Bayesian Inference in Econometric Models Using Monte Carlo Integration
Author(s): John Geweke
Source: *Econometrica*, Vol. 57, No. 6 (Nov., 1989), pp. 1317-1339
Published by: The Econometric Society
Stable URL: http://www.jstor.org/stable/1913710
Accessed: 14/08/2010 13:30

Econometrica, Vol. 57, No. 6 (November, 1989), 1317–1339

BAYESIAN INFERENCE IN ECONOMETRIC MODELS USING MONTE CARLO INTEGRATION

BY JOHN GEWEKE[1]

Methods for the systematic application of Monte Carlo integration with importance sampling to Bayesian inference in econometric models are developed. Conditions under which the numerical approximation of a posterior moment converges almost surely to the true value as the number of Monte Carlo replications increases, and the numerical accuracy of this approximation may be assessed reliably, are set forth. Methods for the analytical verification of these conditions are discussed. Importance sampling densities are derived from multivariate normal or Student t approximations to local behavior of the posterior density at its mode. These densities are modified by automatic rescaling along each axis. The concept of relative numerical efficiency is introduced to evaluate the adequacy of a chosen importance sampling density. The practical procedures based on these innovations are illustrated in two different models.

KEYWORDS: Importance sampling, numerical integration, Markov chain model, ARCH linear model.

1. INTRODUCTION

ECONOMETRIC MODELS are usually expressed in terms of an unknown vector of parameters $\underline{\theta} \in \Theta \subseteq R^k$, which fully specifies the joint probability distribution of the observations $X = \{\underline{x}_1, \ldots, \underline{x}_T\}$. In most cases there exists a probability density function $f(X|\underline{\theta})$, and classical inference then often proceeds from the likelihood function $L(\underline{\theta}) = f(X|\underline{\theta})$. The asymptotic behavior of the likelihood function is well understood, and as a consequence there is a well developed set of tools with which problems of computation and inference can be approached; Quandt (1983) and Engle (1984) provide useful surveys. The analytical problems in a new model are often far from trivial, but there are typically several approaches that can be explored systematically with the realistic anticipation that one or more will lead to classical inference procedures with an asymptotic justification.

Bayesian inference proceeds from the likelihood function and prior information which is usually expressed as a probability density function over the parameters, $\pi(\underline{\theta})$, it being implicit that $\pi(\underline{\theta})$ depends on the conditioning set of prior information. The posterior distribution is proportional to $p(\underline{\theta}) = \pi(\underline{\theta})L(\underline{\theta})$. This formulation restricts the prior probability measure for Θ to be absolutely continuous with respect to Lebesgue measure. Extensions that allow concentrations of prior probability mass are generally straightforward but notationally cumbersome, and are postponed to the concluding section. Most Bayesian inference problems can be expressed as the evaluation of the expectation of a

[1]Financial support for this work was provided by NSF Grant SES-8605867, and by a grant from the Pew Foundation to Duke University. This work benefited substantially by suggestions from an editor, Jean-Francois Richard, Herman van Dijk, Robert Wolpert, and two anonymous referees, who bear no responsibility for any errors of commission or omission.

function of interest $g(\underline{\theta})$ under the posterior,

$$(1) \qquad E[g(\underline{\theta})] = \int_{\Theta} g(\underline{\theta}) p(\underline{\theta})\, d\underline{\theta} \Big/ \int_{\Theta} p(\underline{\theta})\, d\underline{\theta}.$$

Approaches to this problem are nowhere near as systematic, methodical, or general as are those to classical inference problems: classical inference is carried out routinely using likelihood functions for which the evaluation of (1) is at best a major and dubious undertaking and for most practical purposes impossible. If the integration in (1) is undertaken analytically then the range of likelihood functions that can be considered is small, and the class of priors and functions of interest that can be considered is severely restricted. Many numerical approaches, like quadrature methods, require special adaptation for each g, π, or L, and become unworkable if k exceeds, say, three. (Good surveys of these methods may be found in Davis and Rabinowitz, 1975; Hammersley and Handscomb, 1979; and Rubinstein, 1981.)

With the advent of powerful and cheap computing, numerically intensive methods for the computation of (1) have become more attractive. Monte Carlo integration with importance sampling provides a systematic approach that—in principle—can be applied in any situation in which $E[g(\underline{\theta})]$ exists, and is practical for large values of k. It was discussed by Hammersley and Handscomb (1964, Section 5.4) and brought to the attention of econometricians by Kloek and van Dijk (1978). The main idea is simple. Let $\{\underline{\theta}_i\}$ be a sequence of k-dimensional random vectors, i.i.d. in theory and in practice generated synthetically and therefore a very good approximation to an i.i.d. sequence. If the probability distribution function of the $\underline{\theta}_i$ is proportional to $p(\underline{\theta})$, then $n^{-1}\sum_{i=1}^{n} g(\underline{\theta}_i) \to E[g(\underline{\theta})]$; if it is proportional to $L(\underline{\theta})$, then $n^{-1}\sum_{i=1}^{n} g(\underline{\theta}_i)\pi(\underline{\theta}_i) \to E[g(\underline{\theta})]$. (Except in Section 4, all convergence is in n, the number of Monte Carlo replications. Almost sure convergence is indicated " $\to$ ", convergence in distribution " $\Rightarrow$ ".) Only in a very limited set of simple cases is it feasible to generate synthetic variates whose p.d.f. is in proportion to the posterior or the likelihood function (but the set does include common and interesting cases in which classical and analytical Bayesian inference fail; see Geweke, 1986). More generally, suppose that the probability distribution function of the $\underline{\theta}_i$ is $I(\underline{\theta})$, termed the *importance sampling density*. Then, under very weak assumptions given in Section 2,

$$(2) \qquad \bar{g}_n \equiv \sum_{i=1}^{n} [g(\underline{\theta}_i) p(\underline{\theta}_i)/I(\underline{\theta}_i)] \Big/ \sum_{i=1}^{n} [p(\underline{\theta}_i)/I(\underline{\theta}_i)] \to E[g(\underline{\theta})].$$

The rate of almost sure convergence in (2) depends critically on the choice of the importance sampling density. Under stronger assumptions, developed in Section 3,

$$(3) \qquad n^{1/2}\{g_n - E[\bar{g}(\underline{\theta})]\} \Rightarrow N(0, \sigma^2),$$

and σ^2 may be estimated consistently. This result was indicated by Kloek and

van Dijk (1978), but does not follow from the result of Cramer (1946) they cite, and is repeated by Bauwens (1984) but without explicit discussion of the moments whose existence is assumed. Loosely speaking, the importance sampling density should mimic the posterior density, and it is especially important that the tails of $I(\underline{\theta})$ not decay more quickly than the tails of $p(\underline{\theta})$. This is keenly appreciated by those who have done empirical work with importance sampling, including Zellner and Rossi (1984), Gallant and Monahan (1985), Bauwens and Richard (1985), and van Dijk, Kloek, and Boender (1985). These and other investigators have experienced substantial difficulties in tailoring importance sampling densities to the problem at hand. This is an important failing, for the approach is neither attractive nor methodical if special arcane problems in numerical analysis have to be resolved in each application.

This paper approaches these problems analytically, and presents new results which should make the application of Monte Carlo integration by importance sampling much more routine. Section 3 provides sufficient and easily verified conditions for (3). It also introduces a measure of relative numerical efficiency which is natural to use in evaluating the effectiveness of any given importance sampling density. Based on these considerations, Section 4 outlines a systematic approach to the choice of $I(\underline{\theta})$. It utilizes the local behavior of the posterior density at its mode, and a new class of importance sampling densities, the multivariate split-normal and multivariate split-Student. These developments are illustrated in Section 5, with the homogeneous Markov chain model, and in Section 6, with the ARCH linear model. Some directions for future research are taken up in the last section.

2. BAYESIAN INFERENCE WITH IMPORTANCE SAMPLING

We begin with some additional notation, and a set of assumptions basic to what follows. Since expectations are taken with respect to a variety of distributions, it is useful to denote $E_f[h(\underline{\theta})] = [\int_\Theta f(\underline{\theta})\, d\theta]^{-1} \int_\Theta f(\underline{\theta}) h(\underline{\theta})\, d\underline{\theta}$ where $f(\underline{\theta})$ is proportional to a probability density and $h(\underline{\theta})$ is integrable with respect to the probability measure that is induced by $f(\underline{\theta})$. Similarly denote by $\mathrm{var}_f[h(\underline{\theta})]$ and $md_f[h(\underline{\theta})]$ the variance and mean deviation of $h(\underline{\theta})$ under this probability measure, when the respective moments exist.

ASSUMPTION 1: *The product of the prior density, $\pi(\underline{\theta})$, and the likelihood function, $L(\underline{\theta})$, is proportional to a proper probability density function defined on Θ.*

ASSUMPTION 2: *$\{\underline{\theta}_i\}_{i=1}^{\infty}$ is a sequence of i.i.d. random vectors, the common distribution having a probability density function $I(\underline{\theta})$.*

ASSUMPTION 3: *The support of $I(\underline{\theta})$ includes Θ.*

ASSUMPTION 4: *$E[g(\underline{\theta})]$ exists and is finite.*

Assumption 1 is generally satisfied and usually not difficult to verify. It explicitly allows the prior density to be improper, so long as $\int_\Theta p(\underline{\theta})\, d\underline{\theta} \equiv c < \infty$. For $h(\underline{\theta})$ integrable with respect to the posterior probability measure, the simpler notation $\bar{h} \equiv E[h(\underline{\theta})] \equiv E_p[h(\underline{\theta})]$ (and similarly $\mathrm{var}[h(\underline{\theta})]$ and $\mathrm{md}[h(\underline{\theta})]$) will be used. Assumptions 2 and 3 are specific to the method of Monte Carlo integration with importance sampling; their role is obvious from the discussion in the introduction. Most inference problems can be cast in the form of determining $E[g(\underline{\theta})]$, through the function of interest, $g(\underline{\theta})$.

The value of $\bar{g}_n$ in (2) is invariant with respect to arbitrary scaling of the posterior and importance sampling densities. In certain situations it is convenient not to work with normalizing constants for $I(\underline{\theta})$, so we shall re-express $\bar{g}_n$ with $I^*(\underline{\theta}) = d^{-1} I(\underline{\theta})$ in place of $I(\underline{\theta})$. Denote the *weight function* $w(\underline{\theta}) \equiv p(\underline{\theta})/I^*(\underline{\theta})$; then $\bar{g}_n = \sum_{i=1}^n g(\underline{\theta}_i) w(\underline{\theta}_i) / \sum_{i=1}^n w(\underline{\theta}_i)$. Consistency of $\bar{g}_n$ for g is a direct consequence of the strong law of large numbers.

THEOREM 1: *Under Assumptions* 1–4, $\bar{g}_n \to E[g(\underline{\theta})]$.

This simple result provides a foundation for the numerical treatment of a very wide array of practical problems in Bayesian inference. A few classes of examples are worth mention.

(i) If the value of a Borel measurable function of the parameters, $h(\underline{\theta})$, whose expectation exists and is finite, is to be estimated with a quadratic loss function, then $g(\underline{\theta}) = h(\underline{\theta})$.

(ii) If a decision is to be made between actions a_1 and a_2, according to the expectation of loss functions $l(\underline{\theta}; a_1)$ and $l(\underline{\theta}; a_2)$, then $g(\underline{\theta}) = l(\underline{\theta}; a_1) - l(\underline{\theta}; a_2)$.

(iii) Posterior probabilities may be assessed through indicator functions of the form $\chi(\underline{\theta}; S) = 1$ if $\underline{\theta} \in S$, 0 if $\underline{\theta} \notin S$. To evaluate the posterior probability $P[h(\underline{\theta}) \in A]$, let $g(\underline{\theta}) = \chi(\underline{\theta}; \{\underline{\theta}: h(\underline{\theta}) \in A\})$. Assumption 4 is satisfied trivially, so given the other assumptions $\bar{g}_n \to P[h(\underline{\theta}) \in A]$. Construction of posterior c.d.f.'s is a special case that is particularly relevant if $h(\underline{\theta})$ itself has no posterior moments. (An example is provided subsequently in Section 5; see especially Table IV.)

(iv) The application to c.d.f.'s suggests how to construct quantiles. Suppose $P[h(\underline{\theta}) \leqslant x] = \alpha$, and for a given $\varepsilon > 0$ $P[h(\underline{\theta}) \leqslant x - \varepsilon] < \alpha$, $P[h(\underline{\theta}) < x + \varepsilon] > \alpha$. If q_n is defined to be any number such that $\sum_{i:\, h(\underline{\theta}_i) \leqslant q_n} w(\underline{\theta}_i) / \sum_{i=1}^n w(\underline{\theta}_i) \geqslant \alpha$ and $\sum_{i:\, h(\underline{\theta}_i) \geqslant q_n} w(\underline{\theta}_i) / \sum_{i=1}^n w(\underline{\theta}_i) \geqslant 1 - \alpha$, then by Theorem 1 $\lim_{n \to \infty} P[q_n \in (x - \varepsilon, x + \varepsilon)] = 1$. Hence if the c.d.f. of $h(\underline{\theta})$ is strictly increasing at x, the αth quantile can be computed routinely as a byproduct of Monte Carlo integration with importance sampling.

(v) In forecasting problems functions of interest are of the form $g(\underline{\theta}) = P(x_{T+s} \in A | \underline{\theta}, X)$ and in this case $g(\underline{\theta})$ itself is an integral. Monte Carlo integration with an importance sampling density for the parameters may be combined with Monte Carlo integration for future realizations of the random vector $\underline{x}_t$ to form predictive densities (Geweke (1988b)).

3. EVALUATING THE IMPORTANCE SAMPLING DENSITY

Standing alone these results are of little practical value, because nothing can be said about rates of convergence and there is no formal guidance for choosing $I(\underline{\theta})$ beyond the obvious requirement that subsets of the support of the posterior should not be systematically excluded. Moreover, as noted in the introduction, for seemingly sensible choices of $I(\underline{\theta})$, $\bar{g}_n$ can behave badly. Poor behavior is usually manifested by values of $\bar{g}_n$ that exhibit substantial fluctuations after thousands of replications, which in turn can be traced to extremely large values of $w(\underline{\theta}_i)$ that turn up occasionally. The key problem is the distribution of the ratio of two means in $\bar{g}_n$, and this may be characterized by a central limit theorem.

THEOREM 2: *In addition to Assumptions* 1–4, *suppose*

$$E[w(\underline{\theta})] = c^{-1}\int_{\Theta}\left[p(\underline{\theta})^2/I^*(\underline{\theta})\right]d\underline{\theta}$$

and

$$E\left[g(\underline{\theta})^2 w(\underline{\theta})\right] = c^{-1}\int_{\Theta}\left[g(\underline{\theta})^2 p(\underline{\theta})^2/I^*(\underline{\theta})\right]d\underline{\theta}$$

are finite. Let

$$\sigma^2 \equiv d^{-1}E\left\{\left[g(\underline{\theta})-\bar{g}\right]^2 w(\underline{\theta})\right\} = c^{-1}d^{-1}\int_{\Theta}\left[g(\underline{\theta})-\bar{g}\right]^2 w(\underline{\theta})\,p(\underline{\theta})\,d\theta,$$

$$\hat{\sigma}_n^2 \equiv \sum_{i=1}^{n}\left[g(\underline{\theta}_i)-\bar{g}_n\right]^2 w(\underline{\theta}_i)^2 \Bigg/ \left[\sum_{i=1}^{n} w(\underline{\theta}_i)\right]^2.$$

Then

$$n^{1/2}(\bar{g}_n - \bar{g}) \Rightarrow N(0,\sigma^2),$$

$$n\hat{\sigma}_n^2 \to \sigma^2.$$

PROOF: Define $A(\underline{\theta}) \equiv g(\underline{\theta})w(\underline{\theta})$ and observe that $\bar{g}_n = [n^{-1}\Sigma_{i=1}^n A(\underline{\theta}_i)]/[n^{-1}\Sigma_{i=1}^n w(\underline{\theta}_i)]$. From Theorem 1, $n^{-1}\Sigma_{i=1}^n A(\underline{\theta}_i) \to c\,d\bar{g}$, $n^{-1}\Sigma_{i=1}^n w(\underline{\theta}_i) \to cd$. By a Central Limit Theorem (e.g. Billingsley (1979, Theorem 29.5)), $n^{1/2}(\bar{g}_n - \bar{g}) \Rightarrow N(0,\sigma^2)$,

$$\sigma^2 = c^{-2}d^{-2}\left\{\operatorname{var}[A(\underline{\theta})] + \bar{g}^2\operatorname{var}[w(\underline{\theta})] - 2\bar{g}\operatorname{cov}[A(\underline{\theta}), w(\underline{\theta})]\right\}$$

$$= c^{-2}d^{-1}\int_{\Theta}\left[g(\underline{\theta})-\bar{g}\right]^2 w(\underline{\theta})\,p(\underline{\theta})\,d\underline{\theta}. \qquad \textit{Q.E.D.}$$

We shall refer to $\hat{\sigma}_n = (\hat{\sigma}_n^2)^{1/2}$ as the numerical standard error of $\bar{g}_n$.

The conditions of Theorem 2 must be verified, analytically, if that result is to be used to assess the accuracy of $\bar{g}_n$ as an approximation of $E[g(\underline{\theta})]$. Failure of

these conditions does not vitiate the use of $\bar{g}_n$: there is no compelling reason why $n^{1/2}(\bar{g}_n - \bar{g})$ needs to have a limiting distribution. But it is essential to have some indication of $|\bar{g}_n - \bar{g}|$, and this task becomes substantially more difficult if a central limit theorem cannot be applied. Practicality suggests that $I(\underline{\theta})$ be chosen so that Theorem 2 applies. The additional assumptions in Theorem 2 typically can be verified by showing either

(4) $\quad w(\underline{\theta}) < \bar{w} < \infty \quad \forall \underline{\theta} \in \Theta, \quad \text{and} \quad \text{var}[g(\underline{\theta})] < \infty;$

(5) $\quad \theta$ is compact, and $\quad p(\theta) < \bar{p} < \infty, \quad I(\underline{\theta}) > \varepsilon > 0, \quad \forall \underline{\theta} \in \Theta.$

Demonstration of (4) involves comparison of the tail behaviors of $p(\underline{\theta})$ and $I(\underline{\theta})$. Demonstration of (5) is generally simple, although Θ may be compact only after a nontrivial reparameterization of the prior and the likelihood. Meeting these conditions does not establish the reliability of $\bar{g}_n$ and $\hat{\sigma}_n^2$ in any practical sense: examples in Section 5 demonstrate uncontrived cases in which (5) applies but $\hat{\sigma}_n^2$ would be unreliable even for $n \approx 10^{20}$. The strength of these conditions is, rather, that they provide a starting point to use numerical methods in constructing $I(\underline{\theta})$. We shall return to this endeavor in Section 4.

The expression for σ^2 in Theorem 2 indicates that the numerical standard error is adversely affected by large $\text{var}[g(\underline{\theta})]$, and by large relative values of the weight function. The former is inherent in the function of interest and in the posterior density, but the latter can in principle be controlled through the choice of the importance sampling density. A simple benchmark for comparing the adequacy of importance sampling densities is the numerical standard error that would result if the importance sampling density were the posterior density itself, i.e., $I^*(\underline{\theta}) \propto p(\underline{\theta})$. In this case $\sigma^2 = \text{var}[g(\underline{\theta})]$ and the number of replications controls the numerical standard error relative to the posterior standard deviation of the function of interest—e.g., $n = 10{,}000$ implies the former will be one percent of the latter. This is a very appealing metric for numerical accuracy.

Only in special cases is it possible or practical to construct synthetic variates whose distribution is the same as the posterior. But since $\text{var}[g(\underline{\theta})]$ can be computed as a routine byproduct of Monte Carlo integration, it is possible to see what the numerical variance would have been, had it been possible to generate synthetic random vectors directly from the posterior distribution. Define the *relative numerical efficiency* of the importance sampling density for the function of interest $g(\underline{\theta})$,

$$RNE \equiv \text{var}[g(\underline{\theta})]/d^{-1}E\left\{[g(\underline{\theta}) - \bar{g}]^2 w(\underline{\theta})\right\} = \text{var}[g(\underline{\theta})]/\sigma^2.$$

The *RNE* is the ratio of number of replications required to achieve any specified numerical standard error using the importance sampling density $I(\underline{\theta})$, to the number required using the posterior density as the importance sampling density. The numerical standard error for $\bar{g}_n$ is the fraction $(RNE \cdot n)^{-1/2}$ of the posterior standard deviation.

Low values of *RNE* indicate that there exists an importance sampling density (namely, the posterior density itself) that does not have to be tailored specifically

to the function of interest, and provides substantially greater numerical efficiency. Thus they alert the investigator to the possibility that more efficient, yet practical, importance sampling densities might be found. If one is willing to consider importance sampling densities that depend on the function of interest—which will be impractical if expected values of many functions of interest are to be computed—then even greater efficiencies can generally be achieved. A lower bound on σ^2 for all such functions can be expressed.

THEOREM 3: *In addition to Assumptions* 1–4 *suppose that* $md[g(\underline{\theta})]$ *is finite. Then the importance sampling density that minimizes* σ^2 *has kernel* $|g(\underline{\theta}) - \bar{g}|\, p(\theta)$, *and for this choice* $\sigma^2 = \{md[g(\underline{\theta})]\}^2$.

PROOF: From Theorem 2, $\sigma^2 = c^{-2}\int_\Theta K(\underline{\theta})^2/I(\underline{\theta})\, d\underline{\theta}$, when we impose the constraint $d^{-1} = \int_\Theta I(\underline{\theta})\, d\underline{\theta} = 1$ and define $K(\underline{\theta}) = |g(\underline{\theta}) - \bar{g}|\, p(\underline{\theta})$. Following Rubinstein (1981, Theorem 4.3.1; the restriction to bounded Θ is inessential), σ^2 is minimized by $I(\underline{\theta}) \propto |K(\underline{\theta})|$. *Q.E.D.*

Theorem 3 has some interesting direct applications. An important class of cases consists of model evaluation problems: $\underline{\theta} \in \Theta^*$ or $\underline{\theta} \in \overline{\Theta}^*$, $g(\underline{\theta})$ is an indicator function for Θ^*, and so $E[g(\underline{\theta})] = P[\underline{\theta} \in \Theta^*] \equiv p^*$. Then σ^2 is minimized if $I(\underline{\theta}) \propto (1 - p^*)p(\underline{\theta})$ for $\underline{\theta} \in \Theta^*$ and $I(\underline{\theta}) \propto p^* p(\underline{\theta})$ for $\underline{\theta} \in \overline{\Theta}^*$: half the drawings should be made in Θ^* and half in $\overline{\Theta}^*$, and in proportion to the posterior in each case.

It appears impractical to construct the densities with kernel $|g(\underline{\theta}) - \bar{g}|\, p(\underline{\theta})$, in general: a different function would be required for each $g(\underline{\theta})$; a preliminary evaluation of $\bar{g}$ with a less efficient method would be a prerequisite; and methods for sampling from those importance sampling densities would need to be devised. However, Theorem 3 suggests that importance sampling densities with tails thicker than the posterior density (a characteristic of densities with kernel $|g(\underline{\theta}) - \bar{g}|\, p(\underline{\theta})$) might be more efficient than the posterior density itself (and hence have $RNE > 1$) although still suboptimal. This is found in practice with some frequency, and some examples are given in Section 5.

4. CHOOSING THE IMPORTANCE SAMPLING DENSITY

There are two logical steps in choosing the importance sampling density. The first is to determine a class of densities that satisfy the conditions of Theorem 2, usually by using (4) or (5). The second step is to find a density within that class that attains a satisfactory *RNE* for the functions of interest. The first objective can only be achieved analytically. The second is a well-defined problem in its own right, that can be solved numerically.

When Θ is compact the first step will usually be trivial. For example, if the classical regularity conditions for the asymptotic normal distribution of the maximum likelihood estimator $\hat{\underline{\theta}}$ obtain, with $\hat{\underline{\theta}} \dot{\sim} N(\underline{\theta}^*, V)$, then the $N(\hat{\underline{\theta}}, V)$ importance sampling density satisfies (5). If Θ is not compact, it will generally be

necessary to investigate the tail behavior of the likelihood function to verify (4). There are many instances in which $L(\underline{\theta})$ behaves like $k\theta_i^{-q}$ for large values of θ_i and it is not difficult to determine q. Given a multivariate normal prior on $\underline{\theta}$ the tail of the marginal posterior density in θ_i will behave like $\exp(-c\theta_i^2)$ for large values of θ_i and a multivariate normal importance sampling density may be appropriate. On the other hand, if the prior is improper, and flat for one or more parameters, then in these instances no multivariate normal importance sampling density will satisfy Theorem 2. An instructive leading example is the computation of the expected value of a function of the mean of a univariate normal population with unknown mean and variance, using the $N(\hat{\mu}, s^2/n)$ importance sampling density. It is not hard to see that (a) Theorem 2 does not apply; (b) if one computes $\hat{\sigma}_n^2$ and $\hat{\text{var}}(\bar{g}_n) \equiv \sum_{i=1}^n [g(\underline{\theta}_i) - \bar{g}_n]^2 w(\theta_i)/[\sum_{i=1}^n w(\theta_i)]^2$, then the computed *RNE* approaches 0 almost surely as $n \to \infty$; (c) if n is held fixed, computed $RNE \to 1$ as sample size approaches infinity. These results are typical of cases in which the classical asymptotic expansion of the log-likelihood function is valid, and Θ is not compact. They underscore the importance of investigating the tail behavior of the likelihood function analytically rather than numerically.

Once the tail behavior of the likelihood function is characterized a class of importance sampling densities is usually suggested. For example, if the tail is multivariate Student with variance matrix Σ and degrees of freedom ν, a multivariate t importance sampling density with weakly larger variance and weakly smaller degrees of freedom will satisfy Theorem 2. In general let $I(\underline{\theta}; \underline{\alpha})$ be a family of importance sampling densities indexed by the $g \times 1$ vector $\underline{\alpha}$, and let $w(\underline{\theta}; \underline{\alpha}) \equiv p(\underline{\theta})/I(\underline{\theta}; \underline{\alpha})$ be the corresponding family of weight functions. It will not usually be practical to choose a distinct importance sampling density for each function of interest. A reasonable objective is the minimization of $E[w(\underline{\theta}, \underline{\alpha})] \propto \int_\Theta [p(\underline{\theta})^2/I(\underline{\theta}; \underline{\alpha})]\, d\underline{\theta}$. This function appears in the first condition of Theorem 2 and in the denominator of each $E[g(\underline{\theta})]$; it is precisely the function one would minimize if $g(\underline{\theta})$ were an indicator function for Θ^*, $p(\Theta^*) = 1/2$.

Two related classes of importance sampling densities have turned out to be quite useful in our experience with this approach. They involve densities that have not, to our knowledge, been described in the literature. The heuristic idea is to begin with minus the inverse of the Hessian of the log posterior density evaluated at its mode as the variance matrix of a multivariate normal importance sampling density, and shift to the multivariate Student t, if warranted, with degrees of freedom indicated by the tail behavior of the posterior density. If the prior is diffuse, the asymptotic variance of the maximum likelihood estimator may be substituted for minus the inverse of the Hessian of the log posterior density evaluated at the mode. In practice either importance sampling density can be poor, because the Hessian poorly predicts (especially if it underpredicts) the value of the posterior density away from the mode, because the posterior density is substantially asymmetric, or both. Illustrations of these problems are provided in the next two sections. To cope with them we modify the multivariate

normal or Student t, comparing the posterior density and importance sampling density in each direction along each axis, in each case adjusting the importance sampling density to reduce $E[w(\underline{\theta}, \underline{\alpha})]$. With k parameters, there are $2k$ adjustments in the last step, so $\underline{\alpha}$ is a $2k \times 1$ vector.

A little more formally, let $\text{sgn}^+(\ )$ be the indicator function for nonnegative real numbers, and let $\text{sgn}^-(\ ) = 1 - \text{sgn}^+(\ )$. The k-variate split normal density $N^*(\underline{\mu}, T, \underline{q}, \underline{r})$ is readily described by construction of a member $\underline{x}$ of its population:

$$\underline{\varepsilon} \sim N(\underline{0}, I_k);$$

$$\eta_i = [q_i \,\text{sgn}^+(\varepsilon_i) + r_i \,\text{sgn}^-(\varepsilon_i)]\,\varepsilon_i \qquad (i = 1, \ldots, k);$$

$$\underline{x} = \underline{\mu} + T\underline{\eta}.$$

The log-p.d.f. at $\underline{x}$ is (up to an additive constant)

$$-\sum_{i=1}^{n} [\log(q_i)\text{sgn}^+(\varepsilon_i) + \log(r_i)\text{sgn}^-(\varepsilon_i)] - (1/2)\underline{\varepsilon}'\underline{\varepsilon}.$$

In the application here, $\underline{\mu}$ is the posterior mode, and T is a factorization such that the inverse of TT' is the negative of the Hessian of the log posterior density evaluated at the mode. A variate from the multivariate split Student density $t^*(\underline{\mu}, T, \underline{q}, \underline{r}, \nu)$ is constructed the same way, except that $\eta_i = [q_i \,\text{sgn}^+(\varepsilon_i) + r_i \,\text{sgn}^-(\varepsilon_i)]\varepsilon_i(\zeta/\nu)^{-1/2}$, with $\zeta \sim \chi^2(\nu)$. The log-p.d.f. is (up to an additive constant)

$$-\sum_{i=1}^{n} [\log(q_i)\text{sgn}^+(\varepsilon_i) + \log(r_i)\text{sgn}^-(\varepsilon_i)] - [(\nu+k)/2]\log\left(1+\nu^{-1}\underline{\varepsilon}'\underline{\varepsilon}\right).$$

Choosing $\underline{\alpha}' = (\underline{q}', \underline{r}')$ to minimize $E[w(\underline{\theta}; \underline{\alpha})]$ requires numerical integration at each step of the usual algorithms for minimization of a nonlinear function. Solution of this optimization problem with any accuracy would, in most applications, likely be more time consuming than the computation of $E[g(\underline{\theta})]$ itself. A much more efficient procedure can be used to select values for $\underline{q}$ and $\underline{r}$ that, in many applications, produces high RNE's. The intuitive idea is to explore each axis in each direction, to find the slowest rate of decline in the posterior density relative to a univariate normal (or Student) density, and then choose the variance of the normal (or Student) density to match that slowest rate of decline. A little more formally, let $\underline{e}^{(i)}$ be a $k \times 1$ indicator vector, $e_i^{(i)} = \underline{e}^{(i)\prime}\underline{e}^{(i)} = 1$. For the split normal define $f_i(\delta)$ according to

$$\begin{aligned} &p\left(\hat{\underline{\theta}} + \delta T\underline{e}^{(i)}\right)/p\left(\hat{\underline{\theta}}\right) = \exp\left[-\delta^2/2f_i(\delta)^2\right] \\ &\Rightarrow f_i(\delta) = |\delta|\left\{2\left[\log p\left(\hat{\underline{\theta}}\right) - \log p\left(\hat{\underline{\theta}} + \delta T\underline{e}^{(i)}\right)\right]\right\}^{-1/2} \end{aligned} \tag{6}$$

Then take $q_i = \sup_{\delta > 0} f_i(\delta)$ and $r_i = \sup_{\delta < 0} f_i(\delta)$. For the split Student the

procedure is the same except that

$$f_i(\delta) = \nu^{-1/2}|\delta|\left\{\left[p(\hat{\theta})/p\left(\underline{\hat{\theta}} + \delta T\underline{e}^{(i)}\right)\right]^{2/(\nu+k)} - 1\right\}^{-1/2}.$$

In practice, carrying out the evaluation for $\delta = 1/2, 1, \ldots, 6$, seems to be satisfactory.

The two examples that follow illustrate how these methods are applied.

5. EXAMPLES: THE BINOMIAL AND MARKOV CHAINS

In this section we take up a set of successively complex examples in which the parameter space Θ is compact. We begin with situations which can be studied analytically in some detail, and consequently there would be no need for a numerical approach at all. This is strictly for illustrative purposes. As will be seen, not much elaboration is required to generate realistic models in which there are no analytical solutions.

5.1. *A Simple Binomial Model*

Let a sample of size N be drawn from a population in which $P[x = 1] = \theta$, $P[x = 0] = 1 - \theta$, and suppose there are M occurrences of "1". If the prior is $\pi(\theta) = 1$, $0 \leqslant \theta \leqslant 1$, then the posterior is

$$p(\theta) = [B(M+1, N-M+1)]^{-1}\theta^M(1-\theta)^{N-M} \quad (0 \leqslant \theta \leqslant 1). \tag{7}$$

Since $N^{1/2}(\hat{\theta} - \theta) \Rightarrow N(\theta, \theta(1-\theta))$, a normal approximation based on the asymptotic distribution of the m.l.e. $\hat{\theta}$ is

$$I(\theta) = \left[2\pi N^{-1}\hat{\theta}(1-\hat{\theta})\right]^{1/2}\exp\left\{-N(\hat{\theta}-\theta)^2/2\hat{\theta}(1-\hat{\theta})\right\}.$$

Since $0 \leqslant \theta \leqslant 1$ condition (5) is satisfied.

Consider two cases. In Case A, $N = 69$, $M = 6$, $\hat{\theta} = .087$, the asymptotic standard error for $\hat{\theta}$ is .034 and the asymptotic variance is .00115. In Case B, $N = 71$, $M = 54$, $\hat{\theta} = .761$, the asymptotic standard error for $\hat{\theta}$ is .051, and the asymptotic variance is .00256. For the corresponding normal importance sampling densities Figure 1 shows the weight functions $w(\theta) = p(\theta)/I(\theta)$ indicated by the heavy lines. In each case $w(\theta)$ appears badly behaved, attaining a value of over 10^{70} for θ around .9 in Case A, and attaining a value of over 10^4 for θ between .2 and .3 in Case B. For our purposes it is $E[w(\theta)] = \int_\Theta w(\theta)p(\theta)\,d\theta$ and not $w(\theta)$ which matters. The light lines in Figure 1 show $w(\theta)p(\theta)$, which is much less than $w(\theta)$ when $w(\theta)$ is large. In Case A $E[w(\theta)] = 1.65 \times 10^{24}$, and in Case B $E[w(\theta)] = 1.05$. Clearly the importance sampling density $I(\theta)$ is adequate in the latter case but not the former. The difficulty in Case A is that the likelihood function declines more slowly than its asymptotic normal approximation, for $\theta > \hat{\theta}$. A split normal importance sampling density with variance .00115 for $\theta < \hat{\theta}$ and variance greater than .00115 for $\theta > \hat{\theta}$ performs better, as shown in Figure 2. As the latter variance increases, $E[w(\theta)]$ decreases exponentially, until the variance reaches .0016; $E[w(\theta)]$ attains its minimum at .0019, and then increases slowly.

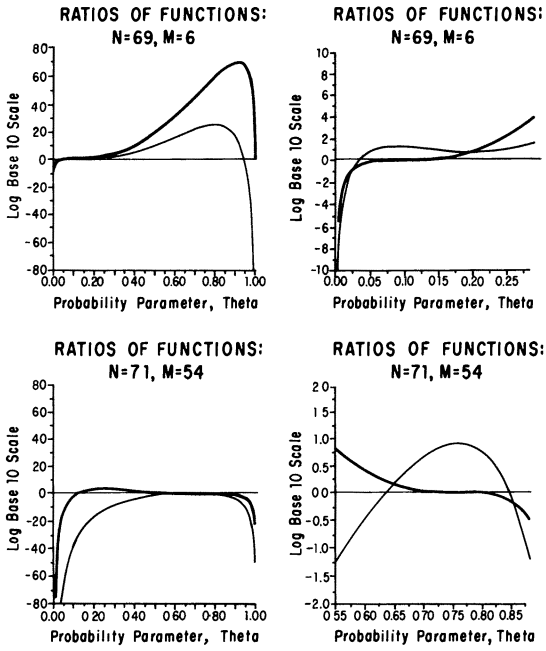

FIGURE 1.—Values of $\log(w(\theta))$ (heavy lines) and $\log(w(\theta)p(\theta))$ (light lines) for the two binomial models discussed in Section 5.1.

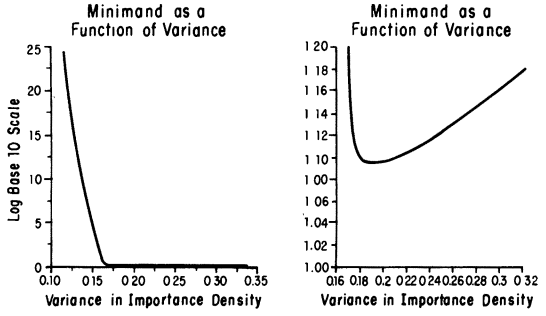

FIGURE 2.—Values of $E(w(\theta))$ as a function of the variance of a split normal importance sampling density, centered at $\hat{\theta}$, for Case A discussed in Section 5.1.

5.2. *The Two-State Homogeneous Markov Chain Model*[2]

Suppose an agent can occupy one of two states, with $P[\text{In state } i \text{ at time } t | \text{In state } j \text{ at time } t-1] = p_j$, $i \neq j$. Let N agents be observed at times $t-1$ and t, with m_{ij} agents in state i at time $t-1$ and state j at time t. With a prior density $\pi(p_1, p_2) = 1$, $0 < p_i < 1$ $(i = 1, 2)$, the posterior p.d.f. is proportional to

$$p(p_1, p_2) = p_1^{m_{12}} p_2^{m_{21}} (1-p_1)^{m_{11}} (1-p_2)^{m_{22}}, \tag{8}$$

[2]For elaboration on the Markov chain model and the problem of embeddability, see Singer and Spilerman (1976) or Geweke, Marshall, and Zarkin (1986a, 1986b).

TABLE I

THREE EXAMPLES FOR THE TWO-STATE HOMOGENEOUS MARKOV CHAIN MODEL[a]

	Case I	Case II	Case III
m_{11}	63	21	68
m_{12}	6	66	28
m_{21}	17	6	17
m_{22}	54	24	4
$\hat{p}_1$	.087 (.034)	.759 (.046)	.292 (.046)
$\hat{p}_2$	.239 (.051)	.200 (.073)	.810 (.086)
$\hat{p}_1 + \hat{p}_2$	.326 (.061)	.959 (.086)	1.102 (.097)

[a] Carats denote m.l.e.'s; asymptotic standard errors are shown parenthetically. Data are taken from Singer and Cohen (1980).

$0 < p_i < 1$ $(i = 1,2)$. If the underlying process of transition takes place continuously and homogeneously through time, it may be described in terms of $\lim_{\delta \to 0} \delta^{-1} P[\text{In state } i \text{ at time } t + \delta | \text{In state } j \text{ at time } t] = r_j$, $i \neq j$. If $p_1 + p_2 < 1$, then $r_j = p_j \log(1 - p_1 - p_2)/(p_1 + p_2)$; otherwise the process of transition cannot be continuous and homogeneous through time. Whenever there exists a continuous time model corresponding to a discrete time Markov chain model, the discrete time model is said to be embeddable.

To provide specific numerical examples, we use part of a data set reported by Singer and Cohen (1980). The data pertain to the incidence of malaria in Nigeria, state 1 being no detectable parasitemia and state 2 being a positive test for parasitemia. Data are reported separately for different demographic groups. The process of infection and recovery is modeled as a homogeneous Markov chain, and Singer and Cohen (1980) focus on whether this model is embeddable. Data for three cases, and the corresponding maximum likelihood estimates and their asymptotic standard errors are provided in Table I. Notice that the Case I data are those used to illustrate the binomial model; a two-state Markov chain model is simply a combination of two binomial models, as reflected in comparison of (7) and (8). In Case I $\hat{p}_1 \hat{p}_2 \ll 1$, and the model is embeddable; in Case II $\hat{p}_1 + \hat{p}_2 < 1$ and in Case III $\hat{p}_1 + \hat{p}_2 > 1$, and the question of embeddability is open.

We conduct inference in this model with two priors and six functions of interest. The first prior is uninformative: $\pi(p_1, p_2) = 1$ for all $p_j \in (0,1)$. The second prior imposes embeddability but otherwise is the same as the first: $\pi(p_1, p_2) = 2$ for all (p_i, p_j): $p_i > 0$, $p_j > 0$, $p_1 + p_2 < 1$. The functions of interest are p_1, p_2, p_1^{-1}, p_2^{-1}, r_1^{-1}, and r_2^{-1}. The function p_j^{-1} is the mean duration in state j in the discrete time model, while r_j^{-1} is the mean duration in state j assuming a continuous time model. Since r_j^{-1} is undefined if $p_1 + p_2 \geq 1$, $E(r_j^{-1})$ is defined only under the second prior; $E[r_j]$ is infinite for either prior, although the posterior distribution of r_j may be obtained readily as suggested in examples (iii) and (iv) of Section 2.

Results obtained using the methods proposed in earlier sections are presented in Table II. Asterisks in column 1 denote the use of the second prior; otherwise the first is used. The first block of columns provides results with the normal

TABLE II

INFERENCE IN THE TWO-STATE HOMOGENEOUS MARKOV CHAIN MODEL[a]

Importance Density $g(\ )$[b]	Normal $E[g]$	Normal $sd[g]$	Normal $100\hat{\delta}_n$	Normal RNE[c]	Split Normal $E[g]$	Split Normal $sd[g]$	Split Normal $100\hat{\delta}_n$	Split Normal RNE[c]	Likelihood Function $E[g]$	Likelihood Function $sd[g]$	Likelihood Function $100\hat{\delta}_n$	Likelihood Function RNE[c]	Actual $E[g]$	Actual $sd[g]$
Case I														
p_1	.098	.034	.051	.441	.099	.035	.033	1.13	.099	.035	.035	1.00	.099	.035
p_2	.246	.049	.056	.769	.247	.051	.050	1.01	.248	.050	.050	1.00	.247	.050
p_1^{-1}	11.7	4.98	5.09	.960	11.7	4.94	4.15	1.41	11.6	5.37	5.37	1.00	11.7	4.98
p_2^{-1}	4.25	.926	.970	.911	4.24	.934	.891	1.10	4.21	.910	.910	1.00	4.24	.925
r_1^{-1*}	9.60	4.35	4.44	.961	9.58	4.32	3.62	1.42	9.55	4.68	4.68	1.00	—	—
r_2^{-1*}	3.49	.887	.932	.905	3.48	.897	.854	1.10	3.45	.874	.874	1.00	—	—
Case II														
p_1	.753	.046	.051	.808	.753	.045	.044	1.05	.752	.046	.046	1.00	.753	.046
p_2	.218	.070	.083	.720	.219	.072	.070	1.05	.220	.072	.072	1.00	.219	.072
p_1^{-1}	1.33	.083	.096	.759	1.33	.083	.081	1.05	1.33	.084	.084	1.00	1.33	.083
p_2^{-1}	5.18	2.09	1.92	1.19	5.16	2.07	1.77	1.37	5.15	2.06	2.06	1.00	5.17	2.08
p_1^*	.739	.043	.060	.526	.740	.043	.052	.686	.739	.044	.054	.647	—	—
p_2^*	.183	.050	.064	.629	.182	.049	.060	.672	.183	.050	.063	.647	—	—
p_1^{-1*}	1.36	.082	.117	.491	1.36	.081	.098	.681	1.36	.082	.103	.647	—	—
p_2^{-1*}	6.00	2.13	2.21	.928	6.01	2.09	2.18	.921	5.97	2.07	2.58	.647	—	—
r_1^{-1*}	.472	.129	.170	.576	.472	.130	.163	.642	.471	.132	.164	.647	—	—
r_2^{-1*}	2.18	1.23	1.30	.891	2.18	1.22	1.30	.880	2.17	1.22	1.51	.647	—	—
Case III														
p_1	.296	.045	.051	.810	.296	.046	.046	1.01	.297	.046	.046	1.00	.296	.046
p_2	.782	.084	.118	.507	.784	.083	.082	1.03	.781	.085	.085	1.00	.783	.084
p_1^{-1}	3.46	.558	.583	.915	3.46	.565	.547	1.06	3.46	.556	.556	1.00	3.46	.562
p_2^{-1}	1.29	.153	.246	.387	1.29	.152	.152	.998	1.30	.156	.156	1.00	1.29	.154
p_1^*	.269	.041	.145	.080	.268	.041	.092	.201	.270	.041	.089	.208	—	—
p_2^*	.669	.062	.257	.057	.670	.060	.138	.189	.668	.062	.136	.208	—	—
p_1^{-1*}	3.80	.600	1.82	.108	3.82	.615	1.32	.217	3.80	.595	1.30	.208	—	—
p_2^{-1*}	1.51	.150	.686	.048	1.51	.147	.340	.185	1.51	.152	.333	.208	—	—
r_1^{-1*}	1.21	.394	.116	.115	1.22	.419	0.93	.206	1.21	.408	.893	.208	—	—
r_2^{-1*}	.488	.170	.698	.060	.488	.173	.401	.185	.490	.179	.391	.208	—	—

[a] Expected values of functions of interest $g(\ )$ were computed by Monte Carlo integration with importance sampling, using 10,000 replications.

[b] Prior is $\pi(p_1, p_2) = 1$ if no asterisk: with asterisk, prior is $\pi(p_1, p_2) = 2$ if $p_1 + p_2 < 1$. In Case I, essentially all the mass of the posterior satisfies $p_1 + p_2 < 1$.

[c] Relative numerical efficiency.

importance sampling density motivated by the asymptotic distribution: p_1 and p_2 are independent, $p_j \sim N(\hat{p}_j, \hat{p}_j(1-\hat{p}_j)/(m_{j1}+m_{j2}))$. The second block of columns provides results with the split normal importance sampling density constructed as described in Section 4. Since the posterior density is the product of a binomial in p_1 and a binomial in p_2, each p_j may be expressed as a ratio involving two independent chi-square random variables (Johnson and Kotz (1970, Section 24.2)). Hence it is straightforward to use the likelihood function itself as the importance sampling density; the third block of columns provides results obtained this way. Finally, the posterior moments of p_j and p_j^{-1} have simple analytical expressions under the first prior, and these are provided in the last block of columns. (Under the second prior the posterior moments of p_j and

p_j^{-1} could probably be computed more efficiently by numerical evaluation of the incomplete beta, but this has not been pursued because it does not extend to multi-state Markov chain models.)

In Case I, $P[p_1 + p_2 < 1] \doteq 1$ under the first prior; none of the 10,000 replications with any importance sampling density turned up $p_1 + p_2 > 1$. In Case II, $P[p_1 + p_2 < 1] \doteq .65$ and in Case III $P[p_1 + p_2 < 1] \doteq .21$, under the first prior. The computed expected values of all functions of interest with each importance sampling density are quite similar to each other and (where available) to the actual values. There are no surprises given the computed numerical standard errors. There are greater relative differences among the computed standard deviations of the functions of interest, and it may be verified analytically that this must be the case. One could, of course, compute numerical standard errors for the posterior standard deviations as well. The relative numerical efficiencies indicate that, with a few small exceptions in Case II, the split normal is a better importance sampling density than the normal. The superiority of the split normal increases systematically as $E(p_j)$ approaches zero and the asymptotic normal becomes a poorer approximation to the likelihood function. Many examples of relative numerical efficiencies in excess of 1.0 turn up with the split normal importance sampling density, precisely where we would expect: $E(p_j)$ closer to zero, and moments of the p_j showing greater dispersion, like p_j^{-1}. By construction RNE is 1.0 for importance sampling from the posterior density, and for importance sampling from the likelihood function it is the posterior probability of an imposed prior, evaluated assuming a prior that is uniform in the parameter space of the likelihood function.

TABLE III

SOME DIAGNOSTICS FOR COMPUTATIONAL ACCURACY IN THE TWO-STATE HOMOGENEOUS MARKOV CHAIN MODEL

	Normal Sampling Density		Split Normal Sampling Density	
	$n = 10{,}000$	$n = 50{,}000$	$n = 10{,}000$	$n = 50{,}000$
Case I				
RNE, p_1	.441	.269	1.137	1.139
RNE, p_2	.769	.657	1.014	1.012
ω_1	186.2	1,774.7	2.5	2.5
ω_{10}	72.9	497.0	2.5	2.5
Case II				
RNE, p_1	.808	.774	1.054	1.047
RNE, p_2	.720	.564	1.054	1.053
ω_1	31.2	277.9	1.9	1.9
ω_{10}	17.5	106.3	1.9	1.9
Case III				
RNE, p_1	.810	.790	1.011	1.009
RNE, p_2	.507	.462	1.032	1.030
ω_1	101.0	348.7	1.8	1.8
ω_{10}	52.8	161.3	1.8	1.8

From the example of Section 5.1, we know that the population *RNE* for the normal importance sampling density in Case I should be on the order of 10^{-24}. The normal is such a bad importance sampling density that, even with 10^4 replications, it will almost certainly not produce the values of the p_j that lead to difficulties, and computed *RNE* will bear no resemblance to population *RNE*. As the number of replications increases computed *RNE*'s tend to decline in situations like this, as is clear from consideration of Figure 1. Table III contrasts some computed *RNE*'s for $n = 10{,}000$ and $n = 50{,}000$ and bears out the difficulties with Case I. In these situations the largest weights $w(\underline{\theta}_i)$ increase rapidly as the number of replications, n, increases, and order statistics pertaining to the $w(\underline{\theta}_i)$ are often useful in identifying very poorly conditioned importance sampling densities. Since the largest $w(\underline{\theta}_i)$'s may be retained systematically as n increases and $\Sigma_{i=1}^{n} w(\underline{\theta}_i)$ and $\Sigma_{i=1}^{n} w(\underline{\theta}_i)^2$ are computed in any event, the diagnostic of the form $\omega_m = (n/m)\Sigma_{i=1}^{m} w^{(n+1-i)}(\underline{\theta})^2 / \Sigma_{i=1}^{n} w(\underline{\theta}_i)^2$ (where superscripts denote order statistics and m is small) can be quite useful as an indication of thin tails in the importance sampling density relative to the posterior. The values of these diagnostics for our examples are provided in Table III, again for $n = 10{,}000$ and $n = 50{,}000$; they indicate—correctly—that the split normal is a much better importance sampling density than is the normal.

5.3. *The Quartile Homogeneous Markov Chain Model*[3]

Consider last a four-state homogeneous Markov chain model, state j denoting the jth quartile of income. Let $p_{ij} \equiv P[\text{In state } j \text{ at time } t | \text{In state } i \text{ at time } t-1]$, and arrange the p_{ij} in a 4×4 matrix P. Since $\Sigma_{i=1}^{4} p_{ij} = \Sigma_{i=1}^{4} p_{ji} = 1$ for $j = 1, \ldots, 4$, the likelihood function has 9 free parameters and is not proportional to a product of multinomial densities. We suppose that the process of transition between income quartiles takes place in continuous time; define

$$r_{ij} = \lim_{\delta \to 0} \delta^{-1} P[\text{In state } j \text{ at time } t+\delta | \text{In state } i \text{ at time } t]$$

for all $i \neq j$, and let $r_{ii} = -\Sigma_{j \neq i} r_{ij}$. If the discrete time model is embeddable, then $R = Q \log[\Lambda] Q^{-1}$, where Q is the matrix whose columns are right eigenvectors of P and Λ is the diagonal matrix of corresponding eigenvalues. A given discrete time model is embeddable if and only if the off-diagonal elements of $Q \log[\Lambda] Q^{-1}$ are all real and nonnegative and the diagonal elements are all real and nonpositive for some combination of the branches of the logarithms of the eigenvalues. The problems of inference subject to the constraint of embeddability is ill suited to classical treatment, and a purely analytical Bayesian approach is precluded by the complexity of the indicator function for embeddability.

[3]Details on the construction of the normal importance sampling density for this example may be found in Geweke, Marshall, and Zarkin (1986a). Interest in this problem in general and in measures of mobility in particular is motivated by Geweke, Marshall, and Zarkin (1986b).

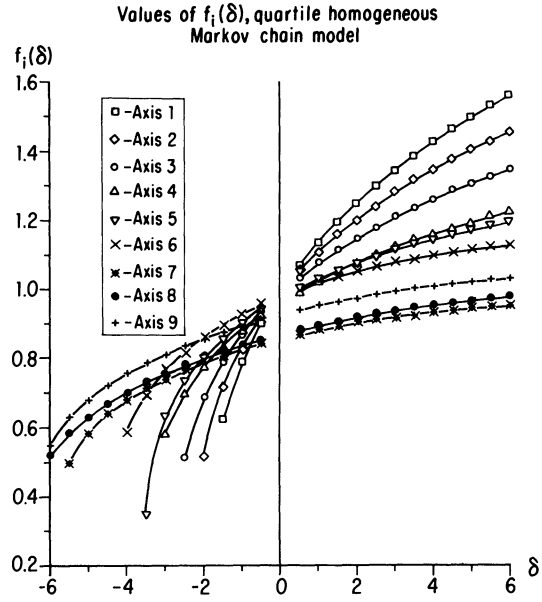

FIGURE 3.—Values of $f_i(\delta)$ along nine orthogonal axes of the likelihood function, constructed as described in Section 5.3.

The 9×1 vector of parameters for the likelihood function consists of those p_{ij} for which $j \neq i$ and $j \neq i+1$, ordered so that the $\hat{p}_{ij}$ appear in ascending order.[4] The split normal importance sampling density was constructed as described in Section 4, with T being the lower triangular Choleski factorization of the aymptotic variance matrix of $\underline{\theta}$. Some idea of the behavior of the likelihood surface is conveyed in Figure 3, which shows values of the $f_i(\delta)$ defined in (6). For the lower-numbered axes, the boundary of the parameter space is only a few standard deviations in the negative direction, and so $f_i(\delta)$ is truncated on the left. For these axes $f_i'(\delta)$ is greater for $\delta > 0$ than it is for the higher-numbered axes, indicating that the likelihood function tapers most slowly relative to the asymptotic normal approximation in these directions. This is not surprising in view of the example in Section 5.1.

The functions of interest are $-r_{jj}^{-1}$, mean duration in each state, r_{jj}, rate of transition to and from each state, and three summary measures of mobility that have been proposed in the literature: $M^*(R) = -.25 \operatorname{tr}(R)$, $M_B^*(R) = .25\Sigma_{i=1}^4 \Sigma_{j=1}^4 |i-j| r_{ij}$, and $M_2^*(R) = -\log(\lambda_2)$ where λ_2 is the second largest eigenvalue of P. We impose the prior $\pi(P) \propto 1$ if P is embeddable and $\pi(P) = 0$ otherwise. Under this prior posterior means of the r_{jj}^{-1}, but not the r_{jj} or the mobility measures, exist. The data set was extracted from the National Longitudinal Survey data file, consisting of those 460 white males who were not enrolled

[4]For details, see the Appendix of Geweke, Marshall, and Zarkin (1986a).

in school, employed full time, and married with spouse present during 1968, 1969, and 1970; and whose family incomes were coded for 1969, 1970, and 1971.

Monte Carlo integration with 10,000 replications, using the normal and split normal importance sampling densities was carried out. The normal importance sampling density performed very poorly: $\omega_1 = 2{,}829.6$. The corresponding value for the split normal was 22.1. The posterior probability of embeddability is .246 (with a numerical standard error of .0009, computed using the split normal importance sampling density). This probability, rather than 1.0, provides the norm against which *RNE*'s should be compared, for any importance sampling density based on the likelihood function rather than the posterior.

The computed values in Table IV are essentially the same, for expected values and medians of functions of interest, with the two importance sampling densities employed. Higher moments, and quantile points farther from the median, show greater differences. The computed *RNE*'s suggest that the split normal is substantially more efficient than the normal. However, for the normal importance sampling density computed *RNE* tends to vary widely with each set of 10,000 replications, and deteriorates as the number of replications increases. In this application, and—we conjecture—in most applications with more than a few dimensions, careful modification of asymptotic normal and other symmetric distributions is required if computed numerical standard errors are to be interpreted reliably using Theorem 2.

TABLE IV

INFERENCE IN THE QUARTILE HOMOGENEOUS MARKOV CHAIN MODEL[a]

Importance Density $g(\ )$	Split Normal $E[g]$	$sd[g]$	$100\hat{\sigma}_n$	RNE[b]	Normal $E[g]$	$sd[g]$	$100\hat{\sigma}_n$	RNE[b]
$-r_{11}^{-1}$	2.35	.338	.786	.184	2.36	.341	.919	.138
$-r_{22}^{-1}$	1.33	.170	.393	.187	1.34	.170	.470	.131
$-r_{33}^{-1}$	1.33	.162	.380	.182	1.33	.169	.546	.095
$-r_{44}^{-1}$	2.28	.324	.786	.169	2.28	.323	1.056	.094

Importance Density $g(\ )$	Split Normal Quantile .01	.25	.50	.75	.99	Normal Quantile .01	.25	.50	.75	.99
$-r_{11}^{-1}$	1.68	2.10	2.33	2.56	3.28	1.68	2.12	2.32	2.56	3.21
$-r_{22}^{-1}$	.987	1.22	1.32	1.45	1.76	.965	1.21	1.32	1.44	1.76
$-r_{33}^{-1}$	.979	1.21	1.32	1.43	1.74	.973	1.21	1.32	1.44	1.75
$-r_{44}^{-1}$	1.60	2.05	2.26	2.48	3.13	1.65	2.05	2.25	2.48	3.14
$-r_{11}$	.304	.391	.428	.477	.594	.311	.390	.430	.472	.589
$-r_{22}$	.566	.692	.759	.821	1.01	.566	.691	.755	.825	1.03
$-r_{33}$	.572	.698	.760	.823	1.02	.570	.694	.757	.823	1.02
$-r_{44}$	.319	.403	.443	.488	.621	.318	.402	.444	.488	.598
$M^*(R)$	.485	.565	.598	.637	.741	.487	.563	.597	.638	.745
$M_B^*(R)$	.586	.675	.718	.760	.881	.586	.676	.716	.759	.864
$M_2^*(R)$	.317	.435	.810	1.03	1.48	.316	.418	.795	1.00	1.49

[a] Expected values, standard deviations, and fractiles of functions of interest $g(\ )$ were computed by Monte Carlo integration with importance sampling from a split normal density, using 10,000 replications.

[b] Relative numerical efficiency.

6. EXAMPLES: ARCH LINEAR MODELS

In this section we take up an example in which the parameter space Θ is not compact. It involves constraints which cannot readily be imposed in classical inference, as discussed in detail elsewhere (Geweke (1988a)). For this, among other reasons, exact Bayesian inference is appealing in this model.

The ARCH linear model was proposed by Engle (1982) to allow persistence in conditional variance in economic time series. Let $\underline{x}_t$: $k \times 1$ and y_t be time series for which the distribution of y_t conditional on $\psi_{t-1} = \{\underline{x}_{t-s}, y_{t-s}, s \geqslant 1\}$ is

$$y_t|\psi_{t-1} \sim N(\underline{x}_t'\underline{\beta}, h_t). \tag{9}$$

Defining $\varepsilon_t = y_t - \underline{x}_t'\underline{\beta}$, take

$$h_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j}^2 = h(\varepsilon_{t-1}, \ldots, \varepsilon_{t-p}; \underline{\gamma}_{q\times 1}).$$

The parameterization in terms of $\underline{\gamma}$ allows restrictions like $\alpha_0 = \gamma_0$, $\alpha_j = \gamma_1(p+1-j)$, the linearly declining weights employed by Engle (1982, 1983). For (9) to be plausible it is necessary that $\alpha_0 > 0$ and $\alpha_j \geqslant 0$ $(j = 1, \ldots, p)$. For $\{\varepsilon_t\}$ to be stationary it is necessary and sufficient that the roots of $1 - \sum_{j=1}^{p} \alpha_j z^j$ all be outside the unit circle.

Given the sample $(\underline{x}_t, y_t; t = 1, \ldots, T)$ the log-likelihood function is (up to an additive constant)

$$l = -(1/2) \sum_{t=p+1}^{T} \log h_t - (1/2) \sum_{t=p+1}^{T} h_t^{-1} \varepsilon_t^2. \tag{10}$$

Engle (1982) has shown that for large samples $\partial^2 l / \partial \underline{\beta}\, \partial \underline{\gamma}' \doteq 0$, and the maximum likelihood estimates $\hat{\underline{\beta}}$ of $\underline{\beta}$ and $\hat{\underline{\gamma}}$ of $\underline{\gamma}$ are asymptotically independent. A scoring algorithm may be used to maximize the likelihood function, and revision of the estimate $\underline{\gamma}^{(i)}$ of $\underline{\gamma}$ and $\underline{\beta}^{(i)}$ of $\underline{\beta}$ may proceed separately at each step. We programmed the procedure described in Engle (1982) and then replicated other results (Engle (1983)) using data furnished by the *Journal of Money, Credit, and Banking*.

Since $\hat{\underline{\beta}}$ and $\hat{\underline{\gamma}}$ are asymptotically independent we construct the importance sampling density as the product of a density for $\underline{\beta}$ and a density for $\underline{\gamma}$. To employ condition (4) and the procedures of Section 4 first compare the tail behavior of the multivariate t density in m dimensions,

$$f(\underline{x}; \underline{\mu}, \Sigma, \nu) \propto \left[1 + \nu^{-1}(\underline{x} - \underline{\mu})'\Sigma^{-1}(\underline{x} - \underline{\mu})\right]^{-(\nu+m)/2},$$

with that of the likelihood function. Allowing $|x_j| \to \infty$ while fixing all other x_i, the log of $f(\underline{x}; \underline{\mu}, \Sigma, \nu)$ behaves like $-(\nu + m)\log|x_j|$. The log likelihood function (10) behaves like $-[(T-p)/2]\log(\gamma_j)$ as $\gamma_j \to \infty$ with all other parameters fixed, and like $-(T-p)\log|\beta_j|$ as $|\beta_j| \to \infty$ with all other parameters fixed. Hence for $\underline{\beta}$ we shall use a split Student importance sampling density with $T - p - k$ degrees

TABLE V

INFERENCE IN THE ARCH LINEAR MODEL[a]

$g(\ )$	$E[g]$	$sd[g]$	$100\delta_n$	RNE[b]
β_1	1.096	.092	.110	.700
β_2	.954	.135	.155	.757
γ_0	1.001	.291	.373	.609
γ_1	.281	.066	.079	.693
P[Stable]	.795	—	.475	.722

[a] Expected values and standard deviations of functions of interest $g(\ )$ were computed by Monte Carlo integration with importance sampling from a split Student density, using 10,000 replications.
[b] Relative numerical efficiency.

of freedom, and for $\underline{\gamma}$ we shall use a split Student importance sampling density with $(T-p)/2-q$ degrees of freedom. The variance matrices are given by expressions (28) and (32), respectively, in Engle (1982).

An artificial sample of size 200 provides a concrete example:

$$y_t = \sum_{j=1}^{2} \beta_j x_{tj} + \varepsilon_t, \qquad \beta_1 = \beta_2 = 1;$$

$$x_{t1} = 1, \qquad x_{t2} = \cos(2\pi t/200);$$

$$\varepsilon_t \sim N(0, h_t), \qquad h_t = 1 + .5\varepsilon_{t-1}^2 + .25\varepsilon_{t-2}^2.$$

The parameterization of the conditional variance process in the model is $h_t = \gamma_0 + \gamma_1(2\varepsilon_{t-1}^2 + \varepsilon_{t-2}^2)$; $\alpha_0 = \gamma_0$, $\alpha_1 = 2\gamma_1$, $\alpha_2 = \gamma_1$. We assume a flat prior on $\underline{\beta}$ and $\underline{\gamma}$, and estimate these four parameters and assess the posterior probability of stability $P[|\gamma_1| < 1/3]$. Results using the split Student importance sampling density constructed as described in Section 4 are presented in Table V.

The local behavior of the log-likelihood function is portrayed in Figure 4. For each of the four axes this figure indicates the actual log-likelihood function (solid thick line), the multivariate "t" approximation to the likelihood function using the information matrix (Engle (1982, (28) and (32))) (dotted line), and the split Student approximation to the likelihood function constructed as described in Section 4 (thin line). (In the case of $\underline{\gamma}$, the curves terminate at $\gamma_0 = 0$ or $\gamma_1 = 0$.) For $\underline{\beta}$ the classical asymptotic theory provides a good approximation to the likelihood function, but for $\underline{\gamma}$ the approximation is poor.

In Table VI we compare diagnostics for computational accuracy using the unmodified and split normal densities (which are unjustified theoretically since $E[w(\underline{\theta})]$ is infinite in these cases) and the unmodified and split Student densities. The unmodified densities perform quite poorly, for reasons made clear in Figure 4. The split normal and split Student sampling densities lead to acceptable performance, in the sense that RNE is apparently of the same order of magnitude as would be achieved if one could sample directly from the likelihood function. In the case of the split normal we know this is an illustration, but from

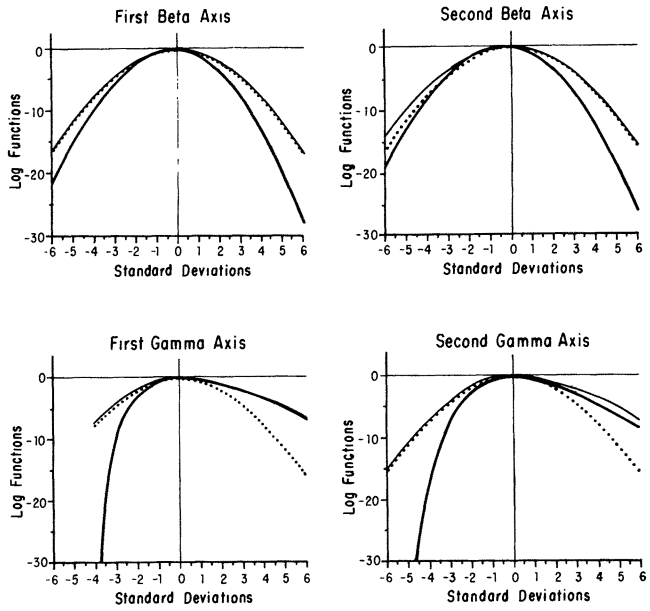

FIGURE 4.—All functions are normalized to have the value 1 at 0 standard deviations, which corresponds to the asymptotic maximum likelihood estimator. The solid thick line is the log likelihood function; the dotted line is the log multivariate t importance sampling density; and the solid thin line is the log split Student importance sampling density.

TABLE VI

SOME DIAGNOSTICS FOR COMPUTATIONAL ACCURACY
ARCH LINEAR MODEL

	Multivariate Normal Sampling Density		Split Normal Sampling Density	
	$n = 10{,}000$	$n = 50{,}000$	$n = 10{,}000$	$n = 50{,}000$
RNE, β_1	.006	.016	.686	.666
RNE, β_2	.005	.014	.682	.708
RNE, γ_0	.001	.002	.564	.529
RNE, γ_1	.006	.012	.748	.714
RNE, p	.018	.044	.744	.738
ω_1	9,872.8	39,893.6	51.0	225.4
ω_{10}	994.4	4,857.2	22.3	66.0
	Multivariate t (99) Sampling Density		Split Student Sampling Density	
	$n = 10{,}000$	$n = 50{,}000$	$n = 10{,}000$	$n = 50{,}000$
RNE, β_1	.036	.068	.700	.707
RNE, β_2	.027	.065	.757	.742
RNE, γ_0	.002	.006	.609	.501
RNE, γ_1	.017	.041	.692	.690
RNE, p	.049	.118	.722	.741
ω_1	9,726.9	37,067.8	19.9	232.3
ω_{10}	979.6	4,476.4	16.0	70.9

TABLE VII

SOME DIAGNOSTICS FOR COMPUTATIONAL ACCURACY
ARCH LINEAR MODEL

Multivariate Sampling Density with Indicated Degrees of Freedom $n = 10{,}000$				
	$\nu = 0.5$	$\nu = 1.0$	$\nu = 1.5$	$\nu = 2.0$
RNE, β_1	.144	.302	.402	.451
RNE, β_2	.150	.307	.410	.455
RNE, γ_0	.127	.243	.303	.313
RNE, γ_1	.129	.262	.339	.365
RNE, p	.127	.282	.368	.417
ω_1	52.2	42.2	41.2	40.2
ω_{10}	47.2	32.3	29.8	34.2
	$\nu = 3.0$	$\nu = 4.0$	$\nu = 6.0$	$\nu = 12.0$
RNE, β_1	.536	.529	.487	.356
RNE, β_2	.540	.525	.537	.430
RNE, γ_0	.336	.317	.271	.197
RNE, γ_1	.410	.399	.366	.292
RNE, p	.493	.504	.522	.495
ω_1	53.4	69.0	136.3	321.4
ω_{10}	34.2	42.7	71.2	130.0

Figure 4 astronomical values of n are required before *RNE* is mainly determined by the relative tail behavior of the likelihood function and importance sampling density. Diagnostics for the split Student and split normal importance sampling densities are essentially the same.

The alternative approach to Monte Carlo integration with importance sampling for posterior densities with tails thick relative to the normal has been to employ multivariate t distributions with small degrees of freedom (Zellner and Rossi (1984), van Dijk, Hop, and Louter (1986)). We conclude this example with an examination of diagnostics for computational accuracy for this procedure. Beginning with the asymptotic variance matrices for $\underline{\beta}$ and $\underline{\gamma}$, multivariate importance sampling densities with $\nu = .5$, 1, 1.5, 2, 3, 4, 6, and 12 degrees of freedom were used. The resulting diagnostics for numerical accuracy are given in Table VII. *RNE* for the four parameters attains a maximum at $\nu = 3$, that for p at $\nu = 6$. For $\nu < 3$ the sampling density is too diffuse, and as ν becomes larger it approaches the multivariate normal which is much too compact. That ω-diagnostics are sensitive to sampling densities with tails that are too thin rather than too thick is clearly indicated in Table VII. The *RNE* of the multivariate t with low degrees of freedom is decidedly less than that of the split Student t, but of the same order of magnitude for proper choice of ν. In the absence of any systematic or theoretical basis for choosing ν, however, costly numerical experimentation is required to find the appropriate value.

7. CONCLUSION

Integration by Monte Carlo with importance sampling provides a paradigm for Bayesian inference in econometric models. Thanks to increasingly cheap comput-

ing it is practical to obtain the exact posterior distributions of any function of interest of the parameters, subject to any diffuse prior. This is done through systematic exploration of the likelihood function, and the asymptotic sampling distribution theory for maximum likelihood estimators provides the organization for this exploration. Diffuse priors may incorporate inequality restrictions which arise frequently in applied work but are impractical if not impossible to handle in a classical setting. Concentrations of prior probability mass may be treated by applying the methods in this paper to both the original and lower dimensional problems, and combining results through the appropriate posterior odds ratios. By choosing functions of interest appropriately, formal and direct answers to the questions that motivate empirical work can be obtained. The investigator can routinely handle problems in inference that would be analytical nightmares if attacked from a sampling theoretic point of view, and is left free to craft his models and functions more according to substantive questions and less subject to the restrictions of what asymptotic distribution theory can provide.

Integration by Monte Carlo is an attractive research tool because it makes numerical problems much more routine than do other numerical integration methods. The principle analytical task left to the econometrician is the characterization of the tail behavior of the likelihood function. This characterization leads to a family of importance sampling densities from which a good choice can be made automatically. Prior distributions and functions of interest can then be specified at will, taking care that computed moments actually exist. There is no *guarantee* that these methods will produce computed moments of reasonable accuracy in a reasonable number (say, 10,000) of Monte Carlo replications. Pathological likelihood functions will demand more analytical work: for example analytical marginalization may be possible, and transformation of parameters to obtain more regular posterior densities can be quite helpful, but these tasks must be approached on a case-by-case basis. It is in precisely such cases that sampling theoretic asymptotic theory and normal or other approximations to the posterior are also likely to be inadequate: the difficulty is endemic to the model and not the approach. Clearly there is much to be learned about more complicated likelihood functions. In approaching these problems, the emphasis should be—as it has been in this paper—on generic rather than specific solutions.

Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27706, U.S.A.

Final manuscript received March, 1989.

REFERENCES

BAUWENS, W. (1984): *Bayesian Full Information Analysis of Simultaneous Equation Models Using Integration by Monte Carlo*. Berlin: Springer-Verlag.

BAUWENS, W., AND J. F. RICHARD (1985): "A 1-1 Poly-*t* Random Variable Generator with Application to Monte Carlo Integration," *Journal of Econometrics*, 29, 19–46.

BILLINGSLEY, P. (1979): *Probability and Measure*. New York: Wiley.

CRAMER, H. (1946): *Mathematical Methods of Statistics*. Princeton: Princeton University Press.

DAVIS, P. J., AND P. RABINOWITZ (1975): *Methods of Numerical Integration*. New York: Academic Press.

ENGLE, R. F. (1982): "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50, 987–1008.

——— (1983): "Estimates of the Variance of U.S. Inflation Based on the ARCH Model," *Journal of Money, Credit, and Banking*, 15, 286–301.

——— (1984): "Wald, Likelihood Ratio and Lagrange Multiplier Tests in Econometrics," Chapter 13 of *Handbook of Econometrics*, Vol. II, ed. by Z. Griliches and M. D. Intriligator. Amsterdam: North-Holland.

GALLANT, A. R., AND J. F. MONAHAN (1985): "Explicitly Infinite-Dimensional Bayesian Analysis of Production Technologies," *Journal of Econometrics*, 30, 171–202.

GEWEKE, J. (1986): "Exact Inference in the Inequality Constrained Normal Linear Regression Model," *Journal of Applied Econometrics*, 1, 127–141.

——— (1988a): "Exact Inference in Models with Autoregressive Conditional Heteroscedasticity," in *Dynamic Econometric Modeling*, ed. by E. Berndt, H. White, and W. Barnett. Cambridge: Cambridge University Press, 73–104.

——— (1988b): "Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference," *Journal of Econometrics*, 38, 73–90.

GEWEKE, J., R. C. MARSHALL, AND G. ZARKIN (1986a): "Exact Inference for Continuous Time Markov Chain Models," *Review of Economic Studies*, 53, 653–699.

——— (1986b): "Mobility Indices in Continuous Time Markov Chains," *Econometrica*, 54, 1407–1423.

HAMMERSLEY, J. M., AND D. C. HANDSCOMB (1964): *Monte Carlo Methods*. London: Methuen (First Edition).

——— (1979): *Monte Carlo Methods*. London: Chapman and Hall (Second Edition).

JOHNSON, N. C., AND S. KOTZ (1970): *Continuous Univariate Distributions-2*. New York: Wiley.

KLOEK, T., AND H. K. VAN DIJK (1978): "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo," *Econometrica*, 46, 1–20.

QUANDT, R. E. (1983): "Computational Problems and Methods," Chapter 12 of *Handbook of Econometrics*, Vol. I, ed. by Z. Griliches and M. D. Intriligator. Amsterdam: North-Holland.

RUBINSTEIN, R. Y. (1981): *Simulation and the Monte Carlo Method*. New York: Wiley.

SINGER, B., AND J. COHEN (1980): "Estimating Malaria Incidence and Recovery Rates from Panel Surveys," *Mathematic Biosciences*, 49, 273–305.

SINGER, B., AND S. SPILERMAN (1976): "The Representation of Social Processes by Markov Models," *American Journal of Sociology*, 82, 1–54.

VAN DIJK, H. K., J. P. HOP, AND A. S. LOUTER (1986): "An Algorithm for the Computation of Posterior Moments and Densities Using Simple Importance Sampling," Erasmus University Econometric Institute Report 8625/E.

VAN DIJK, H. K., T. KLOEK, AND C. G. E. BOENDER (1985): "Posterior Moments Computed by Mixed Integration," *Journal of Econometrics*, 29, 3–18.

ZELLNER, A., AND P. ROSSI (1984): "Bayesian Analysis of Dichotomous Quantal Response Models," *Journal of Econometrics*, 25, 365–394.