

# Using Simulation Methods for Bayesian Econometric Models: Inference, Development and Communication

John Geweke  
University of Minnesota and Federal Reserve Bank of Minneapolis

Original version: June, 1997  
Final version: December, 1998

## Abstract

This paper surveys the fundamental principles of subjective Bayesian inference in econometrics and their implementation using posterior simulation methods. The emphasis is on the combination of models and the development of predictive distributions. Moving beyond conditioning on a fixed number of completely specified models, the paper introduces subjective Bayesian tools for formal comparison of these models with as yet incompletely specified models. The paper then shows how posterior simulators can facilitate communication between investigators (for example econometricians) on the one hand and remote clients (for example decision makers) on the other, enabling clients to vary the prior distributions and functions of interest employed by investigators. A theme of the paper is the practicality of subjective Bayesian methods. To this end, it describes publicly available software for Bayesian inference, model development and communication, and provides illustrations using two simple econometric models.

JEL classification: Primary C15 ; Secondary C11

Keywords: Bayes factor; Diagnostic checking; Importance sampling; Markov chain Monte Carlo; Model development; Model specification; Prior distributions

©1997 by John Geweke. This document may be freely reproduced for educational and research purposes provided that (i) this copyright notice is included with each copy, (ii) no changes are made in the document, and (iii) copies are not sold, but retained for individual use or distributed free of charge.

This paper was originally prepared for the Australasian meetings of the Econometric Society, Melbourne, July 2-4, 1997. Financial support from NSF grant SBR-9600040 is gratefully acknowledged, as is fine research assistance from Hulya Eraslan and John Landon-Lane. Siddhartha Chib, William Griffiths, Michael Keane, Dale Poirier, and four anonymous referees provided helpful comments on earlier versions. My understanding of the issues in this paper has greatly benefited from discussions over the years with these and many other individuals, but misconceptions conveyed in this work are entirely my own. The views expressed herein are those of the author and not necessarily those of the Federal Reserve Bank of Minneapolis.

# 1. Introduction

Subjective uncertainty is a central concept in economic theory and applied economics. In economic theory, it characterizes the beliefs of economic agents about the state of their environment. In applied economics, subjective uncertainty describes the situation of investigators who assess competing models based on their implications for what might be observed, and the circumstances of decision makers who must act given limited information. With the application of the expected utility paradigm in increasingly richer environments explicit distributional assumptions have become common, but closed form analytical expressions for the distribution of observables are typically unobtainable. In this environment simulation methods — the representation of probability distributions by related finite samples — have become important tools in economic theory.

In applied economics the possibility of proceeding strictly analytically is also remote. Even in the simplest typical situation the investigator or decision maker must proceed knowing the observables which are random variables in models of behavior, but not knowing the specification of tastes and technology that the theorist takes as fixed. Bayesian inference formalizes the applied economics problem in exactly this way: given a distribution over competing models and the prediction of each model for observables, the distribution of competing models *conditional* on the observables is well defined. But the technical tasks in moving from even such well specified models and data to the conditional distribution over models are more daunting than those found in economic theory. In the past decade very substantial progress has been made in the development of simulation methods suited to this task. Section 2 of this article reviews the conditional distributions of interest to the investigator or decision maker. Section 3 describes how modern simulation methods permit access to these distributions, using some simple examples and publicly available software to illustrate the methods.

A central issue in any kind of inference, whether or not it is Bayesian or even explicitly based on probability theory of any kind, is that the simple paradigm of theory before measurement is oversimplified. The set of models which theorists and investigators have before them is constantly changing. Some models become fully developed with explicit predictions, others are no more than incomplete notions, and many are somewhere between these two extremes. The process by which some models become more fully developed, others receive little attention and still others are abandoned is driven in large part by data. Section 4 of this article sets forth recently developed numerical procedures for the explicit comparison of fully developed models. Section 5 turns to the practical but more difficult problem of the interaction between data and the development of models. It advances the

thesis that the process of model development is inherently Bayesian. It shows that this process can be implemented in a practical way using two new concepts -- the incomplete model, and limited information marginal likelihood. This process is illustrated in worked examples using public domain software.

The rigor of conditioning on what is known and working through the implications of explicit assumptions for what is unknown has both a rich yield and a substantial cost. The rich yield is the exact distribution of unobservables conditional on the data, rendered accessible by simulation methods. The cost is that models must provide the joint distribution of observables and unobservables explicitly. In part this cost is the real effort expended in formulating this explicit distribution. Perhaps a greater concern is that decision makers may not share in all the distributional assumptions that an investigator makes in this process. In Bayesian inference this concern has focused on the development of prior distributions of parameters, but usually the more serious problem is the restrictions on observables inherent in the parameterization of the model -- a problem faced by Bayesians, non-Bayesians, and those who would abandon formal probability theory altogether in inference. The last section of the article takes up simple, effective ways of simultaneously realizing the rich promise of explicitly Bayesian methods, and dealing with the desire of decision makers to change investigators' assumptions at low cost. The procedures are intimately related to simulation methods and rapid movement of large information sets over the internet. They are illustrated for some simple but realistic examples, using publicly available software.

## 2. Bayesian Inference

This section provides a brief overview of Bayesian inference with reference to its application in economics. The purpose is to set the contribution of simulation methods in an explicit context of concepts and notation. Every attempt has been made to distill a large literature in statistics to what is essential to Bayesian inference as it is usually applied to economic problems. If this endeavor has been successful then this section also provides a sufficient introduction for econometricians with little or no grounding in Bayesian methods to appreciate some of the contributions, both realized and potential, of simulation methods to economic science.

Most of the material here is standard, reflecting much more comprehensive treatments including Jeffreys (1939, 1961), Zellner (1971), Berger (1985), Bernardo and Smith (1994) and Poirier (1995). At two junctures the exposition departs from the usual development. The first is the concept of a complete model (Section 2.1), which is the specification of a proper predictive distribution over an explicit class of events. This concept can be a clarifying analytical device. It also sets the foundation for the concept of an incomplete model (Section 5.2) which provides a proper Bayesian interpretation of the work of economists in improving their models and formulating new ones.

The other deviation from the standard treatment is the decomposition of the marginal likelihood in terms of predictive densities (Section 2.3). This development was first provided explicitly by Geisel (1977) but has largely been ignored in the subsequent literature. The decomposition is the quantitative expression of the fact that predictive power is the scientifically relevant test of the validity of a hypothesis (Jeffreys, 1939; Friedman, 1953).

This review concentrates entirely on exact, finite sample methods. As is the case in non-Bayesian statistics, given suitable regularity conditions there exist useful asymptotic approximations to the exact, finite sample results. Bernardo and Smith (1994, Section 5.3) provide an accessible introduction to these results. Asymptotic methods are complementary to, rather than prerequisite for, the posterior simulation methods taken up subsequently in Section 3.

### 2.1 Basic concepts and notation

Bayesian inference takes place in the context of one or more parametric econometric models. Let  $\mathbf{y}_t$  denote a  $p \times 1$  vector of observable random vectors over a sequence of discrete time units  $t = 1, 2, \dots, K$ . The history of the sequence  $\{\mathbf{y}_t\}$  at time  $t$  is given by  $\mathbf{Y}_t = \{\mathbf{y}_s\}_{s=1}^t \in \Psi_t$ ;  $\mathbf{Y}_0 = \{\emptyset\}$ . A *model*,  $A$ , specifies a corresponding sequence of

probability density functions  $p(\mathbf{y}_t|\mathbf{Y}_{t-1},\theta,A)$  in which  $\theta$  is a  $k \times 1$  vector of unknown parameters,  $\theta \in \Theta \subseteq \mathfrak{R}^k$ , and  $A$  denotes the model.<sup>1</sup> In this section we shall condition on a single model but subsequently, in Section 2.3, several models will be entertained simultaneously.

The probability density function (p.d.f.) of  $\mathbf{Y}_T$ , conditional on the model  $A$  and parameter vector  $\theta$ , is  $p(\mathbf{Y}_T|\theta,A) = \prod_{t=1}^T p(\mathbf{y}_t|\mathbf{Y}_{t-1},\theta,A)$ . Conditional on observed  $\mathbf{Y}_T$ , the *likelihood function* is any function  $L(\theta;\mathbf{Y}_T,A) \propto p(\mathbf{Y}_T|\theta,A)$ . If the model specifies that the  $\mathbf{y}_t$  are independent and identically distributed then  $p(\mathbf{y}_t|\mathbf{Y}_{t-1},\theta,A) = p(\mathbf{y}_t|\theta,A)$  and  $p(\mathbf{Y}_T|\theta,A) = \prod_{t=1}^T p(\mathbf{y}_t|\theta,A)$ . 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.

If in addition the model  $A$  also provides the distribution of  $\theta$ , then it also provides the joint distribution of  $\theta$  and  $\mathbf{Y}_T$ . In particular, if  $p(\theta|A)$  denotes the *prior density* then

$$(2.1.1) \quad p(\mathbf{Y}_T,\theta|A) = p(\theta|A)\prod_{t=1}^T p(\mathbf{y}_t|\mathbf{Y}_{t-1},\theta,A) = p(\theta|A)p(\mathbf{Y}_T|\theta,A).$$

But we also have

$$(2.1.2) \quad p(\mathbf{Y}_T,\theta|A) = p(\theta|\mathbf{Y}_T,A)p(\mathbf{Y}_T|A),$$

in which

$$(2.1.3) \quad p(\mathbf{Y}_T|A) = \int_{\Theta} p(\mathbf{Y}_T|\theta,A)p(\theta|A)d\nu(\theta)$$

is the *marginal likelihood*<sup>2</sup> of model  $A$  and

$$p(\theta|\mathbf{Y}_T,A) = p(\mathbf{Y}_T|\theta,A)p(\theta|A)/p(\mathbf{Y}_T|A) \propto p(\mathbf{Y}_T|\theta,A)p(\theta|A)$$

is the *posterior density* of  $\theta$  in model  $A$ , so long as

$$(2.1.4) \quad \int_{\Theta} p(\mathbf{Y}_T|\theta,A)p(\theta|A)d\nu(\theta)$$

is absolutely convergent. This last condition is typically, but not necessarily, satisfied and easy to verify. For example, boundedness of the likelihood function  $p(\mathbf{Y}_T|\theta,A)$  in  $\theta$  is sufficient, since  $\int_{\Theta} p(\theta|A)d\nu(\theta) = 1$ . But if the likelihood function is unbounded, it is vital to confirm the absolute convergence of (2.1.4). Expressions (2.1.1) and (2.1.2) are central, either explicitly or implicitly, to scientific learning. The former is used to express the reduction of reality to  $\theta$  inherent in the model  $A$ , and the latter is used to learn about reality from the perspective of this particular simplification. This section outlines the basic principles of the explicit, or Bayesian, approach to learning.

---

<sup>1</sup>Throughout,  $p(\cdot)$  denotes a generic probability density function with respect to a measure  $d\nu(\cdot)$  and  $P(\cdot)$  a generic cumulative distribution function. The conditioning set makes clear the specific distribution or density intended. The measure  $d\nu(\cdot)$  permits continuous, discrete, and mixed random variables.

<sup>2</sup>This terminology dates at least to Raiffa and Schlaifer (1961, Section 2.1) which also treats these topics.

In addition to the data density  $p(\mathbf{Y}_T|\theta, A)$  and the prior density  $p(\theta|A)$ , a model also specifies a density  $p(\omega|\mathbf{Y}_T, \theta, A)$  for a vector of interest  $\omega \in \Omega \subseteq \mathbb{R}^1$ . This vector represents entities the model is intended to describe. Whereas  $\theta$  is specific to  $A$ ,  $\omega$  remains the same across models. For example, suppose one model specifies a Cobb-Douglas production function  $\theta_2 y_{1t}^{\theta_1} y_{2t}^{(1-\theta_1)}$  for two inputs  $y_{1t}$  and  $y_{2t}$ . Then the technical rate of substitution,  $\omega$ , is  $\omega = \theta_1/(1-\theta_1)$ . If a second model specifies a constant elasticity of substitution (CES) production function  $(\theta_1 + \theta_2 y_{1t}^{\theta_4} + \theta_3 y_{2t}^{\theta_4})^{1/\theta_4}$ , then the same technical rate of substitution,  $\omega$ , is  $\omega = \theta_4(y_{1t}/y_{2t})^{(\theta_4-1)}(\theta_2/\theta_3)$ . In each case the mapping from  $\theta$  to  $\omega$  is deterministic:  $p(\omega|\mathbf{Y}_T, \theta, A)$  puts unit mass on a single value of  $\omega$ .

As a second example, suppose that one model specifies a first order stationary autoregressive process for  $y_t$ ,  $(y_t - \theta_1) = \theta_2(y_{t-1} - \theta_1) + \varepsilon_t$  with  $\varepsilon_t \stackrel{iid}{\sim} N(0, \theta_3)$ . If  $\omega' = (y_{T+1}, y_{T+2})$ , the first two post-sample observations, then  $p(\omega|\theta, \mathbf{Y}_T, A)$  is a bivariate normal density with mean and variance

$$\begin{bmatrix} \theta_1 + \theta_2(y_T - \theta_1) \\ \theta_1 + \theta_2^2(y_T - \theta_1) \end{bmatrix} \text{ and } \theta_3 \begin{bmatrix} 1 & \theta_2^2 \\ \theta_2^2 & 1 + \theta_2^2 \end{bmatrix},$$

respectively. If a second model specifies a second order stationary autoregressive process for  $y_t$ ,  $(y_t - \theta_1) = \theta_2(y_{t-1} - \theta_1) + \theta_3(y_{t-2} - \theta_1) + \varepsilon_t$  with  $\varepsilon_t \stackrel{iid}{\sim} N(0, \theta_4)$ , then  $p(\omega|\theta, \mathbf{Y}_T, A)$  is again a bivariate normal density, but with mean and variance

$$\begin{bmatrix} \theta_1 + \theta_2(y_T - \theta_1) + \theta_3(y_{T-1} - \theta_1) \\ \theta_1(1 + \theta_2) + \theta_3(1 + \theta_2)(y_T - \theta_1) + \theta_2\theta_3(y_{T-1} - \theta_1) \end{bmatrix} \text{ and } \theta_4 \begin{bmatrix} 1 & \theta_2^2 \\ \theta_2^2 & 1 + \theta_2^2 \end{bmatrix},$$

respectively.

Since  $p(\omega|\mathbf{Y}_T, \theta, A)$  implies marginal distributions for subvectors of  $\omega$ , one need not explicitly elaborate all of  $\omega$ . Indeed, much scientific discourse can be interpreted as specification of  $\omega$ . A *complete model* consists of three components:  $p(\mathbf{Y}_T|\theta, A)$ ,  $p(\theta|A)$  and  $p(\omega|\mathbf{Y}_T, \theta, A)$ .

Without loss of generality, let the objective of inference when there is one model be

$$(2.1.5) \quad E[h(\omega)|\mathbf{Y}_T, A],$$

for suitably chosen  $h(\cdot)$ . This formulation includes several special cases of interest. If a hypothesis restricts  $\theta$  to a set  $\Theta_0$  then by taking  $h(\omega) = \chi_{\Theta_0}(\theta)$ , we have  $E[h(\omega)|\mathbf{Y}_T, A] = P(\theta \in \Theta_0|\mathbf{Y}_T, A)$ , the posterior probability that the hypothesis is true.<sup>3</sup> To illustrate, suppose that in the first example we wish to inquire whether the technical rate of substitution exceeds one when  $y_{1t} = y_{2t}$ . For the Cobb-Douglas production function, take  $\Theta_0 = \{\theta_1 : \theta_1 > .5\}$ , and for the CES production function take  $\Theta_0 = \{\theta_2, \theta_3, \theta_4 : \theta_4\theta_2/\theta_3 > 1\}$ .

<sup>3</sup>Here and throughout,  $\chi_s(z)$  is the indicator function  $\chi_s(z) = 1$  if  $z \in S$  and  $\chi_s(z) = 0$  if  $z \notin S$ .

Note that in each case there is a nuisance parameter:  $\theta_2$  for Cobb-Douglas and  $\theta_1$  for CES. Here, and in general, nuisance parameters pose no particular difficulties.

Another important class of cases arises from prediction problems,  $\omega' = (\mathbf{y}_{T+1,K}, \mathbf{y}_{T+f})$ . Through the appropriate choice of  $h(\omega)$  this category includes expected values, turning point probabilities, and predictive intervals. In the time series example just set forth, suppose that  $y_{T-2} < y_{T-1} < y_T$ . If a turning point at time  $t$  is said to occur if  $y_{t-2} < y_{t-1} < y_t > y_{t+1} > y_{t+2}$ , then a turning point at time  $T$  is the set of events  $\Omega^* = \{\omega : \omega_2 < \omega_1 < y_T\}$ . Hence for  $h(\omega) = \chi_{\Omega^*}(\omega)$ ,  $E[h(\omega)|\mathbf{Y}_T, A]$  is the probability of a turning point at time  $T$ , where  $T$  is the end of the sample.

Yet another useful class of functions is  $h(\omega) = L(a_1, \omega) - L(a_2, \omega)$ , in which  $L(a, \omega)$  denotes the loss incurred if action  $a$  is taken and then the realization of the vector of interest is  $\omega$ . To examine a specific case, suppose that in the second example  $y_t$  is the logarithm of tax revenue in period  $t$ . A policy maker must either commit ( $a_1$ ) or not commit ( $a_2$ ) to a program which utilizes tax revenues in periods  $T+1$  and  $T+2$ . Then the policy maker's loss function  $L(a, \omega)$  might be monotone decreasing in  $\omega_1 + \omega_2$  for  $a = a_1$ , and monotone increasing in  $\omega_1 + \omega_2$  for  $a = a_2$ , and consequently  $h(\omega)$  is monotone decreasing in  $\omega_1 + \omega_2$ . The solution of the decision problem is to commit to the project if  $E[h(\omega)|\mathbf{Y}_T, A] < 0$  and not commit if  $E[h(\omega)|\mathbf{Y}_T, A] > 0$ .

The posterior moment (2.1.5) can be expressed as

$$(2.1.6) \quad E[h(\omega)|\mathbf{Y}_T, A] = \int_{\Theta} \int_{\Omega} h(\omega) p(\omega|\theta, \mathbf{Y}_T, A) p(\theta|\mathbf{Y}_T, A) d\nu(\omega) d\nu(\theta) \\ = \int_{\Theta} \int_{\Omega} h(\omega) p(\omega|\theta, \mathbf{Y}_T, A) p^*(\theta|\mathbf{Y}_T, A) d\nu(\omega) d\nu(\theta) / \int_{\Theta} p^*(\theta|\mathbf{Y}_T, A) d\nu(\theta)$$

where  $p^*(\theta|\mathbf{Y}_T, A) \propto p(\theta|\mathbf{Y}_T, A) \propto p(\theta|A) p(\mathbf{Y}_T|\theta, A)$  is any *posterior density kernel* for  $\theta$ .<sup>4</sup>

It clearly matters not which posterior kernel is used. However, the problem of evaluating integrals -- one in the numerator, the other in the denominator -- remains paramount.

The importance of verifying the absolute convergence of the integral in the denominator of the right side of (2.1.6) has already been noted. It is, of course, equally important to verify the absolute convergence of the numerator of (2.1.6). Together, both conditions are equivalent to the existence of the posterior moment (2.1.5). It is straightforward to verify these convergence conditions in the examples discussed above.

Many of these ideas can be illustrated in the standard linear model. For an observable  $T \times 1$  vector of dependent variables  $\mathbf{y}$  and  $T \times k$  matrix of fixed covariates<sup>5</sup>  $\mathbf{X}$ ,

<sup>4</sup>More generally, any nonnegative function proportional to a probability density is a density kernel.

<sup>5</sup>If instead  $\mathbf{X}$  is random with p.d.f.  $p(\mathbf{X}|\eta)$ ,  $p(\beta, h, \eta|A) = p(\beta, h|A) p(\eta|A)$  and  $p(\omega|\mathbf{y}, \mathbf{X}, \beta, h, \eta) = p(\omega|\mathbf{y}, \mathbf{X}, \beta, h)$ , then  $\mathbf{X}$  is ancillary and the analysis that follows still pertains. For further discussion of ancillarity see Bernardo and Smith (1994, Section 5.1.4). The condition of weak exogeneity in the econometrics literature (Engle *et al.*, 1983; Steel and Richard, 1991) is closely related.

$$(2.1.7) \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}; \quad \boldsymbol{\varepsilon}|\mathbf{X} \sim \mathbf{N}(\mathbf{0}, h^{-1}\mathbf{I}_T); \quad \text{rank}(\mathbf{X}) = k.$$

The parameter  $h$  is the *precision* of the i.i.d. disturbances,  $\varepsilon_{1,K}, \varepsilon_T$ ; it is the inverse of  $\text{var}(\varepsilon_i) = \sigma^2$ .<sup>6</sup> Consider the independent prior distributions for  $\boldsymbol{\beta}$  and  $h$ ,

$$(2.1.8) \quad \boldsymbol{\beta} \sim \mathbf{N}(\underline{\boldsymbol{\beta}}, \underline{\mathbf{H}}^{-1}),$$

$$(2.1.9) \quad \underline{s}^2 h \sim \chi^2(\underline{\nu}),$$

where  $\underline{\mathbf{H}}$  is a fixed precision matrix,  $\underline{\boldsymbol{\beta}}$  is a fixed mean vector, and  $\underline{s}^2$  and  $\underline{\nu}$  are fixed scalars. In any given application (2.1.8)-(2.1.9) is not necessarily an adequate expression of prior beliefs.<sup>7</sup> However, the specification in (2.1.8)-(2.1.9) has attractive analytical properties that will become clear in due course. Moreover, in many cases it is straightforward to modify the posterior distribution implied by the prior distributions (2.1.8)-(2.1.9) to express the posterior distributions corresponding to (2.1.7) and alternative prior distributions, using simple numerical methods described in Section 6.

From (2.1.8),

$$(2.1.10) \quad p(\boldsymbol{\beta}) = (2\pi)^{-k/2} |\underline{\mathbf{H}}|^{1/2} \exp\left[-.5(\boldsymbol{\beta} - \underline{\boldsymbol{\beta}})' \underline{\mathbf{H}}(\boldsymbol{\beta} - \underline{\boldsymbol{\beta}})\right].$$

and from (2.1.9)

$$(2.1.11) \quad p(h) = [2^{\underline{\nu}/2} \Gamma(\underline{\nu}/2)]^{-1} (\underline{s}^2)^{\underline{\nu}/2} h^{(\underline{\nu}-2)/2} \exp(-\underline{s}^2 h/2).$$

Since (2.1.7) is equivalent to the conditional data density

$$(2.1.12) \quad p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, h) = (2\pi)^{-T/2} h^{T/2} \exp\left[-.5h(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right],$$

a posterior density kernel is the product of (2.1.10), (2.1.11) and (2.1.12),

$$(2.1.13a) \quad (2\pi)^{-(T+k)/2} [2^{\underline{\nu}/2} \Gamma(\underline{\nu}/2)]^{-1}$$

$$(2.1.13b) \quad \cdot |\underline{\mathbf{H}}|^{1/2} (\underline{s}^2)^{\underline{\nu}/2}$$

$$(2.1.13c) \quad \cdot h^{(T+\underline{\nu}-2)/2} \exp(-\underline{s}^2 h/2)$$

$$(2.1.13d) \quad \cdot \exp\left\{-.5\left[(\boldsymbol{\beta} - \underline{\boldsymbol{\beta}})' \underline{\mathbf{H}}(\boldsymbol{\beta} - \underline{\boldsymbol{\beta}}) + h(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]\right\}.$$

To simplify this expression, complete the square in  $\boldsymbol{\beta}$  of the term in brackets in (2.1.13d), yielding

$$(\boldsymbol{\beta} - \underline{\boldsymbol{\beta}})' \underline{\mathbf{H}}(\boldsymbol{\beta} - \underline{\boldsymbol{\beta}}) + h(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})' \bar{\mathbf{H}}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) + Q,$$

where

$$(2.1.14) \quad \bar{\mathbf{H}} = \underline{\mathbf{H}} + h\mathbf{X}'\mathbf{X},$$

$$(2.1.15) \quad \bar{\boldsymbol{\beta}} = \bar{\mathbf{H}}^{-1}(\underline{\mathbf{H}}\underline{\boldsymbol{\beta}} + h\mathbf{X}'\mathbf{y}) = \bar{\mathbf{H}}^{-1}(\underline{\mathbf{H}}\underline{\boldsymbol{\beta}} + h\mathbf{X}'\mathbf{X}\mathbf{b}) \text{ and}$$

$$(2.1.16) \quad Q = h\mathbf{y}'\mathbf{y} + \underline{\boldsymbol{\beta}}' \underline{\mathbf{H}}\underline{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}' \bar{\mathbf{H}}\bar{\boldsymbol{\beta}}$$

<sup>6</sup>More generally, the precision of any random variable is the inverse of its variance.

<sup>7</sup>Nor is (2.1.7), necessarily. We return to this important question in greater depth in Sections 2.3 and 4.

$$= hvs^2 + (\mathbf{b} - \bar{\boldsymbol{\beta}})' h\mathbf{X}'\mathbf{X}(\mathbf{b} - \bar{\boldsymbol{\beta}}) + (\underline{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})' \underline{\mathbf{H}}(\underline{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}),$$

with  $\mathbf{b}$  denoting the coefficients in the ordinary least squares fit of  $\mathbf{y}$  to  $\mathbf{X}$ ,  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ ;  $s^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})/v$  and  $v = T - k$ . If (2.1.13) is interpreted as a function of  $\boldsymbol{\beta}$  only, that function must be a posterior density kernel for  $\boldsymbol{\beta}$  conditional on  $h$ , and our square completion shows that  $p(\boldsymbol{\beta}|h, \mathbf{y}, \mathbf{X}) \propto \exp\left\{-.5(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})' \bar{\mathbf{H}}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\right\}$ .

Consequently,

$$(2.1.17) \quad \boldsymbol{\beta}|(h, \mathbf{y}, \mathbf{X}) \sim N(\bar{\boldsymbol{\beta}}, \bar{\mathbf{H}}^{-1}).$$

Interpreting (2.1.13) as a function of  $h$  alone,

$$p(h|\boldsymbol{\beta}, \mathbf{y}, \mathbf{X}) \propto h^{(T+\underline{v}-2)/2} \exp\left\{-.5\left[\underline{s}^2 + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]h\right\}$$

and consequently,

$$(2.1.18) \quad \left[\underline{s}^2 + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]h \mid (\boldsymbol{\beta}, \mathbf{y}, \mathbf{X}) \sim \chi^2(T + \underline{v}).$$

The distributions in (2.1.8) and (2.1.9) are special cases of conditionally conjugate priors (to be defined shortly). They are attractive because they lead to the tractable results (2.1.17) and (2.1.18). Yet these results are not directly useful, for they do not provide distributions conditional only on the data and prior information. However they form the basis of an attractive simulation method discussed in Section 3.3.

In any application of the standard linear model the vector of interest  $\boldsymbol{\omega}$  is likely to include an as yet unobserved  $T^* \times 1$  vector  $\mathbf{y}^*$  corresponding to a situation in which it is hypothesized that  $\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$ ,  $\boldsymbol{\varepsilon}^* | \mathbf{X}^* \sim N(\mathbf{0}, h^{-1}\mathbf{I}_{T^*})$ . If  $\boldsymbol{\varepsilon}$  and  $\boldsymbol{\varepsilon}^*$  are conditionally independent given  $(\mathbf{X}, \mathbf{X}^*, \boldsymbol{\beta}, h)$  then  $\mathbf{y}^* | (\mathbf{X}^*, \mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, h) \sim N(\mathbf{X}^*\boldsymbol{\beta}, h^{-1}\mathbf{I}_{T^*})$  and it is straightforward to show  $\mathbf{y}^* | (\mathbf{X}^*, \mathbf{y}, \mathbf{X}, h) \sim N(\mathbf{X}^*\bar{\boldsymbol{\beta}}, \mathbf{X}^*\bar{\mathbf{H}}^{-1}\mathbf{X}^* + h^{-1}\mathbf{I}_{T^*})$ .

## 2.2 Conjugate and improper prior distributions

The prior distribution  $p(\boldsymbol{\theta}|A)$  is a representation of belief in the context of model  $A$ . In selecting a prior or data distribution, the richer the class of functional forms from which to choose the more adequate the representation of prior beliefs possible. On the other hand, the choice is constrained by the tractability of the posterior density  $p(\boldsymbol{\theta}|\mathbf{Y}_T, A) \propto p(\boldsymbol{\theta}|A)p(\mathbf{Y}_T|\boldsymbol{\theta}, A)$ , which is jointly determined by the choice of functional forms for the data density and prior density. The search for rich tractable classes of prior distributions may be formalized by considering classes of prior densities,  $p(\boldsymbol{\theta}|A) = p(\boldsymbol{\theta}|\boldsymbol{\gamma}, A)$ , where  $\boldsymbol{\gamma}$  is a parameter vector that indexes prior beliefs. We have used

this approach in the linear model: for example, the prior distribution  $\beta \sim N(\underline{\beta}, \underline{\mathbf{H}}^{-1})$  is indexed by  $\underline{\beta}$  and  $\underline{\mathbf{H}}$ .

Suppose the model  $p(\mathbf{Y}_T|\theta, A)$  has sufficient statistic  $\mathbf{s}_T = ((s_T)_{1,K}, (s_T)_r)' = \mathbf{s}_T(\mathbf{Y}_T)$ ,  $r$  is fixed as  $T$  varies, and  $(s_T)_1 = T$ . Then the *conjugate family of prior densities with respect to*  $p(\mathbf{Y}_T|\theta, A)$  is  $\{p(\theta|\gamma, A), \gamma \in \Gamma\}$ , where

$$p(\theta|\gamma, A) \propto p\left[(s_{\gamma_i})_j = \gamma_j (j = 2, K, r) | \theta, A\right]$$

and

$$\Gamma = \left\{ \gamma : \int_{\Theta} p\left[(s_{\gamma_i})_j = \gamma_j (j = 2, K, r) | \theta, A\right] d\theta < \infty \right\}.$$

The kernel of any conjugate prior density may be interpreted as a likelihood function corresponding to a data set  $\mathbf{Z}_{\gamma_i}$  with sufficient statistic  $\mathbf{s}'_{\gamma_i} = (\gamma_{2,K}, \gamma_r)$ . To the extent one can represent prior beliefs arising from notional data with the same probability density functional form as the actual data, a conjugate prior distribution will provide a good representation of belief. By construction  $p(\mathbf{Y}_T|\theta, A) \propto p^*(\mathbf{s}_T|\theta, A)$  and  $p(\theta|A) \propto p^*(\gamma|\theta, A)$ , where the proportionality is in  $\theta$ , and  $p^*(\mathbf{s}_T|\theta, A)$  and  $p^*(\gamma|\theta, A)$  have exactly the same functional form in  $\theta$ . Hence  $p(\theta|\mathbf{Y}_T, A) \propto p^*(\mathbf{s}_T|\theta, A)p^*(\gamma|\theta, A)$ . It is often the case that the functional form of  $p(\theta|\mathbf{Y}_T, A)$  is the same as that of  $p^*(\mathbf{s}_T|\theta, A)$ , and it is this feature that makes the posterior density tractable.<sup>8</sup>

To extend this idea let  $\theta' = (\theta'_1, \theta'_2)$  and fix  $\theta_2 = \theta_2^0$ . Suppose the model  $p(\mathbf{Y}_T|\theta_1, \theta_2 = \theta_2^0, A)$  has sufficient statistic  $\mathbf{s}_T^* = \mathbf{s}_T^*(\mathbf{Y}_T)$ ,  $r^*$  is fixed, and  $(s_T^*)_1 = T$ . Then the *conditionally conjugate family of prior densities with respect to*  $p(\mathbf{Y}_T|\theta_1, \theta_2 = \theta_2^0, A)$  is  $\{p(\theta_1|\gamma^*, A), \gamma^* \in \Gamma^*\}$ , with  $\Gamma^* = \left\{ \gamma^* : \int_{\Theta_1} p\left[(s_{\gamma_i^*})_j = \gamma_j^* (j = 2, K, r^*) | \theta_1, \theta_2 = \theta_2^0, A\right] d\theta_1 < \infty \right\}$  and  $p(\theta_1|\gamma^*, A) \propto p\left[(s_{\gamma_i^*})_j = \gamma_j^* (j = 2, K, r^*) | \theta_1, \theta_2 = \theta_2^0, A\right]$ .

The prior distributions (2.1.8) and (2.1.9) are conditionally conjugate, but not conjugate, in the linear model (2.1.7). In this example the prior density for  $\theta' = (\beta', h)$  is indexed by  $\gamma = \{\underline{\beta}, \underline{\mathbf{H}}, \underline{s}^2, \underline{\mathbf{v}}\}$ . In the linear model, because

$$(2.2.1) \quad p(\mathbf{y}|\mathbf{X}, \beta, h) \propto h^{T/2} \exp\left\{-.5h\left[ \mathbf{v}\mathbf{s}^2 + (\beta - \mathbf{b})' \mathbf{X}'\mathbf{X}(\beta - \mathbf{b}) \right]\right\},$$

<sup>8</sup>Indeed, one can begin with this property as the definition of conjugate; see Berger (1985, Section 4.2.2) and Poirier (1995, Section 6.7). The definition here is that used by Bernardo and Smith (1994, Section 5.2.1) and Zellner (1971, Section 2.3). For the exponential family of distributions (which includes the standard linear model) the two are equivalent (see Bernardo and Smith (1994), Proposition 5.4).

the vector  $\mathbf{s}_T = [T, \mathbf{b}, s^2, \mathbf{X}'\mathbf{X}]$  is a sufficient statistic.<sup>9</sup> Conditioning on  $h = h_0$ ,  $p(\mathbf{y}|\mathbf{X}, \beta) \propto \exp\left[-.5(\beta - \mathbf{b})' h_0 \mathbf{X}'\mathbf{X}(\beta - \mathbf{b})\right]$ . Since  $p(\beta) \propto \exp\left[-.5(\beta - \underline{\beta})' \underline{\mathbf{H}}(\beta - \underline{\beta})\right]$ , the prior density (2.1.10) is conditionally conjugate. Likewise conditioning on  $\beta = \beta_0$ ,  $p(\mathbf{y}|\mathbf{X}, h) \propto h^{T/2} \exp(-\bar{s}^2 h/2)$  where  $\bar{s}^2 = vs^2 + (\beta_0 - \mathbf{b})' \mathbf{X}'\mathbf{X}(\beta_0 - \mathbf{b})$ . Hence the prior density (2.1.11),  $p(h) \propto h^{(v-2)/2} \exp(-\bar{s}^2 h/2)$ , is conditionally conjugate.

In many instances posterior moments (2.1.5) continue to be well defined, as a mathematical formality, even if  $p^*(\theta|A)$  is not the kernel of any probability density function. Particular interest focuses on the case in which  $p^*(\theta|A) \geq 0 \forall \theta \in \Theta$  but  $\int_{\Theta} p^*(\theta|A) d\nu(\theta)$  is divergent. Such a function is said to be the density kernel of an *improper prior distribution*. The kernel  $p^*(\theta|A)$  may often be constructed by considering a sequence of models  $A_1, A_2, \dots, A_K$  that differ only in the specification of the prior density  $p(\theta|A_j)$  and not in the data density or in the conditional distribution of the vector of interest. Suppose the limit of kernels of prior density functions,  $p^*(\theta|A_j)$ , has the property

$$(2.2.2) \quad \lim_{j \rightarrow \infty} \int_{\Theta} \int_{\Omega} h(\omega) p(\omega|\theta, \mathbf{Y}_T, A) p^*(\theta|\mathbf{Y}_T, A_j) d\nu(\omega) d\nu(\theta) / \int_{\Theta} p^*(\theta|\mathbf{Y}_T, A_j) d\nu(\theta) \\ = \int_{\Theta} \int_{\Omega} h(\omega) p(\omega|\theta, \mathbf{Y}_T, A) p^*(\theta|\mathbf{Y}_T, A) d\nu(\omega) d\nu(\theta) / \int_{\Theta} p^*(\theta|\mathbf{Y}_T, A) d\nu(\theta).$$

In the last expression, if the denominator and numerator are absolutely convergent, then  $\lim_{j \rightarrow \infty} E[h(\omega)|\mathbf{Y}_T, A_j] = E[h(\omega)|\mathbf{Y}_T, A]$  may be interpreted as the posterior expectation of  $h(\omega)$  in a complete model with data density  $p(\mathbf{Y}_T|A, \theta)$  and improper prior density with kernel  $p^*(\theta|A)$ . Verifying the absolute convergence conditions can be substantially more difficult for improper priors than for proper priors: in particular, a bounded likelihood function no longer suffices for absolute convergence of the integrals in the denominator of (2.2.2).

As an example in the context of the standard linear model, consider the sequence of prior distributions  $\beta|A_j \sim N(\underline{\beta}, j\underline{\mathbf{H}}^{-1})$  ( $j = 1, 2, \dots, K$ ) conditional on a known value of the disturbance precision  $h$ . A corresponding sequence of kernels is  $p^*(\beta|A_j) = \exp\left[-.5(\beta - \underline{\beta})' j^{-1} \underline{\mathbf{H}}(\beta - \underline{\beta})\right]$ ;  $\lim_{j \rightarrow \infty} p^*(\beta|A_j) = p^*(\beta|A) = 1 \forall \beta$ . The corresponding sequence of posterior distributions is  $\beta|(\mathbf{Y}_T, A) \sim N[\bar{\beta}_j, (\bar{\mathbf{H}}_j)^{-1}]$ , with  $\bar{\mathbf{H}}_j = h\mathbf{X}'\mathbf{X} + j^{-1} \underline{\mathbf{H}}$ ,  $\bar{\beta}_j = \bar{\mathbf{H}}_j^{-1} [h\mathbf{X}'\mathbf{X}\mathbf{b} + j^{-1} \underline{\mathbf{H}}\underline{\beta}]$ . Hence

<sup>9</sup>This follows from the Neyman factorization criterion (Bernardo and Smith, 1994, Section 4.5.2). Less formally, from (2.2.1) it is clear that we only need to know  $\mathbf{s}_T$  to write the likelihood function for  $\beta$  and  $h$ .

$$\begin{aligned}
p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}, h, A_j) &\propto \exp\left[-.5(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_j)' \bar{\mathbf{H}}_j(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}_j)\right] \\
&\rightarrow \exp\left[-.5(\boldsymbol{\beta} - \mathbf{b})' h\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \mathbf{b})\right] \propto p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, h, A) = p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, h, A) p^*(\boldsymbol{\beta}|A).
\end{aligned}$$

The last line shows that the limiting posterior distribution could also have been achieved by carrying out a formal analysis using the “prior density kernel”  $p^*(\boldsymbol{\beta}|A)$ .

It is important to note that while posterior moments (2.1.5) may continue to be defined equivalently as mathematical formalities and as the limit of posterior moments under a sequence of prior distributions, an improper prior distribution and a data density do not together provide a joint distribution of parameters and data. In particular, under a sequence of proper prior distributions  $p(\boldsymbol{\theta}|A_j)$  converging to the improper prior distribution,  $\lim_{j \rightarrow \infty} p(\mathbf{Y}_T|A_j)$  is undefined. To see this important point intuitively note that if  $p(\boldsymbol{\theta}|A)$  is a proper density, one can work out the implications of the model for the data through simulation: first draw  $\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|A)$  and then  $\mathbf{Y}_T \sim p(\mathbf{Y}_T|\boldsymbol{\theta}, A)$ . If  $p(\boldsymbol{\theta}|A)$  is improper this cannot be done.<sup>10</sup>

### 2.3 Model comparison and combination

Often one has under consideration several complete models, say  $A_{1,K}, A_J$ :

$$p(\boldsymbol{\theta}_j|A_j) (\boldsymbol{\theta}_j \in \Theta_j), p(\mathbf{Y}_T|\boldsymbol{\theta}_j, A_j), p(\boldsymbol{\omega}|\mathbf{Y}_T, \boldsymbol{\theta}_j, A_j) \quad (j = 1, K, J).$$

The numbers of parameters in the models need not be the same, and various models may or may not nest one another. If we assign prior probabilities  $P(A_j)$  ( $j = 1, K, J$ ) to the respective models, with  $\sum_{j=1}^J P(A_j) = 1$ , then there is a complete probability structure for  $\{A_j, \boldsymbol{\theta}_j\}_{j=1}^J$ ,  $\mathbf{Y}_T$  and  $\boldsymbol{\omega}$ . There is no essential conceptual distinction between model and prior, since one could just as well regard the entire collection as the model, with  $\{P(A_j), p_j(\boldsymbol{\theta}_j|A_j)\}_{j=1}^J$  as the characterization of the prior distribution. At an operational level the distinction is usually clear and useful in that one may undertake the essential computations one model at a time.

Suppose that the posterior moment (2.1.5) is ultimately of interest. The formal solution is

$$(2.3.1) \quad E[h(\boldsymbol{\omega})|\mathbf{Y}_T] = \sum_{j=1}^J E[h(\boldsymbol{\omega})|\mathbf{Y}_T, A_j] P(A_j|\mathbf{Y}_T),$$

known as *model averaging*. Clearly  $E[h(\boldsymbol{\omega})|\mathbf{Y}_T, A_j]$  is given by (2.1.6) with  $A = A_j$ .

There is nothing new in this part of (2.3.1). From Bayes' rule,

---

<sup>10</sup>We return to this use of a proper prior distribution in Section 5.2.

$$\begin{aligned}
(2.3.2) \quad P(A_j|\mathbf{Y}_T) &= P(A_j)p(\mathbf{Y}_T|A_j)/\sum_{j=1}^J P(A_j)p(\mathbf{Y}_T|A_j) \\
&= P(A_j)\int_{\Theta_j} p(\mathbf{Y}_T|\theta_j, A_j)p(\theta_j|A_j)d\nu(\theta_j)/\sum_{j=1}^J P(A_j)p(\mathbf{Y}_T|A_j) \\
&\propto P(A_j)\int_{\Theta_j} p(\mathbf{Y}_T|\theta_j, A_j)p(\theta_j|A_j)d\nu(\theta_j) = P(A_j)p(\mathbf{Y}_T|A_j),
\end{aligned}$$

where  $p(\mathbf{Y}_T|A_j) = \int_{\Theta_j} p(\mathbf{Y}_T|\theta_j, A_j)p(\theta_j|A_j)d\nu(\theta_j)$  is the marginal likelihood of model  $j$ , consistent with the definition in (2.1.3) in Section 2.1. Notice it is important that the properly normalized prior and properly normalized data density, and not arbitrary kernels of these densities, be used in forming the marginal likelihood.

Model averaging thus involves three steps. First, obtain the posterior moments (2.1.6) corresponding to each model. Second, obtain the relative values of  $P(A_j|\mathbf{Y}_T)$  from (2.3.2). Finally, obtain the posterior moment using (2.3.1) which now only involves simple arithmetic, recognizing that  $\sum_{j=1}^J P(A_j|\mathbf{Y}_T) = 1$ . Variation of the prior model probabilities  $P(A_j)$  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. On the other hand, the questions of whether to introduce new models, and the formulation of new models, are more difficult. We shall return to these points in Section 5.

From (2.3.2), for any pair of models  $A_j$  and  $A_k$ ,

$$(2.3.3) \quad P(A_j|\mathbf{Y}_T)/P(A_k|\mathbf{Y}_T) = [P(A_j)/P(A_k)] \left[ p(\mathbf{Y}_T|A_j)/p(\mathbf{Y}_T|A_k) \right].$$

This ratio of probabilities is the *posterior odds ratio* in favor of model  $j$  versus model  $k$ . It is invariant with respect to the addition and deletion of models from the set  $\{A_j\}_{j=1}^J$  under consideration, so long as the prior probabilities  $\{P(A_j)\}_{j=1}^J$  are changed in a logically consistent fashion — that is, ratios  $P(A_j)/P(A_k)$  remain unchanged for all included models.<sup>11</sup> The posterior odds ratio is expressed in (2.3.3) as the product of the *prior odds ratio* in favor of model  $j$  versus model  $k$ ,  $P(A_j)/P(A_k)$ , and the *Bayes factor* in favor of model  $j$  versus model  $k$ ,  $p(\mathbf{Y}_T|A_j)/p(\mathbf{Y}_T|A_k)$ .

In the case of the standard linear model it is straightforward to work out the marginal likelihood and Bayes factors if  $h$  is fixed. The product of the properly normalized prior and data densities is

---

<sup>11</sup>This property is analogous to the independence of irrelevant alternatives in the qualitative choice literature; see Poirier (1997).

$$(2.3.4) \quad p(\beta)p(\mathbf{y}|\mathbf{X}, \beta, h) = (2\pi)^{-(T+k)/2} h^{T/2} |\mathbf{H}|^{1/2} \cdot \exp\left\{-.5\left[h(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + (\beta - \underline{\beta})' \mathbf{H}(\beta - \underline{\beta})\right]\right\}.$$

The term in brackets may be expressed as

$$(2.3.5) \quad (\beta - \bar{\beta})' \bar{\mathbf{H}}(\beta - \bar{\beta}) + Q,$$

with  $\bar{\beta}$ ,  $\bar{\mathbf{H}}$  and  $Q$  as defined in (2.1.14)-(2.1.16). Substituting (2.3.5) in (2.3.4), the marginal likelihood is

$$(2.3.6) \quad \int_{\mathbb{R}^k} p(\beta)p(\mathbf{y}|\mathbf{X}, \beta, h) d\beta = (2\pi)^{-(T+k)/2} h^{T/2} |\mathbf{H}|^{1/2} \int_{\mathbb{R}^k} \exp\left\{-.5\left[(\beta - \bar{\beta})' \bar{\mathbf{H}}(\beta - \bar{\beta}) + Q\right]\right\} d\beta = (2\pi)^{-T/2} h^{T/2} |\mathbf{H}|^{1/2} |\bar{\mathbf{H}}|^{-1/2} \exp(-Q/2) = (2\pi)^{-T/2} h^{T/2} \left(\frac{|\mathbf{H}|}{|\bar{\mathbf{H}}|}\right)^{1/2} \cdot \exp\left\{-.5\left[hv s^2 + (\mathbf{b} - \bar{\mathbf{b}})' h\mathbf{X}'\mathbf{X}(\mathbf{b} - \bar{\mathbf{b}}) + (\beta - \bar{\beta})' \mathbf{H}(\beta - \bar{\beta})\right]\right\}.$$

From the last expression it is apparent that the marginal likelihood of a linear model depends on more than the least squares fit of  $\mathbf{y}$  to  $\mathbf{X}$ , which is measured by the sum of squared residuals  $vs^2$ . It also depends on the squared Euclidean distance of the least squares fit  $\mathbf{b}$  from the posterior mean  $\bar{\mathbf{b}}$ , using the data-based norm  $h\mathbf{X}'\mathbf{X}$ ; the squared distance of the prior mean  $\underline{\beta}$  from the posterior mean  $\bar{\beta}$ , using the prior-based norm  $\mathbf{H}$ ; and the fraction of posterior precision accounted for by prior precision, as measured by  $|\mathbf{H}|/|\bar{\mathbf{H}}|$ .

Expression (2.3.2) shows that the marginal likelihood of model  $j$ ,  $p(\mathbf{Y}_T|A_j)$ , is the measure of how well model  $A_j$  predicted the observed data  $\mathbf{Y}_T$  that is relevant for the comparison of model  $j$  with any other models. In fact there is a more formal link between the marginal likelihood of a model, and the adequacy of the model's predictions, that underscores the predictive interpretation of  $p(\mathbf{Y}_T|A_j)$ .<sup>12</sup> To establish this link, first consider the distribution of  $\mathbf{y}_{u+1, K}, \mathbf{y}_t$  conditional on  $\mathbf{Y}_u$  and model  $j$ ,

$$(2.3.7) \quad p(\mathbf{y}_{u+1, K}, \mathbf{y}_t | \mathbf{Y}_u, A_j) = \int_{\Theta_j} p(\theta_j | \mathbf{Y}_u, A_j) \prod_{s=u+1}^t p(\mathbf{y}_s | \mathbf{Y}_{s-1}, \theta_j, A_j) d\nu(\theta_j).$$

As a function of  $\mathbf{y}_{u+1, K}, \mathbf{y}_t$ , after observing  $\mathbf{Y}_u$  and before observing  $\mathbf{y}_{u+1, K}, \mathbf{y}_t$ , expression (2.3.7) is the *predictive density* of  $\mathbf{y}_{u+1, K}, \mathbf{y}_t$  conditional on  $\mathbf{Y}_u$  and model  $A_j$ . Following the observation of  $\mathbf{y}_{u+1, K}, \mathbf{y}_t$ , it is a real number known as the *predictive*

<sup>12</sup>The formal demonstration that follows dates at least from Geisel (1977), but the more recent literature has largely ignored Geisel's result. (Thanks to Jacek Osiewalski for bringing Geisel's thesis to my attention.)

likelihood of  $\mathbf{y}_{u+1, \mathcal{K}}, \mathbf{y}_t$  conditional on  $\mathbf{Y}_u$  and model  $A_j$ . Note that  $p(\mathbf{y}_{1, \mathcal{K}}, \mathbf{y}_t | \mathbf{Y}_0, A_j) = p(\mathbf{Y}_t | A_j)$ , since  $\mathbf{Y}_0 = \{\emptyset\}$ . Substituting for the posterior density in (2.3.7),

$$\begin{aligned} & p(\mathbf{y}_{u+1, \mathcal{K}}, \mathbf{y}_t | \mathbf{Y}_u, A_j) \\ &= \int_{\Theta_j} \left\{ \frac{p(\theta_j | A_j) \prod_{s=1}^u p(\mathbf{y}_s | \mathbf{Y}_{s-1}, \theta_j, A_j)}{\int_{\Theta_j} p(\theta_j | A_j) \prod_{s=1}^u p(\mathbf{y}_s | \mathbf{Y}_{s-1}, \theta_j, A_j) d\nu(\theta_j)} \right\} \prod_{s=u+1}^t p(\mathbf{y}_s | \mathbf{Y}_{s-1}, \theta_j, A_j) d\nu(\theta_j) \\ &= \frac{\int_{\Theta_j} p(\theta_j | A_j) \prod_{s=1}^t p(\mathbf{y}_s | \mathbf{Y}_{s-1}, \theta_j, A_j) d\nu(\theta_j)}{\int_{\Theta_j} p(\theta_j | A_j) \prod_{s=1}^u p(\mathbf{y}_s | \mathbf{Y}_{s-1}, \theta_j, A_j) d\nu(\theta_j)} = \frac{p(\mathbf{Y}_t | A_j)}{p(\mathbf{Y}_u | A_j)}. \end{aligned}$$

Hence for any  $0 \leq u = s_0 < s_1 < \mathcal{K} < s_q = t$ , we have

$$\begin{aligned} (2.3.8) \quad p(\mathbf{y}_{u+1, \mathcal{K}}, \mathbf{y}_t | \mathbf{Y}_u, A_j) &= \frac{p(\mathbf{Y}_{s_1} | A_j)}{p(\mathbf{Y}_{s_0} | A_j)} \cdot \frac{p(\mathbf{Y}_{s_2} | A_j)}{p(\mathbf{Y}_{s_1} | A_j)} \cdot \dots \cdot \frac{p(\mathbf{Y}_{s_q} | A_j)}{p(\mathbf{Y}_{s_{q-1}} | A_j)} \\ &= \prod_{\tau=1}^q p(\mathbf{y}_{s_{\tau-1}+1, \mathcal{K}}, \mathbf{y}_{s_\tau} | \mathbf{Y}_{s_{\tau-1}}, A_j). \end{aligned}$$

This decomposition shows that the marginal likelihood ( $u = 0, t = T$ ) summarizes the out-of-sample prediction record of the model as expressed in the predictive likelihoods  $p(\mathbf{Y}_T | A_j) = \prod_{\tau=1}^q p(\mathbf{y}_{s_{\tau-1}+1, \mathcal{K}}, \mathbf{y}_{s_\tau} | \mathbf{Y}_{s_{\tau-1}}, A_j)$ . In the sense made precise by (2.3.8) and the use of  $p(\mathbf{Y}_T | A_j)$  in posterior model probability and model averaging, there is no distinction between a model's adequacy and its out-of-sample prediction record.<sup>13</sup>

Hypothesis testing is the problem of choosing one model from several. In the context of model combination this problem is somewhat artificial, but nonetheless it may be cast as a formal Bayesian decision problem. With no real loss of generality assume there are only two models in the choice set. Treating model choice as a Bayes action, suppose that the loss incurred in choosing model  $i$  depends only on which model is true, and so may be denoted  $L(i|j)$ . Further, suppose that  $L(i|i) = 0$  and  $L(i|j) > 0$  ( $j \neq i$ ). Given the data  $\mathbf{Y}_T$  the expected posterior loss from choosing model  $i$  is  $P(A_j | \mathbf{Y}_T) L(i|j)$  ( $j \neq i$ ) and so the Bayes action, based on the criterion of minimizing expected posterior loss, is to choose model 1 if

---

<sup>13</sup>The decomposition (2.3.8) may be interpreted as a formal expression of Milton Friedman's well known identification of a model's evaluation with its predictive performance: "Theory is to be judged by its predictive power ... the only relevant test of the *validity* of a hypothesis is comparison of its predictions with experience" (Friedman, 1953, pp. 8-9; emphasis in original). There are striking similarities between Friedman (1953) and Jeffreys (1939, 1961). The third edition (Jeffreys, 1961) contains, in Chapter 1, essentially the results presented here for the very special case of deterministic dichotomous outcomes.

$$(2.3.9) \quad \frac{P(A_1|\mathbf{Y}_T)}{P(A_2|\mathbf{Y}_T)} = \frac{P(A_1)p(\mathbf{Y}_T|A_1)}{P(A_2)p(\mathbf{Y}_T|A_2)} > \frac{L(1|2)}{L(2|1)}.$$

The value  $L(1|2)/L(2|1)$  is known as the *Bayes critical value*. One chooses Model 1 if the posterior odds ratio in favor of it exceeds the Bayes critical value. For reasons of economy an investigator may therefore report only the marginal likelihood, leaving it to her clients -- i.e, the users of the investigator's research -- to provide their own prior model probabilities and loss functions. The steps of simply reporting marginal likelihoods and Bayes factors are sometimes called hypothesis testing as well. The Bayes factor itself can be seen to serve as a test statistic, by rearranging (2.3.9) as

$$\frac{p(\mathbf{Y}_T|A_1)}{p(\mathbf{Y}_T|A_2)} > \frac{L(1|2)P(A_2)}{L(2|1)P(A_1)}.$$

That is, the Bayes action can be viewed as choosing model 1 if the sample evidence in its favor (as measured by the Bayes factor) is greater than the *prior* expected loss associated with its choice.

It is instructive to consider briefly the choice between two models given a sequence of prior distributions  $p(\theta_1|A_1^j)$  in Model 1 in which  $\lim_{j \rightarrow \infty} p(\theta_1|A_1^j) = 0 \forall \theta_1 \in \Theta_1$  but  $p(\mathbf{Y}_T|\theta_1, A_1^j)$  is the same for all  $j$ . It was seen in Section 2.2 that limiting posterior moments in Model 1 can be well-defined in this case, and may be found conveniently using a corresponding sequence of convergent prior density kernels. If the likelihood function satisfies a mild regularity condition, like  $\{\theta_1 : p(\mathbf{Y}_T|\theta_1, A_1^j) > c\}$  is a compact set of finite  $dv(\theta_1)$  measure for all  $c > 0$ , then  $\lim_{j \rightarrow \infty} p(\mathbf{Y}_T|A_1^j) = 0$ . This ensures  $\lim_{j \rightarrow \infty} p(A_1^j|\mathbf{Y}_T) = 0$ . 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). It can be observed explicitly in the linear model with  $h$  fixed, for which the marginal likelihood is (2.3.3) If  $\underline{\beta}$  is fixed but  $\underline{\mathbf{H}} \rightarrow \mathbf{0}$ , then  $p(\beta|A) \rightarrow 0 \forall \beta$  and  $\int_{\mathfrak{R}^k} p(\beta|A)p(\mathbf{y}|\mathbf{X}, \beta, h, A)d\beta \rightarrow 0$  as well.

## 2.4 Hierarchical priors and latent variables

A *hierarchical prior distribution* expresses the prior in two or more steps. The two-step case specifies a model

$$(2.4.1) \quad p^{(1)}(\mathbf{Y}_T|\theta, \lambda, A)$$

with a prior density for  $\theta \in \Theta$  conditional on a vector of *hyperparameters*  $\phi \in \Phi$ ,

$$(2.4.2) \quad p^{(2)}(\theta|\phi, A).$$

and a prior density for  $\phi$  and  $\lambda \in \Lambda$

$$(2.4.3) \quad p^{(3)}(\phi, \lambda|A),$$

it being understood in (2.4.1) that  $p^{(1)}(\mathbf{Y}_T|\theta, \lambda, A) = p(\mathbf{Y}_T|\theta, \lambda, \phi, A)$ .

The full prior density for all parameters and hyperparameters is

$$(2.4.4) \quad p(\theta, \phi, \lambda|A) = p^{(3)}(\phi, \lambda|A) p^{(2)}(\theta|\phi, A).$$

There is no fundamental difference between this prior density and the one described in Section 2.1, since

$$p(\theta, \lambda) = \int_{\Phi} p^{(2)}(\theta|\phi, A) p^{(3)}(\phi, \lambda|A) d\nu(\phi).$$

However the hierarchical formulation is often so convenient as to render fairly simple the analysis of posterior densities that would otherwise be quite difficult. Given a hierarchical prior, one may express the full posterior density

$$(2.4.5) \quad p(\theta, \lambda, \phi|\mathbf{Y}_T, A) \propto p^{(1)}(\mathbf{Y}_T|\theta, \lambda, A) p^{(2)}(\theta|\phi, A) p^{(3)}(\phi, \lambda|A).$$

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

$$(2.4.6) \quad p^{(1)}(\mathbf{Y}_T|\mathbf{Z}_T^*, \lambda, A)$$

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

$$(2.4.7) \quad p^{(2)}(\mathbf{Z}_T^*|\phi, A)$$

and the prior density for  $\phi \in \Phi$  and  $\lambda$  is

$$(2.4.8) \quad p^{(3)}(\phi, \lambda|A).$$

The full prior density for all parameters and unobservable variables is

$$(2.4.9) \quad p(\mathbf{Z}_T^*, \lambda, \phi|A) = p^{(3)}(\phi, \lambda|A) p^{(2)}(\mathbf{Z}_T^*|\phi, A)$$

and the full posterior density is

$$(2.4.10) \quad p(\mathbf{Z}_T^*, \lambda, \phi|\mathbf{Y}_T, A) \propto p^{(1)}(\mathbf{Y}_T|\mathbf{Z}_T^*, \lambda, A) p^{(2)}(\mathbf{Z}_T^*|\phi, A) p^{(3)}(\phi, \lambda|A).$$

Comparing (2.4.1)-(2.4.5) with (2.4.6)-(2.4.10), it is apparent that the latent variable model is formally identical to a model with a two-stage hierarchical prior, the latent variables corresponding to the intermediate level of the hierarchy. With appropriate marginalization of (2.4.10) one may obtain  $p(\mathbf{Z}_T^*|\mathbf{Y}_T, A)$ , which fully reflects uncertainty about the parameters. If one is interested only in  $\lambda$  and  $\phi$ , these distributions may also be obtained by marginalization of (2.4.10). Marginalization requires integration over  $\mathbf{Z}_T^*$ , which is possible analytically only in special cases. If the problem is approached using the simulation methods described beginning in the next section, then this integration simply amounts to discarding simulated values of  $\mathbf{Z}_T^*$ .

A simple example of a latent variable model is provided by the textbook probit model,

$$(2.4.11) \quad \mathbf{y}^* = \mathbf{X}\beta + \varepsilon, \quad \varepsilon|\mathbf{X} \sim \mathbf{N}(\mathbf{0}, \mathbf{I}_T), \quad \text{rank}(\mathbf{X}) = k, \quad d_t = \chi_{[0, \infty)}(y_t^*),$$

in which the  $T \times k$  matrix of covariates  $\mathbf{X} = [\mathbf{x}'_1, \dots, \mathbf{x}'_T]'$  and decision vector  $\mathbf{d}' = (d_1, \dots, d_T)$  are observed, but  $\mathbf{y}^{*'} = (y_1^*, \dots, y_T^*)$  is latent. To complete the model take

$$(2.4.12) \quad \beta \sim \mathbf{N}(\underline{\beta}, \mathbf{H}^{-1}).$$

In the equivalent formulation of this model using a hierarchical prior the parameter vector is  $(\mathbf{y}^*, \beta)$ . The first level of the hierarchical prior is  $\beta \sim \mathbf{N}(\underline{\beta}, \mathbf{H}^{-1})$ , corresponding to  $p^{(3)}$  with  $\phi = \beta$ . The second level is  $\mathbf{y}^* | (\beta, \mathbf{X}) \sim \mathbf{N}(\mathbf{X}\beta, \mathbf{I})$ , with  $\theta = \mathbf{y}^*$  in the hierarchical prior interpretation and  $\mathbf{Z}_t^* = \mathbf{y}^*$  in the latent variable interpretation. (There is no analog of  $\lambda$  in this example.) The data distribution is

$$p(\mathbf{d} | \mathbf{y}^*) = \prod_{t=1}^T [\chi_{[0, \infty)}(y_t^*) d_t + \chi_{(-\infty, 0)}(y_t^*) (1 - d_t)].$$

Either formulation leads to the same joint distribution for  $\beta, \mathbf{y}^*$  and  $\mathbf{d}$ ,

$$(2.4.13) \quad p(\beta, \mathbf{y}^*, \mathbf{d} | \mathbf{X}) = (2\pi)^{-(T+k)/2} |\mathbf{H}|^{1/2} \exp \left[ -0.5 (\beta - \underline{\beta})' \mathbf{H} (\beta - \underline{\beta}) \right] \\ \cdot \prod_{t=1}^T \exp \left[ -0.5 (y_t^* - \beta' \mathbf{x}_t)^2 \right] [\chi_{[0, \infty)}(y_t^*) d_t + \chi_{(-\infty, 0)}(y_t^*) (1 - d_t)].$$

The main conceptual point is that since Bayesian inference conditions on the observables  $(\mathbf{d}, \mathbf{X})$ , parameters and latent variables have the same standing as unknown entities whose joint distribution with the observables is given by the model. As we shall see in Section 3.3, this formulation provides a basis for computations as well.

### 3. Posterior Simulation Methods

The objective of inference in a single model,

$$E[h(\omega)|\mathbf{Y}_T, A] = \int_{\Theta} \int_{\Omega} h(\omega) p(\omega|\theta, \mathbf{Y}_T, A) p(\theta|\mathbf{Y}_T, A) d\nu(\omega) d\nu(\theta),$$

can be evaluated analytically only in a few specific simple cases. This section describes simulation methods for obtaining a sequence of strongly consistent approximations to  $E[h(\omega)|\mathbf{Y}_T, A]$ , and the following section will take up the process of model averaging. In most applications, it is generally straightforward to find a function  $g(\mathbf{Y}_T, \theta)$ , possibly random, with the property

$$(3.0.1) \quad E[g(\mathbf{Y}_T, \theta)|\mathbf{Y}_T, \theta, A] = E[h(\omega)|\mathbf{Y}_T, \theta, A] = \int_{\Omega} h(\omega) p(\omega|\mathbf{Y}_T, \theta, A) d\omega = \bar{g}.$$

Finding this function is trivial if  $h(\omega)|(\mathbf{Y}_T, \theta, A)$  is deterministic.<sup>1</sup> This was the case in the production function examples discussed in Section 2.1. If  $h(\omega)$  is random then it is often straightforward to take  $\omega \sim p(\omega|\mathbf{Y}_T, \theta, A)$  and then  $g(\mathbf{Y}_T, \theta) = h(\omega)$ . This was the case in the tax revenue forecasting example in Section 2.1.

More generally one may be able to find a function satisfying (3.0.1), but for which

$$(3.0.2) \quad \text{var}[g(\mathbf{Y}_T, \theta)|\mathbf{Y}_T, \theta, A] < \text{var}[h(\omega)|\mathbf{Y}_T, \theta, A].$$

The turning point example of Section 2.1 provides an illustration. Recall that in this example the objective was to evaluate  $P(y_{T+2} < y_{T+1} < y_T | \mathbf{Y}_T)$  and to this end we took  $\omega' = (y_{T+1}, y_{T+2})$ . One could draw  $\omega|(\mathbf{Y}_T, \theta)$  and use the random function  $g(\mathbf{Y}_T, \theta) = h(\omega) = \chi_{\Omega'}(\omega)$ . Alternatively, one could draw only  $\omega_1|(\mathbf{Y}_T, \theta)$  and use the random function  $g(\mathbf{Y}_T, \theta) = P(\omega_2 < \omega_1 | \mathbf{Y}_T, \theta)$ , which requires only the ability to evaluate the univariate standard normal c.d.f. Yet a third alternative is to employ the deterministic function  $g(\mathbf{Y}_T, \theta) = P(\omega_2 < \omega_1 < y_T | \mathbf{Y}_T, \theta)$  using bivariate quadrature. In each case  $E[g(\mathbf{Y}_T, \theta)|\mathbf{Y}_T, \theta] = P(y_{T+2} < y_{T+1} < y_T | \mathbf{Y}_T, \theta)$  but  $\text{var}[g(\mathbf{Y}_T, \theta)|\mathbf{Y}_T, \theta]$  is greatest in the first alternative, less in the second, and zero in the third.<sup>2</sup>

Throughout we shall make use of the notation  $g(\mathbf{Y}_T, \theta)$ , it always being implicit that (3.0.1) is satisfied.

If one could also make a sequence of independent draws  $\{\theta^{(m)}\}$  from the posterior distribution, then by choosing  $\omega^{(m)} \sim p(\omega|\mathbf{Y}_T, \theta^{(m)}, A)$  one could guarantee  $M^{-1} \sum_{m=1}^M h(\omega^{(m)}) \xrightarrow{a.s.} E[h(\omega)|\mathbf{Y}_T, A]$ . But direct simulation from the posterior distribution is rarely possible. This section describes methods for obtaining a sequence

<sup>1</sup>The *evaluation* of  $g(\mathbf{Y}_T, \theta)$  may not be trivial at all. For example, Bajari (1997) has functions of interest whose evaluation requires the solution of a system of nonlinear differential equations.

<sup>2</sup>In some cases the left side of (3.0.2) can be made quite small indeed, and asymptotically it may be made to approach zero (Geweke, 1988).

$\{\theta^{(m)}\}_{m=1}^{\infty}$ , and an associated weighting function  $w(\theta)$ , with the property that if  $E[h(\omega^{(m)})|\theta^{(m)}] = E[h(\omega)|\mathbf{Y}_T, \theta^{(m)}, A]$  for a corresponding sequence  $\{\omega^{(m)}\}_{m=1}^{\infty}$ , then

$$\frac{\sum_{m=1}^M w(\theta^{(m)})h(\omega^{(m)})}{\sum_{m=1}^M w(\theta^{(m)})} \xrightarrow{a.s.} E[h(\omega)|\mathbf{Y}_T, A].$$

The ability to generate such sequences has improved greatly in the past ten years, due in large part to the development of Markov chain Monte Carlo (MCMC) methods and the dramatic decrease in the cost of computing. We begin by reviewing two more established methods, acceptance and importance sampling, and then move on to the Gibbs sampler and the Hastings-Metropolis algorithm as examples of MCMC. This is followed by a more abstract development of MCMC theory, a description of some of the hybrid procedures that make MCMC a powerful tool for posterior simulation, and a discussion of the evaluation of approximation error. The section concludes with a description of some public domain software for posterior simulation and two simple examples. The emphasis here is on concepts and practicality. With one exception we provide only references to proofs of theorems. A more general and extensive introduction is provided by Gelman et al. (1995). A concise presentation of the relevant continuous state space Markov chain theory that underlies MCMC procedures is Tierney (1994).

*A word of caution.* Sections 2.1 and 2.2 emphasized the importance of verifying the absolute convergence of integrals in the denominator and numerator in the generic expression (2.1.6) for the posterior expectation of a function of interest. If either condition is violated, then the simulation methods discussed below in this section have absolutely no justification, because the posterior expectation allegedly being approximated does not exist. In this circumstance there is often no indication of difficulty in the output of the posterior simulator, which may appear reasonable. *Absolute convergence of integrals must be verified analytically before using a posterior simulator.* This is often quite simple: e.g., if the likelihood function  $p(\mathbf{Y}_T|\theta, A)$  is bounded and the prior distribution is proper, then the denominator of (2.1.6) is absolutely convergent; and if in addition the prior expectation  $E[h(\omega)|A]$  exists then the numerator of (2.1.6) is absolutely convergent. If the prior is improper, or the likelihood function is unbounded, or the prior expectation does not exist, then the extra effort to verify existence of the posterior expectation at hand must be expended before proceeding with posterior simulations.

### 3.1 Acceptance sampling

Acceptance sampling is the algorithm that underlies the generation of random variables from most familiar univariate distributions like the normal and the gamma (Press *et al.*,

1992). The idea behind acceptance sampling is to generate a random vector<sup>3</sup> from a distribution that is similar, in an appropriate sense, to the posterior distribution, and then to accept that drawing with a probability that depends on the drawn value of the vector. If this acceptance probability function is chosen correctly then the accepted values will have the desired distribution.

*Theorem 3.1.1.* Suppose that  $p^*(\theta|\mathbf{Y}_T, A)$  is any kernel of the posterior density  $p(\theta|\mathbf{Y}_T, A)$ . Let  $s^*(\theta)$  be a source density kernel with respect to the same measure  $d\nu(\theta)$  as  $p(\theta|A)$ , with support  $S$  and the property

$$(3.1.1) \quad 0 \leq p^*(\theta|\mathbf{Y}_T, A)/s^*(\theta) \leq a < \infty \forall \theta \in \Theta .$$

Suppose that the sequence  $\{\theta^{(m)}\}$  is generated as follows:

- (a) Set  $m = 1$ ;
- (b) Generate  $u \sim U(0, 1)$ ;
- (c) Generate  $\tilde{\theta}$  from the source density;
- (d) If  $u > p^*(\tilde{\theta}|\mathbf{Y}_T, A)/[as^*(\tilde{\theta})]$ , go to (b); otherwise,
- (e)  $\theta^{(m)} = \tilde{\theta}$ ;
- (f) Increment  $m$  and go to (b).

Then  $\theta^{(m)} \stackrel{iid}{\sim} p(\theta|\mathbf{Y}_T, A)$ .

*Proof.* Given  $\tilde{\theta}$  from (c), the probability of proceeding directly from step (d) to step (e) is  $p^*(\tilde{\theta}|\mathbf{Y}_T, A)/as^*(\tilde{\theta})$ . To obtain the unconditional probability of proceeding directly from step (d) to step (e), integrate the product of this expression and the source density of  $\tilde{\theta}$ ,

$$(3.1.2) \quad \int_{\Theta} [p^*(\theta|\mathbf{Y}_T, A)/as^*(\theta)] s^*(\theta) d\nu(\theta) / \int_S s^*(\theta) d\nu(\theta) \\ = \int_{\Theta} p^*(\theta|\mathbf{Y}_T, A) d\nu(\theta) / a \int_S s^*(\theta) d\nu(\theta).$$

The unconditional probability of proceeding from step (d) to step (e) with  $\theta \in \Theta_1 \subseteq \Theta$  is

$$(3.1.3) \quad \int_{\Theta_1} [p^*(\theta|\mathbf{Y}_T, A)/as^*(\theta)] s^*(\theta) d\nu(\theta) / \int_S s^*(\theta) d\nu(\theta) \\ = \int_{\Theta_1} p^*(\theta|\mathbf{Y}_T, A) d\nu(\theta) / a \int_S s^*(\theta) d\nu(\theta).$$

The probability that  $\theta \in \Theta_1 \subseteq \Theta$ , conditional on arriving at step (e), is the ratio of (3.1.3) to (3.1.2), which is

$$\int_{\Theta_1} p^*(\theta|\mathbf{Y}_T) d\nu(\theta) / \int_{\Theta} p^*(\theta|\mathbf{Y}_T) d\nu(\theta) = P(\theta \in \Theta_1 | \mathbf{Y}_T). \quad \#\#$$

---

<sup>3</sup>We ignore the distinction between the mathematical properties of a sequence of random variables and the properties of (what is properly called) a pseudo-random variable sequence created using a computer. For a discussion of these issues see Geweke (1996) and references therein.

A successful application of acceptance sampling has three requirements. First, there must be a source density corresponding to a distribution from which it is efficient and convenient to make i.i.d. draws. Second, there must be a known upper bound on the ratio of the posterior density to the source density. Finally, the frequency of rejection (moving to step (b) from step (d)) must not be so great that the whole algorithm is impractical. The upper bound must be established analytically, whereas efficiency can be evaluated through experimentation. Notice that draws from the source density may (and usually do) involve acceptance sampling: for example, if the source density is a normal or gamma density, the software used to draw from this density very likely employs acceptance sampling, a fact typically transparent to the software user.

Acceptance sampling produces an i.i.d. sequence  $\{\theta^{(m)}\}$ . Given (3.0.1) it follows from the strong law of large numbers that  $\bar{g}_M = M^{-1} \sum_{m=1}^M g(\mathbf{Y}_T, \theta^{(m)}) \xrightarrow{a.s.} \bar{g}$ . If in addition  $\sigma^2 = \text{var}[g(\mathbf{Y}_T, \theta) | \mathbf{Y}_T, A]$  exists then from the Lindberg-Levy central limit theorem  $M^{1/2} [g(\mathbf{Y}_T, \theta^{(m)}) - \bar{g}] \xrightarrow{d} N(0, \sigma^2)$ , and a second application of the strong law of large numbers yields  $\hat{\sigma}^2 = M^{-1} \sum_{m=1}^M [g(\mathbf{Y}_T, \theta^{(m)}) - \bar{g}_M]^2 \xrightarrow{a.s.} \sigma^2$ . Thus, if the posterior variance of the function of interest exists, a central limit theorem may be used in the usual way to assess the numerical accuracy of the approximation of  $E[h(\omega) | \mathbf{Y}_T, A]$  by  $M^{-1} \sum_{m=1}^M g(\mathbf{Y}_T, \theta^{(m)})$ .

### 3.2 Importance sampling

Rather than accept only a fraction of the draws from the source density, it is possible to retain all of them, and consistently approximate the posterior moment by appropriately weighting the draws. The probability density function of the source distribution is then called the *importance sampling density*, a term due to Hammersly and Handscomb (1964), who were among the first to propose the method. It appears to have been introduced to the econometrics literature by Kloek and van Dijk (1978). To help distinguish between acceptance and importance sampling we shall indicate the importance sampling distribution by its density  $j(\theta)$  with respect to the same measure  $d\nu(\theta)$  as the prior density  $p(\theta|A)$ . Let  $j^*(\theta)$  be any kernel of  $j(\theta)$ , and let  $p^*(\theta | \mathbf{Y}_T, A)$  be any kernel of  $p(\theta | \mathbf{Y}_T, A)$ .

*Theorem 3.2.1.* Suppose  $E[g(\mathbf{Y}_T, \theta) | \mathbf{Y}_T, A]$  exists, and the support of  $j(\theta)$  includes  $\Theta$ . Then

$$\bar{g}_M = \sum_{m=1}^M g(\mathbf{Y}_T, \theta^{(m)}) w(\theta^{(m)}) / \sum_{m=1}^M w(\theta^{(m)}) \xrightarrow{a.s.} E[g(\mathbf{Y}_T, \theta) | \mathbf{Y}_T, A] = \bar{g},$$

where  $w(\theta) = p^*(\theta|\mathbf{Y}_T, A)/j^*(\theta)$  is the corresponding *weighting function*. If in addition both  $E[w(\theta)|\mathbf{Y}_T, A] = \int_{\Theta} w(\theta)p(\theta|\mathbf{Y}_T, A)d\nu(\theta)$  and  $\text{var}[g(\mathbf{Y}_T, \theta)|\mathbf{Y}_T, A]$  exist, then

$$(3.2.1) \quad M^{1/2}(\bar{g}_M - \bar{g}) \xrightarrow{d} N(0, \sigma^2),$$

and

$$(3.2.2) \quad \hat{\sigma}_M^2 = M \sum_{m=1}^M [g(\mathbf{Y}_T, \theta^{(m)}) - \bar{g}_M]^2 w(\theta^{(m)})^2 / \left[ \sum_{m=1}^M w(\theta^{(m)}) \right]^2 \xrightarrow{a.s.} \sigma^2.$$

*Proof reference.* Geweke (1989b, Theorems 1 and 2).

##

This result provides a practical way to assess approximation error and also indicates conditions in which the method of importance sampling will work well. Small variance in  $w(\theta)$ , perhaps reflecting close upper and lower bounds on  $w(\theta)$ , will lead to small values of  $\sigma^2$  relative to  $\text{var}[g(\mathbf{Y}_T, \theta)|\mathbf{Y}_T, A]$ . Of course, the existence of  $E[w(\theta)|\mathbf{Y}_T, A]$  and  $\text{var}[g(\mathbf{Y}_T, \theta)|\mathbf{Y}_T, A]$  must be verified analytically. The following implication of Theorem 3.2.1 is often useful in the latter undertaking.

*Corollary 3.2.2.* If  $\text{var}[g(\mathbf{Y}_T, \theta)|\mathbf{Y}_T, A]$  exists and the weighting function  $w(\theta) = p^*(\theta|\mathbf{Y}_T, A)/j^*(\theta)$  is bounded, then (3.2.1) and (3.2.2) are true.

##

The hypothetical special case  $j(\theta) \propto p(\theta|\mathbf{Y}_T, A)$  corresponds to i.i.d. sampling from the posterior distribution, since the weighting function is then constant. In this case,  $\sigma^2 = \text{var}[g(\mathbf{Y}_T, \theta)|\mathbf{Y}_T, A]$ , which can serve as a benchmark in evaluating the adequacy of  $j(\theta)$  in all other cases. The ratio  $\text{var}[g(\theta, \mathbf{Y}_T)|\mathbf{Y}_T, A]/\sigma^2$  has been termed the *relative numerical efficiency* of the importance sampling approximation to  $E[g(\mathbf{Y}_T, \theta)|\mathbf{Y}_T, A]$  (Geweke, 1989b): it indicates the ratio of iterations using  $p(\theta|\mathbf{Y}_T, A)$  itself as the importance sampling density, to the number using  $j(\theta)$ , required to achieve the same accuracy of approximation of  $\bar{g}$ . Since both the numerator and denominator of the ratio  $\text{var}[g(\theta, \mathbf{Y}_T)|\mathbf{Y}_T, A]/\sigma^2$  can be approximated consistently as the number of draws  $M$  increases, this is a practical indication of the computational efficiency of importance sampling. Relative numerical efficiency much less than 1.0 (less than 0.1, certainly less than 0.01) indicates poor imitation of  $p(\theta|\mathbf{Y}_T, A)$  by  $j(\theta)$ , possibly the existence of a better importance sampling distribution or the failure of the underlying convergence conditions for (3.2.2).

Acceptance and importance sampling are closely related. If (3.1.1) is satisfied, then the source density used in acceptance sampling can be an importance sampling density in importance sampling and the weighting function  $w(\theta)$  will be bounded as assumed in

Corollary 3.2.2. Which procedure should be used depends on computation time and the acceptance probability in acceptance sampling. If drawing  $\theta^{(m)}$  and evaluating the relevant densities is expensive relative to evaluation of the functions  $g(\mathbf{Y}_T, \theta)$ , and if acceptance probability is low, then importance sampling is more attractive; and conversely.

Importance sampling is an important useful tool in modifying prior distributions. Suppose that models  $A_1$  and  $A_2$  are distinguished only by their prior densities  $p(\theta|A_j)$ ,  $j = 1, 2$ . Suppose that one has available an i.i.d. sample from the posterior density  $p(\theta|\mathbf{Y}_T, A_1) \propto p(\theta|A_1)p(\mathbf{Y}_T|\theta, A_1)$ . If  $p(\theta|A_2)/p(\theta|A_1)$  is bounded above then  $p(\theta|\mathbf{Y}_T, A_1)$  is an importance sampling density for  $p(\theta|\mathbf{Y}_T, A_2)$  that satisfies the conditions of Corollary 3.2.2. The weighting function is  $w(\theta) = p(\theta|A_2)/p(\theta|A_1)$ . Thus, one may change the prior distribution without reworking the entire problem. The ability to do so makes conditionally conjugate prior distributions, of the kind discussed in Section 2 in conjunction with the standard linear model, attractive as reporting devices because an investigator's results, produced with such priors, may be modified by a client with different priors. This idea will be developed more fully in Section 6.

### 3.3 The Gibbs sampler

The Gibbs sampler is an algorithm that has been used with noted success in many econometric models. It is one example of a wider class of procedures known as *Markov chain Monte Carlo* (MCMC). In these procedures the idea is to construct a Markov chain with state space  $\Theta$ , and unique invariant distribution  $p(\theta|\mathbf{Y}_T, A)$ . Following an initial transient or *burn-in* phase, simulated values from the chain are used to approximate  $E[g(\mathbf{Y}_T, \theta)|\mathbf{Y}_T, A]$ .

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 in econometrics. Since 1990 application of MCMC methods has grown rapidly (Chib and Greenberg, 1996).

This section and the next concentrate on a heuristic development of two widely used variants of these methods, the Gibbs sampler and the Hastings-Metropolis algorithm. The

general theory of convergence is taken up in Section 3.5. Section 3.6 details some useful specific variants and combinations of these methods. Section 3.7 turns to the assessment of numerical accuracy.

The Gibbs sampler begins with a partition, or *blocking*, of  $\theta, \theta' = (\theta'_{(1),K}, \theta'_{(B)})$ . In applications, the blocking is chosen so that it is possible to draw from each of the conditional p.d.f.'s,  $p(\theta_{(b)} | \mathbf{Y}_T, \theta_{(a)} (a < b), \theta_{(a)} (a > b), A)$ . This blocking can arise naturally, if the prior distributions for the  $\theta_{(b)}$  are independent and each is conditionally conjugate. To motivate the key idea underlying the Gibbs sampler suppose — contrary to fact — that there existed a single drawing  $\theta^{(0)}, \theta'^{(0)} = (\theta'^{(0)},K, \theta'^{(0)})$ , from  $p(\theta | \mathbf{Y}_T, A)$ .

Successively make drawings from the conditional distributions as follows:

$$\begin{aligned}
 \theta_{(1)}^{(1)} &\sim p(\cdot | \mathbf{Y}_T, \theta_{(2)}^{(0)},K, \theta_{(B)}^{(0)}, A) \\
 \theta_{(2)}^{(1)} &\sim p(\cdot | \mathbf{Y}_T, \theta_{(1)}^{(1)}, \theta_{(3)}^{(0)},K, \theta_{(B)}^{(0)}, A) \\
 &\vdots \\
 \theta_{(b)}^{(1)} &\sim p(\cdot | \mathbf{Y}_T, \theta_{(1)}^{(1)},K, \theta_{(b-1)}^{(1)}, \theta_{(b+1)}^{(0)},K, \theta_{(B)}^{(0)}, A) \\
 &\vdots \\
 \theta_{(B)}^{(1)} &\sim p(\cdot | \mathbf{Y}_T, \theta_{(1)}^{(1)},K, \theta_{(B-1)}^{(1)}, A).
 \end{aligned}
 \tag{3.3.1}$$

This defines a transition process from  $\theta'^{(0)}$  to  $\theta'^{(1)} = (\theta'^{(1)},K, \theta'^{(1)})$ . Since  $\theta^{(0)} \sim p(\theta | \mathbf{Y}_T, A)$ ,

$$(\theta_{(1)}^{(1)},K, \theta_{(b-1)}^{(1)}, \theta_{(b)}^{(1)}, \theta_{(b+1)}^{(0)},K, \theta_{(B)}^{(0)}) \sim p(\theta | \mathbf{Y}_T, A)$$

at each step in (3.3.1) by definition of the conditional density. In particular,  $\theta^{(1)} \sim p(\theta | \mathbf{Y}_T, A)$ .

Iteration of this algorithm produces a sequence  $\theta^{(0)}, \theta^{(1)},K, \theta^{(m)},K$  which is a realization of a Markov chain with probability density function kernel for the transition from point  $\theta^{(m)}$  to point  $\theta^{(m+1)}$  given by

$$(3.3.2) \quad K_G(\theta^{(m)}, \theta^{(m+1)}) = \prod_{b=1}^B p[\theta_{(b)}^{(m+1)} | \mathbf{Y}_T, \theta_{(a)}^{(m)} (a > b), \theta_{(a)}^{(m+1)} (a < b), A].$$

Any single iterate  $\theta^{(m)}$  retains the property that it is drawn from the posterior distribution. 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 in an efficient manner. In econometrics the blocking is often natural and the conditional distributions familiar. In making the drawings (3.3.1) acceptance sampling is often useful.

The appeal of the Gibbs sampler is easy to illustrate with the standard linear model (2.1.7)-(2.1.9): The results (2.1.17) and (2.1.18) indicate that the blocking

$\theta_{(1)} = \beta$ ,  $\theta_{(2)} = h$  meets the criterion that drawings can be made in an efficient manner. The probit model introduced in Section 2.4 is a further example, as noted in Albert and Chib (1993). From (2.4.11)-(2.4.12) it is evident that conditional on the vector of latent variables  $\mathbf{y}^*$  the distribution of  $\beta$  is given by (2.1.17) if we use  $\mathbf{y}^*$  in place of  $\mathbf{y}$  and set  $h = 1$ . Examination of the kernel of (2.4.13) in  $\mathbf{y}^*$  shows that given  $\beta$  and  $\mathbf{X}$  the  $y_i^*$  are conditionally independent, with  $y_i^* \sim N(\beta' \mathbf{x}_i, 1)$  truncated to  $[0, \infty)$  if  $d_i = 1$  and truncated to  $(-\infty, 0)$  if  $d_i = 0$ . An efficient algorithm for drawing from truncated normal distributions is given in Geweke (1991). In both cases, given drawings for the parameters it is straightforward to produce numerical approximations to  $E[h(\omega)|\mathbf{Y}_T, A]$ , as indicated at the start of this section. And as discussed in Section 2.1 the evaluation of  $E[h(\omega)|\mathbf{Y}_T, A]$  subsumes most of the uses to which these models are put.

Of course, if it really were possible to make an initial draw from the posterior distribution, then independence Monte Carlo would also be possible. An important remaining task is to elucidate conditions for the distribution of  $\theta^{(m)}$  to converge to the posterior for any  $\theta^{(0)} \in \Theta$ . This is not trivial, because even if  $\theta^{(0)}$  were drawn from  $p(\theta|\mathbf{Y}_T, A)$ , the argument just given establishes only that any single  $\theta^{(m)}$  is also drawn from the posterior distribution. It does not establish that a single sequence  $\{\theta^{(m)}\}_{m=1}^{\infty}$  is representative of the posterior distribution. For example, if  $\Theta$  consists of two disjoint subsets  $\Theta_1$  and  $\Theta_2$  with  $\theta_1 > \theta_2 \forall \theta_j \in \Theta_j$ , then a Gibbs sampler that begins in  $\Theta_1$  will never visit  $\Theta_2$  and vice versa (see Figure 3.3.1.) This situation clearly does not arise in the Gibbs samplers for the standard linear and probit models just described, but evidently a careful development of conditions under which  $\{\theta^{(m)}\}$  converges in distribution to the posterior distribution is needed. We outline these developments in Section 3.5.

### 3.4 The Hastings-Metropolis algorithm

The Hastings-Metropolis algorithm begins with an arbitrary transition probability density function  $q(\mathbf{x}, \mathbf{y})$  indexed by  $\mathbf{x} \in \Theta$  and with density argument  $\mathbf{y} \in \Theta$ , and with an arbitrary starting value  $\theta^{(0)} \in \Theta$ . The random vector  $\theta^*$  generated from  $q(\theta^{(m)}, \theta^*)$  is a candidate value for  $\theta^{(m+1)}$ . The algorithm actually sets  $\theta^{(m+1)} = \theta^*$  with probability

$$(3.4.1) \quad \alpha(\theta^{(m)}, \theta^*) = \min \left\{ \frac{p(\theta^*|\mathbf{Y}_T, A)q(\theta^*, \theta^{(m)})}{p(\theta^{(m)}|\mathbf{Y}_T, A)q(\theta^{(m)}, \theta^*)}, 1 \right\} = \min \left\{ \frac{p(\theta^*|\mathbf{Y}_T, A)/q(\theta^{(m)}, \theta^*)}{p(\theta^{(m)}|\mathbf{Y}_T, A)/q(\theta^*, \theta^{(m)})}, 1 \right\};$$

otherwise, the algorithm sets  $\theta^{(m+1)} = \theta^{(m)}$ . This defines a Markov chain with a generally mixed continuous-discrete transition probability from  $\theta^{(m)}$  to  $\theta^{(m+1)}$  given by

$$K_H(\theta^{(m)}, \theta^{(m+1)}) = \begin{cases} q(\theta^{(m)}, \theta^{(m+1)}) \alpha(\theta^{(m)}, \theta^{(m+1)}) & \text{if } \theta^{(m+1)} \neq \theta^{(m)} \\ 1 - \int_{\Theta} q(\theta^{(m)}, \theta) \alpha(\theta^{(m)}, \theta) d\nu(\theta) & \text{if } \theta^{(m+1)} = \theta^{(m)}. \end{cases}$$

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

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

where  $w(\theta) = p(\theta | \mathbf{Y}_T, A) / k(\theta)$ . The independence chain is closely related to acceptance sampling and importance sampling. In acceptance sampling, if the posterior density is low (high) relative to the source density the probability of acceptance is low (high). In importance sampling, if the posterior density is low (high) relative to the importance sampling the weight assigned to the draw is low (high). In the independence chain, to the extent the posterior density is lower (higher) relative to the proposal than was the case in the previously accepted draw, the probability of accepting the proposed vector is lower (one).

There is a simple two-step argument that motivates the convergence of the sequence  $\{\theta^{(m)}\}$  generated by the Hastings-Metropolis algorithm to the posterior. (This approach is due to Chib and Greenberg, 1995.) First, observe that if the transition probability function  $p(\theta^{(m)}, \theta^{(m+1)})$  satisfies the *reversibility condition*

$$(3.4.2) \quad p(\theta^{(m)}) p(\theta^{(m)}, \theta^{(m+1)}) = p(\theta^{(m+1)}) p(\theta^{(m+1)}, \theta^{(m)}),$$

for stated  $p(\cdot)$ , then it has  $p(\cdot)$  as an invariant distribution. To see this, note that if (3.4.1) holds then

$$\begin{aligned} \int_{\Theta} p(\theta^{(m)}) p(\theta^{(m)}, \theta^{(m+1)}) d\nu(\theta^{(m)}) &= \int_{\Theta} p(\theta^{(m+1)}) p(\theta^{(m+1)}, \theta^{(m)}) d\nu(\theta^{(m)}) \\ &= p(\theta^{(m+1)}) \int_{\Theta} p(\theta^{(m+1)}, \theta^{(m)}) d\nu(\theta^{(m)}) = p(\theta^{(m+1)}). \end{aligned}$$

For  $\theta^{(m+1)} = \theta^{(m)}$ , (3.4.2) is satisfied trivially. For  $\theta^{(m+1)} \neq \theta^{(m)}$ , suppose without loss of generality that  $p(\theta^{(m+1)}) / q(\theta^{(m)}, \theta^{(m+1)}) > p(\theta^{(m)}) / q(\theta^{(m+1)}, \theta^{(m)})$ . Then

$$p(\theta^{(m)}, \theta^{(m+1)}) = q(\theta^{(m)}, \theta^{(m+1)})$$

and

$$p(\theta^{(m+1)}, \theta^{(m)}) = q(\theta^{(m+1)}, \theta^{(m)}) \cdot \frac{p(\theta^{(m)}) / q(\theta^{(m+1)}, \theta^{(m)})}{p(\theta^{(m+1)}) / q(\theta^{(m)}, \theta^{(m+1)})} = p(\theta^{(m)}) q(\theta^{(m)}, \theta^{(m+1)}) / p(\theta^{(m+1)})$$

whence (3.4.2) is satisfied.

In implementing the Hastings-Metropolis algorithm the transition probability density function must share two important properties. First, it must be possible to generate  $\theta^*$

efficiently from  $q(\theta^{(m)}, \theta^*)$ . 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  $\theta^{(m)}$  is too great.

In the case of the independence chain the Hastings-Metropolis algorithm will be efficient under essentially the same conditions that the corresponding importance sampling algorithm with the same  $j(\theta)$  will be efficient. If there are values of  $\theta$  for which  $p^*(\theta|\mathbf{Y}_T, A)/j(\theta)$  is very much greater than at other values, then the importance sampling algorithm will place very high weights on these values, which are drawn infrequently relative to  $p^*(\theta|\mathbf{Y}_T, A)$ . The Hastings-Metropolis independence chain will tend to remain at such values for many successive iterations. In either case relative numerical efficiency will, as a consequence, be low.

Another variant of the Hastings-Metropolis algorithm is the random walk chain, in which  $q(\theta^{(m)}, \theta^*) = f(\theta^{(m)} - \theta^*) = f(\theta^* - \theta^{(m)})$ . For example  $f$  could be multivariate normal, with mean  $\mathbf{0}$  and a constant variance matrix. If the variance matrix is chosen to reflect the shape of  $p^*(\theta|\mathbf{Y}_T, A)$  at least roughly, then this algorithm can be quite efficient.

### 3.5 Some MCMC theory

Much of the treatment here draws heavily on the work of Tierney (1994), who first used the theory of general state space Markov chains to demonstrate convergence, and Roberts and Smith (1994), who elucidated sufficient conditions for convergence that turn out to be applicable in a wide variety of problems in econometrics.

Let  $\{\theta^{(m)}\}_{m=0}^{\infty}$  be a Markov chain defined on  $\Theta \subseteq \mathfrak{R}^k$  with transition density  $K: \Theta \times \Theta \rightarrow \mathfrak{R}^+$  such that, for all  $\nu$ -measurable  $\Theta_0 \subseteq \Theta$ ,

$$P(\theta^{(m)} \in \Theta_0 | \theta^{(m-1)}) = \int_{\Theta_0} K(\theta^{(m-1)}, \theta) d\nu(\theta) + r(\theta^{(m-1)}) \chi_{\Theta_0}(\theta^{(m-1)}),$$

$$\text{where } r(\theta^{(m-1)}) = 1 - \int_{\Theta} K(\theta^{(m-1)}, \theta) d\nu(\theta).$$

The transition density  $K$  is substochastic: it defines only the distribution of accepted candidates. Assume that  $K$  has no absorbing states, so that  $r(\theta) < 1 \forall \theta \in \Theta$ . The corresponding substochastic kernel over  $m$  steps is then defined iteratively,

$$\begin{aligned} \mathbf{K}^{(m)}(\theta^{(0)}, \theta^{(m)}) &= \int_{\Theta} \mathbf{K}^{(m-1)}(\theta^{(0)}, \theta) \mathbf{K}(\theta, \theta^{(m)}) d\nu(\theta) \\ &+ \mathbf{K}^{(m-1)}(\theta^{(0)}, \theta^{(m)}) r(\theta^{(m)}) + [r(\theta^{(0)})]^{m-1} \mathbf{K}(\theta^{(0)}, \theta^{(m)}). \end{aligned}$$

This describes all  $m$ -step transitions that involve at least one accepted move. As a function of  $\theta^{(m)}$  it is the p.d.f. with respect to  $\nu$  of  $\theta^{(m)}$ , excluding realizations with  $\theta^{(n)} = \theta^{(0)} \forall n = 1, \dots, m$ . For any  $\nu$ -measurable  $\Theta_0$  let  $\mathbf{P}^{(m)}(\theta^{(0)}, \Theta_0)$  denote the  $m$ 'th iterate of  $\mathbf{P}$ ,

$$\mathbf{P}^{(m)}(\theta^{(0)}, \Theta_0) = \int_{\Theta_0} \mathbf{K}^{(m)}(\theta^{(0)}, \theta) d\nu(\theta) + [r(\theta^{(0)})]^m \chi_{\Theta_0}(\theta^{(0)}).$$

An invariant distribution of the transition density  $\mathbf{K}$  is a function  $p(\theta)$  that satisfies

$$\begin{aligned} \mathbf{P}(\Theta_0) &= \int_{\Theta_0} p(\theta) d\nu(\theta) = \int_{\Theta} \mathbf{P}(\theta^{(m)} \in \Theta_0 | \theta^{(m-1)} = \theta) p(\theta) d\nu(\theta) \\ &= \int_{\Theta} \left\{ \int_{\Theta_0} \mathbf{K}(\theta, \theta^*) d\nu(\theta^*) + r(\theta) \chi_{\Theta_0}(\theta) \right\} p(\theta) d\nu(\theta) \end{aligned}$$

for all  $\nu$ -measurable  $\Theta_0$ . Let  $\Theta^* = \{\theta \in \Theta : p(\theta) > 0\}$ . The density  $\mathbf{K}$  is *p-irreducible* if for all  $\theta^{(0)} \in \Theta^*$ ,  $\mathbf{P}(\Theta_0) > 0$  implies that  $\mathbf{P}^{(m)}(\theta^{(0)}, \Theta_0) > 0$  for some  $m \geq 1$ . Return to Figure 3.3.1, where the support is disconnected and the Markov chain is the Gibbs sampler. Note that if  $\theta^{(0)} \in \tilde{\Theta}_i$ , it is impossible that  $\theta^{(m)} \in \tilde{\Theta}_j$  ( $j \neq i$ , any  $m > 0$ ). Thus the transition density is not irreducible in this case. There are two invariant distributions, one for  $\tilde{\Theta}_1$  (reached if  $\theta^{(0)} \in \tilde{\Theta}_1$ ) and one for  $\tilde{\Theta}_2$  (reached if  $\theta^{(0)} \in \tilde{\Theta}_2$ ).

The transition density  $\mathbf{K}$  is *aperiodic* if there exists no  $\nu$ -measurable partition  $\Theta = \bigcup_{s=0}^{r-1} \tilde{\Theta}_s$  ( $r \geq 2$ ) such that

$$\mathbf{P}(\theta^{(m)} \in \tilde{\Theta}_{m \bmod(r)} | \theta^{(0)} \in \tilde{\Theta}_0) = 1 \quad \forall m.$$

It is *Harris recurrent* if  $P[\theta^{(m)} \in \Theta_0 \text{ i.o.} | \theta^{(0)}] = 1$  for all  $\nu$ -measurable  $\Theta_0$  with  $\int_{\Theta_0} p(\theta) d\nu(\theta) > 0$  and all  $\theta^{(0)} \in \Theta$ .<sup>4</sup> It follows directly that if a kernel is Harris recurrent, then it is  $p$ -irreducible. A kernel whose invariant distribution is proper, and that is both aperiodic and Harris recurrent, is *ergodic* by definition (Tierney, 1994, pp. 1712-1713).

A useful metric in what follows is the total variation norm for signed and bounded measures  $\mu$  defined over the field of all  $\nu$ -measurable sets  $S_\nu$  on  $\Theta$ :  $\|\mu\| = \sup_{\Theta_0 \in S_\nu} \mu(\Theta_0) - \inf_{\Theta_0 \in S_\nu} \mu(\Theta_0)$ .

*Theorem 3.5.1. Convergence of continuous state Markov chains.* Suppose  $p(\theta | \mathbf{Y}_T, A)$  is an invariant distribution of the transition density  $K(\theta, \theta^*)$ .

- (A) If  $K$  is  $p(\theta | \mathbf{Y}_T, A)$ -irreducible, then  $p(\theta | \mathbf{Y}_T, A)$  is the unique invariant distribution.
- (B) If  $K$  is  $p(\theta | \mathbf{Y}_T, A)$ -irreducible and aperiodic, then except possibly for  $\theta^{(0)}$  in a set of posterior probability 0,  $\|P^{(m)}(\theta^{(0)}, \cdot) - P(\cdot | \mathbf{Y}_T, A)\| \rightarrow 0$ .

If  $K$  is ergodic (that is, it is also Harris recurrent) then this occurs for all  $\theta^{(0)}$ .

- (C) If  $K$  is ergodic with invariant distribution  $p(\theta | \mathbf{Y}_T, A)$ , then for all  $g(\mathbf{Y}_T, \theta)$  absolutely integrable with respect to  $p(\theta | \mathbf{Y}_T, A)$  and for all  $\theta^{(0)} \in \Theta$ ,

$$M^{-1} \sum_{m=1}^M g(\mathbf{Y}_T, \theta^{(m)}) \xrightarrow{a.s.} \int_{\Theta} g(\mathbf{Y}_T, \theta) p(\theta | \mathbf{Y}_T, A) d\nu(\theta).$$

*Proof.* (A) and (B) follow immediately from Theorem 1 and (C) from Theorem 3 in Tierney (1994). ##

For the Gibbs sampling algorithm we argued informally in Section 3.3 that  $p(\theta | \mathbf{Y}_T, A)$  is an invariant distribution. More formally, from (3.3.2) we have for the blocking  $\theta' = (\theta'_{(1)}, \theta'_{(2)})$ ,

$$\begin{aligned} \int_{\Theta} K_G(\theta, \theta^*) p(\theta | \mathbf{Y}_T, A) d\nu(\theta) &= \int_{\Theta} p(\theta_{(1)}^* | \mathbf{Y}_T, \theta_{(2)}, A) p(\theta_{(2)}^* | \mathbf{Y}_T, \theta_{(1)}, A) p(\theta | \mathbf{Y}_T, A) d\nu(\theta) \\ &= p(\theta_{(2)}^* | \mathbf{Y}_T, \theta_{(1)}, A) \int_{\Theta} p(\theta_{(1)}^* | \mathbf{Y}_T, \theta_{(2)}, A) p(\theta | \mathbf{Y}_T, A) d\nu(\theta) \\ &= p(\theta_{(2)}^* | \mathbf{Y}_T, \theta_{(1)}, A) p(\theta_{(1)}^* | \mathbf{Y}_T, A) = p(\theta^* | \mathbf{Y}_T, A). \end{aligned}$$

The general result for more than two blocks follows by induction. Thus, it is the uniqueness of the invariant state that is at issue in establishing convergence of the Gibbs sampler. The following result is immediate and is often easy to apply.

---

<sup>4</sup>The expression “i.o.” in  $P[\theta^{(m)} \in \tilde{\Theta} \text{ i.o.} | \theta^{(0)}] = 1$  means “infinitely often.” The condition is that  $\lim_{M \rightarrow \infty} P[\sum_{m=1}^M \chi_{\tilde{\Theta}}(\theta^{(m)}) \leq L] = 0 \forall L$ .

*Corollary 3.5.2. A first sufficient condition for convergence of the Gibbs sampler.* Suppose that for every point  $\theta^* \in \Theta$  and every  $\Theta_0 \subseteq \Theta$  with the property  $P(\theta \in \Theta_0 | \mathbf{Y}_T, A) > 0$ , it is the case that  $P_G(\theta^{(m+1)} \in \Theta_0 | \mathbf{Y}_T, \theta^{(m)} = \theta^*, A) > 0$ , where  $P_G(\cdot)$  is the probability measure induced by the Gibbs sampler. Then the Gibbs transition kernel is ergodic.

*Proof.* The conditions ensure that  $P_G$  is aperiodic and absolutely continuous with respect to  $p(\theta | \mathbf{Y}_T, A)$ . The result follows from Corollary 1 of Tierney (1994). ##

A complement to Corollary 3.5.2 is provided by Roberts and Smith (1994).

*Theorem 3.5.3. A second sufficient condition for convergence of the Gibbs sampler.* Suppose that  $p(\theta | \mathbf{Y}_T, A)$  is lower semicontinuous<sup>5</sup> at 0 and  $\int_{\Theta^{(b)}} p(\theta | \mathbf{Y}_T) d\nu(\theta^{(b)})$  is locally bounded ( $b = 1, \kappa, B$ ). Suppose also that  $\Theta$  is connected. Then the Gibbs transition kernel is ergodic. ##

Theorem 3.5.3 rules out situations like the one shown in Figure 3.5.1, where the posterior density is uniform on a closed set. For any point  $\theta$  on the boundary there is no open neighborhood  $N_\theta$  such that for all  $\theta^* \in N_\theta$ ,  $p(\theta^* | \mathbf{Y}_T, A)$  is bounded away from 0. The point  $A$  is absorbing. Tierney (1994) discusses weaker conditions for convergence of the Gibbs sampler. However, the conditions of Corollary 3.5.2 or Theorem 3.5.3 are satisfied for a very wide range of problems in econometrics and are easier to verify.

Tierney (1994) and Roberts and Smith (1994) show that the convergence properties of the Hastings-Metropolis algorithm are inherited from those of  $q(\theta, \theta^*)$ : if  $q$  is aperiodic and  $p(\theta | \mathbf{Y}_T, A)$ -irreducible, then so is the Hastings-Metropolis algorithm. This feature leads to a sufficient condition for convergence analogous to Corollary 3.5.2.

*Theorem 3.5.4. A first sufficient condition for convergence of the Hastings-Metropolis algorithm.* Suppose that for every point  $\theta^* \in \Theta$  and every  $\Theta_0 \subseteq \Theta$  with the property  $P(\theta \in \Theta_0 | \mathbf{Y}_T, A) > 0$ , it is the case that  $\int_{\Theta_0} q(\theta, \theta^*) \alpha(\theta, \theta^*) d\nu(\theta^*) + r(\theta) \chi_{\Theta_0}(\theta) > 0$ . Then the Hastings-Metropolis density  $K(\theta, \theta^*) = q(\theta, \theta^*) \alpha(\theta, \theta^*)$  is ergodic.

*Proof.* The conditions ensure that the transition kernel is aperiodic and  $p(\theta^* | \mathbf{Y}_T, A)$ -irreducible. Thus, by Corollary 2 of Tierney (1994), the Hastings-Metropolis density is Harris recurrent. Since the kernel is both aperiodic and Harris recurrent, it is ergodic. ##

---

<sup>5</sup>A function  $h(\mathbf{x})$  is lower semicontinuous at 0 if, for all  $\mathbf{x}$  with  $h(\mathbf{x}) > 0$ , there exists an open neighborhood  $N_{\mathbf{x}} \supset \mathbf{x}$  and  $\varepsilon > 0$  such that for all  $\mathbf{y} \in N_{\mathbf{x}}$ ,  $h(\mathbf{y}) \geq \varepsilon > 0$ .

A complementary sufficient condition for convergence of Hastings-Metropolis chains is provided by the following result, which is analogous to Theorem 3.5.3 for the Gibbs sampler.

*Theorem 3.5.5.* A second sufficient condition for convergence of the Hastings-Metropolis algorithm. Suppose that for every  $\theta \in \Theta$ ,  $p(\theta | \mathbf{Y}_T, A) > 0$ , and for all pairs  $(\theta^{(m)}, \theta^{(m+1)}) \in \Theta \times \Theta$ ,  $p(\theta^{(m)} | \mathbf{Y}_T, A)$  and  $q(\theta^{(m)}, \theta^{(m+1)})$  are positive and continuous. Then the Hastings-Metropolis kernel  $K_H$  is ergodic.

*Proof.* See Chib and Greenberg (1995) or Mengersen and Tweedie (1993). ##

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

### 3.6 Variants

There are many variations on these methods, and alone or in combination with each other they provide a powerful source of flexibility that can be drawn upon in construction posterior simulators. Here we briefly review two. Further discussion can be found in Tierney (1994) and Gelman et al. (1995).

#### *Mixtures and combinations*

Suppose  $K^{(j)}(\theta^{(m)}, \theta^{(m+1)})$  ( $j = 1, \dots, J$ ) are Markov chain kernels, each with unique invariant distribution  $p(\theta | \mathbf{Y}_T, A)$ . In a mixture, positive probabilities  $\gamma_1, \dots, \gamma_J$  with  $\sum_{j=1}^J \gamma_j = 1$  are specified, and at each step one of the kernels is selected accordingly. The candidate  $\theta^*$  is drawn from the transition probability density selected, and the acceptance probability (3.4.1) is based upon the  $q(\cdot, \cdot)$  of the kernel selected. Observe that if one of the kernels in a mixture is Harris recurrent, then so is the mixture, and if one of the kernels in the mixture is aperiodic, so is the mixture. Hence if one of the kernels in a mixture is ergodic, so is the mixture kernel.

A combination is a variant on this strategy: construct a single transition density  $q(\theta, \theta^*) = \sum_{j=1}^J \gamma_j q^{(j)}(\theta, \theta^*)$  where each  $q^{(j)}(\theta, \theta^*)$  is a probability density function in  $\theta^*$ ,  $\gamma_j > 0$  ( $j = 1, \dots, J$ ) and  $\sum_{j=1}^J \gamma_j = 1$ . If a single transition density  $q^{(j)}(\theta, \theta^*)$  is ergodic, then so is the combination.

The use of mixtures or combinations is often key in successful applications of the Hastings-Metropolis algorithm. For example, if the log likelihood function and its first two

derivatives can be evaluated in closed form, then generic versions of the Hastings-Metropolis algorithm can be constructed that work well in a wide variety of applications (Geweke, 1998). The idea is that a candidate can be chosen from one of several distributions: for example, the mixture could include a normal or Student- $t$  distribution fit to the global posterior mode; a similar random walk component, with location vector equal to the current value and scale matrix determined from the Hessian of the log posterior at the current value; and the prior distribution. “Local” components of the mixture, like the random walk, adjusted to the local shape of the posterior, tend to concentrate candidate draws in regions where acceptance is likely. “Global” components of the transition, like the prior, have lower acceptance probability but cause the algorithm to explore distant regions of the parameter space sooner than would otherwise be the case. As a specific example, in the case of the probit model it is straightforward to integrate the latent variables explicitly and write the posterior density kernel in standard form,

$$(3.6.1) \quad (2\pi)^{-k/2} |\underline{\mathbf{H}}_{\beta}|^{1/2} \exp\left[-.5(\beta - \underline{\beta})' \underline{\mathbf{H}}_{\beta}(\beta - \underline{\beta})\right] \prod_{t=1}^T \{d_t \Phi(\beta' \mathbf{x}_t) + (1 - d_t)[1 - \Phi(\beta' \mathbf{x}_t)]\}.$$

The gradient and Hessian of the log posterior density kernel are easily derived (see Greene (1997, Section 19.4) for the relevant portions from the likelihood) and a Hastings-Metropolis algorithm for this model is straightforward to implement. We shall return to a comparison of the Gibbs sampler and Hastings-Metropolis algorithm for this model in Section 3.9.

### *Metropolis within Gibbs*

Another variant in MCMC is to use conditioning and then apply a more basic strategy to the conditional distribution. For example, draws from a multivariate transition density entail both conditioning and acceptance sampling, although this process is transparent in most software (Geweke, 1996, Section 2). Another such strategy, that is quite useful in Bayesian econometrics, is the Metropolis within Gibbs algorithm (Zeger and Karim (1991), Chib and Greenberg (1996)). In a two-block Gibbs sampler, suppose that it is straightforward to sample from  $p(\theta_{(1)} | \mathbf{Y}_T, \theta_{(2)}, A)$ , but the distribution corresponding to  $p(\theta_{(2)} | \mathbf{Y}_T, \theta_{(1)}, A)$  is intractable. The Hastings-Metropolis algorithm can be used in these circumstances, and it often provides an efficient solution to the problem. In what has become known as the Metropolis-within-Gibbs procedure, at the  $(m+1)$ 'th iteration first draw  $\theta_{(2)}^*$  from a proposal density  $q(\theta_{(2)}^{(m)}, \theta_{(2)}^* | \theta_{(1)}^{(m+1)})$ . Accept this draw with probability

$$\min \left\{ \frac{\mathbb{P}(\boldsymbol{\theta}_{(1)}^{(m+1)}, \boldsymbol{\theta}_{(2)}^* | \mathbf{Y}_T, A) / \mathbb{Q}(\boldsymbol{\theta}_{(2)}^{(m)}, \boldsymbol{\theta}_{(2)}^* | \boldsymbol{\theta}_{(1)}^{(m+1)})}{\mathbb{P}(\boldsymbol{\theta}_{(1)}^{(m+1)}, \boldsymbol{\theta}_{(2)}^{(m)} | \mathbf{Y}_T, A) / \mathbb{Q}(\boldsymbol{\theta}_{(2)}^*, \boldsymbol{\theta}_{(2)}^{(m)} | \boldsymbol{\theta}_{(1)}^{(m+1)})}, 1 \right\}.$$

If is  $\boldsymbol{\theta}_{(2)}^*$  accepted then  $\boldsymbol{\theta}_{(2)}^{(m+1)} = \boldsymbol{\theta}_{(2)}^*$ , and if not then  $\boldsymbol{\theta}_{(2)}^{(m+1)} = \boldsymbol{\theta}_{(2)}^{(m)}$ . The extension of this procedure to multi-block Gibbs samplers, with a Hastings-Metropolis algorithm used at some (or even all) of the blocks is clear. For further discussion see Chib and Greenberg (1994), and for a proof that the posterior distribution is an invariant state of this Markov chain see Chib and Greenberg (1996).

### 3.7 Assessing numerical accuracy in Markov chain Monte Carlo

In any practical application one is concerned with the discrepancy  $\bar{g}_M - \bar{g}$ . A leading analytical tool for assessing this discrepancy is a central limit theorem, if one can be obtained. This was accomplished in Section 3.1 for i.i.d. sampling from the posterior distribution, and in Section 3.2 for importance sampling. The assumption of independence, key to those results, does not apply in Markov chain Monte Carlo. The weaker assumption of uniform ergodicity yields a central limit theorem, however. Let  $\mathbb{P}^{(m)}(\boldsymbol{\theta}^{(0)}, \Theta_0)$  denote  $\mathbb{P}(\boldsymbol{\theta}^{(m)} \in \Theta_0 | \boldsymbol{\theta}^{(0)})$  for any  $\boldsymbol{\theta}^{(0)} \in \Theta$  and for any  $\Theta_0 \subseteq \Theta$  for which  $\mathbb{P}(\boldsymbol{\theta} \in \Theta_0 | \mathbf{Y}_T, A)$  is defined. The Markov chain is *uniformly ergodic* if  $\sup_{\boldsymbol{\theta} \in \Theta} \|\mathbb{P}^{(m)}(\boldsymbol{\theta}, \cdot) - \mathbb{P}(\cdot | \mathbf{Y}_T, A)\| \leq Mr^m$  for some  $M > 0$  and some positive  $r < 1$ .

Tierney (1994, p. 1714) demonstrates two results that are quite useful in establishing uniform ergodicity. First, an independence Metropolis kernel with bounded weighting function  $w(\boldsymbol{\theta}) = \mathbb{p}(\boldsymbol{\theta} | \mathbf{Y}_T) / j(\boldsymbol{\theta})$  is uniformly ergodic. (Recalling the similarity between the independence Metropolis kernel and importance sampling and Corollary 3.5.3 this result is not surprising.) Second, if one kernel in a mixture of kernels is uniformly ergodic, then the mixture kernel itself is uniformly ergodic.

The interest in uniform ergodicity stems from the following central limit theorem. Note how close this result is to Corollary 3.2.2.

*Theorem 3.7.1. A central limit theorem for Markov chain Monte Carlo.* Suppose  $\{\boldsymbol{\theta}^{(m)}\}$  is uniformly ergodic with equilibrium distribution  $\mathbb{p}(\boldsymbol{\theta} | \mathbf{Y}_T, A)$ . Suppose further that  $\mathbb{E}[g(\mathbf{Y}_T, \boldsymbol{\theta}) | \mathbf{Y}_T, A] = \bar{g}$  and  $\text{var}[g(\mathbf{Y}_T, \boldsymbol{\theta}) | \mathbf{Y}_T, A]$  exist and are finite, and let  $\bar{g}_M = M^{-1} \sum_{m=1}^M g(\mathbf{Y}_T, \boldsymbol{\theta}^{(m)})$ . Then there exists finite  $\sigma^2$  such that

$$(3.7.1) \quad M^{1/2}(\bar{g}_M - \bar{g}) \xrightarrow{d} \mathbb{N}(0, \sigma^2).$$

*Proof.* Tierney (1994, Theorem 5), attributed to Cogburn (1972, Corollary 4.2(ii)).

##

Thus for any Markov chain  $\{\theta^{(m)}\}$  with invariant distribution  $p(\theta|Y_T, A)$ , one can guarantee (3.7.1) by mixing the chain with an independence Metropolis kernel with a bounded weighting function, so long as the posterior mean and variance are known to exist. If the likelihood function is bounded, then the prior distribution itself will provide such an independence transition kernel.

Nevertheless, some practical concerns remain. One difficulty is that useful conditions sufficient for approximation of the unknown constant  $\sigma^2$  have not yet been developed. That is, there is no  $\hat{\sigma}_M^2$  for which  $\hat{\sigma}_M^2 \rightarrow \sigma^2$  as there is for independence and importance sampling. A second difficulty is assessing the sensitivity of  $\theta^{(m)}$  to the initial condition  $\theta^{(0)}$ . For example, consider the Gibbs sampler in the case of a multimodal posterior density. In the limiting case of Figure 3.3.1 the Markov chain is reducible. As that case is approached sensitivity to the initial condition increases, as does serial correlation, since the probability that  $\theta^{(m)}$  will be in one region conditional on  $\theta^{(m-1)}$  being in the other goes to zero. Assessing convergence given the possibility of such problems is clearly nontrivial.

There is an extensive literature on this problem. A good introduction is provided by the papers of Gelman and Rubin (1992) and Geyer (1992) and their discussants. Geweke (1992) developed a consistent estimator of  $\sigma^2$  in (3.7.1), under the strong condition that conventional time series mixing conditions (for example Hannan, 1970, pp. 207-210) apply to  $\{\theta^{(m)}\}$ . There is no analytical foundation for this assumption, but these methods are now widely used and have proven reliable in the sense that they predict well the behavior of the Markov chain when it is restarted with a new initial condition, in econometric models.

In practice, some robustness to initial conditions is achieved by discarding initial iterations: 10% to 20% is common. By drawing  $\theta^{(0)}$  from the prior distribution, using a random number generator with a fresh seed each time, several runs may provide some indication of whether the results are sensitive to initial conditions as they might be, for example, given near-reducibility of the kind that may arise from severe multimodality. A formal test for sensitivity to initial conditions was developed by Gelman and Rubin (1992) and is described in Section 3.8. For other tests for sensitivity to initial conditions see Geweke (1992) and Zellner and Min (1995).

### 3.8 Software

Posterior simulation software for some econometric models is publicly available at [www.econ.umn.edu/~bacc](http://www.econ.umn.edu/~bacc). This site also provides software that facilitates the approximation of the investigator's posterior moments (described here), the approximation

of marginal likelihoods (described in Section 4.5), the approximation of moments not recorded by the investigator (Section 6.2), modification of the investigator's prior distribution (Section 6.2), and other computations based on posterior simulator output. Posterior simulation software is available as Fortran source code and DOS executable files. All other software is available in six languages: Fortran, c, Gauss, Matlab, Mathematica, and Splus.<sup>6</sup>

All the software is organized around the creation and subsequent use of posterior simulator files. A posterior simulator file is initially the output of a posterior simulator designed for a particular econometric model. For each iteration it records, at a minimum, the full parameter vector. In general every  $s$ 'th iteration of a posterior simulator is recorded.

The initial record of a posterior simulator file consists of two integers: the first is the number of iterations, and the second is the number of entries in the vector written in each iteration.

For each iteration, two records are written. The first record is an integer followed by three real constants. The integer is the iteration number; it reflects the number of skips ( $s-1$ ), if any, between iterations. (This integer is only for convenience in examining the posterior simulator file and is not used in any way by any of the software.) The first real constant is the logarithm of the weighting function, that is, the log ratio of posterior density kernel to importance sampling kernel; for many MCMC methods, this value is zero. The second real constant is the logarithm of the prior density (not merely the kernel),  $\log p(\theta|A)$ , at the parameter vector for the iteration. The third real constant is the logarithm of the data density (not merely the kernel),  $\log p(\mathbf{Y}_T|\theta, A)$ , at the parameter vector for the iteration.

The second record for each iteration is a vector of parameters and (perhaps) functions of these parameters, written five entries per line and in general occupying multiple lines. The organization of this vector is specific to the particular application, and it is necessary to know how the vector has been set up in order to make sense of the posterior simulator file.

The program **moment** calculates posterior means and posterior standard deviations, assesses the numerical accuracy of the posterior means, and optionally writes a machine readable file for subsequent use by the program **apm** described below.<sup>7</sup> Each column of

---

<sup>6</sup>Complete documentation for all software is provided at the website. Since this software will continue to be developed and improved, some details provided in this article will become outdated. Users should rely on the website documentation for actual use rather than the descriptions in this article, which are intended to provide concrete examples of how Bayesian inference, development and communication can proceed.

<sup>7</sup>The structure of inputs to this and all other programs is specific to the language in which the program is written. Technical details are provided at the website.

the posterior simulator matrix corresponds to a function of interest  $g(\theta, \mathbf{Y}_T)$ . For each column indicated, **moment** computes a numerical approximation to the posterior mean of this function, ignoring the first  $r$  iterations of the  $M$  posterior simulations and using only the last  $M - r$ .

The numerical approximation of the posterior mean of the function of interest is

$$\tilde{g} = \sum_{m=r+1}^M w(\theta^{(m)}) g(\mathbf{Y}_T, \theta^{(m)}) / \sum_{m=r+1}^M w(\theta^{(m)})$$

where  $g(\mathbf{Y}_T, \theta^{(m)})$  is the evaluation of  $g(\mathbf{Y}_T, \theta)$  in the  $m$ 'th iteration.

The numerical approximation of the posterior standard deviation of the function of interest is

$$\left\{ \sum_{m=r+1}^M w(\theta^{(m)}) \left[ g(\mathbf{Y}_T, \theta^{(m)}) - \tilde{g} \right]^2 / \sum_{m=r+1}^M w(\theta^{(m)}) \right\}^{1/2}.$$

Four variants of a numerical standard error for the accuracy of the approximation of the true posterior moment  $E[g(\mathbf{Y}_T, \theta) | \mathbf{Y}_T, A]$  by the numerical approximation  $\tilde{g}$  are provided. To describe these, let  $n_M = (M - r)^{-1} \sum_{m=r+1}^M w(\theta^{(m)}) g(\theta^{(m)}, \mathbf{Y}_T)$  denote the numerator of  $\tilde{g}$  and let  $d_M = (M - r)^{-1} \sum_{m=r+1}^M w(\theta^{(m)})$  denote the denominator of  $\tilde{g}$ . Using the conventional asymptotic expansion (“delta method”),

$$\text{var}(n_M/d_M) \approx \begin{bmatrix} d_M^{-1} & -n_M d_M^{-2} \end{bmatrix} \begin{bmatrix} \text{var}(n_M) & \text{cov}(n_M, d_M) \\ \text{cov}(d_M, n_M) & \text{var}(d_M) \end{bmatrix} \begin{bmatrix} d_M^{-1} \\ -n_M d_M^{-2} \end{bmatrix}$$

The four variants of the numerical standard error,  $\text{s.e.}(n_M/d_M) = [\text{var}(n_M/d_M)]^{1/2}$ , are based on different approximations of  $\text{var}(n_M)$ ,  $\text{cov}(n_M, d_M)$  and  $\text{var}(d_M)$ .

The first method assumes no serial correlation in  $\{\theta^{(m)}\}$ , and is appropriate for independence or importance sampling. Following Geweke (1989b) this leads to

$$\text{var}(n_M/d_M) \approx \sum_{m=r+1}^M w(\theta^{(m)})^2 \left[ g(\theta^{(m)}, \mathbf{Y}_T) - \tilde{g} \right]^2 / \left[ \sum_{m=r+1}^M w(\theta^{(m)}) \right]^2.$$

(The square root of the value on the right hand side is reported by **moment**.)

In the other three methods  $\text{var}(n_M)$ ,  $\text{cov}(n_M, d_M)$  and  $\text{var}(d_M)$  are approximated using conventional time series methods for a wide sense stationary process similar to those described in Geweke (1992). In the case of  $\text{var}(d_M)$ ,

$$(3.8.1) \quad \text{var}(d_M) \approx (M - r)^{-1} \sum_{s=-L+1}^{L-1} [(L - s)/L] c(s)$$

where for  $s \geq 0$ ,  $c(s) = c(-s) = (M - r)^{-1} \sum_{m=r+s+1}^M \left[ w(\theta^{(m)}) - d_M \right] \left[ w(\theta^{(m-s)}) - d_M \right]$ . For  $\text{var}(n_M)$ , the approximation is the right hand side of (3.8.1) but with

$$c(s) = c(-s) = (M - r)^{-1} \sum_{m=r+s+1}^M \left[ g(\theta^{(m)}) w(\theta^{(m)}) - n_M \right] \left[ g(\theta^{(m-s)}) w(\theta^{(m-s)}) - n_M \right],$$

where  $s \geq 0$ . For  $\text{cov}(n_M, d_M)$  the approximation is the right hand side of (3.8.1) but with

$$c(s) = (M - r)^{-1} \sum_{m=\max(r+1, r+s+1)}^{\min(M, M-s)} \left[ g(\theta^{(m)}) w(\theta^{(m)}) - n_M \right] \left[ w(\theta^{(m-s)}) - d_M \right].$$

The three variants differ in the value of  $L$  chosen. In the first,  $L = .04(M - r)$ ; in the second  $L = .08(M - r)$ ; and in the third,  $L = .15(M - r)$ .<sup>8</sup>

The program **apm** combines posterior moments from two or more machine-readable **moment** output files. It provides numerical standard errors and conventional test statistics for the equality of these moments, under the assumption that the posterior simulations from which the **moment** output files were created, are independent of one another. The user specifies the number of machine-readable output files created by **moment**, the number of moments in each file, and the names of the machine-readable **moment** output files. The number of moments must be the same in each file.

The program **apm** produces these results four different times, corresponding to the four variants of the numerical standard error for a posterior moment just discussed. If there are  $J$  **moment** output files, the combined posterior moment approximation is  $\tilde{g} = \sum_{j=1}^J v_j \tilde{g}_j / \sum_{j=1}^J v_j$ , where  $\tilde{g}_j$  is the moment in the  $j$ 'th file and  $v_j$  is the inverse square of its numerical standard error. The program **apm** provides the conventional chi-square test of  $\tilde{g}_1 = \dots = \tilde{g}_J$  (in four variants), and the marginal significance level of this test statistic. If the  $J$  **moment** output files were created using  $J$  independent initial conditions for the same posterior simulator, then this test is essentially the convergence test proposed by Gelman and Rubin (1992).

Evidence of different values of the moments from different files is an indication that there may be sensitivity to starting values, or -- almost equivalently -- that an insufficient number of burn-in iterations were taken in approximating the moments.

### 3.9 Examples

Two examples illustrate the use of these methods, and will be used in subsequent portions of this paper as well.<sup>9</sup> The first example is based on the hedonic model of residential real estate prices discussed by Anglin and Gencay (1996). Their baseline model is a linear regression of the logarithm of sales prices on an intercept and eleven attributes. The attributes are indicated in the leftmost column of Table 3.1. All variables beginning with “#” are positive integers,  $\log(\text{Lot size})$  is continuous, and all other variables are dichotomous (1 if present and 0 if not). The data consist of 546 transactions during July,

---

<sup>8</sup>Small values of  $L$  assume a more rapid rate of decay in the autocovariance function of  $\{g(\mathbf{Y}_T, \theta^{(m)})\}$ . In practice results are usually about the same for the three values. Substantial differences indicate that serial correlation may persist across a substantial fraction of the iterations, and a longer simulation may be warranted.

<sup>9</sup>Data for both examples are available at <http://www.econ.umn.edu/~geweke/papers.html>. Data for the first example are also available at <http://qed.econ.queens.ca:80/jae/1996-v11.6/anglin-gencay>.

August and September 1987 in metropolitan Windsor, Ontario. The least squares estimates match those reported in Anglin and Gencay (1996).

The form of the prior distribution is the one discussed in Section 2.1 for the normal linear model. The normal distribution for the coefficient vector has mean  $\mathbf{0}$ , the precision matrix is diagonal, and standard deviations are chosen to allow reasonable values of the coefficients. The prior distribution of the precision parameter is  $.12h \sim \chi^2(3)$  so that  $h$  has prior mean 25 and standard deviation about 20. The posterior simulator is the Gibbs sampling algorithm described in Section 3.3, based on the conditional distributions (2.1.17) and (2.1.18). The posterior simulator file was created using the program **uvr1**, available at the indicated website both as an executable DOS file and as portable Fortran code. Creation of 10,000 records in the posterior simulator file required 69 seconds.<sup>10</sup> The last four columns of Table 3.1 report results from **moment** using this posterior simulator file, discarding the first 1,000 iterations. Initial values here, and in all other examples discussed in this paper, were drawn from the prior distribution, and in every case the first 1,000 draws were discarded. The **moment** computations took 27 seconds. Posterior means and standard deviations are close to the least squares values, reflecting the lack of information in the prior relative to the data set. The numerical standard error (NSE) of each posterior mean is given for  $L = .08(M - r)$  as described in the previous section.<sup>11</sup> The NSEs imply accuracy of more than two figures past the decimal in the posterior means, and the relative numerical efficiency (RNE) indicates that numerical accuracy is comparable to what would have been achieved with i.i.d. drawings directly from the posterior distribution.

The second example is a probit model of women's labor force participation, based on the one presented in Geweke and Keane (1998). The data consist of 1,555 observations of women in the 1987 Panel Survey of Income Dynamics. The choice variable is 1 if a woman reports positive hours of work for 1987. The covariates are indicated in the left column of Table 3.2. Black, Married and Kids are dichotomous. Age is measured in years and interacted with Married and its negation. Education is years of completed schooling. Spouse\$ is husband's income and Family\$ is unearned household income, in dollars for the year 1987. Work experience is measured in cumulative hours since the woman became a household head or spouse. For each unmarried woman with children, AFDC is the monthly cash support she would receive if she did not work. Food\$ is the monthly food stamp allotment to which a woman's household would be entitled if she did not work. The

---

<sup>10</sup>All execution times are given for a Sun Ultra 200 Sparcstation 20. This machine is about twice as fast as the fastest pentium processors.

<sup>11</sup>This value of  $L$  is used to report NSE and RNE in all other examples in the article, as well. Results using other values of  $L$  are similar.

last two variables differ according to the state of residence. The prior distributions of the coefficients are independent normal, each with mean zero and standard deviation chosen to permit large but reasonable values to be within two standard deviations of zero. Details of sample screening, variable descriptions and prior construction are given in Geweke and Keane (1998).

Conventional maximum likelihood estimates, the posterior mode, and approximate posterior standard deviations based in the usual way on the Hessian of the log-posterior at the mode<sup>12</sup> are presented in Table 3.2. Computation of these values using analytical gradient and Hessian required 9 seconds. It is clear from Table 3.2 that the prior distribution is informative, relative to the data: prior standard deviations of six of the thirteen coefficients are smaller than the corresponding maximum likelihood asymptotic standard errors. Posterior standard deviations are smaller than asymptotic standard errors in every case, and for several coefficients the posterior mode is closer to the prior mode than to the likelihood mode.

Two alternative posterior simulation algorithms were used to construct posterior simulator files. The first is the Gibbs sampler for the probit model developed in Albert and Chib (1993) and described in Section 3.7, available as program **pbt1** at the website. Creation of a posterior simulator file with 10,000 records required 680 seconds. Posterior means, standard deviations and measures of numerical accuracy, based on the last 9000 records, are presented in the left half of Table 3.3. The posterior moments are quite close to the approximation at the posterior mode, very likely reflecting a posterior distribution that is close to multivariate normal. The average relative numerical efficiency for the coefficients indicates that the same accuracy could have been achieved with about 35% the number of iterations, had an i.i.d. sample been drawn directly from the posterior distribution.

The second posterior simulator is a Hastings-Metropolis algorithm, constructing a transition density as the combination of two densities as described in Section 3.6. The first density is the prior, with a weight of .2. The second density is a multivariate Student-*t* distribution centered at the posterior mode with 10 degrees of freedom and scale matrix set to the Hessian of the log posterior at the mode, and a weight of .8. Out of 10,000 iterations 1,977 candidates were drawn from the prior, of which 1 was accepted, and 8,023 were drawn from the multivariate Student-*t*, of which 5,767 were accepted. As indicated in Table 3.3, numerical accuracy is comparable to the Gibbs sampler with the same number of draws. Execution time was 572 seconds, about 15% faster than the Gibbs sampler.

---

<sup>12</sup>For details and the asymptotic justification for this approximation see Bernardo and Smith (1994, Section 5.3) and references given there.

The two sets of results in Table 3.3 provide an opportunity to check on the adequacy of the assumptions underlying the implicit use of a central limit theorem in evaluating numerical accuracy. The last column of that table provides the conventional “*t*” statistic for equality of posterior means using the reported NSEs. The values obtained are consistent with the joint assumptions that the invariant distribution of the Markov chain was reached by the 1,000’th iteration, and that the central limit approximation described in Section 3.7 is valid.

## 4. Model Comparison

Section 2.3 demonstrated that given prior probabilities over models, and prior probability distributions for parameter vectors within models, there is a complete theory of model combination and model comparison. The central technical task in implementing the theory is calculation of the marginal likelihood  $p(\mathbf{Y}_T|A) = \int_{\Theta} p(\mathbf{Y}_T|\theta, A)p(\theta|A)d\nu(\theta)$ . The marginal likelihood cannot, in general, be cast in the form of a posterior moment (2.1.5), and therefore the posterior simulation methods of Section 3, which have proven useful in obtaining posterior moments in a single model, are not directly applicable to this problem. A decade ago, there were essentially no methods developed for the numerical approximation of marginal likelihoods or Bayes factors and results were limited to a handful of cases for which there were analytical results or asymptotic approximations. Now, it is possible to attain good and generic approximations to marginal likelihoods in most cases; however, some models with large numbers of latent variables remain troublesome.

This section provides several approaches to the approximation of marginal likelihoods, with an emphasis on generic methods that are consistent as the number of simulations increases. Generic methods exclude those that are ingenious but specific to particular situations as well as methods that rely on asymptotic approximations rather than simulation. Many of these methods are discussed in a comprehensive review article by Kass and Raftery (1995). We discuss here a method that works well with importance sampling and the Hastings-Metropolis algorithm; a method specific to the Gibbs sampler; and a generic method that works well with most posterior simulators regardless of the algorithm employed. For the last method, we describe publicly available software designed to work with the most commonly used computing platforms in econometrics. Some examples illustrate the numerical accuracy that can be attained, and provide comparisons of some of the different methods of approximating marginal likelihoods.

### 4.1 Importance sampling and the Hastings-Metropolis algorithm

Suppose that  $j(\theta)$ , with support  $\Theta$ , is the probability density function (not just a kernel) with respect to the measure  $d\nu(\theta)$  of an importance sampling distribution for the posterior density  $p(\theta|\mathbf{Y}_T, A) \propto p(\theta|A)p(\mathbf{Y}_T|\theta, A)$ , where  $p(\theta|A)$  is the properly normalized prior density and  $p(\mathbf{Y}_T|\theta, A)$  is the properly normalized data density. Define the weighting function  $w(\theta) = p(\theta|A)p(\mathbf{Y}_T|\theta, A)/j(\theta)$ .

*Corollary 4.1.1.* Let  $j(\theta)$  be the importance sampling density in an importance sampling algorithm. Suppose the support of  $j(\theta)$  includes  $\Theta$ . Then

$$\bar{w}_M = M^{-1} \sum_{m=1}^M w(\theta^{(m)}) \xrightarrow{a.s.} \int_{\Theta} w(\theta) j(\theta) d\nu(\theta) = \int_{\Theta} p(\theta|A) p(\mathbf{Y}_T|\theta, A) d\nu(\theta) = p(\mathbf{Y}_T|A).$$

If  $w(\theta)$  is bounded above then

$$M^{1/2} [\bar{w}_M - p(\mathbf{Y}_T|A)] \xrightarrow{d} N(0, \sigma^2), \quad M^{-1} \sum_{m=1}^M [w(\theta^{(m)}) - \bar{w}_M]^2 \xrightarrow{a.s.} \sigma^2.$$

*Proof.* Immediate from Theorem 3.2.1 and Corollary 3.2.2. ##

The first application of this idea is Geweke (1989a); see also Gelfand and Dey (1994) and Raftery (1995). Since  $w(\theta^{(m)})$  must be computed each iteration of the importance sampling algorithm in any event and the normalizing constant for  $j(\theta)$  is usually known, this simulation-consistent approximation of  $p(\mathbf{Y}_T|A)$  may be obtained at essentially no additional cost.

In the case of the Hastings-Metropolis algorithm there is a similar result. To motivate this result let  $q(\theta^{(m)}, \theta^*)$  be the transition probability density function, and denote the candidate draw on the  $m$ 'th iteration by  $\theta^{*(m)}$  whether it is accepted or not. Define  $w(\theta^{(m)}, \theta^{*(m)}) = p(\theta^{*(m)}|A) p(\mathbf{Y}_T|\theta^{*(m)}, A) / q(\theta^{(m)}, \theta^{*(m)})$ . If the support of  $q(\theta^{(m)}, \theta^*)$  is  $\Theta$  for all  $\theta^{(m)}$ , then  $E[w(\theta^{(m)}, \theta^{*(m)}) | \theta^{(m)}] = \int_{\Theta} p(\theta^*|A) p(\mathbf{Y}_T|\theta^*, A) d\nu(\theta^*) = p(\mathbf{Y}_T|A)$ . This motivates the following result.

*Theorem 4.1.2.* Let  $q(\theta^{(m)}, \theta^*)$  be the transition probability density function for  $\theta^*$  given  $\theta^{(m)}$  in a Hastings-Metropolis algorithm, and let  $\theta^{*(m)}$  denote the proposal drawn on the  $m$ 'th iteration. Suppose the support of  $q(\theta^{(m)}, \theta^*)$  is  $\Theta$  for all  $\theta^{(m)}$ , and that the Hastings-Metropolis Markov chain  $\{\theta^{(m)}\}$  is ergodic. Define the weighting function  $w(\theta^{(m)}, \theta^{*(m)}) = p(\theta^{*(m)}|A) p(\mathbf{Y}_T|\theta^{*(m)}, A) / q(\theta^{(m)}, \theta^{*(m)})$ . Then

$$\bar{w}_M = M^{-1} \sum_{m=1}^M w(\theta^{(m)}, \theta^{*(m)}) \xrightarrow{a.s.} \int_{\Theta} p(\theta|A) p(\mathbf{Y}_T|\theta, A) d\nu(\theta) = p(\mathbf{Y}_T|A).$$

If  $\{\theta^{(m)}\}$  is uniformly ergodic and  $w(\theta, \theta^*)$  is uniformly bounded above, then

$$M^{1/2} [\bar{w}_M - p(\mathbf{Y}_T|A)] \xrightarrow{d} N(0, \sigma^2).$$

*Proof.* See Geweke (1998). ##

The conditions of Theorem 4.1.2 are not as strong as they might appear. Recall from Section 3.6 that if one kernel in a mixture of kernels is uniformly ergodic, then the mixture kernel itself is uniformly ergodic. If the likelihood function is bounded above, and one of the kernels in the mixture (or a combination) is the prior distribution, then all the conditions of Theorem 4.1.2 will be met, and moreover there will be a central limit theorem. This

result is remarkably similar to the central limit theorem in Corollary 4.1.1 for importance sampling. In each case boundedness of the ratio of the posterior to candidate generating density leads to a strong result on approximation of the marginal likelihood.

## 4.2 The Gibbs sampler

In the case of the Gibbs sampler there is a different procedure due to Chib (1995) that provides accurate evaluations of the marginal likelihood, at the cost of additional simulations. Suppose that the output from the blocking  $\theta' = (\theta'_{(1),\mathcal{K}}, \theta'_{(B)})$  is available, and that the conditional p.d.f.'s  $p(\theta_{(j)} | \theta_{(i)} (i \neq j), \mathbf{Y}_T, A)$  can be evaluated in closed form for all  $j$ . (This latter requirement is generally satisfied.)

From (2.1.1)-(2.1.2),

$$(4.2.1) \quad p(\mathbf{Y}_T | A) = p(\tilde{\theta} | A) p(\mathbf{Y}_T | \tilde{\theta}, A) / p(\tilde{\theta} | \mathbf{Y}_T, A)$$

for any  $\tilde{\theta} \in \Theta$ . Typically  $p(\mathbf{Y}_T | \tilde{\theta}, A)$  and  $p(\tilde{\theta} | A)$  can be evaluated in closed form, but  $p(\tilde{\theta} | \mathbf{Y}_T, A)$  cannot. A marginal/conditional decomposition of  $p(\tilde{\theta} | \mathbf{Y}_T, A)$  is

$$(4.2.2) \quad p(\tilde{\theta} | \mathbf{Y}_T, A) = p(\tilde{\theta}_{(1)} | \mathbf{Y}_T, A) p(\tilde{\theta}_{(2)} | \mathbf{Y}_T, \tilde{\theta}_{(1)}, A) \cdot \mathcal{K} \cdot p(\tilde{\theta}_{(B)} | \mathbf{Y}_T, \tilde{\theta}_{(1),\mathcal{K}}, \tilde{\theta}_{(B-1)}, A).$$

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

$$M^{-1} \sum_{m=1}^M p(\tilde{\theta}_{(1)} | \mathbf{Y}_T, \theta_{(2)}^{(m),\mathcal{K}}, \theta_{(B)}^{(m)}, A) \xrightarrow{a.s.} p(\tilde{\theta}_{(1)} | \mathbf{Y}_T, A).$$

To approximate  $p(\tilde{\theta}_{(b)} | \mathbf{Y}_T, \tilde{\theta}_{(1),\mathcal{K}}, \tilde{\theta}_{(b-1)}, A)$ , first execute the Gibbs sampler with the parameters in the first  $b-1$  blocks fixed at the indicated values, thus producing a sequence  $\{\theta_{(b),(b+1),\mathcal{K}}^{(m)}, \theta_{(b),(B)}^{(m)}\}$  from the conditional posterior. Then

$$M^{-1} \sum_{m=1}^M p(\tilde{\theta}_b | \mathbf{Y}_T, \tilde{\theta}_{(1),\mathcal{K}}, \tilde{\theta}_{(b-1)}, \theta_{(b),(b+1),\mathcal{K}}^{(m)}, \theta_{(b),(B)}^{(m)}, A) \xrightarrow{a.s.} p(\tilde{\theta}_{(b)} | \mathbf{Y}_T, \tilde{\theta}_{(1),\mathcal{K}}, \tilde{\theta}_{(b-1)}, A).$$

These approximations are then used in (4.2.1) and (4.2.2) to obtain the approximation to the marginal likelihood. In general this method is more efficient the greater is  $p(\tilde{\theta} | \mathbf{Y}_T, A)$ , so in many applications it is natural to choose  $\tilde{\theta}$  near the posterior mode. It is straightforward to apply the methods of Section 3.7 to evaluate the numerical accuracy of the final approximation to the marginal likelihood, using standard delta methods. See Chib (1995) on these and other important practical details.

## 4.3 Modified harmonic mean

Gelfand and Dey (1994) observe that for any p.d.f.  $f(\theta)$  whose support is contained in  $\Theta$ ,

$$\begin{aligned}
(4.3.1) \quad \mathbb{E} \left[ \frac{f(\theta)}{p(\theta|A)p(\mathbf{Y}_T|\theta,A)} \middle| \mathbf{Y}_T, A \right] &= \int_{\Theta} \frac{f(\theta)}{p(\theta|A)p(\mathbf{Y}_T|\theta,A)} p(\theta|\mathbf{Y}_T, A) d\nu(\theta) \\
&= \int_{\Theta} \frac{f(\theta)}{p(\theta|A)p(\mathbf{Y}_T|\theta,A)} \cdot \frac{p(\theta|A)p(\mathbf{Y}_T|\theta,A)}{\int_{\Theta} p(\theta|A)p(\mathbf{Y}_T|\theta,A) d\nu(\theta)} d\nu(\theta) \\
&= \frac{\int_{\Theta} f(\theta) d\nu(\theta)}{\int_{\Theta} p(\theta|A)p(\mathbf{Y}_T|\theta,A) d\nu(\theta)} = p(\mathbf{Y}_T|A)^{-1}.
\end{aligned}$$

Thus the posterior mean of the function of interest  $f(\theta)/p(\theta|A)p(\mathbf{Y}_T|\theta,A)$  is  $p(\mathbf{Y}_T|A)^{-1}$ . It is therefore a candidate for approximation by a posterior simulator. If  $f(\theta)/p(\theta|A)p(\mathbf{Y}_T|\theta,A)$  is bounded above, then the approximation is simulation consistent and the rate of convergence is likely to be practical.

It is not difficult to guarantee the boundedness condition in (4.3.1). Consider first the case in which  $\Theta = \mathfrak{R}^k$ . From the output of the posterior simulator define<sup>1</sup>

$$\hat{\theta}_M = \sum_{m=1}^M w(\theta^{(m)}) \theta^{(m)} / \sum_{m=1}^M w(\theta^{(m)})$$

and

$$\hat{\Sigma}_M = \sum_{m=1}^M w(\theta^{(m)}) (\theta^{(m)} - \hat{\theta}_M) (\theta^{(m)} - \hat{\theta}_M)' / \sum_{m=1}^M w(\theta^{(m)}).$$

(It is not essential that the posterior mean and variance of  $\theta$  exist.) Then for some  $p \in (0,1)$ , define  $\hat{\Theta}_M = \left\{ \theta : (\theta - \hat{\theta}_M)' \hat{\Sigma}_M^{-1} (\theta - \hat{\theta}_M) \leq \chi_{1-p}^2(k) \right\}$  and take

$$(4.3.2) \quad f(\theta) = p^{-1} (2\pi)^{-k/2} |\hat{\Sigma}_M|^{-1/2} \exp \left[ -0.5 (\theta - \hat{\theta}_M)' \hat{\Sigma}_M^{-1} (\theta - \hat{\theta}_M) \right] \chi_{\hat{\Theta}_M}(\theta).$$

If the posterior density is uniformly bounded away from 0 on every compact subset of  $\Theta$ , then the function  $f(\theta)/p(\theta|A)p(\theta|\mathbf{Y}_T, A)$  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. In most situations smaller values of  $p$  will result in better behavior of  $f(\theta)/p(\theta|A)p(\mathbf{Y}_T|\theta,A)$  over the domain  $\hat{\Theta}_M$ , but greater simulation error due to a smaller number of  $\theta^{(m)} \in \hat{\Theta}_M$ ; there is almost no incremental cost in carrying out the computations for several values of  $p$  rather than a single value of  $p$ .

So long as  $\hat{\Theta}_M \subseteq \Theta$ ,  $\int_{\hat{\Theta}_M} f(\theta) d\nu(\theta) = 1$ . If not, the domain of integration must be redefined to be  $\hat{\Theta}_M \cap \Theta$ . In this case a new normalizing constant for  $f(\theta)$  can be well

---

<sup>1</sup>The weighting function  $w(\theta)$  is defined in Theorem 3.2.1 in the case of importance sampling. For MCMC algorithms  $w(\theta) = 1$ .

approximated by taking a sequence of i.i.d. draws  $\{\theta^{(1)}\}$  from the original distribution (4.3.2) with domain  $\hat{\Theta}_M$ , and then averaging  $\chi_{\circ}(\theta^{(1)})$ .

Frequently the behavior of  $f(\theta)/p(\theta|A)p(\mathbf{Y}_T|\theta,A)$  can be improved by reparameterization of  $\theta$  to  $\zeta = h(\theta)$ , where  $h$  is a one-to-one function. Of course, the prior density must then be adjusted by the Jacobian of transformation. If the support  $Z$  of  $p(\zeta|A)$  is  $\mathfrak{R}^k$ , then  $\int_Z f(\zeta)d\nu(\zeta) = 1$  for  $f$  constructed as indicated in (4.3.2). For example, if this method is used to approximate the marginal likelihood in the standard linear model (2.1.7)-(2.1.9), transformation of  $h$  to  $\log(h)$  guarantees the support condition, and generally results in more accurate approximation of  $p(\mathbf{Y}_T|A)$ .

The numerical accuracy of the approximation can be evaluated using the methods of Section 3.8, as detailed below in Section 4.5.

#### 4.4 Improving numerical approximations

In many instances a portion of the marginal likelihood  $P(\mathbf{Y}_T|A) = \int_{\Theta} p(\theta|A)p(\mathbf{Y}_T|\theta,A)d\nu(\theta)$  may be evaluated analytically. Suppose

$$\int_{\Theta} p(\theta|A)p(\mathbf{Y}_T|\theta,A)d\nu(\theta) = \int_{\Theta_1} \int_{\Theta_2} p(\theta_1, \theta_2|A)p(\mathbf{Y}_T|\theta_1, \theta_2, A)d\nu(\theta_2)d\nu(\theta_1) = \int_{\Theta_1} r(\theta_1)d\nu(\theta_1)$$

where  $r(\theta_1) = \int_{\Theta_2} p(\theta_1, \theta_2|A)p(\mathbf{Y}_T|\theta_1, \theta_2, A)d\nu(\theta_2)$  can be evaluated analytically. Then the modified harmonic mean method can be applied directly to the simulated values  $\theta_1^{(m)}$ , using  $r(\theta_1^{(m)})$  in lieu of  $p(\theta^{(m)}|A)p(\mathbf{Y}_T|\theta^{(m)}, A)$  and tailoring  $f(\theta)$  to  $r(\theta)$  rather than to  $p(\theta|A)p(\mathbf{Y}_T|\theta, A)$ . Similar adjustments can be made for importance sampling. Because the dimension of integration is lower, the resulting approximation will typically be more accurate. In the case of the method employing the Gibbs sampler described in Section 4.2 this preliminary evaluation will eliminate at least one of the blocks for which the auxiliary simulations must be undertaken.

An example of this procedure is provided by earlier results for the standard linear model. The entire posterior kernel in standard form is (2.1.13). But in Section 2.3 we found the marginal likelihood conditional on  $h$ , (2.3.6). The latter expression is a function of a single unknown parameter, whereas the former is a function of  $k+1$  unknown parameters.

The probit model described in Section 2.4 provides a second example. In this case there are  $T+k$  unknown parameters ( $T$  latent variables and  $k$  coefficients). The modified harmonic mean method in this case is completely unwieldy, since it would require storing a very large amount of posterior simulator output, and generation of the requisite  $T+k$  truncated normal random variables would require the factorization of a matrix of the same

order. In this case, integration of the  $T$  latent variables is straightforward, and leads to the product of the prior density for the coefficients and the likelihood function as typically written,

$$(2\pi)^{-k/2} |\underline{\mathbf{H}}_{\beta}|^{1/2} \exp\left[-.5(\beta - \underline{\beta})' \underline{\mathbf{H}}_{\beta}(\beta - \underline{\beta})\right] \prod_{t=1}^T \{d_t \Phi(\beta' \mathbf{x}_t) + (1 - d_t)[1 - \Phi(\beta' \mathbf{x}_t)]\}.$$

More generally, in models with latent variables accurate evaluation of the marginal likelihood requires that it be possible to perform the integration over the space of latent variables analytically.

#### 4.5 Software

The program **mlike**, available in six languages at [www.econ.umn.edu/~bacc](http://www.econ.umn.edu/~bacc), provides approximations of the log marginal likelihood using the modified harmonic mean posterior simulation method, given a posterior simulator file. The program renormalizes the density  $f(\theta)$  if the condition  $\hat{\Theta}_M \subseteq \Theta$  is violated, as described in Section 4.3. The program uses the values  $p = 0.9, 0.8, \kappa, 0.1$  in (4.3.2). For each of these, it computes

$$\left[\sum_{m=1}^M w(\theta^{(m)})\right]^{-1} \left[\sum_{m=1}^M w(\theta^{(m)}) f(\theta^{(m)}) / p(\theta^{(m)}|A) p(\mathbf{Y}_T|\theta^{(m)}, A)\right]$$

and then it reports minus the logarithm of this value. However, there are two features of **mlike** that are specific to the model: first, the position of model parameters within the posterior simulator file must be communicated to **mlike**; second, any reparameterization of the model from  $\theta$  to  $\zeta = h(\theta)$  must also be communicated to **mlike**. This is accomplished through three auxiliary procedures.

The procedure `repar0` sets any parameters required by `repar`, needed to organize the parameter vector. For example, in a multivariate regression model the number of equations and number of covariates is not evident from the number of columns in the posterior simulator matrix. In this case procedure `repar0` sets up the requisite pointers to indicate which columns are the coefficients and which are the elements of the disturbance variance matrix.

The procedure `repar` accomplishes the reparameterization. It maps the parameters of the posterior simulator file into the transformed parameters for **mlike**, and modifies the prior density by the appropriate Jacobian of transformation. For example, in the standard linear model the disturbance precision is replaced by its logarithm. The Jacobian of transformation is evaluated, and then the prior density is modified by this value.

The procedure `lrange` indicates whether or not a given parameter vector is within the support of the prior distribution. If  $\Theta_A = \mathcal{R}^k$  this is always the case, and `lrange` communicates this through a logical variable `lall` with value `true` and **mlike** then does not undertake the simulations to appropriately adjust the normalization constant of

$f(\theta)$ . If `lall` is set to `false`, then this procedure must determine whether the parameter vector is within the support of the prior, and then communicate `lrange=true` if it is and `lrange=false` if it is not.

The **mlike** output file provides no direct information on the accuracy of the numerical approximation of the log marginal likelihood. (There is some indirect information provided, by looking at the differences in the nine alternative computations of the log marginal likelihood provided.) To find the numerical standard error of the approximation it is necessary to create an **mlike** posterior simulator file. The posterior simulator file created by **mlike** contains a pair of records for each simulation used in the approximation of the log marginal likelihood. The first record in each pair specifies the iteration number, the log of the weighting function, and two dummy entries each zero (to make the structure match that of all posterior simulator files). The second record in each pair has nine entries, corresponding to the values of  $f(\theta^{(m)})/p(\theta^{(m)}|A)p(\mathbf{Y}_T|\theta^{(m)},A)$  for each of the nine values of  $p$  in (4.3.2). These values will have been normalized by the constant given in the **mlike** output file, to prevent exponent overflow or underflow. The **mlike** posterior simulator file, used as input to `moment`, will then provide the posterior means and numerical standard errors of the normalized  $f(\theta)/p(\theta|A)p(\mathbf{Y}_T|\theta,A)$ . The numerical standard error of the corresponding log marginal likelihood is this numerical standard error divided by the posterior mean, in this application of `moment` to the **mlike** posterior simulator file.

## 4.6 Examples

For the regression model described in Section 3.9, the marginal likelihood was approximated using **mlike**, which is available with **uvr1** at the website. The software incorporates a reparameterization of the precision from  $h$  to  $\log h$ . After this reparameterization  $\Theta_A = \mathfrak{R}^k$ , and for this case **mlike** execution time is roughly proportional to the number of records in the posterior simulator file, about 8 seconds in the examples discussed here. Computation of numerical standard error with `moment` takes another 4 seconds.

The top panel of Table 4.1 provides results using  $p = .9$ ,  $.5$  and  $.1$  in expression (4.3.2). Computation with  $p = .9$  provides the most accurate assessment. In view of the good approximation of the posterior by a multivariate normal distribution, it is not surprising that a more inclusive  $f(\theta)$  yields more accurate results. Differences in approximations for different values of  $p$  are consistent with the numerical standard errors (NSEs).

To illustrate the use of the approximated marginal likelihood in the construction of Bayes factors two variants on the model set forth in Section 3.9 were constructed by making two changes in the prior distribution. In the first change the mean of the prior distributions of all slope coefficients is shifted to the value of the standard deviation which in turn is unchanged: that is, the prior distribution is changed from  $N(0, .1^2)$  to  $N(.1, .1^2)$  for all covariates except  $\log(\text{Lot size})$ , for which the prior distribution is shifted from  $N(0, .3^2)$  to  $N(.3, .3^2)$ . Log marginal likelihood approximations for this model are given in the middle panel of Table 4.1. Finally, the standard deviations in these priors are reduced by half: now all priors are  $N(.1, .05^2)$  except for  $\log(\text{Lot size})$  which is  $N(.3, .15^2)$ . Log marginal likelihoods for this model are shown in the lower panel of Table 4.1.

Using the approximation based on  $p = .9$ , the log Bayes factor in favor of the last model, versus the first, is approximately 10.285 and the associated NSE is .005. Thus the Bayes factor is almost certainly (based on  $\text{NSE} \times 3$ ) between 28,853 and 29,733.

In the probit model example, for both the Gibbs sampler and the Hastings-Metropolis algorithm the marginal likelihood can be approximated using the modified harmonic mean method implemented in `mlike`. In the case of the Gibbs sampler the evaluation of the likelihood function for the probit model discussed at the end of Section 4.4 is used. These results are shown in the top two panels of Table 4.2. For the same reasons as in the regression model the approximation is quite accurate and is better for larger values of  $p$ . For the probit model Hastings-Metropolis algorithm the marginal likelihood can also be approximated using Theorem 4.1.2. This approximation, given in the last line of Table 4.2, is consistent with the other assessments (as measured by NSE) and is as accurate as the most accurate of the harmonic mean approximations.

## 5. Model development

In the preceding sections it has been assumed that a collection of complete models  $A_1, \dots, A_J$  is available, each model specifying a parametric data density  $p(\mathbf{Y}_T | \theta_j, A_j)$ , a prior distribution for parameters  $p(\theta_j | A_j)$ , and a conditional distribution of a vector of substantive variables of interest,  $p(\omega | \mathbf{Y}_T, \theta_j, A_j)$ . In addition there is a probability  $P(A_j)$  associated with each model, and  $\sum_{j=1}^J P(A_j) = 1$ . The specification of  $p(\mathbf{Y}_T | \theta_j, A_j)$  and  $p(\omega | \mathbf{Y}_T, \theta_j, A_j)$  are familiar tasks to economists. This section takes up some ways that simulation methods can assist in what may be less familiar, and are often less formal, aspects of model development: expression of prior distributions for parameters, and specification of a set of models  $A_1, \dots, A_J$  adequate to the task at hand.

### 5.1 Prior elicitation and specification

Any complete model  $A$  implies a prior, or predictive, distribution

$$p(\omega | A) = \int_{\Theta} \int_{\Psi_T} p(\omega | \mathbf{Y}_T, \theta, A) p(\mathbf{Y}_T | \theta, A) d\nu(\mathbf{Y}_T) p(\theta | A) d\nu(\theta).$$

Generally it will not be possible to access  $p(\omega | A)$  analytically. On the other hand, i.i.d sampling from  $p(\theta | A)$ ,  $p(\mathbf{Y}_T | \theta, A)$  and  $p(\omega | \mathbf{Y}_T, \theta, A)$  will generally be straightforward. These tasks may be trivial. For example, in the probit model taken up initially at the end of Section 2.4, suppose that one is interested in the effect of a change in some covariate on the probability of the outcome  $d_i = 1$ . Given the complete probit model specification in Section 2.4, sampling from the prior density  $p(\beta | A)$  entails drawing from a multivariate normal distribution; sampling from  $p(\mathbf{Y}_T | \beta, A)$  amounts to drawing the latent variables  $\mathbf{y}^*$  defined in (2.4.11) from univariate normal distributions, followed by mapping  $y_i^* \geq 0$  into  $d_i = 1$  and  $y_i^* < 0$  into  $d_i = 0$ ; and the vector  $\omega = P(d_i = 1 | \mathbf{x}_i, \beta, A)$  can be computed directly. On the other hand, these tasks need not be trivial. For example, simulating  $\mathbf{Y}_T | (\theta, A)$  may require solution of a model that cannot be carried out in closed form<sup>1</sup>, and simulating  $\omega | (\mathbf{Y}_T, \theta, A)$  may demand ingenious forecasting algorithms. In Bajari (1997),  $\mathbf{Y}_T$  includes bids submitted under conditions of asymmetric information, draws from  $p(\mathbf{Y}_T | \theta, A)$  involve solution of a system of nonlinear differential equations, and the vector  $\omega$  includes revenue realized by an auctioneer. But these sorts of exercises are routinely carried out by economists. In general, model simulation is much simpler than posterior simulation.

While a complete model demands  $p(\theta | A)$ , it is often difficult to elicit (“think about”) a prior distribution of a parameter vector  $\theta$  directly. But unless this task is taken seriously,

---

<sup>1</sup>For methods and extensive references see Amman, Kendrick and Rust (1996).

the claim to an exact evaluation of  $E[h(\omega)|\mathbf{Y}_T, A]$  is not secure. In the comparison or averaging of models, careless development of  $p(\theta|A)$  will more often than not lead to posterior odds ratios that reflect the relative plausibility of two arbitrary prior distributions in different models. The outcome may simply convey the information that some models have absurd prior distributions and others do not, and not the relative plausibility of the models with more carefully considered prior distributions of parameters.

It is typically easier to elicit prior distributions about  $\omega$  than about  $\theta$ . For example, in an earnings model involving high order polynomials in age and education it is natural to consider reasonable ranges for earnings ratios at different age and education levels, and nearly impossible to think about individual coefficients of the polynomial (Geweke and Keane, 1997). Moreover, formulation of prior distributions over the substantive, model invariant elements of  $\omega$  provides considerable discipline in developing prior distributions  $p(\theta_j|A_j)$  that are at least reasonably consistent across models.

In some cases it may be possible to obtain  $p(\theta|A)$  analytically from  $p(\omega|A)$ . In general, however, the relationship between  $\theta$  and  $\omega$  will be sufficiently complicated that this is precluded. But if simulation from  $p(\omega|A)$  is cheap, then  $p(\theta|A)$  that approximates prior beliefs about  $\omega$  may be obtained through trial and error. This process may reveal that for some functions  $h(\omega)$ ,  $p[h(\omega)|A]$  cannot be well approximated by any choice of  $p(\theta|A)$ . This indicates that the data distribution  $p(\mathbf{Y}_T|\theta, A)$  is incapable of expressing  $p(\omega)$ , and in this case the model  $A$  should be discarded a priori from consideration. If no model  $A_j$  conveys  $p(\omega|A_j)$  approximating prior beliefs then it is necessary to develop other models. Of course no formal procedure will indicate what such a model will entail, but results obtained for  $p(\omega|A_j)$  over the models  $A_1, \dots, A_J$  may provide nutritious food for thought.

Comparison of  $h(\omega)|A$  with  $h(\omega)|(\mathbf{Y}_T, A)$  can reveal important ways in which the data  $\mathbf{Y}_T$  change prior beliefs about  $h(\omega)$ . At one extreme the function  $h(\omega)$  may not be identified by the data, in which case the prior and posterior density functions are equivalent: i.e.,  $p[h(\omega)|\mathbf{Y}_T, A] = p[h(\omega)|A]$  for some  $\mathbf{Y}_T \in \Psi_T$ .<sup>2</sup> For these  $\mathbf{Y}_T \in \Psi_T$ , the data do not change prior beliefs about  $h(\omega)$  at all because of weakness in the data with regard to  $h(\omega)$ . A classic example is the standard linear model (2.1.7)-(2.1.9) with  $\mathbf{X}\mathbf{c}=\mathbf{0}$  for some vector  $\mathbf{c}$ , and  $h(\omega) = \mathbf{c}'\beta$ . An overplot of  $p[h(\omega)|\mathbf{Y}_T, A]$  and  $p[h(\omega)|A]$ , or of  $P[h(\omega)|\mathbf{Y}_T, A]$  and  $P[h(\omega)|A]$ , will exhibit curves that differ only by simulation noise, and a plot of  $P^{-1}\{P[h(\omega)|\mathbf{Y}_T, A]|A\}$  will differ from a 45-degree line only by simulation noise.

---

<sup>2</sup>This definition coincides with the classical treatment of identification (e.g. Poirier, 1995, p. 256). An alternative, weaker definition of identification is that the posterior distribution of  $h(\omega)$  exist (Richard, 1970, pp. 3-9). In a complete model  $h(\omega)$  is always identified under the weaker definition.

At another extreme, no set of data can change prior beliefs about  $h(\omega)$  because prior beliefs about  $h(\omega)$  are dogmatic. A classic example is the population first order autocorrelation of the disturbances  $\varepsilon_t$  in the standard linear model. Since the disturbances are dogmatically i.i.d.,  $E(\varepsilon_t \varepsilon_{t-1} | \beta, h, \mathbf{X}, A) / \text{var}(\varepsilon_t | \beta, h, \mathbf{X}, A) \equiv 0$ . Overplotting of  $p[h(\omega) | \mathbf{Y}_T, A]$  and  $p[h(\omega) | A]$ , or of  $P[h(\omega) | \mathbf{Y}_T, A]$  and  $P[h(\omega) | A]$ , will show vertical lines at 0. In this situation the sample counterpart will be even more revealing: let  $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$ ,  $u_t = y_t - \mathbf{b}'\mathbf{x}_t$  and  $h(\omega) = \sum_{t=2}^T u_t u_{t-1} / \sum_{t=1}^T u_t^2$ . Then the prior distribution of  $h(\omega)$  will be concentrated near 0. (Informally, the latter situation suggests that another model might be preferred to the standard linear model, and Box (1980) suggests this approach. We return to further consideration of this possibility below.)

Intermediate cases include those in which the data contribute strongly to knowledge of  $h(\omega)$  in a manner consistent with the model:  $p[h(\omega) | \mathbf{Y}_T, A]$  is more concentrated than  $p[h(\omega) | A]$  and is well within the support of  $p[h(\omega) | A]$ . For example, a prior distribution for the standard linear model (2.1.7) might specify  $p(\beta_j) = \chi_{[0,1]}(\beta_j)$ , and the posterior distribution of  $\beta_j$  is concentrated almost entirely between .75 and .76. A second intermediate case is one in which the data contribute strongly to knowledge of  $h(\omega)$ , but  $p[h(\omega) | \mathbf{Y}_T, A]$  is not well within the support of  $p[h(\omega) | A]$ . In the context of the previous example, the posterior distribution of  $\beta_j$  might be nearly collapsed about the left side of  $\beta_j = 1$ , and in this case the prior and posterior distributions of  $h(\omega) = b_j$  will differ markedly.

## 5.2 Incomplete models and partial information inference

Model development is costly. A new complete model can easily require years to create; millions of dollars and careers may be devoted to the effort. The process of scientific investigation entails working with a limited array of complete models and incompletely formed ideas about alternative models, developing the latter only when there is substantial evidence that they might be preferred to the former.<sup>3</sup>

Examining the ways in which a model represents observed data poorly is a standard part of sound scientific practice, often going under the name diagnostic checking in the statistics literature or misspecification testing in econometrics. There are divergent but well established approaches to this task. At one extreme, classical pure significance testing identifies poor representation with a function of the data that is greater or smaller than what might have occurred. At the other, formal Bayesian model comparison requires the

---

<sup>3</sup>On a grand scale, this familiar process is spelled out in the classic work of Kuhn (1970). This is an intentionally Bayesian statement of part of Kuhn's thesis.

formulation of all models that are plausible. In practice it would be too costly even if conceptually possible to formulate the complete set of plausible models, but on the other hand there are at least vague ideas of what other models might be and this affects the choice of the data function in pure significance testing. A thorough and still timely discussion of these issues is Box (1980) and the accompanying discussion. A portion of Box's article argues that non-Bayesian methods are required for diagnostic checking of a set of models. Some of the discussants, including Barnard, Bernardo, and Dawid argue for a Bayesian interpretation of Box's argument. The procedures set forth here may be viewed as an explicit implementation of Barnard's ideas.

To formalize the notion of incompletely formed ideas about other models, return to the environment described in Section 2.1. Let  $f(\mathbf{Y}_T): \mathfrak{R}^{pT} \rightarrow \mathfrak{R}^q$  be a function of the observable data, where  $q$  is a small integer (often  $q = 1$ ). An *incomplete model*  $\tilde{A}$  is a specification  $p[f(\mathbf{Y}_T)|\tilde{A}]$ . It is incomplete because it does not state the predictive distribution  $p(\mathbf{Y}_T|\tilde{A})$  and because it takes no stand on the definition or distribution of any vector of substantive variables  $\omega$ . The model  $\tilde{A}$  is simply a formal statement of what a certain aspect of the observable data might look like, given models not as yet articulated.

A complete model  $A$  and an incomplete model  $\tilde{A}$  can be compared through their common prediction of the observable  $f(\mathbf{Y}_T)$ :

$$p[A|f(\mathbf{Y}_T)] = p(A) \int_{\Theta} p(\theta|A) p[f(\mathbf{Y}_T)|\theta, A] d\nu(\theta) / p[f(\mathbf{Y}_T)],$$

$$p[\tilde{A}|f(\mathbf{Y}_T)] = p(\tilde{A}) p[f(\mathbf{Y}_T)|\tilde{A}] / p[f(\mathbf{Y}_T)].$$

The *partial information Bayes factor* in favor of  $A$  versus  $\tilde{A}$  is thus

$$\frac{p[f(\mathbf{Y}_T)|A]}{p[f(\mathbf{Y}_T)|\tilde{A}]} = \frac{\int_{\Theta} p(\theta|A) p[f(\mathbf{Y}_T)|\theta, A] d\nu(\theta)}{p[f(\mathbf{Y}_T)|\tilde{A}]}.$$

In this definition "partial information" refers to the fact that conditioning is on  $f(\mathbf{Y}_T)$  rather than on all of the observed data  $\mathbf{Y}_T$ . Since  $\tilde{A}$  only predicts  $f(\mathbf{Y}_T)$ , this is as it must be: models can be compared only on the basis of their common predictions. Correspondingly,  $p[f(\mathbf{Y}_T)|A]$  is the *partial information marginal likelihood* (PIML) of the complete model and  $p[f(\mathbf{Y}_T)|\tilde{A}]$  is the partial information marginal likelihood of the incomplete model.

Typically  $p[f(\mathbf{Y}_T)|\tilde{A}]$  is specified directly and its evaluation is no problem. Evaluation of  $p[f(\mathbf{Y}_T)|A]$ , if undertaken along the lines described in Sections 3 and 4, could be arduous. With rare exception a new posterior simulator for the parameter vector  $\theta$  would be required, and the marginal likelihood would then have to be evaluated. On the other hand it is straightforward to make multiple, i.i.d. drawings  $\theta^{(m)}$  from the prior distribution

$p(\theta|A)$ , draw  $\mathbf{Y}_T^{(m)} \sim p(\mathbf{Y}_T|\theta^{(m)}, A)$ , and form  $f(\mathbf{Y}_T^{(m)})$ . If  $q$  is small, and in particular if  $q=1$  so that  $f(\mathbf{Y}_T)$  is a scalar,  $p[f(\mathbf{Y}_T)|A]$  can be approximated by standard kernel smoothing methods from  $\{f(\mathbf{Y}_T^{(m)})\}_{m=1}^M$ . The relative ease of this procedure has significant implications for the conduct of research. One can construct partial information Bayes factors comparing complete and incomplete models before (a) developing posterior simulators of other procedures for formal Bayesian inference in the complete models or (b) further developing the incomplete models into complete models. Thus both the conceptual effort of fully articulating complete models, and the technical work of formal Bayesian inference, can be concentrated on those models that will ultimately have nonnegligible posterior probability.<sup>4</sup>

Some familiar examples in the standard linear model (2.1.7)-(2.1.9) will illustrate this approach. In any given situation this model alone hardly constitutes a reasonably representative set of models  $A_1, K, A_j$ . Other models might, for example, replace the normal distributional assumption for the disturbance term with a distribution having different third or fourth moments. Consider the functions

$$f^{(1)}(\mathbf{Y}_T) = T^{1/2} \sum_{t=1}^T u_t^3 / \left[ \sum_{t=1}^T u_t^2 \right]^{3/2}, \quad f^{(2)}(\mathbf{Y}_T) = T \sum_{t=1}^T u_t^4 / \left[ \sum_{t=1}^T u_t^2 \right]^2 - 3,$$

the skewness and excess kurtosis based on the ordinary least squares residuals. Given  $\beta^{(m)} \stackrel{iid}{\sim} N(\underline{\beta}, \mathbf{H}^{-1})$ , let

$$\mathbf{y}^{(m)} \sim N\left[\mathbf{X}\beta^{(m)}, (h^{(m)})^{-1} \mathbf{I}_T\right], \quad \mathbf{b}^{(m)} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}^{(m)}, \quad \mathbf{u}^{(m)} = \mathbf{y}^{(m)} - \mathbf{X}\mathbf{b}^{(m)}.$$

Then the partial information marginal likelihood  $p[f^{(1)}(\mathbf{Y}_T)|A]$  may be approximated by a kernel density estimate applied to  $\left\{ T^{1/2} \sum_{t=1}^T (u_t^{(m)})^3 / \left[ \sum_{t=1}^T (u_t^{(m)})^2 \right]^{3/2} \right\}_{m=1}^M$ , evaluated at the point  $T^{1/2} \sum_{t=1}^T (u_t)^3 / \left[ \sum_{t=1}^T (u_t)^2 \right]^{3/2}$ , where  $\mathbf{u} = \mathbf{y} - \mathbf{X}\mathbf{b}$ . If the skewness coefficient is about .5 and there are several hundred observations then  $p[f^{(1)}(\mathbf{Y}_T)|A] < 10^{-12}$ . Supposing  $p[f^{(1)}(\mathbf{Y}_T)|\tilde{A}] \sim N(0, \tau^2)$ , then for  $\tau$  in the range of (say) .2 to  $10^4$ , the partial information Bayes factor against the standard linear model is over 1000. Similar calculations may be made for  $f^{(2)}(\mathbf{Y}_T)$ .

The technical steps involved in computing a partial information marginal likelihood are superficially similar to bootstrapping a sampling theoretic test statistic. First, find a

---

<sup>4</sup>The emphasis here is on what can be done before construction and execution of a posterior simulator, as well as before the completion of other models. Given a set of complete models, one of which is nested by all the others, one may save considerable time by constructing and executing a posterior simulator for the nested model and using score function (Lagrange multiplier) approximations to Bayes factors; on this approach see Poirier (1988b, 1996).

function  $f(\mathbf{Y}_T)$  for which the predictive distribution under the complete and incomplete models are not the same. Then, use simulation methods to evaluate the partial information marginal likelihood in the complete model. The first step is superficially similar to finding a test statistic with good power properties, the second to bootstrapping a critical value. It should be clear, however, that the procedure described here conditions on the observed data, the known properties of the (as yet) incomplete model, and the prediction held in common by the complete and incomplete models. This procedure is consistent with the likelihood principle and is entirely Bayesian given the assumptions about the information at hand.

In the example given there are many functions  $f(\mathbf{Y}_T)$  that could be considered, and this will generally be the case. The usual non-Bayesian list of alternative hypotheses and corresponding test statistics is a rich group of candidates. If the incomplete models specified the joint distribution of several such functions  $f(\mathbf{Y}_T)$ , then the partial information Bayes factor could be modified accordingly with the appropriate multivariate  $f(\mathbf{Y}_T)$ . But there are several reasons why this modification is not likely to be worth pursuing. First, specification of joint predictive distributions of the  $f(\mathbf{Y}_T)$  moves one rapidly toward the specification of a complete model, if for no other reason than to maintain logical consistency. The procedures of Sections 2.3 and 4 then apply. Second, if kernel smoothing methods are to be employed in the evaluation of  $p[f(\mathbf{Y}_T)|A]$  then the number of draws  $\mathbf{Y}_T^{(m)}|A$  increases exponentially in the dimension of  $f(\mathbf{Y}_T)$  for reliable evaluation. This is strictly a technical problem, but it is serious. Third, evaluation of  $p[f(\mathbf{Y}_T)|A]$  separately for specific unidimensional  $f(\mathbf{Y}_T)$  is likely to provide informal as well as formal guidance in the elaboration of incomplete into complete models.

### 5.3 Examples

To provide an illustration of how the comparison of a complete model with incomplete models might work in practice, return again to the hedonic price regression example introduced in Section 3.9. Several functions of the least squares residuals  $\mathbf{u}$  and the explanatory variables  $\mathbf{X}$ ,  $f(\mathbf{u}, \mathbf{X})$ , were evaluated. Next, 1,000 draws  $\beta^{(m)} \sim p(\beta)$  and  $h^{(m)} \sim p(h)$ , were made, each followed by  $\mathbf{y}^{(m)} \sim N(\mathbf{X}\beta^{(m)}, h^{(m)-1}\mathbf{I}_T)$ , computation of  $\mathbf{u}^{(m)} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^{(m)}$  and evaluation of  $f(\mathbf{u}^{(m)}, \mathbf{X})$ . (This required about 20 seconds.) Finally, standard kernel smoothing methods were used to approximate the predictive density at the value of  $f(\mathbf{u}, \mathbf{X})$  computed using the data, which is the partial information marginal likelihood (PIML) of the complete model.

The results of this exercise are displayed in Table 5.1. For each function  $f(\mathbf{u}, \mathbf{X})$  the approximate .05 and .95 quantiles of the predictive distribution are given, followed by the data value, and finally by the approximate predictive density evaluated at the data value. The function “skewness + kurtosis” is the sum of the squared skewness coefficient and one-fourth of the squared kurtosis coefficient.<sup>5</sup> The “nonlinear regression” functions are the simple correlation of the least squares residuals  $\mathbf{u}$  and the squared values of the indicated regressors. The “conditional heteroscedasticity” functions are the simple correlation coefficients of the squared least squares residuals and the square of the indicated explanatory variables.<sup>6</sup>

To make sense of the PIMLs for the regression model (the complete model) reported in the last column of Table 5.1, it is necessary to think about what that value might be under other models not yet formulated (incomplete models). In the case of skewness, one might proceed as follows, using as a reference the chi-squared distribution which is skewed and has a shape familiar to most econometricians. Suppose that the predictive distribution for the absolute value of skewness conditional on the set of incomplete models corresponds to that in the family of chi-squared distributions with a uniform prior on degrees of freedom in the interval (2, 8). Taking the distribution of skewness to be symmetric about zero, standard calculations show that the implied density at a skewness of -.184 is then about 2.0, or roughly twice that for the regression model. Similarly if one contemplates a predictive distribution for excess kurtosis equivalent to that in the Student- $t$  distribution with a uniform prior on degrees of freedom in the interval (4,10), then the implied density at .517 is about .85, almost five times greater than the predictive density of .177 under the standard linear model. In these cases, the partial information Bayes factor against the linear model is approximately 2 and 5, respectively.

Similar methods may be used to interpret other values in Table 5.1. For example, one could think about other models in which the precision of the disturbances depends on the dichotomous air conditioning variable. Let the ratio of precisions with and without central airconditioning be  $c$  and suppose  $\log(c) \sim N(0, [\log(1.5)]^2)$ . Using the fact that 31.6% of the sample has central air conditioning, the correlation coefficient -.075 corresponds to a value  $c = 1.25$ . Standard methods show that the implied marginal likelihood of the

---

<sup>5</sup>This function is motivated by the classical test against normality developed in Kiefer and Salmon (1983).

<sup>6</sup>The table demonstrates the difference between what is proposed in Section 5.2 and the “predictive distribution of checking functions” advanced by Box (1980). Box compares the function of the observed data with the distribution of that function conditional on the model  $A$ , and concludes against the model if the function of the observed data lies in the tails of the predictive distribution. This can be done with the information in Table 5.1, if by “tails” one means 5%.

correlation coefficient  $-.075$ , under the presumed predictive distribution, is  $4.66$ , implying a partial information Bayes factor of almost  $6$  against the linear model.

Comparison of the probit model with alternative incomplete models can proceed in similar fashion. In this example a natural set of predictive statistics is based on the maximum likelihood estimate  $\hat{\beta}$  of  $\beta$ , the predictive probabilities  $\hat{p}_t = \Phi(\hat{\beta}'\mathbf{x}_t)$  associated with this estimate, and the residuals  $d_t - \Phi(\hat{\beta}'\mathbf{x}_t)$ . These statistics were formed from the data set. To approximate the predictive intervals of the complete probit model,  $\beta$  was drawn from the prior distribution, corresponding choices were formed corresponding to each  $\mathbf{x}_t$ , the maximum likelihood estimate was computed, and the predictive statistics were formed. (For 1000 replications this required about 70 minutes.) Standard kernel smoothing methods were again used to approximate the PIMLs.

Evaluation of PIMLs and associated quantities are given in Table 5.2. “Fraction choosing” is the fraction of the sample that participates in the labor force. Since the data p.d.f. of the probit model can generate any such fraction, evaluating the PIML for this function amounts to a check that the prior distribution of  $\beta$  does not dogmatically declare all women in the sample to be labor force (non)participants. The value of the PIML,  $.279$ , reflects the fact that the event that all women would (not) participate is reasonable under the prior, but so are all rates of participation between  $0$  and  $1$ .

The “fraction in” functions measure the fraction of women participating in the labor force, for whom the predictive probability based on the MLE,  $\hat{p}_t$ , was between  $p_1$  and  $p_2$ . Ten combinations of  $p_1$  and  $p_2$  are chosen in Table 5.2. While the notion that these functions might be reasonable indications of actual participation probability motivates the construction of these statistics, its (in)adequacy as a predictor is irrelevant to the evaluation of the PIML. What matters is the predictive density for the actual proportion of women participating for whom  $\hat{p}_t$  is between  $p_1$  and  $p_2$ , evaluated at the proportion observed in the sample; the support of this predictive density is  $[0, 1]$  and not  $[p_1, p_2]$ . This is nicely illustrated for the first case,  $p_1 = 0$  and  $p_2 = .1$ . The predictive distribution for the fraction participating is concentrated at the low end of this interval, reflecting the fact that for most women for whom  $\hat{p}_t < .1$ ,  $\hat{p}_t$  is in fact very low. But in the data the observed fraction is  $.222$ : there were nine women for whom  $\hat{p}_t < .1$ , and two of these were in fact labor force participants. This outcome is nearly impossible in the complete probit model set forth in Section 3.2. For the other combinations of  $p_1$  and  $p_2$  the PIML is not substantially lower than what we would expect under alternative, incomplete models with the possible exception of  $p_1 = .4$ ,  $p_2 = .5$ .

The correlation of the squared residuals,  $(d_i - \hat{p}_i)^2$ , with the squared covariates,  $x_{ii}^2$ , is a means of comparing this complete probit model with alternative incomplete models for which the functional relation between the covariate and labor force participation probability is different. The results for education, number of children and work experience suggest that elaboration of the probit model allowing for nonlinearity or conditional heteroscedasticity in those variables might be worth pursuing.

## 6. Bayesian communication

For a subjective Bayesian decision maker the computation of the posterior moments  $E[h(\omega)|\mathbf{Y}_T, A]$  for suitable models, priors and functions of interest is typically the final objective of inference. For an investigator reporting results for other potential decision makers, however, the situation is different. In the language of Hildreth (1963) these decision makers are *remote clients*, whose priors and functions of interest are not known to the investigator.

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

### 6.1 Posterior reweighting

An investigator will have carried through formal inference for a set of models  $A_{1,K}, A_J$ . This collection will reflect the process of model development, and a public report of the investigator's work should at least summarize this process. In the ideal situation described by Poirier (1988a), clients have agreed to disagree in terms of the prior. Since the set of models that exist in any meaningful sense is the set publicly reported, collectively investigators will have provided the grand model in which variation of the prior is the basis of formal discourse in normal science.<sup>1</sup>

Corresponding to each model  $A_{1,K}, A_J$  included in an investigation, there is a posterior simulator file of the form described in Section 3.8. It is a simple matter to make these files available at an ftp or web site, and for any client to obtain them for the purpose of the manipulations described here.

Given the posterior simulator file a client can immediately compute numerical approximations to posterior moments not reported or even considered by the investigator. Specifically, suppose a client wishes to know  $E[h(\omega)|\mathbf{Y}_T, A]$  where  $\omega \sim p(\omega|\mathbf{Y}_T, A)$  is specified by the client and  $p(\mathbf{Y}_T|\theta, A)$  and  $p(\theta|A)$  have been specified by the investigator. Corresponding to each  $\theta^{(m)}$  reported in the posterior simulator file the client forms  $g(\mathbf{Y}_T, \theta^{(m)})$  with the property  $E[g(\mathbf{Y}_T, \theta)|\mathbf{Y}_T, \theta, A] = E[h(\omega)|\mathbf{Y}_T, \theta, A]$ , and then computes

---

<sup>1</sup>The term normal science is used here as in Kuhn (1970). In this framework revolutionary science may be interpreted as the search for new models in the light of limited information Bayes factors in favor of incomplete models (see Section 5.2).

$\bar{g}_M = \sum_{m=1}^M w(\theta^{(m)})g(\theta^{(m)}) / \sum_{m=1}^M w(\theta^{(m)})$ . If the investigator's posterior simulator is ergodic then  $\bar{g}_M \xrightarrow{a.s.} \bar{g} = E[h(\omega)|\mathbf{Y}_T, A]$  and if it is uniformly ergodic then  $M^{1/2}(\bar{g}_M - \bar{g}) \xrightarrow{d} N(0, \sigma^2)$ . For simple functions  $h(\omega)$  computation of  $\bar{g}_M$  amounts to spreadsheet arithmetic. More elaborate functions of interest may involve simulations, but in all cases these computations are precisely those which economists undertake as a routine matter when investigating the implications of a model.

For example, a client reading a research report might be skeptical that the investigator's model, prior and data set provide much information about the effects of an interesting change in a policy variable on the outcome in question. If the simulator output matrix is available electronically the client can obtain the exact (up to numerical approximation error, which can also be evaluated) answer to his query without arising from his office chair in considerably less time than required to read the research report.

The social contribution of the investigator in this context is clear. She enables clients to incorporate the effects of uncertainty about parameters in a specified model consisting of  $p(\mathbf{Y}_T|\theta, A)$  and  $p(\theta|A)$ , in reaching conclusions or decisions of the client's choosing, that can be addressed by the model. This contribution extends in an obvious way to uncertainty about models, so long as a posterior simulator matrix has been provided by an investigator for each model considered.

With a small amount of additional effort the client can modify many of the investigator's assumptions. Suppose the client wishes to evaluate  $E[h(\omega)|\mathbf{Y}_T, A^*]$ , where the model  $A^*$  differs from the model  $A$  only in the specification of the prior distribution  $p(\theta|A^*) \neq p(\theta|A)$ ; that is,  $p(\mathbf{Y}_T|\theta, A) = p(\mathbf{Y}_T|\theta, A^*)$ . Suppose further that the support of the investigator's prior distribution includes the support of the client's prior. Then the investigator's posterior density may be regarded as an importance sampling density for the client's posterior density. The client reweights the investigator's  $\{\theta^{(m)}\}_{m=1}^M$  using the function

$$w(\theta; A^*) = \frac{p(\theta|\mathbf{Y}_T, A^*)}{p(\theta|\mathbf{Y}_T, A)} = \frac{p(\theta|A^*)p(\mathbf{Y}_T|\theta, A^*)}{p(\theta|A)p(\mathbf{Y}_T|\theta, A)} = \frac{p(\theta|A^*)}{p(\theta|A)}.$$

The client then approximates his posterior moment  $E[g(\theta)|\mathbf{Y}_T, A^*]$  by

$$\bar{g}_M^* = \frac{\sum_{m=1}^M w(\theta^{(m)}; A^*) w(\theta^{(m)}) g(\mathbf{Y}_T, \theta^{(m)})}{\sum_{m=1}^M w(\theta^{(m)}; A^*) w(\theta^{(m)})} \xrightarrow{a.s.} E[g(\mathbf{Y}_T, \theta)|\mathbf{Y}_T, A^*] = \bar{g}^*.$$

If the investigator has employed importance sampling, this result is simply Theorem 3.2.1. For the case in which the investigator has employed Markov chain Monte Carlo, the result can be formalized as follows.

*Theorem 6.1.1.* Let  $p(\mathbf{Y}_T|\theta, A) = p(\mathbf{Y}_T|\theta, A^*)$ . Suppose that  $\{\theta^{(m)}\}$  is ergodic with invariant distribution  $p(\theta|\mathbf{Y}_T, A)d\nu(\theta)$ , and  $E[g(\mathbf{Y}_T, \theta)|\mathbf{Y}_T, A^*]$  exists and is finite. Suppose the support of  $p(\theta|A)$  includes the support of  $p(\theta|A^*)$ , and let  $w(\theta; A^*) = p(\theta|A^*)/p(\theta|A)$ .

Then for all  $\theta^{(0)} \in \Theta$ ,

$$\bar{g}_M^* = \sum_{m=1}^M g(\mathbf{Y}_T, \theta^{(m)}) w(\theta^{(m)}; A^*) / \sum_{m=1}^M w(\theta^{(m)}; A^*) \xrightarrow{a.s.} \int_{\Theta^*} g(\mathbf{Y}_T, \theta) p(\theta|\mathbf{Y}_T, A^*) d\theta = \bar{g}^*.$$

*Proof.* Since  $p(\theta|\mathbf{Y}_T, A^*)$  is integrable with respect to  $d\nu(\theta)$ ,  $w(\theta; A^*)$  is integrable with respect to  $p(\theta|\mathbf{Y}_T, A)d\nu(\theta)$ . From Theorem 3.5.1(C),

$$\begin{aligned} M^{-1} \sum_{m=1}^M w(\theta^{(m)}; A^*) &\xrightarrow{a.s.} E[w(\theta; A^*)|\mathbf{Y}_T, A] \\ &= \int_{\Theta^*} [p(\theta|A^*)/p(\theta|A)] p(\theta|\mathbf{Y}_T, A) d\nu(\theta) = \int_{\Theta^*} p(\theta|A^*) p(\mathbf{Y}_T|\theta, A) d\nu(\theta) \end{aligned}$$

and

$$M^{-1} \sum_{m=1}^M w(\theta^{(m)}; A^*) g(\mathbf{Y}_T, \theta^{(m)}) \xrightarrow{a.s.} \int_{\Theta^*} g(\mathbf{Y}_T, \theta) p(\theta|A^*) p(\mathbf{Y}_T|\theta, A) d\nu(\theta). \quad \#\#$$

Obviously this result is true if  $p(\theta|A^*)$  and  $p(\theta|A)$  are kernels rather than densities; they need not employ the same factor of proportionality. But as discussed in Section 2.3, if  $p(\theta|A^*)$  and  $p(\theta|A)$  are the properly normalized prior densities, then  $M^{-1} \sum_{m=1}^M w(\theta^{(m)}; A^*)$  is a consistent approximation of the Bayes factor in favor of the client's model in comparison with the investigator's model. This fraction, together with the marginal likelihood of the investigator's model, provides the marginal likelihood of the client's model.

In Theorem 3.7.1 uniform ergodicity was one of the sufficient conditions for a central limit theorem. If the investigator's algorithm produces uniformly ergodic  $\{\theta^{(m)}\}$ , and if ratio of the client's prior to the investigator's prior is bounded, then there is a central limit theorem under the client's prior as well, so long as the client's function of interest has finite posterior variance using his prior. This condition is strikingly similar to the sufficient conditions for a central limit theorem for importance sampling in Corollary 3.2.2. This is not surprising: the client is using the investigator's posterior as his importance sampling distribution.

*Theorem 6.1.2.* Given the notation and assumptions of Theorem 6.1.1, suppose also that  $\{\theta^{(m)}\}$  is uniformly ergodic and that  $\text{var}[g(\mathbf{Y}_T, \theta)|\mathbf{Y}_T, A]$  exists and is finite, and  $w(\theta; A^*) \leq \bar{w}^* < \infty \forall \theta \in \Theta$ . Then there exists  $\sigma^2 > 0$  such that

$$M^{1/2} (\bar{g}_M^* - \bar{g}^*) \xrightarrow{d} N(0, \sigma^2).$$

*Proof.* The vector  $\mathbf{v}(\theta)' = [\mathbf{w}(\theta^{(m)}; A^*) \mathbf{g}(\mathbf{Y}_T, \theta^{(m)}), \mathbf{w}(\theta^{(m)}; A^*)]$  is uniformly ergodic, with  $\bar{\mathbf{v}}'_M = M^{-1} \sum_{m=1}^M \mathbf{v}(\theta^{(m)})' \xrightarrow{a.s.} [c \mathbb{E}[\mathbf{g}(\mathbf{Y}_T, \theta) | \mathbf{Y}_T, A^*], c] = \bar{\mathbf{v}}'$ . From Cogburn (1972, Corollary 4.2(ii)), there exists a positive definite matrix  $\Sigma$  such that  $M^{1/2}(\bar{\mathbf{v}}_M - \bar{\mathbf{v}}) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \Sigma)$ . A standard application of the delta method yields the result.##

Efficiency of the reweighting scheme requires some similarity of  $p(\theta|A^*)$  and  $p(\theta|A)$ , as illustrated subsequently in Section 6.3. In particular, both reasonable convergence rates and the use of a central limit theorem to assess numerical accuracy essentially require that  $p(\theta|A^*)/p(\theta|A)$  be bounded. Across a set of diverse clients this condition is more likely to be satisfied the more diffuse is  $p(\theta|A)$ , and is trivially satisfied for the (possibly improper) prior  $p(\theta|A) \propto \chi_{\Theta}(\theta)$  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). Relative numerical efficiency will indicate situations in which the reweighting scheme is inefficient. If the investigator chooses to use an improper prior for reporting, it is of course incumbent on her to verify the existence of the posterior distribution and convergence of her posterior simulator.

Including  $p(\theta^{(m)}|A)$  in the standard posterior simulator file avoids the need for every client who wishes to impose his own priors to re-evaluate the investigator's prior. Of course,  $p(\theta|A^*)$  need not be the client's subjective prior: it may simply be a device by which the client, functioning as another investigator, explores robustness of results with respect to alternative reasonable priors.

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

$$\begin{aligned} p(\theta | \mathbf{Y}_{T+f}, A) &\propto p(\theta|A) p(\mathbf{Y}_{T+f} | \theta, A) = p(\theta|A) p(\mathbf{Y}_T | \theta, A) \prod_{s=T+1}^{T+f} p(\mathbf{y}_s | \mathbf{Y}_{s-1}, \theta, A) \\ &\propto p(\theta | \mathbf{Y}_T, A) \prod_{s=T+1}^{T+f} p(\mathbf{y}_s | \mathbf{Y}_{s-1}, \theta, A). \end{aligned}$$

The client therefore forms the approximation to the updated posterior moment  $\mathbb{E}[\mathbf{g}(\mathbf{Y}_{T+f}, \theta) | \mathbf{Y}_{T+f}, A]$ ,

$$\bar{\mathbf{g}}_M^* = \frac{\sum_{m=1}^M \mathbf{w}(\theta^{(m)}; \mathbf{Y}_{T+f}) \mathbf{w}(\theta^{(m)}) \mathbf{g}(\mathbf{Y}_{T+f}, \theta^{(m)})}{\sum_{m=1}^M \mathbf{w}(\theta^{(m)}; \mathbf{Y}_{T+f}) \mathbf{w}(\theta^{(m)})} \xrightarrow{a.s.} \mathbb{E}[\mathbf{g}(\theta) | \mathbf{Y}_{T+f}, A] = \bar{\mathbf{g}}^*.$$

with  $\mathbf{w}(\theta; \mathbf{Y}_{T+f}) = \prod_{s=T+1}^{T+f} p(\mathbf{y}_s | \mathbf{Y}_{s-1}, \theta, A)$ . If  $f$  is small relative to  $T$ , and there is no major change in the data generating process between  $T$  and  $T+f$ , the new approximation will be

efficient. But as  $f$  grows, efficiency diminishes and at some point the approximation  $\bar{g}^*$  becomes too inaccurate to be useful. This process also requires evaluation of the likelihood function, which usually involves more technical difficulties than evaluation of priors or functions of interest.

## 6.2 Software

The program `reweight` transforms the parameter vectors in a posterior simulator file, and/or modifies the weights associated with each iteration, and then writes another posterior simulator file incorporating these changes. By transforming the parameter vectors, the program can be used to examine posterior moments other than those corresponding to the posterior expectations of the parameters or functions of interest in the original posterior simulator file, through subsequent use of `moment`. By changing the weights the program can be used to change the original prior distribution, and then examine the effect of the change on posterior moments through subsequent use of `moment`.

The actual transformation and changing of the weighting function is accomplished through an auxiliary procedure called `client`. (The exact form of this procedure depends on the programming language.) This procedure takes as input a parameter vector read from the old posterior simulator file, and the corresponding log prior density and log weight. It returns as output the new parameter vector, new log prior density, and new log weight.

## 6.3 Examples

Return to the regression model of hedonic pricing introduced in Section 3.9. Suppose that an investigator wishes to provide a posterior simulator file with the intention that clients will impose their own priors by reweighting the output. To this end the investigator should use a prior distribution that is uninformative relative to the data. Illustrating what such an investigator might do, prior distributions  $N(0, 1^2)$  for all slope coefficients, except  $N(0, 3^2)$  for  $\log(\text{Lot size})$ , and  $.04h \sim \chi^2(1)$  for precision, were chosen and the corresponding posterior simulator file with 10,000 replications was created. To mimic a client, the tightest of the three prior distributions described in Section 4.6 was chosen:  $N(.1, .05^2)$  for all slope coefficients except  $N(.3, .15^2)$  for lot size, and  $.12h \sim \chi^2(3)$  for precision. Using `reweight` a posterior simulator file with the corresponding weights was created; this required less than two seconds. Then `moment` was used to obtain posterior moments for the coefficients; this required 28 seconds.

The results of this exercise are displayed in Table 6.1. The left panel provides the results that would have been obtained had the client directly executed the posterior simulator corresponding to his prior. The accuracy of the numerical approximation of the

posterior moments is similar to that exhibited for the less informative prior in Table 3.1. The right panel displays the results the client obtains by reweighting the investigator's simulator output. The “*t*” statistics (last column) comparing the posterior means approximated in these two different ways indicate no difficulties with the assessment of numerical accuracy through the numerical standard errors. Relative numerical efficiency for the coefficient posterior means ranges from .18 to .60: overall, the client obtains the same numerical accuracy he would have achieved executing the simulator directly with about 3000 iterations. This would have required 21 seconds and the `uvr1` software, whereas the reweighting took 2 seconds and the simpler `reweight` software. Reweighting of the simulator output succeeds, in this example, because both the investigator's and the client's prior are uninformative relative to the sample and their posterior distributions are therefore similar.

A similar exercise was conducted for the probit example introduced in Section 3.9. To mimic what an investigator might do, a posterior simulator file for the posterior distribution with a prior distribution in which all standard deviations were 10 times larger than those indicated in Table 3.2, was created. This simulator output was then reweighted to reflect the prior distribution in Table 3.2. This attempt failed completely -- a single draw received more than 99.99% of the total weight. To appreciate the reason for this failure recall that the prior distribution used in Section 3.9 is highly informative relative to the sample. This is evident from inspection of Table 3.2.

To complete this example, the investigator was recast in the role of a client — that is, the client now has the priors indicated in Table 3.2 except that the standard deviations are ten times larger than shown there. Imagine an investigator who uses a prior distribution with standard deviations ten times larger yet — that is, the standard deviations are 100 times those shown for the prior distribution in Table 3.2. This client's reweighting of this investigator's simulator output yields the results displayed in the right half of Table 6.2. The “*t*” statistic for comparison of the posterior means again indicates no problem with numerical standard errors (NSEs). The relative numerical efficiencies (RNEs) show that the investigator's simulator output with 10,000 records provides about the same information the client would have obtained with 1000 records directly from simulator output for his posterior. This would have required the `pbt1` software and about 70 seconds of execution time, whereas the reweighting required 2 seconds and much simpler software. (Notice that the posterior means, in Table 6.2, are much closer to the maximum likelihood statistics in Table 3.2, than are the posterior means in the same table which correspond to the previous more informative prior.)

These examples underscore both the potential efficiency of Bayesian communication through posterior reweighting, and its limitations. The efficiency comes about because reweighting software is simple and generic, whereas posterior simulators are model specific and impose greater computational demands. (This advantage increases dramatically in more complex models.) The limitations arise from the need for some similarity of the investigator's and client's posterior distributions. As argued and illustrated here, the investigator should use a prior distribution that is uninformative relative to the sample. In this situation a client's reweighting will be successful for priors that are moderately, but not greatly, informative relative to the sample.

**Table 3.3**  
Probit model: Posterior moments

Coefficient	Posterior (Gibbs)				Posterior (Hastings-Metropolis)				“t”
	Mean	Stan. dev.	NSE	RNE	Mean	Stan. dev.	NSE	RNE	
Intercept	1.2114	.174	.0025	.533	1.2194	.176	.0035	.285	-1.86
Black	.01387	.0767	.00135	.358	.01445	.0792	.00165	.254	-.27
Age-Single	-.010323	.00373	$5.94 \times 10^{-5}$	.440	-.010365	.00382	$5.41 \times 10^{-5}$	.554	.52
Age-Married	-.028064	.00648	$1.27 \times 10^{-4}$	.289	-.028037	.00643	$1.11 \times 10^{-4}$	.372	-.16
Education	.0023627	.00410	$3.75 \times 10^{-5}$	1.330	.0021866	.00416	$4.99 \times 10^{-5}$	.774	2.82
Married	.18227	.119	.00123	1.030	.17997	.115	.00171	.504	1.09
Kids	-.36985	.133	.00285	.242	-.37198	.130	.00218	.398	.59
#Kids	-.15142	.0444	$7.24 \times 10^{-4}$	.418	-.15165	.0438	$8.01 \times 10^{-4}$	.333	.21
Spouse\$	$-7.2693 \times 10^{-6}$	$2.28 \times 10^{-6}$	$5.06 \times 10^{-8}$	.224	$-7.2936 \times 10^{-6}$	$2.27 \times 10^{-6}$	$3.05 \times 10^{-8}$	.614	.41
Family\$	$-1.0994 \times 10^{-6}$	$2.90 \times 10^{-6}$	$4.99 \times 10^{-8}$	.376	$-1.1187 \times 10^{-6}$	$2.98 \times 10^{-6}$	$4.99 \times 10^{-8}$	.395	.27
AFDC	$-5.6986 \times 10^{-4}$	$3.11 \times 10^{-4}$	$5.93 \times 10^{-6}$	.306	$-5.6895 \times 10^{-4}$	$3.09 \times 10^{-4}$	$6.55 \times 10^{-6}$	.247	-.10
Food\$	-.0012122	$4.40 \times 10^{-4}$	$5.08 \times 10^{-6}$	.831	-.0012110	$4.34 \times 10^{-4}$	$7.66 \times 10^{-6}$	.356	-.13
WorkExp	$1.1747 \times 10^{-4}$	$8.19 \times 10^{-6}$	$3.12 \times 10^{-7}$	.077	$1.1720 \times 10^{-4}$	$8.28 \times 10^{-6}$	$1.26 \times 10^{-7}$	.478	.80

**Table 6.1**

Regression model: Comparison of direct MCMC with client's reweighting

Coefficient	Posterior (Tightest prior, direct Gibbs)				Posterior (Reweighting of MCMC from more diffuse prior)				“t”
	Mean	Stan. dev.	NSE	RNE	Mean	Stan. dev.	NSE	RNE	
Intercept	7.7280	.2100	.0018	1.526	7.7170	.2079	.0038	.327	2.61
Driveway	.10774	.02484	.00030	.775	.10747	.02442	.00046	.312	.49
Rec room	.068375	.02265	.00045	.949	.068712	.023079	.00039	.379	-.57
Fin basement	.10335	.01962	.00021	.952	.10338	.01951	.00042	.239	-.06
Gas hot water	.14335	.03329	.00046	.571	.14319	.03266	.00066	.270	.20
Central air	.15407	.01943	.00014	2.177	.15407	.01933	.00039	.280	.00
#Garage stalls	.052000	.01117	.00011	1.234	.052095	.011402	.00026	.217	-.34
Good nbhd	.12585	.02064	.00022	1.003	.12561	.020970	.00052	.180	.43
log(Lot size)	.30468	.02574	.00024	1.292	.30609	.025485	.00051	.286	-2.50
#Bedrooms	.040620	.013536	.00017	.698	.040376	.013362	.00025	.313	.81
#Full baths	.15545	.018749	.00019	1.064	.15523	.018589	.00041	.227	.49
#Stories	.093635	.012025	.00010	1.530	.093804	.012037	.00016	.595	-.90

**Table 6.2**

Probit model: Comparison of direct MCMC with client's reweighting

Coefficient	Posterior (Gibbs, direct MCMC)				Posterior (Reweighting of MCMC from more diffuse prior)				“t”
	Mean	Stan. dev.	NSE	RNE	Mean	Stan. dev.	NSE	RNE	
Intercept	1.548	.4640	.0068	.517	1.606	.4712	.0138	.129	-3.77
Black	.1077	.1051	.0028	.160	.1043	.1056	.0037	.088	.73
Age-Single	-.05702	.01227	.00022	.356	-.05752	.01217	.00035	.131	1.21
Age-Married	-.08360	.01212	.00021	.361	-.08375	.01201	.00037	.119	.35
Education	.07898	.02240	.00042	.316	.07581	.02277	.00074	.106	3.73
Married	.6570	.4694	.0068	.528	.6439	.4832	.0164	.096	.74
Kids	-.5093	.1652	.0052	.112	-.5249	.1667	.0054	.107	2.08
#Kids	-.05441	.05453	.00129	.231	-.05433	.05353	.00180	.098	-.04
Spouse\$	$-1.687 \times 10^{-5}$	$3.455 \times 10^{-6}$	$7.52 \times 10^{-8}$	.235	$-1.662 \times 10^{-5}$	$3.407 \times 10^{-6}$	$1.09 \times 10^{-7}$	.108	-2.18
Family\$	$8.975 \times 10^{-7}$	$6.246 \times 10^{-6}$	$9.93 \times 10^{-8}$	.440	$9.052 \times 10^{-7}$	$6.240 \times 10^{-6}$	$2.12 \times 10^{-7}$	.096	-.02
AFDC	$-5.397 \times 10^{-4}$	$3.836 \times 10^{-4}$	$6.81 \times 10^{-6}$	.352	$-5.252 \times 10^{-4}$	$3.731 \times 10^{-4}$	$1.07 \times 10^{-5}$	.135	-1.14
Food\$	-.002196	$6.845 \times 10^{-4}$	$1.14 \times 10^{-5}$	.400	-.002190	$6.719 \times 10^{-4}$	$2.13 \times 10^{-5}$	.110	.25
WorkExp	$1.367 \times 10^{-4}$	$9.543 \times 10^{-6}$	$4.37 \times 10^{-7}$	.053	$1.376 \times 10^{-4}$	$9.072 \times 10^{-6}$	$3.89 \times 10^{-7}$	.060	1.54

**Table 3.1**

Regression model: Posterior moments

Coefficient	Ordinary least squares		Prior		Posterior			
	Estimate	s.e.	Mean	s.d.	Mean	s.d.	NSE	RNE
Intercept	7.745	.216	0	11	7.726	.217	.0015	2.09
Driveway	.110	.028	0	.1	.104	.027	.0002	1.59
Recreation room	.058	.026	0	.1	.058	.025	.0003	1.05
Finished basement	.104	.021	0	.1	.103	.021	.0002	0.96
Gas hot water	.179	.043	0	.1	.149	.040	.0004	0.97
Central air	.166	.021	0	.1	.159	.020	.0001	2.16
#Garage stalls	.048	.011	0	.1	.049	.011	.0001	1.49
Good nbhd	.132	.023	0	.1	.127	.022	.0002	1.04
log(Lot size)	.303	.027	0	.3	.307	.027	.0002	1.98
#Bedrooms	.034	.014	0	.1	.036	.014	.0001	1.14
#Full bathrooms	.166	.020	0	.1	.161	.020	.0002	1.34
#Stories	.092	.013	0	.1	.093	.013	.0001	2.07

**Table 3.2**

Probit model: Likelihood mode, prior and posterior mode

Coefficient	Maximum likelihood		Prior		Posterior	
	Mode	Asymptotic s.e.	Mean	Stan. dev.	Mode	Approx. s.d.
Intercept	1.22	.520	0	4	1.21	.177
Black	.109	.105	0	.125	.0151	.0773
Age-Single	-.0611	.0132	0	.00417	-.0102	.00381
Age-Married	-.0874	.0128	0	.03333	-.0279	.00652
Education	.113	.0274	0	.00417	.00228	.00411
Married	.682	.522	0	.125	.180	.118
Kids	-.488	.171	0	.250	-.365	.131
#Kids	-.0505	.0552	0	.125	-.151	.0441
Spouse\$	$-1.78 \times 10^{-5}$	$3.46 \times 10^{-6}$	0	$3.57 \times 10^{-6}$	$-7.18 \times 10^{-6}$	$2.28 \times 10^{-6}$
Family\$	$1.86 \times 10^{-6}$	$7.25 \times 10^{-6}$	0	$3.57 \times 10^{-6}$	$-9.51 \times 10^{-7}$	$2.92 \times 10^{-6}$
AFDC	$-5.02 \times 10^{-4}$	$3.85 \times 10^{-4}$	0	$6.25 \times 10^{-4}$	$-5.68 \times 10^{-4}$	$3.05 \times 10^{-4}$
Food\$	-.00211	$6.94 \times 10^{-4}$	0	$6.25 \times 10^{-4}$	-.00121	$4.35 \times 10^{-4}$
WorkExp	$1.36 \times 10^{-4}$	$9.50 \times 10^{-6}$	0	$6.25 \times 10^{-5}$	$1.16 \times 10^{-4}$	$8.28 \times 10^{-6}$

**Table 4.1**  
Regression model: Marginal likelihoods

	Log marginal likelihood	NSE
First prior (Zero center)		
$p = .9$	46.077	.003
$p = .5$	46.069	.011
$p = .1$	46.063	.047
Second prior (Nonzero center)		
$p = .9$	52.145	.004
$p = .5$	52.132	.012
$p = .1$	52.122	.029
Third prior (Nonzero center, higher precision)		
$p = .9$	56.362	.004
$p = .5$	56.372	.011
$p = .1$	56.383	.036

**Table 4.2**  
Probit Model: Marginal likelihoods

	Log marginal likelihood	NSE
Gibbs algorithm based on Gelfand-Dey		
$p = .9$	-564.72	.0053
$p = .5$	-564.72	.0151
$p = .1$	-564.69	.0393
Hastings-Metropolis based on Gelfand-Dey		
$p = .9$	-564.72	.0070
$p = .5$	-564.71	.0185
$p = .1$	-564.68	.0544
Hastings-Metropolis based on weights	-564.69	.0054

**Table 5.1**

Regression model: Limited information marginal likelihood

	Predictive c.d.f.		Data	Limited information marginal likelihood
	.05	.95		
Data function	.05	.95	Data	
Skewness	-.160	.175	-.184	1.17
Excess kurtosis	-.319	.340	.517	.177
Skewness + excess kurtosis	.0009	.074	.101	*
Kolmogorov-Smirnov	.024	.238	.044	7.24
Nonlinear regression				
#Garage stalls	-.0235	.0235	-.0349	2.534
log(Lot size)	-.00216	.00221	-.00141	185.4
#Bedrooms	-.0123	.0125	-.0074	29.103
#Full bathrooms	-.0135	.0131	-.0023	36.8
#Stories	-.0143	.0144	.0020	47.8
Conditional heteroscedasticity				
Driveway	-.070	.067	-.009	8.33
Recreation room	-.071	.066	-.011	8.49
Finished basement	-.072	.070	.015	8.86
Gas hot water	-.070	.069	.059	2.27
Central air	-.075	.069	-.075	.81
#Garage stalls	-.067	.072	.082	1.15
Good nbhd	-.071	.066	-.113	*
log(lot size)	-.065	.073	-.017	8.82
#Bedrooms	-.074	.065	.067	1.89
#Full bathrooms	-.071	.070	-.027	9.48
#Stories	-.070	.073	-.044	5.76

\*Value too small to be approximated reliably by kernel smoothing methods.

**Table 5.2**

Probit model: Limited information marginal likelihood

Data function	Predictive c.d.f.		Data	Limited information marginal likelihood
	.05	.95		
Fraction choosing	.000	1.000	.803	.279
Fraction in:				
.00 - .10	.000	.037	.222	*
.10 - .25	.000	.222	.208	1.088
.25 - .40	.000	.400	.338	3.942
.40 - .50	.273	.571	.362	.321
.50 - .60	.454	.667	.557	1.74
.60 - .75	.571	1.000	.663	4.675
.75 - .90	.772	.941	.830	11.43
.90 - .95	.900	1.000	.953	1.564
.95 - .99	.960	1.000	.983	4.204
.99 - 1.00	.995	1.000	.990	5.657
Correlation of squared residual with squared covariate:				
Black	-.958	.116	.022	6.57
Age-Single	-.206	.226	-.252	6.52
Age-Married	-.254	.242	.038	2.15
Education	-.095	.070	-.154	.244
Married	-.269	.236	.028	5.66
Kids	-.131	.105	.213	.098
#Kids	-.102	.169	.163	.701
Spouse\$	-.078	.061	.024	3.22
Family\$	-.030	.006	-.012	16.9
AFDC	-.108	.177	.079	2.19
FoodStamps	-.104	.190	.092	1.07
WorkExp	-.202	.238	-.264	.318

\*Value too small to be approximated reliably by kernel smoothing methods.