Markov Switching Rationality

Florens Odendahl* Barbara Rossi† Tatevik Sekhposyan‡

October 31, 2021

Abstract

We propose novel tests for the detection of Markov switching deviations from forecast rationality. Existing forecast rationality tests either focus on constant deviations from forecast rationality over the full sample or are constructed to detect smooth deviations based on non-parametric techniques. In contrast, our proposed tests are parametric and have an advantage in detecting abrupt departures from unbiasedness and efficiency, which we demonstrate with Monte Carlo simulations. Using the proposed tests, we investigate whether Blue Chip Financial Forecasts for the Federal Funds Rate are unbiased. Our tests find evidence of a state-dependent bias: forecasters tend to systematically overpredict interest rates during periods of monetary easing, while the forecasts are unbiased otherwise. We show that a similar state-dependent bias is also present in market-based forecasts of interest rates, but not in the forecasts of real GDP growth and GDP deflator-based inflation. Our results emphasize the special role played by monetary policy in shaping survey interest rate expectations above and beyond macroeconomic fundamentals.

*Banco de España, Address: Calle de Alcalá 48, 28014 Madrid, Spain. Email: florens.odendahl@bde.es.

†ICREA-Universitat Pompeu Fabra, Barcelona School of Economics and CREI, c/Ramon Trias Fargas 25/27, Barcelona 08005, Spain; Email: barbara.rossi@upf.edu.

‡Texas A&M University, 4228 TAMU, College Station, TX 77843, USA; Email: tatevik.sekhposyan@gmail.com.

We thank the editor Yoosoon Chang and the three anonymous referees for their constructive comments and suggestions. Part of this research was carried out while Tatevik Sekhposyan was a Visiting Fellow at the Federal Reserve Bank of San Francisco, whose hospitality is greatly acknowledged. The views expressed are those of the authors and do not necessarily reflect the views of the Banco de España, the Eurosystem, the Federal Reserve Bank of San Francisco or the Board of Governors of the Federal Reserve System. Barbara Rossi acknowledges Financial support from the Spanish Ministry of the Economy and Competitiveness and from the Spanish Agencia Estatal de Investigación, through the Severo Ochoa Programme for Centres of Excellence in R&D (CEX2019-000915-S).

1 Introduction

Producing accurate forecasts for economic variables is an important task for both researchers and policymakers alike. A desirable property that forecasts should have is optimality. When evaluating forecasts using a quadratic loss, forecast optimality implies that the forecast error should not be predictable by a constant (unbiasedness), the forecast itself (efficiency), or any information available at the time the forecast is made; otherwise, it is reasonable to conclude that the forecast is suboptimal and can be improved.

However, it is well known that forecasting performance can be unstable over time, and changes in the forecast's quality may be associated with recurring periods of economic importance. For instance, Joutz and Stekler (2000) find that the Federal Reserve Board's (Fed) forecasts overestimated output growth in slowdowns and recessions and underestimated it in recoveries. Similarly, inflation is typically underpredicted when it is rising and overpredicted when it is declining. Granziera et al. (2021) reach similar conclusions for the European Central Bank's (ECB) inflation forecasts: the ECB tends to overpredict (underpredict) inflation when inflation is below (above) target. Sinclair et al. (2010) show that information on real and inflationary cycles, though incorporated in the Fed's nowcast, are not incorporated into one-quarter-ahead forecasts.

Evaluating forecasts that may have state-dependent forecast rationality (unbiasedness and efficiency) properties requires particular care since standard tests of absolute and relative forecast evaluations are misleading in the presence of instabilities (Rossi, 2013). As a consequence, the literature on the evaluation of the absolute (and relative) performance of forecasting models has developed techniques to robustify inference, where instabilities are accounted for non-parametrically; see for instance Rossi and Sekhposyan (2016).

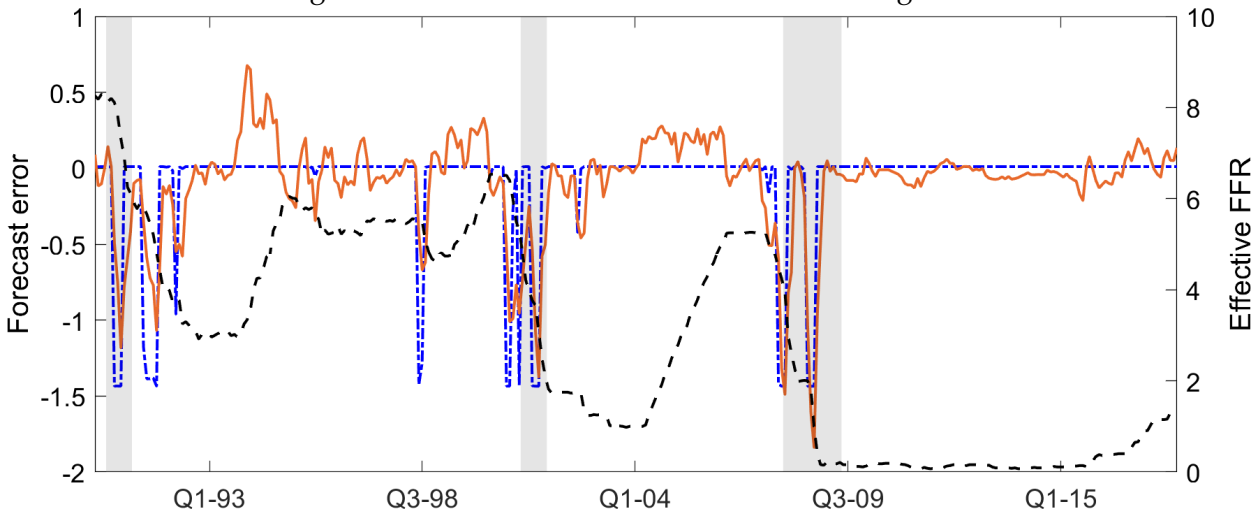
We contribute to this literature by proposing two novel tests that are robust to the presence of time variation. The novelty is that we assume a parametric form of time variation driven by unobserved states; namely, we assume that the forecast errors follow a Markov switching process. Consequently, our tests are more powerful than non-parametric tests of forecast rationality when the forecasting performance varies over time in a regime-switching manner. Our approach is relevant in situations where the rationality of the forecasts depends on exogenous, unobserved (or *a priori* unknown but observable) cycles. An interesting extension would be to adapt the tests to models of endogenous regime-switching, where the regime depends on whether a latent factor takes a value above or below some threshold (Chang et al., 2017).

To demonstrate the empirical relevance of our approach consider Figure 1, which displays the

forecast errors of three-month-ahead (effective) Federal Funds Rate (FFR) forecasts from the Blue Chip Financial Forecasts (BCFF) survey; the forecast error is measured by the difference between the realization and the forecast. The figure also depicts our estimated Markov switching forecast error mean (dashed-dotted line).[1] The estimation results imply deviations from unbiasedness during periods of monetary easing (when interest rates, depicted by the dashed black line, are decreasing); this state dependence was identified by the Markov switching model and did not require *ex-ante* specification of the state variable.

Importantly, our proposed tests can uncover recurring periods of deviations from forecast rationality, even when traditional tests do not reject, on average, over the full sample. Deviations from forecast rationality could be linked to important economic events, such as recessions, financial distress, or other economic circumstances. For instance, Bullard (2016) announced that the St. Louis Fed has abandoned the view of the economy having a single steady-state in favor of a regime-switching world with several steady states. Alternatively, occasional deviations from forecast rationality could be potentially related to information frictions or the way agents learn at different times. Our procedure allows researchers to treat the periods when the absolute performance varies as an unknown state variable, identify the periods when the forecast was biased as well as quantify the bias.

Figure 1: Forecast errors with Markov switching bias



Note: The solid line (left-hand side y-axis) shows the forecast errors: the difference between the FFR realizations and the respective BCFF's three-month-ahead forecasts. The dashed-dotted line (left-hand side y-axis) shows a regime-switching unconditional mean estimated using the smoothed state probabilities implied by a two-state Markov switching model; see Section 5 for the description of the model. The dashed line (right-hand side y-axis) shows the FFR level. Grey shaded areas display NBER recession periods.

In this paper, we propose forecast rationality tests that build on Hansen (1992)'s tests for detecting Markov switching and on the bootstrap procedure proposed in Qu and Zhuo (2021) (we

[1] See Section 5 for more details.

also consider extensions of Garcia (1998) in the Online Appendix). Testing for Markov switching requires non-standard inference because of several problems. First, the hyper-parameters of the switching process (for instance, the state-to-state transition probabilities) are not identified under the null hypothesis of parameter stability (Davies, 1977, 1987). Second, under the null of parameter stability, the score with respect to the restricted parameters is identically zero, which violates standard regularity conditions imposed to derive an asymptotic chi-squared distribution via a usual second-order Taylor expansion. Consequently, standard Likelihood ratio (LR), Wald, and Lagrange multiplier tests do not have a chi-squared distribution, even asymptotically. Third, the conditional regime probabilities follow a stochastic process that can only be represented recursively, thus making higher-order Taylor approximations infeasible. Furthermore, there are multiple ways to impose the null of a single state, which further complicates inference, in addition to creating a boundary parameter problem for the null parameter space. Hansen (1992), Garcia (1998), Cho and White (2007), Carter and Steigerwald (2012), and Qu and Zhuo (2021) discuss these issues in detail and make significant contributions, thus shaping our knowledge on how to test for the number of regimes in Markov-switching models.

Our paper builds on the existing literature and proposes a Markov switching test in a forecast rationality framework, where we impose a joint null hypothesis that there is a single regime and that the relevant parameters are restricted to zero under the null hypothesis. More specifically, we rely on Hansen (1992) who treated the likelihood function as a stochastic process and obtained a lower bound for the likelihood ratio test. In addition, we use the bootstrap procedure of Qu and Zhuo (2021) to test our null. The bootstrap addresses several of the difficulties associated with testing for Markov switching outlined above and performs well in finite samples. Qu and Zhuo's (2021) bootstrap, though building on Cho and White (2007), does not explicitly address the boundary parameter issue of the composite null that the latter focused on, even though they analyze situations when the transition probabilities are close to the boundary.

In particular, we adapt these tests to our absolute forecast evaluation context and refer to them as the "absolute forecast evaluation - Hansen" (AFE-H) and the "bootstrap" (AFE-BS). In addition, the Online Appendix also investigates the "absolute forecast evaluation - Garcia" (AFE-G) test inspired by Garcia (1998). The main difference between our proposed tests relative to the ones in the literature is that Hansen (1992) and Qu and Zhuo (2021) test for Markov switching in the parameters of a model, leaving the values of the parameters unspecified under the null. In contrast, we test for Markov switching directly in the forecast errors by specifying parameter

values such that forecast rationality is satisfied under the null. That is, we test for forecast rationality in the full out-of-sample portion of the data against local, regime-switching deviations. More specifically, consider a standard forecast unbiasedness test (Mincer and Zarnowitz, 1969), which evaluates whether the forecast error has a zero mean against the alternative that the mean differs from zero. Instead, under the alternative of our approach, we let the forecast error evolve according to a Markov switching process, and jointly test that the mean of the forecast error is time-invariant and equal to zero.

Aside from the literature on testing for Markov switching, we also relate to a large literature on the evaluation of the absolute predictive performance (Mincer and Zarnowitz, 1969; West and McCracken, 1998; Rossi and Sekhposyan, 2016). Mincer and Zarnowitz (1969) and West and McCracken (1998) assume a constant mean of the forecast error and a constant efficiency parameter, an assumption that is violated in the presence of time variation. Rossi and Sekhposyan (2016) propose a non-parametric test for forecast rationality that is robust to instabilities, which is useful in the situations where rationality is time-varying yet, on average, holds in the full sample. Given the non-parametric nature, their test performs well when time-variation is smooth and persistent, while our proposed tests have stronger power in the presence of abrupt, short-lived and recurring deviations from rationality. In fact, our tests can detect deviations from rationality that occur for short periods of time, as long as the deviations occur repeatedly.[2] Note that in our work we focus on model-free or survey-based forecasts, as well as forecasts obtained either with a rolling window with finite size or a recursive window where the contribution of parameter estimation error can be reasonably ignored (for instance, when the estimation sample size is relatively large compared to the evaluation sample size).

We investigate the finite sample properties of our proposed tests with Monte Carlo simulations. The simulations show empirical rejection frequencies close to the nominal size when testing the null hypothesis of unbiasedness and efficiency using the AFE-BS test. For unbiasedness, the AFE-H is well sized in medium to large samples and somewhat undersized when testing for efficiency. In terms of power, the rejection frequencies are similar to the test of West and McCracken (1998) and the Fluctuation rationality test of Rossi and Sekhposyan (2016) for the alternative of a constant deviation from rationality, and clearly outperform both under a Markov switching alternative.

Turning to the empirical analysis, we investigate potential biases in the BCFF survey pre-

[2] In a two-state model, the expected duration of regime j is given by $1/(1-p_{jj})$, where p_{jj} is the state-to-state transition probability of regime j. Therefore, for instance, the expected duration of a state with a state-to-state transition probability as high as 90% is only ten periods.

dictions for the FFR. When we consider the three-month- and six-month-ahead forecast errors, our test rejects unbiasedness in favor of a two-regime model. The estimated regimes indicate that the forecasts are unbiased in the first regime, the one that is prevalent most of the time. However, there is evidence of a second regime in which the forecasters overestimate the FFR. The occurrence of the second regime is associated with monetary policy easing and is not limited to recessionary periods. The biases are present not only in survey forecasts but also in market-based forecasts, suggesting that the lack of forecast rationality is not specific to the survey but inherent to difficulty in forecasting. We investigate the role of disagreement among panelists as well as the role of monetary policy uncertainty based on newspaper articles (Baker et al., 2016) in explaining this regime-dependent behavior. We find no clear association with disagreement, while the regimes appear to be weakly associated with monetary policy uncertainty. Our findings on state-dependent biases can be used to improve the forecasts; for instance, by adjusting for a bias only in monetary easing episodes.

The paper is organized as follows. Section 2 introduces the econometric framework and formalizes the null hypothesis. Section 3 introduces the proposed test statistics. Section 4 provides a Monte Carlo analysis of the size and power of our proposed procedures, while Section 5 illustrates the usefulness of our test in an empirical analysis. Section 6 concludes.

2 Econometric framework

We consider the situation where the researcher has a series of out-of-sample predictions, $y_{t,h}$, made at time t, h-periods into the future, whose corresponding realizations are denoted by y_{t+h}, for $t = 1, ..., T$. Let $\epsilon_{t,h} \equiv y_{t+h} - y_{t,h}$ denote the forecast error. We are interested in testing whether the forecast error is unbiased, efficient, and rational (i.e. jointly unbiased and efficient) — while being able to detect regime-switching deviations from the respective forecast rationality property. For simplicity, consider the leading case of a forecast rationality regression with two regimes:

$$\epsilon_{t,h} = \beta x_{t+h} + S_{t+h}\beta_s x_{t+h} + \sum_{i=1}^{d} \phi_i \epsilon_{t-i,h} + e_{t+h}, \tag{1}$$

where S_{t+h} is a latent, stationary Markov chain with $S_{t+h} \in \{0, 1\}$; e_{t+h} is a mean zero error term with a constant variance; $x_{t+h} = [1, y_{t,h}]'$; and ϕ_i are the lag coefficients included to control for potential autocorrelation.[3] The stationary Markov chain S_{t+h} is characterized by the two

[3]The constant variance assumption could be relaxed by allowing for conditional heteroskedasticity. More specifically, the case where the variance follows a Markov switching process is discussed in Section 4.3.

state-to-state transition probabilities (p, q), which take values between zero and one. The vectors $\beta = (\mu, \gamma)$ and $\beta_s = (\mu_s, \gamma_s)$ contain the relevant parameters for rationality regressions. The parameters μ and μ_s are the relevant parameters for the unbiasedness test, while γ and γ_s (the regression coefficients on the forecasts) are the relevant parameters for the efficiency test. Note that the vector x_{t+h} can be extended to contain more regressors, making our test applicable, in general, to all regression-based tests of predictive ability. For notational simplicity, we drop the autoregressive coefficients (ϕ_i) in what follows. This simplification is inconsequential since they are neither parameters of interest nor state-dependent.

Rationality: The test for rationality is a joint test of unbiasedness and efficiency, and our null and alternative hypothesis are:

$$\mathrm{H}_0^R : \beta = \beta_s = 0 \;\text{ vs. }\; \mathrm{H}_A^R : \beta \neq 0,\ \beta_s \neq 0,\ \text{or } (\beta, \beta_s) \neq 0. \tag{2}$$

In contrast, traditional tests of Markov switching (Hansen, 1992; Garcia, 1998; Carrasco et al., 2014; Qu and Zhuo, 2021) consider the null hypothesis

$$\mathrm{H}_0^{\mathrm{MS}} : \beta_s = 0,$$

but leave the value of β unspecified under the null. Traditional tests of forecast rationality, on the other hand, (Mincer and Zarnowitz, 1969; West and McCracken, 1998) consider the model

$$\epsilon_{t,h} = \beta x_{t+h} + e_{t+h},$$

and restrict the value of β to be equal to zero under the null hypothesis, while β_s is not part of the model's parameter space.

Unbiasedness: In the special case of unbiasedness tests, $x_{t+h} = 1$, our null and alternative hypotheses are:

$$\mathrm{H}_0^U : \mu = \mu_s = 0 \;\text{ vs. }\; \mathrm{H}_A^U : \mu \neq 0,\ \mu_s \neq 0,\ \text{or } (\mu, \mu_s) \neq 0. \tag{3}$$

Existing tests for Markov switching test the null hypothesis of no time variation:

$$\mathrm{H}_0^{\mathrm{MS}} : \mu_s = 0,$$

but do not impose $\mu = 0$. However, the additional restriction of

$$\mathrm{H}_0^{U} : \mu_s = \mu = 0$$

is important in order to have power against a constant deviation from forecast rationality. Traditional tests for unbiasedness (Mincer and Zarnowitz, 1969; West and McCracken, 1998) implement the regression

$$\epsilon_{t,h} = \mu + e_{t+h},$$

where the null hypothesis is that μ is equal to zero, and μ_s is not part of the model's parameter space. Consequently, the tests lack power in the case of Markov switching deviations from forecast unbiasedness.

Efficiency: In the special case of efficiency tests, $x_{t+h} = [y_{t,h}]$, our null and alternative hypotheses are

$$\mathrm{H}_0^{E} : \gamma = \gamma_s = 0 \;\text{ vs. }\; \mathrm{H}_A^{E} : \gamma \neq 0,\ \gamma_s \neq 0,\ \text{or } (\gamma, \gamma_s) \neq 0. \tag{4}$$

The null in existing tests for Markov switching is

$$\mathrm{H}_0^{\mathrm{MS}} : \gamma_s = 0,$$

and the value of γ is unrestricted under the null hypothesis. On the other hand, traditional forecast efficiency tests (Mincer and Zarnowitz, 1969; West and McCracken, 1998) implement the regression

$$\epsilon_{t,h} = y_{t,h}\gamma + e_{t+h},$$

where the value of γ is restricted to be to be zero under the null hypothesis, and γ_s is not part of the model's parameter space.

3 Testing for Markov switching rationality

This section introduces our Markov switching forecast rationality tests, inspired by Hansen (1992) and Qu and Zhuo (2021). We also consider a test inspired by Garcia (1998) in the Online Appendix.

3.1 AFE-H test for rationality

Let α_0 denote the parameter vector under our null hypotheses, formulated in Section 2, and let $\alpha \in A$, with A being a compact metric space, denote a given alternative. Hansen (1992) considers the likelihood ratio as an empirical process indexed by the parameters of interest, $\alpha = (\beta_s, p, q)$, where (p, q) are transition probabilities, and depending further on the nuisance parameters, $\theta = (\beta, \phi_1, ..., \phi_d, \sigma)$. To use the strategy of Hansen (1992) for testing our joint null hypothesis, we need to partition the parameter space differently since our null hypotheses specify both β and β_s. Therefore, we cannot treat β as a nuisance parameter, instead, we must add it to the vector of parameters of interest.

Therefore, the three relevant parameter vectors for us are $\alpha = (\beta, \beta_s, p, q)$, $\alpha = (\mu, \mu_s, p, q)$, and $\alpha = (\gamma, \gamma_s, p, q)$, for testing rationality, unbiasedness, and efficiency respectively. The vector of nuisance parameters reduces to the lag coefficients and the standard deviation, $\theta = (\phi_1, ..., \phi_d, \sigma)$. The subsequent derivation follows closely Hansen (1992). Let us define

$$\widehat{\theta} = \max_{\theta \in \Theta} \mathrm{L}_T(\alpha_0, \theta)$$

to be the the maximum likelihood estimation (MLE) of the nuisance parameters under the null, α_0, and let

$$\widehat{\theta}(\alpha) = \max_{\theta \in \Theta} \mathrm{L}_T(\alpha, \theta(\alpha))$$

denote the MLE of the nuisance parameters under the alternative α. The likelihood ratio is defined as

$$\widehat{\mathrm{LR}}_T(\alpha) = \mathrm{L}_T(\alpha, \widehat{\theta}(\alpha)) - \mathrm{L}_T(\alpha_0, \widehat{\theta}),$$

with

$$\mathrm{L}_T(\alpha, \widehat{\theta}(\alpha)) = \sum_{t=1}^{T} \ell_t(\alpha, \widehat{\theta}(\alpha)) \quad \text{and} \quad \mathrm{L}_T(\alpha_0, \widehat{\theta}) = \sum_{t=1}^{T} \ell_t(\alpha_0, \widehat{\theta}),$$

where ℓ_t denotes the log likelihood of observation t, which is allowed to exhibit serial correlation

and heterogeneity.[4] As in Hansen (1992), the likelihood ratio is split into its expected value, $R_T(\alpha)$, and its deviation from that expectation, $Q_T(\alpha)$,

$$\widehat{LR}_T(\alpha) = R_T(\alpha) + Q_T(\alpha) + O_p(1),$$

where

$$R_T(\alpha) = E\big[L_T\big(\alpha, \theta(\alpha)\big) - L_T\big(\alpha_0, \theta\big)\big] = E\Big[\sum_{t=1}^{T}\big[\ell_t\big(\alpha, \theta(\alpha)\big) - \ell_t\big(\alpha_0, \theta\big)\big]\Big],$$

and

$$Q_T(\alpha) = \sum_{t=1}^{T} q_t(\alpha) = \sum_{t=1}^{T}\Big[\ell_t\big(\alpha, \theta(\alpha)\big) - \ell_t\big(\alpha_0, \theta\big) - E\big[\ell_t\big(\alpha, \theta(\alpha)\big) - \ell_t\big(\alpha_0, \theta\big)\big]\Big],$$

while θ and $\theta(\alpha)$ denote the large sample values of the MLE of $\widehat{\theta}$ and $\widehat{\theta}(\alpha)$. Note that $Q_T(\alpha)$ has a mean of zero. The parameter estimation error that is present in the sample analog of $R_T(\alpha)$ and $Q_T(\alpha)$ is included in the term $O_p(1)$. Please see the Online Appendix for details.

Under the null, $R_T(\alpha) \leq 0$ since the value of $R_T(\alpha)$ is maximized at the true parameter α_0 (under the null). It follows that

$$\frac{1}{\sqrt{T}}\widehat{LR}_T(\alpha) \leq \frac{1}{\sqrt{T}}Q_T(\alpha) + o_p(1).$$

Let $V_T(\alpha)$ denote the variance of the $q_t(\alpha)$. For a fixed α, and by standardizing with $V_T(\alpha)$, the zero mean process $Q_T(\alpha)$ converges to a standard Normal distribution by a Central Limit Theorem (CLT):

$$\frac{1}{\sqrt{T}}\frac{Q_T(\alpha)}{V_T^{1/2}(\alpha)} = \frac{1}{\sqrt{T}}Q_T^*(\alpha) \to_d \ N(0,1). \tag{5}$$

The asymptotic distribution of the bound, $\frac{1}{\sqrt{T}}Q_T^*(\alpha)$, uniformly over $\alpha \in A$, can be derived by applying an empirical process CLT, using the assumptions stated in Hansen (1992), to eq. (5):

$$\sup_{\alpha\in A}\ \frac{1}{\sqrt{T}}\widehat{LR}_T^*\big(\alpha, \theta(\alpha)\big) \leq \sup_{\alpha\in A}\ \frac{1}{\sqrt{T}}Q_T^*(\alpha) + o_p(1) \to_d \sup_{\alpha\in A}\ Q^*(\alpha). \tag{6}$$

The process $Q^*(\alpha)$ is Gaussian with covariance function:

$$K^*(\alpha_1, \alpha_2) = \frac{\sum_{k=-\infty}^{\infty} E q_t(\alpha_1) q_{t+k}(\alpha_2)}{V(\alpha_1)^{\frac{1}{2}} V(\alpha_2)^{\frac{1}{2}}},$$

where $V(\alpha_i)$ denotes the probability limit of the sample analog $V_T(\alpha_i)$.

Since the covariance function, $K^*(\cdot,\cdot)$, depends through $q_t(\alpha_i)$ on the data, critical values

[4] The serial correlation is restricted to near-epoch dependence.

cannot be tabulated for the general case. Instead, analog to Hansen (1992), critical values can be approximated by drawing independently and identically distributed (iid) Gaussian processes that have covariance function $\widehat{K}^*(\cdot,\cdot)$, the empirical counterpart of the unknown $K^*(\cdot,\cdot)$. Doing so is straightforward and requires the simulation of

$$\mathrm{Q}_T^{*j}(\alpha_i) = \frac{\sum_{m=0}^{M}\sum_{t=1}^{T}\widehat{q}_t(\alpha_i)v_{t+m}^j}{\sqrt{1+M}V_T(\alpha_i)^{\frac{1}{2}}},$$

based on J replications, where the v_{t+m}^j, for $j = 1, ..., J$, are iid $N(0,1)$ variates, and $\widehat{q}_t(\alpha_i)$ is the empirical counterpart of $q_t(\alpha_i)$. The $\mathrm{Q}_T^{*j}(\alpha_i)$ have $\widehat{K}^*(\cdot,\cdot)$ as a covariance function and, hence, approximate the asymptotic distribution. Moreover, Hansen (1996) points out that the likelihood components $q_t(\alpha_i)$ are serially correlated even if the data is iid and, therefore, a Bartlett kernel is used to account for the autocorrelation. The Bartlett's bandwidth parameter, M, can be data dependent; typical choices are $M = T^{1/4}$ or $M = [4(T/100)^{2/9}] + 1$.[5] Critical value are then obtained as percentiles from the distribution of $\{\mathrm{Q}_T^{*j}\}_{j=1}^J$, with $\mathrm{Q}_T^{*j} = \sup_{\alpha\in A}\mathrm{Q}_T^{*j}(\alpha_i)$.

To obtain a set of $\mathrm{Q}_T^*(\alpha_i)$, ones has to estimate the model under the alternative over a grid of values for $\alpha = (\beta, \beta_s, p, q)$. Table 1 displays the average critical values we obtain when we implement the AFE-H test of unbiasedness and efficiency. Let the data be generated by $y_t = e_t$, where $e_t \sim N(0,1)$. The partitions of the parameter vectors are in this case $\alpha = (\mu, \mu_s, p, q)$, and $\alpha = (\gamma, \gamma_s, p, q)$ respectively. The column denoted by 'H' shows the critical values for the original Hansen (1992) null, $\mathrm{H}_0^{\mathrm{MS}}$: $\mu_s = 0$ and $\mathrm{H}_0^{\mathrm{MS}}$: $\gamma_s = 0$, with $\alpha_U = (\mu_s, p, q)$ and $\alpha_E = (\gamma_s, p, q)$. As the approximation of the asymptotic distribution is data dependent, the numbers are obtained by averaging the critical values over all Monte Carlo replications. The aim of this exercise is not to tabulate critical values, which would be invalid due to the data dependence of the asymptotic distribution, but to show how the additional parameter restriction changes the critical values we obtain. As expected, the critical values of the AFE-H test are larger on average, reflecting the additional restriction of the AFE-H on the null parameter space.

As mentioned, in order to implement the AFE-H test in practice, the researcher needs to decide on the grid values for (p, q) and (β, β_s). When evaluating the likelihood ratio process under the Markov switching alternative, for each point on the grid, the researcher will optimize a constrained (imposing the grid point values) likelihood to obtain $\widehat{\theta}(\alpha)$. The model under the null hypothesis, on the other hand, is estimated with the constraint that $(\beta = 0, \beta_s = 0)$, i.e. there is

[5] In addition, in applications of the AFE-H, the researcher can easily simulate the asymptotic distribution for different values of M to gauge the impact of the serial correlation on the critical values. Our results where not sensitive to the choice of M, which is in line with the findings of Hansen (1996).

Table 1: Average critical values

	Unbiasedness		Efficiency	
Nominal Size	AFE-H	H	AFE-H	H
1%	3.60	2.80	3.44	3.24
5%	3.01	2.51	2.84	2.64
10%	2.72	2.18	2.53	2.33

Note: The table shows the average critical values based on our simulations for the proposed AFE-H test and the original Hansen (1992), labeled 'H', test for a standard Normal DGP, a sample size of $T = 500$, and 500 Monte Carlo simulations.

no Markov switching present by assumption. The grid points under the alternative serve as a basis for the construction of the test statistics as well the limiting distribution, thus deserving particular interest.

Since (p, q) are bounded below by zero and above by one, the grid choice for (p, q) is about how many grid points to consider; we used 12 grid points in our Monte Carlo and did not experience the results to be very sensitive to slightly different choices. In addition, state-to-state transition probabilities in Markov switching models are often well above 0.5, such that the researcher may as well restrict the grid of (p, q) accordingly. The grid choice for (β, β_s) is somewhat more difficult since their domain is not restricted to be between zero and one. Although the grid values for (β, β_s) will vary with the empirical application, the researcher can typically rule out large values for the grid since the left hand side variable is the forecast error, which tends to be small. In general, we recommend to plot the data, and to estimate an unrestricted Markov switching rationality regression to get an idea of where to set the grid for (β, β_s) in practice. Relative to Hansen (1992), our proposed procedure could be somewhat more computationally intensive, since it requires evaluating the constrained likelihood with a grid structure for an additional parameter, β. In our simulations the performance of the tests are not very sensitive to the choice of the grid points.

The estimation of the Markov switching model (with our without grid) is implemented efficiently using the expectation-maximization (EM) algorithm described in Hamilton (1990). However, the researcher can rely on other approaches, as for instance, on the filtering techniques for endogeneous switching outlined in (Chang et al., 2017).

3.2 AFE-BS test for forecast rationality

Qu and Zhuo (2021) showed that when testing for Markov-switching, a parametric bootstrap consistently approximates the asymptotic distribution of the likelihood ratio test as long as it

correctly reproduces the covariance function of the limiting distribution derived in Qu and Zhuo's (2021) Proposition 1.

Recall that the model under the null hypothesis reduces to an AR(d). Following Qu and Zhuo (2021), let $\Lambda_{(p,q)} = \{(p,q) : 0.02 \geq p,q \leq 0.98 \quad \text{and} \quad p+q \geq 1.02\}$ denote the set of feasible values for (p,q). Let $\text{LL}_{0,T}$ and $\text{LL}_{A,T,\Lambda_{(p,q)}}$ (note that we subsequently drop the subscript $\Lambda_{(p,q)}$ for notational convenience) denote the log-likelihood under the null and under the alternative hypothesis, respectively.[6] Let $\text{LR}_T = 2\big(\text{LL}_{0,T} - \text{LL}_{A,T,\Lambda_{(p,q)}}\big)$ denote the likelihood-ratio.

Parametric bootstrap for testing forecast rationality: In order to construct the likelihood ration test, we need to evaluate the likelihood under both the null and the alternative. This requires us to re-sample both $\epsilon_{t,h}$ and x_{t+h}, respecting their covariance structure.[7] Let $\widehat{\phi}_{i,0}$, for $i = 1,...,d$, denote the parameter estimates of the autoregressive coefficients under the null. Let $\widehat{\phi}_{i,x}$ and $\widehat{\phi}_{i,\epsilon_x}$, for $i = 1,..,d_x$, denote the parameter estimates of a regression x_{t+h} on d_x lags of $\epsilon_{t,h}$ and x_{t+h} jointly. Let $\widehat{e}_{t+h}$ and $\widehat{e}_{x,t+h}$ denote the estimated error term of the regression of $\epsilon_{t,h}$ on its own lags and of the regression of x_{t+h} on its own lags and lags of $\epsilon_{t,h}$, respectively. Further, let $\widehat{\Sigma}_e$ denote the covariance matrix of $[\widehat{e}_{t+h}, \widehat{e}_{x,t+h}]'$, and $d^* = max(d, d_x)$. Then, for $j = 1,...,J$, we proceed with the following steps:

1. Draw $T + d^*$ random variables from $N(0, \widehat{\Sigma}_e)$ and denote by $\{v^*_{t,j}\}_{t=-d^*+1}^{T}$ the set of draws.
2. Construct a series $\epsilon^*_{t,h,j}$ and $x^*_{t+h,j}$, for $t = 1,...,T$ using $v^*_{t,j}$, $\widehat{\phi}_{i,0}$ for $i = 1,...,d$, and $(\widehat{\phi}_{i,x}, \widehat{\phi}_{i,\epsilon_x})$, for $i = 1,...,d_x$. We elaborate details about this step below.
3. Using $\{\epsilon^*_{t,h,j}, x^*_{t+h,j}\}_{t=1}^{T}$, compute the bootstrap log-likelihood under the null, $\text{LL}^*_{0,T,j}$, and under the alternative, $\text{LL}^*_{A,T,j}$.
4. Store the bootstrapped likelihood ratio: $\text{LR}^*_{T,j} = 2\big(\text{LL}^*_{A,T,j} - \text{LL}^*_{0,T,j}\big)$.

After J iterations, we obtain a set of the bootstrapped likelihood ratio statistic, $\{\text{LR}^*_{T,j}\}_{j=1}^{J}$, that approximates the asymptotic distribution.

For the case of forecast unbiasedness, $x_{t+h} = 1$, the researcher only has to re-sample $\epsilon_{t,h}$ and, therefore, $v^*_{t,j}$ is a scalar, drawn from $N(0, \widehat{\sigma}^2_{e,0})$, where $\widehat{\sigma}^2_{e,0}$ denotes the estimated variance of

[6] To reduce the computational costs of the estimation procedure under the alternative, we proceed as follows. In a first step, we maximize the log-likelihood of the model under the alternative without taking into account the restrictions on (p,q) embedded in $\Lambda_{(p,q)}$. If the obtained maximum implies values for (p,q) outside of the feasible set $\Lambda_{(p,q)}$, we resort to estimating the model over a 2-tuple of 10 equally-spaced grid values for (p,q) in $[0.02, 0.98]$; otherwise, we proceed with the maximum obtained in the first step.

[7] We assume normal errors to illustrate the bootstrap procedure, since a normality assumption is the leading case for Markov switching applications.

$\widehat{e}_{t+h}$. Then, if $d > 0$, set $(\epsilon^*_{-d+1,h,j}, ..., \epsilon^*_{0,h,j})$ equal to $(1 - \sum_i^d \widehat{\phi}_{i,0})^{-1}(v^*_{-d+1,h,j}, ..., v^*_{0,h,j})$. We further generate $\epsilon^*_{t,h,j} = \sum_{i=1}^d \widehat{\phi}_{i,0}\epsilon^*_{t-i,j,h} + v^*_{t,h,j}$ recursively for $t = 1, ..., T$; if $d = 0$, set $\epsilon^*_{t,h,j} = v^*_{t,h,j}$.

For the case of forecast efficiency, we need to resample $x_{t+h} = 1$ and $\epsilon_{t,h}$ jointly. The bootstrap procedure of Qu and Zhuo (2021) does not directly apply in this case, and they do not recommend using a fixed-regressor bootstrap. Instead, we implement the following procedure following the recommendation of Qu and Zhuo (2021). If the DGP is $y_t = \psi y_{t-1} + u_t$, the forecasting model will be $y_{t,1} = \psi y_t$, such that $x_{t+1} = y_{t,1}$. The forecast error subsequently is $\epsilon_{t,1} = y_{t+1} - y_{t,1} = u_{t+1}$. Then, the researcher can re-sample $\epsilon_{t,1}$ and x_{t+1} as follows: $\epsilon^*_{t,1,j} \sim N(0, \widehat{\sigma}^2_{e,0})$, where $\widehat{\sigma}^2_{e,0}$ is estimated using $\widehat{e}_{t+h}$ with $d = 0$, and $x^*_{t+1,j} = \psi(x^*_{t,1,j} + \epsilon^*_{t-1,1,j})$, and $x^*_{0,j} \sim N(0, \frac{\psi^2}{1-\psi^2}\widehat{\sigma}^2_{e,0})$.

4 Monte Carlo simulation results

This section provides Monte Carlo evidence on the finite sample size and power of the unbiasedness and efficiency tests. In all instances, the estimation of the Markov switching model is based on the EM algorithm (Hamilton, 1990).

4.1 Monte Carlo results — unbiasedness

The DGP is the following in this section:

$$y_t = \psi y_{t-1} + u_t, \tag{7}$$

where $|\psi| < 1$ and $u_t \sim N(0, 1)$. We then consider three different forecasting situations that lead to three different cases of forecast errors.

Case 1: Forecasting one-step-ahead with an AR(1)

$$y_{t,1} = \psi y_t \quad \text{and} \quad \epsilon_{t,1} = u_{t+1}, \tag{8}$$

where we model the forecast error as $\epsilon_{t,1} = \mu + S_{t+1}\mu_s + e_{t+1}$ and set $\psi = 0.5$.

Case 2: Forecasting one-step-ahead with a constant

$$y_{t,1} = c, \quad \epsilon_{t,1} = \psi y_t + u_{t+1}, \tag{9}$$

where we model the forecast error as $\epsilon_{t,1} = \mu + S_{t+1}\mu_s + \phi_1\epsilon_{t-1,1} + e_{t+1}$, set $c = 0$, and set $\psi = 0.5$

.

Case 3: Forecasting multi-step-ahead

When forecasting two-periods-ahead with an AR(1) model, the errors will have a MA(1) dynamic in population:

$$y_{t,2} = \psi^2 y_t, \quad \epsilon_{t,2} = (1 + \psi L) u_{t+2}, \tag{10}$$

where we set $\psi = 0.25$.

In the application of our AFE-H and AFE-BS tests, we approximate the MA(1) error dynamics with a Markov switching AR(1) process: $\epsilon_{t,2} = \mu + S_{t+1}\mu_s + \phi_1 \epsilon_{t-1,2} + e_{t+2}$.

In all cases, μ denotes the intercept and μ_s is the parameter that changes with the Markov switching regime. S_{t+1} is a stationary Markov chain and $e_t \sim N(0, \sigma^2)$. Note that the Markov switching specification correctly approximates the forecast error's dynamics in Case 1 and Case 2, while the Markov switching model's dynamics are misspecified in Case 3. This case intends to emulate a realistic forecast situation where the researcher has a multiple-step-ahead forecast at hand and controls for potential serial correlation in the forecast error via an autoregressive specification, which is typically easier to estimate than a MA specification.

The null hypothesis of the AFE-H and AFE-BS tests imposes the restriction $\mu = \mu_s = 0$. Panel A of Table 2 shows the size results for AFE-H and AFE-BS tests for a nominal size of 5%. Panel B, instead, shows results for a nominal size of 10%. We also include results for the tests of West and McCracken (1998) (labeled "WM") and Rossi and Sekhposyan (2016) (labeled "Fluctuation"), which test for constant unbiasedness and time-varying unbiasedness, respectively. As discussed previously, Rossi and Sekhposyan (2016) capture time variation non-parametrically, based on a rolling window estimation.[8] Overall, the size results of the AFE-H and AFE-BS tests are good, although AFE-BS performs better in small and medium-sized samples relative to the AFE-H. The AFE-H test overrejects for small and medium-sized samples; however, the size distortions are of a similar magnitude as in Hansen (1992). The mild misspecification in the forecast error dynamics in Case 3 only leads to small size distortions for the AFE-H and AFE-BS tests.[9]

To study power, we consider first the alternative of a constant deviation from unbiasedness. The DGP takes the form of

$$y_t = \tilde{\mu} + \psi y_{t-1} + u_t, \tag{11}$$

[8]Note that the size distortions in small samples for the Fluctuation test come from the fact that although the true DGP is an AR(1), we use a HAC estimator of Newey and West (1987) with a bandwidth equal to $T^{(1/4)}$ to control for the autocorrelation.

[9]Markov switching tests are generally not robust to misspecification under the null hypothesis. In unreported results, we found that for a more severe misspecification, i.e. large values of the MA(1) coefficient in Case 3, both the AFE-H and the AFE-BS show size distortions.

Table 2: Size results - forecast unbiasedness test

Panel A. *Nominal size 5%*

	Case 1			Case 2			Case 3		
Test	T = 100	T = 200	T = 500	T = 100	T = 200	T = 500	T = 100	T = 200	T = 500
WM	0.046	0.048	0.054	0.107	0.110	0.070	0.062	0.062	0.056
Fluct.	0.062	0.062	0.060	0.218	0.187	0.126	0.134	0.128	0.081
AFE-H	0.122	0.082	0.046	0.130	0.058	0.026	0.144	0.076	0.056
AFE-BS	0.048	0.047	0.058	0.045	0.043	0.045	0.046	0.026	0.032

Panel B. *Nominal size 10%*

	Case 1			Case 2			Case 3		
Test	T = 100	T = 200	T = 500	T = 100	T = 200	T = 500	T = 100	T = 200	T = 500
WM	0.099	0.091	0.099	0.178	0.167	0.133	0.121	0.115	0.103
Fluct.	0.112	0.112	0.115	0.299	0.262	0.193	0.208	0.207	0.138
AFE-H	0.176	0.122	0.086	0.182	0.100	0.060	0.198	0.114	0.096
AFE-BS	0.097	0.109	0.110	0.093	0.097	0.088	0.077	0.068	0.064

Note: T denotes the sample size. Cases 1, 2, and 3 refer to the various simulation designs described in Section 4.1. Results are based on 1000 Monte Carlo replications, except for AFE-H: due to the computational time, these Monte Carlo replications are limited to 500. The results for AFE-H test are based on a 4-tuple of 12 equally-spaced grid points for $(p, q) \in [0.05, 0.95]$ and 20 equally-spaced grid points for $(\mu, \mu_s) \in [-1, 1] \times [-2, 2]$. Results for the AFE-BS test are based on 200 bootstrap replications and $(p, q) \in \Lambda_{(p,q)}$. The window size m for the Fluctuation test is set to $m = \frac{T}{2}$.

with $\psi = 0.5$, where the forecasting model is $y_{t,1} = \psi y_t$, such that the forecast error becomes $\epsilon_{t,1} = \tilde{\mu} + u_{t+1}$. The different values for $\tilde{\mu}$ are $[0.20, 0.25, 0.30, 0.35, 0.375, 0.40, 0.45, 0.50]$.

Panel A of Table 3 shows size-adjusted power results for a sample size of $T = 100$ and a nominal size of 5%. As expected, the WM test has the highest power against a constant deviation from the null hypothesis of unbiasedness. However, the AFE-H and AFE-BS test exhibits good power as well and the power increases rapidly with the magnitude of the deviation from unbiasedness. The power of the Fluctuation test is comparable and only slightly worse than that of the AFE-BS test against the constant alternative.

To test for power against the alternative of Markov switching, the DGP takes the form of

$$y_t = \tilde{\mu} + S_t \tilde{\mu}_s + \psi y_{t-1} + u_t, \tag{12}$$

with $\psi = 0.5$, where the forecasting model is $y_{t,1} = \psi y_t$, such that the forecast error becomes $\epsilon_{t,1} = \tilde{\mu} + S_{t+1}\tilde{\mu}_s + u_{t+1}$.

We set the state-to-state transition probabilities of the Markov chain S_t to be $(p, q) = (0.9, 0.9)$ and impose $\tilde{\mu} = -\tilde{\mu}_s/2$. These parameter choices ensure that the unconditional mean of the series is zero, i.e. $\mathrm{E}(\epsilon_{t+1|t}) = 0$, such that we can compute the power against

Table 3: Power results - unbiasedness

	Panel A. *Constant bias* Values of $\tilde{\mu}$							
	0.20	0.25	0.30	0.35	0.375	0.40	0.45	0.50
WM	0.49	0.71	0.82	0.92	0.97	0.98	0.99	0.99
Fluct	0.33	0.49	0.65	0.76	0.84	0.88	0.95	0.95
AFE-H	0.32	0.50	0.66	0.78	0.87	0.89	0.96	0.99
AFE-BS	0.28	0.47	0.63	0.78	0.88	0.90	0.96	0.98
	Panel B. *Markov switching bias* Values of $\tilde{\mu}_s$							
	0.80	1.00	1.20	1.40	1.50	1.60	1.80	2.00
WM	0.13	0.16	0.17	0.22	0.21	0.27	0.28	0.23
Fluct	0.20	0.27	0.32	0.37	0.39	0.42	0.48	0.44
AFE-H	0.14	0.24	0.33	0.53	0.58	0.67	0.80	0.88
AFE-BS	0.28	0.50	0.73	0.86	0.91	0.96	0.98	1.00

Note: The values denote the size-adjusted empirical rejection frequency based on 500 Monte Carlo replications. The values for μ and μ_s are given in the first rows of Panel A and B respectively. The nominal size is 5%. The results for AFE-H are based on a 4-tuple of 12 equally-spaced grid points for $(p, q) \in [0.05, 0.95]$ and 20 equally-spaced grid points for $(\mu, \mu_s) \in [-1, 1] \times [-2, 2]$. Results for the AFE-BS test are based on 200 bootstrap replications and $(p, q) \in \Lambda_{(p,q)}$.

the alternative of Markov switching only. The different values that we explore for $\tilde{\mu}_s$ are $[0.80, 1.00, 1.20, 1.40, 1.50, 1.60, 1.80, 2.00]$.

Panel B of Table 3 displays the size-adjusted rejection frequencies at a nominal size of 5%. The AFE-BS and AFE-H tests exhibit strong power against the alternative of Markov switching.

The rejection frequency of the WM test would theoretically be expected to remain at the nominal level of 5%. However, in small samples, it is quite likely to sample one of the states more often than the other, even if the unconditional state probabilities are 0.5, which shifts the sample mean away from zero (this only occurs in small samples with a high regime persistence).

When looking at the Fluctuation test, we find that it does not have strong power against Markov-switching type of time variation. This result is driven by the non-parametric approach of the test, i.e. it has less power against parametric discrete switches. Note, however, that the power results of the Fluctuation test depend to some extent on the window size — smaller windows would likely improve the tests' power under Markov switching. AFE-BS exhibits a lower power than AFE-H; however, note that the grid size used for the test statistic of μ and μ_s could influence the results (though it did not seem to be sensitive to the grid choice in our Monte Carlo exercises).

4.2 Monte Carlo results — efficiency

We now turn to test forecast efficiency. Under the null, the DGP is the same as in eq. (7), i.e.

$$y_t = \psi y_{t-1} + u_t, \tag{13}$$

where we set $\psi = 0.5$ and $u_t \sim N(0, \sigma_e^2)$. The forecasting model takes the form of $y_{t,1} = \psi y_t$ such that the forecast error becomes

$$\epsilon_{t,1} = u_{t+1}. \tag{14}$$

We use the following Markov switching specification

$$\epsilon_{t,1} = \gamma y_{t,1} + S_{t+1}\gamma_s y_{t,1} + e_{t+1}. \tag{15}$$

to test the null hypothesis: $\gamma = \gamma_s = 0$ in the following regression, where S_{t+1} is a stationary Markov chain and $e_{t+1} \sim N(0,1)$.

Table 4 shows the size results for AFE-H and AFE-BS tests as well as the WM and the Fluctuation tests. The size results of AFE-BS test are good, even in small samples; the AFE-H somewhat underrejects for small and large samples. The reason for the distortions could be that the critical values are taken from a bound instead of an exact distribution and, therefore, the test is more conservative, or that the test is more sensitive to the choice of the grid for (γ, γ_s).

Table 4: Size results - efficiency

Test	T = 100	T = 200	T = 500	T = 100	T = 200	T = 500
	Nominal size 5%			Nominal size 10%		
WM	0.059	0.067	0.053	0.114	0.120	0.107
Fluct.	0.050	0.049	0.046	0.094	0.092	0.095
AFE-H	0.020	0.014	0.012	0.036	0.030	0.028
AFE-BS	0.055	0.058	0.053	0.098	0.113	0.094

Note: T denotes the sample size. Results are based on 1000 Monte Carlo replications, except for AFE-H: due to the computational time, these Monte Carlo replications are limited to 500. The results for AFE-H test are based on a 4-tuple of 12 equally-spaced grid points for $(p, q) \in [0.05, 0.95]$ and 20 equally-spaced grid points for $(\gamma, \gamma_s) \in [-1, 1] \times [-2, 2]$. Results for the AFE-BS test are based on 200 bootstrap replications and $(p, q) \in \Lambda_{(p,q)}$. The window size m for the Fluctuation test is $m = \frac{T}{2}$.

Table 5: Power results - efficiency

	Constant deviation Values of $\tilde{\gamma}$					*Markov switching deviation* Values of $\tilde{\gamma}_s$				
	0.15	0.20	0.25	0.30	0.35	0.30	0.40	0.50	0.60	0.70
WM	0.46	0.67	0.87	0.96	1.00	0.05	0.09	0.12	0.15	0.18
Fluct	0.44	0.58	0.74	0.88	0.98	0.08	0.12	0.15	0.20	0.27
AFE-H	0.38	0.58	0.80	0.94	1.00	0.06	0.15	0.32	0.59	0.82
AFE-BS	0.30	0.51	0.74	0.91	0.99	0.07	0.22	0.45	0.74	0.92

Note: The values denote the size-adjusted empirical rejection frequency based on 500 Monte Carlo replications. The values for γ and γ_s are given in the first row of Panel A and B respectively. The nominal size is 5%. The results for AFE-H test are based on a 4-tuple of 12 equally-spaced grid points for $(p, q) \in [0.05, 0.95]$ and 20 equally-spaced grid points for $(\gamma, \gamma_s) \in [-1, 1] \times [-2, 2]$. Results for the AFE-BS test are based on 200 bootstrap replications and $(p, q) \in \Lambda_{(p,q)}$.

To study power, we proceed as follows. Under the alternative of a constant, but non-zero efficiency coefficient, the DGP takes the form

$$y_t = (\psi + \tilde{\gamma})y_{t-1} + u_t, \tag{16}$$

with the forecasting model being $y_{t,1} = \psi y_t$, such that the forecast error becomes $\epsilon_{t,1} = \tilde{\gamma} y_t + u_{t+1}$, where we let $\tilde{\gamma}$ take the following values $[0.15, 0.20, 0.25, 0.30, 0.35]$.

Panel A of Table 5 shows the size-adjusted power results for a sample size of $T = 100$ at a nominal size of 5%. Again the MW test outperforms the other tests in terms of power. However, the AFE-H and AFE-BS have good power against the null of a constant deviation as well.

To test for the alternative of Markov switching, the DGP takes the form

$$y_t = (\psi + \tilde{\gamma})y_{t-1} + S_t \tilde{\gamma}_s y_{t-1} + u_t, \tag{17}$$

where the forecasting model is $y_{t,1} = \psi y_t$, such that the forecast error becomes $\epsilon_{t,1} = u_{t+1} + \tilde{\gamma} y_t + S_{t+1} \tilde{\gamma}_s y_t$. We set the state-to-state transition probabilities to be $(p, q) = (0.9, 0.9)$, set $\tilde{\gamma} = -\tilde{\gamma}_s/2$, and let $\tilde{\gamma}_s$ take the following values $[0.30, 0.50, 0.70, 0.90, 1.10]$.[10]

Results are shown in Panel B of Table 5. We find that the traditional WM and Fluctuation tests have less power against the alternative of Markov switching efficiency than the AFE-H and AFE-BS tests.

[10] Again, these parameter choices ensure that we can compute the power against the alternative of Markov switching efficiency only.

4.3 Discussion

Testing for Markov switching is challenging and both of the proposed tests have advantages and disadvantages.

When using the AFE-H test, the researcher needs to carefully set the grid of parameters of interest. When testing unbiasedness, we recommend plotting the forecast error to decide the grid values. When testing efficiency, setting a grid around the full sample efficiency parameter could be a natural starting point. In addition, the AFE-H has the drawback of being computationally intensive and displaying size distortions in small samples, but the presence of an additional control variable (beyond the autoregressive lags) does not require the researcher to specify a law of motion for this variable.

While the parametric bootstrap procedure requires the researcher to make this additional assumption, which might be difficult in some situations, such as when testing for forecast efficiency of survey forecasts, it addresses many of the technical issues associated with testing for Markov switching. In addition, it has good small sample properties and the researcher only has to search over a grid for the state-to-state transition probabilities.

So far, we have let the variance of e_t be constant; however, that can be relaxed. For instance, the variance can follow a Markov switching process itself. If the variance shares the same regime dynamics as the rationality coefficients, then it can help identify the regime. However, in this case, a rejection of a null hypothesis would not indicate whether the rejection is due to violations of rationality or switches in the variance. Instead, if the variance has its own Markov switching dynamics, then it should be modeled separately. Though testing for switches in the variance might not be of first-order interest in the context of our proposed rationality tests, our framework allows for it nonetheless.

In general, Markov switching models are mixture models and, therefore, the misspecification of the likelihood can lead to size distortions when using likelihood-based tests. Misspecification is less prevalent in the context of rationality tests than when comparing forecasts since forecast error distributions are often reasonably well approximated by a Normal distribution.

5 A Markov switching bias in the Federal Funds Rate forecasts

This section investigates the forecast unbiasedness of the Blue Chip Financial Forecasts (BCFF) survey's predictions for the FFR. Significant deviations from forecast unbiasedness by survey participants are important since a state-dependent bias in the interest rate expectations implies

that it might be possible to improve prediction in specific periods in time and policymakers such as the central banks can help in the process by improving the communication strategies.

Previous work that found state dependence in forecast errors includes Joutz and Stekler (2000), Sinclair et al. (2010), and Granziera et al. (2021) for various forecasts of the Federal Reserve and the ECB. Studies of forecast rationality of private-sector survey predictions include Croushore (2012) and Rossi and Sekhposyan (2016), who investigate the forecast rationality of U.S. Survey of Professional Forecasters' predictions; the latter find that forecast rationality is time-varying and depends on the sub-sample considered. Dahlhaus and Sekhposyan (2020) consider the BCFF predictions of the FFR, and test forecast rationality in sub-samples, conditional on whether the economy is in a monetary easing or tightening regime; in their work, the regime is observable and measured by lagged interest rate decreases and increases. Our empirical analysis contributes to this literature by revealing state dependence in the BCFF, without having to restrict our consideration to a specific state variable *ex-ante*. Additionally, we show that the state-dependent bias extends to the forecasts implied by the Federal Funds Futures (FFF) markets, i.e. it is not an idiosyncratic feature of the BCFF survey.

The BCFF is conducted monthly and consists of approximately fifty participants in the private financial sector. We focus on the consensus forecast, which is the cross-sectional average of all participants. The predictions are fixed-event forecasts, and we follow Dahlhaus and Sekhposyan (2020) (see also Chun, 2011) to convert the survey predictions to fixed-horizon forecasts.

In total, the survey data ranges from 1983:M4, the start of the survey, to 2018:M2. In the analysis, we focus on the period starting in 1990:M1 for two reasons. First, the data in the 1980s is quite volatile and contains several outliers. Second, an increase in the Fed's transparency in monetary policy communication at the beginning of the 1990s (Woodford, 2005) gives rise to potentially confounding structural changes in the forecast error dynamics relative to the earlier period. Lastly, the effective sample size depends on the forecasting horizon.

Let FFR_{t+h} denote the average of the effective Federal Funds Rate in month $t+h$, and let $\text{BCFF}_{t,h}$ denote the h-step prediction of the FFR provided by the Blue Chip Financial Forecasts at time t. Then, the forecast error is given by $\epsilon_{t,h} = \text{FFR}_{t+h} - \text{BCFF}_{t,h}$, i.e. it is the difference of the realization and the forecast. To test for unbiasedness, we specify the following model:

$$\epsilon_{t,h} = \mu + S_{t+h}\mu_s + \sum_{i=1}^{d} \phi_i \epsilon_{t-i,h} + e_{t+h}, \tag{18}$$

where $S_t \in \{0,1\}$ is a stationary first-order Markov chain, and $e_t \sim \text{N}(0,\sigma^2)$. In the following, we

denote the case of $S_t = 0$ as regime one and the case of $S_t = 1$ as regime two.

In our baseline specification, we focus on the three-month-ahead forecast error, i.e. $h = 3$. Results for the six-month-ahead are very similar and are reported in the Online Appendix.

Table 6 and Table 7 display the results of the AFE-BS and AFE-H test for unbiasedness, i.e. $\mu = \mu_s = 0$ in eq. (18), for $d = 0, 1, 2, 3$.[11] For the AFE-BS test we let $(p, q) \in \Lambda_{(p,q)}$ as defined in Section 3.2. For the AFE-H test, we used a 4-tuple of 12 equally-spaced grid points for $(p, q) \in [0.04, 0.96] \times [0.04, 0.96]$, and 20 equally-spaced grid points for $(\mu, \mu_s) \in [-1, 0.2] \times [-2, 0.4]$.For all lag lengths, the AFE-BS and AFE-H test reject the null hypothesis of an unbiased forecast at a significance value below 0.01.

The coefficients p and q, displayed in Table 6 and Table 7, show the state-to-state transition probabilities of regime one and two respectively. Across different lag length specifications, regime one is persistent, with a state-to-state transition probability of 96% to 97%, and the forecasts appear to be unbiased as $\mu \approx 0$; a subsequent t-test on μ does not reject the null hypothesis of $\mu = 0$. However, in the second regime, which is considerably less persistent (in Table 6) when controlling for lags of the forecast error, the forecasters overestimate the future FFR, as the coefficient $\mu + \mu_s$ is large, negative, and significantly different from zero. The forecast bias in absolute terms, i.e. $|\mu + \mu_s|$, is estimated to be around 18 to 50 basis points. Note that while the results are not identical across Table 6 and Table 7, they have the same implications.[12]

Figure 2 plots the smoothed regime probabilities (solid lines) of the MS-AR(3) model of Table 6 against the forecast error and the FFR. The left y-axis denotes the scale of the regime probability, whereas the right y-axis denotes the scale of the forecast error and the FFR. The dashed line displays the forecast error (rescaled by a factor of two to increase the legibility of the plot). The dotted line displays the FFR, while grey shaded areas display NBER recession periods. An increase of the probability of regime two is associated with monetary easing, but is not limited to recessionary periods. In particular, in the early 1990s, around 1998, and before the Great Recession in 2007-2009, the probability co-moves with changes in the FFR although the economy was not in a recession according to the NBER. Overall, the regimes appear to be well identified, in the sense that most regime probabilities are close to zero or one.

Figure 3 plots the forecast error, $\epsilon_{t,3}$, against the time-varying unconditional mean of the MS-AR(3) model of Table 6, given by $\widehat{\mu}(1 - \widehat{\phi}_1 - \widehat{\phi}_2 - \widehat{\phi}_3)^{-1} + \widehat{S}_{t+3}\widehat{\mu}_s(1 - \widehat{\phi}_1 - \widehat{\phi}_2 - \widehat{\phi}_3)^{-1}$, using

[11] A rational forecast would exhibit maximum serial correlation length of h-1, i.e. in this case two. We show results for a maximum of three lags to be robust against rejections of the null hypothesis due to other type of misspecifications.

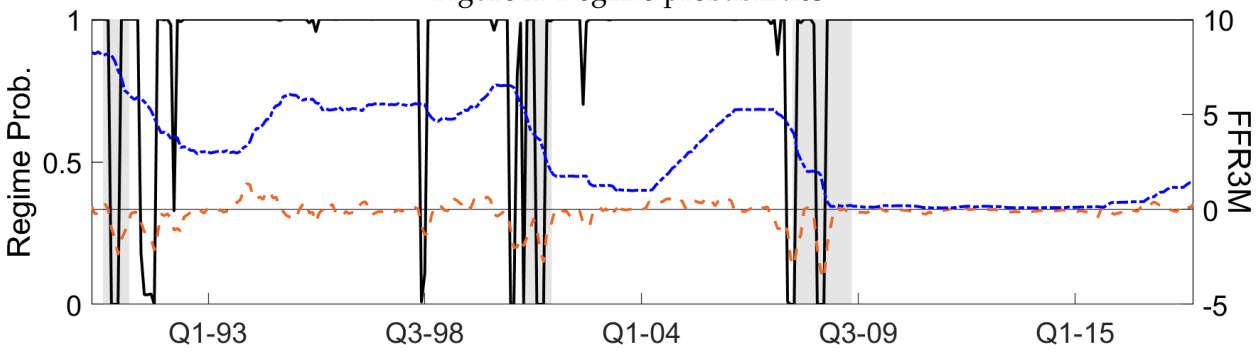
[12] Remember that AFE-H is estimated over a finite grid for both (p, q) and μ, μ_s, which may lead to a slightly different maximum than the estimation that is not based on a finite grid.

Table 6: AFE-BS test results — three-month-ahead forecast error

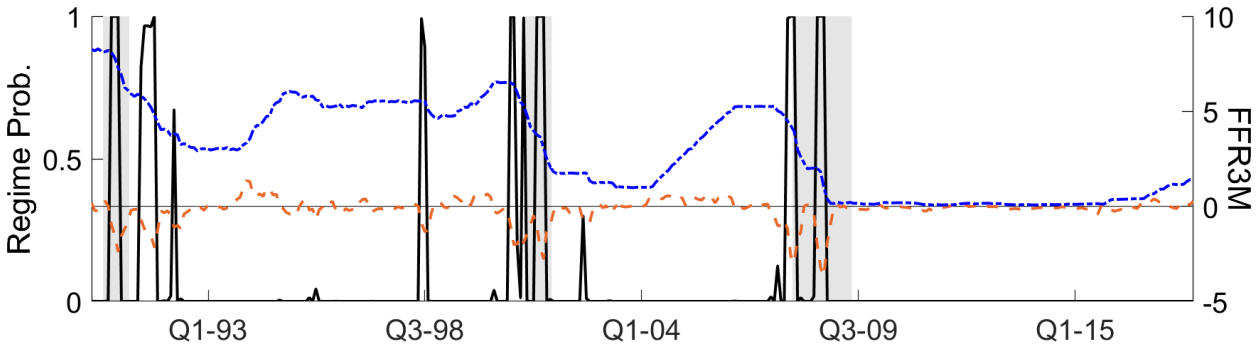
Model	$\widehat{p}$	$\widehat{q}$	$\widehat{\mu}$	$\widehat{\mu}+\widehat{\mu}_s$	$\widehat{\phi}_1$	$\widehat{\phi}_2$	$\widehat{\phi}_3$	LR-value	pvalue
AR(0)	0.98	0.83	0.01	-0.78	-	-	-	146.06	< 0.01
	(0.01)	(0.06)	(0.01)	(0.02)					
AR(1)	0.97	0.62	0.01	-0.53	0.67	-	-	77.17	< 0.01
	(0.01)	(0.11)	(0.01)	(0.02)	(0.01)				
AR(2)	0.97	0.60	0.00	-0.50	0.84	-0.17	-	63.55	< 0.01
	(0.01)	(0.12)	(0.01)	(0.02)	(0.03)	(0.03)			
AR(3)	0.97	0.60	0.00	-0.50	0.81	-0.09	-0.07	64.87	< 0.01
	(0.01)	(0.11)	(0.01)	(0.02)	(0.03)	(0.04)	(0.03)		

Note: The sample size is $T = 338$. maximum obtained under the alternative, using the restriction that $(p, q) \in \Lambda_{(p,q)}$. The column labelled 'LR-value' denotes the value of the likelihood ratio. Numbers in parentheses denote robust standard errors. The column 'pvalue' denotes the p-value obtained using the approximated asymptotic distribution based on 200 bootstrap replications. p denotes the state-to-state transition probability for regime one and q denotes the state-to-state transition probability for regime two.

Figure 2: Regime probabilities



(a) Regime One



(b) Regime Two

Note: The left y-axis denotes the regime probability. The right y-axis denotes the value of the forecast error and the FFR. The solid line displays the smoothed regime probabilities of the Markov switching model with three lags, defined in eq. (18). The dashed line displays the forecast error. We rescaled the forecast error by a factor of two, to increase the legibility of the plot. The dashed-dotted line displays the FFR and grey shaded areas display NBER recession periods.

the smoothed state probabilities for S_{t+3}. The figure shows that the switches in the unconditional mean alone can account for much of the recurring negative realizations of the forecast error.

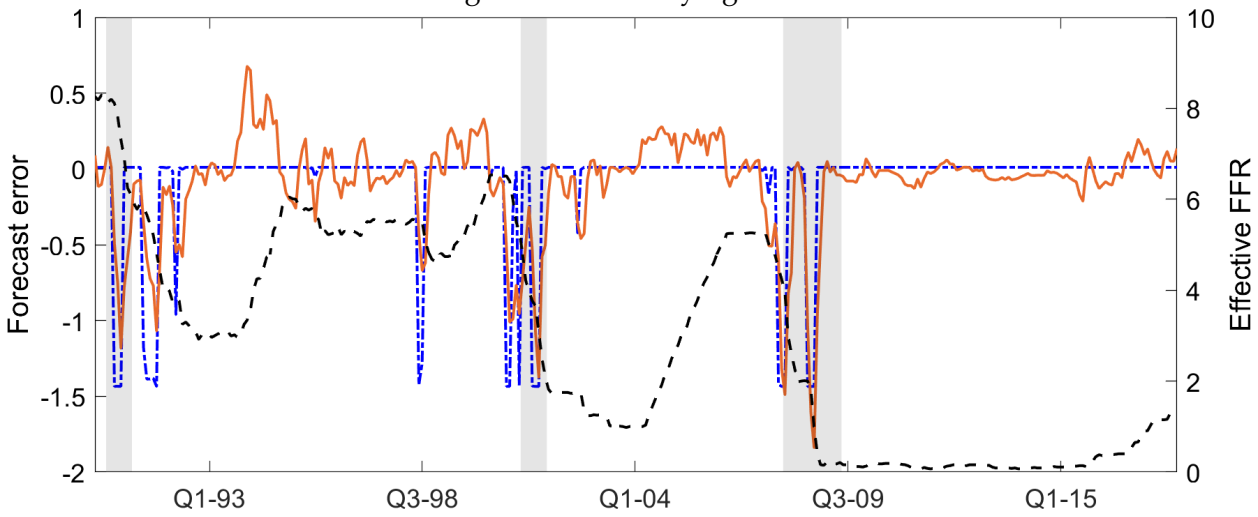
In comparison to eq. (18), West and McCracken's (1998) full-sample test for unbiasedness considers the null hypothesis H_0: $\mu = 0$ in the model $\epsilon_{t,h} = \mu + e_{t+h}$, where e_t is a zero mean

Table 7: AFE-H test results — three-month-ahead forecast error

Model	$\widehat{p}$	$\widehat{q}$	$\widehat{\mu}$	$\widehat{\mu}+\widehat{\mu}_s$	$\widehat{\phi}_1$	$\widehat{\phi}_2$	$\widehat{\phi}_3$	AFE-H	pvalue
AR(0)	0.96	0.92	-0.94	0.20	-	-	-	11.31	< 0.01
	(0.04)	(0.02)	(0.02)	(0.03)					
AR(1)	0.96	0.88	-0.31	0.14	0.74	-	-	8.49	< 0.01
	(0.04)	(0.02)	(0.01)	(0.03)	(0.03)				
AR(2)	0.96	0.88	0.01	-0.18	1.03	-0.31	-	7.74	< 0.01
	(0.08)	(0.02)	(0.03)	(0.01)	(0.04)	(0.04)			
AR(3)	0.96	0.88	0.01	-0.18	1.02	-0.26	-0.05	8.25	< 0.01
	(0.08)	(0.02)	(0.03)	(0.01)	(0.04)	(0.05)	(0.03)		

Note: The sample size is $T = 338$. The displayed coefficients correspond to the coefficients obtained when maximizing the likelihood over the finite grid of (p, q, μ, μ_s) of the AFE-H statistic. Numbers in parentheses denote robust standard errors. 'AFE-H' denotes the value of the test statistic. The column 'pvalue' denotes the p-value obtained from the simulated asymptotic distribution. The results for AFE-H are based on a 4-tuple of 12 equally-spaced grid points for $(p, q) \in [0.04, 0.96]$ and 20 equally-spaced grid points for $\mu, \mu_s \in [-1, 0.2] \times [-2, 0.4]$. p denotes the state-to-state transition probability for regime one and q denotes the state-to-state transition probability for regime two.

Figure 3: Time-varying mean



Note: The solid line (left hand side y-axis) displays the forecast error. The dashed line (left hand side y-axis) displays the time-varying unconditional mean of the MS-AR(3) model of Table 6, i.e. $\widehat{\mu}(1-\widehat{\phi}_1-\widehat{\phi}_2-\widehat{\phi}_3)^{-1}+\widehat{S}_{t+3}\widehat{\mu}_s(1-\widehat{\phi}_1-\widehat{\phi}_2-\widehat{\phi}_3)^{-1}$, using the smoothed state probabilities for $\widehat{S}_{t+3}$. The dashed line (right hand side y-axis) shows the FFR level and grey shaded areas display NBER recession periods.

error term. Applying West and McCracken (1998) to the three-month-ahead forecast error does not reject the null of $\mu = 0$; the p-value is around 0.6.[13]

The non-parametric Fluctuation test by Rossi and Sekhposyan (2016) rejects the null hypothesis of unbiasedness at the 5% level with the rolling window size m chosen to be at the $\frac{1}{3}$ of the total out-of-sample period. However, Markov switching model results can identify the potential states driving the bias, making our results more useful when trying to bias-correct forecasts or make

[13]These results hold when additionally controlling for lags of the forecast error and using HAC standard errors of Newey and West (1987) with a bandwidth $T^{(1/4)}$.

policy decisions in particular states of the world.

Our empirical results are closely related to Dahlhaus and Sekhposyan (2020). The authors evaluate forecast unbiasedness of FFR forecast errors of the BCFF and find that the bias seems to be mainly present in periods of monetary easing. However, since there is no common definition of "periods of monetary easing", the authors first have to define a state variable to identify their subsamples. In contrast, although the Markov switching approach proposed here finds similar periods of a negative forecast bias, it does so without having to define the state variable *ex-ante*. Instead, the periods are identified via the latent state of the regime-switching model.

Regime switching bias in market-based forecast errors: So far, our analysis focused on the BCFF forecasts, which are collected from prominent forecasters working in the financial sector. If their forecasts are indeed representative of what major financial institutions expect, then their forecasts could be correlated with the futures market's expectation of the FFR. Thus, we might expect similar regime switches and deviations for forecast unbiasedness in FFF as well.

To investigate whether this is the case, we constructed three- and six-month-ahead monthly forecast errors using prices of FFF for the period of January 1995 to February 2018.[14] FFF settle on the average effective FFR of the respective h-step-ahead target month and, therefore, provide a benchmark market-based measure of FFR expectations. We compute the h-step-ahead forecast error as the average effective FFR in month $t+h$ minus the FFF settlement price of the last trading day of month t. For instance, the March 31, 2006, settlement price of the three-month-ahead FFF is evaluated against the average effective FFR of June 2006.[15] Figure A.1 plots the forecast error based on the BCFF prediction (labeled FE-BCFF) against the forecast error based on the FFF prices (labeled FE-FFF) for the three-month-ahead periods. Note that the forecast errors of the BCFF and the FFF are very similar, suggesting that the information set of the BCFF panelists and the financial agents in futures market are indeed very similar. In the Online Appendix, we show that this is also the case for the six-month-ahead forecast errors.

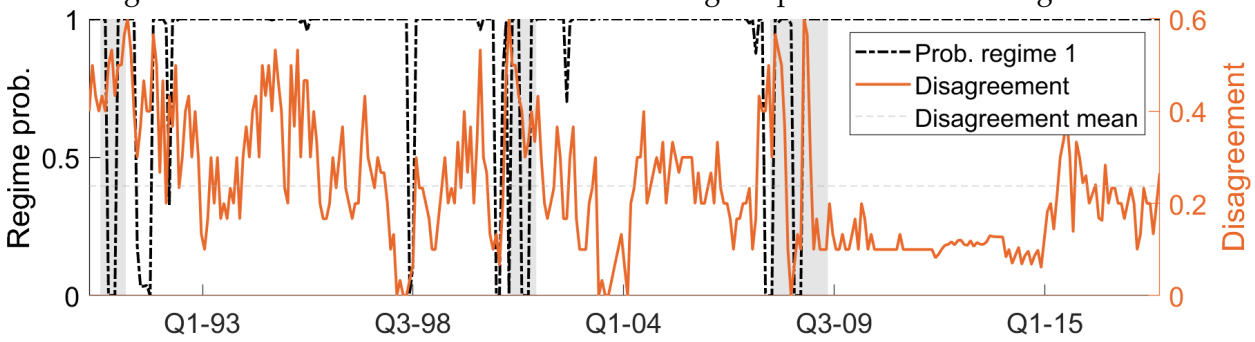
Table A.1 and Table A.2 in the Appendix show the results for testing for a Markov switching bias in the FFF implied forecast error. Results are very similar to the BCFF results displayed in Table 6 and Table 7, respectively. In fact, Figure A.2 shows that the respective regime probabilities estimated on the three-month-ahead forecast error produced by the FFF (solid line, RP-FFF) and by the BCFF (dashed line, RP-BCFF) are very similar.

[14] We start our sample in 1995 due to data availability.

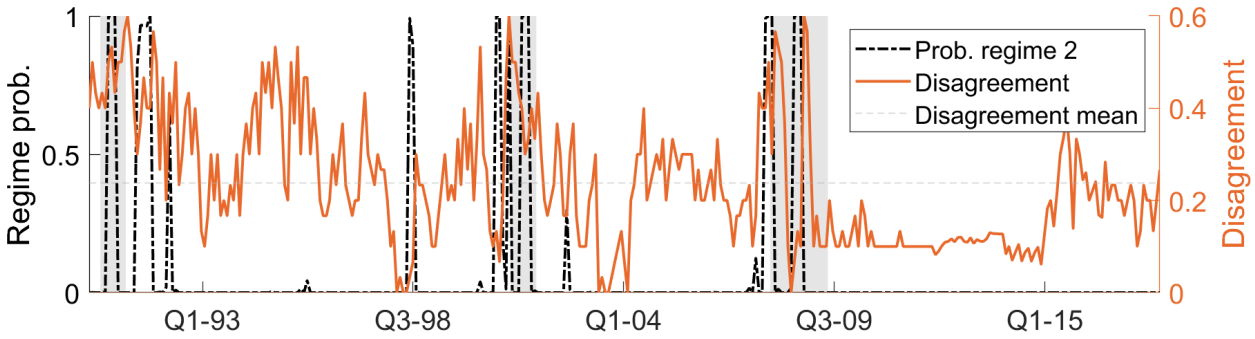
[15] Prices on non-trading days are substituted by the respective price of the most recent previous trading day (Swanson, 2006).

Regime switches and forecast disagreement: We also investigate whether the bias is related to the panelists' disagreement about the future FFR. In fact, if the forecast error biases are systematically correlated with disagreement, then point forecasts may reflect a shift in the marginal forecaster between "hawks", who always over-predict the interest rate, and "doves", who always under-predict the interest rate. To that end, we computed the difference between the top-10-average and bottom-10-average forecasts of the panelists as in Andrade et al. (2016) and Dahlhaus and Sekhposyan (2020). Figure 4 shows plots the disagreement of the forecasters (dashed line, right y-axis) against the regime probabilities (solid line, left y-axis). The figure suggests that while sometimes the disagreement and the regime probabilities co-move, there is no systematic correlation between disagreement and point forecast errors.

Figure 4: Three-month-ahead forecast error: regime probabilities vs disagreement



(a) Regime one vs disagreement

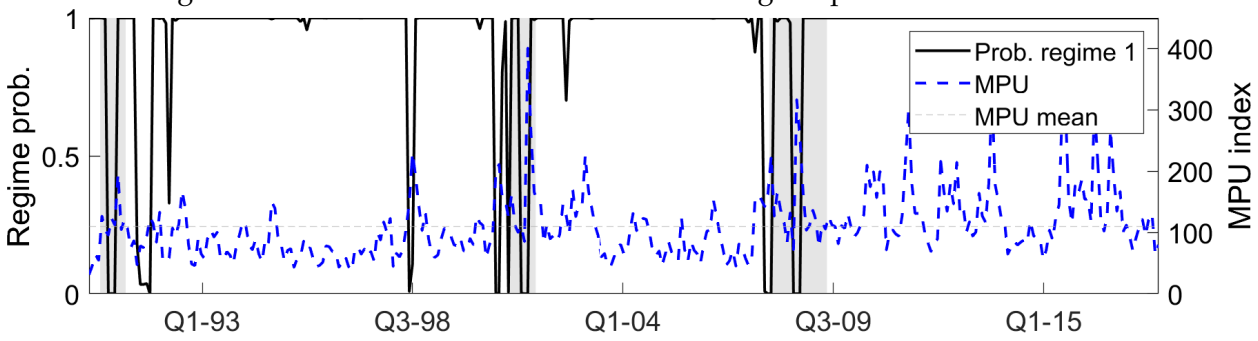


(b) Regime two vs disagreement

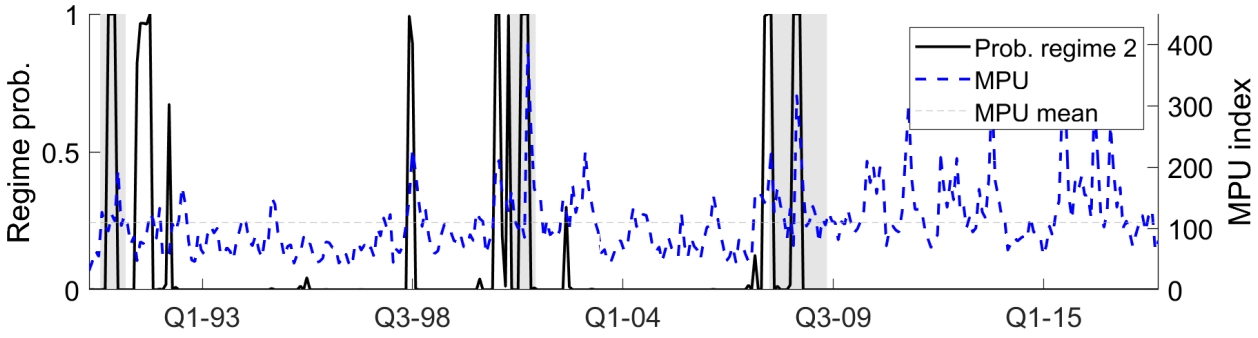
Note: The left y-axis shows the scale of the smoothed regime probability, estimated using the model defined in eq. (18), with three lags, on the BCFF FFR forecast error. The right y-axis shows the disagreement between BCFF panelists. Grey shaded areas display NBER recession periods.

Regime switches and monetary policy uncertainty: Moreover, we analyze whether the bias of BCFF panelists' FFR forecast is more generally related to uncertainty about monetary policy. Our measure of monetary policy uncertainty is the "MPU" index of Baker et al. (2016), which is constructed using newspaper articles of the 10 major U.S. newspapers. Figure 5 plots the estimated regime probabilities (solid line, left y-axis) against the MPU index (dashed line, right

Figure 5: Three-month-ahead forecast error: regime probabilities vs MPU



(a) Regime one vs MPU



(b) Regime two vs MPU

Note: The left y-axis shows the scale of the smoothed regime probability, estimated using the model defined in eq. (18), with three lags, on the BCFF FFR forecast error. The right y-axis shows the MPU index. Grey shaded areas display NBER recession periods.

y-axis). Spikes in the MPU index before the zero lower bound (ZLB) period tend to coincide with an increase in the probability of the second regime, notably around the two U.S. recessions in our sample as well as in 1998 around the Fed intervention triggered by the collapse of Long Term Capital Management.

Forecast errors of real GDP and GDP deflator growth rates: The BCFF panelists also provide forecasts for U.S. real GDP and the GDP deflator growth (from hereon referred to as inflation). To investigate whether the bias in the FFR forecast is associated with a bias in the corresponding macroeconomic forecasts, we computed the average forecast error of real GDP growth and inflation conditional on the regimes estimated on the BCFF FFR forecast errors, denoted by $\widehat{S}^{\text{FFR}}_{t+h}$. Then, we estimate the following regression for the forecast error of real GDP growth and inflation:

$$\epsilon_{t,h} = \mu + \mu_s \widehat{S}^{\text{FFR}}_{t+h} + \sum_{i=1}^{3} \phi_i \epsilon_{t-i,h} + e_{t+h}, \tag{19}$$

where $e_{t+h} \sim N(0, \sigma_e^2)$. Results are reported in Table 8, which shows the estimated coefficients, $\widehat{\mu}$ and $\widehat{\mu} + \widehat{\mu}_s$, and the p-value of a t-test of $\mu = 0$ and an F-test on $\mu + \mu_s = 0$. The point estimates

of $\widehat{\mu} + \widehat{\mu}_s$ are negative for both the real GDP growth and inflation forecast errors, i.e. periods of overestimation of the FFR coincide with periods of overestimation of real GDP growth and inflation. Note, however, that we cannot reject the null hypothesis of $\mu + \mu_s = 0$ at conventional significance levels and that part of these results are driven by the large negative forecast error during the Great Recession of 2008 to 2009. For the GDP deflator we find that $\widehat{\mu}$ is also negative and significantly different from zero, pointing to a potential constant bias in the forecast error.

Table 8: Results for output growth and inflation

	GDP growth		Inflation	
	$\widehat{\mu}$	$\widehat{\mu} + \widehat{\mu}_s$	$\widehat{\mu}$	$\widehat{\mu} + \widehat{\mu}_s$
Parameter values	0.030	-0.426	-0.112	-0.375
	(0.316)	(0.117)	(0.002)	(0.122)

Note: The sample size is $T = 338$. Values in parenthesis denote p-values based on the HAC variance estimator of Newey and West (1987) with a bandwidth of $T^{1/4}$. Inflation measures the growth rate of the GDP deflator.

Additional robustness analyses: In the Online Appendix, we show that our results are robust to both the exclusion of the ZLB period after the Great Recession and the use of the six-month-ahead (instead of the three-month-ahead) forecast errors.

Besides addressing the question of what causes the bias, the results also have potential implications for monetary policy communication. While prior to the 1990s many in the Fed believed that policy effectiveness depended on surprising the market (Poole, 2005), the current consensus is rather along the opposite lines: it is a central bank's job to transparently manage expectations (see Woodford, 2005 for a discussion). In the words of Goodfriend (1991): "By making itself more predictable to the markets, the central bank makes market reactions to monetary policy more predictable to itself. And that makes it possible to do a better job of managing the economy." From that perspective, a systematic overprediction of the policy rate during monetary easings suggests that there is room for improvements in the Fed's communication strategy.

6 Conclusion

Despite ample evidence on state dependence in prediction errors, existing forecast rationality tests either rely on non-parametric techniques to account for the time-variation or treat the states as observable when thinking of forecast optimality. We propose a framework for forecast evaluation that is able to detect state-dependent deviations from forecast rationality where the states are unknown *a priori*. Overall, our tests exhibit good size and power properties in Monte Carlo

simulations, although they somewhat underreject when testing forecast efficiency. We show that, in the presence of Markov switching, the new tests outperform available alternatives, which in general have weak power when the time-variation takes the form of regime-switching.

While we focus on a two-state Markov switching structure, we expect our results to generalize to n-states. We leave this analysis to future work due to the fact that, in practice, Markov switching models are most commonly estimated in a two-state environment and that the computational costs make implementation in the presence of more than two states difficult in practice .

In an empirical investigation of the forecast unbiasedness of the Blue Chip Financial Forecasts survey, for the sample period from 1990 to 2018, our results show that the predictions exhibit a Markov switching bias when forecasting the three- and six-month-ahead Federal Funds Rate. While we find no evidence in favor of a constant deviation from unbiasedness in the full sample, we do provide evidence that participants tend to systematically overestimate the Federal Funds Rate in monetary easing episodes. We show that a similar state-dependent bias is also present in market-based forecasts of interest rates, but not in the forecasts of real GDP growth and GDP deflator-based inflation.

References

Andrade, P., Crump, R.K., Eusepi, S., Moench, E., 2016. Fundamental disagreement. Journal of Monetary Economics 83, 106–128.

Baker, S., Bloom, N., Davis, S.J., 2016. Measuring Economic Policy Uncertainty. Quarterly Journal of Economics 131, 1593–1636.

Bullard, J., 2016. The St. Louis Fed's New Characterization of the Outlook for the U.S. Economy. Commentary, Federal Reserve Bank of St. Louis.

Carrasco, M., Hu, L., Ploberger, W., 2014. Optimal Test for Markov Switching Parameters. Econometrica 82, 765–784.

Carter, A., Steigerwald, D., 2012. Testing for regime switching. Econometrica 80, 1809–1812.

Chang, Y., Choi, Y., Park, J.Y., 2017. A New Approach to Model Regime Switching. Journal of Econometrics 196, 127–143.

Cho, J.S., White, H., 2007. Testing for Regime Switching. Econometrica 75, 1671–1720.

Chun, A.L., 2011. Expectations, Bond Yields, and Monetary Policy. Review of Financial Studies 24, 208–247.

Croushore, D., 2012. Forecast Bias in Two Dimensions. Federal Reserve Bank of Philadelphia Working Paper 12-9.

Dahlhaus, T., Sekhposyan, T., 2020. Survey-based Monetary Policy Uncertainty and its Asymmetric Effects. Working Paper.

Davies, R.B., 1977. Hypothesis Testing When a Nuisance Parameter is Present Only under the Alternative. Biometrika 64, 247–254.

Davies, R.B., 1987. Hypothesis Testing When a Nuisance Parameter is Present Only under the Alternative. Biometrika 74, 33–43.

Garcia, R., 1998. Asymptotic Null Distribution of the Likelihood Ratio Test in Markov Switching Models. International Economic Review 39, 763–788.

Goodfriend, M., 1991. Interest Rates and the Conduct of Monetary Policy. Carnegie-Rochester Conference on Public Policy 34, 7–30.

Granziera, E., Jalasjoki, P., Paloviita, M., 2021. The Bias and Efficiency of the ECB Inflation Projections: a State Dependent Analysis. Norges Bank Working Paper 1/2021.

Hamilton, J.D., 1990. Analysis of Time Series Subject to Changes in Regime. Journal of Econometrics 45, 39–70.

Hansen, B.E., 1992. The Likelihood Ratio Test under Non-Standard Conditions: Testing the Markov switching model of GNP. Journal of Applied Econometrics 7, 61–82.

Hansen, B.E., 1996. Erratum: The Likelihood Ratio Test under Non-Standard Conditions: Testing the Markov switching model of GNP. Journal of Applied Econometrics 11, 195–198.

Joutz, F., Stekler, H., 2000. An Evaluation of the Predictions of the Federal Reserve. International Journal of Forecasting 16, 17–38.

Mincer, J., Zarnowitz, V., 1969. The Evaluation of Economic Forecasts, in: JA, M. (Ed.), Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance. National Bureau of Economic Research: New York.

Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica 55, 703–708.

Poole, W., 2005. Remarks: Panel on "After Greenspan: Whither Fed Policy?", remarks delivered at the Western Economics Association International Conference, San Francisco, California.

Qu, Z., Zhuo, F., 2021. Likelihood Ratio Based Tests for Markov Regime Switching. The Review of Economic Studies 88, 937–968.

Rossi, B., 2013. Advances in Forecasting under Model Instability, in: Elliot, G., Tmmermann, A. (Eds.), Handbook of Economic Forecasting. Elsevier Publications. volume 2b. chapter 21, pp. 1203–1324.

Rossi, B., Sekhposyan, T., 2016. Forecast Rationality Tests in the Presence of Instabilities, With Applications to Federal Reserve and Survey Forecasts. Journal of Applied Econometrics 31, 507–532.

Sinclair, T.M., Joutz, F., Stekler, H., 2010. Can the Fed Predict the State of the Economy? Economic Letters 108, 28–32.

Swanson, E., 2006. Have increases in Federal Reserve transparency improved private sector interest rate forecasts? Journal of Money, Credit, and Banking 38, 791–819.

West, K.D., McCracken, M.W., 1998. Regression-based Tests of Predictive Ability. International Economic Review 39, 817–840.

Woodford, M., 2005. Central Bank Communication and Policy Effectiveness. Proceedings - Economic Policy Symposium - Jackson Hole, Federal Reserve Bank of Kansas City , 399–474.

A Empirical results using Federal Funds Futures

Table A.1: AFE-BS test results — three-month-ahead forecast error by federal funds futures

Model	$\widehat{p}$	$\widehat{q}$	$\widehat{\mu}$	$\widehat{\mu}+\widehat{\mu}_s$	$\widehat{\phi}_1$	$\widehat{\phi}_2$	$\widehat{\phi}_3$	LR-value	pvalue
AR(0)	0.98	0.98	-0.75	0.01	-	-	-	112.54	< 0.01
	(0.01)	(0.01)	(0.02)	(0.01)					
AR(1)	0.98	0.51	0.01	-0.63	0.59	-	-	80.51	< 0.01
	(0.01)	(0.17)	(0.01)	(0.02)	(0.01)				
AR(2)	0.97	0.51	0.00	-0.59	0.73	-0.16	-	73.69	< 0.01
	(0.01)	(0.16)	(0.01)	(0.02)	(0.03)	(0.03)			
AR(3)	0.97	0.52	0.00	-0.59	0.71	-0.12	-0.04	72.80	< 0.01
	(0.01)	(0.15)	(0.01)	(0.02)	(0.03)	(0.04)	(0.03)		

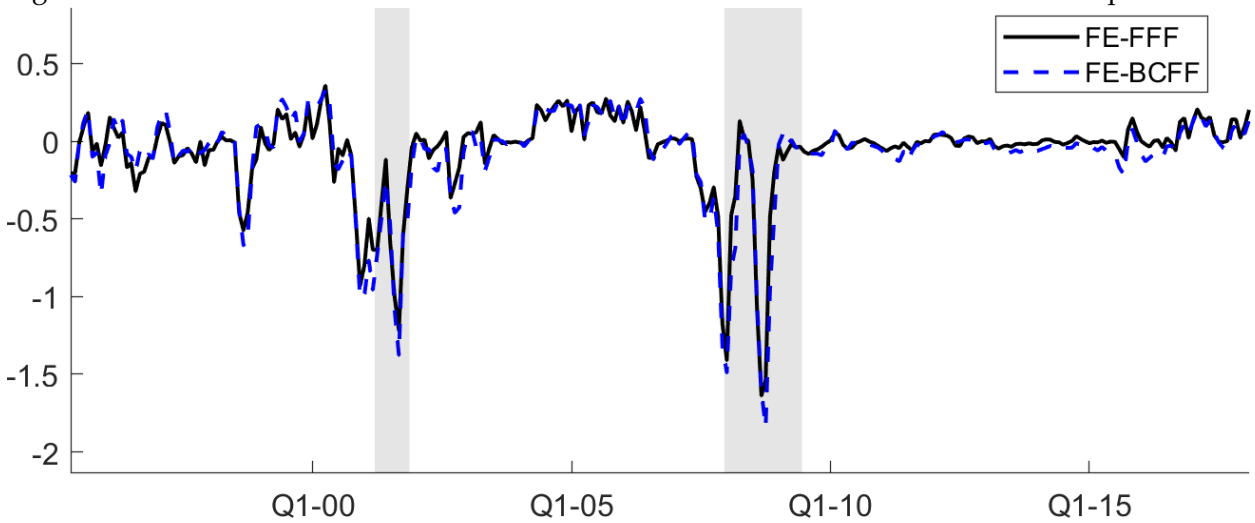
Note: The sample size is $T = 277$. The displayed coefficients correspond to the maximimum obtained under the alternative, using the restriction that $(p, q) \in \Lambda_{(p,q)}$. The column labelled 'LR-value' denotes the value of the likelihood ratio. Numbers in parentheses denote robust standard errors. The column 'pvalue' denotes the p-value obtained using the approximated asymptotic distribution based on 200 bootstrap replications. p denotes the state-to-state transition probability for regime one and q denotes the state-to-state transition probability for regime two.

Table A.2: AFE-H test results — three-month-ahead forecast error by federal funds futures

Model	$\widehat{p}$	$\widehat{q}$	$\widehat{\mu}$	$\widehat{\mu}+\widehat{\mu}_s$	$\widehat{\phi}_1$	$\widehat{\phi}_2$	$\widehat{\phi}_3$	AFE-H	pvalue
AR(0)	0.92	0.88	-0.05	0.01	-	-	-	11.01	< 0.01
	(0.96)	(0.75)	(0.21)	(0.10)					
AR(1)	0.96	0.88	-0.31	-0.49	0.54	-	-	9.11	< 0.01
	(1.61)	(0.82)	(0.80)	(0.25)	(0.27)				
AR(2)	0.96	0.88	-0.05	0.01	1.03	-0.33	-	9.10	< 0.01
	(0.40)	(0.28)	(0.13)	(0.09)	(0.06)	(0.04)			
AR(3)	0.96	0.88	0.01	-0.68	0.66	-0.09	-0.04	9.22	< 0.01
	(0.07)	(0.02)	(0.02)	(0.01)	(0.03)	(0.05)	(0.03)		

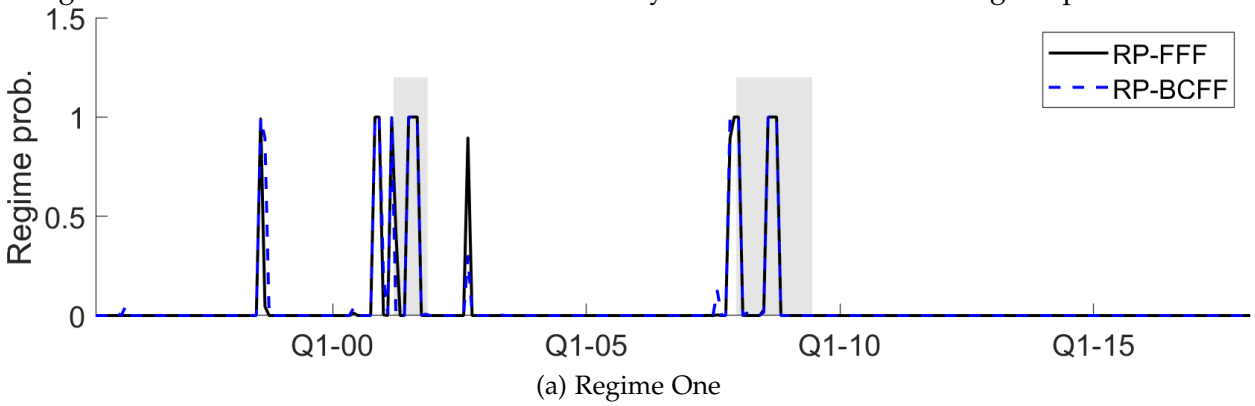
Note: The sample size is $T = 277$. The displayed coefficients correspond to the coefficients obtained when maximizing the likelihood over the finite grid of (p, q, μ, μ_s) of the AFE-H statistic. Numbers in parentheses denote robust standard errors. 'AFE-H' denotes the value of the test statistic. The column 'pvalue' denotes the p-value obtained from the simulated asymptotic distribution. The results for AFE-H are based on a 4-tuple of 12 equally-spaced grid points for $(p, q) \in [0.04, 0.96]$ and 20 equally-spaced grid points for $\mu, \mu_s \in [-1, 0.2] \times [-2, 0.4]$. p denotes the state-to-state transition probability for regime one and q denotes the state-to-state transition probability for regime two.

Figure A.1: Three-month-ahead forecast error based on federal funds futures and BCFF predictions

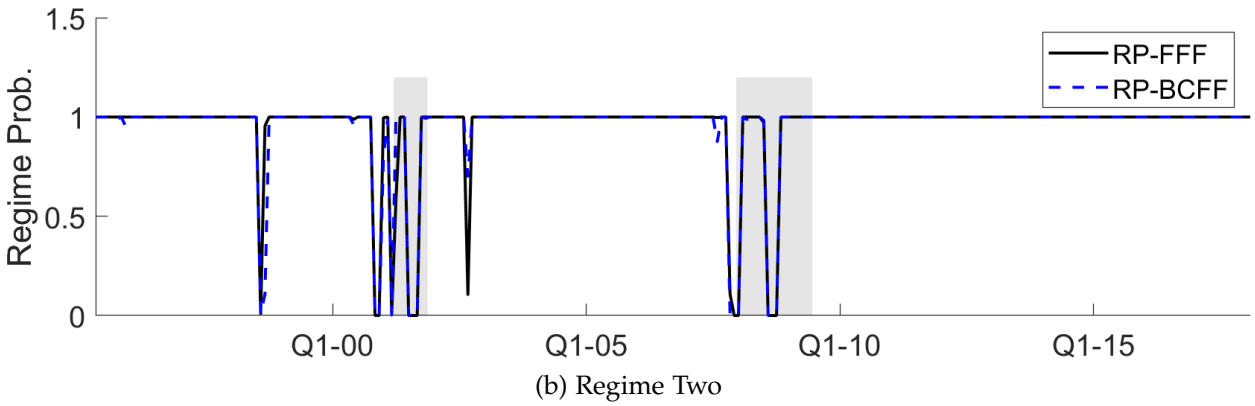


Note: The solid line displays the forecast error implied by the FFF, denoted by 'FE-FFF', whereas the dashed line displays the forecast error when using the BCFF survey forecasts, denoted by 'FE-BCFF'. Grey shaded areas display NBER recession periods.

Figure A.2: Three-month-ahead forecast error by federal funds futures: regime probabilities



(a) Regime One



(b) Regime Two

Note :The solid lines display the smoothed probabilities of the regimes estimated on the forecast errors implied by the FFF, denoted by 'RP-FFF', whereas the dashed lines display the regimes estimated on the forecast errors of the BCFF survey forecasts. In both cases we obtained the smoothed regime probabilities from the Markov switching model, defined in eq. (18), with three lags. Grey shaded areas display NBER recession periods.