implications of adopting a conditional perspective by proposing and illustrating a simple two-step decision rule for forecast selection that tests for equal performance of the competing forecasts and then—in case of rejection—uses currently available information to select the best forecast for the future date of interest.

One useful application of our tests is the evaluation of different parameter-reduction methods for forecasting with a large number of predictors. We consider three popular methods: a sequential model selection approach, the diffusion indexes approach of Stock and Watson (2002), and Bayesian shrinkage estimation. Previous techniques are not capable of comparing these forecasting methods. We find that the sequential model-selection method performs worst, probably due to its tendency to select large models. A second result is that the predictors are less useful for price indexes than real variables. For these variables, Bayesian shrinkage is the best method.

Much work remains to be done. A significant area for future research is the exploration of procedures for selecting the best forecasting method or for optimally combining the methods in case of rejection of equal conditional predictive ability. A further generalization of our tests is to consider multiple comparison methods that are more sophisticated than the Hochberg–Bonferroni

bounds of Section 6, for example, by adapting the "reality check" approach of White (2000) to the conditional framework. Finally, it may be possible to obtain asymptotic refinements of the tests presented here by using bootstrap resampling techniques; for example, by establishing whether the results of Andrews (2002) can be extended to heterogeneous data.

Dept. of Economics, University of California, at Los Angeles, 405 Hilgard Avenue, Box 951477, CA 90095, U.S.A.; giacomin@econ.ucla.edu

and

Dept. of Economics, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0508, U.S.A.; hwhite@weber.ucsd.edu.

Manuscript received April, 2003; final revision received April, 2006.

APPENDIX: PROOFS

PROOF OF THEOREM 1: Under H_0, $\{Z_{m,t}, \mathcal{F}_t\}$ is a MDS and we can apply a MDS central limit theorem (CLT) to show that $\hat{\Omega}_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} \xrightarrow{d} N(0, I)$ as $n \to \infty$, from which it follows that $T_{m,n}^h \xrightarrow{d} \chi_q^2$ as $n \to \infty$. The MDS CLT we use requires conditions such that $\hat{\Omega}_n - \Omega_n \xrightarrow{p} 0$, where $\Omega_n = \text{var}[\sqrt{n}\bar{Z}_{m,n}]$. Write $Z_{m,t+1}Z'_{m,t+1} = f(h_t, W_{t+1}, \ldots, W_{t-m})$, where $f(\cdot)$ is a measurable function. Since $\{W_t\}$ and $\{h_t\}$ are mixing by (i), and f is a function of only a finite number of leads and lags of W_t and h_t, it follows from Lemma 2.1 of White and Domowitz (1984) that $\{Z_{m,t+1}Z'_{m,t+1}\}$ is also mixing of the same size as W_t. To apply the law of large numbers (LLN) to $Z_{m,t+1}Z'_{m,t+1}$, we further need to ensure that each of its elements has absolute $r+\delta$ moment bounded uniformly in t. By the Cauchy–Schwarz inequality and (ii), $E|Z_{m,t+1,i}Z_{m,t+1,j}|^{r+\delta} \leq [E|Z_{m,t+1,i}^2|^{r+\delta}]^{1/2}[E|Z_{m,t+1,j}^2|^{r+\delta}]^{1/2} < \Delta^{1/2}\Delta^{1/2} < \infty$, $i, j = 1, \ldots, q$ and for all t. That $\hat{\Omega}_n - \Omega_n \xrightarrow{p} 0$ then follows from McLeish's (1975) LLN as in Corollary 3.48 of White (2001). The variable Ω_n is finite by (ii) and it is uniformly positive definite by (iii). We apply the Cramér–Wold device (e.g., Proposition 5.1 of White (2001)) and show that for all $\lambda \in \mathbb{R}^q$, $\lambda'\lambda = 1$, $\lambda'\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} \xrightarrow{d} N(0,1)$, which implies that $\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} \xrightarrow{d} N(0, I)$. Consider $\lambda'\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} = n^{-1/2} \times \sum_{t=m}^{T-1} \lambda'\Omega_n^{-1/2}Z_{m,t+1}$ and write $\lambda'\Omega_n^{-1/2}Z_{m,t+1} = \sum_{i=1}^q \tilde{\lambda}_i Z_{m,t+1,i}$. The variable $\tilde{\lambda}_i Z_{m,t+1,i}$ is measurable with respect to $\mathcal{F}_t$, and we have that $E[\lambda'\Omega_n^{-1/2}Z_{m,t+1}| \mathcal{F}_t] = \sum_{i=1}^q \tilde{\lambda}_i E[Z_{m,t+1,i}|\mathcal{F}_t] = 0$, given (3). Hence $\{\lambda'\Omega_n^{-1/2}Z_{m,t+1}, \mathcal{F}_t\}$ is a MDS. The asymptotic variance is $\bar{\sigma}_n^2 = \text{var}[\lambda'\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n}] = \lambda'\Omega_n^{-1/2}\text{var}[\sqrt{n}\bar{Z}_{m,n}] \times \Omega_n^{-1/2}\lambda = 1$ for all n sufficiently large. We have

$$n^{-1}\sum_{t=m}^{T-1} \lambda'\Omega_n^{-1/2}Z_{m,t+1}Z'_{m,t+1}\Omega_n^{-1/2}\lambda - 1$$

$$= \lambda'\Omega_n^{-1/2}\hat{\Omega}_n\Omega_n^{-1/2}\lambda - \lambda'\Omega_n^{-1/2}\Omega_n\Omega_n^{-1/2}\lambda = g(\hat{\Omega}_n) - g(\Omega_n) \xrightarrow{p} 0,$$

because $\hat{\Omega}_n - \Omega_n \xrightarrow{p} 0$ and by using Proposition 2.30 of White (2001). Furthermore, by Minkowski's inequality,

$$E|\lambda'\Omega_n^{-1/2}Z_{m,t+1}|^{2+\delta} = E\left|\sum_{i=1}^{q}\tilde{\lambda}_i Z_{m,t+1,i}\right|^{2+\delta}$$
$$\leq \left[\sum_{i=1}^{q}\tilde{\lambda}_i (E|Z_{m,t+1,i}|^{2+\delta})^{1/(2+\delta)}\right]^{2+\delta} < \infty,$$

the last inequality following from (ii). Hence, the sequence $\{\lambda'\Omega_n^{-1/2}Z_{m,t+1}, \mathcal{F}_t\}$ satisfies the conditions of Corollary 5.26 of White (2001) (CLT for MDS), which implies that $\lambda'\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} \xrightarrow{d} N(0,1)$. By the Cramér–Wold device, $\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} \xrightarrow{d} N(0, I)$, from which the desired result follows by consistency of $\hat{\Omega}_n$ for Ω_n. *Q.E.D.*

PROOF OF THEOREM 2: By arguments similar to those used in the proof of Theorem 1, $\{Z_{m,t+1}\}$ is mixing of the same size as W_t. Furthermore, each element of $Z_{m,t+1}$ is bounded uniformly in t by (ii). McLeish's (1975) LLN (cf. White (2001, Cor. 3.48)) then implies $\bar{Z}_{m,n} - E[\bar{Z}_{m,n}] \xrightarrow{p} 0$. Under $H_{A,h}$ there exists $\varepsilon > 0$ such that $E[\bar{Z}'_{m,n}]E[\bar{Z}_{m,n}] > 2\varepsilon$ for all n sufficiently large. Then

$$P[\bar{Z}'_{m,n}\bar{Z}_{m,n} > \varepsilon] \geq P\big[\bar{Z}'_{m,n}\bar{Z}_{m,n} - E[\bar{Z}'_{m,n}]E[\bar{Z}_{m,n}] > -\varepsilon\big]$$
$$\geq P\big[|\bar{Z}'_{m,n}\bar{Z}_{m,n} - E[\bar{Z}'_{m,n}]E[\bar{Z}_{m,n}]| < \varepsilon\big] \to 1. \tag{15}$$

By arguments identical to those used in the proof of Theorem 1, $\{Z_{m,t+1}Z'_{m,t+1}\}$ is mixing of the same size as W_t by (i) and each of its elements is bounded uniformly in t by (ii). McLeish's (1975) LLN then implies that $\hat{\Omega}_n - \Omega_n \xrightarrow{p} 0$, with Ω_n uniformly positive definite by (iii). The conditions of Theorem 8.13 of White (1994) are then satisfied, and the theorem implies that for any constant $c \in \mathbb{R}$, $P[T^h_{m,n} > c] \to 1$ as $n \to \infty$. *Q.E.D.*

PROOF OF THEOREM 3: (a) Under H_0, we show that $\tilde{\Omega}_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} \xrightarrow{d} N(0, I)$ as $n \to \infty$, from which (a) follows. First, we apply the Cramér–Wold device and show that for all $\lambda \in \mathbb{R}^q$, $\lambda'\lambda = 1$, $\lambda'\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} \xrightarrow{d} N(0,1)$, where $\Omega_n = \text{var}[\sqrt{n}\bar{Z}_{m,n}]$, using the fact that $E[Z_{m,t+\tau}|\mathcal{F}_t] = 0$. The variable Ω_n is finite by (ii) and it is uniformly positive definite by (iii). Write $\lambda'\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} = n^{-1/2}\sum_{t=m}^{T-\tau}\lambda'\Omega_n^{-1/2}Z_{m,t+\tau}$. We verify that $\{\lambda'\Omega_n^{-1/2}Z_{m,t+\tau}\}$ satisfies the conditions of the Wooldridge and White (1988) CLT for mixing processes. By arguments identical to those used in the proof of Theorem 1, $\{\lambda'\Omega_n^{-1/2}Z_{m,t+\tau}\}$ is mixing of the same size as W_t. Furthermore, $\bar{\sigma}_n^2 = \text{var}[\lambda'\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n}] = \lambda'\Omega_n^{-1/2}\text{var}[\sqrt{n}\bar{Z}_{m,n}]\Omega_n^{-1/2}\lambda = 1 > 0$ for all n sufficiently large. Finally, by

Minkowski's inequality,

$$E|\lambda'\Omega_n^{-1/2}Z_{m,t+\tau}|^{2+\delta} = E\left|\sum_{i=1}^{q}\tilde{\lambda}_i Z_{m,t+\tau i}\right|^{2+\delta}$$
$$\leq \left[\sum_{i=1}^{q}\tilde{\lambda}_i (E|Z_{m,t+\tau i}|^{2+\delta})^{1/(2+\delta)}\right]^{2+\delta} < \infty,$$

the last inequality following from (ii). Hence, $\{\lambda'\Omega_n^{-1/2}Z_{m,t+\tau}\}$ satisfies the conditions of Corollary 3.1 of Wooldridge and White (1988), which implies that $\lambda'\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} \xrightarrow{d} N(0,1)$. By the Cramér–Wold device, we then have $\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} \xrightarrow{d} N(0,I)$. It remains to show that $\tilde{\Omega}_n - \Omega_n \xrightarrow{p} 0$, which completes the proof. We have

$$\begin{aligned}\tilde{\Omega}_n - \Omega_n &= n^{-1}\sum_{t=m}^{T-\tau}[Z_{m,t+\tau}Z'_{m,t+\tau} - E(Z_{m,t+\tau}Z'_{m,t+\tau})] \\ &\quad + n^{-1}\sum_{j=1}^{\tau-1} w_{n,j}\sum_{t=m+j}^{T-\tau}\big[Z_{m,t+\tau}Z'_{m,t+\tau-j} - E(Z_{m,t+\tau}Z'_{m,t+\tau-j}) \\ &\qquad + Z_{m,t+\tau-j}Z'_{m,t+\tau} - E(Z_{m,t+\tau-j}Z'_{m,t+\tau})\big].\end{aligned}$$

For $j = 0, \ldots, \tau-1$, $\{Z_{m,t+\tau}Z'_{m,t+\tau-j}\}$ is mixing of the same size as W_t and each of its elements is bounded uniformly in t by (ii). Applying McLeish's (1975) LLN (e.g., Corollary 3.48 of White (2001)) and using the fact that $w_{n,j} \to 1$ for $n \to \infty$, it follows that $n^{-1}w_{n,j}\sum_{t=m+j}^{T-\tau}[Z_{m,t+\tau}Z'_{m,t+\tau-j} - E(Z_{m,t+\tau}Z'_{m,t+\tau-j})] \xrightarrow{p} 0$ for each $j = 0, \ldots, \tau-1$ (with $w_{n,0} \equiv 1$), implying $\tilde{\Omega}_n - \Omega_n \xrightarrow{p} 0$.

(b) Using the same arguments as in the proof of Theorem 1, $\{Z_{m,t+\tau}\}$ is mixing of the same size as W_t. Furthermore, each element of $Z_{m,t+\tau}$ is bounded uniformly in t by (ii). McLeish's (1975) LLN then implies that $\bar{Z}_{m,n} - E[\bar{Z}_{m,n}] \xrightarrow{p} 0$. By definition, under $H_{A,h}$ there exists $\varepsilon > 0$ such that $E[\bar{Z}'_{m,n}]E[\bar{Z}_{m,n}] > 2\varepsilon$ for all n sufficiently large. We then have

$$\begin{aligned}P[\bar{Z}'_{m,n}\bar{Z}_{m,n} > \varepsilon] &\geq P\big[\bar{Z}'_{m,n}\bar{Z}_{m,n} - E[\bar{Z}'_{m,n}]E[\bar{Z}_{m,n}] > -\varepsilon\big] \\ &\geq P\big[|\bar{Z}'_{m,n}\bar{Z}_{m,n} - E[\bar{Z}'_{m,n}]E[\bar{Z}_{m,n}]| < \varepsilon\big] \to 1.\end{aligned} \tag{16}$$

By arguments identical to those used in part (a), which for this particular result do not require the time dependence structure imposed under the null hypothesis, it follows that $\tilde{\Omega}_n - \Omega_n \xrightarrow{p} 0$ with Ω_n uniformly positive definite by (iii). Theorem 8.13 of White (1994) then implies that for any constant $c \in \mathbb{R}$, $P[T^h_{m,n,\tau} > c] \to 1$ as $n \to \infty$. *Q.E.D.*

PROOF OF THEOREM 4: (a) We separately show that under H_0, $\sqrt{n}(\Delta\bar{L}_{m,n}/\sigma_n) \xrightarrow{d} N(0,1)$, where $\sigma_n^2 = \text{var}[\sqrt{n}\Delta\bar{L}_{m,n}]$, and that $\hat{\sigma}_n - \sigma_n \xrightarrow{p} 0$, from which the result follows. The variable σ_n^2 is finite by (ii) and it is positive for all n sufficiently large by (iii). Write $\sqrt{n}(\Delta\bar{L}_{m,n}/\sigma_n) = n^{-1/2}\sum_{t=m}^{T-\tau}\sigma_n^{-1}\Delta L_{m,t+\tau}$. We verify that the sequence $\{\sigma_n^{-1}\Delta L_{m,t+\tau}\}$ satisfies the conditions of Wooldridge and White's (1988) CLT for mixing processes. By arguments similar to those used in the proof of Theorem 1, $\{\sigma_n^{-1}\Delta L_{m,t+\tau}\}$ is mixing of the same size as W_t. Furthermore, by (ii), $E|\sigma_n^{-1}\Delta L_{m,t+\tau}|^{2+\delta} < \infty$. Hence, $\{\sigma_n^{-1}\Delta L_{m,t+\tau}\}$ satisfies the conditions of Corollary 3.1 of Wooldridge and White (1988), which implies that $\sqrt{n}(\Delta\bar{L}_{m,n}/\sigma_n) \xrightarrow{d} N(0,1)$. By arguments similar to the preceding, $\{\Delta L_{m,t+\tau}\}$ is mixing of the same size as W_t, which implies that $\{\Delta L_{m,t+\tau}\}$ is also mixing with ϕ of size $-r/(r-1)$ or α of size $-2r/(r-2)$. This, together with assumption (ii) and with the fact that $E(\Delta L_{m,t+\tau}) = 0$ under H_0, implies that the conditions of Theorem 6.20 of White (2001) are satisfied, and thus $\hat{\sigma}_n - \sigma_n \xrightarrow{p} 0$.

(b) As shown in (a), $\{\Delta L_{m,t+\tau}\}$ is mixing of the same size as W_t. Furthermore, $\Delta L_{m,t+\tau}$ is bounded uniformly in t by (ii). McLeish's (1975) LLN (as in Corollary 3.48 of White (2001)) then implies that $\Delta\bar{L}_{m,n} - E[\Delta\bar{L}_{m,n}] \xrightarrow{p} 0$. Under H_A there exists $\varepsilon > 0$ such that $(E[\Delta\bar{L}_{m,n}])^2 > 2\varepsilon$ for all n sufficiently large. We then have

$$
\begin{aligned}
(17) \qquad P[\Delta\bar{L}_{m,n}^2 > \varepsilon] &\geq P\big[\Delta\bar{L}_{m,n}^2 - (E[\Delta\bar{L}_{m,n}])^2 > -\varepsilon\big] \\
&\geq P\big[|\Delta\bar{L}_{m,n}^2 - (E[\Delta\bar{L}_{m,n}])^2| < \varepsilon\big] \to 1.
\end{aligned}
$$

By arguments identical to those used in part (a), $\hat{\sigma}_n^2 - \sigma_n^2 \xrightarrow{p} 0$ and by (iii), $\sigma_n^2 > 0$ for all n sufficiently large. From Theorem 8.13 of White (1994), it follows that for any constant $c \in \mathbb{R}$, $P[n\Delta\bar{L}_{m,n}^2/\hat{\sigma}_n^2 > c^2] = P[t_{m,n,\tau}^2 > c^2] \to 1$ as $n \to \infty$, which implies that $P[|t_{m,n,\tau}| > c] \to 1$ as $n \to \infty$. *Q.E.D.*

PROOF OF PROPOSITION 5: We have

$$
\begin{aligned}
&E\left[\frac{1}{n}\sum_t (Y_{t+1} - \hat{f}_{t,m}^{(i)})^2\right] \\
&\quad = \frac{1}{n}\sum_t \big\{(E[Y_{t+1} - \hat{f}_{t,m}^{(i)}])^2 + \text{Var}(Y_{t+1} - \hat{f}_{t,m}^{(i)})\big\}, \quad i = 1, 2.
\end{aligned}
$$

For $i = 1$, the bias term is

$$
\begin{aligned}
(E[Y_{t+1} - \hat{\beta}_{t,m}X_{t+1}])^2 &= \big(c - X_{t+1}(E[\hat{\beta}_{t,m}] - 1)\big)^2 \\
&= c^2\left(1 - \frac{\sum_j X_j}{\sum_j X_j^2}X_{t+1}\right)^2
\end{aligned}
$$

and the variance term is

$$\operatorname{Var}(Y_{t+1} - \hat{\beta}_{t,m} X_{t+1}) = \sigma^2 \left(1 + \frac{X_{t+1}^2}{\sum_j X_j^2}\right).$$

For $i = 2$, the bias term is

$$(E[Y_{t+1} - \hat{\delta}_{t,m} - \hat{\gamma}_{t,m} X_{t+1}])^2 = 0$$

and the variance term is

$$\begin{aligned}
&\operatorname{Var}(Y_{t+1} - \hat{\delta}_{t,m} - \hat{\gamma}_{t,m} X_{t+1}) \\
&\quad = \sigma^2 + \operatorname{Var}(\hat{\delta}_{t,m}) + X_{t+1}^2 \operatorname{Var}(\hat{\gamma}_{t,m}) + 2X_{t+1} \operatorname{cov}(\hat{\delta}_{t,m}, \hat{\gamma}_{t,m}) \\
&\quad = \sigma^2 \left(1 + \frac{\sum_j X_j^2}{m S_{xx}} + \frac{X_{t+1}^2}{S_{xx}} - 2\frac{\bar{X}}{S_{xx}} X_{t+1}\right).
\end{aligned}$$

Letting $E[\frac{1}{n}\sum_t (Y_{t+1} - \hat{f}_{t,m}^{(1)})^2] = E[\frac{1}{n}\sum_t (Y_{t+1} - \hat{f}_{t,m}^{(2)})^2]$ gives c in (11) as a solution. *Q.E.D.*

PROOF OF PROPOSITION 6: Given the assumption of normality, we have

$$\begin{aligned}
&E\left[\frac{1}{n}\sum_t L(Y_{t+1}, \hat{f}_{t,m}^{(i)})\right] \\
&\quad = \frac{1}{n}\sum_t \left\{E[\exp(Y_{t+1} - \hat{f}_{t,m}^{(i)})] - E[Y_{t+1} - \hat{f}_{t,m}^{(i)}] - 1\right\} \\
&\quad = \frac{1}{n}\sum_t \left\{\exp\left(E[Y_{t+1} - \hat{f}_{t,m}^{(i)}] + \frac{1}{2}\operatorname{Var}(Y_{t+1} - \hat{f}_{t,m}^{(i)})\right)\right. \\
&\qquad\qquad \left. - E[Y_{t+1} - \hat{f}_{t,m}^{(i)}] - 1\right\}.
\end{aligned}$$

Substituting the expressions for $E[Y_{t+1} - \hat{f}_{t,m}^{(i)}]$ and $\operatorname{Var}(Y_{t+1} - \hat{f}_{t,m}^{(i)})$, $i = 1, 2$, from the proof of Proposition 5 and letting $F(c) = E[\frac{1}{n}\sum_t L(Y_{t+1}, \hat{f}_{t,m}^{(1)})] - E[\frac{1}{n}\sum_t L(Y_{t+1}, \hat{f}_{t,m}^{(2)})]$ gives (12). *Q.E.D.*

REFERENCES

AMISANO, G., AND R. GIACOMINI (2006): "Comparing Density Forecasts via Weighted Likelihood Ratio Tests," *Journal of Business & Economic Statistics*, in press. [1553]

ANDREWS, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59, 817–858. [1554,1556,1557,1560]

——— (2002): "Higher-Order Improvements of a Computationally Attractive k-Step Bootstrap for Extremum Estimators," *Econometrica*, 70, 119–162. [1572]

BIERENS, H. B. (1990): "A Consistent Conditional Moment Test of Functional Form," *Econometrica*, 58, 1443–1458. [1556]

CHAO, J. C., V. CORRADI, AND N. R. SWANSON (2001): "An Out-of-Sample Test for Granger Causality," *Macroeconomic Dynamics*, 5, 598–620. [1545]

CLARK, T. E. (1999): "Finite-Sample Properties of Tests of Equal Forecast Accuracy," *Journal of Forecasting*, 18, 489–504. [1558]

CLARK, T. E., AND M. W. MCCRACKEN (2001): "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics*, 105, 85–110. [1545,1547,1559,1560,1571]

CLARK, T. E., AND K. D. WEST (2005): "Using Out-of-Sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis," NBER Technical Working Paper #305. [1546]

CLEMENTS, M. P., AND D. F. HENDRY (1998): *Forecasting Economic Time Series*. Cambridge, U.K.: Cambridge University Press. [1551]

——— (1999): *Forecasting Non-Stationary Economic Time Series*. Cambridge, MA: MIT Press. [1551]

CORRADI, V., N. R. SWANSON, AND C. OLIVETTI (2001): "Predictive Ability with Cointegrated Variables," *Journal of Econometrics*, 104, 315–358. [1545]

DIEBOLD, F. X., AND R. S. MARIANO (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13, 253–263. [1545,1546,1557,1558,1570,1571]

DIEBOLD, F. X., AND J. A. LOPEZ (1996): "Forecast Evaluation and Combination," in *Handbook of Statistics*, Vol. 14: Statistical Methods in Finance, ed. by G. S. Maddala and C. R. Rao. Amsterdam: North-Holland, 241–268. [1553]

FAMA, E. F., AND J. D. MACBETH (1973): "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy*, 81, 607–636. [1548,1550]

GIACOMINI, R., AND I. KOMUNJER (2005): "Evaluation and Combination of Conditional Quantile Forecasts," *Journal of Business & Economic Statistics*, 23, 416–431. [1553,1555]

GONEDES, N. (1973): "Evidence on the Information Content of Accounting Massages: Accounting-Based and Market-Based Estimate of Systematic Risk," *Journal of Financial and Quantitative Analysis*, 8, 407–444. [1548,1550]

GRANGER, C. W. J., AND P. NEWBOLD (1977): *Forecasting Economic Time Series*. London: Academic Press. [1546]

HARVEY, D. I., S. J. LEYBOURNE, AND P. NEWBOLD (1997): "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting*, 13, 281–291. [1546]

HOCHBERG, Y. (1988): "A Sharper Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, 75, 800–802. [1569]

HOOVER, K. D., AND S. J. PEREZ (1999): "Data Mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search," *Econometrics Journal*, 2, 167–191. [1548,1564]

LEITCH, G., AND J. E. TANNER (1991): "Economic Forecast Evaluation: Profits versus the Conventional Error Measures," *American Economic Review*, 81, 580–590. [1546,1552]

LITTERMAN, R. B. (1986): "Forecasting with Bayesian Vector Autoregressions—Five Years of Experience," *Journal of Business & Economic Statistics*, 4, 25–38. [1548,1564,1565]

MCCRACKEN, M. W. (1999): "Asymptotics for Out-of-Sample Tests of Granger Causality," Working Paper, University of Missouri, Columbia. [1559-1561,1571]

——— (2000): "Robust Out-of-Sample Inference," *Journal of Econometrics*, 99, 195–223. [1545, 1554]

MCLEISH, D. L. (1975): "A Maximal Inequality and Dependent Strong Laws," *The Annals of Probability*, 3, 826–836. [1572-1575]

NEWEY, W. K., AND K. D. WEST (1987): "A Simple, Positive Semidefinite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708. [1556]

PESARAN, M. H., AND A. TIMMERMANN (2006): "Selection of Estimation Window in the Presence of Breaks," *Journal of Econometrics*, forthcoming. [1548,1552]

STINCHCOMBE, M. B., AND H. WHITE (1998): "Consistent Specification Testing with Nuisance Parameters Present Only under the Alternative," *Econometric Theory*, 14, 295–325. [1556]

STOCK, J. H., AND M. W. WATSON (2002): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business & Economic Statistics*, 20, 147–162. [1548,1564,1571]

WEST, K. D. (1996): "Asymptotic Inference about Predictive Ability," *Econometrica*, 64, 1067–1084. [1545,1546,1549,1550,1554,1570]

WEST, K. D., H. J. EDISON, AND D. CHO (1993): "A Utility-Based Comparison of Some Models of Exchange Rate Volatility," *Journal of International Economics*, 35, 23–45. [1546,1552]

WHITE, H. (1994): *Estimation, Inference and Specification Analysis*. New York: Cambridge University Press. [1551,1573-1575]

——— (2000): "A Reality Check for Data Snooping," *Econometrica*, 68, 1097–1126. [1572]

——— (2001): *Asymptotic Theory for Econometricians*. San Diego: Academic Press. [1572-1575]

WHITE, H., AND I. DOMOWITZ (1984): "Nonlinear Regression with Dependent Observations," *Econometrica*, 52, 143–162. [1572]

WOOLDRIDGE, J. M., AND H. WHITE (1988): "Some Invariance Principles and Central Limit Theorems for Dependent Heterogeneous Processes," *Econometric Theory*, 4, 210–230. [1573-1575]