PROBABILITY MODELS FOR MONETARY POLICY DECISIONS

CHRISTOPHER A. SIMS

ABSTRACT. To be written.

I. THE STATE OF CENTRAL BANK POLICY MODELING

I.1. **Not probability models.**

I.2. **Ad hoc "econometrics" layered on a "long-run" core.**

I.3. **Lack of a language to discuss uncertainty about models and parameters.**

I.4. **Why models are nonetheless used, and are useful.** The need to bring data to bear on policy discussions. There's a lot of data, and it's hard to maintain accounting and statistical consistency without a model.

I.4.1. *Decentralization, model size.*

I.5. **The impact of flexible inflation targeting.**

II. DIRECTIONS FOR PROGRESS

There are promising new directions, but as we get closer to actually using these methods in real-time decision-making we will need to solve some implementation problems and broaden understanding of the contributions and limitations of these methods.

II.1. **Existing work.**

II.1.1. *Using bad models.* Schorfheide, Geweke, Brock-Durlauf-West?

II.1.2. *MCMC methods for parameter uncertainty.* Smets and Wouters

Date: July 9, 2003.

II.1.3. *Bayes factors, odds ratios, model averaging.* Brock-Durlauf-West, Smets and Wouters

II.1.4. *Perturbation Methods.*

II.2. **Problems and prospects.**

II.2.1. *Using bad models: wasted energy?*

II.2.2. *MCMC methods.*

II.2.3. *Nonlinearities.* Stickiness, welfare evaluation, the zero bound.

III. Odds ratios and model averaging

Though most central banks focus much more attention on one primary model than on others, the banks' economists and policy makers are well aware that other models exist, in their own institutions as well as outside them, and that the other models might have different implications and might in some sense fit as well or better than their primary model. So one aspect of the Bayesian DSGE modeling approach that central bank economists find most appealing is its ability to deal with multiple models, characterizing uncertainty about which models fit best and prescribing how to use results from multiple models in policy analysis and forecasting.[1]

But practical experience with Bayesian approaches to handling multiple models has frequently turned out to be disappointing or bizarre. 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".

[1]Some discussions of Bayesian methods by econometricians have asserted that Bayesian model selection methods can be useful in choosing among "false models". This is basically not true. It is true that non-Bayesian approaches cannot put probability distributions across models, conditional on the data, any more than they can put probability distributions on continuously distributed parameters. But, as Schorfheide (2000) explains, choosing among, or putting weight on, false models by fit criteria or probability calculations that assume one of the models is true can lead to large errors. Bayesian methods are likelihood-based and therefore as subject to this kind of error as any other approach to inference. Schorfheide proposes ways to mitigate this problem.

There are several related ways the Bayesian model comparison methods tend to misbehave.

- Results are sensitive to prior distributions on parameters within each model's parameter space, even when the priors attempt to be "uninformative".
- Results can be sensitive to seemingly minor aspects of model specification.
- Results tend to be implausibly sharp, with posterior probabilities of models mostly very near zero or one.

When we work with a single model, with a prior distribution specified by a continuous density function over the parameter space, under mild regularity conditions in large samples posterior distributions over parameters cease to depend on the prior. There are no such results available for model comparison, when we hold the set of models fixed as sample size increases.

I will argue that these pathologies of Bayesian model comparison are not inherent in the methodology, but instead arise from the ways we generate and interpret collections of parametric models. Once this is understood, the model comparison methodology can be useful, but as much for guiding the process of generating and modifying our collections of models as for choosing among or weighting a given collection of models.

III.1. **When the Set of Models is Too Sparse.** Even in simple situations with a small number of parameters, model comparison methods will misbehave when the discrete collection of models is serving as a proxy for a more realistic continuous parameter space. For example, suppose we observe a random variable X_t distributed as $N(\mu, .01)$. One theory, model 0, asserts that $\mu = 0$, while another, model 1, asserts that $\mu = 1$. With equal prior probabilities, if the observed X is bigger than .55 or smaller than .45, the posterior odds ratio on the two models is greater than 100 to 1. This is a correct conclusion if we in fact know that one of the two models must be true. The practical problem is that often we will have proposed these two models as representatives of "μ is small" and "μ is what is predicted by a simple theory" *classes* of models. It then is worrisome when an observation such as $X = .6$, which might seem to slightly favor the "μ is big" class, but actually is highly unlikely under either model 0 or model 1, implies a result that suggests near-certainty, rather than skepticism.

Here it is quite clear what the problem is and how to fix it: treat μ as a continuous parameter and report the entire likelihood. Or, having noticed that likelihood concentrates almost entirely in a region far from either of our representative models, generate new representative models that are more realistic. If we had a collection of models with $\mu = 0, .1, .2, \ldots, .9, 1$, odds ratios across models would produce roughly the same sensible implications as use of a continuous parameter space. This is the kind of example that Gelman, Carlin, Stern, and Rubin (1995) have in mind when they cite, as a condition for posterior probabilities on models being "useful", the requirement that "each of the discrete models makes scientific sense, and there are no obvious scientific models in between".

III.2. **Changing Posteriors by Changing Priors.** Suppose we have a set of models $i = 1, \ldots, q$ for a vector of observations X, with model i specifying the pdf of X as $p_i(X \mid \theta_i)$, and the prior pdf on θ_i as $\pi_i(\theta_i)$, over parameter space Θ_i. We will assume for simplicity that priors on parameter spaces are independent across models. Take the prior probabilities on models to be equal. The posterior weight on model i is then the marginal data density

$$w_i(X) = \int_{\Theta_i} p_i(X \mid \theta_i)\pi_i(\theta_i)\, d\theta_i \tag{1}$$

and the posterior probability on model i is $w_i / \sum_j w_j$. In order to understand reports of posterior odds on models in the context of scientific reporting, it is important to understand the range of results possible from a given set of likelihood functions $p_i(X \mid \cdot)$ as priors $\pi_i(\cdot)$ are varied. Since w_i is a weighted average of the likelihood function for the i'th model, it is clear that it is maximized when the prior pdf is concentrated entirely on a small region near the peak of the likelihood function, in which case we have $w_i = p_i(X \mid \hat{\theta}_{MLE})$.

It is usual, when sample sizes are not very small, for likelihood values to be arbitrarily close to zero in parts of the parameter space. This occurs because usually the data are able to rule out as very unlikely very extreme values of parameters. Likelihood will always approach zero when the parameter space is unbounded and the likelihood function itself is integrable. Under these conditions, w_i can be driven to zero in two ways. Most obviously, the prior can concentrate almost entirely on some small region of the parameter space where likelihood is very small.

That is, the prior can express great certainty about a value of the parameter that is ruled out by the data. But the same result emerges if the prior expresses what is conventionally thought of as great *un*certainty. If the parameter space is unbounded and the likelihood is integrable, w_i can be pushed to zero by choosing π_i to be nearly "flat", that is to put nearly equal small probabilities on similar-sized subsets of Θ_i dispersed widely over a large region of the parameter space. In the limit as it becomes flatter, such a prior makes $\pi(\theta_i)$ arbitrarily small over the entire region in which the likelihood $p(X \mid \theta_i)$ is non-trivially positive. This is less paradoxical than it may seem. Such a "flat" prior is actually expressing great certainty that the true parameter value is extremely far from the region of high likelihood, and is thus incorrect in the same way as a prior that concentrates too sharply on an unlikely finite parameter value.

So we see that with most models and samples (because most have integrable likelihoods and unbounded parameter spaces) it is possible to make any given model have as low a posterior probability as we like by making the prior on its parameter space extremely flat. Of course a corollary, since posterior probabilities are determined by relative w_i's, is that we can make any given model have as high a posterior probability as we like by choosing very flat priors for its competitors. For a $N(\bar{\theta}_i, \nu_i \Sigma_i)$ prior, for example, as a function of ν the posterior pdf is, for large values of ν, approximately proportional to ν^{-d_i}, where d_i is the dimension of the θ_i vector. By choosing large values of ν for all models, but choosing those for the favored model many times smaller than those for the others, we can make the posterior probability of the favored model as large as we like, even though all models have been given "flat" priors.

This arbitrariness can be mitigated if the models under consideration can be merged, so that the discrete "model number" parameter $i = 1, \ldots, q$ disappears or is replaced by a continuous parameter. For example, if we have two competing regression models,

$$\{Y_t \mid \{Y_s, s \neq t; X_s, Z_s, s = 1, \ldots, T\}\} \sim N(X_t \theta_1, \sigma^2) \tag{2}$$

$$\{Y_t \mid \{Y_s, s \neq t; X_s, Z_s, s = 1, \ldots, T\}\} \sim N(Z_t \theta_2, \sigma^2)\,, \tag{3}$$

we can replace them with a mixture model

$$p(Y_t \mid Y_s, s \neq t; \{X_s, Z_s, s = 1, \ldots, T\})$$
$$= \frac{1}{\sigma}\big(\lambda\varphi(Y_t - X_t\theta_1) + (1-\lambda)\varphi(Y_t - Z_t\theta_2)\big)\,, \qquad \lambda \in [0,1] \quad (4)$$

(where φ is the pdf for $N(0,1)$) or with a regression model including both variables

$$\{Y_t \mid \{Y_s, s \neq t; X_s, Z_s, s = 1, \ldots, T\}\} \sim N(X_t\theta_1 + Z_t\theta_2, \sigma^2)\,. \quad (5)$$

Either of these options allows us to report a continuous posterior pdf jointly over θ_1 and θ_2. Such a joint pdf is not subject to easy manipulation via flat priors. This resolution of the problem assumes, though, that the merged model is taken seriously. An artificial merger of the two models, where all interest is still focused on the restricted part of the merged parameter space where only one of the two models is operative, will still have the original difficulties.

III.3. **Understanding the asymptotic results.** Under mild regularity conditions, the shape of a likelihood function on a Euclidean parameter space will be well approximated in large samples as normal, and with any prior with continuous density the posterior pdf will be well approximated by the likelihood itself. This result seems in sharp contrast to the situation with a discrete collection of models, where similar regularity conditions simply imply that the posterior probability of the true model converges to one, and where, as we have seen, in any given sample posterior probabilities of models can be driven to zero or one by manipulation of "uninformative" prior distributions within each model's parameter space.

But as we saw in III.1, the distinction between the continuous and discrete cases is not as sharp as this result makes it seem. The asymptotic normality and insensitivity to the prior of the posterior distribution emerges because the posterior is rescaled as sample size increases, so our attention focuses and smaller and smaller subsets of the parameter space. The continuous-pdf prior automatically adapts itself to the rescaling, and in the process becomes closer and closer to a constant over the relevant range. The reason conventional asymptotics gives no such result for comparison of finite sets of models is that asymptotically only one model is in the "relevant range". But in a given sample, if we have the option of refining our collection of models in the light of the likelihood, we can ensure that we have

a well-articulated prior over the likelihood-relevant part of the parameter space, even though it may be a discrete prior.

But while this will take care of the problem of unrealistically large posterior odds ratios because of sparsity of the set of models, it does not directly address the issue of sensitivity of results to priors within models' parameter spaces. This problem also, however, yields to an appropriate "filling in" of the parameter space with discrete models. In our two-regression-model case (2)-(3), filling in the space between the two models might be interpreted as considering a family of models

$$\left.\begin{aligned} p_{1j}(Y_t \mid Y_s, s \neq t; \{X_s, Z_s, s = 1, \ldots, T\}) &= N(X_t\theta_1 + \kappa_{1j}Z_t, \sigma^2) \\ p_{2j}(Y_t \mid Y_s, s \neq t; \{X_s, Z_s, s = 1, \ldots, T\}) &= N(Z_t\theta_2 + \kappa_{2j}X_t, \sigma^2) \end{aligned}\right\} \quad j = 1, \ldots, m\,, \tag{6}$$

where the $\{\kappa_{ij}\}$ sequences are sets of values that include several within the ranges where the likelihood function is non-trivially positive. If we try to use this collection of models to assess the probability of the $\{p_{1j}\}$ models vs. the $\{p_{2j}\}$ models, the ability to manipulate results by varying the heights of "flat" priors remains unabated. But if instead we consider inference about the strength of, say, the Z effect, the filling in has resolved the problem. The relative heights of the flat priors within the continuous components of the overall parameter space can affect the degree to which the posterior appears dominated by discrete or continuous components, but cannot strongly influence the location and dispersion of the posterior distribution of the Z effect, where this is a mixture of the distribution over κ_j from the p_{2j} models and over θ_1 from the p_{2j} models.

It might be objected that the recommendation to choose discrete collections of models so that they fill in the region of the grand parameter space (over "models", and "parameters") with high likelihood amounts to letting the data affect our prior distribution, and thus undermines the internal consistency of Bayesian procedures. This would be true if we began with a collection of models that had been rigidly determined by *a priori* considerations and that could be claimed with certainty to contain the true model. But that is not a common situation in econometric modeling. Usually, when we use a finite set of models, this is an expedient. The models are chosen from a much larger class of models we might have considered. Our discrete prior over the finite collection of models we work with is meant as

an approximation to our prior over this larger collection. From this perspective, it is essential that the collection of models we work with adapt to the nature of the likelihood function at hand, and there is nothing illogical in proceeding this way. Of course this also means we accept that the individual models themselves are not the focus of interest. We can make any one model have low posterior probability by surrounding it with very similar models with equal prior weight. But the posterior probability on the resulting class of similar models in this part of the parameter space will not be reduced by doing so, in fact it will be increased. This leads to another important principle in specifying classes of models: equal prior weights across models can in fact strongly favor a particular substantive conclusion, if the models are chosen so that there are many more that imply that conclusion than that contradict it.

The implication of this discussion is that to avoid paradoxical results from model comparison exercises it is best where possible to think of the models as representative points in a larger continuous parameter space. Results that show overwhelming odds in favor of one model in the collection then imply that the collection is not well chosen: some rethinking of specification and filling in of the space of models is in order.

IV. Model comparisons among VAR's and DSGE's: A case study

In a path-breaking series of articles Frank Smets and Raf Wouters 2002; 2003c; 2003b; 2003a have shown that a linearized DSGE model, endowed with enough sources of stochastic disturbance and rich enough dynamics, can fit approximately as well as VAR models. A centerpiece of their discussion is a Bayesian model comparison among their DSGE model, three Bayesian VAR models that invoke a "Minnesota-like" prior, and three VAR models that use what is known as a "training sample" prior. What they have done illustrates some of the difficulties in setting up and interpreting such a model comparison.

IV.1. **The ill-defined boundary between models and priors.** Smets and Wouters consider seven models — three VAR's, of orders 1, 2 and 3, three BVAR's of the same orders, and their DSGE model. It is easy to see how these all naturally are thought of as embedded in the same large parameter space — that of linear ARMA

models of the seven time series they explain. The BVAR's and VAR's are obviously special cases of finite order ARMA models, but the DSGE is also. It is a linearized model, so it results in a linear ARMA model for the observed data, though with the coefficients in the ARMA model nonlinear functions of the 32 underlying free parameters in the DSGE.

Are the BVAR's and VAR's different models? Perhaps being unsure of this, Smets and Wouters never construct a posterior distribution over all seven models jointly, only 4-model posteriors, over the DSGE and the three VAR's or the DSGE and the three BVAR's. Bayesian model comparison proceeds by, for each model, constructing an implied unconditional density function for the observed data, using the prior to integrate out unknown parameters. Model weights are then the heights of the marginal density functions at the observed point in the data space. Whether different marginal distributions for the data are arrived at via a fixed model $p(Y \mid \theta)$ and differing piors $\pi_i(\theta)$, or at the opposite extreme, from fixed priors $\pi(\theta)$ and varying models $p_i(Y \mid \theta)$ makes no difference to the analysis. So it is perfectly legitimate to treat BVAR's and VAR's as different models.

In fact, as research proceeds in the direction of a richer menu of DSGE models, it will be important to keep in mind that priors, generated from substantive knowledge about parameters, are an essential part of a model's specification. This is well illustrated by a result found by Schorfheide (2000). Though he concludes that VAR's fit the data better than the DSGE models he considers, he notes (p.660) that one of the DSGE models does fit as well as a VAR(4) if the prior on the DSGE is loosened enough so that its capital share parameter goes to .7 and the real rate of return to capital goes to 10%. He nonetheless sticks with the tighter prior, rightly treating parameter estimates that seem substantively unreasonable under the model's economic interpretation as evidence against the DSGE model.

When we consider several richly parameterized DSGE models, they are likely to concentrate probability on identical or nearby submanifolds of the general ARMA parameter space. Differences among the models could easily arise mainly from the priors they imply, via the substantive interpretations of their parameters.

But the BVAR and VAR priors that Smets and Wouters use are not substantively motivated. The BVAR priors are a natural counterfoil to a DSGE model. They

deliberately ignore any substantive knowledge about relations among variables, making the prior as symmetric in the variables as the VAR model itself. They do use the knowledge that economic aggregate time series tend to be persistent. If a BVAR fits as well as a DSGE, it calls into question whether the data lend any support to the DSGE model's substantive interpretation of the relations among the variables.

IV.1.1. *Training sample priors.* What Smets and Wouters call a plain "VAR" is actually also a Bayesian VAR, but with a different prior. It is in fact not possible to generate a meaningful posterior across a set of models that includes models that are not accompanied by a proper (i.e. integrable) prior.

For these models they use training sample priors, with 1970:2-1980:2 as the training sample. The mechanics of the training sample method run as follows. One takes the product of likelihood $p(Y_T \mid \theta)$ and prior $\pi(\theta)$ generated by the model, which may involve an improper prior or even no prior (i.e. the flat prior $\pi(\theta) \equiv 1$), and splits it into two pieces,

$$p(Y_T \mid \theta)\pi(\theta) = [p(Y_{T_1} \mid \theta, Y_{T_0})] \cdot [p(Y_{T_0} \mid \theta)\pi(\theta)], \tag{7}$$

Where Y_T is the full data matrix and Y_{T_0} (the training sample) and Y_{T_1} are a partition of it into an earlier and a later segment.

Whether or not the prior is proper, it is likely that the product $p(Y_T \mid \theta)\pi(\theta)$ is integrable and hence, normalized to integrate to one, can be used as a posterior distribution. This is common practice in Bayesian scientific reporting. However, as we have already discussed, when this model is one among several being compared, if we try to use the integral of $p \cdot \pi$ as a weight w_i to form posterior probabilities over models, the results are meaningless if the prior is not proper.

The training sample method uses

$$\frac{p(Y_{T_0} \mid \theta)\pi(\theta)}{\int p(Y_{T_0} \mid \theta)\pi(\theta)\, d\theta} \tag{8}$$

as if it were the prior pdf and $p(Y_{T_1} \mid \theta, Y_{T_0})$ as if it were the likelihood. This obviously results in exactly the same posterior pdf over the parameter space Θ as the standard analysis using the full sample. But because it uses a proper "prior",

this approach to forming a weight for model comparison purposes is not obviously meaningless.

The heuristic argument in favor of training samples is that, in a situation where some priors are improper, or where we worry that we may have made priors too tight or too loose, thereby penalizing some models, the training sample method levels the playing field. The prior pdf's are scaled so that at the end of the training sample, the posterior probabilities on all models are equal. The posterior odds on models after the full sample has been brought to bear are then interpreted as the weight of evidence in the remainder of the sample, Y_{T_1}.

The heuristic argument has some plausibility when all models being compared are being given priors with the training sample method. That is, even models that have proper priors are being handled according to the training sample method. Then all models are indeed being measured against the standard of what additional evidence is provided by Y_{T_1} relative to Y_{T_0}. When, as in Smets and Wouters' use of training sample priors for VAR's in a comparison with the DSGE model that is not handled with training sample methods, it is not clear that the procedure is any less arbitrary than the naive procedure of treating integrated posteriors, based on flat priors, as if they produced legitimate model weights.

Even when all models are handled by training sample methods, there can be systematic biases for or against more heavily parameterized models. If T_0 is chosen just barely large enough to make $p \cdot \pi$ integrable for all models, the larger models will have few degrees of freedom. Their likelihood functions are therefore likely to peak at noise-ridden estimates, and their likelihood functions will be spread out. Both these effects will tend to penalize large models. On the other hand, if a fixed fraction of the sample is used as a training sample as $T \to \infty$, large models are favored, and to such an extent that the consistency property of Bayesian model comparison is lost.

Since this last point may not be widely understood, it may be helpful to explain it in more detail. The log of the integrated likelihood for a regression model with Gaussian errors has the form

$$-\frac{T-k}{2}\log(T\hat{\sigma}^2) - \frac{1}{2}\log\left|X'X\right| + \log(\Gamma\left(\frac{T-k}{2}\right) - \frac{T-k}{2}\log(\pi)\,. \tag{9}$$

When we compare two regression models with different X variable lists, but with one model having a better fit than the other (smaller limiting value for $\hat{\sigma}^2$, the difference between the two models' integrated posterior pdf's will be dominated by the first term, which is $O(T)$. If the X_t vector is stationary and ergodic and if, as when the additional variables in the larger model are redundant, $\hat{\sigma}^2$ has the same limiting value for both models, then the difference in the models' first, second and third terms are all $O((k-k')\log T$, where k and k' are the numbers of variables in the larger and smaller models, respectively. In particular the first term behaves asymptotically like $\frac{1}{2}(k-k')\log T$ and the latter two like minus this quantity, so the sum of these three components behaves asymptotically like $-\frac{1}{2}(k-k')\log T$. In other words, if both models fit equally well, posterior probability concentrates asymptotically on the model without redundant variables.

But if we use a training sample of size $T_0 = \alpha T$, with $\alpha \in (0,1)$ fixed, we lose this behavior. Posterior probability no longer concentrates asymptotically on the smaller model when both have the same residual variance. The training sample method will replace each $p(Y_T \mid \theta)\pi(\theta)$ by the same function divided by $\int p(Y_{T_0} \mid \theta)\pi(\theta)\, d\theta$. That is, the expression (9) is replaced by the difference between it and the same expression with αT replacing T. It is easy to see that this cancels out all terms that depend on T, so there is no reliable tendency for the posterior probability on the smaller model to go to infinity with T when the smaller model is true.

To conclude, training sample methods can be a handy shortcut when we apply them using the same training sample for each of a list of models. But they are always somewhat arbitrary, and are particularly dubious when the models under consideration are of widely different size or when only some of the models are given the training sample treatment.

IV.2. **Erratic posterior odds.** Smets and Wouters have used a single model with 32 parameters as their DSGE model, and have compared it to much more densely parameterized VAR's with up to 182 parameters. They have not tried to "fill in the gap" by creating models intermediate between their DSGE and VAR's. We might expect, then, that they run the risk of finding pathologies like those we expect from Bayesian model comparison with over-sparse sets of models. Their original

Model	Base	Training	Detrended	S&W detrend
VAR(3)		−330.21		
BVAR(6)	−280.87			
BVAR(5)	−277.85			
BVAR(4)	−281.23		−292.19	
BVAR(3)	−280.15	−251.26	−290.66	−266.71
Martingale	−312.06	−311.2	−272.17	
S&W DSGE				−269.2

TABLE 1. Marginal Data Densities (w_i's)

European paper (2002 found, when comparing the DSGE to BVAR's, a posterior probability for the DSGE of 0.07. Their US paper finds posterior probability for the DSGE of 1.00. Their recent paper on forecasting with the European data, which adds three years of data to the sample, finds a posterior probability on the DSGE of 0.00. These erratic and suspiciously sharp results suggest that the collection of models considered may indeed be over-sparse. As a report of research results such odds ratios are useful. They are a diagnostic suggesting that there is work to be done in expanding the range of model specifications considered. But such odds ratios would not be a useful characterization of model uncertainty for decision-making — and indeed there is little danger that a result that some particular model is the correct one with probability one will be taken seriously in policy discussion.

IV.3. **Unraveling the effects of priors and data filters on the Smets and Wouters results.** Table 1 shows variants on the marginal data density calculations Smets and Wouters carried out with the European data. The last column reproduces results from their paper. The other columns use exactly their raw data set. In the final version of this paper I hope to have the bottom row of the table filled in, using a DSGE model that is capable of handling data that has not been pre-detrended. As it is, the results can only show the order of magnitude of the effects of variations in the data set and the choice of priors, without showing clearly how these effects balance out in the DSGE vs. BVAR horse race.

Smets and Wouters "pre-detrended" their data. That is, they extracted from each logged series (in the European paper) a linear trend, fitted to the entire sample, and proceeded with the rest of their analysis as if the detrended data were raw data. This kind of pre-processing can very substantially distort measures of likelihood and of forecasting performance. It is particularly worrisome here, because the DSGE model used assumes stationarity of the data series, while the priors on the Bayesian VAR's concentrate near the *non*-stationary region of the parameter space. The preprocessing makes the series mean-reverting, whereas for most of the data series in the model the raw data appear non-stationary, or nearly so. The preprocessing thus makes it conform to the assumptions of the DSGE, while making the BVAR prior less reasonable.

The column of the table labeled "Base" shows results obtained without detrending and without training samples for several BVAR models and for a naive random walk model that simply forecasts $Y_{it} = Y_{i,t-1}$ for every variable. The random walk model does have a prior on its residual covariance matrix.

The BVAR's use a variant on the Minnesota prior, but not quite the variant that Smets and Wouters use. Here the "decay" parameter is set to one, and "sum of coefficients" (with weight 1) and "co-persistence" (with weight 5) prior components are imposed. The "decay" parameter, which determines the rate at which prior variances decline as lag increases, is set to 1, and the overall tightness is set to .3. These values are in the range of what has been found to work well most often in forecasting applications. Smets and Wouters do not include the sum-of-coefficients and co-persistence prior components (apparently) and set overall tightness to .05 and decay to 2. Their prior is therefore more concentrated on the random walk mean than is the BVAR prior used in this paper. The marginal data density values are computed analytically and non-recursively.[2] This paper also, as suggested by Sims and Zha (1998), uses a dummy observation prior component to favor a diagonal reduced form covariance matrix and pre-multiplies all components of the prior by the Jeffreys-like improper density $|\Sigma|^{-(m+1)/2}$.

[2]Software that carries out these calculations in Matlab will be available with this paper at `www.princeton.edu/~sims`

Comparing the Base and Detrended columns shows that the preliminary detrending has very substantial effects on marginal likelihood. The BVAR models, which are set up to look for complex near-unit-root behavior and cointegration, have their posterior weights reduced by a factor of $e^{10} \doteq 2 \times 10^5$ by the detrending, while the simple random walk model has its marginal likelihood even more drastically increased.

Comparing the Training column to the Base column shows that use of training priors also drastically affects results. The BVAR(3) model fits thousands of times better (in terms of posterior weight) when given a training sample "prior", while the martingale model (whose prior is only over the covariance matrix of disturbances) is almost completely unaffected.

The Base column itself shows that BVAR performance improves when additional lags are admitted, at least up to order 5. Even with the smaller decay parameter 1, the prior standard deviation of the 6th lag is only one sixth of the standard deviation of the first lag coefficient. Thus it is not surprising that beyond a certain point additional lags do not change the marginal data density much.

V. Conclusion

This paper has aimed to make suggestions for specific directions in which we can make progress toward usable characterizations of uncertainty for monetary policy analysis. It has focused particular attention on the problems that arise in using Bayesian model comparison to validate models and to characterize uncertainty across models. A still incomplete reanalysis of the model comparison calculations in one paper by Smets and Wouters suggests that the results are quite sensitive to particular choices concerning prior distributions and preliminary filtering of the data. There is apparently plenty of room for progress, to some extent in developing familiarity with and software for model comparison, but probably more importantly in expanding the list of plausible DSGE models, so that a posterior distribution over the list can provide a realistic characterization of model uncertainty.

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.

SIMS, C. A., AND T. ZHA (1998): "Bayesian Methods for Dynamic Multivariate Models," *International Economic Review*.

SMETS, F., AND R. WOUTERS (2002): "An Estimated Stochastic Dynamic General Equilibrium Model of the Euro Area*," working paper, European Central Bank and National Bank of Belgium, Frank.Smets@ecb.int,Rafael.Wouters@nbb.be.

——— (2003a): "Forecasting with a Bayesian DSGE model: an application to the euro area," Discussion paper, European Central Bank and National Bank of Belgium, Frank.Smets@ecb.int,Rafael.Wouters@nbb.be.

——— (2003b): "Shocks and Frictions in US and Euro Area Business Cycles: A Bayesian DSGE Approach," Discussion paper, European Central Bank and National Bank of Belgium, Frank.Smets@ecb.int,Rafael.Wouters@nbb.be.

——— (2003c): "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," Discussion paper, Eurpean Central Bank and National Bank of Belgium, Frank.Smets@ecb.int,Rafael.Wouters@nbb.be.

DEPARTMENT OF ECONOMICS, PRINCETON UNIVERSITY
E-mail address: sims@princeton.edu