

Non-nested model selection in unstable environments

Raffaella Giacomini and Barbara Rossi

University of California, Los Angeles and Duke University

May 2006

PRELIMINARY AND INCOMPLETE

Abstract

We consider non-nested model selection tests in the presence of possible data and parameter instabilities. The novelty of our approach is that we allow the models' relative performance to be varying over time, whereas existing model selection techniques look for an overall best model. We argue that the time path of the models' relative performance may contain useful information that is lost when seeking a model that performs best "on average". We consider two competing non-nested nonlinear and possibly misspecified dynamic models and provide statistical and graphical methods to: (1) analyze the evolution of the models' relative performance over historical samples; and (2) monitor the models' relative performance in real time, as new data becomes available. Our methods are valid under general data assumptions and can be applied to multivariate models that are estimated by a variety of techniques including ML, GMM and Bayesian methods. An empirical application provides insights into the time variation in the performance of Smets and Wouters' (2003) DSGE model of the European economy relative to that of Bayesian VARs.

Keywords: Model Selection Tests, Fluctuation Tests, Sequential Tests, Structural Change, Forecast Evaluation, Misspecification

Acknowledgments: We are grateful to F. Smets and R. Wouters and A. Justiniano for providing their codes. We also thank Giorgio Primiceri and seminar participants at the Empirical Macro Study Group at Duke University for comments.

J.E.L. Codes: C22, C52, C53

1 Introduction

We consider the problem of model selection in an environment that is possibly characterized by structural instability and model misspecification. We focus on the problem of comparing the performance of two competing non-nested nonlinear dynamic models by assessing the statistical significance of the relative value of the objective function used to estimate the model's parameters. Previous approaches to model comparison either rely on in-sample hypothesis testing (e.g., Vuong, 1989, Rivers and Vuong, 2002, Sin and White, 1994 etc.) or on out-of-sample predictive ability testing (e.g., Diebold and Mariano, 1995, West, 1996, McCracken, 2000). In both cases, the alternative hypothesis of the test is that one of the two models is preferable, either because it better fits the data in the full sample or because it forecasts better, *on average*, over the out-of-sample period. In the presence of structural instability, however, the relative performance of the two models may itself be time-varying, and thus the existing approaches - which average out this evolution over time - may involve a loss of information. The possibility of time variation in relative performance is supported by recent empirical evidence (e.g., Stock and Watson, 2003), showing that even though certain models have good out-of-sample performance in certain periods, they do not necessarily perform well in other periods. This suggests that methods that provide insight into the evolution over time in the performance of competing models may be of interest to both forecasters and policy makers.

In this paper, we attempt to provide a solution to the problem by proposing two approaches. The first is a “fluctuation” test that provides insight into the evolution of the relative performance over time using historical data. The second is a “sequential test” that can be used to monitor the relative performance of the models in real time, and detect any deviation of the relative performance from that observed over the historical sample.

Our methods are valid under general conditions on the data-generating process, on the models under scrutiny and on the estimation techniques. In particular, the data can be characterized by structural instability and the models can be multivariate, non-linear, and estimated by a variety of techniques typically used in empirical applications, including OLS, ML, GMM, IV and Bayesian estimation. Importantly, our tests allow both models to be misspecified under the null hypothesis.

This paper sits at the intersection of various strands of the literature on forecast evaluation and model selection. Our test is related to both the in-sample non-nested model selection tests of Vuong (1989), Rivers and Vuong (2002) and to the out-of-sample predictive ability tests of West (1996) and McCracken (2000) because of our focus on relative performance given a general objective (loss) function. The main difference is that we test a different null hypothesis. Vuong (1989) tests the null that the two models have equal in-sample fit in expectation, where the expectation is constant over time because of the assumption of i.i.d. data. West (1996) and McCracken (2000) test the

null that the two models have equal out-of-sample accuracy in expectation, where the expectation is constant because of the assumption of stationarity in the data. Rivers and Vuong (2002) relax the assumption of identical distribution - implying that the expected relative performance may be time varying - but test the null that the two models are equally accurate *asymptotically*, which amounts to postulating that any time variation in relative performance that is present in finite samples eventually disappears as the sample size increases. In all the above cases, the test is based on computing the average relative performance for the competing models, either over the full sample or over the out-of-sample period only. Contrary to these contributions, we test the null hypothesis that the two models are equally accurate at every point in time, which is a stronger requirement than that in Rivers and Vuong (2002). Our approach implicitly recognizes that the entire time path of relative performance may contain information that is of interest to the user. In the context of out-of-sample predictive ability testing, Giacomini and White (2003) similarly argue that relative forecast performance may differ in different states of the economy. They take however a different approach, which involves assessing whether one can construct a forecasting model for the time series of the out-of-sample relative losses using economic variables. In the context of in-sample model selection tests, Rossi (2005) also stresses the importance of parameter instabilities; her approach, however, focuses exclusively on the case of nested and correctly specified models. We instead propose capturing the possible time variation in relative performance by considering sequences of test statistics a' la Rivers and Vuong (2002) for non-nested and possibly misspecified models computed over expanding or rolling samples. We also provide boundary lines that are crossed by the sequence of test statistics under the null hypothesis with known probability, so that instability is detected when any of such statistics crosses the boundary lines. We call this test a "fluctuation test", which highlights the relation of our test to the fluctuation tests for parameter instability proposed by Ploberger, Kramer and Kontrus (1989), the main difference being that we focus on stability of relative performance of possibly misspecified models rather than on stability of parameters within a correctly specified model.

What all the above tests have in common is the focus on analyzing past behavior in relative performance. We further propose a sequential test that allows the user to assess, in real time, whether a model selection decision reached in the past is reversed or confirmed by the arrival of new data. Sequential tests have been considered in the literature, but to address the more specific problems of monitoring structural change in model's parameters (Chu, Stinchcombe and White, 1996) or for assessing whether some variables have predictive content for other variables in real time (Inoue and Rossi, 2005). Our test answers a different question (we compare models based on general measures of fit), and it is valid under less restrictive assumptions (e.g., we allow both models to be misspecified under the null hypothesis).¹

¹If a researcher is worried about instabilities, he might take a stand and explicitly estimate a time-varying parame-

The methods proposed in this paper have many useful applications. The recent developments in macroeconomics (Smets and Wouters, 2003, Del Negro et al., 2004) have shown that it is possible to estimate dynamic stochastic general equilibrium (DSGE) models whose performance is comparable to that of VARs. However, the measures of relative performance used in these papers are average measures over the full sample, which might hide important changes in the relative performance of such models over time. We select one such representative DSGE model – the Smets and Wouters’ (2003) DSGE model for the European area – and offer some insight into the time variation in the performance of their model relative to that of Bayesian VARs.

2 Motivation

Suppose a researcher is interested in comparing the relative performance of two competing non-nested models (more precise definitions will be introduced later), where the models are estimated by maximizing some objective function, $Q_T(\cdot)$ over a sample of size T . For example, for Maximum Likelihood estimation (ML), the objective function for the first model is $Q_T(\theta) = \frac{1}{T} \sum_{t=1}^T \ln f_t(\theta)$, where $\ln f_t(\theta)$ is the conditional log-likelihood of the model at time t , θ is the parameter and $T = 100$ is the sample size. Similarly, for the second model, the objective function is $Q_T(\gamma)$, where γ denotes the parameters of the second model.

A natural way of proceeding is to use Vuong’s (1989) test for model comparison. The test is a rescaled version of the likelihood ratio test statistic, $Q_T(\theta) - Q_T(\gamma) = \frac{1}{T} \sum_{t=1}^T [\ln f_t(\theta) - \ln f_t(\gamma)]$. Vuong (1989) calculates its asymptotic distribution for the i.i.d. case, whereas Rivers and Vuong (2002) extend these results to the more general case of non-nested dynamic models.

Figure 1 shows a possible sample path of the relative performance of the two model, so that the first model is better over the initial quarter of the sample and the second model is better afterwards. The solid line represents the actual relative performance *at each point in time*, $\ln f_t(\theta) - \ln g_t(\gamma)$, whereas the dashed line shows the average measure of the relative performance *up to time t* based on a *recursive procedure*, $Q_t(\theta) - Q_t(\gamma) = \frac{1}{t} \sum_{j=1}^t [\ln f_j(\theta) - \ln f_j(\gamma)]$. In the figure, Vuong’s (1989) likelihood ratio test statistic is a rescaled version of $Q_T(\theta) - Q_T(\gamma)$, the value of the dashed line at $T = 100$, which in this case is close to zero. Therefore, if the policymaker selected the model based on a measure of the average performance over the full sample, he would likely conclude that the two models are equivalent.

The actual plot of the relative performance shows instead that the second model is a better description of the data at time T . Limiting the analysis to the average performance therefore hides

ter model. However, note that our main interest is in evaluating the relative performance of two possibly misspecified models over time, not to estimate the models’ parameters. Also, our procedure allows us not to restrict ourselves to a particular time-varying parameter model, but rather to allow the models’ parameters to be recursively re-estimated over time over an expanding or rolling window.

important information otherwise contained in the data, and would therefore lead the researcher to wrongly conclude that the two models are equally good, whereas the second model is actually the one that should be used for policy analysis as well as forecasting at time T .

The goal of this paper is to provide methodologies that overcome such a problem by going beyond focusing on the average performance, and instead track the relative performance of the models over time, by re-estimating the test statistic for relative performance over time, on either an expanding or on a rolling window of data. We will therefore provide confidence bands (like those showed by the dotted lines in Figure 1) that use the information contained in the full time path of the relative performance of the two models (as opposed to its average relative performance at time T) to allow the researcher to select the best model in such problematic situations.

INSERT FIGURE 1 HERE

3 Econometric methodology

We consider two different scenarios. In the first, the researcher is interested in learning about the relative performance of the two models over a historical sample. In this case, we propose a *fluctuation* test that plots the relative performance of the two candidate models over time together with appropriate boundary lines, which, if crossed, signal instability. In the second, the researcher wants to monitor the relative performance of two models as new data becomes available, in order to establish whether some model selection decision reached on the historical sample (e.g., that one model was more accurate on average) is confirmed or reversed by new evidence. For this purpose, we propose a *sequential* test that plots the evolution of the relative performance in the post-historical sample period, together with appropriate boundary lines which, if crossed, signal a change in the model selection decision.

The derivation of our test builds on existing results for in-sample model selection tests, in particular Rivers and Vuong (2002), who generalize the non-nested model selection tests of Vuong (1989) to a setting permitting data heterogeneity and dependence. The main difference is that we test the more restrictive null hypothesis that the competing models are equally accurate at every point in time, whereas the focus in the above papers is on whether the two models are equally accurate asymptotically.

3.1 Notation and assumptions

We now introduce the notation and discuss the assumptions on the data, the models and the estimation procedures. We are interested in selecting a model for y_t , which could be a scalar or a vector, using a collection of other variables z_t (possibly containing lags of y_t). We let $x_t = (y_t', z_t)'$.

The assumptions (formally stated in Assumption 1 below) are rather weak. In particular, we treat the data generating process as unknown, and only assume that the data are weakly dependent, which rules out unit roots. A key feature of our approach is that we allow the data generating process to be changing over time. This assumption adds notational and computational complexity, but is necessary given our focus on the dynamic nature of the model selection procedure.

We consider two competing non-nested nonlinear dynamic models for y_t . Importantly, we allow both models to be potentially misspecified. The only assumption is that, at time t , each model is estimated by maximizing the same scalar objective function $Q_t(\cdot)$ that depends on the sample (x_1, \dots, x_t) and on parameters θ ($p \times 1$) $\in \Theta$ for model 1 and γ ($q \times 1$) $\in \Gamma$ for model 2:

$$\hat{\theta}_t \equiv \arg \max_{\theta \in \Theta} Q_t(\theta) \text{ and } \hat{\gamma}_t \equiv \arg \max_{\gamma \in \Gamma} Q_t(\gamma). \quad (1)$$

This broad class of extremum estimators includes linear and nonlinear least squares, linear and nonlinear GMM and ML as special cases. Using the results of Fernandez-Villaverde and Rubio-Ramirez (2004), we can further incorporate Bayesian estimation, in which case the objective function corresponds to the log marginal data density for the model divided by the sample size.

The fluctuation test is performed by considering the historical sample of data from time $t = 1$ to $t = T$, whereas the sequential test begins in the post-historical sample $t = T + 1, T + 2, \dots$. The fluctuation test involves estimating the models recursively starting from observation $R < T$ using either expanding samples of size R, \dots, T (which we call the “recursive approach”) or on rolling samples of fixed size R (the “rolling approach”). R is a user-defined constant which, in the derivation of the asymptotic results, is allowed to grow with the sample size. For the sequential test, one similarly re-estimates the models recursively using expanding samples of size $T + 1, T + 2, \dots$. For both tests, the procedure yields sequences of differences in the estimated objective functions for the two models, $Q_t(\hat{\theta}_t) - Q_t(\hat{\gamma}_t)$, which - appropriately normalized - will constitute the basis for our tests. For convenience, below we define the estimated objective functions for model 1 (similar definitions hold for model 2) for the different estimation methods and estimation schemes:

	Recursive	Rolling	
ML	$Q_t(\hat{\theta}_t) = \frac{1}{t} \sum_{j=1}^t \ln f(x_j; \hat{\theta}_t)$	$Q_t(\hat{\theta}_t) = \frac{1}{R} \sum_{j=t-R+1}^t \ln f(x_j; \hat{\theta}_t),$	(2)
<i>GMM</i>	$Q_t(\hat{\theta}_t) = -\frac{1}{2} \bar{g}_t(\hat{\theta}_t)' \widehat{W}_{\theta, t} \bar{g}_t(\hat{\theta}_t)$ $\bar{g}_t(\hat{\theta}_t) \equiv \frac{1}{t} \sum_{j=1}^t g(x_j; \hat{\theta}_t)$	$Q_t(\hat{\theta}_t) = -\frac{1}{2} \bar{g}_t(\hat{\theta}_t)' \widehat{W}_{\theta, t} \bar{g}_t(\hat{\theta}_t)$ $\bar{g}_t(\hat{\theta}_t) \equiv \frac{1}{R} \sum_{j=t-R+1}^t g(x_j; \hat{\theta}_t)$	(3)

$$\text{Bayesian} \quad Q_t(\cdot) = \int_{\Theta} \frac{1}{t} \sum_{j=1}^t \ln f(x_j; \theta) \pi(\theta) d\theta \quad Q_t(\cdot) = \int_{\Theta} \frac{1}{R} \sum_{j=t-R+1}^t \ln f(x_j; \theta) \pi(\theta) d\theta, \quad (4)$$

where $\ln f(x_j; \hat{\theta}_t)$ is the conditional log likelihood at time j ; $g(x_j; \hat{\theta}_t)$ is a $(k \times 1)$ moment condition; $\widehat{W}_{\theta,t}$ is a $(k \times k)$ symmetric and positive definite matrix;² and $\pi(\theta)$ is the prior.

Let θ_t^* and γ_t^* denote the pseudo-true values of the parameter estimates (e.g., White, 1994), which are indexed by t to indicate that they may differ in different samples due to possible instability in the data. We also define $Q_t^*(\theta_t^*) = E\left(t^{-1} \sum_{j=1}^t \ln f(x_j; \theta_t^*)\right)$ for MLE and Bayesian methods,³ and $Q_t^*(\theta_t^*) = (1/2) E(\bar{g}_t(\theta_t^*))' W E(\bar{g}_t(\theta_t^*))$ for GMM.

The test statistic is normalized by an estimator of the square root of the variance σ_t^2 of the rescaled relative fit, evaluated at the pseudo-true values:

$$\sigma_t^2 \equiv \text{var}\left(\sqrt{t}(Q_t(\theta_t^*) - Q_t(\gamma_t^*))\right). \quad (5)$$

We derive results based on the following two high-level assumptions (only stated for the first model), that essentially guarantee consistency of the parameter estimates and weak convergence of the objective functions. In what follows, we let ∇ and ∇^2 denote the gradient and the second derivative operators with respect to the parameter, \implies denote weak convergence on $[0, 1]$ and \xrightarrow{p} denote convergence in probability.

Assumption 1. (a) $\Theta \subseteq \mathbb{R}^p$ and $\Gamma \subseteq \mathbb{R}^q$ are compact; (b1) $\left\{\frac{1}{\sqrt{T}} \sum_{j=1}^t \ln f(x_j; \theta)\right\}$ obeys a FCLT with $\lim_{t \rightarrow \infty} t^{-1} E\left[\left(\sum_{j=1}^t \ln f(x_j; \theta)\right)^2\right]$ positive definite; (b2) $\left\{\sqrt{T} \sum_{j=1}^t g(x_j; \theta)'\right\}$ obeys a FCLT with $\lim_{t \rightarrow \infty} t^{-1} E\left[\sum_{j=1}^t g(x_j; \theta) \sum_{j=1}^t g(x_j; \theta)'\right]$ positive definite; (c) $\sqrt{T}(\hat{\theta}_t - \theta_t^*) = O_p(1)$ uniformly in $\tau \equiv t/T \in [0, 1]$; (d) $\nabla^2 Q_t(\theta) - \nabla^2 Q_t^*(\theta) = o_p(1)$ uniformly on Θ in $\tau \in [0, 1]$; (e) There exists a sequence of symmetric and positive definite matrices $W_{\theta,t}$ such that $\widehat{W}_{\theta,t} - W_{\theta,t} = o_p(1)$ uniformly in $\tau \in [0, 1]$; (f1) $\bar{g}_t(\theta) - E(\bar{g}_t(\theta)) \xrightarrow{p} 0$ uniformly in Θ ; (f2) $t^{-1} \sum_{j=1}^t \ln f(x_j; \theta) - E\left(t^{-1} \sum_{j=1}^t \ln f(x_j; \theta)\right) \xrightarrow{p} 0$ uniformly in Θ ; (g) $R/T \rightarrow \rho$ as $T \rightarrow \infty$.

²Possible weighting matrices $\widehat{W}_{\theta,t}$ include the optimal weighting matrix $\left[\left(\frac{1}{t} \sum_{j=1}^t g(x_j; \hat{\theta}_t)\right) \left(\frac{1}{t} \sum_{j=1}^t g(x_j; \hat{\theta}_t)\right)'\right]^{-1}$.

³As we will explain in detail later, we rely on the asymptotic equivalence result between the average log marginal data density and the loglikelihood evaluated at the pseudo true parameter values, as shown in Fernandez-Villaverde and Rubio-Ramirez (2004).

3.2 The fluctuation test

We test the null hypothesis:⁴

$$H_0 : Q_t^*(\theta_t^*) - Q_t^*(\gamma_t^*) = 0 \text{ for all } t = 1, \dots, T, \quad (6)$$

by considering a sequence of recursive estimates of the average relative performance $(Q_t(\hat{\theta}_t) - Q_t(\hat{\gamma}_t))$, $t = R, \dots, T$. The starting period $t = R$ for the computation of the first measure of average relative performance is user-defined, although, in practice, R should be chosen as to guarantee that the parameters are consistently estimated in order for the test to have desirable size properties. At each time t , we further normalize the average relative performance by its standard deviation.

The sample path of the recursively estimated measures of relative performance contains information on their evolution over time. We provide a way to test whether this sample path departs from its hypothesized value of zero by plotting it together with boundary lines that are crossed with probability α . The basic intuition is that, by using standard results on FCLT, under the null hypothesis the sequence of test statistics is well approximated by functions of Brownian Motions.

Proposition 1 (Fluctuation test) *If the models are non-nested and estimated by either rolling or recursive GMM, MLE, or Bayesian estimation methods, the statistics for each $t = R, \dots, T$ are*

$$F_t^{rec} = \hat{\sigma}_t^{-1} \sqrt{t} (Q_t(\hat{\theta}_t) - Q_t(\hat{\gamma}_t)), \quad (7)$$

$$F_t^{roll} = \hat{\sigma}_t^{-1} \sqrt{R} (Q_t(\hat{\theta}_t) - Q_t(\hat{\gamma}_t)), \quad (8)$$

where $\hat{\sigma}_t^2$ is given in Corollary 2. The critical values at time t for a significance level α are:

$$c_{\alpha,t}^{rec} = \pm k_{\alpha}^{rec} \sqrt{\frac{T-R}{t}} \left(1 + 2\frac{t-R}{T-R}\right), \quad (9)$$

$$c_{\alpha,t}^{roll} = \pm k_{\alpha}^{roll} \quad (10)$$

where k_{α}^{rec} solves $P\{\sup_{\tau} |\mathcal{B}(\tau)|/\tau > k_{\alpha}^{rec}\} = \alpha$, and k_{α}^{roll} solves $P\{\sup_{\tau} |[\mathcal{B}(\tau) - \mathcal{B}(\tau - \rho)]/\sqrt{\rho}| > k_{\alpha}^{roll}\} = \alpha$, for τ such that $t = \lceil \tau T \rceil = R, \dots, T$, $R = \lceil \rho T \rceil$, and $\mathcal{B}(\cdot)$ a standard univariate Brownian motion. Typical values of $(\alpha, k_{\alpha}^{rec})$ are $(0.01, 1.143)$, $(0.05, 0.948)$ and $(0.10, 0.850)$, whereas typical values of $(\alpha, k_{\alpha}^{roll})$ are reported in Table 1.

⁴Note that under a more restrictive null hypothesis where also the parameters are constant over time then (6) reduces to $E[\ln f_t(x_t; \theta^*) - \ln f_t(x_t; \gamma^*)] = 0 \forall t$. One could, in such cases, design optimal tests along the line of Rossi (2005).

Corollary 2 (Variance estimation) (a) (MLE estimator) Under Assumption 1, let $\hat{\sigma}_t^2$ be a HAC estimator of the asymptotic variance $\text{var}(\sqrt{t}(Q_t(\theta_t^*) - Q_t(\gamma_t^*)))$, for example

$$\begin{aligned} \hat{\sigma}_t^2 &\equiv S^{-1} \sum_{j=t-S+1}^t \left(\ln f(x_j; \hat{\theta}_t) - \ln f(x_j; \hat{\gamma}_t) \right)^2 \\ &+ 2 \left[S^{-1} \sum_{j=t-S+1}^{l_t} w_{t,j} \sum_{i=j}^t \left(\ln f(x_i; \hat{\theta}_t) - \ln f(x_i; \hat{\gamma}_t) \right) \left(\ln f(x_{i-j}; \hat{\theta}_t) - \ln f(x_{i-j}; \hat{\gamma}_t) \right) \right], \end{aligned} \quad (11)$$

where: $S = t$ for the recursive case and $S = R$ for the rolling case; $\{l_t\}$ is a sequence of integers such that $l_t \rightarrow \infty$ as $T \rightarrow \infty$, $l_t = o(T)$ and $\{w_{t,j} : t = 1, 2, \dots; j = 1, \dots, l_t\}$ is a triangular array such that $|w_{t,j}| < \infty$, $t = 1, 2, \dots, j = 1, \dots, l_t$ and $w_{t,j} \rightarrow 1$ as $T \rightarrow \infty$ for each $j = 1, \dots, l_t$ (cf. Andrews, 1991, and Newey and West, 1987). Then $\hat{\sigma}_t^2 - \sigma_t^2 \xrightarrow{p} 0$.

(b) (GMM estimator) Under Assumption 1, let $\hat{\sigma}_t^2$ be a HAC estimator of the asymptotic variance $\text{var}(\sqrt{t}(Q_t(\theta_t^*) - Q_t(\gamma_t^*)))$, for example

$$\hat{\sigma}_t^2 = \begin{bmatrix} \bar{g}_t(\hat{\theta}_t) \widehat{W}_{\theta,t} & -\bar{g}_t(\hat{\gamma}_t) \widehat{W}_{\gamma,t} \end{bmatrix} \widehat{V}_t \begin{bmatrix} \bar{g}_t(\hat{\theta}_t) \widehat{W}_{\theta,t} & -\bar{g}_t(\hat{\gamma}_t) \widehat{W}_{\gamma,t} \end{bmatrix}', \quad (12)$$

where \widehat{V}_t is a HAC estimator of the asymptotic variance $\text{var}(\sqrt{t}(\bar{g}_t(\theta_t^*)', \bar{g}_t(\gamma_t^*)'))'$:

$$\begin{aligned} \widehat{V}_t &\equiv S^{-1} \sum_{j=t-S+1}^t \left(g(x_j; \hat{\theta}_t) - g(x_j; \hat{\gamma}_t) \right)^2 \\ &+ 2 \left[S^{-1} \sum_{j=t-S+1}^{l_t} w_{t,j} \sum_{i=j}^t \left(g(x_i; \hat{\theta}_t) - g(x_i; \hat{\gamma}_t) \right) \left(g(x_{i-j}; \hat{\theta}_t) - g(x_{i-j}; \hat{\gamma}_t) \right) \right]. \end{aligned} \quad \text{Then } \hat{\sigma}_t^2 - \sigma_t^2 \xrightarrow{p} 0.$$

(c) (Bayesian estimator) Under Assumption 1 and covariance stationarity, let $\hat{\sigma}^2$ be:

$$\hat{\sigma}^2 \equiv \lambda^{-1} \text{var} \left(\sqrt{(T-R+1)} \sum_{t=R}^T \left(Q_t(\hat{\theta}_t) - Q_t(\hat{\gamma}_t) \right)^2 \right), \quad (13)$$

where λ equals: $2 \left[1 - \left(\frac{T-R}{R} \right)^{-1} \ln \left(\frac{T}{R} \right) \right]$ in the recursive case; $\left[\left(\frac{T-R}{R} \right) - \left(\frac{T-R}{R} \right)^2 / 3 \right]$ in the rolling case if $R \geq T/2$; and $\left[1 - \left(3 \frac{T-R}{R} \right)^{-1} \right]$ in the rolling case if $R < T/2$. Then $\hat{\sigma}^2 - \sigma^2 \xrightarrow{p} 0$.

Remarks. (i) Note that, unlike in Rivers and Vuong (2002), we are able to provide a consistent estimator of the asymptotic variance without imposing any restrictions on the heterogeneity of the expectations $Q_t^*(\theta_t^*)$ and $Q_t^*(\gamma_t^*)$. This is because under our null hypothesis (6) the sequence $\{Q_t(\theta_t^*) - Q_t(\gamma_t^*)\}$ has zero mean for all t , whereas the null hypothesis considered by Rivers and Vuong (2002) only requires that the sequence (appropriately normalized) has zero mean asymptotically. As a result, under our null hypothesis the asymptotic variance of $\sqrt{T}(Q_t(\theta_t^*) - Q_t(\gamma_t^*))$ is consistently estimated by $\hat{\sigma}_t^2$ even in the presence of arbitrary misspecification and data heterogeneity.

(ii) Note that (6) is the same null hypothesis as in Diebold and Mariano (1995) and West (1996). These out-of-sample tests have extensively been used in the empirical literature to deal with the

problem of testing which theoretical model is a better description of the observed data when the models can be possibly misspecified or there might be an underlying problem of parameter instability. In fact, these tests can detect parameter instability because they use rolling or sequential methods to recursively estimate the parameters. The reason why they can detect model misspecification is because they compare some loss function of the forecast errors of the unrestricted model with those of the restricted model. Even if these out-of-sample tests can potentially detect both model misspecification and parameter instability, they only test that the two models have equal predictive ability on average over the out of sample period. Therefore, there could be situations where the best model changes over time, but the test does not have any power to detect that (see the Monte Carlo section for an example). This paper, instead, proposes in-sample tests for the null hypothesis that the two models' losses are equal at each point in time.

3.3 The sequential test

In this section, we derive a procedure that allows the researcher to monitor the model selection decision in the post-historical sample period. Suppose that model 1 was selected as the best performing model in the historical sample of data up to time T , based on the fact that it yielded a significantly superior performance, i.e., that $Q_T^*(\theta_T^*) > Q_T^*(\gamma_T^*)$.

We test the hypothesis that model 1 continues to be the best model for all subsequent periods:

$$H_0 : Q_t^*(\theta_t^*) - Q_t^*(\gamma_t^*) \geq 0 \text{ for } t = T + 1, T + 2, \dots, \quad (14)$$

against the one-sided alternative $H_1 : Q_t^*(\theta_t^*) - Q_t^*(\gamma_t^*) < 0$ at some $t \geq T$.

We test this hypothesis sequentially, that is, by considering a sequence of test statistics, together with appropriate critical values that control the overall size of the procedure. The following proposition provides the test statistic and its critical values.

Proposition 3 (Sequential test) *The test statistic for each $t = T + 1, T + 2, \dots$ is*

$$J_t = \hat{\sigma}_t^{-1} \sqrt{t} \left(Q_t(\hat{\theta}_t) - Q_t(\hat{\gamma}_t) \right), \quad (15)$$

where $\hat{\sigma}_t^2$ is as in Corollary 2. The critical value at time t for a level α test is:

$$c_{\alpha,t} = -\sqrt{k_\alpha^2 + \ln(t/T)}, \quad (16)$$

where the exact expression for k_α is given in the Appendix. Typical values of (α, k_α) are $(0.05, 2.7955)$ and $(0.10, 2.5003)$.

4 Monte Carlo simulation

This section provides a simple analysis of the size and power properties of the procedures proposed in this paper. We compare them with the standard test for model selection proposed by Vuong (1989), and commonly used in the literature, and with standard out of sample tests of forecast comparison.

We consider the following two scenarios: first, the researcher analyzes the relative performance over the historical sample, i.e. he tests whether the two models are equally good in a given sample of T observations; then, for each observation $T+1$, $T+2$, ... he tests the null hypothesis again in real time.

For the tests based on the historical performance (the fluctuation and the joint tests), we consider the following Data Generating Process:

$$y_t = \beta_t x_t + \gamma_t z_t + \varepsilon_t, \quad t = 1, 2, \dots, T, \quad (17)$$

where $T = 400$, $R = 200$, $\{x_t, z_t, \varepsilon_t\}$ are independent standard normal random variables. Furthermore:

$$\beta_t = \beta_0 + \beta_{11} \cdot 1(R + 20 < t \leq R + 50) + (1 - \beta_{11}) \cdot 1(t > R + 50)$$

$$\gamma_t = \gamma_0 + \gamma_{11} \cdot 1(R + 20 < t \leq R + 50) + (1 - \gamma_{11}) \cdot 1(t > R + 50)$$

The two competing models are the following. The first model is: $y_t = \beta x_t + u_{1,t}$, whereas the second model is: $y_t = \gamma z_t + u_{2,t}$. We let $\beta_0 = \gamma_0$. Note that the null hypothesis (that the models are equally good) holds when $\beta_{11} = \gamma_{11} = 0.5$, in which case the empirical rejection rates of the tests should be equal to the nominal size. Table 1 part (a) compares empirical rejection probabilities for the fluctuation tests, eq. (7) and (8), denoted respectively by " F_t^{rec} " and " F_t^{roll} ", the full sample Vuong (1989) test, denoted by "Full sample test"; and tests of out of sample forecasting ability based on split-sample, recursive, or rolling estimation techniques, denoted respectively by " OOS_{fix} ", " OOS_{rec} " and " OOS_{roll} ".

For the sequential tests, the Data Generating Process is the same as (17). However, we now suppose that, after having done the in-sample analysis described above, and having found that the two models are equally good, the researcher moves forward in time as new data come in, and recursively tests in real time whether the two models are equivalent or not. That is, at observation $T+1$ the researcher starts testing recursively at each period whether the two models are equally good. Table 2, part (b), shows the empirical rejection probabilities after additional 200, 300 and 400 periods ($t/T=1.5$ ($=600/400$), 1.75, 2) of both the J_t and the Vuong tests implemented at each period.

In all experiments, the number of Monte Carlo replications is 5,000, and the nominal size of the tests is $\alpha = 0.05$. Results are reported in Table 2.

The Monte Carlo results in Table 1(a) show that the Fluctuation tests proposed in this paper have the correct size. The power comparison shows that there are situations in which the full sample and out of sample tests have no power at all to distinguish between two models whose relative performance changes over time (see the line $\beta_{11} = 0.95$, $\gamma_{11} = 0.4$), whereas our Fluctuation tests successfully reject the null hypothesis. Panel (b) in Table 2 shows that the recursive test successfully controls size, whereas the full sample test clearly overrejects.

5 Empirical application: time-variation in the performance of DSGE vs. VAR models

In a highly influential paper, Smets and Wouters (2003) (henceforth SW) show that a DSGE model of the European economy - estimated using Bayesian techniques over the period 1970:2-1999:4 - is able to fit the data as well as atheoretical Bayesian VARs (BVARs). Furthermore, they find that the parameter estimates from the DSGE model have the expected sign. Perhaps for these reasons, this model has attracted a lot of interest from forecasters and central banks. SW's model features include sticky prices and wages, habit formation, adjustment costs in capital accumulation and variable capacity utilization. The model is estimated using seven variables: GDP, consumption, investment, prices, real wages, employment, and the nominal interest rate. Their conclusion that the DSGE fits the data as well as BVARs is based on the fact that the full-sample marginal data densities for the two models are of comparable magnitudes over the full sample. However, given the changes that have characterized the European economy over the sample analyzed by SW - for example, the creation of the European Union in 1993, changes in productivity and in the labor market, to name a few - it is plausible that the relative performance of theoretical and atheoretical models may itself have varied over time. In this section, we first investigate whether the DSGE parameters have been stable and then apply the techniques proposed in this paper to assess whether the performance of the DSGE model has been consistently better than that of a BVAR over the sample considered by SW.

To begin, we re-estimate the SW's model to informally gauge the extent of the variation in some of the key parameters of the model. The total sample has size $T = 118$, and we estimate both models recursively over either expanding windows with an initial sample of size $R = 70$ (recursive scheme) or a rolling window of size $R = 70$ (rolling scheme). As in SW, the first 40 data points in each sample are used to initialize the estimates of the DSGE model and as training samples for the BVAR priors. We consider both a BVAR(1) and a BVAR(2), both of which use a variant of

the Minnesota prior, as suggested by Sims (2003).⁵ The log marginal data density for the DSGE is computed as in SW (using the Laplace approximation), and is computed analytically for the BVAR.

Figure 2 displays the evolution of the mode of some representative parameters estimated by using the rolling scheme. Figure 2(a) focuses on parameters that describe the evolution of the persistence of some representative shocks (productivity, investment, government spending, and labor supply); Figure 2(b) focuses on parameters that describe the standard deviation of the same shocks; and Figure 2(c) plots monetary policy parameters. The figures also report 95% confidence bands constructed by using the estimated standard deviation in the full sample, as Smets and Wouters (2003) do. Overall, Figure 2 shows only moderate evidence of parameter variation. In particular, Figure 2(a) reveals some increase in the persistence of the productivity shock around 1990, followed by a gradual decline towards its value in the late Eighties, and an increase in the persistence of the labor supply shock in mid-Nineties. It also is worth noticing the changes in the standard deviation of the shocks in Figure 2(b) around the start of the European Union. For example, we observe a moderate stabilization (associated with a decrease) of the standard deviation of the labor supply shock around 1993, as well a moderate increase in the standard deviation of the government spending shocks around 1993 and in that of the investment shock between 1994-1997. Finally, Figure 2(c) shows some decrease in the coefficient of the lagged output gap in the empirical monetary reaction function (panel labeled “d(output gap) coeff.”), and some increase in the standard deviation of the interest rate shock (the monetary policy shock) after 1993.

INSERT FIGURE 2 HERE

Then, we apply the fluctuation test of Section 3, implemented with both the recursive and the rolling schemes, to test the hypothesis that the DSGE model and the BVAR have equal performance at every point in time over the historical sample. Figures 3 (recursive case) and 4 (rolling case) show the sequences of test statistics for the fluctuation tests, eqs. (7) and (8), constructed using the variance estimator of Corollary 2, together with confidence bands. In all cases, we find evidence that the relative performance of the DSGE and the BVARs was not constant over the sample analyzed by SW. A general conclusion that emerges from the figures is that the performance of the DSGE relative to that of the BVARs improves over time. In particular, the recursive fluctuation test concludes that the DSGE is significantly worse than the BVARs in the late 1980’s and early 1990’s, but that it outperforms a BVAR(1) and is as good as a BVAR(2) from the mid-1990s onwards. The rolling fluctuation test in Figure 4 concludes instead that the DSGE is generally worse than BVARs,

⁵The marginal data densities for the BVAR were carried out using the software provided by Chris Sims at www.princeton.edu/~sims. As in Sims (2003), for the Minnesota prior we set the decay parameter to 1 and the overall tightness to .3. We also included sum-of-coefficients (with weight $\mu = 1$) and co-persistence (with weight $\lambda = 5$) prior components.

but with a general improvement in the performance of the DSGE's from the mid-1990s that results in the two models performing equally well at the end of the sample. Importantly, Figure 3 confirms SW's conclusion that the DSGE outperforms a BVAR(1) and that it is as good as a BVAR(2) over the full sample (which corresponds to the last point in the recursive plot in Figure 3). However, our methods allow us to uncover a considerable amount of time variation in relative performance, and show that the performance of the DSGE model has progressively improved over the last decade, suggesting that this class of estimated DSGE model may be a useful tool for analyzing the recent and - possibly - future behavior of macroeconomic variables. A possible conclusion that emerges from the analysis is that, in spite of only moderate evidence of time variation in the parameters, the economic restrictions imposed by DSGE seem to be a better description of the most recent data.

INSERT FIGURES 3 AND 4 HERE

6 Conclusions

This paper provides new tests for non-nested model selection in the presence of possible data and parameter instabilities.

This paper focused on model selection techniques applicable to non-nested models. If the models of interest are instead nested, the researcher has the following options. A possible counterpart for the in-sample test (7) would be the joint test for nested model selection in the presence of underlying parameter instability proposed by Rossi (2005). The counterpart of the sequential test (15) for nested models is discussed instead in Inoue and Rossi (2005). Both tests' null hypotheses can be expressed as zero restrictions on the parameters of the largest model, and they both test jointly this null hypothesis as well as the maintained assumption that the smallest model is correctly specified.

References

- [1] Andrews, D.W.K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation", *Econometrica* 59, 817-858.
- [2] Brown, R.L., J. Durbin and J.M. Evans (1975), "Techniques for Testing the Constancy of Regression Relationships over Time with Comments", *Journal of the Royal Statistical Society, Series B*, 37, 149-192.
- [3] Chu, C. J., M. Stinchcombe and H. White (1996), "Monitoring Structural Change", *Econometrica*, 64, 1045-1065.
- [4] Del Negro, M., F. Schorfheide, F. Smets and R. Wouters (2004), "On the Fit and Forecasting Performance of New Keynesian Models", *mimeo*.
- [5] Diebold, F. X., R. S. Mariano (1995), "Comparing Predictive Accuracy", *Journal of Business and Economic Statistics*, 13, 253-263.
- [6] Fernandez-Villaverde, J., and J.F. Rubio Ramirez (2004), "Comparing Dynamic Equilibrium Models to Data: a Bayesian Approach", *Journal of Econometrics* 123, 153-187.
- [7] Giacomini, R., and H. White (2003), "Tests of Conditional Predictive Ability", manuscript, UCLA and UCSD
- [8] Inoue, A. and B. Rossi (2005), "Recursive Predictability Tests for Real-Time Data", *Journal of Business and Economic Statistics*, 23, 336-345
- [9] McCracken, M. W. (2000), "Robust Out-of-Sample Inference", *Journal of Econometrics*, 99, 195-223.
- [10] Newey, W. and D. McFadden (1994), "Large Sample Estimation and Hypothesis Testing", in Engle, R. and D. McFadden, *Handbook of Econometrics*, Vol. IV, Amsterdam: Elsevier-North Holland.
- [11] Newey, W., and K. West (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica* 55, 703-708.
- [12] Ploberger, W., W. Kramer and K. Kontrus (1989), "A New Test for Structural Stability in the Linear Regression Model", *Journal of Econometrics*, 40, 307-318.
- [13] Rivers, D. and Q. Vuong (2002), "Model Selection Tests for Nonlinear Dynamic Models", *Econometrics Journal*, 5, 1-39.

- [14] Rossi, B. (2005), “Optimal Tests for Nested Model Selection with Underlying Parameter Instabilities”, *Econometric Theory*.
- [15] Sin, C.Y. and H. White (1996), “Information Criteria for Selecting Possibly Misspecified Parametric Models”, *Journal of Econometrics*, 71, 207-225.
- [16] Smets, F. and R. Wouters (2003), “An Estimated Stochastic Dynamic General Equilibrium Model of the Euro Area”, *Journal of the European Economic Association*, 1, 1123-1175.
- [17] Stock, J. H. and M. W. Watson (2003), “Combination Forecasts of Output Growth in a Seven-Country Data Set,” forthcoming *Journal of Forecasting*
- [18] Vuong, Q. H. (1989), “Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses”, *Econometrica*, 57, 307-333.
- [19] West, K. D. (1996), “Asymptotic Inference about Predictive Ability”, *Econometrica*, 64, 1067-1084.
- [20] White, H. (1994), *Estimation, Inference and Specification Analysis*, Cambridge University Press, New York.

7 Appendix

Sketch of proof for Proposition 1. For the recursive case, by applying a Mean Value expansion, we have

$$\begin{aligned} & \sigma_t^{-1} \sqrt{T} (Q_t(\theta_t^*) - Q_t(\gamma_t^*)) \\ = & \sigma_t^{-1} \sqrt{T} \left(Q_t(\hat{\theta}_t) - Q_t(\hat{\gamma}_t) \right) \\ & + \sigma_t^{-1} \sqrt{T} \left\{ \left(\hat{\theta}_t - \theta_t^* \right)' \nabla^2 Q_t(\tilde{\theta}_t) \left(\hat{\theta}_t - \theta_t^* \right) - \left(\hat{\gamma}_t - \gamma_t^* \right)' \nabla^2 Q_t(\tilde{\gamma}_t) \left(\hat{\gamma}_t - \gamma_t^* \right) \right\} \end{aligned}$$

where $\tilde{\theta}_t$ ($\tilde{\gamma}_t$) denote intermediate values between $\hat{\theta}_t$ and θ_t^* ($\hat{\gamma}_t$ and γ_t^*), and where we used the fact that $\nabla Q_t(\hat{\theta}_t) = 0$ and $\nabla Q_t(\hat{\gamma}_t) = 0$. The second term is $o_p(1)$ by Assumptions 1-(c) and 1-(d). Further, $\sigma_t^{-1} \sqrt{T} (Q_t(\theta_t^*) - Q_t(\gamma_t^*)) \implies \mathcal{B}(\tau)$ by Assumption 1-(b1) or 1-(b2), by the Continuous Mapping Theorem and by Assumption 1-(e) in the case of GMM estimators. The desired result follows from the fact that, under H_0 in (6), $\hat{\sigma}_t - \sigma_t = o_p(1)$ uniformly in $\tau \in [0, 1]$ (Andrews, 1991), where $\hat{\sigma}_t$ is one of the estimators in Corollary 2. Note that the behavior of the test statistic between R and T is the same as that of a standard Brownian Motion between 1 and $T - R$. The critical values are therefore obtained from results in Brown et al. (1975). The result for the Bayesian estimation method requires an intermediate step: from a similar result to that in Lemma 3 in Fernandez-Villaverde and Rubio-Ramirez (2004), we have that, under covariance stationarity and bounded priors:

$$\int_{\Theta} \sum_{j=1}^t \ln f(x_j; \theta) \pi(\theta) d\theta = K(\pi(\theta_t^*), \Sigma_T) + \sum_{j=1}^t \ln f(x_j; \theta_t^*) + o_p(1)$$

where $K(\pi(\theta_t^*), \Sigma_T) \equiv |\Sigma_T|^{-1/2} (2\pi)^{-k/p} \pi(\theta_t^*)$ is a constant that depends on the prior and on the estimate of minus the inverse of the second derivative of the loglikelihood, Σ_T . Therefore, $\int_{\Theta} t^{-1} \sum_{j=1}^t \ln f(x_j; \theta) \pi(\theta) d\theta = t^{-1} \sum_{j=1}^t \ln f(x_j; \theta_t^*) + o_p(1)$, and the result follows from the steps above.

For the rolling case, by using the same argument and by Assumption 1(g), we have $\sigma_t^{-1} \sqrt{R} (Q_t(\theta_t^*) - Q_t(\gamma_t^*)) \implies [\mathcal{B}(\tau) - \mathcal{B}(\tau - \rho)] / \sqrt{\rho}$ and critical values were obtained by Monte Carlo simulation (based on 8,000 Monte Carlo replications and by approximating the Brownian Motion with 400 observations).

■

Sketch of proof of Corollary 2. (a) Follows directly from Andrews (1991) and Newey and West (1987).

(b) Take a Mean Value expansion of $Q_t(\theta_t^*) \equiv -\frac{1}{2} \bar{g}_t(\theta_t^*)' W_{\theta,t} \bar{g}_t(\theta_t^*)$ around $E\bar{g}_t(\theta_t^*) : Q_t(\theta_t^*) = -\frac{1}{2} [E\bar{g}_t(\theta_t^*)]' W_{\theta,t} [E\bar{g}_t(\theta_t^*)] - [E\bar{g}_t(\theta_t^*)]' W_{\theta,t} [\bar{g}_t(\theta_t^*) - E\bar{g}_t(\theta_t^*)] - [\bar{g}_t(\theta_t^*) - \tilde{g}^*(\theta_t^*)]' W_{\theta,t} [\bar{g}_t(\theta_t^*) - \tilde{g}^*(\theta_t^*)]$, where $\tilde{g}^*(\theta_t^*)$ is an intermediate point between $\bar{g}(\theta_t^*)$ and $E\bar{g}_t(\theta_t^*)$; we have

$$Q_t(\theta_t^*) - Q_t(\gamma_t^*) = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} [E\bar{g}_t(\theta_t^*)]' W_{\theta,t} [E\bar{g}_t(\theta_t^*)] \\ \frac{1}{2} [E\bar{g}_t(\gamma_t^*)]' W_{\gamma,t} E\bar{g}_t(\gamma_t^*) \end{bmatrix}$$

$$+ \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} [E\bar{g}_t(\theta_t^*)]' W_{\theta,t} [\bar{g}_t(\theta_t^*) - E\bar{g}_t(\theta_t^*)] \\ [E\bar{g}_t(\gamma_t^*)]' W_{\gamma,t} [\bar{g}_t(\gamma_t^*) - E\bar{g}_t(\gamma_t^*)] \end{bmatrix} + o_p(\sqrt{t}),$$
 where the equality follows by Assumption 1(f1). The null hypothesis (6) guarantees that the first term on the right end side equals zero and, therefore, $\text{var}(\sqrt{t}(Q_t(\theta_t^*) - Q_t(\gamma_t^*))) = \begin{bmatrix} [E\bar{g}_t(\theta_t^*)]' W_{\theta,t} & -[E\bar{g}_t(\gamma_t^*)]' W_{\gamma,t} \end{bmatrix} \times \text{var}(\sqrt{t}(\bar{g}_t^* - E\bar{g}_t(\theta_t^*))) \begin{bmatrix} [E\bar{g}_t(\theta_t^*)]' W_{\theta,t} & -[E\bar{g}_t(\gamma_t^*)]' W_{\gamma,t} \end{bmatrix}$. By Assumption 1(b1), 1(e) and 1(f1), the first and third terms on the right end side can be consistently estimated by $\begin{bmatrix} \bar{g}_t(\hat{\theta}_t) \widehat{W}_{\theta,t} & -\bar{g}_t(\hat{\gamma}_t) \widehat{W}_{\gamma,t} \end{bmatrix}$, whereas the second term can be consistently estimated by using a HAC estimator for the vector of moment conditions, and the result follows.

(c) Under covariance stationarity, from McCracken (2000, expression 6, p. 203), we have that: $\text{var}(\sqrt{t}(Q_t(\theta^*) - Q_t(\gamma^*))) = \lambda \text{var}(\ln f(x_j; \theta^*) - \ln f(x_j; \gamma^*))$, where λ is the same as McCracken's λ_{hh} . The result follows directly. ■

Proposition 4 Sketch of proof for Proposition 3. *From the Proof of Proposition 1, it follows that $\sqrt{t}\hat{\sigma}_t^{-1}(Q_t(\hat{\theta}_t) - Q_t(\hat{\gamma}_t)) \Rightarrow B(\delta)/\sqrt{\delta}$, where $B(\delta)$ is a Brownian Motion defined on $[1, \infty)$. The critical value is then determined from the hitting probability of the Brownian Motion, as in Chu et al. (1996, p.1053):*

$$P\left\{|B(\delta)|/\sqrt{\delta} \geq \sqrt{(k_\alpha^2 + \ln \delta)}, \text{ for some } t \geq 1\right\} = 2[1 - \Phi(k_\alpha) = k_\alpha \phi(k_\alpha)],$$
 where $t = [\delta T]$, and $\phi(\cdot)$ and $\Phi(\cdot)$ are, respectively, the pdf and cdf of a standard normal distribution. ■

8 Tables and Figures

Table 1. Critical values for the Rolling Fluctuation test (k_{α}^{roll})

R/T	α	
	0.05	0.10
0.1	3.393	3.170
0.2	3.179	2.948
0.3	3.012	2.766
0.4	2.890	2.626
0.5	2.779	2.500
0.6	2.634	2.356
0.7	2.560	2.252
0.8	2.433	2.130
0.9	2.248	1.950

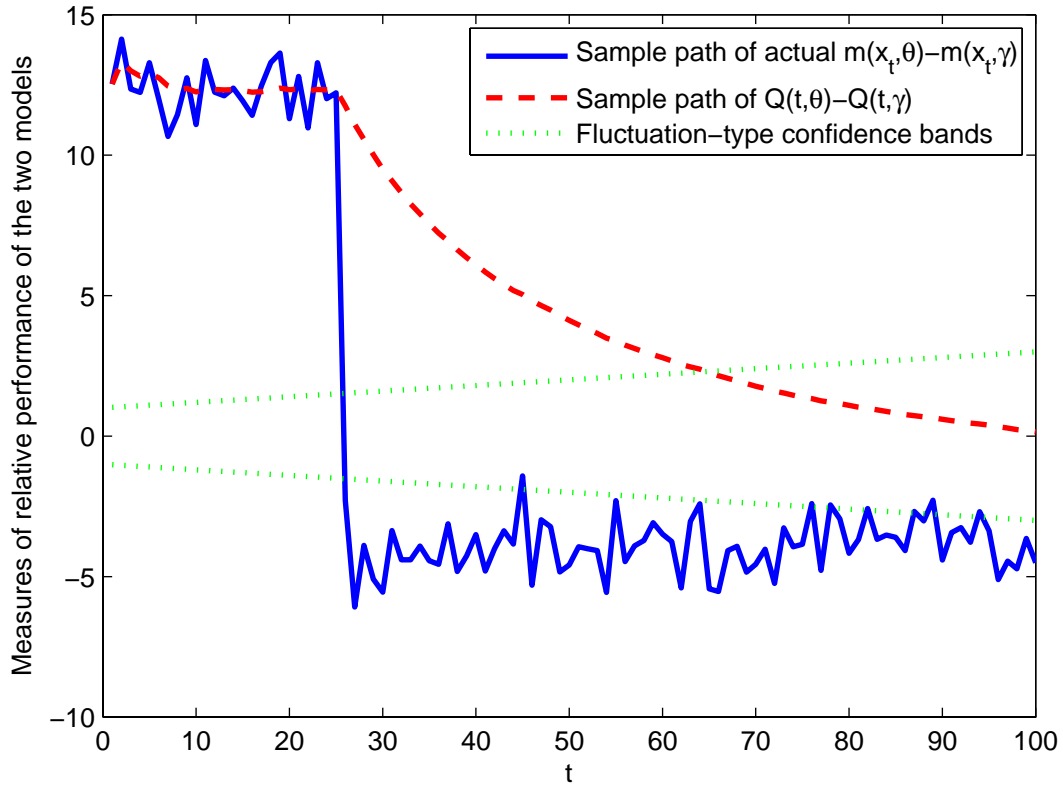
Note. Table 1 reports critical values for F_t^{roll} , the Fluctuation test implemented with a rolling scheme, eq. (8).

**Table 2. Empirical rejection rates of the tests proposed in this paper,
the standard model selection test, and the out of sample tests**

Designs:	Parameter values			Empirical rejection rates					
	β_0	β_{11}	γ_{11}	F_t^{rec}	F_t^{roll}	Full sample test	OOS _{fix}	OOS _{rec}	OOS _{roll}
(a)	1	0.5	0.5	0.051		0.047	0.089	0.044	0.034
		0.95	0.4	0.449		0.047	0.056	0.026	0.022
		0.7	0.2	0.518		0.589	0.914	0.864	0.802
	0.5	0.5	0.5	0.051		0.049	0.053	0.049	0.050
		0.95	0.4	0.041		0.058	0.053	0.050	0.052
		0.7	0.2	0.763		0.646	0.905	0.900	0.905
Real-time date				Empirical rejection rates					
		t/T	J_t	Full sample test					
(b)	1	1.5	0.010	0.121					
		1.75	0.020	0.152					
		2	0.032	0.179					

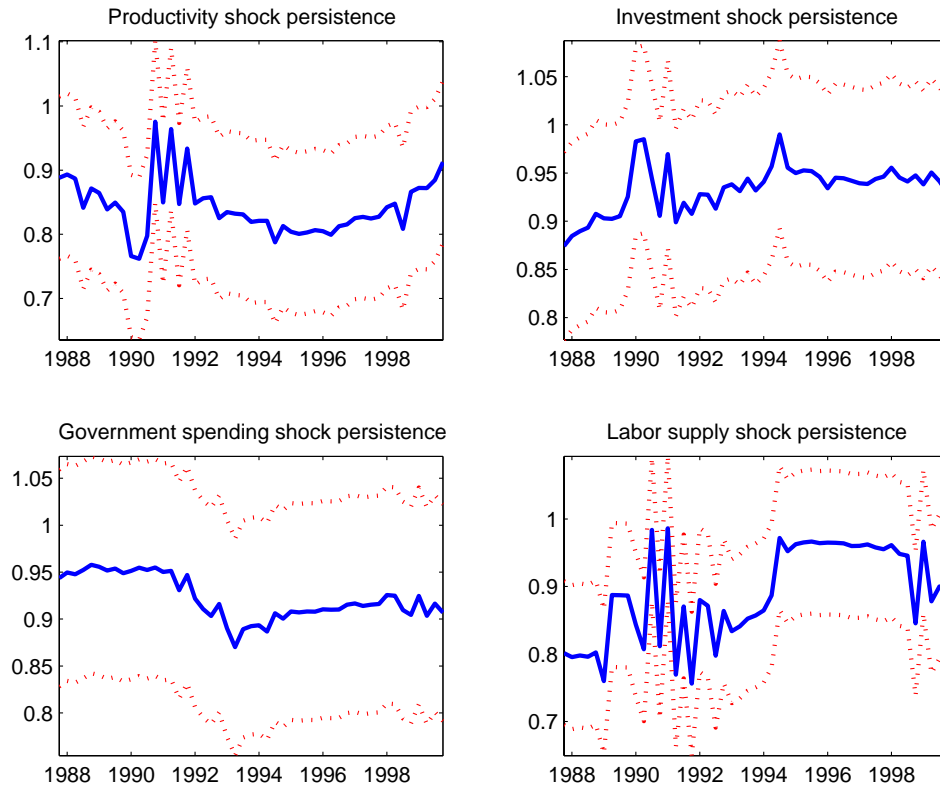
Note to Table 2. The table reports empirical rejection rates for the following test statistics: F_t^{rec} and F_t^{roll} respectively denote the test statistics described in (7) and (8); “Full sample test” denotes Vuong’s (1989) test; “OOS_{fix}”, “OOS_{rec}” and “OOS_{roll}” denote, respectively, out of sample forecasting tests implemented with a fixed, recursive and rolling scheme. The DGP is (17), and the Monte Carlo experiment is described in detail in Section 4.

Figure 1. Motivating example



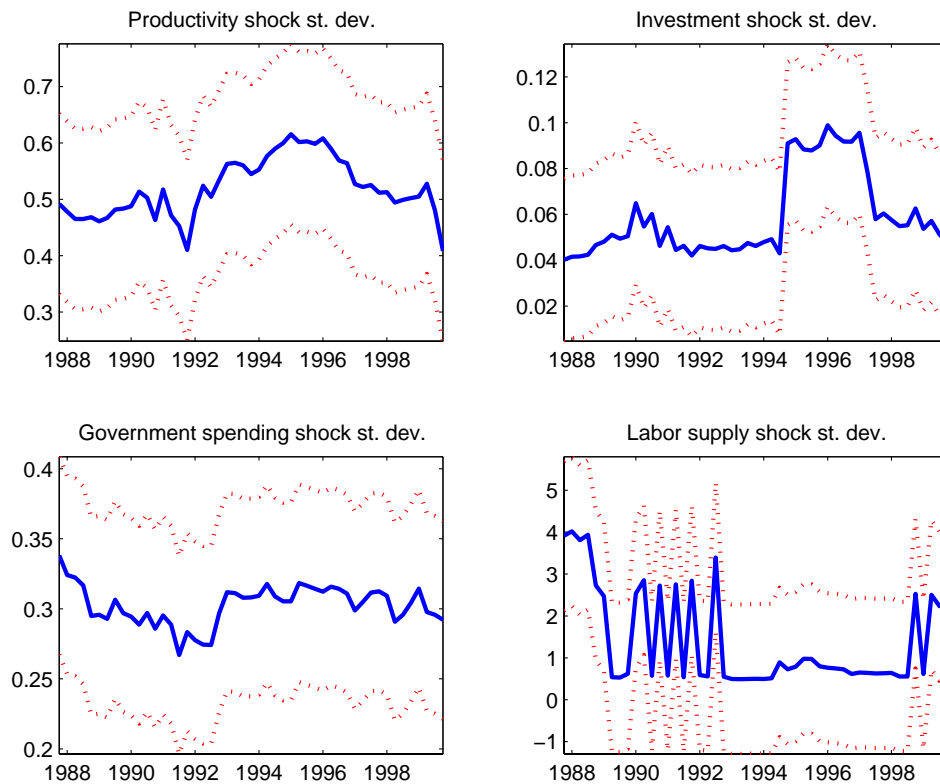
Notes to Figure 1. The figure shows the sample path of the relative performance of two models over time. The solid line represents the actual relative performance at each point in time, measured by the relative log-likelihoods at time t , $\ln f_t(\theta) - \ln g_t(\gamma)$, whereas the dashed line shows the average measure of the relative performance (i.e. the Likelihood Ratio) up to time t based on a recursive procedure.

Figure 2(a). Parameter Evolution over Time. Persistence of the shocks.



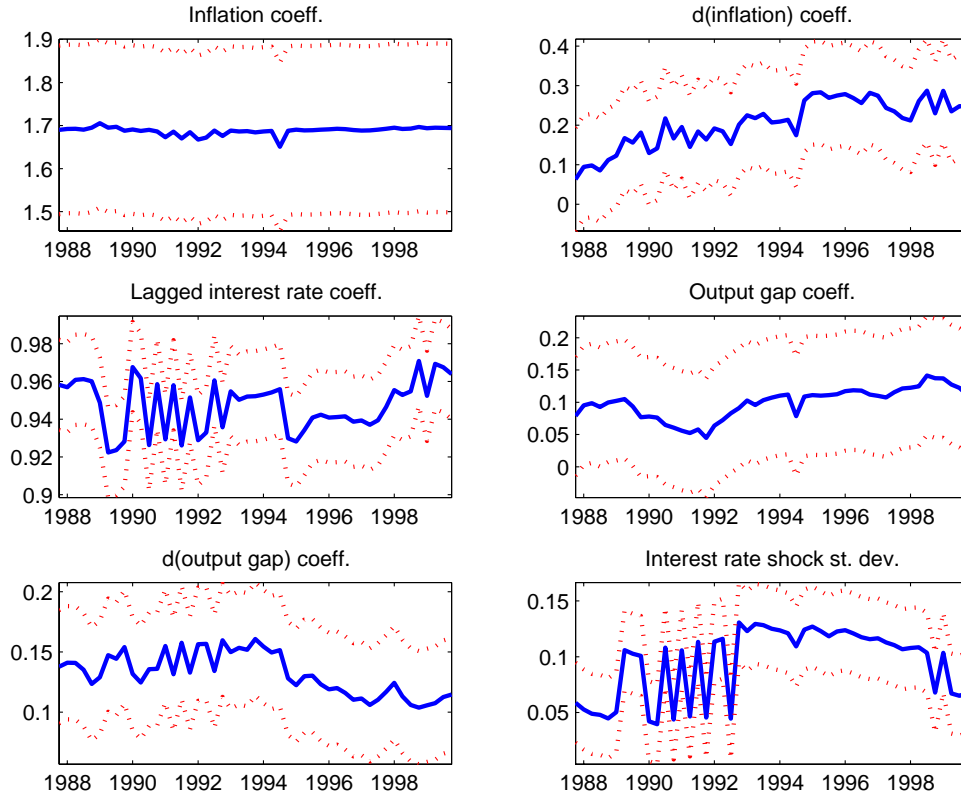
Notes to Figure 2(a). The figure plots rolling estimates of some parameters in Smets and Wouter's (2002) model. See Smets and Wouter's Table 1, p. 1142 for a description.

Figure 2(b). Parameter Evolution over Time. Standard deviation of the shocks.



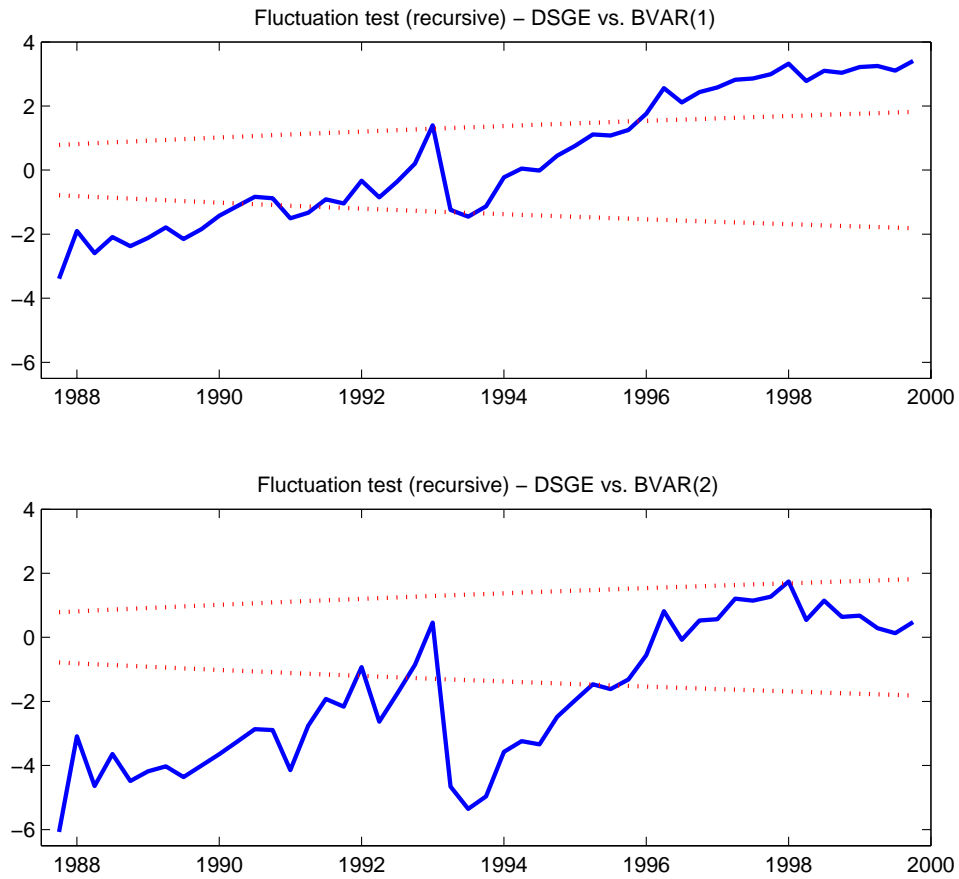
Notes to Figure 2(b). The figure plots rolling estimates of some parameters in Smets and Wouter's (2002) model. See Smets and Wouter's Table 1, p. 1142 for a description.

Figure 2(c). Parameter Evolution over Time. Standard deviation of the shocks.



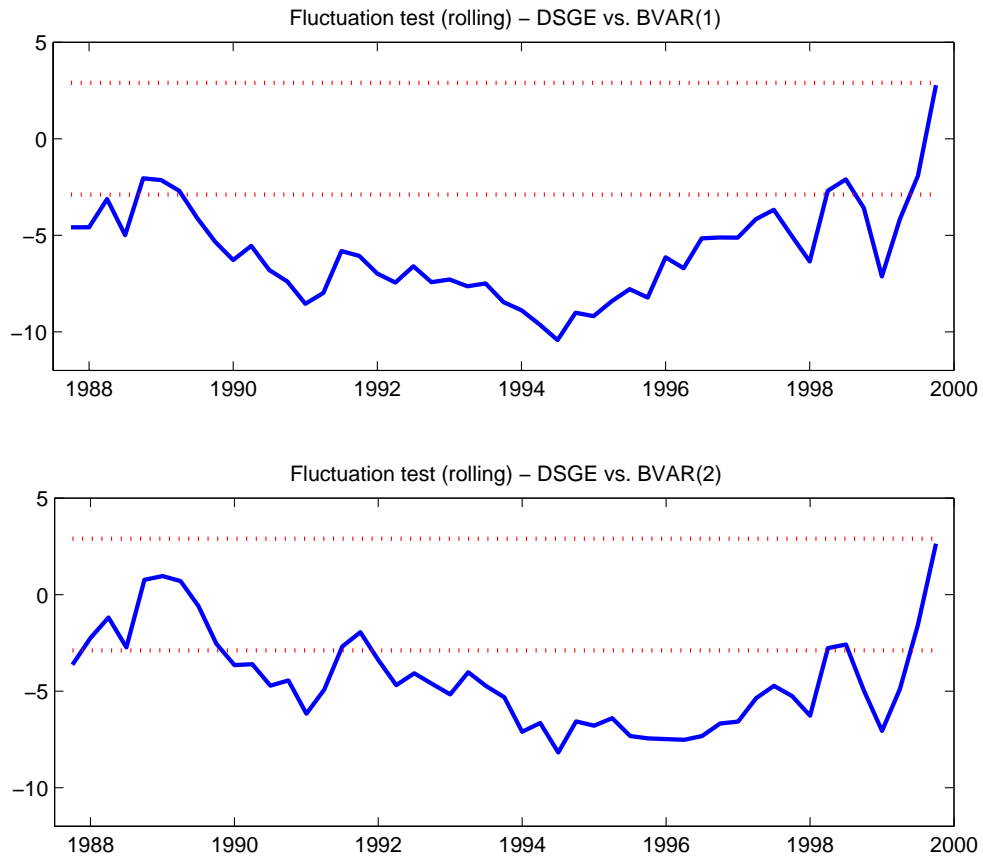
Notes to Figure 2(c). The figure plots rolling estimates of the parameters in the monetary policy reaction function described in Smets and Wouters' (2002) eq. (36), here reported for convenience: $\widehat{R}_t = \rho \widehat{R}_{t-1} + (1 - \rho) \left\{ \overline{\pi}_t + r_\pi (\widehat{\pi}_{t-1} - \overline{\pi}_t) + r_Y (\widehat{Y}_{t-1} - \widehat{Y}_t^p) \right\} + r_{\Delta\pi} (\widehat{\pi}_t - \widehat{\pi}_{t-1}) + r_{\Delta Y} \left((\widehat{Y}_t - \widehat{Y}_t^p) - (\widehat{Y}_{t-1} - \widehat{Y}_{t-1}^p) \right) + \eta_t^R$, $\overline{\pi}_t = \rho_\pi \overline{\pi}_{t-1} + \eta_t^\pi$. The figure plots: inflation coefficient (r_π), d(inflation) coefficient ($r_{\Delta\pi}$), lagged interest rate coefficient (ρ), output gap coefficient (r_Y), d(output gap) coefficient ($r_{\Delta Y}$), and standard deviation of the interest rate shock ($\sqrt{\text{var}(\eta_t^R)}$).

Figure 3. Recursive Fluctuation Tests



Notes to Figure 3. The figure plots the difference between the objective function (4) for the DSGE model and that of the bayesian VAR(1) (top panel) and bayesian VAR(2) (lower panel). All models are estimated by using a recursive scheme.

Figure 4. Rolling Fluctuation Tests



Notes to Figure 4. The figure plots the difference between the objective function (4) for the DSGE model and that of the bayesian VAR(1) (top panel) and bayesian VAR(2) (lower panel). All models are estimated by using a rolling scheme.