Tests of conditional predictive ability

Raffaella Giacomini and Halbert White*

University of California, San Diego

This version: April 2003

Abstract

We argue that the current framework for predictive ability testing (e.g.,West, 1996) is not necessarily useful for real-time forecast selection, i.e., for assessing which of two competing forecasting methods will perform better in the future. We propose an alternative framework for out-of-sample comparison of predictive ability which delivers more practically relevant conclusions. Our approach is based on inference about *conditional* expectations of forecasts and forecast errors rather than the *unconditional* expectations that are the focus of the existing literature. We capture important determinants of forecast performance that are neglected in the existing literature by evaluating what we call the forecasting method (the model and the parameter estimation procedure), rather than just the forecasting model. Compared to previous approaches, our tests are valid under more general data assumptions (heterogeneity rather than stationarity) and estimation methods, and they can handle comparison of both nested and non-nested models, which is not currently possible. To illustrate the usefulness of the proposed tests, we compare 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. We also wish to thank Lutz Kilian for insightful suggestions and Farshid Vahid 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 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. Corresponding author: Raffaella Giacomini, Department of Economics, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093-0508, U.S.A. Web page: http://www.econ.ucsd.edu/~rgiacomi. E-mail: rgiacomini@ucsd.edu.

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 that economic forecasters often face is how to select the best forecasting method from a set of two (or more) alternatives. The econometric answer to this problem is to develop tests for comparing the predictive ability of two alternative forecast methods, given the forecaster's loss function. The literature on forecast comparison has witnessed a renaissance in recent years, and a number of authors have proposed econometric techniques for forecast comparison under general loss functions, known as out-of-sample predictive ability testing. 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), Chao, Corradi and Swanson (2001), among others, and it represents a generalization of several existing evaluation techniques which typically restricted attention to a particular loss function (e.g., Granger and Newbold 1977, McCulloch and Rossi 1990, Leitch and Tanner 1991, West, Edison and Cho 1993, Harvey, Leybourne and Newbold 1997).

In this paper, we argue that the current framework for out-of-sample predictive ability testing (which in the remainder of the paper we consider to be represented by West, 1996) is not necessarily appropriate for real-time forecast selection, i.e., for assessing which of two competing forecasting methods will give better forecasts in the future. We propose an alternative approach to out-of-sample predictive ability testing that delivers inferences that are more relevant to economic forecasters. Our tests can be applied to multi-step point, interval, probability or density forecasting, and they can be viewed as a generalization of the tests of West (1996) since they are applicable in all cases in which his tests are applicable and in many more besides.

From a methodological point of view, the main idea of the paper is to view the problem of forecast evaluation as a problem in inference about *conditional* expectations of forecasts and forecast errors rather than the *unconditional* expectations that are the focus of the approach of West (1996).

An important distinction between our approach and the existing literature is that we consider the object of the evaluation to be what we call the "forecasting method", which includes not only the forecast model but also a number of choices that must be made by the forecaster at the time of the prediction, such as which estimation procedure to choose and which data to use for estimation. The current approach to forecast evaluation focuses instead solely on the forecast model. The reason for evaluating the forecasting method and not just the model is that all elements of the method can affect future forecast performance: a good model can produce bad forecasts if its parameters are not precisely estimated or if they change over time. The fact that we consider the forecasting method rather than the model implies that our tests can lead to a different conclusion than West's (1996) tests. Suppose for example that one of the two models is correctly specified but has a large number of parameters, while the competitor is a simpler, misspecified model. West's (1996) test will tend to choose the large model, while our tests may choose the forecasting method that uses the small model, especially if we are in the presence of high estimation uncertainty.

Our approach is applicable in many situations where the tests of West (1996) are not valid. One such case is when the data are heterogeneous, in the form of time-varying underlying processes for the series of interest. As emphasized by Clements and Hendry (1998, 1999), this is a more realistic assumption for economic forecasting contexts than the assumption of stationarity that is typically made in the literature. The assumption of heterogeneity also affects the approach to estimation. In this context, instead of considering a recursive forecasting scheme, where the estimation window expands over time, it makes sense to consider a rolling window forecast procedure where the forecasts are based on a moving window of the data which discards old observations. The size of the estimation window can itself be time-varying, as in the procedure suggested by Pesaran and Timmermann (2002). A fundamental difference with the existing literature is that here we consider the estimation window to be a component of the forecasting method under evaluation. In the existing literature, instead, the sample split between estimation and evaluation samples is arbitrary and typically there is little guidance on how to choose it in practice.

Another situation where the tests of West (1996) are not applicable is in comparing forecasts based on nested models. This is an important comparison because many models considered for forecasting are naturally derived as generalizations of existing models and it is often of interest to test if a larger, more sophisticated model can outperform a simple, nested benchmark model. Our framework permits a unified treatment of nested and non-nested models.

Finally, the current framework for predictive ability testing is not valid when the forecasts are obtained by using estimation methods such as Bayesian estimation, semi-parametric, or nonparametric estimation. Our framework, instead, can accommodate such estimation procedures and can thus be used to compare the impact on forecast performance of using different estimation techniques, a question that cannot be answered within the current framework.

A final, practical advantage of our tests is that they are easily computed using standard regression packages, whereas the existing tests can be quite difficult to compute or have limiting distributions that are context-specific (e.g., Clark and McCracken, 2001).

To illustrate the usefulness of the conditional predictive ability tests, we consider the problem of macroeconomic forecasting using a large number of predictors and compare forecasts of eight macroeconomic variables obtained by employing leading methods for parameter reduction: a sequential procedure which is a simplified version of the general-to-specific model selection approach implemented by Hoover and Perez (1999), the "diffusion indexes" approach of Stock and Watson

(2002) and the use of Bayesian shrinkage estimators (Litterman, 1986). We use the data set of Stock and Watson (2002), including monthly U.S. data on 146 macroeconomic variables and evaluate 1-, 6- and 12-month ahead forecasts of four measures of real activity and four price indexes obtained using the different forecasting methods. The general conclusion is that for the price indexes the forecast performance of the three methods is indistinguishable from that of a univariate autoregression. For the real variables, instead, Bayesian shrinkage appears to be the preferred method. Finally, the sequential model selection approach performs poorly for most variables and forecast horizons, and it is often outperformed by the naive autoregressive and random walk benchmarks.

2 Unconditional and conditional approaches to predictive ability testing

To illustrate the differences between conditional and unconditional out-of-sample predictive ability testing, suppose we are interested in comparing the accuracy of two competing forecasting models $f_t(\beta_1)$ and $g_t(\beta_2)$ for the conditional mean of the variable of interest Y_{t+1}, given a squared error loss function. The dependence of the forecasts on parameters β_1 and β_2 indicates that in this example the forecasting models are parametric. The approach of West (1996) consists of testing the null hypothesis of equal accuracy of the two forecasts formulated as

$$H_0 : E[(Y_{t+1} - f_t(\beta_1^*))^2 - (Y_{t+1} - g_t(\beta_2^*))^2] = 0, \tag{1}$$

where β_1^* and β_2^* are population values of the parameters (i.e., probability limits of the parameter estimates). The null hypothesis (1) can be interpreted as saying that the two forecast models are equally accurate on average. If the null hypothesis is rejected, one would choose the model yielding the lower loss. Notice that a test of the null hypothesis (1) will tend to choose a forecast based on a correctly specified model (i.e., the test will choose the correctly specified model asymptotically with probability $1 - \alpha$, where α is the level of the test).[1] A focus on the null hypothesis (1) is thus justifiable if one is interested in establishing which of two models better approximates the data-generating process. However, even a model that well approximates the data-generating process may forecast poorly, for example in the case that its parameters are imprecisely estimated. If the question is which model will give better forecasts in the future, therefore, it is not clear that (1) is in fact the appropriate null hypothesis.

The central idea of this paper is to test a null hypothesis different than (1), where the expectation is conditional on the information set $\mathcal{F}_t$ available at time t and the losses depend on the parameter

[1] To see why, suppose forecast 1 is based on a correctly specified model, which implies that $f_t(\beta_1^*)$ is the true conditional mean of Y_{t+1}. Also assume for simplicity that the two forecasts are based on the same information set. Since the conditional mean is the optimal forecast for a squared error loss function, $f_t(\beta_1^*)$ minimizes the expected loss: $E[(Y_{t+1} - f_t(\beta_1^*))^2] < E[(Y_{t+1} - f_t')^2]$ for any other forecast f_t', and thus in particular for $g_t(\beta_2^*)$.

estimates at time t, $\hat{\beta}_{1t}$ and $\hat{\beta}_{2t}$, rather than on their probability limits:

$$H_0 : E[(Y_{t+1} - f_t(\hat{\beta}_{1t}))^2 - (Y_{t+1} - g_t(\hat{\beta}_{2t}))^2 | \mathcal{F}_t] = 0 \text{ almost surely, } t = 1, 2, ... \qquad (2)$$

We call a test of the hypothesis (2) a test of equal *conditional predictive ability.* The motivation for conducting inference about a conditional, rather than an unconditional, expectation is that it more closely represents the real-time problem of a forecaster. In particular, we can view this hypothesis as saying that the forecaster cannot predict which of the two forecasts will be more accurate, given what is known today.

Further motivation for expressing the null hypothesis in terms of time-t parameter estimates rather than probability limits is that they are more relevant for the forecaster: since the population parameters are not known and must be estimated, it is the actual future loss that is of interest to the forecaster, rather than that based on some population value that is only attained in the limit. As a result, whereas the unconditional tests restrict attention to the forecast model, the conditional approach allows evaluation of the forecasting method, which includes the model, the estimation procedure and the possible choice of estimation window. Viewed this way, it appears obvious that considering the forecasting method as a whole is appropriate, as each of its components can have a potential impact on future forecast performance.

In the following subsections, we outline in detail the directions along which the conditional testing framework represents a more realistic environment for forecast evaluation and discuss how it directly accounts for different determinants of forecast performance that are neglected by the unconditional framework.

2.1 Heterogeneity of economic data

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 underlying process generating the series to be forecast. It is thus of fundamental importance to develop evaluation techniques that take into account the possibly heterogeneous nature of economic variables. In this paper, we therefore work with the assumption that the data generating process is heterogeneous rather than stationary.[2] In our view, this is a realistic and practical assumption for economic forecasting contexts and more plausible than the perhaps idealistic assumption of stationarity typically made in the unconditional predictive ability literature. Specific sources for 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

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

definition of the measured variables may change from time to time; which entities are measured in constructing the variables measured changes; budgets for the economic, demographic, and statistical agencies measuring the processes of interest change, leading to the possibility of greater or lesser care in producing the official numbers on strict time schedules; and directors and other key personnel of these agencies regularly join and leave, leading to intentional or unintentional variations in the processes and procedures that produce the official time series. Heterogeneity in even one of these sources would produce heterogeneity in the observed series. 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 themselves are comprised of a variety of forces that operate as further sources of heterogeneity, affecting either the nature of the commodity itself, or the way the commodity is traded. With regard to the latter, the laws and regulations governing trade change, and the technologies used by buyers and sellers of the commodities change. (The onward march of both computing and software technology is an obvious example.) With regard to the former, the laws and regulations governing for example the behavior of firms represented by equity assets change, as do market conditions and technologies used by such firms. Taken together, these factors make it plausible in our view that the relations between variables of interest this month and next relevant for forecasting are somewhat 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 accepted as an accurate description of economic time series, appropriate methods for model-based forecasting and forecast evaluation need to be applied. In general, it seems appropriate in a time-varying environment to consider estimators with finite memory, rather than basing forecasts on an expanding window of data. An example is the practice of specifying and estimating forecast models over a rolling window of the data, as a way to accommodate a data generating process that varies slowly over time (e.g., Fama and McBeth, 1973, Gonedes, 1973). The size of the estimation window may itself be time-varying, as was recently suggested by Pesaran and Timmermann (2002), who propose a recursive procedure which detects breaks in real time and then uses an appropriate subset of the data for estimation. Further, the estimators can usefully incorporate time weights which may assign decreasing importance to observations from the more distant past. The use of these methods in the production of forecasts has important implications for the evaluation procedure. The approach to out-of-sample testing in the unconditional predictive ability framework is to arbitrarily split the data into an estimation and an evaluation sample, and to obtain the asymptotic distribution of the test statistic under the assumption that both the in-sample and the out-of-sample sizes diverge to infinity. The choice of sample split is thus a finite sample artifice. In contrast, in our conditional framework the size of the estimation window and the possible time weighting are treated as choice variables of the forecast method, and as such they can be evaluated along with the forecast model and the estimation procedure as parts of the forecasting method under analysis.

2.2 Estimation uncertainty

As emphasized by Clements and Hendry (1998, 1999) and Ericsson (2002), parameter estimation uncertainty is an important determinant of forecast performance. Our conditional tests directly account for the effects of estimation uncertainty on forecast performance by expressing the null hypothesis in terms of parameter estimates and by considering finite window estimation, which leads to asymptotically non-vanishing estimation uncertainty. In contrast, the unconditional framework does not take into account differing model complexities, unless explicitly incorporated into the accuracy measure (e.g., AIC or BIC); further, the presence of probability limits in the unconditional null hypothesis (1) means that different estimators that converge to the same limit will lead to the same conclusion. As a result, the unconditional tests are not able to detect superior forecasting performance which is due to reduced estimation uncertainty. For example, consider the case of comparing the accuracy of nested models in the unconditional framework. If the smaller model is correctly specified, the forecast errors from the two models calculated at the probability limits of the parameters are identical, and the null hypothesis (1) is automatically satisfied (this would hold for any loss function). In other words, one would conclude that the two models yield equally accurate forecasts, regardless of the number of excess parameters contained in the larger model. A test of this sort may thus lead to misleading conclusions if the goal is real-time forecast selection.

2.3 Out-of-sample versus in-sample testing

The literature on forecast evaluation has long argued that out-of-sample, rather than in-sample, testing is the "true" test of a forecast model. As Granger (1999, p. 65) observes, the potentially large number of parameters compared with the relative scarcity of macroeconomic data "leads to worries that a model presented for consideration is the result of considerable specification searching ..., data mining, or data snooping (in which data are used several times). Such a model might well appear to fit the data, in sample, rather well but will often not perform satisfactorily out-of-sample". This is related to the problem of overfitting: a good fit may result from explaining not only the stable relationships that are useful for forecasting but also possible accidental relationships that are specific to the sample. Out-of-sample evaluation of forecast performance, on the contrary, simulates a real-time forecast scenario where the quality of a forecasting method is directly measured against the actual data. The unreliability of in-sample testing is particularly evident if the data in the sample are generated by a time-varying process, whereas out-of-sample testing can incorporate this heterogeneity through recursive specification and estimation of the model. Our conditional testing framework is fully congruent with this motivation for out-of-sample testing: it is valid under heterogeneity of the data-generating process, and it is based on the forecasts and forecast errors actually observed, rather than viewing them as estimates of some population quantities. In the unconditional predictive ability framework, on the other hand, it is less clear why one should use out-of-sample rather than in-sample testing if the goal is to test hypotheses about population parameters under the assumption of stationarity. This point is well made by Inoue and Kilian (2002), who argue that if the goal of the testing procedure is to assess population predictability (corresponding to testing a null hypothesis of equal *unconditional* predictive ability of two nested models, e.g., as in (1)), the use of out-of-sample testing involves an unnecessary loss of information, whereas the in-sample test of the same hypothesis utilizes all the information available and thus leads to power gains in finite samples.

2.4 Practical advantages of the conditional tests

In addition to the methodological considerations just articulated, there are also significant practical advantages to the approach advocated here. The main benefit of our approach is that it allows a unified treatment of nested and non-nested models, while the existing testing framework of West (1996) is only valid under non-nestedness. Unconditional tests of predictive ability for nested models have been proposed by Clark and McCracken (2001), among others, but they lack the general applicability of West (1996)'s results, as the test statistics have complicated limiting distributions that are context-specific. As discussed in section 3.2, the different treatment of nested and non-nested models in the unconditional framework is due to the fact that the asymptotic distribution of the test statistic relies on convergence of the parameter estimates to their probability limits, and this limiting behavior differs in the two cases. The fact that we don't rely on such convergence in the conditional approach instead makes it possible to consider nested and non-nested models in the same framework. A second advantage of the conditional tests is that they do not impose restrictions on the estimation procedure utilized to produce the forecasts, while the approach of West (1996) rules out, e.g., Bayesian, semi-parametric, and non-parametric estimation. Finally, our tests are simple to compute due to the imposition of a particular time dependence structure under the null hypothesis (e.g., martingale difference sequences for the one-step-ahead forecasts), which leads to a computationally simple expression for the asymptotic variance estimator.

3 Theory

3.1 Description of the environment

Consider a stochastic process $W \equiv \{W_t : \Omega \longrightarrow \mathbb{R}^{s+1}, s \in \mathbb{N}, t = 1, \ldots, T\}$ 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 $Y_t : \Omega \rightarrow \mathbb{R}$ is the variable of interest and $X_t : \Omega \rightarrow \mathbb{R}^s$ 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. Consider a situation where two alternative models are used to forecast the variable of interest τ steps ahead, $Y_{t+\tau}$. The forecasts are formulated at time t and are based on the information set $\mathcal{F}_t$. Denote the two forecasts 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 forecast formulated at time t is a measurable function of a sample of at most size m, consisting of the m most recent observations. Recall that we do not restrict attention to point forecasting. Our framework accommodates evaluation of point, interval, probability, and density forecasts. 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 forecasts. The estimator $\hat{\beta}_{m,t}$ can be further selected to minimize a weighted loss function over the estimation period, where smaller weights are typically assigned to observations from the more distant past. For example, for a linear model $Y_t = X_t\beta + u_t$ and a quadratic loss function, we can consider the family of weighted least squares estimators $\hat{\beta}_{m,t} = \min_\beta \sum_{s=t-m+1}^{t}(y_s - x_s\beta)^2 w_{m,s}$, where $\{w_{m,s}\}$ is a sequence of weights assigned to the observations in the estimation sample that can be selected by the user (for example one may assign exponentially decreasing weights to the observations further away from t).

We emphasize that the estimators may be parametric, semi-parametric or non-parametric. The only requirement here is that m (the maximum estimation window size) must be finite. All of the elements listed above - the model, the estimation procedure, the size of the estimation window and the estimation weight function - are treated as dimensions of choice by the user and are part of what we call the "forecasting method" under evaluation.

The evaluation is performed in a simulated out-of-sample fashion. Let T be the size of the sample available. Since the data indexed $1, ..., m$ are used for estimation of the first set of parameters, the first $\tau-$step ahead forecasts are formulated at time m and compared to the realization $y_{m+\tau}$. The second set of forecasts is produced by moving the estimation window forward one step and estimating the parameters on data indexed $2, ..., m+1$. These forecasts are compared to the realization $y_{m+1+\tau}$. The procedure is thus iterated and the last forecasts are generated at time $T - \tau$, by estimating the parameters on data indexed $T - \tau - m + 1, ..., T - \tau$, and they are compared to y_T. This rolling window procedure yields a sequence of $n \equiv T - \tau - m + 1$ forecasts and relative forecast errors.

The sequence of out-of-sample forecasts thus produced is evaluated by selecting a loss function $L_{t+\tau}(Y_{t+\tau}, \hat{f}_{m,t})$, which 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), Lopez (2001), Giacomini and Komunjer (2002) and Giacomini (2002). For simplicity, let $f_t \equiv \hat{f}_{m,t}$ and consider $\tau = 1$.

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: $L_{t+1}(Y_{t+1}, f_t) = \exp(a(Y_{t+1} - f_t)) - a(Y_{t+1} - f_t) - 1$, $a \in \mathbb{R}$.
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: $L_{t+1}(Y_{t+1}, f_t) = \log f_t(Y_{t+1})$, where f_t is in this case the density forecast of Y_{t+1}

3.2 One-step conditional predictive ability test

For a given loss function, we write the null hypothesis of equal conditional predictive ability of forecasts f and g as

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

Due to certain computational issues, we consider separately the cases of one-step and multi-step forecast horizons.

3.2.1 Null hypothesis

When $\tau = 1$, the null hypothesis claims that the out-of-sample sequence $\{\Delta L_{m,t+1}, \mathcal{F}_t\}$ is a martingale difference sequence (*mds*). In this case, the conditional moment restriction (3) is equivalent to stating that $E[h_t \Delta L_{m,t+1}] = 0$, for all $\mathcal{F}_t-$ measurable functions h_t. Let us 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 $H_{0,h} : E[h_t \Delta L_{m,t+1}] = 0$. The standard unconditional approach to predictive ability testing corresponds to testing the hypothesis $H_{0,h}$ with $h_t = 1$ and with the parameter estimate $\hat{\beta}_{m,t}$ replaced with its probability limit β^*.

For fixed m, standard asymptotic normality arguments suggest using a Wald-type test statistic of the form

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

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 $T^h_{n,m} > \chi^2_{q,1-\alpha}$, 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 (Conditional predictive accuracy test) *For forecast horizon $\tau = 1$, maximum estimation window size $m < \infty$ and $q \times 1$ test function sequence $\{h_t\}$ suppose:*

(i) $\{W_t\}$, $\{h_t\}$ are mixing sequences 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), $T^h_{n,m} \xrightarrow{d} \chi^2_q$ 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.

2. The asymptotic distribution is obtained for the number of out-of-sample observations going to infinity, whereas the estimation sample size m remains finite. This leads to asymptotically non-vanishing estimation uncertainty. In contrast, in the unconditional framework of West (1996), both the in-sample and the out-of-sample sizes grow, causing estimation uncertainty to vanish asymptotically. A result of letting both m and n grow is that the choice of how to split the available sample into in-sample and out-of-sample portions is arbitrary, while in our framework the choice of estimation window (up to some maximum m) is part of the forecasting method under evaluation. Also notice that the requirement of finite estimation window rules out the use of a recursive forecasting scheme, which utilizes an expanding estimation window.

3. Assumption (iii), imposing positive definiteness of the asymptotic variance of the test statistic, is related to a similar requirement made in the existing literature about predictive ability testing (e.g., West, 1996, McCracken, 2000), but it differs in a fundamental way. In that literature, the size

of the estimation window is assumed to grow at the same rate or faster than the out-of-sample size, which means that the asymptotic variance of the test statistic is computed at the probability limits of the parameters. Because of the focus on this limiting behavior, the asymptotic variance matrix may be singular when the forecasts are based on nested models. In contrast, in the conditional framework the size of the estimation window remains finite as the prediction sample size n grows to infinity, which prevents the parameter estimates from reaching their probability limits. This makes our tests applicable to both nested and non-nested models.

4. The test statistic for conditional predictive ability test is straightforward to compute. A further simplifying feature is the fact that the null hypothesis imposes a particular time dependence structure (in this case that of a martingale difference sequence), which implies that the asymptotic variance can be consistently estimated by the sample variance.

The following results provide computationally convenient ways to obtain the test statistic for the conditional predictive ability test using standard regression packages.

Corollary 2 *Under the assumptions of Theorem 1, the test statistic $T^h_{n,m}$ 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 $t = m, ..., T-1$.*

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 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 in a given application, the true distribution of the regression-based test statistic in Corollary 3 may be better approximated by its asymptotic distribution than the statistic of Corollary 2, and it might thus deliver better inference.

3.2.2 Alternative hypothesis

We now analyze the behavior of the test statistic $T^h_{n,m}$ under a form of global alternative to the null hypothesis 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, we will be able to obtain tests 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 $T^h_{n,m}$ 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 $c \in \mathbb{R}$, $P[T^h_{n,m} > c] \to 1$ as $n \to \infty$.*

Notice that H_0 and $H_{A,h}$ are not necessarily exhaustive. For a given test function sequence $\{h_t\}$, 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*. The resulting test may thus have no power against alternatives for which $\Delta L_{m,t+1}$ is uncorrelated with the chosen test function (and thus $E[\bar{Z}'_{m,n'}]E[\bar{Z}_{m,n'}] = 0$) but it is correlated with some element of the information set $\mathcal{F}_t$ that is not contained in h_t. In other words, the properties of the test will depend on the chosen test function. The flexibility in the choice of test function is both a shortcoming and an advantage of our testing framework. On the one hand, for any given selection of 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 specific direction. Further, using methods developed recently in the statistics literature, one may be able to identify with some confidence which elements of h_t are responsible for rejection of the null hypothesis using the notion of False Discovery Rate for multiple comparison testing procedures (Benjamini and Hochberg, 1995).

In practice, the test function is chosen by the researcher to embed elements of the information set $\mathcal{F}_t$ that are believed to have potential explanatory power for the future difference in predictive ability. Examples are, e.g., indicators of past relative performance or other variables that may help distinguish between the forecast performance of the two methods, such as business cycle indicators that may capture possible asymmetries in relative performance during booms and recessions. When choosing the number of elements for the test function h_t, it is important to 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 have little or no power against the alternative for which $\Delta L_{m,t+1}$ is not *mds*. As a consequence, the test would 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 truly 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 $\tau > 1$, the null hypothesis (3) of equal conditional predictive ability of forecasts f and g implies in particular 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 $cov(h_t \Delta L_{m,t+\tau}, h_{t-j}\Delta L_{t+\tau-j}(\hat{\beta}_{m,t-j})) = 0$ for all $j \geq \tau$. Similarly to the previous section, we are able to exploit this simplifying feature in the derivation of the test statistic. Using reasoning that mirrors the development of the test for the one-step horizon, we construct a test of

$$H_{0,\tau} : E[\Delta L_{m,t+\tau}|\mathcal{F}_t] = 0 \tag{6}$$

against the global alternative

$$H_{A,h,\tau} : E[\bar{Z}'_{m,n}]E[\bar{Z}_{m,n}] \geq \delta > 0 \text{ for all } n \text{ sufficiently large,} \tag{7}$$

where h_t is a $q \times 1$ $\mathcal{F}_t-$ measurable test function and $\bar{Z}_{m,n} \equiv n^{-1}\sum_{t=m}^{T-\tau} Z_{m,t+\tau}$, $Z_{m,t+\tau} \equiv h_t \Delta L_{m,t+\tau}$. For a fixed maximum estimation window length m, the test is based on the statistic

$$T^h_{n,m,\tau} = n(n^{-1}\sum_{t=m}^{T-\tau} h_t \Delta L_{m,t+\tau})'\tilde{\Omega}_n^{-1}(n^{-1}\sum_{t=m}^{T-\tau} h_t \Delta L_{m,t+\tau}) = n\bar{Z}'_{m,n}\tilde{\Omega}_n^{-1}\bar{Z}_{m,n} \tag{8}$$

where $\tilde{\Omega}_n \equiv n^{-1}\sum_{t=m}^{T-\tau} Z_{m,t+\tau}Z'_{m,t+\tau} + n^{-1}\sum_{j=1}^{\tau-1} w_{n,j}\sum_{t=m+j}^{T-\tau}[Z_{m,t+\tau}Z'_{m,t+\tau-j} + Z_{m,t+\tau-j}Z'_{m,t+\tau}]$, with $w_{n,j}$ a weight function such that $w_{n,j} \to 1$ as $n \to \infty$ for each $j = 1, ..., \tau - 1$ (see, e.g., Newey and West, 1987 and Andrews, 1991).

A level α test rejects the null hypothesis of equal conditional predictive ability whenever $T^h_{n,m,\tau} > \chi^2_{q,1-\alpha}$, 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 accuracy test) *For given forecast horizon* $\tau > 1$, *maximum estimation window size* $m < \infty$ *and a* $q \times 1$ *test function sequence* $\{h_t\}$ *suppose:*

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

(ii) $E|Z_{m,t+1,i}|^{r+\delta} < \Delta < \infty$ *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 $H_{0,\tau}$ *in (6),* $T^h_{n,m,\tau} \xrightarrow{d} \chi^2_q$ *as* $n \to \infty$ *and (b) under* $H_{A,h,\tau}$ *in (7), for any constant* $c \in \mathbb{R}$, $P[T^h_{n,m,\tau} > c] \to 1$ *as* $n \to \infty$.

3.4 A decision rule for forecast selection

If the null hypothesis of equal conditional predictive ability of forecast methods f and g is rejected, this raises the possibility that one might be able to select at time T a best forecasting method for time $T+\tau$. Rejection of the null hypothesis is caused by the fact that the test functions $\{h_t\}$ have predictive power for the loss differences $\{\Delta L_{m,t+\tau}\}$ over the out-of-sample period. This suggests that the test function at time T, h_T, can be used to predict which forecast method will yield lower loss at time $T+\tau$, resulting, for example, in the following decision rule:

- Let $\hat{\alpha}_n$ denote the coefficient obtained by regressing $\Delta L_{m,t+\tau} = L_{t+\tau}(Y_{t+\tau}, \hat{f}_{m,t}) - L_{t+\tau}(Y_{t+\tau}, \hat{g}_{m,t})$ on h_t over the out-of-sample period $t = m, ..., T-\tau$. Then choose g if $h_T'\hat{\alpha}_n > c$ and choose f if $h_T'\hat{\alpha}_n < c$, where c is a user-specified threshold.

In general, the plot of the predicted loss differences over the out-of-sample period $\{h_t'\hat{\alpha}_n\}_{t=m}^{T-\tau}$ contains useful information for assessing the relative performance of f and g. For example, one could consider the indicator $I_{n,c} = n^{-1}\sum_{t=m}^{T-\tau} 1\{h_t'\hat{\alpha}_n > c\}$, where $1\{A\}$ is the indicator variable taking the value 1 if A is true and 0 otherwise. $I_{n,c}$ represents the proportion of times that the above decision rule would have chosen forecast method g over the out-of-sample period. In the empirical application in section 5, we utilize the indicator $I_{n,c}$ with $c = 0$ to summarize the relative performance of the forecast methods under analysis.

4 Monte Carlo evidence

In this section, we investigate the size and power properties of our conditional predictive ability test in finite samples of the sizes typically available in macroeconomic forecasting applications. For simplicity, we restrict attention to a squared error loss function and to the one-step forecast horizon.

4.1 Size properties

In order to construct a series of data and forecasts that satisfy the null hypothesis, we exploit the following result.

Proposition 6 $E[(Y_{t+1} - \hat{f}_{m,t})^2 - (Y_{t+1} - \hat{g}_{m,t})^2 | \mathcal{F}_t] = 0$ *if and only if either* $\hat{f}_{m,t} = \hat{g}_{m,t}$ *a.s. or* $E[Y_{t+1}|\mathcal{F}_t] = (\hat{f}_{m,t} + \hat{g}_{m,t})/2$.

We can thus generate data under the null hypothesis

$$H_0 : E[(Y_{t+1} - \hat{f}_{m,t})^2 - (Y_{t+1} - \hat{g}_{m,t})^2 | \mathcal{F}_t] = E[\Delta L_{m,t+1}|\mathcal{F}_t] = 0 \tag{9}$$

by first constructing forecasts $\{\hat{f}_{m,t}, \hat{g}_{m,t}\}$ and then letting $Y_{t+1} = (\hat{f}_{m,t} + \hat{g}_{m,t})/2 + \varepsilon_{t+1}$, where $\varepsilon_{t+1} \,\tilde{}\, i.i.d.\, N(0, \sigma^2)$. One of the important features of our testing framework is its ability to handle heterogeneous data. To create data that exhibits interesting behavior we consider an actual macroeconomic time series $\{W_t\}$, which corresponds to one of the measures of inflation that we consider in the empirical application. Specifically, we let W_t be the second (log) difference of the monthly U.S. Consumer Price Index measured over the period 1959:1-1998:12, for a total sample size $T = 468$. We construct the forecasts $\hat{f}_{m,t}$ and $\hat{g}_{m,t}$ by a rolling window procedure; the first forecast is simply the unconditional mean of the estimation sample, while the second is the forecast implied by an AR(1) model for W_t:

$$\begin{aligned} \hat{f}_{m,t} &= (W_t + ... + W_{t-m+1})/m \\ \hat{g}_{m,t} &= \hat{\alpha}_{m,t} + \hat{\beta}_{m,t} W_t. \end{aligned} \tag{10}$$

We consider a range of values for the size of the estimation sample m and for the variance of the disturbances $\sigma^2 : m = (36, 60, 120, 240, 360)$ and $\sigma^2 = (.1, 1, 3)$. For each pair (m, σ^2) we generate $10,000$ Monte Carlo replications of the time series $\{Y_{t+1}, \hat{f}_{m,t}, \hat{g}_{m,t}\}$ and compute the proportion of rejections of the null hypothesis (9) at the 10% nominal level. The test function is $h_t = (1, \Delta L_{m,t})'$. The results are collected in Table 1.

[TABLE 1 HERE]

From the analysis of Table 1, the test appears to be reasonably well-sized, with a mild tendency to under-reject. The size properties of the test are seemingly unaffected by varying the length of the estimation window and the error variances.

4.2 Power properties

We investigate the power of the CPA test against serially correlated alternatives. In particular, we consider the following alternative hypothesis

$$H_{a,\rho} : E[\Delta L_{m,t+1} | \mathcal{F}_t] = \rho \Delta L_{m,t}, \tag{11}$$

which occurs when $E[Y_{t+1}|\mathcal{F}_t] = (\hat{f}_{m,t} + \hat{g}_{m,t})/2 - \rho \Delta L_{m,t}/(2(\hat{f}_{m,t} - \hat{g}_{m,t}))$. We consider a number of different values for the AR coefficient ρ, ranging from $\rho = 0.05$ to $\rho = 0.5$, at increments of 0.05. For a given ρ, we generate data under the alternative hypothesis (11) by first constructing the forecasts $\{\hat{f}_{m,t}, \hat{g}_{m,t}\}$ as in (10) and then letting $Y_{t+1} = (\hat{f}_{m,t} + \hat{g}_{m,t})/2 - \rho \Delta L_{m,t}/(2(\hat{f}_{m,t} - \hat{g}_{m,t})) + \varepsilon_{t+1}$, where $\varepsilon_{t+1} \tilde{} \, i.i.d.\, N(0,1)$ and the initial value $\Delta L_{m,m}$ is drawn from a standard normal distribution. For each parameterization, we generate $10,000$ Monte Carlo replications of the time series $\{Y_{t+1}, \hat{f}_{m,t}, \hat{g}_{m,t}\}$ and compute the proportion of rejections of the null hypothesis (9) at the 10% nominal level.[3] Figure 1 plots the power curves for $m = (60, 120, 240)$.

[3] We drop the first 100 observations of the generated time series $\{Y_{t+1}, \hat{f}_{m,t}, \hat{g}_{m,t}\}$ to reduce the dependence on the initial observation, which leaves us with a total sample size $T = 368$.

[FIGURE 1 HERE]

The test displays good power properties. For example, more than 50% of the time the test is able to detect the presence of moderately low serial correlation (i.e., an AR coefficient between 0.15 and 0.2). As expected, the power of the test increases as the size of the out-of-sample evaluation data set increases.

5 Application: comparing parameter-reduction methods in macroeconomic forecasting

A problem that often arises in macroeconomic forecasting is the selection of 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 of several parameter-reduction schemes that have been considered in the literature to overcome this so-called "curse of dimensionality". We will consider three leading methods; a sequential model-selection approach based on a simplified general-to-specific modelling strategy (see the overview of Mizon, 1995), the "diffusion indexes" approach of Stock and Watson (2002) and the use of Bayesian shrinkage estimation (Litterman, 1986, Sims and Zha, 1998). We also compare each method to benchmark forecasts. The existing framework for comparison of predictive ability is not appropriate for addressing these issues, since it does not easily accommodate, for example, Bayesian estimation or the presence of estimated regressors. Further, some of the comparisons are between nested models, in which case the existing techniques are not readily applicable. In contrast, our conditional predictive ability approach is naturally well suited for comparison of 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.[4] We refer the reader to SW for a complete description of the data.

[4] These variables coincide with the variables forecasted by SW, with the exception of the consumer price index

5.1 Parameter-reduction methods

Following SW, our approach to multistep-ahead forecasting is to consider forecast models that project the $\tau-$step ahead variable $Y_{t+\tau}^{\tau}$ onto predictor variables measured at time t. Both the dependent variable and the predictors are transformations of the original data that render these variables $I(0)$. In particular, the real variables are modeled as being $I(1)$ in logarithms and the price indexes as $I(2)$ in logarithms. If RAW_t is the original datum at time t, this implies that $Y_{t+\tau}^{\tau}$ is generated as

$$\begin{aligned}\text{Real variables} &: \quad Y_{t+\tau}^{\tau} = (1200/\tau)\log(RAW_{t+\tau}/RAW_t) \\ \text{Price indexes} &: \quad Y_{t+\tau}^{\tau} = (1200/\tau)\log(RAW_{t+\tau}/RAW_t) - 1200\log(RAW_t/RAW_{t-1}).\end{aligned} \tag{12}$$

For ease of notation, we denote the one-step ahead variable Y_{t+1}^{1} as Y_{t+1}.We consider the following forecasting methods.

5.1.1 Sequential model selection

This method considers the entire set of 145 predictors, together with lags of the dependent variable and performs a sequential selection search on each estimation sample that retains only variables that are statistically significant. The subset of significant variables is then used for forecasting. The initial model specification is

$$Y_{t+\tau}^{\tau} = \alpha + \beta' X_t + \gamma(L)Y_t + \varepsilon_{t+\tau}. \tag{13}$$

where X_t indicates the vector containing the 145 predictors and $\gamma(L)$ is an autoregressive polynomial of order 6. We overcome the problem of multicollinearity in the original X_t matrix by removing the groups of variables whose correlation is greater than .98 and replacing them with an average of all the highly correlated variables. After this procedure, the new matrix X_t^* contains a total of 130 regressors. Our sequential modeling approach begins by estimating the full model and then applies a series of sequential tests until a more parsimonious restriction is found that conveys most of the information contained in the initial model.[5] We apply a simplified version of the search algorithm described by Hoover and Perez (1999, p.175), which consists of reducing 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. The simplification adopted here considers only one reduction path rather than multiple paths and performs only a subset of the

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).

[5] See Hoover and Perez, (1999) and the ensuing discussion for relevant references, a thorough description of the methodology, and an account of the heated debate about the merits and shortcomings of the so-called LSE approach to econometric modeling.

sequential tests in Hoover and Perez (1999). As suggested by these authors, we use a significance level $\alpha = 0.01$ for all the tests, which should encourage parsimony of the final model. A complete description of the particular algorithm that we utilize is contained in Appendix B.

5.1.2 Diffusion indexes

This is a new method proposed by SW. The forecasts are constructed using a two-step procedure. First, the method of principal components is used to estimate the factors F_t from the predictors X_t. Second, the forecasting model is constructed as

$$Y_{t+\tau}^{\tau} = \alpha + \beta'\hat{F}_t + \gamma(L)Y_t + \varepsilon_{t+\tau}, \tag{14}$$

where both the number of factors k retained in $\hat{F}_t$ and the order p of $\gamma(L)$ are selected by BIC, with $1 \leq k \leq 12$ and $0 \leq p \leq 6$.

5.1.3 Bayesian shrinkage estimation

We consider the full model (13) and apply Bayesian estimation of its coefficients using the Litterman (1986) prior. The Litterman prior, when applied to variables expressed in differences, shrinks all coefficients in (13) towards zero, except that for the intercept term a diffuse prior is used. Formally, the variance-covariance matrix V for the prior distribution of $\theta \equiv (\alpha, \beta', \gamma')'$ is diagonal, with $\alpha \sim N(0, 10^8)$, $\beta_i \sim N(0, (w \cdot \lambda \cdot \hat{\sigma}_y/\hat{\sigma}_{x_i})^2)$, $i = 1, ..., k$ and $\gamma_j \sim N(0, (\lambda/j))^2)$, $j = 1, ..., p$. There are two hyperparameters that must be selected *a priori*: λ and w. The parameter λ is the prior standard deviation of the first autoregressive coefficient (that is, the coefficient of Y_t). The prior standard deviation of the subsequent lags of Y_t is further divided by the lag length to reflect an increasing confidence in the prior mean for longer lags. The parameter w is a number between zero and one that reflects the belief that the predictors collected in X_t are less useful for forecasting than lagged values of the dependent variable. Further, the prior standard deviation of β_i is 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 the effects of differences in scale. The Bayesian estimate of θ is then given by

$$\theta^B = (X'X + \hat{\sigma}^2 V^{-1})^{-1}(X'Y^{\tau}), \tag{15}$$

where X is the $m \times 151$ matrix (m is the size of the estimation sample) with rows $(X_t', Y_t, Y_{t-1}, ..., Y_{t-5})$, Y^{τ} is the $m \times 1$ vector with rows $Y_{t+\tau}^{\tau}$ and $\hat{\sigma}$ is the estimated standard error of the residuals in a univariate autoregression for $Y_{t+\tau}^{\tau}$. As suggested by Litterman (1986), we set $w = 0.2$ and $\lambda = 0.2$.[6]

[6] The results were generally robust to a number of different choices for w and λ.

5.1.4 Benchmarks

In addition to the three methods above, we consider two benchmarks. The first is a forecasting method based on an autoregressive (AR) model

$$Y^{\tau}_{t+\tau} = \alpha + \gamma(L)Y_t + \varepsilon_{t+\tau}, \tag{16}$$

where the lag order p of the lag polynomial $\gamma(L)$ is selected by BIC with $0 \leq p \leq 6$. The second benchmark is based on a random walk hypothesis for the levels of the variable; this amounts to specifying the following forecast equation for the variable in differences:

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

5.2 Real-time forecasting experiment

The five methods described above are used to simulate real-time forecasting. The available sample has size $T = 468$, and we choose a maximum estimation window $m = 150 + \tau$, which is the minimal length that allows us to estimate and test the full model in the sequential model selection approach. For comparability, we apply the same transformations to the original series as those documented in Appendix B of SW. The first estimation sample we consider ranges from 1960:1 through 1972:6 + τ (the first 12 data were used as initial observations). The data in this sample are first screened for outliers, which we replace with the unconditional mean of the corresponding variable. We then standardize the regressors, estimate the diffusion indexes and select the autoregressive lag lengths and number of diffusion indexes by BIC. Finally, we run the regressions (13), (14), (16), (17) and apply the Bayesian shrinkage method for t =1960:1,...,1972:6. We use the values of the regressors at time t =1972:6 + τ to generate a set of forecasts for $Y^{\tau}_{1972:6+2\tau}$. We then move the estimation window forward one period and repeat all of the above steps (outlier detection, standardization, specification, estimation and so forth) on data from 1960:2 through 1972:7 + τ. This generates the set of forecasts for $Y^{\tau}_{1972:7+2\tau}$. The final forecasts are produced at t =1998:12 − τ for the variable $Y^{\tau}_{1998:12}$.

5.3 Results of the conditional predictive ability tests

We apply the conditional predictive ability test of Theorem 1 to evaluate the accuracy of the different forecast methods. We take the series of 1-, 6- and 12-month-ahead forecast errors e calculated above for each of the five models and conduct a number of pairwise tests using absolute error and squared error loss functions: $L_1(e) = |e|$ and $L_2(e) = e^2$. For $\tau = 1, 6$ and 12, the null hypotheses of equal conditional predictive ability for the two loss functions are given by

$$\begin{aligned} H_0^1 &: E[|Y_{t+\tau} - \hat{f}_{m,t}| - |Y_{t+\tau} - \hat{g}_{m,t}|\ |\mathcal{F}_t] \equiv E[\Delta L_{1t+\tau}|\mathcal{F}_t] = 0 \text{ and} \\ H_0^2 &: E[(Y_{t+\tau} - \hat{f}_{m,t})^2 - (Y_{t+\tau} - \hat{g}_{m,t})^2|\mathcal{F}_t] \equiv E[\Delta L_{2t+\tau}|\mathcal{F}_t] = 0. \end{aligned} \tag{18}$$

The hypothesis test H_0^i makes use of test function: $h_t = (1,\, \Delta L_{it})'$, $i = 1, 2$. As discussed in section 3.4, in case of rejection of the null hypothesis of equal conditional predictive ability, we consider the sequence of predicted loss differences over the out-of-sample period $\{h_t' \hat{\alpha}_n\}_{t=m}^{T-\tau}$, to establish which method would have been selected at each point in time by the decision rule described in that section. To illustrate, Figure 2 plots the sequence of predicted absolute error loss differences for 1-month ahead forecasts of industrial production, for the two comparisons sequential model selection versus AR and Bayesian shrinkage versus AR.

[FIGURE 2 HERE]

The predicted loss for the sequential method is greater than the predicted loss for the AR for the vast majority of the sample dates, while the predicted loss for the Bayesian shrinkage is always smaller than that of the AR. Further, notice that the predicted loss differences for the pair sequential-AR are several orders of magnitude higher and more volatile than those for the pair Bayesian shrinkage-AR. Provided the test rejects the null hypothesis of equal conditional predictive ability, these considerations lead to the conclusion that Bayesian shrinkage would have been invariably a better method than the AR for forecasting one-month ahead industrial production over the years 1972-1998.

The results of the test for all pairwise comparisons, loss functions, and forecast horizons are contained in Tables 2-5. Tables 2 and 3 present the results for the real variables forecasts, whereas Tables 4 and 5 consider the price indexes forecasts. The entries in each table are the p-values of the tests of equal conditional predictive ability of the two methods. The number within parentheses below each entry is the indicator $I_{n,c}$ discussed in section 3.4 (for $c = 0$) which represents the proportion of times the method in the column would have been preferred to the method in the row over the out-of-sample period using the decision rule described in that section. To facilitate interpretation of the tables, we use a plus sign to indicate rejection of the null hypothesis of equal conditional predictive ability of the two methods at the 10% level and to signal that the method in the column would have been chosen more often than the method in the row (as suggested by an entry $I_{n,c} > .5$). Similarly, a minus sign denotes rejection of the null hypothesis at the 10% level and it indicates that the method in the column would have been chosen more often than the method in the row (i.e., $I_{n,c} < .5$).

[TABLES 2 - 5 HERE]

A sharp result that emerges from Tables 2-5 is that the sequential model selection method is characterized by the worst performance across all forecast horizons, especially for the real variables. In the majority of cases, it is outperformed by every other method, including the naive random walk forecast. The likely explanation for this is the tendency of the method to select over-parameterized models (cases with 40 or more predictors in the final model were not uncommon), in spite of the use of a small confidence level for the sequential tests. Further, performing a new sequential search on each of the rolling estimation windows means that we typically select a different model at each iteration, in spite of the fact that consecutive windows only differ by two observations. This suggests that improvements on the performance of the sequential method may be obtained by updating the model less frequently than every month.

A second general observation is that the information contained in the predictors seems to be less useful for forecasting price indexes than real variables. For the price indexes, there are only a few cases where the AR benchmark is outperformed (by the diffusion index method). In the majority of cases, the parameter-reduction methods, while outperforming the naive random walk forecasts, are indistinguishable from the AR benchmark. Further, the Bayesian shrinkage method is outperformed by the diffusion indexes and by the AR method mainly at the 6- and 12-month forecast horizons.

The Bayesian shrinkage and the diffusion indexes methods appear to fare better for forecasting real variables. Bayesian shrinkage, in particular, outperforms the AR in 11 of the 12 comparisons, while the diffusion indexes method outperforms the AR in 7 of the 12 comparisons. It is interesting to note that for the majority of variables and forecast horizons the AR forecasts are not distinguishable from the random walk forecasts. This suggests that the predictors do contain useful information for forecasting real variables beyond what can be captured by the variable's own lags. Bayesian shrinkage emerges in this case as the best method for reducing the estimation uncertainty of the system, while still conveying its information content.

6 Conclusion

We propose a general framework for out-of-sample predictive ability testing which, as we argue, represents a more realistic setting for economic forecasting than the existing framework, exemplified by West (1996). We start from the premise that the forecaster not only cares about whether two competing forecasts do equally well on average, but also whether she can predict which forecast will do better tomorrow. We implement this different focus by conducting inference about conditional, rather than unconditional moments of forecasts and forecast errors. Recognizing that even a good model may produce bad forecasts due to estimation uncertainty or model instability, we make the object of evaluation the entire forecasting method (including the model, the estimation procedure and the size of the estimation window), whereas the existing literature concentrates solely on the model. In so doing, we are also able to handle more general data assumptions (heterogeneity rather than stationarity) and estimation methods, as well as providing a unified framework for comparing forecasts based on nested or non-nested models, which was not previously available.

One useful application of the conditional predictive ability tests is in evaluating different methods for model selection and parameter estimation. We considered in particular the case of macroeconomic forecasting with a large number of predictors and compared the forecast performance of different parameter-reduction 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 generated 1-, 6- and 12-month ahead forecasts of four measures of real activity and four price indexes using the different forecast methods. The conditional predictive ability tests led to the conclusion that the sequential model selection was the worst performing method, probably due to its tendency to select large models. A second general result was that the information contained in the predictors appeared to be less useful for forecasting price indexes than real variables. For the price indexes, the performance of the various methods was mostly indistinguishable from the one of a simple autoregression. For the real variables, instead, we found that the predictors did contain useful information beyond what is contained in the variable’s own lags. For these variables Bayesian shrinkage seemed to be the best method for reducing the estimation uncertainty of the system. We emphasize that the results of the empirical application are specific to the situation where the number of parameters is very large relative to the sample size and thus one should be careful in generalizing our conclusions to other situations. The fact that shrinkage estimation methods work best in such an environment should come as no surprise. Likewise, it could be argued that the sequential model selection approach was originally conceived for the case where there are enough observations per parameter to make the results of the sequential tests credible. Viewed in this light, our experiments may be unduly hard on the sequential methodology.

Much work remains to be done. A refinement that we are currently exploring is to consider a richer set of decision rules for selecting the best forecasting method or for optimally combining the information embedded in each method once the null hypothesis of equal conditional predictive ability is rejected. A further natural generalization of the tests proposed in the paper is to consider multiple comparisons, for example by adapting the approach of White (2000) to our 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 the case of heterogeneous data.

7 Appendix A. Proofs

Proof of Theorem 1. Under the null hypothesis H_0 in (3), $\{Z_{m,t+1}, \mathcal{F}_t\}$ is an *mds*, and we can apply an *mds* central limit theorem (CLT) to show that

$$\hat{\Omega}_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} \xrightarrow{d} N(0, I) \tag{19}$$

as $n \to \infty$, from which it follows that $T^h_{n,m} \xrightarrow{d} \chi^2_q$ as $n \to \infty$. The *mds* CLT we use requires conditions such that the sample variance $\hat{\Omega}_n$ is a consistent estimator of $\Omega_n = var(\sqrt{n}\bar{Z}_{m,n})$, i.e., such that $\hat{\Omega}_n - \Omega_n \xrightarrow{p} 0$. 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 from (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 (1975)'s 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 $\bar{\sigma}^2_n = var(\lambda'\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n}) = \lambda'\Omega_n^{-1/2}var(\sqrt{n}\bar{Z}_{m,n})\Omega_n^{-1/2}\lambda = 1$ for all n sufficiently large. We have that

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

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 (19) 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 (1975)'s 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 that

$$P[\bar{Z}'_{m,n}\bar{Z}_{m,n} > \varepsilon] \geq P[\bar{Z}'_{m,n}\bar{Z}_{m,n} - E[\bar{Z}'_{m,n}]E[\bar{Z}_{m,n}] > -\varepsilon] \geq P[|\bar{Z}'_{m,n}\bar{Z}_{m,n} - E[\bar{Z}'_{m,n}]E[\bar{Z}_{m,n}]| < \varepsilon] \to 1. \tag{20}$$

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 (1975)'s 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 $c \in \mathbb{R}$, $P[T^h_{n,m} > c] \to 1$ as $n \to \infty$. ■

Proof of Theorem 5. (a) Under the null hypothesis H_0 in (3), we show that

$$\tilde{\Omega}_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} \xrightarrow{d} N(0,I) \tag{21}$$

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

$E[Z_{m,t+\tau}|\mathcal{F}_t]=0$. The asymptotic variance Ω_n is finite by (ii) and it is uniformly positive definite by (iii), Write $\lambda'\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} = n^{-1/2}\sum_{t=m}^{T-\tau}\lambda'\Omega_n^{-1/2}Z_{m,t+\tau}$ and consider the scalar sequence $\{\lambda'\Omega_n^{-1/2}Z_{m,t+\tau}\}$. We verify that the sequence satisfies the conditions of the Wooldridge and White (1988) CLT for mixing processes. For each t, $\lambda'\Omega_n^{-1/2}Z_{m,t+\tau} = f(h_t,\, W_{t+\tau}, ..., W_{t-m})$, where $f(\cdot)$ is a measurable function. Since $\{\ W_t\}$ and $\{h_t\}$ are mixing from (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 $\{\lambda'\Omega_n^{-1/2}Z_{m,t+\tau}\}$ is also 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,

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

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 then conclude that $\Omega_n^{-1/2}\sqrt{n}\bar{Z}_{m,n} \xrightarrow{d} N(0,I)$. It remains to show that $\tilde{\Omega}_n - \Omega_n \xrightarrow{p} 0$, from which (21) follows. We have that

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

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 (1975)'s 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 $w_{n,0} \equiv 1$), which in turn implies that $\tilde{\Omega}_n - \Omega_n \xrightarrow{p} 0$ and the proof is complete.

(b) Given Assumption (i), it follows from Lemma 2.1 of White and Domowitz (1984) that $\{Z_{m,t+\tau}\}$ 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 $Z_{m,t+\tau}$ is bounded uniformly in t by (ii). McLeish (1975)'s 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 $H_{A,h,\tau}$ 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 that

$$P[\bar{Z}'_{m,n}\bar{Z}_{m,n} > \varepsilon] \ge P[\bar{Z}'_{m,n}\bar{Z}_{m,n} - E[\bar{Z}'_{m,n}]E[\bar{Z}_{m,n}] > -\varepsilon] \ge P[|\bar{Z}'_{m,n}\bar{Z}_{m,n} - E[\bar{Z}'_{m,n}]E[\bar{Z}_{m,n}]| < \varepsilon] \to 1. \tag{22}$$

By arguments identical to those used in part (a) - which for this particular result did not necessitate 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 $c \in \mathbb{R}$, $P[T^h_{n,m,\tau} > c] \to 1$ as $n \to \infty$. ■

Proof of Proposition 6. We have $E[(Y_{t+1} - \hat{f}_{m,t})^2 - (Y_{t+1} - \hat{g}_{m,t})^2|\mathcal{F}_t] = E[-2Y_{t+1}(\hat{f}_{m,t} - \hat{g}_{m,t}) + \hat{f}^2_{m,t} - \hat{g}^2_{m,t}|\mathcal{F}_t] = -2(\hat{f}_{m,t} - \hat{g}_{m,t})E[Y_{t+1}|\mathcal{F}_t] + \hat{f}^2_{m,t} - \hat{g}^2_{m,t} = (\hat{f}_{m,t} - \hat{g}_{m,t})(-2E[Y_{t+1}|\mathcal{F}_t] + \hat{f}_{m,t} + \hat{g}_{m,t})$ which is zero (a.s.) if and only if either one of the two factors is zero (a.s.). ■

8 Appendix B. Sequential model selection algorithm

The following is our modification of the search algorithm described by Hoover and Perez (1999, pp.175-176). All the tests are conducted for a significance level $\alpha = 0.01$ and use heteroskedasticity and autocorrelation consistent standard errors and covariance matrices (e.g., Newey and West, 1987, Andrews, 1991).

1. Estimate the full model on the available sample and run the following tests:

   a. Autocorrelation of residuals up to sixth order (LM test; see Breusch and Pagan, 1980).

   b. Stability test (Chow predictive test leaving out the last 10 observations in the sample; see Fisher, 1970).

   If the full model fails any one of the tests (i.e., autocorrelation and/or structural breaks are detected), do not use this test in the following steps.

2. Eliminate the variable of the general specification that has the lowest $t-$statistic and re-estimate the model, which becomes the current model.

3. On each current model, perform the two tests in step 1 together with

   c. F-test of the hypothesis that the coefficients of the variables in the full model that are not included in the current model are jointly insignificant. If the hypothesis cannot be rejected, the current model can be considered a valid restriction of the full model (in this case we say that the model passes the test).

4. If the current model passes all three tests, eliminate the variable with the next lowest $t-$statistic and perform the tests on the new current model. If this model fails any one of the tests, restore the last variable eliminated and remove the variable with the next lowest $t-$statistic. Continue in this fashion until the current model passes all the tests and either all the variables are significant or the elimination of any residual insignificant variable would lead to failing one of the tests.

5. Estimate the final model from step 4.

5.1. If all remaining variables are significant terminate the algorithm.

5.2. If there are remaining insignificant variables, remove all of them and perform the three tests on the restricted model.

a. If the restricted model passes all the tests and all the variables are significant, terminate the algorithm.

b. If the restricted model fails any of the tests, restore the block of eliminated insignificant variables and terminate the algorithm.

c. If the restricted model passes all the tests but there are some remaining insignificant variables, go back to step 5.2.

References

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

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

[3] Benjamini, Y., Hochberg, Y. (1995): "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing", *Journal of the Royal Statistical Society Series B,* 57, 289-300.

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

[5] Breusch, T., Pagan, A. (1980): "The LM Test and its Applications to Model Specification in Econometrics", *Review of Economic Studies*, 47, 239-254.

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

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

[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] Ericsson, N. R. (2002): "Predictable Uncertainty in Economic Forecasting", in Clements, M. P. and Hendry, D. F. (eds) *A Companion to Economic Forecasting,* Blackwell Publishers, Oxford.

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

[15] Fisher, F. (1970): "Tests of Equality between Sets of Coefficients in Two Linear Regressions: an Expository Note", *Econometrica,* 28, 361-366.

[16] Giacomini, R., Komunjer, I. (2002): "Evaluation and Combination of Conditional Quantile Forecasts", UCSD discussion paper 2002-11.

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

[18] 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.

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

[20] Granger, C. W. J. (1999): *Empirical Modeling in Economics: Specification and Evaluation,* Cambridge University Press, New York.

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

[22] 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.

[23] Inoue, A., Kilian, L. (2002): "In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?", manuscript.

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

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

[26] Lopez, J. A. (2001): "Evaluation of Predictive Accuracy of Volatility Models", *Journal of Forecasting,* 20, 87-109

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

[28] McCulloch, R., Rossi, P. E. (1990): "Posterior, Predictive and Utility-Based Approaches to Testing the Arbitrage Pricing Theory", *Journal of Financial Economics,* 28, 7-38.

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

[30] Mizon, G. E. (1995): "Progressive Modelling of Macroeconomic Time Series: The LSE Methodology", in K. D. Hoover (eds.), *Macroeconometrics: Developments, Tensions and Prospects,* 107-170. Kluwer, Boston.

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

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

[33] Sims, C., Zha, T.: "Bayesian Methods for Dynamic Multivariate Models", *International Economic Review*, 39, 949-968.

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

[35] Stock, J. H. (1999): "Forecasting Economic Time Series", forthcoming in B. Baltagi (eds.), *A Companion to Theoretical Econometrics,* Blackwell Publishers, Oxford.

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

[37] 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.

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

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

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

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

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

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

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

Table 1. Empirical size of nominal .1 tests

	m				
σ^2	36	60	120	240	360
.1	.098	.097	.097	.088	.093
1	.089	.093	.094	.093	.095
3	.098	.099	.096	.089	.096

Notes: The table reports the empirical size of the test of equal conditional predictive ability for a nominal size .1 for the Monte Carlo experiment described in Section 4.1. Entries represent the rejection frequencies over 10,000 Monte Carlo replications of the null hypothesis $H_0 : E[(Y_{t+1} - \hat{f}_{m,t})^2 - (Y_{t+1} - \hat{g}_{m,t})^2|\mathcal{F}_t] = 0$, where the forecasts $\hat{f}_{m,t}, \hat{g}_{m,t}$ and the DGP are defined in Section 4.1. Each cell corresponds to a pair of estimation window length m and variance of the DGP disturbances σ^2.

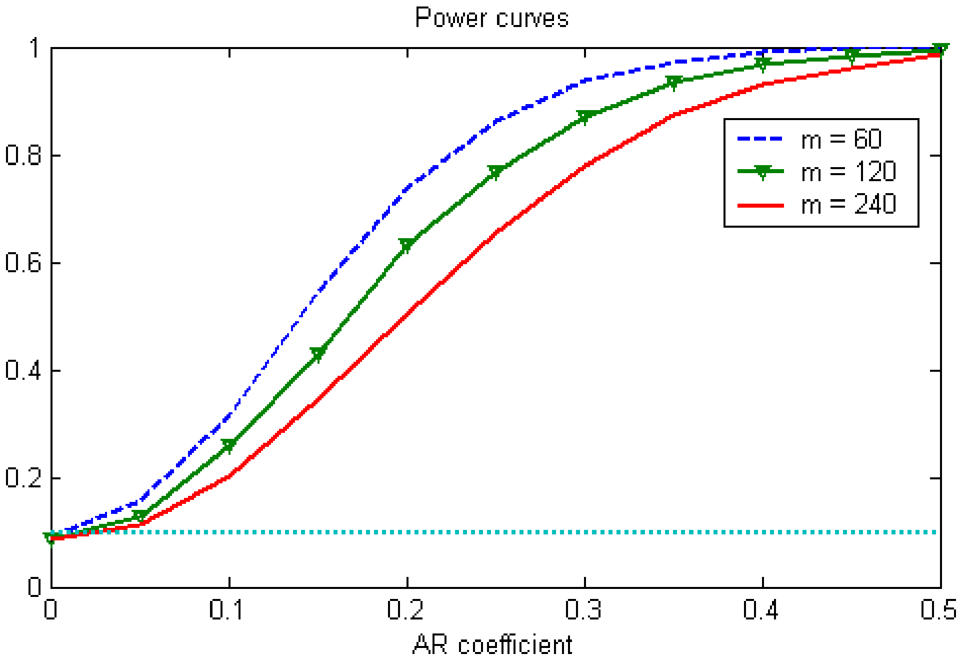

Figure 1: Power curves of the one-step CPA test in the Monte Carlo experiment discussed in Section 4.2. Each curve represents the rejection frequencies over 10,000 Monte Carlo replications of the null hypothesis $H_0 : E[\Delta L_{m,t+1}|\mathcal{F}_t] \equiv E[(Y_{t+1} - \hat{f}_{m,t})^2 - (Y_{t+1} - \hat{g}_{m,t})^2|\mathcal{F}_t] = 0$, where the forecasts $\hat{f}_{m,t}, \hat{g}_{m,t}$ and the DGP are defined in Section 4.2. The DGP is such that $E[\Delta L_{m,t+1}|\mathcal{F}_t] = \rho\Delta L_{m,t}$. The horizontal axis represents the AR coefficient ρ. The total sample size is $T = 368$.

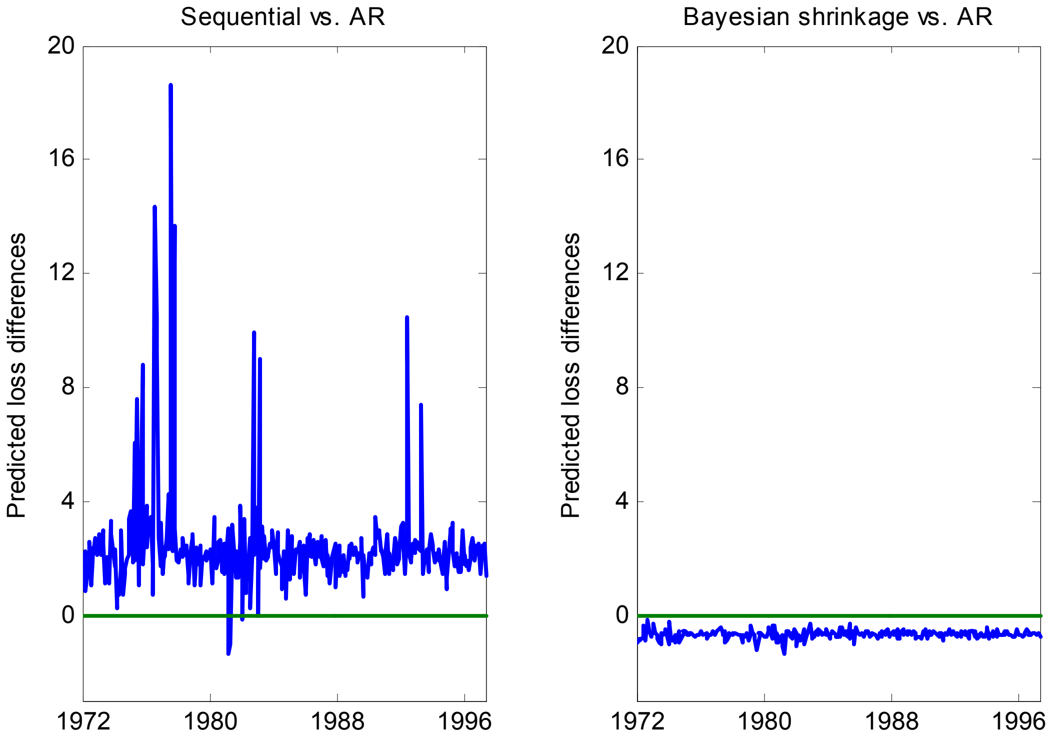

Figure 2: The figure shows the sequences of predicted loss differences over the out-of-sample period $\{h_t' \hat{\alpha}_n\}_{t=1972:8}^{1998:11}$ for 1-month ahead forecasts of industrial production, as described in section 5.3. The left panel shows the predicted loss differences for the sequential model selection versus the AR method and the right panel represents the loss differences for Bayesian shrinkage versus AR. The decision rule of section 3.4 specifies selecting the benchmark model when the loss difference is positive and the competitor model otherwise.

Table 2. Relative performance for absolute error loss function. 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	0.001				0.001				0.013				0.001			
	(0.00)$^-$				(0.00)$^-$				(0.00)$^-$				(0.00)$^-$			
Bayes	0.000	0.109			0.001	0.633			0.007	0.129			0.000	0.090$^-$		
	(0.00)$^-$	(0.01)			(0.00)$^-$	(0.59)			(0.00)$^-$	(0.00)			(0.00)$^-$	(0.00)		
AR	0.006	0.051	0.001		0.035	0.011	0.173		0.020	0.244	0.002		0.004	0.969	0.081	
	(0.01)$^-$	(1.00)$^+$	(1.00)$^+$		(0.06)$^-$	(0.97)$^+$	(1.00)		(0.00)$^-$	(0.89)	(0.99)$^+$		(0.00)$^-$	(0.39)	(0.75)$^+$	
RW	0.021	0.013	0.001	0.020	0.136	0.020	0.016	0.029	0.021	0.260	0.001	0.945	0.036	0.000	0.000	0.000
	(0.02)$^-$	(0.80)$^+$	(0.83)$^+$	(0.69)$^+$	(0.15)	(0.82)$^+$	(0.79)$^+$	(0.68)$^+$	(0.00)$^-$	(0.97)	(0.97)$^+$	(0.87)	(0.40)$^-$	(0.76)$^+$	(0.80)$^+$	(0.92)$^+$
									B. Horizon = 6 months							
Diff Ind	0.016				0.000				0.000				0.000			
	(0.10)$^-$				(0.00)$^-$				(0.00)$^-$				(0.01)$^-$			
Bayes	0.000	0.024			0.000	0.082			0.000	0.381			0.000	0.073		
	(0.03)$^-$	(0.00)$^-$			(0.00)$^-$	(0.00)$^-$			(0.00)$^-$	(0.51)			(0.00)$^-$	(0.08)$^-$		
AR	0.695	0.346	0.013		0.003	0.181	0.049		0.007	0.007	0.004		0.020	0.194	0.083	
	(0.00)	(0.94)	(0.98)$^+$		(0.01)$^-$	(0.82)	(0.99)$^+$		(0.00)$^-$	(1.00)$^+$	(1.00)$^+$		(0.00)$^-$	(0.84)	(0.92)$^+$	
RW	0.939	0.266	0.016	0.173	0.006	0.460	0.110	0.932	0.004	0.018	0.011	0.414	0.875	0.019	0.001	0.000
	(0.12)	(1.00)	(1.00)$^+$	(0.99)	(0.01)$^-$	(1.00)	(1.00)	(0.73)	(0.00)$^-$	(1.00)$^+$	(1.00)$^+$	(0.00)	(0.58)	(0.99)$^+$	(0.99)$^+$	(1.00)$^+$
									C. Horizon = 12 months							
Diff Ind	0.000				0.003				0.000				0.000			
	(0.00)$^-$				(0.00)$^-$				(0.00)$^-$				(0.00)$^-$			
Bayes	0.000	0.748			0.001	0.085			0.000	0.236			0.000	0.217		
	(0.00)$^-$	(0.00)			(0.00)$^-$	(0.00)$^-$			(0.00)$^-$	(0.18)			(0.00)$^-$	(0.04)		
AR	0.062	0.112	0.042		0.056	0.384	0.195		0.006	0.042	0.003		0.117	0.384	0.138	
	(0.04)$^-$	(1.00)	(1.00)$^+$		(0.00)$^-$	(0.87)	(1.00)		(0.07)$^-$	(0.98)$^+$	(1.00)$^+$		(0.01)	(0.95)	(1.00)	
RW	0.062	0.069	0.022	0.699	0.084	0.462	0.292	0.483	0.006	0.042	0.003	1.000	0.706	0.056	0.006	0.002
	(0.07)$^-$	(1.00)$^+$	(1.00)$^+$	(0.87)	(0.00)$^-$	(0.90)	(1.00)	(0.52)	(0.07)$^-$	(0.98)$^+$	(1.00)$^+$	(0.00)	(0.22)	(1.00)$^+$	(1.00)$^+$	(1.00)$^+$

This table contains the results of pairwise tests of equal conditional predictive accuracy of the forecast methods described in section 5.1 for the real variables and for an absolute error loss function. The entries in the table are the p-values of the tests of equal conditional predictive ability for the methods in the corresponding row and column.

The numbers within parentheses are the proportion of times the forecast method in the column outperformed the method in the row over the out-of-sample period, for forecast horizons of 1, 6 and 12 months, according to the decision rule described in Section 3.4. The test function is $h_t = (1, |e_{ft}| - |e_{gt}|)$.

An entry greater than .5 indicates that the method in the column outperformed the method in the row most of the time, and a plus sign represents significance at the 10% level.

An entry less than .5 means that the method in the column was outperformed by the method in the row and a minus sign indicates significance at the 10% level.

For example, for industrial production at the 1-month ahead horizon, the Bayesian shrinkage forecasts outperformed the AR forecasts 100% of the time and a p-value of 0.001 indicates that this superior performance was significant at typical levels of confidence.

Table 3. Relative performance for squared error loss function. 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	0.026				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)$^+$

This table contains the results of pairwise tests of equal conditional predictive accuracy of the forecast methods described in section 5.1 for the real variables and for a squared error loss function. The entries in the table are the p-values of the tests of equal conditional predictive ability for the methods in the corresponding row and column.

The numbers within parentheses are the proportion of times the forecast method in the column outperformed the method in the row over the out-of-sample period, for forecast horizons of 1, 6 and 12 months, according to the decision rule described in Section 3.4. The test function is $h_t = (1,\, e^2_{ft} - e^2_{gt})$.

An entry greater than .5 indicates that the method in the column outperformed the method in the row most of the time, and a plus sign represents significance at the 10% level.

An entry less than .5 means that the method in the column was outperformed by the method in the row and a minus sign indicates significance at the 10% level.

For example, for industrial production at the 1-month ahead horizon, the Bayesian shrinkage forecasts outperformed the AR forecasts 93% of the time and a p-value of 0.001 indicates that this superior performance was significant at typical levels of confidence.

Table 4. Relative performance for absolute error loss function. 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	0.007				0.042				0.007				0.000			
	(0.00)$^-$				(0.00)$^-$				(0.01)$^-$				(0.00)$^-$			
Bayes	0.024	0.292			0.079	0.092			0.018	0.437			0.004	0.097		
	(0.01)$^-$	(0.86)			(0.00)$^-$	(0.97)			(0.00)$^-$	(0.92)			(0.03)$^-$	(0.99)		
AR	0.011	0.493	0.698		0.046	0.801	0.163		0.004	0.979	0.407		0.000	0.038	0.021	
	(0.00)$^-$	(0.74)	(0.03)		(0.00)$^-$	(0.92)	(0.03)		(0.00)$^-$	(0.22)	(0.00)		(0.01)$^-$	(0.32)	(0.00)	
RW	0.047	0.344	0.671	0.384	0.144	0.123	0.752	0.148	0.094	0.080	0.218	0.057	0.001	0.168	0.432	0.066
	(0.00)$^-$	(0.95)	(0.82)	(0.79)	(0.00)	(1.00)	(1.00)	(1.00)	(0.06)$^-$	(1.00)	(0.96)	(1.00)	(0.01)$^-$	(0.78)	(0.26)	(0.85)
									B. Horizon = 6 months							
Diff Ind	0.000				0.000				0.000				0.000			
	(0.00)$^-$				(0.00)$^-$				(0.01)$^-$				(0.00)$^-$			
Bayes	0.406	0.000			0.015	0.009			0.039	0.004			0.353	0.000		
	(0.05)	(1.00)$^+$			(0.00)$^-$	(1.00)$^+$			(0.05)$^-$	(1.00)$^+$			(0.64)	(1.00)$^+$		
AR	0.006	0.135	0.000		0.000	0.275	0.016		0.000	0.396	0.003		0.000	0.718	0.000	
	(0.00)$^-$	(0.64)	(0.06)$^-$		(0.00)$^-$	(0.78)	(0.08)$^-$		(0.00)$^-$	(0.34)	(0.08)$^-$		(0.00)$^-$	(0.18)	(0.00)$^-$	
RW	0.800	0.000	0.008	0.000	0.205	0.000	0.002	0.000	0.895	0.000	0.000	0.000	0.809	0.000	0.648	0.000
	(0.91)	(1.00)$^+$	(0.96)$^+$	(0.97)$^+$	(0.00)	(1.00)$^+$	(1.00)$^+$	(1.00)$^+$	(0.82)	(1.00)$^+$	(1.00)$^+$	(1.00)$^+$	(1.00)	(1.00)$^+$	(0.99)	(1.00)$^+$
									C. Horizon = 12 months							
Diff Ind	0.000				0.000				0.000				0.000			
	(0.00)$^-$				(0.01)$^-$				(0.00)$^-$				(0.00)$^-$			
Bayes	0.055	0.000			0.007	0.000			0.001	0.000			0.237	0.000		
	(0.00)$^-$	(0.98)$^+$			(0.00)$^-$	(0.98)$^+$			(0.00)$^-$	(1.00)$^+$			(0.00)	(1.00)$^+$		
AR	0.004	0.292	0.552		0.000	0.826	0.004		0.000	0.330	0.041		0.000	0.200	0.000	
	(0.00)$^-$	(0.99)	(0.00)		(0.00)$^-$	(0.94)	(0.13)$^-$		(0.00)$^-$	(0.93)	(0.00)$^-$		(0.00)$^-$	(0.55)	(0.00)$^-$	
RW	0.142	0.000	0.070	0.000	0.033	0.000	0.036	0.000	0.035	0.000	0.002	0.000	0.235	0.000	0.253	0.000
	(0.00)	(0.98)$^+$	(1.00)$^+$	(1.00)$^+$	(0.00)$^-$	(1.00)$^+$	(0.98)$^+$	(0.98)$^+$	(0.00)$^-$	(1.00)$^+$	(1.00)$^+$	(1.00)$^+$	(0.11)	(1.00)$^+$	(0.92)	(1.00)$^+$

This table contains the results of pairwise tests of equal conditional predictive accuracy of the forecast methods described in section 5.1 for the price indexes and for an absolute error loss function. The entries in the table are the p-values of the tests of equal conditional predictive ability for the methods in the corresponding row and column.

The numbers within parentheses are the proportion of times the forecast method in the column outperformed the method in the row over the out-of-sample period, for forecast horizons of 1, 6 and 12 months, according to the decision rule described in Section 3.4. The test function is $h_t = (1, |e_{ft}| - |e_{gt}|)$.

An entry greater than .5 indicates that the method in the column outperformed the method in the row most of the time, and a plus sign represents significance at the 10% level.

An entry less than .5 means that the method in the column was outperformed by the method in the row and a minus sign indicates significance at the 10% level.

For example, for the Consumer price index at the 1-month ahead horizon, the sequential forecasts outperformed the AR forecasts 0% of the time and a p-value of 0.011 indicates that this superior performance was significant at typical levels of confidence.

Table 5. Relative performance for squared error loss function. 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)$^+$

This table contains the results of pairwise tests of equal conditional predictive accuracy of the forecast methods described in section 5.1 for the price indexes and for a squared error loss function. The entries in the table are the p-values of the tests of equal conditional predictive ability for the methods in the corresponding row and column.

The numbers within parentheses are the proportion of times the forecast method in the column outperformed the method in the row over the out-of-sample period, for forecast horizons of 1, 6 and 12 months, according to the decision rule described in Section 3.4. The test function is $h_t = (1,\, e^2_{ft} - e^2_{gt})$.

An entry greater than .5 indicates that the method in the column outperformed the method in the row most of the time, and a plus sign represents significance at the 10% level.

An entry less than .5 means that the method in the column was outperformed by the method in the row and a minus sign indicates significance at the 10% level.

For example, for the Producer price index at the 1-month ahead horizon, the sequential forecasts outperformed the diffusion index forecasts 2% of the time and a p-value of 0.000 indicates that this superior performance was significant at typical levels of confidence.