# A Test for State-Dependent Predictive Ability based on a Markov-Switching Framework

Sebastian Fossati*
University of Alberta

This version: May 17, 2018

**Abstract**

This paper proposes a new test for comparing the out-of-sample forecasting performance of two competing models for situations in which the predictive content may be state-dependent (for example, expansion and recession states or low and high volatility states). To apply this test the econometrician is not required to observe when the underlying states shift. The test is simple to implement and accommodates several different cases of interest. An out-of-sample forecasting exercise for US output growth using real-time data illustrates the improvement of this test over previous approaches to perform forecast comparison.

*Keywords*: Forecast Evaluation, Testing, Regime Switching, Structural Change
*JEL Codes*: C22, C53

*Contact: Department of Economics, University of Alberta, Edmonton, AB T6G 2H4, Canada. Email: sfossati@ualberta.ca. Web: http://www.ualberta.ca/~sfossati/. A previous version of this paper circulated under the title "Testing for State-Dependent Predictive Ability". I thank Peter Fuleky, Kundan Kishor, Tatevik Sekhposyan, Rodrigo Sekkel, and conference and seminar participants at the University of Alberta, SNDE 2016, Central Bank of Uruguay, IAAE 2016, and SEU 2017 for helpful comments.

# 1 Introduction

A well-established empirical fact in the macroeconomic forecasting literature is that predictability is unstable over time. For example, many individual indicators exhibit significant out-of-sample predictive content for output growth and inflation but only sporadically. This result has been documented in Stock and Watson (2003), Giacomini and Rossi (2010), Rossi and Sekhposyan (2010), Rossi (2013), and Granziera and Sekhposyan (2017), among others. A recent literature has established a new (but related) empirical fact: predictability varies across economic states. For example, Dotsey et al. (2015) report that out-of-sample Phillips curve forecasts of the inflation rate tend to be more accurate, relative to a benchmark model, during economic recessions but less accurate during expansions (see also Gibbs and Vasnev, 2017). Similarly, Chauvet and Potter (2013) find that most output growth forecasting models exhibit a similar performance during economic expansions but one model performs significantly better during recessions. Evidence of state-dependent predictability has also been documented in the empirical finance literature. For example, work by Rapach et al. (2010), Henkel et al. (2011), and Dangl and Halling (2012) shows stronger evidence for out-of-sample stock return predictability during economic recessions than during expansions. Likewise, Gargano et al. (2016) find that the degree of predictability of bond excess returns rises during recessions, while Gargano and Timmermann (2014) find similar results for commodity prices.

The most common approach to test for state-dependent predictive ability relies on exogenously provided shift dates and the results on conditional predictive ability of Giacomini and White (2006). For example, the literature cited above uses peak and trough dates determined by the Business Cycle Dating Committee of the National

Bureau of Economic Research (NBER) to identify recession and expansion periods in the US economy. The relative performance of the two forecasting models is then evaluated using a test of unconditional predictive ability (Diebold and Mariano, 1995; West, 1996; Clark and West, 2006; Giacomini and White, 2006) applied to recession and expansion periods separately (see, for example, Chauvet and Potter, 2013). But this testing strategy is not feasible if the econometrician is not able to observe when the underlying states shift (for example, low and high volatility states).[1] In addition, existing methods for evaluating relative forecasting performance in unstable environments (Giacomini and Rossi, 2010; Martins and Perron, 2016) are inadequate when one of the states is constrained to just a few observations. This is the case of recessions in the US economy which tend to be short-lived, typically lasting less than four quarters. For example, the Fluctuation test of Giacomini and Rossi (2010) is based on a rolling average of loss differences and, as a result, it works best when the relative performance of the two models changes abruptly (a structural break).

In this paper I propose a new test for comparing the out-of-sample forecasting performance of two competing models for situations in which the predictive content may be state-dependent. The main advantage of this test is that the econometrician is not required to observe when the underlying states shift. Similar to Giacomini and Rossi (2010) and Martins and Perron (2016), I use the framework of Giacomini and White (2006) to treat the sequence of out-of-sample loss differences as observed data. Next, forecast loss differences are modeled using a Markov-switching mean plus noise process which can be used to test for state-dependent predictive ability. The test is simple to implement and accommodates several different cases that can be of interest. For example, two competing models (or predictors) can exhibit equal predictive ability

[1] I am assuming here that the NBER dating provides a reasonably accurate approximation to the underlying recession and expansion states which are, of course, not observable.

in one state, whereas one of the models is more accurate in the other. Alternatively, one model could be more accurate in the first state while the other model could be more accurate in the second. In addition, there can be differences in the degree of persistence or expected duration of the two states. For example, both states could be very persistent or one state could be very persistent (expansions) while the other state is constrained to shorter periods (recessions). Finally, the test can also accommodate a permanent structural break (one level shift) in the loss differences, as the one-time reversal test of Giacomini and Rossi (2010).

In the next section I use an illustrative example to show that current tests of equal predictive ability (Giacomini and White, 2006; Giacomini and Rossi, 2010) can fail to reject the null hypothesis if the superior performance of one of the competing models is constrained to very short periods (just a few observations), even if observed repeatedly. This finding is consistent with results documented in Casini (2018). Next, I present an (asymptotic) heteroskedasticity and autocorrelation consistent (HAC) Wald test of the null hypothesis of equal and constant predictive ability against the alternative of state-dependent predictive ability. In section 3 I investigate the small-sample properties of the tests using Monte Carlo simulations. Size and power properties of the tests are evaluated in two situations: (i) allowing for unequal but constant relative forecasting performance of the two competing models; (ii) allowing for different relative forecasting performance of the two models in different states (the business cycle). In section 4 I use an out-of-sample forecasting exercise for US output growth to illustrate the usefulness of the proposed test over previous approaches to forecast comparison in unstable environments. Finally, section 5 concludes.

# 2 State-Dependent Predictive Ability

In this section I present the test for comparing the out-of-sample forecasting performance of two models for situations in which the predictive content may be state-dependent. Of special interest is the case in which one of the states is of short duration, maybe just a few observations. The literature on testing for equal predictive ability is substantial. The focus of this paper is on finite-sample predictive ability (Giacomini and White, 2006), unstable environments (Giacomini and Rossi, 2010), and forecast instabilities of short duration (Casini, 2018). See Clark and McCracken (2011), Giacomini (2011), Rossi (2013), and Casini and Perron (2018) for detailed surveys of the literature.

## 2.1 Environment

The objective is to compare sequences of $h$-step ahead out-of-sample forecasts for the variable $y_t$ obtained from two competing models. The models are characterized by parameters $\delta$ and $\gamma$, respectively. It is assumed that a sample of size $T$ has been divided into an in-sample portion of size $R$ and an out-of-sample portion of size $P$. For a given loss function $L(\cdot)$, we have a sequence of $P$ out-of-sample forecast loss differences $\{\Delta L_t(\hat{\delta}_{t-h,R}, \hat{\gamma}_{t-h,R})\}_{t=R+h}^{T}$ with $\Delta L_t(\hat{\delta}_{t-h,R}, \hat{\gamma}_{t-h,R}) = L^{(1)}(y_t, \hat{\delta}_{t-h,R}) - L^{(2)}(y_t, \hat{\gamma}_{t-h,R})$. As in Giacomini and White (2006), the parameters of the models are estimated using a fixed scheme (that is, estimated once using observations $1, \ldots, R$) or a rolling scheme (that is, re-estimated at each $t = R+h, \ldots, T$ using observations $t-h-R+1, \ldots, t-h$). We are interested in testing the null hypothesis of equal predictive ability

$$H_0^{(1)} : E\left[\Delta L_t(\hat{\delta}_{t-h,R}, \hat{\gamma}_{t-h,R})\right] = 0 \text{ for all } t = R+h, \ldots, T, \tag{1}$$

which is tested against the alternative hypothesis that one of the models exhibits

superior predictive ability (that is, $E[\Delta L_t(\hat{\delta}_{t-h,R}, \hat{\gamma}_{t-h,R})] \neq 0$). Giacomini and White (2006) note that since the null hypothesis $H_0^{(1)}$ is stated in terms of the estimated parameters $\hat{\delta}_{t-h,R}$ and $\hat{\gamma}_{t-h,R}$, not the population parameters $\delta$ and $\gamma$, this is a test of predictive ability in the finite sample (that is, given models, estimation windows, estimation procedures, etc.).

In this context, Giacomini and White (2006) propose an asymptotic test of equal (unconditional) predictive ability that can be applied to both nested and non-nested models. The statistic is a t-test of the hypothesis that $\mu = 0$ in the simple linear regression $\Delta L_t = \mu + u_t$ for $t = R + h, \ldots, T$ and is given by

$$\text{GW} = \hat{\sigma}_P^{-1} P^{1/2} \overline{\Delta L}_P, \tag{2}$$

where $\overline{\Delta L}_P = P^{-1} \sum_{t=R+h}^{T} \Delta L_t$ and $\hat{\sigma}_P^2$ is the sample variance of $\Delta L_t$ if $h = 1$ or a HAC estimator of the long-run variance if $h > 1$. A common choice for the latter is the kernel-based estimator

$$\hat{\sigma}_P^2 = \sum_{j=-(q-1)}^{q-1} k(j/q) P^{-1} \sum_{t=R+h}^{T} \Delta L_t^* \Delta L_{t-j}^* \tag{3}$$

where $k(\cdot)$ is a kernel weight function, for example the Bartlett kernel of Newey and West (1987), $q$ is a bandwidth that grows with $P$, and $\Delta L_t^* = \Delta L_t - \overline{\Delta L}_P$ (see, for example, Andrews, 1991). Under (1), GW $\xrightarrow{d}$ $N(0,1)$ as $P \to \infty$ and the null hypothesis of equal unconditional predictive ability is rejected at the 5% level if $|\text{GW}| > 1.96$. As a result, the test of equal unconditional predictive ability of Giacomini and White (2006) coincides with the test proposed in Diebold and Mariano (1995).[2]

[2]In addition to this test of equal unconditional predictive ability, Giacomini and White (2006) also propose a test of conditional predictive ability. The intuition is that the econometrician could use available information to predict which of the competing models will be more accurate at a given time. In this case, the statistic is a Wald test of the hypothesis that $\beta = 0$ in the regression $\Delta L_t = \beta' X_{t-h} + u_t$, where $X_{t-h}$ includes a constant and other variables (for example, a business cycle indicator).

In related work, Giacomini and Rossi (2010) propose asymptotic tests of the joint hypothesis of equal and constant performance of the two models, $H_0^{(1)}$, against the alternative of local predictive ability. In the regression with time-varying parameters $\Delta L_t = \mu_t + u_t$, they propose tests of the null hypothesis that $\mu_t = 0$ for all $t$ against different specifications of the alternative hypothesis. For example, the Fluctuation test is based on a rolling average of loss differences and the statistic is given by

$$\text{Fluct}_{t,m} = \hat{\sigma}_P^{-1} m^{-1/2} \sum_{j=t-m/2}^{t+m/2+1} \Delta L_j. \tag{4}$$

The statistic is computed for $t = R + h + m/2, \ldots, T - m/2 + 1$, with $m$ the window size and $\hat{\sigma}_P^2$ a HAC estimator of the long-run variance of $\Delta L_t$, for example the kernel-based estimator (3). The test rejects when $\max_t |\text{Fluct}_{t,m}| > k_\alpha$ with the critical value $k_\alpha$ obtained by simulation. In addition, Giacomini and Rossi (2010) and Martins and Perron (2016) consider tests based on structural break models that can accommodate one or more sudden breaks in relative forecasting performance.

## 2.2 An illustrative example

A substantial recent literature on macroeconomic forecasting employs the tests of equal predictive ability described above to evaluate the predictive content of different indicators or models (see Rossi, 2013, for an extensive survey of this literature). Here I use a simple simulated example to illustrate how these tests can fail to uncover the superior performance of one of the two competing models when the predictive ability is state-dependent.

Consider a data generating process (DGP) given by

$$\begin{aligned} y_t &= -\beta s_t + \sigma_\varepsilon \varepsilon_t, \\ x_t &= \delta s_t + \sigma_\nu \nu_t, \end{aligned}$$

where $\varepsilon_t$ and $\nu_t$ are i.i.d.$N(0,1)$ and $s_t$ is a state variable identifying, for example, recession ($s_t = 1$) and expansion ($s_t = 0$) periods. The econometrician does not observe $s_t$ but rather $x_t$, which is correlated with $s_t$. For example, if $x_t \in [0,1]$ for all $t$, then $x_{t+1}$ could be a predicted probability of recession for quarter $t+1$. I consider two competing forecasting models: (i) $y_{t+1} = u_{1t+1}$ and (ii) $y_{t+1} = \gamma x_{t+1} + u_{2t+1}$. As a result, the time-$t$ one-step ahead forecasts of $y_{t+1}$ are

$$\begin{aligned} \hat{f}^1_{t,R} &= 0, \\ \hat{f}^2_{t,R} &= \hat{\gamma}_{t,R} x_{t+1}, \end{aligned}$$

where $\hat{\gamma}_{t,R}$ is the in-sample rolling estimate of $\gamma$ based on the last $R$ observations and $x_{t+1}$ is assumed known at time $t$. Note that during expansions both models yield forecasts that are on average 0. On the other hand, during recessions $\hat{f}^2_{t,R}$ will be (on average) closer to the actual value $y_{t+1}$. As a result, while the two models should exhibit similar predictive content during expansions, model (ii) should yield more accurate forecasts during recessions. For this exercise, the variable $s_t$ is the actual quarterly time series of NBER recession dates for the sample 1960Q1-2015Q4 (that is, 224 quarters), $\beta = 2$, $\delta = 1$, and $\sigma_\varepsilon = \sigma_\nu = .5$. Based on this set-up, I simulate *one sample* of $y_t$ and $x_t$ of size $T = 224$ and generate the sequence of one-step ahead forecasts of $y_{t+1}$ with $R = 100$ and $P = 124$.

Figure 1 shows the simulated time series (actual) and one-step ahead out-of-sample

forecasts from the two competing models (top), as well as the forecast loss differences computed using a quadratic loss function (bottom). As we can observe, loss differences in recessions are generally large and positive suggesting that forecasts from model (ii) are more accurate during these periods. In expansions, on the other hand, loss differences are typically smaller and on average negative. Descriptive statistics and tests of equal predictive ability are reported in Table 1. For the full hold-out sample (OOS), the average loss difference is close to 0 and positively autocorrelated. In this case, the GW and $\text{Fluct}_{t,m}$ tests fail to reject the null hypothesis of equal predictive ability and we conclude that both models exhibit similar predictive content over the full hold-out sample. Recently, researchers have used peak and trough dates determined by the NBER to evaluate the performance of the two forecasting models during recession and expansion periods separately. In this case, using the GW test we reject the null hypothesis of equal predictive ability for both sub-periods and conclude that model (ii) is more accurate during recessions ($\text{GW} > 0$) but less accurate during expansions ($\text{GW} < 0$).[3]

[ FIGURE 1 ABOUT HERE ]

[ TABLE 1 ABOUT HERE ]

In sum, this exercise illustrates the following issues. First, the GW and $\text{Fluct}_{t,m}$ tests can fail to reject the null hypothesis of equal predictive ability if the superior performance of one of the two competing models is constrained to short periods as in

[3]This approach is used in Rapach et al. (2010), Henkel et al. (2011), Dangl and Halling (2012), Chauvet and Potter (2013), Gargano and Timmermann (2014), Gargano et al. (2016), Dotsey et al. (2015), and Gibbs and Vasnev (2017) to test for state-dependent predictive ability over the business cycle and is equivalent to implementing the test of conditional predictive ability of Giacomini and White (2006).

the case of recessions (just a few observations, even if observed repeatedly). This inability to reject extends to other tests of unconditional predictive ability (for example, West, 1996). Second, state-dependent predictive content can be uncovered using tests of unconditional predictive ability and exogenously provided shift dates if the econometrician is able to observe when the underlying states shift. In many cases, however, such a testing strategy would not be feasible since the underlying states may not be observed by the econometrician.

## 2.3 State-dependent loss differences

To test for state-dependent predictive ability without exogenously provided shift dates, the out-of-sample forecast loss differences can be modeled using the Markov-switching mean plus noise regression

$$\Delta L_t(\hat{\delta}_{t-h,R}, \hat{\gamma}_{t-h,R}) = \mu_{s_t} + \sigma_{s_t} u_t, \tag{5}$$

where $s_t$ $(= 0, 1)$ is an unobserved two-state first-order Markov process with transition probabilities

$$\text{Prob}(s_t = j \,|\, s_{t-1} = i) = p_{ij} \quad i, j = 0, 1, \tag{6}$$

and $u_t$ is an unobservable moving-average (MA) process with zero mean and non-zero autocorrelations up to lag $h - 1$.

Test equation (5) can accommodate several cases of interest. For example, the two models can exhibit equal predictive ability in one regime ($\mu_0 = 0$), whereas one of the models is more accurate in the other regime ($\mu_1 \neq 0$). Alternatively, one model could be more accurate in the first regime ($\mu_0 > 0$) while the other model could be more accurate in the second regime ($\mu_1 < 0$). In addition, there can be differences in

the degree of persistence or expected duration of the two regimes. For example, both regimes could be very persistent (both $p_{00}$ and $p_{11}$ are large) or one regime could be very persistent while the other regime is constrained to shorter periods ($p_{00}$ large and $p_{11}$ small). Finally, (5) can also accommodate a permanent structural break in the relative forecasting performance ($p_{11} = 1$ and the second regime is absorbing), as the one-time reversal test of Giacomini and Rossi (2010). As a result, several hypotheses of interest can be formulated. For example, $H_0^{(1)}$ is the null hypothesis of *equal and constant predictive ability* of the two forecasting models considered in Giacomini and Rossi (2010) and implies testing $\mu_0 = \mu_1 = 0$. In contrast, Martins and Perron (2016) suggest testing the null hypothesis of *constant predictive ability* given by

$$H_0^{(2)} : E\left[\Delta L_t(\hat{\delta}_{t-h,R}, \hat{\gamma}_{t-h,R})\right] = c \text{ for all } t = R+h, \ldots, T, \tag{7}$$

for some constant $c$ against the alternative of changing or, in this case, state-dependent predictive ability. $H_0^{(2)}$ implies testing $\mu_0 = \mu_1$ in (5).[4]

While tests of the null hypothesis of equal predictive ability are about $\mu_{s_t}$, in the test equation I allow for differences in the mean ($\mu_{s_t}$) and also the variance ($\sigma_{s_t}$) of the forecast loss differences across regimes. There are at least two reasons for not imposing $\sigma = \sigma_0 = \sigma_1$ in (5). A practical one is that under $\mu_0 = \mu_1 = 0$ and $\sigma_0 = \sigma_1$ some parameters are only identified under the alternative hypothesis. A more important reason, however, is that allowing for different variances across regimes is empirically relevant as shown both in the illustrative example above (section 2.2) and the empirical application below (section 4). As a result, while tests that impose a constant variance across regimes can be constructed, allowing for different variances can in fact help with

[4]In addition, Martins and Perron (2016) suggest the following testing strategy. If $H_0^{(2)}$ is rejected, conclude that the relative forecasting ability of the two models is not constant. If $H_0^{(2)}$ is not rejected, apply the GW test of unequal (but constant) predictive ability.

the empirical identification of the two regimes.[5]

## 2.4 Asymptotic predictive ability tests

An important result of Giacomini and White (2006) is that under their framework we can treat the sequence of $P$ out-of-sample loss differences $\{\Delta L_t(\hat{\delta}_{t-h,R}, \hat{\gamma}_{t-h,R})\}_{t=R+h}^{T}$ as observed data. For this result to hold, however, the estimation sample size $R$ must remain finite as the out-of-sample size $P$ grows to infinity. As a result, the parameters of the forecasting models ($\hat{\delta}$ and $\hat{\gamma}$) are usually estimated using a rolling scheme with a finite estimation window $R$, while a recursive scheme with an expanding estimation window cannot be used.

In this context, the regime-switching regression (5) can be estimated by (quasi) maximum likelihood (ML) following Hamilton (1989, 1990) or Kim (1994). Equation (5) is characterized by six parameters $\theta = (\mu_0, \mu_1, \sigma_0, \sigma_1, p_{00}, p_{11})'$ and under standard asymptotic normality arguments the asymptotic distribution of the ML estimator $\hat{\theta}$ is $\sqrt{P}\big(\hat{\theta} - \theta_0\big) \xrightarrow{d} N\big(0, \Omega\big)$ as $P \to \infty$. Linear hypotheses of the form $R_0\theta = 0$ can be tested by constructing the Wald test statistic

$$\text{Wald} = P\big(R_0\hat{\theta}\big)'\big(R_0\hat{\Omega}R_0'\big)^{-1}\big(R_0\hat{\theta}\big), \tag{8}$$

where $R_0$ is a $r \times 6$ matrix with $r$ is the number of restrictions, and $\hat{\Omega}$ is a consistent estimator of $\Omega$. Under the null hypothesis, Wald $\xrightarrow{d} \chi^2(r)$ as $P \to \infty$. Wald statistics for $H_0^{(1)}$ and $H_0^{(2)}$, SD-Wald hereafter, can easily be constructed by appropriately selecting the matrix of restrictions $R_0$. For example, to test the null hypothesis of equal and constant predictive ability $H_0^{(1)}$, $R_0 = (I_2, 0_2, 0_2)$ with $I_2$ an identity matrix and

[5]For example, in section 4 I show that in the case of real GDP growth the parameter estimates of the test equation (5) are: $\hat{\mu}_0 = 24.72$ (6.10), $\hat{\mu}_1 = -0.06$ (0.42), $\hat{\sigma}_0^2 = 129.13$ (57.90), $\hat{\sigma}_1^2 = 9.37$ (1.70), $\hat{p}_{00} = 0.48$ (0.25), and $\hat{p}_{11} = 0.98$ (0.02). As a result, we observe a significant difference in variance across regimes with $\hat{\sigma}_0^2/\hat{\sigma}_1^2 \approx 13.78$. See Table 3, Figure 7, and the empirical example below.

$0_2$ a matrix of zeros both of dimension $2 \times 2$. As a result, SD-Wald $\overset{a}{\sim} \chi^2(2)$ and $H_0^{(1)}$ is rejected at the 5% level if SD-Wald $> 5.99$. Similarly, to test the null hypothesis of constant predictive ability $H_0^{(2)}$, $R_0 = (1, -1, 0, 0, 0, 0)$ and the SD-Wald statistic is $P(\hat{\mu}_0 - \hat{\mu}_1)^2/(\hat{\omega}_{11} + \hat{\omega}_{22} - 2\hat{\omega}_{21}) \overset{a}{\sim} \chi^2(1)$, where $\hat{\omega}_{ij}$ is the $(i, j)$ element of $\hat{\Omega}$. As a result, $H_0^{(2)}$ is rejected at the 5% level if SD-Wald $> 3.84$. The rest of this paper will focus on testing the null hypothesis $H_0^{(1)}$.

Construction of the SD-Wald statistic requires a consistent estimator of $\Omega$. Based on the information matrix condition, two estimators are typically used. The first estimator of the covariance matrix is minus the inverse of the average Hessian evaluated at the ML estimates

$$\hat{\Omega}^H = -\left[\frac{1}{P}\sum_{t=R+h}^{T} H_t(\hat{\theta})\right]^{-1}, \tag{9}$$

where $H_t(\theta)$ is the Hessian for observation $t$ and $\sum_{t=R+h}^{T} H_t(\hat{\theta})$ is usually obtained numerically as in Hamilton (1989). The second estimator of the covariance matrix is the inverse of the average outer-product of the score vector evaluated at the ML estimates

$$\hat{\Omega}^{OP} = \left[\frac{1}{P}\sum_{t=R+h}^{T} g_t(\hat{\theta}) g_t(\hat{\theta})'\right]^{-1}, \tag{10}$$

where $g_t(\theta)$ is the score vector for observation $t$ as in Hamilton (1996). For multi-step forecasts (that is, $h > 1$), however, the loss differences are autocorrelated up to lag $h - 1$ and (5) is misspecified.[6] In this case, the information matrix condition does not hold, (9) and (10) are not valid, and a HAC estimator of the covariance matrix is required (Diebold and Mariano, 1995). In addition, if there are instabilities (for example, state-dependency), the forecast loss differences are also autocorrelated under the null hypothesis of linearity (see the illustrative example above) and, as a result,

[6]For $h > 1$, (5) is estimated by quasi-ML ignoring the MA$(h - 1)$ structure.

a HAC estimator is required even if $h = 1$ (Morley and Rabah, 2014; Martins and Perron, 2016).

An alternative estimator of the covariance matrix that remains valid when the information matrix condition does not hold is the 'sandwich' estimator

$$\hat{\Omega}^S = \left[\frac{1}{P}\sum_{t=R+h}^{T} H_t(\hat{\theta})\right]^{-1} \hat{\Sigma}(\hat{\theta}) \left[\frac{1}{P}\sum_{t=R+h}^{T} H_t(\hat{\theta})\right]^{-1}, \tag{11}$$

where $\Sigma$ is the long-run variance of the score vector (see, for example, Hayashi, 2000). A HAC estimate of $\Sigma$ can be obtained using the kernel-based estimator

$$\hat{\Sigma}(\hat{\theta}) = \sum_{j=-(q-1)}^{q-1} k(j/q) P^{-1} \sum_{t=R+h}^{T} g_t(\hat{\theta}) g_{t-j}(\hat{\theta})', \tag{12}$$

where $k(\cdot)$ is a kernel weight function, $q$ is a bandwidth that grows with $P$, and $g_t(\theta)$ is the score vector for observation $t$ (see, for example, Andrews, 1991). For serially uncorrelated scores $q = 1$ and (11) is the same estimator suggested in Hamilton (1996) to calculate robust standard errors in a regime-switching regression.

# 3 Monte Carlo Evidence

In this section I evaluate the size and power properties of the three tests of equal predictive ability (GW, $\text{Fluct}_{t,m}$, and SD-Wald) using Monte Carlo simulation. The performance of the quasi-ML estimator of $\theta$ and the HAC estimators of the covariance matrix $\Omega$ is reported in the appendix.

## 3.1 Size properties

To evaluate the empirical size of the tests of the null hypothesis of equal predictive ability, I consider three DGPs. The first DGP considers the case of a fundamental-based

model and a random walk benchmark for first differences (with no state-dependency) and it is the same set up as in Giacomini and Rossi (2010) and Martins and Perron (2016). DGP-1 is given by

$$y_t = \beta x_t + \sigma_\varepsilon \varepsilon_t, \tag{13}$$

$$x_t = \phi x_{t-1} + \sigma_\nu \nu_t, \tag{14}$$

where $\varepsilon_t$ and $\nu_t$ are i.i.d.$N(0,1)$, with $\phi = .5$ and $\sigma_\varepsilon = \sigma_\nu = 1$. I consider one-step ahead out-of-sample forecasts from two models: (1) $y_{t+1} = u_{1t+1}$ and (2) $y_{t+1} = \beta x_{t+1} + u_{2t+1}$. As a result, the time-$t$ one-step ahead out-of-sample forecasts are

$$\begin{aligned} \hat{f}^1_{t,R} &= 0, \\ \hat{f}^2_{t,R} &= \hat{\beta}_{t,R} x_{t+1}, \end{aligned}$$

where $\hat{\beta}_{t,R}$ is the in-sample estimate of $\beta$ based on the last $R$ observations and $x_{t+1}$ is known at time $t$.

Under the null hypothesis $H_0^{(1)}$ the forecasting models are equally accurate in the finite sample. Setting $\beta = 0$, however, implies that the competing models are equally accurate in the population and, as a result, the smaller model (model 1) would be preferred in the finite sample. To obtain properly sized tests, I follow Giacomini and White (2006) and Giacomini and Rossi (2010) and select $\beta$ such that the two models have equal expected mean-squared errors

$$E\left[\left(y_{t+1} - \hat{f}^1_{t,R}\right)^2\right] = E\left[\left(y_{t+1} - \hat{f}^2_{t,R}\right)^2\right]. \tag{15}$$

Since $E[(y_{t+1} - \hat{f}^1_{t,R})^2] = \sigma^2_\varepsilon + \beta^2 x^2_{t+1}$ and $E[(y_{t+1} - \hat{f}^2_{t,R})^2] = \sigma^2_\varepsilon + \sigma^2_\varepsilon \frac{x^2_{t+1}}{\sum_{j=t-R+1}^{t} x_j^2}$, we have that $\beta^0 = \sqrt{\sigma^2_\varepsilon / \sum_{j=t-R+1}^{t} x_j^2}$. Given that $\sigma_\varepsilon = \sigma_\nu = 1$ and $\phi = 0.5$, after some

simplifications we have $\sum_{j=t-R+1}^{t} x_j^2 \approx R4/3$ and the value of $\beta$ that satisfies (15) is

$$\beta^0 \approx 1/\sqrt{R4/3}. \tag{16}$$

The second DGP considers the case of state-dependent predictive ability and is the same set up as the illustrative example of section 2.2 but with a randomly generated state variable. DGP-2 is given by

$$\begin{aligned} y_t &= -\beta s_t + \sigma_\varepsilon \varepsilon_t, & (17)\\ x_t &= \delta s_t + \sigma_\nu \nu_t, & (18) \end{aligned}$$

where $\varepsilon_t$ and $\nu_t$ are i.i.d.$N(0,1)$, $s_t$ $(= 0, 1)$ is an unobserved two-state first-order Markov process with transition probabilities $\text{Prob}(s_t = j \mid s_{t-1} = i) = p_{ij}$ for $i, j = 0, 1$. The specification is completed by assuming $\delta = 1$, $\sigma_\varepsilon = \sigma_\nu = 0.5$, and $p_{00} = p_{11} = 0.8$. As before, the two competing models are (1) $y_{t+1} = u_{1t+1}$ and (2) $y_{t+1} = \gamma x_{t+1} + u_{2t+1}$, and the time-$t$ one-step ahead out-of-sample forecasts are

$$\begin{aligned} \hat{f}_{t,R}^1 &= 0, \\ \hat{f}_{t,R}^2 &= \hat{\gamma}_{t,R} x_{t+1}, \end{aligned}$$

where $\hat{\gamma}_{t,R}$ is the in-sample rolling estimate of $\gamma$ based on the last $R$ observations and $x_{t+1}$ is known at time $t$.

Again, to obtain properly sized tests, I select $\beta$ such that the two models have equal expected mean-squared errors. Since $E[(y_{t+1} - \hat{f}_{t,R}^1)^2] = \sigma_\varepsilon^2 + \beta^2 S_{t+1}^2$ and $E[(y_{t+1} - \hat{f}_{t,R}^2)^2] = \sigma_\varepsilon^2 + \delta^2 S_{t+1}^2 \frac{\sigma_\varepsilon^2 + (\beta/\delta)^2 \sigma_\nu^2}{\sum_{j=t-R+1}^{t} x_j^2}$, after some simplifications we have that $\sum_{j=t-R+1}^{t} x_j^2 \approx R(\delta^2 p_1 + \sigma_\nu^2)$, $p_1 = \frac{1-p_{00}}{2-p_{00}-p_{11}}$, and the value of $\beta$ that satisfies (15) is

$$\beta^0 = \sqrt{\delta^2 \left[1 - \frac{\sigma_\nu^2}{R(\delta^2 p_1 + \sigma_\nu^2)}\right]^{-1} \frac{\sigma_\varepsilon^2}{R(\delta^2 p_1 + \sigma_\nu^2)}}. \tag{19}$$

Finally, the third DGP also considers the case of state-dependent predictive ability but, in this case, the state variable $s_t$ is the actual quarterly time series of NBER recession dates for the sample 1960Q1-2015Q4 (224 quarters) and equals 1 if the observation corresponds to a recession. DGP-3 is also given by (17), (18), and (19) with the unconditional probability of recession given by $p_1 = T^{-1} \sum_{t=1}^{T} s_t$.

Table 2 reports rejection frequencies of the GW, $\text{Fluct}_{t,m}$, and SD-Wald tests based on 1,000 replications. The three tests are constructed using the HAC estimator of the covariance matrix given by (11) and (12) and the Bartlett kernel of Newey and West (1987). The $\text{Fluct}_{t,m}$ test is constructed using a window size $m$ of $.3P$. The GW test exhibits small size distortions under DGP-1 and DGP-2 but is oversized under DGP-3. The $\text{Fluct}_{t,m}$ test exhibits small size distortions under all three DGPs. The SD-Wald test also exhibits small size distortions under all three DGPs as long as the hold-out sample $P$ is sufficiently large (at least 74 observations). The test is oversized when $P = 50$ under DGP-1 and DGP-2.

[ TABLE 2 ABOUT HERE ]

## 3.2 Power properties

Next, I evaluate the empirical power of the tests under the three DGPs discussed in section 3.1. To obtain the power curves, rejection frequencies are computed under the alternative hypothesis where data is generated assuming $\beta = \beta^0 + \beta^+$ with $\beta^0$ given by (16) or (19) and $\beta^+ > 0$. Under the Giacomini and White (2006) asymptotic framework, the estimation sample size $R$ must remain finite as the out-of-sample size $P$ grows to infinity. As a result, the parameters of the forecasting models need to be estimated using a rolling scheme. In practice, however, researchers appear to favor

a recursive scheme with an expanding estimation window. So, in addition to results obtained using rolling estimation, I also report power curves using recursive estimation. As before, the three tests are constructed using the HAC estimator of the covariance matrix and the Bartlett kernel. All the results are based on 1,000 replications.

DGP-1 is given by (13), (14), and (16) and under the alternative hypothesis implies a case of unequal performance of the forecasting models that is constant over time (that is, no time variation or state-dependency). In this set up, model 2 is unconditionally more accurate on average. Figure 2 reports power curves for the three tests as $\beta^+$ is increased from 0 to 1 (with .1 increments) and $R = P = 100$. When the relative performance of the models is constant over time the SD-Wald test has power that is comparable to the $\text{Fluct}_{t,m}$ test with $m = .3P$ but lower than the GW test. Higher power of the GW test is expected as this is the set up for which the GW test is proposed (see, Giacomini and White, 2006). Rolling and recursive estimation generate similar results.

[ FIGURE 2 ABOUT HERE ]

DGP-2 and DGP-3 are given by (17), (18), and (19) and under the alternative hypothesis both DGPs imply a case where the relative performance of the models varies across states. In this set up, model 1 (the smaller model) is on average more accurate when $s_t = 0$ (for example, economic expansions) but less accurate when $s_t = 1$ (for example, economic recessions). Figure 3 reports power curves for the three tests as $\beta^+$ is increased from 0 to 1 (with .1 increments) for $p_{00} = p_{11} = 0.8$ and $R = P = 100$. Relative to the results under unequal but constant performance (Figure 2), the power curves of the GW and $\text{Fluct}_{t,m}$ tests shift down. In contrast, the SD-Wald test exhibits an improvement in power.

[ FIGURE 3 ABOUT HERE ]

Recessions in the US economy tend to be short-lived, typically lasting less than four quarters, while expansions are more persistent. Therefore, to account for this difference in persistence, Figure 4 reports power curves for the three tests with $p_{00} = 0.9$, $p_{11} = 0.75$ and $R = P = 100$. This DGP implies economic expansions ($s_t = 0$) with an expected duration of ten quarters and economic recessions ($s_t = 1$) with an expected duration of four quarters.[7] In this case, the GW and $\text{Fluct}_{t,m}$ tests exhibit a substantial deterioration in power (power curves shift down). In contrast, while the SD-Wald test also exhibits a reduction in power, the test is still able to frequently reject the null hypothesis of equal predictive ability. Similarly, Figure 5 reports power curves for the three tests but, in this case, $s_t$ is the actual time series of NBER recession dates for the period 1960Q1 to 2015Q4, with $R = 100$ and $P = 124$. The power curves show that the GW and $\text{Fluct}_{t,m}$ tests are now unable to reject the null hypothesis when the superior performance of one of the two competing models is constrained to very short periods as in the case of recessions in the US economy. As a result, the GW and $\text{Fluct}_{t,m}$ tests exhibit no power. In contrast, the SD-Wald test is still able to detect the change in relative performance over the business cycle and exhibits superior power.

[ FIGURE 4 ABOUT HERE ]

[ FIGURE 5 ABOUT HERE ]

In sum, the simulation results suggest that the SD-Wald test has relatively more power when the superior performance of one of the two competing forecasting models is constrained to short periods as in the case of recessions in the US economy. In contrast,

[7] The expected duration (in quarters) of regime $j$ is computed as $1/(1 - p_{jj})$ for $j = 0, 1$.

the GW and $\text{Fluct}_{t,m}$ tests can exhibit no power at all, generally failing to reject the (incorrect) hypothesis of equal predictive ability. When the relative performance of the models is constant over time, the SD-Wald test has power that is comparable to the $\text{Fluct}_{t,m}$ test (with $m = .3P$) but lower than the GW test. Overall, the tests exhibit more power using recursive estimation than rolling estimation.

# 4 Empirical Example: Forecasting Output

Chauvet and Potter (2013) survey the recent literature on real-time forecasting of US output growth and evaluate the accuracy of forecasts obtained from many different models (linear and nonlinear models, reduced-form and structural models, survey based forecasts, etc.). Among other results, they document that most output growth forecasting models exhibit a similar performance during economic expansions but *one model performs significantly better during recessions*. In this section I use this result to illustrate the improvement of the proposed test over previous approaches to perform forecast comparisons.

## 4.1 Forecasting models and data

The out-of-sample forecasting exercise is implemented as follows. The first forecasting model, the benchmark, is a simple linear autoregressive (AR) model for output growth. Let $y_t$ be 400 times the log difference of quarterly real GDP, then the AR($k$) model for output growth is given by

$$y_{t+1} = \alpha + \sum_{i=0}^{k-1} \beta_i y_{t-i} + \varepsilon_{t+1}. \tag{20}$$

Chauvet and Potter (2013) find that this model is the most accurate during expansions (or at least as good as any of the other models considered). For recessions, on the

other hand, they find that the AR model augmented with an estimated real activity (dynamic) factor and an estimated probability of recession (AR-DF) exhibits the best performance. As a result, the alternative forecasting model is given by

$$y_{t+1} = \alpha + \sum_{i=0}^{k-1} \beta_i y_{t-i} + \delta \hat{g}_t + \gamma \hat{p}_t + \varepsilon_{t+1}, \tag{21}$$

where $\hat{g}_t$ is the estimated real activity factor and $\hat{p}_t$ is the estimated probability of recession.

Following Chauvet and Potter (2013), I estimate the factor and probabilities of recession from a set of four monthly real activity indicators previously used in Stock and Watson (1991), Diebold and Rudebusch (1996), Chauvet (1998), Chauvet and Piger (2008), Camacho et al. (2015), and Fossati (2015, 2016), among many others. This panel includes industrial production, real manufacturing sales, real personal income less transfer payments, and employment. It is assumed that the series in the panel have a factor structure of the form

$$x_{it} = \lambda_i g_t + e_{it}, \tag{22}$$

where $i = 1, \ldots, 4$, $t = 1, \ldots, T$, $g_t$ is an unobserved common factor, $\lambda_i$ is the factor loading, and $e_{it}$ is the idiosyncratic error. The dynamics of the common factor are driven by $\phi(L)g_t = \eta_t$ with $\eta_t \sim$ i.i.d.$N(0,1)$, while the dynamics of the idiosyncratic errors are driven by $\psi_i(L)e_{it} = \nu_{it}$ with $\nu_{it} \sim$ i.i.d.$N(0, \sigma_i^2)$ for $i = 1, \ldots, 4$. Identification is achieved by assuming that all shocks are independent and the specification is completed with all autoregressive processes including two lags. Next, I use a Markov-switching model to generate recession probabilities directly from the dynamic factor as in Diebold and Rudebusch (1996) and Camacho et al. (2015). Assume that the factor $\hat{g}_t$ switches between expansion and contraction regimes following a mean plus noise

specification given by

$$\hat{g}_t = \mu_{s_t} + \epsilon_t, \tag{23}$$

where $s_t$ is defined such that $s_t = 0$ during expansions and $s_t = 1$ during recessions, and $\epsilon_t \sim$ i.i.d.$N(0, \sigma_\epsilon^2)$. As usual, $s_t$ is an unobserved two-state first-order Markov process with transition probabilities given by $\text{Prob}(s_t = j \,|\, s_{t-1} = i) = p_{ij}$, with $i, j = 0, 1$. The models can be estimated by maximum likelihood following Hamilton (1989, 1990) and Kim and Nelson (1999).

Finally, we need to discuss the timing of the forecasting exercise. First, the dynamic factor model is estimated recursively, using (standardized) real-time monthly data from Camacho et al. (2015). In order to account for publication delay and important revisions that are usually observed in the first and the second release (see Chauvet and Piger, 2008), the four indicators are lagged two months. Next, the recession probabilities are estimated by fitting the Markov-switching model to the estimated dynamic factor. The AR models are estimated recursively using real-time data, a rolling window of 80 quarters, and with the lag order ($p$) selected using the Akaike information criterion. In this case, I use real-time quarterly data obtained from the Federal Reserve Bank of Philadelphia's Real Time Data Set for Macroeconomists. The initial estimation sample is from 1967:Q1 to 1986:Q4, the first one-quarter-ahead forecast is for 1987:Q1, and the last forecast corresponds to 2010:Q4 (that is, $R = 80$ and $P = 96$). As an example, in late March 1987 the econometrician would have access to the February release of real GDP with data up to December 1986. The monthly real activity indicators would be available up to January 1987 (a two-month publication delay). These series are used to estimate the models described above and to generate a one-quarter-ahead forecast for 1987:Q1, effectively a nowcast.[8]

[8]This exercise is similar but not exactly the same as the one in Chauvet and Potter (2013).

## 4.2 Results

Figure 6 shows the actual time series of real-time real GDP growth (the second release) and the one-quarter-ahead out-of-sample forecasts from the two competing models (top), as well as the forecast loss differences computed using a quadratic loss function for the hold-out sample 1987Q1-2010Q4 (bottom). As we can observe, both forecasts track the actual time series reasonably well during expansions and, as a result, loss differences are typically small and close to zero. In contrast, loss differences in recessions are generally large and positive suggesting that forecasts from the AR-DF model are more accurate during these periods. Descriptive statistics and tests of equal predictive ability are reported in Table 3. For the full hold-out sample (OOS), the average loss difference is greater than zero and positively autocorrelated. In this case, the GW test fails to reject the null hypothesis of equal predictive ability. As a result, we conclude that both models exhibit similar predictive content over the full hold-out sample. Chauvet and Potter (2013) then use peak and trough dates determined by the NBER to evaluate the performance of the two forecasting models during recession ($OOS_1$) and expansion ($OOS_0$) periods separately. In recessions, the GW test rejects the null hypothesis of equal predictive ability and we conclude that the AR-DF model is more accurate ($GW > 0$). In expansions, however, we fail to reject the null hypothesis of equal predictive ability and conclude that both models exhibit a similar performance ($GW \approx 0$).

[ FIGURE 6 ABOUT HERE ]

The main differences are: (1) Chauvet and Potter (2013) estimate the dynamic factor and recession probabilities simultaneously following Chauvet (1998) while I follow the two-step approach of Diebold and Rudebusch (1996). See also Camacho et al. (2015). (2) Chauvet and Potter (2013) estimate the AR models using a recursive scheme while I estimate the models using a rolling scheme with a fixed window of 80 quarters. (3) The hold-out sample in this paper is 1987:Q1-2010Q4 and includes the last three US recessions. In contrast, Chauvet and Potter (2013) use the sample 1992:Q1-2010Q4 and include only the last two recessions. As a result, there are some differences in the results reported.

[ TABLE 3 ABOUT HERE ]

Next, we turn our attention to the Fluctuation test of Giacomini and Rossi (2010). The $\max_t|\text{Fluct}_{t,m}|$ test statistics are reported in Table 3 and show that with a window size set to $m = .3P$ the test fails to reject the null hypothesis of equal predictive ability over the full hold-out sample. Given that recessions in the US are typically short-lived events, a smaller window size may be more appropriate. For example, setting the window size to $m = .1P$ we find that the test rejects the null hypothesis at the 5% level. Figure 7 (top) reports the time series of $\text{Fluct}_{t,m}$ test statistics and the corresponding two-sided critical values at the 5% level. Positive values of the test statistic indicate that the AR-DF model is more accurate. The results suggest that most of the time the two models exhibit a similar predictive ability ($\text{Fluct}_{t,m} \approx 0$), but there is some evidence that the AR-DF model is better during the 2007-09 recession ($\text{Fluct}_{t,m} > 0$).

[ FIGURE 7 ABOUT HERE ]

Finally, the results reported in Table 3 show that the SD-Wald test rejects the null hypothesis of equal predictive ability. The parameter estimates (standard errors in parentheses) of the test equation (5) are: $\hat{\mu}_0 = 24.72$ (6.10), $\hat{\mu}_1 = -0.06$ (0.42), $\hat{\sigma}_0^2 = 129.13$ (57.90), $\hat{\sigma}_1^2 = 9.37$ (1.70), $\hat{p}_{00} = 0.48$ (0.25), and $\hat{p}_{11} = 0.98$ (0.02). The results suggest the presence of a regime in which the AR-DF model is more accurate ($\hat{\mu}_0 > 0$) and that this regime has an expected duration of (about) two quarters. In the other regime, the two models exhibit a similar performance ($\hat{\mu}_1 \approx 0$). In addition, we observe a significant difference in variance across regimes with $\hat{\sigma}_0^2/\hat{\sigma}_1^2 \approx 13.78$. The estimated state probabilities reported in Figure 7 (bottom) show that the AR-DF

model performed better than the AR model during the 1990-91 and 2007-09 recessions, but not during the 2001 recession.

In sum, in this exercise I show that tests of the null hypothesis of equal predictive ability based on averages (unconditional or rolling) are inadequate when predictability is constrained to just a few observations. For example, the $\text{Fluct}_{t,m}$ test only rejects the null hypothesis during the longest recession in the sample and only if a very small window is used ($m = .1P$). But using such a small window typically yields oversized tests (Giacomini and Rossi, 2010). Using NBER peak and through dates and a test of unconditional predictive ability as in Chauvet and Potter (2013) I find that the AR-DF model performs significantly better during recessions. The SD-Wald test proposed in this paper is able to uncover this result without making assumptions about the state of the economy. In addition, we learn that the rejection of the null hypothesis for recessions is driven by two recessions (the 1990-91 and 2007-09 recessions) and just four observations.

# 5 Conclusion

In the macroeconomic forecasting literature, predictability has been shown to be both unstable over time and state-dependent. In this paper I show that tests of the null hypothesis of equal predictive ability based on averages (unconditional or rolling) are inadequate when the predictive content is constrained to just a few observations because the superior accuracy of a model can be averaged out. As a result, two models may appear to be equally accurate when they are not. This finding is consistent with results documented in Casini (2018). To address this issue, this paper proposed a new test for comparing the out-of-sample forecasting performance of two competing mod-

els for situations in which the predictive content may be state-dependent and of short duration. The main improvement over previous approaches to conditional forecast comparison is that the econometrician is not required to observe when the underlying states shift and, as a result, the test can be applied in situations in which the states are not observed (for example, expansion and recession states, low and high volatility states, etc.). I illustrate these results by analyzing the real-time performance of two forecasting models for US output growth discussed in Chauvet and Potter (2013).

The results discussed in this paper have implications for applied macroeconomic forecasting. For example, Gibbs and Vasnev (2017) show that if one model is found to be more accurate in one state but the other model is more accurate in the other state, a forecast combination strategy that weighs forecasts based on the predicted probability of being in each state (forward-looking weights) yields more accurate predictions of the inflation rate. The methods proposed in this paper may uncover states in which one model exhibits superior predictive ability and these results could be used to improve the overall accuracy of macroeconomic forecasts.

# Appendix

## HAC estimators and asymptotic confidence intervals

To evaluate the finite sample properties of the HAC estimators of the covariance matrix (discussed in section 2.4), sequences of loss differences $\Delta L_t$ are generated for the same parameterization of (5) considered in Hamilton (1996): $\mu_0 = -2$, $\mu_1 = 2$, $\sigma_0^2 = \sigma_1^2 = 1$, and $p_{00} = p_{11} = 0.8$. In this case, however, the error term is the MA(1) process

$$u_t = (1 + \theta^2)^{-1/2}(1 + \theta L)\varepsilon_t, \tag{24}$$

where $\varepsilon_t$ is i.i.d.$N(0, 1)$. Multiplying by $(1 + \theta^2)^{-1/2}$ normalizes the unconditional variance of $u_t$ to 1. I consider sample sizes $T = 50, 100, 250$ and parameter values $\theta = 0, 0.5, 0.9$. For each sample, quasi-ML estimates are obtained using the EM algorithm described in Hamilton (1990). Next, the covariance matrix of $\hat{\theta}$ is estimated using the Hessian (H) estimator (9), the outer-product (OP) estimator (10), and the 'sandwich' estimator (11) implemented using three common kernels: the Bartlett kernel (NW), the Parzen kernel (PK), and the quadratic spectral kernel (QS). The results reported below are based on 1,000 replications, with the same set of random errors used across values of $\theta$.

Table 4 reports the finite sample properties of the ML estimates of $\mu_0$. The bias of the estimator is calculated as the average value of $(\hat{\mu}_0^i - \mu_0)$ across the 1,000 replications (that is, $i = 1, \ldots, 1000$). The actual variation (SD) is the square root of the average value of $(\hat{\mu}_0^i - \mu_0)^2$. Table 4 also summarizes the average standard errors and the exact confidence levels of the nominal 95% confidence intervals using each of the five estimators of the covariance matrix. For $T = 50$ and $\theta = 0$, estimates of $\mu_0$ are slightly biased and the average standard errors from all five estimators slightly understate

the actual variation of $\hat{\mu}_0$. As a result, exact confidence levels are just below the nominal level (between 0.920 and 0.936). As $\theta$ is increased from 0 to 0.5 to 0.9 the actual variation of $\hat{\mu}_0$ increases while the average standard errors using the H and OP estimators remain mostly constant and, as a result, exact confidence levels drop substantially. This is expected as these commonly used estimators assume i.i.d errors. In contrast, average standard errors from the three HAC estimators increase and exact confidence levels remain (somewhat) closer to the nominal level. When the sample size is increased from 50 to 100 we observe a similar pattern but with improved confidence levels. For $T = 250$, the difference between the average standard errors from any of the three HAC estimators and the actual variation of $\hat{\mu}_0$ almost disappears and exact confidence levels are very close to the nominal 95% level.

[ TABLE 4 ABOUT HERE ]

In sum, when the model is correctly specified and no residual autocorrelation is present ($\theta = 0$), the H and OP estimators of the covariance matrix work well and generate average standard errors that accurately measure the actual variation of $\hat{\mu}_0$ and exact confidence levels that are close to the nominal level. On the other hand, when autocorrelation is present ($\theta > 0$), the H and OP estimators understate the actual variation of $\hat{\mu}_0$ and exact confidence levels can be substantially below the nominal level. In contrast, the three HAC estimators of the covariance matrix are more accurate (on average) and exact confidence levels are closer to the nominal level.

# References


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

Casini, A. (2018): "Tests for Forecast Instability and Forecast Failure under a Continuous Record Asymptotic Framework", unpublished manuscript, Department of Economics, Boston University.

Casini, A., and Perron, P. (2018): "Structural Breaks in Time Series", unpublished manuscript, Department of Economics, Boston University.

Camacho, M., Perez-Quiros, G., and Poncela, P. (2015): "Extracting Nonlinear Signals from Several Economic Indicators", *Journal of Applied Econometrics*, 30, 1073–1089.

Chauvet, M. (1998): "An Econometric Characterization of Business Cycle Dynamics with Factor Structure and Regime Switches", *International Economic Review*, 39(4), 969–996.

Chauvet, M., and Piger, J. (2008): "A Comparison of the Real-Time Performance of Business Cycle Dating Methods", *Journal of Business and Economic Statistics*, 26, 42–49.

Chauvet, M., and Potter, S. (2013): "Forecasting output", *Handbook of Forecasting*, 81, 608–616.

Clark, T., and McCracken M.W. (2011): "Testing for unconditional predictive ability", Oxford Handbook of Economic Forecasting, ed. M. Clements and D. Hendry, Oxford University Press.

Clark, T., and West K.D. (2006): "Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis", *Journal Econometrics*, 135, 155–186.

Dangl, T., and Halling, M. (2012): "Predictive regressions with time-varying coefficients", *Journal of Financial Economics*, 106, 157–181.

Diebold, F.X., and Mariano, R.S. (1995): "Comparing predictive accuracy", *Journal of Business and Economic Statistics*, 13, 253–263.

Diebold, F., and Rudebusch, G. (1996): "Measuring Business Cycles: A Modern Perspective", *Review of Economics and Statistics*, 78, 66–77.

Dotsey, M., Fujita, S., and Stark, T. (2015): "Do Phillips curves conditionally help to forecast inflation?", Working Paper 15-16, Federal Reserve Bank of Philadelphia.

Fossati, S. (2015): "Forecasting US Recessions with Macro Factors", *Applied Economics*, 47, 5726–5738.

Fossati, S. (2016): "Dating US Business Cycles with Macro Factors", *Studies in Nonlinear Dynamics & Econometrics*, 20, 529–547.

Gargano, A., Pettenuzzo, D., and Timmermann, A. (2016): "Bond return predictability: economic value and links to the macroeconomy", *Management Science*, forthcoming.

Gargano, A., and Timmermann, A. (2014): "Forecasting commodity price indexes using macroeconomic and financial predictors", *International Journal of Forecasting*, 30, 825–843.

Giacomini, R., and White, H. (2006): "Tests for conditional predictive ability", *Econometrica*, 74, 1545–1578.

Giacomini, R., and Rossi, B. (2010): "Forecast comparison in unstable environments", *Journal of Applied Econometrics*, 25, 595–620.

Giacomini, R. (2011): "Testing conditional predictive ability", Oxford Handbook of Economic Forecasting, ed. M. Clements and D. Hendry, Oxford University Press.

Gibbs, C.G., and Vasnev, A.L. (2017): "Conditionally optimal weights and forward-looking approaches to combining forecasts", Discussion Papers 2017-10, School of Economics, The University of New South Wales.

Granziera, E., and Sekhposyan, T. (2017): "How to Predict Your Next Forecasting Model: Conditional Predictive Ability Approach", unpublished manuscript.

Hamilton, J.D. (1989): "A new approach to the economic analysis of nonstationary time series and the business cycle", *Econometrica*, 57, 357–384.

Hamilton, J.D. (1990): "Analysis of time series subject to changes in regime", *Journal of Econometrics*, 45, 39–70.

Hamilton, J.D. (1996): "Specification testing in Markov-switching time-series models", *Journal of Econometrics*, 70, 127–157.

Hayashi, F. (2000): *Econometrics*, Princeton University Press, Princeton, New Jersey.

Henkel, S.J., Martin, J.S., and Nardari, F. (2011): "Time-varying short-horizon predictability", *Journal of Financial Economics*, 99, 560–580.

Kim, C.J. (1994): "Dynamic linear models with Markov-switching", *Journal of Econometrics* , 60, 1–22.

Kim, C.J., and Nelson, C.R. (1999): *State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications*, The MIT Press.

Martins, L.F., and Perron, P. (2016): "Improved tests for forecast comparisons in the presence of instabilities", *Journal of Time Series Analysis*, 37, 650–659.

Morley, J., and Rabah, Z. (2014): "Testing for a Markov-switching mean in serially correlated data", in J. Ma and M. Wohar (eds.), Recent Advances in Estimating Nonlinear Models (Springer, Berlin, 2014, 85–97).

Newey, W.K., and West, K.D. (1987): "A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix", *Econometrica*, 55, 703–708.

Rapach, D.E., Strauss, J.K., and Zhou, G. (2010): "Out-of-sample equity premium prediction: combination forecasts and links to the real economy", *The Review of Financial Studies*, 23(2), 821–862.

Rossi, B. (2013): "Advances in forecasting under instability", *Handbook of Economic Forecasting*, Volume 2, Part B, 1203–1324.

Rossi, B., and Sekhposyan, T. (2010): "Have economic models' forecasting performance for US output growth and inflation changed over time, and when?", *International Journal of Forecasting*, 26, 808–835.

Stock, J.H., and Watson, M.W. (1991): "A Probability Model of the Coincident Economic Indicators", in *Leading Economic Indicators: New Approaches and Forecasting Records*, edited by K. Lahiri and G, Moore, Cambridge University Press.

Stock, J.H., and Watson, M.W. (2003): "Forecasting output and inflation: The role of asset prices", *Journal of Economic Literature*, volume XLI, 788–829.

West, K.D. (1996): "Asymptotic inference about predictive ability", *Econometrica*, 64, 1067–1084.

Table 1: Loss difference statistics for a simulated time series

| | OOS | $OOS_0$ | $OOS_1$ |
|---|---|---|---|
| Observations | 124 | 110 | 14 |
| Average | 0.092 | -0.179 | 2.217 |
| Standard Dev. | 0.969 | 0.492 | 1.169 |
| AR(1) | 0.407** | -0.137 | -0.005 |
| GW | 1.056 | -3.809** | 7.096** |
| GW(HAC) | 0.683 | | |
| Fluctuation(HAC,$m = .1P$) | 2.704 | | |
| Fluctuation(HAC,$m = .3P$) | 2.568 | | |

Notes: Significance of the AR(1) coefficients is tested based on the asymptotic result $\sqrt{T}\hat{\rho} \xrightarrow{d} N(0,1)$. HAC tests constructed using the Bartlett kernel of Newey and West (1987). The 5% (10%) critical value is 1.96 (1.645) for a two-sided GW test, 3.012 (2.766) for a two-sided Fluctuation test with a rolling window size of $m = .3P$, and 3.393 (3.170) for $m = .1P$. ** (*) denotes rejection of the null hypothesis at the 5% (10%) level.

Table 2: Empirical size under quadratic loss

| $R$ | $P$ | GW | Fluct | SD-W | GW | Fluct | SD-W |
|---|---|---|---|---|---|---|---|
| | | DGP-1 | | | DGP-2 | | |
| 50 | 50 | 0.044 | 0.043 | 0.084 | 0.053 | 0.056 | 0.093 |
| | 100 | 0.055 | 0.048 | 0.044 | 0.057 | 0.040 | 0.045 |
| | 250 | 0.031 | 0.037 | 0.034 | 0.076 | 0.044 | 0.055 |
| 100 | 50 | 0.059 | 0.043 | 0.103 | 0.047 | 0.053 | 0.097 |
| | 100 | 0.038 | 0.045 | 0.053 | 0.067 | 0.062 | 0.049 |
| | 250 | 0.038 | 0.051 | 0.030 | 0.064 | 0.030 | 0.046 |
| 250 | 50 | 0.054 | 0.052 | 0.118 | 0.075 | 0.068 | 0.114 |
| | 100 | 0.061 | 0.065 | 0.056 | 0.052 | 0.063 | 0.052 |
| | 250 | 0.047 | 0.071 | 0.041 | 0.073 | 0.062 | 0.045 |
| | | | | | DGP-3 | | |
| 50 | 174 | | | | 0.104 | 0.053 | 0.056 |
| 100 | 124 | | | | 0.081 | 0.061 | 0.051 |
| 150 | 74 | | | | 0.058 | 0.065 | 0.062 |

Notes: The $\text{Fluct}_{t,m}$ test constructed using a window size of $.3P$. Nominal size 0.05.

Table 3: Loss difference statistics for real GDP growth

| | OOS | $OOS_0$ | $OOS_1$ |
|---|---|---|---|
| Observations | 96 | 82 | 14 |
| Average | 1.032 | -0.159 | 8.005 |
| Standard Dev. | 6.388 | 3.019 | 13.422 |
| AR(1) | 0.279** | -0.136 | 0.121 |
| GW | 1.582 | -0.477 | 2.232** |
| GW(HAC) | 1.360 | | |
| Fluctuation(HAC,$m = .1P$) | 3.517** | | |
| Fluctuation(HAC,$m = .3P$) | 2.232 | | |
| SD-Wald(HAC) | 16.484** | | |

Notes: Significance of the AR(1) coefficients is tested based on the asymptotic result $\sqrt{T}\hat{\rho}_1 \xrightarrow{d} N(0,1)$. HAC tests constructed using the Bartlett kernel of Newey and West (1987). The 5% (10%) critical value is 1.96 (1.645) for a two-sided GW test, 3.012 (2.766) for a two-sided Fluctuation test with a rolling window size of $m = .3P$, and 3.393 (3.170) for $m = .1P$. The 5% (10%) critical value for a SD-Wald test is 5.99 (4.61). ** (*) denotes rejection of the null hypothesis at the 5% (10%) level.

Table 4: Small sample properties of the ML estimates of $\mu_0$

| | $\theta$ | Bias | SD | H | OP | NW | PK | QS |
|---|---|---|---|---|---|---|---|---|
| $T = 50$ | 0 | 0.008 | 0.256 | 0.223 | 0.235 | 0.247 | 0.243 | 0.245 |
| | | | | 0.920 | 0.929 | 0.936 | 0.920 | 0.928 |
| | 0.5 | 0.007 | 0.314 | 0.219 | 0.231 | 0.266 | 0.275 | 0.276 |
| | | | | 0.830 | 0.841 | 0.875 | 0.886 | 0.887 |
| | 0.9 | 0.025 | 0.345 | 0.226 | 0.228 | 0.291 | 0.301 | 0.302 |
| | | | | 0.798 | 0.810 | 0.867 | 0.868 | 0.871 |
| $T = 100$ | 0 | 0.006 | 0.159 | 0.154 | 0.156 | 0.163 | 0.162 | 0.163 |
| | | | | 0.933 | 0.938 | 0.946 | 0.942 | 0.945 |
| | 0.5 | 0.015 | 0.220 | 0.161 | 0.156 | 0.202 | 0.210 | 0.210 |
| | | | | 0.846 | 0.838 | 0.896 | 0.909 | 0.910 |
| | 0.9 | 0.010 | 0.233 | 0.163 | 0.157 | 0.211 | 0.219 | 0.220 |
| | | | | 0.824 | 0.817 | 0.904 | 0.908 | 0.913 |
| $T = 250$ | 0 | -0.004 | 0.096 | 0.098 | 0.099 | 0.101 | 0.100 | 0.101 |
| | | | | 0.962 | 0.959 | 0.963 | 0.961 | 0.963 |
| | 0.5 | 0.006 | 0.128 | 0.101 | 0.098 | 0.120 | 0.125 | 0.125 |
| | | | | 0.869 | 0.856 | 0.925 | 0.939 | 0.937 |
| | 0.9 | -0.001 | 0.135 | 0.102 | 0.098 | 0.129 | 0.134 | 0.134 |
| | | | | 0.874 | 0.856 | 0.932 | 0.942 | 0.943 |

Notes: The true model is $\Delta L_t(\hat{\delta}_{t-h,R}, \hat{\gamma}_{t-h,R}) = \mu_{s_t} + \sigma_{s_t} u_t$ with $s_t = 0, 1$, $\mu_0 = -2$, $\mu_1 = 2$, $\sigma_0^2 = \sigma_1^2 = 1$, and $p_{00} = p_{11} = 0.8$. The error term is an MA(1) process $u_t = (1+\theta^2)^{-1/2}(1+\theta L)\varepsilon_t$ where $\varepsilon_t \sim$ i.i.d.$N(0,1)$. The covariance matrix estimators are: Hessian (H), outer-product (OP), Bartlett (NW), Parzen (PK), and quadratic spectral kernel (QS). Bias is the average value of $(\hat{\mu}_0^i - \mu_0)$. SD is the square root of the average value of $(\hat{\mu}_0^i - \mu_0)^2$.

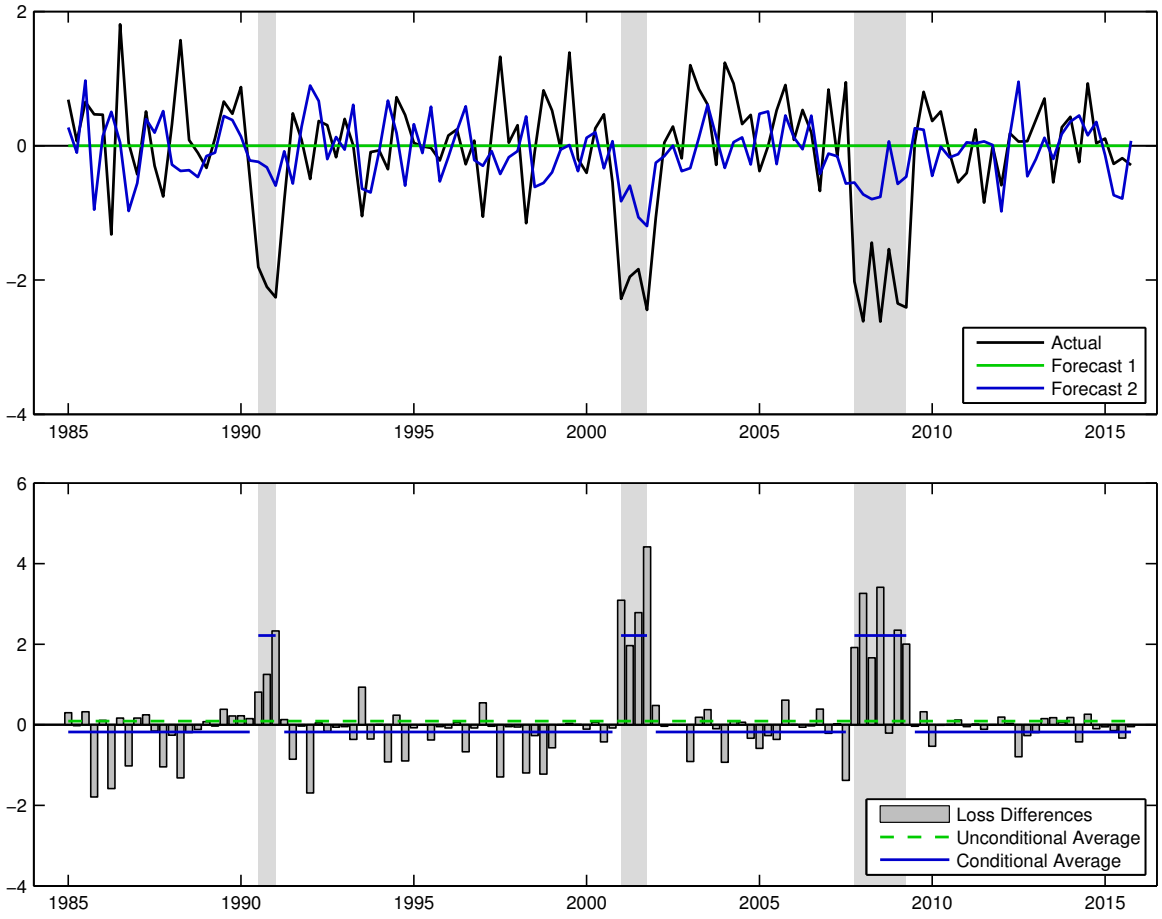


Figure 1: Shows the simulated time series and forecasts from the two competing models (top) and loss differences with sample averages (bottom). The two forecasting models are: (i) $\hat{f}^1_{t,R} = 0$ and (ii) $\hat{f}^2_{t,R} = \hat{\gamma}_{t,R} x_{t+1}$. Shaded areas denote NBER recessions.

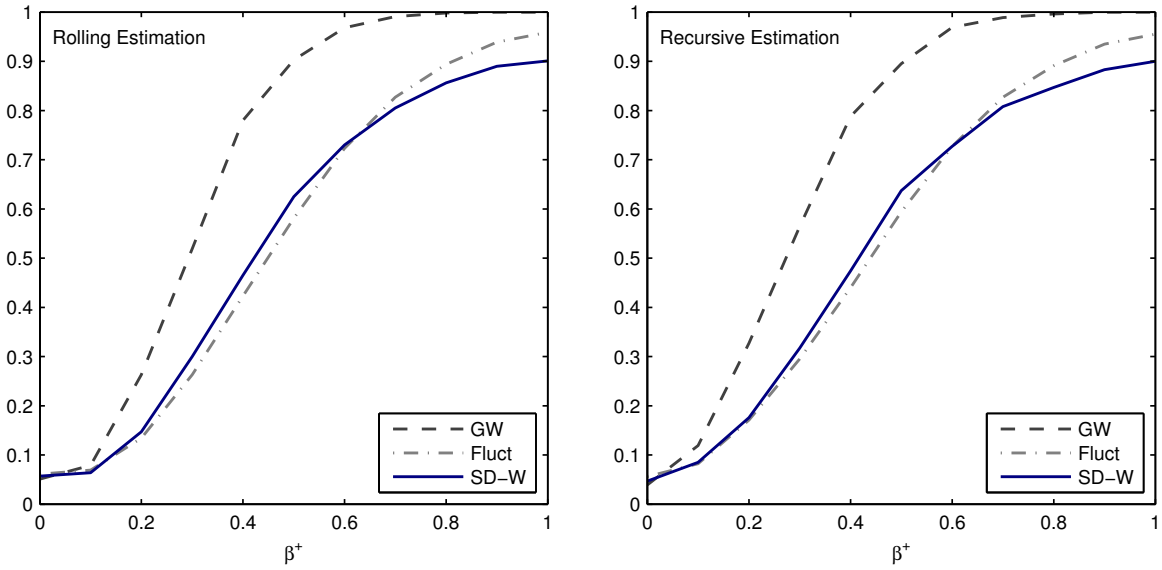


Figure 2: Empirical power (rejection frequencies) of the GW, $\text{Fluct}_{t,m}$ ($m = .3P$), and SD-Wald tests in the case of unequal but constant relative forecasting performance of the two models with $R = P = 100$.

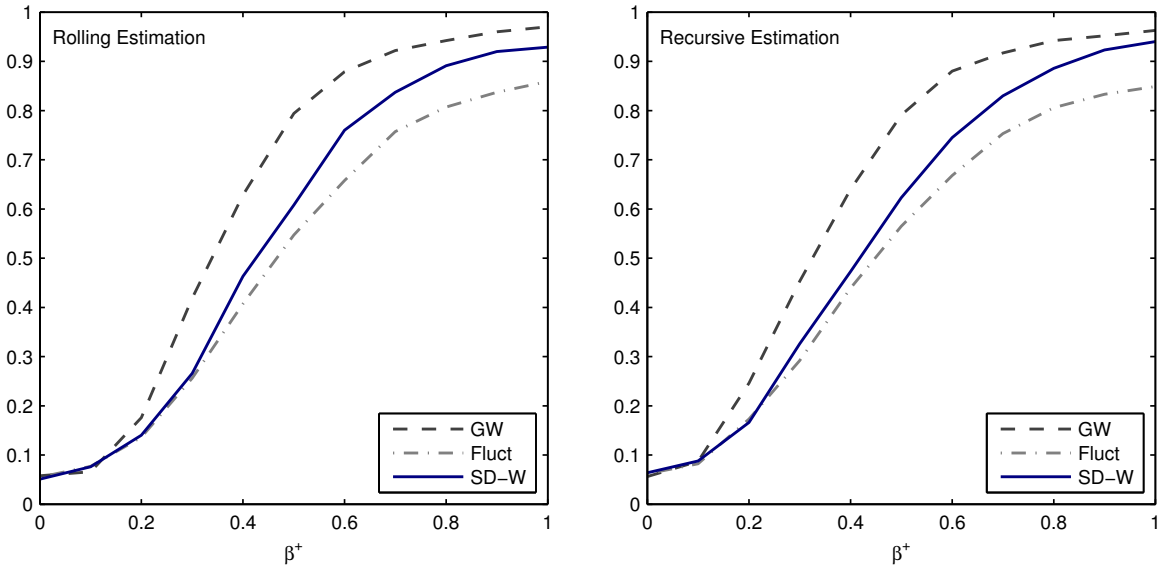


Figure 3: Empirical power (rejection frequencies) of the GW, $\text{Fluct}_{t,m}$ ($m = .3P$), and SD-Wald tests in the case of different relative forecasting performance of the two models in different states with $p_{00} = p_{11} = 0.8$ and $R = P = 100$.

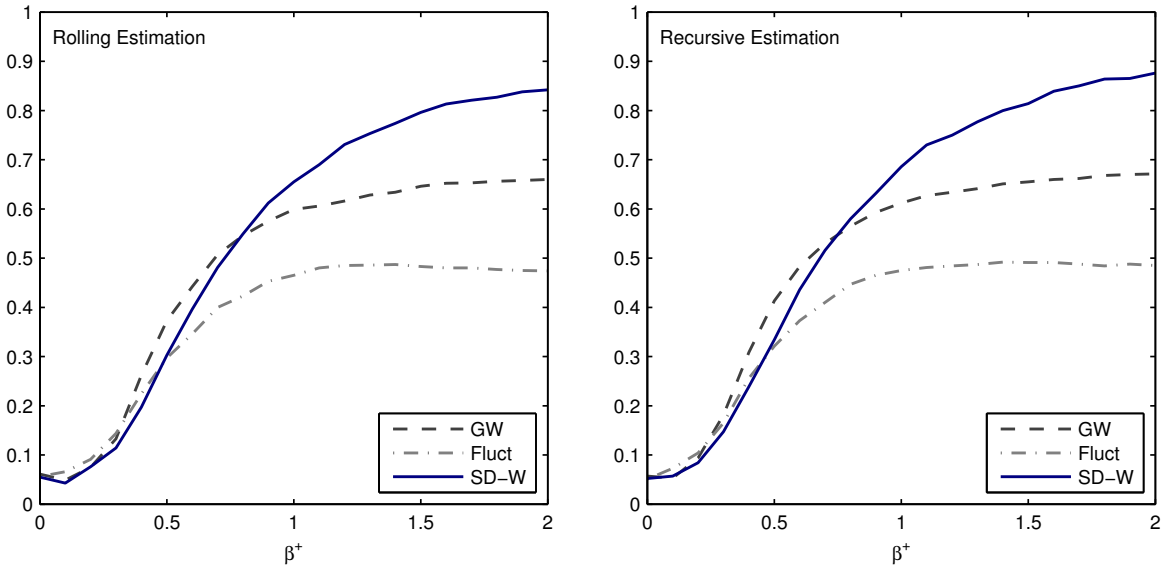


Figure 4: Empirical power (rejection frequencies) of the GW, $\text{Fluct}_{t,m}$ ($m = .3P$), and SD-Wald tests in the case of different relative forecasting performance of the two models in different states with $p_{00} = 0.90$, $p_{11} = 0.75$, and $R = P = 100$.

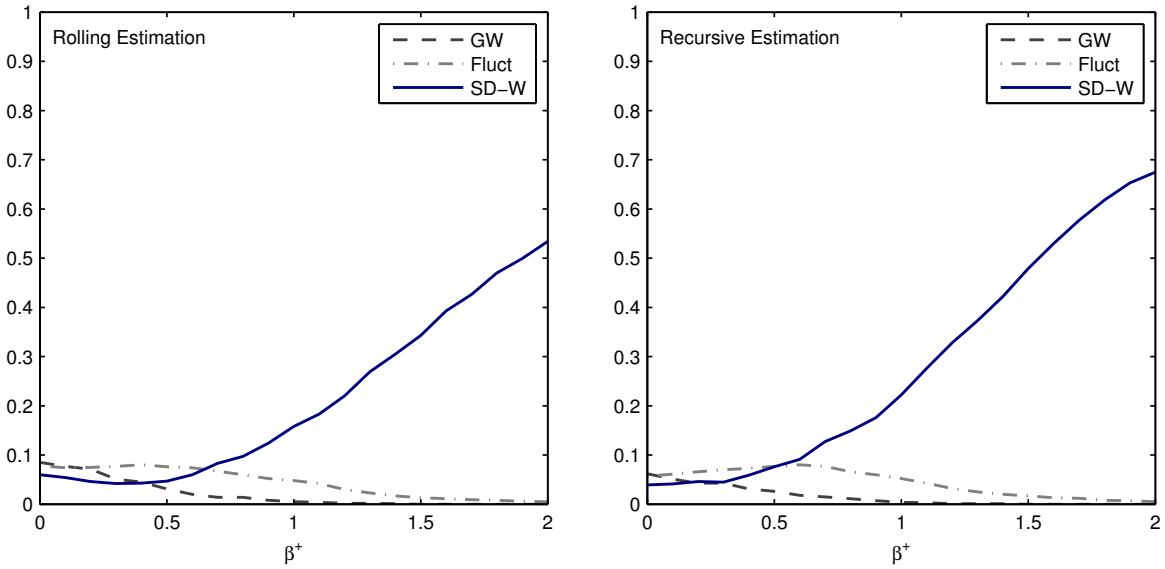


Figure 5: Empirical power (rejection frequencies) of the GW, $\text{Fluct}_{t,m}$ ($m = .3P$), and SD-Wald tests in the case of different relative forecasting performance of the two models in different states with $s_t$ the actual time series of NBER recession dates for the period 1960Q1 to 2015Q4, $R = 100$, and $P = 124$.

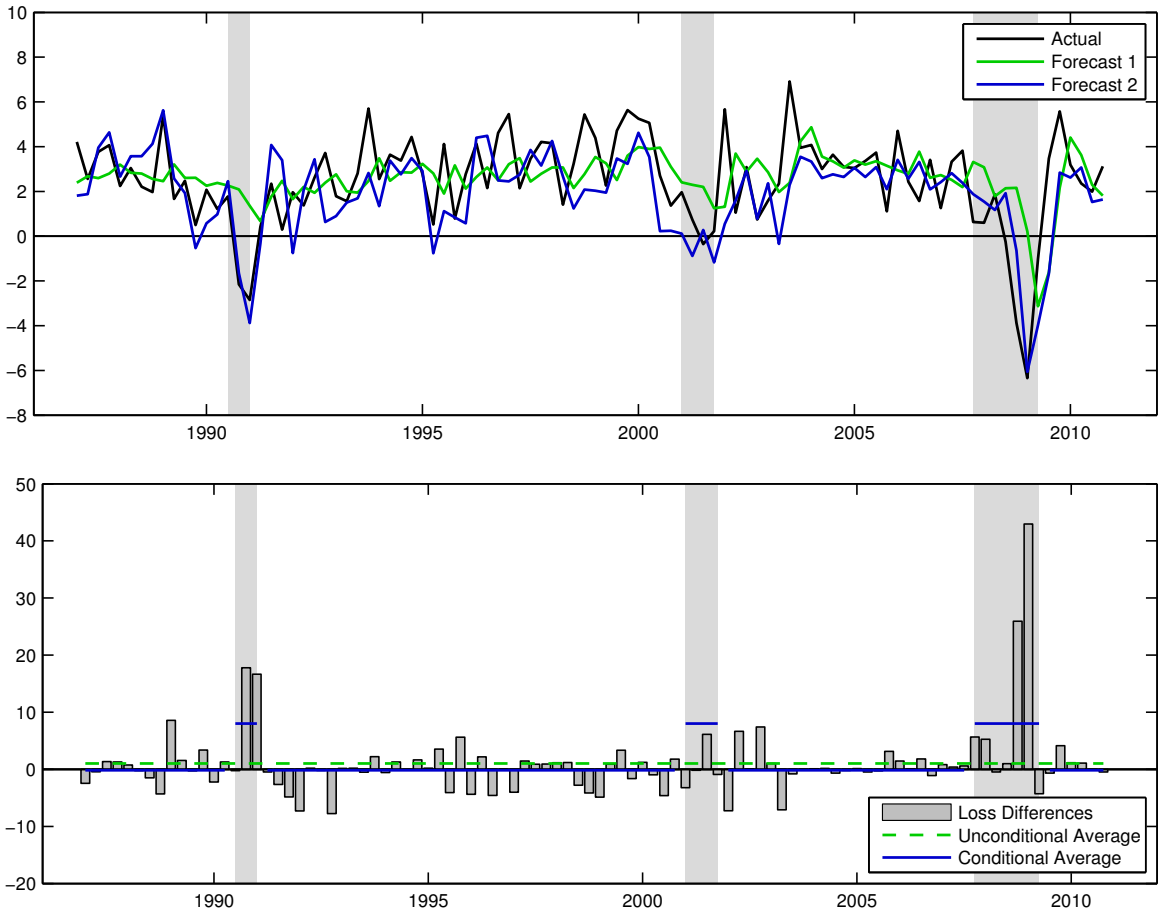


Figure 6: Shows real-time GDP growth and forecasts from the two competing models (top) and loss differences with sample averages (bottom). The two forecasting models are: (i) $\hat{f}^1_{t,R}$ = AR and (ii) $\hat{f}^2_{t,R}$ = AR-DF. Shaded areas denote NBER recessions.

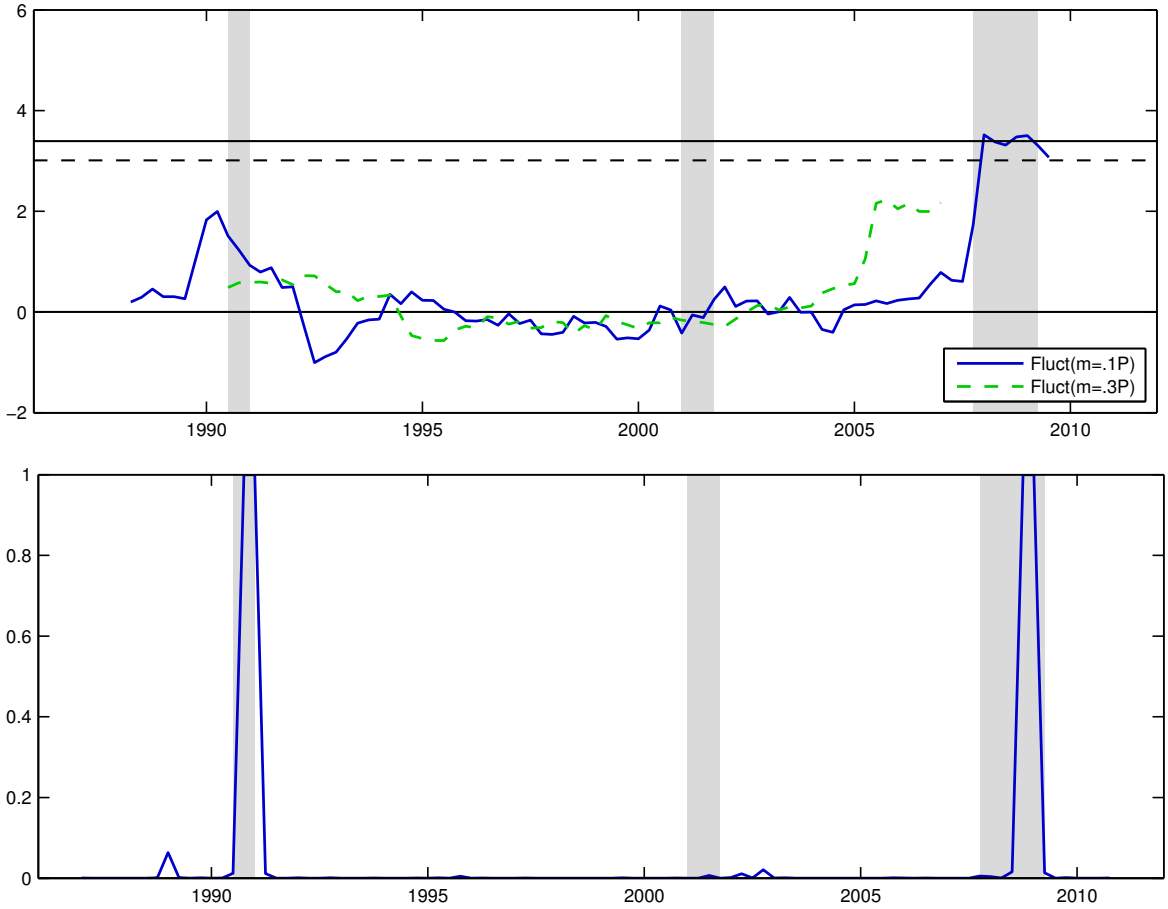


Figure 7: Shows the $\text{Fluct}_{t,m}$ test statistics for $m = .1P$ and $m = .3P$ with the corresponding critical values (top) and the state probabilities from the SD-Wald test (bottom). Shaded areas denote NBER recessions.