Federal Reserve Bank of Minneapolis
Research Department

Posterior Simulators in Econometrics

John Geweke*

Working Paper 555

September 1995

*Geweke, Federal Reserve Bank of Minneapolis and University of Minnesota. Paper prepared for invited symposium, Seventh World Congress of the Econometric Society, Tokyo, August 22–29, 1995. Partial financial support from NSF grant SES-9210070 is gratefully acknowledged. The views expressed herein are those of the author and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System.

1. Introduction

Econometrics is the discipline of using data to revise beliefs about economic issues. In Bayesian econometrics the revision is conducted in accordance with the laws of probability, conditional on what has been observed. The normative appeal of Bayesian econometrics is the same as that of expected utility maximization and Bayesian learning, the dominant paradigms in economic theory. The questions that econometrics ultimately addresses are similar to those faced by economic agents in models, as well. Given the observed data, what decisions should be made? After bringing data to bear on two alternative models, how is their relative plausibility changed? Or more narrowly, having updated a data set should portfolio composition be changed? Any survey of the introductory and concluding sections of papers in the academic literature should provide more examples and illustrate the process of formally or informally updating beliefs.

Until quite recently applied Bayesian econometrics was undertaken largely by those primarily concerned with contributing to the theory, and the proportion of applied work that was formally Bayesian was rather small (Poirier, 1989, 1992). There are several reasons for this. First, Bayesian econometrics demands both a likelihood function and a prior distribution, whereas non-Bayesian methods do not. Second, the subjective prior distribution has to be defended, and if the reader (or worse, the editor) does not agree then the work may be ignored. Third, most posterior moments can't be obtained anyway because the requisite integrals can't be evaluated.

The development of posterior simulators in the last decade has revised beliefs about the foregoing three propositions held by many econometricians who have followed these developments closely. The purpose of this chapter is to convey these innovations and their significance for applied econometrics, to econometricians who have not followed the relevant mathematical and applied literature. There are four substantive sections. The next section reviews aspects of Bayesian inference essential to understanding the implications of posterior simulators for Bayesian econometrics. Section 3 describes these simulators and provides the essential convergence results. Implications of these procedures for some selected econometric models are drawn in Section 4. This is done to indicate the range of tasks to which posterior simulators are well suited, rather than provide a representative survey of the recent Bayesian econometric literature. [The surveys of Koop (1994), Chib and Greenberg (1994), and Geweke (1995b) take up additional models.] Finally, the chapter turns to some implications for model comparison, and for communication between those who do applied work and their

audiences, that are beginning to emerge from the use of posterior simulators in Bayesian econometrics.

2. Bayesian Inference

This section provides a quick review of the principles of Bayesian inference. The purpose is three-fold: to set up notation for the chapter; to provide an introduction for econometricians unfamiliar with Bayesian methods; and to set forth the technical challenges that posterior simulators largely overcome. Much of the notation is standard for econometric models, but differs in some important respects from that used in non-Bayesian approaches because those approaches do not condition on observables.

The introduction here is very concise and provides only the analytic essentials for the subsequent development of posterior simulators. There are few examples and at a number of points the exposition touches lightly on concepts of great depth. Those versed in Bayesian methods at the level of Berger (1985) or Bernardo and Smith (1994) can easily skip to Section 3 and use this section as a reference. Those seeking a complete introduction can consult these references, perhaps supplemented by DeGroot (1970) and Berger and Wolpert (1988) on the distinction between Bayesian and non-Bayesian methods. On Bayesian econometrics in particular, see Zellner (1971) and Poirier (1995).

The results presented in this section are not operational. In particular they all involve integrals that rarely can be evaluated analytically, and the dimensions of integration are typically greater than the four or five for which quadrature methods are practical. The balance of the chapter shows how the theory developed in this section can be implemented in applied econometrics using posterior simulators.

2.1 Basics

Inference takes place in the context of one or more models. A model describes the behavior of a $p \times 1$ vector of observables $\mathbf{y}_t$ over a sequence of discrete time units $t = 1,2,\ldots$. The history of the sequence $\{\mathbf{y}_t\}$ at time t is given by $\mathbf{Y}_t = \{\mathbf{y}_s\}_{s=1}^{t}$. A *model* is a corresponding sequence of probability density functions

(2.1.1) $$f_t(\mathbf{y}_t|\mathbf{Y}_{t-1}, \theta)$$

in which θ is a $k \times 1$ vector of unknown parameters, $\theta \in \Theta \subseteq \mathrm{R}^k$. The function "$\mathrm{p}(\cdot)$" will be used to denote a generic probability density function (p.d.f.). The p.d.f. of $\mathbf{Y}_T$, conditional on the model and parameter vector θ, is

(2.1.2) $$\mathrm{p}(\mathbf{Y}_T|\theta) = \prod_{t=1}^{T} f_t(\mathbf{y}_t|\mathbf{Y}_{t-1}, \theta)$$

The *likelihood function* is any function $L(\theta;\mathbf{Y}_T) \propto p(\mathbf{Y}_T|\theta)$.

[If the model specifies that the $\mathbf{y}_t$ are independent and identically distributed then $f_t(\mathbf{y}_t|\mathbf{Y}_{t-1},\theta) = f_t(\mathbf{y}_t|\theta)$ and $p(\mathbf{Y}_T|\theta) = \prod_{t=1}^{T} f_t(\mathbf{y}_t|\theta)$. More generally, the index "t" may pertain to cross sections, to time series, or both, but time series models and language are used here for specificity. Likewise it is assumed that $\mathbf{y}_t$ is continuously distributed for specificity and brevity.]

The objective of Bayesian inference can in general be expressed

(2.1.3) $\qquad E[g(\theta)|\mathbf{Y}_T],$

in which $g(\theta)$ is a *function of interest.* There are several broad categories of functions of interest that between them encompass most applied econometric work. Clearly the function of interest can be a parameter or a function of parameters. Another category is $g(\theta) = L(a_1,\theta) - L(a_2,\theta)$ in which $L(a,\theta)$ is the loss function pertaining to action a, parameter vector θ, and (implicitly, through (2.1.3)) the model itself. A third category is $g(\theta) = \chi_{\Theta_0}(\theta)$ which arises when a hypothesis restricts θ to a set Θ_0. [Here $\chi(\cdot)$ is the characteristic function $\chi_S(z) = 1$ if $z \in S$, $\chi_S(z) = 0$ if $z \notin S$.] Then $E[g(\theta)|\mathbf{Y}_T] = P(\theta \in \Theta_0|\mathbf{Y}_T)$. Yet another important category arises from predictive densities. Denote $\mathbf{y}^* = (y_{T+1},\ldots,y_{T+f})'$. If $g(\theta) = E[h(\mathbf{y}^*)|\mathbf{Y}_T,\theta]$, then $E[g(\theta)|\mathbf{Y}_T] = E[h(\mathbf{y}^*)|\mathbf{Y}_T]$ Through the appropriate choice of $h(\mathbf{y}^*)$ this category includes point prediction, turning point probabilities, and predictive intervals.

The subject of this chapter is generic, numerical methods of evaluating (2.1.3). To the extent a method is generic, it can be applied to many models without requiring special adaptation, and most of the questions of applied econometrics can be addressed directly. This chapter concentrates on numerical methods because the combinations of models and functions of interest for which (2.1.3) can be evaluated analytically are quite limited. It takes up posterior simulators in particular because this approach is quite general and is especially attractive when the parameter space is high dimensional (i.e., k is large) or the model involves latent variables.

The specification of the model (2.1.1) is completed with a *prior density* $p(\theta)$. It may be shown that given (2.1.1) and a density $p(\mathbf{Y}_T)$ (i.e., a density for the data *un*conditional on θ) a prior density must exist; see Bernardo and Smith (1994, Section 4.2). It is more direct to place the specification of the prior density on the same logical footing as the specification of (2.1.1). Thus a *complete model* specifies

(2.1.4) $\qquad P(\theta \in \tilde{\Theta}) = \int_{\tilde{\Theta}} p(\theta)d\theta, \quad P(\mathbf{Y}_T \in \tilde{Y}|\theta) = \int_{\tilde{Y}} \prod_{t=1}^{T} f_t(\mathbf{y}_t|\mathbf{Y}_{t-1},\theta)d\mathbf{Y}_T,$

where $\tilde{\Theta}$ is any Lebesgue-measurable subset of Θ and $\tilde{Y}$ is any Lebesgue-measurable subset of R^{pT}. [To keep the notation simple, a strictly continuous prior probability distribution for θ is assumed.]

By Bayes Theorem the *posterior density* of θ is

$$\begin{aligned} p(\theta|\mathbf{Y}_T) &= p(\mathbf{Y}_T|\theta)p(\theta)/p(\mathbf{Y}_T) \\ &\propto p(\mathbf{Y}_T|\theta)p(\theta) \\ &\propto L(\theta;\mathbf{Y}_T)p(\theta). \end{aligned}$$

Thus

$$E[g(\theta)|\mathbf{Y}_T] = \int_\Theta g(\theta)p(\theta|\mathbf{Y}_T)d\theta = \frac{\int_\Theta g(\theta)L(\theta;\mathbf{Y}_T)p(\theta)d\theta}{\int_\Theta L(\theta;\mathbf{Y}_T)p(\theta)d\theta}. \tag{2.1.5}$$

In the representation (2.1.5), one may substitute for $p(\theta)$ any function $p^*(\theta) \propto p(\theta)$. The function $p^*(\theta)$ is a *kernel* of the prior density $p(\theta)$. Posterior moments in a given model are invariant to any arbitrary scaling of either the likelihood function or the prior density.

2.2 Sufficiency, ancillarity, and nuisance parameters

The vector $\mathbf{s}_T = s_T(\mathbf{Y}_T)$ is a sufficient statistic in the model (2.1.2) given any of the following equivalent conditions:

$$p[\mathbf{Y}_T|s_T(\mathbf{Y}_T),\theta] = p[\mathbf{Y}_T|s_T(\mathbf{Y}_T)] \;\forall\; \theta \in \Theta; \tag{2.2.1}$$

$$p(\theta|\mathbf{Y}_T) = p[\theta|s_T(\mathbf{Y}_T)] \;\forall\; \theta \in \Theta \text{ for all realizations } \mathbf{Y}_T; \tag{2.2.2}$$

$$p(\mathbf{Y}_T|\theta) = h[s_T(\mathbf{Y}_T),\theta]r(\mathbf{Y}_T) \text{ for some } h(\cdot) \text{ and } r(\cdot). \tag{2.2.3}$$

Condition (2.2.3), the *Neyman factorization criterion*, is the condition usually verified to demonstrate sufficiency of $\mathbf{s}_T = s_T(\mathbf{Y}_T)$. Sufficiency implies that one may use the (sometimes much simpler) expression $h[s_T(\mathbf{Y}_T),\theta]$ in lieu of the likelihood function in (2.1.5).

If $s_T(\mathbf{Y}_T)' = [s_{1T}(\mathbf{Y}_T)', s_{2T}(\mathbf{Y}_T)']$ and $p[s_{1T}(\mathbf{Y}_T)|\theta] = p[s_{1T}(\mathbf{Y}_T)]$, then $s_{1T}(\mathbf{Y}_T)$ is *ancillary with respect to* θ. As a consequence, it suffices to use any function proportional to $p[s_{2T}(\mathbf{Y}_T)|\theta]$ in lieu of the likelihood function in (2.1.5).

If $\theta' = (\theta_1', \theta_2')$ and $g(\theta) = g(\theta_1)$ then θ_2 is a *nuisance parameter* for the function of interest $g(\theta)$. A nuisance parameter presents no special problems in (2.1.5).

2.3 Point estimation and credible sets

Let the $q \times 1$ vector $\omega \in \Omega$ represent an unknown state of the world: for example, ω could be the parameter vector θ itself, a function of interest $g(\theta)$, or a vector of future

values $\mathbf{y}^* = (y_{T+1},\ldots,y_{T+f})'$. Let $\tilde{\omega} \in \tilde{\Omega} \subseteq \Omega$ represent an estimate of ω. The *Bayes estimate* of ω corresponding to the loss function $L(\tilde{\omega},\omega)$ is

(2.2.1) $$\hat{\omega} = \arg\min_{\tilde{\omega}} E[L(\tilde{\omega},\omega)|\mathbf{Y}_T].$$

[Clearly, the estimate $\hat{\omega}$ depends on the complete model (2.1.4) as well as the loss function $L(\tilde{\omega},\omega)$. But given the model and loss function, there is no ambiguity about the Bayes estimate.]

Three loss functions are notable for the simplicity of the Bayes estimates $\hat{\omega}$ that they imply:

given *quadratic loss* $L(\tilde{\omega},\omega) = (\tilde{\omega}-\omega)'\mathbf{Q}(\tilde{\omega}-\omega)$ (where $\mathbf{Q}$ p.d., $\tilde{\omega} \in R^q$), $\hat{\omega} = E(\omega|\mathbf{Y}_T)$;

given *quantile loss* $L(\omega,\tilde{\omega}) = c_1(\tilde{\omega}-\omega)\chi_{(-\infty,\tilde{\omega})}(\omega) + c_2(\omega-\tilde{\omega})\chi_{(\tilde{\omega},\infty)}(\omega)$ (where $c_1 > 0$, $c_2 > 0, q = 1$), $\hat{\omega} = \tilde{\omega}: P(\omega \le \tilde{\omega}|\mathbf{Y}_T) = c_2/(c_1+c_2)$ and hence if $c_1 = c_2$ the Bayes estimate of ω is the median of its posterior distribution;

given *0/1 loss* $L(\tilde{\omega},\omega) = 1 - \chi_{N_\varepsilon(\tilde{\omega})}(\omega)$ (where $N_\varepsilon(\tilde{\omega})$ is an ε-neighborhood of $\tilde{\omega}$), as $\varepsilon \to 0$, $\hat{\omega}$ converges to the global mode of $p(\omega|\mathbf{Y}_T)$ if a global mode exists.

All three estimators are derived in most texts in Bayesian statistics, e.g. Berger (1985, Section 2.4.2) or Bernardo and Smith (1994, Proposition 5.2)

A $100(1-\alpha)\%$ *credible set for* ω is any set C such that $\int_C p(\omega|\mathbf{Y}_T)d\omega = 1-\alpha$. The credible set depends on the complete model (2.1.4) but is defined without reference to a loss function because it does not involve a Bayes action. In general a credible set can be defined with reference to any distribution for ω, not just the posterior distribution. In most cases (always, for continuous distributions) the credible set is not unique.

If $p(\omega_1|\mathbf{Y}_T) \ge p(\omega_2|\mathbf{Y}_T)\ \forall\ (\omega_1,\omega_2): \omega_1 \in C, \omega_2 \in \Omega - C$, except possibly for a subset of Ω with posterior probability 0, then C is a *highest posterior density (HPD) credible set for* ω. It can be shown that HPD sets provide the credible sets with smallest Lebesgue measure. Therefore the choice of a HPD set is a Bayes action if loss is proportional to the Lebesgue measure of the credible set.

Since credible sets are defined with respect to a probability measure they are invariant under one-to-one transformations: i.e., if $\nu = h(\omega)$, $h(\cdot)$ is one-to-one, and C is a $100(1-\alpha)\%$ credible set for ω, then $D = \{\nu: \nu = h(\omega), \omega \in C\}$ is a $100(1-\alpha)\%$ credible set for ν. However, HPD credible sets are not invariant under transformation. [The technical step involves the Jacobian of transformation. For demonstration and

further discussion see Berger (1985, pp. 144-145) or Bernardo and Smith (1994, pp. 261-262).]

2.4 Prior distributions

The complete model (2.1.4) provides a representation of belief. The choice of model is always a judicious compromise between realistic richness in form and the effort required to obtain posterior moments $E[g(\theta)|\mathbf{Y}_T]$. To this end, it has proven useful to employ classes of prior densities, $p(\theta|\tau)$ where τ is an indexing parameter, just as it has proven useful to index the conditional density $f_t(\mathbf{y}_t|\mathbf{Y}_{t-1},\theta)$ by θ.

Suppose that $p(\mathbf{Y}_T|\theta), \theta \in \Theta$ has sufficient statistic $\{T, s_T(\mathbf{Y}_T)\}$, where $s_T(\mathbf{Y}_T)$ is a vector whose dimension is independent of T and $\mathbf{Y}_T$. Then the *conjugate family of prior densities for* θ *with respect to* $p(\mathbf{Y}_T|\theta)$ is

$$\{p(\theta|\tau), \tau \in \mathbf{T}; \tau_0\}$$

where

$$\mathbf{T} = \left\{\tau: \int_\Theta p[s_T(\mathbf{Y}_{\tau_0}) = \tau|\theta]d\theta\right\} < \infty$$

and

$$p(\theta|\tau) = p[s_T(\mathbf{Y}_{\tau_0}) = \tau|\theta] \Big/ \int_\Theta p[s_T(\mathbf{Y}_{\tau_0}) = \tau|\theta]d\theta.$$

A conjugate prior distribution for θ is thus proportional to a likelihood function composed of τ_0 observations whose sufficient statistics are given in the vector τ. Less formally, the information about θ in a conjugate prior distribution is equivalent to the information about θ in a likelihood function with τ_0 imaginary observations and sufficient statistic τ.

There is an extensive literature providing conjugate families of prior distributions corresponding to various specifications of $f_t(\mathbf{y}_t|\mathbf{Y}_{t-1},\theta)$. A strong practical reason for this effort is that in the presence of a conjugate prior distribution, the posterior distribution will retain the same mathematical tractability that characterizes $p(\mathbf{Y}_T|\theta)$ and was likely an important reason for the choice of $f_t(\mathbf{y}_t|\mathbf{Y}_{t-1},\theta)$ in the first place. For example, in the regular *exponential family of distributions*

$$p(\mathbf{Y}_T|\theta) = [s(\theta)]^T \prod_{t=1}^T r(\mathbf{y}_t) \exp\left\{\sum_{i=1}^m c_i \phi_i(\theta)\left[\sum_{t=1}^T h_i(\mathbf{y}_t)\right]\right\}$$

the conjugate family for θ is

$$p(\theta|\tau) \propto [s(\theta)]^{\tau_0} \exp\left[\sum_{i=1}^m c_i \phi_i(\theta)\tau_i\right],$$

$$\tau \in \mathbf{T} = \left\{\tau: \int_\Theta [s(\theta)]^{\tau_0} \exp\left[\sum_{i=1}^m c_i \phi_i(\theta)\tau_i\right] < \infty\right\}$$

and then

(2.4.1) $$p(\theta|\mathbf{Y}_T) \propto [s(\theta)]^{\tau_0+T} \exp\left\{\sum_{i=1}^{m} c_i \phi_i(\theta)\left[\sum_{t=1}^{T} \tau_i + h_i(\mathbf{y}_t)\right]\right\}.$$

If $\theta' = (\theta_1', \theta_2')$ and the value of $\theta_2 = \theta_2^0$ is fixed, then one may define the *conditionally conjugate family of prior densities for* θ_1 *with respect to* $p(\mathbf{Y}_T|\theta_1, \theta_2^0)$ in precisely the same way. Given purely analytical approaches to Bayesian inference the use of conjugate prior distributions is almost always essential. With the advent of the numerical approaches that are the focus of this chapter conjugate prior distributions are no longer essential, but are often useful as belief representations and can simplify computation. Numerical approaches have rendered Bayesian inference practical in models so complex that conjugate prior distributions do not provide simple belief representations. In these cases, conditionally conjugate priors are often more useful and provide computational advantages, as will be seen in Section 4.

The prior distribution, even if it is restricted to a conjugate family, provides a flexible representation of prior beliefs. It is tempting to characterize prior distributions by the extent to which they provide information about parameters. At one extreme, a prior distribution with all its mass at a single point $\theta^* \in \Theta$ is clearly quite informative; such a prior is said to be *dogmatic*. At the other extreme, what (if anything) constitutes an uninformative prior distribution is less clear.

The desire to work with less informative prior distributions leads to an extension of prior distributions that can be useful if applied carefully. Consider a sequence of prior density *kernels* $p_j^*(\theta)$: i.e., $\int_\Theta p_j^*(\theta)d\theta < \infty$ and the corresponding prior density is $p_j(\theta) = p_j^*(\theta) \Big/ \int_\Theta p_j^*(\theta)d\theta$. Suppose further that $\lim_{j\to\infty} p_j(\theta) = 0$ and $\lim_{j\to\infty} p_j^*(\theta) = p^*(\theta)\ \forall\ \theta \in \Theta$, but that $\int_\Theta p^*(\theta)d\theta$ is divergent. It is often the case that $\int_\Theta L(\theta;\mathbf{Y}_T)p^*(\theta)d\theta$ and $\int_\Theta g(\theta)L(\theta;\mathbf{Y}_T)p^*(\theta)d\theta$ are convergent and furthermore

$$\lim_{j\to\infty} \frac{\int_\Theta g(\theta)L(\theta;\mathbf{Y}_T)p_j^*(\theta)d\theta}{\int_\Theta L(\theta;\mathbf{Y}_T)p_j^*(\theta)d\theta} = \frac{\int_\Theta g(\theta)L(\theta;\mathbf{Y}_T)p^*(\theta)d\theta}{\int_\Theta L(\theta;\mathbf{Y}_T)p^*(\theta)d\theta}.$$

In this case the formal use of the "prior density" $p^*(\theta)$ has an unambiguous interpretation and provides correct posterior moments. If $p^*(\theta)$ is the limit of kernels of conjugate prior densities then it generally retains the analytical advantages of the conjugate family. For example in the regular exponential family with conjugate priors, if $\tau^{(j)} = (\tau_0^{(j)}, \ldots, \tau_m^{(j)}) \xrightarrow[j\to\infty]{} \mathbf{0}$ then the limiting posterior distribution is given by (2.4.1) with $\tau_i = 0\ (i = 0, \ldots, m)$. Formal analysis with $p^*(\theta) = 1$ would have led to the same result.

2.5 Robustness

An important part of any thorough econometric investigation is establishing the sensitivity of key conclusions to various aspects of model specification. To the extent the conclusions in question are insensitive to specification the model is *robust.* A key step in robustness analysis is setting up the aspects of the specification to be varied. In a closed robustness analysis one specifies a finite number of alternative models (2.1.4) and compares posterior moments over specifications. Such a comparison is of limited usefulness, mainly because one is typically concerned with variations in specification that are not well captured by a few models. If a few models suffice then the method of model averaging described below is applicable.

In an open robustness analysis one specifies an entire class of models and determines the corresponding range of posterior moments. For example, if a class of prior distributions is indexed by a parameter vector τ then one may define

$$\text{(2.5.1)} \qquad \mathrm{E}\left[\mathrm{g}(\theta)\middle|\mathbf{Y}_T,\tau\right]=\frac{\int_\Theta \mathrm{g}(\theta)\mathrm{L}(\theta;\mathbf{Y}_T)\mathrm{p}(\theta|\tau)d\theta}{\int_\Theta \mathrm{L}(\theta;\mathbf{Y}_T)\mathrm{p}(\theta|\tau)d\theta}.$$

In this approach it is necessary, first, to specify an appropriate range for τ, and second, to determine the corresponding range of the posterior moment (2.5.1).

Robustness analysis may also be nonparametric. An example is provided by the *density ratio class* of prior distributions. This class consists of prior densities whose kernels $\mathrm{p}^*(\theta)$ may be chosen to satisfy $\mathrm{a}(\theta)\le \mathrm{p}^*(\theta)\le \mathrm{b}(\theta)$, where $\mathrm{a}(\theta)$ and $\mathrm{b}(\theta)$ are specified bounding functions. Constraints of this form place upper and lower bounds on $\mathrm{P}(\theta\in\Theta_A)/\mathrm{P}(\theta\in\Theta_B)$ for all pairs of Lebesgue measurable Θ_A and Θ_B contained in Θ. The posterior moment $\mathrm{E}\left[\mathrm{g}(\theta)\middle|\mathbf{Y}_T,\mathrm{p}^*(\theta)\right]$ is maximized over all prior densities in the density ratio class by setting the kernel

$$\text{(2.5.2)} \qquad \mathrm{p}^*(\theta)=\overline{\mathrm{p}}(\theta)=\begin{cases}\mathrm{a}(\theta) \text{ if } \mathrm{g}(\theta)<g^*\\ \mathrm{b}(\theta) \text{ if } \mathrm{g}(\theta)>g^*\end{cases}$$

where g^* satisfies the fixed-point condition

$$\text{(2.5.3)} \qquad g^*=\mathrm{E}\left[\mathrm{g}(\theta)\middle|\mathbf{Y}_T,\overline{\mathrm{p}}(\theta)\right].$$

[The result is due to DeRobertis and Hartigan (1981). See also Lavine (1991a, 1991b).]

The density ratio class is one of the most convenient for studying robustness with respect to the prior distribution, but many other classes have been studied. Berger (1985) provides a good introduction, and a thorough survey is given in Berger (1994).

2.6 Model averaging

Typically one has under consideration several complete models of the form (2.1.4). For specificity suppose there are J models, and distinguish model M_j by the subscript "j":

$$\mathrm{P}_j\left(\theta_j \in \tilde{\Theta}_j\right) = \int_{\tilde{\Theta}_j} \mathrm{p}_j\left(\theta_j\right) d\theta_j, \quad \mathrm{P}_j\left(\mathbf{Y}_T \in \tilde{Y} \middle| \theta_j\right) = \int_{\tilde{Y}} \prod_{t=1}^{T} \mathrm{f}_{jt}\left(\mathrm{y}_t \middle| \mathbf{Y}_{t-1}, \theta_j\right) d\mathbf{Y}_T .$$

The J models are related by their description of a common set of observations $\mathbf{Y}_T$ and a common vector of interest ω. The number of parameters in the models may or may not be the same and various models may or may not nest one another. The vector of interest ω -- e.g., the outcome of a change in policy, or actual future values of y_t -- is substantively the same in all models although its representation in terms of θ_j may vary greatly from one model to another. Each model specifies its conditional p.d.f. for ω, $\mathrm{p}_j\left(\omega \middle| \theta_j, \mathbf{Y}_T\right)$. The specification of the collection of J models is completed with the prior probabilities $p_j \ (j = 1, \ldots, J)$, $\sum_{j=1}^{J} p_j = 1$.

There are now three levels of conditioning. Given model j and θ_j, the p.d.f. of $\mathbf{Y}_T$ is $\mathrm{p}_j\left(\mathbf{Y}_T \middle| \theta_j\right)$. Given only model j, the p.d.f. of θ_j is $\mathrm{p}_j\left(\theta_j\right)$. And given the collection of models $M_1, \ldots, M_J$ the probability of model j is p_j. If the collection of models changes then the p_j will change in accordance with the laws of conditional probability. There is no essential conceptual distinction between model and prior: one could just as well regard the entire collection as the model, with $\left\{p_j, \mathrm{p}_j\left(\theta_j\right)\right\}_{j=1}^{J}$ as the characterization of the prior distribution. At an operational level the distinction is usually quite clear and useful: one may undertake the essential computations one model at a time.

Suppose that the posterior moment $\mathrm{E}\left[\mathrm{h}(\omega) \middle| \mathbf{Y}_T\right]$ is ultimately of interest. (This expression is just as general as (2.1.3) and encompasses the particular cases discussed there.) The formal solution is

$$\mathrm{E}\left[\mathrm{h}(\omega) \middle| \mathbf{Y}_T\right] = \sum_{j=1}^{J} \mathrm{E}\left[\mathrm{h}(\omega) \middle| \mathbf{Y}_T, M_j\right] \mathrm{P}\left(M_j \middle| \mathbf{Y}_T\right). \tag{2.6.1}$$

From (2.1.5),

$$\mathrm{E}\left[\mathrm{h}(\omega) \middle| \mathbf{Y}_T, M_j\right] = \frac{\int_{\Theta_j} \mathrm{g}\left(\theta_j\right) \mathrm{L}_j\left(\theta_j; \mathbf{Y}_T\right) \mathrm{p}_j\left(\theta_j\right) d\theta_j}{\int_{\Theta_j} \mathrm{L}_j\left(\theta_j; \mathbf{Y}_T\right) \mathrm{p}_j\left(\theta_j\right) d\theta_j} \tag{2.6.2}$$

with $\mathrm{g}\left(\theta_j\right) = \int_{\omega} \mathrm{h}(\omega) \mathrm{p}_j\left(\omega \middle| \theta_j, \mathbf{Y}_T\right) d\omega$. There is nothing new in this part of (2.6.1). From Bayes' rule,

$$\begin{aligned} \mathrm{P}(M_j|\mathbf{Y}_T) &= \mathrm{p}(\mathbf{Y}_T|M_j)\mathrm{P}(M_j)/\mathrm{p}(\mathbf{Y}_T) \\ &= p_j \int_{\Theta_j} \mathrm{p}_j(\mathbf{Y}_T|\theta_j)\mathrm{p}_j(\theta_j)d\theta_j / \mathrm{p}(\mathbf{Y}_T) \\ &\propto p_j \int_{\Theta_j} \mathrm{p}_j(\mathbf{Y}_T|\theta_j)\mathrm{p}_j(\theta_j)d\theta_j = p_j M_{jT} \end{aligned} \tag{2.6.3}$$

The value M_{jT} is known as the *marginalized likelihood* of Model j. The name reflects the fact that one can write

$$M_{jT} = \int_{\Theta_j} \mathrm{L}_j(\theta_j;\mathbf{Y}_T)\mathrm{p}_j(\theta_j)d\theta_j . \tag{2.6.4}$$

Expression (2.6.4) must be treated with caution, because the likelihood function typically introduces convenient, model-specific proportionality constants: $\int_{\tilde{Y}} \mathrm{p}_j(\mathbf{Z}_T|\theta_j)d\mathbf{Z}_T = 1$ but $\int_{\tilde{Y}} \mathrm{L}_j(\theta_j;\mathbf{Z}_T)d\mathbf{Z}_T \neq 1$. Whereas (2.6.2), like (2.1.5), is invariant to arbitrary renormalizations of $\mathrm{p}_j(\mathbf{Y}_T|\theta_j)$ and $\mathrm{p}_j(\theta_j)$, (2.6.3) is valid only with the conditional p.d.f.'s themselves, not their kernels. As a simple corollary, model averaging cannot be undertaken using improper prior distributions, a point related to Lindley's paradox described below.

Model averaging thus involves three steps. First, obtain the posterior moments (2.6.2) corresponding to each model. Second, obtain the marginalized likelihood M_{jT} from (2.6.3). Finally, obtain the posterior moment using (2.6.1) which now only involves simple arithmetic. Variation of the prior model probabilities p_j is a trivial step, as is the revision of the posterior moment following the introduction of a new model or deletion of an old one from the conditioning set of models, if (2.6.2) and (2.6.4) for those models are known.

2.7 Hypothesis testing

Formally, *hypothesis testing* is the problem of choosing one model from several. With no real loss of generality assume there are only two models in the choice set. Treating model choice as a Bayes action, let $\mathrm{L}(i|j)$ denote the loss incurred in choosing model i when model j is true and suppose that $\mathrm{L}(i|i) = 0$ and $\mathrm{L}(i|j) > 0\ (j \neq i)$. Given the data $\mathbf{Y}_T$ the expected loss from choosing model i is $\mathrm{P}(M_j|\mathbf{Y}_T)\mathrm{L}(i|j)$ $(j \neq i)$ and so the Bayes action is to choose model 1 if and only if

$$\frac{\mathrm{P}(M_1|\mathbf{Y}_T)}{\mathrm{P}(M_2|\mathbf{Y}_T)} = \frac{p_1 M_{1T}}{p_2 M_{2T}} > \frac{\mathrm{L}(1|2)}{\mathrm{L}(2|1)}.$$

The value $\mathrm{L}(1|2)/\mathrm{L}(2|1)$ is known as the *Bayes critical value*. The data bear on model choice only through the ratio M_{1T}/M_{2T}, known as the *Bayes factor* in favor of Model 1. The term $p_1 M_{1T}/p_2 M_{2T}$ is the *posterior odds ratio* in favor of Model 1. For reasons of

economy an investigator may therefore report only the marginalized likelihood, leaving it to his or her *clients* -- i.e, the users of the investigator's research -- to provide their own prior model probabilities and loss functions. The steps of reporting marginalized likelihoods and Bayes factors are sometimes called hypothesis testing as well.

It is instructive to consider briefly the choice between two models given a sequence of prior distributions $p_{1j}(\theta_1)$ in Model 1 in which $\lim_{j\to\infty} p_{1j}(\theta_1) = 0 \; \forall \; \theta_1 \in \Theta_1$. It was seen in Section 2.4 that the limiting posterior moment in Model 1 can be well-defined in this case, and that it may be found conveniently using a corresponding sequence of convergent prior density kernels. The condition $\lim_{j\to\infty} p_{1j}(\theta_1) = 0 \; \forall \; \theta_1 \in \Theta_1$ ensures $\lim_{j\to\infty} M_{jT} = 0$, however. Therefore, if the prior distribution in Model 1 is improper whereas that in Model 2 is proper, the hypothesis test cannot conclude in favor of Model 1. This result is widely known as *Lindley's paradox*, after Lindley (1957) and Bartlett (1957).

As will be seen, the computation of marginalized likelihoods has been a substantial technical challenge. The reason is that in general M_{jT} cannot be cast as a special case of (2.1.5). In specific settings, however, (2.1.5) may be used to express Bayes factors. A common one is that in which models 1 and 2 have a common likelihood function and differ only in their prior densities $p_j(\theta)$. Then the Bayes factor in favor of Model 1 is

$$\frac{M_{1T}}{M_{2T}} = \frac{\int_\Theta g(\theta) L(\theta; \mathbf{Y}_T) p_2(\theta) d\theta}{\int_\Theta L(\theta; \mathbf{Y}_T) p_2(\theta) d\theta} \tag{2.7.1}$$

with

$$g(\theta) = p_1(\theta)/p_2(\theta). \tag{2.7.2}$$

2.8 Hierarchical priors and latent variable models

A *hierarchical prior distribution* expresses the prior in two or more steps. The two-step case specifies a model $p_A(\mathbf{Y}_T|\theta) \; (\theta \in \Theta)$ and a prior density for θ conditional on a *hyperparameter* ϕ, $p_B(\theta|\phi) \; (\phi \in \Phi)$. The model is completed with a prior density for ϕ, $p_C(\phi)$. There is no fundamental difference between this prior density and the one described in Section 2.4, since

$$p(\theta) = \int_\Phi p_B(\theta|\phi) p_c(\phi) d\phi. \tag{2.8.1}$$

As will be seen, however, the hierarchical formulation is often so convenient as to render fairly simple problems that otherwise would be essentially impossible. Given a hierarchical prior, one may express both a posterior density for θ,

$$\text{(2.8.2)}\qquad \begin{aligned} p(\theta|\mathbf{Y}_T) &\propto \int_\Phi p_A(\mathbf{Y}_T|\theta)\,p_B(\theta|\phi)\,p_C(\phi)\,d\phi \\ &= p_A(\mathbf{Y}_T|\theta)\int_\Phi p_B(\theta|\phi)\,p_C(\phi)\,d\phi = p_A(\mathbf{Y}_T|\theta)\,p(\theta), \end{aligned}$$

and a posterior density for ϕ,

$$\text{(2.8.3)}\qquad p(\phi|\mathbf{Y}_T) \propto \int_\Theta p_A(\mathbf{Y}_T|\theta)\,p_B(\theta|\phi)\,p_C(\phi)\,d\theta \propto p(\mathbf{Y}_T|\phi)\,p_C(\phi).$$

A *latent variable model* expresses the likelihood function in two or more steps. In the two-step case the likelihood function may be written $p_A(\mathbf{Y}_T|\mathbf{Z}_T^*)$ $(\mathbf{Z}_T^* \in \tilde{\mathbf{Z}})$ where $\mathbf{Z}_T^*$ is a matrix of latent variables. The model for $\mathbf{Z}_T^*$ is $p_B(\mathbf{Z}_T^*|\phi)$ $(\phi \in \Phi)$ with prior density $p_C(\phi)$. The prior density induces an unconditional density for $\mathbf{Z}_T^*$,

$$\text{(2.8.4)}\qquad p(\mathbf{Z}_T^*) = \int_\Phi p_B(\mathbf{Z}_T^*|\phi)\,p_C(\phi)\,d\phi.$$

The posterior density for the latent variables is

$$\text{(2.8.5)}\qquad \begin{aligned} p(\mathbf{Z}_T^*|\mathbf{Y}_T) &\propto \int_\Phi p_A(\mathbf{Y}_T|\mathbf{Z}_T^*)\,p_B(\mathbf{Z}_T^*|\phi)\,p_C(\phi)\,d\phi \\ &= p_A(\mathbf{Y}_T|\mathbf{Z}_T^*)\int_\Phi p_B(\mathbf{Z}_T^*|\phi)\,p_C(\phi)\,d\phi = p_A(\mathbf{Y}_T|\mathbf{Z}_T^*)\,p(\mathbf{Z}_T^*), \end{aligned}$$

The posterior density is

$$\text{(2.8.6)}\qquad p(\phi|\mathbf{Y}_T) \propto \int_{\tilde{\mathbf{Z}}} p_A(\mathbf{Y}_T|\mathbf{Z}_T^*)\,p_B(\mathbf{Z}_T^*|\phi)\,p_C(\phi)\,d\mathbf{Z}_T^* \propto p(\mathbf{Y}_T|\phi)\,p_C(\phi).$$

Comparing (2.8.4) with(2.8.1), (2.8.5) with (2.8.2), and (2.8.6) with (2.8.3), it is apparent that the latent variable model is formally identical to a model with a two-stage hierarchical prior: the latent variables correspond to the intermediate level of the hierarchy. The formal identity continues to hold if the p.d.f. for $\mathbf{Y}_T$ is expressed $p_A(\mathbf{Y}_T|\theta,\phi)$ in the hierarchical prior and $p_A(\mathbf{Y}_T|\mathbf{Z}_T^*,\phi)$ in the latent variable model.

In the latent variable model, (2.8.6) reflects the uncertainty about $\mathbf{Z}_T^*$, which is a matrix of nuisance parameters if one is interested only in ϕ. The density $p(\mathbf{Z}_T^*|\mathbf{Y}_T)$ reflects the uncertainty in ϕ, which is a vector of nuisance parameters if one is interested only in $\mathbf{Z}_T^*$. These advantages of Bayesian methods for latent variable models in general are supplemented with their computational convenience, as will be seen in Section 4.

The duality between the hierarchical prior and latent variable models often suggests formulations that decompose more complex problems into simpler ones. For example,

$$y_t \sim t(0,\sigma^2;\nu)$$

is formally equivalent to the latent variable model

$$y_t = \omega_t \varepsilon_t,$$

with ω_t a latent variable, $\nu/\omega_t^2 \sim \chi^2(\nu)$, and $\varepsilon_t \sim N(0,1)$ independent of ω_t. The equivalent hierarchical prior formulation is the p.d.f. specification

$$y_t \big| (\omega_t, \sigma^2) \sim \mathrm{N}(0, \sigma^2 \omega_t)$$

and the conditional prior distribution

$$v/\omega_t^2 \sim \chi^2(v).$$

3. Simulation[1]

Bayesian methods are operational only to the extent that posterior moments (2.1.5) can actually be computed. There are three ways in which this can be done. If the posterior distribution and the function of interest are sufficiently simple, the posterior moment may be obtained analytically. Most results in this category in econometrics may be found in Zellner (1971); few further analytical results for posterior moments in econometrics have been obtained since that work was published. If the required integration takes place in fewer than (say) six dimensions then classical deterministic methods of numerical analysis, principally quadrature, are often practical. (A standard reference for these methods is Davis and Rabinowitz (1984).) In the remaining cases, which constitute the preponderance of applied econometrics, posterior simulators are the approach of choice.

Posterior simulators have a single characteristic principle: generate a sequence of vectors $\{\theta_m\}$ with the property that if $\mathrm{E}[\mathrm{g}(\theta)|\mathbf{Y}_T]$ exists then there is a weighting function $\mathrm{w}(\theta)$ such that

$$(3.0.1) \qquad \overline{g}_M = \sum_{m=1}^{M} \mathrm{g}(\theta_m)\mathrm{w}(\theta_m) \Big/ \sum_{m=1}^{M} \mathrm{w}(\theta_m) \to \mathrm{E}[\mathrm{g}(\theta)|\mathbf{Y}_T] = \overline{g}$$

(Here and throughout this chapter, "$\to$" denotes almost sure convergence.) Many simulators produce $\{\theta_m\}$ that -- at least asymptotically in M -- all have the posterior distribution, and in this case $\overline{g}_M = M^{-1}\sum_{m=1}^{M} \mathrm{g}(\theta_m)$.

Posterior simulators have several attractions. First and foremost, they are often straightforward to construct, even in quite elaborate models. This includes models sufficiently complex that non-Bayesian methods like maximum likelihood are impossible or impractical. Second, posterior simulators can take advantage of the structure of latent variable models as set forth in Section 2.8, simulating parameters and latent variables jointly. This often renders them operational even when the likelihood function cannot be evaluated. Third, posterior simulators are well suited to situations in which $\mathrm{g}(\theta)$ cannot be evaluated in closed form, but unbiased simulators are available, because $\mathrm{g}(\theta)$ may then be replaced by its simulator. Leading examples are forecasting and discrete choice

[1]This section draws heavily on Geweke (1995a).

models. Finally, posterior simulators are practical: they can be executed in reasonable time using desktop equipment, and their very construction often provides further insight into the statistical properties of the model.

All this comes at some cost. The proper use of posterior simulators requires analytical work on the part of the econometrician. First and foremost, the investigator must verify that the posterior distribution exists. A proper prior and a bounded likelihood function are sufficient for the existence of the posterior distribution, but if the prior is improper then the existence of the posterior must be demonstrated. Simulators can appear well-behaved over a finite number of iterations even though the product of the prior and the likelihood is not a probability density kernel in θ. Second, the investigator must verify analytically that the posterior moment of interest exists. In this section it is implicitly assumed that this has been done for the problem at hand; expectation operators used here all apply to moments that exist under the posterior. Third, the investigator must verify (3.0.1). This section provides conditions for the convergence in (3.0.1) for a variety of simulators.

3.1 Pseudorandom number generation

All pseudorandom number generators begin with a pseudorandom sequence $\{u_i\}$ in which the u_i are assumed to be independently and uniformly distributed on the unit interval (0, 1). In fact the sequence $\{u_i\}$ is deterministic: most software employs a multiplicative congruential generator which generates integers $J_i = (aJ_{i-1}) \bmod m$ and takes $u_i = J_i/m$. The constants a and m are chosen carefully so that $\{u_i\}$ has good properties: e.g., $a = 16807$ and $m = 2^{31} - 1$ are common choices. The design and testing of uniform pseudorandom number generators is an important part of numerical analysis with a substantial literature: see Geweke (1995a, Section 3.1) for an overview and citations, and suggestions regarding the use of multiplicative congruential generators. For the purposes at hand it is assumed that the sequence $\{u_i\}$ is a satisfactory approximation to an i.i.d. sequence with a uniform distribution on the unit interval. In what follows "u" will denote a realization from this distribution, and "$\{u_i\}$" a sequence of such i.i.d. realizations.

Given $\{u_i\}$, one can in principle generate random variables from any univariate distribution whose inverse cumulative distribution function (c.d.f.) can be evaluated. Suppose x is continuous, and consequently the inverse c.d.f. $\mathrm{F}^{-1}(p) = \{c\colon \mathrm{P}(x \le c) = p\}$ exists. Then x and $\mathrm{F}^{-1}(u)$ have the same distribution: $\mathrm{P}[\mathrm{F}^{-1}(u) \le d] = \mathrm{P}[u \le \mathrm{F}(d)] = \mathrm{F}(d)$. Hence pseudorandom drawings $\{x_i\}_{i=1}^N$ of x may be

constructed as $F^{-1}(u_i)$, where $\{u_i\}_{i=1}^{N}$ is a sequence of pseudorandom uniform numbers. A simple example is provided by the exponential distribution with probability density $f(x) = \lambda \exp(-\lambda x), x \geq 0$. Then $F(x) = 1 - \exp(-\lambda x), F^{-1}(p) = -\log(1-p)/\lambda$, and consequently, $x = -\log(u)/\lambda$. The inverse c.d.f. method is very easy to apply if an explicit, closed form expression for the inverse c.d.f. is available. Since most inverse c.d.f.'s require the evaluation of transcendental functions, the method may be inefficient relative to others.

Acceptance methods are widely used as a simpler and more efficient alternative to the inverse c.d.f. method. Suppose that x is continuous with p.d.f. $f(x)$ and support C. Let g be the p.d.f. of a different continuous random variable z with p.d.f. $g(z)$ which has a distribution from which it is possible to draw i.i.d. random variables and for which

$$\sup_{x \in C}[f(x)/g(x)] = a < \infty.$$

The function g is known as an *envelope* or *majorizing density* of f, and the distribution with p.d.f. g is known as the *source distribution*. To generate x_i,

(a) Generate u;
(b) Generate z;
(c) If $u > f(z)/[a\,g(z)]$, go to (a);
(d) $x_i = z$.

The unconditional probability of proceeding from step (c) to step (d) in any pass is

$$\int_{-\infty}^{\infty} \{f(z)/[a\,g(z)]\}\, g(z)dz = a^{-1},$$

and the unconditional probability of reaching step (d) with value at most c in any pass is

$$\int_{-\infty}^{c} \{f(z)/[a\,g(z)]\}\, g(z)dz = a^{-1} F(c).$$

Hence the probability that x_i is at most c at step (d) is $F(c)$.

A key advantage of acceptance methods is that they often can be tailored to idiosyncratic univariate distributions that arise in the posterior distributions for specific econometric models. This frequently happens in conjunction with the Gibbs sampler (Section 3.4.1); some examples are provided in Geweke and Keane (1995). In this use of acceptance sampling it is often useful to consider a family of source densities $g(x;\alpha)$ indexed by a parameter vector α. It is then usually easy to choose α to maximize the probability of acceptance from the source density (Geweke, 1995a, Section 3.2).

Composition methods decompose a random variable into two or more components, each of which is easy to generate. For example, $x \sim t(0, 1; 2)$ can be generated in the obvious way from three independent standard normals; if $x \sim B(m,n)$ then $x = z_1/(z_1 + z_2)$ with z_1 and z_2 independent, $z_1 \sim \chi^2(2m+2)$, $z_2 \sim \chi^2(2n+2)$ (Johnson and Kotz, 1972, Section 40.5).

The univariate normal distribution arises repeatedly in posterior distributions, usually as the distribution of a subset of parameters conditional on others. Both inverse c.d.f. and acceptance methods for generating univariate normal pseudo-random vectors are well developed. Good software libraries implement both. The gamma distribution with scale parameter λ and shape parameter a has p.d.f.

$$f(x) = \lambda \exp(-\lambda x)(\lambda x)^{a-1}/\Gamma(a), x \geq 0.$$

In general, random variables from this distribution may be generated efficiently using composition algorithms and acceptance methods. Fast and accurate methods are complicated but readily available in statistical software libraries.

Two multivariate distributions are especially important in posterior simulators. The generation of a multivariate normal random vector $\underset{m\times 1}{\mathbf{x}}$ from the distribution $N(\mu,\Sigma)$ is based on the familiar decomposition

$$\mathbf{z} \sim N(\mathbf{0},\mathbf{I}_m), \quad \mathbf{x} = \mu + \mathbf{A}\mathbf{z} \text{ with } \mathbf{A}\mathbf{A}' = \Sigma.$$

While any factorization $\mathbf{A}$ of Σ will suffice, it is most efficient to make $\mathbf{A}$ upper or lower triangular so that $m(m+1)/2$ rather than m^2 products are required in the transformation from $\mathbf{z}$ to $\mathbf{x}$. The Choleski decomposition, in which the diagonal elements of the upper or lower triangular $\mathbf{A}$ are positive, is typically used.

If $\underset{m\times 1}{\mathbf{x}_i} \overset{IID}{\sim} N(\mathbf{0},\Sigma)$, the distribution of $\mathbf{A} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ is Wishart, with p.d.f.

$$f(\mathbf{A}) = \frac{|\mathbf{A}|^{\frac{1}{2}(n-m)} \exp\left(-\frac{1}{2}\operatorname{tr}\Sigma^{-1}\mathbf{A}\right)}{2^{\frac{1}{2}(n-1)m}\pi^{m(m-1)/4}|\Sigma|^{\frac{1}{2}(n-1)}\prod_{i=1}^{m}\Gamma\left[\frac{1}{2}(n-i)\right]}; \tag{3.1.1}$$

for brevity, $\mathbf{A} \sim W(\Sigma, n-1)$. Direct construction of $\mathbf{A}$ through generation of $\{\mathbf{x}_i\}_{i=1}^{n}$ becomes impractical for large n. A more efficient indirect method follows Anderson (1984). Let Σ have lower triangular Choleski decomposition $\Sigma = \mathbf{L}\mathbf{L}'$, and suppose $\mathbf{Q} \sim W(\mathbf{I}_m, n-1)$. Then $\mathbf{L}\mathbf{Q}\mathbf{L}' \sim W(\Sigma, n-1)$ (Anderson, 1984, pp. 254-255). Furthermore $\mathbf{Q}$ has representation

$$\mathbf{Q} = \mathbf{U}\mathbf{U}' \qquad u_{ij} = 0 \ (i < j < m)$$

$$u_{ij} \sim N(0,1) \qquad u_{ii} \sim \chi^2(n-i)$$

$(i = 1,\ldots,m)$, with the u_{ij} mutually independent for $i \geq j$ (Anderson, 1984, p. 247). Even if n is small, this indirect construction is much more efficient than the direct construction.

3.2 Independence simulation

The simplest possible posterior simulator can be constructed if one can generate the i.i.d. sequence $\{\theta_m\}$ with common p.d.f. $p(\theta|\mathbf{Y}_T)$. Denoting $\bar{g} = E[g(\theta)|\mathbf{Y}_T]$ and $\bar{g}_m = M^{-1}\sum_{m=1}^{M} g(\theta_m)$, by the strong law of large numbers

(3.2.1) $\qquad \bar{g}_m \to \bar{g}$.

If the *posterior variance* of $g(\theta)$, $\sigma_g^2 = \text{var}[g(\theta)|\mathbf{Y}_T] = E\{[g(\theta) - \bar{g}]^2|\mathbf{Y}_T\} < \infty$, then by the Lindberg-Levy central limit theorem

(3.2.2) $\qquad M^{-1/2}(\bar{g}_M - \bar{g}) \Rightarrow N(0, \sigma_g^2)$.

(Here and in what follows "$\Rightarrow$" denotes convergence in distribution.)

The leading simple example of a posterior simulator based on independence sampling in econometrics is the normal linear model with conjugate prior distribution,

(3.2.3) $\qquad \underset{T\times 1}{\mathbf{y}} = \underset{T\times k}{\mathbf{X}}\,\underset{k\times 1}{\beta} + \underset{T\times 1}{\varepsilon}, \qquad \varepsilon|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_T)$,

(3.2.4) $\qquad \underline{\nu}\underline{s}^2/\sigma^2 \sim \chi^2(\underline{\nu}), \qquad \beta|\sigma^2 \sim N(\underline{\beta}, \sigma^2\underline{\mathbf{H}}_\beta^{-1})$.

[The matrix $\sigma^{-2}\underline{\mathbf{H}}_\beta$ is the *precision* of the conditional prior distribution for β -- i.e., the inverse of its variance matrix.] Straightforward manipulation shows

(3.2.5) $\qquad (\bar{\nu}\bar{s}^2/\sigma^2)|(\mathbf{y},\mathbf{X}) \sim \chi^2(\bar{\nu})$,

(3.2.6) $\qquad \beta|(\sigma^2,\mathbf{y},\mathbf{X}) \sim N(\bar{\beta}, \sigma^2\bar{\mathbf{H}}_\beta^{-1})$,

where $\qquad \bar{\nu} = \underline{\nu} + T - k$, $\bar{s}^2 = \bar{\nu}^{-1}\left[\underline{\nu}\underline{s}^2 + (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})\right]$, $\bar{\mathbf{H}}_\beta = \underline{\mathbf{H}}_\beta + (\mathbf{X}'\mathbf{X})^{-1}$,

$\bar{\beta} = \bar{\mathbf{H}}_\beta^{-1}\left[\underline{\mathbf{H}}_\beta\underline{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{b}\right]$ with $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. [For derivations see Zellner (1971, Section 3.2.3) or Poirier (1995, Theorem 9.9.1).] Since the marginal posterior distribution of β is multivariate Student-*t*, closed-form expressions for the moments of β exist. But many functions of interest are nonlinear if β. For example, if the explanatory variables include lagged dependent variables then conditional on the presample lagged dependent variables the posterior distribution is given by (3.2.5) and (3.2.6), but functions of interest like predictors of future values and spectral densities involve nonlinear transformation of β and σ^2.

The generation of pseudorandom vectors following (3.2.5) and (3.2.6) in fact involves acceptance sampling, as explained in Section 3.1, although this feature will be transparent to the user of a mathematical software library or a higher-level language. The acceptance sampling algorithm is quite general and can in principle be used to produce an independent sample from any posterior density $p(\theta|\mathbf{Y}_T)$. The essential requirement is that one be able to draw pseudorandom vectors from a distribution whose p.d.f. $r(\theta)$ is an envelope of $p(\theta|\mathbf{Y}_T)$. One then proceeds as in Section 3.1. The advantages of the

procedure are that it requires only specification of the kernels of the two p.d.f.'s, and that it produces i.i.d. pseudorandom vectors from the posterior distribution. The disadvantages are that it is often difficult to find an envelope and determine $\sup_{\theta\in\Theta}\left[p(\theta|\mathbf{Y}_T)/r(\theta)\right]$, and that acceptance probabilities may be so low as to render the whole algorithm impractical. The potential for these difficulties generally increases with the dimension of θ (although the strucutre of the posterior density is also important). When acceptance sampling succeeds, however, (3.2.1) always applies, and (3.2.2) applies if the posterior variance exists.

A simulator closely related to acceptance sampling is importance sampling. Let $j(\theta)$ be a probability density kernel corresponding to a distribution from which an i.i.d. sequence $\{\theta_m\}$ can be drawn conveniently, and whose support includes Θ. Define the corresponding weight function $w(\theta) = p(\theta|\mathbf{Y}_T)/j(\theta)$. (In this expression, $p(\theta|\mathbf{Y}_T)$ need only be the kernel of the posterior density.) Then

(3.2.7) $$\bar{g}_M = \sum_{m=1}^{M} g(\theta_m) w(\theta_m) \Big/ \sum_{m=1}^{M} w(\theta_m) \to \bar{g}$$

If both

(3.2.8) $$E[w(\theta)] = \int_\Theta \left[p(\theta|\mathbf{Y}_T)^2 / j(\theta)\right] d\theta$$

and

$$E\left[g(\theta)^2 w(\theta)|\mathbf{Y}_T\right] = \int_\Theta \left[g(\theta)^2 p(\theta|\mathbf{Y}_T)^2 / j(\theta)\right] d\theta$$

are absolutely convergent, then

(3.2.9) $$M^{1/2}(\bar{g}_M - \bar{g}) \Rightarrow N(0, \sigma^2)$$

and

$$s_M^2 = M \sum_{m=1}^{M} \left[g(\theta_m) - \bar{g}\right]^2 w(\theta_m) \Big/ \left[\sum_{m=1}^{M} w(\theta_m)\right]^2 \to \sigma^2$$

where

$$\sigma^2 = E\left\{\left[g(\theta) - \bar{g}\right]^2 w(\theta)\right\}.$$

(For proofs see Geweke (1989b).)

In importance sampling the simulated θ_m are independent but the sample must be weighted to produce a simulation-consistent approximation of the posterior moment $\bar{g}$ from an "incorrectly drawn" sample. The intuition underlying (3.2.7) is that if θ_m is drawn from an area that is undersampled, relative to the posterior distribution, then that drawing must receive a large weight to compensate, and conversely. Neither (3.2.7) nor (3.2.9) requires that $w(\theta)$ be bounded, but as a practical matter if $w(\theta)$ is bounded and $\mathrm{var}\left[g(\theta)|\mathbf{Y}_T\right] < \infty$ then (3.2.8) is satisfied, and without this condition establishing (3.2.8) is usually tedious. Experience suggests that when $w(\theta)$ is unbounded convergence in (3.2.7) is so slow as to make the method impractical.

In many circumstances one therefore can choose between acceptance and importance sampling. The choice depends on the computational demands of the problem. If evaluation of $g(\theta)$ is trivial relative to the generation of θ_m and computation of $w(\theta)$ then importance sampling is preferred; conversely, acceptance sampling is the method of choice. Geweke (1995a, Section 4.4) provides elaborations on the comparison, as well as a mixture of acceptance and importance sampling that can be optimized for each problem.

3.3 Variance reduction

In many instances it is possible to modify independence sampling to produce a sequence of drawings each of which is identically distributed as in the original algorithm, but with dependence between draws that substantially lowers the sampling variance of the mean, thereby increasing the accuracy of $\overline{g}_m$ as an approximation of $\overline{g}$.

Antithetic acceleration (Geweke, 1988) is based on a technique originally due to Hammersly and Morton (1956). The essential properties are most easily conveyed in the case where the sequence $\{\theta_m\}$ can be drawn directly from the posterior distribution. In this method the sample drawn can be described $\{\theta_{mi}\}_{i=1 m=1}^{2 \quad M/2}$ with the θ_{mi} identically distributed and the only mutual dependence being that arising between θ_{m1} and θ_{m2}. Let $\overline{g}_M = M^{-1}\sum_{m=1}^{M/2}\sum_{i=1}^{2} g(\theta_{mi})$ and suppose $\mathrm{var}[g(\theta)|\mathbf{Y}_T] < \infty$. Then

$$M^{-1/2}(\overline{g}_M - \overline{g}) \Rightarrow N(0, \sigma^{*2}), \quad \sigma^{*2} = \mathrm{var}[g(\theta_{mi})] + \mathrm{cov}[g(\theta_{m1}), g(\theta_{m2})].$$

As long as $\mathrm{cov}[g(\theta_{m1}), g(\theta_{m2})] < 0$, antithetic acceleration with $M/2$ replications will have smaller variance of approximation error than importance sampling with M replications, and the computational requirements will be about the same.

To focus further on the properties of antithetic acceleration, consider the situation in which $p(\theta|\mathbf{Y}_T)$ is symmetric about the point μ. In this case $\theta_{m1} = \mu + \varepsilon_m$, $\theta_{m2} = \mu - \varepsilon_m$ describes a pair of variables drawn from the posterior distribution, with correlation matrix $-\mathbf{I}$. If $g(\theta)$ were a linear function, then $\mathrm{var}\{\frac{1}{2}[g(\theta_{m1}) + g(\theta_{m2})]\} = 0$, and variance reduction would be complete. At the other extreme, if $g(\theta)$ is also symmetric about μ, then $\mathrm{var}\{\frac{1}{2}[g(\theta_{m1}) + g(\theta_{m2})]\} = \mathrm{var}[g(\theta)]$: antithetic simple Monte Carlo integration will require double the number of computations of simple Monte Carlo for the same information. As an intermediate case, suppose that $d(y) = g(\theta y)$ is either monotone nondecreasing or monotone nonincreasing for all θ. Then $g(\theta_{m1}) - \overline{g}$ and $g(\theta_{m2}) - \overline{g}$ must be of opposite sign if they are nonzero. This implies $\mathrm{cov}[g(\theta_{m1}), g(\theta_{m2})] < 0$, whence $\sigma^{*2} \leq \mathrm{var}[g(\theta)] = \sigma^2$, and so antithetic acceleration produces gains in efficiency.

As T increases, the posterior distribution generally becomes increasingly symmetric and concentrated about the true value of the vector of unknown parameters, reflecting the operation of a central limit theorem. (For an overview and citations, see Bernardo and Smith (1994, Section 5.3).) In these circumstances $g(\theta)$ is increasingly well described by a linear approximation of itself over most of the support the posterior distribution as T increases. Let σ_T^2 indicate the accuracy of simple Monte Carlo and σ_T^{*2} the accuracy of antithetic Monte Carlo. Given some weak side conditions, it may be shown that $\sigma_T^{*2}/\sigma_T^2 \to 0$, and under somewhat stronger conditions that $T\sigma_T^{*2}/\sigma_T^2$ converges to a constant (Geweke, 1988).

To introduce another method of variance reduction, suppose there is an approximation to the original problem that can be solved exactly with reasonable effort: i.e., one can determine $\bar{h} = \tilde{E}[h(v)|\mathbf{Y}_T] = \int_N h(v)\tilde{p}(v|\mathbf{Y}_T)dv$ exactly. Suppose that the sequence $\{\theta_m, v_m\}$ can be drawn, $\{\theta_m\}$ an i.i.d. sequence from the original posterior distribution and $\{v_m\}$ an i.i.d. sequence from the approximating distribution, but with θ_m and v_m constructed from the same underlying random numbers so that $g(\theta_m)$ and $h(v_m)$ are correlated. Let $\bar{g}_M = M^{-1}\sum_{m=1}^{M} g(\theta_m)$ and $\bar{h}_M = M^{-1}\sum_{m=1}^{M} h(v_m)$, and consider approximations of the form

$$\bar{g}'_M = \bar{g}_M + \beta(\bar{h}_M - \bar{h}).$$

Clearly $E(g'_M) = \bar{g}$. One can easily verify that $\mathrm{var}(\bar{g}'_M)$ is minimized by

$$\beta = -\mathrm{cov}[g(\theta_m), h(v_m)]/\mathrm{var}[h(\theta_m)]$$

and that in this case

$$\mathrm{var}(\bar{g}'_M) = \mathrm{var}(\bar{g}_M)\{1 - \mathrm{corr}^2[g(\theta_m), h(\theta_m)]\}.$$

The parameter β may be estimated in the obvious way from the replications. This is an example of the use of *control variates*, introduced by Kahn and Marshall (1953) and Hammersly and Handscomb (1964).

Yet a third method of variance reduction is the use of conditional expectations If $\theta' = (\theta'_{(1)}, \theta'_{(2)})$ and $g(\theta) = g(\theta_{(1)})$, it may be the case that $E[g(\theta_{(1)})|\theta_{(2)}, \mathbf{Y}_T]$ can be evaluated analytically. If so, then by the Rao-Blackwell Theorem the variance of approximation error can be reduced by using the function of interest $E[g(\theta_{(1)})|\theta_{(2)m}, \mathbf{Y}_T]$ rather than $g(\theta_{(1)m})$ in any posterior simulator. Extensions of this idea are developed in Casella and Robert (1994).

3.4 Markov chain Monte Carlo

This section takes up a recently developed class of posterior simulators that have collectively become known as *Markov chain Monte Carlo*. The idea is to construct a Markov chain with state space Θ and invariant distribution with p.d.f. $p(\theta|\mathbf{Y}_T)$. Following an initial transient or *burn-in* phase, simulated values from the chain form a basis for approximating $E[g(\theta)|\mathbf{Y}_T]$. What is required is to construct an appropriate algorithm and verify that its invariant distribution is unique, with p.d.f. $p(\theta|\mathbf{Y}_T)$.

Markov chain methods have a history in mathematical physics dating back to the algorithm of Metropolis *et al.* (1953). This method, which is described in Hammersly and Handscomb (1964, Section 9.3) and Ripley (1987, Section 4.7), was generalized by Hastings (1970), who focused on statistical problems, and was further explored by Peskun (1973). A version particularly suited to image reconstruction and problems in spatial statistics was introduced by Geman and Geman (1984). This was subsequently shown to have great potential for Bayesian computation by Gelfand and Smith (1990). Their work, combined with data augmentation methods (Tanner and Wong, 1987), has proven very successful in the treatment of latent variables and other unobservables in econometric models. Since about 1990 application of Markov chain Monte Carlo methods has grown rapidly; new refinements, extensions, and applications appear almost continuously.

3.4.1 The Gibbs sampler

The Gibbs sampler begins with a partition, or *blocking*, of $\underset{k\times 1}{\theta}$, $\theta' = (\theta'^{(1)},\ldots,\theta'^{(B)})$. For $b = 1,\ldots,B$, $\theta'^{(b)} = (\theta_1^{(b)},\ldots,\theta_{k(b)}^{(b)})$ where $k(b) \geq 1$; $\sum_{b=1}^{B} k(b) = k$; and the $\theta_i^{(b)}$ are the components of θ. Let $p(\theta^b|\theta^{(-b)},\mathbf{Y}_T)$ denote the conditional p.d.f.'s induced by $p(\theta|\mathbf{Y}_T)$, where $\theta^{(-b)} = \{\theta^{(a)}, a \neq b\}$.

Suppose a single drawing θ_0, $\theta_0' = (\theta_0'^{(1)},\ldots,\theta_0'^{(B)})$, from the posterior distribution is available. Consider successive drawings from the conditional distribution as follows:

$$
\begin{aligned}
&\theta_1^{(1)} \sim p(\theta^{(1)}|\theta_0^{(-1)},\mathbf{Y}_T)\\
&\theta_1^{(2)} \sim p(\theta^{(2)}|\theta_1^{(1)},\theta_0^{(3)},\ldots,\theta_0^{(B)},\mathbf{Y}_T)\\
&\quad\vdots\\
&\theta_1^{(j)} \sim p(\theta^{(j)}|\theta_1^{(1)},\ldots,\theta_1^{(j-1)},\theta_0^{(j+1)},\ldots,\theta_0^{(B)},\mathbf{Y}_T)\\
&\quad\vdots\\
&\theta_1^{(B)} \sim p(\theta^{(B)}|\theta_1^{(-B)},\mathbf{Y}_T).
\end{aligned}
\tag{3.4.1}
$$

This defines a transition process from θ_0 to $\theta_1' = \left(\theta_1'^{(1)},\ldots,\theta_1'^{(B)}\right)$. The Gibbs sampler is defined by the choice of blocking and the forms of the conditional densities induced by $p(\theta|\mathbf{Y}_T)$ and the blocking. Since $\theta_0 \sim p(\theta|\mathbf{Y}_T)$, $\left(\theta_1^{(1)},\ldots,\theta_1^{(j-1)},\theta_1^{(j)},\theta_0^{(j+1)},\ldots,\theta_0^{(B)}\right)$ $\sim p(\theta|\mathbf{Y}_T)$ at each step in (3.4.1) by definition of the conditional density. In particular, $\theta_1 \sim p(\theta|\mathbf{Y}_T)$.

Iteration of the algorithm produces a sequence $\theta_1, \theta_2,\ldots,\theta_m,\ldots$ which is a realization of a Markov chain with probability density function kernel for the transition from point θ_j to point θ_{j+1} given by

(3.4.2) $$K_G(\theta_j,\theta_{j+1}) = \prod_{b=1}^{B} p\left[\theta_{j+1}^{(b)} \mid \theta_j^{(a)}(a>b), \theta_{j+1}^{(a)}(a<b), \mathbf{Y}_T\right].$$

Any single iterate θ_j retains the property that it is drawn from the distribution with p.d.f. $p(\theta|\mathbf{Y}_T)$.

For the Gibbs sampler to be practical, it is essential that the blocking be chosen in such a way that one can make the drawings (3.4.1) in an efficient manner. For many problems in economics, the blocking is natural and the conditional distributions are familiar; Section 4 provides several examples. In making the drawings (3.4.1) all the methods of this section are at one's disposal.

The informal argument just given assumes that it is possible to make an initial draw from the posterior distribution. That is generally not possible; otherwise, one could use independence sampling. Even if it were, the argument potentially establishes only that given a collection of independent initial draws from the posterior distribution, one can generate a collection of independent final draws by iterating (3.4.1) on each initial draw. What is needed for application is a demonstration that one can consistently approximate a posterior moment with successive realizations of a single chain that begins with arbitrary $\theta_0 \in \Theta$. The stylized examples in Figures 1 and 2 show that this need not be the case.

Conditions for this sort of convergence are based on the mathematics of continuous state space Markov chains. Brief overviews for econometricians are presented in Chib and Greenberg (1994) and Geweke (1995a); from there the reader may turn to Tierney (1991), and to Tierney (1994) for a rigorous treatment based on Numelin (1994). There are two sets of convergence conditions emerging from this literature that are most directly useful in Bayesian econometric models. If either set holds, then $\bar{g}_M = M^{-1}\sum_{m=1}^{M} g(\theta_m)$ $\rightarrow E[g(\theta)|\mathbf{Y}_T]$.

Gibbs sampler convergence condition 1 (after Tierney, 1994). For every point $\theta^* \in \Theta$ and every $\Theta_1 \subseteq \Theta$ with the property $P(\theta \in \Theta_1|\mathbf{Y}_T) > 0$, it is the case that

$P_G\left(\theta_{j+1} \in \Theta_1 \middle| \theta_j = \theta^*, \mathbf{Y}_T\right) > 0$, where $P_G(\cdot)$ is the probability measure induced by the transition kernel (3.4.2).

Gibbs sampler convergence condition 2 (after Roberts and Smith, 1994). The density $p\left(\theta|\mathbf{Y}_T\right)$ is lower semicontinuous at 0, $\int_{\Theta^{(b)}} p\left(\theta|\mathbf{Y}_T\right)d\theta^{(b)}$ is locally bounded $(b = 1,\ldots,B)$, and Θ is connected. [A function $h(\mathbf{x})$ is lower semicontinuous at 0 if, for all $\mathbf{x}$ with $h(\mathbf{x}) > 0$, there exists an open neighborhood $N_{\mathbf{x}} \supset \mathbf{x}$ and $\varepsilon > 0$ such that for all $\mathbf{y} \in N_{\mathbf{x}}$, $h(\mathbf{y}) \geq \varepsilon > 0$. This condition rules out situations like the one shown in Figure 2.]

These conditions are by no means necessary for convergence of the Gibbs sampler; Tierney (1994) provides substantially weaker conditions. However, the conditions stated here are satisfied for a very wide range of posterior distributions in econometrics and are much easier to verify than the weaker conditions. Furthermore, the appropriate blocking is usually inherent in the structure of the posterior density, as will be seen in several examples in Section 4.

3.4.2 The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm begins with an arbitrary transition probability density function $q\left(\theta_m, \theta^*\right)$ and a starting value θ_0. The random vector θ^* generated from $q\left(\theta_m, \theta^*\right)$ is considered as a candidate value for θ_{m+1}. The algorithm actually sets $\theta_{m+1} = \theta^*$ with probability

$$\alpha\left(\theta_m, \theta^*\right) = \min\left\{\frac{p\left(\theta^*|\mathbf{Y}_T\right)q\left(\theta^*, \theta_m\right)}{p\left(\theta_m|\mathbf{Y}_T\right)q\left(\theta_m, \theta^*\right)}, 1\right\};$$

otherwise, the algorithm sets $\theta_{m+1} = \theta_m$. This defines a Markov chain with a generally mixed continuous-discrete transition probability from θ_m to θ_{m+1} given by

$$K_{MH}\left(\theta_m, \theta_{m+1}\right) = \begin{cases} q\left(\theta_m, \theta_{m+1}\right)\alpha\left(\theta_m, \theta_{m+1}\right) \text{ if } \theta_{m+1} \neq \theta_m \\ 1 - \int_D q\left(\theta_m, \theta\right)\alpha\left(\theta_m, \theta\right)d\theta \text{ if } \theta_{m+1} = \theta_m \end{cases}.$$

This form of the algorithm is due to Hastings (1970). The Metropolis *et al.* (1953) form takes $q\left(\theta_j, \theta^*\right) = q\left(\theta^*, \theta_j\right)$. A simple variant that is often useful is the independence chain (Tierney, 1991, 1994), $q\left(\theta_j, \theta^*\right) = j\left(\theta^*\right)$. Then

$$\alpha\left(\theta_j, \theta^*\right) = \min\left\{\frac{p\left(\theta^*|\mathbf{Y}_T\right)j\left(\theta_j\right)}{p\left(\theta_j|\mathbf{Y}_T\right)j\left(\theta^*\right)}, 1\right\} = \min\left\{\frac{w\left(\theta^*\right)}{w\left(\theta_j\right)}, 1\right\},$$

where $w(\theta) = p\left(\theta|\mathbf{Y}_T\right)/j(\theta)$. The independence chain is closely related to acceptance sampling and importance sampling. But rather than place a low (high) probability of acceptance or a low (high) weight on a draw that is too likely (unlikely) relative to

$p(\theta|\mathbf{Y}_T)$, the independence chain assigns a high (low) probability of accepting the candidate for the next draw.

There is a simple two-step argument that motivates the convergence of the sequence $\{\theta_m\}$ generated by the Metropolis-Hastings algorithm to the posterior distribution. (This approach is due to Chib and Greenberg, 1994.) First, observe that if any transition probability function $p(\theta_m, \theta_{m+1})$ satisfies the reversibility condition

$$p(\theta_m|\mathbf{Y}_T)\,p(\theta_m, \theta_{m+1}) = p(\theta_{m+1}|\mathbf{Y}_T)\,p(\theta_{m+1}, \theta_m),$$

then it has the posterior as its invariant distribution. To see this, note that

$$\begin{aligned}\int p(\theta|\mathbf{Y}_T)\,p(\theta, \theta_{m+1})\,d\theta &= \int p(\theta_{m+1}|\mathbf{Y}_T)\,p(\theta_{m+1}, \theta)\,d\theta \\ &= p(\theta_{m+1}|\mathbf{Y}_T)\int p(\theta_{m+1}, \theta)\,d\theta = p(\theta_{m+1}|\mathbf{Y}_T).\end{aligned}$$

The second step is to consider the implications of the requirement that $K_{MH}(\theta_m, \theta_{m+1})$ be reversible: $p(\theta_m|\mathbf{Y}_T)K_{MH}(\theta_m, \theta_{m+1}) = p(\theta_{m+1}|\mathbf{Y}_T)K_{MH}(\theta_{m+1}, \theta_m)$. For $\theta_{m+1} \neq \theta_m$ it implies that

$$p(\theta_m|\mathbf{Y}_T)\,q(\theta_m, \theta^*)\,\alpha(\theta_m, \theta^*) = p(\theta^*|\mathbf{Y}_T)\,q(\theta^*, \theta_m)\,\alpha(\theta^*, \theta_m).$$

Suppose (without loss of generality) that $p(\theta_m|\mathbf{Y}_T)\,q(\theta_m, \theta^*) \geq p(\theta^*|\mathbf{Y}_T)\,q(\theta^*, \theta_m)$. If we take $\alpha(\theta^*, \theta_m) = 1$ and $\alpha(\theta_m, \theta^*) = p(\theta^*|\mathbf{Y}_T)\,q(\theta^*, \theta_m)/p(\theta_m|\mathbf{Y}_T)\,q(\theta_m, \theta^*)$, this equality is satisfied.

In implementing the Metropolis-Hastings algorithm, the transition probability density function must share two important properties. First, it must be possible to generate θ^* efficiently from $q(\theta_m, \theta^*)$. All the methods of this and the previous section are potential tools for these drawings. (Once again, acceptance sampling is attractive relative to importance sampling.) A second key characteristic of a satisfactory transition process is that the unconditional acceptance rate not be so low that the time required to generate a sufficient number of distinct θ_m is too great.

The convergence properties of the Metropolis-Hastings algorithm are inherited from those of $q(\theta_m, \theta^*)$ (Roberts and Smith, 1994). In particular the following condition guarantees $M^{-1}\sum_{m=1}^{M} g(\theta_m) \to E[g(\theta)|\mathbf{Y}_T]$:

Metropolis-Hastings algorithm convergence condition 1 (after Tierney, 1994). For every point $\theta^* \in \Theta$ and every $\Theta_1 \subseteq \Theta$ with the property $P(\theta \in \Theta_1|\mathbf{Y}_T) > 0$, it is the case that $P_q(\theta_{m+1} \in \Theta_1|\theta_m = \theta^*, \mathbf{Y}_T) > 0$, where $P_q(\cdot)$ is the probability measure induced by the transition kernel $q(\theta_m, \theta^*)$.

Metropolis-Hastings algorithm convergence condition 2 (after Chib and Greenberg (1994) and Mengersen and Tweedie (1993)). For every $\theta \in \Theta$, $p(\theta|Y_T) > 0$, and for all pairs $(\theta_j, \theta_{j+1}) \in \Theta \times \Theta$, $p(\theta_j|Y_T)$ and $q(\theta_j, \theta_{j+1})$ are positive and continuous.

Once again, the conditions are sufficient but not necessary, but weaker conditions are typically much more difficult to verify. On weaker conditions, see Tierney (1994).

3.4.3 *Caveats*

In any practical application one is concerned with numerical accuracy. Markov chain Monte Carlo methods present two characteristic potential difficulties in assessing numerical accuracy: slow convergence, and the formal inapplicability of central limit theorems.

A leading cause of slow convergence is multimodality of the posterior distribution, for example, as shown in Figure 3 for a Gibbs sampler. In the limit multimodality approaches disconnectedness of the support, and increasingly large values of M are required for a good approximation. This difficulty is essentially undetectable given a single Markov chain: for a chain of any fixed length, one can imagine multimodal distributions for which the probability of leaving the neighborhood of a single mode is arbitrarily small. This sort of convergence problem is precisely the same as the multimodality problem in optimization, where iterations from a finite collection of starting values cannot guarantee the determination of a global optimum. Multimodal disturbances are difficult to manage by any method, including independence sampling. In the context of the Markov chain Monte Carlo algorithms, the question may be recast as one of sensitivity to initial conditions: θ_A^0, θ_B^0, and θ_C^0 will lead to quite different chains, in Figure 3, unless the simulations are sufficiently long.

A Markov chain Monte Carlo algorithm can be made more robust against sensitivity to initial conditions by constructing many very long chains. Just how one should trade off the number of chains against their length for a given budget of computation time is problem specific and as a practical matter not yet full understood. Many of the issues involved are discussed by Gelman and Rubin (1992), Geyer (1992), and their discussants and cited works. In an extreme variant of the multiple chains approach, the chain is restarted many times, with initial values chosen independently and identically distributed from an appropriate distribution. But finding an appropriate distribution may be difficult: one that is too concentrated reintroduces the difficulties exemplified by Figure 3; one that is too diffuse may require excessively long chains for convergence. These problems aside, proper use of the output of Markov chain Monte Carlo in a situation of multimodality requires specialized diagnostics; Zellner and Min (1995) have obtained

some interesting results of this kind. At the other extreme a single starting value is used. This approach provides the largest number of iterations toward convergence, but diagnostics of the type of problem illustrated in Figure 3 will not be as clear.

If one assumes standard mixing conditions for the serially correlated process $g(\theta_m)$ (e.g. Hannan, 1970, 207-210) then well-established central limit theorems apply to the distribution of $\bar{g}_m$. The resulting assessment of numerical accuracy (Geweke, 1992) has proven reliable in econometric models in the sense that it provides good forecasts of the output of repeated simulations. This approach is fundamentally unsatisfactory, however, because it assumes properties that should be derived from the known structure of the algorithm, and/or are strictly not true. For example, if the posterior variance exists, then in a stationary Metropolis-Hastings algorithm a standard central limit result applies (Geyer, 1992; Kipnis and Varadhan, 1986). But since a Metropolis-Hastings algorithm begins with an arbitrary initial condition it is not stationary. In addition, there is no central limit theorem applicable to Markov chain Monte Carlo in which it has been shown that the variance parameter can be estimated consistently in M, to the author's knowledge. Given the success of both Markov chain Monte Carlo algorithms in econometrics and statistics and the apparent reliability of assumed central limit theorems, these questions are clearly prime candidates for future research.

4. Some models

Recent innovations in posterior simulators have made possible routine and practical applications of Bayesian methods in statistics. This section reviews the implementation of posterior simulators in some common econometric models. The survey is selective. It concentrates on generic or "textbook" models to introduce approaches that can be applied in many specific settings. In so doing the interrelatedness of specific approaches is emphasized. All of the methods presented here can be combined, used in more elaborate models, and be tailored to more specific models implied by the theory and data in a given application.

In keeping the number of topics manageable the examples exclude time series models, largely because of another survey in preparation (Geweke, 1995b) on that topic. The reader will note that most of the posterior simulators presented rely principally on the Gibbs sampler. In part that reflects the exclusion of time series models, where the Metropolis-Hastings algorithm is more important. But it also reflects the fact that more elaborate econometric models are typically constructed through the use of conditional

distributions which can usually be undone by the Gibbs sampler to exploit the simpler conditionals. That is especially so in models involving latent variables.

4.1 Normal linear regression

The normal linear model with conjugate prior distribution was discussed in the context of independence simulation (Section 3.2). This prior distribution links dispersion in prior beliefs about β and σ^2, since $\beta \sim t\left(\underline{\beta}, \underline{s}^2 \underline{\mathbf{H}}_{\beta}^{-1}; \underline{\nu}\right)$. Suppose instead that prior beliefs about β are represented by $\beta \sim N\left(\underline{\beta}, \underline{\mathbf{H}}_{\beta}^{-1}\right)$ independent of σ^2. Then the prior density kernel is

$$(4.1.1) \qquad \left(\sigma^2\right)^{-(\underline{\nu}+2)/2} \exp\left(-\underline{\nu}\underline{s}^2/2\sigma^2\right) \exp\left[-\tfrac{1}{2}\left(\beta-\underline{\beta}\right)' \underline{\mathbf{H}}_{\beta}\left(\beta-\underline{\beta}\right)\right]$$

and the likelihood function may be expressed either

$$(4.1.2) \qquad \left(\sigma^2\right)^{-T/2} \exp\left[-(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)/2\sigma^2\right]$$

or

$$(4.1.3) \qquad \exp\left(-\nu s^2/2\sigma^2\right) \exp\left[-(\beta-\mathbf{b})'\mathbf{X}'\mathbf{X}(\beta-\mathbf{b})/2\sigma^2\right],$$

with $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, $\nu = T-k$, $s^2 = \nu^{-1}(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})$. Forming the posterior density kernel as the product of (4.1.1) and (4.1.2) it is immediate that

$$(4.1.4) \qquad \left\{\left[\underline{\nu}\underline{s}^2 + (\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)\right]/2\sigma^2\right\} \mid (\beta, \mathbf{y}, \mathbf{X}) \sim \chi^2(T+\underline{\nu}).$$

Forming the posterior density kernel as the product of (4.1.1) and (4.1.3) and completing the square,

$$(4.1.5) \qquad \beta \mid \left(\sigma^2, \mathbf{y}, \mathbf{X}\right) \sim N\left[\left(\underline{\mathbf{H}}_{\beta} + \sigma^{-2}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\underline{\mathbf{H}}_{\beta}\underline{\beta} + \sigma^{-2}\mathbf{X}'\mathbf{X}\mathbf{b}\right), \left(\underline{\mathbf{H}}_{\beta} + \sigma^{-2}\mathbf{X}'\mathbf{X}\right)^{-1}\right].$$

In this model the prior for β is conjugate conditional on σ^2, and the prior for σ^2 is conjugate conditional on β. Hence the conditional posterior distribution of each is of the same family as its conditional prior distribution. Moreover, (4.1.4) and (4.1.5) indicate the obvious construction of a Gibbs sampler to draw from the posterior distribution. It is trivial to verify either Gibbs sampler convergence condition (Section 3.4.1) in this model.

Several general principles are at work in this simple but important model.

(1) The decomposition of the posterior distribution into mutually conditional distributions can provide a convenient description of a nonstandard distribution.

(2) Conditionally conjugate prior distributions are convenient representations of belief when their blocking coincides with that of the Gibbs sampler.

(3) Further blocking of the Gibbs sampler is clearly possible (by means of conditional normal posterior distributions for subvectors of β) but is counterproductive because it generally increases serial correlation in the Gibbs sampler thereby slowing the rate of convergence (Geweke, 1992, Section 3.6).

To provide more flexibility in the representation of prior beliefs, suppose in lieu of $\beta \sim \mathrm{N}\left(\underline{\beta}, \underline{\mathbf{H}}_{\beta}^{-1}\right)$ that $\beta \sim \mathrm{t}\left(\underline{\beta}, \underline{\mathbf{H}}_{\beta}^{-1}; \lambda\right)$. If the corresponding density kernel is substituted appropriately in (4.1.1) the resulting expressions are formidable. Instead, write the prior distribution in hierarchical form

(4.1.6) $$\lambda/w \sim \chi^2(\lambda), \quad \beta|w \sim \mathrm{N}\left(\underline{\beta}, w\underline{\mathbf{H}}_{\beta}^{-1}\right).$$

Conditional on w the model has not changed: (4.1.4) remains true as does (4.1.5) once $\underline{\mathbf{H}}_{\beta}$ is replaced with $w^{-1}\underline{\mathbf{H}}_{\beta}$. It remains only to find the conditional posterior density kernel for w,

(4.1.7) $$w^{-(\lambda+2)/2}\exp(-\lambda/2w)\left|w^{-1}\underline{\mathbf{H}}_{\beta}\right|^{1/2}\exp\left[-\left(\beta-\underline{\beta}\right)'\underline{\mathbf{H}}_{\beta}\left(\beta-\underline{\beta}\right)\Big/2w\right]$$

which implies

$$\left[\lambda+\left(\beta-\underline{\beta}\right)'\underline{\mathbf{H}}_{\beta}\left(\beta-\underline{\beta}\right)\right]\Big/w \mid \left(\sigma^2,\beta,\mathbf{y},\mathbf{X}\right) \sim \chi^2(\lambda+k).$$

This modest extension of the model illustrates three further principles that hold more generally in using posterior simulators in econometrics.

(4) Decomposing a prior distribution into a hierarchy of simple distributions can simplify the model and help in constructing the posterior simulator.

(5) Correspondingly, conditional distributions are natural building blocks for more elaborate models and simulator design.

(6) While conditional posterior distributions may be obvious, there is no substitute for deliberately writing the entire posterior kernel in detail, and then establishing kernels for conditionals.

The last point was honored in (4.1.1)-(4.1.3) but violated in the extension (4.1.6): it would have been easy to carelessly neglect the term $\left|w^{-1}\underline{\mathbf{H}}_{\beta}\right|$ in (4.1.7) because the corresponding term vanished in the kernel of the simpler model. While point (6) is essential to research practice editorial constraints generally demand its violation in published scientific papers.

4.2 Normal linear regression with constraints

It is often the case that coefficients in the linear model are assumed to satisfy constraints that are not well represented by multivariate normal or Student-t distributions, or there is a substantial prior probability that these constraints may be satisfied. Examples include the restriction of β to a subset of R^k, and the event that some coefficients take on specified values, in particular 0.

There is a long history of formal treatment of restrictions of this kind in linear models in econometrics including Judge and Takayama (1966), Lovell and Prescott (1970), Gourieroux, Holly and Monfort (1982), and Wolak (1987). Analytical Bayesian treatments include Chamberlain and Leamer (1976), Leamer and Chamberlain (1976), and Davis (1978). Non-Bayesian approaches are technically awkward and lead to estimators with unappealing properties because of their *ex ante* conditioning (Poirier, 1995, Section 9.8). Analytical Bayesian approaches produce useful results in one dimension but fail in higher dimensions.

The earliest treatment using a posterior simulator is Geweke (1986). That work considered the model (3.2.3) and (3.2.4) with $\underline{\nu} \to 0$, $\underline{\mathbf{H}}_\beta \to \mathbf{0}$, combined with the restriction $\beta \in Q \subseteq \mathrm{R}^k$. A posterior independence simulator in this case generates a candidate using (3.2.5)-(3.2.6) and accepts it if and only if $\beta \in Q$. The advantage of the procedure is its simplicity and ability to handle constraints expressed implicitly as well as explicitly. Its disadvantage is that the rate of acceptance may be so low as to render it impractical. When k is large (exceeding 8, say) computational efficiency can be quite poor even for restrictions that are reasonable when compared to the likelihood function. Nevertheless, the method works well for many problems and no better method has been developed that applies to a general restriction set.

When the restrictions are linear inequalities, substantial improvements in efficiency are possible. Beginning with the model (3.2.3) and (4.1.1) suppose the constraints

(4.2.1) $$\mathbf{a} \le \underset{k\times k}{\mathbf{D}}\beta \le \mathbf{w}$$

are added, where the elements of $\mathbf{a}$ and $\mathbf{w}$ are extended real numbers. Since this constraint has no effect if $a_i = -\infty$ and $w_i = +\infty$ fewer than k linear inequality restrictions may actually be involved. In particular, this model includes as specific cases sign restrictions on coefficients. Rewrite the model

$$\mathbf{y} = \mathbf{Z}\alpha + \varepsilon, \quad \varepsilon \sim \mathrm{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_T\right),$$

$$\underline{\nu s}^2/\sigma^2 \sim \chi^2(\underline{\nu}), \quad \alpha \sim \mathrm{N}\left(\underline{\alpha}, \underline{\mathbf{H}}_\alpha^{-1}\right),$$

where $\mathbf{Z} = \mathbf{X}\mathbf{D}^{-1}$, $\alpha = \mathbf{D}\beta$, $\underline{\alpha} = \mathbf{D}\underline{\beta}$, $\underline{\mathbf{H}}_\alpha = \mathbf{D}'^{-1}\underline{\mathbf{H}}_\beta\mathbf{D}^{-1}$. The Gibbs sampler may be applied to this model just as it was in (4.1.4) and (4.1.5) to (3.2.3) and (4.1.1), except that the conditional distribution of α is truncated normal:

$$\alpha\left|\left(\sigma^2, \mathbf{y}, \mathbf{Z}\right) \sim \mathrm{N}\left[\left(\underline{\mathbf{H}}_\alpha + \sigma^{-2}\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\underline{\mathbf{H}}_\alpha\underline{\alpha} + \sigma^{-2}\mathbf{Z}'\mathbf{Z}\mathbf{a}\right),\ \left(\underline{\mathbf{H}}_\alpha + \sigma^{-2}\mathbf{X}'\mathbf{X}\right)^{-1}\right],\quad \mathbf{a} \le \alpha \le \mathbf{w},\right.$$

with $\mathbf{a} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$. The algorithm of Geweke (1991) for the truncated normal then applies directly to **a**. Decompose α into k blocks of one parameter each, and draw each element successively conditional on the other. The conditional distributions involved are all univariate truncated normal, so this procedure is straightforward. Since no draw is ever rejected, the procedure does not suffer from the potential inefficiency of the more general acceptance algorithm.

In a variant on this method (Geweke, 1995c), $\beta_i = 0$ with prior probability $\underline{p}_i$; conditional on $\beta_i \neq 0$ the prior distribution of β_i is $\mathrm{N}\left(\underline{\beta}_i, \tau_i^2\right)$, possibly truncated to the interval (λ_i, υ_i). These priors are independent across the k coefficients. This model characterizes the ubiquitous variable selection problem in regression. With a Bayesian treatment, problems of regression strategies and pretest estimators do not arise: the interpretation of the posterior distribution is unambiguous.

The posterior simulator for this model again has complete blocking, but now the conditional posterior distribution of each coefficient is mixed: the coefficient is either zero or is drawn from a possibly truncated normal distribution. The respective probabilities are proportional to the conditional marginalized likelihoods for each event. This amounts to evaluating the posterior distribution at $\beta_i = 0$ in the one case, and integrating it over the permitted range of the coefficient in the other. The Gibbs sampler satisfies condition 1 for convergence, and the posterior probability of any configuration of regressors being in the model is the posterior expectation of the corresponding indicator function.

A closely related procedure is stochastic search variable selection (SVSS), introduced by George and McCulloch (1993, 1994). (A related work is Clyde and Parmigiani (1994).) The prior distribution in this model is

$$\beta_i = \gamma_i\delta_{i1} + (1-\gamma_i)\delta_{i2},$$

$$\delta_{ij} \sim \mathrm{N}\left(0, \tau_{ij}^2\right)\quad \left(j = 1,2;\ \tau_{i2}^2 >> \tau_{i1}^2\right),$$

$$\mathrm{P}(\gamma_i = 0) = p_i,\quad \mathrm{P}(\gamma_i = 1) = 1 - p_i.$$

Here a regressor is selected if $\gamma_i = 1$ and not selected if $\gamma_i = 0$, but "not selected" means that the corresponding coefficient is small in absolute value, not 0. Posterior moments are obtained using a Gibbs sampler. Conditional on $(\gamma_1, \ldots, \gamma_k)$ the relevant conditional

distributions are of the form (3.2.5) and (3.2.6), and the conditional posterior distribution of each γ_i is Bernoulli. Condition 1 for convergence of the Gibbs sampler again applies. But note that as $\tau_{i1}^2 \to 0$ the probability that γ_i will change from 0 to 1 in successive iterations also goes to 0.

Markov chain Monte Carlo posterior simulators for variable selection in regression exhibit increasing serial correlation as the degree of multicollinearity increases, and as the number of regressors grows more iterations are generally required to represent all models with nonnegligible posterior probability. George and McCulloch (1994) have addressed this problem in two ways. First, by using a conjugate prior conditional on the included variables, closed form expressions for the probabilities of different configurations can be obtained, which obviates the problem of serial correlation in the presence of multicollinearity. Second, a Metropolis algorithm that generates candidates far from the current selection of regressors appears promising in finding models with substantially different regressors than the current model in the Markov chain.

Raftery, Madigan and Hoeting (1993) take up the variable selection problem using a posterior simulator, but their approach is distinctly different. The "priors" employed are data dependent. The computational algorithm uses the Occam's window algorithm of Madigan and Raftery (1994) and therefore does not provide a simulation-consistent approximation of the posterior probability of all combinations of regressors.

These methods for normal linear models have much wider applicability than the normal linear model itself. The reason is principle (5) stipulated above: linear model posteriors appear as conditionals in many other models. In particular, the conditional posterior distribution of β as it appears in these models arises repeatedly in Bayesian econometrics.

4.3 Seemingly unrelated regressions

The seemingly unrelated regressions (SUR) model of Zellner (1962) has been extensively applied in economics, especially in neoclassical models of production and consumption. It appears repeatedly conditional on other parameters or latent variables in other models as well, including the multinomial probit model (Section 4.6), linear instrumental variables models (Section 4.7), factor analysis models, and vector autoregressions. The simplest form of this model is

$$\underset{T\times 1}{\mathbf{y}_j} = \underset{T\times k_j}{\mathbf{X}_j}\ \underset{k_j\times 1}{\beta_j} + \underset{T\times 1}{\varepsilon_j} \quad (j = 1,\ldots,m). \tag{4.3.1}$$

Let $\varepsilon' = (\varepsilon_1',\ldots\varepsilon_m')$ and $\mathbf{X} = [\mathbf{X}_1,\ldots,\mathbf{X}_m]$; then

$$\varepsilon|\mathbf{X} \sim \mathrm{N}(\mathbf{0}, \Sigma \otimes \mathbf{I}_T). \tag{4.3.2}$$

Defining $\mathbf{y}' = (\mathbf{y}_1', \ldots, \mathbf{y}_m')$ the model may be expressed

(4.3.3) $\qquad \mathbf{y} = \mathbf{Z}\beta + \varepsilon, \quad \varepsilon \sim \mathrm{N}(\mathbf{0}, \Sigma \otimes \mathbf{I}_T).$

If

$$\beta' = (\beta_1', \ldots, \beta_m'), \qquad \mathbf{Z} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{X}_m \end{bmatrix}$$

then (4.3.3) is equivalent to (4.3.1) and (4.3.2). With the appropriate reorganization of $\mathbf{Z}$, it also includes cases in which there are exact cross-equation restrictions on the coefficient vector, a situation that arises commonly. The independent priors

(4.3.4) $\qquad \beta \sim \mathrm{N}(\underline{\beta}, \underline{\mathbf{H}}_\beta^{-1}), \quad \Sigma^{-1} \sim \mathrm{W}(\underline{\mathbf{S}}^{-1}, \underline{\nu})$

are conditionally conjugate. (The Wishart distribution is briefly presented at the end of Section 3.1.)

The complete SUR model is (4.3.3) and (4.3.4). The kernel of $\mathrm{p}(\beta, \Sigma)$ from (4.3.4) is

(4.3.5) $\qquad |\Sigma|^{-(\nu+m+1)/2} \exp\left[-\tfrac{1}{2}\mathrm{tr}\,\Sigma^{-1}\underline{\mathbf{S}}\right] \exp\left[-\tfrac{1}{2}(\beta-\underline{\beta})' \underline{\mathbf{H}}_\beta^{-1}(\beta-\underline{\beta})\right].$

Defining $s_{ij}(\beta) = (\mathbf{y}_i - \mathbf{X}_i\beta_i)'(\mathbf{y}_j - \mathbf{X}_j\beta_j)$ and $\mathbf{S}(\beta) = [s_{ij}(\beta)]$ then following Zellner (1971, Section 8.5) the likelihood function may be expressed either

(4.3.6) $\qquad |\Sigma|^{-T/2} \exp\left[-\tfrac{1}{2}\mathrm{tr}\,\Sigma^{-1}\mathbf{S}(\beta)\right]$

or

(4.3.7) $\qquad |\Sigma|^{-T/2} \exp\left\{\left[\beta - \hat{\beta}(\Sigma)\right]' \mathbf{Z}'(\Sigma^{-1} \otimes \mathbf{I}_T)\mathbf{Z}\left[\beta - \hat{\beta}(\Sigma)\right]\right\}$

with $\hat{\beta}(\Sigma) = \left[\mathbf{Z}'(\Sigma^{-1} \otimes \mathbf{I}_T)\mathbf{Z}\right]^{-1}\mathbf{Z}'(\Sigma^{-1} \otimes \mathbf{I}_T)\mathbf{y}$. Forming the posterior density kernel as the product of (4.3.5) and (4.3.6) and comparing the result with (3.1.1),

(4.3.8) $\qquad \Sigma^{-1}|(\beta, \mathbf{y}, \mathbf{X}) \sim \mathrm{W}\left(\left[\underline{\mathbf{S}} + \mathbf{S}(\beta)\right], \; T + \underline{\nu}\right).$

Forming the posterior density kernel as the product of (4.3.5) and (4.3.7) and completing the square,

(4.3.9) $\qquad \beta|(\Sigma, \mathbf{y}, \mathbf{X}) \sim \mathrm{N}\left[\overline{\beta}(\Sigma), \overline{\mathbf{H}}_\beta(\Sigma)^{-1}\right]$

with $\overline{\mathbf{H}}_\beta(\Sigma) = \underline{\mathbf{H}}_\beta + \mathbf{Z}'(\Sigma^{-1} \otimes \mathbf{I}_T)\mathbf{Z}$, $\overline{\beta}(\Sigma) = \left[\overline{\mathbf{H}}_\beta(\Sigma)\right]^{-1}\left[\underline{\mathbf{H}}_\beta\underline{\beta} + \mathbf{Z}'(\Sigma^{-1} \otimes \mathbf{I}_T)\mathbf{Z}\hat{\beta}(\Sigma)\right]$. A Gibbs sampling algorithm is defined by (4.3.8) and (4.3.9); both conditions 1 and 2 for convergence are satisfied trivially. Early publications of this algorithm are Blattberg and George (1991) and Percy (1992).

The prior density (4.3.4) may be an insufficiently flexible representation of prior beliefs about the vector β. For example, in many applications the vectors β_j of (4.3.1)

are thought to be similar but not identical (Stein, 1966; Ghosh, Saleh and Sen, 1989). In such situations a hierarchical prior often provides a good representation of beliefs that is also convenient because the Gibbs sampling algorithm can be extended to the implied complete model. Chib and Greenberg (1995) consider the hierarchy

$$\beta = \mathbf{A}_0\beta_0 + \mathbf{u}, \quad \mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{B}_0)$$
$$\beta_0 = \mathbf{A}_1\mu + \eta, \quad \eta \sim \mathrm{N}(\mathbf{0}, \mathbf{B}_1)$$
$$\mu \sim \mathrm{N}(\mu_0, \mathbf{M}_0)$$

and derive the required conditional posterior (normal) distributions for β, β_0, and μ. Modest variants of this model accommodate time-varying parameters (Chib and Greenberg, 1995; Min and Zellner, 1993; Gammerman and Migon, 1993). An alternative to hierarchical models is the recursive extended natural conjugate prior distribution of Richard and Steel (1988) in which some of the integrations can be performed analytically.

Since the conditional distribution of β is multivariate normal, the methods for coping with inequality constraints described in Section 4.2 apply in the SUR model as well. Such constraints can be a fundamental part of the economics underlying the model. Perhaps the leading example is quasiconcavity constraints in neoclassical microeconomics, the imposition of which has spawned a substantial econometrics literature (e.g., Barnett and Lee (1985), Diewert and Wales (1987), and references therein). Chalfant and Wallace (1993) and Terrell (1995) illustrate that simple acceptance sampling, in conjunction with a posterior simulator, can be a simple practical method for imposing these constraints.

4.4 Nonnormality

Analytical approaches to Bayesian inference in econometrics rest heavily on normality assumptions: e.g. Zellner (1971) uses this distribution exclusively for continuously distributed disturbances. (The same is very nearly true of analytical non-Bayesian methods for which there exists a finite sample theory.) The combination of hierarchical models and posterior simulators has removed this constraint, so greatly expanding the scope for practical work that ideas about what the effective limitations of this approach might be are not yet well formed.

There are at least three compelling reasons why distributional assumptions are important in Bayesian econometrics.

(1) Distributional assumptions are central to Bayesian inference and its claim of exact finite sample results. Flexible representations of beliefs about the shapes of distributions are essential to reliable results.

(2) Many of the posterior moments of interest that motivate applied econometrics are sensitive to distributional assumptions. This is perhaps most evident in prediction of future events and the consequences of policy changes. (For a compelling example see Geweke and Keane (1995).)

(3) Utility functions typically assumed in general equilibrium models make equilibrium outcomes sensitive to the assumed distribution of shocks. Flexible assumptions about these distributions are therefore necessary in empirical work if the implications of these models for prices, welfare and dynamics are to be evaluated.

The leading Bayesian approach to nonnormality is the application of *normal mixture models*. The most general such model in the univariate case may be written

$$\varepsilon\left|(\mu,\sigma^2)\right. \sim \mathrm{N}(\mu,\sigma^2),$$

$$\mu \sim d\mathrm{P}_\mu(\mu;\theta_\mu), \tag{4.4.1}$$

$$\sigma^2 \sim d\mathrm{P}_{\sigma^2}(\sigma^2;\theta_{\sigma^2}). \tag{4.4.2}$$

Then

$$\mathrm{p}\left(\varepsilon\left|\theta_\mu,\theta_{\sigma^2}\right.\right) = (2\pi)^{-1/2}\int_{-\infty}^{\infty}\int_0^{\infty}\sigma^{-1}\exp\left[-(\varepsilon-\mu)^2/2\sigma^2\right]d\mathrm{P}_{\sigma^2}(\sigma^2;\theta_{\sigma^2})\,d\mathrm{P}_\mu(\mu;\theta_\mu)$$

and the model is completed with prior distributions for the parameter vectors θ_μ and θ_{σ^2}.

In practice the mixture may be either discrete or continuous. Continuous mixture models often mix only with respect to the scale parameter σ: if (4.4.1) degenerates to $\mu = \theta_\mu$ but (4.4.2) is nondegenerate, the model is said to be a *scale mixture of normals.* Scale mixture normal models have a substantial history in the Bayesian hierarchical modeling literature, e.g. West (1984). In constructing such models one has available a rich set of results on the genesis of continuous distributions, which often involve scale mixtures of normals. (A classic and quite useful reference is Johnson and Kotz (1970, 1972).) For example, it was seen at the end of Section 2.8 that for the specific case of (4.4.2),

$$\theta_{\sigma^2} = (\nu,\sigma^{*2}), \quad \nu\sigma^{*2}/\sigma^2 \sim \chi^2(\nu)$$

it follows that

$$\varepsilon \sim \mathrm{t}(\mu,\sigma^{*2};\nu).$$

A convenient flexible completion of the model is provided by the independent prior distributions

$$\underline{\nu s}^2/\sigma^{*2} \sim \chi^2(\underline{\nu}), \quad \nu \sim \exp(\lambda).$$

The implementation of this model is discussed fully in Geweke (1993) which finds $\nu < 10$ in autoregressive representations of a variety of U.S. macroeconomic time series. Gamma mixing distributions yield a variety of other distributions for ε including the Erlang and LaPlace (Tsionas, 1994, Section 3.2).

Another class of nonnormal distributions with increasing application in Bayesian econometrics is the additive mixture model. The general formulation is

$$\eta_t = \zeta_t + \varepsilon_t$$

where ζ_t and ε_t are mutually and serially independent, $\varepsilon_t \sim \mathrm{N}(0, \sigma^2)$, and ζ_t has p.d.f. $\mathrm{p}(\zeta_t|\theta_\zeta)$. Assume that η_t is observed. (In fact η_t may be the disturbance in a regression equation, or some similar unobservable, but in the context of a posterior simulator that applies the Gibbs sampler this assumption is typically innocuous.) Treating ζ_t as a latent variable, or equivalently a parameter in the first stage of a two stage hierarchy,

$$\mathrm{p}[\zeta_t|(\eta_t, \sigma^2)] \propto \exp[-(\eta_t - \zeta_t)^2/2\sigma^2]\mathrm{p}(\zeta_t|\theta_\zeta),$$
$$\mathrm{p}(\theta_\zeta|\zeta_1, \ldots, \zeta_T) \propto \prod_{t=1}^{T} \mathrm{p}(\zeta_t|\theta_\zeta).$$

This procedure has found considerable application in stochastic frontier models in which distributions have constituents with sign constraints (van den Broek, Koop, Osiewalski and Steel, 1994; Koop, Steel and Osiewalski, 1995). For an application to heterogeneity in panel data, see Geweke and Keane (1995).

The development of nonnormal multivariate distributions along the same lines is clearly straightforward and should be practical, but there are as yet no published applications to the author's knowledge. A leading case is the multivariate Student-*t* distribution in the context of the seemingly unrelated regressions model.

In implementing new nonnormal distributions using normal or additive mixtures it is especially important to verify the existence of posterior distributions and moments of interest, and to be cognizant of the role of prior distributions. If the likelihood function is bounded and prior distributions in all stages of the hierarchy are proper, the posterior distribution exists. If not, there may be no posterior distribution despite the existence of well defined conditionals for each block of a Gibbs sampler. (In this case, no invariant distribution exists.) Verification of the existence of posterior expectations of unbounded functions of interest is typically more difficult and must proceed on a case-by-case basis. In any event, the econometrician will often find it enlightening to compare prior and posterior moments to assess, informally, the informativeness of the data.

4.5 Censored regression

The standard Tobit censored regression model (Tobin, 1958) is

(4.5.1) $$y_t^* = \mathbf{x}_t'\beta + \varepsilon_t, \quad \varepsilon_t \overset{IID}{\sim} \mathrm{N}(0, \sigma^2),$$

(4.5.2) $$y_t = \max(y_t^*, 0).$$

The observed data are $\{\mathbf{x}_t, y_t\}_{t=1}^T$. The model may be completed with the independent prior distributions

(4.5.3) $$\beta \sim \mathrm{N}(\underline{\beta}, \underline{\mathbf{H}}_\beta^{-1}), \quad \underline{\nu s}^2/\sigma^2 \sim \chi^2(\nu).$$

This is one of the simplest latent variable models in econometrics, and is also a simple example of a limited dependent variable model (Maddala, 1983). Censored regression models are widely applied; Amemiya (1984) provides a survey.

The only econometric novelty in (4.5.1)-(4.5.3) is that y_t^* is unobserved. If it were observed, the model would revert to normal linear regression (Section 4.1). On the other hand, conditional on β, σ^2, and the data the distribution of the y_t^* is simple. These latent variables are conditionally independent,

$$y_t^* = y_t \text{ if } y_t > 0,$$

$$y_t^* \sim \mathrm{N}(\mathbf{x}_t'\beta, \sigma^2) \text{ s.t. } y_t^* \le 0 \text{ if } y_t = 0.$$

Bayesian inference may proceed in the Tobit censored regression model by means of a posterior simulator employing the Gibbs sampler. Condition 1 for convergence applies, although condition 2 does not. The earliest published implementations of this algorithm appear to be Chib (1992) and Geweke (1992), but see Wei and Tanner (1991) for a similar approach.

The step of drawing $\{y_t^*\}_{t=1}^T$ is known as *data augmentation* after Tanner and Wong (1987). In a Bayesian approach there is no formal reason to distinguish between latent variables and parameters (Section 2.8). The single step of drawing from the conditional distribution of the latent variables has been used in non-Bayesian approaches as well, and hence the distinction. (For a closely related procedure, see Rubin (1987).)

Generalizations of the Tobit censored regression model in which Bayesian inference is practical are available immediately from the developments in preceding sections, including a multivariate Student-*t* rather than a normal prior distribution for β, linear inequality restrictions on elements of β, selection of regressors, and generalizations of the assumed normal distribution for ε_t through mixture and additive models.

4.6 Probit models

Situations in which individuals make a single choice from among a number of alternatives are very common. The study of this behavior is the subject of a vast literature in economics, psychology, political science, marketing, and other disciplines. One of the earliest and still most important statistical models is the multinomial probit model developed by Thurstone (1927). Until quite recently the application of this model was impractical for more than three choices because of technical problems to be described shortly.

An alternative, the multinomial logit model, presents no such problems and in particular maximum likelihood is simple even when there are scores of choices. However, this model in its common form is inconsistent with choice theory. (For a classic and thorough discussion of this and related issues, see Manski and McFadden (1981).) Bayesian methods, some based on importance sampling, have been developed for these models: see Zellner and Rossi (1984) and Koop and Poirier (1993, 1994).

To describe the probit model, suppose

(4.6.1) $$u_{jt} = \mathbf{x}'_{jt}\beta + \varepsilon_{jt} \quad (j = 1,\ldots,J)$$

where $\mathbf{x}_{jt}$ is a vector of individual characteristics for individual t and choice attributes for choice j, u_{jt} is individual t's utility from choice j, and $\varepsilon_t = (\varepsilon_{1t},\ldots,\varepsilon_{Jt})' \sim \mathrm{N}(\mathbf{0}, \Sigma)$. The econometrician does not observe u_{jt} of course, but only the choice corresponding to the highest utility. Correspondingly denote $d_{jt} = 1$ if individual t makes choice j and $d_{jt} = 0$ otherwise. Since the model has implications only for observed choices, any translation and positive scaling of (4.6.1) will be produce an equivalent model. It has become conventional to normalize (4.6.1) with $u_{Jt} = 0$, $\sigma_{11} = \mathrm{var}(\varepsilon_{1t}) = 1$. In the presence of suitable variation in individual characteristics and choice attributes (Heckman and Sedlacek, 1985), (4.6.1) is identified. These conditions are usually satisfied in practice.

If individuals are observed over several time periods in a panel data set it is typically unreasonable to assume that shocks are serially uncorrelated. Then the unit of observation remains the individual, but if there are Q choices in each of S periods, there are $J = Q^S$ possible choices that may be observed. In this way the number of choices can become quite large. (A model for panel data has other changes in structure as well; see Geweke, Keane and Runkle (1994b).)

In the case of dichotomous choice the normalization leaves a single equation which may be expressed

$$u_t = \mathbf{x}_t'\beta + \varepsilon_t, \quad \varepsilon_t \overset{IID}{\sim} \mathrm{N}(0,1),$$
$$u_t \geq 0 \text{ if } d_{1t} = 1, \quad u_t < 0 \text{ if } d_{2t} = 1.$$

If the model is completed with a prior distribution of the form $\beta \sim \mathrm{N}\left(\underline{\beta}, \underline{\mathbf{H}}_\beta^{-1}\right)$ it is immediately clear that a posterior simulator based on the Gibbs sampler may be employed. Conditional on $\{u_t\}_{t=1}^T$ the model reverts to the normal linear regression model. Conditional on β and the data

$$u_t \sim \mathrm{N}(\mathbf{x}_t'\beta, 1) \text{ s.t. } \begin{cases} u_t \geq 0 \text{ if } d_{1t} = 1, \\ u_t < 0 \text{ if } d_{2t} = 1. \end{cases}$$

This algorithm was first published by Albert and Chib (1993). Clearly the various extensions of linear models discussed in Sections 4.1, 4.3 and 4.4 apply immediately to this model. Of these, the most interesting are nonnormal distributions.

When there are more than two choices the model is more complicated. Writing the normalized model by individual, with J choices,

$$\underset{(J-1)\times 1}{\tilde{\mathbf{u}}_t} = \underset{(J-1)\times k}{\tilde{\mathbf{X}}_t} \underset{k\times 1}{\beta} + \underset{(J-1)\times 1}{\tilde{\varepsilon}_t}, \quad \tilde{\varepsilon}_t \sim \mathrm{N}(\mathbf{0}, \Sigma^*) \quad (t = 1,\ldots,T)$$

If the model is organized by equation instead of by observation,

$$\underset{T\times 1}{\mathbf{u}_j} = \underset{T\times k}{\mathbf{X}_j} \underset{k\times 1}{\beta} + \underset{T\times 1}{\varepsilon_j} \quad (j = 1,\ldots,J-1).$$

Defining $\mathbf{u}' = (\mathbf{u}_1',\ldots\mathbf{u}_{J-1}')$ and $\varepsilon' = (\varepsilon_1',\ldots,\varepsilon_{J-1}')$ the model may be expressed

$$\mathbf{u} = \mathbf{Z}\beta + \varepsilon, \quad \varepsilon \sim \mathrm{N}(\mathbf{0}, \Sigma \otimes \mathbf{I}_T).$$

This is precisely the seemingly unrelated regressions model (4.3.3). The essential econometric differences are (1) the unobservability of $\mathbf{u}$ and (2) the restriction $\sigma_{11}^* = 1$.

The latency of the utilities in the multinomial probit model prevented its practical application for many years. The difficulty is that all approaches, including non-Bayesian procedures like maximum likelihood and method of moments, involve expressions for the probability of choice conditional on the data and parameters. This entails integration over the appropriate orthant of a $(J-1)$-dimensional multivariate normal distribution, and there are T such integrations for each evaluation of the likelihood function or moment conditions. For up to five choices, an approximation due to Clark (1961) was generally used. Only with the advent of simulation methods for assessing these probabilities (beginning with McFadden (1989) and Pakes and Pollard (1989)) did the multinomial probit model for more than five choices become feasible.

In a posterior simulator based on Gibbs sampling the step of assessing the probabilities is avoided all together. Conditional on all other parameters and the data the $\tilde{\mathbf{u}}_t$ are independent,

(4.6.2) $$\tilde{\mathbf{u}}_t \sim \mathrm{N}(\mathbf{x}_t'\beta, \Sigma^*), \text{ s.t. } u_{jt} \geq u_{it} \text{ if } d_{jt} = 1,$$

where the restriction is taken over $j = 1,\ldots,J$ and $u_{Jt} = 0$. Since there are exactly $J-1$ linear inequality constraints on the $(J-1)$-dimensional multivariate normal random vector $\tilde{\mathbf{u}}_t$, the algorithm of Geweke (1991) described in Section 4.2 can be applied, drawing the individual elements of $\tilde{\mathbf{u}}_t$ in succession, each from the appropriate univariate normal distribution truncated appropriately as indicated by (4.6.2).

The restriction $\sigma_{11}^* = 1$ is not accommodated if the model is completed with a Wishart prior distribution for Σ^{*-1} in the same way that the seemingly unrelated regressions model was completed. The earliest implementations of posterior simulators for the multinomial probit model (McCulloch and Rossi, 1995; Geweke, Keane and Runkle, 1994a) use the Wishart prior ignoring the restriction $\sigma_{11}^* = 1$ and report $\beta\sigma^{-1/2}$ in lieu of β. More recently McCulloch, Polson and Rossi (1995) present a convenient algorithm for drawing from the conditional posterior distribution for Σ^* when the prior distribution is $\Sigma^* \sim \mathrm{W}(\underline{S}^{-1}, \underline{\nu})$ subject to $\sigma_{11}^* = 1$.

While the posterior simulator does not require the evaluation of choice probabilities, functions of interest often do. In this step the Bayesian econometrician has available all the methods developed in conjunction with non-Bayesian procedures that require probability evaluations for estimation. A thorough review and comparison of methods is Hajivassiliou, McFadden and Ruud (1995) which concludes that the Geweke-Hajivassiliou-Keane (GHK) simulator performs best. Geweke, Keane and Runkle (1995) provide a self-contained description, algorithm and code for this procedure; see also Hajivassiliou and McFadden (1990) and Keane (1990).

Besides the general argument in favor of Bayesian procedures, in the multinomial probit model there is evidence that posterior means of parameters computed using the posterior simulator described here have better sampling properties than do non-Bayesian estimates using the best available technology for approximating choice probabilities. Geweke, Keane and Runkle (1994a, 1994b) compare the performance of various estimators in Monte Carlo studies with from 7 to 30 choices and 5,000 to 30,000 observations. With very few exceptions, the posterior means have smaller root mean square errors than do maximum likelihood or simulated method of moments estimators.

4.7 Linear simultaneous equation models

The linear simultaneous equation model has a long and rich history in theoretical econometrics, approached from both Bayesian and non-Bayesian viewpoints. The canonical model

$$\underset{1\times L}{\mathbf{y}_t'}\,\underset{L\times L}{\Gamma} + \underset{1\times k}{\mathbf{x}_t'}\,\underset{k\times L}{\mathrm{B}} = \underset{1\times L}{\varepsilon_t'}, \quad \varepsilon_t \sim \mathrm{N}(\mathbf{0}, \Sigma)$$

consists of L equations with L endogenous variables ($\mathbf{y}_t$) and k predetermined variables ($\mathbf{x}_t$). The system is normalized by $\gamma_{jj} = -1 \quad (j = 1,\ldots,L)$ and with this normalization one may write

$$\mathbf{y}_t = \underset{L\times M}{\mathbf{A}}\,\mathbf{z}_t + \boldsymbol{\varepsilon}_t$$

in which $M = K + L$, $\mathbf{z}_t' = (\mathbf{y}_t', \mathbf{x}_t')$, $a_{jj} = 0 \ (j = 1,\ldots,L)$, and for all $i = 1,\ldots,L$, $a_{ij} = \gamma_{ij}$ $(j \neq i, j = 1,\ldots,L)$ and $a_{i,L+j} = \beta_{ij} \ (j = 1,\ldots,k)$. The corresponding reduced form is

$$\mathbf{y}_t' = \mathbf{x}_t'\Pi + \mathbf{v}_t', \quad \Pi = -\mathbf{B}\Gamma^{-1}, \quad \mathbf{v}_t \sim \mathrm{N}\left(\mathbf{0}, \Gamma'^{-1}\Sigma\Gamma^{-1}\right).$$

The likelihood function is

$$\left|\Gamma'^{-1}\Sigma\Gamma^{-1}\right|^{-T/2} \exp\left[-\tfrac{1}{2}\sum_{t=1}^{T}\left(\mathbf{y}_t + \Gamma'^{-1}\mathbf{B}'\mathbf{x}_t\right)' \Gamma\Sigma^{-1}\Gamma'\left(\mathbf{y}_t + \Gamma'^{-1}\mathbf{B}'\mathbf{x}_t\right)\right]$$

$$= |\Gamma|^{T}|\Sigma|^{-T/2} \exp\left[-\tfrac{1}{2}\sum_{t=1}^{T}\left(\Gamma'\mathbf{y}_t + \mathbf{B}'\mathbf{x}_t\right)' \Sigma^{-1}\left(\Gamma'\mathbf{y}_t + \mathbf{B}'\mathbf{x}_t\right)\right] \tag{4.7.1}$$

$$= |\Gamma|^{T}|\Sigma|^{-T/2} \exp\left[-\tfrac{1}{2}\sum_{t=1}^{T}\left(\mathbf{y}_t - \mathbf{A}\mathbf{z}_t\right)' \Sigma^{-1}\left(\mathbf{y}_t - \mathbf{A}\mathbf{z}_t\right)\right]. \tag{4.7.2}$$

Even before taking up the completion of the model with the prior distribution for the parameters in Γ, B and Σ, the essential technical difficulty with the posterior density is evident in the presence of the term $|\Gamma|^{T}$ in (4.7.2). The problems in using the posterior density that derive from any one of a variety of priors have been extensively studied. Richard (1973) and Rothenberg (1975) took up the general question, Dreze (1976) studied the analogous limited information problem, and Kloek and van Dijk (1978) approached the problem using importance sampling. A thorough survey of this work is Dreze and Richard (1983). In approaches using independence sampling poly-t densities have proven useful (Dreze, 1977; Bauwens, 1984; Bauwens and Richard, 1985). Approaches using improper priors have proven especially troublesome, because of the ill-conditioned likelihood function (4.7.1); Chao and Phillips (1994) and Kleibergen and van Dijk (1994) are recent examples. No method emerging from this research has been used widely in applied work.

From (4.7.2) observe that if Γ is lower (or upper) triangular then $|\Gamma| = 1$ and the likelihood function is precisely that of the seemingly unrelated regressions model. If in addition Σ is diagonal then the likelihood function factors equation-by-equation and if prior distributions are also independent across equations then the methods of Section 4.1 apply. This corresponds to the well-known fact that maximum likelihood is equivalent to least squares in a recursive simultaneous equation system. But the essential simplification requires *only* that $|\Gamma| = 1$. The significance of this point for applied work was first noted, to the author's knowledge, in Zellner, Min and Dallaire (1994).

The leading instance of this case in applied econometrics is the incomplete simultaneous equation model that specifies only the first equation, leaving the rest of the system in reduced form:

$$\Gamma = \begin{bmatrix} 1 & \mathbf{0}' \\ \gamma & \mathbf{I}_{L-1} \end{bmatrix}, \quad \mathrm{B} = \begin{bmatrix} \beta & \underset{k\times(L-1)}{\Pi_{(1)}} \end{bmatrix};$$

the variance matrix Σ is unrestricted. This is perhaps more commonly known as "the instrumental variables model" after the popular method of estimation. If the model is completed with a multivariate normal prior for β and $\Pi_{(1)}$ and (independently) a Wishart prior for Σ^{-1} then the posterior simulator for the seemingly unrelated regressions model described in Section 4.3 can be used with no change at all. The commonly employed improper prior $p\left(\beta, \Pi_{(1)}, \Sigma\right) \propto |\Sigma|^{-(L+1)/2}$ corresponds to $\underline{\nu} = 0$ in (4.3.8) and $\underline{\mathbf{H}}_{\beta} = \mathbf{0}$ in (4.3.9).

While the difficulties surrounding the presence of $|\Gamma|^{T}$ in (4.7.1) have received most of the attention in the theoretical literature, most applications involve the incomplete (or instrumental variables) model in which $|\Gamma| = 1$. For these applications, the posterior simulator based on the Gibbs sampler provides a practical basis for Bayesian inference.

5. Model comparison and communication

For a subjective Bayesian decision maker the computation of the posterior moment (2.5.1) for a suitable model, prior and function of interest is the final objective of inference. For an investigator reporting results for other potential decision makers, however, the situation is quite different. In the language of Hildreth (1963) these decision makers are *remote clients*, who ideally have agreed to disagree in terms of the prior (Poirier, 1988). Clients may also have different uses for the model. In general, therefore, the investigator will not know either the priors or the functions of interest of her clients.

What should the investigator report? Traditionally, published papers report a few posterior moments, and more rarely some indication of sensitivity to prior distributions and alternative data densities may be given. Such information is generally much too limited. At the other extreme, the investigator may simply report some likelihood functions, but this leaves most of the work to the client. Investigators almost never report marginalized likelihoods, thereby leaving unrealized the promise inherent in model averaging.

This section first takes up the model comparison question, which has been intensively studied in the past five years in the wake of the rapid innovations in posterior simulators. It surveys some of these developments and argues that one method in particular is most promising for the generic comparison of models to which posterior simulators apply. It then turns to the more general question of Bayesian communication. Here it appears that posterior simulators, coupled with current storage, communication and computation capabilities (to say nothing of future developments in these areas) offer the potential to revolutionize applied econometrics.

5.1 Model comparison

Posterior odds ratios are the basis of model comparison, by which is meant both model averaging and model choice. The essential technical task in model comparison is obtaining the marginalized likelihood M_{jT} defined in (2.6.3). In describing how the marginalized likelihood can be obtained using a posterior simulator it is convenient to drop the subscript j denoting the model. For reasons discussed in Section 2.6 it is essential to distinguish between probability distribution functions and their kernels in the marginalized likelihood. In what follows, $p(\theta)$ always denotes the properly normalized prior density and $p(\mathbf{Y}_T|\theta)$ the properly normalized data density.

There are three conditions that a good approach to the computation of the marginalized likelihood M_T should satisfy.

(1) Given a large number of models it is much easier to summarize the comparative evidence through the marginalized likelihood than through pairwise Bayes factors. Therefore, the approach should provide a simulation-consistent approximation of M_T alone, rather than the Bayes factor comparing two models. For example, it is sometimes easy to compute a Bayes factor using (2.7.1) and (2.7.2), but that does not meet this criterion.

(2) The development of a posterior simulator, its execution, and the organization of simulator output all require real resources. Therefore, the numerical approximation of M_T should require only the original simulator output and not any additional, auxiliary simulations.

(3) Accurate approximations are always desirable. The accuracy of the approximation of M_T should be of the same order as the approximation of posterior moments in the model. Ideally, it should be convenient to assess numerical accuracy using a central limit theorem.

For posterior simulators based on independence sampling it is generally straightforward to satisfy all three criteria. In the case of importance sampling let $j(\theta)$

denote the p.d.f. of the importance sampling distribution, not merely the kernel. Since importance sampling distributions are chosen in part with regard to the convenience of generating draws from them, their normalizing constants are generally known. So long as the support of the importance sampling distribution includes the support of the posterior distribution,

$$\text{(5.1.1)} \qquad \hat{M}_T^{(M)} = M^{-1}\sum_{m=1}^{M} \mathrm{p}(\theta_m)\mathrm{p}(\mathbf{Y}_T|\theta_m)/\mathrm{j}(\theta_m) = M^{-1}\sum_{m=1}^{M} \mathrm{w}(\theta_m) \rightarrow \int_\Theta \mathrm{p}(\theta)\,\mathrm{p}(\mathbf{Y}_T|\theta)d\theta = M_T.$$

And if

$$\text{(5.1.2)} \qquad \int_\Theta \left[\mathrm{p}(\theta)^2\,\mathrm{p}(\mathbf{Y}_T|\theta)^2/\mathrm{j}(\theta)\right]d\theta = \int_\Theta \mathrm{w}(\theta)^2\,\mathrm{j}(\theta)d\theta < \infty$$

then

$$M^{1/2}\left(\hat{M}_T^{(M)} - M_T\right) \Rightarrow \mathrm{N}(0, \sigma^2)$$

where

$$\sigma^2 = \int_\Theta \left[\mathrm{p}(\theta)\mathrm{p}(\mathbf{Y}_T|\theta)/j(\theta) - M_T\right]^2 \mathrm{j}(\theta)d\theta$$

and

$$\hat{\sigma}^2 = M^{-1}\sum_{m=1}^{M}\left[\mathrm{p}(\theta_m)\mathrm{p}(\mathbf{Y}_T|\theta_m)/\mathrm{j}(\theta_m) - \hat{M}_T\right]^2 \rightarrow \sigma^2.$$

A sufficient condition for these results is that the weight function $\mathrm{w}(\theta)$ be bounded above, the same condition that is most useful in establishing the simulation-consistency of importance sampling simulators.

This approximation to the marginalized likelihood was used in Geweke (1989a). More recently it has been proposed by Gelfand and Dey (1994); see also Raftery (1995). The practical considerations involved are the same as those in the approximation of posterior moments using importance sampling. For the sake of efficiency the importance sampling distribution should not be too diffuse relative to the posterior distribution. For example $\mathrm{j}(\theta) = \mathrm{p}(\theta)$ satisfies (5.1.2) and leads to the very simple approximation $\hat{M}_T^{(M)} = M^{-1}\sum_{m=1}^{M}\mathrm{p}(\mathbf{Y}_T|\theta_m)$. But the prior distribution works well as an importance sampler only if sample size is quite small and θ is of very low dimension (Kloek and van Dijk, 1978). For an evaluation of the use of the prior in this way, see McCulloch and Rossi (1991).

Acceptance sampling from a source density $\mathrm{r}(\theta)$ is so similar to importance sampling that exactly the same procedure can be used to produce $\hat{M}_T^{(M)}$. The ratio $\mathrm{p}(\theta_m)\mathrm{p}(\mathbf{Y}_T|\theta_m)/\mathrm{r}(\theta_m)$ is needed for the acceptance probability in any event. The only additional work is to record $\mathrm{p}(\theta_m)\mathrm{p}(\mathbf{Y}_T|\theta_m)/\mathrm{r}(\theta_m)$ whether the draw is accepted or not,

and then to set $\hat{M}_T^{(M)} = M^{-1}\sum_{m=1}^{M} p(\theta_m)p(\mathbf{Y}_T|\theta_m)/r(\theta_m)$, the summation being taken over all candidate draws.

Simulation-consistent approximation of the marginalized likelihood from the output of a Markov chain Monte Carlo posterior simulator is a greater challenge, and has spawned a substantial recent literature. No method will fully meet the three criteria stipulated above, without more fundamental progress on the application of central limit theorems. Many methods are specialized to particular kinds of models and require at least two models for the computations because they provide Bayes factors rather than marginalized likelihoods. One example was presented in Section 4.2: the prior distribution that places mass at zero on regression coefficients produces a posterior distribution which provides Bayes factors for models in which the regressors are subsets of the superset. Methods have been developed for approximation of Bayes factors when the dimension of the parameter vectors in the two models is the same (Meng and Wong, 1993; Gelman and Meng, 1994; Chen and Shao, 1994), or the models are nested (Chen and Shao, 1995). A more general procedure is due to Carlin and Chib (1995) but this requires simultaneous simulation of two models. Methods that exploit the decomposition of the marginalized likelihood into predictive likelihoods (Kass and Raftery, 1995, Section 3.2) in effect require the consideration of many models (Gelfand, Dey and Chang, 1992; Geweke, 1994; Min, 1995).

Many straightforward approaches yield procedures with impractically slow convergence rates. A leading example is the "harmonic mean of the likelihood function" suggested by Newton and Raftery (1994): if $g(\theta) = [p(\theta)p(\mathbf{Y}_T|\theta)]^{-1}$ then $E[g(\theta)] = M_T^{-1}$. But $g(\theta)$ generally has no higher moments and consequently numerical approximations are poor.

At this juncture the procedure for approximating the marginalized likelihood from the output of a Markov chain Monte Carlo posterior simulator that comes closest to satisfying all three criteria is a modification of the harmonic mean of the likelihood function, suggested in Gelfand and Dey (1994). They observed that

(5.1.3) $$E[f(\theta)/p(\theta)p(\mathbf{Y}_T|\theta)] = M_T^{-1}$$

for any p.d.f. $f(\theta)$ whose support is contained in Θ. One can approximate (5.1.3) from the output of any posterior simulator in the obvious way, but for this approximation to have a practical rate of convergence $f(\theta)/p(\theta)p(\mathbf{Y}_T|\theta)$ should be uniformly bounded. Gelfand and Dey (1994) and Raftery (1995) interpret this condition as requiring that $f(\theta)$ have "thin tails" relative to the likelihood function.

It is not difficult to guarantee both the boundedness and thin tail condition in (5.1.3). Consider first the case in which $\Theta = \mathrm{R}^k$. From the output of the posterior simulator define $\hat{\theta}_M = M^{-1}\sum_{m=1}^{M}\theta_m$ and $\hat{\Sigma}_M = M^{-1}\sum_{m=1}^{M}\left(\theta_m - \hat{\theta}_M\right)\left(\theta_m - \hat{\theta}_M\right)'$. [Since the posterior simulator is a Markov chain Monte Carlo algorithm, it is assumed that $\mathrm{w}(\theta_m) = 1$. If the posterior simulator is an importance sampler, then (5.1.1) can be applied directly.] It is not essential that the posterior mean and variance of θ exist. Then take

$$\text{(5.1.4)} \qquad \mathrm{f}(\theta) = 2(2\pi)^{-k/2}\left|\hat{\Sigma}_M\right|^{-1/2}\exp\left[-\tfrac{1}{2}\left(\theta_m - \hat{\theta}_M\right)'\Sigma^{-1}\left(\theta_m - \hat{\theta}_M\right)\right]\chi_{\hat{\Theta}_M}(\theta),$$

$$\hat{\Theta}_M = \left\{\theta : \left(\theta_m - \hat{\theta}_M\right)'\Sigma^{-1}\left(\theta_m - \hat{\theta}_M\right) \le \chi^2_{.5}(k)\right\}.$$

If the posterior is uniformly bounded away from 0 on every compact subset of Θ, then the function of interest $\mathrm{f}(\theta)/\mathrm{p}(\theta)\mathrm{p}(\theta|\mathbf{Y}_T)$ possesses posterior moments of all orders. For a wide range of regular problems, this function will be approximately constant on $\hat{\Theta}_M$, which is nearly ideal.

If $\hat{\Theta}_M$ is not included in Θ some modifications of this procedure are required. In some cases it may be easy to reparameterize the model so that $\Theta = \mathrm{R}^k$. If not, the domain of integration for the function of interest $\mathrm{f}(\theta)/\mathrm{p}(\theta)\mathrm{p}(\mathbf{Y}_T|\theta)$ can be redefined to be $\hat{\Theta}_M \cap \Theta$ or a subset of $\hat{\Theta}_M \cap \Theta$, and a new normalizing constant for $\mathrm{f}(\theta)$ can be well approximated by taking a sequence of i.i.d. draws $\{\theta_t\}$ from the original distribution with p.d.f. (5.1.4) and averaging $\chi_\Theta(\theta_t)$, at the cost of an additional, but simple, simulation.

In the case of the Gibbs sampler there is an entirely different procedure due to Chib (1995) that provides quite accurate evaluations of the marginalized likelihood, at the cost of additional simulations. Suppose that the output from the blocking $\theta' = \left(\theta'^{(1)},\ldots,\theta'^{(B)}\right)$ is available, and that the conditional p.d.f.'s $\mathrm{p}\left(\theta^{(j)}\middle|\theta^{(i)}(i \ne j),\mathbf{Y}_T\right)$ can be evaluated in closed form for all j. [This latter requirement is generally satisfied.] Suppose further that condition 1 or 2 for convergence of the Gibbs sampler is satisfied.

From the identity $\mathrm{p}(\theta|\mathbf{Y}_T) = \mathrm{p}(\theta)\mathrm{p}(\mathbf{Y}_T|\theta)/M_T$, $M_T = \mathrm{p}(\theta^*)\mathrm{p}\left(\mathbf{Y}_T|\theta^*\right)/\mathrm{p}\left(\theta^*|\mathbf{Y}_T\right)$ for any $\theta^* \in \Theta$. [In all cases, $\mathrm{p}(\cdot)$ denotes a properly normalized density and not merely a kernel.] Typically $\mathrm{p}\left(\mathbf{Y}_T|\theta^*\right)$ and $\mathrm{p}\left(\theta^*\right)$ can be evaluated in closed form, but $\mathrm{p}\left(\theta^*|\mathbf{Y}_T\right)$ cannot. A marginal/conditional decomposition of $\mathrm{p}\left(\theta^*|\mathbf{Y}_T\right)$ is

$$\mathrm{p}\left(\theta^*|\mathbf{Y}_T\right) = \mathrm{p}\left(\theta^{*(1)}\middle|\mathbf{Y}_T\right)\mathrm{p}\left(\theta^{*(2)}\middle|\theta^{*(1)},\mathbf{Y}_T\right)\cdots\mathrm{p}\left(\theta^{*(B)}\middle|\theta^{*(1)},\ldots,\theta^{*(B-1)},\mathbf{Y}_T\right).$$

The first term in the product of B terms can be approximated from the output of the posterior simulator because

$$M^{-1}\sum_{m=1}^{M} \mathrm{p}\left(\theta^{*(1)} \middle| \theta_m^{(2)}, \ldots, \theta_m^{(B)}, \mathbf{Y}_T\right) \rightarrow \mathrm{p}\left(\theta^{*(1)} \middle| \mathbf{Y}_T\right).$$

To approximate $\mathrm{p}\left(\theta^{*(j)} \middle| \theta^{*(1)}, \ldots, \theta^{*(j-1)}, \mathbf{Y}_T\right)$, first execute the Gibbs sampling algorithm with the parameters in the first j blocks fixed at the indicated values, thus producing a sequence $\left\{\theta_{jm}^{(j+1)}, \ldots, \theta_{jm}^{(B)}\right\}$ from the conditional posterior. Then

$$M^{-1}\sum_{m=1}^{M} \mathrm{p}\left(\theta^{*(j)} \middle| \theta^{*(1)}, \ldots, \theta^{*(j-1)}, \theta_{jm}^{(j+1)}, \ldots, \theta_{jm}^{(B)}, \mathbf{Y}_T\right) \rightarrow \mathrm{p}\left(\theta^{*(j)} \middle| \theta^{*(1)}, \ldots, \theta^{*(j-1)}, \mathbf{Y}_T\right).$$

Extension to include latent variables, so long as the vector of latent variables is not blocked, is straightforward; see Chib (1995) for details and applications.

5.2 Bayesian communication

An investigator cannot anticipate the uses to which her work will be put, or the variants on her model that may interest a client. Different uses will be reflected in different functions of interest. Variants will often revolve around changes in the prior distribution. Any investigator who has publicly reported results has confronted the constraint that only a few representative findings can be conveyed in written work.

Posterior simulators provide a clear answer to the question of what the investigator should report, and in the process remove the constraint that only a few representative findings can be communicated. What should be reported is the $M \times (k+2)$ *simulator output matrix*,

$$\begin{bmatrix} \theta_1' & \mathrm{w}(\theta_1) & \mathrm{p}(\theta_1) \\ \vdots & \vdots & \vdots \\ \theta_m' & \mathrm{w}(\theta_m) & \mathrm{p}(\theta_m) \end{bmatrix},$$

by making it publicly and electronically available. In a reasonably large problem ($M = 10{,}000$ and $k = 100$) the corresponding file occupies about 3.2 megabytes of storage (at a current capital cost of about US$1.40) and can be moved over the internet in about a minute.

Given the simulator output matrix the client can immediately compute approximations to posterior moments not reported or even considered by the investigator. For example, a client reading a research report might be skeptical that the investigator's model, prior and data set provide much information about the effects of an interesting change in a policy variable on the outcome in question. If the simulator output matrix is available via FTP anonymous the client can obtain the exact (up to numerical

approximation error, which can also be evaluated) answer to his query without arising from his office chair in considerably less time than required to read the research report.

With a small amount of additional effort the client can modify many of the investigator's assumptions. Suppose the client wishes to evaluate $E[g(\theta)|\mathbf{Y}_T]$ using his own prior density $p^*(\theta)$ rather than the investigator's prior density $p(\theta)$. Suppose further that the support of the investigator's prior distribution includes the support of the client's prior. Then the investigator's posterior distribution may be regarded as an importance sampling distribution for the client's posterior density. The client reweights the investigator's $\{\theta^m\}_{m=1}^M$ using the function

$$w^*(\theta) = \frac{p^*(\theta|\mathbf{Y}_t)}{p(\theta|\mathbf{Y}_t)} = \frac{p^*(\theta)L(\theta|\mathbf{Y}_t)}{p(\theta)L(\theta|\mathbf{Y}_t)} = \frac{p^*(\theta)}{p(\theta)},$$

where $p^*(\theta|\mathbf{Y}_t)$ denotes the client's posterior distribution. The client then approximates his posterior moment $E^*[g(\theta)|\mathbf{Y}_t]$ by

$$\bar{g}_M^* \equiv \sum_{m=1}^M w^*(\theta_m)w(\theta_m)g(\theta_m) \Big/ \sum_{m=1}^M w^*(\theta_m)w(\theta_m) \to E^*[g(\theta)|\mathbf{Y}_t] \equiv \bar{g}^*.$$

The result $\bar{g}_M^* \to \bar{g}^*$ follows almost at once from Tierney (1994); see Geweke (1995d).

The efficiency of the reweighting scheme requires some similarity of $p^*(\theta)$ and $p(\theta)$. In particular, both reasonable convergence rates and the use of a central limit theorem to assess numerical accuracy essentially require that $p^*(\theta)/p(\theta)$ be bounded. Across a set of diverse clients this condition is more likely to be satisfied the more diffuse is $p(\theta)$, and is trivially satisfied for the improper prior $p(\theta) \propto$ constant if the client's prior is bounded. In the latter case the reweighting scheme will be efficient so long as the client's prior is uninformative relative to the likelihood function. This condition is stated precisely in Theorem 2 of Geweke (1989b). Diagnostics described there will detect situations in which the reweighting scheme is inefficient, as will standard errors of numerical approximation as well. If the investigator chooses to use an improper prior for reporting, it is of course incumbent on her to verify the existence of the posterior distribution and convergence of her posterior simulator.

Including $p(\theta_m)$ in the standard simulator output file avoids the need for every client who wishes to impose his own priors to re-evaluate the investigator's prior. Of course, the $p^*(\theta)$'s need not be the client's subjective priors: they may simply be devices by which clients explore robustness of results with respect to alternative reasonable priors.

The client can also undertake a nonparametric analysis of prior robustness for his posterior moments, using the density ratio class described in Section 2.5. There is an efficient, generic algorithm to determine the maximum value of the posterior moment of the density ratio class described in Geweke and Petrella (1995), which extends earlier

work of Wasserman and Kadane (1992). First, order $g_m = g(\theta_m)$ in monotone nondecreasing order, and define

$$(5.2.1) \qquad Q_\ell = \frac{\sum_{m=1}^{\ell} g_m u_m + \sum_{m=\ell+1}^{M} g_m v_m}{\sum_{m=1}^{\ell} u_m + \sum_{m=\ell+1}^{M} v_m} \quad (\ell = 0,\ldots,M)$$

where $u_m = w(\theta_m)a(\theta_m)/\tilde{p}(\theta_m)$ and $v_m = w(\theta_m)b(\theta_m)/\tilde{p}(\theta_m)$ $(m = 1,\ldots,M)$. Using successive bisection determine an index ℓ such that $g_\ell \le Q_\ell \le g_{\ell+1}$. Increment this index ℓ until $g_\ell > Q_\ell$ and then set $m^* = \ell - 1$. This provides the global maximum of (5.2.1). Under weak conditions (Geweke and Petrella (1995), Proposition 4),

$$Q_{m^*} \to \sup_{p:a\le p\le b} \int_\Theta g(\theta)L(\theta)p(\theta)d\theta \Big/ \int_\Theta L(\theta)p(\theta)d\theta.$$

The reweighting scheme permits updating of the investigator's results at low cost. If observations $T+1,\ldots,F$ beyond the T originally used have become available then

$$\begin{aligned} p^F(\theta|\mathbf{Y}_T) \propto p(\theta)L[\theta|\mathbf{Y}_F] &= p(\theta)L[\theta|\mathbf{Y}_t]\prod_{s=T+1}^{F} f_s(\mathbf{y}_s|\mathbf{Y}_{s-1},\theta) \\ &= p[\theta|\mathbf{Y}_t]\prod_{s=T+1}^{F} f_s(\mathbf{y}_s|\mathbf{Y}_{s-1},\theta). \end{aligned}$$

The client therefore forms the approximation to the updated posterior moment $E[g(\theta)|\mathbf{Y}_F]$,

$$\bar{g}_M^F \equiv \sum_{m=1}^{M} w^F(\theta_m)w(\theta_m)g(\theta_m) \Big/ \sum_{m=1}^{M} w^F(\theta_m)w(\theta_m) \to E[g(\theta)|\mathbf{Y}_F] \equiv \bar{g}^F$$

with $w^F(\theta) \equiv \prod_{s=T+1}^{F} f_s(\mathbf{y}_s|\mathbf{Y}_{s-1},\theta)$. Rare pathological cases aside the likelihood function and therefore $w^F(\theta)$ is bounded. If $F-T$ is small relative to T, and there is no major change in the data generating process between T and $T+F$, the new approximation will be efficient. But as F grows, efficiency diminishes and at some point the approximation $\bar{g}^F$ becomes too inaccurate to be useful.

The potential for clients to alter investigators' priors, update their results, and examine alternative posterior moments, exists given current technology. All that is required is for Bayesian investigators to begin making their results available in a conventional format, in the same way that many now provide public access to text and data. Once this is done, colleagues, students, and policy makers may employ the results to their own ends much more flexibly than has heretofore been possible, with modest technical requirements.

6. Conclusion

The introduction set forth three propositions which, if believed, would keep most econometricians from using Bayesian methods in most applications. The intervening part of this chapter has presented some developments that, to the extent an econometrician

was previously unaware of them, might well revise beliefs about these propositions. I conclude with a personal revision, taking the propositions in reverse order.

The statement that most posterior moments are unobtainable because of technical difficulties with integration was true for many models a decade ago, although steady inroads had been made (Zellner, 1971; Richard, 1973; Dreze, 1977; Kloek and van Dijk, 1978; Bauwens, 1984). With breakthroughs in importance sampling (Geweke, 1988, 1989) and especially in Markov chain Monte Carlo (Gelfand and Smith, 1990; Tierney, 1994) the statement is false. Econometric models in which any posterior moment of interest that exists cannot be obtained using a posterior simulator are now the exception, not the rule. In the past two years there have emerged important cases in which the posterior moments are more easily and reliably obtained than are non-Bayesian estimates. This is especially the case for models with latent variables; an example was presented in Section 4.6.

The implications of posterior simulators for model comparison and the communication of results bear on the subjectivity of the prior distribution. It is now the case that the reader -- or more generally the client, as described in Section 5 -- need not be passive and can conveniently take a role in the specification of econometric models and their application. The reader is free to explore posterior moments of his choice and examine the implications of revisions of the investigator's prior distribution for those moments. Indeed, the investigator can choose her prior to facilitate this process, as described in Section 5.2.

If exploration of priors by readers becomes commonplace, then questions about the impact of subjective choices made by the econometrician shifts from the prior distribution to the functional form of the data distribution. This choice is made subjectively in Bayesian and many non-Bayesian procedures alike, and when it is not made explicitly in non-Bayesian procedures then implicit restrictions on functional form exist in the assumed applicability of a central limit theorem. Alternative functional forms for data distributions can be compared using Bayes factors; no such general comparison is possible using non-Bayesian methods. Within the past two years reliable methods of approximation for the marginalized likelihoods that constitute Bayes factors have become available, and rapid further progress is currently being made.

None of the innovations in posterior simulators relieves the Bayesian econometrician of the burden of specifying a likelihood function and a prior distribution. To the contrary, there are three reasons for the econometrician to subject himself to this discipline, two of which have been made more compelling by developments in posterior simulators. First, specification of the likelihood function and prior distribution make assumptions explicit,

and this has clear benefits in interpreting what the econometrician has done. [For example, in non-Bayesian approaches an alternative to specifying a likelihood function is to assume the applicability of a central limit theorem or a particular nonparametric expansion, and an alternative to stipulating a prior distribution is to discard or discount results that don't look right.] Posterior simulators have no implications here. Second, if one specifies a likelihood function and prior distribution then one can obtain useful results, not merely expressions. Posterior simulators have made this possible to the point that it is now increasingly easier to obtain posterior moments than to compute non-Bayesian estimates. Finally, and most important, economic theory that addresses decision making under uncertainty -- which is to say, most economic theory -- requires distributional assumptions, and results in theory are generally not robust to changes in these distributions. Decisions and policy recommendations depend on distributional assumptions. Econometricians cannot address these matters without being concerned with likelihood functions and prior distributions.

References

Albert, J. and S. Chib, 1993, "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association* **88**: 669-679.

Albert, J. and S. Chib, 1994, "Computation in Bayesian Econometrics: An Introduction to Markov Chain Monte Carlo," in T. Fomby and R.C. Hill (eds.), *Advances in Econometrics: Bayesian Computational Methods and Applications*, forthcoming. Greenwich: JAI Press.

Amemiya, T., 1984, "Tobit Models: A Survey," *Journal of Econometrics*, 3-61.

Anderson, T.W., 1984, *An Introduction to Multivariate Statistical Analysis*. New York: Wiley. (Second edition)

Barnett, W.A. and Y.W. Lee, 1985, "The Global Properties of the Minflex Laurent, Generalized Leontief, and Translog Functional Forms," *Econometrica* **53**: 1421-1437.

Bartlett, M.S., 1957, "A Comment on D.V. Lindley's Statistical Paradox," *Biometrika* **44**: 533-534.

Bauwens, L., 1984, *Bayesian Full Information Analysis of Simultaneous Equation Models Using Integration by Monte Carlo*. Berlin: Springer-Verlag.

Bauwens, L. and J.F. Richard, 1985, "A 1-1 Poly-*t* Random Variable Generator with Applications to Monte Carlo Integrations," *Journal of Econometrics* **29**: 19-46.

Berger, J.O., 1985, *Statistical Decision Theory and Bayesian Analysis* (Second Edition). New York: Springer-Verlag, 1985.

Berger, J.O., 1994, "An Overview of Robust Bayesian Analysis" (with discussion), *Test* **3**: 5-124.

Berger, J.O. and M. Oh, 1993, "Integration of Multimodal Functions by Monte Carlo Importance Sampling," *Journal of the American Statistical Association* **88**: 450-456.

Berger, J.O. and R.L. Wolpert, 1988, *The Likelihood Principle*. Hayward: Institute of Mathematical Statistics. (Second edition)

Bernardo, J.M., and A.F.M. Smith, *Bayesian Theory*. New York: Wiley, 1994.

Blattberg, R.C. and E.I. George, 1991, "Shrinkage Estimation of price and Promotional Elasticities: Seemingly Unrelated Equations," *Journal of the American Statistical Association* **86**: 304-315.

van den Broeck, J., G. Koop, J. Osiewalski and M.F.J Steel, 1994, "Stochastic Frontier Models: A Bayesian Perspective," *Journal of Econometrics* **61**: 273-303.

Carlin, B.P. and N.G. Polson, 1991, "Inference for Nonconjugate Bayesian Models Using the Gibbs Sampler," *Canadian Journal of Statistics* **19**: 399-405.

Carlin, B. and S. Chib, 1995, "Bayesian Model Choice via Markov Chain Monte Carlo," *Journal of the Royal Statistical Society Series B* **57**: 473-484.

Casella, G. and E.I. George, 1992, "Explaining the Gibbs Sampler," *The American Statistician 46*: 167-174.

Casella, G. and C.P. Robert, 1994, "Rao-Blackwellization of Sampling Schemes," Cornell University Biometrics Unit technical report BU-1252-M.

Chalfant, JA. and N.E. Wallace, 1993, "Bayesian Analysis and Regularity Conditions on Flexible Functional Forms: Application to the U.S. Motor Carrier Industry," in W.E. Griffiths, H. Lütkepohl and M.E. Bock (eds.), *Readings in Econometric Theory and Practice*. Amsterdam: Elsevier-North Holland

Chamberlain, G., and E. Leamer, 1976, "Matrix Weighted Averages and Posterior Bounds," *Journal of the Royal Statistical Society, Series B*, **38**, 73-84.

Chao, J.C. and P.C.B. Phillips, 1994, "Bayesian Posterior Distributions in Limited Information Analysis of the Simultaneous Equations Model," Yale University Cowles Foundation working paper.

Chen, M. and Q. Shao, 1994, "On Monte Carlo Methods for Estimating Ratios of Normalizing Constants," National University of Singapore Department of Mathematics Research Report No. 627.

Chen, M. and Q. Shao, 1995, "Estimating Ratios of Normalizing Constants for Densities with Different Dimensions," Worcester Polytechnical Institute technical report.

Chib, S., 1992, "Bayes Inference in the Tobit Censored Regression Model," *Journal of Econometrics* **51**: 79-99.

Chib, S., 1995, "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association, Journal of the American Statistical Association*, forthcoming. Also Washington University Olin School of Business working paper.

Chib, S. and E. Greenberg, 1994a, "Markov Chain Simulation Methods in Econometrics," Washington University Olin School of Business working paper.

Chib, S. and E. Greenberg, 1994b, "Understanding the Metropolis-Hastings Algorithm," Washington University Olin School of Business working paper.

Chib, S. and E. Greenberg, 1995, "Hierarchical Analysis of SUR Models with Extensions to Correlated Serial Errors and Time Varying Parameter Models," *Journal of Econometrics*, forthcoming. Also Washington University Olin School of Business working paper.

Clark, C., 1961, "The Greatest of a Finite Set of Random Variables," *Operations Research* **9**: 145-162.

Clyde, M.. and G. Parmigiani, 1994, "Bayesian Variable Selection and Prediction with Mixtures," *Journal of Biopharmaceutical Statistics*, forthcoming. (Also Duke University ISDS discussion paper.)

Davis, P.J., and P. Rabinowitz, 1984, *Methods of Numerical Integration*. Orlando: Academic Press. (Second edition)

Davis, W. W., 1978, "Bayesian Analysis of the Linear Model Subject to Linear Inequality Constraints, *Journal of the American Statistical Association*, **73**, 573-579.

DeGroot, M., 1970, *Optimal Statistical Decisions*. New York: McGraw-Hill.

DeRobertis, L., and J. A. Hartigan, 1981, "Bayesian Inference Using Intervals of Measures," *The Annals of Statistics* **9**: 235-244.

Diewert, W.E. and T.J. Wales, 1987, "Flexible Functional Forms and Global Curvature Conditions," *Econometrica* **55**: 43-88.

Dreze, J.H., 1976, "Bayesian Limited Information Analysis of the Simultaneous Equation Model," *Econometrica* **46**: 1045-1075.

Dreze, J.H., 1977, "Bayesian Regression Analysis Using Poly-*t* Densities," *Journal of Econometrics* **6**: 329-354.

Dreze, J.H. and J.F. Richard, 1983, "Bayesian Analysis of Simultaneous Equation Systems," in Z. Griliches and M. Intriligator (eds.), *Handbook of Econometrics*, 369-377. Amsterdam: North-Holland. CHECK PAGES

Gammerman , D. and H.S. Migon, 1993, "Dynamic Hierarchical Models," *Journal of the Royal Statistical Society Series B*, **55**: 629-642.

Gelfand, A.E., D.K. Dey and H. Chang, 1992, "Model Determination Using predictive Distributions with Implementation via Sampling-Based Methods," in J.M. Bernardo, J.O. Berger, A.P. Dawid and AF.M. Smith (eds.), *Bayesian Statistics 4*. Oxford: Oxford University Press.

Gelfand, A.E., and D.K. Dey, 1994, "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society Series B* **56**: 501-514.

Gelfand, A.E., and A.F.M. Smith, 1990, "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association* **85**: 398-409.

Gelman, A., and D.B. Rubin, 1992, "Inference from Iterative Simulation Using Multiple Sequences," *Statistical Science* **7**: 457-472.

Gelman, A. and X.L. Meng, 1994, "Path Sampling for Computing Normalizing Constants: Identities and Theory," University of Chicago Department of Statistics Technical Report No. 377.

Geman, S., and D. Geman, 1984, "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**: 721-741.

George, E.I. and R.E. McCulloch, 1993, "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association* **88**: 881-889.

George, E.I., and R.E. McCulloch, 1994, "Fast Bayes Variable Selection," University of Texas CSS Technical Report #94-01.

George, E.I., R.E. McCulloch and R. Tsay, 1994, "Two Approaches to Bayesian Model Selection with Applications," in D. Berry, K. Chaloner and J. Geweke (eds.), *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner.* New York: Wiley.

Geweke, J., 1986, "Exact Inference in the Inequality Constrained Normal Linear Regression Model," *Journal of Applied Econometrics*, **1**, 127-141.

Geweke, J., 1988, "Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference," *Journal of Econometrics* **38**: 73-89.

Geweke, J, 1989a, "Exact Predictive Densities in Linear Models with ARCH Disturbances," *Journal of Econometrics, 1989*, **40**: 63-86.

Geweke, J., 1989b, "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica* **57**: 1317-1340.

Geweke, J., 1991, "Efficient Simulation from the Multivariate Normal and Student-*t* Distributions Subject to Linear Constraints," in E. M. Keramidas (ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 571-578. Fairfax, VA: Interface Foundation of North America.

Geweke, J, 1992, "Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments," in J.O. Berger, J.M. Bernardo, A.P. Dawid, and A.F.M. Smith (eds.), *Proceedings of the Fourth Valencia International Meeting on Bayesian Statistics*, 169-194. Oxford: Oxford University Press, 1992.

Geweke, J., 1993, "Bayesian Treatment of the Student-*t* Linear Model," *Journal of Applied Econometrics*, 1993, **8**, S19-S40.

Geweke, J., 1994, "Bayesian Comparison of Econometric Models," Federal Reserve Bank of Minneapolis working paper No. 532.

Geweke, J., 1995a, "Monte Carlo Simulation and Numerical Integration," in H. Amman, D. Kendrick and J. Rust (eds.), *Handbook of Computational Economics.* Amsterdam: North-Holland, forthcoming. Also Federal Reserve Bank of Minneapolis Staff Report No. 192.

Geweke, J., 1995b, "Simulation-Based Bayesian Inference for Economic Time Series," in preparation.

Geweke, J., 1995c, "Bayesian Inference for Linear Models Subject to Linear Inequality Constraints," in W.O. Johnson, J.C. Lee and A. Zellner (eds.), *Forecasting, Prediction and Modeling in Statistics and Econometrics: Bayesian and non-Bayesian Approaches.* New York: Springer-Verlag, forthcoming. Also Federal Reserve Bank of Minneapolis working paper No. 552.

Geweke, J., 1995d, "Bayesian Communication," in preparation.

Geweke, J. and M. Keane, 1995, "An Empirical Analysis of the Male Income Dynamics in the PSID: 1986-1989," University of Minnesota Department of Economics working paper.

Geweke, J., M. Keane and D. Runkle, 1994a, "Alternative Computational Approaches to Statistical Inference in the Multinomial Probit Model," *Review of Economics and Statistics*, 1994, **76**, 609-632.

Geweke, J., M. Keane and D. Runkle, 1994b, "Statistical Inference in Multinomial Multiperiod Probit Models," Federal Reserve Bank of Minneapolis Staff Report No. 177.

Geweke, J., M. Keane and D. Runkle, 1995, "Recursively Simulating Multinomial Multiperiod Probit Probabilities," *American Statistical Association 1994 Proceedings of the Business and Economic Statistics Section.*

Geweke, J. and L. Petrella, 1995, "Prior Density Ratio Class Robustness in Econometrics," Federal Reserve Bank of Minneapolis working paper No. __.

Geyer, C.J., 1992, "Practical Markov Chain Monte Carlo," *Statistical Science* **7**: 473-481.

Ghosh, M., A.K. Saleh and P.K. Sen, 1989, "Empirical Bayes Subset Information in Regression Models," *Statistics and Decisions* **7**: 15-36.

Gourieroux, C., A. Holly, and A. Monfort, 1982, "Likelihood Ratio Test, Wald Test, and Kuhn-Tucker Test in Linear Models with Inequality Constraints on the Regression Parameters, *Econometrica*, **50**, 63-80.

Hajivassiliou, V. and D. McFadden, "The Method of Simulated Scores for the Estimation of LDV Models with an Application to External Debt Crises," Cowles Foundation Discussion Paper 967, Yale University.

Hajivassiliou, V., D. McFadden and P. Ruud, 1995, "Simulation of Multivariate Normal Orthant Probabilities: Methods and Programs," *Journal of Econometrics*, forthcoming.

Hammersly, J.M., and D.C. Handscomb, 1964, *Monte Carlo Methods*. London: Methuen and Company.

Hammersly, J.M., and K.W. Morton, 1956, "A New Monte Carlo Technique: Antithetic Variates," *Proceedings of the Cambridge Philosophical Society* **52**: 449-474.

Hannan, E.J., 1970, *Multiple Time Series*. New York: Wiley.

Hastings, W.K., 1970, "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika* **57**: 97-109.

Heckman, J. and G. Sedlacek , 1985, "Heterogeneity, Aggregation, and Market Wage Functions: An Empirical Model of Self-Selection in the Labor Market," *Journal of Political Economy* **93**: 1077-1125.

Hildreth, C., 1963, "Bayesian Statisticians and Remote Clients," *Econometrica* **31**: 422-438.

Johnson, N.L., and S. Kotz, 1970, *Distributions in Statistics: Continuous Univariate Distributions*. (In two volumes) New York: Wiley.

Johnson, N.L., and S. Kotz, 1972, *Distributions in Statistics: Continuous Multivariate Distributions*. New York: Wiley.

Judge, G. G., and T. Takayama, 1966, Inequality restrictions in regression analysis, *Journal of the American Statistical Association*, **61**, 166-181.

Kahn, M., and A.W. Marshall, 1953, "Methods of Reducing Sample Size in Monte Carlo Computations," *Operations Research* **1**: 263-278.

Kass, R.E. and A.E. Raftery, 1995, "Bayes Factors," *Journal of the American Statistical Association* **90**: 773-795.

Keane, M., 1990, Four Essays in Empirical Macro and Labor Economics. Unpublished Ph.D. dissertation, Brown University.

Kipnis, C., and S.R.S. Varadhan, 1986, "Central Limit Theorem for Additive Functionals of Reversible Markov Processes and Applications to Simple Exclusions," *Communications in Mathematical Physics* **104**: 1-19.

Kleibergen, F. and H.K van Dijk, 1994, "Bayesian Analysis of Simultaneous Equation Models Using Noninformative Priors," Tinbergen Institute discussion paper 94-134.

Kloek, T. and H.K. van Dijk, 1978, "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo," *Econometrica* **46**: 1-19.

Koop, G., 1994, "Recent Progress in Applied Bayesian Econometrics," *Journal of Economic Surveys* **8**: 1-34.

Koop, G. and D.J. Poirier, 1993, "Bayesian Analysis of Logit Models Using Natural Conjugate Priors," *Journal of Econometrics* **56**: 323-340.

Koop, G. and D.J. Poirier, 1994, "Rank-Ordered Logit Models: An Empirical Analysis of Ontario Voter Preferences," *Journal of Applied Econometrics* **9**: 369-388.

Koop, G., M.F.J. Steel and J. Osiewalski, 1995, "Posterior Analysis of Stochastic Frontier Models Using Gibbs Sampling," *Computational Statistics*, forthcoming.

Lavine, M., 1991a, "Sensitivity in Bayesian Statistics: The Prior and the Likelihood," *Journal of the American Statistical Association* **86**: 396-399.

Lavine, M., 1991b, "An Approach to Robust Bayesian Analysis for Multidimensional Parameter Spaces," *Journal of the American Statistical Association* **86**: 400-403.

Leamer, E., and G. Chamberlain, 1976, "A Bayesian Interpretation of Pretesting," *Journal of the Royal Statistical Society, Series B*, **38**, 85-94.

Lindley, D.V., 1957, "A Statistical Paradox," *Biometrika* **44**: 187-192.

Lovell, M. C., and E. Prescott, 1970, "Multiple Regression with Inequality Constraints: Pretesting Bias, Hypothesis Testing, and Efficiency," *Journal of the American Statistical Association*, **65**, 913-925.

Maddala, G.S., 1983, *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Madigan, D. and A.E. Raftery, 1994, "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association* **89**: 1535-1546.

Manski, C.F. and D. McFadden (eds.), 1981, *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge: MIT Press.

McCulloch, R.E. and P.E. Rossi, 1991, "A Bayesian Approach to Testing the Arbitrage Pricing Theory," *Journal of Econometrics* **49**: 141-168.

McCulloch, R.E. and P.E. Rossi, 1995, "An Exact Likelihood Analysis of the Multinomial Probit Model," *Journal of Econometrics* **64**: 207-240.

McCulloch, R.E., N.G. Polson and P.E. Rossi, 1995, "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters," University of Chicago Graduate School of Business working paper.

McFadden, D., 1989, "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica* **57**: 995-1026.

Meng, X.L. and W.H Wong, 1993, "Simulating Ratios of Normalizing Constants via a Simple identity," University of Chicago Department of Statistics Technical Report No. 365.

Mengersen, K.L and R.L. Tweedie, 1993, "Rates of Convergence of the Hastings and Metropolis Algorithms," Colorado State University Department of Statistics working paper.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller, 1953, "Equation of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics* **21**: 1087-1092.

Min. C., 1995, "Forecasting the Adoptions of New Consumer Durable Products," George Mason University School of Business Administration working paper.

Min, C., and A. Zellner, 1993, "Bayesian and non-Bayesian Methods for Combining Models and Forecasts with Applications to Forecasting International Growth Rates," *Journal of Econometrics* **56**: 89-118.

Newton, M.A. and A.E. Raftery, 1994, "Approximate Bayesian Inference by the Weighted Likelihood Bootstrap" (with discussion), *Journal of the Royal Statistical Society Series B* **56**: 3-48.

Numelin, E., 1984, *General Irreducible Markov Chains and Non-negative Operators*. Cambridge: Cambridge University Press.

Pakes, A. and D. Pollard, 1989, "Simulation and the Asymptotics of Optimization Estimators," *Econometrica* **57**: 1027-1058.

Percy, D.F., 1992, "Prediction for Seemingly Unrelated Regressions," *Journal of the Royal Statistical Society Series B*, **54**: 243-252.

Peskun, P.H., 1973, "Optimum Monte-Carlo Sampling using Markov Chains," *Biometrika* **60**: 607-612.

Poirier, D.J., 1988, "Frequentist and Subjectivist Perspectives on the Problem of Model Building in Economics" (with discussion). *Journal of Economic Perspectives* **2**: 120-170.

Poirier, D.J., 1989, "A Report from the Battlefront," *Journal of Business and Economic Statistics* **7**: 137-139.

Poirier, D.J., 1992, "A Return to the Battlefront," *Journal of Business and Economic Statistics* **10**: 473-474.

Poirier, D.J., 1995, *Intermediate Statistics and Econometrics: A Comparative Approach.* Cambridge: MIT Press.

Raftery, A.E., 1995, "Hypothesis Testing and Model Selection Via Posterior Simulation," University of Washington working paper.

Raftery, A.E. and S.M. Lewis, "Implementing MCMC," in W.R. Gilks, S. Richardson and D.J. Spiegelhalter (eds.), *Practical Markov Chain Monte Carlo*, forthcoming. Also University of Washington working paper, "The Number of Iterations, Convergence Diagnostics and Generic Metropolis Algorithms."

Raftery, A., D. Madigan and J. Hoeting, 1993, "Model Selection and Accounting for Model Uncertainty in Linear Regression Models," University of Washington Department of Statistics Technical Report No. 262.

Richard, J.F., 1973, *Posterior and Predictive Densities for Simultaneous Equation Models.* Berlin: Springer-Verlag.

Richard, J.F. and M.F.J. Steel, 1988, "Bayesian Analysis of Systems of Seemingly Unrelated Regression Equations Under a Recursive Extended Natural Conjugate Prior Density," *Journal of Econometrics 38*: 7-37.

Ripley, R.D., 1987, *Stochastic Simulation.* New York: Wiley.

Roberts, G.O., and A.F.M. Smith, 1994, "Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis-Hastings Algorithms," *Stochastic Processes and Their Applications* **49**: 207-216.

Rothenberg, T.J., 1975, "Bayesian Analysis of Simultaneous Equations Models," in S.E. Fienberg and A. Zellner (eds.), *Studies in Bayesian Econometrics and Statistics.* Amsterdam: North-Holland.

Rubin, D.B., 1987, *Multiple Imputation for Nonresponse in Surveys* New York: Wiley.

Stein, C., 1966, "An Approach to the Recovery of Inter-block Information in Balanced Incomplete Block Designs," in F.N. David (ed.), *Festschrift for J. Neyman*, 351-366. New York: Wiley.

Tanner, M.A., and W.-H. Wong, 1987: "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association* **82**, 528-550.

Terrell, D., 1995, "Incorporating Monotonicity and Concavity Conditions in Flexible Functional Forms," *Journal of Applied Econometrics*, forthcoming. Also Kansas State University Department of Economics working paper.

Thurstone, L., 1927, "A Law of Comparative Judgment," *Psychological Review* **34**: 273-286.

Tierney, L., 1991, "Exploring Posterior Distributions Using Markov Chains," in E.M. Keramaidas (ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, 563-570. Fairfax: Interface Foundation of North America, Inc.

Tierney, L., 1994, "Markov Chains for Exploring Posterior Distributions" (with discussion and rejoinder), *Annals of Statistics* **22**: 1701-1762. (Also Technical Report No. 560, University of Minnesota School of Statistics.)

Tobin, J., 1958, "Estimation of Relationships for Limited Dependent Variables," *Econometrica* **26**: 24-36.

Tsionas, E.G., 1994, Asset Returns in General Equilibrium with Scale-Mixture-of-Normals Endowment Processes. Unpublished Ph.D. dissertation, University of Minnesota.

Wasserman, L., and J. B. Kadane, 1992, "Computing Bounds on Expectations," *Journal of the American Statistical Association* **87**: 516-522.

Wei, C.G. and M.A. Tanner, 1990, "Posterior Computations for Censored Regression Data," *Journal of the American Statistical Association* **85**: 829-839.

West, M., 1984. "Outlier Models and Prior Distributions in Bayesian Linear Regression," *Journal of the Royal Statistical Society Series B* **46**: 431-439.

Wolak, F.A., 1987, "An Exact Test for Multiple Inequality and Equality Constraints in the Linear Regression Model," *Journal of the American Statistical Association*, **82**, 782-793.

Zellner, A., 1962, "An Efficient Method of Estimating Seemingly Unrelated Regressions and Test of Aggregation Bias," *Journal of the American Statistical Association* **57**: 500-509.

Zellner, A., 1971, *Bayesian Inference in Econometrics*. New York: Wiley.

Zellner, A. and C. Min, 1995, "Gibbs Sampler Convergence Criteria," *Journal of the American Statistical Association*, forthcoming.

Zellner, A., C. Min and D. Dallaire, 1994, "Bayesian Analysis of Simultaneous Equation, Asset-Pricing and Related Models Using Markov Chain Monte Carlo Techniques and Convergence Checks," University of Chicago Graduate School of Business working paper.

Zellner, A. and P.E. Rossi, 1984, "Bayesian Analysis of Dichotomous Quantal Response Models," *Journal of Econometrics* **25**: 365-393.

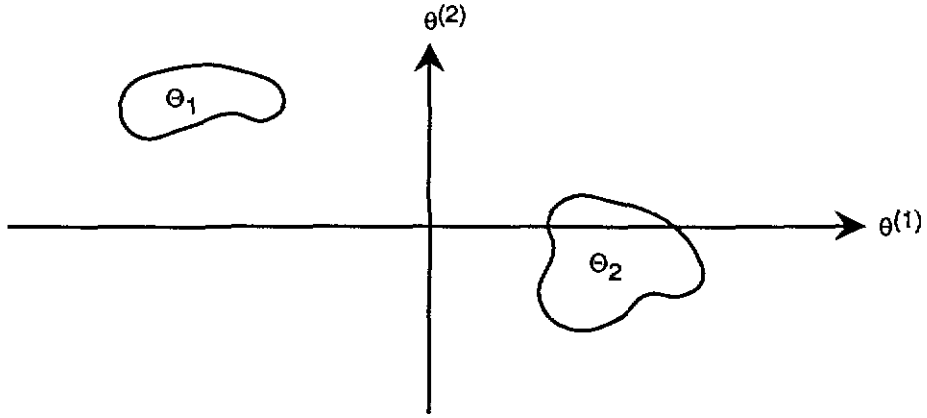

Figure 1. The disconnected support $\Theta = \Theta_1 \cup \Theta_2$ for the probability distribution implies that a Gibbs sampler with blocking $(\theta^{(1)}, \theta^{(2)})$ will not have the probability distribution as its invariant distribution, for any starting value.

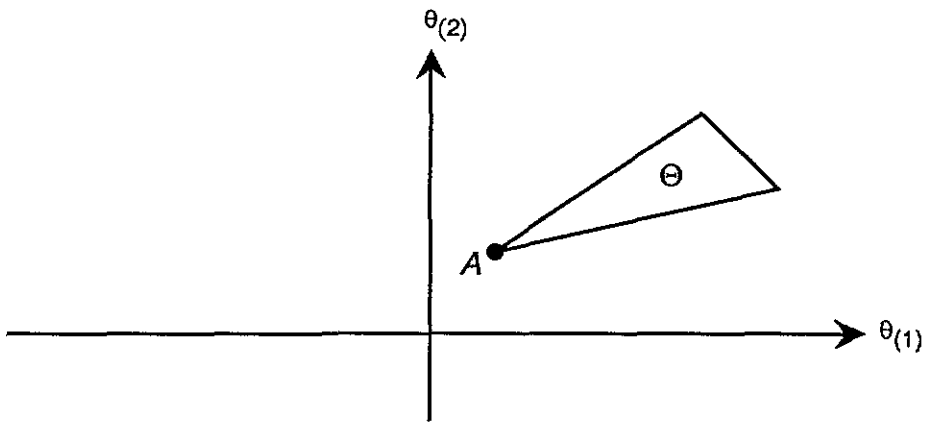

Figure 2. The probability density $p(\theta)$ is uniform on the closed set Θ and consequently is not lower semicontinuous at 0. The point A is absorbing for the Gibbs sampler with blocking $(\theta^{(1)}, \theta^{(2)})$, so if $\theta_0 = A$ convergence will not occur.

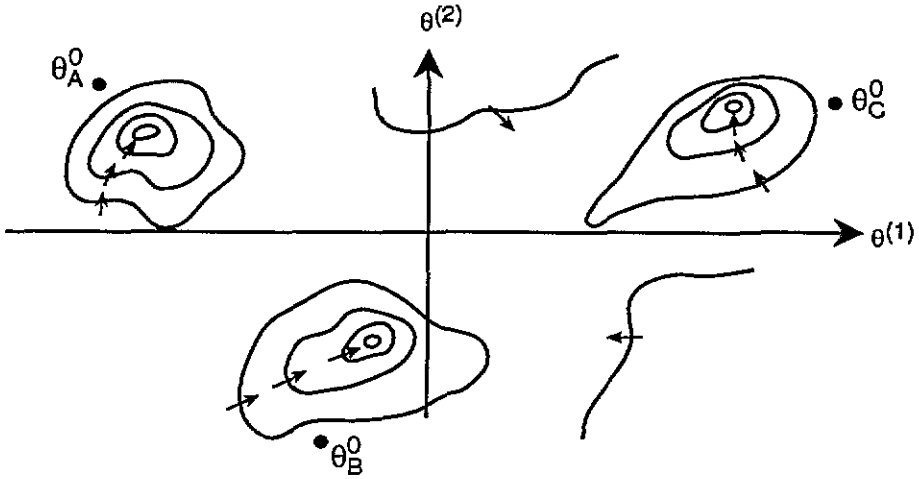

Figure 3. Iso-probability density contours of a multimodal bivariate distribution are shown. (Arrows indicate directions of increased density.) Given sufficiently steep gradients the Gibbs sampler will converge very slowly.