

# Model Fit and Model Selection

Narayana R. Kocherlakota

This paper uses an example to show that a model that fits the available data perfectly may provide worse answers to policy questions than an alternative, imperfectly fitting model. The author argues that, in the context of Bayesian estimation, this result can be interpreted as being due to the use of an inappropriate prior over the parameters of shock processes. He urges the use of priors that are obtained from explicit auxiliary information, not from the desire to obtain identification. (JEL C11, E40, E60)

Federal Reserve Bank of St. Louis *Review*, July/August 2007, 89(4), pp. 349-60.

In an influential recent paper, Smets and Wouters (2003) construct a dynamic stochastic general equilibrium (DSGE) model with a large number of real and nominal frictions and estimate the unknown parameters of the model using sophisticated Bayesian techniques. They document that the estimated model has out-of-sample forecasting performance superior to that of an unrestricted vector autoregression. They write of their findings (p. 1125), “This suggests that the current generation of SDGE [stochastic dynamic general equilibrium] models with sticky prices and wages is sufficiently rich to capture the stochastics and the dynamics in the data, as long as a sufficient number of structural shocks is considered. *These models can therefore provide a useful tool for monetary policy analysis*” (italics added for emphasis). The European Central Bank (ECB) agrees. They are planning to begin using models with explicit micro-foundations for the first time in their analyses of monetary policy. In doing so, they are explicitly motivated by the Smets and Wouters (2003) analysis.<sup>1</sup>

Smets and Wouters and the ECB are adherents to what one might call the *principle of fit*. According to this principle, models that fit the available data well should be used for policy analysis; models that do not fit the data well should not be. The principle underlies much of applied economic analysis. It is certainly not special to sophisticated users of econometrics: Even calibrators who use little or no econometrics in their analyses believe in the principle of fit. Indeed, there are literally dozens of calibration papers concerned with figuring out what perturbation in a given model will lead it to fit one or two more extra moments (like the correlation between hours and output or the equity premium).

In this paper, I demonstrate that the principle of fit does not always work. I construct a simple example economy that I treat as if it were the true world. In this economy, I consider an investigator who wants to answer a policy question of interest and estimates two models to do so. I show that

<sup>1</sup> See [www.ecb.int/home/html/researcher\\_swm.en.html](http://www.ecb.int/home/html/researcher_swm.en.html) for details.

Narayana R. Kocherlakota is a professor of economics at the University of Minnesota, a consultant at the Federal Reserve Bank of Minneapolis, and a research associate at the National Bureau of Economic Research. The author thanks Ricardo DiCecio, Lee Ohanian, Tom Sargent, Adam Slawski, Hakki Yazici, Stan Zin, and especially Barbara McCutcheon for conversations about this paper. He learned much of what is in this paper from joint work that he did in the 1990s with Beth Ingram and Gene Savin when they were colleagues at the University of Iowa. The views expressed herein are the author's and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

© 2007, The Federal Reserve Bank of St. Louis. Articles may be reprinted, reproduced, published, distributed, displayed, and transmitted in their entirety if copyright notice, author name(s), and full citation are included. Abstracts, synopses, and other derivative works may be made only with prior written permission of the Federal Reserve Bank of St. Louis.

model 1, which has a perfect fit to the available data, may actually provide worse answers than model 2, which has an imperfect fit.

The intuition behind this result is quite simple. The policy question of interest concerns how labor responds to a change in the tax rate. The answer depends on the elasticity of the labor supply. In both models, the estimate of this parameter hinges on a particular non-testable assumption about how stochastic shocks to the labor-supply curve covary with tax rates. When model 2's identification restriction is closer to being correct than model 1's, model 2 provides a better answer to the policy question, even though its fit is always worse.

In the second part of the paper, I consider a potential fix. I enrich the class of possible models by discarding the non-testable assumption mentioned above. The resultant class of models is, by construction, only *partially identified*; there is a continuum of possible parameter estimates that are consistent with the observed data. I argue that, from the Bayesian perspective, a user of model 1 essentially has an incorrect prior over the set of parameters of this richer third model. As a solution, I suggest using a prior that is carefully motivated from auxiliary information, so that it does not assign zero probability to positive-probability events.

In general, there is much prior information available about behavioral parameters, such as those governing preferences and technology. However, there is much less prior information about the parameters governing shock processes. One possible response to this problem is to be what I will term *agnostic*—that is, to be fully flexible about the specification of the prior concerning the shock-process parameters. I argue in the context of the example that if one takes such an agnostic approach, the data themselves reveal no information about the behavioral parameters. I interpret this result as an indirect argument for the procedure commonly called *calibration*, in which an investigator picks a plausible range for technology and preference parameters based only on prior auxiliary information.

In the final part of the paper, I return to Smets and Wouters (2003). Using the above analysis, I

offer a critique of their approach to estimation and model evaluation. I suggest how one might change their estimation and evaluation approach to ensure more reliable policy analyses.

The second part of the paper is about Bayesian estimation of models given limited a priori information, which is, as far as I know, novel. However, the first part is not new: It is well-known that there are potential problems with the principle of fit. In early contributions, Marschak (1950) and Hurwicz (1950) emphasize that multiple structures (mappings between interventions and outcomes) may be consistent with a given reduced form (probabilistic description of available data). Liu (1960) argues that this potential problem is, in fact, endemic: The available data never serve to identify the true structure uniquely. In perhaps the most related work, Sims (1980) argues explicitly that large-scale models may fit the data well and yet provide misleading answers to (some) questions because their estimates are based on incredible identification restrictions.

Though it lacks novelty, my discussion about the principle of fit serves three purposes. First, the principle remains a dominant one among policymakers and others (as my opening paragraphs indicate). Given the recent excitement about Smets and Wouters (2003) and other related papers, it is worthwhile (I believe) to remind everyone of the principle's limitations.

Second, I want to make absolutely clear that we cannot resolve the problem by using structural models. Most macroeconomists are highly cognizant of the Lucas critique (1976). It correctly emphasizes that to assess a policy intervention a model's parameters should be structural—that is, invariant to the intervention. In response, most macroeconomists now use structural models to analyze policy interventions. My paper demonstrates that this response is not a panacea. In particular, I show that even if it is *structural* and *well-fitting*, a model may provide misleading answers to policy questions.

Finally, my argument is not just that better fit *can* lead to worse answers but that we should expect that obtaining better fit *will* lead to worse answers. Archetypal macroeconomic models are usually under-shocked relative to the data under

consideration. Hence, to get macroeconomic models that fit well, we need to add shocks. But we generally know little about these shocks. It should not be at all surprising if adding them were to create new, possibly substantial, sources of error.

## AN ARTIFICIAL WORLD AND POLICY INTERVENTIONS

The basic structure of this paper is akin to a Monte Carlo study. I first set up an artificial world over which I have complete control. I introduce an investigator (econometrician) into this artificial world who does not know the structure of the artificial world but is instead limited to using one of two possible (classes of) models. Both are false because the artificial world is not a special case of either model; however, the investigator does not know that they are false. Based on data from the artificial world, the investigator uses a variety of possible methods to determine which model has superior fit.

In this section, I describe the artificial world and a class of policy interventions under consideration in that world. In the artificial world, agents decide how much to work at each date. Their decisions are influenced by shocks to labor productivity, taxes, and preferences. The means of these random variables are hit by observable shocks in each period. Preference shocks and tax rates covary; it is this covariance that makes estimation of the parameters of the model challenging.

### The Artificial World

Time is discrete and continues forever. There is a unit measure of agents who live forever, and preferences are given by

$$\sum_{t=1}^{\infty} \delta^{t-1} \left[ \ln c_t - \exp(\psi_t) n_t^* / \gamma^* \right], 0 < \delta < 1,$$

where  $c_t$  is consumption in period  $t$  and  $n_t$  is labor in period  $t$ . Technology is given by

$$y_t = \exp(A_t) n_t,$$

where  $y_t$  is the amount of consumption produced

in period  $t$ . Agents are taxed at rate  $\tau_t$ , where  $\tau_t$  is governed by the policy rule

$$\begin{aligned} \ln(1 - \tau_t) &= -\beta \psi_t + \varepsilon_t \\ \beta &> 0. \end{aligned}$$

The proceeds of the taxes are handed back lump-sum to the agents.

The random variables  $(A_t, \psi_t, \varepsilon_t)$  are i.i.d., over time. There is another random variable  $\lambda_t$ , which is equally likely to be 0 or 1. Conditional on  $\lambda_t = i$ , the random variables  $(A, \psi, \varepsilon)$  are all Gaussian and mutually independent, with means  $(\mu_A(i), \mu_\psi(i), \mu_\varepsilon(i))_{i=0}^1$  and positive variances  $(\sigma_A^2, \sigma_\psi^2, \sigma_\varepsilon^2)$ . Note that the means depend on  $i$ , but the variances do not.

It is easy to prove that, in this economy, there is a unique equilibrium of the form

$$\begin{aligned} \ln n_t &= (\ln(1 - \tau_t) - \psi_t) / \gamma^* \\ \ln y_t &= A_t + \ln n_t. \end{aligned}$$

### Interventions

Consider the following class of *interventions*, indexed by the real variable  $\Delta$ . With intervention  $\Delta$ , the tax rate follows the rule

$$\ln(1 - \tau_t(\Delta)) = \Delta + \ln(1 - \tau_t).$$

The *policy question* is this: How much does average logged output change in response to a change in the tax rate? Mathematically, let  $y_t(\Delta)$  denote per capita output under intervention  $\Delta$ . What is  $E(\ln(y_t(\Delta^*))) - E(\ln(y_t))$ , where  $\Delta^*$  is a given intervention? The true answer to this question is  $\Delta^* / \gamma^*$ .

## TWO (IDENTIFIED) MODELS

There is an investigator who wants to know the answer to the given policy question. The investigator does not know the structure of the artificial world, but does observe the following data:

$$(\ln y_t, \ln n_t, \ln(1 - \tau_t), \lambda_t)_{t=1}^{\infty}.$$

The investigator has two possible models to use to answer the question. The basic economic ele-

ments of the models are the same as that of the artificial world itself. In each model, there is a unit measure of identical agents who work to produce output. The agents face a linear tax on output, and the proceeds of this tax are handed out lump-sum. However, the shock-generation processes in the two models are different from each other and from the artificial world.

### Model 1

In model 1, preferences are of the form

$$\sum_{t=1}^{\infty} \delta^{t-1} [\ln c_t - \exp(\psi_{1t}) n_t^{\gamma_1} / \gamma_1];$$

technology is given by  $y_t = \exp(A_{1t}) n_t$ ; and agents are taxed at rate  $\tau_t$ , where  $\ln(1 - \tau_t) = \varepsilon_{1t}$ . The random variables  $(A_{1t}, \psi_{1t}, \varepsilon_{1t})$  are i.i.d. over time and mutually independent. The random variable  $\lambda_t$  has support  $\{0, 1\}$ ; the probability that  $\lambda_t$  equals 1 is given by  $p_1$ . Conditional on  $\lambda_t = i$ , the random variables  $(A_{1t}, \psi_{1t}, \varepsilon_{1t})$  are Gaussian, with means  $(\mu_{1A}(i), \mu_{1\psi}(i), \mu_{1\varepsilon}(i))_{i=0}^1$  and variances  $(\sigma_{1A}^2, \sigma_{1\psi}^2, \sigma_{1\varepsilon}^2)$ . The investigator does not know these means and variances; they will have to be estimated in some fashion from the data. Put another way, this is actually a class of models indexed by the 11 parameters  $(\gamma_1, (\mu_{1A}(i), \mu_{1\psi}(i), \mu_{1\varepsilon}(i))_{i=0}^1, \sigma_{1\varepsilon}^2, \sigma_{1A}^2, \sigma_{1\psi}^2, p_1)$ .

Model 1 implies that in equilibrium

$$\begin{aligned} \ln(n_t) &= [\ln(1 - \tau_t) - \psi_{1t}] / \gamma_1 \\ \ln(y_t) &= A_{1t} + \ln(n_t). \end{aligned}$$

How does model 1 differ from the artificial world? It is alike in all respects except one: In model 1, the parameter  $\beta$  has been set to zero. As we shall see, this additional restriction allows the investigator to estimate  $\gamma_1$  from the available data.

### Model 2

In model 2, preferences are given by

$$\sum_{t=1}^{\infty} \delta^{t-1} [\ln c_t - \exp(\psi_2) n_t^{\gamma_2} / \gamma_2]$$

and technology is given by  $y_t = \exp[A_2(1)\lambda_t + A_2(0)(1 - \lambda_t)] n_t$ . Here,  $\psi_2$ ,  $A_2(1)$ , and  $A_2(0)$  are all

constants;  $\lambda_t$  is a random variable; and agents are taxed at rate  $\tau_t$ , where

$$\ln(1 - \tau_t) = \varepsilon_2(1)\lambda_t + \varepsilon_2(0)(1 - \lambda_t).$$

The random variable  $\lambda_t$  is i.i.d. over time, with support  $\{0, 1\}$ , and the probability that  $\lambda_t$  equals 1 is given by  $p_2$ . The parameters  $\varepsilon_2(1)$  and  $\varepsilon_2(0)$  are both constants. Hence, in model 2, there are seven unknown parameters,  $(\gamma_2, \psi_2, A_2(1), A_2(0), \varepsilon_2(1), \varepsilon_2(0), p_2)$ . The model implies that

$$\begin{aligned} \ln(n_t) &= [\ln(1 - \tau_t) - \psi_2] / \gamma_2 \\ \ln(y_t) &= A_2(1)\lambda_t + A_2(0)(1 - \lambda_t) + \ln(n_t). \end{aligned}$$

How does model 2 differ from the artificial world? In model 2, tastes do not vary at all. As well, the variances of the other shocks around their means are both set to zero. Like many modern macroeconomic models, model 2 has relatively few sources of uncertainty compared with what is true of the (artificial) world.

## THE FALLACY OF FIT

The investigator has two models available. He wants to use his infinitely long sample to decide which model to use in order to answer the policy question. The sample  $(\ln(y_t), \ln(n_t), \ln(1 - \tau_t))_{t=1}^{\infty}$  is jointly Gaussian conditional on  $\lambda_t = i$ , for  $i = 0, 1$ . The means of the conditional distributions depend on  $\lambda_t$ ; the conditional distributions have the same variance-covariance matrix. Hence, the sample can be fully summarized by 13 moments: the probability  $p$  that  $\lambda_t$  equals 1, the means  $(\mu_y(i), \mu_n(i), \mu_\tau(i))$  of  $(\ln(y), \ln(n), \ln(1 - \tau))$  conditional on  $\lambda_t = i$ , and the variance-covariance matrix  $\Sigma$  of  $(\ln(y), \ln(n), \ln(1 - \tau))$  conditional on  $\lambda$ .

Note that in these data, there are two distinct kinds of variation. The first kind is because of  $\lambda$ . Movements in  $\lambda$  generate changes in  $(\mu_y(i), \mu_n(i), \mu_\tau(i))$ ; these changes can provide information about the unknown parameters of the two models. At the same time,  $(\ln(y), \ln(n), \ln(1 - \tau))$  vary around these fluctuating means. This information is summarized by the six moments of  $\Sigma$ . The goal

of the investigator is to use these two sources of variation to estimate the unknown parameter  $\gamma$ .

Given this information, the investigator has available three methods of estimating/evaluating the models.

### Method 1: Maximum Likelihood

In this subsection, I suppose that the investigator estimates the unknown parameters of each model by maximum likelihood, and then compares the models' abilities to fit the 13 population moments.

Model 2 implies that, conditional on  $\lambda_t = i$ , the data is deterministic. In other words, according to model 2, the conditional variance-covariance matrix of  $(\ln(y_t), \ln(n_t), \ln(1-\tau_t))$  contains only zeros. It follows that the likelihood of the data, conditional on any specification of model 2, is zero.<sup>2</sup>

For model 1, the maximum-likelihood estimates of the 11 unknown parameters are given by

$$\begin{aligned}
 \hat{p}_1 &= 1/2 \\
 \hat{\gamma}_1 &= \Sigma_{\tau\tau} / \Sigma_{n\tau} \\
 \hat{\sigma}_{1A}^2 &= \Sigma_{yy} - \Sigma_{nn} \\
 \hat{\sigma}_{1\epsilon}^2 &= \Sigma_{\tau\tau} \\
 \hat{\sigma}_{1\psi}^2 &= (\hat{\gamma}_1)^2 \Sigma_{nn} - \Sigma_{\tau\tau} \\
 \hat{\mu}_{1A}(i) &= \mu_y(i) - \mu_n(i), \quad i = 0, 1 \\
 \hat{\mu}_{1\epsilon}(i) &= \mu_\tau(i) \\
 \hat{\mu}_{1\psi}(i) &= \mu_\tau(i) - \mu_n(i) \hat{\gamma}_1, \quad i = 0, 1.
 \end{aligned}
 \tag{1}$$

Given the infinitely long sample, these estimates are very precise; the likelihood of the sample is 1 given this parameter setting and zero given all others. Note that under this parameter setting, the model fits all 13 moments of the data exactly.

Hence, according to maximum likelihood, only model 1 should be used to answer the policy question (with the parameter estimates (1)); no specification of model 2 should be used. The lack of fit is because model 2 is "under shocked" rela-

tive to the data. The world has four distinct shocks generating the data, but model 2 has only one. Maximum likelihood punishes this kind of discrepancy severely; from a statistical point of view, it is the most readily detectable form of misspecification.

### Method 2: Bayesian Estimation

In this subsection, I suppose that the investigator applies Bayesian estimation methods to the available data from the artificial world. Obviously, models 1 and 2 are nested—if model 2 is true, model 1 is also true. Consider an econometrician who has a prior over the 11 unknown parameters of model 1. The prior is such that it puts probability  $q$  on the parameters being consistent with model 2 and puts probability  $(1 - q)$  on the parameters being inconsistent with model 2.

Now suppose the econometrician observes an infinite sequence of data from the artificial world. The data bleaches out the effect of the initial prior; the econometrician's posterior will be concentrated on the parameter estimates (1). A Bayesian econometrician with an infinitely long sample will reach the same policy conclusions as does a classical econometrician using maximum likelihood.

### Method 3: Method of Moments

As we have seen, maximum likelihood and Bayesian estimation simply discard all under-shocked models. Now, we consider a less severe measure of fit: method of moments, by which I mean the following. Consider the 13 population moments that characterize the sample. Pick 13 positive weights that sum to 1. Estimate the unknown parameters in each model by minimizing the weighted sum of squared deviations between model-generated moments and sample moments. Then compare model 1 and model 2 by the value of the minimized objective.

Note again that the models are nested. Because we are minimizing the same objective for each model, model 1 must do at least as well as model 2. By setting the parameters in model 1 according to (1), the objective is set equal to zero. Because model 2 (incorrectly) generates a non-invertible variance-covariance matrix for any parameter

<sup>2</sup> It is worth noting that model 2 implies that the sample variance-covariance matrix, conditional on  $\lambda_t = i$ , is noninvertible in any finite sample. Hence, the likelihood of any finite sample, conditional on model 2, is zero.

## Kocherlakota

setting, the objective must be strictly larger than 0. Model 1 must fit the data better than model 2, according to this measure of fit.

However, using method of moments, we can now actually estimate parameters for model 2, as opposed to simply discarding it as maximum likelihood does. For model 2, regardless of the weights, the estimated seven parameters are

$$\begin{aligned}\hat{p}_2 &= 1/2 \\ \hat{A}_2(i) &= [\mu_y(i) - \mu_n(i)], i = 0, 1 \\ \hat{\varepsilon}_2(i) &= \mu_\tau(i), i = 0, 1 \\ \hat{\gamma}_2 &= \frac{\mu_\tau(1) - \mu_\tau(0)}{\mu_n(1) - \mu_n(0)} \\ \hat{\psi}_2 &= [\mu_\tau(1) - \hat{\gamma}_2 \mu_n(1)] = [\mu_\tau(0) - \hat{\gamma}_2 \mu_n(0)].\end{aligned}$$

The seven parameters are set so that the model generates the values in the data for the moments  $(p, \mu_y(1), \mu_y(0), \mu_\tau(1), \mu_\tau(0), \mu_n(1), \mu_n(0))$ . Model 2 predicts that the other moments are zero for any choice of parameters, so that part of the minimization problem is irrelevant for parameter estimation.

## Using the Estimated Models to Answer the Policy Question

Recall that the policy question is this: What is the value of

$$E(\ln(y_t)(\Delta^*)) - E(\ln(y_t))$$

when taxes are changed so that  $\ln(1 - \tau_t(\Delta^*)) - \ln(1 - \tau_t) = \Delta^*$ ? The true answer to this question is given by  $\Delta^*/\gamma^*$ .

Here is what the two models deliver. Under model 1, the answer is  $\Delta^*/\hat{\gamma}_1 = \Delta^* \Sigma_{n\tau} / \Sigma_{\tau\tau}$ , where  $\Sigma_{n\tau}$  is the population covariance of  $\ln(n_t)$  and  $\ln(1 - \tau_t)$  in the artificial world, and  $\Sigma_{\tau\tau}$  is the population variance of  $\ln(1 - \tau_t)$  in the true world. (This is  $\Delta^*$  multiplied by the population regression coefficient.) We can calculate these population moments to find that that answer in model 1 is given by

$$ANS_1 = \Delta^* \left( 1/\gamma^* + \beta \gamma^{*-1} \sigma_\psi^2 / \{ \beta^2 \sigma_\psi^2 + \sigma_\varepsilon^2 \} \right),$$

which is too large in absolute value relative to the true answer of  $\Delta^*/\gamma^*$ .

Under model 2, the answer is given by

$$\begin{aligned}ANS_2 &= \frac{[\mu_n(1) - \mu_n(0)]}{[\mu_\tau(1) - \mu_\tau(0)]} \Delta^* \\ &= \Delta^* \frac{[\mu_\varepsilon(1) - \mu_\varepsilon(0)]/\gamma^* - \beta[\mu_\psi(1) - \mu_\psi(0)]/\gamma^* - [\mu_\psi(1) - \mu_\psi(0)]/\gamma^*}{\mu_\varepsilon(1) - \mu_\varepsilon(0) - \beta[\mu_\psi(1) - \mu_\psi(0)]} \\ &= \Delta^* / \gamma^* - (\Delta^* / \gamma^*) \frac{[\mu_\psi(1) - \mu_\psi(0)]}{\mu_\varepsilon(1) - \mu_\varepsilon(0) - \beta[\mu_\psi(1) - \mu_\psi(0)]}.\end{aligned}$$

Note that if  $(\mu_\psi(1) - \mu_\psi(0))$  is sufficiently close to zero in absolute value,  $ANS_2$  is nearer to  $1/\gamma^*$  than is  $ANS_1$ . Even though model 2's fit is worse than that of model 1, model 2 may still deliver a superior answer to the policy question.

## Why Doesn't Fit Work?

In the above discussion, we have seen that the model that fits better—indeed, the model that fits the available data perfectly—may well deliver a worse answer to the policy question. What is going on here? The policy question is this: What happens to hours worked if we increase the tax rate on labor? The answer is wholly governed by the elasticity of labor, which is equal to  $1/\gamma^*$  in the artificial world. To answer the question the investigator has to estimate  $\gamma^*$  well, but there is a traditional difficulty associated with doing so. If there are no shifts in the labor supply, then the comovement in hours and tax rates will pin down the elasticity of the labor supply. However, if the labor supply shifts (that is, movements in  $\psi$ ) are correlated with the variation in tax rates, then the investigator will achieve biased estimates of  $1/\gamma^*$ .

How do the two models estimate  $\gamma^*$ ? In the artificial world, there are two sorts of variation in the data. The first is that the means of the distributions of  $(A, \psi, \ln(1 - \tau))$  fluctuate over time. The second is that the realizations of the random variables fluctuate around their means. The good news is that, because  $\lambda$  is observable, the two kinds of variations are distinct. The bad news is that both kinds of fluctuations feature potential comovement between  $\psi_t$  and  $\tau_t$ —comovement that makes our task of estimating  $\gamma$  more difficult.

The two models differ in their estimates of  $\gamma$  because each one relies on a different type of

fluctuation to pin down  $\gamma$ . Model 1 assumes (incorrectly) that the fluctuations of  $\ln(1-\tau_t)$  and  $\psi_t$  around their means are independent. It then exploits the fluctuations in tax rates and hours around their means to estimate  $\gamma$ . Model 2 assumes (incorrectly) that the mean of  $\psi_t$  does not fluctuate at all. It then uses the shifts in the means of hours and tax rates over time to estimate  $\gamma$ . Which one works better depends on which incorrect assumption is a better approximation to reality. Nothing in the data answers this question.

The key point is that the relative fit of the models does not tell us which of these assumptions is closer to being right. More generally, parameter estimation of any kind always relies on two sources of information: the data and nontestable identification assumptions. The fit of a model tells us nothing about the reliability of the latter.<sup>3</sup> Yet their reliability is essential if one is to obtain accurate parameter estimates.

### PRIOR CARE

The problem with model 1 is that it includes a false restriction, which is included solely to identify the unknown parameter. In this section, I consider a richer model than model 1, in which I dispense with the false identification restriction. By construction, this model is only partially identified. I argue that one way to interpret the problem with model 1 is that the investigator is using an incorrect prior over the larger parameter space of this richer model. I suggest a simple fix to these problems: estimate the larger, partially identified model in a Bayesian fashion while being meticulous in building the prior explicitly from auxiliary information.

As before, assume that there is an investigator who has an infinite sample  $(\ln(y_t), \ln(n_t), \ln(1-\tau_t), \lambda_t)_{t=1}^\infty$  from the artificial world described

in the first section. The investigator does not use model 1 or model 2 though. Instead, the investigator uses a new model, model 3.

### Model 3

In model 3, preferences are of the form

$$\sum_{t=1}^\infty \delta^{t-1} \left[ \ln c_t - \exp(\psi_{3t}) n_t^{\gamma_3} / \gamma_3 \right], \gamma_3 \geq 0;$$

technology is given by  $y_t = \exp(A_{3t})n_t$ ; and agents are taxed at rate  $\tau_t$ , where

$$\ln(1 - \tau_t) = -\beta_3 \psi_{3t} + \varepsilon_{3t}, \beta_3 \in R.$$

The random variables  $(A_{3t}, \psi_{3t}, \varepsilon_{3t})$  are i.i.d. over time and mutually independent. The collection of random variables  $\{\lambda_t\}_{t=1}^\infty$  are i.i.d., with support  $\{0, 1\}$ ; the probability that  $\lambda_t$  equals 1 is given by  $p_3$ . Conditional on  $\lambda_t = i$ , the random variables  $(A_{3t}, \psi_{3t}, \varepsilon_{3t})$  are Gaussian, with means  $(\mu_{3A}(i), \mu_{3\psi}(i), \mu_{3\varepsilon}(i))_{i=0}^1$  and variances  $(\sigma_{3A}^2, \sigma_{3\psi}^2, \sigma_{3\varepsilon}^2)$ . In this class of models, there are 12 unknown parameters  $(\gamma_3, \beta_3, (\mu_{3A}(i), \mu_{3\psi}(i), \mu_{3\varepsilon}(i))_{i=0}^1, \sigma_{3A}^2, \sigma_{3\psi}^2, \sigma_{3\varepsilon}^2, p_3)$ .

Model 3 implies that in equilibrium

$$\begin{aligned} \ln(n_t) &= \left[ \ln(1 - \tau_t) - \psi_{3t} \right] / \gamma_1 \\ \ln(y_t) &= A_{3t} + \ln(n_t). \end{aligned}$$

Model 3 is exactly the same as model 1, except that, in model 3, tax rates may be correlated with the preference shock  $\psi_3$ . This change means that model 3 is sufficiently rich—to nest both model 1 and the artificial world.

Model 3 is only partially identified. Suppose  $\gamma_3 = \hat{\gamma}_3$ . Then, there is a unique specification of the other 11 parameters so that model 3 fits the available data exactly. In particular, let

$$\begin{aligned} \hat{p}_3 &= 1/2 \\ \hat{\sigma}_{3\psi}^2 &= (0|\hat{\gamma}_3|-1)' \Sigma(0|\hat{\gamma}_3|-1) \\ \hat{\beta}_3 &= \frac{\hat{\gamma}_3 \Sigma_{n\tau} - \Sigma_{\tau\tau}}{\hat{\sigma}_{3\psi}^2} \\ \hat{\sigma}_{3\varepsilon}^2 &= \frac{(\hat{\gamma}_3)^2 \left[ \Sigma_{nn} \Sigma_{\tau\tau} - \Sigma_{n\tau}^2 \right]}{\hat{\sigma}_{3\psi}^2} \\ \hat{\sigma}_{3A}^2 &= \Sigma_{yy} - \Sigma_{nn} \end{aligned} \tag{2}$$

<sup>3</sup> Model 1 is a just-identified model; the number of identifying restrictions is equal to the number of estimated parameters. More generally, there may be more identifying restrictions than unknown parameters. It is commonplace to construct tests of the overidentifying restrictions in such models. However, it is important to keep in mind that these are tests only of some linear combinations of the restrictions. The other linear combinations are being used to estimate the parameters and are, as in the just-identified case, nontestable.

$$\begin{aligned} \hat{\mu}_{3A}(i) &= \mu_y(i) - \mu_n(i), \quad i = 0, 1 \\ \hat{\mu}_{3\psi}(i) &= \mu_\tau(i) - \mu_n(i)\hat{\gamma}_3 \\ \hat{\mu}_{3\varepsilon}(i) &= \mu_\tau(i) + \hat{\beta}_3\hat{\mu}_{3\psi}(i), \end{aligned}$$

and then the 13 moments generated by model 3 correspond to the moments of the sample. (Note that all parameter estimates that are supposed to be non-negative [that is, variances] are in fact non-negative.) Hence, for each specification of the parameter  $\hat{\gamma}_3$ , there exists a specification of the other 11 parameters so that the model exactly fits the data.

Recently, there has been a great deal of work on classical methods to estimate partially identified models (see Manski, forthcoming, for a useful survey). However, I believe it is most useful to consider the estimation of model 3 from a Bayesian perspective.<sup>4</sup> Specifically, let  $\theta = (\beta_3, (\mu_{3A}(i), \mu_{3\psi}(i), \mu_{3\varepsilon}(i))_{i=0}^1, \sigma_{3A}^2, \sigma_{3\psi}^2, \sigma_{3\varepsilon}^2, p_3)$  represent the parameters of the model other than  $\gamma_3$ . Suppose that the parameter space for  $(\gamma_3, \theta_3)$  is given by  $R_+ \times \Theta$ , where  $\Theta = R^7 \times R_+^3 \times [0, 1]$ . This parameter space is a 12-dimensional manifold. I assume that the investigator has a prior density over this manifold such that  $\gamma_3$  is stochastically independent from  $\theta$ . I will let the marginal prior density over  $\gamma_3$  be denoted by  $f$  and the prior density over  $\theta$  be denoted by  $g$ .

A basic intuition in Bayesian estimation/learning is that the prior is essentially irrelevant if one has a large amount of data. Intuitively, the impact of the data is sufficiently large to bleach out the initial information in the prior. However, this intuition applies only when the model is identified. As we shall see, when one uses the partially identified model 3, the prior over  $(\gamma_3, \theta)$  affects the posterior distribution over  $\gamma_3$ , even though the investigator has access to an infinite sample.

<sup>4</sup> Lubik and Schorfheide (2004) use a Bayesian procedure to estimate a partially identified model. As Schorfheide (forthcoming) emphasizes, identification problems—that is, the presence of ridges or multiple peaks in the likelihood—do not create any problems for Bayesian estimation: “Regardless, the posterior provides a coherent summary of pre-sample and sample information and can be used for inference and decision making.”

## A Mistaken Prior: The Case of Model 1

Suppose  $g$  is such that the prior puts probability 1 on the event  $\beta_3 = 0$ . In this case, after seeing the available data, the investigator’s posterior is concentrated on the vector (1). With this kind of prior, model 3 is equivalent to model 1.

We have seen that using model 1 gives misleading answers to the policy question. A prior like  $g$  implicitly contains a great deal of information, because no amount of data can shift the investigator from his belief that  $\beta_3 = 0$ . It should not be used unless the investigator actually has this information about the world.

## An Arbitrary Prior

Suppose instead that  $g$  and  $f$  are such that the support of the investigator’s prior is the entire parameter space. Let  $h(\hat{\gamma}_3; Data)$  represent the parameter estimates (2) when  $\gamma_3 = \hat{\gamma}_3$ . Then, after seeing the infinite sample, the investigator’s posterior is concentrated on a one-dimensional manifold  $(\hat{\gamma}_3, h(\hat{\gamma}_3; Data))$  indexed by  $\hat{\gamma}_3 \in [\gamma_L, \gamma_H]$ . His posterior over this one-dimensional manifold is proportional to  $\phi(\hat{\gamma}_3)$ , where

$$\phi(\hat{\gamma}_3) = f(\hat{\gamma}_3)g(h(\hat{\gamma}_3; Data)) \prod_{n=1}^{11} \frac{\partial h_n(\hat{\gamma}_3; Data)}{\partial \gamma_3}.$$

(Here,  $h_n$  represents the  $n$ th component of the function  $h$ .) Given this posterior uncertainty, the investigator’s answer to the policy question is no longer a single number. Instead, the investigator’s answer is now a random variable, with support equal to the interval  $[\Delta^*/\gamma_H, \Delta^*/\gamma_L]$  and density proportional to

$$\phi(1/x) / x^2,$$

where  $x$  represents the answer to the policy question.

Because the model is only partially identified, the investigator’s posterior over the answer to the policy question is influenced by his marginal prior  $f$  over the preference parameter  $\gamma_3$  and his prior  $g$  over the other parameters. This dependence exists even though he sees an infinite sample.<sup>5</sup>

<sup>5</sup> Note that in (2), the estimates  $(\hat{\sigma}_{3A}^2, \hat{\mu}_{3A}^0, \hat{\mu}_{3A}^1, \hat{\beta}_3)$  depend only on the



This means that the investigator cannot count on the available data to eventually correct all misinformation in his initial prior. Instead, he must be sure that his prior truly represents information about these parameters derived from auxiliary sources.

### No Prior Information About Shock Processes

It is easy to see how to construct a prior  $f$  over the preference parameters (or over technology parameters in a more general context). We can derive information about such behavioral parameters from other data sources, from introspection, or from experiments. However, it is more difficult to obtain this information about the joint shock process  $(A, \tau, \psi)$ . In at least some, and perhaps most, cases, there will be no auxiliary information available about these processes. What should be done?<sup>6</sup>

In this subsection, I assume that the investigator has information that leads to a prior  $f$  with support  $[\gamma_L, \gamma_H]$ , where  $\gamma_L > 0$ . The investigator has no auxiliary information about the shock processes.

**The Bayesian Approach.** One possible response to this no-information situation is to formulate a purely subjective prior belief over the 11 parameters of the shock process and then proceed in a standard Bayesian fashion. In doing so, it is important to keep two issues in mind. First, as we have seen above, when the model is partially identified, the prior impacts the answer to the policy question regardless of how large the sample is. The subjective beliefs always matter.

Second, every prior—regardless of how neutral it may seem—has some information embedded in it. To appreciate this last point, suppose there is a parameter  $\alpha$ . All that an investigator

truly knows about  $\alpha$  is that  $\alpha$  lies in  $[0, 1]$ ; he wants his prior over  $\alpha$  to be neutral over its location within that interval. It is tempting to conclude that we can capture this neutrality by using a uniform distribution over  $[0, 1]$ . But now consider  $y = \alpha^{1000}$ . What does the investigator know about  $y$ ? Presumably, all that the investigator knows about  $y$  is that it lies in  $[0, 1]$ —if he knew more, he would have known more about  $\alpha$ . However, if the investigator has a uniform prior over  $\alpha$ , then the investigator's prior over  $y$  is proportional to  $y^{-999/1000}$ . This density is far from uniform over  $[0, 1]$ ; it places a lot more weight on low values of  $y$  than on high values of  $y$ . The uniform density over  $\alpha$  actually does smuggle information about  $\alpha$  into the analysis.

**An Agnostic Approach.** The Bayesian approach wedds the investigator to a single prior  $g$ . As I suggest above, this prior contains information that the investigator does not literally have. One response to this problem is to use what I would call an *agnostic* approach: Be flexible about the choice of  $g$  and compute a posterior density for each possible prior  $g$  over  $\gamma_3$ . By doing so, the investigator's answer to the policy question is no longer a single number, or even a single posterior, but rather a collection of posteriors generated by varying  $g$ . All of these posteriors have support  $[\Delta^*/\gamma_H, \Delta^*/\gamma_L]$ .

The resulting collection of posteriors is large. In particular, let  $p$  be any continuous probability density function over  $[\Delta^*/\gamma_H, \Delta^*/\gamma_L]$ . Let  $g_p$  be a continuous function mapping  $\Theta$  into  $R_+$  such that

$$g_p(h(1/x; Data)) = \frac{x^2 p(\Delta^{*-1}x)}{f(1/x) \prod_{n=1}^N h_n(1/x; Data)}$$

for all  $x$  in  $[1/\gamma_H, 1/\gamma_L]$ . (This pins down the behavior of  $g_p$  only on a given line in  $\Theta$ .) If the investigator's prior over  $\Theta$  is given by  $g_p$ , then his posterior over  $[\Delta^*/\gamma_H, \Delta^*/\gamma_L]$  is given by  $p$ . Thus, the agnostic approach imposes no discipline on the question of interest beyond the upper or lower bounds on  $\gamma_3$  imposed by the prior  $f$ .

This kind of agnostic analysis is reminiscent of calibration. Under calibration, an investigator uses information from auxiliary sources to pin

---

data and not on  $\hat{\gamma}_3$ . Hence, the posterior over these four parameters is concentrated on a single vector after the investigator sees an infinite sample.

<sup>6</sup> Recently, del Negro and Schorfheide (2006) have suggested using prior beliefs about endogenous variables (such as output and inflation) as a way to construct legitimate priors about exogenous shocks. This approach is potentially interesting. One concern is that usually our prior beliefs about endogenous variables come from the macroeconomic data that will, in fact, be used for estimation.

down a range of possible values for behavioral parameters. He then reports answers to the policy question for all of the parameter settings in this range. It is exactly this information that the investigator ends up reporting under the agnostic approach: a range of possible values for the policy question given the range of possible values for the behavioral parameters.

It is important to emphasize that this conclusion does not mean that the data is useless under the agnostic approach. Estimation collapses the support of the original joint prior from a 12-dimensional manifold to the 1-dimensional support of the posterior. Hence, the prior information about  $\gamma_3$ , combined with the data, does help the investigator learn a great deal about the nature of the shocks hitting the economy. It is true that this information is irrelevant given the policy question posed. For other potential questions, though, this information may well be valuable.

#### The Agnostic Approach and Decisionmaking.

Of course, more generally, the investigator may have some information about the underlying shocks that restricts the possible specifications of  $g$ . Then, the agnostic approach is not equivalent to calibration. In this general case, the agnostic approach implies that each policy intervention (each  $\Delta$ ) leads to a set of posterior probability distributions over outcomes.

It is interesting to consider the problem of choosing  $\Delta$  in this setting. Suppose there is a social welfare  $W(p)$  associated with a given posterior  $p$  over the set of outcomes. Let  $\Pi(\Delta)$  represent the set of possible posteriors implied by a given  $\Delta$ . Then, choosing  $\Delta$  is akin to optimizing under Knightian uncertainty, as opposed to risk. It is standard in such settings to use a maximin approach, under which the choice of  $\Delta$  solves the problem:

$$\max_{\Delta} \min_{p \in \Pi(\Delta)} W(p).$$

Hurwicz (1950, p. 257) provides a similar resolution to the problem of decisionmaking with partially identified models.<sup>7</sup>

<sup>7</sup> See Gilboa and Schmeidler (1989) for an axiomatization of this approach to uncertainty.

## RELATIONSHIP TO SMETS AND WOUTERS (2003)

As reported in the introduction, Smets and Wouters (2003) estimate a DSGE monetary model. They note that their model is highly similar to that of Christiano, Eichenbaum, and Evans (2005). The big difference between the two specifications is in the number of shocks: Smets and Wouters allow for 10 different shock processes. None of these shock processes represent measurement error. Instead, they all play a substantive economic role.

Smets and Wouters use a Bayesian procedure to estimate their model. As argued above, the prior plays an important role in this kind of estimation. Smets and Wouters correctly spend a great deal of time in their paper discussing the specification of the prior over the preference and technology parameters. They motivate this part of the prior thoroughly using explicit auxiliary information.

The motivation for their choice of prior over the 10 shock processes is quite different. They write (p. 1140), "Identification is achieved by assuming that each of the structural shocks [is] uncorrelated and that four of the ten shocks... follow a white noise process." In other words, they choose the prior over the 10 shock processes in order to achieve identification, not because of auxiliary information. The first example makes clear the problems with this approach. Like Smets and Wouters, the user of model 1 chooses the prior over the shock processes to achieve identification. Because this prior does not truly reflect auxiliary information, the resulting estimates are severely biased, even though the model fits the data exactly. Smets and Wouters give us no reason to believe why the same should not be true of the estimates of their model.<sup>8</sup>

The second part of the current paper suggests an alternative approach. The investigator should not pick a prior that is designed to achieve iden-

<sup>8</sup> In their recent discussion of identification of DSGE models, Canova and Sala (2006, p. 40) write that "resisting the temptation to arbitrarily induce identifiability is the only way to make DSGE models verifiable and knowledge about them accumulate on solid ground." I agree.

tification. Instead, the prior—or collection of priors—over the shock processes should reflect the investigator’s actual beliefs about those processes. The resulting set of posteriors will naturally contain less information—but also be more reliable. The key property of model 3 is that it is sufficiently rich to include as a special case the artificial world that is actually generating the data. It is certainly difficult to build such a class of models in the real world. Nonetheless, Bayesian estimation techniques (or any other for that matter) are only reliable if one does so.<sup>9</sup>

## CONCLUSIONS

A model-based analysis of a policy intervention has two steps. The first is to figure out the key parameters that shape the quantitative impact of the intervention. The second is to gather information about these parameters. This information can come in two forms: prior information and information derived from estimating the model using a particular data set. The first part of this paper argues that the fit of a model tells us little about the quality of information coming from either source. The second part of the paper argues that the latter source of information (estimation) is not useful unless the investigator has reliable prior information about shock processes.

There is an important lesson for the analysis of monetary policy. Simply adding shocks to models in order to make them fit the data better should not improve our confidence in those models’ predictions for the impact of policy changes. Instead, we need to find ways to improve our information about the models’ key parameters (for example, the costs and the frequency of price adjustments). It is possible that this improved information may come from estimation of model parameters using macroeconomic data. However, as we have seen, this kind of estimation is only useful if we have reliable a priori evidence about the shock processes. My own belief is that this kind of a priori evidence is unlikely to be avail-

able. Then, auxiliary data sources, such as the microeconomic evidence set forth by Bils and Klenow (2004), will serve as our best source of reliable information about the key parameters in monetary models.

## REFERENCES

- Bils, Mark and Klenow, Peter J. “Some Evidence on the Importance of Sticky Prices.” *Journal of Political Economy*, October 2004, 112(5), pp. 947-85.
- Canova, Fabio and Sala, Luca. “Back to Square One: Identification Issues in DSGE Models.” Working Paper 583, European Central Bank, 2006.
- Christiano, Lawrence J; Eichenbaum, Martin and Evans, Charles. “Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy.” *Journal of Political Economy*, February 2005, 113(1), pp. 1-45.
- del Negro, Marco and Schorfheide, Frank. “Forming Priors for DSGE Models and How It Affects the Assessment of Nominal Rigidities.” Working paper, University of Pennsylvania, 2006.
- Gilboa, Itzhak and Schmeidler, David. “Maxmin Expected Utility with a Non-unique Prior.” *Journal of Mathematical Economics*, 1989, 18(2), pp. 141-53.
- Hurwicz, Leonid. “Generalization of the Concept of Identification,” in Tjalling Koopmans, ed., *Statistical Inference in Dynamic Economic Models*. New York: John Wiley and Sons, 1950.
- Liu, Ta-Chung. “Underidentification, Structural Estimation, and Forecasting.” *Econometrica*, October 1960, 28(4), pp. 855-65.
- Lubik, Thomas A. and Schorfheide, Frank. “Testing for Indeterminacy: An Application to U.S. Monetary Policy.” *American Economic Review*, March 2004, 94(1), pp. 190-217.
- Lucas, Robert E. “Econometric Policy Evaluation: A Critique,” in K. Brunner and A. Meltzer, eds., *The Phillips Curve and Labor Markets*. Amsterdam: North-Holland, 1976.

<sup>9</sup> See Schorfheide (2000) for a discussion of how to augment Bayesian techniques to allow for the possibility that no model under consideration is true.

## **Kocherlakota**

- Manski, Charles F. "Partial Identification in Econometrics," in Stephen Durlauf and Lawrence Blume, eds., *The New Palgrave Dictionary of Economics*. Second Edition (forthcoming).
- Marschak, Jacob. "Statistical Inference in Economics," in T. Koopmans, ed., *Statistical Inference in Dynamic Economic Models*. New York: John Wiley and Sons, 1950.
- Schorfheide, Frank. "Loss Function Based Evaluation of DSGE Models." *Journal of Applied Econometrics*, November-December 2000, 15(6), pp. 645-70.
- Schorfheide, Frank. "Bayesian Methods in Macroeconometrics," in Stephen Durlauf and Lawrence Blume eds., *The New Palgrave Dictionary of Economics*. Second Edition (forthcoming).
- Sims, Christopher A. "Macroeconomics and Reality." *Econometrica*, January 1980, 48(1), pp. 1-48.
- Smets, Frank and Wouters, Raf. "An Estimated Stochastic Dynamic General Equilibrium Model of the Euro Area." *Journal of the European Economic Association*, September 2003, 1(5), pp. 1123-75.