Comparing Density Forecasts via Weighted Likelihood Ratio Tests*

Gianni Amisano† and **Raffaella Giacomini**‡

University of Brescia, Italy and University of California, Los Angeles

This version: February 2005

Abstract

We propose a test for comparing the out-of-sample accuracy of competing density forecasts of a variable. The test is valid under general conditions: the data can be heterogeneous and the forecasts can be based on (nested or non-nested) parametric models or produced by semi-parametric, non-parametric or Bayesian estimation techniques. The evaluation is based on scoring rules, which are loss functions defined over the density forecast and the realizations of the variable. We restrict attention to the logarithmic scoring rule and propose an out-of-sample 'weighted likelihood ratio' test that compares weighted averages of the scores for the competing forecasts. The user-defined weights are a way to focus attention on different regions of the distribution of the variable. For a uniform weight function, the test can be interpreted as an extension of Vuong (1989)'s likelihood ratio test to time series data and to an out-of-sample testing framework. We apply the tests to evaluate density forecasts of US inflation produced by linear and Markov Switching Phillips curve models estimated by either maximum likelihood or Bayesian methods. We conclude that a Markov Switching Phillips curve estimated by maximum likelihood produces the best density forecasts of inflation.

*We are deeply indebted to Clive W. J. Granger for many interesting discussions. We also thank Carlos Capistran, Roberto Casarin, Graham Elliott, Ivana Komunjer, Andrew Patton, Kevin Sheppard, Allan Timmermann and seminar participants at the University of Brescia for valuable comments. This paper was previously circulated under the title "Comparing Density Forecasts via Weighted Likelihood Ratio Tests: Asymptotic and Bootstrap Methods", by Raffaella Giacomini.

†amisano@eco.unibs.it

‡giacomin@econ.ucla.edu

1 Introduction

A density forecast is an estimate of the future probability distribution of a random variable, conditional on the information available at the time the forecast is made. It thus represents a complete characterization of the uncertainty associated with the forecast, as opposed to a point forecast, which provides no information about the uncertainty of the prediction.

Density forecasting is receiving increasing attention in both macroeconomics and finance (see Tay and Wallis, 2000 for a survey). A famous example of density forecasting in macroeconomics is the 'fan-chart' of inflation and GDP published by the Bank of England and by the Sveriges Riksbank in Sweden in their quarterly Inflation Reports (for other examples of density forecasting in macroeconomics, see also Diebold, Tay and Wallis, 1999 and Clements and Smith, 2000). In finance, where the wide availability of data and the increasing computational power make it possible to produce more accurate estimates of densities, the examples are numerous. Leading cases are in risk management, where forecasts of portfolio distributions are issued with the purpose of tracking measures of portfolio risk such as the Value-at-Risk (see, e.g., Duffie and Pan, 1996) or the Expected Shortfall (see, e.g., Artzner *et al.,* 1997*).* Another example is the extraction of density forecasts from option price data (see, e.g. Soderlind and Svensson, 1997). The vast literature on forecasting volatility with GARCH-type models (see Bollerslev, Engle and Nelson, 1994) and its extensions to forecasting higher moments of the conditional distribution (see Hansen, 1994) can also be seen as precursors to density forecasting. The use of sophisticated distributions for the standardized residuals of a GARCH model and the modeling of time dependence in higher moments is in many cases an attempt to capture relevant features of the data to better approximate the true distribution of the variable. Finally, in the multivariate context, a focus on densities is the central issue in the literature on copula modeling and forecasting, that is gaining interest in financial econometrics (Patton, 2001).

With density forecasting becoming more and more widespread in applied econometrics, it is necessary to develop reliable techniques to evaluate the forecasts' performance. The literature on evaluation of density forecasts (or, equivalently, predictive distributions) is still young, but growing at fast speed. In one of the earliest contributions, Diebold, Gunther and Tay (1998) suggested evaluating a sequence of density forecasts by assessing whether the probability integral transforms of the realizations of the variable with respect to the forecast densities are independent and identically distributed (i.i.d.) $U(0,1)$. While Diebold *et al.* (1998) adopted mainly qualitative tools for testing the i.i.d. $U(0,1)$ behavior of the transformed data, formal tests of the same hypothesis have been suggested by Berkowitz (2001), Hong and White (2000), Hong (2001). Tests that account for parameter estimation uncertainty have been proposed by Hong and Li (2001), Bai (2003) and Corradi and Swanson (2003) (the latter further allowing for dynamic misspecification under the null hypothesis).

It is important to emphasize that the above methods focus on absolute evaluation, that is, on evaluating the 'goodness' of a given sequence of density forecasts, relative to the data-generating

process. In practice, however, it is likely that any econometric model used to produce the sequence of density forecasts is misspecified, and an absolute test will typically give no guidance to the user as to what to do in case of rejection. In this situation a more practically relevant question is how to decide which of two (or more) competing density forecasts is preferable, given a measure of accuracy. The comparative evaluation of density forecasts has been relatively less explored in the literature. A number of empirical works have considered comparisons of density forecasts (e.g., Clements and Smith, 2000; Weigend and Shi, 2000), but they have predominantly relied on informal assessment of predictive accuracy. More recently, Corradi and Swanson (2004a, 2004b), proposed a bootstrap-based test for evaluating multiple misspecified predictive models, based on a generalization of the mean squared error.

In this paper, we contribute to the literature about comparative evaluation of density forecasts by proposing formal out-of-sample tests for ranking competing density forecasts that are valid under very general conditions. Our method is an alternative approach to Corradi and Swanson (2004a, 2004b), since it uses a different measure of accuracy, and is further valid under more general data and estimation assumptions.

We consider the situation of a user who is interested in comparing the out-of-sample accuracy of two competing *forecast methods,* where we define the forecast method to be the set of choices that the forecaster makes at the time of the prediction, including the density model, the estimation procedure and the estimation window. We impose very few restrictions on the forecast methods. The density forecasts can be based on parametric models, either nested or non-nested, whose parameters are known or estimated. The forecasts could be further produced using semi-parametric, non-parametric or Bayesian estimation techniques. As in Giacomini and White (2004), the key requirement is that the forecasts are based on a finite estimation window. This assumption is motivated by our explicit allowance for a data-generating process that may change over time (unlike all of the existing literature, which assumes stationarity), and it further allows us to derive our tests in an environment with asymptotically non-vanishing estimation uncertainty. Note that the finite estimation window assumption has important consequences for the forecasting schemes used in the out-of-sample evaluation exercise. In particular, an expanding estimation window scheme is not allowed, whereas fixed- or rolling estimation window schemes satisfy the requirement.

We follow the literature on probability forecast evaluation and measure the relative performance of the density forecasts using scoring rules, which are loss functions defined over the density forecast and the outcome of the variable. We restrict attention to the logarithmic scoring rule, which has an intuitively appealing interpretation and is mathematically convenient. We consider the out-of-sample performance of the two forecasts and rank them according to the relative magnitude of a weighted average of the logarithmic scores over the out-of-sample period. Our weighted likelihood ratio test establishes whether such weighted averages are significantly different from each other. We show how the use of weights gives our tests flexibility by allowing the user to compare the performance of the

density forecasts in different regions of the unconditional distribution of the variable, distinguishing for example predictive ability in 'normal' periods from that in 'extreme' periods. In the equal-weights case our test is related to Vuong's (1989) likelihood ratio test for non-nested hypotheses, the differences being that (1) we compare forecast methods rather than models; (2) we allow the underlying models to be either nested or non-nested; (3) we perform the evaluation out-of-sample rather than in-sample; (4) we allow the data to be heterogeneous and dependent rather than i.i.d.

We conclude with an application comparing the performance of density forecasts produced by a linear Phillips curve model of inflation versus a Markov-switching Phillips curve, estimated by either maximum likelihood or Bayesian methods. The focus on the density forecast - rather than point forecast - performance of Markov-switching versus linear models of inflation is empirically relevant and can shed new light on the relative forecast accuracy of linear and non-linear models since the two models imply different densities for inflation (a mixture of normals for the Markov-switching versus a normal density for the linear model).

Our paper is organized as follows: Section 2 describes the notation and the testing environment; Section 3 introduces the loss functions and Section 4 introduces our weighted likelihood ratio test. Section 5 contains a Monte Carlo experiment investigating the empirical size and power properties of our test. In Section 6 we apply our tests to the evaluation of competing density forecasts of inflation obtained by a Markov-switching or a linear Phillips curve model, estimated by either classical or Bayesian methods. Section 7 concludes. The proofs are in Section 8.

2 Description of the environment

Consider a stochastic process $Z \equiv \{Z_t : \Omega \longrightarrow \mathbb{R}^{s+1}, s \in \mathbb{N}, t = 1, \ldots, T\}$ defined on a complete probability space $(\Omega, \mathcal{F}, P)$ and partition the observed vector Z_t as $Z_t \equiv (Y_t, X_t')'$, where $Y_t : \Omega \to \mathbb{R}$ is the variable of interest and $X_t : \Omega \to \mathbb{R}^s$ is a vector of predictors. Let $\mathcal{F}_t = \sigma(Z_1', ..., Z_t')'$ be the information set at time t and suppose that two competing models are used to produce density forecasts of the variable of interest, Y_{t+1}, using the information in $\mathcal{F}_t$. Denote the forecasts by $\hat{f}_{m,t} \equiv f(Z_t, Z_{t-1}, ..., Z_{t-m+1}; \hat{\beta}_{m,t})$ and $\hat{g}_{m,t} \equiv g(Z_t, Z_{t-1}, ..., Z_{t-m+1}; \hat{\beta}_{m,t})$, where f and g are measurable functions. The subscripts indicate that the time-t forecasts are measurable functions of the m most recent observations, where $m < \infty$.

The $k \times 1$ vector $\hat{\beta}_{m,t}$ collects the parameter estimates from both models. Note that the only requirement that we impose on how the forecasts are produced is that they are measurable functions of a finite estimation window. In particular, this allows the forecasts to be produced by parametric as well as semi-parametric, non-parametric or Bayesian estimation methods.

We perform the out-of-sample evaluation using a "rolling window" estimation scheme. Let T be the total sample size. The first one-step-ahead forecasts are produced at time m, using data indexed $1, ..., m$ and they are compared to Y_{m+1}. The estimation window is then rolled forward one step and

the second set of forecasts are obtained using observations $2, ..., m+1$ and compared to Y_{m+2}. This procedure is thus iterated and the last forecasts are obtained utilizing observations $T-m, ..., T-1$ and they are compared to Y_T. This yields a sequence of $n \equiv T-m$ out-of-sample density forecasts.[1]

Note that the condition that m be finite rules out an expanding estimation window forecasting scheme.[2] This condition is however compatible with a fixed estimation sample forecasting scheme, where all n out-of-sample forecasts depend on the same parameters estimated once on the first m observations. For clarity of exposition, we hereafter restrict attention to a rolling window forecasting scheme but all the results remain valid for a fixed estimation sample scheme.

All of the above elements - the model, the estimation method and the size of the estimation window - constitute what we call the "forecast method", which is the object of our evaluation.

3 Loss functions and density forecasting

There is a large literature on loss functions for evaluation of point forecasts (e.g., Christoffersen and Diebold, 1997). In this section, we explore the possibility of incorporating loss functions into the evaluation of density forecasts, and argue that the standard framework of loss functions for forecast evaluation is not appropriate when the object to be forecasted is a conditional density.

The incorporation of loss functions into the forecasting problem has until now focused on the definition of classes of loss functions of the form $L(\hat{f}_{t,\tau}, Y_{t+\tau})$, where $\hat{f}_{t,\tau}$ is a τ-step-ahead point forecast of $Y_{t+\tau}$. In the vast majority of cases, the loss function is assumed to only depend on the forecast error, as for quadratic loss or general asymmetric loss (e.g., Christoffersen and Diebold, 1997, Weiss, 1996). Weiss (1996) shows that, in this framework, the optimal predictor is some summary measure of the true conditional density of the variable $Y_{t+\tau}$ (the mean for quadratic loss, the median for absolute error loss, etc.). This means that a user with, say, a quadratic loss function is only concerned with the accuracy of the mean prediction and will be indifferent among density forecasts that yield the same forecast for the conditional mean. As a consequence, in this situation it becomes unnecessary to issue a density forecast in the first place, and the forecaster should only concentrate on accurately forecasting the relevant summary measure of the true density. The discussion of loss functions relevant for density forecasting must thus involve a shift of focus.

Since a density forecast can be seen as a collection of probabilities assigned by the forecaster to all attainable events, the tools developed in the probability forecasting evaluation literature can be readily employed. In particular, we will make use of so-called scoring rules (see, e.g., Winkler, 1967, Diebold and Lopez, 1996, Lopez, 2001), which are loss functions whose arguments are the density

[1] In principle, the estimation window lengths can vary over time, but for simplicity we express each forecast as a function of m, which can be thought of as the maximum.

[2] In principle, we could allow for a recursive forecasting scheme with estimation window whose size grows more slowly than the out-of-sample size, but at the cost of added technical difficulty. See Giacomini and White (2003) for further discussion.

forecast and the actual outcome of the variable. We restrict attention to the logarithmic scoring rule $S(f, Y) = \log f(Y)$, where Y is the observed value of the variable and $f(\cdot)$ the density forecast.[3] Intuitively, the logarithmic score rewards a density forecast that assigns high probability to the event that actually occurred. The logarithmic score is also mathematically convenient, being the only scoring rule that is solely a function of the value of the density at the realization of the variable.

When a sequence of alternative density forecasts and of realizations of the variable is available, one can rank the density forecasts by comparing the average scores for each forecast. For the two sequences of density forecasts f and g introduced in Section 2, one would compute the average logarithmic scores over the out-of-sample period as $n^{-1}\sum_{t=m}^{T-1}\log \hat{f}_{m,t}(Y_{t+1})$ and $n^{-1}\sum_{t=m}^{T-1}\log \hat{g}_{m,t}(Y_{t+1})$, and select the forecast yielding the highest score.

In this paper, we suggest a more general approach which involves considering a *weighted* average of the scores over the out-of-sample period. The idea is that a user might be especially interested in a density forecast that is accurate in predicting events that lay in a particular region of the unconditional distribution of the variable. An example could be a user who is interested in predicting (loosely defined) tail events, as in the case when different investment strategies or policy implications would arise if the future realizations of the variable fall into the tails of the distribution. If the user is presented with two alternative density forecasts, he might then want to place greater emphasis on the performance of the competing models in the tails of the distribution, and less emphasis on what happens in the center. Another situation that might be of interest is a focus on predicting events that fall near the unconditional mean of the variable, as a way to ignore the influence of possible outliers on predictive performance. Finally, one might want to separate the predictive performance of the models in the right and in the left tail of the distribution, as in the case, e.g., of forecasting models for risk management, where losses have different implications than gains.

For each of the above situations, we can define an appropriate weight function $w(\cdot)$ and compare the weighted average scores $n^{-1}\sum_{t=m}^{T-1} w(Y_{t+1})\log \hat{f}_{m,t}(Y_{t+1})$ and $n^{-1}\sum_{t=m}^{T-1} w(Y_{t+1})\ \log \hat{g}_{m,t}(Y_{t+1})$. The weight function $w(\cdot)$ can be arbitrarily chosen by the forecaster to select the desired region of the unconditional distribution of Y_t. The only requirement imposed on the weight function are that it is positive and bounded. For example, when the data have unconditional mean 0 and variance 1, one could consider the following weight functions.

- Center of distribution: $w_1(Y) = \phi(Y)$, ϕ standard normal density function (or pdf)
- Tails of distribution: $w_2(Y) = 1 - \phi(Y)/\phi(0)$, ϕ standard normal pdf
- Right tail: $w_3(Y) = \Phi(Y)$, Φ standard normal distribution function (or cdf)
- Left tail: $w_4(Y) = 1 - \Phi(Y)$, Φ standard normal cdf

[3] A scoring rule is usually expressed as a gain, rather than a loss. In spite of this, we will continue referring to scoring rules as loss functions.

Plots of $w_1 - w_4$ are shown in Figure 1.

[FIGURE 1 HERE]

A formal test for comparing the weighted average logarithmic scores is proposed in the following section.

4 Weighted likelihood ratio tests

For a given weight function $w(\cdot)$ and two alternative conditional density forecasts f and g for Y_{t+1}, let

$$WLR_{m,t+1} \equiv w(Y_{t+1}^{st}) \left(\log \hat{f}_{m,t}(Y_{t+1}) - \log \hat{g}_{m,t}(Y_{t+1}) \right), \tag{1}$$

where $Y_{t+1}^{st} \equiv (Y_{t+1} - \hat{\mu}_{m,t})/\hat{\sigma}_{m,t}$ is the realization of the variable at time $t+1$, standardized using estimates of the unconditional mean and standard deviation of Y_t, $\hat{\mu}_{m,t}$ and $\hat{\sigma}_{m,t}$, computed on the same sample on which the density forecasts are estimated. A test for equal performance of density forecasts f and g can be formulated as a test of the null hypothesis

$$H_0 \quad : \quad E[WLR_{m,t+1}] = 0, \ t = 1, 2, ... \text{ against} \tag{2}$$

$$H_A \quad : \quad E[\overline{WLR}_{m,n}] \neq 0 \text{ for all } n \text{ suff. large}, \tag{3}$$

where $\overline{WLR}_{m,n} = n^{-1} \sum_{t=m}^{T-1} WLR_{m,t+1}$. Note that this formulation of the null and alternative hypotheses reflects the fact that we do not impose the requirement of stationarity of the data. We call a test of H_0 a weighted likelihood ratio test.

Our test is based on the statistic

$$t_{m,n} = \frac{\overline{WLR}_{m,n}}{\hat{\sigma}_n/\sqrt{n}} \tag{4}$$

where $\hat{\sigma}_n^2$ is a heteroskedasticity and autocorrelation consistent (HAC) estimator of the asymptotic variance $\sigma_n^2 = var[\sqrt{n}\overline{WLR}_{m,n}]$:

$$\hat{\sigma}_n^2 \equiv n^{-1} \sum_{t=m}^{T-1} WLR_{m,t+1}^2 + 2 * \left[n^{-1} \sum_{j=1}^{p_n} b_{n,j} \sum_{t=m+j}^{T-\tau} WLR_{m,t+1} WLR_{m,t+1-j} \right],$$

with $\{p_n\}$ a sequence of integers such that $p_n \to \infty$ as $n \to \infty$, $p_n = o(n^{1/4})$ and $\{b_{n,j} : n = 1, 2, ...; j = 1, ..., p_n\}$ a triangular array such that $|b_{n,j}| < \infty$, $n = 1, 2, ..., j = 1, ..., p_n$ and $b_{n,j} \to 1$ as $n \to \infty$ for each $j = 1, ..., p_n$. In practice, the truncation lag p_n is user-defined (see Newey and West, 1987 and Andrews, 1991 for discussion).

A level α test rejects the null hypothesis of equal performance of forecasts f and g whenever $|t_{m,n}| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the $(1-\alpha/2)-$quantile of a standard normal distribution. In case of rejection, one would choose f if $\overline{WLR}_{m,n}$ is positive and g if $\overline{WLR}_{m,n}$ is negative. The following theorem provides the asymptotic justification for our test.

Theorem 1 (Weighted likelihood ratio test) *For a given estimation window size $m < \infty$ and a weight function $w(\cdot)$, $0 \leq w(Y) < \infty$ for all Y suppose:*

(i) $\{Z_t\}$ is a mixing sequence with ϕ of size $-r/(2r-2)$, $r \geq 2$ or α of size $-r/(r-2)$, $r > 2$;

(ii) $E|\log \hat{f}_{m,t}(Y_{t+1})|^{2r} < \infty$ and $E|\log \hat{g}_{m,t}(Y_{t+1})|^{2r} < \infty$ for all t;

(iii) $\sigma_n^2 \equiv var[\sqrt{n}\overline{WLR}_{m,n}] > 0$ for all n sufficiently large.

Then, (a) under H_0 in (2), $t_{m,n} \xrightarrow{d} N(0,1)$ as $n \to \infty$ and (b), under H_A in (3), for any constant $c \in \mathbb{R}$, $P[|t_{m,n}| > c] \to 1$ as $n \to \infty$.

Comments: 1. Note that we do not require stationarity of the underlying data-generating process. Instead, assumption (i) allows the data to be characterized by considerable heterogeneity and dependence, in particular permitting structural changes at unknown dates.

2. Assumption (ii) requires existence of at least four moments of the log-likelihoods, as functions of estimated parameters. The plausibility of this requirement depends on the models, the underlying data-generating process and the estimators on which the forecasts may depend, and thus it should be verified on a case by case basis. For example, for normal density forecasts, this assumption imposes existence of at least eight moments of the variable of interest and existence of the finite sample moments of the conditional mean and variance estimators.

3. As in Giacomini and White (2004), we derive our test using an asymptotic framework where the number of out-of-sample observations n goes to infinity, whereas the estimation sample size m remains finite. Besides being motivated by the presence of underlying heterogeneity, the use of finite-m asymptotics is a way to create an environment with asymptotically non-vanishing estimation uncertainty.

4. As discussed by Giacomini and White (2004), in an environment with asymptotically non-vanishing estimation uncertainty, assumption (iii) does not rule out the case where the competing forecasts are based on nested models. As a result, our tests are applicable to both nested and non-nested models.

5. For the case $w(\cdot) = 1$, the weighted likelihood ratio test is related to Vuong (1989)'s likelihood ratio test for non-nested hypotheses. In that case, (1) the objects of comparison are competing non-nested models; (2) the evaluation is performed in-sample; and (3) the data are assumed to be independent and identically distributed. In contrast, in this paper, (1) we evaluate competing forecast methods (i.e., not only models but also estimation procedures and estimation windows), where the underlying models can be either nested or non-nested; (2) we perform the evaluation out-of-sample; and (3) we allow the data to be characterized by heterogeneity and dependence.

The weighted likelihood ratio test above is conditional on a particular choice of weight function. To reduce dependence on the functional form chosen for the weight function, one might consider generalizing the test (1) to take into account possibly different specifications for $w(\cdot)$. For example, if the null hypothesis of equal performance is rejected in favour of, say density forecast f, a test of superior predictive ability of f relative to g could be constructed by considering a sequence of

J weight functions $\{w_j(\cdot)\}_{j=1}^J$ spanning the whole support of the unconditional distribution of Y_{t+1} and testing whether $E[\overline{WLR}_{m,n}] = 0$ for all w_j. The theoretical underpinnings of such a test are not further considered in this paper, and are left for future research.

5 Monte Carlo experiment

In this section we analyze the finite sample properties of the weighted likelihood ratio test in samples of the sizes typically available in macroeconomic applications. As discussed by Giacomini and White (2004), the design of the Monte Carlo experiment poses problems due to the fact that the null hypothesis depends on estimated parameters, which makes it impossible to ensure that different models yield equal weighted scores in each Monte Carlo replication. Instead, we adopt a similar device to that considered by Giacomini and White (2004). We consider the case $w(y) = 1$ for all y and let the two competing density forecasts be normal densities with different conditional mean specifications and unit variance, $\hat{f}_{m,t} = N(\hat{\mu}^f_{m,t}, 1)$ and $\hat{g}_{m,t} = N(\hat{\mu}^g_{m,t}, 1)$. We exploit the following result.

Proposition 2 *If* $E[Y_{t+1}|\mathcal{F}_t] = \left(\hat{\mu}^f_{m,t} + \hat{\mu}^g_{m,t}\right)/2$, *then* $E\left[\log \hat{f}_{m,t}(Y_{t+1}) - \log \hat{g}_{m,t}(Y_{t+1})\right] = 0$.

We thus generate data under the null hypothesis by first constructing forecasts $\{\hat{f}_{m,t}, \hat{g}_{m,t}\}$, $\hat{f}_{m,t} = N(\hat{\mu}_{1t}, 1)$ and $\hat{g}_{m,t} = N(\hat{\mu}_{2t}, 1)$, for a variable X_{t+1} and then letting $Y_{t+1} = \left(\hat{\mu}^f_{m,t} + \hat{\mu}^g_{m,t}\right)/2 + \varepsilon_{t+1}$, where $\varepsilon_{t+1} \sim i.i.d.\, N(0,1)$. To create data characterized by realistic heterogeneous behavior, and to remain as close as possible to the empirical application, we define X to be the second log-difference of the monthly U.S. consumer price index. For an estimation sample of size m, we then let $\hat{\mu}^f_{m,t}$ be sample mean of the estimation window and $\hat{\mu}^g_{m,t}$ be the conditional mean forecast implied by a Phillips-curve-type model for X_{t+1}:

$$\begin{aligned} \hat{\mu}^f_{m,t} &= (X_t + ... + X_{t-m+1})/m \\ \hat{\mu}^g_{m,t} &= \hat{\alpha}_{m,t} + \hat{\beta}_{m,t} X_t + \hat{\gamma}_{m,t} u_t, \end{aligned} \tag{5}$$

where u is monthly U.S. unemployment. The subscripts denote the fact that the parameters are estimated using data $(X_t, ..., X_{t-m+1})$. We consider a number of in-sample and out-of-sample sizes $m = (50, 100, 150)$ and $n = (25, 50, 100, 150)$ and for each (m, n) pair we generate 5000 Monte Carlo replications of the series $\{Y_{t+1}, \hat{f}_{m,t}, \hat{g}_{m,t}\}$ using a rolling window forecasting scheme on the sample of size $m + n$ ending with the observations for May 2004. For each iteration, we evaluate the densities at Y_{t+1} and compute the score difference $\log \hat{f}_{m,t}(Y_{t+1}) - \log \hat{g}_{m,t}(Y_{t+1})$. The recursion generates n score differences, that we utilize to compute the test statistic (4). We then compute the proportion of rejections of the null hypothesis H_0 at the 5% nominal level, using the test of theorem 1. In the computation of the test statistic 4, we consider two choices for the truncation lag of the HAC estimator: $p_n = 2$ and $p_n = 12$. The empirical size of the test is reported in Table 1.

[TABLE 1 HERE]

The test is oversized for small out-of-sample sizes ($n = 25$) and long truncation lags p_n for the HAC estimator. The test is well-sized for out-of-sample sizes greater than 100, regardless of the choice of p_n.

We next investigate the power of the weighted likelihood ratio test against the alternative hypothesis that $WLR_{m,t+1}$ has non-zero mean:

$$H_a : E[WLR_{m,t+1}] = \mu, \tag{6}$$

for which a sufficient condition is $E[Y_{t+1}|\mathcal{F}_t] = \left(\hat{\mu}^f_{m,t} + \hat{\mu}^g_{m,t}\right)/2 - \mu/\left(2\left(\hat{\mu}^f_{m,t} - \hat{\mu}^g_{m,t}\right)\right)$. We consider 5000 Monte Carlo replications, and let $m = 150$, $n = (25, 50, 100, 150)$ and $\mu = (0.05,\ 0.1..., 0.5)$. For each (n, μ) pair and for each replication, we generate data under the alternative hypothesis by first constructing the conditional mean forecasts $\hat{\mu}^f_{m,t}$ and $\hat{\mu}^g_{m,t}$ as in (5) and then letting $Y_{t+1} = \left(\hat{\mu}^f_{m,t} + \hat{\mu}^g_{m,t}\right)/2 - \mu/\left(2\left(\hat{\mu}^f_{m,t} - \hat{\mu}^g_{m,t}\right)\right) + \varepsilon_{t+1}$, where $\varepsilon_{t+1} \sim\ i.i.d.\, N(0, 1)$. Table 2 reports the proportion of rejections of the null hypothesis at the 5% nominal level using the weighted likelihood ratio test. In all cases, we set $p_n = 2$.

[TABLE 2 HERE]

The test displays good power properties. For example, the test rejects more than 50 % of the time when the expected difference in log-likelihood is as little as 0.4, provided an out-of-sample size larger than 100 is used.

6 Application: Density forecasts of inflation from linear and Markov-switching Phillips curves

6.1 Motivation

A keystone framework for forecasting inflation is the Phillips Curve (henceforth PC), a model relating inflation to some measure of the level of real activity, in most cases the unemployment rate. Stock and Watson (1999) investigate the point forecasting accuracy of the PC, comparing it to that of competing models, such as simple autoregressions and multivariate models. In this study, the Phillips curve is found to produce more accurate forecasts than competing models, particularly when the activity variable is carefully chosen. Another interesting finding is the presence of parameter variation across different subsamples.

We follow Stock and Watson (1999) in considering the PC as a reference for our analysis, and contribute to the literature assessing its forecast performance in several ways. First, we evaluate the accuracy of density forecasts rather than point forecasts, and compare a linear PC specification to an alternative specification that allows for Markov Switching (henceforth MS) of its parameters[4].

[4] For an introduction to Markov Switching models, see Kim and Nelson (1999).

We choose the MS model as a credible competitor to the linear PC model both because the MS mechanism allows for parameter variation over the business cycle and because the predictive distribution generated by the MS model is non-Gaussian, which is of potential interest when comparing density forecast performance. Our second contribution is the evaluation of the impact on forecast performance of using different estimation techniques when producing the forecasts. In particular, we consider estimating both the linear and the MS model by either classical or Bayesian methods. Note that a formal comparison of these forecast methods could not be conducted using previously available techniques, since they do not easily accommodate nested model comparisons and Bayesian estimation. This application thus gives a flavor of the generality of the forecasting comparisons that can be performed using our testing framework.

6.2 Linear and Markov-switching Phillips curve

Following Stock and Watson (1999), our linear model is a PC model in which changes of inflation depend on their lags and on lags of the unemployment rate:

$$\begin{aligned} \Delta\pi_t &= \alpha + \beta(L)\Delta\pi_t + \gamma(L)u_t + \sigma \cdot \varepsilon_t \\ \beta(L) &= \sum_{i=1}^{p_\pi} \beta_i L^i, \gamma(L) = \sum_{i=1}^{p_u} \gamma_i L^i, \varepsilon_t \backsim N.i.d.(0,1) \\ \pi_t &= 100 \times \ln(CPI_t/CPI_{t-12}), \end{aligned} \tag{7}$$

where CPI_t is the consumer price index and u_t is the unemployment rate. This specification is consistent with the "natural rate hypothesis", since the natural rate of unemployment (NAIRU) is $u^* = -\frac{\alpha}{\gamma(1)}$. Note that we implicitly assume that $\Delta\pi_t$ and u_t do not have a unit root[5]. We started from a general model with maximum lag orders p_π and p_u set to 12. Standard model reduction techniques[6] allowed us to constrain the starting model and settle for a more parsimonious specification:

$$\Delta\pi_t = \alpha + \beta_1\Delta\pi_{t-1} + \beta_{12}\Delta\pi_{t-2} + \beta_{12}\Delta\pi_{t-12} + \gamma u_{t-1} + \sigma \cdot \varepsilon_t \tag{8}$$

$$\equiv X'_{t-1}\delta + \sigma \cdot \varepsilon_t \tag{9}$$

This specification seems to be appropriate also across different non-overlapping 15 year subperiods (1959:01-1973:12, 1974:01-1988:12, 1989:01-2004:07) and generates non-correlated residuals[7].

In order to allow for potentially non-Gaussian density forecasts, we consider Markov Switching models as competitors to the linear model (8). We do this by using the parameterization (8) and assuming that some of the parameters vary depending on the value of an unobserved discrete variable

[5] We have run the usual ADF tests to check for the presence of a unit root in π_t and u_t. The testing results (available on request) confirm that π_t has a unit root, whereas u_t does not. These results hold across subperiods.

[6] We estimated the model over the entire sample and then used the general-to-specific approach to eliminate all insignificant regressors, testing the restrictions being imposed.

[7] Note the inclusion of the 12th lag of the dependend variable: the significance of this lag is robust across subperiods.

s_t (which can be given a structural interpretation such as "expansion" or "recession") evolving according to a finite Markov Chain. We consider the two-state MS-PC relationship

$$\begin{aligned} \Delta\pi_t &= \alpha^{s_t} + \beta_1^{s_t}\Delta\pi_{t-1} + \beta_2^{s_t}\Delta\pi_{t-2} + \beta_{12}^{s_t}\Delta\pi_{t-12} + \gamma^{s_t}u_{t-1} + \sigma\cdot\varepsilon_t = \\ &\equiv X_{t-1}^{'}\delta^{s_t} + \sigma\cdot\varepsilon_t \\ s_t &= \begin{cases} 1 \\ 2 \end{cases}, \ \Pr(s_t = j|s_{t-1} = i) = p_{ij}. \end{aligned} \tag{10}$$

Note that the specification imposes that the conditional variance of the dependent variable does not depend on the hidden state. We deem this specification appropriate in dealing with macro relationships, while in financial applications a MS mechanism with different variances across states may also be considered. Further note that we assume the Markov property on s_t, $\Pr(s_t = j|s_{t-1} = i, \underline{s_{t-2}} = s, \mathcal{F}_{t-1}) = \Pr(s_t = j|s_{t-1} = i) = p_{ij}$, where $\underline{s_t}$ is the history of s up to time t. This assumption leads to the possibility of filtering out the latent variables $\underline{s_T}$, which allows us to promptly obtain the likelihood[8].

We consider two variants of the MS model: Model MS1 is (10), in which all conditional mean parameters vary across states. The second variant, MS2, is obtained by imposing the constraints that the intercept and the coefficient on lagged unemployment are constant across states:

$$\alpha^{s_t} = \alpha \tag{11}$$

$$\gamma^{s_t} = \gamma \tag{12}$$

These restrictions are introduced to induce constancy of the NAIRU across states.[9] In this way we have only different speeds of adjustment of the (changes of) inflation with respect to a fixed equilibrium. The equation becomes:

$$\Delta\pi_t = \alpha + \beta_1^{s_t}\Delta\pi_{t-1} + \beta_2^{s_t}\Delta\pi_{t-2} + \beta_{12}^{s_t}\Delta\pi_{t-12} + \gamma u_{t-1} + \sigma\cdot\varepsilon_t \tag{13}$$

In synthesis, we compare density forecasts generated by 3 different models:

- Model LIN: the LR PC model, i.e. equation (8) in which there is no parameter variation across states.

[8] For the details of the filtering procedure, and the way in which the likelihood function is obtained, see Kim and Nelson, (1999), chapter 4.

[9] This restriction is sufficient to ensure NAIRU constancy but it is not necessary: it would suffice to impose

$$\frac{\alpha^1}{\gamma_1^1} = \frac{\alpha^2}{\gamma_1^2}$$

The reason why we did non use this constraint is that it leads to a slightly more involved implementation of the ML and Bayesian estimation procedures.

- Model MS1: the MS - PC equation (10), in which only σ is constant across states;
- Model MS2: the MS - PC equation (13), in which σ, α and γ are constant across states;

It is possible to think of MS1 as the most general model; MS2 is obtained by imposing on MS1 the constraints (11) and (12), whereas LIN is obtained by imposing the constraints that all coefficients are equal across states. Hence we compare models that are nested.

6.3 Estimation and forecasting

Given that our testing framework allows for comparison of different forecasting methods, i.e. sets of choices regarding model specification and estimation, we estimate the three competing models LIN, MS1, MS2 by both classical and Bayesian methods. In this way we can compare across different specifications (LIN, MS1 and MS2) and/or across different estimation and forecasting techniques (classical ML vs. Bayesian estimation).

We use monthly data spanning the period 1958:01-2004:07 obtained from FRED®II[10]. CPI_t and u_t in (7) are, respectively, the Consumer Price Index For All Urban Consumers: All Items and the Civilian Unemployment Rate. Both series are seasonally adjusted.

The forecasts are generated using a rolling window forecasting scheme. The total sample size is $T = 547$ and we used a rolling estimation window of size $m = 360$, leaving us with $n = 187$ out-of-sample observations. This choice of estimation window is the result of two competing instances: to properly allow for possible heterogeneity of the underlying data and to include enough observation to ensure meaningful estimation of a 2-state MS model. A too small m would imply too few (if any) regime switches in each sample used for estimation, hence leading to unreliable estimates of transition probabilities. It is worth pointing out that we also considered a window of $m = 300$ observations and the testing results we obtained in nearly all cases coincide with those obtained for $m = 360$.

6.3.1 The classical approach

We estimate the LIN model by OLS and models MS1 and MS2 by ML. The one-step-ahead density forecasts from the linear model (8) are given by:

$$\hat{f}_{m,t} = \phi_{\left(X_t'\widehat{\delta}_{m,t}\ ,\ \widehat{\sigma}^2_{m,t}\right)},\ t = m, ..., T-1,$$

where $\phi_{(\mu,\sigma^2)}$ is the probability density function of a normal with mean μ and variance σ^2 and $\widehat{\delta}'_{m,t}$, $\widehat{\sigma}^2_{m,t}$ are OLS estimates at time t based on the most recent m observations.

For the MS models (10) and (13), we estimate the parameter vector θ, where

[10] http://research.stlouisfed.org/fred2/.

$$\theta = \left[\alpha^1, \beta_1^1, \beta_2^1, \beta_{12}^1, \gamma^1, \alpha^2, \beta_1^2, \beta_2^2, \beta_{12}^2, \gamma^2, \sigma, p_{11}, p_{21}\right]', \text{ (MS1)}$$
$$\theta = \left[\alpha, \beta_1^1, \beta_2^1, \beta_{12}^1, \gamma, \beta_1^2, \beta_2^2, \beta_{12}^2, \sigma, p_{11}, p_{21}\right]' \text{ (MS2)},$$

over a rolling estimation window of size m by maximizing the conditional likelihood:[11]

$$\begin{aligned}
\widehat{\theta}_{m,t} &= \arg\max_{\theta} \prod_{\tau=t-m+1}^{t} p(\Delta\pi_\tau|\mathcal{F}_{m,\tau-1};\theta), \text{ where} \\
p(\Delta\pi_\tau|\mathcal{F}_{m,\tau-1};\theta) &= \sum_{j=1}^{2} p(\Delta\pi_\tau|s_\tau = j, \mathcal{F}_{m,\tau-1};\theta) \cdot \Pr(s_\tau = j|\mathcal{F}_{m,\tau-1};\theta) \\
\mathcal{F}_{m,\tau} &= \{\Delta\pi_v, u_v : v = \tau - m + 1, ..., \tau\}
\end{aligned}$$

where $p(\Delta\pi_\tau|s_\tau = j, \mathcal{F}_{m,\tau-1};\theta)$ is the conditional density (on parameters, state and past information) of $\Delta\pi_\tau$ as implied by (10), i.e.

$$p(\Delta\pi_\tau|s_\tau = j, \mathcal{F}_{m,\tau-1};\theta) = \phi_{\left(X'_{\tau-1}\delta^j\ ,\ \sigma^2\right)} \tag{14}$$

and $\Pr(s_\tau = j|\mathcal{F}_{m,\tau-1};\theta)$ is obtained by the usual filtering recursion[12]

$$\text{(prediction)} \quad : \quad \Pr(s_\tau = j|\mathcal{F}_{m,\tau-1};\theta) = \sum_{i=1}^{2} p_{ij} \cdot \Pr(s_{\tau-1} = i|\mathcal{F}_{m,\tau-1};\theta) \tag{15}$$

$$\text{(update)} \quad : \quad \Pr(s_\tau = j|\mathcal{F}_{m,\tau-1};\theta) = \frac{p(\Delta\pi_\tau|s_\tau = j, \mathcal{F}_{m,\tau-1};\theta) \cdot \Pr(s_\tau = j|\mathcal{F}_{m,\tau-1};\theta)}{\sum_{j=1}^{2} p(\Delta\pi_\tau|s_\tau = j, \mathcal{F}_{m,\tau-1};\theta) \cdot \Pr(s_\tau = j|\mathcal{F}_{m,\tau-1};\theta)} \tag{16}$$

initialized with the ergodic probabilities of the Markov Chain

$$\Pr(s_{t-m+1} = 1|\mathcal{F}_{m,t-m};\theta) = \frac{1 - p_{11}}{2 - p_{11} - p_{22}}, \tag{17}$$

$$\Pr(s_{t-m+1} = 2|\mathcal{F}_{m,t-m};\theta) = \frac{1 - p_{22}}{2 - p_{11} - p_{22}}. \tag{18}$$

We then obtain the one-step ahead density forecasts by plugging estimates of the parameters in the filtering recursive formulae (15) and (16) for the unobserved states and in the conditional density (14) as follows:

$$\begin{aligned}
\hat{f}_{m,t} &= \sum_{j=1}^{2} \phi_{\left(X_t'\widehat{\delta}^j_{m,t}\ ,\ \widehat{\sigma}^{\,2}_{m,t}\right)} \cdot \widehat{\Pr}(s_{t+1} = j|\mathcal{F}_{m,t};\widehat{\theta}_{m,t}), \\
\widehat{\Pr}(s_{t+1} &= j|\mathcal{F}_{m,t};\widehat{\theta}_{m,t}) = \sum_{i=1}^{2} \widehat{p}^{ij}_{m,t} \cdot \widehat{\Pr}(s_t = i|\mathcal{F}_{m,t};\widehat{\theta}_{m,t}),\ t = m, ..., T-1.
\end{aligned}$$

[11] The model was reparameterised to achieve an unrestricted domain for the parameter space. Then an unconstrained quasi-Newton maximisation routine was implemented. The code is available on request from the first author.

[12] See Kim and Nelson (1999), section (4.1.2).

6.3.2 The Bayesian approach

An alternative approach is to use simulation-based Bayesian inferential techniques. We start from weakly informative priors (i.e., loose but proper priors), we combine them with the likelihood and obtain the joint posterior. We Use Markov Chain Monte Carlo techniques to perform posterior simulation. For the details, see Geweke (1999) and Kim and Nelson (1999, Ch. 7).

Bayesian estimation of the LIN model In the LIN model we give δ and $h = \sigma^{-2}$ conditionally conjugate priors:

$$\delta \backsim N(\underline{\mu}_\delta, \underline{H}_\delta^{-1}) \tag{19}$$

$$\underline{s} \cdot h \backsim \chi^2_{\underline{\nu}} \tag{20}$$

i.e. priors that generate conditional posteriors for δ and h that take the same analytical form as the priors. These priors depend on 4 hyperparameters: $\underline{\mu}_\delta$ and $\underline{H}_\delta$ are respectively the prior mean and precision of the δ vector; $\underline{s}$ and $\underline{\nu}$ define a Gamma prior for $h = \frac{1}{\sigma^2}$ [13].Hence, the posterior distribution of $\theta = [\delta', h]'$ can be simulated by a simple 2-step Gibbs sampling algorithm, using the posterior distribution of δ conditional on h and the posterior distribution of h conditional on δ.

Given a sample of draws from the joint posterior distribution

$$\theta^{(i)} \backsim p(\theta|\mathcal{F}_{m,t}), i = 1, 2, ..., M \tag{21}$$

we can obtain density forecasts in two different ways:

- The Fully Bayesian (FB) way, i.e. by integrating unknown parameters out

$$\hat{f}_{m,t} = \frac{1}{M}\sum_{i=1}^{M} p(Y_{t+1}|\mathcal{F}_{m,t}; \theta^{(i)}) \xrightarrow{d} \int p(Y_{t+1}|\mathcal{F}_{m,t}; \theta) p(\theta|\mathcal{F}_{m,t}) d\theta \tag{22}$$

- The "Empirical Bayes" (EB or plug-in) way, i.e. by plugging a parameter configuration $\widehat{\theta}$:

$$\hat{f}_{m,t} = p(Y_{t+1}|\mathcal{F}_{m,t}; \widehat{\theta}) \tag{23}$$

  where $\widehat{\theta}$ is taken to synthesize the whole posterior distribution $p(\theta|\mathcal{F}_{m,t})$, i.e. it could be an estimate of the posterior mode, mean or median.

One could argue that the FB way is conceptually superior to the alternative, in that it takes into account the role of parameter uncertainty, whereas the EB way ignores the uncertainty around point estimates, as the density forecasts obtained using the ML approach do. Nonetheless, we chose to also report EB density forecast evaluations in order to compare Bayesian with non-Bayesian approaches.

[13] The choice of the hyperparameters $\underline{s}$ and $\underline{\nu}$ can be flexibly used to calibrate the prior mean and prior variance of $h = \frac{1}{\sigma^2}$ since

$$E(h) = \frac{\underline{\nu}}{\underline{s}}, V(h) = \frac{2\underline{\nu}}{\underline{s}^2}.$$

Note that non-dogmatic settings for these hyperparameters lead to a conditional posterior for h which is dominated by data evidence.

Bayesian estimation of MS models Bayesian inference in MS models is complicated by the fact that these models involve a latent variable. For the details, see Kim and Nelson (1999, chapter 9) and Geweke and Amisano (2004). We use priors (19) and (20) on the regression coefficients

$$\delta = \left[\delta_1', \delta_2'\right]' \tag{24}$$

and on $h = \sigma^{-2}$. In the partition (24) for δ, δ_1 are the first order parameters that are fixed across regimes and δ_2 are the first order parameters that vary across states. For example, in MS1 we have $\delta_1 = 0$ and $\delta_2 = [\alpha^1, \beta_1^1, \beta_2^1, \beta_{12}^1, \gamma^1, \alpha^2, \beta_1^2, \beta_2^2, \beta_{12}^2, \gamma^2]'$, while in MS2 $\delta_1 = [\alpha, \gamma]'$ and $\delta_2 = [\beta_1^1, \beta_2^1, \beta_{12}^1, \beta_1^2, \beta_2^2, \beta_{12}^2]'$.We collect the transition probabilities in the 2×2 matrix P and impose a Beta (Dirichlet prior): $(p_{11}, p_{12}) \backsim Dir(r_{11}, r_{12})$ and $(p_{22}, p_{22}) \backsim Dir(r_{21}, r_{22})$. An easy Bayesian MCMC analysis of the MS models is based on a conceptually simple Gibbs sampling-data augmentation algorithm in which latent variables are sequentially simulated like a block of parameters. The algorithm (for details, see chapter 9 of Kim and Nelson, 1999, or Geweke and Amisano, 2004) is initialized by drawing δ, h, P from their prior distributions, and then is based on the cyclical repetition of the following steps: (1) draw $\underline{s}_{t,m} = \{s_\tau, \tau = t-m+1, ..., t\}$ from its conditional distribution conditional on δ, h, P (data-augmentation step in which latent variables are simulated); (2) draw δ from its conditional distribution conditional on h and on $\underline{s}_{t,m}$; (3) draw h from its conditional distribution conditional on δ and on $\underline{s}_{t,m}$; (4) draw P from its conditional distribution conditional on $\underline{s}_{t,m}$. The resulting Markov Chain converges in distribution to the joint posterior distribution of $\delta, h, P, \underline{s}_{t,m}$. The models MS1 and MS2 are simulated in the same way.

To produce density forecasts, we use a FB method:

$$\begin{aligned}
\hat{f}_{m,t} &= \frac{1}{M}\sum_{i=1}^{M}\sum_{j=1}^{2}\phi_{\left(X_t'\delta_{(i)}^j, \sigma_{(i)}^2\right)} \cdot p(s_{t+1} = j|\mathcal{F}_{m,t}; \theta^{(i)}) && (25)\\
&\xrightarrow{d} \sum_{j=1}^{2}\int p(\Delta\pi_{t+1}|\mathcal{F}_{m,t}, s_{t+1} = j; \theta) \cdot p(s_{t+1} = j|\mathcal{F}_{m,t}; \theta) \cdot p(\theta|\mathcal{F}_{m,t})d\theta = \\
&= p(\Delta\pi_{t+1}|\mathcal{F}_{m,t}) && (26)
\end{aligned}$$

which is based on the marginalization with respect to the unknown parameters and the unknown state variables.

Alternatively, we also consider an EB method

$$\hat{f}_{m,t} = \sum_{j=1}^{2} p(\Delta\pi_{t+1}|s_{t+1} = j, \mathcal{F}_t; \widehat{\theta}) \cdot p(s_{t+1} = j|\mathcal{F}_t; \widehat{\theta}) \tag{27}$$

in which $\widehat{\theta}$ is a synthetic value taken from the posterior density of θ. In this case parameter uncertainty is ignored whereas uncertainty about the latent variables $\underline{s}_{t+1,m+1}$is properly accounted for through marginalization.

6.4 Discussion of the results

We performed ML and Bayesian estimation of the three models LIN, MS1 and MS2 in (8), (10) and (13). Estimation was carried out by Matlab[14]. With reference to (19) and (20), in model LIN we use as hyperparameters

$$\underline{\mu}_\delta = \begin{bmatrix} 0.1 \\ 0.0 \\ 0.0 \\ 0.0 \\ -0.04 \end{bmatrix}, \ \underline{H}_\delta = \left(\frac{1}{0.2}\right)^2 \cdot I_5, \ \underline{s} = 0.3, \ \underline{\nu} = 3. \tag{28}$$

For the MS1 and MS2 models we used the same hyperparameters $\underline{s} = 0.3$, $\underline{\nu} = 3$ as in the LIN model; for the regression coefficients we used the following prior hyperparameters

$$MS1 \quad : \quad \underline{\mu}_\delta = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} 0.1 \\ 0.0 \\ 0.0 \\ 0.0 \\ -0.04 \end{bmatrix}, \ \underline{H}_\delta = \left(\frac{1}{0.2}\right)^2 \cdot I_{10} \tag{29}$$

$$MS2 \quad : \quad \underline{\mu}_\delta = \begin{bmatrix} 0.1 \\ -0.4 \\ \begin{bmatrix} 1 \\ 1 \end{bmatrix} \otimes \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix} \end{bmatrix}, \ \underline{H}_\delta = \left(\frac{1}{0.2}\right)^2 \cdot I_8 \tag{30}$$

whereas for the Dirichlet prior for the rows of P we used

$$R = \begin{bmatrix} 8 & 2 \\ 2 & 8 \end{bmatrix}. \tag{31}$$

All the prior distributions are proper, quite loose, symmetric across unobserved states, and the prior on P imparts relevant persistence on the unobservable states.

In terms of point parameter estimates, we did not find many differences between Bayesian and ML estimates, once the label switching problem of the MS models is properly accounted for (see for details, Geweke and Amisano, 2004). For this reason we decided to report only point ML estimates of the parameters in Figures 2 to 4.

[14] The codes can be requested to the first author. They require the availability of the `Matlab` toolboxes `Optimization` and `Statistics`. ML is carried out by reparameterizing the model in a way that parameters are defined on an unrestricted domain and then minus the loglikelihood is minimized using the `Matlab` function `fminunc`. Also Bayesian inference is carried out by using `Matlab` routines based on the availability of the same toolboxes. The filtering code used for filtering out the latent variables was coded in `Fortran` and linked up as a `dll` in Matlab (`MMfilter.dll`).

[FIGURE 2 HERE]

Figure 2 shows the estimation results of the LIN model. Note that individual parameters seem to vary quite a lot, but the estimated NAIRU (bottom left) $\widehat{NAIRU} = \frac{-\widehat{\alpha_0}}{\widehat{\gamma_1}}$ and the estimated "speed of adjustment" (SP, degree of mean reversion, bottom center) $\widehat{sp} = \widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\beta}_{12} - 1$, tend to be more stable. In particular, the estimated NAIRU is quite high, hovering around 6%, with only a slight drop in the second half of the 1990s. This is in contrast to the commonly held belief[15] that the PC, and the NAIRU in particular, changed radically in the prolonged expansion that took place in the 1990s. The estimated speed of adjustment, measuring the degree of mean reversion, can be viewed as indicating how easy it is to forecast changes of inflation in the short term: when SP becomes more negative, as from 1999 onwards, it becomes harder to forecast $\Delta\pi_t$, at least in the short term.

[FIGURE 3 HERE]

In Figure 3 we report ML parameter estimates of the conditional mean parameters δ and the precision parameter $h = \sigma^{-2}$ for the MS1 model. Our main remarks are: (1) the MS model parameter estimates are more stable than their LIN model counterpart; (2) the interpretation of state 1 being "expansion" and state 2 being "recession" is coherent with the findings that the NAIRU in state 2 is higher than in state 1 and that the persistence of state 1, p_{11}, is higher than the persistence of state 2, p_{22} (consistent with recessions having shorter duration than expansions); (3) in state 2 NAIRU is almost constant whereas in state 1 it drops sharply in the second half of the 1990s.

Figure 4 reports results of the estimation of MS2 model.

[FIGURE 4 HERE]

Note that: (1) it is problematic to identify states: they have the same persistence and they are associated with the same NAIRU. State 1 is associated with higher speed of adjustment; (2) NAIRU estimates are in the same range as for the LIN model, and they have the same time evolution; (3) h has the same time evolution as in the LIN and MS1 cases.

Turning to the evaluation of the density forecast performance of the competing models, we consider the sequence of one-step-ahead density forecasts implied by the three models, using the ML, FB and EB estimation approaches. We conduct pairwise WLR tests and report the results in Tables 3A-3C. In the computation of the test statistic (4) we use a Newey-West estimator of $\hat{\sigma}_n^2$ with bandwidth $\tau = 6$. Different values for τ have no impact on the results.

Table 3A reports the unweighted (ω_0) case.

[TABLE 3A HERE]

[15] See for instance the results presented in Staiger et al. (2001), where the NAIRU is found to have dropped of more than 1% in the 1995-2000 period.

Table 3A shows that LIN is significantly worse than either MS1 or MS2. This finding is robust to the estimation method: using both the ML and the FB approach, LIN fares significantly worse than both competitors. Using the EB approach, conclusions are not that clear-cut, since the differences are not significant. As for comparison between MS1 and MS2, they are not significantly different for the EB case, but MS1 outperforms MS2 in both ML and FB estimation case. In general, we conclude that the winner seems to be the MS1 model and that the best performance is achieved by using ML estimation.

Table 3B shows results for the Center-weighted case (ω_1).

[TABLE 3B HERE]

The conclusions are nearly the same as in the previous case: MS1 dominates MS2 and LIN, both in the classical estimation framework (ML) and in a Bayesian context (FB). Using the EB approach generally yields insignificant differences.

For the Tails-weighted case (ω_2) the results are in Table 3C.

[TABLE 3C HERE]

Here we have partially different results: in fact, using the ML results, LIN is significantly worse than MS2 only, but the MS1-MS2 difference is not significant. These conclusions are not robust to estimation and forecasting method: FB differences are not significant at all while EB shows dominance of the LIN model. We believe that these findings can be due to the fact that the weighting scheme assigns negligible weights to observations near the sample mean of the dependent variable inflation, in this way blurring the differences among models. This interpretation is confirmed by visual inspection of the sample weights (see Figure 5, bottom center), in which it is evident that nearly half of the observations are assigned weights close to zero.

[FIGURE 5 HERE]

For the Right Tail-weighted case (ω_3) the results are contained in Table 3D.

[TABLE 3D HERE]

This weighting scheme can be viewed as assigning importance to good density forecast properties when inflation is rising fast. The conclusions in this case are very similar to the unweighted case: LIN is significantly outperformed by MS1 and MS2 and this is robust to the estimation method (ML and FB). MS1 outperforms MS2, robust with respect to estimation method (ML, FB). The MS1 appears to be the best model and the best way to obtain density forecasts is to use ML estimation.

For the Left Tail weighted case (ω_4) the results are contained in Table 3E.

[TABLE 3E HERE]

Also in this case, LIN is significantly worse than MS1 and MS2, both with the ML and FB approaches and the evidence for EB estimation is not significant. Note also that in this case MS1 is significantly better than MS2 only in the FB approach.

Summing up the whole evidence, it appears that the Markov-switching MS1 model outperforms both alternatives. This happens in the unweighted case and in most of the weighted cases. Only in the Tails weighted (ω_2) case, we have a less clear picture, but this is likely to be due to excessive penalty attributed to the values near the center of the distribution of the dependent variable. We also generally conclude that maximum likelihood estimation yields better forecasts than Bayesian estimation.

6.5 Comparison with predictive posterior odds ratios

As a final piece of evidence regarding the relative performance of the three models, we consider Bayesian model comparison based on posterior odds ratios (POR):

$$POR_{i,j} = \frac{p(M_i|Y)}{p(M_j|Y)} = \frac{p(M_i)}{p(M_j)}\frac{p(Y|M_i)}{p(Y|M_j)} \tag{32}$$

$$\frac{p(Y|M_i)}{p(Y|M_j)} = Bayes\ factor = \frac{p(Y_1|Y_0, M_i)}{p(Y_1|Y_0, M_j)} \cdot \frac{p(Y_0|M_i)}{p(Y_0|M_j)} \tag{33}$$

where Y_0 contains all data up to observation m and Y_1 contains the data after m. A $POR_{i,j}$ greater than one indicates that model i outperforms model j. The predictive Bayes factor $PBF = \frac{p(Y_1|Y_0,M_i)}{p(Y_1|Y_0,M_j)}$ is the ratio of the predictive densities of Y_1 based on the two competing models, which can be computed by recursively simulating the posterior distribution of the parameters:

$$p(Y_1|Y_0, M_j) = \prod_{t=m+1}^{T} p(Y_t|Y_{t-1}, M_j) \tag{34}$$

$$p(Y_t|Y_{t-1}, M_j) = \frac{1}{M}\sum_{i=1}^{M} p(Y_t|Y_{t-1}, M_j; \theta^{(i)}) \xrightarrow{d} p(Y_t|Y_{t-1}, M_j) \tag{35}$$

$$\theta^{(i)} \backsim p(\theta|Y_{t-1}, M_j). \tag{36}$$

In our application, the PBF from LIN versus MS1 and MS2 are respectively 2.73E-07 and 7.55E-06, indicating that the linear model is clearly outperformed by the Markov-switching models. The PBF for MS1 versus MS2 is 27.6, suggesting that MS1 is the clear winner. The conclusion from Bayesian predictive model comparison is thus in concordance with that from the WLR tests. We should however point out that, unlike the WLR testing procedure, the comparison based on the predictive Bayes factor does not allow for time variation in the data generating process.

7 Conclusion

We introduced a weighted likelihood ratio test for comparing the out-of-sample performance of competing density forecasts of a variable. We proposed measuring the performance of the density forecasts by scoring rules, which are loss functions defined over the probability forecast and the outcome of the variable. In particular, we restricted attention to the logarithmic scoring rule, and suggested ranking the forecasts according to the relative magnitude of a weighted average of the scores measured over the available sample. We showed that the use of weights introduces flexibility by allowing the user to isolate the performance of competing density forecasts in different regions of the unconditional distribution of the variable of interest. Loosely speaking, the test can help distinguish, for example, the relative forecast performance in 'normal' days from that in days when the variable takes on 'extreme' values. The special case of equal weights is also of interest, since in this case our test is related to Vuong's (1989) likelihood ratio test for non-nested hypotheses. Unlike Vuong's (1989) test, however our test is performed out-of-sample rather than in-sample, it is valid for both nested and non-nested forecast models, and it considers time series rather than i.i.d. data.

Our test can be applied to time-series data characterized by a considerable amount of heterogeneity and dependence (including possible structural changes) and it is valid under general conditions. In particular, the underlying forecast models can be either nested or non-nested and the forecasts can be produced by utilizing parametric as well as semi-parametric, non-parametric or Bayesian estimation procedures. The only requirement that we impose is that the forecasts are based on a finite estimation window, as in Giacomini and White (2004).

We applied our test to a comparison of density forecasts produced by different versions of a univariate Phillips Curve, based on monthly US data over the last 40 years: a linear regression, a 2-state Markov Switching regression and a 2-state Markov Switching regression in which the natural rate of unemployment was constrained to be equal across states. Note that these models imply different shapes for the density of inflation: a normal density for the linear model and a mixture of normals for the Markov-switching models. In the comparison we also considered versions of each model estimated by maximum likelihood or by Bayesian techniques.

Our general conclusion was that density forecasts from the MS1 model outperformed all alternatives. This happened in the unweighted case and in most of the weighted cases. Only in the Tails-weighted case, the three models could not be discriminated, but this is likely to be due to excessive penalization of the values near the center of the distribution of the dependent variable. We also found that maximum likelihood estimation seemed to yield superior forecasts than Bayesian alternatives.

8 Proofs

Proof of Theorem 1. We show that assumptions (i)-(iii) of Theorem 6 of Giacomini and White (2004) (which we denote by GW1-GW3) are satisfied by letting $W_t \equiv Z_t$ and $\Delta L_{m,t+1} \equiv WLR_{m,t+1}$, from which *(a)* and *(b)* follow.

GW1 coincides with assumption (i).

GW2 imposes the existence of $2r$ moments of $WLR_{m,t+1} = w(Y_{t+1})(\log \hat{f}_{m,t}(Y_{t+1}) - \log \hat{g}_{m,t}(Y_{t+1}))$ for some $r > 2$. From assumption (ii), there exists a $r' > 2$ such that $E|\log \hat{f}_{m,t}(Y_{t+1})|^{2r'} < \infty$. Define $r = \frac{r'}{1+\varepsilon}$, for some $\varepsilon > 0$, so that $r > 2$. Since $w(Y_{t+1}^{st}) \geq 0$, and by applying Hölder's inequality, we have

$$\begin{aligned} E|w(Y_{t+1}^{st}) \log \hat{f}_{m,t}(Y_{t+1})|^{2r} &= E\left| w(Y_{t+1}^{st})^{2r} |\log \hat{f}_{m,t}(Y_{t+1})|^{2r} \right| \\ &\leq \left(E w(Y_{t+1}^{st})^{2r\frac{1+\varepsilon}{\varepsilon}} \right)^{\frac{\varepsilon}{1+\varepsilon}} \left(E|\log \hat{f}_{m,t}(Y_{t+1})|^{2r(1+\varepsilon)} \right)^{\frac{1}{1+\varepsilon}} \\ &= \left(E w(Y_{t+1}^{st})^{\frac{2r'}{\varepsilon}} \right)^{\frac{\varepsilon}{1+\varepsilon}} \left(E|\log \hat{f}_{m,t}(Y_{t+1})|^{2r'} \right)^{\frac{1}{1+\varepsilon}} < \infty, \end{aligned}$$

where the last inequality holds since the first term is finite because $w(\cdot)$ is bounded, and the second term is finite by (ii). Similarly, $E|w(Y_{t+1}^{st}) \log \hat{g}_{m,t}(Y_{t+1})|^{2r} < \infty$. By Minkowski's inequality we thus have

$$E|WLR_{m,t+1}|^{2r} \leq \left[\left(E|w(Y_{t+1}^{st}) \log \hat{f}_{m,t}(Y_{t+1})|^{2r} \right)^{\frac{1}{2r}} + \left(E|w(Y_{t+1}^{st}) \log \hat{g}_{m,t}(Y_{t+1})|^{2r} \right)^{\frac{1}{2r}} \right]^{2r} < \infty$$

Finally, GW3 coincides with assumption (iii). ■

Proof of Proposition 2. We have

$$\begin{aligned} E[\log \hat{f}_{m,t}(Y_{t+1}) - \log \hat{g}_{m,t}(Y_{t+1})] &= \\ E\left[E[\log \hat{f}_{m,t}(Y_{t+1}) - \log \hat{g}_{m,t}(Y_{t+1}) | \mathcal{F}_t] \right] &= \\ E\left[E\left[-\frac{1}{2}\left(Y_{t+1} - \hat{\mu}_{m,t}^f \right)^2 + \frac{1}{2}(Y_{t+1} - \hat{\mu}_{m,t}^g)^2 | \mathcal{F}_t \right] \right] &= \\ E\left[\frac{1}{2} E\left[2Y_{t+1}\left(\hat{\mu}_{m,t}^f - \hat{\mu}_{m,t}^g \right) - \left(\hat{\mu}_{m,t}^f \right)^2 + (\hat{\mu}_{m,t}^g)^2 | \mathcal{F}_t \right] \right] &= \\ E\left[\left(\hat{\mu}_{m,t}^f - \hat{\mu}_{m,t}^g \right) \left(E[Y_{t+1}|\mathcal{F}_t] - \left(\hat{\mu}_{m,t}^f + \hat{\mu}_{m,t}^g \right)/2 \right) \right] &= 0, \end{aligned}$$

where the last equality follows from the fact that $E[Y_{t+1}|\mathcal{F}_t] = \left(\hat{\mu}_{m,t}^f + \hat{\mu}_{m,t}^g \right)/2$. ■

References

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

[2] Artzner, P., Delbaen, F., Eber, J. M., Heath, D. (1997): "Thinking Coherently', *Risk*, 10, 68-71.

[3] Berkowitz, J. (2001): "The Accuracy of Density Forecasts in Risk Management', *Journal of Business and Economic Statistics*, 19, 465-74.

[4] Christoffersen P. F., Diebold, F. X. (1997): "Optimal Prediction under Asymmetric Loss', *Econometric Theory*, 13, 808-817.

[5] Clements, M. P., Smith, J. (2000): "Evaluating the Forecast Densities of Linear and Non-linear Models: Applications to Output Growth and Unemployment', *Journal of Forecasting*, 19, 255-276.

[6] Corradi, V. and N. Swanson (2003): "Bootstrap Conditional Distribution Tests in the Presence of Dynamic Misspecification', *Journal of Econometrics*, forthcoming

[7] Corradi, V. and N. Swanson (2004a): "Predictive Density and Conditional Confidence Interval Accuracy Tests', Rutgers University working paper

[8] Corradi, V. and N. Swanson (2004b): "A Test for Comparing Multiple Misspecified Conditional Distributions', Rutgers University working paper

[9] Diebold, F. X., Gunther, T. A., Tay, A. S. (1998): "Evaluating Density Forecasts with Applications to Financial Risk Management', *International Economic Review,* 39, 863-883.

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

[11] Diebold, F. X., Tay, A.S., Wallis, K. F. (1999): "Evaluating Density Forecasts of Inflation: The Survey of Professional Forecasters ", in R. Engle and H. White (eds.), *Festschrift in Honour of C.W.J. Granger*, 76-90, Oxford University Press.

[12] Duffie, D., Pan, J. (1996): "An Overview of Value at Risk', *Journal of Derivatives*, 4, 13-32.

[13] Geweke, J. and G. Amisano (2004):'Compound Markov Mixture Models with Applications in Finance, mimeo.

[14] Giacomini, R. and H. White (2004): 'Tests of conditional predictive ability', mimeo, University of California, Los Angeles and University of California, San Diego working paper.

[15] Hong, Y. (2000): "Evaluation of Out-of-Sample Density Forecasts with Applications to S&P 500 Stock Prices', Cornell University manuscript.

[16] Hong, Y., White, H. (2000): "Asymptotic Distribution Theory for Nonparametric Entropy Measures of Serial Dependence', University of California, San Diego manuscript.

[17] Hong, Y. and H. Li (2003): "Out of Sample Performance of Spot Interest Rate Models', *Review of Financial Studies,* forthcoming

[18] Kim, C.J. and C. Nelson (1999): *State Space Models with Regime Switching*, MIT Press, Cambridge, USA.

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

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

[21] Patton, A. J. (2001): "Modelling Asymmetric Exchange Rate Dependence', *International Economic Review,* forthcoming.

[22] Soderlind, P., Svensson, L. (1997): "New Techniques to Extract Market Expectations from Financial Instruments', *Journal of Monetary Economics,* 40, 383-429.

[23] Steiger, D., J.H. Stock and M.W. Watson (2001): "Prices, wages and the U.S. NAIRU in the 1990s", NBER working paper # 8320.

[24] Stock, J.H. and M.W. Watson (1999): "Forecasting Inflation", *Journal of Monetary Economics,* 44, 293-335 .

[25] Tay, A. S., Wallis, K. F. (2000): "Density Forecasting: A Survey', *Journal of Forecasting*, 19, 235-254.

[26] Vuong, Q. H. (1989): "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses', *Econometrica*, 57, 307-333.

[27] Weigend, A. S., Shi, S. (2000): "Predicting Daily Probability Distributions of S&P500 Returns', *Journal of Forecasting*, 19, 375-392.

[28] Weiss, A. A. (1996): "Estimating Time Series Models Using the Relevant Cost Function', *Journal of Applied Econometrics*, 11, 539-560.

[29] Winkler, R. L. (1967): "The Quantification of Judgement: Some Methodological Suggestions', *Journal of the American Statistical Association*, 62, 1105-1120.

Table 1. Size of nominal 0.05 tests

A. $p_n = 2$				
	n			
m	25	50	100	150
50	0.068	0.038	0.042	0.045
100	0.077	0.065	0.061	0.053
150	0.074	0.049	0.048	0.051
B. $p_n = 12$				
	n			
m	25	50	100	150
50	0.165	0.069	0.059	0.053
100	0.206	0.123	0.088	0.059
150	0.212	0.056	0.051	0.053

Notes: The two panels report the empirical size of the weighted likelihood ratio test discussed in Section 4, for different choices of truncation lag p_n for the HAC estimator. Entries represent the rejection frequencies over 5000 Monte Carlo replications of the null hypothesis $H_0 : E[\log \hat{f}_{m,t}(Y_{t+1}) - \log \hat{g}_{m,t}(Y_{t+1})] = 0$, where the density forecasts f, g and the data-generating process are defined in section 5. The nominal size is 0.05. Each cell corresponds to a pair of in-sample and out-of-sample sizes (m, n).

Table 2. Power of test

	n			
μ	25	50	100	150
0.05	0.066	0.084	0.064	0.066
0.10	0.074	0.091	0.089	0.122
0.15	0.098	0.101	0.136	0.208
0.20	0.121	0.133	0.186	0.351
0.25	0.147	0.146	0.270	0.470
0.30	0.170	0.190	0.364	0.610
0.35	0.220	0.231	0.455	0.730
0.40	0.260	0.276	0.559	0.825
0.45	0.313	0.310	0.644	0.892
0.50	0.348	0.357	0.712	0.946

Notes: The table reports the rejection frequencies over 5000 Monte Carlo replications of the null hypothesis $H_0 : E[\log \hat{f}_{m,t}(Y_{t+1}) - \log \hat{g}_{m,t}(Y_{t+1})] = 0$ using the test of section 4, with $p_n = 2$. The density forecasts f, g and the data-generating process are defined in section 5. The in-sample size is $m = 150$. n is the out-of-sample size and $\mu = E[\log \hat{f}_{m,t}(Y_{t+1}) - \log \hat{g}_{m,t}(Y_{t+1})]$.

Table 3A: Weighted Likelihood Ratio tests. Unweighted (W0)

Benchmark	LIN_ML	LIN_FB	LIN_EB	MS1_ML	MS1_FB	MS1_EB	MS2_ML	MS2_FB
LIN_FB	0.0004							
	(0.8664)							
LIN_EB	-0.0011	-0.0014						
	(0.5773)	(0.1362)						
MS1_ML	-0.0886*	-0.089*	-0.0875*					
	(0)	(0)	(0)					
MS1_FB	-0.0487*	-0.0491*	-0.0476*	0.0399*				
	(0)	(0)	(0)	(0)				
MS1_EB	0.0131	0.0128	0.0142	0.1017*	0.0618*			
	(0.5843)	(0.5782)	(0.5356)	(0)	(0.0034)			
MS2_ML	-0.0478*	-0.0482*	-0.0467*	0.0408*	0.0009	-0.061*		
	(0)	(0)	(0)	(0.0024)	(0.9437)	(0.0248)		
MS2_FB	-0.0221*	-0.0224*	-0.021*	0.0665*	0.0266*	-0.0352	0.0258*	
	(0)	(0)	(0)	(0)	(0.0004)	(0.1225)	(0.005)	
MS2_EB	0.0058	0.0055	0.0069	0.0944*	0.0545*	-0.0073	0.0536*	0.0279*
	(0.6711)	(0.6839)	(0.5887)	(0)	(0.0002)	(0.7288)	(0.0113)	(0.03)

Table 3B: Weighted Likelihood Ratio tests. Center (W1)

Benchmark	LIN_ML	LIN_FB	LIN_EB	MS1_ML	MS1_FB	MS1_EB	MS2_ML	MS2_FB
LIN_FB	-0.0002							
	(0.6592)							
LIN_EB	-0.0013*	-0.0011*						
	(0.0124)	(0)						
MS1_ML	-0.0338*	-0.0336*	-0.0325*					
	(0)	(0)	(0)					
MS1_FB	-0.0196*	-0.0193*	-0.0182*	0.0143*				
	(0)	(0)	(0)	(0)				
MS1_EB	-0.0061	-0.0059	-0.0048	0.0277*	0.0134*			
	(0.2764)	(0.274)	(0.3745)	(0)	(0.006)			
MS2_ML	-0.0128*	-0.0126*	-0.0115*	0.021*	0.0067*	-0.0067		
	(0)	(0)	(0)	(0)	(0.0236)	(0.2573)		
MS2_FB	-0.0093*	-0.009*	-0.0079*	0.0245*	0.0103*	-0.0032	0.0036*	
	(0)	(0)	(0)	(0)	(0)	(0.5483)	(0.0292)	
MS2_EB	-0.0121*	-0.0119*	-0.0107*	0.0217*	0.0075*	-0.006	0.0008	-0.0028
	(0)	(0)	(0)	(0)	(0.0083)	(0.224)	(0.8232)	(0.157)

Table 3C: Weighted Likelihood Ratio tests. Tails (W2)

Benchmark	LIN_ML	LIN_FB	LIN_EB	MS1_ML	MS1_FB	MS1_EB	MS2_ML	MS2_FB
LIN_FB	0.001							
	(0.5162)							
LIN_EB	0.0023*	0.0013						
	(0.0397)	(0.1432)						
MS1_ML	-0.0038	-0.0047	-0.0061					
	(0.7038)	(0.6336)	(0.5335)					
MS1_FB	0.0003	-0.0006	-0.002	0.0041				
	(0.9587)	(0.913)	(0.7299)	(0.4697)				
MS1_EB	0.0285*	0.0275*	0.0262*	0.0323*	0.0282*			
	(0.0193)	(0.0175)	(0.0226)	(0.0274)	(0.0125)			
MS2_ML	-0.0156*	-0.0166*	-0.0179*	-0.0118	-0.0159*	-0.0441*		
	(0.0406)	(0.0235)	(0.0218)	(0.1823)	(0.0329)	(0.0042)		
MS2_FB	0.0012	0.0003	-0.0011	0.005	0.0009	-0.0273*	0.0168*	
	(0.6556)	(0.9244)	(0.6629)	(0.5153)	(0.8296)	(0.0203)	(0.0091)	
MS2_EB	0.0361*	0.0352*	0.0338*	0.0399*	0.0358*	0.0076	0.0517*	0.0349*
	(0.0003)	(0.0006)	(0.0004)	(0.0034)	(0.0011)	(0.4818)	(0.001)	(0.0004)

Table 3D: Weighted Likelihood Ratio tests. Right tail (W3)

Benchmark	LIN_ML	LIN_FB	LIN_EB	MS1_ML	MS1_FB	MS1_EB	MS2_ML	MS2_FB
LIN_FB	-0.0002							
	(0.6592)							
LIN_EB	-0.0013*	-0.0011*						
	(0.0124)	(0)						
MS1_ML	-0.0338*	-0.0336*	-0.0325*					
	(0)	(0)	(0)					
MS1_FB	-0.0196*	-0.0193*	-0.0182*	0.0143*				
	(0)	(0)	(0)	(0)				
MS1_EB	-0.0061	-0.0059	-0.0048	0.0277*	0.0134*			
	(0.2764)	(0.274)	(0.3745)	(0)	(0.006)			
MS2_ML	-0.0128*	-0.0126*	-0.0115*	0.021*	0.0067*	-0.0067		
	(0)	(0)	(0)	(0)	(0.0236)	(0.2573)		
MS2_FB	-0.0093*	-0.009*	-0.0079*	0.0245*	0.0103*	-0.0032	0.0036*	
	(0)	(0)	(0)	(0)	(0)	(0.5483)	(0.0292)	
MS2_EB	-0.0121*	-0.0119*	-0.0107*	0.0217*	0.0075*	-0.006	0.0008	-0.0028
	(0)	(0)	(0)	(0)	(0.0083)	(0.224)	(0.8232)	(0.157)

Table 3E: Weighted Likelihood Ratio tests. Left tail (W4)

Benchmark	LIN_ML	LIN_FB	LIN_EB	MS1_ML	MS1_FB	MS1_EB	MS2_ML	MS2_FB
LIN_FB	0.0006							
	(0.7419)							
LIN_EB	0.0003	-0.0003						
	(0.8608)	(0.7199)						
MS1_ML	-0.0548*	-0.0554*	-0.055*					
	(0)	(0)	(0)					
MS1_FB	-0.0291*	-0.0297*	-0.0294*	0.0256*				
	(0.0007)	(0.0002)	(0.0001)	(0.0004)				
MS1_EB	0.0193	0.0187	0.019	0.074*	0.0484*			
	(0.3088)	(0.3022)	(0.2924)	(0.0002)	(0.0039)			
MS2_ML	-0.035*	-0.0356*	-0.0352*	0.0198	-0.0058	-0.0542*		
	(0.0002)	(0.0001)	(0.0003)	(0.0802)	(0.5634)	(0.0138)		
MS2_FB	-0.0128*	-0.0134*	-0.013*	0.042*	0.0164*	-0.032	0.0222*	
	(0.0008)	(0.0002)	(0.0001)	(0)	(0.005)	(0.0758)	(0.0053)	
MS2_EB	0.0179	0.0173	0.0177	0.0727*	0.0471*	-0.0013	0.0529*	0.0307*
	(0.1368)	(0.1483)	(0.1187)	(0)	(0.0002)	(0.9356)	(0.0047)	(0.0077)

Figure 1: weight functions
(w_1 = center of distribution, w_2 = tails of distribution, w_3 = right tail, w_4 = left tail)

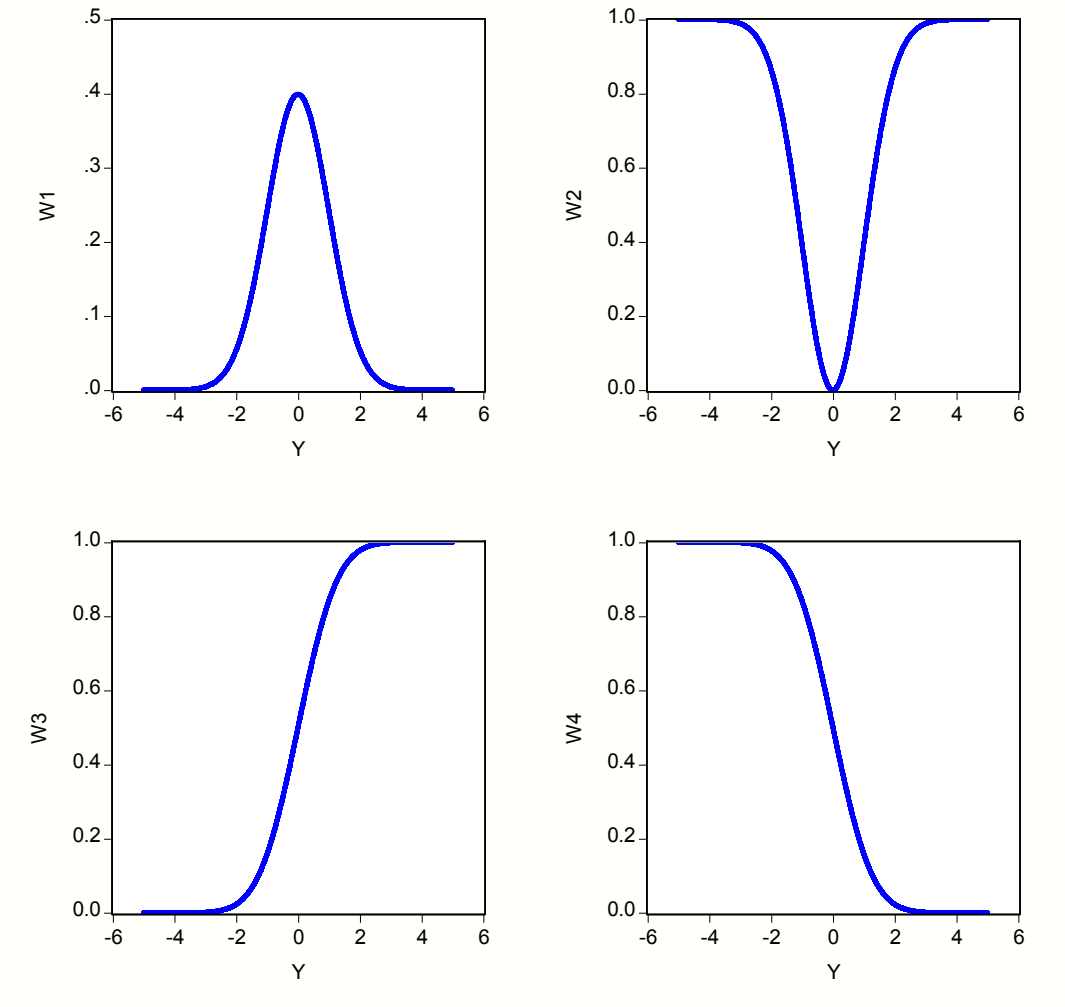

Figure 2: Linear Model (LIN) ML parameter estimates

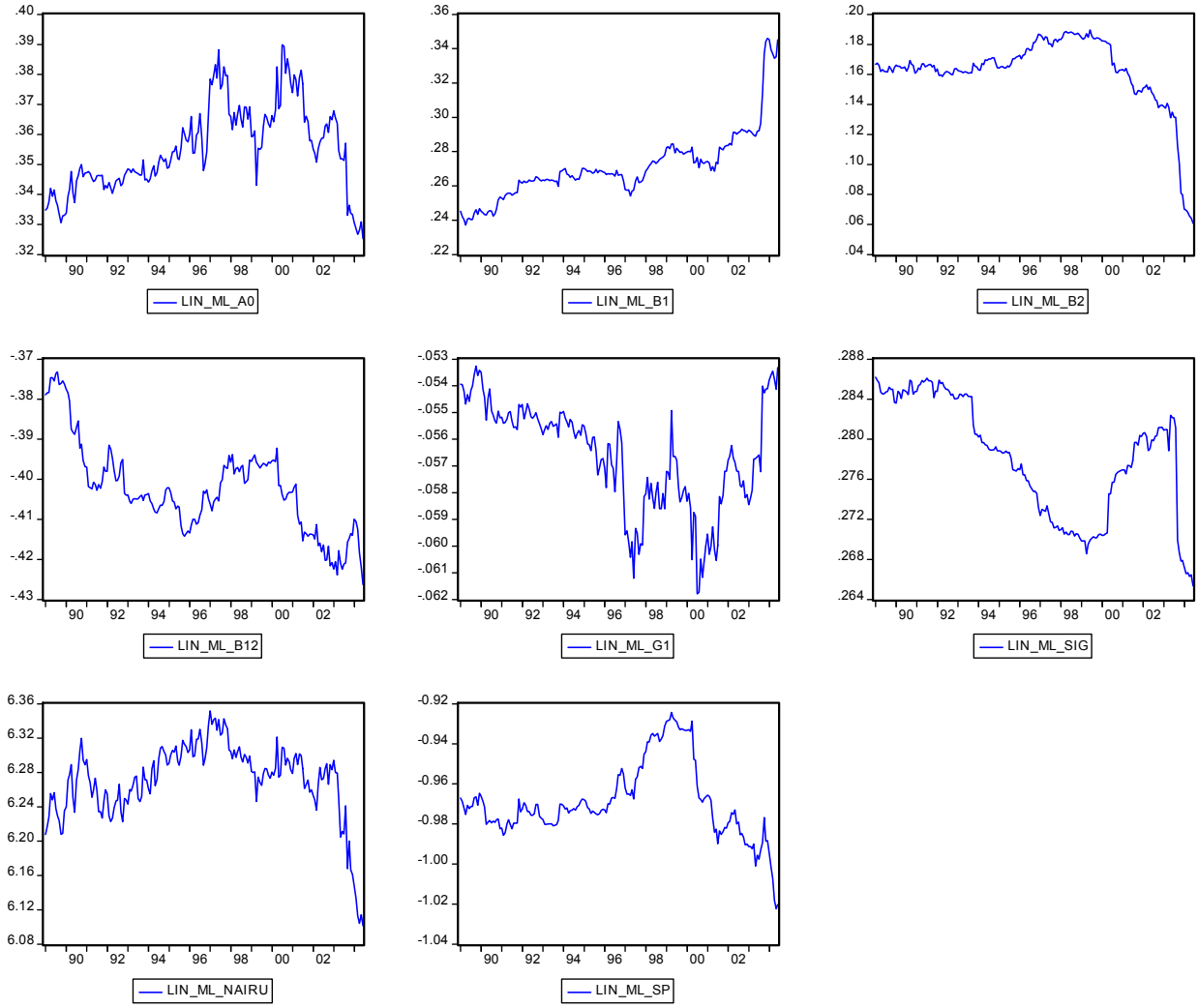

Figure 3: Markov Switching MS1 model ML parameter estimates

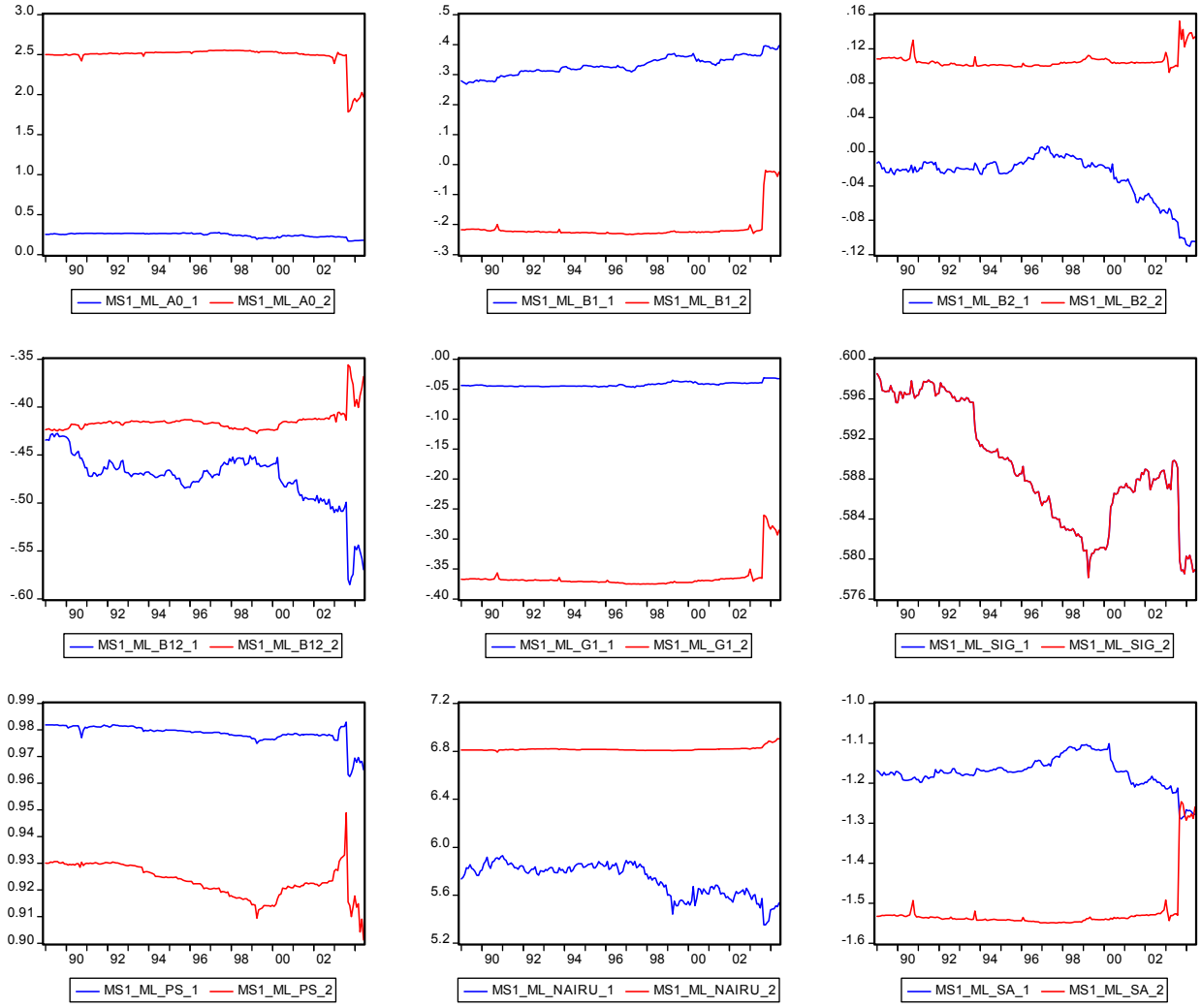

Figure 4: Markov Switching MS2 model ML parameter estimates

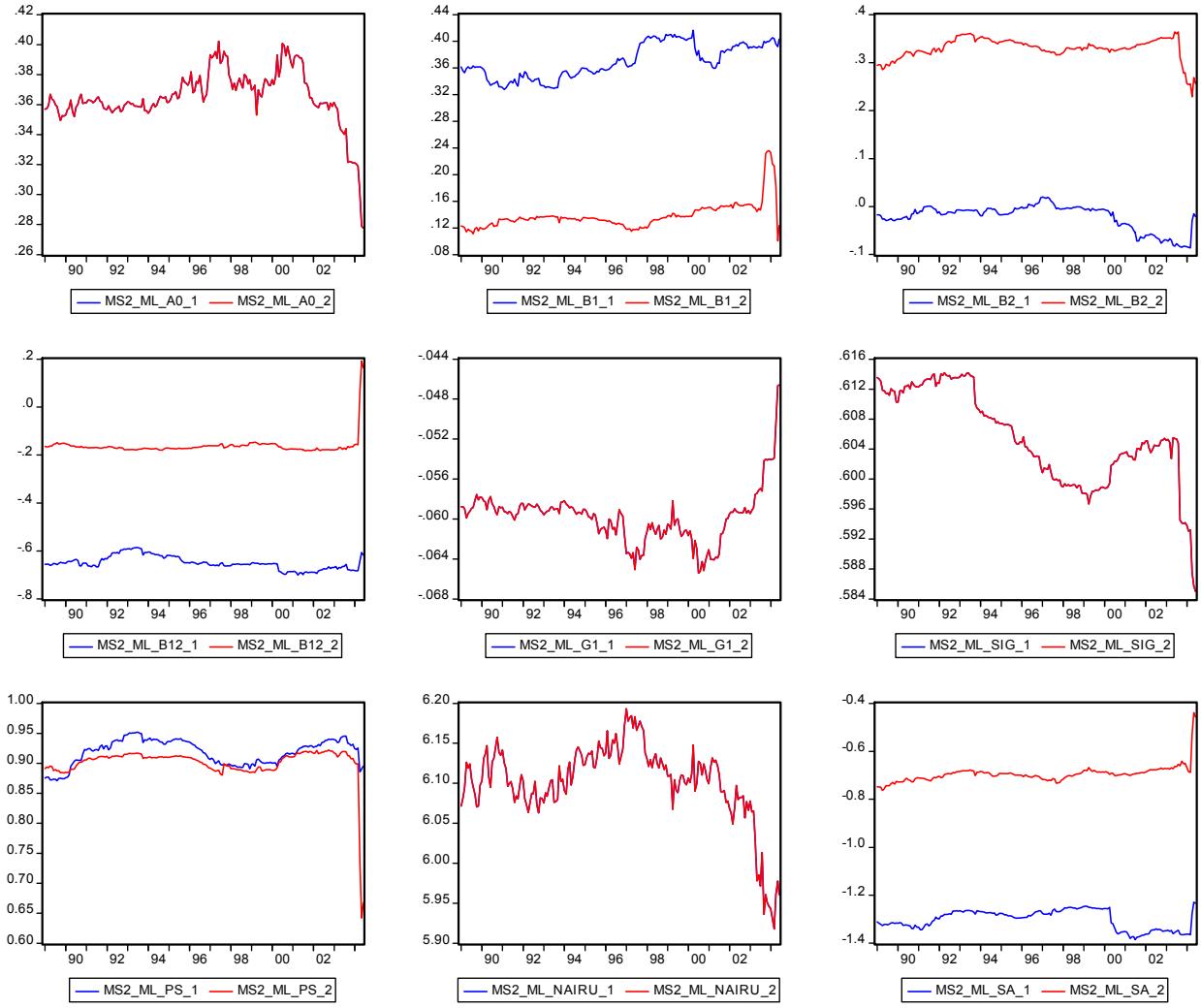

Figure 5: sample weights used in DF comparison tests.