BAYESIAN INference IN econometric models

using Monte Carlo integration

by

John Geweke†

DP#: 87-02

†Department of Economics
Duke University
Durham, NC 27706
(919) 684-2723

July, 1986
Revised, April 1987

Financial support for this work was provided by NSF Grant SES-8605667, and by a grant from the Pew Foundation to Duke University. Jean-Francois Richard provided helpful comments on an earlier version of this paper.
Abstract

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 the asymptotic sampling theoretic densities of the maximum likelihood estimators, with the asymptotic density modified by possible substitution of the multivariate normal by a multivariate $t$ whose degrees of freedom are determined analytically, and by automatic rescaling of the density 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.
1. Introduction

Econometric models are usually expressed in terms of an unknown vector of parameters \( \theta \in \mathbb{R}^k \), which fully specifies the joint probability distribution of the observations \( X = \{X_1, \ldots, X_T\} \). In most cases there exists a probability density function \( f(X|\theta) \), and classical inference then often proceeds from the likelihood function \( L(\theta|X) = f(X|\theta) \). 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 (possibly improper) probability density function over the parameters. Denote the kernel of the prior \( \pi(\theta|Q) \), where \( Q \) is the conditioning set of prior information. The posterior distribution is proportional to \( p(\theta|Q,X) \equiv \pi(\theta|Q)L(\theta|X) \). Most Bayesian inference problems can be expressed as the evaluation of the expectation of a function of interest \( g(\theta) \) under the posterior,

\[
E[g(\theta)|Q,X] = \int g(\theta)p(\theta|Q,X)\,d\theta / \int p(\theta|Q,X)\,d\theta.
\]

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 $g$, $\pi$, 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 $E[g(\theta) \mid Q,X]$ 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 $\{\theta_i\}$ 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 $\theta_i$ is proportional to $p(\theta \mid Q,X)$ then

$$n^{-1} \sum_{i=1}^{n} g(\theta_i) \rightarrow E[g(\theta) \mid Q,X];$$

if it is $L(\theta \mid X)$ then

$$n^{-1} \sum_{i=1}^{n} g(\theta_i) \pi(\theta_i \mid Q) \rightarrow E[g(\theta) \mid Q,X].$$

[Except in Section 4, all convergence is in $n$, the number of Monte Carlo replications.] 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, 1986a). More generally, suppose that the probability distribution function of the $\theta_i$ is $I(\theta)$, termed the importance sampling density. Then, under very weak assumptions given in Section 2,
\[ \tilde{g}_n = \sum_{i=1}^{\infty} \frac{g(\theta_i)p(\theta_i|Q,X)}{I(\theta_i)} \sum_{i=1}^{\infty} \frac{p(\theta_i|Q,X)}{I(\theta_i)} = E[g(\theta)|Q,X]. \]

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,

\[ n^{1/2}(\tilde{g}_n - E[\tilde{g}(\theta)|Q,X]) \Rightarrow N(0,\sigma^2), \]

and \( \sigma^2 \) may be estimated consistently. This result was indicated by Kloek and van Dijk (1978), but does not follow from the result of Cramer (1945) 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 \( I(\theta) \) not decay more quickly than the tails of \( p(\theta|Q,X) \).

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 $I(\theta)$. It utilizes the asymptotic sampling theoretic density of the maximum likelihood estimator, 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. (For reference, a summary of notation may be found at the end of this section.) Since expectations are taken with respect to a variety of distributions, it is useful to denote $E_{f(\theta)}[h(\theta)] = \int_{\Theta} f(\theta) h(\theta) d\theta$ where $f(\theta)$ is proportional to a probability density and $h(\theta)$ is integrable with respect to the probability measure that induced $f(\theta)$. Similarly denote by $\text{var}_{f(\theta)}[h(\theta)]$ and $\text{md}_{f(\theta)}[h(\theta)]$ the variance and mean deviation of $h(\theta)$ under this probability measure, when the respective moments exist.

**Assumption I.** The product of the prior density, $\pi(\theta|Q)$, and the likelihood function, $L(\theta|X)$, is proportional to a proper probability density function defined on $\Theta$.

Assumption I is generally satisfied and usually not difficult to verify. It explicitly allows the prior density to be improper, so long as $\int_{\Theta} p(\theta|Q,X) d\theta \equiv c < \infty$. For $h(\theta)$ integrable with respect to the posterior probability measure, the simpler notation $\bar{h} = E[h(\theta)] = E_p(\theta|Q,X)[h(\theta)]$ (and similarly $\text{var}[h(\theta)]$ and $\text{md}[h(\theta)]$) will be used.
Assumption II. \( \{ \theta_i \}_{i=1}^{\infty} \) is a sequence of i.i.d. random vectors, the common distribution having a probability density function \( I(\theta) \).

Assumption III. The support of \( I(\theta) \) includes \( \theta \).

Assumptions II and III 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 \( E[g(\theta)] \), through the function of interest, \( g(\theta) \).

Assumption IV. \( E[g(\theta)] \) exists.

Examples of \( g(\theta) \) include \( \theta_i \) if \( \theta_i \) is to be estimated with a quadratic loss function; a loss function under a particular action if a decision is to be made; or the indicator function \( \chi(\theta; \theta^*) = \{ 1 \text{ if } \theta \in \theta^*, \ 0 \text{ if } \theta \notin \theta^* \} \) if \( P(\theta \in \theta^* | Q,X) \) is to be determined.

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 \( I(\theta) \), so we shall re-express \( \bar{g}_n \) with \( I^*(\theta) = d^{-1}I(\theta) \) in place of \( I(\theta) \). Denote the weight function \( w(\theta) = p(\theta | Q,X)I^*(\theta) \); then \( \bar{g}_n = \sum_{i=1}^{n} g(\theta_i)w(\theta_i)/\sum_{i=1}^{n} w(\theta_i) \).

**Theorem 1.** Under assumptions I-IV, \( \bar{g}_n \rightarrow E[g(\theta) | Q,X] \).

**Proof.** Since \( w(\theta) \) is a nonnegative function defined on \( \theta \), \( E_{I}(g(\theta)w(\theta)) = c\bar{g} \) (Billingsley, 1979, Theorem 16.10). By Khinchine's strong law of large numbers (Billingsley, 1979, Theorem 22.5), \( n^{-1}\sum_{i=1}^{n} g(\theta_i)w(\theta_i) \rightarrow c\bar{g} \). Similarly, \( n^{-1}\sum_{i=1}^{n} w(\theta_i) \rightarrow cd \). ###
In many instances \( g(\theta) \) is bounded and then Assumption IV is trivially satisfied. A leading example is the determination of the posterior c.d.f. of function \( h(\theta) \).

**Corollary 1.** Let \( h(\theta) \) be a function defined on \( \Theta \) and measurable with respect to the posterior distribution, and let \( F(h^*) = P[h(\theta) \leq h^*] \) be the posterior c.d.f. of \( h(\theta) \). Let \( g(\theta; h^*) = \chi(\theta; \Theta; h(\theta) \leq h^*) \). Given Assumptions I-III,

\[
\tilde{g}_n(h^*) = \sum_{i=1}^{n} g(\theta_i; h^*) w(\theta_i) / \sum_{i=1}^{n} w(\theta_i) \to F(h^*).
\]

While formal inference problems can always be cast in terms of \( E[g(\theta)] \) for suitably chosen \( g(\theta) \), it is sometimes useful to describe the posterior distribution of \( h(\theta) \) by its quantiles rather than by its cdf. Quantile estimation is not a special case of determining \( E[g(\theta)] \).

**Theorem 2.** Extending the notation and assumptions of Corollary 1, denote \( \alpha = F(h^*) \) and suppose that \( F \) is strictly increasing in an open neighborhood \( \mathcal{H} \) of \( h^* \). Define the inverse cdf \( F^{-1}(\alpha) \) on \( \mathcal{Q} = \{ \alpha : F(h^*) = \alpha \} \) for some \( h^* \in \mathcal{H} \). Let \( T_n = \sum_{i=1}^{n} w(\theta_i) \) and let \( h^*_n(\theta) \) be any real number \( q \) such that \( \sum_{i=1}^{n} \chi_{\theta_i \leq h^*_n(\theta)} > q T_n \) and \( \sum_{i=1}^{n} \chi_{\theta_i > h^*_n(\theta)} > (1-q)T_n \). Then \( h^*_n(\theta) \to F^{-1}(\alpha) \).

**Proof.** Let \( \varepsilon \) be any positive constant such that \( (h^*-\varepsilon, h^*+\varepsilon) \subseteq \mathcal{H} \).

Let \( E_n(\varepsilon) = 1 \) if \( |h^*_n(\alpha) - h^*| < \varepsilon \), and 0 otherwise. From Corollary 1, \( \tilde{g}_n(h^* - \varepsilon) \to F(h^* - \varepsilon) < F(h^*) \) and \( \tilde{g}_n(h^* + \varepsilon) \to F(h^* + \varepsilon) > F(h^*) \). Hence \( E_n(\varepsilon) \to 1 \). Since this is true for arbitrarily small \( \varepsilon \), \( h^*_n(\alpha) \to h^* \). # # #
The \( \alpha \)'th empirical quantile is uniquely defined unless there is some \( q \) such that \( \sum_{i=1}^{n} w_i q \theta_i = \alpha T_n \). This generally will not occur, but Theorem 2 implies that any reasonable way of resolving the ambiguity will assure \( \frac{h_n^*(\alpha)}{n} \rightarrow F^{-1}(\alpha) \). The situation is less sanguine if \( F \) is not monotone increasing. In this case a set mapping \( F^{-1}(\cdot) \) is defined, but the limiting distribution of \( \frac{h_n^*(\alpha)}{n} \) has all its mass on the endpoints of the line segment \( F^{-1}(\alpha) \). This is exceptional: most posterior distributions have positive density on a convex subset of the real line.

---

### Summary of Notation

<table>
<thead>
<tr>
<th>Term or Concept</th>
<th>Notation or Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td>X</td>
</tr>
<tr>
<td>Prior information</td>
<td>Q</td>
</tr>
<tr>
<td>Parameter vector</td>
<td>( \theta )</td>
</tr>
<tr>
<td>Parameter space</td>
<td>( \Theta )</td>
</tr>
<tr>
<td>Likelihood function</td>
<td>( \propto L(\theta</td>
</tr>
<tr>
<td>Prior density</td>
<td>( \propto \pi(\theta</td>
</tr>
<tr>
<td>Posterior density</td>
<td>( \propto p(\theta</td>
</tr>
<tr>
<td>Function of interest</td>
<td>( g(\theta) )</td>
</tr>
<tr>
<td>Moments under the posterior</td>
<td>( E, \text{var}, \text{etc.} )</td>
</tr>
<tr>
<td>Moments under density</td>
<td>( E_f(\theta), \text{var}_f(\theta), \text{etc.} )</td>
</tr>
<tr>
<td>Importance sampling density</td>
<td>( I(\theta) )</td>
</tr>
<tr>
<td>( I^*(\theta) )</td>
<td>( \propto I(\theta) )</td>
</tr>
<tr>
<td>( d )</td>
<td>( I^*(\theta) = d^{-1}I(\theta) )</td>
</tr>
<tr>
<td>( \tilde{\theta} )</td>
<td>( E[g(\theta)</td>
</tr>
</tbody>
</table>
3. Evaluating the Importance Sampling Distribution

These results provide the paradigm for inference with importance sampling. Standing alone they are of little practical value, because nothing can be said about rates of convergence and there is no formal guidance for choosing I(\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(\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(\theta_i) that turn up occasionally. The obvious place to begin is to apply a central limit theorem to \bar{g}_n. We use a fact easily derived from Taylor's Theorem.

**Lemma 1.** Suppose \( x_n \rightarrow y, \sqrt{n}(x_n - y) \Rightarrow N(0, \Sigma) \). If \( f(\cdot) \) has continuous first and second partial derivatives in an open neighborhood of \( y \) and \( f'(y) \neq 0 \) then \( n^{1/2}[f(x_n) - f(y)] \Rightarrow N[0, f'(y)'\Sigma f'(y)] \).

The following result may be used to evaluate the numerical accuracy of \( \bar{g}_n \) for large \( n \).
Theorem 3. In addition to assumptions I-IV, suppose
\[
E[w(\theta)] = c^{-1} \int \theta \left[ p(\theta | Q, X)^2 / I^*(\theta) \right] d\theta
\]
and
\[
E[g^2(\theta) w(\theta)] = c^{-1} \int \theta \left[ g^2(\theta) p(\theta | Q, X)^2 / I^*(\theta) \right] d\theta
\]
exist and are finite. Let
\[
\sigma^2 = (cd)^{-1} E\left\{ [g(\theta) - \bar{g}]^2 w(\theta) \right\} = c^{-2} d^{-1} \int \theta \left[ g(\theta) - \bar{g} \right]^2 w(\theta)^2 p(\theta | Q, X) d\theta.
\]
Then
\[
n^{1/2} (\bar{g}_n - \bar{g}) \Rightarrow N(0, \sigma^2),
\]
and if
\[
\frac{\hat{\sigma}^2_n}{n} = \frac{\sum_{i=1}^{n} [g(\theta_i) - \bar{g}_n] w(\theta_i)^2 / \sum_{i=1}^{n} w(\theta_i)^2}{n^{1/2}}
\]
then \( n^{1/2} \hat{\sigma}^2_n \to \sigma^2. \)

Proof. Define \( A(\theta) \equiv g(\theta) w(\theta) \) and observe that
\[
\bar{g}_n = \left( \frac{1}{n} \sum_{i=1}^{n} A(\theta_i) \right) / \left( \frac{1}{n} \sum_{i=1}^{n} w(\theta_i) \right). \] From Theorem 1, \( n^{-1} \sum_{i=1}^{n} A(\theta_i) \to cd \bar{g}, \)
\( n^{-1} \sum_{i=1}^{n} w(\theta_i) \to cd. \) By the Central Limit Theorem (Billingsley, 1979, Theorem 29.5)

\[
n^{1/2} \begin{bmatrix} - \frac{1}{n} \sum_{i=1}^{n} A(\theta_i) - cd \bar{g} \\ - \frac{1}{n} \sum_{i=1}^{n} w(\theta_i) - cd \end{bmatrix} \Rightarrow N \begin{bmatrix} - \var_{I(\theta)}[A(\theta)] \\ - \var_{I(\theta)}[w(\theta)] \end{bmatrix},
\]

with
\[
\var_{I(\theta)}[A(\theta)] = \int \theta \left[ g(\theta) w(\theta) - cd \bar{g} \right]^2 I(\theta) d\theta,
\]
\[
\var_{I(\theta)}[w(\theta)] = \int \theta \left[ w(\theta) - cd \right]^2 I(\theta) d\theta.
\]
\[
\text{cov}_{I(\theta)}[A(\theta), w(\theta)] = \int \theta [g(\theta)w(\theta) - cd\theta][w(\theta) - cd]I(\theta)d\theta,
\]
so long as \( \text{var}_{I(\theta)}[w(\theta)] > 0 \). The convergence of \( n^{1/2}(\tilde{g}_n - \tilde{g}) \) follows from the application of Lemma 1 to \( f(A, w) = A/w \), and that of \( \tilde{\sigma}_n^2 \) follows by an argument parallel to that for Theorem 1. It is easily verified that the result applies when \( \text{var}_{I(\theta)}[w(\theta)] = 0 \), as well. 

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

The conditions of Theorem 3 must be verified, analytically, if that result is to be used to assess the accuracy of \( \tilde{g}_n \) as an approximation of \( E[g(\theta)] \). Failure of these conditions does not vitiate the use of \( \tilde{g}_n \): there is no compelling reason why \( n^{1/2}(\tilde{g}_n - \tilde{g}) \) needs to have a limiting distribution. But it is essential to have some indication of \( |\tilde{g}_n - \tilde{g}| \), and this task becomes substantially more difficult if a central limit theorem cannot be applied. Practicality suggests that \( I(\theta) \) be chosen so that Theorem 3 applies; and this can usually be accomplished in one of two ways.

**Corollary 2.** Given Assumptions I-IV,

\[
n^{1/2}(\tilde{g}_n - \tilde{g}) \Rightarrow N(0, \sigma^2) \quad \text{and} \quad n\tilde{\sigma}_n^2 + \sigma^2
\]

if either

(A) \( w(\theta) < \bar{w} < \infty \forall \theta \in \Theta \) and \( \text{var}[g(\theta)] < \infty \),

or

(B) \( \Theta \) is compact and \( I(\theta) > \varepsilon > 0 \forall \theta \in \Theta \).

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

The expression for \( \sigma^2 \) in Theorem 3 indicates that the numerical standard 
error is adversely affected by large \( \text{var}[g(\theta)] \), 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 distribution. A simple benchmark for comparing the adequacy of importance sampling distributions is the 
numerical standard error that would result if the importance sampling distrib-
ution were the posterior distribution itself, i.e., \( I^*(\theta) = p(\theta|Q,X) \). In 
this case \( \sigma^2 = \text{var}[g(\theta)] \) 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 per-
cent 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 \( \text{var}[g(\theta)] \) 
can be computed numerically, 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 
\( g(\theta) \),

\[
\text{RNE} = \text{var}[g(\theta)]/d^{-1}\mathbb{E}\left[\left[g(\theta) - \bar{g}\right]^2w(\theta)\right] = \text{var}[g(\theta)]/\sigma^2
\]
The RNE is the ratio of the number of replications required to achieve any specified numerical standard error using the importance sampling density \( I(\theta) \), to the number required using the posterior density as the importance sampling density. The numerical standard error for \( \bar{g}_n \) is the fraction \((\text{RNE} \cdot n)^{-1/2}\) of the posterior standard deviation. Low values of RNE indicate potentially large increases in numerical efficiency by choosing a different importance sampling density.

While the posterior is an appealing importance sampling distribution, there are distributions that produce lower numerical standard errors and therefore RNE may exceed 1. The following result shows that such improvements on the posterior may be made by placing additional weight in the tails of the posterior distribution of \( g(\theta) \).

**Theorem 4.** In addition to Assumptions I-IV suppose that \( \text{md}[g(\theta)] \) and \( E[|g(\theta) - \bar{g}|^{-1}] \) exist. Then the importance sampling density that minimizes \( \sigma^2 \) is proportional to \( I^{**}(\theta) = |g(\theta) - \bar{g}| p(\theta | Q, X) \). For this choice, \( \sigma^2 = \{\text{md}[g(\theta)]\}^2 \).

**Proof.** Let \( K^{*}(\theta) \) be any other integrable function such that \( \int_{\theta} K^{*}(\theta) d\theta = \int_{\theta} I^{**}(\theta) d\theta = d^{-1} \). Without loss of generality \( K^{*}(\theta) = I^{**}(\theta) + \varepsilon J(\theta) \), where \( \int_{\theta} J(\theta) d\theta = 0 \) and \( \int_{\theta} |J(\theta)| d\theta = 1 \). With the importance sampling density \( K^{*}(\theta) \),

\[
\sigma^2 = d^{-1} \int_{\theta} \left[ |g(\theta) - \bar{g}|^2 p(\theta | Q, X)^2 / K^{*}(\theta) \right] d\theta
\]

\[
= d^{-1} \int_{\theta} \left[ |g(\theta) - \bar{g}| p(\theta | Q, X)[1 + \varepsilon I^{**}(\theta)^{-1} J(\theta)]^{-1} \right] d\theta \equiv W(\varepsilon).
\]

Then
\[ W'(\varepsilon) = -d^{-1} \int_{\Theta} J(\theta) [1 + \varepsilon I^{**}(\theta)^{-1} J(\theta)]^{-2} d\theta, \]

\[ W''(\varepsilon) = 2d^{-1} \int_{\Theta} I^{**}(\theta)^{-1} J(\theta)^{2} [1 + \varepsilon I^{**}(\theta)^{-1} J(\theta)]^{-3} d\theta > 0 \quad \forall \; \varepsilon, \]

\[ W'(0) = 0. \]

Direct application of Theorem 4 is limited by the assumption \( E[|g(\theta) - \bar{g}|^{-1}] < \infty, \) and the fact that the importance sampling density which minimizes \( \sigma^2 \) differs depending on \( g(\theta). \) The most interesting leading cases are model evaluation problems: \( \theta \in \Theta^* \) or \( \theta \in \Theta^\dagger, \) \( g(\theta) = \chi(\theta^*), \) and so \( E[g(\theta)] = P[\theta \in \Theta^*]. \) If \( p^* \equiv P[\theta \in \Theta^*], \) then \( \sigma^2 \) is minimized if \( I(\theta) = (1-p^*)p(\theta|Q,X) \) for \( \theta \in \Theta^* \) and \( I(\theta) \propto p^*(\theta|Q,X) \) for \( \theta \in \Theta^\dagger: \)

half the drawings should be made in \( \Theta^* \) and half in \( \Theta^\dagger, \) and in proportion to the posterior in each case.

Improvements on importance sampling are essentially always possible, under the weaker conditions of Theorem 3. Take the sampling density to be the convex combination \( (1-\lambda)c^{-1}p(\theta|Q,X) + \lambda \text{md}[g(\theta)]^{-1} |g(\theta) - \bar{g}| p(\theta|Q,X), \) and express the variance \( \sigma^2 \) of Theorem 3 as \( f(\lambda): \) then

\[ f'(0)/f(0) = 1 - \frac{E|g(\theta) - \bar{g}|^3}{E|g(\theta) - \bar{g}|E|g(\theta) - \bar{g}|^2} < 0, \]

and in most cases the inequality will be strict. It is impractical to tailor the importance sampling distribution to each function of interest, since unknown moments are involved and a typical investigation will have many functions of interest. Theorem 4 and the inequality (4) suggest that there may be something to be gained from importance densities with tails thicker
than those of the posterior density. Potential improvements in numerical efficiency are greatest with functions of interest whose posterior distributions exhibit thick tails, and for these functions it may be possible to achieve \( RNE > 1 \).

4. Choosing the Importance Sampling Distribution

There are two logical steps in choosing the importance sampling distribution. The first is to determine a class of distributions that satisfy the conditions of Theorem 3, usually by using Corollary 2. The second step is to find a distribution 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 numerical problem in its own right, that can be solved numerically.

When \( \Theta \) 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 \( \hat{\theta} \) obtain, with \( \hat{\theta} \sim N(\theta^*, V) \), then the \( N(\hat{\theta}, V) \) importance sampling distribution satisfies Corollary 2(B). If \( \Theta \) is not compact, it will generally be necessary to investigate the tail behavior of the likelihood function to verify the conditions of Corollary 2(A). There are many instances in which \( L(\theta) \) behaves like \( k\theta^{-q} \) for large values of \( \theta' \theta \), and it is not difficult to determine \( i \) and \( q \). In these instances no multivariate normal importance sampling distribution will satisfy Theorem 3. 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 distribution. It is not hard to see that (a) Theorem 3 does not apply; (b) if one computes \( \hat{\sigma}_n^2 \) and
\( \hat{\text{var}}(\hat{g}_n) = \frac{1}{n} \sum_{i=1}^{n} \left[ g(\theta_i) - \hat{\theta}_n \right]^2 w(\theta_i) / \sum_{i=1}^{n} w(\theta_i) \) then the computed RNE approaches 0 almost surely as \( n \to \infty \); (c) if \( n \) is held fixed, RNE \( \to 1 \) as sample size approaches infinity because \( w(\hat{\theta}_n) \to 1 \) for all \( i = 1, \ldots, n \). These results are typical of cases in which the classical asymptotic expansion of the log-likelihood function is valid, and \( \Theta \) 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 \( \Sigma \) and degrees of freedom \( \nu \), a multivariate t importance sampling distribution with weakly larger variance and/or weakly small degrees of freedom will satisfy Theorem 3.

In general let \( I(\hat{\theta}; g) \) be a family of importance sampling densities indexed by the \( g \times 1 \) vector \( g \), and let \( w(\hat{\theta}; g) \equiv p(\hat{\theta} | Q, X) / I(\hat{\theta}; g) \) be the corresponding family of weight functions. It will not usually be practical to choose a distinct importance sampling distribution for each function of interest. A reasonable objective is the minimization of

\[ E[w(\hat{\theta}; g)] = \int_{\Theta} \left[ p(\hat{\theta} | Q, X)^2 / I(\hat{\theta}; g) \right] d\Theta. \]

This function appears in the first condition of Theorem 3 and in the denominator of each \( E[g(\hat{\theta})] \); it is precisely the function one would minimize for \( g(\hat{\theta}) = \chi(\hat{\theta}^*) \), \( p(\hat{\theta}^*) = 1/2 \).

The first and second order conditions for minimization are

\[ E[w(\hat{\theta}; g) \partial \log I(\hat{\theta}; g) / \partial g] = 0 \]

and

\[ E[w(\hat{\theta}; g)] \left[ 2 \partial \log I(\hat{\theta}; g) / \partial g \right] \partial \log I(\hat{\theta}; g) / \partial g' - \partial^2 \log I(\hat{\theta}; g) / \partial g \partial g' \right] p.d., \]
respectively.

Two related classes of importance sampling distributions have turned out to be quite useful in our experience with this approach. Both involve distributions that have not, to our knowledge, been described in the literature. The heuristic idea is to begin with the classical, asymptotic normal distribution of the maximum likelihood estimator as a first approximation to the likelihood function itself; use the corresponding multivariate Student, if warranted, with the degrees of freedom that satisfy Corollary 2 indicated by the tail behavior of the likelihood function; choose a set of asymptotically orthogonal axes using the variance matrix of the asymptotic distribution of the maximum likelihood estimator; and adjust separately along each axis and in each direction to reduce $E[w(\theta_0; \alpha)]$. With $k$ parameters, there are $2k$ adjustments in the last step, so $\alpha$ 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 $N^k(\mu, \Sigma, \theta, r)$ is readily described by construction of a member $\mathbf{x}$ of its population:

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

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

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

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

$$-\sum_{i=1}^{2k} \left[ \log(q_i)\text{sgn}^+(\varepsilon_i) + \log(r_i)\text{sgn}^-(\varepsilon_i) \right] - (1/2)\mathbf{\varepsilon}'\Sigma^{-1}\mathbf{\varepsilon},$$
and satisfies (6) globally. In the application here, $\theta$ is the m.l.e. of $\theta$, and $T$ is a factorization such that $TT'$ provides an asymptotically valid expression for the variance of the m.l.e. A variate from the multivariate split Student density $t^*(\eta, T, \varpi, \tau, \nu)$ is constructed the same way, except that $
_i = [q_i \text{sgn}^+ (\varepsilon_i) + r_i \text{sgn}^- (\varepsilon_i)]\varepsilon_i (\tau/\nu)^{-1/2}$, with $\tau \sim \chi^2(\nu)$. The log-p.d.f. is (up to an additive constant)

$$-\sum_{i=1}^{k} [\log(q_i) \text{sgn}^+ (\varepsilon_i) + \log(r_i) \text{sgn}^- (\varepsilon_i)] - [(\nu+k)/2] \log(1+\nu^{-1} \varepsilon_i^2),$$

and again (6) is satisfied globally.

Choosing $\varpi' = (\varpi', \tau')$ to minimize $E[w(\varpi; g)]$, or equivalently to satisfy (5), 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 $E[g(\varpi)]$ itself. A much more efficient procedure can be used to select values for $\varpi$ and $\tau$ that, in our applications to date, produce high RNE's. The intuitive idea is to explore each axis in each direction, to find the lowest rate of decline in the posterior density relative to a univariate normal (or Student), and then choose the variance of the normal (or Student) to match that lowest rate of decline. A little more formally, let $e^{(i)}$ be a $k \times 1$ indicator vector, $e^{(i)}_i = e^{(i)}_i = 1$. For the split normal define $f_i(\delta)$ according to

$$p(\theta + \delta T^{(i)}; Q, X)/p(\theta; Q, X) = \exp[-\delta^2/2f_i(\delta)^2],$$
Then take \( q_i = \min_{\delta>0} f_i(\delta) \) and \( r_i = \min_{\delta<0} f_i(\delta) \). For the split Student the procedure is the same except that

\[
f_i(\delta) = \nu^{-1/2} |\delta| \{ [p(\hat{\theta};Q,X)/p(\hat{\theta} + \delta \Sigma^{(i)};Q,X)]^{2/(\nu+k)} - 1 \}^{-1/2}.
\]

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 \( \theta \) 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 \( \pi(\theta|Q) = 1, 0 < \theta < 1 \), then the posterior is

\[
p(\theta|Q,X) = [B(M+1,N-M+1)]^{-1} \theta^M (1-\theta)^{N-M} \quad (0 < \theta < 1).
\]
Since \( N^{1/2}(\hat{\theta} - \theta) \rightarrow N(0, \theta(1 - \theta)) \), a normal approximation based on the asymptotic distribution of the m.l.e. \( \hat{\theta} \) is \( I(\theta) = [2\pi N^{-1/2}(1 - \theta)]^{1/2} \exp\{-N(\hat{\theta} - \theta)^2 / 2\theta(1 - \theta)\} \). Since \( 0 < \theta < 1 \) the conditions of Corollary 2(B) are 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. Figure 1 shows the weight functions \( w(\theta;Q,X) = L(\theta;Q,X)/I(\theta) \) for these two cases, indicated by the heavy line. In each case \( w(\theta;Q,X) \) appears badly behaved, attaining a value of over \( 10^{70} \) for \( \theta \) around .9 in Case A, and attaining a value of over \( 10^{4} \) for \( \theta \) between .2 and .3 in Case B. For our purposes it is \( E[w(\theta;Q,X)] = \int_0^\theta w(\theta;Q,X)p(\theta;Q,X)d\theta \) and not \( w(\theta;Q,X) \) which matters. The light lines in Figure 1 show \( w(\theta;Q,X)p(\theta;Q,X) \), which is much less than \( w(\theta;Q,X) \) when \( w(\theta;Q,X) \) is large. In Case A \( E[w(\theta;Q,X)] = 1.65 \times 10^{-24} \), and in Case B \( E[w(\theta;Q,X)] = 1.05 \). 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 distribution 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, \( E[w(\theta|Q,X)] \) decreases exponentially, until the variance reaches .0016; \( E[w(\theta|Q,X)] \) attains its minimum at .0019, and then increases slowly. Figure 3 indicates the variance \( \sigma^2 \) that equates the ratio of a \( N(\hat{\theta}, \sigma^2) \) density evaluated at \( \theta \) to that at \( \hat{\theta} \), with the likelihood ratio \( L(\theta;Q,X)/L(\hat{\theta};Q,X) \). Following the procedure suggested in Section 4, the indicated variance would increase monotonically until \( \theta = .65 \).
some 16 asymptotic standard deviations above \( \hat{\theta} \). For \( \theta \) above .18 — about 2.7 standard deviations above \( \hat{\theta} \) — changes in the indicated variance have essentially no effect on \( E[w(\theta)|Q,X] \), as comparison of Figures 2 and 3 indicates.

5.2 The Two-State Homogeneous Markov Chain Model

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 \( \pi(p_1,p_2|Q,X) = 1 \), \( 0 < p_1 < 1 \) (\( i=1,2 \)), the posterior p.d.f. is proportional to

\[
p(p_1,p_2|Q,X) = p_1^{m_{12}} p_2^{m_{21}} (1-p_1)^{m_{11}} (1-p_2)^{m_{22}} ,
\]

\( 0 < p_1 < 1 \) (\( i=1,2 \)). 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

---

\(^1\) For elaboration on the Markov chain model and the problem of embeddability, see Singer and Spilerman (1976) or Geweke, Marshall and Zarkin (1986a, 1986b).
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 embed-
dable. Data for three cases, and the corresponding maximum likelihood estimates and their asymptotic standard errors are provided in Table 1. 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 (8) and (9). In Case I \( \hat{p}_1 + \hat{p}_2 << 1 \), and the model is embed-
dable; 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: \( \pi(p_1, p_2 | Q) = 1 \) for all \( p_j \in (0,1) \). The second prior imposes embeddability but otherwise is the same as the first: \( \pi(p_1, p_2 | Q) = 2 \) for all \( (p_1, p_j): p_1 > 0, p_j > 0, p_1 + p_2 < 1 \). The functions of interest are \( p_1, p_2, p_{1\text{-}1}, p_{2\text{-}1}, r_1^{-1}, \) and \( r_2^{-1} \). The function \( p_j^{-1} \) is the mean duration in state \( j \) in the discrete model, while \( r_j^{-1} \) is the mean duration in state \( j \) assuming a continuous time model. Since \( r_j^{-1} \) is undefined if \( p_1 + p_2 > 1 \), \( E[r_j^{-1} | Q, X] \) is defined only under the second prior; \( E[r_j^{-1} | Q, X] \) does not exist for either prior, although the posterior distribution of \( r_j \) may be obtained readily using Corollary 1.

Results obtained using the methods proposed in earlier sections are pre-
sented in Table 2. 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 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 distribution, constructed as described in Section 4. Since the posterior distribution is the product of a binomial in \( p_1 \) and a binomial in
p_j, 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, \( P[p_1 + p_2 < 1 \mid Q,X] = 1 \) under the first prior: none of the 10,000 replications with any sampling distribution turned up \( p_1 + p_2 > 1 \). In Case II, \( P[p_1 + p_2 < 1 \mid Q,X] = .65 \) and in Case III \( P[p_1 + p_2 < 1 \mid Q,X] = .21 \), under the first prior. The computed expected values of all functions of interest with each importance sampling distribution 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 distribution than the normal. The superiority of the split normal increases systematically as \( E[p_j \mid Q,X] \) 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 sampling distribution, precisely where we would expect: \( E[p_j \mid Q,X] \) 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, 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.

From the example of Section 5.1, we know that the population RNE for the normal importance sampling distribution in Case I should be on the order of $10^{-24}$. The normal is such a bad sampling distribution 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 3 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 $w(\theta_i)$ increase rapidly as the number of replications, $n$, increases, and order statistics pertaining to the $w(\theta_i)$ are often useful in identifying very poorly conditioned importance sampling distributions. Since the largest $w(\theta_i)$'s may be retained systematically as $n$ increases and $\sum_{i=1}^{n} w(\theta_i)$ and $\sum_{i=1}^{n} w(\theta_i)^2$ are computed in any event, the diagnostic of the form $\omega_m = [n/(n-m+1)] \sum_{i=m}^{n} w^{(i)}(\theta)^2 / \sum_{i=1}^{n} w(\theta_i)^2$ (where superscripts denote order statistics and $m$ is close to $n$) 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 3, again for $n=10,000$ and $n=50,000$; they indicate -- correctly -- that the split normal is a much better importance sampling distribution than is the normal.
5.3 The Quartile Homogeneous Markov Chain Model\textsuperscript{2}

Consider last a four-state homogeneous Markov chain model, state \( j \) denoting the \( j \)'th quartile of income. Let \( p_{ij} \equiv P[\text{In state } j \text{ at time } t \mid \text{In state } i \text{ at time } t-1] \), and arrange the \( p_{ij} \) in a 4x4 matrix \( P \). Since \( \sum_{i=1}^{4} p_{ij} = \sum_{i=1}^{4} p_{ji} = 1 \) for \( j=1,\ldots,4 \), the likelihood function has 9 free parameters and cannot be written as a product of multinomial distributions.

We suppose that the process of transition between income quartiles takes place in continuous time; define

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

for all \( i \neq j \), and let \( r_{ii} = -\sum_{j \neq i} r_{ij} \). If the discrete time model is embeddable then \( R = Q \log[A] Q^{-1} \), where \( Q \) is the matrix whose columns are right eigenvectors of \( P \) and \( A \) is the diagonal matrix of corresponding eigenvalues. A given discrete time model is embeddable if and only if the off-diagonal elements of \( Q \log[A] Q^{-1} \) 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 problem 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.

The 9x1 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 \( p_{ij} \)

\textsuperscript{2}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).
appear in ascending order. The split normal importance sampling distribution was constructed as described in Section 4, with \( T \) being the lower triangular Choleski factorization of the asymptotic variance matrix of \( \hat{\Theta} \). Some idea of the behavior of the likelihood surface is conveyed by Figure 4, which shows values of the \( f_i(\delta) \) defined in (7). 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: \( M^R(R) = -0.25\text{tr}(R), \)
\( M^B(R) = 0.25 \sum_{j=1}^{4} \sum_{j=1}^{4} (i-j)r_{ij} \), and \( M^Z(R) = -\log(\lambda_2) \) where \( \lambda_2 \) is the second largest eigenvalue of \( P \). We impose the prior \( \pi(P|Q,X) = 1 \) if \( P \) is embeddable and \( \pi(P|Q,X) = 0 \) 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 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 distributions, was carried out. The normal importance sampling density performed very poorly: \( \omega_1 = 2,329.6 \). The

---

\(^3\)For details, see the Appendix of Geweke, Marshall, and Zarkin (1986a).
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 sampling distribution). This probability, rather than 1.0, provides the norm against which RNE's should be compared, for any importance sampling distribution based on the likelihood function rather than the posterior.

The computed values in Table 4 are essentially the same, for expected values and medians of functions of interest, with the two importance sampling distributions employed. Higher moments, and fractile 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 sampling distribution 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 3.

6. Examples: ARCH Linear Models

In this section we take up an example in which the parameter space \( \Theta \) is not compact. It involves constraints which cannot readily be imposed in classical inference, as discussed in detail elsewhere (Geweke, 1986b). 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 \( x_t : k \times 1 \) and \( y_t \) be
time series for which the distribution of \( y_t \) conditional on \( \psi_{t-1} = \{x_{t-s}, y_{t-s}, s > 1\} \) is

\[
y_t | \psi_{t-1} \sim N(x_t \beta, \gamma_t).
\]

Defining \( \varepsilon_t = y_t - x_t \beta \), take

\[
h_t = \alpha_0 + \sum_{j=1}^{p} \alpha_j \varepsilon_{t-j}^2 = h(\varepsilon_{t-1}, \ldots, \varepsilon_{t-p}; \gamma)
\]

The parameterization in terms of \( \gamma \) allows restrictions like \( \alpha_0 = \gamma_0 \), \( \alpha_j = \gamma_1(p+1-j) \), the linearly declining weights employed by Engle (1982, 1983). For (10) to be plausible it is necessary that \( \alpha_0 > 0 \) and \( \alpha_j > 0 \) (j=1,...,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 \( (x_t, y_t; t=1,\ldots,T) \) the log-likelihood function is (up to an additive constant)

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

Engle (1982) has shown that for large samples \( \frac{\partial^2 l}{\partial \gamma \gamma'} = 0 \), and the maximum likelihood estimates \( \hat{\beta} \) of \( \beta \) and \( \hat{\gamma} \) of \( \gamma \) are asymptotically independent. A scoring algorithm may be used to maximize the likelihood function, and revision of the estimate \( \hat{\gamma}^{(i)} \) of \( \gamma \) and \( \hat{\beta}^{(i)} \) of \( \beta \) 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 \( \hat{\beta} \) and \( \hat{\gamma} \) are asymptotically independent we construct the
importance sampling density as the product of a density for $\theta$ and a density for $\gamma$. To employ Corollary 2(A) and the procedures of Section 4 first compare the tail behavior of the multivariate $t$ density in $m$ dimensions,

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

with that of the likelihood function. Allowing $|x_j| \to \infty$ while fixing all other $x_i$, the log of $f(x; \mu, \varepsilon, \nu)$ behaves like $-(\nu+m)\log|x_j|$. The log likelihood function (11) 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 $\theta$ we shall use a split Student sampling distribution with $T-p-k$ degrees of freedom, and for $\gamma$ we shall use a split Student sampling distribution 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:

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

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

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

The parameterization of the conditional variance process in the model is

$$h_t = \gamma_0 + \gamma_1 (2\varepsilon_{t-1}^2 + \varepsilon_{t-2}^2); \quad \alpha_0 = \gamma_0, \quad \alpha_1 = 2\gamma_1, \quad \alpha_2 = \gamma_1.$$ 

We assume a flat prior on $\theta$ and $\gamma$, 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 5.

A representation of the global behavior of the log-likelihood function for this sample is provided in Figure 5. There are two orthogonal axes in the decomposition $TT'$ of the classical asymptotic variance matrix of $\hat{g}$, and two in the decomposition of the variance matrix of $\hat{Y}$. In the case of $\hat{g}$, global behavior of the likelihood function is quite similar in each direction along the respective axes. In the case of $\hat{Y}$, $\gamma_0$ and $\gamma_1$ remain positive only in one direction along one axis. The log-likelihood in each of these three directions is shown in the figure, as a function of the logarithm of the argument. The log-likelihood along the $\hat{g}$-axes has an asymptotic slope of 198, and along the $\hat{Y}$-axis the slope is 99.

The local behavior of the log-likelihood function is portrayed in Figure 6. 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)) (shaded thick line), and the split Student approximation to the likelihood function constructed as described in Section 4 (thin line). (In the case of $\hat{Y}$, the curves terminate at $\gamma_0 = 0$ or $\gamma_1 = 0$.) For $\hat{g}$ the classical asymptotic theory provides a good approximation to the likelihood function, but for $\hat{Y}$ the approximation is poor.

In Table 6 we compare diagnostics for computational accuracy using the unmodified and split normal densities (which are unjustified theoretically since $E[w(\hat{g})]$ does not exist in these cases) and the unmodified and split Student densities. The unmodified densities perform quite poorly, for reasons made clear in Figure 6. The split normal and split Student sampling densities lead to acceptable performance, in the sense that RME 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 illusion, but from Figure 5 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 $\theta$ and $\gamma$, multivariate importance sampling densities with $v = .5, 1, 1.5, 2, 3, 4, 6,$ and 12 degrees of freedom were used. The resulting diagnostics for numerical accuracy are given in Table 7. RNE for the four parameters attains a maximum at $v = 3$, that for $p$ at $v = 6$. For $v < 3$ the sampling density is too diffuse, and as $v$ becomes larger it approaches the multivariate normal which is much too compact. That $\omega$-diagnostics are sensitive to sampling densities with tails that are too thin rather than too thick is clearly indicated in Table 7. 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 $v$. In the absence of any systematic or theoretical basis for choosing $v$, 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 computing it is practical to obtain the exact posterior distribution 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. 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 distributions 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, and the potential for pathological likelihood functions rises with the number of parameters. 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.
References


1986b, "Mobility Indices in Continuous Time Markov Chains," Econometrica. (Forthcoming.)


TABLE 1

Three Examples for the Two-State Homogeneous Markov Chain Model$^a$

<table>
<thead>
<tr>
<th></th>
<th>Case I</th>
<th>Case II</th>
<th>Case III</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_{11}$</td>
<td>63</td>
<td>21</td>
<td>68</td>
</tr>
<tr>
<td>$m_{12}$</td>
<td>5</td>
<td>66</td>
<td>28</td>
</tr>
<tr>
<td>$m_{21}$</td>
<td>17</td>
<td>6</td>
<td>17</td>
</tr>
<tr>
<td>$m_{22}$</td>
<td>54</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>$\hat{p}_1$</td>
<td>0.087 (0.034)</td>
<td>0.759 (0.046)</td>
<td>0.292 (0.046)</td>
</tr>
<tr>
<td>$\hat{p}_2$</td>
<td>0.239 (0.051)</td>
<td>0.200 (0.073)</td>
<td>0.810 (0.086)</td>
</tr>
<tr>
<td>$\hat{p}_1 + \hat{p}_2$</td>
<td>0.326 (0.061)</td>
<td>0.959 (0.086)</td>
<td>1.102 (0.097)</td>
</tr>
</tbody>
</table>

$^a$Carats denote m.l.e.'s; asymptotic standard errors are shown parenthetically.

Data are taken from Singer and Cohen (1980).
### Table 2

Inference in the Two-State Homogenous Markov Chain Model

<table>
<thead>
<tr>
<th>Importance density</th>
<th>Normal</th>
<th>Split Normal</th>
<th>Likelihood Function</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{E}[g]$</td>
<td>sd[g]</td>
<td>$100 \hat{\sigma}_n^2$</td>
<td>RNE$^c$</td>
</tr>
<tr>
<td>Case I</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_1$</td>
<td>.098</td>
<td>.034</td>
<td>.051</td>
<td>.441</td>
</tr>
<tr>
<td>$p_2$</td>
<td>.246</td>
<td>.049</td>
<td>.056</td>
<td>.769</td>
</tr>
<tr>
<td>$r_1^{-1}$</td>
<td>11.7</td>
<td>4.98</td>
<td>5.09</td>
<td>.960</td>
</tr>
<tr>
<td>$p_2^{-1}$</td>
<td>.262</td>
<td>.926</td>
<td>.970</td>
<td>.911</td>
</tr>
<tr>
<td>$r_1^{-1}*$</td>
<td>9.60</td>
<td>4.35</td>
<td>4.44</td>
<td>.961</td>
</tr>
<tr>
<td>$r_2^{-1}*$</td>
<td>3.49</td>
<td>.887</td>
<td>.932</td>
<td>.905</td>
</tr>
<tr>
<td>Case II</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_1$</td>
<td>.753</td>
<td>.046</td>
<td>.051</td>
<td>.808</td>
</tr>
<tr>
<td>$p_2$</td>
<td>.218</td>
<td>.070</td>
<td>.083</td>
<td>.720</td>
</tr>
<tr>
<td>$p_1^{-1}$</td>
<td>1.33</td>
<td>.083</td>
<td>.096</td>
<td>.759</td>
</tr>
<tr>
<td>$p_2^{-1}$</td>
<td>5.18</td>
<td>2.09</td>
<td>1.92</td>
<td>1.19</td>
</tr>
<tr>
<td>$p_1^{-1}*$</td>
<td>.739</td>
<td>.043</td>
<td>.060</td>
<td>.526</td>
</tr>
<tr>
<td>$p_2^{-1}*$</td>
<td>.183</td>
<td>.050</td>
<td>.064</td>
<td>.629</td>
</tr>
<tr>
<td>$p_1^{-1}*$</td>
<td>1.36</td>
<td>.082</td>
<td>.117</td>
<td>.491</td>
</tr>
<tr>
<td>$p_2^{-1}*$</td>
<td>6.00</td>
<td>2.13</td>
<td>2.21</td>
<td>.928</td>
</tr>
<tr>
<td>$r_1^{-1}$</td>
<td>.472</td>
<td>.129</td>
<td>.170</td>
<td>.576</td>
</tr>
<tr>
<td>$r_2^{-1}$</td>
<td>2.18</td>
<td>1.23</td>
<td>1.30</td>
<td>.891</td>
</tr>
<tr>
<td>Importance density</td>
<td>Normal</td>
<td>Split Normal</td>
<td>Likelihood Function</td>
<td>Actual</td>
</tr>
<tr>
<td>--------------------</td>
<td>---------</td>
<td>--------------</td>
<td>---------------------</td>
<td>--------</td>
</tr>
<tr>
<td></td>
<td>$E[g]$</td>
<td>$sd[g]$</td>
<td>$100 , \sigma_n$ RNE$^c$</td>
<td>$E[g]$</td>
</tr>
<tr>
<td>$g(\cdot)^b$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case III</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p_1$</td>
<td>.296</td>
<td>.045</td>
<td>.051</td>
<td>.810</td>
</tr>
<tr>
<td>$p_2$</td>
<td>.782</td>
<td>.084</td>
<td>.118</td>
<td>.507</td>
</tr>
<tr>
<td>$p_1^{-1}$</td>
<td>3.46</td>
<td>.558</td>
<td>.583</td>
<td>.915</td>
</tr>
<tr>
<td>$p_2^{-1}$</td>
<td>1.29</td>
<td>.153</td>
<td>.246</td>
<td>.387</td>
</tr>
<tr>
<td>$p_1^*$</td>
<td>.269</td>
<td>.041</td>
<td>.145</td>
<td>.080</td>
</tr>
<tr>
<td>$p_2^*$</td>
<td>.669</td>
<td>.062</td>
<td>.257</td>
<td>.057</td>
</tr>
<tr>
<td>$p_1^{-1}$</td>
<td>3.80</td>
<td>.600</td>
<td>1.82</td>
<td>.108</td>
</tr>
<tr>
<td>$p_2^{-1}$</td>
<td>1.51</td>
<td>.150</td>
<td>.686</td>
<td>.048</td>
</tr>
<tr>
<td>$r_1^{-1}$</td>
<td>1.21</td>
<td>.394</td>
<td>.116</td>
<td>.115</td>
</tr>
<tr>
<td>$r_2^{-1}$</td>
<td>.488</td>
<td>.170</td>
<td>.698</td>
<td>.060</td>
</tr>
</tbody>
</table>

$^a$Expected values of functions of interest $g(\cdot)$ 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.
TABLE 3

Some Diagnostics for Computational Accuracy in the Two-State Homogeneous Markov Chain Model

<table>
<thead>
<tr>
<th></th>
<th>Normal Sampling Distribution</th>
<th></th>
<th>Split Normal Sampling Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=10,000</td>
<td>n=50,000</td>
<td>n=10,000</td>
</tr>
<tr>
<td>Case I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNE, ( p_1 )</td>
<td>.441</td>
<td>.269</td>
<td>1.137</td>
</tr>
<tr>
<td>RNE, ( p_2 )</td>
<td>.769</td>
<td>.657</td>
<td>1.014</td>
</tr>
<tr>
<td>( \omega_1 )</td>
<td>186.2</td>
<td>1,774.7</td>
<td>2.5</td>
</tr>
<tr>
<td>( \omega_{10} )</td>
<td>72.9</td>
<td>497.0</td>
<td>2.5</td>
</tr>
<tr>
<td>Case II</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNE, ( p_1 )</td>
<td>.808</td>
<td>.774</td>
<td>1.054</td>
</tr>
<tr>
<td>RNE, ( p_2 )</td>
<td>.720</td>
<td>.564</td>
<td>1.054</td>
</tr>
<tr>
<td>( \omega_1 )</td>
<td>31.2</td>
<td>277.9</td>
<td>1.9</td>
</tr>
<tr>
<td>( \omega_{10} )</td>
<td>17.5</td>
<td>106.3</td>
<td>1.9</td>
</tr>
<tr>
<td>Case III</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNE, ( p_1 )</td>
<td>.810</td>
<td>.790</td>
<td>1.011</td>
</tr>
<tr>
<td>RNE, ( p_2 )</td>
<td>.507</td>
<td>.462</td>
<td>1.032</td>
</tr>
<tr>
<td>( \omega_1 )</td>
<td>101.0</td>
<td>348.7</td>
<td>1.8</td>
</tr>
<tr>
<td>( \omega_{10} )</td>
<td>52.8</td>
<td>161.3</td>
<td>1.8</td>
</tr>
</tbody>
</table>
TABLE 4

Inference in the Quartile Homogeneous Markov Chain Model\textsuperscript{a}

<table>
<thead>
<tr>
<th>Importance Density</th>
<th>Split Normal</th>
<th></th>
<th>Normal</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E[g]$</td>
<td>sd[g]</td>
<td>100 $\sigma_n$</td>
<td>RNE\textsuperscript{b}</td>
</tr>
<tr>
<td>$g(\cdot)$</td>
<td>2.35</td>
<td>.338</td>
<td>.786</td>
<td>.184</td>
</tr>
<tr>
<td>$-r_{11}^{-1}$</td>
<td>1.33</td>
<td>.170</td>
<td>.393</td>
<td>.187</td>
</tr>
<tr>
<td>$-r_{22}^{-1}$</td>
<td>1.33</td>
<td>.162</td>
<td>.380</td>
<td>.182</td>
</tr>
<tr>
<td>$-r_{33}^{-1}$</td>
<td>2.28</td>
<td>.324</td>
<td>.786</td>
<td>.169</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Importance Density</th>
<th>Split Normal Fractile</th>
<th></th>
<th>Normal Fractile</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$g(\cdot)$</td>
<td>.01</td>
<td>.25</td>
<td>.50</td>
</tr>
<tr>
<td></td>
<td>$-r_{11}^{-1}$</td>
<td>1.68</td>
<td>2.10</td>
<td>2.33</td>
</tr>
<tr>
<td></td>
<td>$-r_{22}^{-1}$</td>
<td>.987</td>
<td>1.22</td>
<td>1.32</td>
</tr>
<tr>
<td></td>
<td>$-r_{33}^{-1}$</td>
<td>.979</td>
<td>1.21</td>
<td>1.32</td>
</tr>
<tr>
<td></td>
<td>$-r_{44}^{-1}$</td>
<td>1.60</td>
<td>2.05</td>
<td>2.26</td>
</tr>
<tr>
<td></td>
<td>$-r_{11}$</td>
<td>.304</td>
<td>.391</td>
<td>.428</td>
</tr>
<tr>
<td></td>
<td>$-r_{22}$</td>
<td>.566</td>
<td>.692</td>
<td>.759</td>
</tr>
<tr>
<td></td>
<td>$-r_{33}$</td>
<td>.572</td>
<td>.698</td>
<td>.760</td>
</tr>
<tr>
<td></td>
<td>$-r_{44}$</td>
<td>.319</td>
<td>.403</td>
<td>.443</td>
</tr>
<tr>
<td></td>
<td>$M_1(R)$</td>
<td>.485</td>
<td>.565</td>
<td>.598</td>
</tr>
<tr>
<td></td>
<td>$M_2(R)$</td>
<td>.586</td>
<td>.675</td>
<td>.718</td>
</tr>
<tr>
<td></td>
<td>$M_2(R)$</td>
<td>.317</td>
<td>.435</td>
<td>.810</td>
</tr>
</tbody>
</table>

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

\textsuperscript{b} Relative numerical efficiency.
TABLE 5

Inference in the ARCH Linear Model\(^a\)

<table>
<thead>
<tr>
<th>(g(\cdot))</th>
<th>(E[g])</th>
<th>(sd[g])</th>
<th>(\hat{\sigma}_n)</th>
<th>(RNE^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\beta_1)</td>
<td>1.096</td>
<td>.092</td>
<td>.110</td>
<td>.700</td>
</tr>
<tr>
<td>(\beta_2)</td>
<td>.954</td>
<td>.135</td>
<td>.155</td>
<td>.757</td>
</tr>
<tr>
<td>(\gamma_0)</td>
<td>1.001</td>
<td>.291</td>
<td>.373</td>
<td>.609</td>
</tr>
<tr>
<td>(\gamma_1)</td>
<td>.281</td>
<td>.066</td>
<td>.079</td>
<td>.693</td>
</tr>
<tr>
<td>P[Stable]</td>
<td>.795</td>
<td>-</td>
<td>.475</td>
<td>.722</td>
</tr>
</tbody>
</table>

\(^a\) Expected values and standard deviations of functions of interest \(g(\cdot)\) were computed by Monte Carlo integration with importance sampling from a split Student density, using 10,000 replications.

\(^b\) Since there are 10,000 replications, this is 100 times the numerical standard error.

\(^c\) Relative numerical efficiency.
TABLE 6

Some Diagnostics for Computational Accuracy
ARCH Linear Model

<table>
<thead>
<tr>
<th></th>
<th>Multivariate Normal Sampling Distribution</th>
<th></th>
<th>Split Normal Sampling Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n=10,000</td>
<td>n=50,000</td>
<td>n=10,000</td>
</tr>
<tr>
<td>RNE, $\beta_1$</td>
<td>.006</td>
<td>.016</td>
<td>.686</td>
</tr>
<tr>
<td>RNE, $\beta_2$</td>
<td>.005</td>
<td>.014</td>
<td>.682</td>
</tr>
<tr>
<td>RNE, $\gamma_0$</td>
<td>.001</td>
<td>.002</td>
<td>.564</td>
</tr>
<tr>
<td>RNE, $\gamma_1$</td>
<td>.006</td>
<td>.012</td>
<td>.748</td>
</tr>
<tr>
<td>RNE, $\rho$</td>
<td>.018</td>
<td>.044</td>
<td>.744</td>
</tr>
<tr>
<td>$\omega_1$</td>
<td>9,872.8</td>
<td>39,893.6</td>
<td>51.0</td>
</tr>
<tr>
<td>$\omega_{10}$</td>
<td>994.4</td>
<td>4,857.2</td>
<td>22.3</td>
</tr>
</tbody>
</table>

<p>|                  | Multivariate t Sampling Distribution |                  | Split Student Sampling Distribution |
|                  | n=10,000       | n=50,000       | n=10,000       | n=50,000       |
| RNE, $\beta_1$  | .036          | .068           | .700           | .707           |
| RNE, $\beta_2$  | .027          | .065           | .757           | .742           |
| RNE, $\gamma_0$ | .002          | .006           | .609           | .501           |
| RNE, $\gamma_1$ | .017          | .041           | .692           | .690           |
| RNE, $\rho$     | .049          | .118           | .722           | .741           |
| $\omega_1$      | 9,726.9       | 37,067.8       | 19.9           | 232.3          |
| $\omega_{10}$   | 979.6         | 4,476.4        | 16.0           | 70.9           |</p>
<table>
<thead>
<tr>
<th></th>
<th>$\nu=0.5$</th>
<th>$\nu=1.0$</th>
<th>$\nu=1.5$</th>
<th>$\nu=2.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNE, $\beta_1$</td>
<td>0.144</td>
<td>0.302</td>
<td>0.402</td>
<td>0.451</td>
</tr>
<tr>
<td>RNE, $\beta_2$</td>
<td>0.150</td>
<td>0.307</td>
<td>0.410</td>
<td>0.455</td>
</tr>
<tr>
<td>RNE, $\gamma_0$</td>
<td>0.127</td>
<td>0.243</td>
<td>0.303</td>
<td>0.313</td>
</tr>
<tr>
<td>RNE, $\gamma_1$</td>
<td>0.129</td>
<td>0.262</td>
<td>0.339</td>
<td>0.365</td>
</tr>
<tr>
<td>RNE, $\rho$</td>
<td>0.127</td>
<td>0.282</td>
<td>0.368</td>
<td>0.417</td>
</tr>
<tr>
<td>$\omega_1$</td>
<td>52.2</td>
<td>42.2</td>
<td>41.2</td>
<td>40.2</td>
</tr>
<tr>
<td>$\omega_{10}$</td>
<td>47.2</td>
<td>32.3</td>
<td>29.8</td>
<td>34.2</td>
</tr>
<tr>
<td></td>
<td>$\nu=3.0$</td>
<td>$\nu=4.0$</td>
<td>$\nu=6.0$</td>
<td>$\nu=12.0$</td>
</tr>
<tr>
<td>RNE, $\beta_1$</td>
<td>0.536</td>
<td>0.529</td>
<td>0.487</td>
<td>0.356</td>
</tr>
<tr>
<td>RNE, $\beta_2$</td>
<td>0.540</td>
<td>0.525</td>
<td>0.537</td>
<td>0.430</td>
</tr>
<tr>
<td>RNE, $\gamma_0$</td>
<td>0.336</td>
<td>0.317</td>
<td>0.271</td>
<td>0.197</td>
</tr>
<tr>
<td>RNE, $\gamma_1$</td>
<td>0.410</td>
<td>0.399</td>
<td>0.366</td>
<td>0.292</td>
</tr>
<tr>
<td>RNE, $\rho$</td>
<td>0.493</td>
<td>0.504</td>
<td>0.522</td>
<td>0.495</td>
</tr>
<tr>
<td>$\omega_1$</td>
<td>53.4</td>
<td>69.0</td>
<td>136.3</td>
<td>321.4</td>
</tr>
<tr>
<td>$\omega_{10}$</td>
<td>34.2</td>
<td>42.7</td>
<td>71.2</td>
<td>130.0</td>
</tr>
</tbody>
</table>
Figure 2. Values of $E[w(\theta, x)]$ as a function of the variance of a normal importance sampling distributions, centered at $\theta$, for Case A discussed in Section 5.1.

Figure 3. Variance $\sigma^2$ that equates the ratio of a $N(\theta, \sigma^2)$ density at $\theta$ to that at $0$, with the likelihood ratio $L(\theta|x)/L(0|x)$, for Case A discussed in Section 5.1.
Values of $f_1(\delta)$, quartile homogeneous Markov chain model

Figure 4. Values of $f_1(\delta)$ along nine orthogonal axes of the likelihood function, constructed as described in the text.
Exploration of ARCH LLF

Natural Log of Standard Deviation

Figure 5. Values along the two beta axes are indicated by the solid heavy and light lines, respectively. The shaded line shows values along the single gamma axis that extends away from the boundary of the parameter space.
Figure 6. 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 shaded line is the log multivariate t importance sampling density; and the solid thin line is the log split Student importance sampling density.