A New Strategy for Simulating From Mixture Distributions With Applications to Bayesian Model Selection

Giovanni Petris
University of Arkansas
Fayetteville, AR 72701
GPetris@uark.edu

Luca Tardella
Università di Roma "La Sapienza"
Roma, ITALY
Tardella@pow2.sta.uniroma1.it

November 15, 2000

Abstract

We present a method of generating random vectors from a distribution having an absolutely continuous component and a discrete component. The method is then extended to more general mixture distributions that arise quite naturally when dealing with nested models within a Bayesian framework. The main idea is to transform the mixture distribution of interest into an absolutely continuous one, in a way that does not require the explicit calculation of the relative weights of the various components of the mixture. For nested models, the proposed method represents a simple alternative to Reversible Jump MCMC schemes. Its distinguishing features are the absence of a proposal step to reduce/increase the dimension of the current space and the fact that in order to assess the convergence of the chain, one can use all the standard tools available for MCMC on a space of fixed dimension.

KEY WORDS: Bayesian inference; model averaging; Markov chain Monte Carlo.

1 Introduction

Bayesian statisticians often have to consider, and typically to draw samples from, mixture distributions having one component that is absolutely continuous with respect to Lebesgue measure and the other component concentrated on a point or more generally a linear subspace of $\mathbb{R}^n$. (We use the term *mixture* to denote a convex combination of mutually singular probability measures. In the statistical literature the term is sometimes used to denote a convex combination of probability measures *tout court*, as in *mixture of normals*.) An example familiar to everybody is that of matched pairs. Suppose a researcher has available a random sample $(X_{1,j}, X_{2,j})$ $(j = 1, \ldots, N)$ of matched pairs, and she wants to know if the two marginal distributions have the same mean. If it is assumed that $X_{1,j} - X_{2,j} \sim \mathcal{N}(\theta, \sigma^2)$,

with σ^2 known, she can specify a prior for θ and then base her conclusions on the posterior. The probability that the two means are equal is the probability of the event $\{\theta = 0\}$. Therefore a convenient prior π will take the form:

$$\pi(A) = w_0 \cdot \delta(A) + w_1 \cdot \nu(A), \qquad A \in \mathcal{B}(\mathbb{R}),$$

with $\delta(A) = 1$ if $0 \in A$, and 0 otherwise, $\nu(\cdot)$ an absolutely continuous distribution with density p, w_0 and w_1 positive weights with $w_0 + w_1 = 1$. The posterior distribution of θ is again a mixture of an absolutely continuous component and a discrete component concentrated on the event $\{\theta = 0\}$. In fact, if $\ell(\theta)$ is the likelihood function corresponding to the observed differences and $K = \int_{\mathbb{R}} \ell(\theta)p(\theta)\, d\theta$,

$$\pi(A|\text{ Data }) = w_0^* \cdot \delta(A) + w_1^* \cdot \nu^*(A), \qquad A \in \mathcal{B}(\mathbb{R}),$$

where

$$w_0^* = \frac{w_0 \ell(0)}{w_0 \ell(0) + w_1 K}\,,$$
$$\nu^*(A) = \frac{1}{K} \int_A \ell(\theta)p(\theta)\, d\theta,$$

and $w_1^* = 1 - w_0^*$. Apart from a few special cases, e.g. when ν is a Normal distribution, the continuous part of the posterior, ν^*, does not have a nice analytic form. Usually, a draw from the posterior in those cases is obtained in the following way:

1. compute w_0^* and w_1^* using a numerical approximation for K;
2. choose which component to draw from running an auxiliary Bernoulli experiment with probabilities w_0^* and w_1^*;
3. draw from the selected component, possibly using a Markov chain Monte Carlo (MCMC) simulation scheme.

Alternatively, when numerical approximations for K are hard to obtain, for instance in high dimensional problems, one can resort on recent MCMC methods such as the Reversible Jump (Green, 1995). In this paper we propose an approach that avoids the numerical integration needed to compute K, while providing an effective method whose implementation is simpler than Reversible Jump. The core of the idea is to get the desired posterior distribution through an auxiliary absolutely continuous distribution corresponding, up to the appropriate normalizing constant, to a density $g(\cdot)$ obtained as follows (see Figure 1). First, starting with the unnormalized continuous component of the posterior, make room around the origin by shifting the mass away from zero:

$$g(\theta) = \begin{cases} \ell(\theta - h)p(\theta - h) & \text{if } \theta > h, \\ \ell(\theta + h)p(\theta + h) & \text{if } \theta < -h, \end{cases}$$

h being a positive quantity to be determined. Second, spread the mass concentrated on

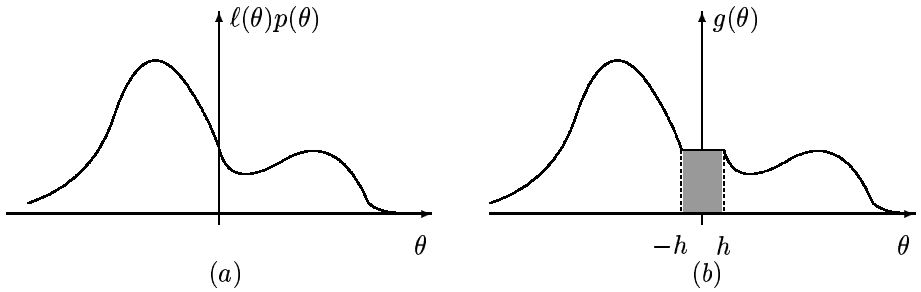

FIGURE 1: *(a) Unnormalized posterior density (absolutely continuous part) and (b) its transform.*

$\theta = 0$ onto the interval $(-h, h)$, putting

$$g(\theta) = \ell(0)p(0) \qquad \text{if } |\theta| < h.$$

If $2h = w_0/(w_1 \ell(0) p(0))$ and θ^* is drawn from the density $cg(\cdot)$, it is easily verified that the transformation

$$\theta = \phi(\theta^*) := \begin{cases} \theta^* - h & \text{if } \theta^* > h, \\ 0 & \text{if } |\theta^*| < h, \\ \theta^* + h & \text{if } \theta^* < -h \end{cases}$$

has distribution $\pi(\cdot|$ Data $)$.

In the rest of the article we extend to multidimensional settings the basic idea just described and then we show how it can be effectively employed in statistical applications such as Bayesian model selection and model averaging. The setup of the paper is as follows. In Section 2 we present a multidimensional generalization of the idea sketched above. Section 3 deals with the case of mixtures of probability distributions whose components are supported by a family of nested hyperplanes. Since that section is the most relevant for applications, we have included a piece of pseudo-code describing explicitly all the steps needed in order to evaluate the auxiliary density and the associated transformation. In Section 4 we provide two examples of statistical applications, one probing the method with simulated data and the other dealing with a real data set. Section 5 concludes, highlighting the distinguishing features of the proposed strategy and discussing how it compares with alternative existing methods.

2 Main result

Let δ be the degenerate probability measure on zero and η be Lebesgue measure on $\mathbb{R}$, and consider, for a fixed $k \in \{1, \ldots, n\}$, a finite measure given by

$$d\mu = f_0(w,x)\eta^{n-k}(dw)\eta^k(dx) + f_k(w)\eta^{n-k}(dw)\delta^k(dx), \tag{1}$$

i.e. a mixture of an absolutely continuous measure on $\mathbb{R}^n$ and a measure supported by an $(n-k)$-dimensional hyperplane and absolutely continuous with respect to Lebesgue measure on that hyperplane. The example in the introduction corresponds to $n = k = 1$. Theorem 1 below shows how to construct, up to a normalizing constant, the density g of an absolutely continuous n-dimensional random vector Y and a function ϕ from $\mathbb{R}^n$ in itself so that $X = \phi(Y)$ is a random vector with the desired distribution $\tilde{\mu}$, obtained by normalizing the measure in (1),

$$d\tilde{\mu} = d\mu/\mu(\mathbb{R}^n). \tag{2}$$

Consider any $d > 0$ and let $B_d^{(n)} = \{x \in \mathbb{R}^n : |x| \leq d\}$ be the n-dimensional closed ball of radius d, centered at the origin. The main instrument used in the constructions that follow is the function $\psi_n(\cdot; d)$ defined for $x \in \mathbb{R}^n$, $x \neq 0$, by

$$\psi_n(x;d) = \frac{x}{|x|}\left(|x|^n + d^n\right)^{1/n}.$$

The inverse function, defined for $x \notin B_d^{(n)}$, is

$$\psi_n^{-1}(x;d) = \frac{x}{|x|}\left(|x|^n - d^n\right)^{1/n}.$$

It is easy to see, considering polar coordinates in $\mathbb{R}^n$, that for any d, both ψ_n and ψ_n^{-1} preserve Lebesgue measure η^n, i.e. the Jacobian of the transformation ψ_n (or ψ_n^{-1}) is identically equal to one. The function ψ_n can be used to extend the example in Section 1

to a multidimensional setting. In fact, suppose that one wants to draw an n-dimensional random vector with distribution proportional to

$$f_0(x)\eta^n(dx) + f_n\delta^n(dx).$$

Then one can start by drawing an absolutely continuous random vector Y from the density proportional to

$$g(y) = \begin{cases} f_n & \text{if } y \in B_d^{(n)}, \\ f_0(\psi_n^{-1}(y;d)) & \text{if } y \notin B_d^{(n)} \end{cases}$$

and then consider

$$X = \phi(Y) := \begin{cases} 0 & \text{if } Y \in B_d^{(n)}, \\ \psi_n^{-1}(Y;d) & \text{if } Y \notin B_d^{(n)}. \end{cases}$$

If d is chosen so that $\eta^n(B_d^{(n)}) = 1$, then X has the required distribution. This can be deduced easily from Theorem 1 below, as a special case when $k = n$.

Theorem 1. *Consider a random vector (Z, Y) taking values in $\mathbb{R}^{n-k} \times \mathbb{R}^k$, whose distribution has density proportional to*

$$g(z,y) = \begin{cases} c(z)f_k(z) & \text{if } y \in B_{d(z)}^{(k)}, \\ f_0(z, \psi_k^{-1}(y;d(z))) & \text{if } y \notin B_{d(z)}^{(k)} \end{cases}$$

with respect to Lebesgue measure η^n. If $c(z)$ and $d(z)$ satisfy the two conditions

$$c(z)f_k(z) = f_0(z,0),$$

$$c(z)\eta^k(B_{d(z)}^{(k)}) = 1$$

for every $z \in \mathbb{R}^{n-k}$, *then the random vector* (W, X), *defined by*

$$W = Z,$$

$$X = \begin{cases} 0_k & \text{if } Y \in B^{(k)}_{d(Z)}, \\ \psi_k^{-1}(Y; d(Z)) & \text{if } Y \notin B^{(k)}_{d(Z)}, \end{cases}$$

has distribution (2). *Moreover, if* f_0 *and* f_k *are continuous, so is* g.

Proof. Let $K = \int g \, d\eta^n$ and consider two Borel sets A_1, A_2 in $\mathbb{R}^{n-k}$ and $\mathbb{R}^k$, respectively, with $0 \notin A_2$. Then, since the Jacobian of the transformation $(z, y) \mapsto (z, \psi_k^{-1}(y; d(z)))$ is one,

$$\begin{aligned} P\big(W \in A_1, X \in A_2\big) &= P\big(Z \in A_1, \psi_k^{-1}(Y; d(Z)) \in A_2\big) \\ &= K^{-1} \int_{\mathbb{R}^{n-k} \times \mathbb{R}^k} \mathbf{I}_{A_1 \times \mathbb{R}^k}(z, y) \mathbf{I}_{\mathbb{R}^{n-k} \times A_2}(\psi_k^{-1}(y; d(z)), z) \, f_0(z, \psi_k^{-1}(y; d(z))) \, \eta^{n-k}(dz) \, \eta^k(dy) \\ &= K^{-1} \int_{\mathbb{R}^{n-k} \times \mathbb{R}^k} \mathbf{I}_{A_1}(z) \mathbf{I}_{A_2}(\psi_k^{-1}(y; d(z))) \, f_0(z, \psi_k^{-1}(y; d(z))) \, \eta^{n-k}(dz) \, \eta^k(dy) \\ &= K^{-1} \int_{\mathbb{R}^{n-k} \times \mathbb{R}^k} \mathbf{I}_{A_1}(w) \mathbf{I}_{A_2}(x) \, f_0(w, x) \, \eta^{n-k}(dw) \, \eta^k(dx) \\ &= K^{-1} \mu(A_1 \times A_2). \end{aligned}$$

Similarly,

$$\begin{aligned} P\big(W \in A_1, X = 0\big) &= P\big(Z \in A_1, Y \in B^{(k)}_{d(Z)}\big) \\ &= K^{-1} \int_{\mathbb{R}^{n-k} \times \mathbb{R}^k} \mathbf{I}_{A_1}(z) \mathbf{I}_{B^k_{d(z)}}(y) c(z) f_k(z) \, \eta^{n-k}(dz) \, \eta^k(dy) \\ &= K^{-1} \int_{\mathbb{R}^{n-k}} \left\{ \int_{\mathbb{R}^k} \mathbf{I}_{B^k_{d(z)}}(y) \, \eta^k(dy) \right\} \mathbf{I}_{A_1}(z) c(z) f_k(z) \, \eta^{n-k}(dz) \\ &= K^{-1} \int_{\mathbb{R}^{n-k}} \mathbf{I}_{A_1}(z) \eta^k(B^k_{d(z)}) c(z) f_k(z) \, \eta^{n-k}(dz) \\ &= K^{-1} \int_{\mathbb{R}^{n-k}} \mathbf{I}_{A_1}(w) f_k(w) \, \eta^{n-k}(dw) \\ &= K^{-1} \mu(A_1 \times \{0\}). \end{aligned}$$

The two equalities imply that $K = \mu(\mathbb{R}^n)$ and, for any Borel set A in $\mathbb{R}^n$,

$$P\big((X,Y) \in A\big) = \tilde{\mu}(A).$$

Continuity of g follows easily from the continuity of f_0, f_k, $d(\cdot)$ and the joint continuity of $\psi_k^{-1}(y;d)$ in y and d. □

If either one of f_0 or f_k is not continuous, or one is not interested in preserving the continuity of f_0 and f_k in the resulting density g, then in Theorem 1 one can drop the requirement that $c(z)f_k(z) = f_0(z,0)$ taking, for example, $d(z)$ and $c(z)$ constant. Note also that the proof of the theorem implies that μ and the measure with density g have the same total mass:

$$\mu(\mathbb{R}^n) = \int_{\mathbb{R}^n} g\, d\eta^n.$$

This fact will be used in the following.

3 Applications to general nested models

Theorem 1 can be applied repeatedly to deal with the case of a finite measure μ on $\mathbb{R}^n$, given by

$$\begin{aligned} d\mu = {} & f_0(x_1,\dots,x_n)\,\eta(dx_1)\dots\eta(dx_n) \\ & + f_1(x_1,\dots,x_{n-1})\,\eta(dx_1)\dots\eta(dx_{n-1})\delta(dx_n) \\ & + \dots \\ & + f_n\,\delta(dx_1)\dots\delta(dx_n). \end{aligned} \tag{3}$$

In fact, one can first replace the first two terms in (3), $f_0(x_1,\dots,x_n)\,\eta(dx_1)\dots\eta(dx_n) + f_1(x_1,\dots,x_{n-1})\,\eta(dx_1)\dots\eta(dx_{n-1})\delta(dx_n)$, with an absolutely continuous measure with

density g_1, having the same total mass, then combine this g_1 and the third term $f_2(x_1, \ldots, x_{n-2})$ $\eta(dx_1) \ldots \eta(dx_{n-2})\ \delta(dx_{n-1})\ \delta(dx_n)$ into another absolutely continuous measure with density g_2, and so forth. The result of this iterative process is an absolutely continuous finite measure on $\mathbb{R}^n$ with density $g = g_n(y_1, \ldots, y_n)$, together with a function ϕ from $\mathbb{R}^n$ in itself, with the property that if a random vector Y has a density proportional to g, then $\phi(Y)$ has distribution proportional to μ. To illustrate how the procedure works, let us con-

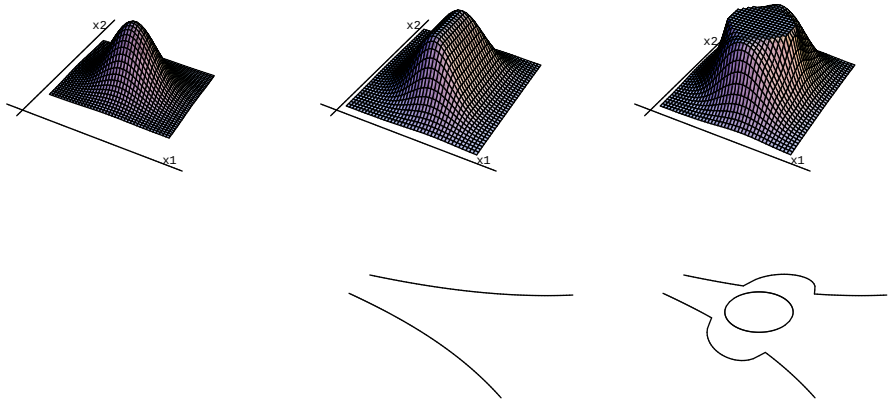

FIGURE 2: *The densities f_0, g_1 and g_2.*

sider the following example. Let $n = 2$, $f_0(x_1, x_2)$ be a bivariate standard normal density, $f_1(x_1)$ be a univariate normal density with mean 0.5 and variance 0.5, and $f_2 = 1$. The upper part of Figure 2 shows, from left to right, the densities f_0, g_1, $g_2 = g$. The lower part shows, for the corresponding densities, the boundaries of the different regions corresponding to the components being mixed. In particular, the bottom-right plot contains the regions

corresponding to all of the components. This means that if (y_1, y_2) is a draw from g and $(x_1, x_2) = \phi(y_1, y_2)$ is the appropriate transformation then we have the following cases: i) (x_1, x_2) will be the origin $(0, 0)$ when (y_1, y_2) falls in the inner circle and hence it corresponds to a draw from the component with density f_2, ii) x_2 will be zero when (y_1, y_2) falls in the region enclosing the circle, corresponding to the component $f_1(x_1)$ and iii) (x_1, x_2) will both be nonzero when (y_1, y_2) falls in the outer region, corresponding to a draw from the component $f_0(x_1, x_2)$.

Going back to the general case, let us describe explicitly an algorithm to evaluate the density g and the transformation ϕ at any given point $(y_1, \ldots, y_n)$. The following pseudocode takes as input a vector $(y_1, \ldots, y_n)$ in $\mathbb{R}^n$. It returns the triple (k, g, x), where k is the number of coordinates degenerate on zero (identifying one of the component of the mixture), the variable g contains $g(y_1, \ldots, y_n)$ and the vector $x = (x_1, \ldots, x_n)$ is $\phi(y_1, \ldots, y_n)$. Straightforward modifications are needed to treat the case when one or more of the f_k's in (3) are identically zero.

```
/* input */
```

$(y_1, \ldots, y_n)$;

```
/* initialize */
```

$(z_1, \ldots, z_n) := (y_1, \ldots, y_n)$;

$d := \left(\eta(B_1^{(n)}) \dfrac{f_0(0, \ldots, 0)}{f_n} \right)^{-\frac{1}{n}}$;

$k := n$;

```
/* loop */
```

```
while k > 0 and (z_{n-k+1}, ..., z_n) ∉ B_d^{(k)} do {
```

$(z_{n-k+1}, \ldots, z_n) := \psi_k^{-1}(z_{n-k+1}, \ldots, z_n; d)$;

$k := k - 1$;

`if` $k > 0$ `then`

$$d := \left(\eta(B_1^{(k)}) \frac{f_0(z_1, \ldots, z_{n-k}, 0, \ldots, 0)}{f_k(z_1, \ldots, z_{n-k})} \right)^{-\frac{1}{k}};$$

}

`/* output */`

k ;

$g := f_0(x_1, \ldots, x_n)$;

$x := (z_1, \ldots, z_{n-k}, 0, \ldots, 0)$;

Theorem 1 can be iteratively applied in many other ways, producing different variations of the above algorithm. For example, one could start the iterative procedure by mixing f_0 with f_n and then proceed by mixing the resulting intermediate density with f_{n-1} and so on. Alternatively, one could start by mixing f_n with f_{n-1} and then the resulting one dimensional density with the two-dimensional f_{n-2} and so forth. Some of these variations have been explored and applied in Petris and Tardella (1998) and Tardella (1999). The new method proposed here in the previous pseudo-code privileges simplicity of coding and continuity properties of the resulting g. In addition one can show that if f_0 is star-unimodal around zero, then g preserves the same shape property.

We have thus transformed the problem of generating a random vector from a mixture distribution like (3) into the much simpler one of sampling from an absolutely continuous distribution having a density explicitly known up to a normalizing constant. A variety of efficient methods to perform the last step and get a sample from g are currently available. In the examples reported in Section 4, we use a Metropolis-within-Gibbs sampling scheme,

in which each coordinate is drawn in turn from its current full conditional distribution using Metropolis algorithm (see Tierney, 1994). In order to further automatize the procedure, at each Metropolis step we use the Adaptive Rejection Metropolis Sampling scheme (Gilks et al., 1995), which neither requires the explicit specification and tuning of a proposal distribution nor the calculation of Jacobians.

4 Examples

We simulated $N = 200$ observations from the following model:

$$y_i = X\beta + \varepsilon_i \qquad i = 1, ..., N$$

where $\beta' = (6, 5, 4, 3, 2, 1, 0, 0, 0, 0) \in \mathbb{R}^{10}$ and X is a matrix of covariates. We fixed a prior distribution placing weight $w_j = 1/11$ on the event $H_j := \{\beta_{10-j} \neq 0, \beta_{11-j} = \ldots = \beta_{10} = 0\}$, $j = 0, 1, ..., 10$. Within each H_j, we took the non degenerate part of the parameter vector to follow a $(10-j)$-dimensional normal distribution with mean parameter zero and covariance matrix equal to the identity matrix. In this simple setting the posterior probabilities of the 11 regions and the posterior distributions therein can be determined analytically. Hence the exact values of the underlying posterior probabilities can be displayed in Table 1 together with their Monte Carlo estimates obtained by 15000 MCMC draws from g as described in the pseudo-code of Section 3. Figure 3 contains the histograms representing the Monte

TABLE 1: *Exact posterior probabilities and Monte Carlo estimates.*

	H_{10}	H_9	H_8	H_7	H_6	H_5	H_4	H_3	H_2	H_1	H_0
true	0.000	0.000	0.000	0.000	0.000	0.005	0.417	0.415	0.116	0.031	0.015
estimate	0.000	0.000	0.000	0.000	0.000	0.005	0.423	0.423	0.105	0.028	0.015

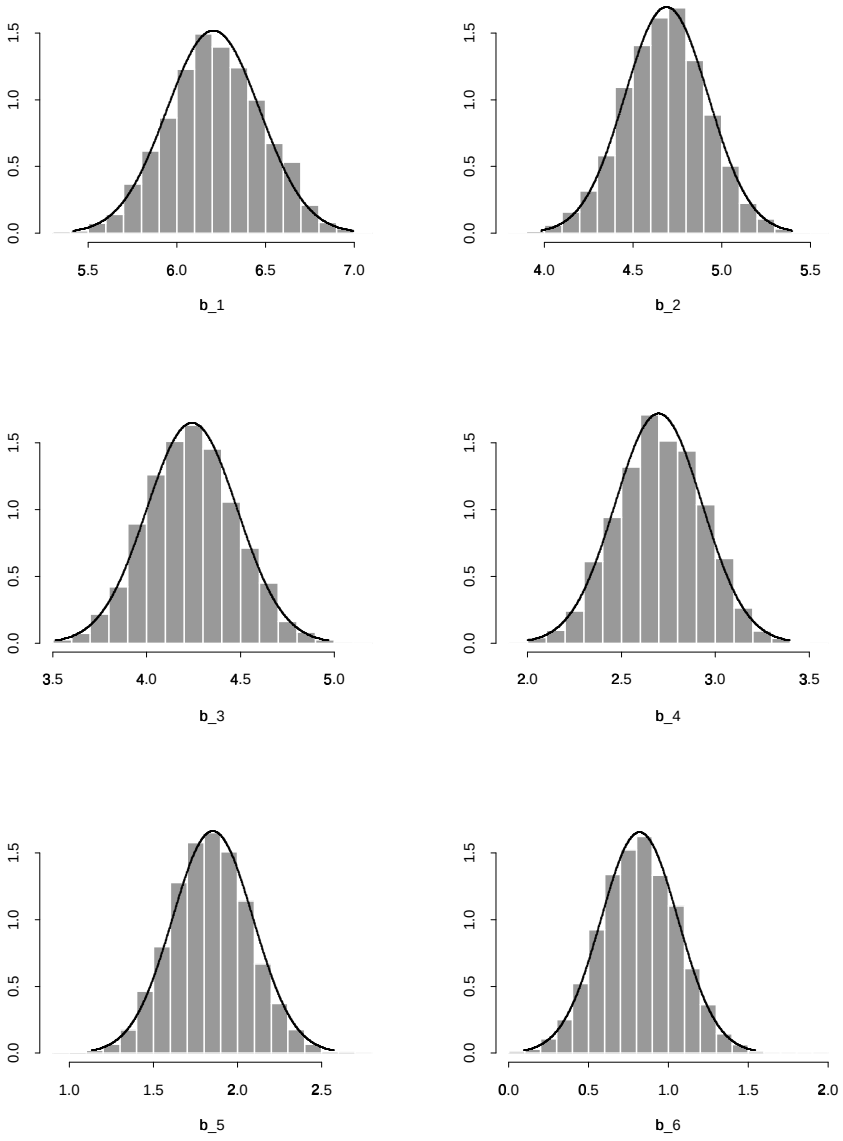

FIGURE 3: *Histograms and true marginal posteriors.*

Carlo estimate of the densities of the marginal posterior distributions of the parameters $\beta_1, \ldots, \beta_6$ given $\beta_7 = ... = \beta_{10} = 0$, $\beta_j \neq 0$, $j = 1, \ldots, 6$. In each plot the continuous line gives the corresponding exact density.

For our second example we used the Wölfer sunspot numbers data, reported as Series E in Box et al. (1994). We modeled the data as a mixture of Gaussian autoregressive processes of order p (AR(p)), with p ranging from zero to ten. In an attempt to represent a vague opinion about the parameters of the model, we specified the prior distribution in two steps in the following way. Given p, the data were assumed to satisfy the autoregressive relation

$$x_t = \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + \epsilon_t, \tag{4}$$

where, conditionally on an additional parameter σ_ϵ^2, the innovations ϵ_t's are independent identically distributed Normal random variables with mean zero and variance σ_ϵ^2. We took the parameters of the AR(p) process (4) to be uniformly distributed in the stationarity region

$$S_p = \left\{(\phi_1, \ldots, \phi_p) : z \in \mathbb{C}, |z| \leq 1 \Rightarrow 1 - \phi_1 z - \ldots - \phi_p z^p \neq 0\right\}.$$

Note that, for $p < \tilde{p}$, one has the inclusion

$$S_p \times \{0_{\tilde{p}-p}\} \subset S_{\tilde{p}}.$$

The variance of the observations, σ^2, which is equal to σ_ϵ^2 times a rational function of the ϕ_j's (Box et al. 1994, Sec. 3.2.2), was taken to be independent of the ϕ_j's with an Inverse Gamma distribution. Note that, setting σ^2 as the scale parameter, instead of σ_ϵ^2, makes it more easily interpretable across models of different order. In practice, instead of using the autoregressive parameters, it is easier to work with the partial autocorrelations $(\pi_1, \ldots, \pi_p)$. As shown in Barndorff-Nielsen and Schou (1973), the vector of partial autocorrelations, as a function of

$(\phi_1, \ldots, \phi_p)$, determines a one-to-one differentiable correspondence between S_p and the p-dimensional cube $(-1, 1)^p$. Moreover, a uniform distribution over S_p corresponds to a product of Beta distributions for the π_j's (Jones, 1987). For the order p we assumed a uniform prior distribution over $\{0, 1, \ldots, 10\}$. After transforming the posterior distribution into a 12-dimensional absolutely continuous distribution as described in the previous section, we ran a Gibbs sampler using the freely distributed code by Gilks for Adaptive Rejection Metropolis Sampling (`http://www.mrc-bsu.cam.ac.uk/pub/methodology/adaptive_rejection/`) on each full conditional distribution. We used `BOA`, another freely available piece of software (`http://www.public-health.uiowa.edu/boa/`) that interactively performs a number of convergence tests on the output of a Markov chain, to assess the convergence of our chain to its stationary distribution. We ran a chain for 100000 iterations and saved only one draw every five. Of the 20000 saved draws we used the first 5000 as burn in and then repeated the same procedure to obtain five different chains starting from five different (over-dispersed) points. Standard convergence tests showed no particular evidence of global convergence troubles for our simulations, (see for instance Brooks and Gelman test displayed in Figure 4) so we used the combined output of the Markov chains to estimate posterior quantities of interest. Table 2 displays the posterior probability of each model S_p, and also the Bayes

TABLE 2: *Posterior probabilities and Bayes factors in favor of S_2*

	S_0	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
Posterior probability	0.000	0.000	0.436	0.394	0.119	0.026	0.007	0.004	0.012	0.002	0.000
Bayes Factor for S_2	$> 10^4$	$> 10^4$	–	1.1	3.7	17	66	121	36	182	1090

factors in favor of S_2, which has the highest posterior probability, against the other models. Figure 5 displays the contour plot of the posterior distribution of the autoregressive

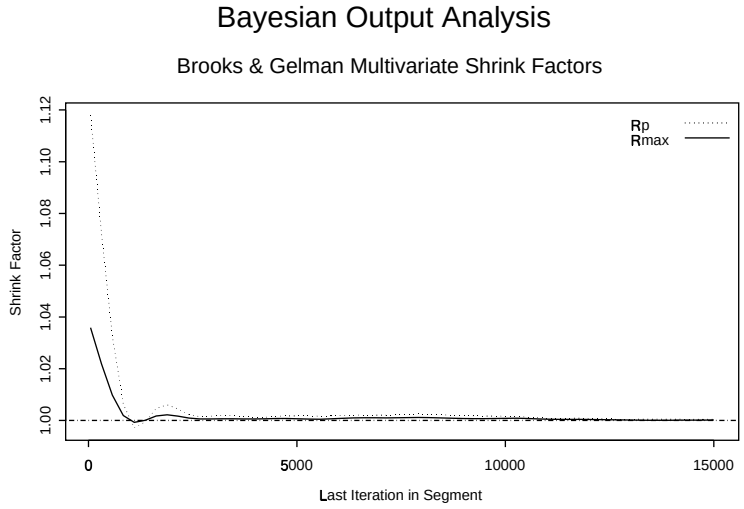

FIGURE 4: *Convergence Diagnostics*

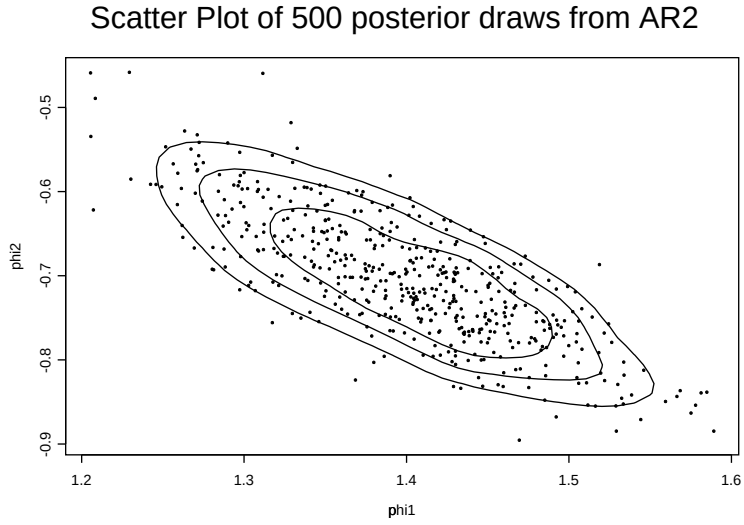

FIGURE 5: *Estimated contour plot of the posterior density of the parameters in S_2: probability regions at level 0.95, 0.75, 0.50.*

parameters conditionally on S_2.

The results of our analysis show an overall agreement with the findings of Box et al. (1994).

Let us stress the fact that, after a simple transformation of the density, the analysis could be carried out effectively, including convergence assessment of the Markov chain, using standard techniques and publicly available software. A Bayesian treatment of mixtures of autoregressive models using reversible jump MCMC can be found in Barbieri and O'Hagan (1996), Troughton and Godsill (1997) and Huerta and West (1999).

5 Discussion

We have proposed a new strategy of simulating from mixtures with nested components whose basic tenet relies on the fact that generating a random vector from an absolutely continuous distribution, known up to a multiplicative constant, is a well studied problem for which many efficient solutions exist. General purpose algorithms to sample from an absolutely continuous distribution include Acceptance/Rejection sampling, Adaptive Rejection Sampling (Gilks and Wild, 1992), Adaptive Rejection Metropolis Sampling (Gilks, Best and Tan, 1995), Metropolis-Hastings and many other MCMC algorithms. In our numerical applications we used a hybrid MCMC method known as Metropolis-within-Gibbs and our method proved itself easy to implement and effective with simulated data as well as with a real data set.

Since other methods exist to solve the problem tackled in this paper, we wish to spend a few words to compare our approach to some popular alternatives. The most immediate and standard way of generating from a mixture distribution, known up to a constant, requires the preliminary evaluation of the relative weights of the components of the mixture. The

evaluation of these weights might require formidable efforts of numerical integration, especially in high dimensions, possibly incurring in numerical instability. All these issues are avoided by our simulation method. Another powerful simulation method that can be effectively applied to sample from mixture distributions is the reversible jump MCMC (Green, 1995). A detailed and accurate comparison between the reversible jump sampler and our strategy would be probably meaningless depending necessarily on particular examples and the fine tuning of the specific implementations. However, in the special case of nested models, we believe that our method can compete with it on the basis of the following appealing features:

- The strategy proposed here is more suited for a simple automatic use, avoiding the delicate step of efficiently selecting the moves between different components of the mixture. The proposal moves to be implemented in the reversible jump sampler are sometimes devised with difficulty without relying on any intuition.

- Often in the actual implementation of the reversible jump sampler at each iteration of the chain one moves to components whose dimensions differ by one from the current one, while our method automatically produces an algorithm where one can move at each iteration from one component to any one of the others. This last feature should produce a faster mixing.

- With our method all the standard convergence diagnostic techniques developed for fixed dimension MCMC can be used, while convergence monitoring of a reversible jump MCMC sampler is much more difficult (see, for example, Richardson and Green, 1997) and still matter of investigation (see Brooks and Giudici, 1998).

We do not believe our method can always be competitive in terms of computational burden

but it can provide a simple automatic resource (or alternative) in situations where numerical approximations are out of discussion and reversible jump proposals are difficult to devise.

Another method to sample from a mixture distribution is the one proposed by Carlin and Chib (1995). Carlin and Chib's approach, although in spirit more similar to ours than reversible jump, aiming at reducing the problem in a space of fixed dimension, is likely to be less automatic and efficient. In fact, again, it requires the fine tuning of what they call pseudo-priors, based on preliminary posterior analyses conducted separately for all of the different models considered. Moreover, at each sweep the sampler must draw all the parameters of all the models, not only those of the one currently being visited by the chain.

To conclude, let us point out that the method proposed in the present paper to deal with mixtures of nested models can be extended to more general settings where component models are not necessarily nested. In fact, in principle one can embed all the models into a single parameter space of fixed dimension as we did in the nested case. Efficient ways of realizing the embedding and extensions of the algorithm in Section 3 to non-nested settings are the object of current investigation.

References

Barbieri, M. M. and O'Hagan, A., (1996), "A reversible jump MCMC sampler for Bayesian analysis of ARMA time series," *Unpublished manuscript*, Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università "La Sapienza", Roma.

Barndorff-Nielsen, O. and Schou, G., (1973), "On the parametrization of autoregressive models by partial autocorrelations," *J. Multivariate Anal.*, 3, 408–419.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C., (1994), *Time Series Analysis Forecasting and Control (Third Edition)*, Prentice-Hall.

Brooks, S. P. and Giudici, P., (1998), "Convergence assessment for Reversible Jump MCMC Simulations," *Bayesian Statistics 6,* Oxford University Press.

Carlin, B. and Chib, S., (1995), "Bayesian model choice via Markov chain Monte Carlo methods," *Journal of the Royal Statistical Society, Series B, Methodological* 57, 473–484.

Gilks, W. R., Best, N. G. and Tan, K. K. C., (1995), "Adaptive rejection Metropolis sampling within Gibbs sampling," *Applied Statistics,* 44, 455–472.

Gilks, W. R. and Wild, P., (1992), "Adaptive rejection sampling for Gibbs sampling," *Applied Statistics,* 41, 337–348.

Green, P. J., (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika* 82, 711–732.

Huerta, G. and West, M. (1999), "Priors and component structures in autoregressive time series models," *Journal of the Royal Statistical Society, Series B, Methodological,* 61, 881–899.

Jones, M. C., (1987), "Randomly choosing parameters from the stationarity and invertibility region of autoregressive-moving average models," *Applied Statistics,* 36, 134–138.

Mardia, K. V., Kent, J. T. and Bibby, J. M., (1979), *Multivariate Analysis*, Academic.

Petris, G., and Tardella, L., (1998), "Simulating from mixture distributions with applications to Bayesian model selection," *Technical Report 681,* Department of Statistics, Carnegie Mellon University.

Richardson, S. and Green, P., (1997), "On Bayesian analysis of mixtures with an un-

known number of components," *J. Roy. Statist. Soc. B,* 59, 731–792. (With discussion).

Tardella, L., (1999), "Some topics in Bayesian methodology", Ph.D. Thesis, Duke University.

Tierney, L., (1994), "Markov chains for exploring posterior distributions," *Ann. Statist.* **22**, 1701–1762. (With discussion).

Troughton, P. T. and Godsill, S. J., (1997), "A reversible jump sampler for autoregressive time series, employing full conditionals to achieve efficient model space moves," *Technical Report 304,* Department of Engineering, University of Cambridge.