Vo Phuong Mai Le; Meenagh, David; Minford, Patrick; Wickens, Michael

Working Paper

Testing DSGE models by indirect inference and other methods: Some Monte Carlo experiments

Cardiff Economics Working Papers, No. E2012/15

Provided in Cooperation with:
Cardiff Business School, Cardiff University

Suggested Citation: Vo Phuong Mai Le; Meenagh, David; Minford, Patrick; Wickens, Michael (2012) : Testing DSGE models by indirect inference and other methods: Some Monte Carlo experiments, Cardiff Economics Working Papers, No. E2012/15, Cardiff University, Cardiff Business School, Cardiff

This Version is available at:
https://hdl.handle.net/10419/65790

CARDIFF BUSINESS SCHOOL
WORKING PAPER SERIES

**Cardiff Economics
Working Papers**

Vo Phuong Mai Le, David Meenagh, Patrick Minford and Michael Wickens

Testing DSGE models by Indirect inference and other methods: some Monte Carlo experiments

E2012/15

Cardiff Business School
Cardiff University
Colum Drive
Cardiff CF10 3EU
United Kingdom
t: +44 (0)29 2087 4000
f: +44 (0)29 2087 4419
www.cardiff.ac.uk/carbs

ISSN 1749-6101
June 2012

Testing DSGE models by Indirect inference and other methods: some Monte Carlo experiments

Vo Phuong Mai Le (University of Sheffield)* David Meenagh (Cardiff University)†

Patrick Minford (Cardiff University and CEPR)‡

Michael Wickens (Cardiff University, University of York and CEPR)§

June 2012

Abstract

Using Monte Carlo experiments, we examine the performance of Indirect Inference tests of DSGE models, usually versions of the Smets-Wouters New Keynesian model of the US postwar period. We compare these with tests based on direct inference (using the Likelihood Ratio), and on the Del Negro-Schorfheide DSGE-VAR weight. We find that the power of all three tests is substantial so that a false model will tend to be rejected by all three; but that the power of the indirect inference tests are by far the greatest, necessitating re-estimation by indirect inference to ensure that the model is tested in its fullest sense.

JEL Classification: C12, C32, C52, E1,

Keywords: Bootstrap, DSGE, New Keynesian, New Classical, indirect inference, Wald statistic, likelihood ratio, DSGE-VAR weight

*V.Le@sheffield.ac.uk; University of Sheffield, Department of Economics, 9 Mappin Street, Sheffield, UK

†Meenaghd@cf.ac.uk; Cardiff Business School, Cardiff University, Aberconway Building, Colum Drive, Cardiff, CF10 3EU, UK

‡Patrick.minford@btinternet.com; Cardiff Business School, Cardiff University, Aberconway Building, Colum Drive, Cardiff, CF10 3EU, UK

§Cardiff Business School, Cardiff University, Aberconway Building, Colum Drive, Cardiff, CF10 3EU, UK

1 Introduction

An unresolved issue in macroeconomics is the best way to evaluate the empirical performance of DSGE models. This problem has become more pressing due to the increasing use of DSGE models estimated using Bayesian methods. The aim of this paper is to compare a number of different ways of evaluating DSGE models, including carrying out formal statistical tests.

DSGE models emerged largely as a response to the perceived shortcomings of previous formulations of macroeconometric models. The main complaint was that these macroeconometric models were not structural and so were subject to Lucas' critique (Lucas, 1976). They were also criticised for incorporating "incredible" identifying restrictions (Sims, 1960) and for over-fitting the data due to data-mining. In contrast, the emphasis in DSGE models - particularly their early RBC manifestation - is on the logical coherence of their theoretical structure and choosing parameter values through calibration rather than conventional econometric estimation.

For all their theoretical advantages, the strong simplifying restrictions on the structure of DSGE models resulted in a severe deterioration of fit. This had not been such a problem with the previous generation of macro models in which ad hoc supply and demand functions with lagged adjustment had been sufficiently flexible to pass tests of the models; indeed they still did so even when rational expectations were added.

There have been various reactions to this empirical failure of DSGE models. Supporters of DSGE models - such as Prescott and Lucas (see Canova, 1994, for a discussion of their position) - argued that these models were 'misspecified' or 'false' and so should not be tested by the usual econometric methods which would always reject them. Rather they should be evaluated by whether in simulation they could generate behaviour mimicing the 'stylised facts'. This mimicry could not be evaluated by statistical inference; rather it was an informal comparison. Others, following Sims's criticisms of earlier macroeconometric models, thought that DSGE models were far too simple to capture the economy's real complexity and preferred instead to build time-series models of the economy, such as VARs. A third reaction has been to replace calibration with Bayesian estimation, a weaker way of incorporating prior beliefs, and to specify a more flexible lag structure determined more by the data, for example, by including disturbances that are allowed to be serially correlated. As a result, modern DSGE models tend to fit the data much better. Nonetheless,

the concern remains that there would be no need to use Bayesian estimation if the model was correct, as classical estimation would yield similar parameter values. The need to evaluate DSGE models empirically therefore remains.

Various econometricians — for example Watson (1993), Canova (1994, 1995, 2005), Del Negro and Schorfheide (2004, 2006) — have shown an interest in evaluating DSGE models statistically. They accepted the argument that these models were misspecified and, instead of testing these models, they suggested measures of their overall 'closeness' to the data. The role of such closeness measures was not to reject the model, rather to act as a ranking device for alternative models (all of which implicitly would be rejected). Bayesians in particular suggested the use of marginal likelihood testing, whereby features of the model that could be varied as alternative specifications could be ranked by their likelihood values. Overall closeness measures could then be seen as linked to these marginal tests since a model with less marginally rejected features should be closer overall to the data. To quote Canova (1994), these econometricians were asking about a DSGE model: 'given that it is false, how true is it?' Bayesian methods could seek out marginal features that could enhance the extent of its truth.

Measuring the closeness of the predictions of a DSGE model to actual data rather than carrying out a formal statistical test of the model might be justified on the grounds that since DSGE models are known to be simplifications they cannot be a true representation of the macroeconomy. Support for this view may be found as far back as Friedman (1953) who laid down generally accepted foundations of economic testing methodology according to which a model should not be tested on its 'literal truth' but rather on its implications for the data it was designed to explain. Thus a macro model should be tested on its ability to explain macro data, not on whether it is a gross simplification of a complex reality; indeed as Friedman noted, the more simple a theory, the greater its potential explanatory power, in the sense that it has a higher ratio of explanation to theoretical complexity. He gave the example of perfect competition which could, though never actually existent, predict closely the behaviour of industries with a high degree of competition. Nonetheless, it is still more informative to assess closeness of fit to the data by measuring the probability that the data could be generated by the model, using a statistical test.

A standard likelihood ratio test is based on all of the parameters of a model. This appears to be a

stringent criterion.[1] In a recent paper Corradi and Swanson (2007) proposed a likelihood test based on the model's simulated distribution for output alone. This is, perhaps, too weak a criterion. Another criterion, that may reflect better the purposes for which the model is built, is to base a test on key parameters or features of interest such as measures of certain 'stylised facts', for example, key moments and cross-moments. This provides a way of implementing the notion that a DSGE model may be a good model in the sense that it can capture data features of principal interest. The Wald statistic proposed in this paper provides a way of evaluating whether a DSGE model satisfies this criterion but is potentially flexible in its choice of data features to focus on. In this way we provide valid statistical foundations for inference about a model in respect of chosen features of the data rather than the data as a whole.

The distribution of the Wald statistic under the null of the DSGE model can be easily and accurately obtained using the bootstrap; it can be seen as an aspplication of indirect inference. The details of this testing procedure are dealt with in Le et al (2011), together with Monte Carlo experiments gauging the procedure's accuracy and the consistency of the bootstrap in this context. As the test examines more than one feature, the test is based their joint distribution under the null of the model. The results of this test may be very different from that based on matching these features one by one. In the rest of this paper we review the main properties of indirect inference as implemented by our test procedure and compare them with other procedures currently in use to evaluate DSGE models.

We begin in section 2 by reviewing the main features of the indirect inference testing procedure as we implement it; in it we cover the major issues that have arisen in the application of this largely new and unfamiliar testing procedure[2]. In section 3 we compare this testing method with the widely used Likelihood Ratio test used in Direct Inference. In section 4 we go on to compare both these two methods with the evaluation procedure proposed by the Bayesians Del Negro and Schorfheide (2006), in which an unrestricted

[1] In a recent interview Sargent remarked of the early days of testing DSGE models: '...my recollection is that Bob Lucas and Ed Prescott were initially very enthusiastic about rational expectations econometrics. After all, it simply involved imposing on ourselves the same high standards we had criticized the Keynesians for failing to live up to. But after about five years of doing likelihood ratio tests on rational expectations models, I recall Bob Lucas and Ed Prescott both telling me that those tests were rejecting too many good models.' Tom Sargent, interviewed by Evans and Honkapohja (2005)

[2] This discussion draws extensively on Le et al (2011) — q.v. for more details — as well as discussing some issues not covered there.

In all our work reported here we deal with stationary data; in practice most macroeconomic data has to be detrended in some way for this to be possible. We largely ignore issues of how such detrending should be done and what effect it has on the analysis. However in Davidson et al (2010) we show how these testing methods can be extended to non-stationary data and illustrate the approach with an application to UK data. Nevertheless this remains an important area for further research as we have not investigated most of the issues discussed here in a situation of unfiltered data.

VAR is weighted with the restricted VAR implied by the DSGE model; the weight maximising the marginal likelihood is used as an index of misspecification. We show that this weight can be treated as a test statistic, if its distribution is found by bootstrapping[3]. Our final section concludes with some overall lessons of these comparisons.

2 Model evaluation by indirect inference

Indirect inference provides a classical statistical inferential framework for judging a calibrated or already, but maybe partially, estimated model whilst maintaining the basic idea employed in the evaluation of the early RBC models of comparing the moments generated by data simulated from the model with actual data. An extension of this procedure is to posit a general but simple formal model (an auxiliary model) — in effect the conditional mean of the distribution of the data — and base the comparison on features of this model, estimated from simulated and actual data. If necessary these features can supplemented with moments and other measures directly generated by the data and model simulations.

Indirect inference on structural models may be distinguished from indirect estimation of structural models. Indirect estimation has been widely used for some time, see Smith (1993), Gregory and Smith (1991,1993), Gourieroux et al. (1993), Gourieroux and Monfort (1995) and Canova (2005). In estimation the parameters of the structural model are chosen so that when this model is simulated it generates estimates of the auxiliary model similar to those obtained from actual data. The optimal choice of parameters for the structural model are those that minimise the distance between a given function of the two sets of estimated coefficients of the auxiliary model. In the use of indirect inference for model evaluation the parameters of the structural model are taken as given. The aim is to compare the performance of the auxiliary model estimated on simulated data derived from the given estimates of a structural model — which is taken as the true model of the economy, the null hypothesis — with the performance of the auxiliary model when estimated from

[3] Testing an overall model against the data is not something that Bayesian analysis does because it involves questioning the priors and their distribution. However, in economics priors cannot be considered as at all firm in many cases; a given theory can contain a wide difference in values for many parameters. Thus it is natural to ask whether the whole model fits the data since this could lead one to seek different values for the parameters restricted under the priors. Bayesian estimation appears to us to be a most useful method for obtaining initial ML estimates of a model which may be hard to obtain without some restrictions on the likelihood surface. However our argument here is that it needs to be supplemented with a test of the overall success of the resulting estimates. The Bayesian λ which is nowadays widely estimated as a measure of closeness can also provide a test statistic, as we show in this section.

actual data. If the structural model is correct then its predictions about the impulse responses, moments and time series properties of the data should match those based on actual data. The comparison is based on the distributions of the two sets of parameter estimates of the auxiliary model, or of functions of these estimates.

Our choice of auxiliary model exploits the fact that the solution to a log-linearised DSGE model can be represented as a restricted VARMA model and this can be closely represented by a VAR. For further discussion on the use of a VAR to represent a DSGE model, see for example Canova (2005), Dave and DeJong (2007), Del Negro and Schorfheide (2004, 2006) and Del Negro et al. (2007a,b) together with the comments by Christiano (2007), Gallant (2007), Sims (2007), Faust (2007) and Kilian (2007). A levels VAR can be used if the shocks are stationary, but a VAR in differences may be needed if the shocks are permanent as is commonly assumed in DSGE models for productivity (real) and money supply shocks (nominal). This is because, if productivity shocks are permanent, then the production function is not cointegrated and so the associated VAR representation in levels would have non-stationary disturbances. The *a priori* structural restrictions of the DSGE model impose restrictions on the VAR; see Canova and Sala (2009) for an example of lack of identification based on a simple three equation model consisting of a new Keynesian IS function, a Phillips Curve and a Taylor Rule. Provided this VAR is over-identified, the DSGE model can be tested by comparing unrestricted VAR estimates (or some function of these estimates such as the value of the log-likelihood function or the impulse response functions) derived using data simulated from the DSGE model with unrestricted VAR estimates obtained from actual data. If the solved VAR is not identified then it is not restricted; the null and alternative hypotheses are then indistinguishable in the VAR. The Smets-Wouters (2007a, SW) model we will use in what follows is clearly over-identified; changes in its parameter values imply quite different simulation properties.[4]

The model evaluation criterion we use is based on the difference between the vector of relevant VAR coefficients from simulated and actual data as represented by a Wald statistic. If the DSGE model is correct (the null hypothesis) then the simulated data, and the VAR estimates based on these data, will not be

[4] We discuss the issue of identification in Le, Minford and Wickens (2012); there we propose a numerical test for identification based on indirect inference and show that the SW model is idetified according to it, as is also an example New Keynesain three-equation model.

significantly different from those derived from the actual data. The method is in essence extremely simple; though it is numerically taxing, with modern computer resources it can be carried out quickly. The idea is to bootstrap the DSGE model with the shocks it and the data imply must have generated the data. These bootstrap simulations represent what the model and its implied shocks could have generated for the sample historical period of the data. The test then compares the VAR coefficients estimated on the actual data with the VAR coefficient distribution coming from VAR estimates on the many simulated samples. The Wald statistic, WS, computes the joint distance of the data coefficients from the mean of the simulation distribution as follows:

$$WS = (a_T - \overline{a_S(\theta_0)})' W(\theta_0)(a_T - \overline{a_S(\theta_0)})$$

where $W(\theta_0)$ is the inverse of the variance-covariance matrix of the distribution of $a_T - \overline{a_S(\theta_0)}$. Here a_T is the set of VAR coefficients found in the data while $\overline{a_S(\theta_0)}$ is the mean of the VAR coefficients found in the simulated samples; the distribution of $a_T - \overline{a_S(\theta_0)}$ is estimated as the bootstrap distribution of a_S around its mean. We are not compelled to use the VAR coefficients in this formula: thus one could use other data 'descriptors' considered to be key features of the data that the model should match — these could be particular impulse response functions (such as to a monetary policy shock) or particular moments (such as the correlations of various variables with output). However, such measures are functions of the VAR coefficients and it seems that a parsimonious set of features is these coefficients themselves. There are still issues about which variables to include in the VAR (or equivalently whether to focus only on a subset of VAR coefficients related to these variables) and what order the VAR should be. Also it is usual to include the variances of the data or of the VAR residuals as a measure of the model's ability to match variation. We discuss these issues further below.

The idea of the test can be seen usefully with a simple example, with just two VAR coefficients being assessed. The diagram below shows the joint distribution of the two coefficients under two assumptions: the top illustrates the case where the two are uncorrelated, the bottom where they are highly correlated. Two data points are shown, one in blue, the other in red. The Wald statistic would not reject the blue point in the top case, but reject it in the bottom case. It would reject the red point in the top case but not in the bottom

case. In the top case the covariance matrix of the coefficients has zero off-diagonal elements, whereas in the bottom case they are non-zero. It is unusual to find zero off-diagonal elements because different features of the data to be correlated across the samples generated by the DSGE model[5].

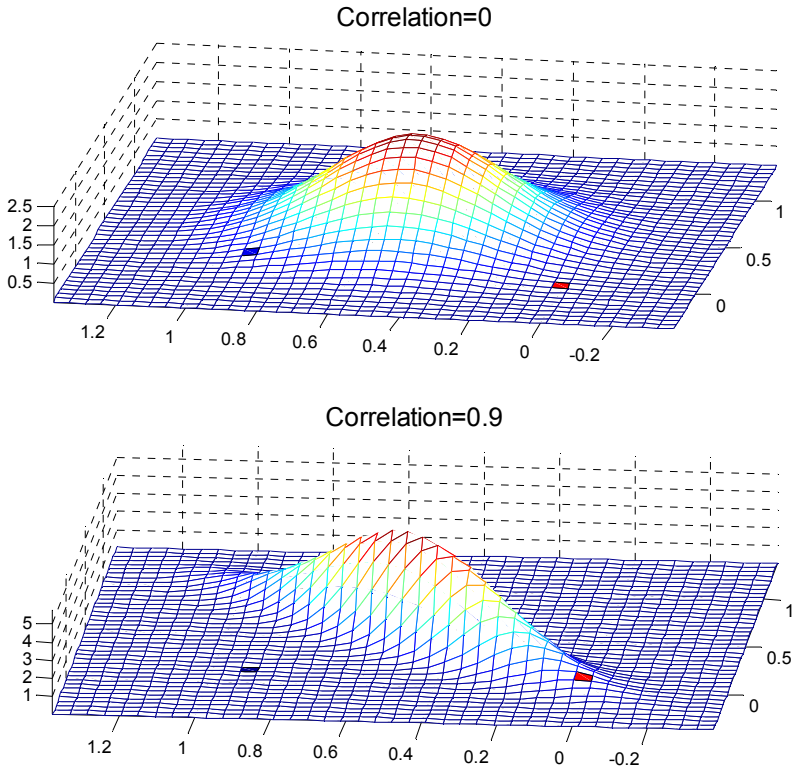


Figure 1: Bivariate Normal Distributions $(0.1, 0.9$ shaded blue and $0, 0$ shaded red) with correlation of 0 and 0.9.

[5] To understand why DSGE models will typically produce high covariances and so distributions like those in the bottom panel of Figure 1, we can give a simple example in the case where the two descriptors are the persistence of inflation and interest rates. If we recall the Fisher equation, we will see that the persistence of inflation and interest rates will be highly correlated. Thus in samples created by the DSGE model from its shocks where inflation is persistent, so will interest rates be; and similarly when the former is non-persistent so will the latter tend to be. Thus the two estimates of persistence under the null have a joint distribution that reflects this high correlation.

In Figure 1, we suppose that the model distribution is centred around 0.5 for each VAR coefficient; and the data-based VAR produced values (the blue ones) for their partial autocorrelations of 0.1 and 0.9 respectively for inflation and interest rates — the two VAR coefficients. We suppose too that the 95% range for each was $0 - 1.0$ (a standard deviation of 0.25) and thus each is accepted individually. If the parameters are uncorrelated across samples, then the situation is as illustrated in the top panel. They will also be jointly accepted.

Now consider the case where there is a high positive covariance between the parameter estimates across samples, as implied by the DSGE model (with its Fisher equation). The lower panel illustrates the case for a 0.9 cross-correlation between the two parameters. The effect of the high covariance is to create a ridge in the density mountain; and the joint parameter combination of $0.1, 0.9$ will be rejected even though individually the two parameters are accepted.

The red data values $(0, 0)$ analogously reject the model in the top case but do not reject it in the more usual bottom case of high correlation.

We can show where in the Wald bootstrap distribution the Wald based on the data lies (the Wald percentile). We can also show the Mahalanobis Distance based on the same joint distribution, normalised as a t-statistic, and also the equivalent Wald p-value, as an overall measure of closeness between the model and the data.[6] In Le et al. (2011) we applied this test to a well-known model of the US, that of Smets and Wouters (2007; qv); we found that the Bayesian estimates of SW were rejected for both the full post-war sample and for a more limited post-1984 (great Moderation) sample. However, we modified the model by adding competitive goods and labour market sectors and we searched with a powerful Simulated Annealing algorithm for other parameter sets that might improve the Wald, succeeding in finding a set for the post-1984 sample. Here we refer to some aspects of that work to illustrate the test's application.

Thus the joint distribution of the VAR(1) coefficients for 5 variables and their data variances in the Smets-Wouters model gave a distribution for the Wald statistic shown in Figure 2:

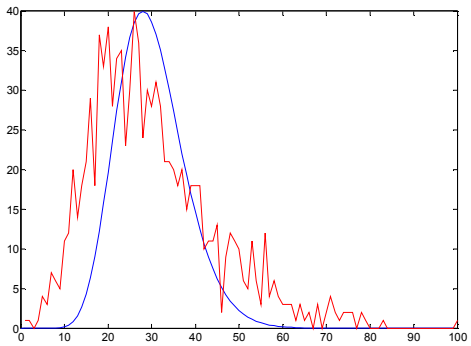


Figure 2: Histogram of Wald statistic for weighted SW, with Chi-squared distribution (30 degrees of freedom)

It may be observed that the small sample distribution given by 1000 bootstraps differs from the asymptotic distribution shown in this figure; the small sample distribution is in principle the one we wish to use, subject to any inaccuracy imported by the bootstrap.

The same Wald statistic transformed to a normalised t-value (Mahalanobis Distance) is shown next. This is arrived at as follows. The $\sqrt{\chi^2}$ is a normal variable; $\sqrt{Wald}$ is the Mahalanobis Distance, MD; by adjusting its mean and standard deviation this can be normalised as a t-value where the 95% point is 1.645

[6] The Mahalanobis Distance is the square root of the Wald value. As the square root of a chi-squared distribution, it can be converted into a t-statistic by adjusting the mean and the size. We normalise this here by ensuring that the resulting t-statistic is 1.645 at the 95% point of the distribution.

as in Figure 3.

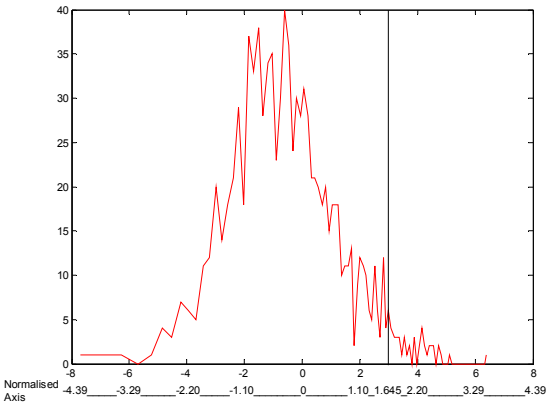


Figure 3: Normalised Mahalanobis Distance

The jointness of the distribution of the VAR coefficients can be illustrated from the two-coefficient case shown in the three-dimensional diagram above. With 2 coefficients, as in that diagram, the Wald statistic for the blue values can be written as $\frac{t_1^2+t_2^2-2t_1t_2\rho_{12}}{1-\rho_{12}^2}$ where $t_1 = -1.6, t_2 = +1.6$ are the data points expressed as the distance from the mean relative to their standard deviation; thus in the top panel of Figure 1 $\rho_{12} = 0.9$ and the Wald=51.2; whereas if it is the case (Figure 1, bottom) that $\rho_{12} = 0$, then the Wald=5.12. The 95% value of the $\chi^2(2)$ is 5.99. Thus we can see how a high correlation affects the Wald statistic and so the joint significance of the two coefficients. In the case of the red values the Wald is 8.0 for $\rho_{12} = 0$ (a rejection) but falls to 4.2 for $\rho_{12} = 0.9$ (non-rejection).

A variety of issues concerning the use of the bootstrap and the robustness of these methods more generally are dealt with in Le at al (2011)[7]. A particular concern with the bootstrap has been its consistency under conditions of near-unit roots. Several authors (e.g. Basawa et al., 1991, Hansen (1999) and Horowitz, 2001a,b) have noted that asymptotic distribution theory is unlikely to provide a good guide to the bootstrap distribution of the AR coefficient if the leading root of the process is a unit root or is close to a unit root. This is also likely to apply to the coefficients of a VAR when the leading root is close to unity and may therefore affect indirect inference where a VAR is used as the auxiliary model, as here. In Le et al. (2011) we carried

[7] Since that paper we have looked at two further issues in more depth: the choice of covariance estimate in the Wald, and the method of error extraction from the DSGE model being tested. This work is shown in the Appendix to this paper.

out a Monte Carlo experiment to check whether this was a problem in models such as SW. We found that the bootstrap was reasonably accurate in small samples, converged asymptotically on the appropriate chi-squared distribution and, thus being asymptotically chi-squared, satisfied the usual requirement for consistency of being asymptotically pivotal.

3 Comparing Indirect Inference with other testing methods in general use

It is useful to consider how other methods of testing are related to indirect inference. If there is a distribution of certain features of the data behaviour under the null and hence of a Wald statistic derived from these features, we would also like to know what the distribution is under the null for other measures of fit, including likelihood ratio tests and the closeness DSGE-VAR λ measure of Del Negro and Schorfheide.[8]

Another question of interest is the relative power of these various tests against misspecified models. That is, suppose we set up the test based on the distribution under the null hypothesis; we want to know how the rejection rate rises as the true model deviates increasingly from the null.

These questions can be answered by Monte Carlo experiments in which we posit a true model and find out the distribution of the various test statistics for this model. We can check power by using test statistics under the null of a particular model but generate the data used in these tests from an alternative true model; we can then see the frequency of rejection of the null (false) model as we move it increasingly further from the true model.

3.1 Some preliminary experiments comparing indirect with direct inference

We would like to know how far the indirect inference tests replicate tests via direct inference. In particular, we compare the DSGE model performance in respect of properties where we could conceivably expect some

[8] The marginal likelihood ratio tests carried out in Bayesian estimation apply to features of the model: they measure the extent to which adding a particular feature improves the closeness of the model to the data. Hence they are not measures of closeness for the whole model like our Wald statistic. Also as measures of improvement they will depend on which features of the model are taken as given already. It can be that a particular feature that we are interested in will reduce the closeness of a model with given parameter values; yet when those parameter values are allowed to change to maximise the model's fit when this feature is added, the model may perform better than when the feature is absent.

relationship. For our comparison we use the Smets-Wouters model of the US, for the whole post-war sample, investigated in Le et al. (2011), with a VAR as the auxiliary model. We treat this model as the true one and generate 1000 bootstraps.

The two tests focus on rather different aspects of a model's performance. Direct inference asks how close the model gets to forecasting current data; indirect inference asks how close the model gets to replicating properties of the auxiliary model we obtain from the data. As the Direct Inference test statistic we take the Likelihood Ratio for the DSGE model against the unrestricted VAR; thus each LR shows how well the DSGE model forecasts the 'data' compared with the VAR estimated for that data.

We check the power of the Wald test by positing a variety of false models. We generate the falseness by introducing a rising degree of numerical mis-specification for the model parameters. Thus we construct a False DSGE model whose parameters were moved $x\%$ away from their true values in both directions (even positive, odd negative); similarly the higher moments of the error processes (standard deviation, skewness and kurtosis) are altered by the same $x\%$. We think of this False Model as having been proposed after either calibration or estimation on some set of actual true data or a combination of both. The essential point is that the False Model is believed by the researcher to be potentially true; the relevant question is how the testing procedure will evaluate this belief. We assume that the degree of error in the beliefs, whether of parameters or errors, cannot be obviously identified from just examining the data sample; as we see in the appendix, from a given sample with a given θ one can extract different error processes with different methods so that there is no uniquely determined set of ρs and error moments; similarly θ can also be variously estimated or calibrated. Thus for any given data sample the information in the sample itself will be consistent with a wide range of parameters and error moments.

Thus for the evaluation of the power of Indirect Inference under our Monte Carlo procedure we generate 10000 samples from some True model (where we take an error distribution with the variance, skewness and kurtosis found in the SW model errors), and find the distribution of the Wald for these True samples. We then generate a set of 10000 samples from the False model θ and calculate the Wald distribution for this False Model. We then calculate how many of the actual samples from the True model would reject the False Model on this calculated distribution with 95% confidence. This gives us the rejection rate for a given percentage

degree x of mis-specification. We use 10000 samples for the Indirect Inference case because the size of the variance-covariance matrix of the VAR coefficients is large for high VAR orders and variable numbers.[9]

For the evaluation of the power of Direct Inference we need to ask how well the DSGE model forecasts within samples generated by the True Model compared with a VAR model fitted to those samples. We use the first 1000 samples as no more are needed in this case. The DSGE model is given a θ and the residuals and ρs are extracted by LIML from each sample. In a forecasting test the model is given at each stage the lagged data including the lagged errors. We assume that since the lagged errors are observed in each sample, the researcher can also estimate the implied ρs for the sample errors and uses these in its forecast. We assume the researcher does this by LIML as this is a robust method — the DSGE model's forecasting capacity is clearly helped by the presence of these autoregressive error processes. We find the distribution of the LR when θ is the true model; and then we apply the 5% critical value from this to the False model LR value for each True sample to obtain the rejection rate for the False Model. A False model is again chosen by changing its θ parameters by + or $-x\%$.

Thus the two methods share the same first 1000 samples generated from the same True model. In Indirect it checks the VAR behaviour of these against the simulated VAR behaviour from the DSGE model; in Direct it checks the relative forecast accuracy of the DSGE model versus that of the VAR model on these samples.

Percent Mis-specified	Indirect Inference	Direct Inference
True	5.0	5.0
1	19.8	6.3
3	52.1	8.8
5	87.3	13.1
7	99.4	21.6
10	100.0	53.4
15	100.0	99.3
20	100.0	99.7

Table 1: Rejection Rates for Wald and Likelihood Ratio for 3 Variable VAR(1)

[9] We assume in this the accuracy of the bootstrap itself as an estimate of the distribution; the bootstrap substitutes repeated drawings from errors in a particular sample for repeated drawings from the underlying population.
Le et al (2011) evaluate the accuracy of the bootstrap for the Wald distribution and find it to be fairly high.

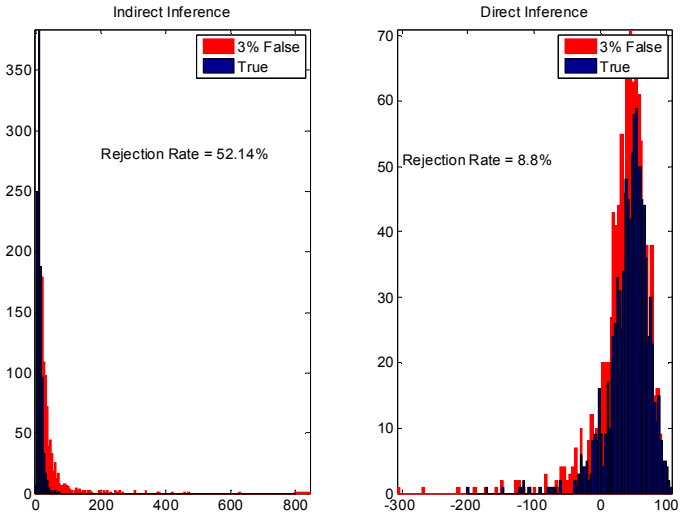


Figure 4: Histograms of true and 3% false models for Indirect and Direct Inference (some outliers deleted from Indirect for the chart)

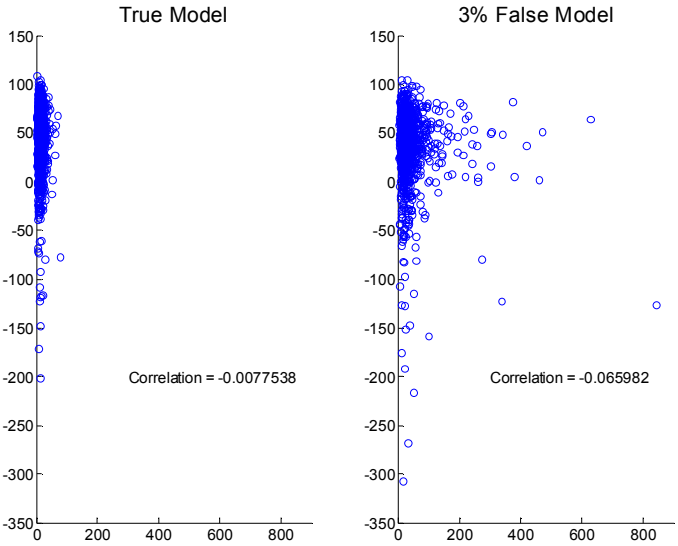


Figure 5: Scatter Plots of Indirect Inference (Wald) v. Direct Inference (LR) for 1000 samples of True Model (3 Variable VAR(1))

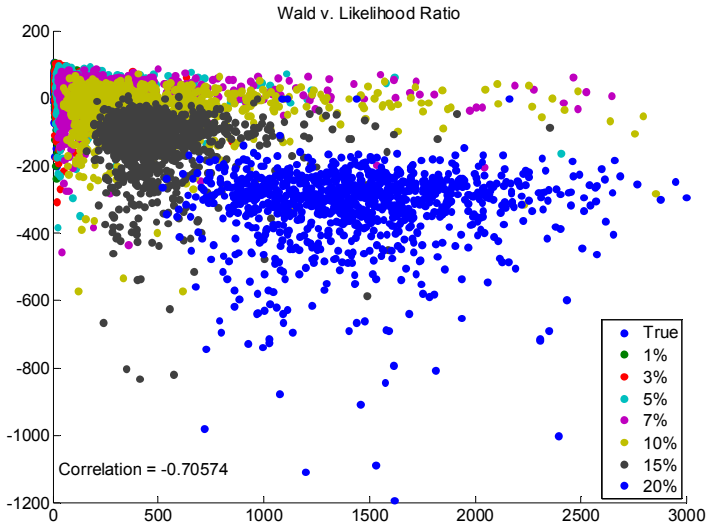


Figure 6: Scatter Plots of Indirect Inference (Wald) v. Direct Inference (LR) for True and False Models (some outliers taken out for clarity of scale)(3 Variable VAR(1))

This comparison shows, first, that there is no correlation between the two measures of 'fit' for a given model (Figure 5). When a sample is well-fitting on one measure it may be well- or badly-fitting on the other measure. Thus in effect these two tests measure entirely different things. When the model forecasts well it does not necessarily match data behaviour and vice versa.

Second, we find that the power of the Indirect Inference Wald test is considerably greater than that of the Direct Inference LR test (Table 1). With 5% mis-specification the Wald rejects 99% of the time (at the 95% confidence level) while the LR test rejects 15% of the time. At a sufficiently high degree of falseness both reject 100% of the time, so that clearly the LR test too has good power. This in turn implies that the two tests though uncorrelated across samples for a given model are reasonably correlated across models; thus as a model becomes more false both tests increase their rejection rate. This can be seen in Figure 6 that shows the same samples' measures across progressively more False models.

What this seems to show is that both methods of testing have a potential role. But they are not equivalent tests. The researcher can choose which test to use depending on the purposes in hand. If the purpose is to have a DSGE model that behaves qualitatively like the data, indirect inference is the testing tool. This purpose also imposes greater demands on a DSGE model as the test's power is very high. If the purpose is to

have a DSGE model that forecasts accurately, then direct inference is the relevant tool; it is also substantially less demanding. Thus in the first case one wishes to know whether a model has a structure that picks up 'typical' responses, even if it is not accurate in any given episode. In the second case one is using forecasting precision as a test of the model.

It seems to us that for researchers or policymakers interested in whether the DSGE model captures the causal system well, the Indirect Inference test will be relevant since it tests whether the model captures the interrelationships within the data. Such people have no interest in the model's forecasting capacity since forecasting is generally done by other means; thus the second test will not be relevant to them.

Nonetheless, one may use a direct inference test as a pure 'specification test'; that is, asking whether the model is the true generating system. As such it will perform with much less power than the comparable Indirect Inference test. Nevertheless one should be aware that in a given sample a DSGE model that was not rejected in its capacity to reflect data behaviour could be rejected in its capacity to forecast, and vice versa. For a DSGE model that is only modestly False will be rejected by two almost entirely different sets of samples under the two tests. Thus a final point to make about conducting these tests is that if a researcher were to use say two tests jointly and insist that the DSGE model passes both (reject if any reject), then the size of each test (since they are uncorrelated under the null) needs to be adjusted downwards; thus to test a model at the 5% level, each test individually must be at the $5/2\% = 2.5\%$ level, so that the chance of rejecting the true model (by one or other) is then 5%. By the same token if the researcher tests on both and only rejects if both reject, each test must be sized at a sufficiently high level that 5% of the samples lie within the joint rejection area.

3.1.1 Comparison of the tests with different VAR variable coverage and VAR order

Tests based on Indirect Inference that use high order VARs, or VARs with more than a few variables, are extremely stringent and they tend to reject uniformly. We proposed in Le et al. (2011) 'directed' Wald tests where the information used in evaluating a DSGE model was deliberately reduced to cover 'essential features' of the data; of course all Wald tests are based on chosen features of the data and therefore are always to some degree 'directed' but we use the term in practice when the Wald is focused on only a small

subset of variables or aspects of their behaviour. Table 2 shows how the Wald rejection rate varies when the order or number of variables is raised. What we see is how the power tends to rise with the number of coefficients in the VAR.

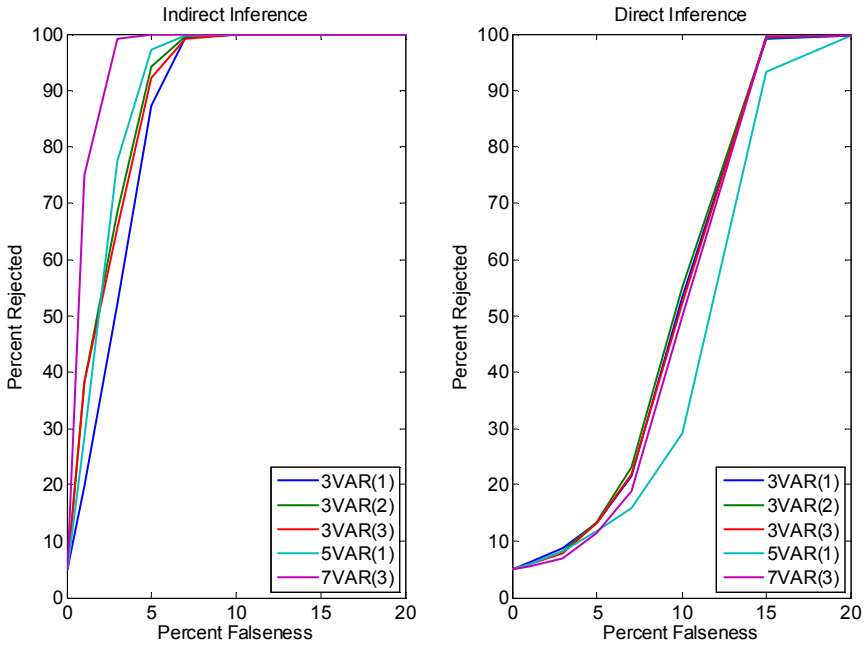


Figure 7: Power Functions for Indirect and Direct Inference for Various VARs.

VAR — no of coeffs	TRUE	1%	3%	5%	7%	10%	15%	20%
3 variable VAR(1) — 9	5.00	19.76	52.14	87.30	99.38	100.00	100.00	100.00
3 variable VAR(2) — 18	5.00	38.24	68.56	84.10	99.64	100.00	100.00	100.00
3 variable VAR(3) — 27	5.00	38.22	65.56	92.28	99.30	100.00	100.00	100.00
5 variable VAR(1) — 25	5.00	28.40	77.54	97.18	99.78	100.00	100.00	100.00
7 variable VAR(3) — 147	5.00	75.10	99.16	99.96	100.00	100.00	100.00	100.00

Table 2: Indirect Inference Rejection Rates at 95% level for varying VARs

From Table 3 we can see that using the Direct Inference Likelihood ratio test (using the LIML method on the residuals) we find that the test's power does not appear to vary in any systematic way with the benchmark VAR used, either in terms of the number of variables included or the order of the VAR. It follows that the LR test may well perform similarly, whatever the order of the VAR and the number of variables that are included. The reason for this seems to be that

a) the absolute forecasting performance of a DSGE model is not worsened by adding variables to be

VAR — no of coeffs	TRUE	1%	3%	5%	7%	10%	15%	20%
3 variable VAR(1) — 9	5.00	6.30	8.80	13.10	21.60	53.40	99.30	99.70
3 variable VAR(2) — 18	5.00	6.00	8.30	13.40	23.10	55.10	99.40	99.70
3 variable VAR(3) — 27	5.00	6.00	7.90	13.10	21.90	52.30	99.50	99.70
5 variable VAR(1) — 25	5.00	6.00	8.20	11.70	15.90	29.30	93.30	99.70
7 variable VAR(3) — 147	5.00	5.50	7.10	11.40	18.80	49.90	99.60	99.70

Table 3: Direct Inference Rejection Rates at 95% level for varying VARs

forecast — each variable in the DSGE model has its own autoregressive error that helps to keep its forecast on track

b) the relative forecasting performance and so the relative likelihood of the DSGE model is indeed worsened by going to a higher order VAR because the higher the VAR order the better the VAR forecasts. However this lowers the relative likelihood of both the True DSGE model and the False DSGE models by improving the forecasting performance of the common comparator. Hence the LR distribution is similarly altered for all DSGE models, whether True or False; and so the rejection rate (which depends on the difference in the distributions) is not much altered.

With Indirect Inference however the addition of variables or VAR detail adds to the complexity of behaviour that the DSGE model must match; the more complexity, the less well can the matching occur when the model is moderately false.

Again, this brings out the essential difference in the two measures of performance.

4 Extension to the Del Negro-Schorfheide measure of closeness, the DSGE-VAR weight

As noted earlier, Del Negro and Shorfheide have proposed as a measure of model closeness, the hyperparameter λ, which gives the maximum-likelihood weight of the model-restricted VAR when combined with the unrestricted VAR. This measure is now widely used in Bayesian analysis of DSGE models to evaluate the extent to which the model is misspecified as a whole — for a recent example, an application to the Smets-Wouters model of the US, see Del Negro et al. (2007a). It is therefore of interest to find the distribution of λ under the model null and compare its power, if used as a testing device, with the other measures considered

already, Wald and LR.

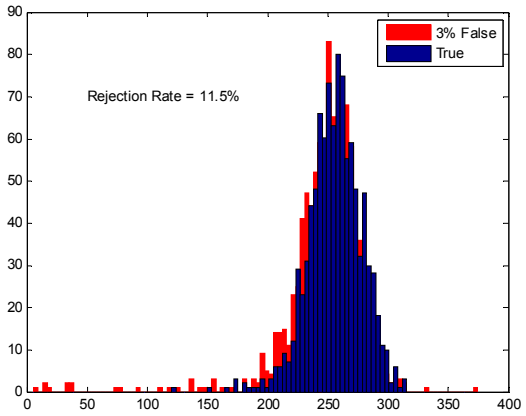
This λ hyperparameter is estimated under the model's priors in essentially the same way as the model's ordinary parameters; the posterior λ is found that maximises the likelihood. It can be done using any set of variables to be used to gauge the fit and with any VAR order wished for. This test is similar to the LR test in that it is also a forecasting test. We might therefore expect the λ test to have rather similar properties to the LR test when applied with the same VAR order and variable number. We therefore initially compare the two in what follows, using a maximum 7 variables and a high-order VAR(3).

To compare the two tests we take a true model (the same as above), generate 1000 Monte Carlo samples from a normal distribution[10] and we estimate on each sample a VAR for all variables at an order that fits best and avoids overfitting; we experiment with a few samples to check what this is — and find in fact that it is a VAR(3). For each sample we find the DSGE-VAR weight for the True Model's parameters (θ, ρ and standard deviations of the errors). The distribution of these 1000 weights is the distribution under the null from which we obtain the 5% critical value. We now evaluate the power against False Models,where the θ, ρ and error standard deviations are false by $+/-x\%$. Thus our procedure mirrors that under direct inference, to which it is essentially similar in that we are asking what combination of the data-based VAR and the DSGE-restricted VAR gives the lowest errors in current period forecasts. Table 4 shows how the power of the two tests differs; the distributions for the True and 3 and 7% False are shown in Figure 8. Here the model we use is the same as in the last section (the SW model) but we make the error distributions normal with no skewness or excess kurtosis as assumed in DYNARE which we use to do this calculation.

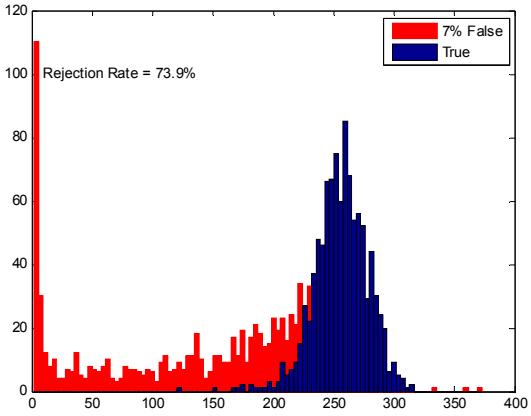
Percent Mis-specified	Direct Inference	Lambda
True	5.0	5.0
1	4.0	5.6
3	4.4	11.5
5	10.3	35.8
7	30.8	73.9
10	89.7	99.8
15	100	100
20	100	100

Table 4: Rejection Rates for Direct Inference and DSGE-VAR Weight, 7 Variable VAR(3)

[10] The usual Bayesian assumption is that errors are normal and hence this assumption is embodied in the null hypothesis for the λ test. The normality restriction on the null is unnecessary for the other two tests.



Histograms of true and 3% false models for λ



Histograms of true and 7% false models for λ

Figure 8: Histograms of true and false models for Lambda

It can be seen that the power of this weight-derived test is rather higher than that of the Likelihood Ratio/direct inference test. There appears to be a small positive correlation between λ and LR for a given model and a strong positive correlation between them across models of differing falseness (Figures 9 and 10). So clearly these two tests are related for a given model; across models they are clearly again strongly related. The Direct LR measures current-period forecasting success and λ also measures forecasting success though in a somewhat different way from Direct.

The question then arises how the λ test varies in power as the number of variables and the VAR order vary. We have seen above that the LR test is essentially insensitive to these variations while the Wald Indirect Inference test is highly sensitive, tending to greater rejection rates as either of these two rise. For the λ test, much like the LR test, this is hardly true; the power of the test rises very little as one moves from the minimum 3VAR(1) to the maximum 7VAR(3). Again this shows that the λ test is fairly close to the LR test in being another forecasting test.

Percent Mis-specified	λ 3VAR(1)	λ 7VAR(3)
True	5.0	5.0
1	6.6	5.6
3	11.7	11.5
5	28.5	35.8
7	63.0	73.9
10	97.9	99.8
15	100	100
20	100	100

Table 5: Rejection Rates for DSGE-VAR Weight, 3 Variable VAR(1) and 7 Variable VAR(3)

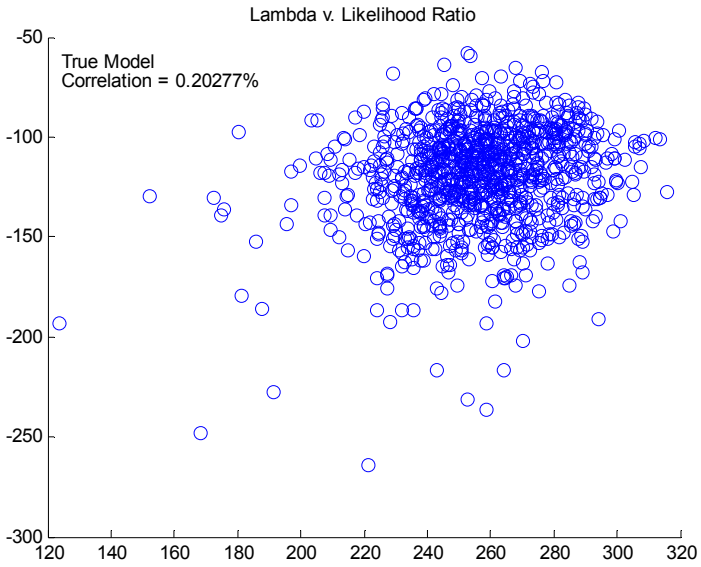


Figure 9: Correlation of Lambda and Likelihood Ratio

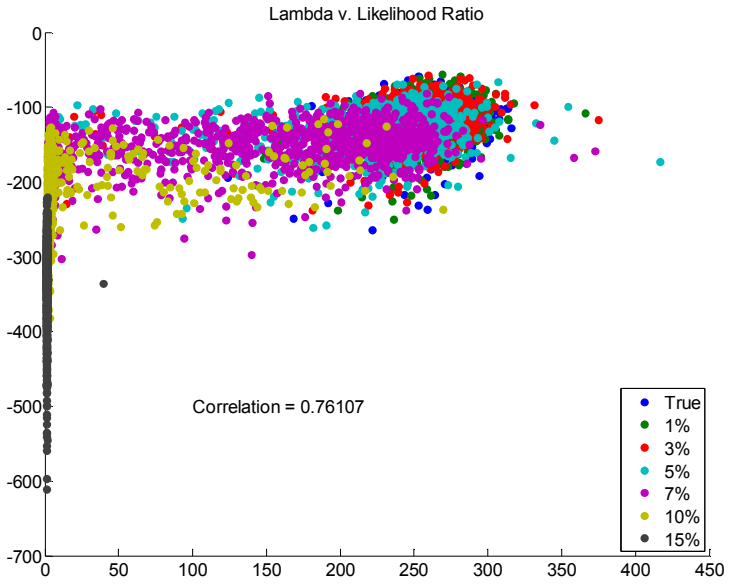


Figure 10: Scatter plot of Likelihood Ratio and Lambda for all models

4.1 An example of test comparison: the SW model post-1984 as re-estimated by Le et al. (2011)

We now examine[11] whether we can use the LR test on the SW model post-1984 that Le et al. (2011) found, after re-estimation by Indirect Inference, passed the Indirect Inference test comfortably. That test used a VAR(1) on three variables — output, inflation and interest rates — as the auxiliary model. With a higher order VAR on these 3 variables, as well as a VAR(1) on more than these three, the model did progressively worse, being rejected on most. Le et al. interpreted this to mean that the model is able to capture the 'truth' about the joint behaviour of these 3 variables but not with fine accuracy; thus it captures the 'broad outlines' of key macro behaviour. We would like to know whether this same model would pass these two other tests.

We choose a VAR(1) on 3 variables as the benchmark for the Likelihood Ratio; as we have seen the power does not vary with the order of the VAR or the number of variables included. The test gives similar results to the VAR(1), which is to pass about as well as it does on the Indirect inference test. In both these cases our procedure is to generate bootstrap data from the model's errors to obtain the distribution of the tests for this model under the null that it is true, to obtain the critical value.

We may note that with more variables and higher VAR orders, the results worsen for Indirect Inference but remain similar for the Direct Inference LR test — Table 6. Out of the two tests, the LR fails to reject at all, while the Wald rejects for any VAR with more that 18 coefficients.

VAR — no. of coefficients	Wald$^+$	LR
3 variable VAR(1) — 9	83.5	71.7
3 variable VAR(2) — 18	99.6	71.4
3 variable VAR(3) — 27	100	67.7
4 variable VAR(1) — 16	90.1	82.8
5 variable VAR(1) — 25	96.6	74.2
7 variable VAR(3) — 147	100	13.4

$^+$For the Wald the test includes the variances of the data in each case

Table 6: Tests using varying VARs

[11] As noted in the last footnote, the usual Bayesian assumption that the errors are normal implies that the λ test cannot be used for a model as here where the errors are skewed and have excess kurtosis. To use it the Bayesian process must be adapted to allow for these higher moments being non-normal; Dynare, which we are using to solve our models, is not adapted for this. We are working on solutions for this problem. We may conjecture, based on the work in the last section, that the λ test would give similar results to the LR test.

We can see that for our main focus on three variables with a VAR(1) both tests give consistent results. We also notice that the LR is totally insensitive to the benchmark VAR used, as discussed above..

5 Conclusions

In comparing these tests we have found that they measure the performance of a model in different ways: two of them, LR and λ, measure its forecasting performance, and one, the Indirect Inference Wald, measures its ability to replicate descriptors of the data. While all three tests are correlated across models of increasing falseness, and the LR is moderately correlated with the λ test for a given model, the Wald measure is quite uncorrelated with the LR test (and therefore also with λ) for a given model; thus the Wald and the other two tests are measuring quite different things.

The power of the Indirect Inference tests depends on the benchmark comparator used. As the number of variables in, and the order of, the VAR comparator are raised the power increases. This is because as the benchmark becomes more demanding the true model manages to match the VAR coefficients as well as before; but False models fail by a greater extent than before because the accuracy demanded is greater. We choose a level of power suitable for our purpose as users. Typically a user of a macro model wants a model to replicate the behaviour of three key macro variables: output, inflation and interest rates. Furthermore, since the model is a guide to main inter-relationships, a VAR(1) on these three giving 9 coefficients, plus the variances to check 'size', 12 coefficients in all, will be sufficiently accurate a descriptive set. This is somewhat like requiring a photograph to have a certain minimum resolution (the pixel density) for it to represent reality sufficiently well for our needs; for our macro models we require a moderate density to be satisfied. Thus 'Truth' or 'realism' is defined by the user practically in relation to the user's needs.

The power of the Direct Inference test, the LR test, does not depend on the benchmark comparator. This is because as the benchmark becomes better at forecasting (with higher order), the true model fails relative to it by about the same extent as the false models. Similarly, when the number of variables being forecast is raised, the relative forecast performance of the model and the VAR alters by a similar extent for both the True and the False models. To a reasonable degree the same statements apply to the λ test, with the proviso that when a maximum-size VAR is used, the power of this test is much larger.

For the single example we have investigated here — the SW model for post-1984 — the Direct and Indirect Inference tests give broadly consistent results in assessing the accuracy of this model for the key three variables. However, as the VAR complexity is raised the tests differ: LR fails to reject however great the complexity, while the Wald test rejects to an increasing degree as the complexity rises..

These results seem to suggest that the reason why early Rational Expectations models were being rejected by LR tests (as in the quotation from Sargent in Evans and Honkapohja, 2005, in footnote 1) was probably that these models did not have the sort of dynamic content that enabled them to meet these stringent forecasting tests nor did they have lavished on them the re-estimation resources that we have used here. Certainly the SW model does pass these tests for the post-1984 sample when re-estimated as above by indirect methods.

References

[1] Basawa, I.V., Mallik, A.K., McCormick, W.P., Reeves, J.H., Taylor, R.L., 1991. Bootstrapping unstable first-order autoregressive processes. Annals of Statistics 19, 1098–1101.

[2] Canova, F., 1994. Statistical Inference in Calibrated Models. Journal of Applied Econometrics 9, S123–144.

[3] Canova, F., 1995. Sensitivity Analysis and Model Evaluation in Dynamic Stochastic General Equilibrium Models. International Economic Review 36, 477–501.

[4] Canova, F., 2005. Methods for Applied Macroeconomic Research, Princeton University Press, Princeton.

[5] Canova, F., Sala, L., 2009. Back to square one: Identification issues in DSGE models. Journal of Monetary Economics 56, 431–449.

[6] Chari, V., Kehoe, P. J., McGrattan, E., 2002. Can Sticky Price Models Generate Volatile and Persistent Real Exchange Rates? Review of Economic Studies 69, 533–563.

[7] Christiano, L. 2007. Comment on 'On the fit of new Keynesian models'. Journal of Business and Economic Statistics 25,143–151.

[8] Dave.C., De Jong, D.N., 2007. Structural Macroeconomics. Princeton University Press.

[9] Davidson, J., Meenagh, D., Minford, P., Wickens, M. 2010. Why crises happen — nonstationary macroeconomics. Cardiff Economics Working Papers E2010/3, Cardiff University.

[10] Del Negro, M., Schorfheide, F., 2004. Priors from General Equilibrium Models for VARs. International Economic Review, 45, 643–673.

[11] Del Negro, M., Schorfheide, F., 2006. How good is what you've got? DSGE-VAR as a toolkit for evaluating DSGE models. Economic Review, Federal Reserve Bank of Atlanta, issue Q2, 21–37.

[12] Del Negro, M., Schorfheide, F., Smets, F., Wouters, R., 2007a. On the fit of new Keynesian models. Journal of Business and Economic Statistics 25,123–143.

[13] Del Negro, M., Schorfheide, F., Smets, F., Wouters, R., 2007b. Rejoinder to Comments on 'On the fit of new Keynesian models'. Journal of Business and Economic Statistics 25,159–162.

[14] Evans, R. and Honkapohja, S. (2005) 'Interview with Thomas J. Sargent', Macroeconomic Dynamics, 9, 2005, 561–583.

[15] Faust, J., 2007. Comment on 'On the fit of new Keynesian models. Journal of Business and Economic Statistics 25,154–156.

[16] Friedman, M. 1953. The methodology of positive economics, in Essays in Positive Economics, Chicago: University of Chicago Press.

[17] Gallant, A.R., 2007. Comment on 'On the fit of new Keynesian models'. Journal of Business and Economic Statistics 25,151–152.

[18] Gourieroux, C., Monfort, A., 1995. Simulation Based Econometric Methods. CORE Lectures Series, Louvain-la-Neuve.

[19] Gregory, A., Smith, G., 1991. Calibration as testing: Inference in simulated macro models. Journal of Business and Economic Statistics 9, 293–303.

[20] Gregory, A., Smith, G., 1993. Calibration in macroeconomics, in: Maddala, G. (Ed.), Handbook of Statistics vol. 11, Elsevier, St. Louis, Mo., pp. 703–719.

[21] Hansen, B.E., 1999. The Grid Bootstrap And The Autoregressive Model. The Review of Economics and Statistics 81, 594–607.

[22] Horowitz, J.L., 2001a. The bootstrap, in: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, vol.5, ch. 52, 3159–3228, Elsevier.

[23] Horowitz, J.L., 2001b. The Bootstrap and Hypothesis Tests in Econometrics. Journal of Econometrics 100, 37–40.

[24] Kilian, L., 2007. Comment on 'On the fit of new Keynesian models'. Journal of Business and Economic Statistics 25,156–159.

[25] Le, V.P.M., Meenagh, D., Minford, P., Wickens, M., 2011. How much nominal rigidity is there in the US economy — testing a New Keynesian model using indirect inference. Journal of Economic Dynamics and Control 35(12), 2078–2104.

[26] Le, V.P.M., Minford, P., Wickens, M., 2010. The 'Puzzles' methodology: en route to indirect inference? Economic Modelling 27, 1417–1428.

[27] Le, V.P.M., Minford, P., Wickens, M., 2012. A Monte Carlo procedure for checking identification in DSGE models. Unpublished manuscript, Cardiff University. (http://patrickminford.net/Academic_Page/Identification.pdf)

[28] Lucas, R.E.,1976, Econometric policy evaluation: A critique, Carnegie Rochester Conference Series on Public Policy No. 1, The Phillips Curve and Labour markets, K. Brunner and A. Meltzer, eds., supplement to Journal of Monetary Economics.

[29] Sims, C.A.,1980. Macroeconomics and reality. Econometrica, 48, 1-48.

[30] Sims, C.A., 2007. Comment on 'On the fit of new Keynesian models', Journal of Business and Economic Statistics 25,152–154.

[31] Smets, F., Wouters, R., 2007. Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach. American Economic Review 97, 586–606.

[32] Smith, A., 1993. Estimating nonlinear time-series models using simulated vector autoregressions. Journal of Applied Econometrics 8, S63–S84.

[33] Watson, M., 1993. Measures of fit for calibrated models. Journal of Political Economy 101, 1011–1041.

6 Appendix 1

6.1 The bootstrap variance-covariance matrix

In our Wald statistic we used the variance-covariance matrix implied by our model bootstrap samples. An alternative procedure would be to use the variance-covariance matrix implied by the VAR based on the data; under the null this would correspond to the true variance-covariance matrix implied by the model. If, however, the model is false then this variance-covariance matrix would not correspond to the true variance-covariance matrix. The bootstrapped DSGE variance-covariance matrix corresponds to the true matrix, as restricted by the model, whether the model is true or false. When the model is false, therefore, we conjecture that the null is more likely to be rejected if we use the restricted variance-covariance matrix than if we use the data variance-covariance matrix. Effectively the Wald statistic would not be based on the behaviour of the model being tested except in respect of the bootstrap mean. Yet we have seen that the model restricts the distribution of the VAR coefficients in very definite ways, with a joint distribution 'footprint'. It will be easier for a false model to get closer to the data if it is allowed to have a footprint from the true model whose restrictions will be different.[12]

To investigate the relative power of the test using the two different covariance matrices we set up a Monte

[12] Indeed we found this to be the case with the SW model here. This was rejected by the data using the model-based covariance matrix and can be regarded as False. In the data-based matrix the off-diagonal elements are small, so that the test becomes close to the Wald based only on the no-covariance case, whereas in the model-based covariance matrix the off-diagonal elements are often rather large, indicating the way in which the model produces restrictions between the implied VAR coefficients. Had we used the data-based covariance matrix the model would have got much closer to passing. This mirrors our basic findings in this and related papers: that DSGE models are rejected because the data does not reflect the fine detail imposed by the model's restrictions. Thus, specifically, the data-based variance-covariance matrix mirrors in its small off-diagonal elements the lack of tight restrictions between VAR coefficients in the data, whereas the model-restricted variance-covariance matrix with its large off-diagonal elements reflects the tight restrictions imposed between the VAR coefficients by the DSGE model (for an illustration of this see the discussion above of Figure 1).

Carlo experiment in which we treated the SW model with unity NK weight. We then test False models where Falseness is created by randomly increasing even parameters in θ and reducing odd parameters by $x\%$ where x goes from 1-20. To ensure that the errors are consistent with these False models given the data, we recompute them from the data using the False parameters.[13]

What we find is that the power of the test is reduced substantially in the range of 1-10% mis-specification. By 20% both covariance matrices are rejecting at 100% because the bootstrapped VAR coefficients are so far away from the true model coefficients that the subtlety of the covariances implied by the model is redundant.

Rejection rates using 5% tail	General Mis-specification				
	1%	3%	5%	7%	20%
DSGE Model Covariance	6.60	8.78	18.07	35.98	100.00
VAR-from-data Covariance	4.07	4.50	5.91	8.53	100.00

Notes: the errors and their ρ(AR) parameters are extracted from the data using LIML

Table 7: Power Function for 95%

6.2 Extracting the model residuals from the data

We find that there is little change in the results when using structural errors obtained from the original data using our Limited Information estimation method, instead of errors derived from SW's Bayesian estimates. This finding is reassuring as it suggests that our test procedure is robust to different ways of extracting the residuals from the data. It also echoes the finding of Le et al. (2010) in their investigation of the results of Chari et al. (2002). We now investigate this issue using a Monte Carlo experiment. We set up a True Model and endow it with some True Errors. We then simulate it once to obtain a data set which we treat as our sample data and compare two ways of extracting residuals from this data. The differences in the errors arise from the way in which expectations are generated. For equations containing no expectations and given the model parameters, both methods give the same as the residual is simply backed out of the data.

One way to extract the residuals of equations with expectations is to find a set of residuals such that

[13] We do this as follows. The True model is endowed with some True errors. One simulated data set is assumed to be the 'data' for our experiment. We extract from this the LIML estimates of the errrors under the True model parameters. We then test the True Model with these estimated errors (and corresponding ρs,timeseries parameters of the errors); thus we treat this as the True Model for the test. We generate 1000 data sets from this True Model. The test based on the bootstrap distribution of the errors will have approximately the right size: thus for example this model will be rejected 5% of the time at the 95% significance level. The $x\%$ False Model is then specified; the same data is used to extract its implied LIML errors. These are bootstrapped to provide the test.

the rational expectations they imply in turn generate the same residuals.This iterative method will give the exact residuals implied by the model if it is given the true ρs since it will use the exact rational expectations implied by the model and the data. When it does not have the true ρs then it has to estimate them and this creates a potential error in both the ρs and the residuals. We have created an iterative algorithm to search for this set of residuals, using OLS to estimate the ρs from the residuals — we call this the Exact Method, even though it is of course inexact in practice when the ρs must be estimated.

Another way to extract the residuals, and one predominantly used in our work both because of its computational speed and because the residuals it produced were invariably well behaved, is to project the expectations using LIML; specifically we use the predictions of the data-generated VAR as the expectations. We call this the LIML Method.

Applying these two methods to the data for our full US sample and the Smets-Wouters model, it can be seen that the two sets of residuals are very close together; with correlations (marked on the graphs) of $0.82 - 0.91$**.** Nevertheless, the ρs they each imply are in several cases startlingly different.

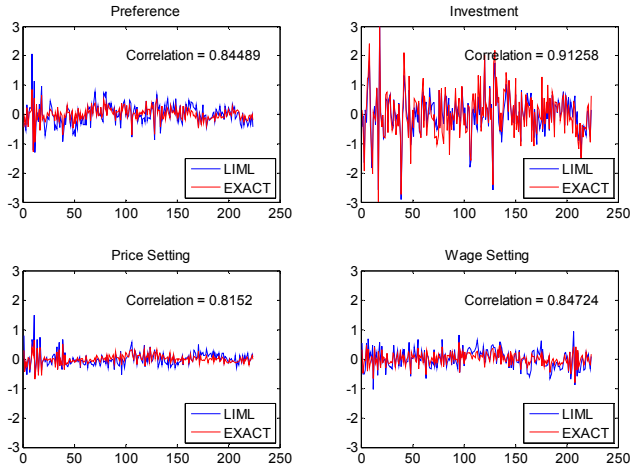


Figure 11: Comparison of LIML and EXACT Residuals

These two sets of extracted residuals can be compared with the True residuals used in a Monte Carlo generated sample where we know the True errors by construction. Again we see that not only are the two extracted sets very close but both are very close to the True errors. It might seem that the True residuals and

	LIML	EXACT
ρ^{gov}	0.9452	0.9445
ρ^{pref}	−0.0658	0.6155
ρ^{inv}	0.5282	0.2506
ρ^{mon}	0.3682	0.3680
ρ^{prod}	0.9750	0.9718
ρ^{price}	0.2052	0.6664
ρ^{wage}	0.1195	0.6283
$\rho^{gov,prod}$	0.6341	0.4831

Table 8: Comparison of LIML and EXACT rhos

the estimated ones must give the same results. However, these small differences create non-trivial differences between the true and the estimated ρs. These in turn make the model behave differently.

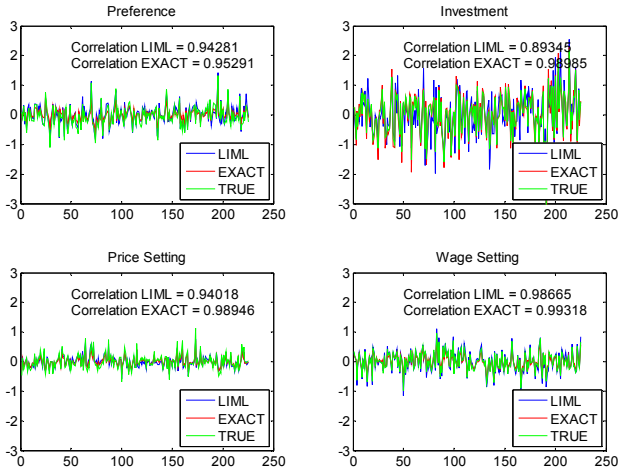


Figure 12: Comparison of LIML and EXACT residuals with True residuals

Comparing the LIML and the Exact Walds we get from around 700 random samples of this Monte Carlo experiment, we find that, on average, they are fairly similar: the average Wald differs by only 15%, indicating that the estimated ρ and η are on average similar for the LIML and the Exact Wald statistics. Somewhat surprisingly, the Exact Wald statistic is less accurate than the LIML Wald statistic when compared with the true errors. The correlation between the two across these samples is 0.61. What this reveals is that while the two methods do not give identical results in all samples, they do give quite similar ones. Thus which method is used to extract the residuals may not affect the initial test of the model. Nevertheless, as we have explained in the last section, this initial test is merely the starting point for testing the model and the

	TRUE	LIML	EXACT
ρ^{gov}	0.9452	0.9407	0.9289
ρ^{pref}	−0.0658	0.0593	0.4163
ρ^{inv}	0.5282	0.6509	0.4465
ρ^{mon}	0.3682	0.3557	0.3500
ρ^{prod}	0.9750	0.9247	0.9520
ρ^{price}	0.2052	0.2539	0.5557
ρ^{wage}	0.1195	0.1236	0.3561
$\rho^{gov,prod}$	0.6341	0.0812	0.4741

Table 9: Comparison of LIML and EXACT rhos with True rhos

ultimate set of parameters that are found to result in the least rejection rate for the model should not be affected by this starting point.

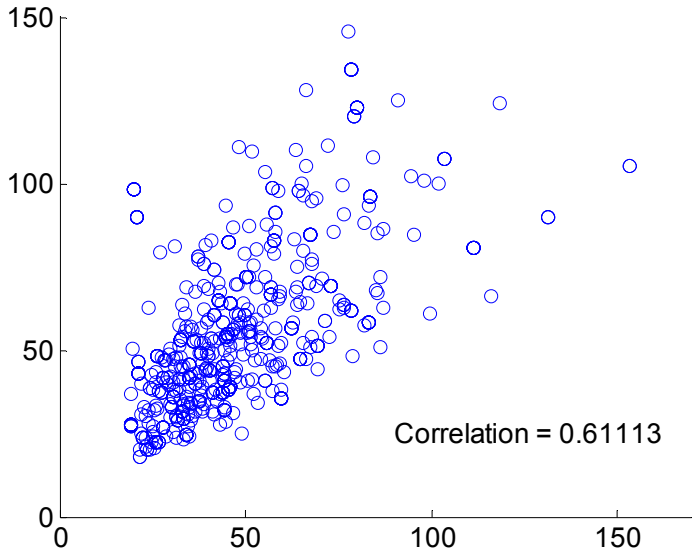


Figure 13: Scatter Plot of LIML versus EXACT Walds

It may help at this point if we recall what our procedure entails. We wish to test some model on an actual sample (we compare the $\widehat{\alpha}$ from the sample with its distribution under the null hypothesis of the model). We create a True Model with vectors of structural parameters θ, of autoregressive error parameters ρ, and of innovations η with some true distribution: the Model is thus described as (θ, ρ, η). We generate the potential sample distribution of this True Model's $\widehat{\alpha}$ in a Monte Carlo experiment. When we test whether a given data sample's $\widehat{\alpha}$ comes from a True Model which is determined by (θ, ρ, η), we choose these parameters

according to our 'best estimates/calibration'.[14]

Thus to repeat what we set down in the previous section, our procedure sets up a unique model as the null (True) hypothesis and tests it on the given sample. The ρ, η used in this are constrained to be consistent with the actual data; thus we cannot choose values that are directly rejected by the sample. In this respect we depart from some other methods of testing simulated models in that we are not at liberty to assume errors and autodependence that are directly rejected by the data when fitted to the DSGE model. As we have noted the DSGE model forces on the data a division between the part that is due to θ, the causal structure, and the unexplained part that is exogenous, the residuals. In testing the model we cannot ignore the nature of this unexplained part since this is dictated by the data when confronted with the model.

We have seen that the two methods of error extraction for given θ give similar values for Wald statistic for given samples generated by a model with that θ and some chosen ρ, η. The power of the test of the True supposed model is quite insensitive to the choice of residual extraction method. We therefore take one of our Monte Carlo samples and set up the True Model to have the estimated ρs and innovations found by the LIML method; we generate 1000 Monte Carlo samples of this model, which by construct will be correctly rejected at a nominal 5% level. Now to test the False models, we fit each to the same actual sample using LIML to extract the residuals and ρs, and check the rejection rate on the 1000 samples. For the Exact method we follow the same procedure, with Exact in place of LIML.[15] It can be seen that the two methods of extraction have very similar power. This exercise shows is, for a given actual sample from which the innovation and ρ estimates are extracted, both the substantial power of the test and its insensitivity to the method of error extraction.

Finally, as expkained in the previous sction, we note that our initial null hypothesis (θ, ρ, η) is just a starting point for the whole testing process for the model. Our initial test simply discovers whether the model can be rejected for this particular set of parameters and innovations. We then proceed, using

[14] This should be distinguished from testing some unknown True Model with estimated parameters drawn from the distribution of these parameters associated with this unknown True Model. As explained in the last section, such a test gives rise to a far more dispersed Wald distribution than the one for a model where the parameters are known and fixed. In particular, if we regard the ρ (and hence also η) as unknown, then while the distribution is not so dispersed as the one where all the parameters are regarded as unknown, it is still substantially more widely-distributed than the one where the parameters are known.

[15] Notice that this exercise differs from the power exercise in the section above on parameter uncertainty. There the False models were created by varying both θ and ρ parameters by the $+/-x\%$ and the true η were kept unchanged. Here we vary θ similarly and extract the implied ρ, η. The former method uses η which are not necessarily compatible with the sample and hence tend to generate slightly higher rejection rates. But the difference is not great.

	Percent General Mis-specification						
	0%	3%	5%	7%	10%	15%	20%
LIML	5	7.18	13.71	31.02	78.12	99.97	100
EXACT	5	6.76	13.36	32.00	76.30	99.86	100

Table 10: Power and size of tests using estimated residuals: Percent Rejection at 95% Confidence

simulated annealing, to find the set of parameters and innovations that fail to be rejected by the sample data, if such exists. If we find a set that we cannot reject as the True parameter set, we can then extract from this parameter set and the sample the Exact innovations implied; if the parameters are the True parameters, then these will be the True innovations. We may then infer that the test still fails to reject when bootstrapping these innovations with these parameters.[16] Note that, wherever we start the process from, the SA algorithm should converge on the same end result.

7 Appendix 2

7.1 The accuracy of the bootstrap

Le et al. (2011) reported extensive tests of the accuracy of the bootstrap under Indirect Inference. In a Monte Carlo experiment, they set up the model they chose as best for the post-1984 sample, with the errors implied by that sample and they gave these a normal distribution with the same variance as estimated. They then drew random samples from the innovations in these error processes, creating 1000 artificial samples of the same length as the original data — 76 observations, a small sample. They then bootstrapped each of these samples 1000 times. For each sample they computed the Wald statistic generated by the bootstraps to check whether the model is accepted or rejected at various confidence levels. The results in Table 11 indicate that the procedure is fairly accurate, with a slight under rejection.

We replicated these results on the model used in our work here — viz the SW model as modified for Le et al. (2011) and obtained very similar results (Table 12).

[16] The Exact and the LIML estimates of ρ, η are neither of them more or less accurate when the model fitted is not in fact the true model; nor when the model has the true θ parameters but we do not impose the true ρ. In fact both provide inaccurate estimates of ρ, η in these circumstances. However, if we know the True θ, ρ the the Exact method will generate the exactly correct innovations for the sample. Hence once we are in the vicinity of the True θ, ρ we must use the Exact method to extract the innovations.

Nevertheless, at the start of the process when Simulated Annealing search has not yet taken place, neither extraction method gives clearly superior performance; LIML is probably at this stage easiest to use from a computational viewpoint.

Nominal RP(%):	10	5	1
True RP (%)	5.1	2.6	0.4

Table 11: True versus nominal Rejection Probabilities for Indirect Inference (n=76; Montecarlo experiment; post-1984 model and sample)

Nominal RP(%):	10	5	1
True RP (%)	4.3	2.6	0.4

Table 12: True versus nominal Rejection Probabilities for Indirect Inference (n=225; Montecarlo experiment; post-war model and sample)

With the Likelihood Ratio we test the bootstrap's accuracy by taking the same model and generating 1000 samples by Monte Carlo. For each one, we generate the bootstrap distribution of the LR (thus we bootstrap that sample's innovations to obtain 1000 data sets; and for each we generate the VAR forecast errors as well as the model's forecast errors, using LIML residuals extracted from the data as in our usual procedure). Then we check whether the true sample LR is rejected on the bootstrap distribution. Over the 1000 true samples we obtain the bootstrap rejection rate (Table 13).

Nominal RP(%):	10	5	1
True RP (%)	8.4	5.1	2.6

Table 13: True versus nominal Rejection Probabilities for Likelihood Ratio (n=225; Montecarlo experiment; post-war model and sample)

We do not use the bootstrap for the λ tests (the distribution of λ is generated by the Bayesian stochastic simulation routine) and so its accuracy does not arise.