Model Selection and Forecast Comparison in Unstable Environments

Raffaella Giacomini and Barbara Rossi

UCL and Duke University

Abstract

We propose new methods for analyzing the relative performance of two competing, misspecified models in the presence of possible data instability. The main idea is to develop a measure of the relative "local performance" for the two models, and to investigate its stability over time by means of statistical tests. The models' performance can be evaluated using either in-sample or out-of-sample criteria. In the former case, we suggest using the local Kullback-Leibler information criterion, whereas in the latter, we consider the local out-of-sample forecast loss, for a general loss function. We propose two tests: a "fluctuation test" for analyzing the evolution of the model's relative performance over historical samples and a "sequential test", that monitors the models' relative performance in real time. Compared to previous approaches to model selection and forecast comparison, which are based on measures of "global performance" (e.g., Vuong (1989) and West (1996)), our focus on the entire time path of the models' relative performance may contain useful information that is lost when looking for a globally best model. Our methods can be applied to nonlinear, dynamic, multivariate models estimated by a variety of techniques. 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 VARs.

Keywords: Model Selection Tests, Misspecification, Structural Change, Forecast Evaluation, Kullback-Leibler Information Criterion

Acknowledgments: We are grateful to F. Smets and R. Wouters for providing their codes. We also thank B. Hansen, M. Jacoviello, U. Muller, M. del Negro, G. Primiceri, T. Zha, and seminar participants at the Empirical Macro Study Group at Duke University, Atlanta Fed, University of Michigan, NYU Stern, University of Montreal, UNC Chapel Hill, University of Wisconsin, UCI, LSE, UCL for useful comments and suggestions. Support by NSF grant 0647770 is gratefully acknowledged.

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

1 Introduction

This paper proposes new techniques for comparing the performance of competing models in the presence of model misspecification and structural instability. This is a realistic and relevant environment for applied macroeconomists, forecasters and policy makers for two reasons. First, policy makers and economic forecasters often face the problem of choosing the best performing model out a number of competing models, which can only be approximations of the truth. Second, the empirical importance of structural instabilities or "breaks" has been widely recognized for macroeconomic data. For example, Stock and Watson (2003) show that instabilities affect most macroeconomic time series; McConnell and Perez-Quiroz (2000) report evidence in favor of a break in the volatility of U.S. GDP and Fernald (2005) and Francis and Ramey (2005) investigate the implications of breaks in hours worked for the debate on the effects of technology shocks. As a consequence, prominent macroeconomists are now recognizing the importance of instabilities and incorporating them in their theoretical models. For example, Cogley and Sargent (2005) consider models with time-varying parameters, Clarida et al. (2000) introduce structural breaks in monetary policy; Justiniano and Primiceri (2007) and Fernandez-Villaverde and Rubio-Ramirez (2005, 2006) consider dynamic stochastic general equilibrium (DSGE) models with time-varying parameters.

The main insight of this paper is that, in unstable environments, it is plausible that the relative performance of competing models may itself change over time. This possibility is supported by recent empirical evidence reported in the forecasting literature (e.g., Stock and Watson, 2003), which shows that, even though some models outperform naive benchmarks in certain periods, this is not necessarily true when considering different periods.

As we discuss below, the existing techniques for model selection and forecast comparison appear inadequate in an environment characterized by instability and model misspecification, because they do not account for the possibility that the performance of the models may be changing. This paper fills the gap in the literature by proposing convenient techniques for analyzing the evolution over time in the performance of competing, misspecified models.

We propose two approaches, which address different evaluation objectives. The first can be used by empirical macroeconomists and forecasters interested in analyzing the evolution in the performance of two competing models over historical samples. The main idea is to develop a measure of the relative "local performance" of the models, and to test its stability over time by means of a "fluctuation test". The test is easily implemented by plotting the (appropriately normalized) sample path of the estimated measure of local performance, together with boundary lines which, if crossed, signal instability. The performance can be evaluated using either in-sample or out-of-sample criteria. In the former case, we introduce a measure that can be interpreted as a "local Kullback-Leibler information criterion (KLIC)", whereas in the latter, we consider what we

call the "local out-of-sample forecast loss", for a general, user-defined loss function. The fluctuation test, although convenient to obtain, does now however have optimality properties. We thus further provide a test for the null hypothesis of equal performance of the two models at each point in time that is optimal against the alternative hypothesis that there is a one-time break in the relative performance, and propose a method for estimating the timing of the break. We call this the "optimal test".

The second evaluation objective that we address is when a researcher is interested in monitoring the relative performance of two competing models in real time, in order to detect any deviation from the relative performance that was observed over the historical sample. To this end, we propose a "sequential test".

To better understand why existing econometric techniques are inadequate in conducting model selection and forecast evaluation in an environment characterized by instability and misspecification, it might be useful to divide the literature into two groups. The first group proposes techniques for model selection and forecast comparison that allow for misspecification, but the approach in this literature is to select the model with the best "global performance", which in practice amounts to selecting the model that performs best on average. The performance can be measured either in terms of in-sample fit (e.g., Vuong (1989); Rivers and Vuong (2002); see Fernandez-Villaverde and Rubio-Ramirez (2006) for an application to the selection between competing macro models), or out-of-sample forecast loss (e.g., Diebold and Mariano (1995); West (1996); McCracken (2000)). In the realistic presence of structural instability, however, the relative performance of the two models may itself be time-varying, and thus averaging this evolution over time may result in a loss of information. For example, a forecaster or policymaker may select the model that performed best on average over a particular historical sample, ignoring the fact that the competing model may be a more accurate description of the recent data or that it may produce more accurate forecasts when considering only the recent past. Such wrong choices would lead to poor forecasts and unsuccessful policymaking. The second strand of the literature is concerned with parameter instability tests. This literature focuses on one specific model, and tests for instability in its parameters under the assumption that the model is correctly specified (e.g., Andrews (1993), Bai and Perron (1998), Hansen (2000), Elliott and Muller (2005)), or for instability in its forecast performance, allowing for misspecification (Giacomini and Rossi (2005)). The example in Section 2 illustrates the relationship between parameter instability and instability in relative performance. The example, inspired by our empirical application, considers the comparison between a linearized Dynamic Stochastic General Equilibrium (DSGE) model and a VAR. The two competing models can be viewed as imposing different sets of misspecified restrictions on the parameters of an ARMA data-generating process (DGP), which are possibly time-varying. We show that the local relative KLIC in this case captures the relative degrees of misspecification of the two models at each point in time, by measuring how

far each misspecified restriction is from the true restriction. Since the true restriction is a function of the DGP parameters, whether the relative performance of the models changes or not depends on whether the parameters vary in a way that makes the true restriction also change. For instance, the parameters may vary but in a way that leaves the true restriction, and thus the relative performance of the models, unchanged. This suggests that a test for instability in the parameters in the DSGE and/or VAR would not necessarily shed light onto the stability in their relative performance. The possibility of a non-constant relative performance between two forecasting models is considered by Giacomini and White (2006), who argue that the relative forecast performance may differ in different states of the economy. They take however a different approach, which involves assessing whether one can relate the out-of-sample relative losses to observable economic variables. In the context of in-sample model selection tests, Rossi (2005) proposes tests to select between two models in the presence of possible parameter instability. She only focuses however on the case of nested and correctly specified models, whereas this paper considers a more general environment.

Our methods have many useful applications, and we show an example in our empirical analysis. Recent developments in empirical macroeconomics (Smets and Wouters, 2003, Del Negro and Schorfheide, 2004) have shown that it is possible to estimate DSGE models whose performance is comparable to that of VARs. However, the measures of relative performance used in these papers are average measures over historical samples, which may hide important changes in the relative performance of the models over time. We select one such representative DSGE model – 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 VARs.

The rest of the paper is organized as follows. The first section discusses a motivating example inspired by our empirical application, namely the comparison of a DSGE model's performance with that of a VAR. There we show interesting cases in which existing tests fail to recognize the time variation in the relative performance of the two models and therefore may induce the applied researcher to select the "wrong" model. The second section describes our methods in detail. In the third section we apply our techniques to analyze the performance of Smets and Wouters' (2003) DSGE model of the European economy relative to the performance of VARs. Interestingly, our techniques show evidence of time variation in the relative performance of the DSGE model versus the VAR over the last decades.

2 Motivating example

The following simple example illustrates the main issues associated with testing for model selection and forecast comparison in the presence of misspecification and structural instability, and motivates our approach.

Suppose that we observe a sample of size T (the "historical" sample) for a variable y_t with true conditional density h_t^{true}:

$$\begin{aligned} h_t^{true} &: N(\theta_t x_t + \gamma_t z_t, 1), \text{ where} \\ x_t &\sim N(0, \sigma_{x_t}^2);\ z_t \sim N(0, \sigma_{z_t}^2) \text{ independent,} \end{aligned} \tag{1}$$

and that two competing models assume the following misspecified conditional densities f_t and g_t for y_t:

$$f_t : N(\theta_t x_t, 1) \text{ and } g_t : N(\gamma_t z_t, 1). \tag{2}$$

2.1 In-sample fluctuation test

The goal of the in-sample fluctuation test is to analyze the relative in-sample performance of the two models over historical samples. In this case, the measure of relative performance for the two models at time t is the difference in the KLIC, which measures the relative distance of f_t and g_t from h_t^{true}:

$$\Delta KLIC = E\left[\log h_t^{true} - \log g_t\right] - E\left[\log h_t^{true} - \log f_t\right] = E\left[\log f_t - \log g_t\right],\ t = 1, ..., T \tag{3}$$

where the expectation is with respect to h^{true}. If $\Delta KLIC > 0$, we conclude that f_t performs better than g_t. Note that selecting the model that is closer to the data-generating process (DGP) is equivalent to selecting the model with the largest expected loglikelihood. It is easy to show that the $\Delta KLIC$ in our example is

$$\Delta KLIC = .5(\theta_t^2 \sigma_{x_t}^2 - \gamma_t^2 \sigma_{z_t}^2),\ t = 1..., T. \tag{4}$$

Intuitively, in this example the $\Delta KLIC$ captures the relative degrees of misspecification of the two models. To see why, note that the term $\gamma_t z_t$ is the component of the error for model f that is due to misspecification, and thus f performs better than g if the contribution of its misspecification term to the variance of the error is smaller than the corresponding quantity for model g. Concerning the possibility of time variation in the relative performance, which is our focus in this paper, note that (4) implies that the relative performance in this example can vary over time because the two models are affected by time variation in the DGP parameters and/or in the unconditional variance of the regressors in different ways (both of which result in time variation in the relative degree of

misspecification for the two models). It can also happen that the two terms in (4) are equal for each t, which shows that the models can have the same performance at each point in time, even though the underlying DGP is unstable.

To illustrate the type of time variation in the relative performance of two misspecified models that could arise in economic applications, the solid line in Figure 1a shows the sample path of the $\Delta KLIC$ (4) in two scenarios: in the first, the variance of the regressors is constant but one of the DGP parameters evolves as a random walk (left panel);[1] in the second, the DGP parameters are constant but the variance of one regressor has a break in the middle of the sample (right panel).[2]

FIGURE 1 HERE

One difficulty that arises when attempting to estimate the $\Delta KLIC$ at time t is that one needs to obtain consistent estimates of θ_t and γ_t, which are unknown. Our solution to this problem is to conduct inference about a "smoothed" version of the $\Delta KLIC$, obtained by computing moving averages of the measure of relative performance over windows of size m. Let $\sum_j = \sum_{j=t-m/2+1}^{t+m/2}$, where, without loss of generality, m is chosen to be an even number. We will define:

$$\text{Smoothed } \Delta KLIC : E\left[m^{-1}\sum_j \left(\log f_j(\theta^*_{t,m}) - \log g_j\left(\gamma^*_{t,m}\right)\right)\right],\ t = m/2+1, ..., T-m/2, \quad (5)$$

where $\theta^*_{t,m}$ and $\gamma^*_{t,m}$ are the pseudo-true parameters for the models estimated over the window of size m, e.g., $\theta^*_{t,m} = \max_\theta E\left[m^{-1}\sum_j \log f_j(\theta)\right]$. Unlike the $\Delta KLIC$, the smoothed $\Delta KLIC$ can be consistently estimated by substituting $\theta^*_{t,m}$ with the maximum likelihood estimates of the model's parameters computed over each moving window. Note that the smoothing implies that one loses the first and last $m/2$ data points, and is thus left with a series of length $n = T - m$.

In the example, we have $\theta^*_{t,m} = \sum_j \theta_j \sigma^2_{x_j} / \sum_j \sigma^2_{x_j}$ (and thus - when the variance of the regressor is constant - $\theta^*_{t,m}$ is the average of the true parameters over the moving window), and the smoothed $\Delta KLIC$ is:

$$\text{Smoothed } \Delta KLIC = .5\left(\theta^{*2}_{t,m} m^{-1}\sum_j \sigma^2_{x_j} - \gamma^{*2}_{t,m} m^{-1}\sum_j \sigma^2_{z_j}\right),\ t = m/2+1, ..., T-m/2. \quad (6)$$

If there is no time variation in the DGP parameters and in the unconditional variance of the regressors, the smoothed $\Delta KLIC$ (6) coincides with the $\Delta KLIC$ (4). The smoothed $\Delta KLIC$ is thus a better approximation of the $\Delta KLIC$ the smaller the variation within the moving window.

The dotted line in Figure 1a shows the time plot of the smoothed $\Delta KLIC$, obtained using a window of size 1/5 of the sample size. This time path, viewed as an approximation of the true

[1] Specifically, we let $\theta_t = \theta_{t-1} + \varepsilon_t$, $\theta_1 = 0$, $\varepsilon_t \sim i.i.d.N(0, .1)$; $\gamma = .5$; $\sigma^2_x = \sigma^2_z = 1$.

[2] Specifically, we let $\theta = \gamma = .5$; $\sigma^2_z = 1$; $\sigma^2_{x_t} = 1.25$ for $t < 100$; $\sigma^2_{x_t} = .75$ for $100 \leq t \leq 200$.

$\Delta KLIC$ that contains information about the relative performance of the models over time, is the object of interest of our analysis. This is in contrast to previous approaches (e.g., Vuong, 1989 and Rivers and Vuong, 2002), whose focus is on the average $\Delta KLIC$, computed over the full sample. Note that in the two scenarios considered in Figure 1a, the average full-sample $\Delta KLIC$ (marked by the dot) is close to zero, indicating that the two models perform equally well, whereas the smoothed $\Delta KLIC$ correctly reveals the presence of time variation in the models' relative performance.

Concerning the implementation of our test, the basic intuition is to consider the sample analog of the smoothed $\Delta KLIC$ (6), and normalize it to obtain the fluctuation statistic:

$$F_{t,m}^{IS} = \hat{\sigma}^{-1} m^{-1/2} \sum_{j} \left(\log f_j(\widehat{\theta}_{t,m}) - \log g_j\left(\widehat{\gamma}_{t,m}\right) \right), \; t = m/2+1, ..., T-m/2, \tag{7}$$

where $\hat{\sigma}^2$ is a suitable estimator of the asymptotic variance and $\widehat{\theta}_{t,m}$ and $\widehat{\gamma}_{t,m}$ are the maximum likelihood estimates of the models' parameters computed over the moving window. Under the assumption that a Functional Central Limit Theorem holds for the partial sums of $\log f_t - \log g_t$, one can characterize the behavior of the sample path of $F_{t,m}^{IS}$ under the null hypothesis that the smoothed $\Delta KLIC$ (5) equals zero at each point in time. In practice, we will provide boundary lines that are crossed by the limiting process with small probability under the null hypothesis, so that rejection occurs if the sample path of the fluctuation statistics crosses such boundaries. For illustration, Figure 1b plots the fluctuation statistics together with boundary lines for the data-generating processes considered in Figure 1a. We see that the sample path of the fluctuation statistics mimics that of the smoothed $\Delta KLIC$, revealing in both scenarios that the first model performs better in the first part of the sample and that the second model performs better in the second part of the sample.

We will show that the fact that the fluctuation statistics (7) depend on estimated parameters whereas the null hypothesis is expressed in terms of pseudo-true parameters does not affect the asymptotic distribution of the test statistics. As in Vuong (1989) and Rivers and Vuong (1989), however, the competing models must be non-nested, in order for the asymptotic distribution to be non-degenerate.

2.2 Out-of-sample fluctuation test

If the goal is to analyze the relative out-of-sample performance of the two models over historical samples, one would use the out-of-sample fluctuation test. This consists of first choosing a forecast horizon (h) and an in-sample size (R), and then estimating the models recursively, using only the in-sample observations, to derive a sequence of h-step ahead out-of-sample forecasts for times $t = R+h, ..., T$, for a total of $P \equiv T-(R+h)+1$ forecasts. The measure of relative performance in this case is the difference of the expected forecast losses computed over the out-of-sample portion.

The analysis is similar to that for the in-sample case, the main difference being that one must now take into account that the forecast losses depend on parameters estimated over a different sample. This issue is handled differently in the asymptotic framework of West (1996) (henceforth W) or that of Giacomini and White (2006) (henceforth GW). We present results for both. To fix ideas, assume a quadratic loss. The measure of relative performance in the two cases is:

$$(GW) \quad E\left[(y_t - \widehat{\gamma}_{t-h,R} z_t)^2 - (y_t - \widehat{\theta}_{t-h,R} x_t)^2\right], \; t = R+h, ..., T. \tag{8}$$

$$(W) \quad E\left[(y_t - \gamma^*_{t-h,R} z_t)^2 - (y_t - \theta^*_{t-h,R} x_t)^2\right], \; t = R+h, ..., T \tag{9}$$

When the expressions in (8) or (9) are positive, we conclude that model f performs better than model g. In the W framework, $\theta^*_{t-h,R}$ and $\gamma^*_{t-h,R}$ are the pseudo-true parameters for the in-sample estimation window (using either a fixed, rolling or expanding estimation window scheme, see definitions below). In the GW framework, the losses depend on the actual in-sample parameter estimates $\widehat{\theta}_{t-h,R}$, $\widehat{\gamma}_{t-h,R}$ (using only a fixed or rolling estimation window scheme). Similarly to the in-sample analysis, we estimate the time path of the models' relative performance by considering a sequence of statistics computed over moving windows of size m :

$$F^{OOS}_{t,m} = \hat{\sigma}^{-1} m^{-1/2} \sum_j \Delta L_j(\widehat{\theta}_{t-h,R}, \widehat{\gamma}_{t-h,R}), \; t = R+h, ..., T, \tag{10}$$

where now $\sum_j$ denotes $\sum_{j=t-m+1}^{t}$, $\Delta L_j(\widehat{\theta}_{t-h,R}, \widehat{\gamma}_{t-h,R}) = (y_j - \widehat{\theta}_{t-h,R} x_j)^2 - (y_j - \widehat{\gamma}_{t-h,R} z_j)^2$.

Note that for the out-of-sample we use a non-centered moving window that only filters past information, instead of the centered moving window considered in the in-sample analysis, which depends on both past and future information. The reason is that, in evaluating the forecasting ability of a model, the researcher is typically interested in assessing ex-ante forecast performance, which is a measure that is not contaminated by future information.

The difference between the W and the GW framework is in the expression for $\hat{\sigma}$, which in W contains terms that capture the effect of estimation uncertainty, whereas in GW it has a simpler form. Moreover, the W framework rules out comparisons between nested models, whereas the GW framework is applicable to both nested and non-nested models. Similarly to the in-sample case, our approach consists of characterizing the sample path of $F^{OOS}_{t,m}$ and deriving boundary lines under the null hypothesis that the measures of relative performance (8) or (9) are equal to zero at each point in time.

2.3 Sequential test

The goal of the sequential test is to provide a tool for monitoring the relative performance of the two models over the post-historical sample $T+1, T+2$ etc., to assess whether previous model selection decisions are reversed by the arrival of new information.

Suppose that the two models performed equally well, on average, in the historical sample. One would like to know whether this continues to be true as new data become available, for example by comparing the models' relative performance on a sample that includes the new observations. The problem with implementing a sequence of tests of equal performance with a fixed significance level is that it would result in size distortions for the overall procedure. The idea behind our approach is to conduct a sequence of full-sample tests, but utilizing modified critical values that control the overall size.

The procedure is implemented as follows. At every point in time $t = T+1, T+2, ...$ the researcher evaluates the measure of relative performance up to that time, that is the sample analog of the rescaled $\Delta KLIC$ at time t:

$$J_t = \hat{\sigma}_t^{-1} t^{-1/2} \sum_{j=1}^{t} \Delta L_j(\widehat{\theta}_t, \widehat{\gamma}_t). \tag{11}$$

where $\hat{\sigma}_t^2$ is given below in (21). The critical values for the J_t statistic at time t are $c_\alpha = \sqrt{r_\alpha^2 + \ln(t/T)}$, where r_a depends on the size of the test, α. Typical values of (α, r_α) are $(0.05, 2.7955)$ and $(0.10, 2.5003)$. The null hypothesis is rejected when $|J_t| > c_\alpha$. The sign of J_t identifies which models is best (for example, if $J_t > 0$ the first model is better).

3 Econometric methodology

3.1 Notation

We first introduce the notation and discuss the assumptions about the data, the models and the estimation procedures. We are interested in selecting a model for y_t, which we assume for simplicity to be a scalar (for the in-sample test, the extension to the multivariate case is straightforward), using a collection of variables z_t, possibly containing lags of y_t. We let $x_t = (y_t', z_t')'$.

For the in-sample analysis, we assume that two competing possibly nonlinear dynamic models for y_t specify different (misspecified) conditional densities f_t and g_t, which depend on parameters $\theta \in \Theta$ and $\gamma \in \Gamma$ that are estimated by Maximum Likelihood (ML). The implementation of the fluctuation test involves estimating the models recursively over moving windows of size $m < T$. Let $\sum_j = \sum_{j=t-m/2+1}^{t+m/2}$. At time t, the sample is $(x_{t-m/2+1}, ..., x_{t+m/2})$ and the parameter estimate for f (the definitions for g are analogous) is $\widehat{\theta}_{t,m} = \arg\max_{\theta\in\Theta} m^{-1} \sum_j \log f_j(x_j, \theta)$, with corresponding pseudo-true parameter $\theta^*_{t,m} = \arg\max_{\theta\in\Theta} m^{-1} \sum_j E\left[\log f_j(x_j, \theta)\right]$. For the in-sample fluctuation test, we thus have $\Delta L_j\left(\widehat{\theta}_{t,m}, \widehat{\gamma}_{t,m}\right) = \log f_j(\widehat{\theta}_{t,m}) - \log g_j\left(\widehat{\gamma}_{t,m}\right)$.

For the out-of-sample analysis, we assume that the researcher has divided the sample into an in-sample portion of size R and an out-of-sample portion of size P and obtained two competing sequences of $h-$step ahead out-of-sample forecasts by estimating the models using either a

fixed or rolling estimation window. For a general loss function L, we thus have sequences of P out-of-sample forecast loss differences, $\left\{L^f(y_t, \widehat{\theta}_{t-h,R}) - L^g(y_t, \widehat{\gamma}_{t-h,R})\right\}_{t=R+h}^{T}$, which depend on the realizations of the variable and on the in-sample parameter estimates for each model $\widehat{\theta}_{t-h,R}$ and $\widehat{\gamma}_{t-h,R}$. Unlike for the in-sample case, for which we restrict attention to maximum likelihood estimation, for the out-of-sample fluctuation test any estimation procedure is allowed. The parameters are estimated recursively, over a sample including data indexed $1, ..., R$ (fixed scheme) or $t-h-m+1, ..., t-h$ (rolling scheme). For the in-sample fluctuation test, we thus have $\Delta L_j\left(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R}\right) = L^f(y_j, \widehat{\theta}_{j-h,R}) - L^g(y_j, \widehat{\gamma}_{j-h,R})$.

3.2 The fluctuation test

3.2.1 In-sample analysis

We make the following assumptions for the in-sample fluctuation test.

Assumption IS: Let τ be s.t. $t = [\tau T]$ and $\tau \in [0,1]$. (a) $\left\{T^{-1/2}\sum_{j=1}^{[\tau T]} \Delta L_j(\theta, \gamma)\right\}$ obeys a Functional Central Limit Theorem (FCLT) for all $\theta \in \Theta$, $\gamma \in \Gamma$; (b) $\widehat{\theta}_{t,m}$ satisfies a Strong Uniform Law of Large Numbers: $\widehat{\theta}_{t,m} \underset{as}{\to} \theta^*_{t,m}$ uniformly over Θ (and similarly for $\widehat{\gamma}_{t,m}$); (c) $\nabla f_j(\theta), \nabla g_j(\gamma)$ satisfy a Uniform Law of Large Numbers; (d) $\sigma^2 = lim_{m\to\infty} E(m^{-1/2}\sum_{j=t-m/2+1}^{t+m/2} \Delta L_j(\theta^*_{t,m}, \gamma^*_{t,m}))^2 > 0$ (e) $m/T \to \mu \in (0,\infty)$ as $m \to \infty, T \to \infty$; (f) Θ, Γ are compact.

Assumption (d) imposes global covariance stationarity for the sequence of loss differences, and it thus limits the amount of heterogeneity permitted under the null hypothesis. This assumption is in principle stronger than necessary, but it facilitates the statement of the FCLT (see Wooldridge and White, 1988 for a general FCLT for heterogeneous mixing sequences). Note that global covariance stationarity allows the variance to change over time, but in a way that ensures that, as the sample size grows, the sequence of variances converges to a finite and positive limit.

The following Proposition provides a justification for the in-sample fluctuation test.

Proposition 1 (In-sample fluctuation test) *Suppose Assumption IS holds. Let $F^{IS}_{t,m} = \hat{\sigma}^{-1} m^{-1/2} \sum_{j=t-m/2+1}^{t+m/2} \left(\log f_j(\widehat{\theta}_{t,m}) - \log g_j\left(\widehat{\gamma}_{t,m}\right)\right)$, $t = m/2+1, ..., T-m/2$, where $\hat{\sigma}^2$ is a HAC estimator of the global asymptotic variance σ^2, for example*

$$\hat{\sigma}^2 = \sum_{i=-q(m)+1}^{q(m)-1} (1 - |i/q(m)|) m^{-1} \sum_{j=t-m/2+1}^{t+m/2} \Delta L_j\left(\widehat{\theta}_{t,m}, \widehat{\gamma}_{t,m}\right) \Delta L_{j-i}\left(\widehat{\theta}_{t,m}, \widehat{\gamma}_{t,m}\right), \qquad (12)$$

with $q(m)$ a bandwidth that grows with m (e.g., Newey and West, 1987). Under the null hypothesis $H_0 : E\left[m^{-1}\sum_j \Delta L_j(\theta^*_{t,m}, \gamma^*_{t,m})\right] = 0$ for all $t = m/2+1, ..., T-m/2$,

$$F^{IS}_{t,m} \Longrightarrow \left[\mathcal{B}(\tau + \mu/2) - \mathcal{B}(\tau - \mu/2)\right] / \sqrt{\mu}, \qquad (13)$$

where $t = [\tau T]$, $m = [\mu T]$ *and* $B(\cdot)$ *is a standard univariate Brownian motion. The boundary lines for a significance level* α *are* $\pm\, k_\alpha$*where* k_α *solves*

$$P\left\{\sup_{\tau} |[\mathcal{B}(\tau + \mu/2) - \mathcal{B}(\tau - \mu/2)]/\sqrt{\mu}| > k_\alpha\right\} = \alpha. \tag{14}$$

Simulated values of (α, k_α) *for various choices of* μ *are reported in Table 1. The null hypothesis is rejected when* $\max_{m/2+1\leq t\leq T-m/2} |F_{t,m}^{IS}| > k_\alpha$.

3.2.2 Out-of-sample analysis

We make the following assumptions for the out-of-sample fluctuation test.

Assumption OOS: Let τ *be s.t.* $t = [\tau P]$ and $\tau \in [0,1]$. *(a)* $\left\{P^{-1/2}\sum_{j=R+h}^{[\tau P]} \Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R})\right\}$ *obeys a FCLT; (b)* $\sigma^2 = lim_{m\to\infty} E(m^{-1/2}\sum_{j=t-m/2+1}^{t+m/2} \Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R}))^2 > 0$ *(c)* $m/P \to \mu \in (0,\infty)$ *as* $m \to \infty, P \to \infty$.

Note that, unlike the in-sample test, which requires the parameters of the two models to be estimated by ML, the out-of-sample test does not impose restrictions on the estimation method used to produce the forecasts for the two models. This is because we use the same asymptotic framework as in Giacomini and White (2006). Giacomini and White (2006) also provide primitive conditions for assumption OOS(a), which allow the data to be mixing and heterogeneous and essentially require the use of a "rolling" or "fixed" estimation window scheme in producing the out-of-sample forecasts.

The procedure for deriving the out-of-sample fluctuation test is analogous to that for the in-sample case. The only difference is that the time variation of the relative forecast performance is only analyzed over the out-of-sample portion of size P, rather than over the full sample of size T. Proposition 1 is thus modified as follows.

Proposition 2 (Out-of-sample fluctuation test) *Suppose Assumption OOS holds. Let* $F_{t,m}^{OOS} = \hat{\sigma}^{-1} m^{-1/2} \sum_{j=t-m/2+1}^{t+m/2} \Delta L_j(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R})$, $t = R + h + m/2, ..., T - m/2$, *where* $\hat{\sigma}^2$ *is a HAC estimator of* σ^2, *for example*

$$\hat{\sigma}^2 = \sum_{i=-q(m)+1}^{q(m)-1} (1 - |i/q(m)|) m^{-1} \sum_{j=t-m/2+1}^{t+m/2} \Delta L_j\left(\widehat{\theta}_{j-h,R}, \widehat{\gamma}_{j-h,R}\right) \Delta L_{j-i}\left(\widehat{\theta}_{j-i-h,R}, \widehat{\gamma}_{j-i-h,R}\right), \tag{15}$$

$q(m)$ *is a bandwidth that grows with* m *(Newey and West, 1987). Under the null hypothesis* $H_0 : E\left[\Delta L_t\left(\widehat{\theta}_{t-h,R}, \widehat{\gamma}_{t-h,R}\right)\right] = 0$ *for all* $t = R + h, ..., T$,

$$F_{t,m}^{OOS} \Longrightarrow [\mathcal{B}(\tau + \mu/2) - \mathcal{B}(\tau - \mu/2)]/\sqrt{\mu}, \tag{16}$$

where $t = [\tau P]$, $m = [\mu P]$ *and* $\mathcal{B}(\cdot)$ *is a standard univariate Brownian motion. The boundary lines for a significance level* α *are* $\pm\, k_\alpha$ *where* k_α *solves*

$$P\left\{\sup_{\tau} |[\mathcal{B}(\tau+\mu/2) - \mathcal{B}(\tau-\mu/2)]/\sqrt{\mu}| > k_\alpha\right\} = \alpha. \tag{17}$$

Simulated values of (α, k_α) *for various choices of* μ *are reported in Table 1.*

3.3 The optimal test

The assumptions that guarantee validity of the optimal test are the same as those for the in-sample fluctuation test.[3] The following proposition gives the justification for the optimal test.

Proposition 3 (Optimal test against a one-time break) *Suppose Assumption IS holds. Let* $QLR_T = \sup_t \Phi_T(t)$, $t \in \{[0.15T], ... [0.85T]\}$, $\Phi_T(t) = LM_1 + LM_2(t)$, *where*

$$\begin{aligned} LM_1 &= \hat{\sigma}^{-2} T^{-1} \left[\sum_{j=1}^{T} \left(\log f_j(\widehat{\theta}_T) - \log g_j(\widehat{\gamma}_T)\right)\right]^2 \\ LM_2(t) &= \hat{\sigma}^{-2} T^{-1} (t/T)^{-1} (1-t/T)^{-1} [\sum_{j=1}^{t} \left(\log f_j(\widehat{\theta}_{1,t}) - \log g_j(\widehat{\gamma}_{1,t})\right) \\ &\quad - (t/T) \sum_{j=1}^{T} \left(\log f_j(\widehat{\theta}_T) - \log g_j(\widehat{\gamma}_T)\right)]^2, \end{aligned}$$

$\widehat{\sigma}^2$ *a HAC estimators of the asymptotic variance* $\sigma^2 = var\left(T^{-1}\sum_{j=1}^{T}(\log f_j(\theta_T^*) - \log g_j(\gamma_T^*))\right)$, *for example*

$$\hat{\sigma}^2 = \sum_{i=-q(T)+1}^{q(T)-1} (1-|i/q(T)|) T^{-1} \sum_{j=1}^{T} \left(\log f_j(\widehat{\theta}_T) - \log g_j(\widehat{\gamma}_T)\right)\left(\log f_{j-i}(\widehat{\theta}_T) - \log g_{j-i}(\widehat{\gamma}_T)\right). \tag{18}$$

Consider the null hypothesis

$$H_0 : E\left[t^{-1/2}\sum_{j=1}^{t}\left(\log f_j(\theta_{1,t}^*) - \log g_j\left(\gamma_{1,t}^*\right)\right) - (T-t)^{-1/2}\sum_{j=t+1}^{T}\left(\log f_j(\theta_{2,t}^*) - \log g_j\left(\gamma_{2,t}^*\right)\right)\right] = 0,$$

for every $t = 1, 2, ..., T$, *where* $\theta_{1,t}^*$ *is the pseudo-true parameter for the sample indexed* $1, ..., t$ *and* $\theta_{2,t}^*$ *is the pseudo-true parameter for the sample indexed* $t+1, ..., T$ *(similar definitions hold for* $\gamma_{1,t}^*$ *and* $\gamma_{2,t}^*$*). We have* $QLR_T \Longrightarrow \sup_\tau \left[\frac{\mathcal{BB}(\tau)'\mathcal{BB}(\tau)}{\tau(1-\tau)} + \mathcal{B}(1)'\mathcal{B}(1)\right]$, *where* $t = [\tau T]$, *and* $\mathcal{B}(\cdot)$ *and* $\mathcal{BB}(\cdot)$ *are, respectively, a standard univariate Brownian motion and a Brownian bridge. The null hypothesis is thus rejected when* $QLR_T > k_\alpha$. *The critical values* (α, k_α) *are:* $(0.05, 9.8257)$, $(0.10, 8.1379)$, $(0.01, 13.4811)$.

[3] **We let** $t = [\tau T]$ **in this section, so Assumption IS(e) should read:** $t/T \to \tau \in (0, \infty)$ **as** $t \to \infty, T \to \infty$**.It is intended that Assumptions IS(a,b,c) hold for both the full sample and the partial sample sums and estimators.**

Among the advantages of this approach, we have that: (i) when the null hypothesis is rejected, it is possible to evaluate whether the rejection is due to instabilities in the relative performance or to a model being constantly better than its competitor; (ii) if such instability is found, it is possible to estimate the time of the switch in the relative performance; (iii) the test is optimal against one time breaks in the relative performance. This is achieved by using the following procedure for a test with overall significance level α:

(i) test the hypothesis of equal performance at each time by using the statistic QLR_T^* from Proposition (3) at α significance level;

(ii) if the null is rejected, compare LM_1 and $\sup_t LM_2(t)$, $t \in \{[0.15T], ...[0.85T]\}$, with the following critical values: $(3.84, 8.85)$ for $\alpha = 0.05$, $(2.71, 7.17)$ for $\alpha = 0.10$, and $(6.63, 12.35)$ for $a = 0.01$. If only LM_1 rejects then there is evidence in favor of the hypothesis that one model is constantly better than its competitor. If only LM_2 rejects, then there is evidence that there are instabilities in the relative performance of the two models but neither is constantly better over the full sample. If both reject then it is not possible to attribute the rejection to a unique source.[4]

(iii) estimate the time of the break by $t^* = \arg\max_{t \in \{0.15T,...,0.85T\}} LM_2(t)$.

(iv) to extract information on which model to choose, we suggest to plot the time path of the underlying relative performance as:

$$\begin{cases} \frac{1}{t^*} \sum_{j=1}^{t^*} \left(\log f_j(\widehat{\theta}_{1,t^*}) - \log g_j\left(\widehat{\gamma}_{1,t^*}\right) \right) & \text{for } t \leq t^* \\ \frac{1}{(T-t^*)} \sum_{j=t^*+1}^{T} \left(\log f_j(\widehat{\theta}_{2,t^*}) - \log g_j\left(\widehat{\gamma}_{2,t^*}\right) \right) & \text{for } t > t^* \end{cases}$$

This approach can be easily generalized to multiple changes in relative performance by following, for example, the sequential procedure suggested by Bai and Perron (1998).

The fluctuation and the optimal tests have trade-offs. If the researcher is willing to specify the alternative of interest (in this case, a one-time break in the relative performance), then the latter test can be implemented and it will have optimality properties. Furthermore, it allows the researcher to estimate the time of the break. The fluctuation test, on the other hand, does not require the researcher to specify an alternative, and therefore might be preferable for researchers who do not have one.

3.4 The sequential test

Suppose that the two models were equally good in the historical sample of data up to time T, based on the fact that they yielded statistically indistinguishable in-sample performance, i.e., that

[4] This procedure is justified by the fact that the two components LM_1 and LM_2 are asymptotically independent – see Rossi (2005). Performing two separate tests does not result in an optimal test, but it is nevertheless useful to heuristically disentangle the causes of rejection of equal performance. The critical values for LM_1 are from a χ_1^2 whereas those for LM_2 are from Andrews (1993).

$E\left[T^{-1}\sum_{j=1}^{T}\Delta L_j(\theta_T^*,\gamma_T^*)\right]=0$. We test the null hypothesis that the two models perform equally well for all subsequent periods in the post-historical sample:

$$H_0: E\left[t^{-1}\sum_{j=1}^{t}\Delta L_j(\theta_t^*,\gamma_t^*)\right]=0 \text{ for } t=T+1,T+2,..., \tag{19}$$

against the alternative $H_1: E\left[t^{-1}\sum_{j=1}^{t}\Delta L_j(\theta_t^*,\gamma_t^*)\right]\neq 0$ at some point $t\geq T$.

We make the following assumptions:

Assumption SEQ: Let τ *be s.t.* $t=[\tau T]$ and $\tau\in[1,n]$; *(a) for every integer* $n>1$, $\left\{T^{-1/2}\sum_{j=1}^{[\tau T]}\Delta L_j(\theta,\gamma)\right\}$ *obeys a FCLT for all* $\theta\in\Theta$, $\gamma\in\Gamma$; *(b)* $\widehat{\theta}_t$ *is consistent for* θ_t^* *uniformly over* Θ *and in* τ; *(c) for every integer* $n>1$, $t^{-1}\sum_{j=1}^{t}\Delta L_j(\theta^*,\gamma^*)=E[t^{-1}\sum_{j=1}^{t}\Delta L_j(\theta^*,\gamma^*)]+o_p(1)$, $t^{-1}\sum_{j=1}^{t}\nabla f_j(\theta)$ *and* $t^{-1}\sum_{j=1}^{t}\nabla g_j(\theta)$ *satisfy a Uniform Law of Large Numbers – uniform in the parameter space and in* τ; *(d)* $\sigma^2=lim_{t\to\infty}E(t^{-1/2}\sum_{j=1}^{t}\Delta L_j(\theta_t^*,\gamma_t^*))^2>0$; *(e)* Θ,Γ *are compact.*

Assumption (b) requires consistency of the parameter estimates for the two models (see Inoue and Rossi (2005) for more primitive conditions that ensure this); (c) ensures uniform convergence for $\tau\in[1,n]$.

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, which are given in the following proposition.

Proposition 4 (Sequential test) *The test statistic for testing the null hypothesis* $E\left[T^{-1}\sum_{j=1}^{T}\Delta L_j(\theta_T^*,\gamma_T^*)\right]=0$ *against the alternative* $H_1: E\left[t^{-1}\sum_{j=1}^{t}\Delta L_j(\theta_t^*,\gamma_t^*)\right]\neq 0$ *at some* $t\geq T$ *is:*

$$J_t=\hat{\sigma}^{-1}t^{-1/2}\sum_{j=1}^{t}\Delta L_j(\widehat{\theta}_t,\widehat{\gamma}_t),\ t=T+1,T+2,..., \tag{20}$$

where $\hat{\sigma}^2$ *is a HAC estimator of* σ, *e.g.,*

$$\hat{\sigma}_t^2=\sum_{i=-q(t)+1}^{q(t)-1}(1-|i/q(t)|)t^{-1}\sum_{j=1}^{t}\Delta L_j\left(\widehat{\theta}_t,\widehat{\gamma}_t\right)\Delta L_{j-i}\left(\widehat{\theta}_t,\widehat{\gamma}_t\right), \tag{21}$$

with $q(m)$ *a bandwidth that grows with* m *(cf. Newey and West, 1987). The critical value at time* t *for a level* α *test is* $c_\alpha=\sqrt{r_\alpha^2+\ln(t/T)}$, *where the exact expression for* r_α *is given in the proof. Typical values of* (α,r_α) *are* $(0.05,2.7955)$ *and* $(0.10,2.5003)$. *The null hypothesis is rejected when* $|J_t|>c_\alpha$.

4 Empirical application: time-variation in the performance of DSGE vs. BVAR 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 - fits 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 new generation of DSGE models 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, and 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 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 apply the techniques proposed in this paper to assess whether the relative performance of the DSGE model and of BVARs was stable over time. We extend the sample considered by SW to include data up to 2004:4, for a total sample of size $T = 145$.

In order to compute the local measure of relative performance, (the local $\Delta KLIC$), we estimate both models recursively over a moving window of size $m = 70$ using Bayesian methods. 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 a BVAR(1) and a BVAR(2), both of which use a variant of the Minnesota prior, as suggested by Sims (2003).[5] We present results for two different transformations of the data. The first applies the same detrending of the data used by SW, which is based on a linear trend fitted on the whole sample (we refer to this as "full-sample detrending"). As cautioned by Sims (2003), this type of pre-processing of the data may unduly favour the DSGE, and thus we further consider a second transformation of the data, where detrending is performed on each rolling estimation window ("rolling-sample detrending").

Figure 2 displays the evolution of the posterior mode of some representative parameters. Figure 2a shows parameters that describe the evolution of the persistence of some representative shocks (productivity, investment, government spending, and labor supply); Figure 2b shows the estimates

[5] The BVAR's were estimated using 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.

of the standard deviation of the same shocks; and Figure 2c plots monetary policy parameters. Overall, Figure 2 reveals evidence of parameter variation. In particular, the figures show some decrease in the persistence of the productivity shock, whereas both the persistence and the standard deviation of the investment shock seem to increase over time. The monetary policy parameters appear to be overall stable over time.

FIGURE 2 HERE

We then apply our in-sample fluctuation test to test the hypothesis that the DSGE model and the BVAR have equal performance at every point in time over the historical sample.

Figure 3 shows the implementation of the fluctuation test for the DSGE vs. a BVAR(1) and BVAR(2), using full-sample detrending of the data. The estimate of the local relative $KLIC$ is evaluated at the posterior modes $\widehat{\theta}_{t,m}$ and $\widehat{\gamma}_{t,m}$ of the models' parameters, using the fact that $\widehat{\theta}_{t,m}$ and $\widehat{\gamma}_{t,m}$ are consistent estimates of the pseudo-true parameters $\theta^*_{t,m}$ and $\gamma^*_{t,m}$ (see, e.g., Fernandez-Villaverde and Rubio-Ramirez, 2004).

FIGURE 3 HERE

Figure 3 suggests that the DSGE has comparable performance to both a BVAR(1) and BVAR(2) up until the early 1990s, at which point the performance of the DSGE dramatically improves relative to that of the reduced-form models.

To assess whether this result is sensitive to the data filtering, we implement the fluctuation test for the DSGE vs. a BVAR(1) and BVAR(2), this time using rolling-window detrended data.

FIGURE 4 HERE

The results confirm the suspicion expressed by Sims (2003) that the pre-processing of the data utilized by SW penalizes the reduced-form models in favour of the DSGE. As we see from Figure 4, once the detrending is performed on each rolling window, the advantage of the DSGE at the end of the sample disappears, and the DSGE performs as well as a BVAR(1) on most of the sample, whereas it is outperformed by a BVAR(2) for all but the last few dates in the sample (when the two models perform equally well).

5 Conclusions

This paper provides new tests for model selection and forecast comparison in the presence of possible misspecification and structural instability. We proposed methods for assessing whether there is time variation in the relative performance of possibly nonlinear dynamic models, where the relative performance could be assessed either in-sample or out-of-sample.

For the in-sample case, our techniques are only applicable if the models are non-nested. If the models of interest are instead nested and misspecification is not a concern, the researcher has the following options. A possible counterpart for the in-sample fluctuation test 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 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 larger model, and they both test jointly this hypothesis as well as the maintained assumption that the small model is correctly specified.

References

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

[2] Andrews, D.W.K. (1993), "Tests for Parameter Instability and Structural Change with Unknown Change Point", *Econometrica* 61, 821–856.

[3] Bai, J. and P. Perron (1998), "Estimating and Testing Linear Models with Multiple Structural Changes", *Econometrica*, 66, 47-78.

[4] 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.

[5] Cavaliere, G. and R. Taylor (2005), "Stationarity Tests Under Time-Varying Second Moments", *Econometric Theory* 21, 1112-1129.

[6] Chu, C. J., M. Stinchcombe and H. White (1996), "Monitoring Structural Change", *Econometrica*, 64, 1045-1065.

[7] Clarida, R., J. Gali, and M. Gertler (2000), "Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory", *The Quarterly Journal of Economics* 115(1), 147-180.

[8] Cogley, T., and T.J. Sargent (2005), "Drifts and Volatilities: Monetary Policies and Outcomes in the Post WWII U.S.", *Review of Economic Dynamics* 8(2), 262-302.

[9] Del Negro, M., F. Schorfheide, F. Smets and R. Wouters (2004), "On the Fit and Forecasting Performance of New Keynesian Models", *mimeo.*

[10] Diebold, F. X., R. S. Mariano (1995), "Comparing Predictive Accuracy", *Journal of Business and Economic Statistics,* 13, 253-263.

[11] Elliott, G. and U. Muller (2003), "Optimally Testing General Breaking Processes in Linear Time Series Models", *The Review of Economic Studies,* forthcoming.

[12] Fernald, J. (2005), "Trend Breaks, Long-Run Restrictions, and the Contractionary Effects of Technology Improvements", *mimeo.*

[13] Fernandez-Villaverde, J., and J.F. Rubio Ramirez (2004), "Comparing Dynamic Equilibrium Models to Data: a Bayesian Approach", *Journal of Econometrics* 123, 153-187.

[14] Fernandez-Villaverde, J., and J. Rubio-Ramirez (2006), "Estimating Macroeconomic Models: A Likelihood Approach", *Review of Economic Studies*, forthcoming.

[15] Fernandez-Villaverde, J. and J. Rubio-Ramirez (2007), "How Structural Are Structural Parameters?", in: D. Acemoglu, K. Rogoff and M. Woodford (eds.), *NBER Macroeconomics Annual*, MIT Press.

[16] Francis, N., and V. Ramey (2005), "A New Measure of Hours Per Capita with Implications for the Technology-Hours Debate", *mimeo*.

[17] Giacomini, R. and H. White (2006), "Tests of Conditional Predictive Ability", *Econometrica,* 74, 1545-1578.

[18] Hansen, B.E. (2000), "Testing for Structural Change in Conditional Models", *Journal of Econometrics,* 97, 93-115.

[19] Inoue, A. and B. Rossi (2005), "Recursive Predictability Tests for Real-Time Data", *Journal of Business and Economic Statistics,* 23, 336-345

[20] Justiniano, A., and G. Primiceri (2007), "The Time Varying Volatility of Macroeconomic Fluctuations", *mimeo*.

[21] McCracken, M. W. (2000), "Robust Out-of-Sample Inference", *Journal of Econometrics,* 99, 195-223.

[22] McConnell, M.M., and G. Perez-Quiroz (2000), "Output Fluctuations in the United States: What Has Changed Since the Early 1980's", *American Economic Review* 90(5), 1464-1476.

[23] 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.

[24] Newey, W., and K. West (1987), "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica* 55, 703-708.

[25] Pesaran, H., D. Pettenuzzo and A. Timmermann (2004), "Forecasting Time Series Subject to Multiple Structural Breaks", *The Review of Economic Studies,* forthcoming.

[26] Ploberger, W., W. Kramer and K. Kontrus (1989), "A New Test for Structural Stability in the Liner Regression Model", *Journal of Econometrics*, 40, 307-318.

[27] Rivers, D. and Q. Vuong (2002), "Model Selection Tests for Nonlinear Dynamic Models", *Econometrics Journal*, 5, 1-39.

[28] Rossi, B. (2005), "Optimal Tests for Nested Model Selection with Underlying Parameter Instabilities", *Econometric Theory.*

[29] Sin, C.Y. and H. White (1996), "Information Criteria for Selecting Possibly Misspecified Parametric Models", *Journal of Econometrics*, 71, 207-225.

[30] 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.

[31] Stock, J. H. and M. W. Watson (2003), "Combination Forecasts of Output Growth in a Seven-Country Data Set," forthcoming *Journal of Forecasting*

[32] Stock, J.H., and M.W. Watson (2003), "Forecasting Output and Inflation: The Role of Asset Prices", *Journal of Economic Literature*.

[33] Van der Vaart, A. and J.A. Wellner (1996), *Weak Convergence and Empirical Processes with Applications to Statistics*, Springer-Verlag: New York.

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

[35] West, K. D. (1996), "Asymptotic Inference about Predictive Ability", *Econometrica*, 64, 1067-1084.

[36] White, H. (1994), *Estimation, Inference and Specification Analysis,* Cambridge University Press, New York.

[37] Wooldridge, J. M. and H. White (1988): "Some Invariance Principles and Central Limit Theorems for Dependent Heterogeneous Processes", *Econometric Theory,* 4, 210-230.

6 Appendix A - Proofs

Proof of Proposition 1. Let $\sum_j \equiv \sum_{j=t-m/1+1}^{t+m/2}$ for $t = m/2+1, ..., T-m/2$. We first show that $\sigma^{-1}m^{-1/2}\sum_j \Delta L_j(\widehat{\theta}_{t,m}, \widehat{\gamma}_{t,m}) = \sigma^{-1}m^{-1/2}\sum_j \Delta L_j(\theta^*_{t,m}, \gamma^*_{t,m}) + o_p(1)$. Applying a Taylor series expansion, we have that

$$\begin{aligned} & \sigma^{-1}m^{-1/2}\sum_j \Delta L_j(\widehat{\theta}_{t,m}, \widehat{\gamma}_{t,m}) \\ = \; & \sigma^{-1}m^{-1/2}\sum_j \Delta L_j(\theta^*_{t,m}, \gamma^*_{t,m}) \\ & -\sigma^{-1}\frac{1}{2}\left\{ E\left[m^{-1}\sum_j \nabla f_j(\ddot{\theta}_{t,m})\right]\sqrt{m}\left(\widehat{\theta}_{t,m} - \theta^*_{t,m}\right)\right. \\ & \left. -E\left[m^{-1}\sum_j \nabla g_j(\ddot{\gamma}_{t,m})\right]\sqrt{m}\left(\widehat{\gamma}_{t,m} - \gamma^*_{t,m}\right)\right\} \\ = \; & \sigma^{-1}m^{-1/2}\sum_j \Delta L_j(\theta^*_{t,m}, \gamma^*_{t,m}) + o_p(1), \end{aligned} \tag{22}$$

where $\ddot{\theta}_{t,m}$is an intermediate point between $\widehat{\theta}_{t,m}$ and $\theta^*_{t,m}$. Assumptions (c) and (b) ensure that $E\left[m^{-1}\sum_j \nabla f_j(\ddot{\theta}_{t,m})\right] \underset{as}{\to} 0$ and Assumption (b) ensures that the second component in the second to last line is $o_p(1)$. Now write

$$\begin{aligned} & \sigma^{-1}m^{-1/2}\sum_j \Delta L_j(\theta^*_{t,m}, \gamma^*_{t,m}) \\ = \; & (m/T)^{-1/2}\left(\sigma^{-1}T^{-1/2}\sum_{j=1}^{t+m/2} \Delta L_j(\theta^*_{t,m}, \gamma^*_{t,m}) - \sigma^{-1}T^{-1/2}\sum_{j=1}^{t-m/2} \Delta L_j(\theta^*_{t,m}, \gamma^*_{t,m})\right). \end{aligned}$$

By Assumptions (a), (d) and (e), we have

$$\sigma^{-1}m^{-1/2}\sum_j \Delta L_j(\theta^*_{t,m}, \gamma^*_{t,m}) \Longrightarrow \left[\mathcal{B}(\tau + \mu/2) - \mathcal{B}(\tau - \mu/2)\right]/\sqrt{\mu},$$

where $t = [\tau T]$, $m = [\mu T]$. The statement in the proposition then follows from the fact that, under H_0, $\widehat{\sigma}$ in (12) is a consistent estimator of σ (Andrews, 1991). Values of k_α in Table 1 are obtained by Monte Carlo simulations (based on 8,000 Monte Carlo replications and by approximating the Brownian Motion with 400 observations). ■

Proof of Proposition 2. Let $\sum_j \equiv \sum_{j=t-m/2+1}^{t+m/2}$ for $t = R+h+m/2, ..., T-m/2$. We have

$$\begin{aligned}
&\sigma^{-1}m^{-1/2}\sum_{j}\Delta L_j(\widehat{\theta}_{j-h,R},\widehat{\gamma}_{j-h,R}) \\
= \;& (m/P)^{-1/2}\left(\sigma^{-1}P^{-1/2}\sum_{j=R+h}^{t+m/2}\Delta L_j(\widehat{\theta}_{j-h,R},\widehat{\gamma}_{j-h,R})-\sigma^{-1}P^{-1/2}\sum_{j=R+h}^{t-m/2}\Delta L_j(\widehat{\theta}_{j-h,R},\widehat{\gamma}_{j-h,R})\right).
\end{aligned}$$

By Assumptions (a), (b) and (c), we have

$$\sigma^{-1}m^{-1/2}\sum_{j}\Delta L_j(\widehat{\theta}_{j-h,R},\widehat{\gamma}_{j-h,R})\Longrightarrow\left[\mathcal{B}\left(\tau+\mu/2\right)-\mathcal{B}\left(\tau-\mu/2\right)\right]/\sqrt{\mu}.$$

The statement in the proposition then follows from the fact that, under H_0, $\widehat{\sigma}$ in (15) is a consistent estimator of σ (Andrews, 1991). ■

Proof of Proposition 3. First we show that: (I) $LM_1=\sigma^{-2}T^{-1}\left[\sum_{j=1}^{T}\left(\log f_j(\theta_T^*)-\log g_j\left(\gamma_T^*\right)\right)\right]^2+o_p\left(1\right)$ and (II)

$$\begin{aligned}
LM_2\left(t\right) &= \sigma^{-2}\left(t/T\right)^{-1}\left(1-t/T\right)^{-1} \\
&\quad [T^{-1/2}\sum_{j=1}^{t}\left(\log f_j(\theta_{1,t}^*)-\log g_j\left(\gamma_{1,t}^*\right)\right)+ \\
&\quad -\left(t/T\right)T^{-1/2}\sum_{j=1}^{T}\left(\log f_j(\theta_T^*)-\log g_j\left(\gamma_T^*\right)\right)]^2+o_p\left(1\right)
\end{aligned}$$

To prove (I), note that by applying a Taylor expansion:

$$\begin{aligned}
&\sigma^{-2}T^{-1}\sum_{j=1}^{T}\left(\log f_j(\widehat{\theta}_T)-\log g_j\left(\widehat{\gamma}_T\right)\right) \\
= \;& \sigma^{-2}T^{-1}\sum_{j=1}^{T}\left(\log f_j(\theta_T^*)-\log g_j\left(\gamma_T^*\right)\right)+ \\
& \frac{1}{2}\sigma^{-2}T^{-1}\sum_{j=1}^{T}\left(E\left[\nabla\log f_j(\ddot{\theta}_T)\right]\left(\widehat{\theta}_T-\theta_T^*\right)-E\left[\nabla\log g_j\left(\ddot{\gamma}_T\right)\right]\left(\widehat{\gamma}_T-\gamma_T^*\right)\right) \\
= \;& \sigma^{-2}T^{-1}\sum_{j=1}^{T}\left(\log f_j(\theta_T^*)-\log g_j\left(\gamma_T^*\right)\right)+o_p\left(1\right)
\end{aligned}$$

where $\ddot{\theta}_T$is an intermediate point between $\widehat{\theta}_T$ and θ_T^* (similarly for $\ddot{\gamma}_T$). Assumptions (c) and (b) ensure that $E\left[\nabla\log f_j(\ddot{\theta}_T)\right]\underset{as}{\to}0$ and Assumption (b) ensures that the second component in the second to last line is $o_p\left(1\right)$. A similar argument proves (II).

By assumptions (a), (d) and (e), under the null hypothesis (**??**):

$$\sigma^{-1}T^{-1/2}\sum_{j=1}^{T}\left(\log f_j(\theta_T^*)-\log g_j\left(\gamma_T^*\right)\right)\Longrightarrow\mathcal{B}\left(1\right) \tag{23}$$

$$\sigma^{-1}(t/T)^{-1/2}(1-t/T)^{-1/2}[T^{-1/2}\sum_{j=1}^{t}\left(\log f_j(\theta^*_{1,t})-\log g_j\left(\gamma^*_{1,t}\right)\right)$$
$$-(t/T)T^{-1/2}\sum_{j=1}^{T}\left(\log f_j(\theta^*_T)-\log g_j\left(\gamma^*_T\right)\right)]$$
$$\Longrightarrow \quad \tau^{-1/2}(1-\tau)^{-1/2}[\mathcal{B}(\tau)-\tau\mathcal{B}(1)] = \tau^{-1/2}(1-\tau)^{-1/2}\mathcal{BB}(\tau) \tag{24}$$

where (23) and (24) are asymptotically independent. Then:

$$\begin{aligned} LM_1+LM_2(t) &= \sigma^{-2}T^{-1}\left[\sum_{j=1}^{T}\left(\log f_j(\theta^*_T)-\log g_j\left(\gamma^*_T\right)\right)\right]^2 \\ &\quad +\sigma^{-2}\left(\frac{t}{T}\right)^{-1}\left(1-\frac{t}{T}\right)^{-1}[T^{-1/2}\sum_{j=1}^{t}\left(\log f_j(\theta^*_{1,t})-\log g_j\left(\gamma^*_{1,t}\right)\right) \\ &\quad -\left(\frac{t}{T}\right)T^{-1/2}\sum_{j=1}^{T}\left(\log f_j(\theta^*_T)-\log g_j\left(\gamma^*_T\right)\right)]^2+o_p(1) \\ &\Longrightarrow \quad \mathcal{B}(1)^2+\tau^{-1}(1-\tau)^{-1}\mathcal{BB}(\tau)^2 \end{aligned}$$

and the result follows by the Continuous Mapping Theorem. ■

Proof of Proposition 4. Suppose that n is a fixed positive integer greater than 1. Using similar reasonings to those in the Proof of Proposition 1, we first show that $\sigma^{-1}t^{-1/2}\sum_{j=1}^{t}\Delta L_j(\widehat{\theta}_t,\widehat{\gamma}_t) = \sigma^{-1}t^{-1/2}\sum_{j=1}^{t}\Delta L_j(\theta^*_t,\gamma^*_t)+o_p(1)$. Applying a Taylor series expansion we have that (22) holds. Assumptions SEQ(b),(c) ensure that $E\left[t^{-1}\sum_{j=1}^{t}\nabla f_j(\ddot{\theta}_t)\right]$ and $E\left[t^{-1}\sum_{j=1}^{t}\nabla g_j(\ddot{\gamma}_t)\right]$ are bounded in probability on $D[1,n]$ and (b) ensures that the second component in the second to last line is $o_p(1)$ for every t on $D[1,n]$. Then, by Assumptions SEQ (a), (d), and (e), we have that $\sigma^{-1}t^{-1/2}\sum_{j=1}^{t}\Delta L_j(\widehat{\theta}_t,\widehat{\gamma}_t) \Rightarrow B(\tau)/\sqrt{\tau}$ on $D[1,n]$. Next, it follows from Thorem 1.6.1 in van der Vaart and Wellner (1996, p. 43) that these convergence also holds on $D[1,\infty]$. The statement in the proposition then follows from the fact that, under the null hypothesis, $\widehat{\sigma}$ in (21) is a consistent estimator of σ (Andrews, 1991). The critical value is then determined from the hitting probability of the Brownian Motion, as in Chu et al. (1996, p.1053): $P\left\{|B(\tau)|/\sqrt{\tau} \geq \sqrt{(r_\alpha^2+\ln\tau)}, \text{ for some } \tau \geq 1\right\} = 2[1-\Phi(r_\alpha) = r_\alpha\phi(r_\alpha)]$, where $t=[\tau T]$, and $\phi(.)$ and $\Phi(.)$ are, respectively, the pdf and cdf of a standard normal distribution. ■

7 Appendix B

Lemma 5 (A bootstrap procedure robust to breaks in variance) *In the presence of breaks in σ satisfying Assumption ν in Cavaliere and Taylor (2005), the following bootstrap à la Hansen (2000) provides the correct p-values. Let $\overline{L}_t \equiv m^{-1} \sum_{j=t-m/2}^{t+m/2} \Delta L_j$ and let z_t denote an independent $N(0,1)$ sequence. At each point in time t the bootstrap sample is defined as $\Delta L_j^{(b)} \equiv \Delta L_j z_j$, $j = 1, .., m$ and the bootstrap statistic is given by $\sigma_b^{-1} m^{-1/2} \sum_{j=t-m/2}^{t+m/2} \Delta L_j^{(b)}$, where $\sigma_b^2 = m^{-1} \sum_{j=t-m/2}^{t+m/2} \left(\Delta L_j^{(b)}\right)^2$. The critical values of the sample path can be obtained by Monte Carlo simulation.*

8 Tables and Figures

Table 1. Critical values for the fluctuation test (k_α)

	α	
μ	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

Notes to Table 1. The table reports critical values for the in-sample and out-of-sample fluctuation tests $F_{t,m}^{IS}$ and $F_{t,m}^{OOS}$ of Propositions 1 and 2.

Figure 1a. Examples of time variation in relative performance

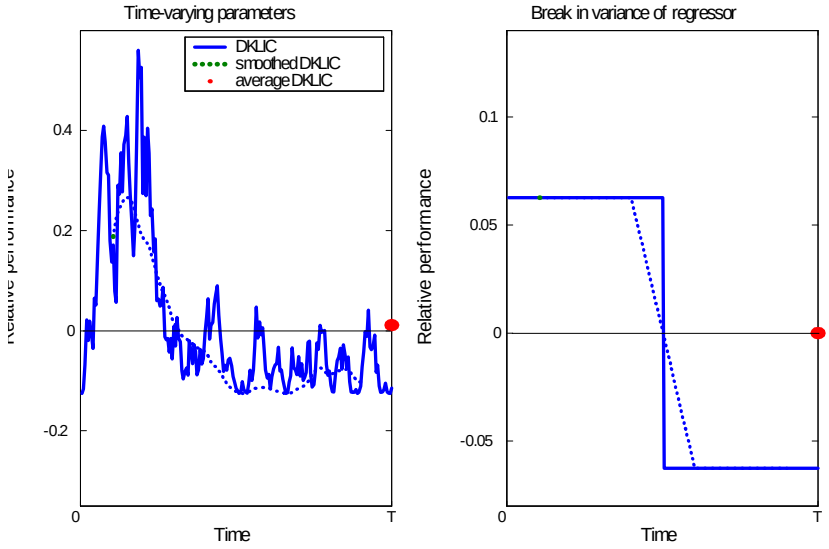
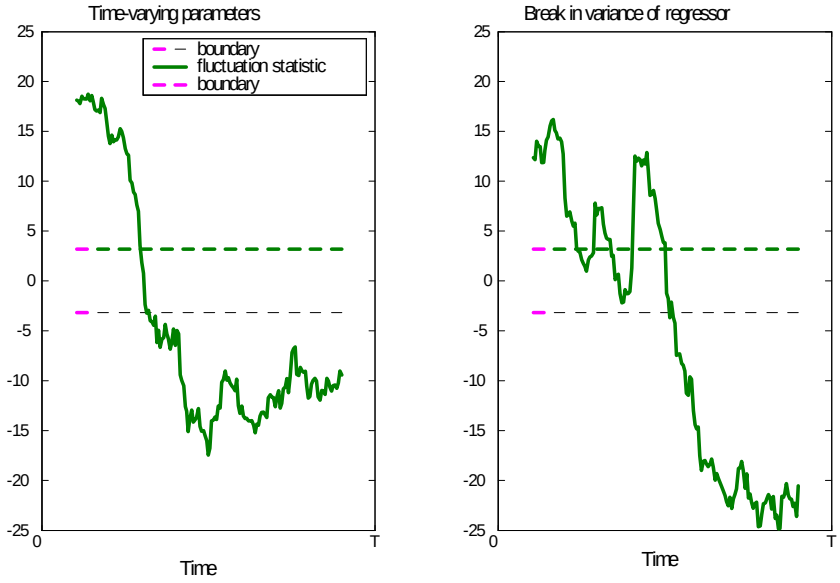
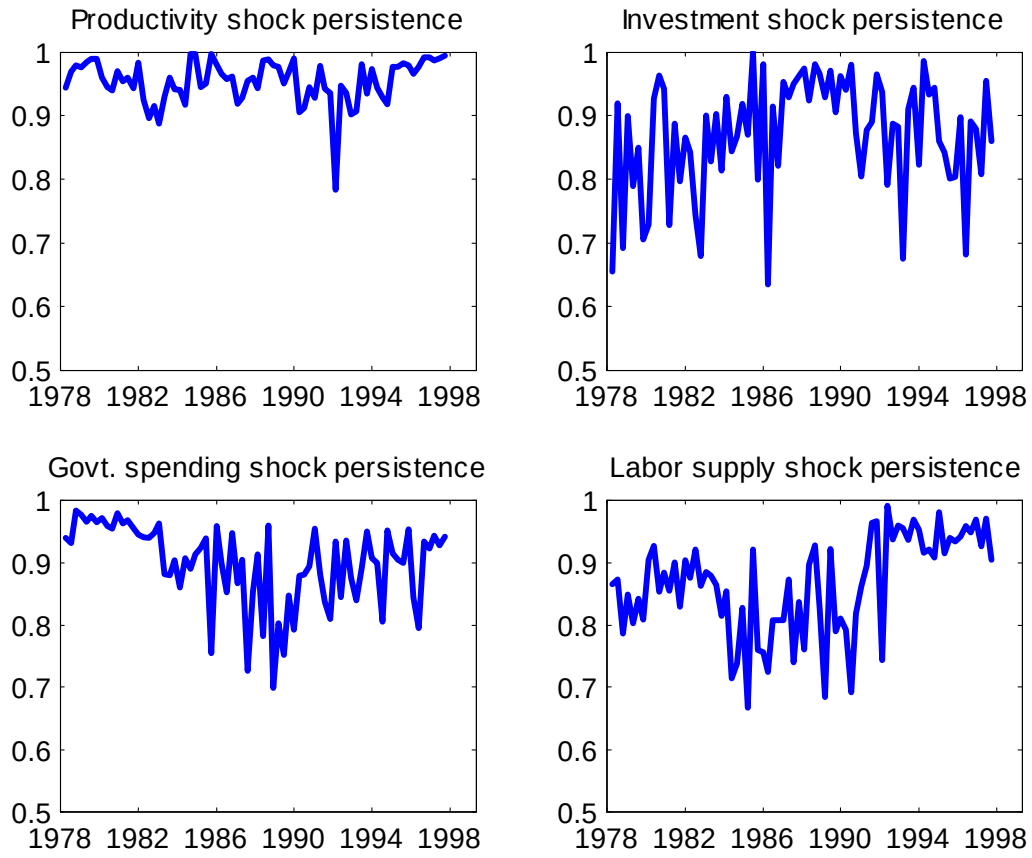


Figure 1b. Implementation of the fluctuation test for the examples in Figure 1a



Notes to Figure 1. Figure 1a shows the sample path for the relative performance of the two models considered in the example in Section 2 for a DGP with time-varying parameters (left panel) or a DGP with constant parameters but with a break in the variance of one regressor (right panel). The solid line represents the relative KLIC at each point in time, the dashed line is the "smoothed" $\Delta KLIC$ computed over moving windows of size 1/5 of the sample size T. The dot represents the full-sample average $\Delta KLIC$. Figure 1b shows the sample path of the in-sample fluctuation test statistics of Proposition 1, together with corresponding boundary

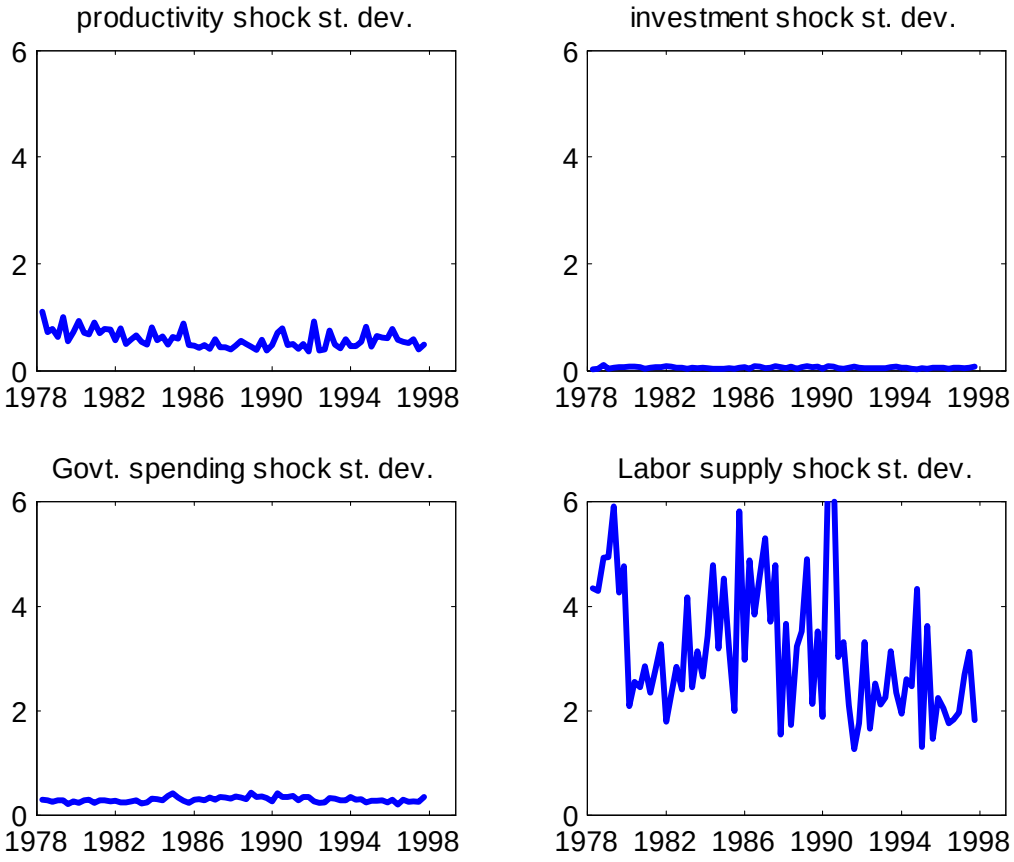
Figure 2a. Rolling estimates of DSGE parameters (persistence of the shocks).



lines.

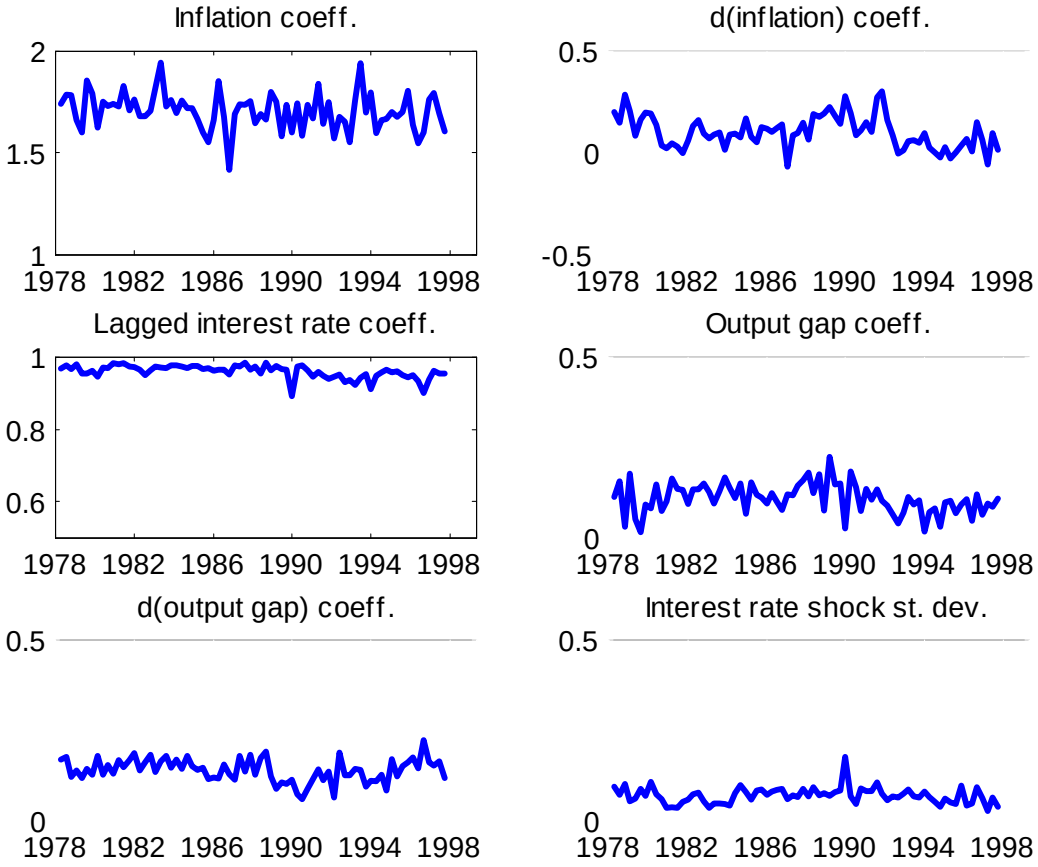
Notes to Figure 2(a). The figure plots rolling estimates of some parameters in Smets and Wouter's (2002) model. See Smets and Wouter's Table 1, p. 1142 for a description.

Figure 2b. Rolling estimates of DSGE parameters (standard deviation of the shocks).



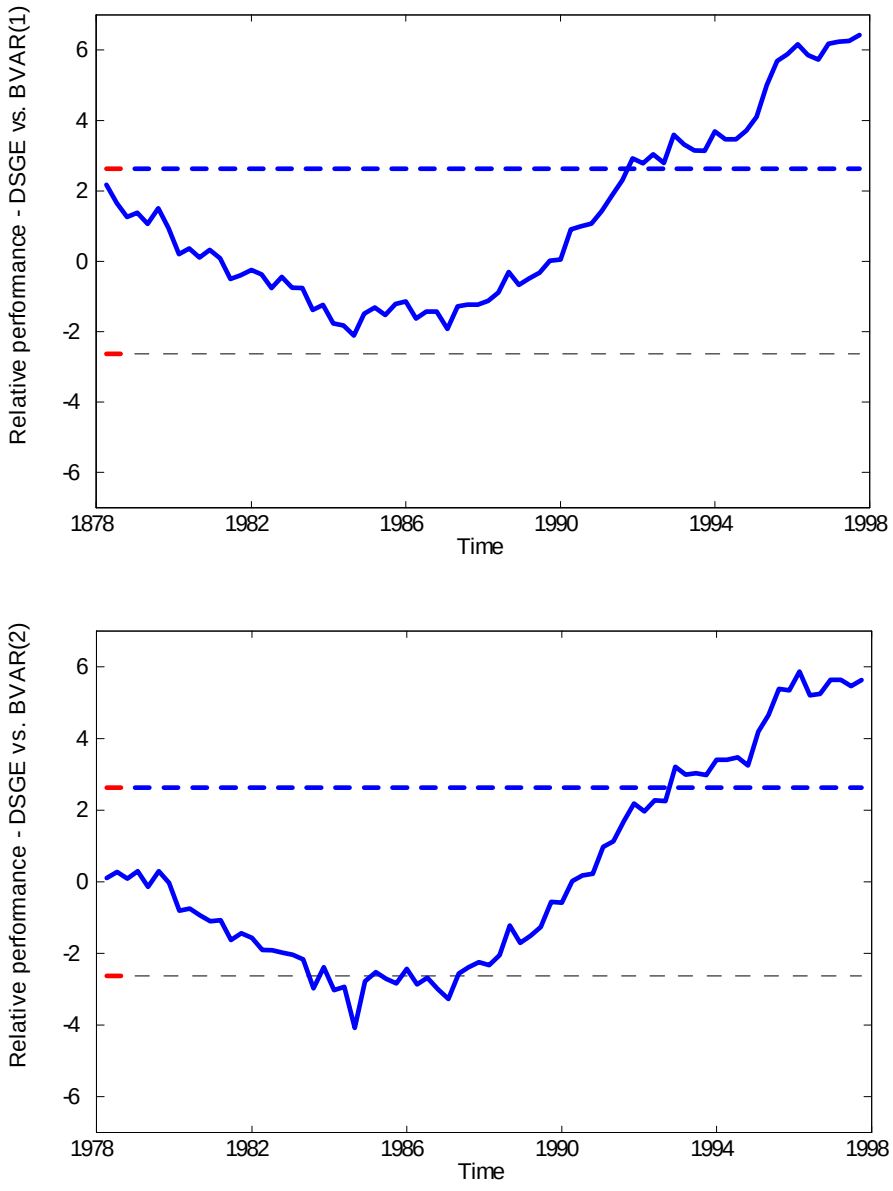
Notes to Figure 2(b). The figure plots rolling estimates of some parameters in Smets and Wouter's (2002) model using full-sample detrended data. See Smets and Wouter's Table 1, p. 1142 for a description.

Figure 2c. Rolling estimates of DSGE parameters (monetary policy parameters).



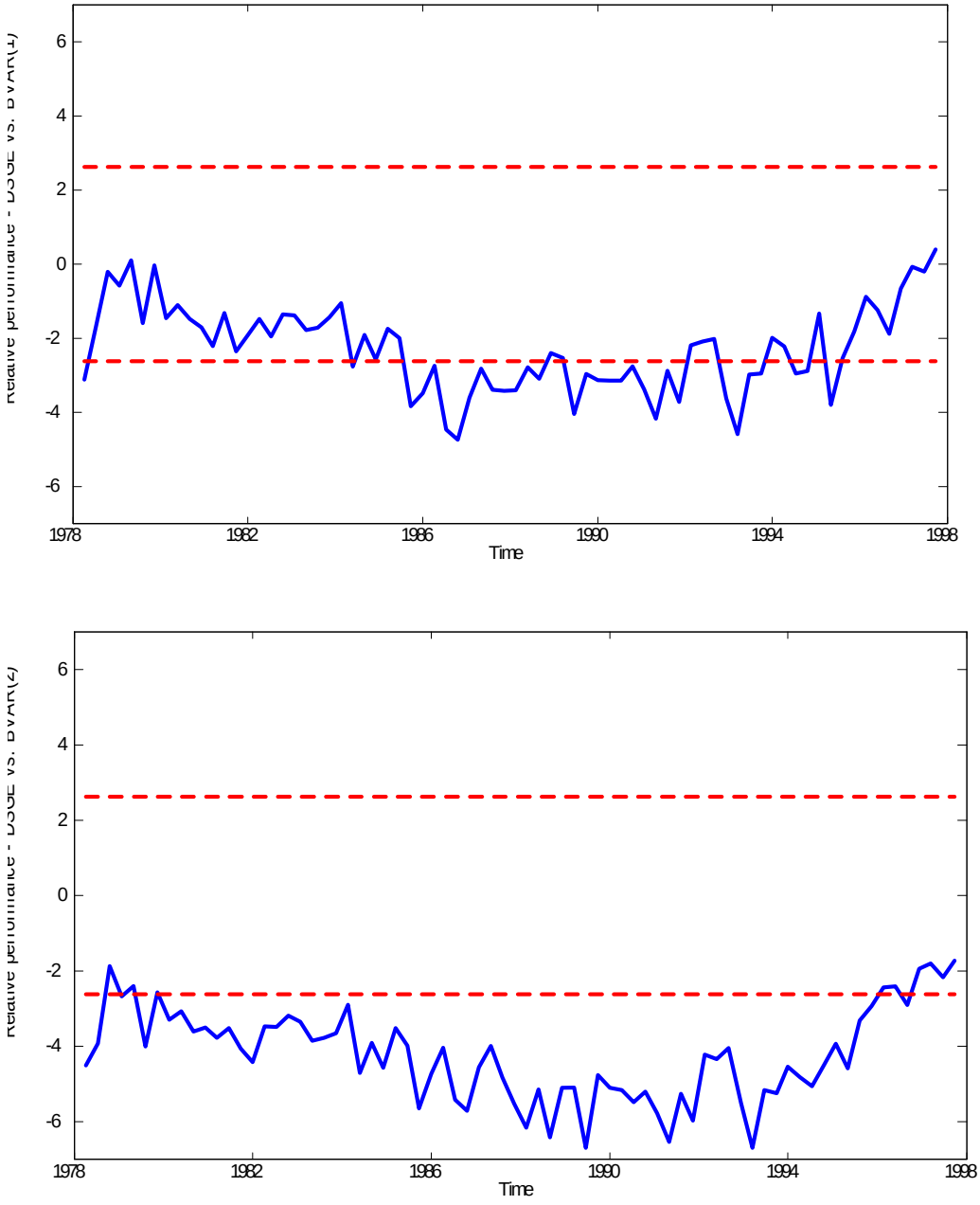
Notes to Figure 2(c). The figure plots rolling estimates of the parameters in the monetary policy reaction function described in Smets and Wouters' (2002) eq. (36), given by: $\widehat{R}_t = \rho \widehat{R}_{t-1} + (1-\rho)\left\{\overline{\pi}_t + r_\pi\left(\widehat{\pi}_{t-1} - \overline{\pi}_t\right) + r_Y\left(\widehat{Y}_{t-1} - \widehat{Y}_t^p\right)\right\} + r_{\Delta\pi}\left(\widehat{\pi}_t - \widehat{\pi}_{t-1}\right) + r_{\Delta Y}\left(\left(\widehat{Y}_t - \widehat{Y}_t^p\right) - \left(\widehat{Y}_{t-1} - \widehat{Y}_{t-1}^p\right)\right) + \eta_t^R$, $\overline{\pi}_t = \rho_\pi \overline{\pi}_{t-1} + \eta_t^\pi$. 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 ($\sqrt{var\left(\eta_t^\pi\right)}$).

Figure 3. Fluctuation test DSGE vs. BVARs. Full-sample detrending



Notes to Figure 3. The figure plots the fluctuation test statistic for testing equal performance of the DSGE and BVARs, using a rolling window of size $m = 70$ (the horizontal axis reports the central point of each rolling window). The 10% boundary lines are derived under the hypothesis that the local $\Delta KLIC$ equals zero at each point in time. The data are detrendend by a linear trend computed over the full sample.The top panel compares the DSGE to a BVAR(1) and the lower panel compares the DSGE to a BVAR(2).

Figure 4. Fluctuation test DSGE vs. BVARs. Rolling sample detrending



Notes to Figure 4. The figure plots the fluctuation test statistic for testing equal performance of the DSGE and BVARs, using a rolling window of size $m = 70$ (the horizontal axis reports the central point of each rolling window). The 10% boundary lines are derived under the hypothesis that the local $\Delta KLIC$ equals zero at each point in time. The data are detrendend by a linear trend computed over each rolling

window.The top panel compares the DSGE to a BVAR(1) and the lower panel compares the DSGE to a BVAR(2).