COMMENTS ON SMETS AND WOUTERS

CHRISTOPHER A. SIMS
PRINCETON UNIVERSITY
SIMS@PRINCETON.EDU

1. Overview

- This paper is part of a suite of three, dealing with US and Euro area data with a common methodology.
- The work is a very substantial advance, suggesting the possibility that DSGE models, combined with Bayesian methods of inference, may before long become the standard framework for macroeconomic policy modeling.
- The work breaks new ground, and has appropriately been carried out and disseminated quickly. There are gaps in the description of the methodology and uncertainties about the results that should stimulate new research.

2. What's new

- Bayesian model comparison.
- Bayesian characterizations of uncertainty about forecasts and about structural parameters.
- Treating a modern DSGE seriously as a probability model for multivariate data.
- Therefore giving the model more complex dynamics, more shocks, than are common in the calibration literature.

3. Proliferation of Adjustment Costs

- DSGE models $\Rightarrow$ responses that are too quick compared to the data.
- Adjustment costs can slow this down, but
- They then tend to imply too much overall smoothness — in the data, own shocks produce quick responses.
- this leads to "adjustment cost shocks" or other routes to introducing idiosyncratic variation in variables that is weakly related to other variables.

Date: June 23, 2003.

4

- "External habit": Do we want to accept the implications of this formulation for welfare evaluation?
- "Equity premium shock": How do we interpret this? Asset market returns are among the data that are most conceptually well matched to theory and the least subject to measurement ambiguities.
- The micro-foundations of stickiness in the entire class of models on which this paper is based are very shaky. They do not in fact allow reliable extrapolation of the effects of persistent changes in monetary policy rule. We need more work on better models of sources of inertia and stickiness.

5. Fiscal structure

- Existing central bank models (FRBUS, AWM) have quirks and inconsistencies in their modeling of fiscal policy and its interaction with private spending plans.
- This model does not, because it does not even attempt to model fiscal variables.
- Yet it is a "cashless limit" economy, in which the price level is only meaningful as a rate of exchange between interest bearing debt and real goods.
- Without a fiscal component, it cannot sensibly model pegged interest rates, which do *not* in general lead to indeterminacy.
- Yet this paper attempts to consider the effects of interest rate pegs and of policy near zero interest rate bounds.

6. Heteroskedasticity

- In the US data, there is a sharply defined period of high interest rate and inflation volatility during 1979:IV-1982:IV.
- There is a less sharply defined trend or shift toward lower volatility in many variables in the US and other countries in the late 80's and early 90's. This is not confined to the US.
- Likelihood is very sensitive to failure to model this heteroskedasticity, and inference about relative model fits and about parameter uncertainty can also be badly affected.

7. Model evaluation anomalies

- BVAR vs. DSGE odds ratios are very different between Euro data models and US models. Even between (2000a) and the current paper. This is hard to understand.
- The BVAR priors are not really the widely tested "Minnesota" priors. They are a bit strange: `othrel=1`, `decay=2`. There are alternatives that are not as computationally burdensome as the true Minnesota prior.
- The DSGE priors are admittedly based in part on data-peeking. This weakens the credibility of the model comparison.

8. Lindley paradox issues

- There is a tendency (not just in this model) for posterior odds on models to emerge as implausibly decisive. Some statisticians argue that it is nearly always better to find a way to replace discrete parameters (model index number) with continuous ones, and then discuss distributions over these continuous parameters.
- Posterior odds are always directly proportional to prior odds, and also, unlike what emerges within a continuous $\mathbb{R}^n$ parameter space, are sensitive to the degree of dispersion of the prior pdf even when the likelihood is concentrated relative to the prior.
- That is, *model comparison* is sensitive to the prior even when the likelihood is such that *posteriors on parameters* are not.

9. Gaps in the papers' specifications of priors

- There is no detail on the nature of the BVAR priors. They are said to be "Minnesota" priors, but the Minnesota priors are improper and hence provide no basis for model comparison.
- Table 2 has a footnote that says that the posterior probabilities for VAR's based on the "prediction error decomposition", which can be done analytically for these models, "treat the first $T0$ observations as given", without saying what $T0$ is.
- Even if we were told $T0$, there is a question as to what "treating the first $T0$ observations as given" means here.

10. Training sample priors

Suppose we have a log posterior that can be written as

$$\ell_1^T(Y_T, \theta) = \sum_{t=1}^{T} \ell_t(Y_t, \theta) \,.$$

With a flat prior, this is just the log likelihood. Otherwise the prior is embedded in ℓ_1.

Even if the prior is improper, $\exp\big(\ell_1^T(Y_T, \theta)\big)$ tends to be integrable in θ once T is above some threshold. For inference on θ, with the model held fixed, we can use ℓ_1^T as the log posterior whether or not the prior is proper, and there is a legitimate interpretation of the result as the limit of inference as priors become spread-out relative to the likelihood.

11

But for model comparison, we need a meaningful normalizing constant for $\exp(\ell_1^T)$.

The **training sample** method picks a T_0 large enough so that $\exp\big(\ell_1^{T_0}\big)$ is integrable and treats the model as if this were proportional to the prior pdf on θ, while the data were available only for $T_0 + 1, \ldots, T$.

In other words, it treats ℓ_1^T as proportional to the posterior, but multiplies by a scale factor such that $\exp(\ell_1^{T_0})$ integrates to 1.

This has no effect on posteriors on parameters within models, but allows meaningful model comparisons.

12

In effect, the procedure asks, "suppose the priors are 'weighted' so that after T_0 observations the marginal likelihoods of all models are the same and all Bayes Factors are 1. How then do the remaining $T - T_0$ observations shift beliefs about model validity?" If the priors are each improper in some dimension, this simply chooses the arbitrary height of the "flat" components so as to achieve this result. If some models have proper priors, these models are being treated as if the priors were dummy observations that are strengthened or weakened by repetition. That is, a proper prior $\pi(\theta)$ is entered into the posterior as a factor $\pi(\theta)^\tau$, where τ is adjusted to give the desired uniform marginal likelihoods after T_0 observations.

13. Pros and Cons of Training Samples

- Biggest (dubious) advantage: They allow us to use improper priors without generating obvious nonsense.
- Training sample methods have an element of arbitrariness: How to pick T_0? Why "train" on $1, \ldots, T_0$ rather than $T_0 + 1, \ldots, T$? Could we leave the prior out of the training sample altogether?
- Training sample priors for small T_0 can easily be eccentric, since they let a few data points take more weight than they would with any reasonable proper prior. This may disadvantage a model given a prior this way.
- But they are a reasonable data-summary device when substantively grounded proper priors are hard to generate.

14. Is a clean training sample procedure what was done here?

It is easily implemented for VAR's or BVAR's with dummy-observation priors, but it is a great deal of work for something like the DSGE model. The reason is that it requires integrating $\exp\big(\ell_1^{T_0}(Y_{T_0}, \theta)\big)$ with respect to θ, which is just as hard as the evaluation of the model's marginal likelihood itself.

Since the calculation is analytic for VAR's and BVAR's with dummy observation priors, it can easily be implemented. But the paper gives no indication that these demanding calculations were performed for the DSGE model, and they were not done in fact.

15. What was actually done

[This slide reflects clarifying discussion with Raf Wouters at the conference, after the slides for my presentation were prepared and was not presented in this form at the conference.]

The VAR's and BVAR's used actual training samples, so comparisons among them have a clean interpretation. The DSGE used the prior described in the first Smets-Wouters European paper, a proper prior, but treated it as if it had arisen as the posterior from the first T_0 observations.

This could advantage the DSGE, because it is protected from the possible eccentricity of a true training sample prior. But it could also disadvantage the DSGE, because its "human-generated" prior may miss, or even unintentionally put low probability on, strong dependencies among parameters that the training-parameter prior recognizes and treats as if known.

16. Making the parameter space more nearly continuous

- Add parameters to the DSGE until it is at least locally just identified.
- It is then equivalent to an unrestricted VAR, except for the prior.
- We can then look at DSGE priors generated from VAR priors and vice versa.
- Model comparison can be carried out by relaxing the restrictions in the DSGE model smoothly, examining both whether fit increases sharply as they are relaxed and whether, if so, posteriors start to drift into unreasonable regions of the parameter space. This procedure also is likely to help with diagnostics as to which parts of the model are responsible for problems with fit.

17. Demeaned and detrended data

- DSGE models apparently do not perform well if their implications about cointegration are forced on the data — at least in European data.
- The response of fitting to data with means and trends removed, while ignoring the data in fixing some parameters that mainly influence steady states, is at best a short-run expedient.
- It should not be a major increase in sophistication to parameterize these models flexibly enough so that the information in means and trends are not thrown away.

18. We need a clearer explanation of how preliminary detrending integrates with measures of forecast performance and of posterior probability.

[This slide also reflects clarifying conversations at the conference and so differs from what was presented.]

- The mean removal and detrending was not repeated each time the DSGE model was re-estimated in the forecast comparisons. It was done once, using the whole sample, and the detrended data were treated as if they were raw data in all subsequent analysis.
- The MCMC posteriors therefore do not take account of uncertainty over means and trends. [The paper on US data is different in this respect, though apparently

only in that trend growth rate, not mean, is recognized as an uncertain parameter value.]

- The VAR's were estimated on the same pre-processed data set as the DSGE.
- The VAR's use a version of the Minnesota prior and "shrink" toward a model with as many unit roots as variables. These VAR's are being fit to data that by construction are stationary with zero mean. The DSGE is by construction a stationary model. It "knows a priori" to make impulse responses return to zero. The VAR's are more erratic at long term forecasts because they look for possible persistence in the data, which by construction is not there. The fact that including a constant term in the VAR's makes them perform better reflects this: The bias toward stationarity of estimated VAR models is stronger when constant terms are included.
- More generally, for many variables uncertainty about means and long run growth rates is a critical component of the forecasting problem. To compare models for forecasting purposes as if these problems were not there makes no sense.

19. Informal model evaluations

- Autocovariance functions are not very helpful. They correspond to no orthogonal decomposition of the data, and thus cannot lead us to eyeball recognition of entropy-reducing patterns in the data.
- Frequency domain statistics don't have this disadvantage.
- Reduced form impulse response functions don't have this disadvantage, and are probably easier to connect to informal reasoning about causal mechanisms than are frequency domain statistics.
- If you don't like the "arbitrariness" of triangular orthogonalization of the disturbance covariance matrix, use the symmetric square root or the singular value decomposition (principal component) square root, though I personally find the arbitrary triangular ordering a more transparent data summary.