Tests of Forecasting Performance and Choice of Estimation Window

Allan Timmermann
Rady School, UCSD

Yinchu Zhu
Rady School, UCSD

July 5, 2016

Abstract

The test of equal forecasting performance for pairs of economic forecasting models proposed by Giacomini and White (2006) has found widespread use in empirical work. However, the test assumes that the parameters of the underlying forecasting models are estimated using a rolling window of fixed width or similar estimation schemes for which the parameter estimates do not converge to their probability limits in large samples. This is an important limitation that, in practice, can significantly reduce the power of the test. This paper uses a martingale central limit theorem to generalize the setup of Giacomini and White to allow for a recursively expanding estimation window. We present evidence from Monte Carlo simulations and empirical applications demonstrating the benefits from conducting tests of equal predictive accuracy with an expanding estimation window.

1 Introduction

Economic forecasts feature prominently in governments' decisions on fiscal policy, central banks' monetary policy, households' consumption and investment decisions and companies' hiring and capital expenditure choices and so it is important to be able to tell a good forecast from a bad one. The importance of the ability to differentiate between forecasts of varying quality is highlighted by the wealth of possible forecast specifications typically on offer to economic forecasters.

Reflecting this need, recent years have seen considerable developments in techniques for comparing the predictive accuracy of competing forecasting models. Diebold and Mariano (1995) develop simple and attractive methods for comparing the performance of pairs of forecasting models under general loss under the assumption that the sequence of loss differentials is covariance stationary.[1] West (1996) develops an approach for comparing the predictive accuracy of non-nested models which applies to situations that can handle non-stationarities in the sequence of loss differentials introduced by recursive updating in parameter estimates. Clark and McCracken (2001), Clark and

[1] For further discussion of this approach, see Diebold (2015).

McCracken (2005) and McCracken (2007) extend the approach of West to allow for nested models.[2] These papers test the null of equal predictive accuracy evaluated at the probability limits of the parameter estimates and show that recursive parameter estimation leads to non-standard distributions for the test statistic of equal predictive performance of nested regression models.[3] Moreover, the distribution of the test statistic depends on "nuisance" parameters such as the fraction of the sample used to estimate the parameters versus the length of the forecast evaluation sample, the number of additional parameters in the big forecasting model, and the estimation scheme (fixed, rolling, or expanding).

The results established in these papers are limited to a relatively narrow set of estimation methods and model specifications (linear regressions) and so do not facilitate comparisons of forecasts generated by other methods in common use such as DSGE models, many Bayesian forecasts, certain nonlinear models, or nonparametric forecasts. In an important paper, Giacomini and White (2006) (GW, henceforth) address these shortcomings by proposing tests of equal predictive accuracy evaluated at the current parameter estimates (as opposed to at their probability limits.) To avoid a degenerate limiting distribution for the test statistic, GW assume that estimation error does not vanish asymptotically by requiring that fixed-width windows are used to estimate the parameters of the forecasting models. Under this assumption, GW show that test statistics for equal predictive accuracy follow a standard chi-squared distribution. The assumption of a rolling estimation window does not come for free, however. Out-of-sample tests of forecasting performance are well-known to have weak power (see, e.g., Inoue and Kilian (2005) and Hansen and Timmermann (2015*b*)) and limiting the estimation window to a restricted range can magnify the effect of parameter estimation error, further weakening the power of such tests and rendering it even more difficult to distinguish between the performance of competing forecasting models. Moreover, as pointed out by Giacomini and White (2006), a hypothesis test based on the predictive accuracy of a model whose parameters are estimated on a rolling window is different from testing the same model's predictive accuracy when its parameters are estimated with an expanding estimation window.[4]

These limitations of existing forecast evaluation methods mean that there are, at present, no methods for comparing the performance of forecasts generated by many commonly used estimation methods that use expanding estimation windows. This paper fills this gap by generalizing the test of equal predictive accuracy proposed by GW to allow for a recursively expanding estimation window. The basic idea of our approach is to apply a martingale central limit theorem (CLT) to a scaled version of the sequence of loss differentials generated by the competing prediction models and used to compute the test statistic.

[2]This extension is particularly important because nested model comparisons arise frequently in economics and finance. Examples include the random walk model for stock prices and the uncovered interest rate parity model in the exchange rate literature. See Pettenuzzo et al. (2014) for tests of predictability of the equity risk premium and Rossi (2013) for an extensive discussion of exchange rate forecasting models.

[3]Two forecasting models are nested if one ("big") model contains all terms in the other ("small") model, plus one or more additional terms. Nested forecasting models pose a particular challenge in this setting because the forecasting performance of the "big" and "small" models are identical under the null of equal predictive accuracy, thus leading to a nonstandard limiting distribution for test statistics based on the average loss differential scaled by its standard error.

[4]The testing framework proposed by Giacomini and White is, thus, better viewed as testing the performance of different forecasting methods (models along with choice of estimation approach), rather than testing the performance of a specific forecasting model.

Moreover, we show that our generalization of the GW test can make a big difference in practice. Estimation error can seriously impact forecasting performance, particularly if the data set used for parameter estimation is restricted to a rolling window with a limited number of observations. A recursively expanding estimation window makes more efficient use of data in covariance stationary environments and can considerably reduce the adverse effect of parameter estimation error on forecasting performance. We illustrate this through Monte Carlo simulations that contrast the power of the GW test in differentiating between the forecasting performance of two nested models under rolling and expanding estimation windows. We also explore the importance of the choice of estimation window through empirical applications to predictability of stock market returns and inflation. In both the simulations and the empirical applications we find considerable improvements in the forecasting performance of a big model measured relative to a small (nested) model as a result of using an expanding rather than a rolling estimation window.

The outline of the paper is as follows. Section 2 introduces the setup used to compare competing models' forecasting performance while Section 3 provides our theoretical results. Section 4 describes our Monte Carlo simulations and Section 5 reports results from empirical applications. Section 6 concludes.

2 Forecast Environment

Forecast comparisons are often conducted using (pseudo) out-of-sample evaluation methods, i.e., by splitting a sample of T observations into an initial sample of n observations used for initial parameter estimation and model selection and a forecast evaluation sample which consists of the remaining p observations. Hence, $T = p + n + 1$ and we view $n = n_T$ and $p = p_T$ as functions of T, although the subscripts are suppressed for notational simplicity. Out-of-sample forecast evaluations can have substantially weaker power than full-sample tests (see, e.g., Inoue and Kilian (2005) and Hansen and Timmermann (2015*b*)), but have the advantage that they are less prone to data mining biases (Hansen and Timmermann (2015*a*)) and can provide important information about the time-series evolution in a prediction model's performance.

Suppose we are interested in comparing the predictive accuracy of a pair of one-step-ahead forecasts of some variable Y_{t+1}, each generated using information available at time t, denoted $\mathcal{F}_t$.[5] Specifically, forecast comparisons are based on observed sequences of forecasts $\{\hat{y}_{1,t+1,n}(\hat{\beta}_{1,t,n})\}_{t=n+1}^{T}$ and $\{\hat{y}_{2,t+1,n}(\hat{\beta}_{2,t,n})\}_{t=n+1}^{T}$, where the estimated parameters $(\hat{\beta}_{i,t,n}), i = 1, 2$ can be based on different weighting schemes for the data observed up to time t. The precision of the forecasts is evaluated using a loss function, $L(y_{t+1}, \hat{y}_{t+1,n})$ which is a mapping from the space of outcomes and forecasts $\mathcal{Y} \times \mathcal{Y}$ to the real line, $\mathcal{R}$. Under squared error loss $L(y_{t+1}, \hat{y}_{t+1,n}) = e_{t+1}^2$, where $e_{t+1} = y_{t+1} - \hat{y}_{t+1,n}$ is the forecast error; this is by far the most common loss function.

Following Diebold and Mariano (1995), let $\Delta L_{t+1,n}(\hat{\beta}_{1,t,n}, \hat{\beta}_{2,t,n}) = L(y_{t+1}, \hat{y}_{1,t+1,n}(\hat{\beta}_{1,t,n})) - L(y_{t+1}, \hat{y}_{2,t+1,n}(\hat{\beta}_{2,t,n}))$ be the loss differential which measures the forecasting performance of model 1 relative to that of model 2. West (1996) proposes testing the null of equal predictive accuracy when the parameters

[5] For simplicity, we restrict the forecast horizon to a single period, but our results are easily generalized to multi-period horizons.

are evaluated at their probability limits, $\beta_i^* = plim_{n\to\infty}(\hat{\beta}_{in})$, $i = 1, 2$. The resulting null hypothesis of equal predictive accuracy evaluated at the probability limits of the parameters is

$$H_0 : E[\Delta L_{t+1}(\beta_1^*, \beta_2^*)] = 0. \tag{1}$$

For non-nested forecasting models whose parameters can be estimated using least squares or similar methods, West establishes that the scaled average loss differential asymptotically follows a normal distribution whose limiting variance depends on the long-run variance of the loss under known model parameters, a term that reflects parameter estimation error and the covariance between these two terms. This result is theoretically elegant and easy to use but has three limitations. First, nested model comparisons are ruled out. This point is addressed by Clark and McCracken (2001) and McCracken (2007) who show that conventional test statistics used in comparisons of nested models based on the null in (1) have non-standard limiting distributions that depend on the proportion of the sample used for forecast evaluation versus the proportion of the sample set aside for initial parameter estimation, $\pi = p/n$. Moreover, the critical values of the test statistic depend on the estimation scheme and differ for fixed, rolling, or expanding estimation windows.[6] Second, the results are limited to a fairly narrow set of forecasting models and estimators, ruling out a variety of forecasting methods involving model selection that changes over time, many non-linear and non-parametric estimation schemes, and many types of Bayesian estimators. Third, the null in (1) compares model performance at the probability limits of the parameter estimates, β_i^*, and so is most relevant if we are interested in using the test of equal predictive accuracy to conduct inference about which model is correct as opposed to which model generates the best forecasts.

2.1 Giacomini-White Approach

In an influential paper Giacomini and White (2006) address these issues. Central to their analysis is that they re-state the null in (1) as a comparison of the two sequences of forecasts evaluated at the current parameter estimates, $\hat{\beta}_{i,t,n}$,

$$H_0 : E\left[\Delta L_{t+1,n}(\hat{\beta}_{1,t,n}, \hat{\beta}_{2,t,n}) \mid \mathcal{F}_t\right] = 0, \tag{2}$$

Equation (2) expresses the null in terms of parameter estimates and so explicitly accounts for the effect of parameter estimation error on the two sets of forecasts. Unlike the null in (1), the null in (2) does not evaluate the models' relative performance at the limiting value of the model parameters which are unknown in practice.

Provided that parameter estimation error affects the sequence of loss differentials even asymptotically, the GW approach can handle nested as well as non-nested model comparisons. GW ensure that estimation error does not vanish by considering forecasts that use estimators with "limited memory", the main example of which is a fixed-

[6] Since the null hypothesis is the mean of unobserved objects (β^*), complicated computations and analysis are required to characterize the effect of estimation errors on the test statistic which can lead to nonstandard asymptotics.

width rolling estimation window but also includes methods such as discounted least squares. GW assume that the (rolling) estimation window, n, is bounded and consider values of $n \in \{25, 75, 125, 150\}$ in their simulations.

The assumption of a bounded estimation window or an estimation scheme that down-weighs observations in the distant past if n is allowed to grow with the sample size poses a serious limitation because it means that an inconsistent estimator is used for the prediction models. In practice, this can lead to less accurate forecasts and a reduction in the power of tests of equal predictive accuracy. To deal with this limitation, we next show how the GW approach can be generalized to allow for an increasing estimation window. Specifically, we maintain our focus on testing the null in (2), but allow the estimation window n to tend to infinity as the sample size expands.[7]

3 A Generalization of the Giacomini-White Test

Our analysis applies the martingale central limit theorem (CLT) to a scaled version of the loss differential,$\Delta L_{t+1,n}$. To this end, define $h_{t,n} = \Delta L_{t,n} Z_{t-1}$, where $Z_{t-1} \in \mathcal{F}_{t-1}$ is an instrument used to evaluate the null of equal predictive accuracy. If $Z_{t-1} = 1$, a test of equal unconditional forecasting performance is obtained, while time-varying instruments can be used to test the null of equal conditional expected accuracy. The estimation window n can be fixed or tend to infinity at any rate. To allow for some types of non-stationarities in which the data generating process can depend on the sample size, T, in what follows we assume an array structure for the data and let $W_{T,t} = \{Y_{T,t}, X_{T,t}\}$ denote the outcome and predictor variables. Moreover, we use $\|\cdot\|_2$ and $\lambda_{\min}$ to denote the Euclidean norm and the smallest eigenvalue, respectively. We work with strong mixing arrays which allow the DGP of $\{W_{T,t}\}$ to depend on the sample size. This general setup allows situations in which $W_{T,t}$ itself is generated from combinations of stationary strong mixing processes. Following Andrews (1988), we impose weak dependence in the form of strong mixing conditions for arrays:

Definition 1. The array $\{W_{T,t}\}_{t=-\infty}^{\infty}$ is strong mixing with coefficient $\alpha(\cdot)$ if

$$\alpha(t) = \sup_{-\infty<i<\infty,\, T\geq 1} \sup_{A\in\mathcal{F}_{-\infty,i}^{T},\, B\in\mathcal{F}_{i+t,\infty}^{T}} \left|P(A\bigcap B) - P(A)P(B)\right| \to 0 \text{ as } t\to\infty,$$

where $\mathcal{F}_{-\infty,i}^{T} = \sigma(\cdots, W_{T,i-1}, W_{T,i})$ and $\mathcal{F}_{i+t,\infty}^{T} = \sigma(W_{T,i+t}, W_{T,i+t+1}, \cdots)$.

Using this definition, we can state our main result.

Theorem 1. *Let $\{W_{T,t}\}_{t=-\infty}^{\infty}$ be an α-mixing array and $h_{t,n} = H_{t,n}(W_{T,t}, W_{T,t-1}, \cdots, W_{T,t-n+1}) \in \mathbb{R}^k$, where $H_{t,n}$ is a measurable function. Suppose that (i) there exist positive constants δ, Δ and a_T such that $E\|a_T h_{t,n}\|_2^{2+\delta} < \Delta$ and $\liminf \lambda_{\min}\left(p^{-1}\sum_{t=n+1}^{n+p} E(a_T^2 h_{t,n} h_{t,n}')\right) > 0$; (ii) there exists a filtration $\mathcal{F}_t$ such that $E(h_{t,n} \mid \mathcal{F}_{t-1}) = 0$. Then we have*

[7] The use of a rolling estimation window is often motivated by reference to non-stationarities in the underlying data generating process. However, this argument is difficult to sustain because a rolling estimation window has not been demonstrated to be optimal (or even robust) in the presence of changes to the data generating process.

$$\left(\sum_{t=n+1}^{n+p} h_{t,n}\right)' \left(\sum_{t=n+1}^{n+p} h_{t,n}h_{t,n}'\right)^{-1} \left(\sum_{t=n+1}^{n+p} h_{t,n}\right) \xrightarrow{d} \chi_h^2. \tag{3}$$

The assumptions we make for Theorem 1 are comparable to and, in fact, weaker than those in Giacomini and White (2006). Assumption (i) requires a moment condition on the normalized loss differential and also imposes uniform positive-definiteness for the normalization matrix as it requires its eigenvalues to be bounded away from zero. We do not need to make any assumptions on the rate at which the α-mixing coefficients decay.[8] In fact, strong mixing, together with moment conditions, typically suffices to derive a CLT and LLN; see Bradley (2007). Assumption (ii) states that the two models' conditionally expected performance is identical at each point in time, t, and so the loss differential is a martingale difference process.

To apply the result in Theorem 1 we do not need to know the scaling sequence a_T; rather, we only need to be assured of the existence of such a sequence since implementation of the test only requires computing the sums $h_{t,n}$ and $h_{t,n}h_{t,n}'$ in (3).

The key point in the proof of Theorem 1 is that the test statistic in (3) is a self-normalized sum. In other words, $\sum_{t=n+1}^{n+p} h_{t,n}$ is normalized by a quantity that is of the same order of magnitude, $\left(\sum_{t=n+1}^{n+p} h_{t,n}h_{t,n}'\right)^{-1}$. Formally, we apply the martingale CLT and LLN to the normalized quantity $a_T h_{t,n}$ and notice that the normalization scalar a_T cancels out, yielding a test statistic that does not depend on a_T.[9]

Theorem 1 allows the magnitude of $h_{t,n}$ to decay to zero which is essential in many empirical applications. For example, when estimation error vanishes in comparisons of nested models, $h_{t,n}$ tends to zero and Theorem 1 guarantees the asymptotic validity of the test for a large class of models.

Theorem 1 bears an interesting relation to the extensive Monte Carlo simulation results reported in Clark and McCracken (2011). Clark and McCracken find that the size properties of the Giacomini-White test appear to be better under an expanding estimation window than under a rolling estimation window. They write (page 26): "Admittedly, though, other aspects of our Monte Carlo results seem to be at odds with the asymptotic results of Giacomini and White (2006), if not their Monte Carlo results. Their asymptotics imply the MSE-t test has an asymptotic distribution that is standard normal for rolling forecasts but not recursive forecasts, suggesting the test should have better size properties in the rolling case than the recursive. But in our Monte Carlo results, the standard normal approximation for MSE-t seems to work better with recursive forecasts than rolling, yielding 1-step ahead rejection rates closer to nominal in the former case than the latter."

We next illustrate Theorem 1 for some commonly encountered cases.

[8] Note that Giacomini and White (2006) impose restrictions on the rate at which the mixing coefficients decay, whereas we only need them to decay.

[9] As pointed out in Giacomini and White (2006), using $\left(\sum_{t=n+1}^{n+p} h_{t,n}h_{t,n}'\right)^{-1}$ as the weighting matrix might lead to better power properties than using a heteroskedasticity and autocorrelation consistent (HAC) estimator. The reason is that, under the alternative hypothesis, it is often plausible that $h_{t,n}$ is positively serially correlated, resulting in a larger variance under a HAC estimator.

Example 1: Bounded n

Giacomini and White (2006) consider the case with a bounded estimation window, n. In our framework this corresponds to setting $a_T = 1$.

Example 2: Unbounded n

Consider the linear data generating process (DGP)

$$y_{t+1} = x_t'\beta + z_t'\gamma + \varepsilon_{t+1}, \tag{4}$$

and suppose that we are interested in comparing forecasts from a small model: $y_{t+1} = x_t'\tilde{\beta} + \varepsilon_{t+1}$ to forecasts generated by a big model, $y_{t+1} = x_t'\beta + z_t'\gamma + \varepsilon_{t+1}$. Both models are estimated using least squares with a rolling window of length n, and $n \to \infty$. The estimation error for γ in the big model will be of order $O_P(n^{-1/2})$. Therefore, if the magnitude of γ is larger than $O_P(n)$, the large model will tend to outperform the small model. To satisfy the martingale condition of Theorem 1, $E(h_{t,n} \mid \mathcal{F}_{t-1}) = 0$, we assume $\gamma = O(n^{-1/2})$. Denote by $(\hat{\beta}_{t,n}', \hat{\gamma}_{t,n}')'$ and $\tilde{\beta}_{t,n}$ the estimates at time t of the parameters in the big and small models, respectively. Consider squared error loss so that we are interested in comparing the forecast errors from the big and small models, $e_{B,t+1} = y_{t+1} - x_t'\hat{\beta}_{t,n} - z_t'\hat{\gamma}_{t,n}$, $e_{S,t+1} = y_{t+1} - x_t'\tilde{\beta}_{t,n}$, and define $h_{t+1,n} = \Delta L_{t+1,n} = e_{B,t+1}^2 - e_{S,t+1}^2$ and $Q_t = (x_t', z_t')'$. We have the following result:

Lemma 1. *Suppose that*

(i) $nE\left[(\hat{\gamma}_{t,n} - \gamma_p)(\hat{\gamma}_{t,n} - \gamma_p)'\right] \to V \neq 0$.

(ii) $\Sigma_Q = E\left[Q_t Q_t'\right]$ *is positive definite.*

(iii) $\{(x_t, z_t, \varepsilon_{t+1})\}_{t=-\infty}^{\infty}$ *is* α*-mixing and* $E\left[\varepsilon_{t+1} \mid \mathcal{F}_t\right] = 0$*, where* $\mathcal{F}_t$ *is the* $\sigma-$*algebra generated by* $\{(x_s, z_s, \varepsilon_s) \mid s \leq t\}$.

(iv) There exists a constant $c_1 > 0$ *with* $E(\varepsilon_{t+1}^2 \mid \mathcal{F}_t) \geq c_1$ *almost surely.*

Then $\liminf_{T\to\infty} a_T^2 E\Delta L_{t+1,n}^2 > 0$*, where* $a_T = \sqrt{n}$.

This result says that for nested models, once we scale $\Delta L_{t+1,n}$ by n, we can apply Theorem 1. Note that $\Delta L_{t+1,n} = (e_{B,t+1} - e_{S,t+1})(e_{B,t+1} + e_{S,t+1})$ and $e_{B,t+1} - e_{S,t+1} = Q_t'(\tilde{\theta}_{t,n} - \hat{\theta}_{t,n})$, where $\hat{\theta}_{t,n} = (\hat{\beta}_{t,n}', \hat{\gamma}_{t,n}')'$ and $\tilde{\theta}_{t,n} = (\tilde{\beta}_{t,n}', 0)'$. Under very mild conditions, $n(\hat{\theta}_{t,n} - \theta)$, $n(\tilde{\theta}_{t,n} - \theta)$ and $e_{B,t+1} + e_{S,t+1}$ are all $O_P(1)$, where $\theta = (\beta', \gamma')'$. Then we have $n\Delta L_{t+1,n} = O_P(1)$. It is not very restrictive to assume that this $O_P(1)$ term has bounded $2 + \delta$ moments, i.e., that $E\left[|n\Delta L_{t+1,n}|^{2+\delta}\right]$ is bounded for some $\delta > 0$, ensuring that the first part of condition (i) in Theorem 1 holds.

4 Monte Carlo Simulations

This section presents results from Monte Carlo simulations that illustrate the theoretical results from Section 3 and shed light on the finite sample properties of tests of equal predictive accuracy for models whose parameters are estimated with rolling or expanding windows. We first use a setup that is sufficiently simple to allow us to study the size properties of tests of equal predictive accuracy. We next proceed to study the finite sample power of tests of equal forecasting performance for nested prediction models using a first-order autoregressive model.

4.1 Size properties

To study the size properties of the test in (3), we follow Giacomini and White (2006) and consider the following DGP:

$$y_t = c + \beta x_t + \varepsilon_t, \quad \varepsilon_t \sim \mathrm{N}(0,\sigma^2). \tag{5}$$

As in GW, x_t is assumed to follow a process fitted to the second log difference in the monthly U.S. consumer price index from 1947:1 to 2015:4. The big forecasting model is the true DGP $y_t = c + \beta x_t + \varepsilon_t$, while the small model takes the form $y_t = \theta x_t + \varepsilon_t$. The associated forecasts from the small and big models are computed as $\hat{y}_{1,t+1,n} = \hat{\theta}_t x_t$ and $\hat{y}_{2,t+1,n} = \hat{c}_t + \hat{\beta}_t x_t$ where $\hat{c}_t$, $\hat{\beta}_t$ and $\hat{\theta}_t$ are computed using data up to time t,$\{(y_s, x_s)\}_{s=1}^{t}$.

The following result shows how c can be chosen in a way that allows us to control the size in the forecast evaluation tests when both models are estimated using an expanding window scheme.

Lemma 2. *Let* $X_t = (1, x_t)' \in \mathbb{R}^2$, $X_{(t)} = (X_1, \cdots, X_t)' \in \mathbb{R}^{t\times 2}$, $x_{(t)} = (x_1, \cdots, x_t)' \in \mathbb{R}^t$ *and* $1_t = (1, \cdots, 1)' \in \mathbb{R}^t$. *If*

$$c = \left[\sum_{t=n+1}^{n+p}\left(X_t'\left(X_{(t)}'X_{(t)}\right)^{-1}X_t - \frac{x_{t+1}^2}{x_{(t)}'x_{(t)}}\right)\right] \Big/ \left[\sum_{t=n+1}^{n+p}\left(1 - \frac{x_{(t)}'1_t}{x_{(t)}'x_{(t)}}x_{t+1}\right)^2\right], \tag{6}$$

then $E\left[p^{-1}\sum_{t=n+1}^{n+p}(y_t - \hat{y}_{B,t|t-1})^2\right] = E\left[p^{-1}\sum_{t=n+1}^{n+p}(y_t - \hat{y}_{S,t|t-1})^2\right]$.

This result, which is similar to Proposition 5 in Giacomini and White (2006), establishes conditions under which the null hypothesis that the loss differential is a martingale difference sequence holds on average, To illustrate Lemma 2, we simulate the DGP in (5) with $\beta = 1$, $\sigma = 0.1$ and c computed from (6). Next, we compute the test statistic

$$J_T = \frac{p^{-1/2}\sum_{t=n+1}^{n+p}\Delta L_{t+1}}{\sqrt{p^{-1}\sum_{t=n+1}^{n+p}\Delta L_{t+1}^2}}, \tag{7}$$

with $\Delta L_{t+1} = (y_t - \hat{y}_{S,t|t-1})^2 - (y_t - \hat{y}_{B,t|t-1})^2$. Our simulations use 5000 random samples to compute $P(J_T^2 > \chi^2_{0.95,1})$, where $\chi^2_{\alpha,1}$ is the α-quantile of the $\chi^2(1)$ distribution. Results are presented in Table 1.[10] As we can see from these simulation results, the test in (7) has good size properties, although it tends to be mildly undersized when $p = 50$. The rejection probabilities are computed using 5000 random samples.

[10] "NA" entries in Table 1 arise for cases where $p \geq T$.

4.2 Power properties

Having verified that the test in (7) has reasonable size properties, we next study the test's power properties using a linear autoregressive model (Section 4.2.1) and a model with regime switching (Section 4.2.2).

4.2.1 Autoregressive models

Consider the following first-order autoregressive DGP:

$$y_{t+1} = \phi y_t + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim iid\ N(0,1).$$

Our simulations assume that the small model forecasts zero, i.e., $\hat{y}_{1,t+1,n} = 0$, while the forecasts from the large model are based on least squares estimates of the AR(1) model $\hat{y}_{2,t+1,n} = \hat{\phi}_t y_t$. The null of equal predictive accuracy takes the form

$$H_0: \ E(\Delta L_{t+1} \mid \mathcal{F}_t) = 0,$$

where $\Delta L_{t+1} = y_{t+1}^2 - (y_{t+1} - \hat{\phi}_t y_t)^2$. For any instrument $Z_t \in \mathcal{F}_t$, we can consider the test statistic

$$J_T = \frac{p^{-1/2} \sum_{t=n+1}^{n+p} \Delta L_{t+1} Z_t}{\sqrt{p^{-1} \sum_{t=n+1}^{n+p} (\Delta L_{t+1} Z_t)^2}}. \tag{8}$$

Setting Z_t to a constant gives an unconditional test of equal predictive accuracy. Notice that in this case, the test can tell us the direction of any violations: if J_T is negative, $E[\Delta L_{t+1}] < 0$ and the small model is best. For a nominal size of 5% we can compare J_T to the critical value (1.96). If, instead, Z_t is time-varying, we can compare $|J_T|$ with 1.96 for a test with a nominal size of 5%.[11]

We consider prediction models estimated with either a rolling or an expanding window. For the rolling window, $\hat{\phi}_t = \hat{\phi}_t^{roll}$ is estimated using the most recent n observations, $\{y_s\}_{s=t-n+1}^{t}$. The associated loss differential and test statistic based on $\hat{\phi}^{roll}$ are denoted by ΔL_{t+1}^{roll} and J_T^{roll}. For the expanding window: $\hat{\phi}_t = \hat{\phi}_t^{expd}$ is estimated using all data points up to time t, $\{y_s\}_{s=1}^{t}$ and the associated loss differential and test statistic are denoted by ΔL_{t+1}^{expd} and J_T^{expd}.

We consider three combinations of the length of the rolling window and the evaluation sample $(n,p) = \{(60,500),(100,500),(240,1000)\}$. For monthly data, a rolling estimation window of 60 observations corresponds to using five years of data, which is a fairly common choice. Increasing this to 100 should reduce estimation error. Estimation error should be less of a concern in the case where $(n,p) = (240,1,000)$. Table 2 presents results based on two instruments, namely (i) $Z_t = constant$; (ii) $Z_t = |y_t|$.[12] The rejection probabilities are computed using 5000 random samples.

[11] This is equivalent to the procedure in Giacomini and White (2006) where J_T^2 is compared with quantiles of a $\chi^2(1)$ variable.

[12] To see that the GW test should have power for $Z_t = |y_t|$, note that, ignoring estimation error, $\Delta L_{t+1}^i = y_{t+1}^2 - (y_{t+1} - \phi y_t)^2 = \phi^2 y_t^2 + 2\phi y_t \varepsilon_{t+1}$. Using that y_t is Gaussian, we have $E[\Delta L_{t+1} z_t] = \phi^2 E|y_t|^3 > 0$.

First consider the case where $Z_t = constant$. For this case, the power of a test of the null of equal predictive accuracy (2) based on the test statistic in (8) tends to be far higher under an expanding than under a rolling estimation window. For example, when $\phi = 0.2$ and $p = 500$, the probability of rejecting the null of equal predictive accuracy is only 13.5% and 25.6% under the rolling estimation window with $n = 60$ and $n = 100$, respectively, compared to 49.9% and 52.5% under the expanding estimation scheme. Similar differences arise in the case with a time varying instrument, $Z_t = |y_t|$, shown in Table 2, for which the frequency of rejections of the null of equal predictive accuracy is 11.4% and 21.7% for $n = 60$ and $n = 100$, respectively, against rejection rates of 42.8% and 44.3% under the expanding estimation scheme, again assuming that $\phi = 0.2$ and $p = 500$.

5 Empirical Applications

We next consider two empirical applications to illustrate our theoretical analysis. We first compare the performance of big and small forecasting models estimated using rolling and expanding windows and two popular predictors of U.S. stock market returns. Our second application considers a simple autoregressive specification for quarterly U.S. inflation augmented with a common factor predictor variable.

5.1 Predictability of Stock Returns

An extensive literature in finance investigates whether the equity risk premium is constant or varies over time. For example, Welch and Goyal (2008) and Campbell and Thompson (2008) contrast the performance of a (constant) "prevailing mean" model with that of a range of univariate forecasting models that use a single time varying predictor. Following this literature, we consider two predictor variables (x_t), namely the dividend-price ratio, measured as dividends over the preceding 12 months divided by the stock price, and the 1-month T-bill rate.[13] Our analysis uses monthly data on returns on the S&P500 index, net of a one-month T-bill rate, as the dependent variable, y_{t+1}. The data, collected by Welch and Goyal (2008), start in 1927 and end in 2013 and so consists of $T = 1044$ observations.

The small forecasting model takes the form $y_{t+1} = \mu + \varepsilon_{t+1}$, while the big model is $y_{t+1} = \mu + \beta x_t + \varepsilon_{t+1}$. To explore predictability of stock returns, define $\Delta L^{i,j}_{t+1} = (y_{t+1} - \hat{y}^i_{S,t+1})^2 - (y_{t+1} - \hat{y}^j_{B,t+1})^2$, where $i, j \in \{roll, expd\}$ and $\hat{y}_{S,t+1}$ is the mean of a rolling (or expanding) window of observations (denoted $benchm^{roll}$ or $benchm^{expd}$). Moreover, $\hat{y}^{roll}_{B,t+1} = \hat{\mu}^{roll}_t + \hat{\beta}^{roll}_t x_t$ and $\hat{y}^{expd}_{B,t+1} = \hat{\mu}^{expd}_t + \hat{\beta}^{expd}_t x_t$, where $(\hat{\mu}^{roll}_t, \hat{\beta}^{roll}_t)$ is the OLS estimator using the most recent n observations $\{(y_{s+1}, x_s)\}^{t-1}_{s=t-n}$, while $(\hat{\mu}^{expd}_t, \hat{\beta}^{expd}_t)$ is the OLS estimator using all observations up to time t, $\{(y_{s+1}, x_s)\}^{t-1}_{s=1}$. Combining the two different estimation schemes for the small model and the big model yields a total of four different model comparisons. To evaluate these different forecasting methods, we compute the

[13] Both predictors have been used extensively in the literature on predictability of stock market returns; see, e.g., Ang and Bekaert (2007), Pettenuzzo et al. (2014), and Rapach et al. (2010).

following test statistic:

$$J_T^{i,j} = \frac{p^{-1/2}\sum_{t=n+1}^{n+p}\Delta L_t^{i,j}}{\sqrt{p^{-1}\sum_{t=n+1}^{n+p}(\Delta L_t^{i,j})^2}}.$$

Table 4 shows results for n ranging from 40 through 400. In all cases the sample used in the forecast comparisons runs from $n+1$ to $n+p$ and so the performance of the models estimated on an expanding window will also change as n changes. Compared to the (small) constant expected return model estimated using either a rolling (first column) or an expanding (third column) window, the rolling window forecasts from the big model result in significantly less accurate forecasts when $n \leq 150$. Although the big model based on the rolling estimation window performs relatively better when n rises above 200, it continues to underperform the small model for all window sizes. In contrast, while the big forecasting model estimated with an expanding window does not produce significantly more accurate forecasts than the small model, it performs better in most comparisons against the small model estimated on a rolling window (second column) and it delivers similar performance when both the big and small models are estimated with an expanding window (fourth column). Similar results hold when the big model includes the T-bill rate as a predictor (panels 5-8).

We conclude from these findings that it can make a material difference in practice whether an expanding or a rolling estimation window is used to estimate the parameters of the forecasting models and compare the predictive accuracy across the big and small models. We find that the performance of the big model, measured relative to that of the small model, can be substantially better when an expanding estimation window is used. Due to its larger number of estimated parameters, parameter estimation error tends to affect the big model more than the small model, so using more observations, as is done by the expanding estimation window, tends to help us in identifying any predictive gains from including additional predictors in the forecasting model.

5.2 Inflation Forecasting

Our second empirical application looks at predictability of U.S. inflation measured as $\pi_t = 400 \times ln(CPI_t/CPI_{t-1})$, where CPI_t is the consumer price index measured at the quarterly frequency. The small model is assumed to be a fourth order autoregressive specification,

$$\pi_t = c + \rho_1\pi_{t-1} + \rho_2\pi_{t-2} + \rho_3\pi_{t-3} + \rho_4\pi_{t-4} + \varepsilon_t,$$

The big model augments the small model by including a common factor, PC_t, the first principal component extracted from the large cross-section of macroeconomic variables used by Jurado et al. (2015), and hence takes the form

$$\pi_t = c + \rho_1\pi_{t-1} + \rho_2\pi_{t-2} + \rho_3\pi_{t-3} + \rho_4\pi_{t-4} + \beta PC_{t-1} + \varepsilon_t,$$

where PC_t denotes the additional predictor variable.

Our analysis uses quarterly data starting from 1960:1 and finishing in 2011:4, yielding a sample of $T = 216$ observations. Once again we compare the big model, estimated under a rolling or an expanding estimation window, to the small AR(4) model estimated on a rolling or an expanding window. Results are reported in in Table 5.

For small window sizes such as $n = 20$ or $n = 40$ (corresponding to five or ten years of quarterly observations), forecasts from the big model based on a rolling estimation window are significantly less accurate than forecasts from the small AR(4) model. The big model's performance measured against that of the small model continues to be poor (though not significantly worse) when the big model is estimated on a rolling window of observations. In contrast, when estimated using an expanding window, the big model performs markedly better against the small model.

6 Conclusion

This paper generalizes the Giacomini-White (2006) framework for testing the (relative) predictive accuracy of competing prediction models to settings with an expanding estimation window. This generalization is important because the Giacomini-White test approach is the only setup flexible enough to allow comparisons of the predictive accuracy of forecasts generated by a range of methods in common use such as Bayesian, DSGE, non-parametric and model selection methods. Using a martingale central limit theorem, we show that our generalization does not require tightening the assumptions of Giacomini and White for the distributional results on their test statistic to remain valid even under a recursively expanding estimation scheme.

Our simulation results and empirical applications show that it can make a material difference in practice whether an expanding or a rolling estimation window is used to compare the predictive accuracy across models. For forecasts of both inflation and stock market returns, we find that the performance of the big model, measured relative to that of the small model, can be substantially better when an expanding estimation window is used. Parameter estimation error tends to affect the big model more than the small model, due to its larger number of estimated parameters, so using more observations, as is done under an expanding estimation window, tends to help us in better identifying any predictive gains from including additional predictors.

Appendix: Proofs

Proof of Theorem 1. Let $D_{t,n} = a_T h_{t,n}$. By the continuous mapping theorem, we need to show that $\hat{\Omega}_p^{-1/2} p^{-1/2} \sum_{t=n+1}^{n+p} D_{t,n} \to^d N(0, I_k)$, where $\hat{\Omega}_p = p^{-1} \sum_{t=n+1}^{n+p} D_{t,n} D_{t,n}'$. By the Cramer-Wold device, this is equivalent to $\lambda' \hat{\Omega}_p^{-1/2} p^{-1/2} \sum_{t=n+1}^{n+p} D_{t,n} \to^d N(0,1)$ for any $\lambda \in \mathbb{R}^k$ with $\|\lambda\|_2 = 1$.

We prove the result assuming the following claims, which will be proved subsequently.

Claim (a): $\hat{\Omega}_p - \Omega_p = o_P(1)$, where $\Omega_p = p^{-1} \sum_{t=n+1}^{n+p} E(D_{t,n} D_{t,n}')$.

Claim (b): $E|\lambda'\Omega_p^{-1/2}D_{t,n}|^{2+\delta}$ is bounded by a finite constant.

By assumption, $\liminf \lambda_{\min}(\Omega_p) > 0$. Thus, Claim (a) implies that $p^{-1}\sum_{t=n+1}^{n+p}(\lambda'\Omega_p^{-1/2}D_{t,n})^2 = \lambda'\Omega_p^{-1/2}\hat{\Omega}_p\Omega_p^{-1/2}\lambda = 1 + o_P(1)$. Hence, by Claim (b) and Corollary 5.26 of White (2014), we have $p^{-1/2}\sum_{t=n+1}^{n+p}\lambda'\Omega_p^{-1/2}D_{t,n} \to^d N(0,1)$. By Slutsky's theorem and Claim (a), we have $\lambda'\hat{\Omega}_p^{-1/2}p^{-1/2}\sum_{t=n+1}^{n+p}D_{t,n} \to^d N(0,1)$.

Next, we proceed to prove Claim (a) and Claim (b). To show Claim (a), let $D_{t,n,i}$ denote the ith entry of $D_{t,n}$ and $\hat{\Omega}_{p,i,j}$ the (i,j) entry of $\hat{\Omega}_p$. Similar notations also apply to Ω_p. Let $\mathcal{G}_{t,s}^T$ denote the σ-algebra generated by $\{W_{T,t},\cdots,W_{T,s}\}$. Let $M_{t,n,i,j} = D_{t,n,i}D_{t,n,j} - E(D_{t,n,i}D_{t,n,j})$ for $i,j \in \{1,\cdots,k\}$. By Equations (2) and (3) of Andrews (1988), $\{M_{t,n,i,j},\mathcal{G}_{-\infty,t}^T\}$ is an L^1-mixingale (see Definition 2 of Andrews (1988)) with constants $c_{p,t} = \|M_{t,n,i,j}\|_{L^r}$ and $\psi_n = 6\alpha([n/2])^{1-r^{-1}}$, where $r = 1+\delta$, $\alpha(\cdot)$ is the $\alpha-$mixing coefficient of $\{W_{T,t}\}$, $[\cdot]$ denotes the integer part of a positive real number and $\|M_{t,n,i,j}\|_{L^r} = (E|M_{t,n,i,j}|^r)^{1/r}$. Since $E\|a_T h_{t,n}\|_2^{2+\delta} < \Delta$, we have $c_{p,t} = \|M_{t,n,i,j}\|_{L^r} < \Delta$. Hence, $\limsup_{p\to\infty} p^{-1}\sum_{t=n+1}^{n+p} c_{p,t} < \infty$. By Theorem 4.2 of Gut (2013), $\{M_{t,n,i,j}\}$ is uniformly integrable. It follows, by Theorem 2 of Andrews (1988), that $p^{-1}\sum_{t=n+1}^{n+p} M_{t,n,i,j} = o_P(1)$. This means that $\hat{\Omega}_{p,i,j} - \Omega_{p,i,j} = p^{-1}\sum_{t=n+1}^{n+p} M_{t,n,i,j} = o_P(1)$. Claim (a) follows.

Claim (b) follows by $|\lambda'\Omega_p^{-1/2}D_{t,n}| \le \|\Omega_p^{-1/2}\lambda\|_2\|D_{t,n}\|_2$, $\|\Omega_p^{-1/2}\lambda\|_2 \le 1/\sqrt{\lambda_{\min}(\Omega_p)} = O(1)$ and $E\|D_{t,n}\|_2^{2+\delta} = E\|a_T h_{t,n}\|_2^{2+\delta} < \Delta$. □

Proof of Lemma 1. Notice that $\Delta L_{t+1,n} = (\tilde{\delta}_t - \hat{\delta}_t)'Q_t(2\varepsilon_{t+1} - Q_t'(\tilde{\delta}_t + \hat{\delta}_t))$, where $\hat{\delta}_t = (\hat{\beta}_{t,n}' - \beta_0', \hat{\gamma}_{t,n}' - \gamma_p')'$ and $\tilde{\delta} = (\tilde{\beta}_{t,n}' - \beta_0', -\gamma_p')'$. Let $D_{t+1,n} = n\Delta L_{t+1,n}$. Then $ED_{t+1,n}^2 = nE\varepsilon_{t+1}^2(\tilde{\delta}_t - \hat{\delta}_t)'Q_tQ_t'(\tilde{\delta}_t - \hat{\delta}_t) + nE((\tilde{\delta}_t - \hat{\delta}_t)'Q_tQ_t'(\tilde{\delta}_t + \hat{\delta}_t))^2 \ge c_1 nE(\tilde{\delta}_t - \hat{\delta}_t)'Q_tQ_t'(\tilde{\delta}_t - \hat{\delta}_t)$.

We claim that $(\tilde{\delta}_t, \hat{\delta}_t)$ is asymptotically independent of Q_t. To see this, consider the OLS estimators using a rolling window with observations from $t-1$ to $t-\kappa_n$, where $n-\kappa_n \to \infty$, $\kappa_n \to \infty$ and $\kappa_n/n \to 0$. We refer to this as the "incomplete estimator". Since $\kappa_n/n \to 0$, the difference between this incomplete estimator and $(\hat{\beta}_{t,n}, \hat{\gamma}_{t,n}, \tilde{\beta}_{t,n})$ is of order $o_P(n^{-1})$. Since $\kappa_n \to \infty$ and there is a κ_n-period difference between the incomplete estimator and Q_t, the mixing condition implies that the incomplete estimator is asymptotically independent of Q_t. Hence, $(\hat{\beta}_{t,n}, \hat{\gamma}_{t,n}, \tilde{\beta}_{t,n})$ is asymptotically independent of Q_t.

It follows that $c_1 nE(\tilde{\delta}_t - \hat{\delta}_t)'Q_tQ_t'(\tilde{\delta}_t - \hat{\delta}_t) = c_1 \text{trace}E[(Q_tQ_t'n(\tilde{\delta}_t - \hat{\delta}_t)(\tilde{\delta}_t - \hat{\delta}_t)'] \to c_1\text{trace}(\Sigma_Q\Omega)$, where $\Omega = \lim_{T\to\infty} nE[(\tilde{\delta}_t - \hat{\delta}_t)(\tilde{\delta}_t - \hat{\delta}_t)']$. Since V is the lower-right submatrix of Ω and is nonzero, the largest eigenvalue of Ω is positive. By the von-Neumann inequality, $\text{trace}(\Sigma_Q\Omega) \ge \lambda_{\min}(\Sigma_Q)\lambda_{\max}(\Omega) > 0$. □

Proof of Lemma 2. To prove this result, we compute the mean squared error for each of the individual models. To this end, define $y_{(t)} = (y_1,\cdots,y_t)' \in \mathbb{R}^t$ and $\varepsilon_{(t)} = (\varepsilon_1,\cdots,\varepsilon_t)' \in \mathbb{R}^t$. For the small model, $y_{t+1} - \hat{y}_{S,t+1} = c + (\beta - \hat{\theta}_t)x_t + \varepsilon_{t+1}$. Using that $\hat{\theta}_t = x_{(t)}'y_{(t)}/(x_{(t)}'x_{(t)})$ and $y_{(t)} = c1_t + x_{(t)}\beta + \varepsilon_{(t)}$, we have

$$y_{t+1} - \hat{y}_{S,t+1} = c\left(1 - \frac{x_{(t)}'1_t}{x_{(t)}'x_{(t)}}x_{t+1}\right) - \frac{x_{(t)}'\varepsilon_{(t)}}{x_{(t)}'x_{(t)}}x_{t+1} + \varepsilon_{t+1}.$$

Simple computations yield

$$E\left(y_{t+1}-\hat{y}_{S,t+1}\right)^2 = c^2\left(1-\frac{x'_{(t)}1_t}{x'_{(t)}x_{(t)}}x_{t+1}\right)^2+\sigma^2\left(1+\frac{x_{t+1}^2}{x'_{(t)}x_{(t)}}\right). \tag{9}$$

For the big model, $y_{t+1}-\hat{y}_{B,t+1} = (c-\hat{c}_t)+(\beta-\hat{\beta}_t)x_t+\varepsilon_{t+1} = \varepsilon_{t+1}-X_t'\begin{pmatrix}\hat{c}_t-c\\ \hat{\beta}_t-\beta\end{pmatrix}$. Since $\begin{pmatrix}\hat{c}_t-c\\ \hat{\beta}_t-\beta\end{pmatrix} = \left(X'_{(t)}X_{(t)}\right)^{-1}X'_{(t)}\varepsilon_{(t)}$, we have $y_{t+1}-\hat{y}_{B,t+1}=\varepsilon_{t+1}-X_t'\left(X'_{(t)}X_{(t)}\right)^{-1}X'_{(t)}\varepsilon_{(t)}$. By simple computations, we obtain

$$E\left(y_{t+1}-\hat{y}_{B,t+1}\right)^2 = \left(1+X_t'\left(X'_{(t)}X_{(t)}\right)^{-1}X_t\right)\sigma^2. \tag{10}$$

The desired result follows by setting $p^{-1}\sum_{t=n+1}^{n+p}E\left(y_{t+1}-\hat{y}_{S,t+1}\right)^2 = p^{-1}\sum_{t=n+1}^{n+p}E\left(y_{t+1}-\hat{y}_{B,t+1}\right)^2$ and using (9) and (10). □

References

Andrews, Donald WK (1988), 'Laws of large numbers for dependent non-identically distributed random variables', *Econometric theory* **4**(03), 458–467.

Ang, Andrew and Geert Bekaert (2007), 'Stock return predictability: Is it there?', *Review of Financial Studies* **20**(3), 651–707.

Bradley, Richard C (2007), *Introduction to strong mixing conditions*, Vol. 1, Kendrick Press Heber City.

Campbell, John Y and Samuel B Thompson (2008), 'Predicting excess stock returns out of sample: Can anything beat the historical average?', *Review of Financial Studies* **21**(4), 1509–1531.

Clark, T.E. and M.W. McCracken (2011), 'Nested forecast model comparisons: A new approach to testing equal accuracy', *Unpublished Working Paper* .

Clark, Todd E and Michael W McCracken (2001), 'Tests of equal forecast accuracy and encompassing for nested models', *Journal of econometrics* **105**(1), 85–110.

Clark, Todd E and Michael W McCracken (2005), 'Evaluating direct multistep forecasts', *Econometric Reviews* **24**(4), 369–404.

Diebold, Francis X (2015), 'Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold-mariano tests', *Journal of Business & Economic Statistics* **33**(1), 1–8.

Diebold, Francis X and Roberto S Mariano (1995), 'Comparing predictive accuracy', *Journal of Business & Economic Statistics* pp. 253–263.

Giacomini, Raffaella and Halbert White (2006), 'Tests of conditional predictive ability', *Econometrica* **74**(6), 1545–1578.

Gut, Allan (2013), *Probability A Graduate Course*, Springer.

Hansen, P. R. and Allan Timmermann (2015*a*), 'Comment on comparing predictive accuracy, twenty years later', *Journal of Business and Economic Statistics* **33**, 17–21.

Hansen, Peter Reinhard and Allan Timmermann (2015*b*), 'Equivalence between out-of-sample forecast comparisons and wald statistics', *Econometrica* **83**(6), 2485–2505.
URL: *http://dx.doi.org/10.3982/ECTA10581*

Inoue, Atsushi and Lutz Kilian (2005), 'In-sample or out-of-sample tests of predictability: Which one should we use?', *Econometric Reviews* **23**(4), 371–402.

Jurado, Kyle, Sydney C Ludvigson and Serena Ng (2015), 'Measuring uncertainty', *The American Economic Review* **105**(3), 1177–1216.

McCracken, Michael W (2007), 'Asymptotics for out of sample tests of granger causality', *Journal of Econometrics* **140**(2), 719–752.

Pettenuzzo, Davide, Allan Timmermann and Rossen Valkanov (2014), 'Forecasting stock returns under economic constraints', *Journal of Financial Economics* **114**(3), 517–553.

Rapach, David E, Jack K Strauss and Guofu Zhou (2010), 'Out-of-sample equity premium prediction: Combination forecasts and links to the real economy', *Review of Financial Studies* **23**(2), 821–862.

Welch, Ivo and Amit Goyal (2008), 'A comprehensive look at the empirical performance of equity premium prediction', *Review of Financial Studies* **21**(4), 1455–1508.

West, Kenneth D (1996), 'Asymptotic inference about predictive ability', *Econometrica: Journal of the Econometric Society* pp. 1067–1084.

White, Halbert (2014), *Asymptotic theory for econometricians*, Academic press.

Table 1: Size of J test of predictive accuracy in Monte Carlo Simulations

$P(J_T^2 > \chi^2_{0.95,1})$						
$n \backslash T$	100	200	300	400	600	800
50	0.043	0.038	0.047	0.042	0.041	0.036
100	NA	0.049	0.045	0.054	0.039	0.040
150	NA	0.045	0.049	0.051	0.047	0.043
250	NA	NA	0.053	0.047	0.056	0.048
500	NA	NA	NA	NA	0.064	0.052
600	NA	NA	NA	NA	NA	0.049

The table shows rejection frequencies for a J test of equal predictive accuracy of a small and a big forecasting model. The assumed data generating process is $y_t = c + \beta x_t + \varepsilon_t$, where $\varepsilon_t \sim N(0, \sigma^2)$. The small forecasting model omits the constant, c, from the model, while the big forecasting model includes both the constant and x_t. Both models use an expanding estimation window. The constant, c, is set in accordance with Lemma 3, ensuring that the expected squared error loss is the same for the two models in a sample of T observations, n of which are used to evaluate the forecasts.

Table 2: Rejection frequencies for models estimated under rolling and expanding estimation windows: unconditional tests

Panel 1: $P(J_T^{roll} > 1.96)$

$(n,p)\backslash\phi$	0.05	0.1	0.15	0.2	0.25
(60,500)	0.000	0.000	0.015	0.135	0.457
(100,500)	0.000	0.004	0.040	0.256	0.654
(240,1000)	0.001	0.038	0.383	0.869	0.994

Panel 2: $P(J_T^{expd} > 1.96)$

$(n,p)\backslash\phi$	0.05	0.1	0.15	0.2	0.25
(60,500)	0.003	0.030	0.184	0.499	0.791
(100,500)	0.005	0.051	0.220	0.525	0.805
(240,1000)	0.016	0.188	0.604	0.905	0.990

This table reports rejection frequencies for the null of equal unconditional expected mean squared error performance of a small and a big forecasting model. The data generating process is an AR(1) model of the form $y_{t+1} = \phi y_t + \varepsilon_{t+1}$, $\varepsilon_{t+1} \sim i.i.d.N(0,1)$, and the small forecasting model always predicts zero, whereas the big model predicts $\hat{y}_{t+1|t} = \hat{\phi}_t y_t$, where $\hat{\phi}_t$ is estimated on a rolling window with m observations (Panel 1) or using an expanding estimation window (Panel 2).

Table 3: Rejection frequencies for models estimated under rolling and expanding estimation windows: conditional tests

Panel 1: $P(|J_T^{roll}| > 1.96)$

$(n,p)\backslash\phi$	0.05	0.1	0.15	0.2	0.25
(60,500)	0.160	0.060	0.024	0.114	0.368
(100,500)	0.102	0.034	0.050	0.217	0.541
(240,1000)	0.060	0.046	0.322	0.767	0.967

Panel 2: $P(|J_T^{expd}| > 1.96)$

$(n,p)\backslash\phi$	0.05	0.1	0.15	0.2	0.25
(60,500)	0.044	0.039	0.165	0.428	0.689
(100,500)	0.051	0.056	0.195	0.443	0.728
(240,1000)	0.042	0.174	0.521	0.831	0.971

This table reports rejection frequencies for the null of equal conditional expected mean squared error performance of a small and a big forecasting model, using $Z_t = |y_t|$ as the predictive instrument. The data generating process is an AR(1) model of the form $y_{t+1} = \phi y_t + \varepsilon_{t+1}$, $\varepsilon_{t+1} \sim i.i.d.N(0,1)$, and the small forecasting model always predicts zero, whereas the big model predicts $\hat{y}_{t+1|t} = \hat{\phi}_t y_t$, where $\hat{\phi}_t$ is estimated on a rolling window with n observations (Panel 1) or using an expanding estimation window (Panel 2).

Table 4: Predictability of stock returns using forecasting models estimated with rolling or expanding windows

	x_t is dp					x_t is tbl			
	$benchmark^{roll}$		$benchmark^{expd}$			$benchmark^{roll}$		$benchmark^{expd}$	
n	J_T^{roll}	J_T^{expd}	J_T^{roll}	J_T^{expd}	n	J_T^{roll}	J_T^{expd}	J_T^{roll}	J_T^{expd}
40	-2.66	-0.18	-2.79	-1.03	40	-1.72	0.35	-2.07	-0.84
80	-2.20	0.79	-2.59	-0.26	80	-1.30	1.43	-1.82	0.00
120	-1.77	0.15	-1.69	0.14	120	-0.92	0.01	-0.80	-0.05
150	-2.18	0.10	-2.65	-0.09	150	-1.11	0.45	-1.02	0.16
200	-0.55	0.32	-1.13	-0.09	200	-0.72	0.75	-0.88	0.20
250	0.17	0.39	-0.41	-0.12	250	-0.97	0.78	-1.20	0.20
300	-0.46	0.15	-1.25	-0.48	300	-0.49	0.81	-0.84	0.21
400	0.16	0.04	-0.27	-0.29	400	-0.25	0.37	-0.46	0.11

This table reports t-statistics comparing the out-of-sample mean squared error performance of a benchmark prevailing mean model versus forecasting models with an additional time-varying predictor, i.e., the dividend price ratio (*dp*, columns 1-4) or the T-bill rate (*tbl* columns 5-8). The benchmark model is estimated either on a rolling window with n observations ($benchmark^{roll}$) or using an expanding window ($benchmark^{expd}$). Similarly, the forecasting models are also estimated using rolling windows (with tests denoted J_T^{roll}) or expanding windows ($J_T^{\exp d}$).

Table 5: Predictability of US inflation using forecasting models estimated with rolling or expanding windows

	$benchmark^{roll}$		$benchmark^{expd}$	
p	J_T^{roll}	J_T^{expd}	J_T^{roll}	J_T^{expd}
20	-2.17	3.29	-3.74	-1.00
40	-1.79	-0.28	-1.84	-0.69
80	-1.18	0.84	-1.78	0.07
120	-0.55	0.35	-1.02	-0.32
150	-0.07	0.55	-0.51	-0.07

This table reports t-statistics comparing the out-of-sample mean squared error performance of a benchmark AR(4) model for US inflation versus a forecasting model that adds a principal component to the AR(4) model. The benchmark AR(4) model is estimated either on a rolling window with n observations ($benchmark^{roll}$) or using an expanding window ($benchmark^{\exp d}$). Similarly, the forecasting models are also estimated using rolling windows (with tests denoted J_T^{roll}) or expanding windows ($J_T^{\exp d}$).