Testing macro models for policy use — an insurrection in applied modelling

Patrick Minford*
Cardiff University and CEPR

September 2015

This lecture is about whether and how we can test the macro-economic models that we use for policy. My colleague and co-author, Mike Wickens, gave a lecture at the MMF last year in which he explained how times had changed in the treatment of macro-economic models. Once upon a time large-scale econometric models were estimated and tested by classical econometric methods. But then the arrival of rational expectations and Lucas' critique meant that we did not trust them for policy-making any more because their parameters were those of aggregate supply and demand curves, which would change with changing policy regimes. The next two decades were spent building micro-founded models whose parameters were thought to be structural. However, when estimated and tested by the classical methods, these models were generally rejected — in a famous quotation from Sargent 'Lucas and Prescott told me we were rejecting too many good models'. There followed two decades or so, lasting until today, where many macromodels were calibrated and compared informally with 'stylised facts'. More recently, Bayesian estimation has brought these models closer to the data, using the calibrated values as priors. But Bayesians make no claims about these models' testability in the classical sense; rather they treat all models as false and evaluate different models' probability.

However, for policymakers this situation, I will argue, is quite unsatisfactory. They would like to know whether the macro models they are using for policy analysis are good enough for this purpose. They would not reject Lucas' critique: their models must contain only structural parameters, that therefore are not affected by the policies they are thinking about. It is for this reason we and they use DSGE models whose parameters we can reasonably claim are structural.

But how are we to choose between the many different DSGE models on offer? Everyone here is familiar with the different schools of thought one meets in macroeconomics — Keynesian, freshwater, Austrian, the list goes on. Since the Great Recession crisis others have been added, with many on the fringes of the subject claiming that it has totally failed and we should go back to ad hoc models built on insights from economic history. This would violate the Lucas Critique and so is a counsel of despair. So what are we to do, faced with this plethora of theories?

There are those who would say: create an unchallengeable theory to which all will then subscribe. We are clearly many miles away from being able to do that. Indeed that was the ambition of the Real Business Cycle theorists thirty years ago. Yet today we face even more disagreement than then.

The Bayesians would say we cannot really judge between them because they are all wrong and so we can come up with some probabilities of each being right. But if one is a policy maker this means that you still have no idea how your policy will turn out since these models contradict each other dramatically. You might succeed very well or fail utterly. So this approach is not much use to you.

So I suggest we need to test these models against the data in the hope that some will be eliminated and one will emerge as closest to the data. This one would then be the survivor in Popper's sense, to go on to be tested on later data. As a policymaker you can then put your faith in this model and think about how your policies will work, confident you have the causal mechanism of their transmission.

In this lecture I am going to explain how one can do this for macro models with a substantial amount of power to reject false models so that when you find the model you do not reject you can have good confidence it is close to the true model for the purposes you have in mind. A group of us at Cardiff have developed these methods so that we can now give anyone a programme that will carry them out for a

*I am grateful to my colleagues and coworkers, Mai Le, David Meenagh, Mike Wickens and Yongdeng Xu, for the substantive issues that this lecture reviews, all of it based on work done with some or all of them. Details of this work are of course to be found in the numerous references in Le et al, 2015, on which this lecture is based. I am also grateful for their comments on and contributions to this lecture.

wide range of models. These methods have been made possible by modern computing capacity. Even ten years ago they would have been too time-consuming to apply. Of course for most of my career we had no decent tools to evaluate models in toto; it is really for this reason that there is so much ongoing controversy about models. There has never been a way to settle disputes by going to the data.

Many of you will be familiar with Likelihood Ratio tests and will perhaps protest that it has been possible to carry these out with existing programmes for some time. This is true. But there are two advantages in the methods we propose over Likelihood Ratio tests. The first is that our method has much more power. The second is that it can be focused on the purposes you want the model for in a way that the Likelihood Ratio test cannot; thus the latter test asks how close the model gets to the data, all the data, for a set of variables, whereas the test we propose asks how close the model gets to the behaviour of a set of variables in particular respects, such as over the business cycle or in its growth aspect. Policymakers want models that capture such behaviour and do not care if they do not capture other behaviour. They want to find a model for policy purposes that is both consistent with the relevant features of the data and also has good power against poor policy outcomes. We will see later how our methods get them closest to this objective.

In this point we encounter what I will call the 'Friedman utility' of tests. You may recall that Friedman in his 1953 paper on methodology argued that we should test models, not on their literal truth, but on their ability to explain the data features we designed them to explain — the ones we were interested in and concerned about. A model was, he said, an 'as if' construct, not meant to be literally true but to capture some essential aspects of behaviour by assumptions that could mimic that behaviour 'as if' it was true. He had in mind that idea that models were gross simplifications of or abstractions from reality, constructed to have 'explanatory power', by which is meant getting a lot of explanation from as simple a construct as possible. Critical to this approach is the choice of the aspects of reality to be explained. Many economists are familiar with 'Likelihood' as the yardstick; but of course this is the likelihood of the model fitting just one aspect of reality, namely the likelihood of observing the data conditional on the model. Effectively this tests whether the model is close to the data in a forecasting sense; the measure is based on the reduced form errors of the model. However a macro policymaker wanting to make good new policy — assumed in this lecture to be the arbiter of taste — is not interested in forecasting performance but rather in whether the model behaves causally like the real world, by which we mean the likelihood of the behaviour of key macro variables in the data conditional on the model. Since that data behaviour is the reduced form of the model (or an approximation to it), we are checking whether the model's reduced form parameters, which are functions of the model's causal structure, are close to the data's. Such a correspondence implies the model's causal structure cannot be rejected as the one generating the data. It is this correspondence that is tested for in Indirect Inference, where we use a Wald-type statistic (an IIW) to measure the gap between what the model says the data behaviour should be and what the data behaviour actually is. For this we usually use the data behaviour as described by the estimated VAR coefficients of the data.

This distinction matters a lot in practice. Different aspects of reality being tested yield different results for the tests of models. The et al. below shows the scatter diagram of Data Likelihood (measured by the Likelihood Ratio of the model to an unrestricted VAR) versus Data Behaviour likelihood (measured by the IIW on the VAR coefficients) for Monte Carlo samples from the SW model. The correlation is 0.008, or essentially zero! This means that had you rejected your true DSGE model on an LR test, you would have been almost certain not to reject it on an IIW test; and vice versa. As the model you are testing becomes more False the test results become slightly more correlated. But even when the model is 10% False they give significantly different results as can be seen from the next diagram, which plots the scatters up to 20% False. The saving grace is that once your model becomes this False, both tests reject it much of the time. But the problem for you as a policymaker is that the success of your policies may depend on the model being much less false than this in terms of your key yardstick.

Another yardstick that has become familiar is Impulse Response Functions; these too could be used to create an IIW. For example if you are on the MPC you will be concerned that your monetary impulse has the desired effects; this yardstick is related to the VAR coefficient yardstick above but is specialised differently in terms of shock and variables affected. If you use it you must be careful to use the joint distribution over the IRFs involved and not evaluate them separately.

What this discussion should reveal is that this is similar to the idea of comparing DSGE models' simulated behaviour with 'stylised facts'. The difference lies in the use of the model's joint distribution over these stylised facts rather than informally deciding whether the group of facts each are 'close' to the simulated facts. But the basic idea of this comparison in effect revives Friedman's ideas about testing, putting the facts that interest the user centre stage in the test.

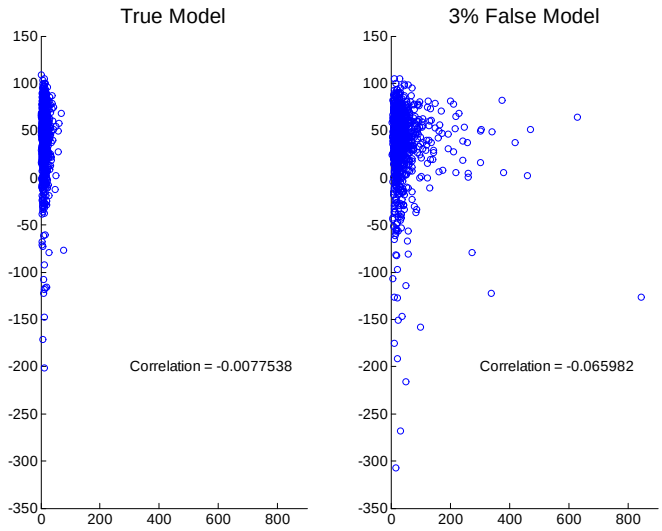

Figure 1: Scatter Plots of Indirect Inference (Wald, horizontal axis) v. Direct Inference (LR, vertical exis) for 1000 samples of True Model (3 Variable VAR(1))

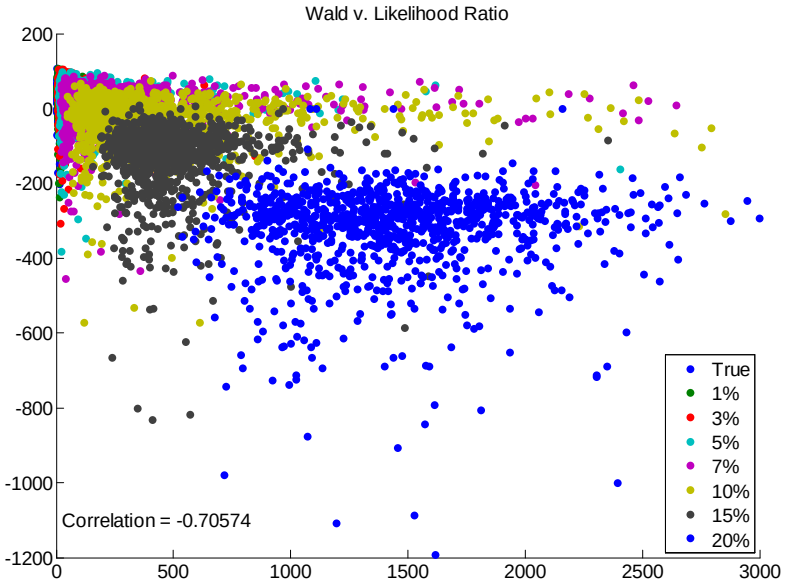

Figure 2: Scatter Plots of Indirect Inference (Wald) v. Direct Inference (LR) for True and False Models (some outliers taken out for clarity of scale)(3 Variable VAR(1))

The methods we have developed are based on and organised around Indirect Inference. This originated with the work of Tony Smith (1993) as an estimation technique for non-linear models where the data behaviour is described in some purely descriptive way, the most usual being a VAR or in the case of non-stationary data a VECM. This description is known as an auxiliary model. Then the model being estimated, which in our case here is a structural DSGE model, is simulated by bootstrapping to enable its predictions for the auxiliary model to be compared with the one found in the data; in estimation the model parameters are varied until the simulated auxiliary model is as close as possible to the one in the data and in testing the auxiliary model coefficients estimated on the data are located in the simulated model's distribution of them, with rejection if they come outside the critical boundary. As already noted the methods can be flexibly focused on whatever features of the data behaviour the user is concerned to explain. In effect the method asks, via the IIW, whether this DSGE model could statistically be the one generating the facts of behaviour we observe. But this is just the starting point for examining the policymakers' model; they also need to know how sure they can be about the model they have at the end of this estmation and testing process. For this they need to discover the power of their test by Monte Carlo experiment; and then check how vulnerable they could be to model mis-specification.

In the rest of this lecture I will describe how we apply this to the widely-used Smets-Wouters model, discuss the power of the method and compare it to some alternatives, and end by reviewing some of the findings we have made so far about macro models.

1 The concept of test power

Once one has decided to test a model in a certain respect, the question then arises of how powerful the test is. This means what percentage of the time (i.e with what probability) will it reject false models in small samples — small because these are the ones we encounter in macro-modelling. We can establish this percentage by replicating the test across many repeated samples. The test is set up (i.e its critical value chosen) so that it will reject the True model at the chosen confidence level, so if at 95% then it will reject 5% of the time. This is the 'size' of the test. Having chosen the size we can then ask how the rejection rate rises as the model becomes more and more false. The faster this rises with falseness the higher the power of the test. For judging this power we construct Monte Carlo experiments where we generate a large number of samples from a True model and construct False models by altering parameter values alternately by + or -x% from their true values. We then check how frequently the data samples reject the False models as x rises.

We could of course ask whether the model can pass a test of mirroring all data behaviour according the fullest possible detailed descriptive VAR. For example the SW model has a reduced form for 7 variables which is a VAR(4) and has 196 coefficients. So if one uses a 7-variable-VAR(4) as the auxiliary model for the SW case we are asking whether the SW model's simulated reduced form is like the data-based reduced form it would have. This is a very demanding test and no model is likely to come anywhere near passing it. If one assumes the SW model is true then the power of the test is massive; even a 1% degree of falseness leads to 100% rejection.

But users really do not care whether the SW model can get all aspects of reality exactly right, especially as they know they will never find such a model. Instead users care about specific aspects of reality. So we could ask whether the SW model captures the data behaviour very approximately (say by a VAR1) of three key macro variables, output, inflation and interest rates. Then the test concentrates on the minimum practical requirements of the user and makes it possible for the user to find a model that gives what is needed. What users would then like is for the test to reject frequently any models that are false to a degree likely to abort their policy changes. How false would that be? We would need to experiment with the policy changes and the tolerances of the policy maker to judge this. But we might say for illustration that a policymaker could well tolerate a falseness in the model's structure of up to 5%. Then a rejection rate at or above the 50-70% range at this level of falseness could well provide some security in choosing a model that passes the test.

What I am trying to describe is a practical procedure for a user to choose a model. Users must discover the tolerance of falseness they can live with, and then they can proceed to discover whether their test will deliver a model that reliably delivers better than that tolerance level — which means it has high power above that falseness level. The toolkit of indirect inference does this job, we believe.

2 An insurrection in testing

The test we are proposing is an unfamiliar type of Wald test, in which we compare the estimates of a VAR on the data with the estimates we would get from the DSGE model when simulated. Statisticians among you will know that a standard Wald test of the distance between an unrestricted parameter estimate and a restricted one is derivable from a Likelihood Ratio test between the unrestricted and the restricted model — essentially they are the same test. How then is it possible for our Wald-type test to have more power than a Likelihood Ratio test?

One reason lies in the way the two tests are carried out in practice. When an LR test is done, it is usual to reestimate at the least the error process in the model to bring them 'on track'. This is not usually done with a Wald test. Unfortunately by bringing the model on track it does artificially better in forecasting the data and this greatly reduces the power of the LR test.

We can put health warnings on LR tests to avoid doing this. Yet even then the Wald-type test, the IIW, we propose has greater power. Let me illustrate this from the Smets-Wouters model. In the Monte Carlo experiment reported in Table we falsify all parameters of the True model (including the AR parameters of the errors) alternately by + or -x% as described above and we do this in exactly the same way for all the methods described. For convenience we keep the True innovations unaltered throughout.

VAR — no of coeffs	TRUE	1%	3%	5%	7%	10%	15%	20%
DIRECT INFERENCE								
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
INDIRECT INFERENCE with unrestricted covariance matrix								
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
INDIRECT INFERENCE with restricted covariance matrix								
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

Table 1: Comparison of rejection rates at 95% level for Indirect Inference and Direct Inference

Table 1 show that at say 5% falseness the power of the test doubles compared with the LR test and also compared with the standard Wald test.

It turns out the reason is that our Wald-type test makes the DSGE model being tested the null hypothesis whereas the standard Wald test makes the unknown true model embodied in the data the null. This is an important difference because we do not know what the true model is; hence to find out the distribution of the VAR coefficients it implies we must use the estimates we have from the data. We find this by bootstrapping the data-based VAR — ie the estimates we got from unrestricted estimation of the VAR on the data. Call this the Unrestricted distribution. What this is doing is reestimating the VAR on new bootstrap data, created each time by redrawing the VAR innovations.

By contrast when we use the DSGE model being tested as the null, we bootstrap its own innovations (which we back out of the data) and simulate the model with all its restrictions every time. The resulting bootstrap samples reflect not just the different innovations but also the effects of these innovations when interacted with the model's restrictions; the latter generates data which will produce different VAR coefficients in each sample, in its own right — the Unrestricted model always uses the same (estimated)

VAR coefficients to generate the bootstrap data[1]. We then estimate the implied VAR coefficients. Call the distribution of these the Restricted distribution. The difference between the Restricted and Unrestricted distributions lies in the way the Restricted uses the model's own restrictions to generate the bootstrap samples while the Unrestricted uses the estimated VAR coefficients. Another way of putting it is that the Restricted distribution has a variance matrix of the VAR coefficients derived from the DSGE model whereas the Unrestricted has one derived from the data sample VAR estimates.

We illustrate the situation in the following two diagrams.

What we show in the first set of graphs below (Figure 3) are two distributions for 2 VAR coeffs taken from a 3-variable VAR1 for the SW model: the two are the own–lag coefficients for the short-term interest rate and for inflation. We generate one sample from the SW model. We then estimate the 3VAR1 coefficients on that true data sample, and we find the distribution for the two coefficients above by bootstrapping the VAR innovations. We then find the same distribution when restricted by the True model, by bootstrapping the *structural* innovations generating that sample. The graphs below show the densities of the joint distribution of the two coefficients. What one observes is that the restricted distribution is both smaller in size and also more elliptical than the unrestricted when the model is true. We then falsify the model structural parameters (including the error AR coefficients) by 5% and 10%; now we bootstrap the same structural innovations and find the resulting (restricted) distributions for the two VAR coefficients. As the model becomes more False the restricted distributions become more elliptical and their variance rises; they also rotate somewhat as the covariance changes. Notice that when the model is False but still close to the True the restricted distribution is both more elliptical and has lower variance than the unrestricted. Both features give it more power at low Falseness as can be seen from the next graph.

The second graph, Figure 4, shows how this impacts on the Wald tests' power when testing a model that is 5% False. We are looking down from above on the true distribution of the two VAR coefficients, generated by Monte Carlo means from the True model using its true structural innovations and parameters; this distribution therefore shows the true sample population. The green dot in the Figure shows the mean of two VAR coefficients implied by the 5% False model. We can now test this False model two ways on a given data sample. One way involves taking that sample's unrestricted VAR1 representation and bootstrapping it; an example of the 5% contour of such a bootstrap distribution is given by the dashed green line. The thick green line shows the frontier at which the 5% False model is just rejected by the data samples on the line; in effect along it we are recentering the same dashed green line. Now consider the red ellipse. This shows the 5% contour of the False model distribution, using the same True innovations.

The two ways of testing the False model give different rejection rates. The data samples to the left of the thick green line are those that reject the False model under the first method — where we use the Unrestricted distribution from the data. The data samples to the left of the red ellipse are those that reject the False model under the second method — where we use the False model Restricted distribution. Plainly the second method gives much greater power. Essentially we are illustrating here what happens when we use the SW model in a Monte Carlo example below (we have assumed for the illustration that the power in respect of these two VAR coefficients is the same as for the whole set of VAR coefficients tested below); there we find that the power roughly doubles as we move from method 1 to method 2 at 5% Falseness.

We can also look at Figure 5 to see how the rotation of the ellipse due to the changing covariance of

[1] This can be seen formally by noting that the α coefficients reestimated from the ith bootstrap of the unrestricted VAR (found from the T data sample) are:

$\widehat{\alpha}_i^{UNR} = f_{OLS}\{\widehat{y}_i^{UNR} = \widehat{y}_i^{UNR}[\widehat{\alpha}_T(\theta, \epsilon_T), \eta_i]\}$

where θ is the vector of structural model coefficients (including those of the error processes), ϵ the vector of structural innovations, η that of VAR innovations, f_{OLS} is the OLS estimator function to obtain the α from the y.

Now compare the analogous estimates with restricted VAR bootstraps:

$\widehat{\alpha}_i^{RES} = f_{OLS}\{\widehat{y}_i^{RES} = \widehat{y}_i^{RES}[\theta, \epsilon_i] = \widehat{y}_i^{RES}[\widehat{\alpha}_i(\theta, \epsilon_i), \eta_i]\}$

We can see that these α OLS estimates come from y simulated directly from the structural model and that these in turn have a VAR representation consisting of two elements, the direct effect of η as before plus the indirect effect of ϵ, θ on α. It is this last extra element that creates the rich variation in resampled data behaviour reflecting the DSGE model's structure interacting with the structural errors.

In terms of the example discussed below in the text where we consider the own-persistence VAR parameters of inflation and interest rates, what is happening is that with restricted bootstraps model-simulated samples in which inflation is not persistent will typically also be those where interest rates are also not persistent, and vice versa, because the model implies a strong connection between the two variables; thus estimated covariation in these own-persistence VAR parameters ($\widehat{\alpha}_i(\theta, \epsilon_i)$) will show up in the resampled data. With unrestricted bootstraps this covariation is not included; instead the VAR parameters generating the data are held constant at those in the data sample, $\widehat{\alpha}_T(\theta, \epsilon_T)$. Notice that the variation due to the direct effect of the innovations, η_i, is the same in both cases.

the two VAR coefficients can raise the power of the IIW test. As the ellipse rotates, it covers less and less of the True model sample points. Thus not just the distance of the model's mean VAR coefficients from the True mean of the data-based ones but also the shape of the model's distribution for these coefficients and its rotation (both due to the model-implied covariance between the coefficients) with rising falseness determine the power of the test — i.e. how many of the data sample points it fails to cover. With the standard UNR Wald the shape and rotation is fixed regardless of Falseness — one is always using the same distribution based on the True data sample — and so only the distance varies with Falseness.

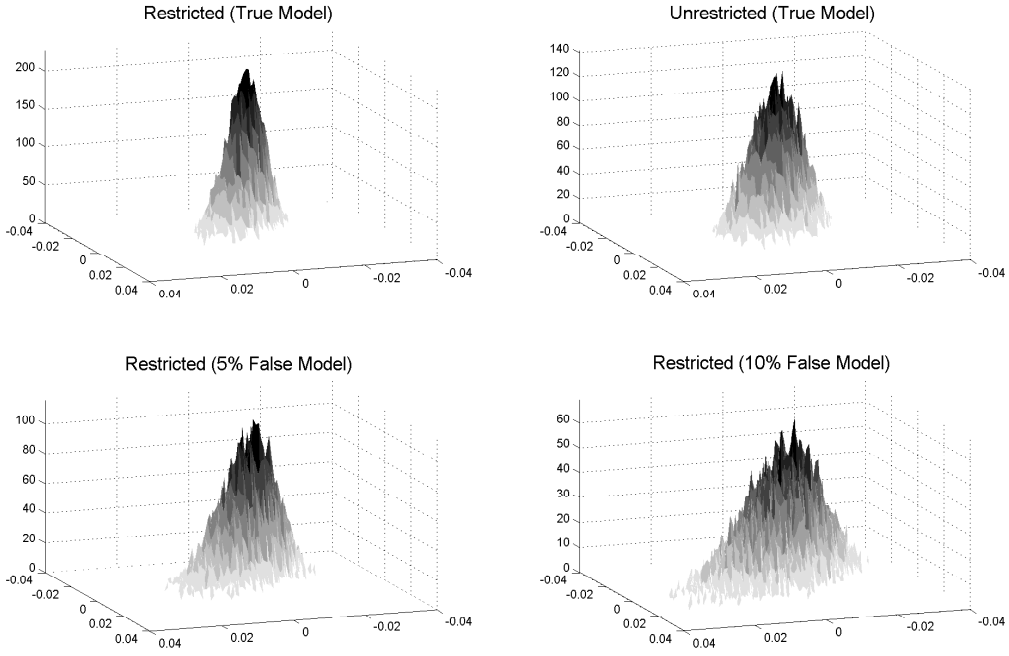

Figure 3: Restricted VAR and Unrestricted VAR Coefficient Distributions

Why would eminent statisticians not have seen this helpful increase in power from this non-standard use of the Wald? I think there may be three reasons. First, asympotically there is no difference between the IIW and the standard Unrestricted Wald when the model is True; this is a small-sample result and this has not typically been a focus for statisticians because analytic results cannot be obtained for small samples. Second, the Restricted distribution can only practically be obtained by the bootstrapping methods used here and these methods have not been popular with statisticians at least until quite recently, whereas the conventional Wald distribution can be obtained by the usual asymptotic methods — the variance matrix is estimated along with the VAR on the data sample. Third, because as we have noted there is a power trade-off between the variance and the restrictions, effectively between the diagonal and the off-diagonal elements in the variance matrix: statisticians may well have assumed that the diagonal elements were the more important.

If we were asked why we discovered the higher-power of the non-standard Wald test, our reply would have be that it it came out of a long process of investigation. We began with a Friedman-style approach to testing in which we explored the distributions implied by the model. This led on to our results and comparisons with other methods. Finally we managed to work out why we were getting these results.

2.1 Exploiting the extra power of the Wald-type test with DSGE-model-restricted variance matrix

Thus when we eliminate the difference in procedures and test like-for-like we found the two tests are reasonably comparable in power when the indirect inference test is performed using the unrestricted Wald test which uses the variance of the unrestricted VAR (auxiliary) model. This turns out to be because the tests are approximately equivalent on a like-for-like basis. However, we showed above that

extra power is delivered by the IIW test set out here, under which the DSGE model being tested is treated as the null hypothesis: in this case the Wald statistic uses the variance restricted by the DSGE model under test. This gives this restricted Wald test still greater power.

It may be possible to raise the power of the Wald test further. We suggest two ways this might be achieved:

1) extending the Wald 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 the restricted Wald test, but not the LR test, as this last 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 "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 Wald test.

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 2: 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 Wald test is reported in Table 3. 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 3: 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.

3 Our methods in practice

I have argued so far that the IIW method described here enables economists with a particular purpose — I have taken it to be policy formation — to find a model that generates the sort of behaviour of interest to them. By focusing the test narrowly on this behaviour these economists can find a model that passes the test and then can be confident that the test would have rejected the model if it was more than x% False, where x is quite low. They can then explore whether this Falseness tolerance satisfies their objectives; In other words they can see whether a model with this degree of Falseness would have potentially misled them or not for their purposes. If x% False is good enough, then they have a model they can use.

They can get this combination of power and focus from the method by deciding carefully the features of interest to them. If they make the focus too broad, then the test's power will be huge but they will never find a model to pass. But with a narrow but suitable focus the power will remain large enough to keep the x% falseness low.

When we look around at the practice of such economists today — eg at central banks — we find that typically either they use no tests at all or they may use standard likelihood or related out-of-sample forecasting tests. These have low power, especially given the way they are implemented in practice; and so the users can only know confidently that they would reject at high Falseness levels. But these may well be too high for them to have any confidence in the policy results.

In the remainder of this talk I am going to discuss the results we have found in using the Smets-Wouters model for monetary and fiscal policy purposes in the context of the recent crisis and its aftermath (Le et al, 2014). This work is all on US data for the period since the mid-1980s; we have not found it possible to mimic US behaviour for earlier data, we think because there has been substantial regime change before then — Le et al (2011, 2014).

I start from the position that the model has credible micro-foundations but that we are searching for a variant of it that a) can allow for a banking system with the monetary base (M0) as an input into it b) can integrate the zero bound on the risk-free interest rate and QE/bank regulation as policy tools; and c) can explain the behaviour of the three key macro variables: output, inflation and interest rates. This is because we want to find a model within which we can reliably explore policies that would improve these variables' behaviour, especially their crisis behaviour. There is of course a large macro literature in which claims are made for the efficacy of a variety of policy prescriptions; but here we just focus on the set of policies investigated for this model, to illustrate the power of our methods.

I will discuss the model's properties with these policies in a moment. But first notice that we can test it two ways — by a Likelihood Ratio test for three key macro variables, inflation, output and interest rates and also by an IIW test on the same three variables. We choose these because they are focused on the behaviour of the three variables of interest to us as policymakers. The LR test measures how close the model gets to the data — essentially a forecasting test; notice at once that this not really our interest but we are using it as a general specification test. It turns out that the LR test is not sensitive, at least for the SW model, to what variables are included in the test, no doubt because if a model forecasts some variables well, it must be forecasting the other variables well that are closely linked to them. We carry out the LR test in the usual way, allowing the ρs to be reestimated on the error processes extracted by LIML. The IIW test looks at how close the model gets to these three variables' data behaviour — which we are deeply interested in matching and represent by a VECM (which we rewrite as a VARX) here as the data is non-stationary. Thus with the IIW test we have carefully chosen its focus to match our policy interests; we could have chosen a broader group of variables which would have raised the test power but at the cost of possibly not finding a model that would fit their broader behaviour. Thus we see here that the focus of the test is a crucial aspect of the IIW test.

I now reproduce, in Table 4, some Monte Carlo experiments for the SW model from Le et al (2015). These follow the same procedures as described above for creating models of increasing Falseness. The

only differences are that the IIW test also falsifies the 2nd, 3rd and 4th moments of the innovations by the same x%; and that we apply the LR test in the usual way it is done, that is after reestimating the residuals' error processes, having extracted them from the model and the data. For the LR test this means that the DSGE model's structural parameters, but excluding the error AR parameters, are falsified by x% while the AR parameters are then determined by reestimation of the implied residuals. As we noted above this reestimation weakens the power of the LR test further compared with the procedure above where the AR parameters too are falsified by x%. However this way of doing LR does correspond to usual practice.

Percent Mis-specified	IIW	LR	IIW	LR
	Stationary data		Non-stationary data	
True	5.0	5.0	5.0	5.0
1	19.8	6.3	7.9	5.2
3	52.1	8.8	49.2	5.8
5	87.3	13.1	97.8	6.2
7	99.4	21.6	100.0	7.4
10	100.0	53.4	100.0	9.6
15	100.0	99.3	100.0	15.6
20	100.0	99.7	100.0	26.5

Table 4: Rejection Rates for Wald and Likelihood Ratio for 3 Variable VAR(1)

The basic point I want to emphasise from this comparison is that if this model passes the IIW test, we can be sure it is less than 7% False whereas if it passes the LR test we can only be sure it is less than 15% False under stationarised data; under non-stationary data, the relevant case here, we cannot even be sure it is less than 20% False — in fact we find that it requires the model to be as much as 50% False for it to be rejected roughly 100% of the time.

When we now apply the two tests to the Monetary model discussed above, it passes both tests. We can now compare how our policy analysis would vary with the two test approaches.

Our basic policy results when we treat the model as True are summarised in the first row of the following Table 5:

Frequency of crisis (expected crises per 1000 years)	Base case	Monetary Reform	PLT	NGDPT	PLT+ Mon.Reform	NGDPT+ Mon.Reform
Policy exercise						
when model is True	20.8	6.62	2.15	1.83	1.41	1.31
when model is 7% False	57.4	18.6	10.3	8.7	11.8	10.3
when model is 15% False	63.6	*Explosive*	19.4	19.6	19.4	17.4
when model is 50% False	70.4	*Explosive*	33.3	33.4	34.4	34.2

Notes:

Base Case: monetary policies as estimated over the sample period;

Monetary Reform: running a Monetary Base rule targeted on the credit premium side by side with a Taylor Rule;

PLT:substituting Price Level Target for Inflation Target in Taylor Rule;

NGDPT: substituting Nominal GDP target for inflation and output targets in Taylor Rule.

Table 5: Policy analysis when model have varying falseness

If we use the IIW test we know that our model could be up to 7% False but no more. We can discover the effect of this degree of Falseness on our policy results by redoing the whole policy exercise with the parameters disturbed by 7%. We obtain the results shown in the second row of Table 5.

In investigating the power of the test, we have simply assumed that we are presented with a False set of parameters somehow from the estimation process. We can then ask what power can we have against a quite mis-specified model whose parameters are simply different. We have looked at this for the model here, by asking what the power is against a quite different model — say a New Classical model versus as assumed True SW model. The power is 100%; it is always rejected. So we can be quite sure the True model is not something quite different.

Between these two things we therefore have a lot of reassurance. First, if the model is not well-specified, it will certainly be rejected. Second, if the model is well-specified, then models up to 7%

distant from it could be True; and our policy conclusions can be tested for robustness within this range as we have done here.

If we use the LR test we know the model could be up to 50% False — we cannot guarantee to reject a model that is less false than this. For example a 15% False model will be rejected only a third of the time. If we now redo the exercise for a 15% disturbance to the parameters we obtain the third row of Table 5. Now our policy is plainly vulnerable. The frequency of crises under the current regime goes up to once every 15 years; with NGDPT+monetary reform it only comes down to once every 50-60 years. This is on the borderline of acceptability.

If we look at the 50% false case, shown in the last row of Table 5, it is disastrous. First, only just under half of the bootstrap simulations have sensible solutions. If we take those that do, we can see that the prevalence of crises under the existing regime would be much greater, at one every 14 years. As with 15% False the monetary reform regime is explosive. The other regimes all generate crisis frequency of around one every 30 years which is far from acceptable.

To make matters worse, we have seen that the LR test has virtually no power against model misspecification, so that we cannot be sure that a misspecified model with yet other, possibly even worse, results might be at work.

What this is showing us is that according to the LR test versions of our model that could be true imply much higher frequency of crises than in the estimated case and the monetary policy regimes suggested as improvements could either give explosive results or produce an improvement in the crisis frequency that is quite inadequate for policy purposes. In other words the policymaker cannot rely on the model policy results. But using the IIW test we can be sure that the recommended policies will deliver the results we claim.

3.1 Can Estimation protect us against Falseness?

But would this vulnerability not be reduced if we take ML estimation seriously? Unfortunately, we have found from Monte Carlo experiments with the SW model that estimation by ML gives us no guarantees of getting close to the true parameters. It is well-known to be a highly biased estimator in small samples — with an average absolute estimation bias across all parameters of nearly 9% in our Monte Carlo experiments — see Table 6. Bearing in mind that our 'falseness' measure assumes x as the absolute bias, alternating plus and minus, this suggests that FIML will on average give us this degree of falseness; in any particular sample it could be much larger therefore.

We also looked above at whether the Indirect Inference estimator could give us any guarantees in this respect. This estimator was much less biased in small samples, with an average absolute bias about half that of FIML — see Table 6. However, again this can give us no guarantees of the accuracy of the estimates in any particular sample.

			Mean Bias (%)		Absolute Mean Bias (%)	
		Starting coef	II	FIML	II	FIML
Steady-state elasticity of capital adjustment	φ	5.74	−0.900	5.297	0.900	5.297
Elasticity of consumption	σ_c	1.38	−5.804	−7.941	5.804	7.941
External habit formation	λ	0.71	−13.403	−21.240	13.403	21.240
Probability of not changing wages	ξ_w	0.70	−0.480	−3.671	0.480	3.671
Elasticity of labour supply	σ_L	1.83	0.759	−8.086	0.759	8.086
Probability of not changing prices	ξ_p	0.66	−1.776	0.027	1.776	0.027
Wage indexation	ι_w	0.58	−0.978	6.188	0.978	6.188
Price indexation	ι_p	0.24	0.483	3.228	0.483	3.228
Elasticity of capital utilisation	ψ	0.54	−13.056	−29.562	13.056	29.562
Share of fixed costs in production (+1)	Φ	1.50	−1.590	2.069	1.590	2.069
Taylor Rule response to inflation	r_p	2.04	7.820	2.815	7.820	2.815
Interest rate smoothing	ρ	0.81	−0.843	−0.089	0.843	0.089
Taylor Rule response to output	r_y	0.08	−4.686	−29.825	4.686	29.825
Taylor Rule response to change in output	$r_{\Delta y}$	0.22	−5.587	0.171	5.587	0.171
Average			−2.861	−5.758	4.155	8.586

Table 6: Small Sample Estimation Bias Comparison (II v. LR)

It follows that we are essentially reliant on the power of the test, in the sense that this can guarantee that our model is both well specified and no more than 7% false under indirect inference, because if it were either it would have been rejected with complete certainty.

The dimension in which we have carried out this examination of the model's reliability in the face of what we might call 'general falseness'. It may be also that the model's performance is sensitive to the

values of one or two particular parameters and if so we would also need to focus on the extent to which these might be false, how far the test's power can protect us against this and how sensitive the model is within this range. This further investigation can be carried out in essentially the same way as the one we have illustrated with general falseness. I should emphasise that the calculations done above should be redone carefully with Monte Carlo power experiments with the same model and data as are being used for the policy analysis; above, for purposes of illustration only, the power estimates and the LR test are taken from earlier versions of the SW model and over varying sample periods.

3.2 Choosing the testing procedure

Thus what I have tried to illustrate in this last section is how macro models can be estimated and tested by a user with a particular purpose in mind. The dilemma a user faces is the trade-off between test power (i.e. the robustness to being false of a model that marginally passes the test) and model tractability (i.e. the relevance for the facts to be explained of a model that marginally passes the test). Different testing procedures give different trade-offs as we have seen and is illustrated in the figure below. Thus the Full Wald test gives the greatest power; but a model that passes this test will have to reflect the full complexity of detailed behaviour and thus be highly intractable. At the other extreme the LR test is easy to pass for a simple and tractable model; but the test has very low power. In between lie Wald statistics with increasing 'narrowness' of focus as we move away from the Full Wald. These offer lower power in return for higher tractability — somewhere along their trade-off will be chosen by the policymaker, as shown in Figure 6 below.

In order for us to find a tractable model we have to allow a degree of falseness in the model with respect to the data features other than those the policymaker prizes. The way to do this is to choose an indirect Inference test that focuses tightly (in a 'directed' way) on the features of the data that are relevant to our modelling purposes.

To apply these methods it is necessary to a) estimate and test the model, b) assess which 'directed' test to choose, c) assess the power in the case of the model being used. We have programmes to do these things which we are making available freely to users — Appendix 2 shows the steps involved in finding the Wald statistic, as carried out in these programmes[2].

It follows that we are essentially reliant on the power of the test, in the sense that this can guarantee that our model is both well specified and no more than 7% false under indirect inference, because if it were either it would have been rejected with complete certainty.

4 Conclusions

I have tried to explain today how users of macro models can once again, as in a bygone era, test their models by classical means, with a view to determining if they can be used reliably for their specified policy purposes. For these users I believe this is of great benefit, since without this they are condemned to a high degree of uncertainty about their models — assuming that they use calibration or Bayesian methods, as is widespread. I have also shown that of the classical tests available, Likelihood Ratio testing (and related tests using out-of-sample forecasts) have quite weak power, which makes it difficult to determine a model's reliability. Here in Cardiff we have developed an estimation and testing procedure using Indirect Inference which offers substantial power even when focused narrowly on the objects of interest for policymakers as in Friedman's original suggestions. This power enables policymakers to determine the bounds within which their model will work and the robustness therefore of their policies. I have given an example of policies proposed for use with the Smets-Wouters model as adapted for the latest decades and how they can be shown to be highly robust. I have ended by making available a suite of programmes (INDIRECT) that will enable users to apply these methods flexibly and easily to their particular models and modelling uses.

[2] Programmes to implement the methods described in this paper can be downloaded freely and at no cost from www.patrickminford.net/indirectinference.

5 References

References

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

[2] 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.

[3] Le, V.P.M., Meenagh, D., Minford, P.„ 2014. Monetarism rides again? US monetary policy in a world of Quantitative Easing, Cardiff working paper No E2014/22,; also CEPR discussion paper 10250.

[4] Le, V.P.M., Minford, P., Wickens, M., 2013, A Monte Carlo procedure for checking identification in DSGE models, working paper E2013/4, Cardiff Economics Working Papers, Cardiff University, Cardiff Business School, Economics Section; also CEPR discussion paper 941.

[5] Le, V.P.M., Meenagh, D., Minford, P., Wickens, M., Xu, Y. (2015) Testing macro models by indirect inference: a survey for users. Cardiff working paper No E2015/9, Cardiff Economics Working Papers from Cardiff University, Cardiff Business School, Economics Section; also CEPR discussion paper.

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

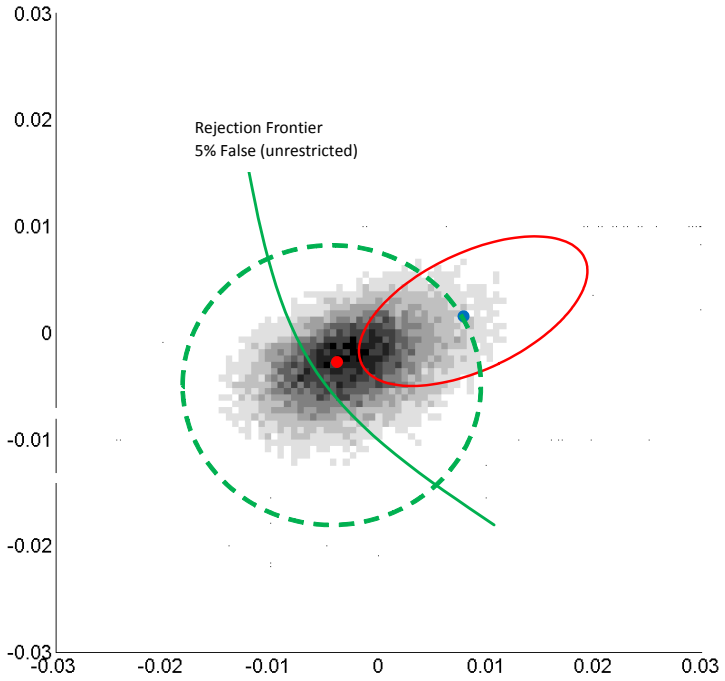

Figure 4: Two 95% contours for tests of 5% False Model- Green=Unrestricted; Red=Restricted.

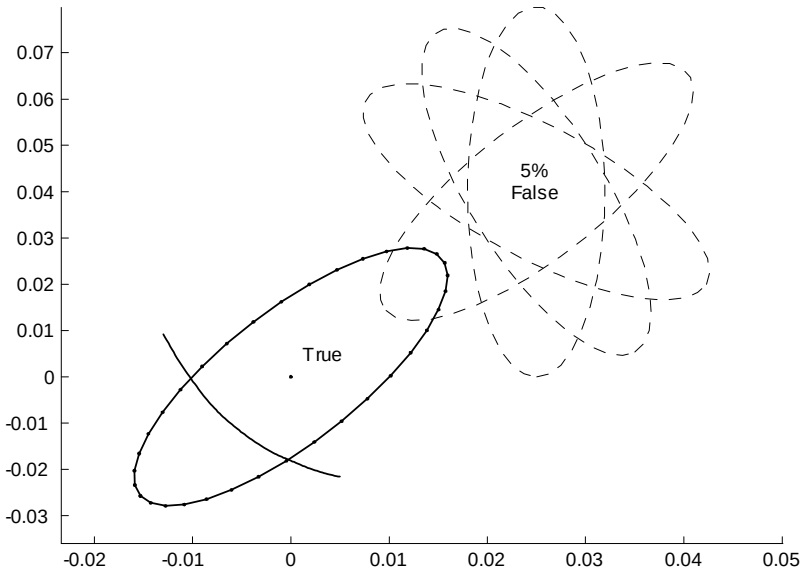

Figure 5: Joint Distribution of VAR coefficients rotates with changing False DSGE parameters

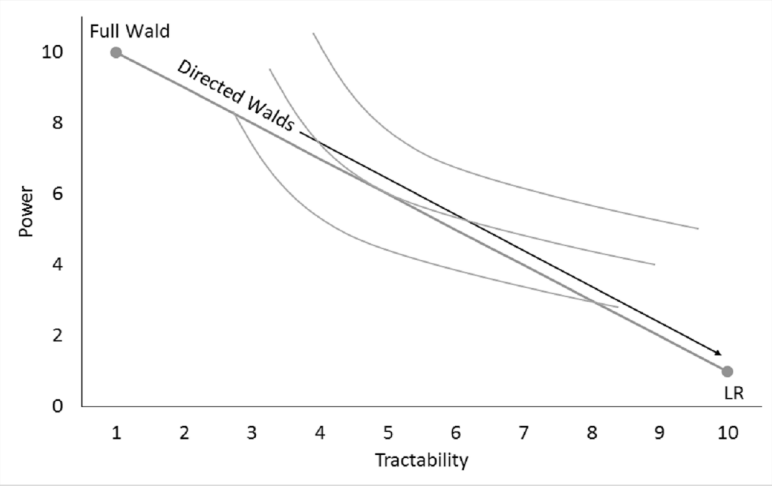

Figure 6: Maximising Friedman utility