Comments on Brock, Durlauf and West’s "Model Uncertainty and Policy Evaluation: Some Theory and Empirics"

Christopher A. Sims
Princeton University
sims@princeton.edu

March 4, 2005

The paper's method

- Considering model uncertainty via posterior probabilities is a good idea.
- But model uncertainty is only a special case of uncertainty about parameters. it is simply uncertainty about a discrete parameter.
- This paper uses the Bayesian idea of forming posterior odds on models, but pays no attention to integrating uncertainty about the "model number" parameter with uncertainty about other parameters.
- It is unclear what the motivation for this approach might be.

Why the mixed methodology?

- Maybe it is thought that handling uncertainty about all parameters jointly is too hard.

- But it isn't. There is a great deal of statistics literature, and recently some economics literature, that develops and applies MCMC sampling schemes to jointly estimate parameters and form posterior odds on models. MCMC sampling within models can be done with many packages now — R's MCMCpack, WinBUGS, Geweke's BACC. And once one can do this, forming posterior odds with Geweke's modified harmonic mean algorithm is straightforward.

- A review in the latest *JASA* of a book by S. James Press points out that he was a Bayesian before being Bayesian was "cool", by which the reviewer means before MCMC methods entered the scene. This paper could benefit from being more cool.

Some non-Bayesian components in order to seem middle-of-the-road?

- The paper asserts that "frequentist methods dominate policy evaluation analysis" and suggests that this makes it a good idea to have frequentist elements in the methodology.
- It's not at all clear that, at least in central banks, frequentist methods dominate, or are even widely used, in policy analysis.

In the interviews I conducted for my 2002 paper that is cited in the paper, I was told explicitly by some central bank staff economists that "econometrics" was not useful in policy discussions. Econometrics was identified with testing hypotheses, sticking with them until they were rejected at a .05 level. Policy makers, I was told, instead want an assessment of the weight of evidence in the data at hand — i.e. (though the interviewees did not make this connection explicitly) posterior probabilities.

> . . . the conduct of monetary policy in the United States has come to involve, at its core, crucial elements of risk management. This conceptual framework emphasizes understanding as much as possible the many sources of risk and uncertainty that policymakers face, quantifying those risks when possible, and assessing the costs associated with each of the risks. In essence, the risk management approach to monetary policymaking is an application of Bayesian decisionmaking.
>
> Alan Greenspan, 1/3/2004

Discrete parameters are treacherous

- Practical experience with Bayesian approaches to handling multiple models has frequently turned out to be disappointing or bizarre. The general phenomenon of conflicts between Bayesian model comparison results and classical tests is labeled the “Lindley paradox”.
- One standard applied Bayesian textbook (Gelman, Carlin, Stern, and Rubin, 1995) has no entry for “model selection” in its index, only an entry for “model selection, why we do not do it”.

- The problem, as Gelman et al make clear, is that more often than not, the discrete parameter is a shortcut where a continuous parameter actually makes more sense. In such cases Bayesian model comparison and model averaging can produce bizarre and mistaken results. Commonly it produces odds ratios that are implausibly close to zero or one and, by eliminating information about which parameter settings are "near" each other, is subject to being distorted by the counts of similar models being considered.

- This paper, though it can be seen as just an exercise illustrating its methodology, is also a good example of bad discretization. The models are actually generated as restrictions on a more general model with a continuous parameterization. When this is the case, it is more useful, unless infeasible, to characterize the posterior on the more general model more directly.

- This is true also for choice of lag lengths. Short-lag models are parametric restrictions of long-lag models. While we can put discrete probability on lag lengths of exactly one, two, three, etc., it makes much more sense to choose a generous lag length and use a prior (like the Minnesota prior) that downweights the more distant lags.

Penalties for model complexity

- The BIC or Bayesian Information Criterion, also called the Schwarz criterion, is widely used as way of penalizing complex model for the overfitting implicit in direct comparisons of maximized likelihood. One of its appeals is that it leads to consistent model choice. (i.e., picking the true model with probability one as $T \to \infty$.

- It is derived by consideration of the asymptotic behavior of fully explicit Bayesian model choice, where parameter spaces are given priors for each model. In large enough samples model comparison is dominated by the BIC term, with terms dependent on the model structure or the prior no longer mattering.

- This paper uses maximized likelihood with a BIC penalty to weight models. This is a worthwhile shortcut when necessary, but it must be recognized that the asymptotics of the BIC do *not* say that in large samples it generates accurate odds ratios *unless the odds ratios are nearly zero or infinity*.

- In finite samples, prior probabilities on models and the shape of the prior on the parameter space always matter. If the odds ratios are in the 1:1 to 20:1 range, we can be sure that the BIC term is not yet dominant and that more careful consideration of priors would be able to produce big effects on rankings.

Treating dimensionality systematically

- The BIC asymptotics deal with overfitting, and they are not the only reason to give attention to model dimensionality.

- Usually, and this is clearly the case here with lag length, we rank models by complexity and act as if it is likely that less complex models will fit well. It is possible to express such beliefs as a prior across an infinite-dimensional space of models. Such a prior, since it has to integrate to one, must downweight more complex models if there are a countable infinity of them. Such downweighting is dominated by the BIC downweighting in large samples, but in practice it can easily dominate the terms that give arise to BIC.

- One can add “trivial” parameters to a model, which will spuriously make the model appear less likely under ML+BIC. This problem again disappears with a complete Bayesian approach to model comparison. Leamer’s discussion of regression models makes these tradeoffs clear in a context that is familiar to econometricians.

The importance of assessing fit

- It is commonly the case when we discuss overidentfied structural models that lurking in the background is a less restricted “reduced form” model that we are confident fits well but that is not as useful as the structural models.
- Checking fit is best thought of as comparing a model to another model, or class of other models, that we are confident fit about as well as is possible, even though they might not use prior information efficiently and might not have structural interpretations.

- If we are dealing with a collection of models that in this sense do not fit well, any likelihood-based method of selecting or averaging models can produce bad results.

- This contradicts a claim seen sometimes in the literature, that these Bayesian methods work even on collections of false models. The claim is based on the result that Bayesian model comparison leads to the model that's closest to the truth in Kullback-Leibler distance.

- Kullback-Leibler distance can be an arbitrarily bad choice from a decision-theoretic perspective. Essentially, KL distance looks for where models make the most different predictions — even if these differences concern aspects of the data behavior that are unimportant to us. This is just right if we are picking out the truth. But if we know the models are all flawed, we don't want the KL distance to arbitrarily focus on some relatively unimportant flaw.

- These points are nicely illustrated in Frank Schorfheide's thesis and in the paper (Schorfheide, 2000) that came out of it.

- In this paper by Brock, Durlauf and West, the reduced form is obvious and easy to construct. Posterior weights for Bayesian VAR's, which is what this paper's reduced form would be, can be constructed analytically with pre-packaged software.

*

References

GELMAN, A., J. B. CARLIN, H. S. STERN, AND D. B. RUBIN (1995): *Bayesian Data Analysis*. Chapman and Hall, London.

SCHORFHEIDE, F. (2000): "Loss Function-Based Evaluation of DSGE Models," *Journal of Applied Econometrics*, 15(6), 645–670.