Tests of conditional predictive ability

Raffaella Giacomini and Halbert White*

University of California, Los Angeles and University of California, San Diego

This version: July 2005

Abstract

We propose a framework for out-of-sample predictive ability testing and forecast selection that is particularly well-suited to the presence of heterogeneity in the data, a plausible feature of many economic time series. Relative to the existing literature (Diebold and Mariano, 1995 and West, 1996), we introduce two main innovations: (1) we derive our tests in an environment where the finite sample properties of the estimators on which the forecasts may depend are preserved asymptotically; (2) 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 (1), 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 non-nested models; they allow the forecasts to be produced by general estimation methods, and they are easy to compute. While 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 the forecast performance of three leading parameter-reduction methods for macroeconomic forecasting using a large number of predictors: a sequential model selection approach, the "diffusion indexes" approach of Stock and Watson (2002), and the use of Bayesian shrinkage estimators.

*Discussions with Clive Granger, Graham Elliott and Andrew Patton were essential to the paper. Useful comments from the 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. The computations in the paper were carried out in the UCSD Experimental and Computational Laboratory, for which we thank Vince Crawford.

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 for comparing 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), West and McCracken (1998), McCracken (2000), Clark and McCracken (2001), Corradi, Swanson and Olivetti (2001) and Chao, Corradi and Swanson (2001), among others. This work represents a generalization of previous evaluation techniques which 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 that can be applied to multi-step 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 DMW), and at the same time they can be viewed as a generalization of the DMW tests since they are applicable in all cases in which those tests are applicable and in many more besides.

We introduce two main methodological innovations: (1) in deriving our tests, we consider an environment where the finite sample properties of the estimators on which the forecasts may depend are preserved asymptotically and (2) we formulate the problem of forecast evaluation as a problem of inference about *conditional* expectations of forecasts and forecast errors, which 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 permitting 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 gives our tests several advantages over existing tests. 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 explicitly incorporated into the loss function (e.g., AIC, BIC)[1]. As a result, our object of evaluation is not simply the forecasting model as in the DMW approach, but

[1]A recent paper by Clark and West (2004) suggests an alternative way to overcome this problem in the context of testing the martingale difference hypothesis.

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 non-nested 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 if forecasts from a given model can outperform those from a nested benchmark model. Third, we can accommodate general estimation procedures in the derivation of the forecasts, including Bayesian and semi- and non-parametric 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, instead, asks whether we can use current information - above and beyond past average behavior - to predict which forecast will be more accurate for a specific future date.

A further contribution is that our tests are derived under the assumption of data heterogeneity, which is plausibly more realistic than the assumption of stationarity typically made in the literature. In particular, we allow the data to be characterized by structural shifts at unknown dates. The assumption of heterogeneity has important consequences for which forecasting methods should be considered. In heterogeneous environments, the use of an expanding estimation window is not appropriate, as observations from the distant past at some point lose their predictive relevance. For this reason, we consider a "rolling window" forecasting scheme as a convenient way to handle instability in the data, and base the forecasts on a (possibly time-varying) moving window of the data that discards old observations. The choice of the estimation window can be data-driven and therefore part of the forecasting method, as in the procedure suggested by Pesaran and Timmermann (2002). Although our main focus is on the rolling window scheme, 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 for making 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, since rejection means that the relative performance of the competing forecasts is predictable, we should exploit current information for predicting 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 multi-step forecasts of eight macroeconomic variables (four measures of real activity and four 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 $\tau-$steps ahead variable $Y_{t+\tau}$, using a loss function $L_{t+\tau}(\cdot)$. The DMW approach tests:

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

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 appearing 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: (1) the losses depend on estimates $\hat{\beta}_{1t}$ and $\hat{\beta}_{2t}$, rather than on their probability limits; and (2) the expectation is conditional on some information set $\mathcal{G}_t$:

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

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. 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 tradeoff 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 tradeoff. 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 non-vanishing estimation uncertainty. This ensures that our tests capture the impact of estimation uncertainty on forecast performance. Further, as we discuss in detail in Section 3.2, this has the important advantage that our tests can handle nested and non-nested models in a unified framework.

We achieve non-vanishing 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 heterogeneous data environments considered here, as they discount or exclude older data that may no longer be informative about the predictive relations of current interest. 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, as it imposes no restrictions on the estimators other than finite memory, whereas the alternatives are comparatively specific; and second, the analysis required for this approach is straightforward, whereas that for the alternatives is more involved, but without a compensating increase in insight.

In stationary environments, limited memory estimators have the disadvantage of inefficiency. It is, however, an empirical question as to whether a given data-generating process is stationary or heterogeneous. Evidence provided by Clements and Hendry (1998, 1999) strongly suggests that heterogeneity is often present in economic time series. Interestingly, our procedures can offer direct evidence as to the advantages or disadvantages of limited memory estimators, thus permitting the user to avoid their inappropriate application. We provide further discussion below.

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 $\sigma-$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, since 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, as it is plausible to expect some predictability in future loss differences. For example, one may expect the relative performance to 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 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 a non-constant data-generating process for the variables to be forecast. In this paper, we therefore work with the assumption that the data generating process is heterogeneous, which, in our view, is a more realistic and practical assumption for economic forecasting contexts than the assumption of stationarity typically made in the predictive ability literature.[2] Specific sources of heterogeneity in the series that economists forecast are several. First, even if the underlying economic processes were stationary, heterogeneity in the observed time series can arise from changes in the measurement process. This source of heterogeneity is one that macroeconomic variables are particularly sensitive to (among other things, the definition of the measured variables and which entities are measured in constructing the variables may change). These sources of heterogeneity are plausibly less a concern for non-aggregated time series, such as the prices of well-defined commodities, as in financial economics. Nevertheless, the underlying economic processes include a variety of forces that affect either the nature of the commodity itself (the laws governing the behavior of firms represented by equity assets change, as do market conditions and technologies used by such firms), or the way the commodity is traded. Taken together, these

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

factors make it plausible that the relations between variables of interest relevant for forecasting could be different now than they were last year, let alone five, ten, or twenty years ago, and are not plausibly identical, as stationarity would require.

If heterogeneity is present in an economic time series, suitable methods for model-based forecasting and forecast evaluation need to be applied. In general, it seems appropriate in a time-varying environment to construct forecasts using estimators with limited memory, rather than using all available data. An example is the popular practice of specifying and estimating forecasting models over a rolling window of the data, which is the approach that we adopt here. A consequence of this is that the estimation window must be treated as a choice variable to be evaluated along with the forecasting model and the estimation procedure as parts of the forecasting method under analysis. In contrast, in the DMW framework the sample split between estimation (in-sample) and evaluation (out-of-sample) portions is arbitrary.

3 Theory

3.1 Description of the environment

Consider a stochastic process $W \equiv \{W_t : \Omega \longrightarrow \mathbb{R}^{s+1}, s \in \mathbb{N}, t = 1, 2, \ldots\}$ 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 $\mathcal{F}_t = \sigma(W_1', ..., W_t', X_{t+1}')'$ (as in, e.g., White, 1994, pg. 96). We adopt the standard convention of denoting random variables by upper case letters and realizations by lower case letters.

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 $\hat{f}_{m,t} \equiv f(w_t, w_{t-1}, ..., w_{t-m+1}; \hat{\beta}_{m,t})$ and $\hat{g}_{m,t} \equiv g(w_t, w_{t-1}, ..., w_{t-m+1}; \hat{\beta}_{m,t})$, where f and g are measurable functions. The subscripts indicate that the time-t forecasts are measurable functions of a sample of size m, consisting of the m most recent observations.

If the forecasts are based on parametric models, the parameter estimates from the two models are collected in the $k \times 1$ vector $\hat{\beta}_{m,t}$. Otherwise, $\hat{\beta}_{m,t}$ represents whatever semi-parametric or non-parametric estimators are used in constructing the forecasts. Note that we allow for general estimation procedures. The only requirement is that the estimation window size must be finite.

We view m as either a method-specific constant or a possibly time-dependent random integer 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, a data-driven choice of m is given by the procedure suggested by Pesaran and Timmermann (2002). 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 observations weights that suitably discount older observations, such as recursive estimators of the exponential smoothing type, with smoothing parameter bounded away from zero. We take this up elsewhere.

We perform the out-of-sample evaluation using a "rolling" window estimation scheme. Let T be the total sample size and m_1 the size of the first estimation window. We formulate the first $\tau-$step ahead forecasts at time m_1, using data indexed $1, ..., m_1$ and compare these forecasts to the realization $y_{m_1+\tau}$.[3] At time m_1+1, we formulate the second set of forecasts using the previous m_2 observations, where m_2 can be different from m_1. We compare the second set of forecasts to the realization $y_{m_1+1+\tau}$. We thus iterate the procedure and obtain the last forecasts at time $T-\tau$, by utilizing the $m_{T-\tau-m_1+1}$ most recent observations and compare the forecasts to y_T. This yields a sequence of $n \equiv T-\tau-m_1+1$ out-of-sample forecasts and relative forecast errors. Even though the estimation window lengths m_t can vary over time, for simplicity we express each forecast as a function of m ($m \leq \bar{m}$), which can be thought of as the maximum of all m_t's.

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 this case $\hat{\beta}_{t,m} = \hat{\beta}_{m_1,m_1}$, $m_1 \leqslant t \leqslant T-1$. For clarity of exposition, in the remainder of the paper we restrict attention to a rolling window forecasting scheme, but all results are valid for a fixed estimation sample scheme.

The elements above - the model, the estimation procedure and the size of the estimation window - are dimensions of choice by the user and are part of the forecasting method under evaluation.

We evaluate the sequence of out-of-sample forecasts by a loss function $L_{t+\tau}(Y_{t+\tau}, \hat{f}_{m,t})$, that depends on the forecasts and on the realizations of the variable. This loss function 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. The following are some examples of statistical loss functions that have been considered in the forecast evaluation literature. Examples of appropriate loss functions for the evaluation of quantile, probability, and density forecasts are also discussed in Diebold and Lopez (1996), Giacomini and Komunjer (in press) and Giacomini (2002). For simplicity, let $f_t \equiv \hat{f}_{m,t}$ and $\tau = 1$.

1. Squared error loss function: $L_{t+1}(Y_{t+1}, f_t) = (Y_{t+1} - f_t)^2$.

2. Absolute error loss function: $L_{t+1}(Y_{t+1}, f_t) = |Y_{t+1} - f_t|$.

3. Asymmetric linear cost function of order α (also known as the lin-lin or "tick function"):

   $L_{t+1}(Y_{t+1}, f_t) = (\alpha - 1(Y_{t+1} - f_t < 0))(Y_{t+1} - f_t)$, for $\alpha \in (0, 1)$.

4. Linex loss function: $L_{t+1}(Y_{t+1}, f_t) = \exp(a(Y_{t+1} - f_t)) - a(Y_{t+1} - f_t) - 1$, $a \in \mathbb{R}$.

[3] If the two forecasts are based on estimation windows of different lengths, we let m_1 be the maximum of the two.

5. Direction-of-change loss function: $L_{t+1}(Y_{t+1}, f_t) = 1\{sign(Y_{t+1} - Y_t) \neq sign(f_t - Y_t)\}$.

6. Predictive log-likelihood: $L_{t+1}(Y_{t+1}, f_t) = \log f_t(Y_{t+1})$, with f_t the density forecast of Y_{t+1}.

For a given loss function and $\sigma-$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

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

In writing (3), we are adopting the convention that $\hat{f}_{m,t}$ and $\hat{g}_{m,t}$ 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 the out-of-sample sequence $\{\Delta L_{m,t}, \mathcal{F}_t\}$ 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 collectively 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 exploiting 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

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

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 $\hat{\Omega}_n \equiv n^{-1} \sum_{t=m}^{T-1} Z_{m,t+1} Z_{m,t+1}{}'$ is a $q \times q$ matrix consistently estimating 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 $m \leq \bar{m} < \infty$ and $q \times 1$ test function sequence $\{h_t\}$ suppose:*

(i) $\{W_t\}$, $\{h_t\}$ are mixing with ϕ of size $-r/(2r-1)$, $r \geq 1$ or α of size $-r/(r-1)$, $r > 1$;

(ii) $E|Z_{m,t+1,i}|^{2(r+\delta)} < \Delta < \infty$ for some $\delta > 0$, $i = 1, ..., q$ 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 $n \to \infty$.

Comments: 1. Assumption (i) is mild, allowing the data to be characterized by considerable heterogeneity as well as dependence. This is in contrast with 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.

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 non-vanishing 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 the forecasting method under evaluation.

3. The use of an expanding window forecasting scheme is ruled out by our assumption of finite estimation window. This assumption is motivated by our explicit allowance for heterogeneity in the data, and it further serves the purpose of creating an environment with asymptotically non-vanishing estimation uncertainty. Such an environment could also be obtained by assuming an expanding window whose size grows more slowly than the out-of-sample size or - as we mentioned in Section 3.1 - by considering an expanding window associated with exponential smoothing recursive estimators with smoothing parameter bounded away from zero.

4. Assumption (iii), imposing 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. In that literature, the asymptotic variance of the test statistic is computed at the probability limits of the parameters, which may cause singularity when the forecasts are based on nested models. In contrast, in our framework the presence of non-vanishing estimation uncertainty prevents such singularity and thus makes our tests applicable to both nested and non-nested models.

5. In the construction of the test statistic we exploit the simplifying feature that the null hypothesis imposes the time dependence structure of an *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 heteroskedasticity 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, as this 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, and 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 correspondingly lower.

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 $\Delta L_{m,t+1}$ by a suitable function of Y_{t+1} and the forecasts. Conditional encompassing for quantile forecasting, in particular, is explored by Giacomini and Komunjer (in press).

The following results provide computationally convenient ways to perform the one-step conditional predictive ability test using standard regression packages.

Corollary 2 *Under the assumptions of Theorem 1, the test statistic $T^h_{m,n}$ 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 the $1 \times q$ vector $(h_t \Delta L_{m,t+1})'$ for $t = m, ..., T-1$. A level α test can be conducted by rejecting the null hypothesis H_0 of equal conditional predictive ability whenever $nR^2 > \chi^2_{q,1-\alpha}$, where $\chi^2_{q,1-\alpha}$ is the $(1-\alpha)-$quantile of a χ^2_q distribution.*

Corollary 3 *Let assumptions (i), (iii) and (iv) of Theorem 1 hold and further assume*

(ii)′ $E|\Delta L_{m,t+1}|^{2(r+\delta_1)} < \Delta_1 < \infty$ and $E|h_{ti}|^{2(r+\delta_2)} < \Delta_2 < \infty$ for some δ_1, $\delta_2 > 0$, $i = 1, ..., q$ and for all t;

(v) $E[(\Delta L_{m,t+1})^2|\mathcal{F}_t] = \sigma^2$ for all t and some $\sigma^2 > 0$.

Then the one-step conditional predictive ability test can be alternatively 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 the $1 \times q$ vector h_t', for $t = m, ..., T-1$. A level α test can be conducted by rejecting the null hypothesis H_0 of equal conditional predictive ability whenever $nR^2 > \chi^2_{q,1-\alpha}$, where $\chi^2_{q,1-\alpha}$ is the $(1-\alpha)-$quantile of a χ^2_q distribution.

If the conditional homoskedasticity assumption (v) can be reasonably expected to hold, the true distribution of the nR^2 statistic in Corollary 3 may be better approximated by its asymptotic distribution than the statistic of Corollary 2, and might thus deliver better inference.

3.2.1 Alternative hypothesis

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

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

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

Theorem 4 *Given Assumptions (i), (ii) and (iii) of Theorem 1, under $H_{A,h}$ in (5) for any constant $c \in \mathbb{R}$, $P[T^h_{m,n} > c] \to 1$ as $n \to \infty$.*

Note that H_0 and $H_{A,h}$ are not necessarily exhaustive. For a given choice of $\{h_t\}$, it may in fact happen that $E[\bar{Z}'_{m,n'}]E[\bar{Z}_{m,n'}] = 0$ for some sequence $\{n'\}$, without $\{\Delta L_{m,t+1}\}$ being an *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. 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 hand, one is left free to choose which test function is more relevant in any situation and thus focus power in that direction.

In practice, the test function is chosen by the researcher to embed elements of the information set $\mathcal{F}_t$ that are thought to help distinguish between the forecast performance of the two methods. Examples are, e.g., 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, one should keep in mind that the properties of the test will be altered if one either includes too few or too many elements. 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), which deliver consistent tests.

3.3 Multi-step 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 $cov(h_t \Delta L_{m,t+\tau}, h_{t-j} \Delta L_{t+\tau-j}) = 0$ 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

$$T^h_{m,n,\tau} = n(n^{-1} \sum_{t=m}^{T-\tau} h_t \Delta L_{m,t+\tau})' \tilde{\Omega}_n^{-1} (n^{-1} \sum_{t=m}^{T-\tau} h_t \Delta L_{m,t+\tau}) = n \bar{Z}'_{m,n} \tilde{\Omega}_n^{-1} \bar{Z}_{m,n} \tag{6}$$

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 $\tilde{\Omega}_n \equiv n^{-1} \sum_{t=m}^{T-\tau} Z_{m,t+\tau} Z'_{m,t+\tau} + n^{-1} \sum_{j=1}^{\tau-1} w_{n,j} \sum_{t=m+j}^{T-\tau} [Z_{m,t+\tau} Z'_{m,t+\tau-j} + Z_{m,t+\tau-j} Z'_{m,t+\tau}]$, with $w_{n,j}$ a weight function such that $w_{n,j} \to 1$ as $n \to \infty$ for each $j = 1, ..., \tau - 1$ (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 4 for the multi-step forecast horizon case.

Theorem 5 (Multi-step 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) $\{W_t\}, \{h_t\}$ are mixing with ϕ of size $-r/(2r-2)$, $r \geq 2$ or α of size $-r/(r-2)$, $r > 2$;

(ii) $E|Z_{m,t+\tau,i}|^{r+\delta} < \Delta < \infty$ for some $\delta > 0$, $i = 1, ..., q$ 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 Multi-step unconditional predictive ability test

When $\mathcal{G}_t$ is the trivial $\sigma-$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, $H_0 : E[\Delta L_{m,t+\tau}] = 0$, $t = 1, 2, ...$, against the alternative

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

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 $\hat{\sigma}^2_n$ is a suitable HAC estimator of the asymptotic variance $\sigma^2_n = var[\sqrt{n} \Delta \bar{L}_{m,n}]$, for example $\hat{\sigma}^2_n \equiv n^{-1} \sum_{t=m}^{T-\tau} \Delta L^2_{m,t+\tau} + 2 \left[n^{-1} \sum_{j=1}^{p_n} w_{n,j} \sum_{t=m+j}^{T-\tau} \Delta L_{m,t+\tau} \Delta L_{m,t+\tau-j} \right]$, with $\{p_n\}$ a sequence of integers such that $p_n \to \infty$ as $n \to \infty$, $p_n = o(n)$ and $\{w_{n,j} : n = 1, 2, ...; j = 1, ..., p_n\}$ a triangular array such that $|w_{n,j}| < \infty$, $n = 1, 2, ..., j = 1, ..., p_n$ and $w_{n,j} \to 1$ as $n \to \infty$ for each $j = 1, ..., p_n$ (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). The following theorem can thus be viewed as providing primitive conditions that not only ensure the validity of Diebold and Mariano's (1995) test, but extend its validity to a framework permitting parameter estimation.

Theorem 6 (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 a mixing with ϕ of size $-r/(2r-2)$, $r \geq 2$ or α of size $-r/(r-2)$, $r > 2$;

(ii) $E|\Delta L_{m,t+\tau}|^{2r} < \Delta < \infty$ for all t;

(iii) $\sigma^2_n \equiv var[\sqrt{n} \Delta \bar{L}_{m,n}] > 0$ 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. Since the loss differences are mixing variables, an 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).[4] Our simulations in Section 5 provide additional evidence on this point.

4 A decision rule for forecast selection

An appealing practical consequence of adopting a conditional perspective when comparing the performance of competing forecasts is that it can provide a basis for making forecast selection decisions. The topic of what to do when equal performance is accepted or rejected is still relatively unexplored in the existing predictive ability testing literature. For example, there is no guidance in that literature as to what to do in case of acceptance of equal unconditional performance. When adopting a conditional perspective, instead, both acceptance and rejection of the null hypothesis may be starting points for real-time forecast selection. In case of acceptance, forecast combination seems to be a natural candidate, as was noted by a referee. If, on the other hand, one of the competing methods is already a forecast combination, the principle of parsimony suggests using the simpler method. We leave further consideration of this issue for future work.

In this section, we focus instead on the implications of rejecting equal conditional predictive ability and describe a simple 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:

1. Regress $\Delta L_{m,t+\tau} = L_{t+\tau}(Y_{t+\tau}, \hat{f}_{m,t}) - L_{t+\tau}(Y_{t+\tau}, \hat{g}_{m,t})$ on h_t over the out-of-sample period $t = m, ..., T - \tau$ and let $\hat{\alpha}_n$ denote the regression coefficient. Apply Theorem 5 to establish whether $\hat{\alpha}_n$ is significantly different from zero. If so, proceed to step 2.

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

We offer this procedure as a simple example of how our tests can be used in forecast selection.

[4] Diebold and Mariano (1995) also acknowledge that $\tau-$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.

More sophisticated approaches immediately suggest themselves, but the subject of forecast selection is a significant topic that deserves extensive focused attention, beyond that possible in the space available here. Accordingly, we restrict attention here to this simple procedure and take up more elaborate methods and their application elsewhere.

In general, the plot of out-of-sample period predicted loss differences $\left\{\hat{\alpha}_n' h_t\right\}_{t=m}^{T-\tau}$ 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 above decision rule chooses g, i.e., $I_{n,c} = n^{-1} \sum_{t=m}^{T-\tau} 1\{\hat{\alpha}_n' h_t > c\}$, where $1\{A\}$ 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.

5.1 Size properties

The goal of our first Monte Carlo experiment is two-fold: 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.

The idea is to consider a situation where the tradeoff 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 (DGP) be:

$$Y_t = \alpha + CPI_t + \varepsilon_t,\ \varepsilon_t \sim i.i.d.N(0, \sigma^2), \tag{9}$$

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 in order to create data that exhibit realistic heterogeneous behavior. [5] The two competing forecasting models are:

$$\begin{aligned} M1 &: Y_t = \beta CPI_t + u_{1t} \\ M2 &: Y_t = \delta + \gamma CPI_t + u_{2t}. \end{aligned}$$

[5] To assess the heterogeneous behavior of CPI_t, we performed the Andrews-Ploberger (1994) test for a structural break in the unconditional mean and in the first autoregressive coefficient of CPI_t. The test detected no breaks in mean but found evidence of a break in the autoregressive coefficient (the p-value for the test of no break was 0.056).

The one-step-ahead forecasts of Y_{t+1} implied by the two models are, respectively,

$$\begin{aligned} \hat{f}^{(1)}_{m,t} &= \hat{\beta}_{m,t} CPI_{t+1} \\ \hat{f}^{(2)}_{m,t} &= \hat{\delta}_{m,t} + \hat{\gamma}_{m,t} CPI_{t+1}, \end{aligned} \tag{10}$$

estimated by OLS over a sample of size m. Here and in the following, we treat CPI as known.

For each pair of estimation window size m and out-of-sample size n in the range $(25, 50, ..., 150)$, we find values of α in (9) such that the two forecasting methods have equal expected MSE, using the following result.

Proposition 7 *Let* $X_t \equiv CPI_t$; $\bar{X} \equiv \frac{1}{m}\sum_{j=t-m+1}^{t} X_j$; $S_{xx} \equiv \sum_{j=t-m+1}^{t} X_j^2 - m\bar{X}^2$; $\sum_t \equiv \sum_{t=m}^{T-1}$ *and* $\sum_j \equiv \sum_{j=t-m+1}^{t}$. *If*

$$\alpha = \sigma \sqrt{\frac{\sum_t \left(\frac{\sum_j X_j^2}{mS_{xx}} + \frac{X_{t+1}^2}{S_{xx}} - 2\frac{\bar{X}}{S_{xx}} X_{t+1} - \frac{X_{t+1}^2}{\sum_j X_j^2} \right)}{\sum_t \left(1 - \frac{\sum_j X_j}{\sum_j X_j^2} X_{t+1} \right)^2}}, \tag{11}$$

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

Using α from Proposition 7, $\sigma = .1$ and the last $T = m + n$ CPI observations, we generate 5000 Monte Carlo replications of Y_t from (9) and compute rolling window forecasts as in (10).

To examine the robustness of the size properties of our test to the choice of loss function and illustrate the flexibility of our method, we further consider a linex loss function. We generate 5000 replications of Y_t from (9) as described above, using values of a such that the two forecasting methods have equal expected average linex loss, obtained as follows.

Proposition 8 *Using the notation of Proposition 7, if* α *solves* $F(\alpha) = 0$, *where*

$$\begin{aligned} F(\alpha) \equiv \sum_t \Bigg\{ &\exp\left[\alpha \left(1 - \frac{\sum_j X_j}{\sum_j X_j^2} X_{t+1} \right) + \frac{\sigma^2}{2} \left(1 + \frac{X_{t+1}^2}{\sum_j X_j^2} \right) \right] - \alpha \left(1 - \frac{\sum_j X_j}{\sum_j X_j^2} X_{t+1} \right) \\ &- \exp\left[\frac{\sigma^2}{2} \left(1 + \frac{\sum_j X_j^2}{mS_{xx}} + \frac{X_{t+1}^2}{S_{xx}} - 2\frac{\bar{X}}{S_{xx}} X_{t+1} \right) \right] \Bigg\}, \end{aligned} \tag{12}$$

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

We find values of α that solve the equation in Proposition 8 by numerical techniques. Table 1 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 non-standard asymptotic distribution. For linex loss, the CM test cannot be applied as it requires the same loss function for estimation and evaluation, whereas we estimate by OLS and not by linex maximum likelihood. [TABLE 1 HERE]

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 appears to be 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, since 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 1 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 predictive ability tests in two directions: (1) against serially correlated loss differences; and (2) against the performance being different in different states of the economy.

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:

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

For each of 5000 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

[6] We also considered selecting p_n using either the data-dependent method of Andrews (1991) or the popular simple alternative $p_n = .75n^{1/3}$, satisfying 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.

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

errors for forecasts of $CPI_{1998:12}$ 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: (1) the loss differences are not serially correlated ($\rho = 0$) but have non-zero unconditional mean; (2) 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: (1) $\rho = 0$, $\mu = (0, 0.05, ..., 1)$; and (2) $\mu = 0$, $\rho = (0, 0.05, ..., 0.9)$.

Figure 1 shows the power curves of the tests of Theorems 1 (conditional) and 6 (unconditional) in scenarios (1) and (2) above computed as the proportion of rejections of the null hypotheses $H_{0,cond}$ and $H_{0,unc}$ 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. [FIGURE 1 HERE]

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 differences become more serially correlated (the power curve is upward sloping, whereas it should be flat since $H_{0,unc}$ 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 5000 Monte Carlo replications, we generate a sequence of loss differences of length $n = 150$ as follows:

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

where $S_t = 1$ with probability p and $S_t = 0$ with probability $1 - p$. We thus have $E\left[\Delta L_{m,t+1}\right] = 0$ but $E\left[\Delta L_{m,t+1}|S_t\right] = \begin{cases} \mu/p & \text{if } S_t = 1 \\ -\mu/(1-p) & \text{if } S_t = 0 \end{cases}$, so that forecast 2 is more accurate in the first state and forecast 1 is more accurate in the second state. Figure 2 shows the rejection frequencies of the null hypotheses $H_{0,cond}$ and $H_{0,unc}$ at the 5% nominal level using the tests of Theorems 1 and 6. The power curves are obtained for $p = .5$ and $d \equiv \frac{\mu}{p(1-p)} = (0, 0.1, ..., 1)$ (d represents the difference in expected loss between the two states). We let $h_t = (1, S_t)'$ for the conditional test and $p_n = 0$ for the unconditional test. [FIGURE 2 HERE]

As expected, the conditional test has power to detect different performance in the different states, whereas the rejection frequencies for the unconditional 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 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 modelling 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 simple autoregressive and random walk benchmark forecasts. The DMW testing framework cannot be used here since some of the comparisons are between nested models and, further, that framework does not easily accommodate Bayesian estimation or the presence of estimated regressors. In contrast, our approach is well suited for comparison of methods based on nested models and for detecting differences in predictive ability arising from use of different modelling 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. We use the different parameter reduction methods to construct 1-, 6- and 12- month-ahead forecasts for eight U.S. macroeconomic variables: four measures of aggregate real activity and four price indexes. The first group includes the components of the Index of Coincident Economic Indicators maintained by the Conference Board: total industrial production; real personal income less transfers; real manufacturing and trade sales; and number of employees on nonagricultural payrolls. The price indexes are: consumer price index; consumer price index less food; personal consumption expenditure implicit price deflator; and producer price index.[8] See SW for full details.

6.1 Parameter-reduction methods

All forecasting models project the $\tau-$step ahead variable $Y^{\tau}_{t+\tau}$ onto time-t predictors X_t and lags of the variable of interest Y_t, $Y_{t-1,\ldots}$. The dependent variable and the predictors are transformations of the original data: if RAW_t is the observation at time t, we define $Y^{\tau}_{t+\tau} = (1200/\tau)\log(RAW_{t+\tau}/RAW_t)$, $X_t = 1200\log(RAW_t/RAW_{t-1})$, $Y_t = 1200\log(RAW_t/RAW_{t-1})$ for the real variables and $Y^{\tau}_{t+\tau} = (1200/\tau)\log(RAW_{t+\tau}/RAW_t){-}1200\log(RAW_t/RAW_{t-1})$, $X_t = 1200\Delta\log(RAW_t/RAW_{t-1})$, $Y_t = 1200\Delta\log(RAW_t/RAW_{t-1})$ for the price indexes. We consider the following forecasting methods.

[8] These variables coincide with the variables forecasted by SW, with the exception of the consumer price index less food which replaces the consumer price index less food and energy series considered by SW (not included in the data set available to the authors).

6.1.1 Sequential model selection

This method considers the full set of 145 predictors, together with lags of the variable of interest and performs a sequential search on each estimation sample that retains only a subgroup of variables that are then used for forecasting. The initial model is

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

where X_t is a vector containing the 145 predictors.[9] We apply a simplified version of the algorithm described by Hoover and Perez (1999, p.175), which reduces the number of regressors in the model by performing a sequence of stability tests, residual autocorrelation tests, and $t-$ and $F-$ tests of significance of the regressor's coefficients. We consider a single reduction path and perform only a subset of the tests used by Hoover and Perez (1999). We use a significance level $\alpha = 0.01$ for all the tests, designed to encourage parsimony of the final model. A complete description of our algorithm is available upon request from the first author.

6.1.2 Diffusion indexes

This two-step method first uses principal component analysis to estimate k factors $\hat{F}_t$ from the predictors X_t $(1 \leq k \leq 12)$ and then considers the model with p lags

$$Y_{t+\tau}^{\tau} = \alpha + \beta' \hat{F}_t + \gamma_1 Y_t + \ldots + \gamma_p Y_{t-p+1} + \varepsilon_{t+\tau}, \tag{15}$$

where both k and p are selected by BIC.

6.1.3 Bayesian shrinkage estimation

This method considers the full model (14) and applies Bayesian estimation of its coefficients using the Litterman (1986) prior. The Litterman prior, when applied to variables expressed in differences, shrinks all coefficients in (14) towards zero, except that for the intercept a diffuse prior is used. Formally, the variance-covariance matrix V for the prior distribution of $\theta \equiv (\alpha, \beta', \gamma')'$ is diagonal, with $\alpha \sim N(0, 10^8)$, $\beta_i \sim N(0, (w \cdot \lambda \cdot \hat{\sigma}_y / \hat{\sigma}_{x_i})^2)$, $i = 1, ..., 145$ and $\gamma_j \sim N(0, (\lambda/j))^2)$, $j = 1, ..., 6$. As suggested by 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.[10] The Bayesian estimate of θ is then $\theta^B = (X'X + \hat{\sigma}^2 V^{-1})^{-1}(X'Y^{\tau})$,

[9] We overcome multicollinearity in X_t by replacing the groups of variables whose correlation is greater than .98 with their average. The new X_t contains 130 regressors.

[10] λ is the prior standard deviation of the Y_t coefficient. The prior standard deviation of subsequent lags of Y_t is further divided by the lag length to reflect higher confidence in the prior mean for longer lags. w is a number between zero and one that reflects the belief that X_t is less useful for forecasting than lagged values of the dependent variable. The prior standard deviation of β_i is further multiplied by the ratio of the sample standard deviations of the dependent variable and of the ith regressor $\hat{\sigma}_y / \hat{\sigma}_{x_i}$, to eliminate differences in scale.

where X is $m \times 152$ (m is the size of the estimation sample) with rows $(1, X_t', Y_t, Y_{t-1}, ..., Y_{t-5})$, Y^τ is $m \times 1$ with elements $Y_{t+\tau}^\tau$, $\hat{\sigma}$ is the estimated standard error of the residuals in a univariate autoregression for $Y_{t+\tau}^\tau$, and V is the covariance matrix implied by our prior.

6.1.4 Benchmarks

The first benchmark is an autoregressive (AR) model

$$Y_{t+\tau}^\tau = \alpha + \gamma_1 Y_t + \ldots + \gamma_p Y_{t-p+1} + \varepsilon_{t+\tau}, \tag{16}$$

where p is selected by BIC with $0 \leq p \leq 6$. The second benchmark is based on a random walk in levels, corresponding to the forecasting model in differences

$$Y_{t+\tau}^\tau = \alpha + \varepsilon_{t+\tau}. \tag{17}$$

6.2 Real-time forecasting experiment

We use the five methods above to simulate real-time forecasting. The available sample has size $T = 468$, and we choose a maximum estimation window $m = 150 + \tau$. For comparability, we apply the same transformations to the original series as those documented in Appendix B of SW. The first estimation sample is from 1960:1 through 1972:6 + τ (the first 12 data were used as initial observations). We screen the data in this sample for outliers, replace the outliers with the unconditional mean of the variable, standardize the regressors, estimate the diffusion indexes and select the AR lag lengths and number of diffusion indexes by BIC. We run the regressions (14), (15), (16), (17) and apply the Bayesian shrinkage method for t =1960:1,...,1972:6 and use the values of the regressors at time t =1972:6 + τ to generate a set of forecasts for $Y_{1972:6+2\tau}^\tau$. We then move the estimation window forward one period and repeat all of the above steps on data from 1960:2 through 1972:7 + τ, which generates the forecasts for $Y_{1972:7+2\tau}^\tau$. The final forecasts for $Y_{1998:12}^\tau$ are produced at t =1998:12 − τ. The out-of-sample size is $n = 318 - \tau$.

In Section 2.1, we mentioned that our procedures can be used to provide direct evidence as to the advantages or disadvantages of limited memory estimators. Specifically, 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 illustrate the information available for assessing the value of rolling window procedures by comparing their performance to that of forecasts of industrial production and consumer price index for all models and forecast horizons using an expanding window of data from 1960:1 onwards. Table 2 reports the relative MSEs of the rolling-window and expanding-window forecasts. [TABLE 2 HERE]

The table shows that MSEs for rolling-window forecasts are most of the time 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). 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 series of forecasts we conduct pairwise tests of equal conditional predictive ability of the five forecasting methods, using a squared error loss[11]. For $\tau = 1, 6$, and 12, we test H_0 : $E[(Y_{t+\tau} - \hat{f}_{m,t})^2 - (Y_{t+\tau} - \hat{g}_{m,t})^2|\mathcal{G}_t] \equiv E[\Delta L_{t+\tau}|\mathcal{G}_t] = 0$, for both $\mathcal{G}_t = \mathcal{F}_t$ (conditional perspective) and $\mathcal{G}_t = \{\emptyset, \Omega\}$ (unconditional perspective).

For the case $\mathcal{G}_t = \mathcal{F}_t$, we use the test function: $h_t = (1,\ \Delta L_t)'$. Tables 3 and 4 show the results of conditional predictive ability tests for real variables and price indexes. The entries in the tables are the p-values of pairwise tests of equal conditional predictive ability, using the test of Theorem 5. 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 .5. [TABLES 3 - 4 HERE]

A sharp result in the tables is that the sequential model selection method is characterized by the worst performance, likely due to its tendency to select over-parameterized 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.

The results for the unconditional case are reported in Tables 5 and 6. The main entries are the p-values of pairwise tests of equal unconditional predictive ability of Theorem 6, and the numbers within parentheses are the ratios of MSE for the method in the column relative to the method in the row. 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. [TABLES 5 - 6 HERE]

The tables reveal that for the real variables the diffusion indexes and the Bayesian shrinkage methods in most cases outperform both benchmarks and that Bayesian shrinkage further outperforms the diffusion indexes method roughly half the time. For price indexes, instead, the parameter-reduction methods cannot generally outperform the AR.

[11] The corresponding results for an absolute error loss function are available upon request.

Two conclusions emerge 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 over the out-of-sample period, 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 either be due to the unconditional test being oversized or to the conditional test having low power. Our Monte Carlo simulations suggest that the more plausible explanation is the size distortion of the unconditional test and its 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 12-months-ahead forecasts of Industrial production obtained by the 5 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 (1) comparing the MSE of the hybrid forecast to the MSE of the individual forecasts; and (2) testing optimality of each forecast for quadratic loss. The entries in Table 7 equal 1 if the MSE of the switching forecast is less than or equal to both the MSEs of the individual forecasts. The table reveals that in 26 out of 30 cases the switching forecast is at least as accurate as the individual forecasts. [TABLE 7 HERE]

We tested forecast optimality by regressing the forecast errors on a constant and one lag of the forecast errors. Optimality is rejected if one rejects that the coefficients are jointly zero at the 5% level. We found that in all but two cases (sequential method vs. AR and RW at the 12-month horizon), if at least one of the individual forecasts is optimal, the switching forecast is also optimal. If they are both suboptimal, so is the switching forecast. To conserve space, we do not tabulate these results here. Detailed results are available from the first author upon request.

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 that is particularly well-suited to the presence of heterogeneity in the data. 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 that of the forecasting model. That is, we consider an environment in which estimation uncertainty in the forecasting model's parameters does not vanish asymptotically, which gives our tests several advantages over the previously available tests: they directly capture the effect of estimation uncertainty on relative forecast performance; they can handle comparison of forecasts based on both nested and non-nested models; and they allow the forecasts to be produced by general parametric, semi- and non-parametric estimation techniques.

Our framework can accommodate both unconditional objectives ("which forecasting method was more accurate on average?"), that 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 permitting 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 comparing 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 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 framework 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 the use of Bayesian shrinkage estimation. Using the data set of Stock and Watson (2002), including monthly U.S. data on a large number of macroeconomic variables, we generate multi-step forecasts of four measures of real activity and four price indexes using the different forecasting methods. Previous techniques are not capable of comparing these forecasting methods. We find that the simplified sequential model selection method performs worst, probably due to its tendency to select large models. A second result is that the predictors appear 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 comparisons, 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.

Appendix. Proofs

Proof of Theorem 1. Under H_0, $\{Z_{m,t}, \mathcal{F}_t\}$ is an *mds*, and we can apply an *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 $T^h_{m,n} \xrightarrow{d} \chi^2_q$ as $n \to \infty$. The *mds* CLT we use requires conditions such that $\hat{\Omega}_n - \Omega_n \xrightarrow{p} 0$, where $\Omega_n = var[\sqrt{n}\bar{Z}_{m,n}]$. Write $Z_{m,t+1}Z'_{m,t+1} = f(h_t, W_{t+1}, ..., W_{t-m})$, 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 a 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), $E|Z_{m,t+1,i}Z_{m,t+1,j}|^{r+\delta} \leq [E|Z^2_{m,t+1,i}|^{r+\delta}]^{1/2}[E|Z^2_{m,t+1,j}|^{r+\delta}]^{1/2} < \Delta^{1/2}\Delta^{1/2} < \infty$, $i,j = 1,...,q$ 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). Ω_n is finite by (ii), and it is uniformly positive definite by (iii). We apply the Cramér-Wold device and show that for all $\lambda \in \mathbb{R}^q$, $\lambda'\lambda = 1$, $\lambda'\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} \xrightarrow{d} N(0,1)$, which implies that $\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} \xrightarrow{d} N(0,I)$. Consider $\lambda'\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} = n^{-1/2}\sum_{t=m}^{T-1}\lambda'\Omega_n^{-1/2}Z_{m,t+1}$, 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 the linearity of conditional expectations implies 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 an *mds*. The asymptotic variance is $\bar{\sigma}_n^2 = var[\lambda'\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n}] = \lambda'\Omega_n^{-1/2} var[\sqrt{n}\bar{Z}_{m,n}]\Omega_n^{-1/2}\lambda = 1$ for all n sufficiently large. We have

$$n^{-1}\sum_{t=m}^{T-1} \lambda'\Omega_n^{-1/2} Z_{m,t+1} Z'_{m,t+1}\Omega_n^{-1/2}\lambda - 1 = \lambda'\Omega_n^{-1/2}\hat{\Omega}_n\Omega_n^{-1/2}\lambda - \lambda'\Omega_n^{-1/2}\Omega_n\Omega_n^{-1/2}\lambda = g(\hat{\Omega}_n) - g(\Omega_n) \xrightarrow{p} 0,$$

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

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

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 (e.g., Proposition 5.1 of White, 2001), $\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} \xrightarrow{d} N(0,I)$, from which the desired result follows by consistency of $\hat{\Omega}_n$ for Ω_n. ■

Proof of Corollary 2. The (constant unadjusted) R^2 for the regression of the constant unity on the variables $Z'_{m,t+1} \equiv (h_t \ \Delta L_{m,t+1})'$ can be written as $R^2 = \iota' Z_m [Z'_m Z_m]^{-1} Z'_m \iota / \iota'\iota$, where ι is an $n \times 1$ vector of ones and Z_m is the $n \times q$ matrix with rows $Z'_{m,t+1}$. Since $\hat{\Omega}_n = Z'_m Z_m / n$, it thus follows that $nR^2 = n(\iota' Z_m/n)\hat{\Omega}_n^{-1}(Z'_m\iota/n) = T^h_{m,n}$. ■

Proof of Corollary 3. The (constant unadjusted) R^2 for the regression of $\Delta L_{m,t+1}$ on h'_t can be written as $R^2 = \Delta L' h [h'h]^{-1} h' \Delta L / \Delta L' \Delta L$, where ΔL is the $n \times 1$ vector with elements $\Delta L_{m,t+1}$ and h is the $n \times q$ matrix with rows h'_t. We thus have $nR^2 = n\bar{Z}'_{m,n}(\hat{\sigma}_n V_n)^{-1}\bar{Z}_{m,n}$, where $\hat{\sigma}_n = \Delta L' \Delta L / n$ and $V_n = h'h/n$. We will show that $\hat{\sigma}_n V_n - \Omega_n \xrightarrow{p} 0$, which implies that the two statistics $T^h_{m,n}$ and nR^2 are asymptotically equivalent and thus the conditional predictive ability test can be alternatively based on the statistic nR^2. By the law of iterated expectations

$$\Omega_n = n^{-1}\sum_{t=m}^{T-1} E[h_t(\Delta L_{m,t+1})^2 h_t{}'] = n^{-1}\sum_{t=m}^{T-1} E[h_t E[(\Delta L_{m,t+1})^2|\mathcal{F}_t]] h_t{}'] = \sigma^2 E[h'h/n],$$

where the last equality follows from assumption (v). Given assumptions (i) and (ii)$'$, the sequences $\{h_t h'_t\}$ and $\{(\Delta L_{m,t+1})^2\}$ satisfy a LLN and it thus follows that $V_n - E[h'h/n] \xrightarrow{p} 0$ and $\hat{\sigma}_n - \sigma^2 = \hat{\sigma}_n - E[\hat{\sigma}_n] \xrightarrow{p} 0$, where the last equality is implied by (v). Hence, $\hat{\sigma}_n V_n - \Omega_n = \hat{\sigma}_n V_n - \sigma^2 E[h'h/n] \xrightarrow{p} 0$, and the proof is complete. ■

Proof of Theorem 4. Given Assumption (i), it follows from Lemma 2.1 of White and Domowitz (1984) that $\{Z_{m,t+1}\}$ is mixing of the same size as W_t, since it is a function of only a finite number of leads and lags of W_t and h_t. Further, 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 $\bar{Z}_{m,n} - E[\bar{Z}_{m,n}] \xrightarrow{p} 0$.

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. Then

$$P[\bar{Z}'_{m,n}\bar{Z}_{m,n} > \varepsilon] \geq P[\bar{Z}'_{m,n}\bar{Z}_{m,n} - E[\bar{Z}'_{m,n}]E[\bar{Z}_{m,n}] > -\varepsilon] \geq P[|\bar{Z}'_{m,n}\bar{Z}_{m,n} - E[\bar{Z}'_{m,n}]E[\bar{Z}_{m,n}]| < \varepsilon] \to 1. \tag{18}$$

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 $n \to \infty$. ■

Proof of Theorem 5. *(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 $\Omega_n = var[\sqrt{n}\bar{Z}_{m,n}]$, using the fact that $E[Z_{m,t+\tau}|\mathcal{F}_t] = 0$. 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 the scalar sequence $\{\lambda'\Omega_n^{-1/2}Z_{m,t+\tau}\}$ satisfies the conditions of the Wooldridge and White (1988) CLT for mixing processes. By arguments identical to those used in the proof of Theorem $\{\lambda'\Omega_n^{-1/2}Z_{m,t+\tau}\}$ is mixing of the same size as W_t. Further, $\bar{\sigma}_n^2 = var[\lambda'\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n}] = \lambda'\Omega_n^{-1/2}var[\sqrt{n}\bar{Z}_{m,n}]\Omega_n^{-1/2}\lambda = 1 > 0$ for all n sufficiently large. Finally, by Minkowski's inequality,

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

the last inequality following from (ii). Hence, the sequence $\{\lambda'\Omega_n^{-1/2}Z_{m,t+\tau}\}$ 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 (e.g., Proposition 5.1 of White, 2001), we have that $\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} \xrightarrow{d} N(0, I)$. It remains to show that $\tilde{\Omega}_n - \Omega_n \xrightarrow{p} 0$, which completes the proof. We have

$$\tilde{\Omega}_n - \Omega_n = n^{-1}\sum_{t=m}^{T-\tau}[Z_{m,t+\tau}Z'_{m,t+\tau} - E(Z_{m,t+\tau}Z'_{m,t+\tau})] +$$

$$+ n^{-1}\sum_{j=1}^{\tau-1}w_{n,j}\sum_{t=m+j}^{T-\tau}[Z_{m,t+\tau}Z'_{m,t+\tau-j} - E(Z_{m,t+\tau}Z'_{m,t+\tau-j}) + Z_{m,t+\tau-j}Z'_{m,t+\tau} - E(Z_{m,t+\tau-j}Z'_{m,t+\tau})].$$

For $j = 0, ..., \tau - 1$, $\{Z_{m,t+\tau}Z'_{m,t+\tau-j}\}$ 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 $j = 0, ..., \tau - 1$ (with $w_{n,0} \equiv 1$), implying $\tilde{\Omega}_n - \Omega_n \xrightarrow{p} 0$.

(b) Using the same arguments as in *(a)*, $\{Z_{m,t+\tau}\}$ is mixing of the same size as W_t. Further, each element of $Z_{m,t+\tau}$ is bounded uniformly in t by (ii). McLeish's (1975) LLN (as in Corollary

3.48 of White, 2001) then implies that $\bar{Z}_{m,n} - E[\bar{Z}_{m,n}] \xrightarrow{p} 0$. 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

$$P[\bar{Z}'_{m,n}\bar{Z}_{m,n} > \varepsilon] \geq P[\bar{Z}'_{m,n}\bar{Z}_{m,n} - E[\bar{Z}'_{m,n}]E[\bar{Z}_{m,n}] > -\varepsilon] \geq P[|\bar{Z}'_{m,n}\bar{Z}_{m,n} - E[\bar{Z}'_{m,n}]E[\bar{Z}_{m,n}]| < \varepsilon] \to 1. \tag{19}$$

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 $n \to \infty$. ■

Proof of Theorem 6. *(a)* We separately show that, under H_0, $\sqrt{n}\frac{\Delta \bar{L}_{m,n}}{\sigma_n} \xrightarrow{d} N(0,1)$, where $\sigma_n^2 = var[\sqrt{n}\Delta \bar{L}_{m,n}]$ and that $\hat{\sigma}_n - \sigma_n \xrightarrow{p} 0$, from which the result follows. σ_n^2 is finite by (ii) and it is positive for all n sufficiently large by (iii). Write $\sqrt{n}\frac{\Delta \bar{L}_{m,n}}{\sigma_n} = n^{-1/2}\sum_{t=m}^{T-\tau}\sigma_n^{-1}\Delta L_{m,t+\tau}$ and consider the scalar sequence $\{\sigma_n^{-1}\Delta L_{m,t+\tau}\}$. We verify that this sequence 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. Further, by (ii), $E|\sigma_n^{-1}\Delta L_{m,t+\tau}|^{2+\delta} < \infty$. Hence, the sequence $\{\sigma_n^{-1}\Delta L_{m,t+\tau}\}$ satisfies the conditions of Corollary 3.1 of Wooldridge and White (1988), which implies that $\sqrt{n}\frac{\Delta \bar{L}_{m,n}}{\sigma_n} \xrightarrow{d} N(0,1)$. By similar arguments as above, $\{\Delta L_{m,t+\tau}\}$ 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 $\hat{\sigma}_n - \sigma_n \xrightarrow{p} 0$, which completes the proof.

(b) As shown in *(a)*, $\{\Delta L_{m,t+\tau}\}$ is mixing of the same size as W_t. Further, $\Delta L_{m,t+\tau}$ 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

$$P[\Delta \bar{L}^2_{m,n} > \varepsilon] \geq P[\Delta \bar{L}^2_{m,n} - (E[\Delta \bar{L}_{m,n}])^2 > -\varepsilon] \geq P[|\Delta \bar{L}^2_{m,n} - (E[\Delta \bar{L}_{m,n}])^2| < \varepsilon] \to 1. \tag{20}$$

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

Proof of Proposition 7. We have

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

For $i = 1$, the bias term is $\left(E\left[Y_{t+1} - \hat{\beta}_{m,t}X_{t+1}\right]\right)^2 = \left(\alpha - X_{t+1}\left(E\left[\hat{\beta}_{m,t}\right] - 1\right)\right)^2 = \alpha^2\left(1 - \frac{\sum_j X_j}{\sum_j X_j^2}X_{t+1}\right)^2$ and the variance term is $Var\left(Y_{t+1} - \hat{\beta}_{m,t}X_{t+1}\right) = \sigma^2\left(1 + \frac{X^2_{t+1}}{\sum_j X_j^2}\right)$. For $i = 2$, the bias term is

$\left(E\left[Y_{t+1}-\hat{\delta}_{m,t}-\hat{\gamma}_{m,t}X_{t+1}\right]\right)^2=0$ and the variance term is

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

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

Proof of Proposition 8. Given the assumption of normality, we have

$$E\left[\frac{1}{n}\sum_t L\left(Y_{t+1},\hat{f}_{m,t}^{(i)}\right)\right]=\frac{1}{n}\sum_t\left\{E\left[\exp\left(Y_{t+1}-\hat{f}_{m,t}^{(i)}\right)\right]-E\left[Y_{t+1}-\hat{f}_{m,t}^{(i)}\right]-1\right\}$$

$$=\frac{1}{n}\sum_t\left\{\exp\left(E\left[Y_{t+1}-\hat{f}_{m,t}^{(i)}\right]+\frac{1}{2}Var\left(Y_{t+1}-\hat{f}_{m,t}^{(i)}\right)\right)-E\left[Y_{t+1}-\hat{f}_{m,t}^{(i)}\right]-1\right\}.$$

Substituting the expressions for $E\left[Y_{t+1}-\hat{f}_{m,t}^{(i)}\right]$ and $Var\left(Y_{t+1}-\hat{f}_{m,t}^{(i)}\right)$, $i=1,2$ from the Proof of Proposition 7 and letting $F(\alpha)=E\left[\frac{1}{n}\sum_t L\left(Y_{t+1},\hat{f}_{m,t}^{(1)}\right)\right]-E\left[\frac{1}{n}\sum_t L\left(Y_{t+1},\hat{f}_{m,t}^{(2)}\right)\right]$ gives (12). ■

References

[1] Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation", *Econometrica*, 59, 817-858.

[2] Andrews, D. W. K. (2002): "Higher-Order Improvements of a Computationally Attractive k-Step Bootstrap for Extremum Estimators", *Econometrica, 70,* 119-162.

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

[4] Chao, J. C., Corradi, V., Swanson, N. R. (2001), "An Out-of-Sample Test for Granger Causality", *Macroeconomic Dynamics,* 5, 598-620.

[5] Clark, T. E. (1999): "Finite-sample Properties of Tests of Equal Forecast Accuracy", *Journal of Forecasting,* 18, 489-504.

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

[7] Clark, T. E., West, K. D. (2004): "Using Out-of-sample Mean Squared Prediction Errors to Test the Martingale Difference Hypothesis", manuscript.

[8] Clements, M. P., Hendry, D. F. (1998): *Forecasting Economic Time Series,* Cambridge University Press, Cambridge.

[9] Clements, M. P., Hendry, D. F. (1999): *Forecasting Non-Stationary Economic Time Series,* MIT Press, Cambridge, Massachusetts.

[10] Corradi, V., Swanson, N. R., Olivetti, C. (2001): "Predictive Ability with Cointegrated Variables", *Journal of Econometrics*, 104, 315-358.

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

[12] Diebold, F. X., Lopez, J. A. (1996): "Forecast Evaluation and Combination", in G. S. Maddala and C. R. Rao (eds) *Handbook of Statistics*, vol. 14: Statistical Methods in Finance, North-Holland, Amsterdam, 241-268.

[13] Fama, E. F., MacBeth, J. D. (1973): "Risk, Return, and Equilibrium: Empirical Tests", *Journal of Political Economy*, 81, 607-636.

[14] Giacomini, R. (2002): "Comparing Density Forecasts via Weighted Likelihood Ratio Tests: Asymptotic and Bootstrap Methods", UCSD discussion paper 2002-12.

[15] Giacomini, R., Komunjer, I. (in press): "Evaluation and Combination of Conditional Quantile Forecasts", *Journal of Business and Economic Statistics,* forthcoming.

[16] 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.

[17] Granger, C. W. J., Newbold, P. (1977): *Forecasting Economic Time Series,* Academic Press Inc., London.

[18] Harvey, D. I., Leybourne, S. J., Newbold, P. (1997): "Testing the Equality of Prediction Mean Squared Errors", *International Journal of Forecasting,* 13*,* 281-291.

[19] Hoover, K. D., Perez, S. J. (1999): "Data Mining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search", *Econometrics Journal*, 2, 167-191.

[20] Leitch, G., Tanner, J. E. (1991): "Economic Forecast Evaluation: Profits Versus the Conventional Error Measures", *American Economic Review,* 81, 580-590.

[21] Litterman, R. B. (1986): "Forecasting with Bayesian Vector Autoregressions - Five Years of Experience", *Journal of Business and Economic Statistics,* 4, 25-38.

[22] McCracken, M. W. (1999): "Asymptotics for Out-of-Sample Tests of Granger Causality", University of Missouri, Columbia working paper.

[23] McCracken, M. W. (2000): "Robust Out-of-Sample Inference", *Journal of Econometrics,* 99, 195-223.

[24] McLeish, D. L. (1975): "A Maximal Inequality and Dependent Strong Laws", *Annals of Probability,* 3, 826-836.

[25] Newey, W. K., West, K. D. (1987): "A Simple, Positive Semidefinite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica,* 55, 703-708.

[26] Pesaran, M. H., Timmermann, A. (2002): "Model Instability and Choice of Observation Window", manuscript.

[27] Stinchcombe, M. B., White, H. (1998): "Consistent Specification Testing with Nuisance Parameters Present Only Under the Alternative", *Econometric Theory,* 14, 295-325.

[28] Stock, J. H., Watson, M. W. (2002): "Macroeconomic Forecasting Using Diffusion Indexes", *Journal of Business and Economic Statistics,* 20, 147-162.

[29] West, K. D. (1996): "Asymptotic Inference about Predictive Ability", *Econometrica*, 64, 1067-1084.

[30] West, K. D., Edison, H. J., Cho, D. (1993): "A Utility-Based Comparison of Some Models of Exchange Rate Volatility", *Journal of International Economics,* 35, 23-45.

[31] West, K. D., McCracken, M. W. (1998): "Regression-Based Tests of Predictive Ability", *International Economic Review,* 39, 817-840.

[32] White, H. (1994): *Estimation, Inference and Specification Analysis,* Cambridge University Press, New York.

[33] White, H. (2000): "A Reality Check for Data Snooping", *Econometrica*, 68, 1097-1126.

[34] White, H. (2001): *Asymptotic Theory for Econometricians,* Academic Press, San Diego.

[35] White, H., Domowitz, I. (1984), "Nonlinear Regression with Dependent Observations", *Econometrica*, 52, 143-162.

[36] Wooldridge, J. M., White, H. (1988): "Some Invariance Principles and Central Limit Theorems for Dependent Heterogeneous Processes", *Econometric Theory,* 4, 210-230.

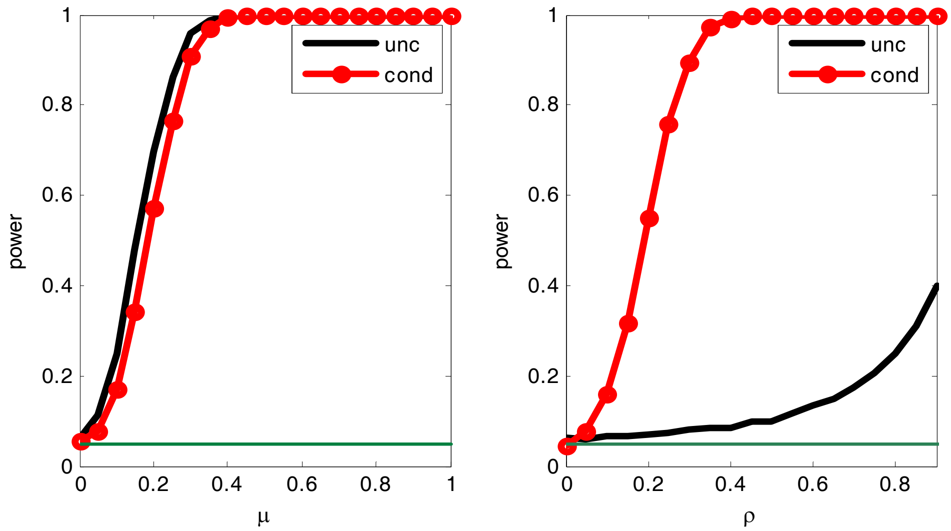

Figure 1: Power curves for the conditional test of Theorem 1 and the unconditional test of Theorem 6. Each curve represents the rejection frequencies over 5000 Monte Carlo replications of the null hypothesis: $H_{0,cond} : E[\Delta L_{m,t+1}|\mathcal{F}_t] = 0$ and $H_{0,unc} : E[\Delta L_{m,t+1}] = 0$. The DGP in the left panel is such that $E[\Delta L_{m,t+1}|\mathcal{F}_t] = \mu$ and the horizontal axis plots different values of μ. The DGP in the right panel is such that $E[\Delta L_{m,t+1}] = 0$ but $E[\Delta L_{m,t+1}|\mathcal{F}_t] = \rho(\Delta L_{m,t})$; ρ is plotted on the horizontal axis.

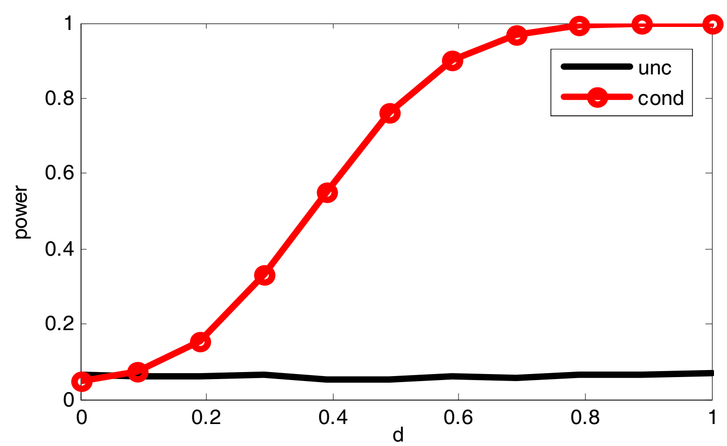

Figure 2: Power curves for the conditional test of Theorem 1 and the unconditional test of Theorem 6. Each curve represents the rejection frequencies over 5000 Monte Carlo replications of the null hypothesis: $H_{0,cond} : E[\Delta L_{m,t+1}|\mathcal{F}_t] = 0$ and $H_{0,unc} : E[\Delta L_{m,t+1}] = 0$. The DGP is such that $E[\Delta L_{m,t+1}] = 0$ but $E[\Delta L_{m,t+1}|\mathcal{F}_t] = d(S_t - p)$ where S_t=1 with probability p and 0 with probability $1 - p$. The horizontal axis plots different values of d.

Table 1. Rejection frequencies of unconditional predictive ability and McCracken (1999) tests

A. Quadratic loss

	Unconditional predictive ability							McCracken (1999)					
	n							n					
m	25	50	75	100	125	150	m	25	50	75	100	125	150
25	.053	.041	.037	.032	.035	.024	25	.087	.284	.360	.428	.481	.525
50	.062	.052	.044	.039	.037	.035	50	.158	.068	.245	.279	.300	.356
75	.062	.053	.048	.050	.040	.037	75	.147	.168	.070	.218	.256	.279
100	.062	.054	.050	.055	.047	.043	100	.118	.152	.152	.062	.216	.241
125	.073	.056	.054	.049	.044	.042	125	.120	.137	.146	.167	.063	.199
150	.061	.055	.056	.049	.048	.046	150	.091	.141	.134	.157	.204	.058

B. Linex loss

Unconditional predictive ability

	n					
m	25	50	75	100	125	150
25	.060	.057	.055	.050	.046	.046
50	.066	.064	.062	.057	.061	.062
75	.069	.067	.065	.066	.064	.058
100	.070	.067	.066	.065	.068	.074
125	.072	.071	.070	.075	.075	.077
150	.075	.077	.075	.076	.077	.073

Rejection frequencies over 5000 Monte Carlo replications 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 .05. The DGP is such that $E\left[\frac{1}{n}\sum_t L(Y_{t+1}, \hat{f}^{(1)}_{m,t})\right] = E\left[\frac{1}{n}\sum_t L(Y_{t+1}, \hat{f}^{(2)}_{m,t})\right]$, where $\hat{f}^{(1)}_{m,t}$ and $\hat{f}^{(2)}_{m,t}$ are defined in (10) and L is either quadratic or linex. m is the estimation window size and n is the out-of-sample size.

Table 2. Relative MSE. Rolling and expanding window

	Industrial production					Consumer price index				
τ	Seq.	Diff Ind	Bayes	AR	RW	Seq.	Diff Ind	Bayes	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

Ratios of MSEs of $\tau-$stepts ahead forecasts obtained by the forecasting methods in the column estimated over either a rolling window of size $m = 150$ or an expanding window with the same initial size.

Table 3. Conditional predictive ability tests. Real variables.

Bench	Industrial production				Personal income				Mfg & trade sales				Nonag. employment			
	Seq.	Diff Ind	Bayes	AR	Seq.	Diff Ind	Bayes	AR	Seq.	Diff Ind	Bayes	AR	Seq.	Diff Ind	Bayes	AR
A. Horizon = 1 month																
Diff Ind	0.026^{-}				0.004^{-}				0.100^{-}				0.171			
	(0.00)				(0.02)				(0.00)				(0.00)			
Bayes	0.020^{-}	0.080^{-}			0.006^{-}	0.731			0.097^{-}	0.209			0.153	0.552		
	(0.00)	(0.02)			(0.02)	(0.90)			(0.00)	(0.09)			(0.00)	(0.16)		
AR	0.034^{-}	0.040^{+}	0.001^{+}		0.027^{-}	0.018^{+}	0.199		0.090^{-}	0.094^{+}	0.001^{+}		0.220	0.241	0.066^{+}	
	(0.00)	(0.91)	(0.93)		(0.03)	(0.97)	(0.93)		(0.00)	(0.95)	(0.97)		(0.00)	(0.96)	(0.99)	
RW	0.043^{-}	0.049^{+}	0.004^{+}	0.163	0.108	0.002^{+}	0.003^{+}	0.023^{+}	0.090^{-}	0.101	0.001^{+}	0.984	0.133	0.000^{+}	0.000^{+}	0.000^{+}
	(0.00)	(0.81)	(0.84)	(0.78)	(0.07)	(0.86)	(0.83)	(0.80)	(0.00)	(0.95)	(0.96)	(0.88)	(0.00)	(0.87)	(0.92)	(0.90)
B. Horizon = 6 months																
Diff Ind	0.010^{-}				0.018^{-}				0.001^{-}				0.001^{-}			
	(0.05)				(0.00)				(0.00)				(0.00)			
Bayes	0.003^{-}	0.432			0.014^{-}	0.037^{-}			0.001^{-}	0.421			0.000^{-}	0.257		
	(0.01)	(0.00)			(0.00)	(0.00)			(0.00)	(0.18)			(0.00)	(0.10)		
AR	0.885	0.167	0.057^{+}		0.031^{-}	0.098^{+}	0.020^{+}		0.013^{-}	0.003^{+}	0.001^{+}		0.172	0.291	0.086^{+}	
	(0.01)	(0.98)	(0.98)		(0.00)	(0.93)	(1.00)		(0.00)	(0.97)	(1.00)		(0.00)	(1.00)	(0.99)	
RW	0.654	0.154	0.038^{+}	0.193	0.035^{-}	0.124	0.024^{+}	0.591	0.012^{-}	0.006^{+}	0.003^{+}	0.476	0.951	0.017^{+}	0.004^{+}	0.001^{+}
	(0.30)	(0.98)	(0.99)	(0.97)	(0.00)	(1.00)	(0.99)	(0.90)	(0.00)	(0.98)	(1.00)	(0.00)	(0.02)	(0.98)	(1.00)	(0.98)
C. Horizon = 12 months																
Diff Ind	0.003^{-}				0.006^{-}				0.000^{-}				0.000^{-}			
	(0.00)				(0.00)				(0.00)				(0.00)			
Bayes	0.001^{-}	0.201			0.003^{-}	0.044^{-}			0.000^{-}	0.189			0.000^{-}	0.082^{-}		
	(0.00)	(0.00)			(0.00)	(0.01)			(0.00)	(0.05)			(0.00)	(0.01)		
AR	0.029^{-}	0.202	0.088^{+}		0.028^{-}	0.314	0.095^{+}		0.010^{-}	0.042^{+}	0.013^{+}		0.228	0.074^{+}	0.048^{+}	
	(0.02)	(0.96)	(0.99)		(0.00)	(0.94)	(1.00)		(0.03)	(0.93)	(0.97)		(0.02)	(0.88)	(0.97)	
RW	0.031^{-}	0.174	0.073^{+}	0.873	0.041^{-}	0.227	0.113	0.096^{+}	0.010^{-}	0.042^{+}	0.013^{+}	1.000	0.508	0.014^{+}	0.005^{+}	0.005^{+}
	(0.05)	(1.00)	(1.00)	(0.01)	(0.00)	(0.92)	(0.99)	(0.85)	(0.03)	(0.93)	(0.97)	(0.00)	(0.07)	(0.96)	(0.99)	(0.93)

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 sign indicates that the test rejects equal conditional predictive ability at the 10% level and that the method in the column outperforms the method in the row more than 50% of the time. A minus sign indicates that the test rejects equal conditional predictive ability at the 10% level and that the method in the column 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 .001 and the Bayesian shrinkage method outperforms the AR method 93% of the time.

Table 4. Conditional predictive ability tests. Price indexes.

	CPI				CPI exc. food				Consumption deflator				Producer price index			
Bench	Seq.	Diff Ind	Bayes	AR	Seq.	Diff Ind	Bayes	AR	Seq.	Diff Ind	Bayes	AR	Seq.	Diff Ind	Bayes	AR
A. Horizon = 1 month																
Diff Ind	0.175				0.175				0.071^-				0.000^-			
	(0.00)				(0.00)				(0.00)				(0.02)			
Bayes	0.146	0.644			0.139	0.570			0.181	0.417			0.046^-	0.067^+		
	(0.01)	(0.99)			(0.00)	(0.99)			(0.00)	(0.91)			(0.01)	(0.99)		
AR	0.192	0.124	0.721		0.184	0.367	0.702		0.072^-	0.479	0.489		0.000^-	0.161	0.047^-	
	(0.00)	(0.74)	(0.18)		(0.00)	(0.94)	(0.06)		(0.00)	(0.55)	(0.03)		(0.02)	(0.22)	(0.01)	
RW	0.193	0.385	0.389	0.452	0.183	0.275	0.360	0.387	0.150	0.057^+	0.118	0.053^+	0.240	0.109	0.578	0.098^+
	(0.00)	(0.84)	(0.78)	(0.80)	(0.00)	(0.97)	(1.00)	(1.00)	(0.00)	(0.94)	(0.88)	(0.91)	(0.01)	(1.00)	(0.76)	(0.98)
B. Horizon = 6 months																
Diff Ind	0.091^-				0.046^-				0.000^-				0.000^-			
	(0.00)				(0.00)				(0.00)				(0.00)			
Bayes	0.261	0.002^+			0.098^-	0.008^+			0.008^+	0.010^+			0.493	0.000^+		
	(0.00)	(0.99)			(0.00)	(0.97)			(0.03)	(1.00)			(0.51)	(0.99)		
AR	0.146	0.082^+	0.055^-		0.064^-	0.090^+	0.061^-		0.000^-	0.149	0.048^-		0.000^-	0.809	0.000^-	
	(0.00)	(0.82)	(0.02)		(0.00)	(0.91)	(0.01)		(0.00)	(0.37)	(0.09)		(0.00)	(0.82)	(0.01)	
RW	0.935	0.000^+	0.004^+	0.000^+	0.122	0.002^+	0.023^+	0.002^+	0.796	0.000^+	0.000^+	0.000^+	0.554	0.003^+	0.404	0.003^+
	(0.00)	(1.00)	(0.96)	(1.00)	(0.00)	(0.98)	(0.99)	(0.98)	(0.48)	(1.00)	(0.99)	(1.00)	(0.95)	(1.00)	(0.97)	(1.00)
C. Horizon = 12 months																
Diff Ind	0.050^-				0.025^-				0.028^-				0.000^-			
	(0.00)				(0.01)				(0.00)				(0.00)			
Bayes	0.271	0.000^+			0.083^-	0.003^+			0.042^-	0.000^+			0.367	0.000^+		
	(0.00)	(0.98)			(0.01)	(0.98)			(0.00)	(1.00)			(0.00)	(1.00)		
AR	0.224	0.059^+	0.634		0.040^-	0.421	0.074^-		0.032^-	0.372	0.049^-		0.001^-	0.152	0.000^-	
	(0.00)	(0.98)	(0.06)		(0.01)	(1.00)	(0.08)		(0.00)	(0.98)	(0.02)		(0.00)	(0.83)	(0.01)	
RW	0.557	0.000^+	0.005^+	0.000^+	0.087^-	0.002^+	0.055^+	0.000^+	0.097^-	0.000^+	0.000^+	0.000^+	0.484	0.001^+	0.187	0.000^+
	(0.03)	(1.00)	(0.99)	(0.99)	(0.01)	(1.00)	(0.98)	(1.00)	(0.00)	(1.00)	(0.99)	(1.00)	(0.08)	(1.00)	(0.96)	(1.00)

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 sign indicates that the test rejects equal conditional predictive ability at the 10% level and that the method in the column outperforms the method in the row more than 50% of the time. A minus sign indicates that the test rejects equal conditional predictive ability at the 10% level and that the method in the column is outperformed by the method in the row more than 50% of the time. For example, for the producer price index at the 1-month horizon, equal conditional predictive ability of the sequential and the diffusion index methods is rejected with a p-value of 0 and the sequential method outperforms the diffusion index method 2% of the time.

Table 5. Unconditional predictive ability tests. Real variables.

	Industrial production				Personal income				Mfg & trade sales				Nonag. employment			
Bench	Seq.	Diff Ind	Bayes	AR	Seq.	Diff Ind	Bayes	AR	Seq.	Diff Ind	Bayes	AR	Seq.	Diff Ind	Bayes	AR
								A. Horizon = 1 month								
Diff Ind	0.036^{-}				0.055^{-}				0.120				0.107			
	(3.82)				(1.83)				(5.37)				(3.23)			
Bayes	0.029^{-}	0.028^{-}			0.054^{-}	0.554			0.114	0.127			0.103	0.502		
	(4.22)	(1.10)			(1.79)	(0.98)			(5.66)	(1.05)			(3.34)	(1.03)		
AR	0.046^{-}	0.027^{+}	0.002^{+}		0.099	0.010^{-}	0.091^{+}		0.126	0.064^{+}	0.003		0.139	0.167	0.091^{+}	
	(3.37)	(0.88)	(0.80)		(1.64)	(0.89)	(0.91)		(4.92)	(0.92)	(0.87)		(2.57)	(0.80)	(0.77)	
RW	0.065^{-}	0.083^{+}	0.029^{+}	0.227	0.160	0.039^{+}	0.058^{+}	0.184	0.126	0.062^{+}	0.002^{+}	0.910	0.383	0.012^{+}	0.008^{+}	0.010^{+}
	(2.97)	(0.78)	(0.70)	(0.88)	(1.50)	(0.82)	(0.84)	(0.92)	(4.91)	(0.91)	(0.87)	(0.01)	(1.63)	(0.50)	(0.49)	(0.63)
								B. Horizon = 6 months								
Diff Ind	0.011^{-}				0.026^{-}				0.002^{-}				0.004^{-}			
	(1.67)				(6.23)				(3.20)				(2.10)			
Bayes	0.001^{-}	0.294			0.022^{-}	0.010^{-}			0.001^{-}	0.368			0.002^{-}	0.051^{-}		
	(1.83)	(1.10)			(7.37)	(1.18)			(3.40)	(1.06)			(2.41)	(1.15)		
AR	0.680	0.175	0.058^{+}		0.037^{-}	0.149	0.025^{+}		0.006^{-}	0.092^{+}	0.028^{+}		0.046^{-}	0.253	0.084^{+}	
	(1.11)	(0.67)	(0.61)		(4.78)	(0.77)	(0.65)		(2.21)	(0.69)	(0.65)		(1.63)	(0.77)	(0.68)	
RW	0.921	0.156	0.060^{+}	0.145	0.039^{-}	0.196	0.057^{+}	0.655	0.005^{-}	0.113	0.038^{+}	0.300	0.729	0.039^{+}	0.013^{+}	0.013^{+}
	(1.03)	(0.62)	(0.56)	(0.93)	(4.62)	(0.74)	(0.63)	(0.97)	(2.25)	(0.70)	(0.66)	(1.02)	(1.09)	(0.52)	(0.45)	(0.67)
								C. Horizon = 12 months								
Diff Ind	0.000^{-}				0.001^{-}				0.000^{-}				0.005^{-}			
	(3.77)				(2.97)				(3.21)				(2.02)			
Bayes	0.000	0.442			0.001	0.009			0.000	0.835			0.001	0.034		
	(4.03)$^{-}$	(1.07)			(3.42)$^{-}$	(1.15)$^{-}$			(3.27)$^{-}$	(1.02)			(2.44)$^{-}$	(1.21)$^{-}$		
AR	0.035	0.064	0.034		0.017	0.076	0.025		0.030	0.023	0.007		0.238	0.070	0.016	
	(1.80)$^{-}$	(0.48)$^{+}$	(0.45)$^{+}$		(2.08)$^{-}$	(0.70)$^{+}$	(0.61)$^{+}$		(1.72)$^{-}$	(0.54)$^{+}$	(0.53)$^{+}$		(1.40)	(0.70)$^{+}$	(0.58)$^{+}$	
RW	0.036^{-}	0.063^{+}	0.032^{+}	0.772	0.034^{-}	0.083^{+}	0.033^{+}	0.212	0.030^{-}	0.023^{+}	0.007^{+}	0.253	0.781	0.008^{+}	0.002^{+}	0.008^{+}
	(1.83)	(0.49)	(0.46)	(1.02)	(1.95)	(0.65)	(0.57)	(0.93)	(1.72)	(0.54)	(0.53)	(1.00)	(1.08)	(0.53)	(0.44)	(0.77)

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 the ratios of MSEs for the method in the column relative to the method in the row. A plus sign indicates that the test rejects equal unconditional predictive ability at the 10% level and that the method in the column has smaller MSE than the method in the row. A minus sign indicates that the test rejects equal unconditional predictive ability at the 10% level and that the method in the column has higher 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 .002 and the Bayesian shrinkage method outperforms the AR method with a MSE ratio of .80.

Table 6. Unconditional predictive ability tests. Price indexes.

	CPI				CPI exc. food				Consumption deflator				Producer price index			
Bench	Seq.	Diff Ind	Bayes	AR	Seq.	Diff Ind	Bayes	AR	Seq.	Diff Ind	Bayes	AR	Seq.	Diff Ind	Bayes	AR
A. Horizon = 1 month																
Diff Ind	0.181				0.249				0.049^-				0.004^-			
	(3.45)				(5.83)				(1.56)				(1.71)			
Bayes	0.196	0.472			0.255	0.194			0.065^-	0.166			0.054^-	0.019^+		
	(3.19)	(0.93)			(5.54)	(0.95)			(1.46)	(0.93)			(1.31)	(0.77)		
AR	0.186	0.287	0.635		0.251	0.194	0.375		0.049^-	0.941	0.150		0.004^-	0.462	0.021^-	
	(3.36)	(0.97)	(1.05)		(5.72)	(0.98)	(1.03)		(1.56)	(1.00)	(1.07)		(1.75)	(1.03)	(1.34)	
RW	0.229	0.301	0.261	0.362	0.262	0.043^+	0.131	0.055^+	0.164	0.015^+	0.030^+	0.009^+	0.030^-	0.108	0.823	0.102
	(2.81)	(0.81)	(0.88)	(0.84)	(5.26)	(0.90)	(0.95)	(0.92)	(1.30)	(0.84)	(0.90)	(0.84)	(1.29)	(0.75)	(0.98)	(0.73)
C. Horizon = 6 months																
Diff Ind	0.098^-				0.030^-				0.000^-				0.000^-			
	(2.67)				(3.29)				(1.93)				(2.30)			
Bayes	0.306	0.000^+			0.079^-	0.003^+			0.005^-	0.002^+			0.988	0.000^+		
	(1.65)	(0.62)			(2.28)	(0.69)			(1.50)	(0.78)			(1.00)	(0.43)		
AR	0.145	0.143	0.003^-		0.037^-	0.101	0.004^-		0.000^-	0.747	0.001^-		0.000^-	0.801	0.000^-	
	(2.27)	(0.85)	(1.37)		(2.94)	(0.89)	(1.29)		(1.97)	(1.02)	(1.32)		(2.24)	(0.98)	(2.25)	
RW	0.742	0.000^+	0.001^+	0.000^+	0.135	0.002^+	0.011^+	0.001^+	0.732	0.000^+	0.000^+	0.000^+	0.476	0.004^+	0.188	0.003^+
	(1.15)	(0.43)	(0.70)	(0.51)	(1.91)	(0.58)	(0.84)	(0.65)	(1.05)	(0.54)	(0.70)	(0.53)	(0.84)	(0.37)	(0.84)	(0.37)
C. Horizon = 12 months																
Diff Ind	0.045^-				0.015^-				0.005^-				0.000^-			
	(3.63)				(4.78)				(5.67)				(2.76)			
Bayes	0.095^-	0.000^+			0.024^-	0.002^+			0.009^-	0.000^+			0.339	0.000^+		
	(2.48)	(0.68)			(3.70)	(0.78)			(4.33)	(0.76)			(1.22)	(0.44)		
AR	0.068^-	0.110	0.389		0.015^-	0.289	0.059^-		0.006^-	0.338	0.146		0.000^-	0.516	0.000^-	
	(2.79)	(0.77)	(1.12)		(4.34)	(0.91)	(1.17)		(4.97)	(0.88)	(1.15)		(2.52)	(0.91)	(2.07)	
RW	0.344	0.000^+	0.001^+	0.000^+	0.032^-	0.002^+	0.036^+	0.001^+	0.026^-	0.000^+	0.000^+	0.000^+	0.642	0.004^+	0.075^+	0.003^+
	(1.52)	(0.42)	(0.61)	(0.54)	(3.07)	(0.64)	(0.83)	(0.71)	(2.81)	(0.50)	(0.65)	(0.56)	(0.88)	(0.32)	(0.72)	(0.35)

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 the ratios of MSEs for the method in the column relative to the method in the row. A plus sign indicates that the test rejects equal unconditional predictive ability at the 10% level and that the method in the column has smaller MSE than the method in the row. A minus sign indicates that the test rejects equal unconditional predictive ability at the 10% level and that the method in the column has higher MSE than the method in the row. For example, for the producer price index at the 1-month horizon, equal unconditional predictive ability of the Bayesian shrinkage and the AR methods is rejected with a p-value of .021 and the Bayesian shrinkage method is outperformed by the AR method with a MSE ratio of 1.34.

Table 7. Decision rule assessment. Performance of the "hybrid" forecast of Industrial production.

	Horizon = 1 month				Horizon = 6 months				Horizon = 12 months			
Bench	Seq.	Diff Ind	Bayes	AR	Seq.	Diff Ind	Bayes	AR	Seq.	Diff Ind	Bayes	AR
Diff Ind	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

The hybrid forecast is obtained by recursively applying the pairwise decision rule described in Section 4 (using a rolling window of size 200) to select between the method in the row and the method in the column. Entries equal 1 if the MSE of the hybrid forecast 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.