Methodology and Computing in Applied Probability, 4, 5–25, 2002
© 2002 Kluwer Academic Publishers. Manufactured in The Netherlands.

A Random-Discretization Based Monte Carlo Sampling Method and its Applications

JAMES C. FU james_fu@umanitoba.ca
Department of Statistics, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2

LIQUN WANG liqun_wang@umanitoba.ca
Department of Statistics, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2

Received January 29, 2001; Accepted January 30, 2002

Abstract. Recently, several Monte Carlo methods, for example, Markov Chain Monte Carlo (MCMC), importance sampling and data-augmentation, have been developed for numerical sampling and integration in statistical inference, especially in Bayesian analysis. As dimension increases, problems of sampling and integration can become very difficult. In this manuscript, a simple numerical sampling based method is systematically developed, which is based on the concept of random discretization of the density function with respect to Lebesgue measure. This method requires the knowledge of the density function (up to a normalizing constant) only. In Bayesian context, this eliminates the ''conjugate restriction'' in choosing prior distributions, since functional forms of full conditionals of posterior distributions are not needed. Furthermore, this method is non-iterative, dimension-free, easy to implement and fast in computing time. Some benchmark examples in this area are used to check the efficiency and accuracy of the method. Numerical results demonstrate that this method performs well for all these examples, including an example of evaluating the small probability values of a high dimensional multivariate normal distribution. As a byproduct, this method also provides an easy way of computing maximum likelihood estimates and modes of posterior distributions.

Keywords: random sample generation, Monte Carlo integration, approximate maximum likelihood estimates, high dimensional distribution, compact support, discretization, contourization, law of large numbers

AMS 1991 Subject Classification: Primary 65C05, 65C60, Secondary 65C10, 62F40, 62F15

1. Introduction

Statistical inferences often involve three types of problems:

1. Generating a set of observations
$$x^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \ldots, x_k^{(i)}\right), i = 1, 2, \ldots, m$$
from a multivariate population F, which has a density function $f(x)$ with respect to Lebesgue measure μ and a support $S(f)$ being a subset of $\mathbb{R}^k$.
2. Integrating a function $H(x)$ with respect to a density function $f(x)$, i.e.,
$$\int_{S(f)} H(x)f(x)\mu(dx). \tag{1}$$

3. Finding the maximizer x_0 of a function $L(x)$ over a subset $A \subset \mathbb{R}^k$, such that

$$L(x_0) = \sup_{x \in A} L(x).$$

These problems are related through the use of sampling-based integration and optimization methods. Recently, several Monte Carlo methods for problems (1) and (2) have been developed, e.g., Markov Chain Monte Carlo (MCMC), data-augmentation and importance sampling. Good introductions and reviews to these topics can be found in, for example, Tanner (1992), Evans and Swartz (1995, 2000), Chib and Greenberg (1996), as well as Robert and Casella (1999). Each of these sampling-based methods has advantages and disadvantages. Compared with numerical analytic methods, the attraction of stochastic sampling-based methods is their conceptual simplicity and ease of implementation. For a given integration problem, however, it is well known that the sampling-based methods may not be as accurate or efficient as numerical analytic methods, especially when the dimension is low or moderate.

While MCMC methods are general tools for multivariate random sample generation, once applied to real problems their actual implementations are often quite involved. Major difficulties are associated with multi-modality of the underlying distribution, ill-shaped sample space, as well as convergence of the iterative process. Consequently, a reparameterization or transformation is often needed to ensure the success of a particular algorithm. The choice of a proper reparameterization or transformation remains a difficult task. Moreover, as dimension k increases, difficulties of both problems (1) and (2) increase significantly.

In this manuscript, we focus on an alternative approach to multivariate random sample generation. This method is based on the combination of numerical and sampling-based approaches. It involves analytical approximation of the density function, random discretization and contourization of an empirical space induced by samples from the uniform distribution on $[0, 1]$, and sampling observations from contours of the empirical space according to the discretized density function. Theoretically speaking, the foundation of this approach is based on the concept of randomly discretizing the density function which is a Radon-Nikodym derivative with respect to Lebesgue measure. One of the great strengths of this method is that it requires only the knowledge of the functional form of the density function up to an unknown normalizing constant. Therefore, it applies equally well to ''well-behaved'' distributions as to more complicated, multi-modal distributions. It is dimension-free and non-iterative, which makes it efficient and fast in computation. Finally, it is very easy to implement and is applicable to most high dimensional multivariate distributions arising in real applications. While the primary purpose of the algorithm in this paper is random sample generation, it can also be applied as by-products to multivariate integration, importance sampling and optimization problems.

The paper is organized as follows: In Section 2, we set up notations and give some preliminary results which lay down a theoretical foundation for our method. Section 3 presents the algorithm for the method. Section 4 gives some examples to illustrate the algorithm. In Section 5, the method is applied to several benchmark examples in statistics, especially in Bayesian analysis. It is also shown here that this method can also be used to search the maximum likelihood estimates (MLE) and modes of posterior distributions.

Section 6 discusses some computational aspects and technical issues associated with random sample generation and integration.

2. Notations and Preliminary Results

We consider a random variable $X : (\Omega, \mathscr{A}, P) \to \left(\mathbb{R}^k, \mathscr{B}, \mu\right)$, where probability measure P is absolutely continuous with respect to Lebesgue measure μ and the Radon-Nikodym derivative, or density function, is $f = dP/d\mu$. Throughout this manuscript, $x = (x_1, x_2, \ldots, x_k)$ represents a point in $\mathbb{R}^k$ and the support $S(f)$ of the density function $f(x)$ is a subset of $\mathbb{R}^k$.

A function $f(x)$ is called a simple function, if $f(x)$ takes l distinct values on $S(f)$, i.e.,

$$f(x) = \sum_{i=1}^{l} a_i I(x, E_i),$$

where $-\infty < a_1 < a_2 < \cdots < a_l < \infty$, the indicator function $I(x, E_i) = 1$, if $x \in E_i$, and zero otherwise, and the μ-measurable (Lebesgue measurable) subsets $E_1, E_2, \ldots, E_l$ form a partition of $S(f)$ in the sense that $\cup_{i=1}^{l} E_i = S(f)$ and $E_i \cap E_j = \emptyset$ for all $i \neq j$.

For a density function $f : S(f) \to (0, \infty)$, there exists a sequence of simple functions, say $\{S_l\}$, on $S(f)$, such that

(1) for every l, $S_l(x)$ takes l distinct values $0 < a_{l1} < a_{l2} < \cdots < a_{ll}$, where $S_l(x) = a_{li}$, for $x \in E_{li}$, $\cup_{i=1}^{l} E_{li} = S(f)$; and
(2) for every $x \in S(f)$, $S_l(x) \to f(x)$ as $l \to \infty$.

Since density function $f(x)$ is Lebesgue measurable and integrable, the above results (1) and (2) are immediate consequences of Theorem 1.17 of Rudin (1966, p. 15). Note that, the sequence $\{S_l\}$ may not be density functions. However, the modified functions

$$g_l(x) = \frac{S_l(x)}{\sum_{i=1}^{l} a_i \mu(E_i)}, \quad \text{for } x \in S(f),\ l = 1, 2, \ldots \tag{2}$$

are simple density functions defined on $S(f)$ and satisfy

$$g_l(x) \to f(x), \quad \text{as } l \to \infty, \quad \text{for all } x \in S(f). \tag{3}$$

In the following, it shows that $g_l(x)$ can be constructed in a special way such that it converges to $f(x)$ uniformly on $S(f)$.

THEOREM 1 *Suppose density function $f(x)$ is continuous on a compact support $S(f) \subset \mathbb{R}^k$ (a closed and bounded subset of $\mathbb{R}^k$) and satisfies $\mu\{x \in S(f) : f(x) = b\} = 0$ for any constant $b \in (0, \infty)$. Then the following statements hold.*

1. *For any given $l \in \mathbb{N}$, there exists constants $\inf_{x \in S(f)} f(x) = b_0 < b_1 < \cdots < b_{l-1} < b_l = \sup_{x \in S(f)} f(x)$, such that subsets $E_i = \{x \in S(f) : b_{i-1} < f(x) \leq b_i\}$, $i = 1, 2, \ldots, l$ form a partition of $S(f)$ and satisfy $\mu(E_i) = \mu(S(f))/l$.*
2. *Define*

$$a_i = \frac{1}{\mu(E_i)} \int_{E_i} f(x)\mu(dx), \quad \text{for } i = 1, 2, \ldots, l. \tag{4}$$

Then the sequence of simple density functions $g_l(x) = \sum_{i=1}^{l} a_i I(x, E_i) \to f(x)$ *uniformly on* $S(f)$ as $l \to \infty$.

3. *If* P_l *is the probability measure corresponding to density* g_l*, then the conditional distribution given* E_i *is a uniform distribution on* E_i *with density*

$$g_l(x \mid E_i) = \frac{l}{\mu(S(f))}. \tag{5}$$

Proof:

1. For any given $l \in \mathbb{N}$, under the conditions of the theorem, $\mu\{x \in S(f) : f(x) \le b\}$ is continuous and monotone increasing in $b \in (b_0, b_l)$. It follows that, there exists a constant $b_0 < b_1$, such that $E_1 = \{x \in S(f) : b_0 < f(x) \le b_1\}$ satisfies $\mu(E_1) = \mu[S(f)]/l$. Let b_1 be the smallest number with this property. By the same reason, there exists a constant $b_1 < b_2$, such that $E_2 = \{x \in S(f) : b_1 < f(x) \le b_2\}$ satisfies $\mu(E_2) = \mu[S(f)]/l$. Again let b_2 be the smallest number with this property. Similarly, constants $b_2 < b_3 < \cdots < b_l$ can be defined. By definition $E_i, i = 1, 2, \ldots, l$ satisfy $\cup_{i=1}^{l} E_i = S(f)$ and $E_i \cap E_j = \emptyset$ for all $i \ne j$.
2. Since $f(x)$ is continuous and $S(f)$ is compact, constants $b_i, i = 1, 2, \ldots, l$ defined above satisfy $\max_{1 \le i \le l} \Delta b_i \to 0$, as $l \to \infty$. The result follows then from the fact that $|g_l(x) - f(x)| \le \max_{1 \le i \le l} \Delta b_i$ for all $x \in S(f)$.
3. Given E_i, the result follows immediately from $P_l(E_i) = a_i \mu(E_i)$ and the definition of the conditional distribution. ■

In view of Equation (5), the simple density function $g_l(x)$ can be written as

$$g_l(x) = \sum_{i=1}^{l} I(x, E_i) P_l(E_i) g_l(x \mid E_i), x \in S(f), \tag{6}$$

where $I(x, E_i) = 1$ if $x \in E_i$ and zero otherwise. The above Equations (5) and (6) lay a foundation for our algorithm of generating an observation x from g_l by, first, generating a subset E_i randomly according to probabilities $\{P_l(E_i)\}_{i=1}^{l}$; and, then, generating an observation x from the uniform distribution on E_i. By Theorem 1(2), for very large l, observation x can be considered approximately as an observation generated from density $f(x)$.

3. The Algorithm

In order to make the algorithm more transparent, we start with the case where density function $f(x)$ is continuous with a compact support $S(f)$ in $\mathbb{R}^k$ and the mathematical form of $f(x)$ is known up to a normalizing constant. Without loss of generality, we assume $S(f) = [a, b]^k$, where $-\infty < a < b < \infty$ are known. To simplify notation, we also assume

that $\mu\{x \in S(f) : f(x) = c\} = 0$ for any constant $c \in (0, \infty)$. The algorithm consists of the following three major steps.

3.1. *Discretization*

First, we generate kn independent points from the uniform distribution on $[a, b]$, $x_1, x_2, \ldots, x_{kn} \sim U[a, b]$. Then, every k consecutive observations are combined as an observation in $[a, b]^k$. Hence observations $x^{(j)} = (x_{(j-1)k+1}, x_{(j-1)k+2}, \ldots, x_{jk})$, $j = 1, 2, \ldots, n$, are independent, identically and uniformly distributed on $[a, b]^k$. Define $S_n(f) = \{x^{(j)}, j = 1, 2, \ldots, n\}$. For a given large n, set $S_n(f)$ can be viewed as a uniformly and randomly discretized $S(f)$. Theoretically speaking, as $n \to \infty$, $S_n(f)$ approximates $S(f)$.

3.2. *Contourization*

Let $x^{[j]}, j = 1, 2, \ldots, n$ be the ordered list of $x^{(j)}, j = 1, 2, \ldots, n$ according to the height of density function, such that if $i > j$, then $f(x^{[i]}) \geq f(x^{[j]})$. Clearly $S_n(f)$ can also be written as $S_n(f) = \{x^{[j]}, j = 1, 2, \ldots, n\}$. Given $l \in \mathbb{N}$, we partition $S_n(f)$ into l contours

$$\tilde{E}_i = \{x^{[j]} : (i-1)u < j \leq iu\}, i = 1, 2, \ldots, l,$$

where $u = n/l$ which is assumed to be an integer without loss of generality. It is easy to see that $\cup_{i=1}^{l} \tilde{E}_i = S_n(f)$ and $\tilde{E}_i \cap \tilde{E}_j = \emptyset$ for all $i \neq j$. Define a sequence of constants

$$\tilde{a}_i = \frac{\sum_{j=(i-1)u+1}^{iu} f(x^{[j]})}{u \sum_{j=1}^{n} f(x^{[j]})}, i = 1, 2, \ldots, l \tag{7}$$

and a discrete probability distribution on $S_n(f)$

$$\tilde{g}_l(x) = \sum_{i=1}^{l} I(x, \tilde{E}_i) P_l(\tilde{E}_i) \tilde{g}_l(x \mid \tilde{E}_i), \tag{8}$$

where $P_l(\tilde{E}_i) = \tilde{a}_i \mu^*(\tilde{E}_i)$, μ^* is the Lebesgue counting measure on $S_n(f)$, and

$$\tilde{g}_l(x \mid \tilde{E}_i) = \frac{1}{u} \tag{9}$$

is the conditional distribution of x given $x \in \tilde{E}_i$.

3.3. *Sampling*

Suppose we would like to generate m independent and identically distributed observations from $S_n(f)$ according to distribution $\tilde{g}_l(x)$. First, we sample m subsets with replacement from $\{\tilde{E}_i\}_{i=1}^{l}$ according to probabilities $\{P_l(\tilde{E}_i)\}_{i=1}^{l}$ respectively. Denote by m_i the number of occurrence of $\tilde{E}_i$ in the m draws, where $\sum_{i=1}^{l} m_i = m$. Then, for each

$i = 1, 2, \ldots, l$, we sample with replacement m_i observations randomly within contour $\tilde{E}_i$ and denote the set of these observations by $\tilde{O}_i$. Finally, the set of observations $R_m(f) = \cup_{i=1}^{l} \tilde{O}_i = \{y^{[j]}, j = 1, 2, \ldots, m\}$ is the desired sample. It is easy to see that each observation $y^{[j]}$ obtained by this two-stage sampling procedure is equivalent to an observation drawn directly from $S_n(f)$ according to distribution $\tilde{g}_l(x)$, and, moreover, all observations are independent.

Note that $\tilde{O}_i$ may not have all the points of $\tilde{E}_i$, but may contain several duplications of one point in $\tilde{E}_i$ due to sampling with replacement. Furthermore, since $\tilde{O}_i \subset \tilde{E}_i$, contours $\{\tilde{O}_i\}_1^l$ form a partition of $R_m(f)$. For the sample generated by the above algorithm, we have the following asymptotic results.

THEOREM 2 *Under the assumptions of Theorem 1, it holds*

1. *For any $l \geq 1$ and $1 \leq i \leq l$, as $n \to \infty$, $\tilde{a}_i \mu^*(\tilde{E}_i) \xrightarrow{P} a_i \mu(E_i)$, where $\xrightarrow{P}$ stands for convergence in probability.*
2. *For any $H(x) \in L^1_\mu(S(f))$, as $n \to \infty$, $m \to \infty$ and $l \to \infty$,*

$$\frac{1}{m} \sum_{j=1}^{m} H\left(y^{[j]}\right) \xrightarrow{P} \int_{S(f)} H(x) f(x) \mu(dx).$$

Proof:

1. For any $l \geq 1$ and $1 \leq i \leq l$, since f is continuous and $x^{(j)}$ are uniformly distributed on $S(f)$,

$$\tilde{a}_i \mu^*(\tilde{E}_i) = \frac{\sum_{j=(i-1)u+1}^{iu} f(x^{[j]})}{\sum_{j=1}^{n} f(x^{[j]})} = \frac{\sum_{j=1}^{n} f(x^{(j)}) I(x^{(j)}, E_i)}{\sum_{j=1}^{n} f(x^{(j)})} + o_p(1).$$

It follows from the law of large numbers and the Slutzky's theorem that

$$\tilde{a}_i \mu^*(\tilde{E}_i) \xrightarrow{P} \frac{\int_{E_i} f(x) \mu(dx)}{\int_{S(f)} f(x) \mu(dx)} = a_i \mu(E_i).$$

2. First, by the law of large numbers, as $m \to \infty$, $n \to \infty$,

$$\frac{1}{m} \sum_{j=1}^{m} H\left(y^{[j]}\right) - \sum_{j=1}^{n} H\left(x^{(j)}\right) \tilde{g}_l(x^{(j)}) = o_p(1).$$

Further, by construction and the law of large numbers,

$$\begin{aligned}\sum_{j=1}^{n} H\left(x^{(j)}\right)\tilde{g}_l(x^{(j)}) &= \sum_{i=1}^{l}\sum_{j=1}^{n} H\left(x^{(j)}\right)\tilde{a}_i I(x^{(j)},\tilde{E}_i)\\ &= \frac{l}{n}\sum_{i=1}^{l}\tilde{a}_i\mu^*(\tilde{E}_i)\sum_{j=1}^{n} H\left(x^{(j)}\right)I(x^{(j)},E_i)+o_p(1)\\ &\xrightarrow{P} \frac{l}{\mu(S(f))}\sum_{i=1}^{l} a_i\mu(E_i)\int_{E_i} H(x)\mu(dx)\\ &= \int_{S(f)} H(x)g_l(x)\mu(dx).\end{aligned}$$

Finally, since $g_l(x)\to f(x)$ uniformly on $S(f)$ by Theorem 1(2),

$$\int_{S(f)} H(x)g_l(x)\mu(dx)\to\int_{S(f)} H(x)f(x)\mu(dx),$$

which completes the proof. ■

There are three basic issues associated with this method (or any other sampling based method):

1. Is the method applicable in moderate to high dimensional problems?
2. Is the method computationally efficient?
3. Does the method provide reasonably accurate approximation?

We have applied our method to a large number of examples, including some benchmark examples which have frequently appeared in the recent literature. Our method performs pretty well with all these examples, some of which are shown in the following sections. We expect that this method will perform well with other real examples, including high dimensional problems. Before providing examples, the following technical remarks are in order, which may offer more insights and details of the algorithm.

REMARK 1 The algorithm basically involves only two computational components, namely sampling from one-dimensional uniform distribution and ordering observations according to the heights of the density function. In this sense the algorithm is dimension-free and also non-iterative. Consequently, this algorithm is computationally very efficient and fast, especially for low or moderate dimensional sampling and integration problems. Another feature of the algorithm is that the contourization step transforms the original distribution f on $S(f)$ to a finite, monotonic step function $\tilde{g}_l$ on a partition of $S_n(f)$. This eliminates problems and difficulties caused by multi-modality and non-connectivity in most Markov Chain Monte Carlo algorithms. In addition, since the re-normalization is build into the algorithm, the normalizing constant of density $f(x)$ needs not be known, which is often convenient and sometimes desirable in many real applications, especially in Bayesian analysis. Furthermore, from

$$1=\int_{S(f)} f(x)dx=\mu(S(f))\int_{S(f)} f(x)/\mu(S(f))dx,$$

the normalizing constant can be estimated using the initial uniform sample as

$$\frac{n}{\mu(S(f))\sum_{j=1}^{n} f(x^{(j)})}.$$

REMARK 2 Our algorithm is applicable to most distributions encountered in real applications. In some cases only minor modifications are needed.

CASE 1 When $f(x)$ is continuous but its support $S(f)$ is unbounded, the algorithm can be modified by applying the Egorov's theorem (Hewitt and Stromberg, 1967, p. 158). Without loss of generality, we assume $S(f) = \mathbb{R}^k$. Since $f(x)$ is continuous, for an arbitrarily small $\varepsilon > 0$, we can construct a compact subset

$$S^*(f) = [-M, M]^k \cap S(f) \tag{10}$$

such that

$$\int_{S^*(f)} f(x)\mu(dx) \geq 1 - \varepsilon \tag{11}$$

for a very large $M > 0$. Then, the reweighted function

$$f^*(x) = \frac{f(x)}{\int_{S^*(f)} f(x)\mu(dx)} \tag{12}$$

is a density function on $S^*(f)$. Similarly, we can define $g_l^*(x)$ on $S^*(f)$ the same way as $g_l(x)$ is defined on $S(f)$. It follows that $g_l^*(x) \to f^*(x)$ uniformly on $S^*(f)$ as $l \to \infty$. In addition, the discretized support $S_n(f)$ should be modified as

$$S_n^*(f) = [-M, M]^k \cap S_n(f) \tag{13}$$

for the given M. More practical issues concerning the choice of M are discussed in Sections 4 and 5.

CASE 2 If support $S(f) \subset [a, b]^k$ but $S(f) \neq [a, b]^k$, then the discretized sample space should be modified as $S_n(f) = \{x^{(j)}\}_{j=1}^{n} \cap S(f)$. The rest part of the algorithm remains the same.

CASE 3 If $f(x)$ is a simple density function, i.e., $f(x) = g_l(x)$ for some $l \in \mathbb{N}$, then function $\tilde{g}_l(x)$ can be taken to be $f(x)$ restricted to $S_n(f)$ and the algorithm becomes simpler.

CASE 4 If $f(x)$ is a mixed type distribution, then for the discrete components the discrete uniform distribution should be used. The rest part of the algorithm can either remain unchanged or can be further modified as in Case 3.

REMARK 3 For integration problem, this method can be used together with importance sampling as

$$\int_{S(f)} H(x)f(x)dx = \int_{S(f)} [H(x)f(x)/g(x)]g(x)dx.$$

In fact, it is the simplicity and efficiency of the algorithm that makes it much easier to identify the importance region and, therefore, to choose importance density $g(x)$. Moreover, our method can also be incorporated with MCMC methods like Gibbs sampler by applying the method sequentially when the full conditional densities are available. However, we don't recommend such combinations when observations can be generated from $f(x)$ directly. Furthermore, if numerical integration is of primary concern, then after identifying importance region, the integral can be approximated by using initial uniform sample $\{x^{(j)}\}$ rather than using sample $\{y^{[j]}\}$.

REMARK 4 If $f(x)$ is a likelihood function or a posterior density function defined on the parameter space $S(f)$ and satisfies $\int_{S(f)} f(x)\mu(dx) < \infty$, then the first two steps of the algorithm can be used to compute the approximate maximum likelihood estimate (AMLE) or the posterior modes. The estimate is any one of the points in the last contour of $S_n(f)$ corresponding to the largest value of the function f. The AMLE is very close to the true MLE, when n, m and l are sufficiently large. Further, if finding the true MLE is desired, then the AMLE can serve as a good initial estimate for an iterative searching procedure such as Newton-Raphson. This method is used to compute the AMLE in Example 6 and modes of various distributions in other examples in Section 5.

4. Examples

In order to illustrate our method and to make our algorithm more transparent, we give four simple examples of two or three dimensional distributions, two of which have bounded support and the other two have unbounded support. More benchmark and real examples of higher dimensional problems are given in the next section. All numerical computations in this paper are carried out using S-PLUS on a Pentium III PC with standard hardware configuration. The actual computing time for each example is much less than one minute, so that we do not report it explicitly.

EXAMPLE 1 Consider two independent random variables $X_1 \sim \text{Beta}(2,2)$ and $X_2 \sim \text{Beta}(3,1)$. The joint density function of (X_1, X_2), up to a normalizing constant, is given by

$$f(x_1, x_2) = x_1(1 - x_1)x_2^2,\ 0 \le x_1,\ x_2 \le 1.$$

This distribution has a compact support $S(f) = [0,1]^2$ and the first two moments $E(X_1) = 1/2$, $\text{Var}(X_1) = 1/20$, $E(X_2) = 3/4$ and $\text{Var}(X_2) = 3/80$. Suppose we wish to draw a sample of size $m = 2{,}000$ and compute the corresponding sample moments.

First, we generate $n = 200{,}000$ uniform random numbers on $[0,1]^2$ to form the discretized support $S_n(f)$. Then $S_n(f)$ is divided into $l = 100$ contours according to the heights of density function $f(x^{[j]})$, with 2,000 points in each contour. Finally, $m = 2{,}000$ sample points are drawn from these contours according to the sampling scheme in Step 3. Figure 1 shows a scatter plot, a contour plot and a surface plot of the sample, together with the histograms of two marginal distributions. The corresponding sample moments are $\overline{X}_1 = 0.4989$, $S_1^2 = 0.0501$, $\overline{X}_2 = 0.7429$ and $S_2^2 = 0.0387$.

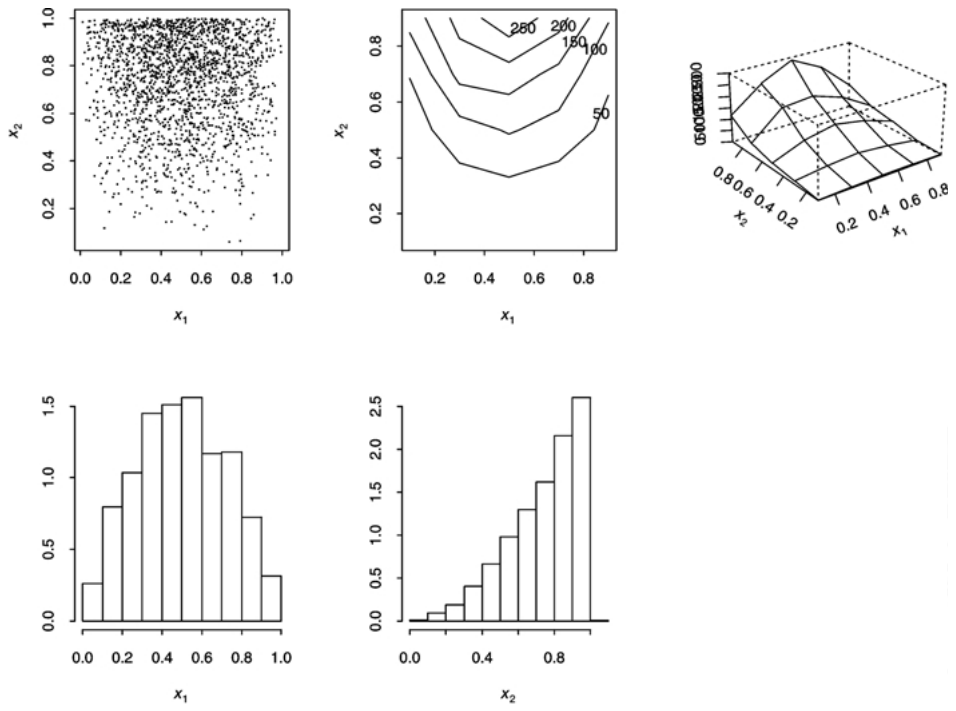

Figure 1. A sample of size $m = 2{,}000$ from the bivariate Beta distribution of Example 1.

EXAMPLE 2 Consider a three dimensional Dirichlet distribution $(X_1, X_2, X_3) \sim D(0.5, 2.5, 4.5, 6.5)$ with density (up to a normalizing constant)

$$f(x_1, x_2, x_3) = x_1^{-0.5} x_2^{1.5} x_3^{3.5} (1 - x_1 - x_2 - x_3)^{5.5}, 0 \le x_1 + x_2 + x_3 \le 1.$$

This distribution has a compact support $S(f) = \{(x_1, x_2, x_3) \in [0,1]^3 : x_1 + x_2 + x_3 \le 1\}$ which is not of the form $[a, b]^3$. Again $n = 200{,}000$ uniform random points are drawn from $S(f)$ to form $S_n(f)$ and then $S_n(f)$ is divided into $l = 200$ contours, such that each contour has 1,000 points. Finally $m = 2{,}000$ sample points are drawn from these contours. Some selected scatter plots and marginal histograms of the sample are shown in Figure 2. The corresponding sample moments are respectively $\overline{X}_1 = 0.0403, \overline{X}_2 = 0.1741, \overline{X}_3 = 0.3241$, $S_1^2 = 0.0025$, $S_2^2 = 0.0094$ and $S_3^2 = 0.0154$.

As indicated in Remark 2 of Section 3, in the case of unbounded support, an initial compact interval $[-M, M]^k$ can be determined, which is large enough to cover the ''high'' probability area of the distribution f. This step is not as complicated as it looks like for most distributions arising in applications. In fact, one may start with a reasonable guess of the interval based on the properties of the given density function. Once a sample has been generated, one can examine the marginal histograms and will realize immediately, if the initial interval is too small or too large and make adjustment correspondingly. This is demonstrated in the next example.

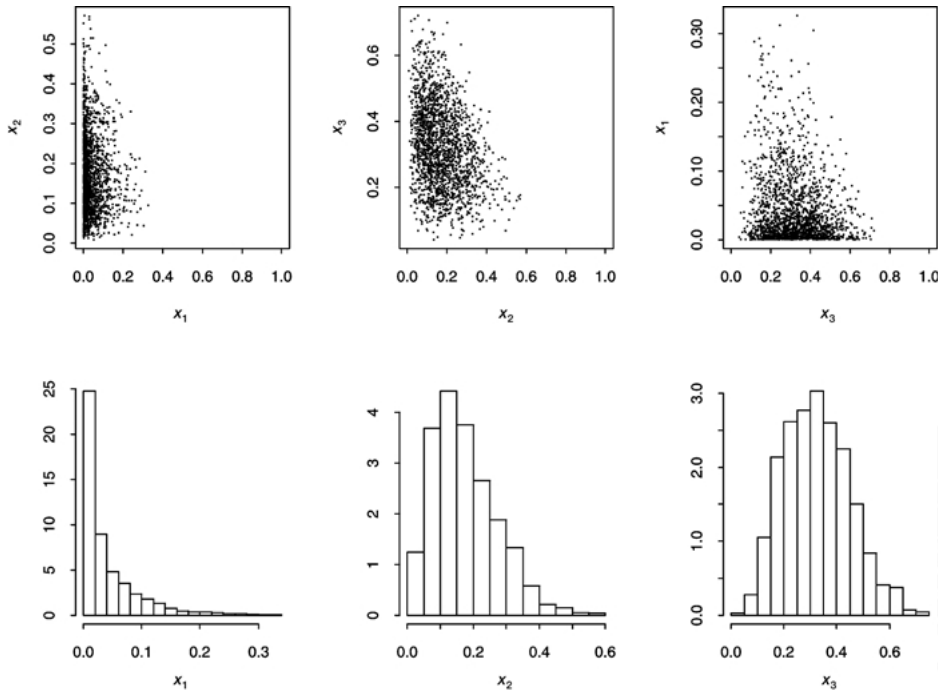

Figure 2. A sample of size $m = 2{,}000$ from the three dimensional Dirichlet distribution.

EXAMPLE 3 Consider a benchmark example which is used to compare different algorithms in Robert (1998). It is the posterior distribution of the location parameter of three i.i.d. Cauchy random variables with unit scale parameter and a constant prior distribution. The density function (up to a normalizing constant) is given by

$$f(x) = \left[(1 + (x+8)^2)(1 + (x-8)^2)(1 + (x-17)^2)\right]^{-1}, -\infty < x < \infty.$$

This distribution is trimodal and has an unbounded support $(-\infty, \infty)$. The difficulty in generating random sample from this distribution is that it has a large dispersion. Following Robert (1998), we also compute the quantities $h_i = EH_i(X)$, $i = 1, 2, 3$, with $H_1(x) = x$, $H_2(x) = (x - 17/3)^2$ and $H_3(x) = I(x \in [4, 8])$, where I is the indicator function.

From the properties of Cauchy distribution and the fact that this distribution has three modes -8, 8 and 17, we take the initial interval to be $[-60, 60]$, from which we generate $n = 500{,}000$ uniform random points to form $S_n(f)$. We then divide $S_n(f)$ into $l = 500$ contours, such that each contour has 1,000 points. Finally we draw $m = 2{,}000$ sample points from these contours. A histogram and a density curve of this sample is shown in the first row of Figure 3.

From the histogram of this sample we see immediately that there is no observation out of interval $[-20, 30]$, meaning that the initial interval $[-60, 60]$ is too large. We thus narrow the interval down to $[-20, 30]$ and repeat the whole procedure. The new sample is

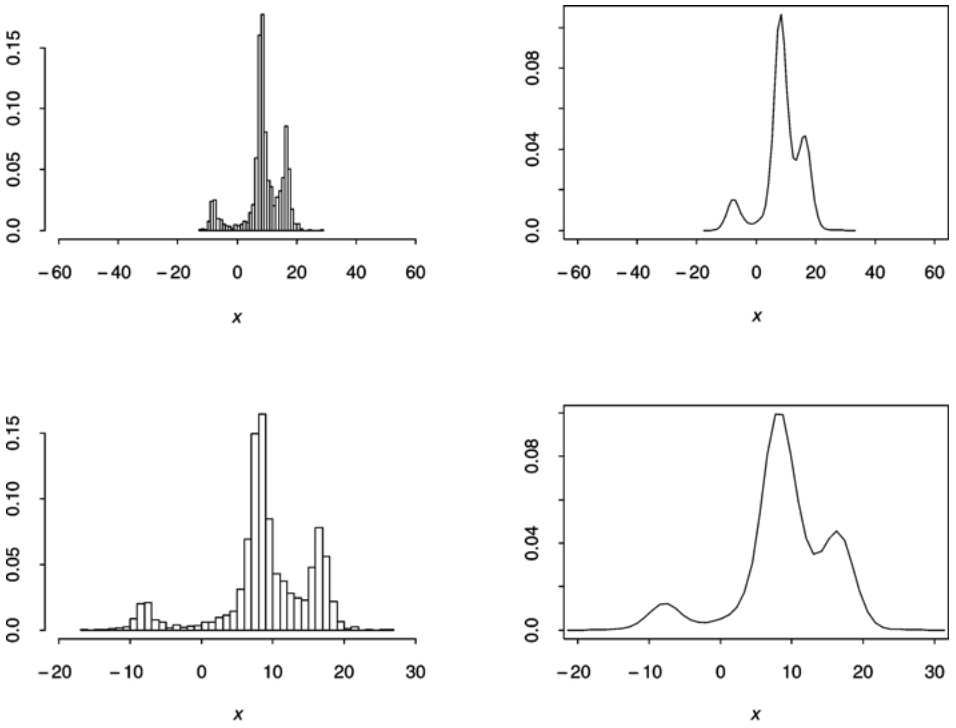

Figure 3. Two samples of size $m = 2{,}000$ from the univariate distribution of Example 3.

shown in the second row of Figure 3. The corresponding sample median is Med $= 8.5401$ and other quantities are respectively $h_1 = 8.9231$, $h_2 = 51.9702$ and $h_3 = 0.2585$. Using the first two steps of the algorithm, the mode of $f(x)$ is found to be Mod $= 8.0485$. These quantities are computed by many authors using much more complicated procedures in the literature (Robert, 1998).

EXAMPLE 4 Consider a mixture of three bivariate normal distributions

$$f(x_1, x_2) = 0.3\, f_1(x_1, x_2) + 0.3\, f_2(x_1, x_2) + 0.4\, f_3(x_1, x_2),$$

where f_1, f_2, f_3 are normal distributions with common standard deviations $(0.4, 0.4)$, but are centered at means $\mu_1 = (-3, 0)$, $\mu_2 = (3, 0)$ and $\mu_3 = (0, 3)$ respectively.

This distribution has support $S(f) = \mathbb{R}^2$. From the property of normal distribution, we start with a large enough initial interval $[-9, 9]^2$, and then narrow it down to $[-6, 6] \times [-3, 6]$. So, we first generate $n = 600{,}000$ uniform random points from $[-6, 6] \times [-3, 6]$ to form $S_n^*(f)$. Then $S_n^*(f)$ is divided into $l = 600$ contours. From these contours $m = 6{,}000$ observations are drawn. The sample is shown in Figure 4. The corresponding sample moments are $\overline{X} = (0.0001, 1.2081)$ and $S^2 = (5.5177, 2.3136)$. The mode is found to be Mod $= (0.0001, 2.9990)$.

Note that the locations of the three component distributions in this example are of equal

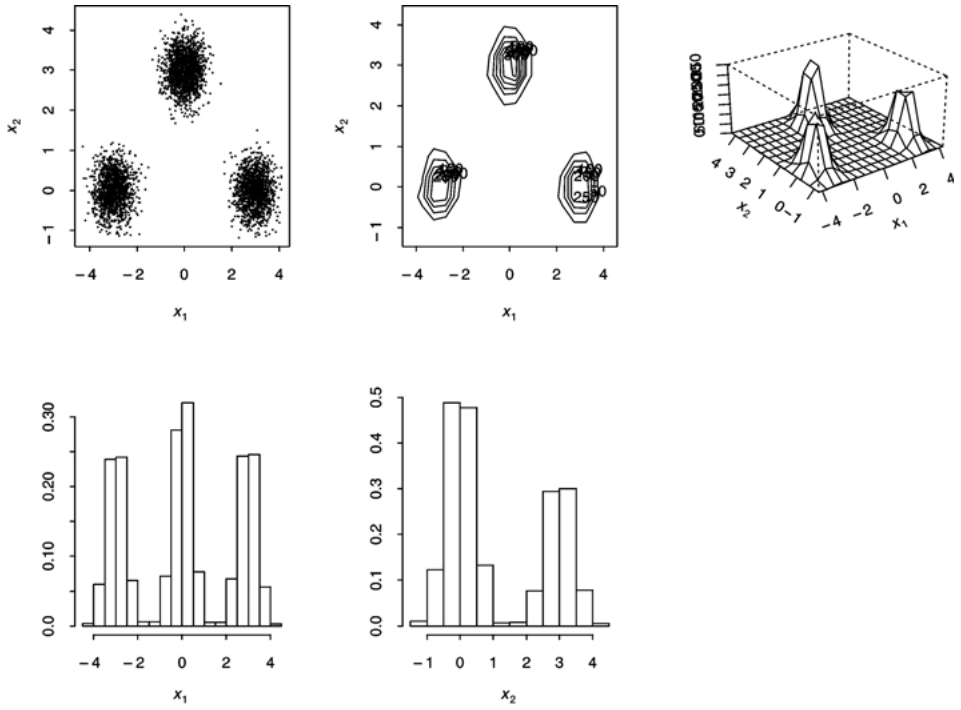

Figure 4. A sample of size $m = 6{,}000$ from the bivariate normal mixture in Example 4.

distance, i.e., $d(\mu_1, \mu_2) = d(\mu_2, \mu_3) = d(\mu_3, \mu_1)$. It is important to see that as the locations of the three distributions move away from each other, the usual MCMC algorithms like Gibbs sampler will become less and less efficient, and will fail when the distributions are far enough from each other. This is due to the fact that the Markov chain, induced by the Gibbs sampler, becomes slower and slower, and collapses eventually as the ratio of the distance between the locations to their variances becomes moderately large. In general, Gibbs sampler will have much trouble to deal with moderate or high dimensional mixture populations, especially when the sub-populations are apart from each other. Note that in this example, this deficiency cannot be overcome by rotating coordinate axes.

5. Applications

In this section we apply our approach to some real problems which have been extensively studied in the recent literature. All previous studies of theses examples are restricted to either conjugate or ''standard'' set-up of prior distributions and likelihoods, for the computational convenience. We demonstrate with these examples that our method is not restricted to those set-up and provides real freedom in choosing prior distributions.

EXAMPLE 5 Gaver and O'Muircheartaigh (1987) analyzed a data set of failure rates of ten pumps at the nuclear power plant Farley 1, UK. They used a Poisson model for the

logarithmic failure rates and log-Student-t and gamma priors respectively for the parameters. This data set has been subsequently re-analyzed by Gelfand and Smith (1990) and Tierney (1994), using different prior distributions. This example has been used as a benchmark example by Robert (1998) to compare different algorithms. Here we analyze the original data (not logarithmic scaled) and use the log-Student-t prior, as originally proposed by Gaver and O'Muircheartaigh (1987). Suppose the pump failure rates $s_i \sim \text{Poisson}(\lambda_i t_i)$, $i = 1, 2, \ldots, 10$. Parameters $\log \lambda_i$ are modeled as a sample from the Student-t distribution with five degrees of freedom,

$$\log \lambda_i \sim t(5; \mu, \tau) = \left[1 + \frac{1}{5}\left(\frac{x-\mu}{\tau}\right)^2\right]^{-3}.$$

Thus, the logarithm of the posterior density is

$$f(\lambda_1, \ldots, \lambda_{10}) = \sum_{i=1}^{10}(s_i - 1)\log \lambda_i - \sum_{i=1}^{10} t_i \lambda_i - 3\sum_{i=1}^{10} \log[5\tau^2 + (\log \lambda_i - \mu)^2].$$

Following Gaver and O'Muircheartaigh (1987), parameters μ and τ are set to their maximum likelihood estimates, $\hat{\mu} = -1.18$ and $\hat{\tau} = 1.29$. This model is estimated using the data given in Gaver and O'Muircheartaigh (1987).

First, from the property of Student-t distribution, we choose the initial interval to be $[0, 20]^{10}$. After several trials, the upper bound 20 has been lowered to different values for different coordinates. Then, from the resulting interval, $n = 600{,}000$ uniform random points are generated. These points are then divided into $l = 1{,}000$ contours, from which $m = 5{,}000$ observations are sampled. The marginal posterior distributions of $\lambda_i, i = 1, 2, \ldots, 10$ are shown in Figure 5. The corresponding posterior modes, means and standard deviations are given in Table 1, together with the real failure rates computed from the data. To our knowledge these graphical and numerical results are most complete compared to all previous studies in the literature.

EXAMPLE 6 Tanner (1992) used a Poisson change point model to analyze a British coalmining disaster data set from 1851–1962. A simplified version of this model has also been considered by Arnold (1993). A time series plot and a bar chart of the data are displayed in Figure 6, which shows a clear structure change around 1890.

Following Tanner (1992), suppose the annual counts of disasters $X_i \sim \text{Poisson}(\theta t_i)$, for $i = 1, 2, \ldots, \kappa$ and $X_i \sim \text{Poisson}(\lambda t_i)$, for $i = \kappa + 1, \kappa + 2, \ldots, N$ with $N = 110$. Then the log-likelihood function is

$$l(\kappa, \theta, \lambda) = \left(\sum_{i=1}^{\kappa} x_i - 1/2\right)\log \theta + \left(\sum_{i=\kappa+1}^{n} x_i - 1/2\right)\log \lambda - \kappa\theta - (n - \kappa)\lambda,$$

where $\kappa \in (1 : N)$, $\theta \in (0, \infty)$ and $\lambda \in (0, \infty)$. Using our algorithm and the steps in the next paragraph, the approximate maximum likelihood estimates (AMLE) for these parameters are computed and the corresponding results are shown in Table 2.

Now let the parameters in this model have the following prior distributions: the change point κ has a discrete uniform distribution on integers $(1 : N)$, θ and λ have gamma

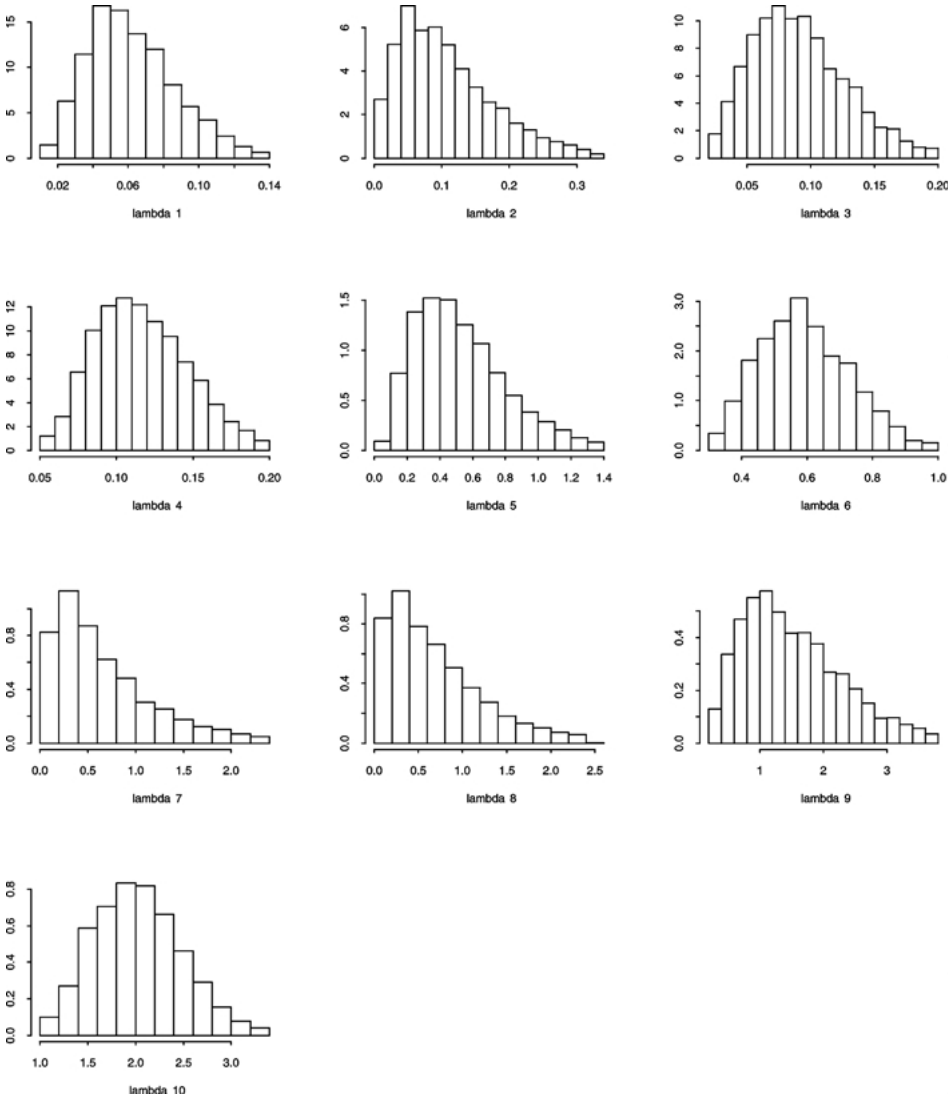

Figure 5. Marginal posterior distributions of pump failure rates λ_i, $i = 1, 2, \ldots, 10$.

Table 1. Posterior modes, means and standard deviations of Farley 1 pump failure rates.

Pump	1	2	3	4	5	6	7	8	9	10
Rate	0.0530	0.0636	0.0795	0.1113	0.5725	0.6043	0.9542	0.9542	1.9084	2.0992
Mode	0.0543	0.0909	0.0810	0.1015	0.2630	0.5744	0.4236	0.2700	1.5815	1.7625
Mean	0.0628	0.1085	0.0913	0.1160	0.5246	0.5914	0.6801	0.6892	1.5431	2.0159
Std.	0.0248	0.0690	0.0364	0.0296	0.2718	0.1345	0.5181	0.5209	0.7653	0.4384

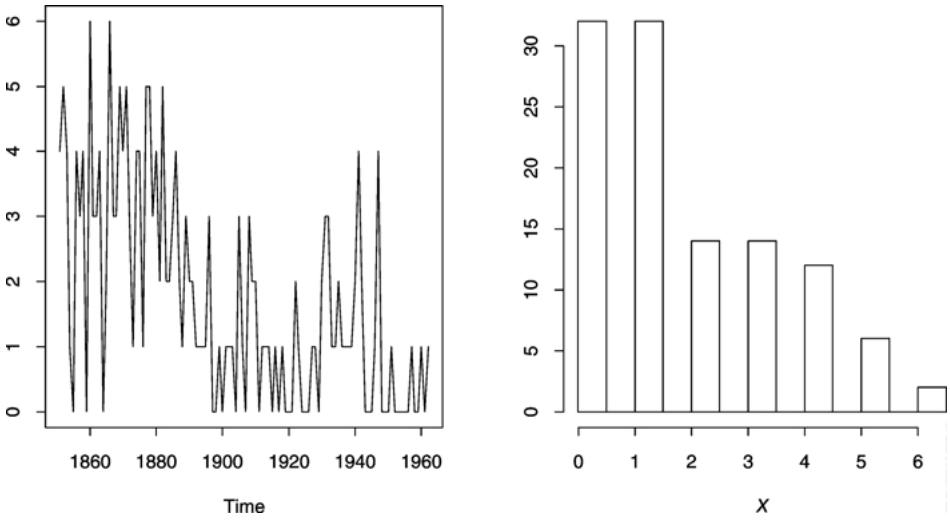

Figure 6. Coalmining disaster data (1851–1962) in Example 6.

distributions $G(1/2, \alpha)$ and $G(1/2, \beta)$ respectively, and hyper-parameters α and β both follow a gamma distribution $G(2, 1)$. Therefore, the logarithm of the posterior density is

$$f(\kappa, \theta, \lambda, \alpha, \beta) = l(\kappa, \theta, \lambda) + 1.5 \log \alpha + 1.5 \log \beta - (\theta + 1)\alpha - (\lambda + 1)\beta.$$

We generate a sample from the posterior distribution for the discrete variable κ and four continuous variables θ, λ, α and β. We start with initial ranges for parameters as $\kappa \in (25 : 55)$, $\theta \in [2, 5]$, $\lambda \in [0, 2]$, $\alpha \in [0, 3]$ and $\beta \in [0, 6]$ respectively, then narrow them down to $\kappa \in (30 : 50)$, $\theta \in [2.2, 4]$, $\lambda \in [0.6, 1.4]$, $\alpha \in [0, 2]$ and $\beta \in [0, 4]$ respectively, from which $n = 600{,}000$ initial points are generated. These points are divided into $l = 600$ contours, from which a sample of size $m = 5{,}000$ is drawn. The marginal posterior distributions are shown in Figure 7. The first plot in Figure 7 shows the discrete posterior distribution for the change point κ, which is very similar to the result obtained by Tanner (1992). The corresponding posterior sample modes, means and standard deviations of all parameters are given in Table 2.

Table 2. The AMLE, posterior modes, means and standard deviations of the change point κ and other parameters in Example 6.

Parameters	κ	θ	λ	α	β
AMLE	41.0000	3.0809	0.9097	–	–
Mode	41.0000	3.0945	0.9090	0.4357	0.8370
Mean	40.0060	3.0807	0.9274	0.6205	1.2971
Std.	2.5224	0.2972	0.1206	0.3815	0.7872

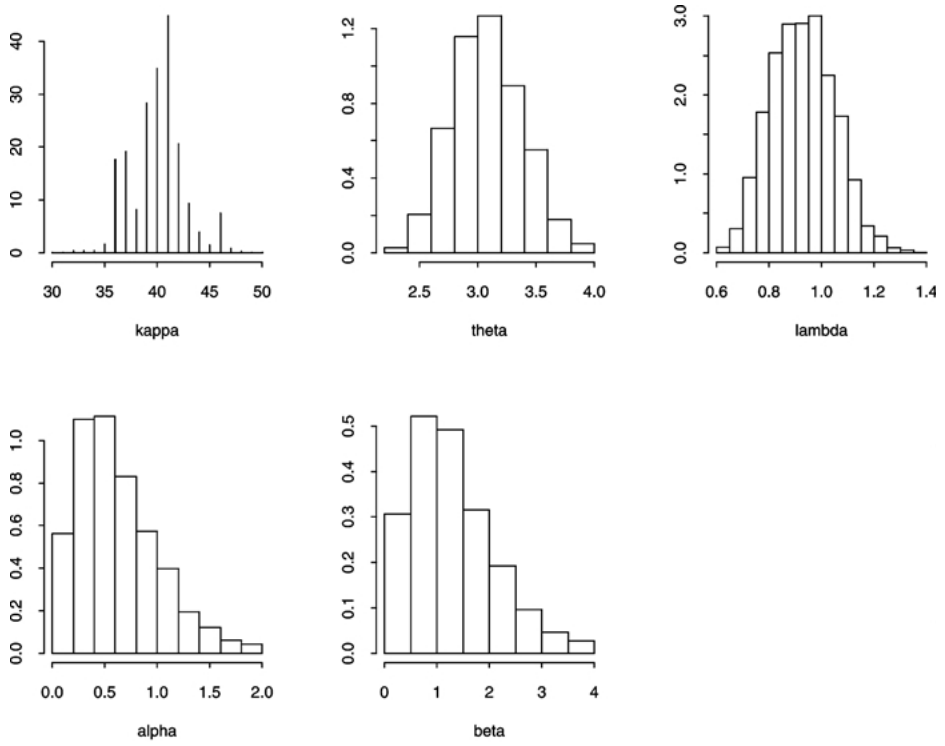

Figure 7. Marginal posterior distributions of κ, θ, λ, α and β in Example 6.

6. Curse of Low Probability

This section is concerned with an important and challenging issue that is closely related to random sample generation and integration. The problem is how to compute accurately an integration over an area where the probability measure is extremely low. In the literature this problem is often regarded as a problem of dimensionality. To our understanding, however, the real difficulty is not (only) high dimension, but rather low probability. We explain the problem with the following three examples.

As indicated in the introduction section, for a given integration problem, it is well known that the sampling-based methods may not be as accurate or efficient as numerical analytic methods, especially when the dimension is low or moderate. However, the attraction of sampling-based methods is their conceptual simplicity and ease of implementation. In this sense our method can provide an indirect solution to some integration problems. More technical and accurate treatments of the numerical integration problems, including evaluating multivariate normal probabilities, can be found in, for example, Evans and Swartz (1995, 2000), Genz (1992, 1993), Genz and Kwong (2000), Schervish (1984) and Somerville (1998).

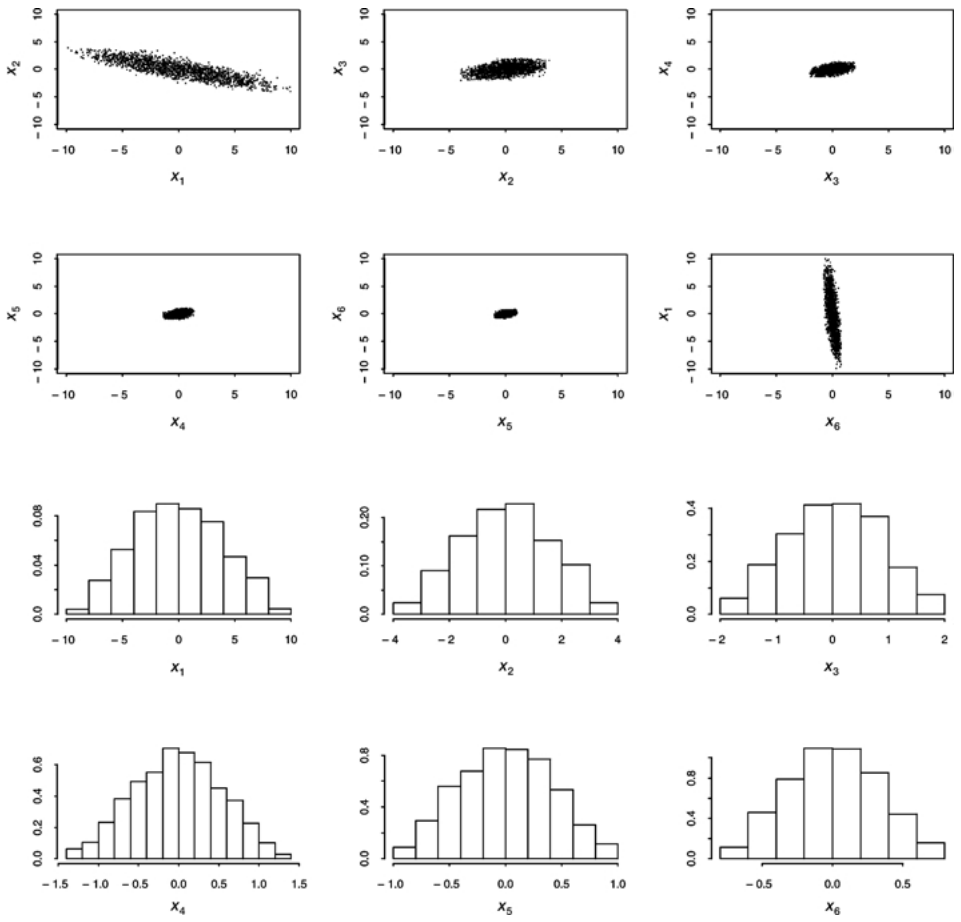

Figure 8. A sample of size $m = 2{,}000$ from the six-dimensional normal distribution $N_6(0, \Sigma)$.

EXAMPLE 7 Evans and Swartz (1995) considered an integral of the density function $f(x)$ of the six dimensional normal distribution $N_6(0, \Sigma)$ over interval $[0, \infty)^6$, where $\Sigma^{-1/2} = \text{diag}(0, 1, 2, 3, 4, 5) + ee'$ and $e = (1, 1, 1, 1, 1, 1)'$. As pointed out by Evans and Swartz (1995), an accurate numerical evaluation of this integral is extremely difficult. To visualize the problem, we use our algorithm to generate a sample of size $m = 2{,}000$ from this distribution and display some selected scatter plots and the marginal histograms in Figure 8.

From these plots we see that, because of the variance-covariance structure of this distribution, the value of $f(x)$ is very small for $x \in [0, \infty)^6$. In fact, the value of this integral is approximately $\int_{[0,\infty)^6} f(x)dx \approx 1/60{,}000$, which means that in about every 60,000 generated observations we expect to have one observation falling in the range $[0, \infty)^6$. This makes it very difficult to use any Monte Carlo method to calculate this integral accurately. However, our method provides an indirect solution to this problem. From the

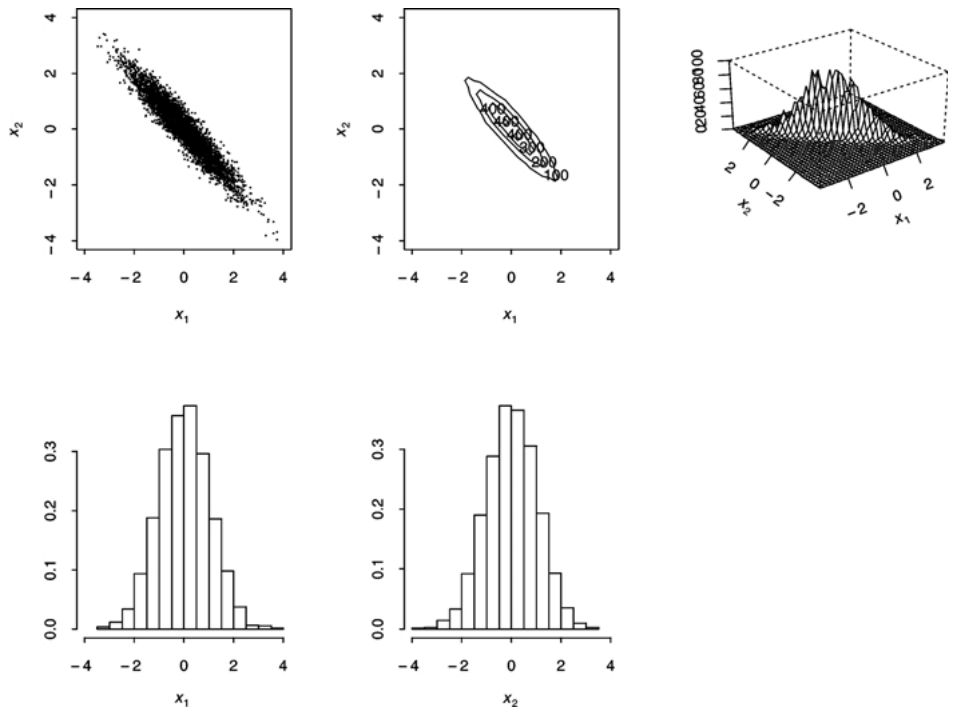

Figure 9. A sample of size 5,000 from the bivariate normal distribution with correlation $r = -0.95$.

drawn sample we can see that in the range of integration $[0, \infty)^6$ the density function highly concentrates around the origin and is almost zero anywhere else. This suggests to approximate the integral using a simple importance sampling procedure. Indeed, a sample of 10^9 uniform observations from interval $[0, 1]^6$ gives a sample mean for $f(x)$ as 1.6658269×10^{-5}. This value is pretty close to the results of Evans and Swartz (1995), obtained by much more sophisticated procedures.

EXAMPLE 8 The difficulty in the above example has in fact nothing to do with dimension. It is easy to see that even an integral of a density function of a bivariate normal distribution over range $[0, \infty)^2$ may be difficult, if the correlation coefficient is very close to -1. This can also be visualized through Figure 9 of a sample from the bivariate normal distribution with means $(0, 0)$, variances $(1, 1)$ and correlation $r = -0.95$. Again, the method in the previous example provides a possible solution for similar problems.

EXAMPLE 9 Sometimes difficulties like one in the above two examples can be overcome through using some advanced techniques and with great care, such as importance sampling techniques (Evans and Swartz, 1995). However, there are situations where even importance sampling does not work properly. This can be seen by the following univariate integration problem. Let $f(x)$ be the density of the standard normal distribution and

$$H(x) = \frac{\sqrt{\pi/2}\exp(x^2/2)}{|x|\log|x|(\log\log|x|)^2} I(|x| \geq m_0),$$

where $m_0 = 9$. The question is how to approximate the integral

$$I = \int_{-\infty}^{\infty} H(x)f(x)dx$$

without using knowledge of the functional forms of f and H. It is easy to see that the integral $1 - \int_{-m_0}^{m_0} f(x)dx$ is a very small number, meaning that an extremely small portion of observations of any sample will fall in the range $(-\infty, -m_0] \cup [m_0, \infty)$. Furthermore, it is easy to calculate, for every $m > m_0$, the difference

$$I - \int_{-m}^{m} H(x)f(x)dx = \frac{1}{\log\log m},$$

which tends to zero very slowly as $m \to \infty$. As a result, the numerical integration of the form $\int_{-m}^{m} H(x)f(x)dx$ will not be a good approximation of I, unless m is extremely large. Indeed, for $m = 10^k$ say, the relative approximation error is

$$\frac{\log\log m_0}{\log\log m} = \frac{\log\log m_0}{\log k + \log\log 10}$$

which is about $1/10$ even for $k = 1{,}000$! It seems to us that any Monte Carlo based simulation method will fail to deal with such kind of problem.

All the above examples show that the real difficulty in calculating integrations over an unbounded low probability area lies in the fact that it is difficult to generate any observation, let alone many, in that area via sampling based methods.

7. Discussions and Conclusions

We proposed a simple, direct numerical sampling method, which is dimension-free, non-iterative and is applicable to almost all multivariate distributions arising in real applications. This method is based on random discretization of the density function and a multivariate version of the inverse probability integral transformation. This approach requires no reparameterization or any other truly restrictive conditions on density functions. Due to its simplicity and efficiency, this algorithm enables flexible and practical choice of prior distributions, which is important in non-conjugate Bayesian analysis of real problems.

This approach is based on an approximation of the density of interest by a simple density with l contours. This approximation becomes more and more accurate, when $l \to \infty$. Based on our experience with numerous real examples, a value of l between 200 and 500 for a density of five dimensions or lower, and a value between 1,000 and 100,000 for a density of higher dimensions usually provide rather satisfactory results.

As pointed out by a referee, there is an interesting connection between the algorithm of this paper and the sampling-importance-resampling (SIR) of Rubin (1988). In the case where density $f(x)$ has a compact support, if the uniform density is used as importance

density in SIR and the number of contours is taken to be $l = n$ in our algorithm, then both procedures are the same.

It is shown that the method of this paper also provides an indirect solution to some numerical integration problems. The examples presented are initial experiments in this direction and much more work is needed for future research.

Acknowledgments

The authors are grateful to two anonymous referees, an associate editor and the editor for helpful comments and suggestions. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.

References

S. F. Arnold, ''Gibbs sampling,'' C. R. Rao ed., *Handbook of Statistics* vol. 9 pp. 599–625, 1993.

S. Chib and E. Greenberg, ''Markov Chain Monte Carlo simulation methods in econometrics,'' *Econometric Theory* vol. 12 pp. 409–431, 1996.

M. Evans and T. Swartz, ''Methods for approximating integrals in statistics with special emphasis on bayesian integration problems,'' *Statistical Sciences* vol. 10 no. 3, pp. 254–272, 1995.

M. Evans and T. Swartz, *Approximating Integrals via Monte Carlo and Deterministic Methods*, Oxford University Press: New York, 2000.

D. P. Gaver and I. G. O'Muircheartaigh, ''Robust empirical Bayes analyses of event rates,'' *Technometrics* vol. 29 pp. 1–16, 1987.

A. E. Gelfand and A. F. M. Smith, ''Sampling-based approaches to calculating marginal densities,'' *Journal of American Statistical Association* vol. 85 pp. 398–409, 1990.

A. Genz, ''Numerical computation of multivariate normal probabilities,'' *Journal of Computational and Graphical Statistics* vol. 1 pp. 141–150, 1992.

A. Genz, ''Comparison of methods for the computation of multivariate normal probabilities,'' *Computing Science and Statistics, Proceedings of the 25th Symposium on the Interface* pp. 400–405, 1993.

A. Genz and K.-S. Kwong, ''Numerical evaluation of singular multivariate normal distributions,'' *Journal of Statistical Computation and Simulation* vol. 1 pp. 1–21, 2000.

E. Hewitt and K. Stromberg, *Real and Abstract Analysis*, Springer-Verlag: New York, 1965.

C. P. Robert, *Discretization and MCMC Convergence Assessment*, Springer: New York, 1998.

C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer: New York, 1999.

D. R. Rubin, ''Using the SIR algorithm to simulate posterior distributions,'' in *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds), Oxford University Press, pp. 395–402, 1988.

W. Rudin, *Real and Complex Analysis*, McGraw-Hill: New York, 1966.

M. Schervish, ''Multivariate normal probabilities with error bound,'' *J. Roy. Statist. Soc. Ser. C* vol. 33 pp. 81–87, 1984.

P. N. Somerville, ''Numerical computation of multivariate normal and multivariate t probabilities over convex regions,'' *J. Comput. Graph. Statist.* vol. 7 no. 4, pp. 529–544, 1998.

M. A. Tanner ''Tools for statistical inference: observaed data and data augmentation methods,'' (2nd printing) *Lecture Notes in Statistics 67*, Springer-Verlag: New York, 1992.

L. Tierney, ''Markov chains for exploring posterior distributions,'' *Annals of Statistics* vol. 22 pp. 1701–1762, 1994.