Federal Reserve Bank of Minneapolis
Research Department

Simulation-Based Bayesian Inference for Economic Time Series

John Geweke*

Working Paper 570

August 1996

ABSTRACT

This paper surveys recently developed methods for Bayesir inference and their use in economic time series models. It begins by reviewing aspects of Bayesian inference essential to understanding the implications of the Bayesian paradigm for time series analysis. It next describes the use of posterior simulators to solve otherwise intractable analytical problems. The theory and the computational advances are brought together in setting forth a practical framework for decision-making and forecasting. These developments are illustrated in the context of the vector autoregressions, stochastic volatility models, and models of changing regimes.

*Geweke, Federal Reserve Bank of Minneapolis and University of Minnesota. Partial financial support from NSF grants SES-9210070 and SBR-9514865 are 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

Econometric time series analysis is the discipline of using data to revise beliefs about economic questions, especially questions about the future. These questions have a common structure. Given data resulting from past behavior, and a set of assumptions about economic behavior (or, several sets of competing assumptions), what decision or action should be taken at the present time? The decision or action might involve public economic policy, a private economic decision, or a choice between competing assumptions.

Unfortunately economic questions are rarely laid out so explicitly. Interactions between assumptions and data are studied by a group of individuals, who (following Hildreth, 1963) we may call investigators. The investigators' tasks are complicated by the facts that data sets are constantly being updated, new models are continually being introduced and old ones modified, and the complete constellation of alternative assumptions is never neatly defined. Decisions are made by another group of individuals, who (again, following Hildreth) we may call clients. An ultimate client may be a public or private sector decision-making body, in the case of policy, or the scholarly community, in the case of choices among assumptions. Investigators typically have at best a vague idea who the clients are, and exactly what use clients will wish to make of their results.

This chapter surveys recently developed methods that hold fresh promise for investigators and their clients. These methods are based in the Bayesian paradigm for the use of economic time series, and in recent advances in simulation methods for the implementation of that paradigm. The purpose is to convey these innovations and their significance for time series 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 the Bayesian paradigm for time series analysis, and of posterior simulators for Bayesian econometrics. Section 3 describes these simulators and provides the essential convergence results. The theory and the computational advances are brought together in Section 4, which sets forth a practical framework for Bayesian investigators report results in a way that is immediately useful for decision-making in general and forecasting in particular. Implementation in some specific time series models is taken up in Section 5. The survey in this section is representative, not complete; see the surveys of Koop (1994), Chib and Greenberg (1994b) and Geweke (1995b) for other models. The key points of the paper are reviewed in the concluding section.

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 $\mathbf{Y}_t = \{\mathbf{y}_s\}_{s=1}^{t}$; $\mathbf{Y}_0 = \{\varnothing\}$. A *model* is a corresponding sequence of probability density functions

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

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

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

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

[If the model specifies that the $\mathbf{y}_t$ are independent and identically distributed then $\mathrm{f}_t(\mathbf{y}_t | \mathbf{Y}_{t-1}, \theta) = \mathrm{f}_t(\mathbf{y}_t | \theta)$ and $\mathrm{p}(\mathbf{Y}_T | \theta) = \prod_{t=1}^{T} \mathrm{f}_t(\mathbf{y}_t | \theta)$. 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

$$\mathrm{E}\left[\mathrm{g}(\theta)|\mathbf{Y}_T\right], \tag{2.1.3}$$

in which $\mathrm{g}(\theta)$ 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 $\mathrm{g}(\theta)=\mathrm{L}(a_1,\theta)-\mathrm{L}(a_2,\theta)$ in which $\mathrm{L}(a,\theta)$ is the loss function pertaining to action a, parameter vector θ, and (implicitly, through (2.1.3)) the model itself. A third category is $\mathrm{g}(\theta)=\chi_{\Theta_0}(\theta)$ which arises when a hypothesis restricts θ to a set Θ_0. [Here $\chi(\cdot)$ is the characteristic function $\chi_S(z)=1 \text{ if } z\in S, \quad \chi_S(z)=0 \text{ if } z\notin S$.] Then $\mathrm{E}\left[\mathrm{g}(\theta)|\mathbf{Y}_T\right]=\mathrm{P}\left(\theta\in\Theta_0|\mathbf{Y}_T\right)$. Yet another important category arises from predictive densities, taken up in detail in Section 4.

The specification of the model (2.1.1) is completed with a *prior density* $\mathrm{p}(\theta)$. It may be shown that given (2.1.1) and a density $\mathrm{p}(\mathbf{Y}_T)$ (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

$$\mathrm{P}\left(\theta\in\tilde{\Theta}\right)=\int_{\tilde{\Theta}}\mathrm{p}(\theta)d\theta,\quad \mathrm{P}\left(\mathbf{Y}_T\in\tilde{Y}|\theta\right)=\int_{\tilde{Y}}\prod\nolimits_{t=1}^{T}\mathrm{f}_t\left(\mathbf{y}_t|\mathbf{Y}_{t-1},\theta\right)d\mathbf{Y}_T, \tag{2.1.4}$$

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

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

Thus

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

In the representation (2.1.5), one may substitute for $\mathrm{p}(\theta)$ any function $\mathrm{p}^*(\theta)\propto\mathrm{p}(\theta)$. The function $\mathrm{p}^*(\theta)$ is a *kernel* of the prior density $\mathrm{p}(\theta)$. 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 $\mathbf{s}_T = \mathrm{s}_T(\mathbf{Y}_T)$ is a sufficient statistic in the model (2.1.2) given any of the following equivalent conditions:

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

(2.2.2) $\quad \mathrm{p}(\theta | \mathbf{Y}_T) = \mathrm{p}[\theta | \mathrm{s}_T(\mathbf{Y}_T)] \; \forall \; \theta \in \Theta$ for all realizations $\mathbf{Y}_T$;

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

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

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

If $\theta' = (\theta_1', \theta_2')$ and $\mathrm{g}(\theta) = \mathrm{g}(\theta_1)$ then θ_2 is a *nuisance parameter* for the function of interest $\mathrm{g}(\theta)$. 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 $\mathrm{g}(\theta)$, 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 $\mathrm{L}(\tilde{\omega}, \omega)$ is

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

[Clearly, the estimate $\hat{\omega}$ depends on the complete model (2.1.4) as well as the loss function $\mathrm{L}(\tilde{\omega}, \omega)$. 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* $\mathrm{L}(\tilde{\omega}, \omega) = (\tilde{\omega} - \omega)' \mathbf{Q} (\tilde{\omega} - \omega)$ (where $\mathbf{Q}$ p.d., $\tilde{\omega} \in \mathrm{R}^q$), $\hat{\omega} = \mathrm{E}(\omega | \mathbf{Y}_T)$;

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

given *0/1 loss* $\mathrm{L}(\tilde{\omega},\omega)=1-\chi_{\mathrm{N}_\varepsilon(\tilde{\omega})}(\omega)$ (where $\mathrm{N}_\varepsilon(\tilde{\omega})$ is an ε-neighborhood of $\tilde{\omega}$), as $\varepsilon \to 0$, $\hat{\omega}$ converges to the global mode of $\mathrm{p}(\omega|\mathbf{Y}_T)$ 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 $\int_C \mathrm{p}(\omega|\mathbf{Y}_T)d\omega = 1-\alpha$. 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 $\mathrm{p}(\omega_1|\mathbf{Y}_T) \geq \mathrm{p}(\omega_2|\mathbf{Y}_T)\ \forall\ (\omega_1,\omega_2): \omega_1 \in C, \omega_2 \in \Omega - C$, 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 $\nu = \mathrm{h}(\omega)$, $\mathrm{h}(\cdot)$ is one-to-one, and C is a $100(1-\alpha)\%$ credible set for ω, then $D = \{\nu: \nu = \mathrm{h}(\omega), \omega \in C\}$ 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 $\mathrm{E}[\mathrm{g}(\theta)|\mathbf{Y}_T]$. To this end, it has proven useful to employ classes of prior densities, $\mathrm{p}(\theta|\tau)$ where τ is an indexing parameter, just as it has proven useful to index the conditional density $\mathrm{f}_t(\mathbf{y}_t|\mathbf{Y}_{t-1},\theta)$ by θ.

Suppose that $\mathrm{p}(\mathbf{Y}_T|\theta), \theta \in \Theta$ has sufficient statistic $\{T, \mathrm{s}_T(\mathbf{Y}_T)\}$, where $\mathrm{s}_T(\mathbf{Y}_T)$ 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* $\mathrm{p}(\mathbf{Y}_T|\theta)$ is

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

where

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

and

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

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 $f_t\left(\mathbf{y}_t|\mathbf{Y}_{t-1},\theta\right)$. 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 $p\left(\mathbf{Y}_T|\theta\right)$ and was likely an important reason for the choice of $f_t\left(\mathbf{y}_t|\mathbf{Y}_{t-1},\theta\right)$ in the first place. For example, in the regular *exponential family of distributions*

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

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

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

If $\theta' = \left(\theta_1', \theta_2'\right)$ 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* $p\left(\mathbf{Y}_T|\theta_1,\theta_2^0\right)$ 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* $\mathrm{p}_j^*(\theta)$: i.e., $\int_\Theta \mathrm{p}_j^*(\theta)d\theta < \infty$ and the corresponding prior density is $\mathrm{p}_j(\theta) = \mathrm{p}_j^*(\theta) \big/ \int_\Theta \mathrm{p}_j^*(\theta)d\theta$. Suppose further that $\lim_{j\to\infty}\mathrm{p}_j(\theta) = 0$ and $\lim_{j\to\infty}\mathrm{p}_j^*(\theta) = \mathrm{p}^*(\theta)\ \forall\ \theta \in \Theta$, but that $\int_\Theta \mathrm{p}^*(\theta)d\theta$ is divergent. It is often the case that $\int_\Theta \mathrm{L}(\theta;\mathbf{Y}_T)\mathrm{p}^*(\theta)d\theta$ and $\int_\Theta \mathrm{g}(\theta)\mathrm{L}(\theta;\mathbf{Y}_T)\mathrm{p}^*(\theta)d\theta$ are convergent and furthermore

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

In this case the formal use of the "prior density" $\mathrm{p}^*(\theta)$ has an unambiguous interpretation and provides correct posterior moments. If $\mathrm{p}^*(\theta)$ 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 $\tau^{(j)} = \left(\tau_0^{(j)},\ldots,\tau_m^{(j)}\right) \xrightarrow[j\to\infty]{} \mathbf{0}$ then the limiting posterior distribution is given by (2.4.1) with $\tau_i = 0\ (i = 0,\ldots,m)$. Formal analysis with $\mathrm{p}^*(\theta) = 1$ would have led to the same result.

2.5 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":

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

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 $\mathbf{y}_t$ -- 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 $\left\{p_j, \mathrm{p}_j(\theta_j)\right\}_{j=1}^{J}$ 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 $\mathrm{E}\left[\mathrm{h}(\omega)|\mathbf{Y}_T\right]$ 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

$$\mathrm{E}\left[\mathrm{h}(\omega)|\mathbf{Y}_T\right] = \sum_{j=1}^{J} \mathrm{E}\left[\mathrm{h}(\omega)|\mathbf{Y}_T, M_j\right] \mathrm{P}\left(M_j|\mathbf{Y}_T\right). \tag{2.5.1}$$

From (2.1.5),

$$\mathrm{E}\left[\mathrm{h}(\omega)|\mathbf{Y}_T, M_j\right] = \frac{\int_{\Theta_j} \mathrm{g}(\theta_j) \mathrm{L}_j(\theta_j; \mathbf{Y}_T) \mathrm{p}_j(\theta_j) d\theta_j}{\int_{\Theta_j} \mathrm{L}_j(\theta_j; \mathbf{Y}_T) \mathrm{p}_j(\theta_j) d\theta_j} \tag{2.5.2}$$

with $\mathrm{g}(\theta_j) = \int_{\omega} \mathrm{h}(\omega) \mathrm{p}_j(\omega|\theta_j, \mathbf{Y}_T) d\omega$. There is nothing new in this part of (2.5.1). From Bayes' rule,

$$\begin{aligned} \mathrm{P}\left(M_j|\mathbf{Y}_T\right) &= \mathrm{p}\left(\mathbf{Y}_T|M_j\right)\mathrm{P}\left(M_j\right)/\mathrm{p}(\mathbf{Y}_T) \\ &= p_j \int_{\Theta_j} \mathrm{p}_j\left(\mathbf{Y}_T|\theta_j\right)\mathrm{p}_j(\theta_j) d\theta_j / \mathrm{p}(\mathbf{Y}_T) \\ &\propto p_j \int_{\Theta_j} \mathrm{p}_j\left(\mathbf{Y}_T|\theta_j\right)\mathrm{p}_j(\theta_j) d\theta_j = p_j M_{jT} \end{aligned} \tag{2.5.3}$$

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

$$M_{jT} = \int_{\Theta_j} \mathrm{L}_j(\theta_j; \mathbf{Y}_T) \mathrm{p}_j(\theta_j) d\theta_j . \tag{2.5.4}$$

Expression (2.5.4) must be treated with caution, because the likelihood function typically introduces convenient, model-specific proportionality constants: $\int_{\tilde{\mathrm{Y}}} \mathrm{p}_j\left(\mathbf{Z}_T|\theta_j\right) d\mathbf{Z}_T = 1$ but $\int_{\tilde{\mathrm{Y}}} \mathrm{L}_j\left(\theta_j; \mathbf{Z}_T\right) d\mathbf{Z}_T \neq 1$. Whereas (2.5.2), like (2.1.5), is invariant to arbitrary renormalizations of $\mathrm{p}_j\left(\mathbf{Y}_T|\theta_j\right)$ and $\mathrm{p}_j(\theta_j)$, (2.5.3) 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 (2.5.2) corresponding to each model. Second, obtain the marginalized likelihood M_{jT} from (2.5.3). Finally, obtain the posterior moment using (2.5.1) 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 (2.5.2) and (2.5.4) for those models are known.

2.6 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

$$\frac{\mathrm{P}(M_1|\mathbf{Y}_T)}{\mathrm{P}(M_2|\mathbf{Y}_T)}=\frac{p_1M_{1T}}{p_2M_{2T}}>\frac{\mathrm{L}(1|2)}{\mathrm{L}(2|1)}.$$

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_1M_{1T}/p_2M_{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 $\mathrm{p}_{1j}(\theta_1)$ in Model 1 in which $\lim_{j\to\infty}\mathrm{p}_{1j}(\theta_1)=0\;\forall\;\theta_1\in\Theta_1$. 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 $\lim_{j\to\infty}\mathrm{p}_{1j}(\theta_1)=0\;\forall\;\theta_1\in\Theta_1$ 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 $\mathrm{p}_j(\theta)$. Then the Bayes factor in favor of Model 1 is

$$\text{(2.6.1)}\qquad \frac{M_{1T}}{M_{2T}}=\frac{\int_\Theta \mathrm{g}(\theta)\mathrm{L}(\theta;\mathbf{Y}_T)\mathrm{p}_2(\theta)d\theta}{\int_\Theta \mathrm{L}(\theta;\mathbf{Y}_T)\mathrm{p}_2(\theta)d\theta}$$

with

(2.6.2) $g(\theta) = p_1(\theta)/p_2(\theta)$.

2.7 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

(2.7.1) $p_A(\mathbf{Y}_T|\theta, \psi)\ (\theta \in \Theta, \psi \in \Psi)$

and a prior density for θ conditional on a *hyperparameter* ϕ,

(2.7.2) $p_B(\theta|\phi)\ (\phi \in \Phi)$. The model is completed with a prior density for ϕ and ψ,

(2.7.3) $p_C(\phi, \psi)$.

The full prior density for all parameters and hyperparmeters is

(2.7.4) $p(\theta, \phi, \psi) = p_C(\phi, \psi) p_B(\theta|\phi)$.

There is no fundamental difference between this prior density and the one described in Section 2.4, since

$$p(\theta, \psi) = \int_\Phi p_B(\theta|\phi) p_c(\phi, \psi) d\phi.$$

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 the full posterior density

(2.7.5) $p(\theta, \psi, \phi|\mathbf{Y}_T) \propto p_A(\mathbf{Y}_T|\theta, \psi) p_B(\theta|\phi) p_C(\phi, \psi)$.

A *latent variable model* expresses the likelihood function in two or more steps. In the two-step case the likelihood function may be written

(2.7.6) $p_A(\mathbf{Y}_T|\mathbf{Z}_T^*, \psi)\ \left(\mathbf{Z}_T^* \in \tilde{\mathbf{Z}}, \psi \in \Psi\right)$

where $\mathbf{Z}_T^*$ is a matrix of latent variables. The model for $\mathbf{Z}_T^*$ is

(2.7.7) $p_B(\mathbf{Z}_T^*|\phi)\ (\phi \in \Phi)$

and the prior density for ϕ and ψ is

(2.7.8) $p_C(\phi, \psi)$.

The full prior density for all parameters and unobservable variables is

(2.7.9) $p(\mathbf{Z}_T^*, \psi, \phi) \propto p_A(\mathbf{Y}_T|\mathbf{Z}_T^*, \psi) p_B(\mathbf{Z}_T^*|\phi) p_C(\phi, \psi)$

and the full posterior density is

(2.7.10) $p(\mathbf{Z}_T^*, \psi, \phi|\mathbf{Y}_T) \propto p_A(\mathbf{Y}_T|\mathbf{Z}_T^*, \psi) p_B(\mathbf{Z}_T^*|\phi) p_C(\phi, \psi)$.

Comparing (2.7.1)-(2.7.5) with (2.7.6)-(2.7.10), 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. With appropriate marginalization of (2.7.10) one may obtain $p(\mathbf{Z}_T^*|\mathbf{Y}_T)$, which fully reflects uncertainty about the parameters. If one is interested only in ψ, or in ψ and ϕ, these distributions may also be obtained by marginalization of (2.7.10). In the latter case the matrix of latent variables $\mathbf{Z}_T^*$ is a group of nuisance parameters, which are treated here as described in Section 2.3.3. Marginalization requires integration, which is generally impossible analytically. If the problem is approached using simulation methods, then marginalization simply amounts to discarding the nuisance parameters.

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

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

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 $\varepsilon_t \sim \mathrm{N}(0,1)$ independent of ω_t. The equivalent hierarchical prior formulation is the p.d.f. specification

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

and the conditional prior distribution

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

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

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

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

(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 $\bar{g}_M = M^{-1}\sum_{m=1}^{M} g(\theta_m)$.

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 2.7, 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 $g(\theta)$ cannot be evaluated in closed form, but unbiased simulators are available, because $g(\theta)$ may then be replaced by its simulator. Leading examples are forecasting and discrete choice 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. $\mathrm{F}^{-1}(p) = \{c: \mathrm{P}(x \le c) = p\}$ 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 $\mathrm{F}^{-1}(u_i)$, 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 $\mathrm{f}(x) = \lambda\exp(-\lambda x), x \ge 0$. Then $\mathrm{F}(x) = 1 - \exp(-\lambda x), \mathrm{F}^{-1}(p) = -\log(1-p)/\lambda$, 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. $\mathrm{f}(x)$ and support C. Let g be the p.d.f. of a different continuous random variable z with p.d.f. $\mathrm{g}(z)$ which has a distribution from which it is possible to draw i.i.d. random variables and for which

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

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 $u > \mathrm{f}(z)/[a\,\mathrm{g}(z)]$, go to (a);

(d) $x_i = z$.

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

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

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

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

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

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 $g(\mathbf{x};\alpha)$ 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, $x \sim \mathrm{t}(0,1;2)$ can be generated in the obvious way from three independent standard normals; if $x \sim \mathrm{B}(m,n)$ 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.

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

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 $\mathrm{N}(\mu,\Sigma)$ is based on the familiar decomposition

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

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

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

for brevity, $\mathbf{A} \sim \mathrm{W}(\Sigma,n-1)$. 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 $\mathbf{Q} \sim \mathrm{W}(\mathbf{I}_m,n-1)$. Then $\mathbf{L}\mathbf{Q}\mathbf{L}' \sim \mathrm{W}(\Sigma,n-1)$ (Anderson, 1984, pp. 254-255). Furthermore $\mathbf{Q}$ has representation

$$\mathbf{Q} = \mathbf{U}\mathbf{U}' \qquad u_{ij} = 0\,(i < j < m)$$
$$u_{ij} \sim \mathrm{N}(0,1) \qquad u_{ii} \sim \chi^2(n-i)$$

$(i = 1,\ldots,m)$, 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. $\mathrm{p}(\theta|\mathbf{Y}_T)$. Denoting $\overline{g} = \mathrm{E}[\mathrm{g}(\theta)|\mathbf{Y}_T]$ and $\overline{g}_m = M^{-1}\sum_{m=1}^{M}\mathrm{g}(\theta_m)$, by the strong law of large numbers

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

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

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

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

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

[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) $\qquad (\overline{vs}^2/\sigma^2)|(\mathbf{y},\mathbf{X}) \sim \chi^2(\overline{v})$,

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

where $\overline{v} = \underline{v} + T - k$, $\overline{s}^2 = \overline{v}^{-1}\left[\underline{vs}^2 + (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})\right]$, $\overline{\mathbf{H}}_\beta = \underline{\mathbf{H}}_\beta + (\mathbf{X}'\mathbf{X})^{-1}$, $\overline{\beta} = \overline{\mathbf{H}}_\beta^{-1}\left[\underline{\mathbf{H}}_\beta\underline{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{b}\right]$ 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 $\text{var}\left[g(\theta)|\mathbf{Y}_T\right] < \infty$ 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 $\bar{g}_m$ as an approximation of $\bar{g}$.

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 $\left\{\theta_{mi}\right\}_{i=1}^{2}{}_{m=1}^{M/2}$ with the θ_{mi} identically distributed and the only mutual dependence being that arising between θ_{m1} and θ_{m2}. Let $\bar{g}_M = M^{-1}\sum_{m=1}^{M/2}\sum_{i=1}^{2} g(\theta_{mi})$ and suppose $\text{var}\left[g(\theta)|\mathbf{Y}_T\right] < \infty$. Then

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

As long as $\text{cov}\left[g(\theta_{m1}), g(\theta_{m2})\right] < 0$, 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 $\text{var}\left\{\frac{1}{2}\left[g(\theta_{m1}) + g(\theta_{m2})\right]\right\} = 0$, and variance reduction would be complete. At the other extreme, if $g(\theta)$ is also symmetric about μ, then $\text{var}\left\{\frac{1}{2}\left[g(\theta_{m1}) + g(\theta_{m2})\right]\right\} = \text{var}\left[g(\theta)\right]$: 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 $g(\theta_{m1}) - \bar{g}$ and $g(\theta_{m2}) - \bar{g}$ 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 $\bar{h} = \tilde{E}[h(\nu)|\mathbf{Y}_T] = \int_N h(\nu)\tilde{p}(\nu|\mathbf{Y}_T)d\nu$ exactly. Suppose that the sequence $\{\theta_m, \nu_m\}$ can be drawn, $\{\theta_m\}$ an i.i.d. sequence from the original posterior distribution and $\{\nu_m\}$ an i.i.d. sequence from the approximating distribution, but with θ_m and ν_m constructed from the same underlying random numbers so that $g(\theta_m)$ and $h(\nu_m)$ are correlated. Let $\bar{g}_M = M^{-1}\sum_{m=1}^{M} g(\theta_m)$ and $\bar{h}_M = M^{-1}\sum_{m=1}^{M} h(\nu_m)$, 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

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

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 $\mathrm{g}\left(\theta_{(1)m}\right)$ 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. $\mathrm{p}\left(\theta|\mathbf{Y}_T\right)$. Following an initial transient or *burn-in* phase, simulated values from the chain form a basis for approximating $\mathrm{E}\left[\mathrm{g}(\theta)|\mathbf{Y}_T\right]$. What is required is to construct an appropriate algorithm and verify that its invariant distribution is unique, with p.d.f. $\mathrm{p}\left(\theta|\mathbf{Y}_T\right)$.

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

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

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

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 $\mathrm{p}(\theta | \mathbf{Y}_T)$ and the blocking. Since $\theta_0 \sim \mathrm{p}(\theta | \mathbf{Y}_T)$, $\left(\theta_1^{(1)}, \ldots, \theta_1^{(j-1)}, \theta_1^{(j)}, \theta_0^{(j+1)}, \ldots, \theta_0^{(B)}\right)$ $\sim \mathrm{p}(\theta | \mathbf{Y}_T)$ at each step in (3.4.1) by definition of the conditional density. In particular, $\theta_1 \sim \mathrm{p}(\theta | \mathbf{Y}_T)$.

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

$$
(3.4.2) \qquad \mathrm{K}_G\left(\theta_j, \theta_{j+1}\right) = \prod\nolimits_{b=1}^{B} \mathrm{p}\left[\theta_{j+1}^{(b)} \,\middle|\, \theta_j^{(a)}(a > b),\, \theta_{j+1}^{(a)}(a < b), \mathbf{Y}_T\right].
$$

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

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 (1994a) 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 $\bar{g}_M = M^{-1}\sum_{m=1}^{M} \mathrm{g}(\theta_m) \to \mathrm{E}[\mathrm{g}(\theta)|\mathbf{Y}_T]$.

Gibbs sampler convergence condition 1 (after Tierney, 1994). For every point $\theta^* \in \Theta$ and every $\Theta_1 \subseteq \Theta$ with the property $\mathrm{P}(\theta \in \Theta_1|\mathbf{Y}_T) > 0$, it is the case that $\mathrm{P}_G(\theta_{j+1} \in \Theta_1|\theta_j = \theta^*, \mathbf{Y}_T) > 0$, where $\mathrm{P}_G(\cdot)$ is the probability measure induced by the transition kernel (3.4.2).

Gibbs sampler convergence condition 2 (after Roberts and Smith, 1994). The density $\mathrm{p}(\theta|\mathbf{Y}_T)$ is lower semicontinuous at 0, $\int_{\Theta^{(b)}} \mathrm{p}(\theta|\mathbf{Y}_T)d\theta^{(b)}$ is locally bounded $(b = 1,\ldots,B)$, and Θ is connected. [A function $\mathrm{h}(\mathbf{x})$ is lower semicontinuous at 0 if, for all $\mathbf{x}$ with $\mathrm{h}(\mathbf{x}) > 0$, there exists an open neighborhood $N_{\mathbf{x}} \supset \mathbf{x}$ and $\varepsilon > 0$ such that for all $\mathbf{y} \in N_{\mathbf{x}}$, $\mathrm{h}(\mathbf{y}) \geq \varepsilon > 0$. 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 $\mathrm{q}(\theta_m, \theta^*)$ and a starting value θ_0. The random vector θ^* generated from $\mathrm{q}(\theta_m, \theta^*)$ is considered as a candidate value for θ_{m+1}. The algorithm actually sets $\theta_{m+1} = \theta^*$ with probability

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

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

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

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

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

where $w(\theta) = p(\theta|\mathbf{Y}_T)/j(\theta)$. 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, 1994b.) 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

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

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 $M^{-1}\sum_{m=1}^{M} g(\theta_m) \rightarrow E[g(\theta)|\mathbf{Y}_T]$:

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 (1994b) and Mengersen and Tweedie (1993)). For every $\theta \in \Theta$, $p(\theta|\mathbf{Y}_T) > 0$, and for all pairs $(\theta_j, \theta_{j+1}) \in \Theta \times \Theta$, $p(\theta_j|\mathbf{Y}_T)$ 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. Bayesian investigation and communication

The elements of the formal problem addressed by a Bayesian investigator are summarized in Section 2.5, which provides the point of departure here. The essentials include a collection of complete models $(j = 1,\ldots,J)$, each of which describes the joint distribution of observed data $\mathbf{Y}_T$ and vector of interest ω, conditional on a vector of parameters θ_j:

$$\mathrm{p}_j\left(\mathbf{Y}_T,\omega|\theta_j\right) = \mathrm{p}_j\left(\omega|\mathbf{Y}_T,\theta_j\right)\mathrm{p}_j\left(\mathbf{Y}_T|\theta_j\right).$$

Each model is completed with the specification of the prior density $\mathrm{p}_j(\theta_j)$ for its parameters, and the collection is completed with the prior model probabilities $p_j(j=1,\ldots,J)$. Within each model, $\mathrm{p}_j(\mathbf{Y}_T|\theta_j)=\prod_{t=1}^{T}\mathrm{f}_{jt}(\mathbf{y}_t|\mathbf{Y}_{t-1},\theta_j)$. Section 4.1 describes some insights into model comparison, predictive densities, and forecasting that arise from this elementary relation.

Since it is required only to specify $\mathrm{p}_j(\omega|\mathbf{Y}_T,\theta_j)$, the vector of interest ω is quite general. For example, it may consist of some future data, $\omega=(\mathbf{y}_{T+1},\ldots,\mathbf{y}_{T+f})'$. In this case,

$$\mathrm{p}_j(\omega|\mathbf{Y}_T,\theta_j)=\prod_{t=1}^{T+f}\mathrm{f}_{jt}(\mathbf{y}_t|\mathbf{Y}_{t-1},\theta_j)\Big/\prod_{t=1}^{T}\mathrm{f}_{jt}(\mathbf{y}_t|\mathbf{Y}_{t-1},\theta_j).$$

Closed form expressions may or may not be readily computed, but it is essentially always the case that simulations from $\mathrm{p}_j(\omega|\mathbf{Y}_T,\theta_j)$ can be carried out in straightforward fashion, and as will be seen this is typically all that is required. In other cases ω may be a vector of latent variables, and then greater ingenuity may be required to simulate from $\mathrm{p}_j(\omega|\mathbf{Y}_T,\theta_j)$: the problem is one of signal extraction, which for many models has been thoroughly investigated.

In a *closed investigation* all of the elements of this problem are completely specified, including models, priors, and vectors of interest. The closed investigation is completed with the specification of a reporting objective in the form of a mapping from the distribution $\omega|\mathbf{Y}_T$ to a real number, $\mathrm{R}[\mathrm{p}(\omega|\mathbf{Y}_T)]$. Very often this mapping may be expressed $\mathrm{E}[\mathrm{h}(\omega)|\mathbf{Y}_T]$, for a known function $\mathrm{h}(\omega)$. Examples include minimum mean square error forecasting, probabilities of turning points, and the evaluation of relative loss for alternative decisions. In other instances quantiles of the distribution are required, as in minimum absolute error forecasting or the reporting of an interquartile range. For simplicity of notation we proceed as if the problem is always of the form $\mathrm{E}[\mathrm{h}(\omega)|\mathbf{Y}_T]$.

Section 4.2 takes up the common elements of forecasting and signal extraction problems in a closed investigation with a single model. Their solution is general, simple and elegant -- especially in comparison with non-Bayesian methods. Section 4.3 takes up the question of why this is so.

In a closed investigation with multiple models the investigator requires

$$\mathrm{p}_j(\omega|\mathbf{Y}_T)=\sum_{j=1}^{J}\mathrm{P}(M_j|\mathbf{Y}_T)\,\mathrm{p}_j(\omega|\mathbf{Y}_T).$$

The essential incremental task is the evaluation of the marginalized likelihood, as explained in Section 2.5. Some methods of evaluating the marginalized likelihood are taken up in Section 4.4.

The closed investigation has two distinguishing features. First, the ultimate consumer -- who, following Hildreth (1963) may be called the *client* -- has provided a complete formal specification of the problem. Second, the investigator's report $\mathrm{E}\left[\mathrm{h}(\omega)|\mathbf{Y}_T\right]$ is formally useless if any aspect of the problem specification is changed, and typically is not then very useful in an informal way either. In an *open investigation* the investigator is missing one or more elements of the problem specification: for example, the investigator may not know the client's priors or vectors of interest, and may not know all of the models the client entertains and their associated prior probabilities. This situation is much more typical of the conditions under which investigations are carried out. Given the methods set forth in this chapter the investigator can in fact do a great deal in this situation, as described in Section 4.5.

4.1 The predictive decomposition

Suppose that $\mathbf{Y}_u = \{y_s\}_{s=1}^{u}$ is available, and a single model has been completely specified. Then the *predictive density* for observations $u+1$ through t is

(4.1.1) $$\mathrm{p}\left(y_{u+1},\ldots,y_t|\mathbf{Y}_u\right) = \int_{\Theta} \mathrm{p}\left(\theta|\mathbf{Y}_u\right) \prod_{s=u+1}^{t} \mathrm{f}_s\left(\mathbf{y}_s|\mathbf{Y}_{s-1},\theta\right) d\theta,$$

where

(4.1.2) $$\mathrm{p}\left(\theta|\mathbf{Y}_t\right) = \mathrm{p}(\theta) \prod_{s=1}^{u} \mathrm{f}_s\left(\mathbf{y}_s|\mathbf{Y}_{s-1},\theta\right) \Big/ \int_{\Theta_j} \prod_{s=1}^{u} \mathrm{f}_s\left(\mathbf{y}_s|\mathbf{Y}_{s-1},\theta\right) d\theta .$$

In this expression, $\left(\mathbf{y}_{u+1},\ldots,\mathbf{y}_t\right)'$ is a random vector; the practical mechanics of working with its distribution are taken up in Section 4.2.

Once the observations $\mathbf{y}_{u+1},\ldots,\mathbf{y}_t$ are known, the *predictive likelihood* for observations $u+1$ through t is

(4.1.3) $$\hat{p}_u^t \equiv \int_{\Theta_j} \mathrm{p}\left(\theta|\mathbf{Y}_u\right) \prod_{s=u+1}^{t} \mathrm{f}_s\left(\mathbf{y}_s|\mathbf{Y}_{s-1},\theta\right) d\theta .$$

The right hand sides of expressions (4.1.1) and (4.1.3) are only formally identical: $\left(\mathbf{y}_{u+1},\ldots,\mathbf{y}_t\right)'$ is a random vector in the former, and is fixed in the latter. The predictive likelihood is the probability density assigned to the observed $\left(\mathbf{y}_{u+1},\ldots,\mathbf{y}_t\right)'$ by the posterior based on observations $1,\ldots,u$. It is a measure of the out-of-sample forecasting performance of the model -- one that fully accounts for parameter uncertainty.

Since $\mathbf{Y}_0 = \{\varnothing\}$, the marginalized likelihood is

$$\hat{p}_0^T = \int_{\Theta} \mathrm{p}(\theta) \prod_{s=1}^{T} \mathrm{f}_s\left(\mathbf{y}_s|\mathbf{Y}_{s-1},\theta\right) d\theta = M_T .$$

Substituting (4.1.2) in (4.1.1),

$$\hat{p}_u^t = \int_\Theta \left\{ \frac{\mathrm{p}(\theta)\prod_{s=1}^{u} \mathrm{f}_s(\mathbf{y}_s|\mathbf{Y}_{s-1},\theta)}{\int_{\Theta_j} \mathrm{p}(\theta)\prod_{s=1}^{u} \mathrm{f}_s(\mathbf{y}_s|\mathbf{Y}_{s-1},\theta)d\theta} \right\} \prod_{s=u+1}^{t} \mathrm{f}_s(\mathbf{y}_s|\mathbf{Y}_{s-1},\theta)d\theta$$

$$= \frac{\int_\Theta \mathrm{p}(\theta)\prod_{s=1}^{t} \mathrm{f}_s(\mathbf{y}_s|\mathbf{Y}_{s-1},\theta)d\theta}{\int_{\Theta_j} \mathrm{p}(\theta)\prod_{s=1}^{u} \mathrm{f}_s(\mathbf{y}_s|\mathbf{Y}_{s-1},\theta)d\theta_j} = \frac{M_t}{M_u}.$$

Hence for any $0 \le u = s_0 < s_1 < \ldots < s_q = t$,

(4.1.4) $$\hat{p}_u^t = \frac{M_{s_1}}{M_{s_0}} \cdot \frac{M_{s_2}}{M_{s_1}} \cdot \ldots \cdot \frac{M_{s_q}}{M_{s_{q-1}}} = \prod_{\tau=1}^{q} \hat{p}_{s_{\tau-1}}^{s_\tau}.$$

This decomposition has several significant implications for time series analysis.

First, specific choices of the s_τ yield

(4.1.5) $$M_T = \hat{p}_0^T = \prod_{s=1}^{T} \hat{p}_{s-1}^{s} = \prod_{s=1}^{T/r} \hat{p}_{r(s-1)}^{rs},$$

where T is an integer multiple of r. This identity links the model's marginalized likelihood with its out-of-sample forecasting record as embodied in the predictive likelihood. Recall that the data bear on model choice only through the marginalized likelihood (Section 2.6). Therefore, (4.1.5) provides a well-defined sense in which model choice is equivalent to the comparison of out-of-sample forecast performance. In a symmetric model choice problem (balanced loss function and equal prior probabilities) the chosen model will be the one with the best out of sample forecasting record as indicated by the right hand side of (4.1.5). The forecasting record can be stated in terms of all one-step-ahead forecasts, or in terms of non-overlapping forecasts of r successive realizations. It cannot be expressed only in terms of r-step-ahead forecasts. This accords well with the common experience in time series analysis, that preferred models provide superior one-step-ahead forecasts, but not necessarily superior r-step-ahead forecasts for $r > 1$.

The decomposition (4.1.4) also provides a general method of computing M_T. If M_T is cast directly in the form of (2.1.3),

$$M_T = \mathrm{E}[\mathrm{g}(\theta)|\mathbf{Y}_0] = \int_\Theta \mathrm{p}(\theta)\mathrm{g}(\theta)d\theta, \quad \mathrm{g}(\theta) = \prod_{s=1}^{T} \mathrm{f}_s(y_s|\mathbf{Y}_{s-1},\theta),$$

the result is not useful for computation: the prior distribution is typically much more diffuse than the likelihood function $\prod_{s=1}^{T} \mathrm{f}_s(y_s|\mathbf{Y}_{s-1},\theta)$, so that draws from the prior are very inefficient, being almost always far removed from the main support of the likelihood function. On the other hand, since

$$\hat{p}_u^t = \mathrm{E}[\mathrm{g}(\theta)|\mathbf{Y}_u] = \int_\Theta \mathrm{p}(\theta|\mathbf{Y}_u)\mathrm{g}(\theta)d\theta, \quad \mathrm{g}(\theta) = \prod_{s=u+1}^{t} \mathrm{f}_s(\mathbf{y}_s|\mathbf{Y}_{s-1},\theta),$$

a simulation method may provide a computationally efficient approximation of $\hat{p}_u^t$. Then, one may appeal to (4.1.5) to approximate M_T. (For further details see Geweke (1994).)

A third use of the decomposition is in reporting. For all $v{:}u \le v < t$,

$$\hat{p}_v^T = \frac{M_t}{M_v} = \frac{M_t/M_u}{M_v/M_u} = \frac{\hat{p}_u^t}{\hat{p}_u^v}.$$

Consequently a plot of $\log(\hat{p}_u^s)$ for $s = u+1,\ldots,t$ is visually revealing of predictive likelihoods for all subintervals.

The decomposition also provides a useful diagnostic in the comparison of models. Introduce an additional subscript "j" or "k" to denote alternative models and define the predictive Bayes factor $\hat{B}_{j|k,u}^t = \hat{p}_{ju}^t / \hat{p}_{ku}^t$; $\hat{B}_{j|k,0}^t = \hat{p}_{j0}^t / \hat{p}_{k0}^t = M_{jT}/M_{kT}$ is the Bayes factor defined in Section 2.6. From (4.1.5),

$$\hat{B}_{j|k,u}^t = \hat{p}_{ju}^t / \hat{p}_{ku}^t = \prod\nolimits_{\tau=1}^q \hat{p}_{js_{\tau-1}}^{s_\tau} \Big/ \prod\nolimits_{\tau=1}^q \hat{p}_{ks_{\tau-1}}^{s_\tau} = \prod\nolimits_{\tau=1}^q \hat{B}_{j|k,s_{\tau-1}}^{s_\tau}.$$

By considering characteristics of observations for which $\hat{B}_{j|k,s_{\tau-1}}^{s_\tau}$ is relatively larger or smaller, it is often possible to get a deeper understanding of model suitability and new models may be suggested.

4.2 Forecasting and signal extraction

The general forecasting and signal extraction problem can be stated as follows. Given a data set $\mathbf{Y}_T$ and a collection of models M_j $(j = 1,\ldots,J)$, corresponding to each model j there is a parameter vector θ_j, a data density $\mathrm{p}(\cdot|\theta_j)$, a prior density $\mathrm{p}(\theta_j)$, and a prior model probability p_j. There is also a common vector of interest $\omega \in \Omega$, and for each model a specified density $\mathrm{p}_j(\omega|\theta_j, \mathbf{Y}_T)$. The problem is to choose a Bayes action $\mathbf{a} \in A$ to minimize

$$\mathrm{E}[\mathrm{C}(\omega,\mathbf{a})|\mathbf{Y}_T] = \int_\Omega \mathrm{C}(\omega,\mathbf{a})\,\mathrm{p}(\omega|\mathbf{Y}_T)\,d\omega$$

$$= \int_\Omega \mathrm{C}(\omega,\mathbf{a}) \sum\nolimits_{j=1}^J \left[\int_{\Theta_j} \mathrm{p}_j(\omega|\theta_j, \mathbf{Y}_T)\,\mathrm{p}_j(\theta_j|\mathbf{Y}_T)\,d\theta_j\right] \mathrm{P}(M_j|\mathbf{Y}_T)\,d\omega,$$

where C is a loss (or "cost") function, $\mathrm{P}(M_j|\mathbf{Y}_T)$ is given by (2.5.3), and $\mathrm{p}_j(\theta_j|\mathbf{Y}_T)$ is given by Bayes theorem for each model.

In many instances the solution of the minimization problem can be expressed $\mathrm{E}[\mathrm{h}(\omega)|\mathbf{Y}_T]$ for a known function $\mathrm{h}(\cdot)$: the leading example is quadratic loss, discussed in Section 2.3. If $\mathrm{g}(\theta) = \int_\Omega \mathrm{h}(\omega)\,\mathrm{p}(\omega|\mathbf{Y}_T, \theta)\,d\omega$ can be evaluated in closed form, then the forecasting and signal extraction problem is a special case of the problem set forth in Section 2.1, to which the simulation methods of Section 3 are directly addressed. This is rarely the case.

In most instances, however, the methods of Section 3 can be adapted to the approximation of $E[h(\omega)|\mathbf{Y}_T]$ through an auxiliary simulation. To cast this method as a special case of the general treatment in Section 3, let $\tilde{\theta}' = (\omega', \theta')$, $\tilde{\theta} \in \tilde{\Theta} = \Omega \times \Theta$, and $g(\tilde{\theta}) = h(\omega)$. Then

$$E[h(\omega)|\mathbf{Y}_T] = \int_{\tilde{\Theta}} g(\tilde{\theta}) p(\tilde{\theta}|\mathbf{Y}_T) d\tilde{\theta}.$$

To draw from $p(\tilde{\theta}|\mathbf{Y}_T)$, first draw θ using an appropriate simulator, and then draw ω from $p(\omega|\theta, \mathbf{Y}_T)$. The second step is generally easy: in forecasting, it amounts to simulation of the model with known parameter values; in signal extraction, it typically involves the appropriate conditional distribution, again with known parameter values. If the convergence conditions of Section 3 are satisfied, and if $E[h(\omega)|\mathbf{Y}_T]$ exists, then

$$\sum_{m=1}^{M} w(\theta_m) h(\omega_m) \Big/ \sum_{m=1}^{M} w(\theta_m) \rightarrow E[h(\omega)|\mathbf{Y}_T]$$

where $w(\theta_m)$ is associated with draw m from $p(\theta|\mathbf{Y}_T)$, and $\omega_m \sim p(\omega|\theta_m, \mathbf{Y}_T)$.

The class of forecasting and signal extraction problems that lead to the approximation of posterior moments of ω is interesting but not exhaustive. It includes minimum mean square error forecasting and signal extraction, taking one of a finite number of Bayes actions, and interval forecasting of the form $P(\omega \in \Omega^* | \mathbf{Y}_T)$. However, solutions of quantile or 0/1 loss forecasting or signal extraction problems, and the formation of credible sets, cannot be expressed as posterior moments.

The set of forecasting and signal extraction problems that can be solved using posterior simulation methods can be widened to include most continuous loss functions using some results of Shao (1989). Shao obtained results for importance sampling algorithms. It is reasonable to conjecture that his results extend to MCMC methods, but this has not yet been shown, to the author's knowledge. Let $r(\mathbf{a}) = \int_{\Omega} C(\omega, \mathbf{a}) p(\omega|\mathbf{Y}_T) d\omega$ denote expected loss corresponding to action $\mathbf{a}$, and let $\mathbf{a}^* = \arg\min_{\mathbf{a} \in A} r(\mathbf{a})$ denote the (possibly set-valued) solution. Correspondingly from the posterior simulator denote

$$r_m(\mathbf{a}) = \sum_{m=1}^{M} C(\omega_m, \mathbf{a}) w(\theta_m) \Big/ \sum_{m=1}^{M} w(\theta_m)$$

and $\mathbf{a}_m = \arg\min_{\mathbf{a} \in A} r_m(\mathbf{a})$. Shao (1989) sets forth two sets of conditions under which $r(\mathbf{a}_m) \rightarrow r(\mathbf{a}^*)$, $r_m(\mathbf{a}_m) \rightarrow r(\mathbf{a}^*)$, and $\mathbf{a}_m \rightarrow \mathbf{a}^*$ if $\mathbf{a}^*$ is unique.

Optimal action convergence conditions 1.

(1) The action space A is compact.

(2) The loss function $C(\omega, \mathbf{a})$ is continuous in $\mathbf{a}$ for all $\omega \in \Omega$.

(3) There exists a measurable function $\mathrm{M}(\omega)$ with the properties $\sup_{\mathbf{a}\in A} \mathrm{C}(\omega,\mathbf{a}) \leq \mathrm{M}(\omega)$ and $\int_{\Omega} \mathrm{M}(\omega)\mathrm{p}(\omega|\mathbf{Y}_T)d\omega < \infty$.

Optimal action convergence conditions 2.

(1) A is convex and $\mathrm{C}(\omega,\mathbf{a})$ is a convex continuous function of $\mathbf{a}$ for all $\omega \in \Omega$.

(2) For all $\mathbf{a} \in A$, $\mathrm{r}(\mathbf{a}) < \infty$, and there exists $\mathbf{a}^*$: $\mathrm{r}(\mathbf{a}^*) = \min_{\mathbf{a}\in A} \mathrm{r}(\mathbf{a})$.

(3) Let A' denote the closure of A and $\mathrm{B}(\mathrm{c}) = \left\{\mathbf{a} \in A' : \left\|\mathbf{a} - \mathbf{a}^*\right\| = c\right\}$; there is nonempty $\mathrm{B}(c)$ such that $\inf_{\mathbf{a}\in \mathrm{B}(c)} \mathrm{r}(\mathbf{a}) > \mathrm{r}(\mathbf{a}^*)$.

(4) Let $\mathrm{N}(\mathrm{c}) = \left\{\mathbf{a} \in A' : \left\|\mathbf{a} - \mathbf{a}^*\right\| \leq c\right\}$; there exists a measurable function $\mathrm{M}(\omega)$ such that $\sup_{\mathbf{a}\in \mathrm{N}(c)} \mathrm{C}(\omega,\mathbf{a}) \leq \mathrm{M}(\omega)$ and $\int_{\Omega} \mathrm{M}(\omega)\mathrm{p}(\omega|\mathbf{Y}_T)d\omega < \infty$.

The second set of conditions admits the quantile loss function. A consequent corollary is that quantiles can be approximated consistently based on posterior simulator output. If the loss function is twice differentiable and it is convenient to evaluate $\partial \mathrm{r}_m / \partial \mathbf{a}$ and $\partial^2 \mathrm{r}_m / \partial \mathbf{a} \partial \mathbf{a}'$, then $\mathbf{a}_m$ can be computed by standard optimization algorithms, subject to the usual caveats about the local shape of the objective function.

4.3 Bayesian and non-Bayesian approaches

In the classical non-Bayesian approach to forecasting and signal extraction an action $\mathbf{a} = \mathrm{a}(\mathrm{Y}_T)$ is taken to minimize

(4.3.1) $$\mathrm{E}[\mathrm{C}(\omega,\mathbf{a})|\theta] = \mathrm{E}\left\{\mathrm{C}[\omega(\mathbf{Y}_T),\mathrm{a}(\mathbf{Y}_T)] | \theta\right\}.$$

In (4.3.1) $\mathbf{Y}_T$ is a set of random vectors whose distribution is indicated by the data density (2.1.2) and an assumed parameter vector θ. (If more than one model is being considered then the conditioning set includes one particular model as well as the parameter vector.) In the classical approach the action $\mathrm{a}(\mathrm{Y}_T)$ minimizes loss on average *over all realizations conditional on a specified model.* In the Bayesian approach the action minimizes loss on average *over all model specifications under consideration conditional on the observed data.* This contrast between *ex ante* and *ex post* approaches is the philosophical heart of the contrast between Bayesian and non-Bayesian methods; e.g., see Berger and Wolpert (1988) and Poirier (1988, 1995).

There are two hurdles that must be overcome in implementing the classical approach to forecasting and signal extraction. Each has spawned a substantial literature.

The first hurdle consists of the technical problems inherent in the minimization of the expression (4.3.1) as stated, i.e., assuming a value for θ. A seminal contribution is

Granger (1969), which obtains analytical results for multistep-ahead forecasts of Gaussian processes. Recent extensions of Granger's approach include Weiss and Anderson (1984) and Weiss (1991) for model selection and estimation, Diebold and Mariano (1994) for forecast evaluation, and Chirstoffersen and Diebold (1995) for conditional heteroscedasticity.

The latter paper proposes a numerical approach to multistep-ahead forecasting,

$$a = \hat{y}_{T+h} = \mathrm{G}\left(\mu_{T+h|T}, \sigma^2_{T+h|T}\right)$$

where $\mu_{T+h|T}$ and $\sigma^2_{T+h|T}$ are the first two conditional moments and G is twice continuously differentiable. A second order Taylor series approximation to G leads to an expression with six unknown coefficients, which are determined by means of a long simulation of $\{y_t\}$ and evaluation of the loss function. This approach requires three levels of approximation: limitation of the conditional distribution to its first two moments; a quadratic approximation to the unknown function of these moments; and the simulation error in approximating the quadratic function. In addition, $\mu_{T+h|T}$ and $\sigma^2_{T+h|T}$ must be determined analytically.

Clearly the approach of Christoffersen and Diebold (1995) can be extended to overcome all of these difficulties except simulation noise (which can be made arbitrarily small with sufficient computing). Since the forecast a is an unknown function of $\{y_{t-s}, s \geq 0\}$, the entire literature on nonparametric minimization provides a good approach. (For technical essentials see Amemiya (1985), and for an application whose essentials are similar to what is being proposed here see Smith (1991).) Thus, the first hurdle in implementing the classical approach to forecasting and signal extraction is entirely technical.

The second hurdle arises from the fact that θ is not, in fact known. This difficulty is fundamental, not merely technical, for the minimization of (4.3.1) is conditional on fixed θ. In all but a handful of trivial problems -- e.g., one-step-ahead minimum mean square error forecasting in a Gaussian first order autoregression -- the unknown parameter vector θ remains in the solution. As a practical matter, θ can be replaced by an estimator $\hat{\theta}$ with desirable asymptotic properties, but good results typically require modification based on an expansion of the distribution and the loss function at hand.

To highlight the fact that conditioning on θ is the fundamental difficulty, consider the modification of the classical problem (4.3.1) in which **a** minimizes

$$\mathrm{E}\left[\mathrm{C}(\omega, \mathbf{a}) \middle| \theta, \mathbf{Y}_T\right].$$

In this problem the first technical hurdle vanishes. Conditional on the data set $\mathbf{Y}_T$, it is no longer necessary to determine the full mapping from all possible $\mathbf{Y}_T$ to the optimal a. Simulation of ω conditional on θ and $\mathbf{Y}_T$ can be employed to find a satisfactory numerical approximation of a, directly. The problem of uncertainty about θ -- the second hurdle -- still remains.

The technical difficulties in the classical approach to forecasting and signal extraction reflect the conditioning on a true model, and the minimization of loss averaged over all possible realizations. Problems arise because this conditioning does not reflect the information the investigator brings to the situation. The Bayesian approach conditions on the data in hand, and loss is minimized conditional on all models under consideration and the corresponding possible values of the signal or future data ω. This reflects the investigator's situation, and simulation methods provide the direct solution of the problem.

4.4 Model averaging

Posterior odds ratios are the basis of model averaging, which via (2.5.1) is fundamental to forecasting and signal extraction when more than one model is under consideration. The essential technical task in model comparison is obtaining the marginalized likelihood M_{jT} defined in (2.5.3). 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 2.5 it is essential to distinguish between probability distribution functions and their kernels in the marginalized likelihood. In what follows, $\mathrm{p}(\theta)$ always denotes the properly normalized prior density and $\mathrm{p}(\mathbf{Y}_T|\theta)$ 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 (2.6.1) and (2.6.2), 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 $\mathrm{j}(\theta)$ 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,

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

And if

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

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 (4.4.2) 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 $\hat{M}_T^{(M)} = M^{-1}\sum_{m=1}^{M} \mathrm{p}(\theta_m)\mathrm{p}(\mathbf{Y}_T|\theta_m)/\mathrm{r}(\theta_m)$, 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. 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. The decomposition of the likelihood function set forth in Section 4.1 provides a fully general approach, but in effect this requires the consideration of many models. On this approach see also Gelfand, Dey and Chang (1992), Geweke (1994), Kass and Raftery (1995, Section 3.2), and 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 $\mathrm{g}(\theta) = [\mathrm{p}(\theta)\mathrm{p}(\mathbf{Y}_T|\theta)]^{-1}$ then $\mathrm{E}[\mathrm{g}(\theta)] = M_T^{-1}$. But $\mathrm{g}(\theta)$ 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

$$\mathrm{E}[\mathrm{f}(\theta)/\mathrm{p}(\theta)\mathrm{p}(\mathbf{Y}_T|\theta)] = M_T^{-1} \tag{4.4.3}$$

for any p.d.f. $\mathrm{f}(\theta)$ whose support is contained in Θ. One can approximate (4.4.3) from the output of any posterior simulator in the obvious way, but for this approximation to have a practical rate of convergence $\mathrm{f}(\theta)/\mathrm{p}(\theta)\mathrm{p}(\mathbf{Y}_T|\theta)$ should be uniformly bounded. Gelfand and Dey (1994) and Raftery (1995) interpret this condition as requiring that $\mathrm{f}(\theta)$ have "thin tails" relative to the likelihood function.

It is not difficult to guarantee both the boundedness and thin tail condition in (4.4.3). 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 $\hat{\Sigma}_M = M^{-1}\sum_{m=1}^{M}(\theta_m - \hat{\theta}_M)(\theta_m - \hat{\theta}_M)'$. [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 (4.4.1) can be applied directly.] It is not essential that the posterior mean and variance of θ exist. Then take

$$(4.4.4) \qquad \mathrm{f}(\theta) = 2(2\pi)^{-k/2}\left|\hat{\Sigma}_M\right|^{-1/2}\exp\left[-\tfrac{1}{2}\left(\theta_m - \hat{\theta}_M\right)'\Sigma^{-1}\left(\theta_m - \hat{\theta}_M\right)\right]\chi_{\hat{\Theta}_M}(\theta),$$

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

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 $\{\theta_\ell\}$ from the original distribution with p.d.f. (4.4.4) and averaging $\chi_\Theta(\theta_\ell)$, 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 $\mathrm{p}(\theta|\mathbf{Y}_T) = \mathrm{p}(\theta)\mathrm{p}(\mathbf{Y}_T|\theta)/M_T$, $M_T = \mathrm{p}(\theta^*)\mathrm{p}(\mathbf{Y}_T|\theta^*)/\mathrm{p}(\theta^*|\mathbf{Y}_T)$ for any $\theta^* \in \Theta$. [In all cases, $\mathrm{p}(\cdot)$ denotes a properly normalized density and not merely a kernel.] Typically $\mathrm{p}(\mathbf{Y}_T|\theta^*)$ and $\mathrm{p}(\theta^*)$ can be evaluated in closed form, but $\mathrm{p}(\theta^*|\mathbf{Y}_T)$ cannot. A marginal/conditional decomposition of $\mathrm{p}(\theta^*|\mathbf{Y}_T)$ is

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

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

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

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

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

Chib (1995) describes an extension to include latent variables.

4.5 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 compute posterior moments and solve signal extraction and forecasting problems not considered by the investigator. In signal extraction or forecasting the client simulates one (or more) values of ω corresponding to each θ_m, from the density $\mathrm{p}(\omega|\theta_m,\mathbf{Y}_T)$. This simulation is typically much easier and faster than is the posterior simulator itself. Given the collection of simulated ω, solution of the formal problem then proceeds as described in Section 4.2.

With a small amount of additional effort the client can modify many of the investigator's assumptions. Suppose the client wishes to evaluate $\mathrm{E}[\mathrm{g}(\theta)|\mathbf{Y}_T]$ using his own prior density $\mathrm{p}^*(\theta)$ rather than the investigator's prior density $\mathrm{p}(\theta)$. 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 $\left\{\theta^m\right\}_{m=1}^{M}$ using the function

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

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

$$\bar{g}_M^* \equiv \sum\nolimits_{m=1}^{M} \mathrm{w}^*(\theta_m)\mathrm{w}(\theta_m)\mathrm{g}(\theta_m) \Big/ \sum\nolimits_{m=1}^{M} \mathrm{w}^*(\theta_m)\mathrm{w}(\theta_m) \rightarrow \mathrm{E}^*\left[\mathrm{g}(\theta)|\mathbf{Y}_t\right] \equiv \bar{g}^*.$$

The result $\bar{g}_M^* \rightarrow \bar{g}^*$ follows almost at once from Tierney (1994).

The efficiency of the reweighting scheme requires some similarity of $\mathrm{p}^*(\theta)$ and $\mathrm{p}(\theta)$. In particular, both reasonable convergence rates and the use of a central limit theorem to assess numerical accuracy essentially require that $\mathrm{p}^*(\theta)/\mathrm{p}(\theta)$ be bounded. Across a set of diverse clients this condition is more likely to be satisfied the more diffuse is $\mathrm{p}(\theta)$, and is trivially satisfied for the improper prior $\mathrm{p}(\theta) \propto$ 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 $\mathrm{p}(\theta_m)$ 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 $\mathrm{p}^*(\theta)$'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 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.

5. Some models

The innovations in methods for simulation from posterior distributions just described have made possible routine and practical applications of Bayesian methods in statistics.

This section reviews the implementation of posterior simulators in a few models for economic time series. The survey concentrates on just a few models in order to provide the technical detail that is essential to the application of these methods, not just their appreciation. All of the method 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.

5.1 Vector autoregressions

The vector autoregression (VAR) was introduced by Sims (1980) and has subsequently been applied extensively in macroeconomics (e.g., Doan, Litterman and Sims, 1984; Blanchard and Quah, 1989) and forecasting (e.g., Litterman, 1986). The canonical model for L time series $\mathbf{y}_t = (y_{1t}, \ldots, y_{Lt})'$ is

$$(5.1.1) \qquad \mathbf{y}_t = \mathbf{B}_0 \mathbf{z}_t + \sum\nolimits_{s=1}^{p} \mathbf{B}_s \mathbf{y}_{t-s} + \varepsilon_t, \quad \varepsilon_t \overset{IID}{\sim} \mathrm{N}(\mathbf{0}, \Sigma) \quad (t = 1,2,\ldots)$$

conditional on $\mathbf{y}_0, \ldots, \mathbf{y}_{1-p}$ and a $k \times 1$ vector of deterministic covariates $\mathbf{z}_t$. There are other, equivalent, representations of the VAR, based on alternative normalizations; these include recursive and block recursive forms, as well as error correction representations. Since these are all renormalizations, it proves convenient to treat them as functions of interest and they are described from that point of view below in Section 5.1.2.

Some extension of notation reveals relationships between the VAR and other econometric models. Let

$$\underset{T\times k}{\mathbf{Z}} = \begin{bmatrix} \mathbf{z}_1' \\ \vdots \\ \mathbf{z}_T' \end{bmatrix}, \quad \underset{T\times L}{\mathbf{Y}_s} = \begin{bmatrix} \mathbf{y}_1' \\ \vdots \\ \mathbf{y}_T' \end{bmatrix}, \quad \underset{T\times L}{\mathbf{E}} = \begin{bmatrix} \varepsilon_1' \\ \vdots \\ \varepsilon_T' \end{bmatrix}$$

and take $\mathbf{X} = [\mathbf{Z}, \mathbf{Y}_0, \ldots, \mathbf{Y}_{1-p}]$, $\mathbf{B}' = [\mathbf{B}_0, \ldots, \mathbf{B}_p]$. Then

$$\underset{T\times L}{\mathbf{Y}_1} = \underset{T\times(k+pL)}{\mathbf{X}} \mathbf{B} + \underset{T\times L}{\mathbf{E}},$$

a multivariate regression (Anderson, 1984). The maximum likelihood estimator of the parameters is $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, $\hat{\Sigma} = T^{-1}\mathbf{S} = T^{-1}(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})$.

Alternatively, let $\mathbf{y} = \mathrm{vec}(\mathbf{Y})$, $\beta = \mathrm{vec}(\mathbf{B})$, and $\varepsilon = \mathrm{vec}(\mathbf{E})$. Then

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

Thus the VAR is a seemingly unrelated regressions model (Zellner, 1962) with the same covariates in each equation.

5.1.1 Prior distributions

Through straightforward manipulations the likelihood function for (5.1.2) can be expressed

(5.1.3) $$|\Sigma|^{-T/2}\exp\left(-\tfrac{1}{2}\operatorname{tr}\Sigma^{-1}\mathbf{S}\right)$$

(5.1.4) $$\cdot\exp\left\{-\tfrac{1}{2}\left(\beta-\hat{\beta}\right)'\left(\Sigma^{-1}\otimes\mathbf{X'X}\right)\left(\beta-\hat{\beta}\right)\right\}$$

where $\hat{\beta}=\operatorname{vec}\left(\hat{\mathbf{B}}\right)$. Integrating (5.1.3) with respect to β, obtain

$$|\Sigma|^{-\left(T-k-pL^2\right)/2}\exp\left(-\tfrac{1}{2}\operatorname{tr}\mathbf{S}\Sigma^{-1}\right),$$

up to a factor of proportionality not involving Σ. The functional form in Σ^{-1} is the same as that of the Wishart distribution (3.1.1). Interpreted as a kernel in β, (5.1.4) implies $\beta\sim\mathrm{N}\left[\hat{\beta},\ \Sigma\otimes\left(\mathbf{X'X}\right)^{-1}\right]$. Hence a fully conjugate prior distribution has the form

$$\Sigma^{-1}\sim\mathrm{W}\left(\underline{\mathbf{S}}^{-1},\underline{\nu}\right),\ \ \beta|\Sigma\sim\mathrm{N}\left(\underline{\beta},\Sigma\otimes\underline{\mathbf{H}}_{\beta}^{-1}\right).$$

Multiplying the kernel of the conjugate distribution by (5.1.3)-(5.1.4), after a little manipulation one obtains the kernel of the posterior distribution

$$\Sigma^{-1}\sim\mathrm{W}\left[\left(\underline{\mathbf{S}}+\mathbf{S}\right)^{-1},\underline{\nu}+T-k-pL^2\right],\ \ \beta|\Sigma\sim\mathrm{N}\left(\overline{\beta},\Sigma\otimes\overline{\mathbf{H}}_{\beta}^{-1}\right)$$

with $\overline{\mathbf{H}}_{\beta}=\underline{\mathbf{H}}_{\beta}+\mathbf{X'X}$ and $\overline{\beta}=\left(\mathbf{I}\otimes\overline{\mathbf{H}}_{\beta}^{-1}\underline{\mathbf{H}}_{\beta}\right)\underline{\beta}+\left(\mathbf{I}\otimes\overline{\mathbf{H}}_{\beta}^{-1}\mathbf{X'X}\right)\hat{\beta}$. Independence simulation from this posterior distribution is very fast and simple.

The fully conjugate prior distribution is often an inconvenient representation of beliefs, since uncertainty about the variance matrix and the coefficients is linked. The prior distribution

(5.1.5) $$\beta\sim\mathrm{N}\left(\underline{\beta},\underline{\mathbf{H}}_{\beta}^{-1}\right),\ \ \Sigma^{-1}\sim\mathrm{W}\left(\underline{\mathbf{S}}^{-1},\underline{\nu}\right)$$

does not have this property, and leads immediately to

(5.1.6) $$\beta\big|\Sigma,\mathbf{Y},\mathbf{X}\ \sim\ \mathrm{N}\left(\overline{\beta},\overline{\mathbf{H}}_{\beta}^{-1}\right)$$

with $\overline{\mathbf{H}}_{\beta}=\underline{\mathbf{H}}_{\beta}+\Sigma^{-1}\otimes\mathbf{X'X}$, $\overline{\beta}=\overline{\mathbf{H}}_{\beta}^{-1}\left(\underline{\mathbf{H}}_{\beta}\underline{\beta}+\overline{\mathbf{H}}_{\beta}\hat{\beta}\right)$, and

(5.1.7) $$\Sigma^{-1}|\mathbf{B},\mathbf{Y},\mathbf{X}\ \sim\ \mathrm{W}\left(\overline{\mathbf{S}}^{-1},\overline{\nu}\right)$$

with $\overline{\mathbf{S}}=\underline{\mathbf{S}}+(\mathbf{Y}-\mathbf{XB})'(\mathbf{Y}-\mathbf{XB})$, $\overline{\nu}=\underline{\nu}+T$. Thus this prior distribution is conditionally conjugate. It is immediately suited to a Gibbs sampling algorithm blocked in β and Σ. The computations here are more demanding, since a linear system of order $k+pL^2$ must be solved at each step.

The prior distribution (5.1.5) as stated involves potentially a very large number of parameters. Simply to organize the representation of prior beliefs about β and Σ within the family (5.1.5) it is necessary to restrict $\underline{\beta}$ and $\underline{\mathbf{H}}_{\beta}$. This can be done conveniently through a system of hierarchical priors (Section 2.7). This approach has been taken by Doan, Litterman and Sims (1984) and by Chib and Greenberg (1995) for a closely related problem in the seemingly unrelated regressions model. A common notation that captures all of these approaches is

$$\beta \sim \mathrm{N}\left[\underline{\beta}(\mu), \underline{\mathbf{H}}_{\beta}(\pi)^{-1}\right]; \quad \mu \sim \mathrm{p}_{\mu}(\cdot), \;\; \pi \sim \mathrm{p}_{\pi}(\cdot).$$

Conditional on $\mu, \pi, \Sigma, \mathbf{X}$ and $\mathbf{Y}$ the distribution of β is (5.1.6). Conditional on β and Σ the posterior distribution of μ and π has kernel

$$\text{(5.1.8)} \qquad \exp\left\{-\tfrac{1}{2}\left[\beta-\underline{\beta}(\mu)\right]' \underline{\mathbf{H}}_{\beta}(\pi)\left[\beta-\underline{\beta}(\mu)\right]\right\} \mathrm{p}_{\mu}(\mu)\mathrm{p}_{\pi}(\pi)$$

since μ and π do not appear directly in the likelihood function. The practicality of this procedure rests on the existence of a suitable method of drawing μ and π from (5.1.8). This is generally not difficult to achieve since μ and π are likely of low dimension. For example, in the spirit of Doan, Litterman and Sims (1984) we might have

$$b_{1jj}|(\mu,\pi) \sim \mathrm{N}(\mu_1, \pi_1), \qquad b_{1ij}|(\mu,\pi) \sim \mathrm{N}(\mu_2, \pi_2) \;\; (i \neq j),$$

$$b_{sjj}|(\mu,\pi,\tau) \sim \mathrm{N}\left(\mu_3, \pi_3\pi_5^{s-1}\right), \quad b_{sij}|(\mu,\pi,\tau) \sim \mathrm{N}\left(\mu_4, \pi_4\pi_6^{s-1}\right) \;\; (s > 1, i \neq j),$$

where all distributions are conditionally independent. The hierarchy might be completed by the seven independent prior distributions,

$$\mu \sim \mathrm{N}\left(\underline{\mu}, \underline{\mathbf{H}}_{\mu}^{-1}\right), \quad \underline{s}_j^2/\pi_j \sim \chi^2\left(\underline{\nu}_j\right) \;\; (j = 1,\ldots 4), \quad \pi_j \sim \mathrm{U}(0,1) \;\; (j = 5,6).$$

It is straightforward to verify that conditional on $\beta, \Sigma, \mathbf{Y}$ and $\mathbf{X}$, β is sufficient for the distribution of μ and π; that the seven posterior distributions of μ and π_j $(j = 1,\ldots 6)$ are conditionally independent; that the conditional distribution of μ is multivariate normal; and π_j $(j = 1,\ldots,4)$ is inverted gamma. The conditional distributions of π_5 and π_6 are unconventional but are easily handled through acceptance sampling along the lines described in Geweke (1995a, Section 3.2).

5.1.2 Functions of interest

The properties of vector autoregressions in the population have been studied extensively. Any set of such properties may be represented through functions of interest of the form $\mathrm{g}\left(\mathbf{B}_0,\ldots,\mathbf{B}_p,\Sigma\right)$, and hence inferences about them can be carried out readily by means of posterior simulation. Examples include questions about stationarity, cointegration, spectral densities, and various decompositions of variance. Here we discuss three kinds of properties: alternative normalizations; transformations of parameters; and problems in prediction.

For many purposes, other normalizations of (5.1.1) are convenient. Once such normalization is the fully recursive form,

$$\text{(5.1.9)} \qquad \mathbf{y}_t = \sum\nolimits_{s=0}^{p} \mathbf{A}_s \mathbf{y}_{t-s} + \eta_t, \quad \eta_t \overset{IID}{\sim} \mathrm{N}(\mathbf{0}, \Phi)$$

in which $\mathbf{A}_0$ is lower triangular, $[\mathbf{A}_0]_{jj} = 0$ $(j = 1,\ldots,L)$, and Φ is a diagonal matrix. The mapping from Σ and the $\mathbf{B}_s$ to Φ and the $\mathbf{A}_s$ can be constructed explicitly by letting $\mathbf{PP}' = \Sigma$ be the unique Choleski decomposition of Σ in which $\mathbf{P}$ is lower triangular and $p_{jj} > 0$ $(j = 1,\ldots,L)$; take $\Phi = \mathrm{diag}(p_{jj}^2)$; $\mathbf{R} = \mathbf{P}\Phi^{-1/2}$; $\mathbf{A}_s = \mathbf{R}^{-1}\mathbf{B}_s$ $(s > 1)$; and $\mathbf{A}_0 = \mathbf{I} - \mathbf{R}^{-1}$.

Given stationarity of $\mathbf{y}$ conditional on $\mathbf{z}$ there are three other standard useful representations of multivariate time series that may be obtained as functions of interest. A necessary and sufficient condition for conditional stationarity is that the roots of

(5.1.10) $\left|\mathbf{I}_L - \sum_{s=1}^{p} \mathbf{B}_s z^s\right| = 0$

all lie outside the unit circle. Stationarity may be imposed by checking this condition directly, and discarding draws corresponding to nonstationary configurations of $\{\mathbf{B}_s\}_{s=1}^{p}$.

The moving average representation corresponding to (5.1.1) is $\mathbf{y}_t = \sum_{s=0}^{\infty} \mathbf{B}_s^* \mathbf{B}_0 \mathbf{z}_{t-s} + \sum_{s=0}^{\infty} \mathbf{B}_s^* \varepsilon_{t-s}$. The sequence $\{\mathbf{B}_s^*\}_{s=0}^{\infty}$ is the inverse of $\{\mathbf{B}_s\}_{s=1}^{p}$ under convolution, i.e.,

$$\left(\sum_{s=0}^{\infty} \mathbf{B}_s^* z^s\right)\left(\mathbf{I}_L - \sum_{s=1}^{\infty} \mathbf{B}_s z^s\right) = \mathbf{I}_L \ \forall\ z{:}|z| \le 1.$$

The terms $\mathbf{B}_s^*$ may be obtained through the recursion

$$\mathbf{B}_0^* = \mathbf{I},\ \mathbf{B}_r^* = \sum_{j=0}^{r-1} \mathbf{B}_j^* \mathbf{B}_{r-j}\ (r = 1,2,\ldots).$$

One may obtain moving averages corresponding to other normalizations, as well. The representation $\mathbf{y}_t = \sum_{s=0}^{\infty} \mathbf{D}_s^* \mathbf{D}_0 \mathbf{z}_{t-s} + \sum_{s=0}^{\infty} \mathbf{D}_s^* \zeta_{t-s}$ corresponding to (5.1.10) is given by the recursion

$$\mathbf{A}_0^* = (\mathbf{I} - \mathbf{A}_0)^{-1},\quad \mathbf{A}_r^* = \left(\sum_{j=0}^{r-1} \mathbf{A}_j^* \mathbf{A}_{r-j}\right)\mathbf{A}_0^*\ \ (r = 1,2,\ldots).$$

This representation has been used extensively to examine the impulse response functions

$$[\mathbf{D}_s^*]_{ij}(\psi_{jj})^{1/2}\quad (s = 0,1,2,\ldots),$$

which trace out the effect of a typical shock of size $(\psi)_{jj}^{1/2}$ in ζ_{jt} on y_{it}. There is a substantial literature on methods for obtaining confidence bands for impulse response functions (e.g., Sims, 1986; Runkle, 1987; Blanchard and Quah, 1989; Sims and Zha, 1994; Koop, 1995; Phillips, 1995). Since the impulse response function is a closed form mapping from the parameters of the VAR, however, there really are no essential difficulties in a Bayesian approach.

The spectral density matrix corresponding to (5.1.1) is

$$\mathbf{S}_y(\lambda) = \left[\mathbf{I} - \sum_{s=1}^{p} \mathbf{B}_s \exp(-i\lambda s)\right]^{-1} \Sigma \left[\mathbf{I} - \sum_{s=1}^{p} \mathbf{B}_s \exp(i\lambda s)\right]^{-1'}.$$

At a given frequency the spectral density matrix is a closed form function of the parameters of the VAR, and it may be computed at many frequencies. Spectral densities have been applied in a wide variety of signal extraction problems (Nerlove, Grether and Carvalho, 1979, Chapters 3-4; Whittle, 1983). The frequency domain representation provides several useful adjuncts to the study of multiple time series. One is that the roots of (5.1.10) all lie outside the unit circle if and only if

$$(5.1.11) \qquad (2\pi)^{-1}\int_{-\pi}^{\pi}\log\left|\mathbf{I}-\sum\nolimits_{s=1}^{p}\mathbf{B}_s\exp(-i\lambda s)\right|^2 d\lambda = 0$$

(Rozanov, 1967, Theorem 4.2). For large systems it is easier to check this condition by computing $\mathbf{I}-\sum_{s=1}^{p}\mathbf{B}_s\exp(-i\lambda s)$ at many frequencies than to determine the roots of (5.1.10) directly. From (5.1.1) and (5.1.11),

$$(2\pi)^{-1}\int_{-\pi}^{\pi}\log\left|\mathbf{S}_{\mathbf{y}}(\lambda)\right|d\lambda = \log|\Sigma|.$$

The autocovariance function of $\mathbf{y}$ conditional on $\mathbf{z}$ cannot be determined in closed form from the VAR parameters. Three approaches are possible, but it is not clear if any is generally more efficient than the others. Since

$$\mathbf{R}_y(r) = (2\pi)^{-1}\int_{-\pi}^{\pi}\mathbf{S}_y(\lambda)\exp(-i\lambda r)d\lambda = \sum\nolimits_{s=r}^{\infty}\mathbf{A}_s\Sigma\mathbf{A}'_{s-r},$$

the autocovariance function may be approximated by computing many spectral density ordinates or terms in the moving average representation. Alternatively the Yule-Walker relations

$$\mathbf{R}_{\mathbf{y}}(0) = \sum\nolimits_{s=1}^{p}\mathbf{B}_s\mathbf{R}_{\mathbf{y}}(-s) + \Sigma$$

$$\mathbf{R}_{\mathbf{y}}(j) = \sum\nolimits_{s=1}^{p}\mathbf{B}_s\mathbf{R}_{\mathbf{y}}(j-s) \quad (j = 1,\ldots,p-1)$$

may be solved for $\mathbf{R}_{\mathbf{y}}(j)$ $(j = 0,\ldots,p-1)$ through iteration to a fixed point, and then $\mathbf{R}_{\mathbf{y}}(j) = \sum_{s=1}^{p}\mathbf{B}_s\mathbf{R}_{\mathbf{y}}(j-s)$ $(j = p, p+1,\ldots)$ may be computed iteratively.

Yet another normalization is the error-correction representation (Davidson, Hendry, Srba and Yeo, 1978) that has proved especially useful in the study of co-integration (Engle and Granger, 1987). Write (5.1.1) in the form

$$(5.1.12) \qquad \Delta\mathbf{y}_t = \mathbf{C}_0\Delta\mathbf{z}_t + \sum\nolimits_{s=1}^{p-1}\mathbf{C}_s\Delta\mathbf{y}_{t-s} + \mathbf{C}_p\mathbf{y}_{t-1} + \varepsilon_t, \quad \varepsilon_t \overset{IID}{\sim} \mathrm{N}(\mathbf{0},\Sigma)$$

in which $\mathbf{C}_0 = \mathbf{B}_0, \mathbf{C}_j = -\sum_{s=j+1}^{p}\mathbf{B}_s$ $(j = 1,\ldots,p-1)$, and $\mathbf{C}_p = -\left(\mathbf{I}-\sum_{s=1}^{p}\mathbf{B}_s\right)$. A necessary and sufficient condition for stationarity of $\mathbf{y}$ conditional on $\mathbf{z}$ is that all roots of (5.1.10) lie outside the unit circle. Hence $\mathrm{rk}\left(\mathbf{C}_p\right) = L$ is necessary for stationarity. Moreover, departures from stationarity thought likely *a priori* typically imply $\mathrm{rk}\left(\mathbf{C}_p\right) < L$.

Interest in the literature has concentrated on the implications of $\mathrm{rk}\left(\mathbf{C}_p\right) < L$. Let

$$\mathbf{w}_t = \begin{pmatrix} \mathbf{w}_{1t} \\ \mathbf{w}_{2t} \end{pmatrix} = \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \end{bmatrix} \mathbf{y}_t = \mathbf{R}\mathbf{y}_t$$

with $\mathbf{R}$ nonsingular and $\mathbf{R}_1\mathbf{R}_2' = \mathbf{0}$. The vector $\mathbf{w}_{1t}$ is taken to be stationary while $\mathbf{w}_{2t}$ is nonstationary with no stationary linear combinations. Then (5.1.12) implies

$$\Delta\mathbf{w}_t = \mathbf{C}_0^*\mathbf{z}_t + \sum\nolimits_{s=1}^{p-1}\mathbf{C}_s^*\Delta\mathbf{z}_{t-1} + \mathbf{C}_p^*\mathbf{z}_{t-1} + \varepsilon_t^*$$

where $\mathbf{C}_0^* = \mathbf{R}\mathbf{C}_0$, $\mathbf{C}_s^* = \mathbf{R}\mathbf{C}_s\mathbf{R}^{-1}\ (s = 1,\ldots,p-1)$, and

$$\mathbf{C}_p^* = \mathbf{R}\mathbf{C}_p\mathbf{R}^{-1} = \begin{bmatrix} \mathbf{R}_1\mathbf{C}_p \\ \mathbf{R}_2\mathbf{C}_p \end{bmatrix} \mathbf{R}^{-1}.$$

If $\mathbf{R}_1\mathbf{C}_p$ is of full row rank and $\mathbf{R}_2\mathbf{C}_p = \mathbf{0}$, then $\mathbf{w}_{1t}$ is stationary but $\mathbf{w}_{2t}$ is nonstationary. Imposition of this condition amounts to assuming that the last r_2 rows of $\mathbf{C}_p^*$ are zero, which is easily accomplished in the context of the Gibbs sampling posterior simulator described above.

Because of the recursive formulation of the VAR sampling from the predictive density is straightforward. Given the parameters of the model, the data, and all future values of the deterministic process $\{\mathbf{z}_t\}$, the recursion

$$\tilde{\mathbf{y}}_t \equiv \mathbf{y}_t \quad (t = 1,\ldots,T)$$

$$\tilde{\mathbf{y}}_{T+j} \sim N\left(\mathbf{B}_0\mathbf{z}_{T+j} + \sum\nolimits_{s=1}^{p}\mathbf{B}_s\tilde{\mathbf{y}}_{T+j-s},\ \Sigma\right) \quad (j = 1,2,\ldots)$$

provides a draw from the conditional distribution. Based on one or more such draws for each simulation of parameter values, forecasting problems can be attacked directly as described in Section 4.2.

5.2 Time-varying volatility models

Models in which the volatility of asset returns varies smoothly over time have received considerable attention in recent years. (For a survey of several approaches see Bollerslev, Chou and Kroner (1992).) Persistent but changing volatility is an evident characteristic of returns data. Since the conditional distribution of returns is relevant in the theory of portfolio allocation, proper treatment of volatility is important. Time-varying volatility also affects the properties of real growth and business cycle models.

The earliest model of time varying volatility is the autoregressive conditional heteroscedasticity (ARCH) model of Engle (1982). This was extended to the generalized ARCH (GARCH) model by Bollerslev (1986). Since then many variants of ARCH models have appeared. The distinguishing characteristic of these models is that the conditional variance of the return is a deterministic function of past conditional variances and past values of the return itself. GARCH models exhibit both time-varying volatility and leptokurtic unconditional distributions, but the two cannot be separated: these models cannot account for leptokurtosis without introducing time-varying volatility.

Stochastic volatility models have been examined by a series of investigators beginning with Taylor (1986). Promising Bayesian methods have been developed by Jacquier, Polson and Rossi (1994). In these models the conditional variance of the return is a stochastic function of its own past values but is unaffected by past returns themselves. Like GARCH models they account for time-varying volatility and leptokurtosis, but unlike GARCH models it is possible to have excess kurtosis without heteroscedasticity.

5.2.1 The GARCH model

The GARCH model of time-varying volatility may be expressed

$$\begin{aligned} & y_t = \beta'\mathbf{x}_t + h_t^{1/2}\varepsilon_t \\ (5.1.1) \qquad & h_t = \alpha + \textstyle\sum_{s=1}^{q} \gamma_s \varepsilon_{t-s}^2 + \sum_{j=1}^{p} \delta_j h_{t-j} \\ & \varepsilon_t \sim \mathrm{IIDN}(0,1) \end{aligned}$$

Here, y_t is the observed return at time t; $\mathbf{x}_t$ is a vector of covariates and β is the corresponding vector of coefficients; h_t is the conditional variance at time t; $\alpha > 0$, $\gamma_s \geq 0$ $(s = 1,\ldots,q)$, $\delta_j \geq 0\,(j = 1,\ldots,p)$. The vector of covariates is typically deterministic, including a constant term and perhaps indicator variables for calendar effects on the mean of y_t.

For the discussion here, assume the GARCH (1,1) model, which is (5.1.1) with $p = q = 1$. (Henceforth, we omit the subscripts on γ_1 and δ_1.) The GARCH (1,1) specification has proven attractive for models of returns. It typically dominates other GARCH models using the Akaike or Schwarz Bayesian information criteria (Bollerslev, Chou and Kroner, 1992). Following the GARCH literature we treat h_1 as a known constant. Then, the likelihood function is

$$\mathrm{L}_u(\beta,\alpha,\gamma,\delta|Y_u) = \textstyle\prod_{s=1}^{u} h_s^{1/2} \exp\left[-(y_s - \mathbf{x}_s'\beta)^2/2h_s\right]$$

where h_s is computed recursively from (5.1.1).

For expressing prior distributions as well as for carrying out the computations it proves useful to work with $a = \log(\alpha)$ rather than α. With this reparameterization a convenient functional form of the prior distribution is

$$\begin{aligned} & a \sim \mathrm{N}(\underline{a},\underline{s}_a^2); \\ (5.1.3) \qquad & \beta \sim \mathrm{N}(\underline{\beta},\underline{S}_\beta); \\ & \pi(\gamma,\delta) = 2\,(\gamma \geq 0, \delta \geq 0,\, \gamma + \delta < 1); \end{aligned}$$

and the distributions are independent. Restriction of $\gamma$and δ to the unit simplex is equivalent to the statement that the variance process is stationary.

A Metropolis independence chain can be construted to produce a sequence of parameters whose unconditional limiting distribution is the posterior distribution. Let $\theta' = (\beta', a, \gamma, \delta)$, and let $p(\theta|\mathbf{Y}_T)$ denote the posterior distribution. The kernel of this distribution is the product of (5.1.2) and the three prior density kernels in (5.1.3). The mode of the log posterior kernel is easily found using analytical expressions for the gradient and Hessian and a standard Newton-Raphson algorithm. Denote the mode by $\hat{\theta}$, and the Hessian at the mode by $\mathbf{H}$. Let $\mathrm{J}(\cdot;\mu, \mathbf{V}, \nu)$ denote the kernel density of a multivariate Student-*t* distribution with location vector μ, scale matrix $\mathbf{V}$, and ν degrees of freedom. For the choices $\mu = \hat{\theta}$, $\mathbf{V} = -(1.2)^2\mathbf{H}^{-1}$, $\nu = 5$, the ratio $p(\theta|\mathbf{Y}_T)/\mathrm{J}(\theta;\mu,\mathbf{V},\nu)$ is typically bounded above.

This multivariate Student-*t* distribution forms a proposal distribution for an independence Metropolis algorithm as follows. At step m, generate a candidate θ^* from $\mathrm{J}(\cdot;\mu, \mathbf{V}, \nu)$. With probability

$$p = \min\left\{\frac{p_{1t}(\theta^*|Y_t)/\mathrm{J}(\theta^*;\mu,\mathbf{V},\nu)}{p_{1t}(\theta^{(m-1)}|Y_t)/\mathrm{J}(\theta^{(m-1)};\mu,\mathbf{V},\nu)},\ 1\right\},$$

$\theta^{(m)} = \theta^*$; and with probability $1-p$, $\theta^{(m)} = \theta^{(m-1)}$. In applications of this proposal distribution, about half the candidate parameter vectors are typically accepted (Geweke, 1994).

5.2.2 The stochastic volatility model

The stochastic volatility model taken up by Jacquier, Polson and Rossi (1994) is

$$\begin{aligned} &y_t = \beta'\mathbf{x}_t + \varepsilon_t, \quad \varepsilon_t = h_t^{1/2}u_t, \\ &\log h_t = \alpha + \delta \log h_{t-1} + \sigma_\nu \nu_t, \\ &\begin{pmatrix} u_t \\ \nu_t \end{pmatrix} \overset{IID}{\sim} \mathrm{N}(0, \mathbf{I}_2), \end{aligned}$$

where $|\delta| < 1$ and $\sigma_\nu > 0$. Following Jacquier, Polson and Rossi, do not condition on h_1 but regard h_1 as a random variable drawn from its unconditional distribution $\mathrm{N}(\alpha/(1-\delta), \sigma_\nu^2/(1-\delta^2))$. Then,

$$\mathrm{L}(\beta,\alpha,\delta,\sigma_\nu;\mathbf{Y}_T) = \int_0^\infty \cdots \int_0^\infty \mathrm{L}^*(\beta,\alpha,\delta,\sigma_\nu,h_1,\ldots,h_u;\mathbf{Y}_T)dh_1\cdots dh_u$$

where

$$\mathrm{L}^*(\beta,\alpha,\delta,\sigma_v,h_1,\ldots,h_u;\mathbf{Y}_T) = \tag{5.2.4}$$
$$\prod_{s=1}^{u} h_s^{-3/2}\exp\left(-\sum_{s=1}^{u}\varepsilon_s^2/2h_s\right)\exp\left[-\sum_{s=2}^{u}(\log h_s-\alpha-\delta\log h_{s-1})^2/2\sigma_v^2\right]$$
$$\cdot\exp\left\{\left[\log h_1-\alpha/(1-\delta)\right]^2/\left[\sigma_v^2/(1-\delta^2)\right]\right\}.$$

The prior distributions for β and σ_v are of the forms

$$\beta\sim\mathrm{N}\left(\underline{\beta},\underline{S}_\beta\right)$$

and

$$\underline{\nu}_v\,\underline{s}_v^2/\sigma_v^2\sim\chi^2(\underline{\nu}_v),$$

respectively. The prior distribution of $(\alpha,\delta)'$ is bivariate normal, induced by independent normal prior distributions on the persistence parameter δ,

$$\delta\sim\mathrm{N}\left(\underline{\delta},\underline{s}_\delta^2\right)$$

and the unconditional mean of $\log h_t$,

$$\alpha/(1-\delta)\sim\mathrm{N}\left(\underline{h},\,\underline{s}_h^2\right).$$

A linearization of $\alpha/(1-\delta)$ yields the corresponding bivariate normal prior distribution,

$$\begin{pmatrix}\alpha\\ \delta\end{pmatrix}\sim\mathrm{N}\left(\begin{bmatrix}\underline{h}(1-\underline{\delta})\\ \underline{\delta}\end{bmatrix},\ \begin{bmatrix}\underline{s}_h^2(1-\underline{\delta})^2+\underline{h}^2\,\underline{s}_\delta^2 & -\underline{h}\underline{s}_\delta^2\\ -\underline{h}\underline{s}_\delta^2 & \underline{s}_\delta^2\end{bmatrix}\right). \tag{5.2.5}$$

Draws from the posterior distribution may be accomplished using Markov chain Monte Carlo algorithms. To describe these procedures, let $\theta'=(\beta',\alpha,\delta,\sigma_v)$ and $\mathbf{h}'=(h_1,\ldots,h_u)$, and note that for any function of interest $\mathrm{g}(\theta,\mathbf{h})$,

$$\mathrm{E}[\mathrm{g}(\theta,\mathbf{h})]=\frac{\int_\Theta \mathrm{g}(\theta)\mathrm{L}_u(\theta|Y_u)\pi(\theta)d\theta}{\int_\Theta \mathrm{L}_u(\theta|Y_u)\pi(\theta)d\theta}=\frac{\int_\Theta\int_H \mathrm{g}(\theta)\mathrm{L}_u^*(\theta,\mathbf{h}|Y_u)\pi(\theta)d\mathbf{h}d\theta}{\int_\Theta\int_H \mathrm{L}_u^*(\theta,\mathbf{h}|Y_u)\pi(\theta)d\mathbf{h}d\theta},$$

where $\pi(\theta)$ is the prior distribution constructed from (5.2.5).

The posterior distribution of β conditional on (α,δ), σ_v and $\mathbf{h}$ is normal, and the posterior distribution of (α,δ) conditional on β, σ_v and $\mathbf{h}$ is normal up to the last term of (5.2.4) which may be accommodated by acceptance sampling; and the distribution of σ_v conditional on β, (α,δ) and $\mathbf{h}$ is inverted gamma.

The nonstandard part of the problem is drawing the vector of latent variables $\mathbf{h}$. The posterior distribution of h_s $(1<s<u)$, conditional on $\{h_r, r\neq s\}$ and θ has density kernel

$$h_s^{-3/2}\exp\left(-\varepsilon_s^2/2h_s\right) \tag{5.1.6}$$

$$\cdot\exp\left[-(\log h_s-\mu_s)^2/2\sigma^2\right] \tag{5.1.7}$$

where

$$\varepsilon_s=y_s-\mathbf{x}_s'\beta,\quad \mu_s=\frac{\alpha(1-\delta)+\delta(\log h_{s-1}+\log h_{s+1})}{1+\delta^2},\quad \sigma^2=\frac{\sigma_v^2}{1+\delta^2}.$$

Drawing **h** may be accomplished in a number of different ways. Jacquier, Polson and Rossi (1994) use a Metropolis chain, generating from a candidate density, and using the procedures described in Section 3.4.2 to either accept or reject the draw. The term (5.1.6) is the kernel of a random variable whose inverse has a gamma distribution. Using this family to approximate (5.1.7) by matching first and second moments, and combining with the first term, yields the candidate density kernel

$$x^{-(\phi+1)}\exp(-\lambda/x);$$
$$\phi=\left[1-2\exp\left(\sigma^2\right)\right]/\left[1-\exp\left(\sigma^2\right)\right]+.5,$$
$$\lambda=(\phi-1)\exp\left(\mu_s+.5\sigma^2\right)+.5\varepsilon_s^2.$$

Alternatively (Geweke, 1994) note that the posterior conditional density kernel for $H_s=\log h_s$ is

$$\exp\left[-\left(H_s-\mu_s^*\right)/2\sigma^2\right]\exp\left[-\varepsilon_s^2/2\exp\left(H_s\right)\right],$$

where $\mu_s^*=\mu_s-.5\sigma^2$. One can draw efficiently from this distribution using acceptance sampling, employing a source $\mathrm{N}\left(\lambda,\sigma^2\right)$ distribution with λ chosen optimally as described in Geweke (1994, Section 3.2). For $H_1=\log h_1$ the conditional posterior density kernel is

$$\exp\left[-\left(H_1-\mu_1^*\right)^2/2\sigma_v^2\right]\exp\left[-\varepsilon_1^2/2\exp\left(H_1\right)\right]$$

where $\mu_1^*=\alpha+\delta H_2-.5\sigma_v^2$. There is a symmetric expression for $H_u=\log h_u$.

This approach to stochastic volatility can be extended in several dimensions, using Markov chain Monte Carlo methods for Bayesian inference. These include leptokurtic or skewed shocks and leverage effects through correlation between u_t and v_t. A multivariate generalization of the model is

$$\underset{L\times 1}{\mathbf{y}_t}=h_t^{1/2}\Sigma^{1/2}\varepsilon_t,$$
$$\log h_t=\alpha+\delta\log h_{t-1}+\sigma_v v_t,$$
$$\begin{pmatrix}\varepsilon_t\\ v_t\end{pmatrix}\overset{IID}{\sim}\mathrm{N}\left(\mathbf{0},\mathbf{I}_{L+1}\right)$$

which is also a stochastic generalization of the discount dynamic model used extensively in Bayesian forecasting (Harrison and Stevens, 1976; West and Harrison, 1989). For an overview of these extensions see Jacquier, Polson and Rossi (1995).

5.3 Changing regime models

Changing regime models provide one means of introducing nonlinear behavior in time series, while still retaining much of the tractability of (5.1.1). The nonlinearity is introduced through a latent process $\left\{s_t\right\}$ whose range is $s_t=j\ (1,\ldots,J)$. If $s_t=j$, then

$$\mathbf{y}_t = \mathbf{B}_0^{(j)}\mathbf{z}_t + \sum_{s=1}^{p} \mathbf{B}_s^{(j)}\mathbf{y}_{t-s} + \varepsilon_t, \quad \varepsilon_t \overset{IID}{\sim} \mathrm{N}\left(\mathbf{0}, \Sigma^{(j)}\right) \quad (t = 1,2,\ldots).$$

Much as before denote $\mathbf{B}^{(j)'} = \left[\mathbf{B}_0^{(j)},\ldots,\mathbf{B}_p^{(j)}\right]$, $\beta^{(j)} = \mathrm{vec}\left[\mathbf{B}^{(j)}\right]$, $\beta' = \left(\beta^{(1)'},\ldots.,\beta^{(J)'}\right)$. A conditionally conjugate family of priors is

$$\beta \sim \mathrm{N}\left(\underline{\beta}, \underline{\mathbf{H}}_\beta^{-1}\right),\ \left[\Sigma^{(j)}\right]^{-1} \sim \mathrm{W}\left(\left[\underline{\mathbf{S}}^{(j)}\right]^{-1}, \underline{\nu}^{(j)}\right) (j = 1,\ldots,J),$$

where the $J+1$ distributions are independent. The need for expression of $\underline{\beta}$ and $\underline{\mathbf{H}}_\beta$ in terms of hyperparameters, or for hierarchical priors, is once again evident.

Since the model is symmetric in the $\left(\beta^{(j)}, \Sigma^{(j)}\right)$, further restrictions are required for identification. These necessarily depend on the particular application. Examples include inequality restrictions on the intercept coefficient in growth rate equations, thus identifying one state as a contraction in a two-state model (McCulloch and Tsay, 1994), restriction of a particular time to a specific state, and inequality constraints on variances (Albert and Chib, 1993). The threshold model, described below, provides yet another means of identification.

Conditional on the process $\{s_t\}$, inference can proceed much as in Section 5.1.1. If the $\beta^{(j)}$ are independent *a priori*, then this amounts to applying the procedures of Section 5.1.1 separately to subsamples of the form $\{t: s_t = j\}\,(j = 1,\ldots,J)$. If not, the posterior distribution of β, conditional on the $\Sigma^{(j)}$ and $\{s_t\}$, is still multivariate normal, and the posterior distributions of the $\Sigma^{(j)}$ are conditionally inverted Wishart.

The changing regime model is completed with specification of the process determining $\{s_t\}$,

$$\mathrm{p}\left(s_t \mid \theta, \mathbf{y}_{t-1},\ldots,\mathbf{y}_{t-r}, s_{t-1}, t\right) \quad (t = 1,\ldots,T),$$

in which the parameter vector θ indexes the class of processes, and a prior distribution $\mathrm{p}(\theta)$. The posterior density kernel is

$$\prod_{t=1}^{T}\Big\langle\left|\Sigma^{(s_t)}\right|^{-1/2} \tag{5.3.1}$$
$$\cdot\exp\left\{-\tfrac{1}{2}\left[\mathbf{y}_t - \mathbf{B}_0^{(s_t)}\mathbf{z}_t - \sum_{s=1}^{p}\mathbf{B}_s^{(s_t)}\mathbf{y}_{t-s}\right]'\left[\Sigma^{(s_t)}\right]^{-1}\left[\mathbf{y}_t - \mathbf{B}_0^{(s_t)}\mathbf{z}_t - \sum_{s=1}^{p}\mathbf{B}_s^{(s_t)}\mathbf{y}_{t-s}\right]\right\}\Big\rangle$$

$$\cdot\prod_{t=1}^{T}\mathrm{p}\left(\mathrm{s}_t \middle| \theta, \mathbf{y}_{t-1},\ldots,\mathbf{y}_{t-r}, s_{t-1}, t\right) \tag{5.3.2}$$

$$\cdot\,\mathrm{p}(\theta) \tag{5.3.3}$$

$$\cdot\prod_{j=1}^{J}\mathrm{p}\left(\Sigma^{(\mathrm{j})}\right)\cdot\mathrm{p}\left(\mathbf{B}_s^{(j)}\,(s = 0,\ldots,p; j = 1,\ldots,J)\right). \tag{5.3.4}$$

Conditional on all other parameters and the latent variables $\{s_t\}$, the kernel density for θ involves only (5.3.2) and (5.3.3). Thus, issues of drawing θ involve only the auxiliary model for the latent variables $\{s_t\}$. Conditional on $s_r\ (r \neq t)$,

$$P(s_t = j) \propto f_t(s_t) \cdot p(s_t | \theta, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-r}, s_{t-1}, t) \cdot p(s_{t+1} | \theta, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-r}, s_t, t)$$

where $\prod_{t=1}^{T} f_t(s_t)$ denotes (5.3.1) as a function of s_t. Thus, s_t can be drawn by evaluating (5.3.5) for $j = 1, \ldots, J$. A practical issue that may arise in this procedure is serial correlation of the algorithm induced in this data augmentation. Generally serial correlation can be reduced by drawing *r* adjacent s_t simultaneously, at the cost of J^r rather than rJ evaluations.

Several variants of (5.3.2)-(5.3.3) have been proposed, and in fact the variety of models that can be applied has been enhanced greatly by the development of Markov chain Monte Carlo methods for Bayesian inference. Here we describe four such models. In each case we concentrate on the conditional posterior distributions of the latent variables s_t, and the conditional distribution of the parameters peculiar to the variant.

5.3.1 Markov switching models

In the Markov switching regime model the evolution of states is described by the first order Markov chain,

$$P(s_t = j | \theta, t = 1) = p_{0j} \ (j = 1, \ldots, J),$$

$$P(s_t = j | \theta, s_t = i) = p_{ij} \ (i, j = 1, \ldots, J; t = 2, 3, \ldots)$$

This model was developed and applied to macroeconomic time series by Hamilton (1989, 1990). Bayesian inference using the Gibbs sampler has been implemented by Albert and Chib (1993), Chib (1994), and McCulloch and Tsay (1994).

The kernel of the likelihood function in the p_{ij} (5.3.3) is the product of $J+1$ kernels of the multinomial distribution indexed by $i = 0, \ldots, J$: $\prod_{j=1}^{J} p_{ij}^{n_{ij}}$ where n_{ij} is the number of transitions from regime *i* to regime *j* in $\{s_t\}_{t=1}^{T}$. A natural conjugate prior distribution for these parameters is the Dirichlet (also known as the multivariate beta) distribution,

$$p_i(p_{ij}) \propto \prod_{j=1}^{J} p_{ij}^{a_{ij}} \quad \left[p_{ij} \geq 0 \ (j = 1, \ldots, J); \sum_{j=1}^{J} p_{ij} = 1; a_{ij} \geq -1 \ (i = 0, \ldots, J)\right].$$

(For further discussion see Zellner, 1971, 38-39.) Hence a conditionally conjugate prior distribution for $\theta = \{p_{ij}, j = 1, \ldots, J; i = 0, \ldots, J\}$ is

$$p(\theta) \propto \prod_{i=0}^{J} \prod_{j=1}^{J} p_{ij}^{a_{ij}} \qquad \left[a_{ij} \geq -1, \, p_{ij} \geq 0 \, , j = 1, \ldots, J; \sum_{j=1}^{J} p_{ij} = 1; (i = 0, \ldots, J)\right]$$

and the corresponding conditional posterior density kernel is

$$\text{(5.3.6)} \qquad p[\theta] \propto \prod_{i=0}^{J} \prod_{j=1}^{J} p_{ij}^{n_{ij}+a_{ij}} \qquad \left[p_{ij} \geq 0 \ j = 1, \ldots, J; \sum_{j=1}^{J} p_{ij} = 1; \ (i = 0, \ldots, J)\right].$$

There is a convenient genesis for this density (Johnson and Kotz, 1972, 232-233). Construct the $J(J+1)$ independent random variables

$$d_{ij} \sim \chi^2\left[2(n_{ij} + a_{ij} + 1)\right] \quad (j = 1, \ldots, J; i = 0, \ldots, J). \qquad (7)$$

Then

$$p_{ij} = d_{ij} \Big/ \sum\nolimits_{k=1}^{m} d_{ik} \quad (j = 1,\ldots,J; i = 0,\ldots,J) \tag{8}$$

has probability density kernel (5.3.6).

Conditioning to denote (5.3.1) as a function of s_t by $\prod_{t=1}^{T} \mathrm{f}_t(s_t)$, the distribution of s_t conditional on s_r $(r \neq t)$ and all other parameters is

$$\mathrm{P}(s_1 = j) \propto \mathrm{f}_1(j) p_{0j} p_{j,s_2};$$
$$\mathrm{P}(s_t = j) \propto \mathrm{f}_t(j) p_{s_{t-1},j} p_{j,s_{t+1}} \; (t = 2,\ldots,T-1);$$
$$\mathrm{P}(s_T = j) \propto \mathrm{f}_T(j) p_{s_{T-1},j}.$$

This completes the set of conditional distributions needed for a Gibbs sampling algorithm for the Markov switching model. One can readily verify that Gibbs sampler convergence condition 1 applies.

The Markov switching model may be extended, by allowing transition probabilities to depend on the time of year. See Ghysels, McCulloch and Tsay (1994) for methods, and evidence for this sort of behavior in macroeconomic time series.

5.3.2 Probit switching model

In the probit switching regime model, the evolution of the state variables is described by means of an auxiliary $J \times 1$ vector of latent variables $\mathbf{v}_t$,

$$(v_{1t},\ldots,v_{J-1,t})' \sim (\Gamma \mathbf{w}_t, \Omega)$$

where $\mathbf{w}_t$ is a subset of $\{\mathbf{y}_{t-1},\ldots,\mathbf{y}_{t-r}\}$, and $v_{Jt} \equiv 0$. Then,

$$\mathrm{P}(s_t = j) = \mathrm{P}(v_{jt} \geq v_{it} \; \forall \; i = 1,\ldots,J).$$

This model has been studied by McCulloch and Tsay (1993).

Conditional on s_t, this is a conventional multinomial probit model. If (Γ, Ω) and $(\mathbf{B}^{(j)}, \Gamma^{(j)}, j = 1,\ldots,J)$ are independent in the prior distribution, then the entire literature on Markov chain Monte Carlo methods for Bayesian inference in the multinomial probit model applies directly to the step of drawing Γ and Ω; e.g., see Geweke, Keane and Runkle (1994a, 1994b), McCulloch and Rossi (1995), and McCulloch, Polson and Rossi (1995). If $J = 2$ there is considerable simplification: see Albert and Chib (1993).

Once again letting $\prod_{t=1}^{T} \mathrm{f}_t(s_t)$ denote (5.3.1) as a function of s_t, the distribution of s_t conditional on s_r $(r \neq t)$ and all parameters is

$$\mathrm{P}(s_t = j) \propto \mathrm{f}_t(j) \cdot \mathrm{P}(\mathrm{v}_{jt} \geq v_{it} | \Gamma, \Omega).$$

When $J = 2$ the second probability in this expression is the c.d.f. of a univariate normal distribution. For $J \geq 3$, a recent literature on evaluation of orthant probabilities for the multivariate normal distribution can be applied; see Hajivassiliou, McFadden and Ruud

(1993) for documentation of the advantages of the GHK probability simulator due to Keane (1990), Geweke (1992), and Hajivassiliou and McFadden (1994). See Geweke, Keane and Runkle (1995) for a description and code.

5.3.3 Threshold autoregressive model

In the threshold autoregressive model introduced by Tong (1978, 1983) and Tong and Lim (1980),

$$s_t = \mathrm{f}\left(y_{t-q}, \theta\right).$$

The simplest leading example one that has been studied from both Bayesian and non-Bayesian perspectives, is $s_t = 1$ if $y_{t-q} \leq r$, $s_t = 2$ if $y_{t-q} > r$; $\theta = \left(q, r\right)'$. Conditional on θ the states are known. If prior distributions for the $\mathbf{B}^{(j)}$ and $\Sigma^{(j)}$ are jointly conditionally conjugate (as described at the start of Section 5.1.1) then analytic marginalilzation of the $\mathbf{B}^{(j)}$ and $\Sigma^{(j)}$ is possible. The marginal posterior in θ can then be evaluated directly, leading to an independence sampling algorithm for the whole posterior. Examples of this approach are Geweke and Terui (1992, 1993). Alternatively the problem can be blocked into the $\mathbf{B}^{(j)}$, the $\Sigma^{(j)}$, q and r. For this approach, see Carlin, Gelfand and Smith (1992) and McCulloch and Tsay (1994).

5.3.4 Pure break models

Perhaps the simplest changing regime model is

$$s_t = 1, t \leq b;\ s_t = 2, t > b,$$

as a special case of (5.3.2), together with a prior distribution for b from (5.3.3). There is a considerable non-Bayesian literature associated with this model in economics, beginning with Peron (1989). Non-Bayesian approaches are complicated by the issue of inference about b; conditioning problems similar to those discussed in Section 4.3 arise. In the Bayesian formulation the parameter b is symmetric with all other parameters in the model. Bayesian inference in the context of (5.3.1)-(5.3.4), using a Gibbs sampling algorithm, is straightforward. See DeJong (1992) for an early study.

The focus in the literature has been on the comparison of models with, and without, breaks. A formal comparison using Bayesian methods has not yet been made, to the author's knowledge. In doing so, it would be important that the prior distribution for the $\mathbf{B}^{(j)}$ and $\Sigma^{(j)}$ be chosen carefully. Given improper priors for these parameters, a finding of a one-regime model would be implied by Lindley's paradox.

The pure break model is ultimately handicapped by its failure to specify a stochastic process that determines breaks. Because of this, forecasts conditional only on the data are not possible in this model, as they are in the other changing regime models considered here. As an incompletely specified model, it is not well suited to serious practical application.

6. Conclusions

The procedures described in this survey can be summarized in three steps for the Bayesian econometric analysis of time series.

(1) *Be explicit about assumptions.* This entails a formal probability distribution over all the models under consideration. As a technical matter, it means describing prior beliefs through a probability for each model and a distribution of plausible parameter values within each model.

(2) *Condition on available information.* Available information consists of the assumptions in (1), and data related to the random variables whose distribution is governed by these assumptions.

(3) *Use posterior simulators to report the logical implications of (1) and (2).* The logical implications are completely summarized by the probability distributions of models and parameters conditional on available information. These implications are drawn using the laws of probability, i.e., Bayes' theorem.

These procedures impose considerable discipline on the econometrician, in all three steps. The discipline is precisely the same as that imposed in the development of ideas in modern economic theory, and on the behavior of rational economic agents in these models. The first two steps are no more than the application of the defining paradigm of the discipline of economics, to the work that economists and econometricians do when they confront their ideas with facts about the real world.

The third step is not essential to drawing the logical implications of the first two steps. However, it has two compelling advantages. The first stems from the fact that drawing the logical implications is technically very demanding. Posterior simulators provide by far the best device currently available for completing this task. As a practical matter they are the only device in most situations. The second compelling advantage is that the output of a posterior simulator, generated by an investigator, provides a simple tool by means of which a remote client can, within reasonable limits, alter the assumptions made in (1), update the data sets used in (2), and examine implications in (3) not considered by the investigator.

The implementation of the Bayesian paradigm made possible by recent innovations in posterior simulators places the formal analysis of economic time series on the same logical footing as economic science in general, and makes the results of that analysis more accessible to scholars and policy makers. The realization of this promise is just beginning, and its pursuit should provide worthy tasks for econometricians for some time.

References

Albert, J.H. and S. Chib, 1993, "Bayes Inference via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean and Variance Shifts," *Journal of Business and Economic Statistics* **11**: 1-15.

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

Amemiya, T., 1985, *Advanced Econometrics.* Cambridge: Harvard University Press.

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

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

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

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.

Blanchard, O.J., and D. Quah, 1989, "The Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review* **79**: 655-673.

Bollerslev, T., 1986, "Generalized Autoregressive Conditional Heteroskedasticity," *Journal of Econometrics* **31**: 307-327.

Bollerslev, T., R. Chou, and K.F. Kroner, 1992, "ARCH Modeling in Finance," *Journal of Econometrics* **52**: 5-59.

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.

Carlin, B., A. Gelfand and A.F.M. Smith, 1992, "Hierarchical Bayesian Analysis of Change Point Problems," *Applied Statistics* **41**: 389-405.

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

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., 1994, "Calculating Posterior Distributions and Modal Estimates in Markov Mixture Models," *Journal of Econometrics*, forthcoming.

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.

Christoffersen, P.F., and F.X. Diebold, 1995, "Optimal Prediction under Asymmetric Loss," NBER Technical Working Paper #167.

Davidson, J.E.H., D.F. Hendry, F. Srba, and S. Yeo, 1978, "Econometric Modeling of the Aggregate Time-Series Relationship between Consumers' Expenditure and Income in the United Kingdom," *Economic Journal* **88**: 661-692.

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

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

DeJong, D.N., 1992, "A Bayesian Search for Structural Breaks in U.S. GNP," in T. Fomby and R.C. Hill (eds.), *Advances in Econometrics: Bayesian Methods Applied to Time Series Data*. JAI Press, forthcoming.

Diebold, F.X. and R.S. Mariano, 1993, "Comparing Predictive Accuracy," University of Pennsylvania Department of Economics manuscript.

Doan, T., R. Litterman and C. A. Sims, 1984, "Forecasting and Conditional Projection Using Realistic Prior Distributions" *Econometric Reviews* **5**: 57-61.

Engle, R., 1982, "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica* **50**: 987-100.

Engle, R.F., and C.W.J. Granger, 1987, "Co-Integration and Error Correction: Representation, Estimation, and Testing," *Econometrica* **55**: 251-276.

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.

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., 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, "Posterior Simulators in Econometrics," in D. Kreps and K.F Wallis (eds.), *Advances in Economics and Econometrics: Theory and Applications.* Cambridge: Cambridge University Press, forthcoming. (Invited symposium paper, Econometric Society Seventh World Congress) Also Federal Reserve Bank of Minneapolis working paper No. 555, September 1995.

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 N. Terui, 1992, "Threshold Autoregressive Models for Macroeconomic Time Series: A Bayesian Approach," *American Statistical Association 1991 Proceedings of the Business and Economic Statistics Section,* 42-50.

Geweke, J., and N. Terui, 1993, "Bayesian Threshold Autoregressive Models for Nonlinear Time Series," *Journal of Time Series Analysis,* 1993, **14**, 441-455.

Ghysels, E., R.E. McCulloch, and R.S. Tsay, 1994, "Bayesian Inference for Periodic Regime-Switching Models," University of Chicago Graduate School of Business mimeo.

Granger, C.W.J., 1969, "Prediction with a Generalized Cost of Error Function," *Operational Research Quarterly* **20**: 199-207.

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

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.

Hamilton, J.D., 1989, "A New Approach to the Economic Analysis of Nonstationary Time Series," *Econometrica* **57**: 357-384.

Hamilton, J.D., 1990, "Analysis of Time Series Subject to Changes in Regime," *Journal of Econometrics* **45**: 39-70.

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.

Harrison, P.J. and C.F. Stevens, 1976, "Bayesian Forecasting," *Journal of the Royal Statistical Society Series B* **38**: 205-247.

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

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

Jacquier, E., N.G. Polson, and P.E. Rossi, 1994, "Bayesian Analysis of Stochastic Volatility Models," *Journal of Business and Economic Statistics* **12**: 371-417.

Jacquier, E., N.G. Polson, and P.E. Rossi, 1995, "Stochastic Volatility: Univariate and Multivariate Extensions," mimeo.

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

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.

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., 1995, "Parameter Uncertainty and Impulse Response Analysis," *Journal of Econometrics,* forthcoming.

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

Litterman, R.B., 1986, "Forecasting with Bayesian Vector Autoregressions - Five Years of Experience," *Journal of Business and Economic Statistics* **4**: 25-38.

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.

McCulloch, R.E., and R.S. Tsay, 1993, "Bayesian Inference and Prediction for Mean and Variance Shifts in Autoregressive Time Series," *Journal of the American Statistical Association* **88**: 968-978.

McCulloch, R.E., and R.S. Tsay, 1994, "Statistical Analysis of Economic Time Series via Markov Switching Models," *Journal of Time Series Analysis* **15**: 523-540.

McCulloch, R.E., and R.S. Tsay, 1995, "Bayesian Analysis of Threshold Autoregressive Processes with a Random Number of Regimes," University of Chicago Graduate School of Business.

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.

Nerlove, M., D.M. Grether, and J.L. Carvalho, 1979, *Analysis of Economic Time Series.* New York: Academic Press.

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.

Peron, P., 1989, "The Great Crash, the Oil Shock, and the Unit Root Hypothesis," *Econometrica* **57**: 1361-1401.

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

Phillips, P.C.B., 1995, "Impulse Response and Forecast Error Variance Asymptotics in Nonstationary VAR's," Yale University Cowles Foundation Discussion paper No. 1102.

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

Pole, A., and A.F.M. Smith, 1985, "Bayesian Analysis of Some Threshold Switching Models," *Journal of Econometrics* **29** 97-119.

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

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.

Rozanov, Y.A., 1967, *Stationary Random Processes.* San Francisco: Holden-Day.

Runkle, D.E., 1987, "Vector Autoregressions and Reality," *Journal of Business and Economic Statistics* **5**: 437-442.

Shao, J., 1989, "Monte Carlo Approximations in Bayesian Decision Theory," *Journal of the American Statistical Association* **84**: 727-732.

Sims, C.A., 1980, "Macroeconomics and Reality," *Econometrica* **48**: 1-48.

Sims, C.A., 1986, "Are Forecasting Models Usable for Policy Analysis?," *Federal Reserve Bank of Minneapolis Quarterly Review* **10**: 2-15.

Sims, C.A., and T. Zha, 1994, "Error Bands for Impulse Responses," Yale University Cowles Foundation Discussion Paper No. 1085.

Smith, A.A., 1991, "Solving Stochastic Dynamic Programming Problems using Rules of Thumb," Queen's University Department of Economics Discussion Paper No. 816.

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.

Taylor, S., 1986, *Modeling Financial Time Series.* New York: John Wiley and Sons.

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.)

Tong, H., 1978, "On a Threshold Model," in C.H. Chan (ed.), *Pattern Recognition and Signal Processing.* Amsterdam: Sijthoff and Noordhoff.

Tong, H., 1983, *Threshold Models in Non-linear Time Series Analysis.* New York: Springer-Verlag.

Tong, H., and K.S. Lim, 1980, "Threshold Autoregression, Limit Cycles and Cyclical Data," *Journal of the Royal Statistical Society Series B* **42**: 245-292.

Weiss, A.A., 1991, "Multi-step Estimation and Forecasting in Dynamic Models," *Journal of Econometrics* **48**: 135-149.

Weiss, A.A., and A.P. Andersen, 1984, "Estimating Forecasting Models Using the Relevant Forecast Evaluation Criterion," *Journal of the Royal Statistical Society Series A* **137**: 484-487.

West, M. and J. Harrison, 1990, *Bayesian Forecasting and Dynamic Models.* Berlin: Springer-Verlag.

Whittle, P., 1983, *Prediction and Regulation by Linear Least-Square Methods* (Second Edition). Minneapolis: University of Minnesota Press.

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.

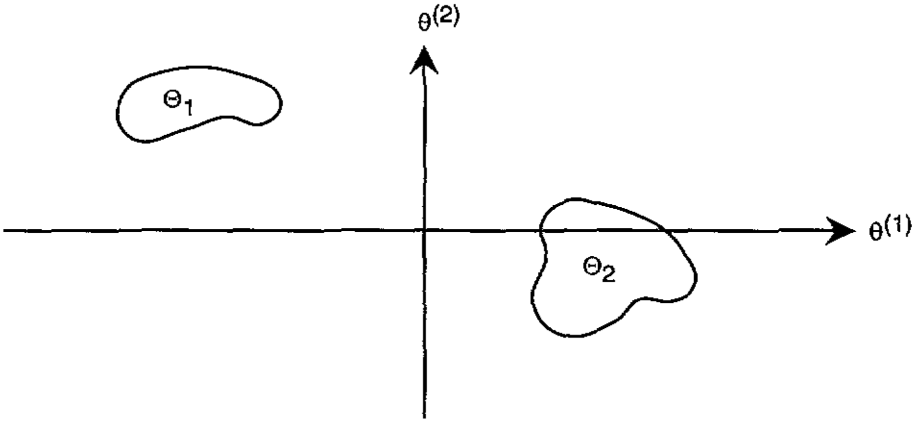

Figure 1. The disconnected support $\Theta = \Theta_1 \cup \Theta_2$ for the probability distribution implies that a Gibbs sampler with blocking $\left(\theta^{(1)}, \ \theta^{(2)}\right)$ will not have the probability distribution as its invariant distribution, for any starting value.

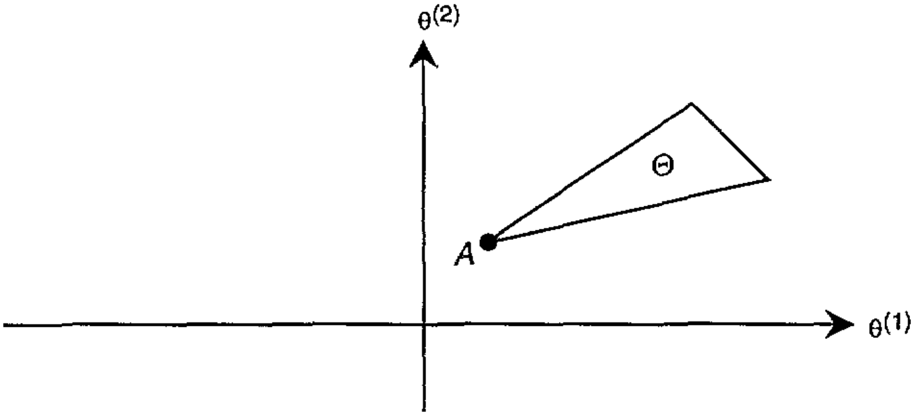

Figure 2. The probability density $p(\theta)$ is uniform on the closed set Θ and consequently is not lower semicontinuous at 0. The point A is absorbing for the Gibbs sampler with blocking $\left(\theta^{(1)}, \theta^{(2)}\right)$, so if $\theta_0 = A$ convergence will not occur.

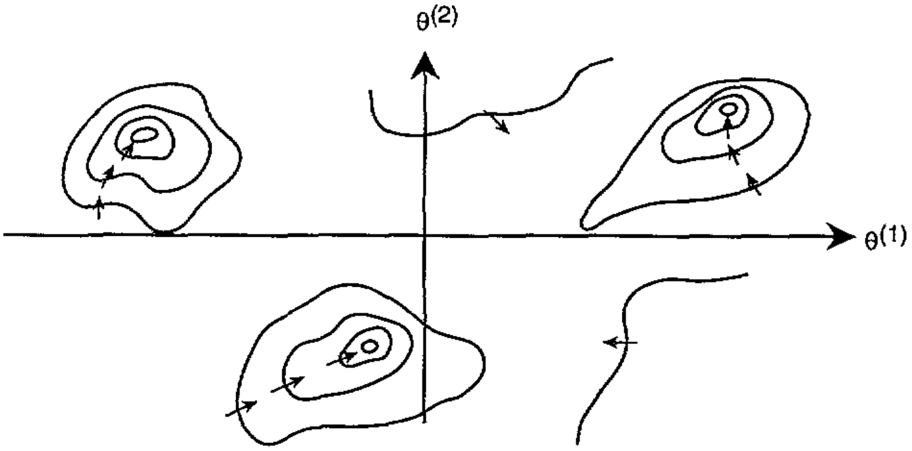

Figure 3. Iso-probability density contours of a multimodal bivariate distribution are shown. (Arrows indicate directions of increased density.) Given sufficiently steep gradients the Gibbs sampler will converge very slowly.