IMPORTANCE WEIGHTED MARGINAL BAYESIAN POSTERIOR DENSITY ESTIMATION *

by

Ming-Hui Chen
Purdue University

Technical Report #92-22

Department of Statistics
Purdue University

May 1992
Revised February 1993

* This research was supported in part by NSF grant DMS-8717799 and DMS-8923071 at Purdue University.

*

Importance Weighted Marginal Bayesian Posterior Density Estimation *

Ming-Hui Chen
Department of Statistics
Purdue University

Revised February 1993

Abstract

Markov chain sampling schemes generate dependent observations $\{\underline{\Theta}_i, 0 \leq i \leq n\}$ from a full joint posterior distribution $\pi(\underline{\theta}|data)$. Frequently, only certain marginals of this full posterior density are of interest; thus an interesting problem is how to estimate the marginal posterior densities based on the dependent observations $\{\underline{\Theta}_i, 0 \leq i \leq n\}$ from $\pi(\underline{\theta}|data)$. We propose a new importance weighted marginal density estimation (IWMDE) method. An IWMDE is obtained by averaging many dependent observations of the ratio of the full joint posterior densities multiplied by a weighting conditional density w. The asymptotic properties for the IWMDE and the guidelines for choosing a weighting conditional density w are also considered. A bivariate normal model and a constrained linear multiple regression model are used to illustrate how to derive the IWMDEs for the marginal posterior densities.

Keywords: Conditional density estimation, Kernel density estimation, Markov chain sampling, Monte Carlo, Simulation.

*This research was supported in part by NSF grant DMS-8717799 and DMS-8923071 at Purdue University.

1 Introduction

In Bayesian inference, for a k-dimensional parameter $\underline{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)$, a joint posterior density $\pi(\underline{\theta}|data)$ with the support $S(\underline{\theta})$ is typically known in the form

$$\pi(\underline{\theta}|data) = c(\underline{x})L(\underline{\theta}, \underline{x})\pi(\underline{\theta}),$$

where the *data* is $\underline{x}$, $L(\underline{\theta}, \underline{x})$ is the likelihood function, $\pi(\underline{\theta})$ is a prior, and $c(\underline{x})$ is an unknown normalization constant.

Without knowing the normalization constant $c(\underline{x})$, a dependent sample

$$\{\underline{\Theta}_i = (\Theta_{1,i}, \cdots, \Theta_{k,i}), 0 \leq i \leq n\}$$

from $\pi(\underline{\theta}|data)$ can be generated by a Markov chain sampler, e.g., the Gibbs sampler (Geman and Geman 1984; Gelfand and Smith 1990), the Hit-and-Run (H&R) sampler (Belisle *et al* 1990; Schmeiser and Chen 1991; Chen and Schmeiser 1992), the Gibbs Hit-and-Run (GH&R) sampler (Chen and Deely 1992), or the general Metropolis sampler (Hastings 1970; Tierney 1991; Müller 1991). Under some regularity conditions, by the *Ergodic* Theorem (e.g., see Gelfand and Smith 1990),

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(\underline{\Theta}_i) \stackrel{a.s.}{=} E^{\pi(\underline{\theta}|data)}(h),$$

where $E^{\pi(\underline{\theta}|data)}(h) = \int_{S(\underline{\theta})} h(\underline{\theta})\pi(\underline{\theta}|data)d\underline{\theta}$.

One purpose of Bayesian posterior inference is the calculation and display of marginal densities. Because of the unknown normalization constant $c(\underline{x})$ and the complexity of the joint posterior density function, closed forms for univariate/joint marginal posterior densities are often not available, or are expensively evaluated numerically. However, we can estimate the marginal posterior densities by using $\{\underline{\Theta}_i = (\Theta_{1,i}, \cdots, \Theta_{k,i}), 0 \leq i \leq n\}$ generated by a Markov chain sampler from the full joint posterior density $\pi(\underline{\theta}|data)$.

The kernel density estimation method (e.g., see Silverman 1986) is often used to estimate the marginal densities based on the dependent Markov chain sample $\{\underline{\Theta}_i, 0 \leq i \leq n\}$. For example, the

kernel density estimator for the 1^{st} marginal posterior density $\pi_1(\theta_1^*|data)$ is of the form

$$\hat{\pi}_1(\theta_1^*|data) \stackrel{def}{=} \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{\theta_1^* - \Theta_{1,i}}{h_n}\right),$$

where the kernel K is a bounded density on R^1, h_n is the *bandwidth*, and θ_1^* is a fixed point. For the *i.i.d.* observations, if as $n \to \infty$, $h_n \to 0$ and $nh_n \to \infty$, then $\lim_{n\to\infty} \hat{\pi}_1(\theta_1^*|data) \stackrel{a.s.}{=} \pi_1(\theta_1^*|data)$. However, it is not clear, especially for the dependent Markov chain sample $\{\Theta_{1,i}, 0 \le i \le n\}$, how to choose good candidates of the kernel K and *bandwidth* h_n so that $\hat{\pi}_1(\theta_1^*|data)$ converges to the true density $\pi_1(\theta_1^*|data)$ suitably fast. Even for obtaining the asymptotic convergence results, many strong conditions are required for the dependent observations.

Gelfand, Smith and Lee (1992) proposed an approach to estimate the j^{th} marginal posterior density $\pi_j(\theta_j^*|data)$ based on the Gibbs sampling observations. Suppose that closed form for the conditional density $\pi(\theta_j^*|\theta_l, l \neq j, data)$ is available. Then Gelfand, Smith and Lee (1992) suggested a conditional marginal density estimator (CMDE) of the form

$$\hat{\pi}_j(\theta_j^*|data) = \frac{1}{n}\sum_{i=1}^{n} \pi(\theta_j^*|\underline{\Theta}_{i,(-j)}, data), \tag{1.1}$$

for a given θ_j^*, where $\underline{\Theta}_{i,(-j)} = (\Theta_{1,i}, \cdots, \Theta_{j-1,i}, \Theta_{j+1,i}, \cdots, \Theta_{k,i})$. Under some regularity conditions, the *Ergodic* Theorem yields

$$\lim_{n\to\infty} \hat{\pi}_j(\theta_j^*|data) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \pi(\theta_j^*|\underline{\Theta}_{i,(-j)}, data) \stackrel{a.s.}{=} \pi_j(\theta_j^*|data).$$

As mentioned in Gelfand, Smith and Lee (1992), the CMDE is better than the kernel density estimator under a wide range of loss functions. However, the disadvantage of the above density estimation method is that the closed form for the conditional marginal density has to be known. Usually it is hard to know the closed forms for many marginal Bayesian posterior densities, especially in Bayesian inference with a constrained parameter space. Furthermore, if we are interested in joint marginal posterior densities, it is even harder.

In this paper, we propose an importance weighted marginal density estimation (IWMDE) method, which does not require knowing closed forms for conditional marginal posterior densities. The values of the IWMDE are obtained by averaging many observations of the ratio of the

full posterior densities multiplied by a weighting conditional density w, based on a Markov chain sample $\{\underline{\Theta}_i, 0 \leq i \leq n\}$ from $\pi(\underline{\theta}|data)$.

The outline of this paper is as follows. In Section 2, we present the IWMDE method and discuss the asymptotic properties for the IWMDE. We point out that a CMDE is a special case of the IWMDE, and prove that the CMDE is the best IWMDE in the sense of minimizing the asymptotic variances. In Section 3, we provide the empirical guidelines for the selection of a weighting conditional density w. In Section 4.1, we use a bivariate normal distribution to illustrate the importance weighted marginal density estimation method in a case where the true answer is known and to illustrate that the support of w can be smaller than that of the conditional marginal posterior density. In Section 4.2, we apply this new method to derive the marginal posterior densities of coefficients for a constrained linear multiple regression model. This example illustrates that the IWMDE works well regardless of dimensionality while existing methods do not. Section 5 is a short discussion.

2 Importance Weighted Marginal Density Estimation

In this section, we give the technical details of IWMDE, including the asymptotic convergence properties. Without loss of generality, here we consider estimating the joint marginal posterior density for only the first j ($\leq k$) components of $\underline{\Theta}$.

We first introduce several notations. Let $S(\underline{\theta})$ denote the support of the full joint posterior density $\pi(\underline{\theta}|data)$. Denote

$$\underline{\Theta}_{(j)} = (\Theta_1, \cdots, \Theta_j) \text{ and } \underline{\Theta}_{(-j)} = (\Theta_{j+1}, \cdots, \Theta_k)$$

to be the respective first j and last $k-j$ components of a random vector $\underline{\Theta}$. Let

$$\underline{\theta}_{(j)} = (\theta_1, \cdots, \theta_j) \in R^j \text{ and } \underline{\theta}_{(-j)} = (\theta_{j+1}, \cdots, \theta_k) \in R^{k-j}$$

denote values of $\underline{\Theta}_{(j)}$ and $\underline{\Theta}_{(-j)}$, respectively. The support of the conditional joint marginal posterior density of $\underline{\Theta}_{(j)}$ given $\underline{\Theta}_{(-j)} = \underline{\theta}_{(-j)}$ is denoted by

$$S_j(\underline{\theta}_{(-j)}) \stackrel{def}{=} \{(\theta_1, \cdots, \theta_j) : (\theta_1, \cdots, \theta_j, \theta_{j+1}, \cdots, \theta_k) \in S(\underline{\theta})\}, \tag{2.1}$$

and the subspace of $S(\underline{\theta})$ given the first j components $\underline{\theta}_j^* = (\theta_1^*, \cdots, \theta_j^*)$ is denoted by

$$S_{-j}(\underline{\theta}_{(j)}^*) \stackrel{def}{=} \left\{(\theta_{j+1}, \cdots, \theta_k) : (\theta_1^*, \cdots, \theta_j^*, \theta_{j+1}, \cdots, \theta_k) \in S(\underline{\theta})\right\}. \tag{2.2}$$

Therefore the joint marginal density of $\underline{\Theta}_{(j)}$ is

$$\pi_j(\underline{\theta}_{(j)}^* | data) \stackrel{def}{=} \int_{S_{-j}(\underline{\theta}_{(j)}^*)} \pi(\underline{\theta}_{(j)}^*, \underline{\theta}_{(-j)} | data) d\underline{\theta}_{(-j)}. \tag{2.3}$$

Now we want to estimate $\pi_j(\underline{\theta}_{(j)}^* | data)$ at a fixed point $\underline{\theta}_{(j)}^*$ by using a dependent Markov chain sample $\{\underline{\Theta}_i, 0 \leq i \leq n\}$ from $\pi(\underline{\theta} | data)$. The IWMDE of $\pi_j(\underline{\theta}_{(j)}^* | data)$ is

$$\hat{\pi}_j(\underline{\theta}_{(j)}^* | data) \stackrel{def}{=} \frac{1}{n} \sum_{i=1}^{n} w(\underline{\Theta}_{(j),i} | \underline{\Theta}_{(-j),i}) \frac{\pi(\theta_{(j)}^*, \underline{\Theta}_{(-j),i} | data)}{\pi(\underline{\Theta}_i | data)}, \tag{2.4}$$

where $\underline{\Theta}_{(j),i}$ and $\underline{\Theta}_{(-j),i}$ are the corresponding first j and last $k-j$ components of the i^{th} observation $\underline{\Theta}_i$; w, which plays the role of a weight function, is a conditional density defined on $S_j(\underline{\theta}_{(-j)})$ for a given point $\underline{\theta}_{(-j)} \in R^{k-j}$. Here we reuse the Markov chain sample $\{\underline{\Theta}_i, 0 \leq i \leq n\}$, which was generated from $\pi(\underline{\theta} | data)$ and had nothing to do with the conditional density w, to evaluate an IWMDE $\hat{\pi}_j(\underline{\theta}_{(j)}^* | data)$; and the dependent sample $\{\underline{\Theta}_i, 0 \leq i \leq n\}$ may also be used for other purposes in Bayesian inference, e.g., estimating the posterior means and variances.

In Appendix A we prove that under some regularity conditions

$$\lim_{n \to \infty} \hat{\pi}_j(\underline{\theta}_{(j)}^* | data) \stackrel{a.s.}{=} \pi_j(\underline{\theta}_{(j)}^* | data).$$

The above result says that the IMWDE is asymptotically valid.

The IWMDE is a generalization of the CMDE in Equation (1.1) suggested by Gelfand, Smith and Lee (1992). This result can be observed by choosing

$$w = w(\underline{\theta}_{(j)} | \underline{\theta}_{(-j)}) = \pi(\underline{\theta}_j | \underline{\theta}_{(-j)}, data),$$

which is the conditional marginal posterior density of $\underline{\Theta}_{(j)}$ given $\underline{\Theta}_{(-j)} = \underline{\theta}_{(-j)}$. Also the IWMDE $\hat{\pi}_j(\underline{\theta}_{(j)}^* | data)$ does not depend on the normalization constant $c(\underline{x})$ since in Equation (2.4) the nor-

malization constant is cancelled out in the ratio

$$\frac{\pi(\theta^*_{(j)}, \underline{\Theta}_{(-j),i}|data)}{\pi(\underline{\Theta}_i|data)}.$$

The only requirement for obtaining the asymptotic convergence of the IWMDE is that w is a conditional density on $S_j(\underline{\theta}_{(-j)})$. Therefore we can have many IWMDEs by choosing different w. Now we will prove that the CMDE is the best among all IWMDEs.

Denote

$$\Pi_{j,i}(\underline{\theta}^*_{(j)}|data) \stackrel{def}{=} w(\underline{\Theta}_{(j),i}|\underline{\Theta}_{(-j),i}) \frac{\pi(\theta^*_{(j)}, \underline{\Theta}_{(-j),i}|data)}{\pi(\underline{\Theta}_i|data)},$$

for $i = 1, 2, \cdots, n$. Then $\hat{\pi}_j(\underline{\theta}^*_{(j)}|data)$ is the sample mean of $\{\Pi_{j,i}(\underline{\theta}^*_{(j)}|data), 1 \leq i \leq n\}$. Let $\hat{V}(\Pi_{j,i}(\underline{\theta}^*_{(j)}|data))$ denote the usual sample variance of $\{\Pi_{j,i}(\underline{\theta}^*_{(j)}|data), 1 \leq i \leq n\}$ and let

$$V_w\left(\pi_j(\underline{\theta}^*_{(j)}|data)\right) \stackrel{def}{=}$$
$$\int_{S(\underline{\theta})} \left[w(\underline{\theta}_{(j)}|\underline{\theta}_{(-j)}) \frac{\pi(\underline{\theta}^*_{(j)}, \underline{\theta}_{(-j)}|data)}{\pi(\underline{\theta}|data)} \right]^2 \pi(\underline{\theta}|data) d\underline{\theta} - \left(\pi_j(\underline{\theta}^*_{(j)}|data)\right)^2, \tag{2.5}$$

which is the variance of $\Pi_{j,i}(\underline{\theta}^*_{(j)}|data)$ when $\underline{\Theta}_i$ has the stationary distribution $\pi(\underline{\theta}|data)$. Therefore the *Ergodic* Theorem implies that if $V_w\left(\pi_j(\underline{\theta}^*_{(j)}|data)\right) < \infty$, then

$$\lim_{n\to\infty} \hat{V}(\Pi_{j,i}(\underline{\theta}^*_{(j)}|data)) \stackrel{a.s.}{=} V_w\left(\pi_j(\underline{\theta}^*_{(j)}|data)\right).$$

To simplify notation, let

$$w_C(\underline{\theta}_{(j)}|\underline{\theta}_{(-j)}) = \pi(\underline{\theta}_{(j)}|\underline{\theta}_{(-j)}, data).$$

The following theorem indicates that w_C is the best conditional density in the sense of minimizing $V_w\left(\pi_j(\underline{\theta}^*_{(j)}|data)\right)$.

Theorem 2.1 *If* $V_{w_C}\left(\pi_j(\underline{\theta}^*_{(j)}|data)\right) < \infty$ *and* w *is an arbitrary conditional density on* $S_j(\underline{\theta}_{(-j)})$, *then*

$$V_{w_C}\left(\pi_j(\underline{\theta}^*_{(j)}|data)\right) \leq V_w\left(\pi_j(\underline{\theta}^*_{(j)}|data)\right). \tag{2.6}$$

Proof: See Appendix B. ■

According to Theorem 2.1, when the CMDE is available, it is the best IWMDE. However the

CMDE is often not available due to complexity of the conditional marginal posterior density; or sometimes it is very expensive to compute the CMDE due to the overflow/underflow in implementing the conditional density function. For those cases, we can use an IWMDE instead of the CMDE by choosing a simple weighting conditional density w. The empirical guidelines for choosing such a weighting conditional densities w are given in the next section.

3 Choosing a Weighting Conditional Density w

According to the expression of an IWMDE given in Equation (2.4), the weighting conditional density w seems to be an importance sampling density. However, $w(\underline{\theta}_{(j)}|\underline{\theta}_{(-j)})$ depends on $\underline{\theta}_{(-j)}$; and by Theorem 2.1, a good w should be chosen to have a shape similar to the conditional marginal density given $\underline{\theta}_{(-j)}$, which implies that a good w will vary from one iteration to another. Therefore, the way how to choose a good W is quite different from that how to to choose a good importance sampling density. Thus we can not directly use the existing importance sampling densities (e.g., see Geweke 1989; Rubinstein 1981) as the candidates of w. On the other hand, empirical results show that if a w is chosen to have a shape roughly similar to the conditional marginal density, the IWMDE will converge to the marginal posterior density at a suitable rate. Therefore, it is not necessary to choose a w that precisely mimics the conditional marginal density.

Since the IWMDE is typically used to estimate the complex marginal posterior density, it is extremely difficult to have a universal procedure to choose a good w. Therefore, we consider two cases: unconstrained and constrained parameter spaces. For each case, we use the pre-processing information to choose a w so that it is good in the overall average sense.

If $S(\underline{\theta})$ is a unconstrained parameter space, we first choose a joint importance sampling density that has a shape roughly similar to the joint posterior density. Then a weighting conditional density w for the parameters of interest is chosen as the conditional marginal density of the joint importance sampling density. Here, any importance sampling density is valid as long as its conditional marginal densities are available in closed form.

Guidelines for choosing an importance sampling density can be found in Geweke (1989). For example, we can choose a multivariate normal $N(\underline{\mu}, \Sigma)$ as a joint importance sampling density. Then the conditional marginal density of the fitted multinormal distribution $N(\underline{\mu}, \Sigma)$ is used as a w. Here, $\underline{\mu}$ and Σ can be chosen as the posterior mode and negative Hessian evaluated at $\underline{\mu}$; or

running the Markov chain for a while to get n_0 observations $\{\underline{\Theta}_i, 0 \leq i \leq n_0\}$, then $\underline{\mu}$ and Σ are set by

$$\begin{aligned} \underline{\mu} &= \frac{1}{n_0}\sum_{i=1}^{n_0}\underline{\Theta}_i, \\ \Sigma &= \frac{1}{n_0(n_0-1)}\sum_{i=1}^{n_0}\sum_{l=1}^{n_0}(\underline{\Theta}_i - \underline{\mu})'(\underline{\Theta}_l - \underline{\mu}). \end{aligned}$$

If $S(\underline{\theta})$ is a constrained parameter space, the choice of w is quite complicated. For illustrative purposes, we start to consider choosing a w for estimating one-dimensional posterior marginal density, say, for Θ_1 ; and for simplicity, we assume the support $S_1(\underline{\theta}_{(-1)})$ defined in Equation (2.1) is a finite/infinite interval.

If the support $S_1(\underline{\theta}_{(-1)})$ is a finite interval with two endpoints a_1 and b_1, which are functions of $\theta_2, \cdots, \theta_k$, then we use a simple power-function distribution to mimic the conditional posterior marginal distribution by the fitting moments. The form for the density of the power-function distribution is

$$w = \frac{\alpha(\theta_1 - a_1)^{\alpha-1}}{(b_1 - a_1)^{\alpha}} \text{ or } w = \frac{\alpha(b_1 - \theta_1)^{\alpha-1}}{(b_1 - a_1)^{\alpha}}, \text{ for } a_1 < \theta_1 < b_1. \tag{3.1}$$

The corresponding means of the above power-function distributions are

$$\mu_w = a_1 + \frac{\alpha}{\alpha+1}(b_1 - a_1) \text{ or } \mu_w = a_1 + \frac{1}{\alpha+1}(b_1 - a_1). \tag{3.2}$$

The parameter α and the form of a power-function density can be determined by

(1) obtaining the estimated posterior means $\hat{a}_1$, $\hat{b}_1$ and $\hat{\theta}_1$ for a_1, b_1 and θ_1 using the first few Markov chain iterations or entire simulated Markov chain if possible;

(2) using $w = \frac{\alpha(\theta_1 - a_1)^{\alpha-1}}{(b_1-a_1)^{\alpha}}$ if $\hat{\theta}_1 \geq (\hat{a}_1 + \hat{b}_1)/2$ and $\alpha = (\hat{\theta}_1 - \hat{a}_1)/(\hat{b}_1 - \hat{\theta}_1)$; otherwise, using the second form for w and $\alpha = (\hat{b}_1 - \hat{\theta}_1)/(\hat{\theta}_1 - \hat{a}_1)$.

If the support $S_1(\underline{\theta}_{(-1)})$ is a half-open interval, say, with the form (a_1, ∞), then an exponential distribution is used to fit the conditional distribution. Then w is chosen as

$$w = \lambda e^{-\lambda(\theta_1 - a_1)},$$

where $\lambda = 1/(\hat{\theta}_1 - \hat{a}_1)$ and $\hat{\theta}_1$ and $\hat{a}_1$ are obtained by fitting moments.

The power-function and exponential distributions are used as the candidates of w since the least information, i.e., posterior means, is used; and such a w is also very easy and cheap to compute. If more information, for example, the posterior covariances matrix, is available, then the weighting conditional density w can be chosen to be a beta or gamma density. But, more computing time is required for such w.

The above procedure can be extended from one dimension to higher dimension. For example, suppose the joint marginal posterior density for (θ_1, θ_2) is of interest. Since

$$w(\theta_1, \theta_2 \mid \theta_3, \cdots, \theta_k) = w(\theta_2 \mid \theta_3, \cdots, \theta_k) w(\theta_1 \mid \theta_2, \cdots, \theta_k), \tag{3.3}$$

then the joint weighting conditional density w could be selected as the product of two one-dimensional weighting conditional densities by applying the empirical procedure for the one-dimensional case. The choice of such w is not unique since there are 2! ways to express the joint conditional density as the product of one-dimensional conditional densities. However, a w is required to be roughly similar to the conditional posterior distribution. Therefore, we can use one of these w's; or we can average 2! w's as the joint weighting conditional density since it is difficult to know which w is better.

In the next section, we will use two examples to illustrate how to derive IWMDEs.

4 Illustration of the IWMDE Method

In this section, we illustrate the IWMDE method with a bivariate normal model (an unconstrained parameter space) and with a constrained linear multiple regression model (a constrained parameter space). The bivariate normal example shows that the support of w can be smaller than the support of the conditional distribution and that the IWMDE works well even for small sample sizes when w is chosen as $U(-2,2)$. The constrained linear multiple regression example demonstrates that the IWMDEs are easily derived. In contrast, existing methods, e.g., the CMDE, suffer from numerical problems and are difficult to apply in multidimensions. The second example also shows the computational efficiency of the IWMDE.

4.1 A Bivariate Normal Model

In this example, we apply the IWMDE method to get the marginal density of a bivariate normal distribution $N(\underline{\mu}, \Sigma)$, where $\underline{\mu} = (0,0)$, and

$$\Sigma = \begin{pmatrix} 1 & 0.1 \times \sqrt{2} \\ 0.1 \times \sqrt{2} & 2 \end{pmatrix}. \tag{4.1}$$

Let $\underline{\Theta} = (\Theta_1, \Theta_2) \sim N(\underline{\mu}, \Sigma)$. We used the Gibbs sampler to generate $\{\underline{\Theta}_i, 0 \leq i \leq n\}$ from the above bivariate normal distribution $N(\underline{\mu}, \Sigma)$. Since the distribution of the i^{th} Gibbs iteration converges to the stationary distribution at a geometric rate, the IWMDE of the marginal density converges to the true marginal density almost surely.

In Figure 1, the IWMDEs of the marginal density of Θ_1 obtained based on $n = 50$, $n = 100$, and $n = 500$ Gibbs iterations are displayed. The conditional density w was chosen as the density of a uniform distribution $U(-2,2)$. The absolute differences between the estimated and true marginal densities are less than 0.035 for $n = 50$, 0.024 for $n = 100$, and 0.009 for $n = 500$. So, the IWMDE does work well even for such small sample sizes. Note that the support of the conditional density of Θ_1 is R^1 while the support of w is $(-2,2)$. Therefore the support of w may differ from that of the true conditional density.

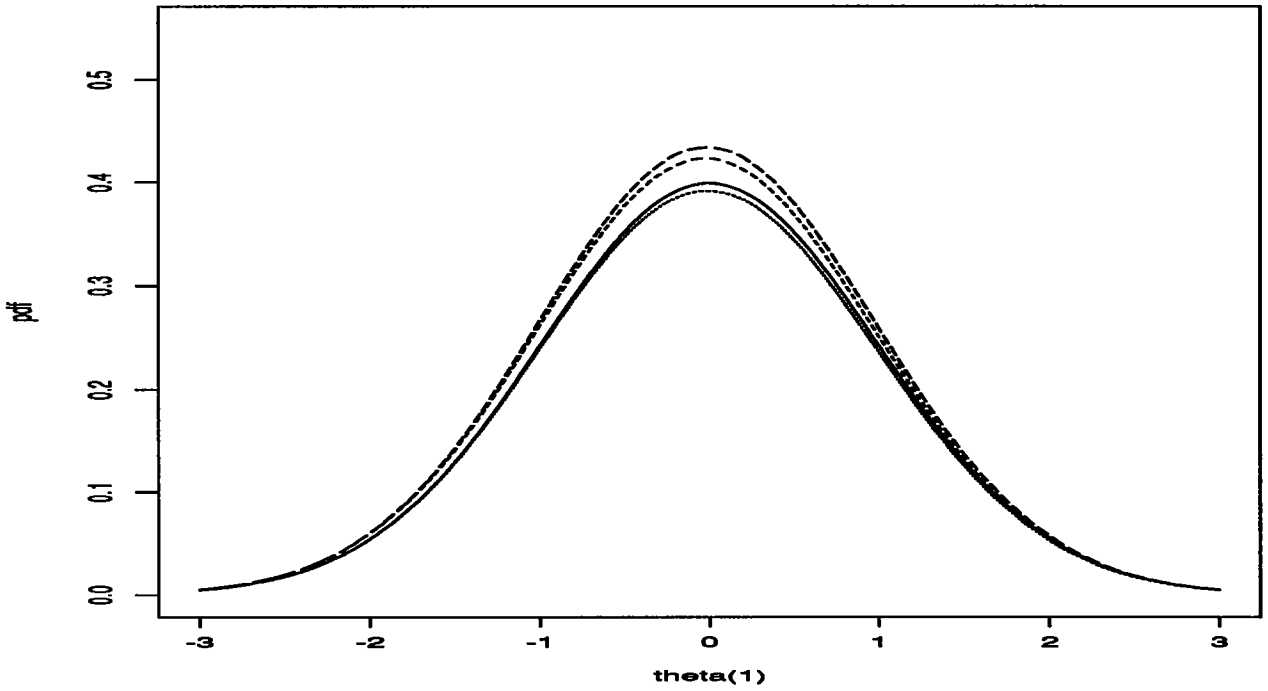

Figure 1: The IWM estimated and true marginal density curves for Θ_1. The solid curve is the true marginal density ($n = \infty$); the dot, dashed, and long dashed curves are the IWMDEs with $n = 500$, 100 and 50 Gibbs iterations, respectively.

4.2 A Constrained Linear Multiple Regression Model

4.2.1 Model and Posterior

A constrained linear multiple regression model considered in Chen and Deely (1992) is

$$y = \sum_{j=1}^{10} \theta_j x_j + \epsilon,$$

with the constraints

$$0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_{10},$$

and $\epsilon \sim N(0, \sigma^2)$.

Denote $\underline{\theta} = (\theta_1, \cdots, \theta_{10}, \theta_{11})$ and $\theta_{11} = \sigma^2$. By choosing a noninformative prior and using the

New Zealand Apple data, Chen and Deely (1992) derived the following full joint posterior density.

$$\pi(\underline{\theta}|data) = c(\underline{x})\frac{1}{(\theta_{11})^{104.5}}\exp\left\{-\frac{1}{2\theta_{11}}\sum_{i=1}^{207}\left(y_i - \sum_{j=1}^{10}\theta_j x_{j,i}\right)^2\right\} I_{S(\underline{\theta})}, \tag{4.2}$$

where $S(\underline{\theta})$ is the support of $\pi(\underline{\theta}|data)$ defined by

$$S(\underline{\theta}) = \{\underline{\theta} : 0 \le \theta_1 \le \theta_2 \le \cdots \le \theta_{10}, \theta_{11} > 0\},$$

$I_{S(\underline{\theta})} = 1$ if $\underline{\theta} \in S(\underline{\theta})$ and 0 otherwise, and $c(\underline{x})$ is an unknown normalization constant. Thus the marginal posterior density of Θ_j is

$$\pi_j(\theta^*_{(j)}|data) = \int_{S_{-j}(\theta^*_{(j)})} \frac{c(\underline{x})}{(\theta_{11})^{104.5}}\exp\left\{-\frac{1}{2\theta_{11}}\sum_{i=1}^{207}\left(y_i - \sum_{l=1}^{10}\theta_l x_{l,i}\right)^2\right\} d\underline{\theta}_{(-j)}, \tag{4.3}$$

where $\underline{\theta}_{(-j)} = (\theta_1, \cdots, \theta_{j-1}, \theta_{j+1}, \cdots, \theta_{11})$ and

$$S_{-j}(\theta^*_{(j)}) = \{\underline{\theta}_{(-j)} : (\theta_1, \cdots, \theta_{j-1}, \theta^*_j, \theta_{j+1}, \cdots, \theta_{11}) \in S(\underline{\theta})\}.$$

4.2.2 The Marginal Posterior Densities for Θ_j

Since $S_{-j}(\theta^*_{(j)})$ is a constrained parameter space, the marginal posterior density of Θ_j is not available in closed form. Therefore the IWMDE method can be applied to obtain the estimator of $\pi_j(\theta^*_{(j)}|data)$. For illustrative purposes, we consider to estimate $\pi_j(\theta^*_{(j)}|data)$ only for $j = 1, 2$.

In this constrained parameter space case, the conditional posterior distribution for Θ_j given the others is a truncated normal distribution with

$$\text{mean} = \mu_j = \frac{\sum_{i=1}^{207}\left(y_i - \sum_{l=1, l\neq j}^{10}\theta_l x_{l,i}\right)}{\sum_{i=1}^{207} x_{j,i}^2} \text{ and } \text{variance} = \sigma_j^2 = \frac{\theta_{11}}{\sum_{i=1}^{207} x_{j,i}^2}. \tag{4.4}$$

Therefore, the normalization constant for this truncated normal density contains the term

$$\left(\Phi\left(\frac{\theta_{j+1} - \mu_j}{\sigma_j}\right) - \Phi\left(\frac{\theta_{j-1} - \mu_j}{\sigma_j}\right)\right)^{-1},$$

where μ_j and σ_j are defined in Equation 4.4, $\theta_{-1} = 0$, and $\Phi(\cdot)$ is the $N(0.1)$ cumulative distribution

function.

When both $(\theta_{j+1}-\mu_j)/\sigma_j$ and $(\theta_{j-1}-\mu_j)/\sigma_j$ are not very large, the CMDE requires to evaluate two $\Phi(\cdot)$ functions. While both $(\theta_{j+1}-\mu_j)/\sigma_j)$ and $(\theta_{j-1}-\mu_j)/\sigma_j)$ are very large with the same sign, then the CMDE can overflow or underflow when computing the conditional density of the truncated normal distribution numerically. But, the IWMDE is more numerically stable since we can choose a simple weighting conditional density w by the empirical procedure given in Section 3. Since the IWMDE requires to know only the properties of the conditional distribution, the IWMDE can be viewed as a "black box" marginal posterior density estimator.

Chen and Deely (1992) derived the posterior means $\hat{\theta}_1 = 0.0131$, $\hat{\theta}_2 = 0.0249$, and $\hat{\theta}_3 = 0.1776$ by using a hybrid Gibbs and Hit-and-Run sampler (GH&R). The support of the conditional posterior density of Θ_1 given $\underline{\Theta}_{(-1)} = \underline{\theta}_{(-1)}$ is

$$S_1(\underline{\theta}_{(-1)}) = \{\theta_1 : 0 \le \theta_1 \le \theta_2\}.$$

For this case, $\hat{a}_1 = 0$, $\hat{b}_1 = 0.0249$, and $\hat{\theta}_1 = 0.0131$. Since $\hat{\theta}_1$ is roughly half of $\hat{\theta}_2$, then w can be chosen as the power-function distribution with $\alpha = 1$, which is a uniform distribution $U(0, \theta_2)$. For Θ_2,

$$S_{\{2\}}(\theta_1, \theta_3, \cdots, \theta_{11}) = \{\theta_2 : 0 \le \theta_1 \le \theta_2 \le \theta_3\}.$$

Then $\hat{a}_1 = 0.0131$, $\hat{b}_1 = 0.1776$, and $\hat{\theta}_2 = 0.0249$. Since $\hat{\theta}_2 < (\hat{a}_1 + \hat{b}_1)/2$, then $\alpha = (\hat{b}_1 - \hat{\theta}_2)/(\hat{\theta}_2 - \hat{a}_1) = 12.94068 \simeq 13$. Thus

$$w(\theta_2|\theta_1, \theta_3, \cdots, \theta_{10}) = \frac{13(\theta_3 - \theta_2)^{12}}{(\theta_3 - \theta_1)^{13}}, \quad \text{for} \quad \theta_1 < \theta_2 < \theta_3. \tag{4.5}$$

In Figure 2, we used 50 GH&R iterations to "warm up" the Markov chain, then used 50,000 GH&R iterations to get the IWMDEs of the marginal posterior densities for Θ_1 and Θ_2 with the uniform $U(0, \theta_2)$ density and the power-function density given in Equation (4.5) as two weighting conditional densities w's. We evaluated the IWMDEs at 101 and 201 griddy points for Θ_1 and Θ_2, respectively. These two w's yielded reasonable convergence rates.

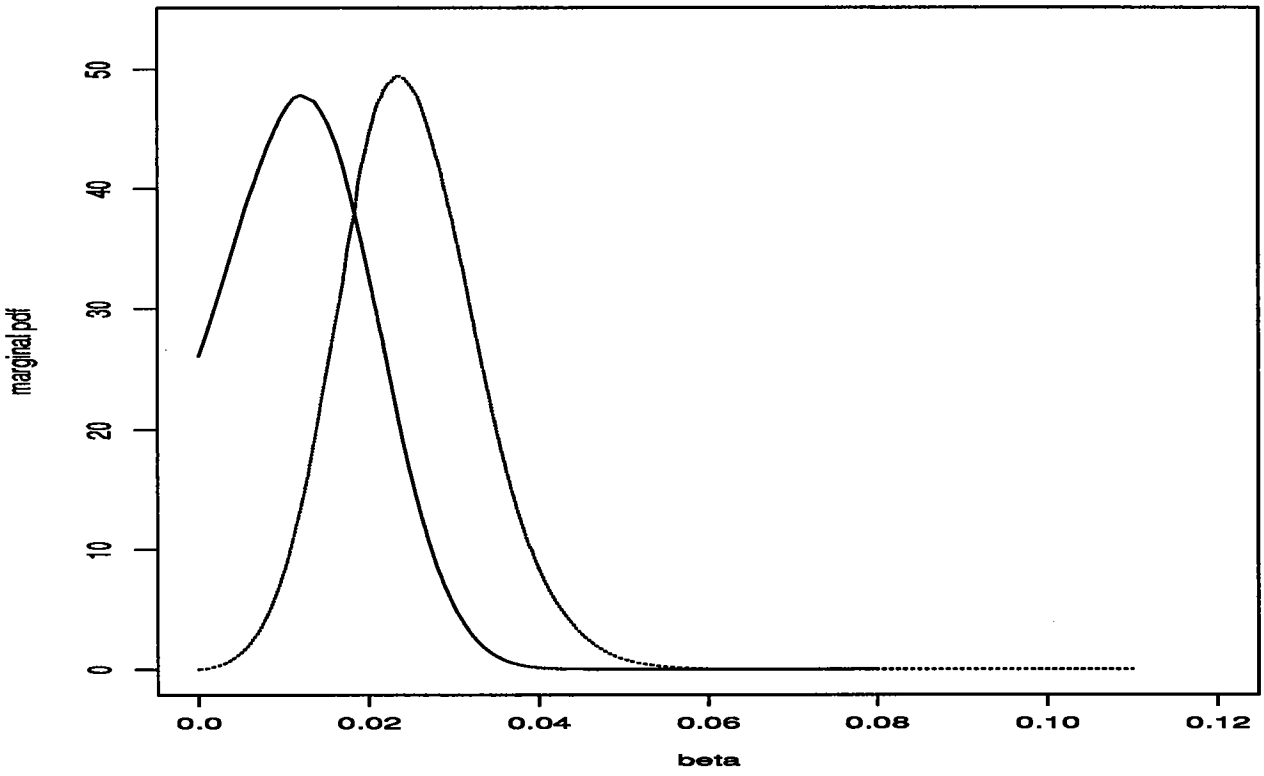

Figure 2: The IWM estimated marginal posterior density curves for Θ_1 and Θ_2. The solid curve is for Θ_1 and the dotted curve is for Θ_2.

4.2.3 The Joint Marginal Posterior Density for Θ_1 and Θ_2

The support of the conditional marginal posterior density for (Θ_1, Θ_2) given $(\Theta_3, \cdots, \Theta_{11}) = (\theta_3, \cdots, \theta_{11})$ is

$$S_2(\theta_3, \cdots, \theta_{11}) = \{(\theta_1, \theta_2) : 0 < \theta_1 \leq \theta_2 \leq \theta_3\}.$$

Even for this two-dimensional normal case, computation of the normalization constant of the conditional marginal posterior density is expensive, and therefore the CMDE is difficult to obtain. In higher dimensions or for nonnormal conditionals, computation is yet worse. However, the IWMDE is still easy to derive.

By using Equation (3.3), a joint weighting conditional density w can be chosen as the product of two one-dimensional weighting conditional densities $w(\theta_1 \mid \theta_2, \cdots, \theta_{11})$ and $w(\theta_2 \mid \theta_3, \cdots, \theta_{11})$.

From Section 4.2.2, we can choose $w(\theta_1 \mid \theta_2, \cdots, \theta_{11}) = 1/\theta_2$. Since $\hat{\theta}_2 = 0.0249$ is less than half of $\hat{\theta}_3 = 0.1776$, then we can use the power-function distribution $\alpha(\theta_3 - \theta_2)^{\alpha-1}/\theta_3^{\alpha}$ as $w(\theta_1 \mid \theta_2, \cdots, \theta_{11})$. By the moment fit, $\alpha = (\hat{\theta}_3/\hat{\theta}_2) - 1 = .1776/.0249 - 1 \simeq 6$. Thus, $w(\theta_1 \mid \theta_2, \cdots, \theta_{11}) = 6(\theta_3 - \theta_2)^5/\theta_3^6$. Therefore

$$w(\theta_1, \theta_2 \mid \theta_3, \cdots, \theta_{11}) = \frac{6(\theta_3 - \theta_2)^5}{\theta_2 \theta_3^6}, \quad \text{for } 0 < \theta_1 \leq \theta_2 \leq \theta_3. \tag{4.6}$$

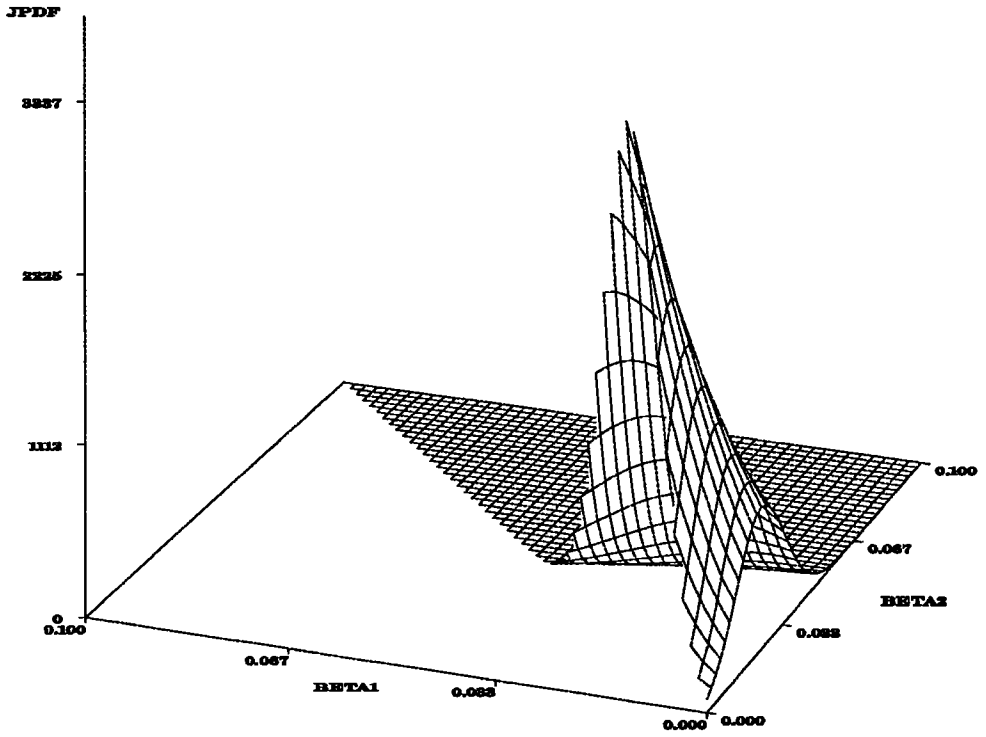

Figure 3: The IWMDE of the joint marginal posterior density for Θ_1 and Θ_2.

In Figure 3, we used 50 GH&R iterations to "warm up" the Markov chain, then used 10,000 GH&R iterations to get the IWMDE of the joint marginal posterior densities for Θ_1 and Θ_2 with the weighting conditional density w given in Equation (4.6). We evaluated the IWMDE at 2500 griddy points for Θ_1 and Θ_2. A similar figure was obtained, but is not shown here, using the w chosen by the product of two one-dimensional densities $w(\theta_2 \mid \theta_1, \theta_3, \cdots, \theta_{11})$ and $w(\theta_1 \mid \theta_3, \cdots, \theta_{11})$.

5 Discussion

When the normalization constant for the full conditional distribution of the parameters of interest can not been evaluated, the CMDE may not be directly applied. Instead of the direct CMDE, the grid method is often applied to estimate the marginal density of these parameters. The procedure is that an appropriate grid of values is selected at which to estimate the density. For each iteration i of the Monte Carlo Markov chain, the unnormalizing conditional density of the parameters of interest given the variates from iteration i for the other parameters is calculated for each point on the grid. The normalization constant for iteration i is estimated as the sum over the grid of the conditional density times the mesh of the grid.

The grid method has several disadvantages. Firstly, it is not easy to choose an appropriate grid, and it is especially difficult to do so in higher dimension cases. Secondly, it is expensive since the normalization constant for each iteration has to be estimated. For example, for the constrained linear multiple regression problem, when the support of the truncated normal density is far away from the mean, it is not easy to accurately evaluate the normal density and the normalization constant. Thirdly, the asymptotic consistency for the grid method is not clear.

Compared to the grid method, IWMDE is easier to implement and has theoretical asymptotical results. Especially, the IWMDE works well for estimating higher-dimensional marginal posterior densities while most existing methods can not.

A final issue is how to determine whether the choice of w is good. In general, it is very difficult to know when a good choice of w has been made. One method is to monitor the area under the estimated density. If it is close to one, then the choice of w can be viewed as a good one. Another method is to estimate the numerical standard deviations at the grid points. The better choice of w will result in the smaller sum of the standard deviations over the grid times the mesh of the grid.

Appendix A: Asymptotic Convergence of the IWMDE

Theorem *If a Markov chain sample* $\{\underline{\Theta}_i, 0 \le i \le n\}$ *from* $\pi(\underline{\theta}|data)$ *is generated by a Markov chain sampler which includes the Gibbs sampler, the H&R sampler, the GH&R sampler, and the Metropolis sampler, and* w *is a conditional density on* $S_j(\underline{\theta}_{(-j)})$*, then for every fixed point* $\underline{\theta}^*_{(j)} \in R^j$*,*

$$\begin{aligned}\lim_{n\to\infty} \hat{\pi}_j(\underline{\theta}^*_{(j)}|data) &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} w(\underline{\Theta}_{(j),i}|\underline{\Theta}_{(-j),i}) \frac{\pi(\underline{\theta}^*_{(j)}, \underline{\Theta}_{(-j),i}|data)}{\pi(\underline{\Theta}_i|data)} \\ &\overset{a.s.}{=} \pi_j(\underline{\theta}^*_{(j)}|data).\end{aligned}$$

Proof: According to *Ergodic* Theorem, e.g., Proposition 4.3 in Schmeiser and Chen [1991],

$$\begin{aligned}\lim_{n\to\infty} \hat{\pi}_j(\underline{\theta}^*_{(j)}|data) &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} w(\underline{\Theta}_{(j),i}|\underline{\Theta}_{(-j),i}) \frac{\pi(\underline{\theta}^*_{(j)}, \underline{\Theta}_{(-j),i}|data)}{\pi(\underline{\Theta}_i|data)} \\ &\overset{a.s.}{=} \int_{S(\underline{\theta})} w(\underline{\theta}_{(j)}|\underline{\theta}_{(-j)}) \frac{\pi(\underline{\theta}^*_{(j)}, \underline{\theta}_{(-j)}|data)}{\pi(\underline{\theta}|data)} \pi(\underline{\theta}|data) d\underline{\theta}. \qquad \text{(A.1)}\end{aligned}$$

Thus by simplifying the right hand side of Equation (A.1) and Fubini's Theorem,

$$\begin{aligned}&\text{The RHS of Equation (A.1)} = \\ &\int_{S_{-j}(\underline{\theta}^*_{(j)})} \pi(\underline{\theta}^*_{(j)}, \underline{\theta}_{(-j)}|data) \left\{ \int_{S_j(\underline{\theta}_{(-j)})} w(\underline{\theta}_{(j)}|\underline{\theta}_{(-j)}) d\underline{\theta}_{(j)} \right\} d\underline{\theta}_{(-j)}.\end{aligned}$$

Since w is a conditional density on $S_j(\underline{\theta}_{(-j)})$, Theorem directly follows from Equation (2.3). ■

Appendix B: Proof of Theorem 2.1

Proof: If $V_w\left(\pi_j(\underline{\theta}^*_{(j)}|data)\right) = \infty$, Inequality (2.6) automatically holds. Now we assume

$$V_w\left(\pi_j(\underline{\theta}^*_{(j)}|data)\right) < \infty.$$

Since the MCDE is a special case of the IWMDE, by Equation (2.5), it suffices to prove that

$$\int_{S(\underline{\theta})} \left[\frac{w_C(\underline{\theta}_{(j)}|\underline{\theta}_{(-j)}) \pi(\underline{\theta}^*_{(j)}, \underline{\theta}_{(-j)}|data)}{\pi(\underline{\theta}|data)} \right]^2 \pi(\underline{\theta}|data) d\underline{\theta}_{(j)} d\underline{\theta}_{(-j)}$$

$$\leq \int_{S(\underline{\theta})} \left[\frac{w(\underline{\theta}_{(j)}|\underline{\theta}_{(-j)})\pi(\underline{\theta}^*_{(j)}, \underline{\theta}_{(-j)}|data)}{\pi(\underline{\theta}|data)} \right]^2 \pi(\underline{\theta}|data) d\underline{\theta}_{(j)} d\underline{\theta}_{(-j)}. \quad \text{(B.1)}$$

Denote $\pi(\underline{\theta}_{(-j)}|data)$ to be the marginal posterior density of $\underline{\Theta}_{(-j)}$. Then

$$w_C(\underline{\theta}_{(j)}|\underline{\theta}_{(-j)}) = \frac{\pi(\underline{\theta}|data)}{\pi(\underline{\theta}_{(-j)}|data)}.$$

Thus

$$\begin{aligned} & \text{The LHS of Inequality (B.1)} \\ = & \int_{S_{-j}(\underline{\theta}^*_{(j)})} \frac{\pi^2(\underline{\theta}^*_{(j)}, \underline{\theta}_{(-j)}|data)}{\pi^2(\underline{\theta}_{(-j)}|data)} \left[\int_{S_j(\underline{\theta}_{(-j)})} \pi(\underline{\theta}|data) d\underline{\theta}_{(j)} \right] d\underline{\theta}_{(-j)} \\ = & \int_{S_{-j}(\underline{\theta}^*_{(j)})} \frac{\pi^2(\underline{\theta}^*_{(j)}, \underline{\theta}_{(-j)}|data)}{\pi(\underline{\theta}_{(-j)}|data)} d\underline{\theta}_{(-j)}. \quad \text{(B.2)} \end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned} 1 = & \left[\int_{S_j(\underline{\theta}_{(-j)})} w(\underline{\theta}_{(j)}|\underline{\theta}_{(-j)}) d\underline{\theta}_{(j)} \right]^2 \\ = & \left[\int_{S_j(\underline{\theta}_{(-j)})} \sqrt{\pi(\underline{\theta}|data)} \frac{w(\underline{\theta}_{(j)}|\underline{\theta}_{(-j)})}{\sqrt{\pi(\underline{\theta}|data)}} d\underline{\theta}_{(j)} \right]^2 \\ \leq & \left[\int_{S_j(\underline{\theta}_{(-j)})} \pi(\underline{\theta}|data) d\underline{\theta}_{(j)} \right] \left[\int_{S_j(\underline{\theta}_{(-j)})} \frac{w^2(\underline{\theta}_{(j)}|\underline{\theta}_{(-j)})}{\pi(\underline{\theta}|data)} d\underline{\theta}_{(j)} \right] \\ = & \pi(\underline{\theta}_{(-j)}|data) \left[\int_{S_j(\underline{\theta}_{(-j)})} \frac{w^2(\underline{\theta}_{(j)}|\underline{\theta}_{(-j)})}{\pi(\underline{\theta}|data)} d\underline{\theta}_{(j)} \right]. \quad \text{(B.3)} \end{aligned}$$

Thus Theorem 2.1 follows from Equations (B.2) and (B.3). ■

Acknowledgement

The author is grateful to Professor James O. Berger (the author's advisor) and Bruce W. Schmeiser of Purdue University and Professor John J. Deely of Canterbury University for many helpful discussions.

References

Belisle, Claude J.P., Romeijn, H. Edwin and Smith, Robert L. (1993), "Hit-and-Run Algorithms for Generating Multivariate Distributions," *Mathematics of Operations Research, 18, 2,* forthcoming.

Chen, Ming-Hui and Deely, John (1992), "Application of a New Gibbs Hit-and-Run Sampler to a Constrained Linear Multiple Regression Problem," Technical Report 92-21, Purdue University, Center for Statistical Decision Sciences and Department of Statistics.

Chen, Ming-Hui and Schmeiser, B.W. (1992), "Performance of the Gibbs, Hit-and-Run, and Metropolis Samplers," *The Journal of Computational and Graphical Statistics*, tentatively accepted.

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

Gelfand, A. E., Smith, A.F.M. and Lee, T.M. (1992), "Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling," *Journal of American Statistical Association, 87*, 523-532.

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

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

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

Müller, P. (1991), "A Generic Approach to Posterior Integration and Gibbs Sampling," Technical Report 91-09, Purdue University, Department of Statistics.

Rubinstein, R. Y. (1981), *Simulation and the Monte Carlo Method*, John Wiley & Sons, New York.

Schmeiser, Bruce W. and Chen, Ming-Hui (1991), "On Hit-and-Run Monte Carlo Sampling for Evaluating Multidimensional Integrals," revision of Technical Report 91-39, Purdue University, Department of Statistics.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Tierney, Luke (1991), "Markov Chains for Exploring Posterior Distributions," Technical Report No. 560, University of Minnesota, School of Statistics.