Cardiff Economics Working Papers

Working Paper No. E2016/14

What is the truth about DSGE models? Testing by indirect inference

David Meenagh, Patrick Minford, Michael Wickens and Yongdeng Xu

December 2016

ISSN 1749-6010

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

What is the truth about DSGE models? Testing by indirect inference

David Meenagh *
Cardiff University, UK

Patrick Minford†
Cardiff University, UK and CEPR

Michael Wickens‡
Cardiff University, University of York and CEPR

Yongdeng Xu§
Cardiff University, UK

December 2016

Abstract

This paper addresses the growing gulf between traditional macroeconometrics and the increasingly dominant preference among macroeconomists to use DSGE models and to estimate them using Bayesian estimation with strong priors, but not to test them as they are likely to fail conventional statistical tests. This is in conflict with the high scientific ideals with which DSGE models were first invested in their aim of finding true models of the macroeconomy. As macro models are in reality only approximate representations of the economy, we argue that a pseudo-true inferential framework should be used to test macro models, especially DSGE models. We find that a Wald test has much greater power than a Likelihood Ratio test when used in indirect inference. We suggest that the power function can be used to provide a measure of the robustness of DSGE models.

Keywords: Pseudo-true inference, DSGE models, Indirect Inference; Wald tests, Likelihood Ratio tests; robustness

JEL classification: C12, C32, C52, E1

*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, Colum Drive, Cardiff, CF10 3EU, UK.

‡Cardiff Business School, Cardiff University, Colum Drive, Cardiff, CF10 3EU, UK.

§Xuy16@cardiff.ac.uk; Cardiff Business School, Cardiff University, Colum Drive, Cardiff, CF10 3EU, UK.

1 An inferential framework for testing DSGE models

Much of classical econometrics was developed with a view to estimating and testing macroeconomic models. It was an integral component of the aim of making economics a science. But since the advent of DSGE macroeconomic modelling formal tests are rarely carried out. Calibration and Bayesian estimation of DSGE models have replaced the classical estimation of traditional macroeconometric models, arguably therefore undermining the status of macroeconomics as a science.

There is an irony in this as the impetus behind the use of DSGE rather than traditional macroeconomic models was Lucas's critique that the latter were essentially reduced form and not structural models and therefore likely to be structurally unstable and unsuited to scientific testing. Moreover, using best-fit classical time series methods of estimation of macroeconometric models with their flexible dynamics has come to be viewed as data-mining and to have undermined the credibility of tests of these models.

DSGE modelling is, however, not without its problems. Lucas and Prescott soon found that when tested using classical methods, DSGE models were invariably rejected. They therefore proposed the use of calibration rather than classical estimation, and their tests consisted of an informal comparison of moments simulated from the calibrated model with those observed in actual data, rather than formal statistical tests.

Bayesian estimation is now widely used instead of calibration or classical estimation. Its attraction is that it is a compromise between using strong priors, as in calibration, and diffuse priors, which would give the same result as classical estimation. In practice, however, the Bayesian posterior estimates are often found to be little different from their prior values but considerably different from their classical estimates, thereby providing prima facie evidence that the prior beliefs are not supported by the data, and that the model may be misspecified. If the mode of the posterior distribution is used as the point estimate then Bayesian estimation is, in effect, a weighted average of the prior values and the maximum likelihood estimates, where the weights are inversely proportional to the strength of the prior beliefs and the precision of the maximum likelihood estimates. The stronger the prior, therefore, the more likely that the posterior estimates will be close to their prior values, and the more like calibration would Bayesian estimation become.

The focus in DSGE models on structural modelling (estimating deep structural rather than "reduced-form" parameters) has resulted in the models being smaller and simpler than traditional macroeconometric models, especially in their dynamic specification. Friedman (1953) regarded the use of simple rather than complicated models as an advantage, but it makes it more likely that DSGE models are misspecified. This is one reason why DSGE models fit the data less well and are frequently rejected. DSGE models often have highly serially correlated structural disturbances, which is a strong signal of potential misspecification. In order to better match actual data, the structural disturbances of DSGE models are commonly fitted with autocorrelated errors. The rejection of DSGE models using conventional testing procedures, the arbitrary weight given to prior distributions, and the practice of modelling dynamic misspecification by highly serially correlated structural disturbances, all undermine the high ideals originally envisaged for the DSGE approach to macroeconomic modelling.

These arguments reveal a fundamental methodological divide between traditional macroeconometric modelling and DSGE macro modelling. Traditional macroeconometric models are not structural but, due to the flexibility this allows, particularly in their dynamic specification, they can be specified in such a way that they pass statistical tests, whereas DSGE models are structural, deliberately simple and, because they are usually rejected using classical inference, strong prior restrictions are imposed in their estimation. DSGE models are also a useful theoretical policy tool and have become the workhorse of modern macroeconomics.

Rather than dismiss DSGE models as 'incredible', as some have done, or accept that there is no point in testing them because they would fail the test, it would be better to find a way of putting them on firmer statistical foundations. In addition to devising suitable tests, and because, being deliberate simplications of reality, all macroeconomic models are "false"

- both DSGE models and conventional macroeconometric models - we might, nonetheless, wish to know the "extent of their falseness" in order to be able to judge how useful they might still be. This has been expressed by the question "how true is your false model?" In order to answer this question we require an inferential framework that reflects the degree of falsity of macroeconomic models. Traditional statistical tests adopt the null hypothesis that a theory is true; the power of a test is the probability of rejecting the theory if it is false. This framework does not fit easily if one starts from the premise that the theory is false and we seek to find how true or false it is. However an alternative to the traditional approach is the null hypothesis that a model is "pseudo-true". The idea, which was developed from testing non-nested hypotheses - Cox (1961, 1962) - is to test an approximation to the "true" - if it exists - but unknown, and probably highly complex, model using the estimates of the parameters of the approximating model. (If estimated by maximum likelihood these are called quasi-maximum likelihood estimates.) In other words, we may treat DSGE models as deliberately simplified representations or approximations of the economy for which it is appropriate to apply a pseudo-true inferential framework rather than classical statistical inference. The same argument can be applied to traditional macroeconometric models. The difference is that instead of testing DSGE models directly we will use indirect inference.

This has been the focus of work of ours and coauthors in the past decade and a half in which we have used indirect inference to test prominent DSGE models estimated by others but not tested by them. Indirect inference can be seen as an example of pseudo-true inference. It involves approximating the DSGE model by an auxiliary model based on its solution, and conducting inference on this. This auxiliary model will also be a pseudo-true representation of the economy. The idea is to simulate the DSGE model and to base a test of the model on a formal comparison of estimates of the auxiliary model derived from the simulated and actual data. This is, in effect, a generalisation and formalisation of the original method used to judge the performance of calibrated DSGE models through a comparison of the moments of simulated and actual data.

This idea is, in effect, a return to Friedman's (Friedman, 1953) 'as if' methodology in which a model is treated as if it is true and is tested on that basis even though it is known to be strictly and literally untrue. The 'as if true' assumption asserts that the model has a data generating mechanism that is a close approximation to the true model, so close that statistical testing will not be able to distinguish between the two. Such a model is 'pseudo-true' in our use of Cox's definition. Strictly, if the structural model is untrue, then so is its reduced form; however, both are hypothesised to be pseudo-true.

It is helpful to illustrate these ideas using Friedman's own favourite example of a pseudo-true model: perfect competition. This, Friedman says without fear of contradiction, cannot truly exist any more than the speed of a falling object can be accurately calculated as if it is in a vacuum- the 'gravity model'. But, Friedman goes on, perfect competition is an excellent model of a highly competitive market. Aspects that are poorly modelled, due to the frictions created by such things as temporary monopoly rents, can be replaced by error terms. These can be modelled as univariate time-series processes which may be autocorrelated because such frictions may persist for some time. Thus the structural model would consist of the systematic demand and supply equations - first order conditions - and the market-clearing condition, together with the structural errors; while the reduced form model can be obtained, for example, as a VAR solution of the structural model with its own reduced form error processes derived from the structural error processes. We assert that the DGP of each model is a close approximation to the true DGP of the corresponding structural and reduced form model: they are both 'pseudo-true'. As we cannot know what the true models are, in practice we cannot check whether any candidate model is true. But we can use normal statistical methods to test whether any candidate pseudo-true model has a DGP that conforms to the actual data where the test is defined in terms of properties of the data relevant to the user. If it passes our test at the chosen confidence level then we treat it as pseudo-true and hence as if it is true.

We show in this paper that a Wald test (the IIW test) that focuses on the parameters of the auxiliary model performs better than a maximum likelihood test which is, in effect, based on predictions from the auxiliary model. We then ask how concerned users should be about the possible mis-specification of their pseudo-true model. To investigate this we generate data from a DSGE model constructed to be more complex than the DSGE model from which we form the pseudo-true model used in our test. We find that in a typical small sample the IIW test will reject any mis-specified pseudo-true model with a probability of virtually 100%. This shows that if a pseudo-true model passes the test, it provides a sufficiently good representation of the generated data to be regarded as if it were the true model. More generally, it implies that the non-rejection of a pseudo-true model is a useful guide to the validity of the DSGE model it is based on. By calculating the power of the test as parameters are moved further away from their estimated values it is possible to establish bounds for their possible numerical falsity.

This implies that a DSGE model can be tested using classical statistical inference as if it were a true representation of the economy even though the economy's "reality" is unknown. The test of the model is whether it is pseudo-true and hence a valid statistical representation of the relevant data properties. In effect as we have said this returns us to Friedman's original methodology whereby a model is a deliberate simplification of the economy's complex reality, which we should test the model as if it is true in order to see whether it can get 'close' to those aspects of reality relevant for the model's user. Under this interpretation traditional macroeconometric models may also be regarded as being only pseudo-true. What distinguishes DSGE models from traditional models is their interpretation as being structural.

The remainder of the paper is organised as follows. In section 2 we discuss the frequentist methods that are used to test DSGE models. In section 3, we introduce the the idea of using indirect inference in carrying out hypothesis tests of estimated structural models and, in particular, DSGE models. We consider, with a simple example, how best to form the auxiliary model for a DSGE model. In section 4, we describe the two tests — the LR and the indirect inference IIW test — and we examine their distributions both analytically and numerically using the Smets-Wouters (2007, SW) model. In section 5, we investigate the power of these two tests using numerical procedures. In section 6, we investigate the power of the IIW test when the model is mis-specified and ask whether we can confidently test a DSGE model as if it is is a pseudo-true hypothesis. Our conclusions are reported in section 7.

2 Testing DSGE models by frequentist methods

We start by addressing the issue of how best to test an already estimated macroeconomic model in a classical or frequentist manner, as judged by the power properties of the test; we do this in the standard way, treating the model as if true and we set on one side the issue of mis-specification to which we return in a later section. This problem has particular relevance for DSGE models. It is rare for these models to be tested because they are commonly estimated by Bayesian methods with the validity of the specification of the model, and the prior information, being taken as given. Both, however, may be incorrect. It would not, for example, be surprising to find that incorporating incorrect prior information would cause the Bayesian-estimated model to be rejected against maximum likelihood estimates of the model. Le et al. (2011), for example, rejected the SW model. There is, however, an argument for not testing DSGE models. As noted by Sargent (see Evans and Honkapohja, 2005), the "rejection of too many good models" was what led Lucas and Prescott to reject classical estimation methods in favour of calibration. The use of Bayesian estimation derives from a similar concern; the difference arises from the weight given to the prior information.

Although it is not common to test DSGE models estimated by Bayesian methods, it is nonetheless possible to do so. One way is to perform a likelihood ratio (LR) test of the model against its unrestricted solution using the observed data set. Another way, proposed by Le et al. (2011), is to use an indirect inference test. This method of testing may be applied to any given set of estimates of a model, and not just DSGE models, or models estimated using Bayesian methods. The basic idea here is to simulate the already estimated model and compare the properties of an auxiliary model — which plays the role of an unrestricted solution — estimated on actual and simulated data. Using Monte Carlo experiments, Le et al. (2016a) found that, in small samples, LR tests of DSGE models may have weaker power than an indirect inference test based on comparing particular features of an auxiliary model estimated on actual and simulated data sets, such as its coefficients or impulse response functions, and using a Wald-type test statistic. The attraction of this approach is that it can be tailored to specific properties of the auxiliary model rather than its overall fit, as in an LR test. In this way it may be possible to test those features of a DSGE model that are thought to be "good" or "important" and avoid rejecting the model on the basis of other features which may be thought to be inessential.

Canova and Sala (2009) have suggested that the low power of LR tests may be due to the likelihood surface of the data being rather flat — a result they put down to poor identification.[1]

Another possible explanation arises from the way that the DSGE model is specified. The equation dynamics of most DSGE models are usually rather simple, having just first-order dynamics. This is probably because the underlying theory usually has little to say about the lag dynamic structure. The estimated equations are, however, often found to have highly serially correlated disturbances. The SW model is a good example; most of the equations have serially correlated errors, some with serial correlations as high as 0.97. Allowing the disturbances to be serially correlated greatly improves fit and so raises the likelihood of the

[1] Identification is a theoretical property of the (DSGE) model when data is unlimited; it exists when the reduced form of a model cannot be generated by a different model. While it is possible that lack of this theoretical property is what lies behind the flat likelihood surface, in a recent paper Le, Minford and Wickens (2013) suggested that two macro models in wide current use, those of Smets and Wouters (2003, 2007) and Clarida, Gali and Gertler (1999), were highly over-identified; they tested both of them by indirect inference to see whether in Monte Carlo samples they generated data whose reduced form (or approximations to it) could also be generated by other DSGE model versions; had it been possible to find such a model it would be rejected the same percent of the time as the true original model. Yet the nearest model they could find in both cases was rejected nearly 100% of the time on a 5% test when of course the true model is only rejected 5% of the time.

model fitting the data.

In effect, due to the form of the solution to DSGE models, an LR test is based on one-period ahead 'forecasts'. It seems possible that the weak power of LR and the flat likelihood surface for DSGE models may come from the way in which false structural parameters may have their 'tracking performance' failure disguised by re-estimated error processes. Thus, a model's false structural parameters will imply different, false, error processes; these processes will give rise to newly estimated autoregressive parameters which will bring the model back on track in its ability to 'forecast' next-period outcomes, this being what likelihood is based on. If modelling a DSGE model's structural errors as autoregressive processes in order to improve its fit is interpreted as an integral part of its dynamic specification, then this would suggest that, when carrying out power calculations — which involves simulating false versions of the model by using alternative values of the structural coefficients — the autoregressive coefficients of the new structural errors should be re-estimated in order to maximise fit. This too will help 'bring the model back on track' in its ability to 'forecast' next-period outcomes, but it will also be likely to reduce the power of the test.

In contrast, the indirect inference Wald (IIW) test does not use tracking performance as a measure of fit. Instead it compares the reduced form — or an auxiliary model that is a close approximation — found in the data with that implied by simulations of the DSGE model derived from the false parameter values generated to make the power calculations. The error processes are not re-estimated and will therefore no longer be best fit for these simulations of the false models. This is likely to improve the power of the test. For example, Le et al. (2016a) allowed the estimated model parameters to be arbitrarily moved towards greater falseness by small percentages; as falseness rose the models were rejected with fast-increasing frequency, showing the power of the test. Le et al. (2016a) also found that making the structural parameters of the model more false caused the autoregressive parameters of the false model to differ significantly from those of the original estimates which made the test reject more powerfully still.

Unless the priors used in Bayesian estimation are uninformative (or diffuse), Bayesian estimates will usually differ from maximum likelihood estimates, and so will not maximise the likelihood function. Consequently, applying an LR test to a model estimated by Bayesian methods is likely to raise the power of the LR test compared to the use of maximum likelihood estimation. In effect, the posterior mode is a weighted average of the mode of the prior distribution and the maximum likelihood estimate with the weights determined roughly by their relative precisions. The greater the influence of the prior information relative to the sample information, the more likely is an LR test to reject the model when the posterior distribution is centred differently from the maximum likelihood estimator.

In this paper we investigate the power of two tests of an already estimated DSGE model. One is an LR test in which the autoregressive processes generating the structural disturbances are re-estimated to maximise fit. The other is the IIW test in which the error autoregressive processes are not re-estimated. We begin by examining their asymptotic or large sample properties. The IIW test is based on the distance between the data descriptors implied by the true model and those implied by the false model; this distance depends on the degree of falseness of the model's structural parameters and error processes — both their AR parameters and their innovation moments. The LR test is based on the distance between the two models' forecasting errors. This depends on the falseness of the structural parameters. It is also affected by re-estimating the AR parameters of the error processes which partly offsets the effect on the overall forecast error of the false parameters. We find that the powers of the two tests turns on two factors. The first is whether the LR test is preceded by re-estimation of the model or of its error processes; if it is, the LR test's power is substantially weakened. The second is the way the IIW test is implemented: whether it is based on the variance matrix of the coefficients of the auxiliary VAR model estimated from the observed data, or, as is done by Le et al. (2011, 2016a), on data simulated from the DSGE model with its false structural parameters. Using the former the powers of the LR and the indirect inference tests are roughly equivalent, but using the latter endows the IIW test with more power. The latter IIW test is the one that is referred to as 'the IIW test' in what follows unless otherwise specified.

3 Indirect inference tests of a DSGE model

The IIW test focuses on specific features of the DSGE model such as particular impulse response functions, rather than on the overall fit of the full model as in an LR test. A justification for this is provided by Lucas and Prescott who objected to likelihood ratio tests of DSGE models on the grounds that "too many good models are being rejected by the data". Their point is that the DSGE model may offer a good explanation of features of interest but not of other features of less interest, and it is the latter that results in the rejection of the model by conventional hypothesis tests.

In an indirect inference test 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 (here a VAR model) when estimated from actual data (the alternative hypothesis). If the DSGE model is correct then the simulated data, and the VAR estimates based on these data, will not be significantly different from those derived from the actual data. The method is in essence extremely simple. The idea is to bootstrap the estimated DSGE model. These bootstraps provide simulations of the data that 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 VAR coefficients estimated using the simulated data.

We have argued that the DSGE model may be regarded as a pseudo-true representation of the economy and that the auxiliary model, which is an approximation to the DSGE model, is therefore also a pseudo-true model of the economy. In the original real business cycle analysis based on calibration the auxiliary model consisted of the moments of the data and a comparison of these moments on observed and simulated data. The drawback with this is that it limits the properties of the DSGE model that can be investigated. Minford, Wickens and Xu (2016) report that using moments instead of VAR coefficients or impulse response functions can lower power substantally. As the solution of a DSGE model can be represented as a VAR, or can be closely approximated by one, we use a VAR as the auxiliary model. In this way we aim to capture the key properties of interest of the DSGE model. A VAR also has the advantage of being easy to estimate. In the next section we show how this VAR may be obtained.

3.1 The auxiliary model: a VAR representation of a DSGE model

There are several ways of deriving a VAR representation of a DSGE model. We make use of the ABCD framework of Fernandez-Villaverde et al. (2007). We consider solely what these authors call the 'square' case, where the number of errors and the number of observable variables are the same. We also consider only DSGE models with no observable exogenous variables. Both the Smets-Wouters model (Smets and Wouters, 2003;2007) and the 3-equation model New Keynesian model used by Le et al. (2013) and Liu and Minford (2014) for their numerous IIW tests fit this framework. (Other classes of models, for example those with 'news shocks', require a different treatment which is beyond our scope here.)

To illustrate, consider the 3-equation New Keynesian model of Le et al. (2013):[2]

$$\begin{aligned}
\pi_t &= \omega E_t \pi_{t+1} + \lambda y_t + e_{\pi t}, \qquad \omega < 1 \qquad (1) \\
y_t &= E_t y_{t+1} - \frac{1}{\sigma}(r_t - E_t \pi_{t+1}) + e_{yt} \\
r_t &= \gamma \pi_t + \eta y_t + e_{rt} \\
e_{it} &= \rho_i e_{i,t-1} + \varepsilon_{it} \qquad (i = \pi, y, r)
\end{aligned}$$

[2]Further lags in both endogenous variables and the errors could be added; but for our main treatment we suppress these. Our results can be extended to deal with them, without essential change.

This has the solution

$$\begin{bmatrix} \pi_t \\ y_t \\ r_t \end{bmatrix} = KH \begin{bmatrix} e_{\pi t} \\ e_{yt} \\ e_{rt} \end{bmatrix} \tag{2}$$

where

$$K = \begin{bmatrix} 1+\frac{\eta}{\sigma}-\rho_\pi & \lambda & -\frac{\lambda}{\sigma} \\ -\frac{1}{\sigma}(\gamma-\rho_\pi) & 1-\omega\rho_y & -\frac{1}{\sigma}(1-\omega\rho_r) \\ \gamma-(\gamma-\frac{\eta}{\sigma})\rho_\pi & \lambda\gamma+\eta-\eta\omega\rho_y & 1-(1+\omega+\frac{\lambda}{\sigma})\rho_r+\omega\rho_r^2 \end{bmatrix},$$

$$H = \begin{bmatrix} H_{11} & 0 & 0 \\ 0 & H_{22} & 0 \\ 0 & 0 & H_{33} \end{bmatrix},$$

$$\begin{aligned} H_{11} &= \frac{1}{1+\frac{\eta+\lambda\gamma}{\sigma}-[\frac{\lambda}{\sigma}+\omega(1+\frac{\eta}{\sigma})]\rho_\pi+\omega\rho_\pi^2} \\ H_{22} &= \frac{1}{1+\frac{\eta+\lambda\gamma}{\sigma}-[\frac{\lambda}{\sigma}+\omega(1+\frac{\eta}{\sigma})]\rho_y+\omega\rho_y^2} \\ H_{33} &= \frac{1}{1+\frac{\eta+\lambda\gamma}{\sigma}-[\frac{\lambda}{\sigma}+\omega(1+\frac{\eta}{\sigma})]\rho_r+\omega\rho_r^2.} \end{aligned}$$

or

$$z_t = \Phi e_t \tag{3}$$

$$e_t = Pe_{t-1}+\varepsilon_t \tag{4}$$

where $z_t' = [\pi_t, y_t, r_t]$, $e_t' = [e_{\pi t}, e_{yt}, e_{rt}]$, $\Phi = K \times H$. Thus the matrix Φ is restricted, having 9 elements but consists of only 5 structural coefficients (the ρ_i can be recovered directly from the error processes), implying that the model is over-identified according to the order condition. The model is not identified, however, if the $\rho_i = 0$ for all i.[3]

[3]Le et al., 2013, also establish that it is identified using the IIW test in unlimited-size sampling.

The solved structural model can be written in ABCD form as follows where y (replacing z above) is now the vector of endogenous variables and x (replacing e above) is the vector of error processes:

$$
\begin{array}{lrcl}
(1) & x_t & = & Ax_{t-1} + B\varepsilon_t \\
(2) & y_t & = & Cx_{t-1} + D\varepsilon_t
\end{array}
$$

where $A = P = \begin{bmatrix} \rho_\pi & 0 & 0 \\ 0 & \rho_y & 0 \\ 0 & 0 & \rho_r \end{bmatrix}; B = I; C = \Phi P; D = \Phi.$

Note that $y_t = \Phi x_t$ is the (solved) structural model. Hence $x_t = \Phi^{-1} y_t$. The VAR representation is [4]

$$y_t = \Phi P \Phi^{-1} y_{t-1} + \Phi \varepsilon_t = V y_{t-1} + \xi_t \tag{5}$$

We may also note that

$$y_t = \Phi \sum_{i=0}^{\infty} P^i \varepsilon_{t-i} = \sum_{i=0}^{\infty} P^i \xi_{t-i}.$$

More generally, the solution of a linearised DSGE model (including the SW model and the 3-equation model) can be summarised by a state-space representation:[5]

$$
\begin{array}{rcl}
x_t & = & Ax_{t-1} + B\varepsilon_t \\
y_t & = & Cx_t
\end{array}
$$

where x_t is an $n \times 1$ vector of possible unobserved state variables, y_t is a $k \times 1$ vector

[4] If the DSGE model also had one-period lags in one or more of the equations so that the solution became $z_t = \Phi e_t + \Lambda z_{t-1}$ then we would obtain a VAR(2) as follows:
(1) $x_t = Ax_{t-1} + B\varepsilon_t$
(2) $y_t = Cx_{t-1} + D\varepsilon_t + \Lambda y_{t-1}$
Using $x_{t-1} = \Phi^{-1}(y_{t-1} - \Lambda y_{t-2})$ we obtain
$y_t = (\Phi P \Phi^{-1} + \Lambda) y_{t-1} - \Phi^{-1} \Lambda y_{t-2} + \Phi \varepsilon_t$

[5] The solution of the model can be obtained by using either Blanchard and Kahn (1980) or Sims (2002) type of algorithms.

of variables observed by an econometrician, and ε_t is an $m \times 1$ vector of economic shocks affecting both the state and the observable variables, i.e., shocks to preferences, technologies, agents' information sets, and economist's measurements. The shocks ε_t are Gaussian vector white noise satisfying $E(\varepsilon_t) = 0, E(\varepsilon_t \varepsilon_t') = I$. The matrices A, B and C are functions of the underlying structural parameters of the DSGE model. Using the ABCD framework of Fernandez-Villaverde et al. (2007), the state-space representation can be written as the VAR

$$y_t = V y_{t-1} + \eta_t \tag{6}$$

where $E(\eta_t \eta_t') = \Phi\Phi' = \Sigma$.

We have assumed that the DSGE model includes no observable exogenous variables. If it does then the solution to the DSGE model contains exogenous variables as well as lagged endogenous variables: in general, lagged, current and expected future exogenous variables. If, however, the exogenous variables are assumed to be generated by a VAR process then the combined solution of both the endogenous and exogenous variables is a purely backward-looking model that can be represented as a VAR.[6]

4 The LR and the IIW test statistics

In indirect inference we do not impose the restrictions on the coefficients of the auxiliary model that are implied by the structural model. Instead, we estimate the auxiliary model on data simulated from the structural model and compare these estimates with those obtained from using the observed data. In both cases the auxiliary model is estimated without any coefficient restrictions. The restrictions imposed by the DSGE model are reflected in the simulated data and not through explicit restrictions on the auxiliary model.

Since both the LR test and the IIW test involve estimation of an unrestricted VAR, first

[6] 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)), and Wickens (2014).

we briefly review the maximum likelihood estimation (MLE) of a standard unrestricted VAR. Consider a randomly generated sample of y_t of size T. If η_t is assumed to be $NID(0, \Sigma)$ then the log-likelihood function is

$$\ln L(V, \Sigma) = -[\frac{Tn}{2}\ln(2\pi) + \frac{T}{2}\ln|\Sigma| + \frac{1}{2}\sum_{t=1}^{T}(y_t - Vy_{t-1})'\Sigma^{-1}(y_t - Vy_{t-1})]$$

Maximising with respect to Σ^{-1} gives

$$\frac{\partial \ln L(V, \Sigma)}{\partial \Sigma^{-1}} = \frac{T}{2}\Sigma - \frac{1}{2}\sum_{t=1}^{T}(y_t - Vy_{t-1})(y_t - Vy_{t-1})'$$

Setting this to zero and solving gives the MLE estimator of Σ as

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}(y_t - Vy_{t-1})(y_t - Vy_{t-1})' \tag{7}$$

Substituting this back into the likelihood function gives the concentrated likelihood

$$\ln L(V, \hat{\Sigma}) = -[\frac{Tn}{2}\ln(2\pi) + \frac{T}{2}\ln|\hat{\Sigma}| + \frac{Tn}{2}]$$

Maximising this with respect to V is identical to minimising $\ln|\hat{\Sigma}|$ with respect to V. Thus

$$\frac{\partial \ln|\hat{\Sigma}|}{\partial V} = 2\hat{\Sigma}^{-1}\sum_{t=1}^{T}(y_t - Vy_{t-1})y_{t-1}' = 0$$

and hence the MLE of V is

$$\widehat{V} = (\sum_{t=1}^{T} y_t y_{t-1}')(\sum_{t=1}^{T} y_{t-1}y_{t-1}')^{-1}$$

and can be calculated by applying OLS to each equation separately. The MLE of Σ becomes

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{V}y_{t-1})(y_t - \widehat{V}y_{t-1})' \tag{8}$$

In order to find the variance matrix of $\widehat{V}$ it is convenient to re-express the VAR. Denoting the T observations on the i^{th} element of y_t as the $T \times 1$ vector y_i and of η_t as η_i, each equation of the VAR may be written as

$$y_i = Zv_i + \eta_i \tag{9}$$

where v_i' is the i^{th} row of V and Z is a $T \times k$ matrix with t^{th} row y_{t-1}. The VAR may now be written in matrix form as

$$Y = Xv + \eta \tag{10}$$

where

$$Y = \begin{bmatrix} y_1 \\ . \\ y_T \end{bmatrix}, \quad X = \begin{bmatrix} Z & ... & 0 \\ .. & .. & .. \\ 0 & ... & Z \end{bmatrix} = I_k \otimes Z, ...\eta = \begin{bmatrix} \eta_1 \\ . \\ \eta_T \end{bmatrix}, ...v = \begin{bmatrix} v_1 \\ . \\ v_k \end{bmatrix}$$

$\otimes$ denotes a Kronecker product. Hence η is $N(0,\Omega)$ where $\Omega = \Sigma \otimes I_T$. Generalised least squares estimation gives the MLE of v as

$$\begin{aligned} \hat{v} &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y \\ &= [I_k \otimes (Z'Z)^{-1}Z']Y \\ &= v + [I_k \otimes (Z'Z)^{-1}Z']\eta \end{aligned}$$

In general $\hat{v}$ is a biased estimate of v as Z consists of lagged endogenous variables, but *plim*

$\hat{v} = v$ and the limiting distribution of $\sqrt{T}(\hat{v} - v)$ is $N(0, W)$ where

$$\begin{aligned} W &= plim\ T[I_k \otimes (Z'Z)^{-1}Z'](\Sigma \otimes I_T)[I_k \otimes (Z'Z)^{-1}Z']' \\ &= \Sigma \otimes (plim\ T^{-1}Z'Z)^{-1} \end{aligned}$$

4.1 The LR test

The LR test for a DSGE model based on the observed data compares the likelihood function of the auxiliary VAR derived from the DSGE model with the likelihood function of the unrestricted VAR computed on the observed data. The former is based on the estimate of the variance matrix of the structural errors from the solution to the DSGE model. On the assumption that the auxiliary model is the solution to the DSGE model and is a VAR, this is also the error variance matrix of a restricted version of the auxiliary VAR. The latter is based on the estimate of the error variance matrix of the unrestricted auxiliary VAR. As the auxiliary model is a VAR, the LR test is, in effect, based on the one-period ahead forecast error matrix. Thus, the logarithm of the likelihood ratio test is

$$\begin{aligned} LR &= 2(\ln L_U - \ln L_R) \\ &= T\left(\ln|\Sigma_R| - \ln\left|\widehat{\Sigma}\right|\right) \end{aligned} \tag{11}$$

where L_R and L_U denote the likelihood values of the restricted and unrestricted VAR, respectively, and Σ_R and $\widehat{\Sigma}$ are the restricted and unrestricted error variance matrices. Note that, given estimates of the DSGE model, we can solve the model for v, and hence we can calculate η_t and $\Sigma_R = T^{-1}\sum_{t=1}^{T}\eta_t\eta_t'$. Note also that the LR test can be routinely transformed into a (direct inference) Wald test between the unrestricted and the restricted VAR coefficients, v.

To obtain the power function of the LR test we endow the structural model with false values of the structural coefficients and compare the restricted VAR with the unrestricted

VAR on the observed data which are assumed to be generated by the true model. The implied false model has the VAR

$$y_t = V_F y_{t-1} + \eta_{Ft} \tag{12}$$

The forecast errors for the false model are

$$\eta_{Ft} = y_t - V_F y_{t-1} = \eta_t + (V - V_F) y_{t-1} = \eta_t + q_t$$

where $q_t = Y_{t-1}(v - v_F)$. If we let

$$\Sigma_F = \frac{1}{T} \sum_{t=1}^{T} \eta_{Ft} \eta'_{Ft} = \frac{1}{T} \sum_{t=1}^{T} (\eta_t + q_t)(\eta_t + q_t)'$$

then the LR test for the false model is given by:

$$LR_F = T[\ln |\Sigma_F| - \ln \left|\widehat{\Sigma}\right|] \tag{13}$$

Thus the power of the test derives from the distance

$$\ln |\Sigma_{_F}| - \ln |\Sigma_{_R}| \,. \tag{14}$$

4.2 The IIW test

In the IIW test we simulate data from the solution to the already estimated DSGE, randomly drawing the samples from the DSGE model's structural errors. We then estimate the auxiliary VAR using these simulated data. We repeat this many times to obtain the average estimate of the coefficients of the VAR which we take as the estimate of the unrestricted

VAR. The simulated VAR may be written

$$y_{S,t} = V_S y_{S,t-1} + \eta_{St}$$

where $y_{S,t}$ is the data simulated from the DSGE model and V_S is the (average estimate of v) or, in the form of equations (9) and (10), as

$$\begin{aligned} y_{S,i} &= Z_S v_{S,i,} + \eta_{S,i} \\ Y_S &= X_S v_S + \eta_S \end{aligned}$$

where $E(\eta_{S,i}\eta'_{S,i}) = \Sigma_S$. The IIW test statistic, which computes the distance of these estimates from the unrestricted estimates based on the observed data, is:

$$IIW = [\widehat{v} - v_S]' W_S^{-1} [\widehat{v} - v_S] \tag{15}$$

where W_S is the covariance matrix of the limiting distribution of v_S, and is given by

$$W_S = \Sigma_S \otimes (plim\ T^{-1} Z_S' Z_S)^{-1} \tag{16}$$

On the null hypothesis that the DSGE model — and hence the auxiliary VAR — are correct, the asymptotic distribution of the estimate of v_S is the same that of the MLE $\widehat{v}$. Moreover, asymptotically, this IIW statistic will have the same distribution as $[\widehat{v} - v]' W^{-1} [\widehat{v} - v]$ and hence will have the same critical values.[7] In general, the IIW statistic differs from a standard Wald statistic in indirect inference which is $[\widehat{v} - v_S]' W^{-1} [\widehat{v} - v_S]$ where W is the covariance matrix of the unrestricted model; we refer to this as the unrestricted IIW statistic.

The power of the IIW test is calculated, like that for the power calculations for the LR test, by simulating the DSGE model using false values of its coefficients and now using

[7] The IIW test can also be carried out for a sub-set of v.

these data to estimate the unrestricted VAR from equation (12). The IIW statistic is then computed from

$$IIW = [\widehat{v} - v_F]' W_F^{-1} [\widehat{v} - v_F] \tag{17}$$

where v_F is the mean vector of coefficients and W_F is their variance matrix, which corresponds to W_S. Consider the decomposition

$$\widehat{v} - v_F = (\widehat{v} - v) + (v - v_F).$$

It follows that the IIW statistic can be decomposed as

$$\begin{aligned}
& [\widehat{v} - v_F]' W_F^{-1} [\widehat{v} - v_F] && (18) \\
= \; & \eta' [I_k \otimes (Z'Z)^{-1} Z']' W_F^{-1} [I_k \otimes (Z'Z)^{-1} Z'] \eta \\
& + [v - v_F]' W_F^{-1} [v - v_F] \\
= \; & \eta' [\Sigma^{-1} \otimes plim\ T(Z'Z)^{-1}] \eta + [v - v_F]' W_F^{-1} [v - v_F] && (19)
\end{aligned}$$

where the last term is based on the difference between the true and the false values of the coefficients. Hence the power of the IIW test derives from the second term on the right-hand side of equation (19).

We note two things. First, both the sign and size of the change in the first term on the right-hand side of equation (19) as the false model changes cannot be evaluated analytically; it depends on how the covariance weighting matrix of the false parameters, W_F^{-1}, changes and interacts with $[I_k \otimes (Z'Z)^{-1} Z']\eta$, the sample differences on the true data of the estimated v from the true v. If W_F were diagonal then the false weighting matrix would, in effect be dividing each element by its false standard deviation, thereby converting them into false t-values; some elements will have t-values that are too large, others that are too small. v_F, the vector of VAR parameters implied by the false model, depends on two false elements: θ (the DSGE model's structural parameters) and P (the time-series parameters of the errors).

Second, in the New Keynesian model above, v_F is a row in the matrix $V = \Phi P \Phi^{-1}$ where Φ depends on θ, P. In small samples we use the mean of the estimated v_F, and hence a third false element — ε (the vector of innovations in the DSGE model's structural errors) — affects the power through its properties (i.e. its variance matrix, skewness and kurtosis). For small samples Le et al. (2016a) were able via Monte Carlo experiments to generate some orders of magnitude for the contribution of the different elements to the power of the IIW test. Essentially they found that they are all of some importance — Table 1 shows their findings for the 3-equation New Keynesian model on stationary data.

ALL ELEMENTS	Level of Falseness						
	1%	3%	5%	7%	10%	15%	20%
2 variable VAR(1)	16.8	82.6	99.6	100.0	100.0	100.0	100.0
3 variable VAR(1)	25.1	97.7	100.0	100.0	100.0	100.0	100.0
3 variable VAR(2)	16.1	77.2	98.4	100.0	100.0	100.0	100.0
3 variable VAR(3)	14.4	73.0	97.5	99.7	100.0	100.0	100.0
STRUCTURAL PARAMETERS							
2 variable VAR(1)	7.3	8.7	12.6	19.3	40.4	76.1	92.7
3 variable VAR(1)	6.2	10.1	25.5	53.7	80.7	99.4	100.0
3 variable VAR(2)	6.8	9.3	12.8	20.6	45.9	77.2	95.0
3 variable VAR(3)	5.8	7.5	12.0	21.7	45.8	74.0	95.5
AR PARAMETERS							
2 variable VAR(1)	16.2	86.3	99.7	100.0	100.0	100.0	100.0
3 variable VAR(1)	18.8	96.8	100.0	100.0	100.0	100.0	100.0
3 variable VAR(2)	16.5	87.3	99.9	100.0	100.0	100.0	100.0
3 variable VAR(3)	18.9	81.6	99.5	100.0	100.0	100.0	100.0
SHOCKS							
2 variable VAR(1)	5.6	6.8	5.7	10.1	15.0	27.3	46.7
3 variable VAR(1)	5.4	6.0	8.4	8.7	11.7	26.7	48.8
3 variable VAR(2)	5.6	5.4	5.1	9.0	13.1	31.0	41.8
3 variable VAR(3)	4.9	6.1	4.1	9.0	12.4	29.5	48.2

Table 1: 3-EQUATION MODEL, STATIONARY DATA: Decomposition of the power of IIW

4.3 Comparing the power of the two tests

We have seen that the LR test compares the one-step ahead forecast error matrix of the unrestricted VAR with that of the model-restricted VAR using the observed data, whereas the IIW test asks whether the distribution of the VAR coefficients based on the simulated data (the restricted model) covers the VAR coefficients based on the observed data (the unrestricted model). We have also found that on the null hypothesis that the DSGE model is true the limiting distributions of the two sets of estimates are the same. It follows from equation (7) that, on the null hypothesis, the error variance matrix using simulated data is

$$\begin{aligned}
\Sigma_S &= \frac{1}{T}\textstyle\sum_{t=1}^{T}(y_{St} - V_S y_{S,t-1})(y_{St} - V_S y_{S,t-1})' \\
&= \frac{1}{T}\textstyle\sum_{t=1}^{T}(y_t - V_S y_{t-1})(y_t - V_S y_{t-1})' + \Delta \\
&= \hat{\Sigma} + (\widehat{V} - V_S)\frac{1}{T}\sum_{t=1}^{T} y_{t-1}y'_{t-1}(\widehat{V} - V_S)' + \Delta
\end{aligned}$$

where $\widehat{\Sigma}$ is the error variance matrix of the unrestricted VAR using the observed data and Δ is $O_p(T^{-\frac{1}{2}})$.

Using the result that $vec(AXB) = (B' \otimes A)vec(X)$, and $vec(V') = v$, it can be shown that

$$vec[(\widehat{V} - V_S)\frac{1}{T}\sum_{t=1}^{T} y_{t-1}y'_{t-1}(\widehat{V} - V_S)'] = v'(I \otimes \frac{1}{T}\sum_{t=1}^{T} y_{t-1}y'_{t-1})v$$

Hence,

$$
\begin{aligned}
LR &= T\left(\ln|\Sigma_S| - \ln\left|\widehat{\Sigma}\right|\right) \\
&= T[\ln|1 + \frac{(\widehat{V} - V_S)\frac{1}{T}\sum_{t=1}^{T} y_{t-1}y'_{t-1}(\widehat{V} - V_S)' + \Delta}{\left|\widehat{\Sigma}\right|}|] \\
&= T[\ln|1 + (\widehat{v} - v_S)'(\widehat{\Sigma} \otimes \frac{1}{T}\sum_{t=1}^{T} y_{t-1}y'_{t-1})^{-1}(\widehat{v} - v_S) + \frac{\Delta}{\left|\widehat{\Sigma}\right|}|] \\
&\rightarrow IIW + O_p(T^{-\frac{1}{2}})
\end{aligned}
$$

In other words, on the null hypothesis that the DSGE model is the true model, the LR test based on observed data is asymptotically equivalent to using the IIW test, which is based on simulated data.

In the power calculations we use

$$
\begin{aligned}
LR &= T\left(\ln|\Sigma_F| - \ln\left|\widehat{\Sigma}\right|\right) \\
&= T\left(\ln|\Sigma_S| - \ln\left|\widehat{\Sigma}\right|\right) + T\left(\ln|\Sigma_F| - \ln|\Sigma_S|\right)
\end{aligned}
$$

The power of the test derives from the last term which reflects the difference between V_S and V_F. This makes Δ of order $O_p(1)$, which does not vanish as $T \rightarrow \infty$, but causes the power of the test to tend to unity.

5 Numerical comparison of the powers of the LR and IIW test

In power calculations we deliberately falsify the DSGE model and seek to discover the probability that the model will be rejected. When the structural model is the true model, and hence the solution is correct, the LR and IIW tests have the same asymptotic distribution and therefore test size. But when the model is made false the two tests are no longer asymptotically equivalent; hence they will have different powers. The IIW test as carried out by Le et al. (2016a) uses the distribution of the model-restricted VAR coefficients. This increases the precision of the variance matrix of the coefficients of the auxiliary model and so improves the power of the IIW test. Thus the IIW test asks whether the distribution of the model-restricted VAR coefficients covers the unrestricted VAR coefficients found in the data. The distribution of the resulting IIW statistic is asymptotically chi–squared. As in practice we are usually dealing with small samples, the distribution of the test statistic will be better determined numerically as below.

5.1 Numerical comparison of the distribution of the estimates

In our numerical comparison of the two tests our structural model is the Smets-Wouters model (2007). This is a DSGE model which has a high degree of over-identification (as established by Le et al., 2013). It has 12 structural parameters and 8 parameters in the error processes. It implies a reduced-form VAR of order 4 with seven observable endogenous variables, i.e. a 7VAR(4), (Wright, 2015). This has 196 coefficients. The size of the VAR in a IIW test and the number of variables is usually lower than a 7VAR(4).

We concentrate on the dynamic response to own shocks of inflation and the short-term nominal interest rate. We focus on the three variables of the above New Keynesian model: inflation, the output gap and the nominal interest rate. We use a 3VAR(1) in these variables as the auxiliary model. We then examine the own-lag coefficients for inflation and the short-term interest rate.

We estimate the coefficients of the 3VAR(1) using the observed data for these three variables. We then find the distribution of the estimates of the two coefficients of interest by bootstrapping the VAR innovations. Next, we estimate the 3VAR(1) using data for these three variables obtained by simulating the full SW model. The distribution of these estimates of the two coefficients is obtained by bootstrapping the structural innovations generating that sample. The graphs below show the densities of the joint distribution of the two coefficients.

Figure 1 displays the joint distributions of the two VAR coefficients based on (i) the observed data (the unrestricted VAR), (ii) simulated data from the original estimates of the structural model (the restricted VAR), and (iii) false specifications of the structural models by 5% and 10% (the 5% false and 10% false restricted VARs). One can see clearly that (ii), the joint distribution based on simulated data from the original structural model, is both more concentrated and more elliptical (implying a higher correlation between the coefficients) which uses the observed data. Increasing the falseness of the model causes (iii), the joint distributions from the 5% and 10% false DSGE model, to become a little more dispersed and more elliptical; they are also located slightly differently but this is not shown as the distribution is centred on zero in all cases.

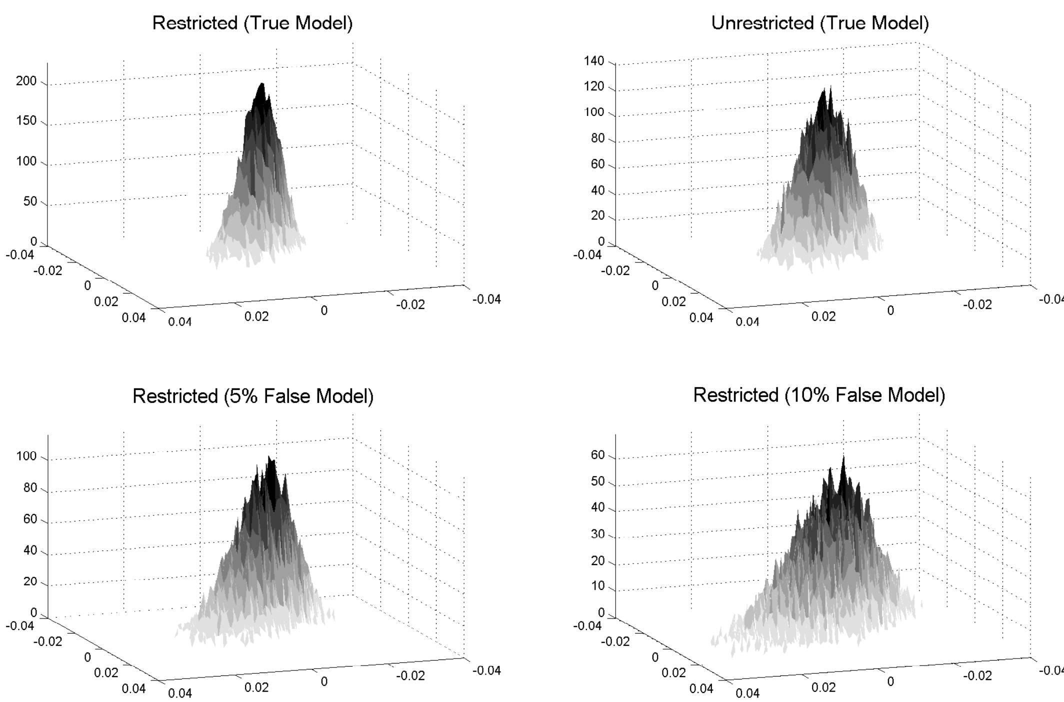

Figure 1: Restricted VAR and Unrestricted VAR Coefficient Distributions

Figure 2 shows how this affects the power of the Wald test for a model that is 5% false.

The green dot is the mean of this false distribution. We have drawn the diagram as if the joint test of two chosen VAR coefficients has the same power as the overall test of all VAR coefficients.

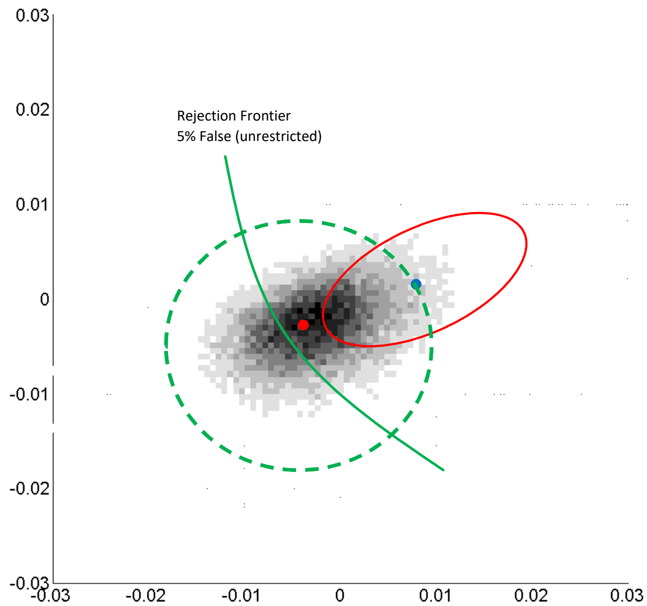

Figure 2: Two 95% contours for tests of 5% False Model- Green=Unrestricted; Red=Restricted.

The test of this false model can be carried out in two ways. First, we employ the unrestricted Wald test, using the observed data to estimate a 3VAR(1) representation and to derive the joint distribution of the two coefficients by bootstrapping. The 5% contour of such a bootstrap distribution is given by the dashed green line; the thick green line shows the critical frontier at which the 5% false model is just rejected. Second, we employ the restricted Wald test, using the distribution implied by the simulated data. The red ellipse shows the 5% contour of the resulting joint distribution. The results show that the second method has nearly double the power of the first. (Increasing the degree of falseness to 10% raises the power of both to 100%.)

5.2 Numerical comparison of the power of the test statistics

In the above comparison of the joint distribution of the two coefficients of interest, the data simulated from the structural model gave serially correlated structural error processes. In order to make the estimates of their joint distribution compatible with the original Smets-Wouters estimation strategy, first-order autoregressive processes were fitted to these structural errors for each bootstrap sample. In calculating the power of the tests we proceed a little differently in order that the tests are based on the same assumptions when the structural model is falsified. We now fix both θ (the vector of structural coefficients of the DSGE model) and ρ (the vector of coefficients of the autoregressive error processes). Each is falsified by $x\%$. We do not, however, falsify the innovations, maintaining them as having the original true distribution. This last point is a matter of convenience as we could extract the exact implied false error innovations, as implied by each data sample, θ and ρ. But this extraction is a long and computationally-intensive process requiring substantial iteration. We simply assume, therefore, that the model is false in all respects except for the innovations. Generally, we find that if only the innovations are false this generates little power under either test (see Le et al., 2016a) so this omission should make no difference to the relative power calculation. We use the SW model as the true model with a sample size of 200 throughout. Our findings are reported in Table 2.

We find, as we would expect, that the two test statistics generate similar power when the IIW test is based on the observed data (the unrestricted VAR). Focusing on the main case, which is a 3VAR(1), and taking 5% falseness as our basic comparison, we see that the rejection rate for the LR test is 38%. For the IIW test based on an unrestricted VAR the rejection rate is 31% while using a restricted VAR (simulated data) for the IIW test it is 85%. The orders of magnitude of the rejection rates are therefore similar for LR test and a IIW test based on the observed data[8], while for the a IIW test based on simulated data

[8]This test uses the variance matrix of the VAR coefficients for the observed data. When this VAR has a very large number of coefficients the variance matrix of the coefficients has a tendency to become unstable; this occurs even when the number of bootstraps is raised massively (e.g. to 10000). This is due to over-fitting

VAR — no of coeffs	TRUE	1%	3%	5%	7%	10%	15%	20%
IIW TEST with unrestricted VAR								
2 variable VAR(1) — 4	5.0	6.2	20.3	69.6	61.0	99.8	100.0	100.0
3 variable VAR(1) — 9	5.0	3.4	7.5	30.7	75.0	97.4	100.0	100.0
3 variable VAR(2) — 18	5.0	3.8	5.2	19.1	57.5	84.3	98.4	99.5
3 variable VAR(3) — 27	5.0	3.9	6.4	21.6	54.5	84.0	97.5	98.7
5 variable VAR(1) — 25	5.0	2.8	3.2	2.6	5.4	6.2	4.5	100.0
7 variable VAR(3) — 147	5.0	5.1	3.4	1.4	0.9	0.2	0.0	100.0
IIW TEST with restricted VAR								
2 variable VAR(1) — 4	5.0	9.8	37.7	80.8	96.8	100.0	100.0	100.0
3 variable VAR(1) — 9	5.0	9.5	36.1	71.0	98.1	100.0	100.0	100.0
3 variable VAR(2) — 18	5.0	8.3	35.5	80.9	96.9	100.0	100.0	100.0
3 variable VAR(3) — 27	5.0	9.2	32.9	78.0	95.1	100.0	100.0	100.0
5 variable VAR(1) — 25	5.0	17.8	85.5	99.8	100.0	100.0	100.0	100.0
7 variable VAR(3) — 147	5.0	77.6	99.2	100.0	100.0	100.0	100.0	100.0
LIKELIHOOD RATIO TEST								
2 variable VAR(1) — 4	5.0	12.0	28.3	45.9	63.4	83.2	97.0	99.7
3 variable VAR(1) — 9	5.0	9.4	21.8	37.5	58.9	84.0	99.0	100.0
3 variable VAR(2) — 18	5.0	8.9	20.7	36.8	57.6	82.9	98.7	100.0
3 variable VAR(3) — 27	5.0	8.9	20.4	36.7	56.7	82.2	98.7	100.0
5 variable VAR(1) — 25	5.0	8.9	22.4	44.3	68.6	89.6	99.6	100.0
7 variable VAR(3) — 147	5.0	5.7	10.6	23.6	46.3	83.2	99.6	100.0

Table 2: Comparison of rejection rates at 95% level for Indirect Inference and Direct Inference

they are very considerably higher, implying greater power. In what follows we will refer to this last, the IIW test based on the restricted VAR, as 'the' IIW test.

5.3 Why does the IIW test have more power in small samples than the Likelihood Ratio test?

Although the LR and IIW tests are asymptotically equivalent when the structural model is the true model and so generating the observed data, the two tests have different power, as we have seen. We consider two possible reasons for this: a) they are carried out with different procedures; b) even when the same procedures are followed, the two tests differ in power by construction.

in small samples (here the sample size is 200); there is then insufficient information to measure the variance matrix of the VAR coefficients.

5.3.1 Reason a): different procedures

We noted earlier that in order to improve the fit of a DSGE model it is usual to respecify the structural errors as being serially correlated by adding to the model the assumption that the errors are generated by first-order autoregressive processes. Accordingly, in calculating the power of the LR test, we re-estimated the error processes in order to 'bring the model back on track'. This will clearly reduce the power of the LR test as it will make it less likely that a false model will be rejected. This can be illustrated by comparing the power of the LR test in which the autoregression coefficients are re-estimated, as above, with an LR test in which the degree of falsification of the autoregressive coefficients is pre-specified, as for the IIW test above. We employ a 3-equation NK model for the comparison. As expected, the results in Table 3 show that the LR test with pre-specified autoregressive coefficients has considerably greater power than the test using re-estimated autoregressive coefficients.

3-equation NK model (no lags) Rejection rate of false models at 95% confidence: T=200		
	Re-estimated $\rho's$	Pre-specified $\rho's$
True	5.0	5.0
1%	5.0	5.0
3%	5.3	9.6
5%	6.1	20.2
7%	8.0	39.1
10%	15.4	63.7
15%	48.1	90.7
20%	75.6	98.9

Table 3: Comparing power due to wrong parameter values

5.3.2 Reason b): comparison when the same procedures are followed

In our numerical comparison above of the LR and the two IIW tests, we noted that the power of the LR and the IIW test using the unrestricted VAR coefficient distribution were similar when testing a DSGE model on a like-for-like basis (i.e. using the same procedure). We also noted the IIW test carried out by Le et al. (2011, 2016a) using the VAR coefficient distribution as restricted by the DSGE model was substantially more powerful than the other form of the IIW test and hence than the LR test even when using the same procedure. This therefore is the second reason for the greater power of the IIW test as they carry it out: that it is carried out using the VAR coefficient distribution that is restricted by the DSGE model.

It may be therefore possible to raise the power of this ('restricted') IIW test further. We suggest two ways in which this might be achieved: 1) extending this IIW test to include elements of the variance matrix of the coefficients of the auxiliary model; 2) including more of the structural model's variables in the VAR, increasing the order of the VAR, or both.

The basic idea here is to extend the features of the structural model that the auxiliary model seeks to match. The former is likely to increase the power of this IIW test, but not the LR test, as the latter can only ask whether the DSGE model is forecasting sufficiently accurately; including more variables is likely to increase the power of both. There is, of course, a limit to the number of features of the DSGE model that can be included in the test. If, for example, we employ the full model then we run into the objection raised by Lucas and Prescott against tests of DSGE models that was noted above, that "too many good models are being rejected by the data". The point is that the model may offer a good explanation of features of interest but not of other features of less interest, and it is the latter that results in the rejection of the model by conventional hypothesis tests. Focusing on particular features is a major strength of the restricted IIW test.

To illustrate the effects of extending the features that are tested, consider again the simple example explored earlier, namely, the 3-equation NK model. This has a 3VAR(1) reduced-form solution with 9 coefficients each of which is a different non-linear combination of the 8 structural and AR parameters. No extra information is therefore obtained by raising the order of VAR. The number of variables in the auxiliary model is not an issue as all three variables are included. The power of the restricted IIW test is reported in Table 4. The results show that increasing the order of the VAR from one lag (the number in the solution of the model) to two lags has no effect on power.

3-equation NK model — no lags (VAR(1) reduced form) Rejection rates at 95% confidence: T=200		
	3 variable VAR(1)	3 variable VAR(2)
True	5.0	5.0
1%	4.9	4.3
3%	7.3	7.1
5%	16.1	21.7
7%	37.0	40.3
10%	73.3	76.3
15%	99.4	99.8
20%	100.0	100.0

Table 4: Comparing power due to VAR order (3-equation NK model with no lags)

Consider now including an indexing lag in the Phillips Curve. This increases the number of structural parameters to 9 and the reduced-form solution is a VAR(2). The power of the restricted IIW test is reported in Table 5. Increasing the number of lags in the auxiliary model has clearly raised the power of the test.

3-equation NK model — with lag (VAR(2) reduced form) Rejection rates at 95% confidence: T=200		
	3 variable VAR(1)	3 variable VAR(2)
True	5.0	5.0
1%	10.6	6.0
3%	20.7	19.5
5%	47.5	57.9
7%	65.6	91.2
10%	89.6	100.0
15%	98.8	100.0
20%	99.9	100.0

Table 5: Comparing power due to VAR order (3-equation NK model with indexing lag)

This additional power is related to the identification of the structural model. The more over-identified the model, the greater the power of the test. Adding an indexation lag has increased the number of over-identifying restrictions exploitable by the reduced form. A DSGE model that is under-identified would produce the same reduced-form solution for different values of the unidentified parameters and would, therefore have zero power for tests involving these parameters.

In practice, most DSGE models will be over-identified, see Le et al. (2013). In particular, the SW model is highly over-identified. The reduced form of the SW model is approximately a 7VAR(4) which has 196 coefficients. Depending on the version used, the SW model has around 15 (estimatable) structural parameters and around 10 ARMA parameters. The 196 coefficients of the VAR are all non-linear functions of the 25 model parameters, indicating a high degree of over-identification.

The over-identifying restrictions may also affect the variance matrix of the reduced-form errors. If true, these extra restrictions may be expected to produce more precise estimates of the coefficients of the auxiliary model and thereby increase its power. It also suggests that the power of the test may be further increased by using these variance restrictions to provide further features to be included in the test.

6 Testing pseudo-true models for mis-specification

In recent years Le et al (2011, 2016a) have suggested a method for testing macroeconomic models that has substantial power in the small samples typically available for macro data. They have exemplified this using Monte Carlo experiments on two major types of macro model, the 3-equation New Keynesian and the Smets-Wouters multi-equation New Keynesian, both DSGE models. Above we have summarised a number of these experiments and explained why the IIW test is so powerful compared to the standard LR test which has up to now dominated econometric thinking.

As we have seen, the IIW test, which we have explored in previous work, has very considerable power in small samples in testing over-identified macroeconomic models, typically of the DSGE variety — Le et al (2016a). Indeeed, one can continuously raise the power of the test by steadily adding to the auxiliary model more separately-identified features implied by the structural model. However, while we can set the power of the test very high, in practice we will choose the level of power to be not so high that we cannot find a tractable model that will pass. Assuming that we do find such a model, we can then ask how false would a model of that type have to be in order to be rejected 100% of the time. We do this by Monte Carlo experiment, creating a model of this type, treating it as pseudo-true, generating many samples from it and then measuring the rejection rate as this model is progressively falsified. Suppose for example that rejection reaches 100% when the model is 7% false. One can then carry out a robustness analysis on any set of policy proposals, asking how they would perform within a 7% range of the model which has passed the test. One can argue that this model has to be within 7% of the pseudo-true model.

Thus the test procedure exploits the test's power in order to inform the users of the margin of effectiveness of their 'policies' on the 'policy targets'. It does so by simulating a potentially pseudo-true model to discover its power function; this model is in the region of the estimated model, so that the power function should apply indifferently to models in this region, as can easily be verified. This power function is then used to analyse the effectiveness margin on the estimated model.

Should users be worried about possible mis-specification of their pseudo-true model? In investigating this we assume that mis-specification may take the form of a failure to include in our pseudo-true model features from a more complex model that is treated as the DGP that in fact generates the data.

We set up a Monte Carlo experiment in which the DGP generating the data is such a more complex model. Our starting point will be as in Le et al (2011, 2016a) the well-known Smets-Wouters model on US data from the early 1980s. Le et al (2011, 2016b) found that this model when modified to allow for a competitive sector and for banking, can explain the main US macro variables, output, inflation and interest rates well. It is this model and versions similar to it that we have used in previous Monte Carlo experiments.

To this model we can add money and a regime shift contingent on the state of the economy, from the Taylor Rule to the zero bound, as in Le et al (2016b). This makes the model's parameters state-contingent so that it has this form of nonlinearity. We then treat this nonlinear model as if it were the DGP generating the data. Using the Indirect Inference test procedure with a VAR as the auxiliary model we estimate the power function for the falseness criterion we described above in order to assess the sensitivity of this function to the presence of greater nonlinearity in the true model than the 'assumed true' DSGE model we started with.

We looked at three very similar models, of varying complexity. All three are based on the Smets-Wouters model as modified in Le et al (2011). Model 1 is that model exactly. Model 2 is that model with the financial shock replaced by the Bernanke, Gertler and Gilchrist model of banking (the 'financial accelerator'). Model 3 is the same model, together with an extension in which collateral is required and base money acts as cheap collateral, and the additional nonlinearity of the zero bound constraint, triggered whenever the Taylor Rule interest rate falls below a low threshold. These last two models are set out in Le et al (2016b).

From the point of view of 'realism' and 'truth' we regard model 3 as the most realistic; model 2 as a linear approximation to it; and model 1 as a simpler approximation to model 2. We investigate whether in each case the simpler, less realistic model can be treated as a valid approximation to the more realistic one.

We carried out the following three experiments with sample sizes between 75 and 200, and with 1000 sample replications. Our IIW test was based in all cases on the coefficients of a three variable (y, π, R) VAR(1) (including the three variances, as is the usual practice in applying these tests; so 12 values in all). Table 6 shows that in all cases there is an overwhelming probability of rejection, close to 100% and falling to 80 with the smallest sample size of 75 and the two models closest in complexity (models 1 and 2).

The models we used for this experiment were those that Le et al (2011, 2016a,b) where the parameters were estimated by indirect estimation using US data. In practice when we carry out the IIW test on a model for a particular sample, in practice, we re-estimate the model. We therefore carried out the test on this basis, re-estimating the tested model on

3 variable VAR(1)	T=200	T=125	T=75
1) Generating data from model 2 as true data, testing model 1 by IIW	99.9	98.2	79.7
2) Generating data from model 3 as true data, testing model 2 by IIW	100	99.3	95.2
3) Generating data from model 3 as true data, testing model 1 by IIW	100	100	98.1

Table 6: Testing mis-specified models: percentage rejection rates using IIW

each sample generated by the true model. As this is highly time-consuming, we did this selectively for two model pairs and different sample sizes: for a sample of 125, model 3 is the complex model and model 1 is the simpler model; with a sample of 75, model 2 is the complex model and model 1 is the simpler model.

Table 7 shows that the rejection rate for model 1 for the first pair is still 100%, even though there is some increase in closeness. Thus, even for a sample as small as 125, the rejection of mis-specification remains virtually 100%.

Transformed Wald	Min	Max	Mean	Rejection rate (Critical value=1.645)
Re-estimated by II	1.686	79.314	21.739	100%
Original estimates	2.459	$9.07E+15$	$2.57E+14$	100%

Table 7: Transformed Wald for model 1 when tested on model 3 samples, T=125

In the second case, Table 8, where model 1 is tested using data from model 2 with a sample of only 75, it is somewhat harder to distinguish model 1 from model 2, the two closest models: the rejection rate falls to 68.6%. However, rejection is still overwhelmingly probable.

Transformed Wald	Min	Max	Mean	Rejection rate (Critical value=1.645)
Re-estimated by II	1.480	8.412	2.409	68.6%
Original estimates	1.291	11.818	2.909	79.7%

Table 8: Transformed Wald for model 1 when tested on model 2 samples, T=75

We would not have found this result had we used a Likelihood Ratio test as this has much less power than the IIW test, even if falsification is done in exactly the same way.

Furthermore, if a mis-specified model is re-estimated on the sample data, the power of the test declines considerably due to the way the re-estimated model is then able to mimic the data. Hence passing a such likelihood ratio test may carry little weight and give the impression that there is little point in testing DSGE models since their misspecification is unlikely to be detected.

Our findings are the opposite using the IIW test. This test is found to give a very high rejection rate of mis-specified DSGE models, even with very low sample size.

What we find therefore is that treating DSGE models as pseudo-true, our IIW test can establish for users a) whether they have a model that can predict relevant features of data behaviour and if so b) the bounds within which they can be sure of its specification and parameter values. With the widely-used DSGE model examined here, we found that if the pseudo-true model passes the test on the behaviour of three key macro variables, the power of the test largely guarantees that no other specification can be pseudo-true and that its parameter values lie within a 7-10% region of the pseudo-true ones.

7 Conclusions

This paper has attempted to address the growing gulf between traditional macroeconometrics and the increasingly dominant preference among macroeconomists to use DSGE models but not to test them as they are likely to fail conventional statistical tests. Instead of using classical estimation and inference procedures, DSGE models are either calibrated or estimated by Bayesian methods in which prior information tends to dominate sample information. The dominance of DSGE models among macroeconomists followed from Lucas' famous Critique of macroeconometric models as being reduced form models - and so subject to structural change - rather than structural models. The strength of traditional macroeconometric models is that they can be specified - often by giving them flexible dynamics - so that they are not rejected by the data. The choice therefore seems to lie between statistically valid representations of the data that are not theoretically valid and more theoretically valid models that are not statistically valid, and therefore not worth testing.

We have argued that both types of model are better regarded as approximations to "reality" and are therefore at best pseudo-true and worth testing if only to evaluate how close they are to providing a valid statistical representation of the data. Based on a decade and a half of previous work, we have described a way to test DSGE models using indirect inference. This involves formulating an auxiliary model that is a pseudo-true approximation to the solution of the DSGE model and comparing estimates based on data simulated from the DSGE model with those derived from actual data. This provides both a generalisation of previous methods of assessing the performance of DSGE models and a formal statistical test of the model.

We addressed the issue of how best to test an already estimated DSGE macroeconomic model as judged by the power properties of the test. A key finding is that, in small samples, a test based on indirect inference (in particular, the IIW test) appears to have much greater power than a likelihood ratio test based on the observed data. This finding is at first sight a little puzzling as under direct inference with the observed data an LR test and a Wald test of all of the coefficients are equivalent, while the IIW test using indirect inference is asymptotically equivalent to the LR test and so has the same power in large samples. We attempted to explain why this result occurs.

We find that the difference in power in small samples of the LR and IIW tests may be attributed to two things. First, in the power calculations, the simulated data is usually obtained differently for the two tests. The structural disturbances of DSGE models are commonly found to be serially correlated. In order to improve the fit of the model, the structural disturbances are specified to allow them to be generated by autoregressive processes. As the simulated structural errors are also serially correlated, in calculating the power of the LR test for the false DSGE model, the autoregressive processes of the resulting simulated structural errors are normally re-estimated. This "brings the model back on track" and as a result undermines the power of the LR test as it is, in effect, based on the relative accuracy of one-step ahead forecasts compared with those obtained from an auxiliary VAR model. The fact that the serial correlation structure of the disturbances changes with each degree of falsification that is imposed shows the data-mining role played by the serially correlated disturbances in the original set of estimates.

Second, the additional power of the IIW test may arise from the use of a covariance matrix of the auxiliary model's coefficients determined from data simulated using the restrictions on the DSGE model. These may both give more precise estimates of these coefficients and provide further features of the model to test. The greater the degree of over-identification of the DSGE model, the stronger this effect. This suggests that for a complex, highly restricted, model like that of Smets and Wouters, the power of the indirect inference IIW test can made very high even in small samples. Because a test of all of the properties of a DSGE model is likely to lead to its rejection, it may be preferable to focus on particular features of the model and their implications for the data. This is where the IIW test has another clear advantage over the LR test.

Finally, we addressed the issue of detecting mis-specification in pseudo-true models. We found that the power of the IIW test on typical small samples is high enough to detect whether DSGE models give significantly different estimates of the auxiliary model from actual data and so are mis-specified. If a DSGE model passes the test, the probability of its being mis-specified, and hence unable to explain estimates of the auxiliary model derived using actual data, seems to be vanishingly small. We concluded that it is possible to test a DSGE model in a normal frequentist way as if it were the true model while recognising that it is, at best, only a good approximation, or pseudo-true, model of the economy. We suggested using the power function of the test statistic to assess the robustness of the DSGE model as a representation of the economy.

References

[1] Blanchard, O. J., and C. M. Kahn. 1980. The solution of linear difference models under rational expectations. Econometrica 48 (5): 1305–11.

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

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

[4] Christiano, L. J., 2007. Comment on 'On the fit of new Keynesian models'.Journal of Business & Economic Statistics, 25,143-151.

[5] Clarida, R., Gali, J. and Gertler, M., 1999. The Science of Monetary Policy: A New Keynesian Perspective. Journal of Economic Literature 37(4), 1661–1707

[6] Cox, D. 1961. Tests of separate families of hypotheses. Proceedings of the Fourth Berkeley Symposium on Mathematical Statististics and Probablity

[7] Cox, D. 1962. Further results on tests of separate families of hypotheses. Journal of the Royal Statistical Society, B24, 406-424.

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

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

[10] 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.

[11] 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.

[12] 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.

[13] Evans, G. W. and S. Honkapohja (2005). "Interview with Thomas J. Sargent". Macroeconomic Dynamics, 9(4): 561–83.

[14] Fernandez-Villaverde, J., Rubio-Ramirez, F., Sargent, T, and Watson, M. ,2007 'ABCs (and Ds) of Understanding VARs', American Economic Review, pp 1021-1026.

[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: 3–43.

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

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

[19] Le, V.P.M., D. Meenagh, P. Minford, M. Wickens, 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.

[20] Le, V.P.M., Meenagh, D., Minford, P.,Wickens, M., and Xu, Y.,2016(a). Testing macro models by indirect inference: a survey for users. Open Economies Review 27, 1-38.

[21] Le, V.P.M., D. Meenagh, P. Minford, M. Wickens,2016(b). Monetarism rides again? US monetary policy in a world of Quantitative Easing. Journal of International Financial Markets, Institutions and Money 44(C), 85–102.

[22] Le, V.P.M, P. Minford, M. Wickens, 2013. A Monte Carlo procedure for checking identification in DSGE models. Cardiff Economics Working Papers E2013/4,

[23] Liu, C. and Minford, P., 2014, How important is the credit channel? An empirical study of the US banking crisis. Journal of Banking and Finance, 41, 119-134.

[24] Minford, P., M. Wickens, and Y. Xu, 2016.Comparing different data descriptors in Indirect Inference tests on DSGE models, Economics Letters, forthcoming.

[25] Sims, C., 2002. Solving linear rational expectations models. Computational Economics 20(-2): 1-20.

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

[27] Smets, F., Wouters, R., 2003. An Estimated Dynamic Stochastic General Equilibrium Model of the Euro Area. Journal of the European Economic Association, 1(5), p1123–1175.

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

[29] Wickens, M., 2014. How useful are DSGE macroeconomic models for forecasting? Open Economies Review, 25, 171–193.

[30] Wright, S., 2015. The predictable and the unpredictable economy. Presentation at European Monetary Forum Conference, Erasmus University, Rotterdam, March 20-21, 2015.