Markov Chain Monte Carlo With Mixtures of Mutually Singular Distributions

Raphael GOTTARDO and Adrian E. RAFTERY

Markov chain Monte Carlo (MCMC) methods for Bayesian computation are mostly used when the dominating measure is the Lebesgue measure, the counting measure, or a product of these. Many Bayesian problems give rise to distributions that are not dominated by the Lebesgue measure or the counting measure alone. In this article we introduce a simple framework for using MCMC algorithms in Bayesian computation with mixtures of mutually singular distributions. The idea is to find a common dominating measure that allows the use of traditional Metropolis–Hastings algorithms. In particular, using our formulation, the Gibbs sampler can be used whenever the full conditionals are available. We compare our formulation with the reversible jump approach and show that the two are closely related. We give results for three examples, involving testing a normal mean, variable selection in regression, and hypothesis testing for differential gene expression under multiple conditions. This allows us to compare the three methods considered: Metropolis–Hastings with mutually singular distributions, Gibbs sampler with mutually singular distributions, and reversible jump. In our examples, we found the Gibbs sampler to be more precise and to need considerably less computer time than the other methods. In addition, the full conditionals used in the Gibbs sampler can be used to further improve the estimates of the model posterior probabilities via Rao–Blackwellization, at no extra cost.

Key Words: Gibbs sampler; Metropolis-Hastings algorithm; Mixture distribution; Rao-Blackwellization; Reversible jump; Singular measures.

1. INTRODUCTION

Mixtures of mutually singular distributions arise quite often in statistics. For example, one could model a process that truly is a mixture of a discrete process and a continuous process. One could also be interested in model selection where the dimension of the parameter space varies, giving rise to singularities in the prior distribution. It seems that mixtures

Raphael Gottardo is Assistant Professor of Statistics, Department of Statistics, University of British Columbia, 333-6356 Agricultural Road, Vancouver, BC, V6T 1Z2 (E-mail: *raph@stat.ubc.ca*; Web: *http://www.stat.ubc.ca/~raph*). Adrian E. Raftery is Blumstein-Jordan Professor of Statistics and Sociology, University of Washington, Department of Statistics Box 354322, Seattle, WA 98195-4322 (E-mail: *raftery@stat.washington.edu*; Web: *www.stat.washington.edu/raftery*).

Journal of Computational and Graphical Statistics, Volume 17, Number 4, Pages 949–975
DOI: 10.1198/106186008X386102

of mutually singular distributions are often avoided because of the measure-theoretic difficulties. Perhaps one reason is that the derivation of the density (i.e., the Radon–Nikodym derivative) is not as intuitive as when the distribution is purely discrete or continuous. However, the difficulty is not great and the goal of this article is to introduce an easily used framework that would facilitate the use of such mixtures. We are particularly interested in Markov chain Monte Carlo methods where the target distribution is of this form. We focus on Bayesian inference, where the target distribution is a posterior distribution, though the method is more general.

The Metropolis–Hastings algorithm (Metropolis et al. 1953; Hastings 1970) is a method for constructing a reversible Markov chain with a specified invariant distribution. The Metropolis–Hastings algorithm has been widely used in Bayesian inference to approximate posterior quantities of interest (Geman and Geman 1984; Gelfand and Smith 1990). In most applications, the Metropolis–Hastings algorithm is used when the dominating measure is the Lebesgue measure, the counting measure, or a product of these. However, the algorithm works for any target distribution with a σ-finite dominating measure (Tierney 1994, 1998)

Even though the theory of MCMC can be used with general dominating measures such as sums of mutually singular distributions, this is rarely done in practice. Simple mixtures of mutually singular distributions such as a point mass and a continuous distribution (with respect to the Lebesgue measure) have been used in Bayesian variable selection (Smith and Kohn 1996; Geweke 1996; George and McCulloch 1997). This approach was not considered in an earlier paper by George and McCulloch (1993). Such mixtures also arise with Dirichlet processes in Bayesian density estimation (Escobar and West 1995; Neal 2000) and identification of regeneration times in MCMC simulation (Brockwell and Kadane 2005). There is a need for a general formulation for MCMC computations with mixtures of mutually singular distributions. Note that the term “mixture of mutually singular distributions” was not used in these previous formulations, perhaps because they were not fully general, considering only point masses.

Here we are interested in more complicated situations where we have several distributions of different dimensions. These include Bayesian model selection where some of the parameters are allowed to lie in a hyperplane, or more generally in a submanifold of $\mathbb{R}^n$. There has been a great deal of work on MCMC algorithms for Bayesian model selection. Madigan and York (1995) and Raftery et al. (1997) integrated over the parameter space analytically and made the MCMC move only in the model space; the resulting method is called MCMC model composition, or MC^3. This avoids the issue of mutually singular distributions, but is not applicable to all such problems. Carlin and Chib (1995) used a product space approach to keep the dimension of the parameter space fixed. Following the pioneering work of Grenander and Miller (1994) and Phillips and Smith (1995) based on jump diffusions, Green (1995) showed how to construct a reversible jump MCMC algorithm to handle cases where the dimension of the parameter space is allowed to vary. Petris and Tardella (2003) introduced a geometric approach to transdimensional MCMC and showed that it can be used to formulate the problem as a mixture of distributions with components supported by subspaces of different dimensions. In this article, we show that reversible jump can be viewed in terms of a mixture of mutually singular distributions. Our formulation is more general than Petris and Tardella (2003) as it can deal with nonnested models, and it is computationally easier as there is no need to transform the parameters. In addition, using three examples, we show how the Gibbs sampler can be used within our framework.

The article is organized as follows. In Section 2, we introduce some notation and show how one can derive densities for mixtures of mutually singular distributions. In Section 3, we briefly review the Metropolis–Hastings algorithm and show how it can be used to form an ergodic chain with a mixture of mutually singular distributions as the invariant distribution. In Section 4, we use three examples to demonstrate the methodology introduced and compare various Metropolis–Hastings algorithms including the Gibbs sampler. In Section 5, we compare our formulation with the reversible jump approach and show that the two are closely related. Finally, in Section 6 we discuss possible extensions and the limitations of our formulation.

2. DENSITIES FOR MIXTURES OF MUTUALLY SINGULAR DISTRIBUTIONS

In this section, we show how one can derive densities for a mixture of singular distributions. Without loss of generality, we let our sample space be the n-dimensional Euclidean space $\mathbb{R}^n$, or a subset of it. We denote the n-dimensional Lebesgue measure by λ_n and the Dirac measure concentrated at $\mathbf{x}$ by $\delta_{\mathbf{x}}$. We will say that a probability measure Π is dominated by a σ-finite measure ν, if Π admits a density with respect to ν. In other words, if we can write $\Pi(\mathbf{dx}) = d\Pi/d\nu(\mathbf{x})\nu(\mathbf{dx})$, where $d\Pi/d\nu$ denotes the density (also known as the Radon–Nikodym derivative) of Π with respect to ν. In cases where the dominating measure to be used is clear from the context, we will just use lower-case letters (e.g., π, q) to denote the densities of the corresponding probability measures denoted in capital letters (e.g. Π, Q).

Before introducing the main result, we first recall that two measures ν_1 and ν_2 are said to be mutually singular, denoted by $\nu_1 \perp \nu_2$, if there exists a (measurable) set A such that $\nu_1(A) = 0$ and $\nu_2(A^c) = 0$ where A^c denotes the complement of A. Basically, the two measures are supported on disjoint subsets. Similarly, a countable collection of measures $\nu_i, i \in I$ are said to be mutually singular if $\nu_i \perp \nu_j$ for each $i, j \in I, i \neq j$. Finally, we will say that two distributions are mutually singular if the corresponding probability measures are mutually singular.

The following theorem, based on the Radon-Nikodym theorem (Billingsley 1995, p. 422), gives a way of explicitly writing down densities for mixtures of mutually singular distributions.

Theorem 1. *Let Π, Π_i, $i \in I$, be probability measures such that $\Pi = \sum_{i\in I} w_i \Pi_i$, $w_i \in [0, 1]$, $\sum_{i\in I} w_i = 1$ and Π_i is dominated by ν_i, where the ν_i's are mutually singular σ-finite measures. Then the density (Radon-Nikodym derivative) of Π with respect to $\nu \equiv$*

$\sum_{i \in I} \nu_i$ *is given by*

$$\frac{d\Pi}{d\nu}(\mathbf{x}) = \sum_{i \in I} w_i \frac{d\Pi_i}{d\nu_i}(\mathbf{x}) \mathbf{1}_{S_i}(\mathbf{x}),$$

where the S_i's are sets such that $\nu_i(S_i^c) = 0$ and $\nu_j(S_i) = 0,\ i \neq j$.

The proof is given in Appendix A.1. If we let π and π_i be the densities of Π and Π_i with respect to ν and ν_i, respectively, then Theorem 1 can be written as

$$\pi(\mathbf{x}) = \sum_{i \in I} w_i \pi_i(\mathbf{x}) \mathbf{1}_{S_i}(\mathbf{x}).$$

Thus, the theorem says that the density π with respect to the global dominating measure ν, can be expressed as a mixture of the "componentwise" densities π_i. In general, the componentwise densities π_i are known or are easy to derive, and therefore π can easily be derived. For example, if ν_i is the Lebesgue measure, π_i is simply the usual Lebesgue density. Note, however, that the indicators $\mathbf{1}_{S_i}$ are crucial to get a proper density. They are there to make sure that a set does not contribute to more than one component, which would not be legitimate since the ν_i's (and thus the Π_i's) are mutually singular. If a set is counted more than once, the density might not even integrate to one! A direct consequence of this is that if $\mathbf{x} \in S_i$, then $\pi(\mathbf{x}) = w_i \pi_i(\mathbf{x})$, and even if I is large we only have to evaluate a single π_i. The sets S_i given in Theorem 1 need to be derived on a case-by-case basis because they depend on the supports of the measures ν_i. However, as we will see in the example below and the ones explored in Section 4, their derivation will be clear from the context. It is also possible to get an explicit (and unique) derivation using the notion of Hausdorff measures (Gottardo and Raftery 2004).

Example 1: Mixture of a point mass and a continuous random variable. Let X_1 be a discrete random variable equal to zero with probability one; thus its induced probability measure is the Dirac mass at zero $\Pi_1 \equiv \delta_0$. Let X_2 be a continuous random variable with probability measure Π_2 dominated by the Lebesgue measure λ; thus X_2 admits a density with respect to the Lebesgue measure. Define a third random variable Y equal to X_1 with probability $(1 - w)$ and to X_2 with probability w.

The probability measure of Y can be written as $\Pi = (1 - w)\Pi_1 + w\Pi_2$. Applying Theorem 1 we obtain the density of Π with respect to $(\delta_0 + \lambda)$, namely

$$\frac{d\Pi}{d(\delta_0 + \lambda)}(x) = (1 - w)\frac{d\Pi_1}{d\delta_0}(x)\mathbf{1}_{S_1}(x) + w\frac{d\Pi_2}{d\lambda}(x)\mathbf{1}_{S_2}(x) \text{ a.e. } \delta_0 + \lambda,$$

where "a.e. $\delta_0 + \lambda$" means almost everywhere with respect to the measure $\delta_0 + \lambda$, that is, everywhere except on a set that has $(\delta_0 + \lambda)$-measure equal to zero. If Y is equal to zero, we know that it comes from the first component, whereas if it is not, it must come from the continuous distribution. Thus, natural choices for S_1 and S_2 are $\{0\}$ and $\mathbb{R} \setminus \{0\}$, respectively. It follows that the density of Y can be written as

$$\pi(x) \equiv \frac{d\Pi}{d(\delta_0 + \lambda)}(x) = (1 - w)\mathbf{1}_{\{0\}}(x) + wf(x)\mathbf{1}_{\mathbb{R}\setminus\{0\}}(x) \text{ a.e. } \delta_0 + \lambda,$$

where $f(x)$ is the Lebesgue density of X_2. Note that it is crucial to remove $\{0\}$ from $\mathbb{R}$ in the second indicator function for the density to be a valid density with respect to $(\delta_0 + \lambda)$. If we did not do that, the density evaluated at 0, which corresponds to the probability that Y is zero, would be $(1 - w) + wf(0)$ and not $1 - w$.

3. MARKOV CHAIN MONTE CARLO WITH MIXTURES OF MUTUALLY SINGULAR DISTRIBUTIONS

We now consider Markov chain Monte Carlo algorithms for mixtures of mutually singular distributions. We show how it is possible to use the regular Metropolis–Hastings and Gibbs sampler algorithms, where Theorem 1 is used to derive the required densities.

Suppose that we wish to sample from a distribution for a variable $\mathbf{x}$ with associated probability measure $\Pi(\mathbf{dx})$ dominated by a σ-finite measure ν. We use π to denote its density, so that $\Pi(\mathbf{dx}) = \pi(\mathbf{x})\nu(\mathbf{dx})$. In this article, we shall be concerned with a posterior distribution with density $\pi(\mathbf{x}|\mathbf{y})$, or more generally a full conditional $\pi(\mathbf{x}|\mathbf{y}, \mathbf{z})$ when blocks of variables are updated in turns, where $\mathbf{y}$ are the data and $\mathbf{z}$ are the variables that are being conditioned upon.

From now on, we use $\pi(\mathbf{x})$ to denote the target distribution, in this case $\pi(\mathbf{x}|\mathbf{y}, \mathbf{z})$. One of the most widely used methods for generating such chains is the Metropolis–Hastings algorithm. MCMC methods have been widely used to generate (dependent) samples from distributions where the normalizing constant is unknown or intractable. Most applications deal with distributions where the dominating measure is either the Lebesgue measure, the counting measure, or a product of those. Not much attention has been given to measures that are not absolutely continuous with respect to the Lebesgue measure or to the counting measure.

To avoid special cases, we assume that $\pi(\mathbf{x}) > 0$ for each $\mathbf{x}$ in the sample space. In Markov chain Monte Carlo, one constructs a Markov chain with invariant distribution π. Following the notation of Tierney (1994), let Q be a Markov transition kernel of the form

$$Q(\mathbf{x}, \mathbf{dx}') = q(\mathbf{x}, \mathbf{x}')\nu(\mathbf{dx}'),$$

and define

$$\alpha(\mathbf{x}, \mathbf{x}') = \begin{cases} \min\left\{\frac{\pi(\mathbf{x}')q(\mathbf{x}', \mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x}, \mathbf{x}')}, 1\right\}, & \text{if} \quad \pi(\mathbf{x})q(\mathbf{x}, \mathbf{x}') > 0 \\ 1, & \text{if} \quad \pi(\mathbf{x})q(\mathbf{x}, \mathbf{x}') = 0. \end{cases} \tag{3.1}$$

Note that, in this definition, both q and π are densities with respect to ν, which in our case will be the sum of several mutually singular measures. If the chain is currently at a point $\mathbf{X}_n = \mathbf{x}$, then a new candidate $\mathbf{x}'$ is generated according to the distribution $Q(\mathbf{x}, \cdot)$ and the new point is accepted with probability $\alpha(\mathbf{x}, \mathbf{x}')$. If the associated Metropolis–Hastings kernel, denoted by K, is Π-irreducible, Harris-recurrent and aperiodic, the Markov chain formed will converge to the unique stationary distribution π with respective probability measure Π (Tierney 1994). Note that the formulation given here includes the Gibbs sampler (Geman and Geman 1984; Gelfand and Smith 1990), where new observations are drawn from the full conditional and the acceptance probability is equal to one. Most applications concern cases where the dominating measure is the Lebesgue measure, the counting measure, or a product of those. However, the general results in Tierney (1994) apply to more general distributions. This was emphasized in a more recent paper (Tierney 1998), where the author described very general conditions under which the Metropolis–Hastings algorithm is reversible.

Thus, the result applies to mixtures of mutually singular distributions. The main difficulty is that the choice of the proposal becomes limited, as mixture components put all their mass on different parts of the parameter space. For example, the usual random walk Metropolis algorithm will not be available in general as symmetric proposals do not exist. This last point will become clearer in Example 2.

Using the notation introduced in Theorem 1, we assume that the target distribution is of the form $\Pi = \sum_{i \in I} w_i \Pi_i$, where Π_i is dominated by ν_i, and the ν_i's are mutually singular σ-finite measures. Thus Π is dominated by $\nu = \sum_i \nu_i$. In a Bayesian context, such singularities would occur when the prior itself is a mixture of mutually singular distributions, whose density with respect to ν can be written as $\pi(\mathbf{x}|\mathbf{z}) = \sum_i w_i \pi_i(\mathbf{x}|\mathbf{z}) \mathbf{1}_{S_i}$ using Theorem 1. Thus, using Bayes theorem it follows that

$$\begin{aligned} \pi(\mathbf{x}|\mathbf{y},\mathbf{z}) &\propto L(\mathbf{y}|\mathbf{x},\mathbf{z})\pi(\mathbf{x}|\mathbf{z}) \\ &= L(\mathbf{y}|\mathbf{x},\mathbf{z}) \sum_i w_i \pi_i(\mathbf{x}|\mathbf{z}) \mathbf{1}_{S_i}(\mathbf{x}) \\ &= \sum_i w_i L(\mathbf{y}|\mathbf{x},\mathbf{z}) \pi_i(\mathbf{x}|\mathbf{z}) \mathbf{1}_{S_i}(\mathbf{x}) \\ &= \sum_i w_i c_i \pi_i(\mathbf{x}|\mathbf{z},\mathbf{y}) \mathbf{1}_{S_i}(\mathbf{x}), \end{aligned} \tag{3.2}$$

where π_i is the "componentwise" full conditional of $\mathbf{x}$ given $(\mathbf{y}, \mathbf{z})$ and that $\mathbf{x}$ belongs to component i (i.e., $\mathbf{x} \in S_i$), and c_i is its normalizing constant, namely $c_i = \int_{\mathbf{x}} L(\mathbf{y}|\mathbf{x},\mathbf{z}) \pi_i(\mathbf{x}|\mathbf{z}) \nu_i(\mathbf{dx})$. Because each $\pi_i \mathbf{1}_{S_i}$ integrates to one with respect to ν it is easy to see that the global normalization constant is $\sum_i w_i c_i$, and it follows that

$$\pi(\mathbf{x}|\mathbf{y},\mathbf{z}) = \sum_i w_i^* \pi_i(\mathbf{x}|\mathbf{z},\mathbf{y}) \mathbf{1}_{S_i}, \tag{3.3}$$

where $w_i^* = c_i w_i / (\sum_i c_i w_i)$. This implies that the posterior Π is a mixture of mutually singular distributions. It also implies that if each c_i is analytically available and π_i can be sampled from directly, one can sample from Π by first selecting a component i at random with probability w_i^* and then sampling from Π_i. So, if Π is a full conditional, and all componentwise full conditionals (the Π_i's) are explicitly available, Gibbs sampling can be used. This will be illustrated with three examples in the next section.

If the full conditional is not available, the Metropolis–Hastings algorithm could be used instead. In order for the Metropolis–Hastings kernel K to be Π-irreducible, it is necessary to construct a Π-irreducible transition kernel Q. In practice, assuming that $\mathbf{x}$ belongs to component i, it will be convenient to write the kernel Q as

$$Q(\mathbf{x}, \mathbf{dx}') = \sum_{j \in I} p_{ij} Q_j(\mathbf{x}, \mathbf{dx}'), \tag{3.4}$$

where the Q_j's are mutually singular transition kernels in the sense that $Q_j(\mathbf{x}, \mathbf{dx}') = q_j(\mathbf{x}, \mathbf{x}')\nu_j(\mathbf{dx}')$, and p_{ij} is the probability of proposing a move to component j from component i. In (3.4), p_{ij} is a between-component move and Q_j is a move within component j. The Π-irreducibility of Q and the associated Metropolis–Hastings kernel will depend on both $P \equiv (p_{ij})$ and the Q_j's, and will need to be verified for each case. However, as with traditional MCMC, this is usually not hard to verify. In our framework, it will be enough to show that there is a positive probability that the chain will reach (i) any component of the mixture and (ii) any set of positive mass (with respect to Π) within that component, from anywhere in the state space, in a finite number of steps. Note that the Gibbs sampler as given in (3.3) is a special case of this where $p_{ij} \equiv w_j^*$ and $Q_j \equiv \Pi_j$.

To fully establish the convergence of the Metropolis–Hastings Markov chain with kernel K, one needs to show that the resulting kernel is also aperiodic and Harris recurrent. Aperiodicity is easily verified and is usually satisfied for the Metropolis–Hastings algorithm even when used as a Gibbs sampler (Tierney 1994; Roberts and Tweedie 1996). Standard MCMC results can be used to verify Harris recurrence (Tierney 1994; Chan and Geyer 1994; Roberts and Tweedie 1996; Roberts and Rosenthal 2004).

4. APPLICATIONS

In this section, we consider three examples that are applications of the Metropolis–Hastings algorithm for Bayesian computation with mixtures of mutually singular distributions. From now on, we denote by $\mathrm{N}(\mu, \sigma^2)$ the Gaussian distribution with mean μ and variance σ^2. The corresponding density evaluated at x is denoted by $\mathrm{N}(x|\mu, \sigma^2)$.

Example 2: Testing a normal mean. Consider the simple Bayesian linear model,

$$\begin{aligned} y_j &= \mu + \epsilon_j, \\ (\epsilon_j|\psi) &\sim \mathrm{N}(0, \psi^{-1}). \end{aligned} \tag{4.1}$$

We might be interested in testing if the mean μ is equal to zero. In order to do so, we need to specify a prior distribution that allows the parameter μ to be equal to zero. We use the prior

$$\mu \sim (1 - w)\delta_0 + w\mathrm{N}(0, \psi_\mu^{-1}), \tag{4.2}$$

which is a mixture of a point mass at 0 and a Gaussian distribution. Using Theorem 1 and Example 1, its density with respect to $(\delta_0 + \lambda)$ is $(1 - w)\mathbf{1}_{\{0\}}(\mu) + w\mathrm{N}(\mu|0, \psi_\mu^{-1})\mathbf{1}_{\mathbb{R}\setminus\{0\}}(\mu)$, where δ_0 is the Dirac mass at zero and λ is the one-dimensional Lebesgue measure.

We let the prior for the precision ψ be $\mathrm{Gamma}(\xi_1, \xi_2)$, which has mean ξ_1/ξ_2 and variance ξ_1/ξ_2^2. The hyperparameters ψ_μ, ξ_1 and ξ_2 will be kept fixed. Here the target distribution can be expressed (up to a normalizing constant) in terms of its density $\pi(\mu, \psi|\mathbf{y})$

with respect to the product measure $(\delta_0 + \lambda) \times \lambda$ as

$$\begin{aligned}\pi(\mu, \psi|\mathbf{y}) &\propto L(\mu, \psi|\mathbf{y})\pi(\mu)\pi(\psi)\\ &\propto \psi^{J/2}\exp(-\frac{\psi}{2}\sum_{j=1}^{J}(y_j-\mu)^2)\\ &\quad\times\left[(1-w)\mathbf{1}_{\{0\}}(\mu) + w\left(\frac{\psi_\mu}{2\pi}\right)^{1/2}\exp(-\frac{\psi_\mu}{2}\mu^2)\mathbf{1}_{\mathbb{R}\setminus\{0\}}(\mu)\right]\\ &\quad\times\psi^{\xi_1-1}\exp(-\xi_2\psi).\end{aligned}$$

Note that here we give the distribution of $\pi(\mu, \psi|\mathbf{y})$ explicitly (up to a normalizing constant), for illustration. However, in practice one does not need to do this, and one would simply update μ and ψ alternatively. The update for ψ presents nothing complicated, as the dominating measure is the Lebesgue measure, and its full conditional distribution is a Gamma distribution whether or not μ is zero. Therefore we shall be concerned only with the update of μ conditional on ψ and $\mathbf{y}$.

The prior density for μ is of the form $w_1\mathbf{1}_{\{0\}}(\mu) + w_2\mathrm{N}(\mu|0, \psi_\mu^{-1})\mathbf{1}_{\mathbb{R}\setminus\{0\}}(\mu)$. Thus, we can use (3.3) with $w_1 = (1-w)$, $w_2 = w$, $\pi_1(\boldsymbol{\mu}|\psi) = 1$ and $\pi_2(\boldsymbol{\mu}|\psi) = \mathrm{N}(\mu|0, \psi_\mu^{-1})$. Now we only need to derive both componentwise full conditionals $\pi_1(\mu|\mathbf{y}, \psi)$ and $\pi_2(\mu|\mathbf{y}, \psi)$. In the case of the first component, we know that $\mu = 0$, and it follows that

$$\pi_1(\mu|\mathbf{y}, \psi) \propto L(\mathbf{y}|0, \psi)\cdot 1,$$

and the normalizing constant is $c_1 = L(\mathbf{y}|0, \psi)$. Similarly, the componentwise full conditional, π_2, is given by

$$\pi_2(\mu|\mathbf{y}, \psi) \propto L(\mathbf{y}|\mu, \psi)\sqrt{\frac{\psi_\mu}{2\pi}}\exp\left(-\frac{\psi_\mu}{2}\mu^2\right),$$

which is just the posterior distribution of a normal mean when the prior is Gaussian. Thus, the normalizing constant, $c_2 = \int L(\mathbf{y}|\mu, \psi)\sqrt{\psi_\mu/(2\pi)}\exp\left(-0.5\psi_\mu\mu^2\right)d\mu$, is available analytically. After straightforward calculation of c_2, the full conditional can be obtained using (3.3) and is given by

$$(\mu|\psi, \mathbf{y}) \sim (1-w^*)\delta_0 + w^*\mathrm{N}\left(\psi\sum_i y_i/(n\psi + \psi_\mu), (n\psi + \psi_\mu)^{-1}\right), \tag{4.3}$$

where

$$w^* = 1 - \frac{1-w}{1-w+w\sqrt{\psi_\mu/(n\psi+\psi_\mu)}\exp(0.5(\psi\sum y_i)^2/(n\psi+\psi_\mu))}.$$

Using this full conditional, a new value for μ can be generated by first selecting a component at random with respective probabilities $1-w^*$ and w^*, and then simulating a new μ from the componentwise full conditional for the selected component. If the first component is selected we simply set μ to zero, and if the second is selected we generate a new value from the corresponding Gaussian distribution.

In some other settings (e.g., non-Gaussian likelihood and/or prior), the componentwise full conditional π_2 might not be explicitly available and the Metropolis–Hastings algorithm provides a good alternative. In order to use Metropolis–Hastings , we need to define an irreducible Markov transition kernel (proposal), $Q(\mu, d\mu') = q(\mu, \mu')\nu(d\mu')$, that is absolutely continuous with respect to $\nu = \delta_0 + \lambda$, where δ_0 is the Dirac measure concentrated at 0, and λ is the Lebesgue measure. In other words, we have to make sure that we propose a move to zero as well as to the remainder of the real line. This can be done using (3.4), which we specify with the density

$$q(\mu, \mu') = (1 - p_i)\mathbf{1}_{\{0\}}(\mu') + p_i q^*(\mu, \mu')\mathbf{1}_{\mathbb{R}\setminus\{0\}}(\mu'), \tag{4.4}$$

where p_i ($i = 0$ if $\mu = 0$ and $i = 1$ if $\mu \neq 0$) is the probability of proposing a move to the continuous component from component i. In other words, with probability $1 - p_i$ we propose a move to zero, and with probability p_i we propose a value according to q^*, the density of a kernel absolutely continuous with respect to the Lebesgue measure. For the Metropolis-Hastings algorithm, the acceptance probability is given by

$$\alpha(\mu, \mu') = \min\left\{\frac{\pi(\mu')q(\mu', \mu)}{\pi(\mu)q(\mu, \mu')}, 1\right\}.$$

In this case it is clear that we cannot find a symmetric proposal q, as this would require that the mass going to zero be the same as the mass leaving zero. The proposal given by (4.4) is clearly not symmetric as $q(0, \mu') \neq q(\mu', 0)$ for $\mu' \neq 0$. Any other proposal will have the same problem because of the singularity between the two measures δ_0 and λ.

We now compare the performance of the Gibbs sampler to two other kernels based on the Metropolis–Hastings algorithm. The first kernel K_1 is a generalization of the random walk Metropolis kernel with proposal density given by

$$q_1(\mu, \mu') = 0.5 \cdot \mathbf{1}_{\{0\}}(\mu') + 0.5\mathrm{N}(\mu'|\mu, \sigma_1^2)\mathbf{1}_{\mathbb{R}\setminus\{0\}}(\mu'),$$

where σ_1^2 is a fixed number. From μ, this proposes 0 with probability 0.5, and proposes a random walk step with probability 0.5. The second kernel, K_2, is the concatenation of two Metropolis–Hastings kernels, with proposal densities

$$q_{21}(\mu, \tilde{\mu}) = \begin{cases} \mathrm{N}(\tilde{\mu}|\hat{\mu}, \sigma_2^2)\mathbf{1}_{\mathbb{R}\setminus\{0\}}(\tilde{\mu}) & \text{if} \quad \mu = 0 \\ \mathbf{1}_{\{0\}}(\tilde{\mu}) & \text{if} \quad \mu > 0, \end{cases}$$

where $\hat{\mu}$ is the sample mean and σ_2^2 is a fixed number, and

$$q_{22}(\tilde{\mu}, \mu') = \begin{cases} \mathbf{1}_{\{0\}}(\mu') & \text{if} \quad \tilde{\mu} = 0 \\ \pi_2(\mu')\mathbf{1}_{\mathbb{R}\setminus\{0\}}(\mu') & \text{if} \quad \tilde{\mu} > 0, \end{cases}$$

where π_2 is the componentwise full conditional of μ for the nonzero component. From μ, this proposes 0 if $\mu \neq 0$ and a nonzero μ' if $\mu = 0$. In the latter case, this nonzero μ' can be simulated in two steps. The first step based on q_{21} proposes $\tilde{\mu}$ from a $\mathrm{N}(\hat{\mu}, \sigma_2^2)$ density where $\hat{\mu}$ is the sample mean and σ_2^2 is a fixed number. If this is accepted, the second step is a componentwise Gibbs step based on the full conditional for the component with nonzero μ. This last step, known as a within-model move in the reversible jump literature, is not necessary for the Markov chain to be ergodic but greatly improves the sampler. It will be convenient when comparing to the reversible jump formulation (Section 5).

We randomly generated 10 observations from a Gaussian distribution with variance 1 and mean 0.5, as follows:

$$0.575, 1.808, 0.532, -0.168, 0.529, 0.888, -1.368, -0.512, 2.667, 0.874.$$

We fitted the model given by Equations (4.1) and (4.2) using $w = 0.5$, that is, each component is equally likely a priori. We fixed $\psi_\mu = 0.01$, $\xi_1 = 1$ and $\xi_2 = 0.05$, corresponding to fairly noninformative priors. Table 1 summarizes the results. The variance proposals $\sigma_1^2 = 0.25$ and $\sigma_2^2 = 1.2$ were chosen to maximize the proportion of moves between components. This was done by tuning the proposal variances using several pilot runs.

The true posterior probability of the model with mean zero is 0.867, conventionally viewed as positive but not strong evidence for the null model (Kass and Raftery 1995). In each case, the null model posterior probability (model with mean zero) can be easily estimated from the MCMC output as the proportion of μ's equal to zero. However, note that in (4.3), $1 - w^*$ corresponds to the probability of $\mu = 0$ given ψ and $\mathbf{y}$, and we can write

$$\Pr(\mu = 0|\mathbf{y}) = \mathbb{E}_\mu[\mathbf{1}_{[\mu=0]}] = \mathbb{E}_\psi\mathbb{E}_\mu[\mathbf{1}_{[\mu=0]}|\psi] = \mathbb{E}_\psi[\Pr(\mu = 0|\mathbf{y}, \psi)] = \mathbb{E}_\psi[1 - w^*],$$

where $\mathbb{E}_T$ is the expectation with respect to the distribution of $T|\mathbf{y}$. Thus, in the case of the Gibbs sampler, one can obtain a more efficient estimate by Rao–Blackwellization when averaging the $1 - w^*$ values computed at each iteration at no extra cost. As a consequence we get four different estimates from K_1, K_2, Gibbs and Gibbs with Rao–Blackwellization.

All four methods considered did well in estimating the posterior probability, with accurate estimates based on 10,000 iterations, and low variability of the estimates. The algorithm based on the proposal K_2 performed better than the one based on K_1. It moved between components almost twice as often, not surprisingly since K_2 forces moves between components. The Gibbs sampler with the simple estimate did essentially as well as Metropolis–Hastings with proposal K_2 in estimating the posterior probability, and it used much less computer time. Note that the proposal K_2 combines two different move types, thus this is not a fair comparison with the Gibbs sampler and this is reflected in the CPU time. Generally speaking, the Metropolis–Hastings algorithm is computationally more expensive than the Gibbs sampler because it requires more evaluations of the target distribution. Finally, the Gibbs sampler with Rao–Blackwellization performed best both in terms of variability and computing time. Overall, therefore, the Gibbs sampler performed best among these three methods for this example.

Example 3: Robust Bayesian variable selection in regression. Variable selection is an important problem whose purpose is to select a group of variables that best predict an outcome variable. Given a dependent variable $\mathbf{Y}$ and a set of potential regressors $\mathbf{X}_1, \ldots, \mathbf{X}_p$, we wish to compare models of the form $\mathbf{Y} = \beta_0 + \mathbf{X}_{i_1}\beta_{i_1} + \cdots + \mathbf{X}_{i_q}\beta_{i_q}$, where $\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_q}$

Table 1. Comparison of the estimated model posterior probabilities computed with each algorithm. The estimates were computed from 10,000 iterations with 1,000 burn-in iterations. The standard deviations were computed by dividing a chain of 10 million iterations into 1,000 batches of 10,000 iterations each. The "truth" was obtained from 10 million iterations on the basis of which the estimates from the three algorithms agreed to within three digits. PMBC is the percentage of time the chain moved from one component to an other. In the case of the Gibbs sampler, RB refers to estimates computed via Rao–Blackwellization.

		MH (K_1)		MH (K_2)		Gibbs			
	Truth	Est.	sd	Est.	sd	Est.	sd	RB-Est.	RB-sd
$\pi(\mu = 0\|\mathbf{y})$	0.867	0.858	0.0054	0.870	0.0031	0.862	0.0034	0.867	0.0005
PMBC(%)		13		25		24			
CPU (μs/iter)		4.9		5.0		1.3			

is a selected subset of $\mathbf{X}_1, \ldots, \mathbf{X}_p$. For this problem we take an approach similar to that of George and McCulloch (1997). However, our model explicitly tries to take account of outliers, by using t-distributed errors. We assume a standard linear model to describe the relationship between the response and dependent variable and the set of predictors, namely

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j + \frac{\epsilon_i}{\sqrt{\varpi_i}}, \\ \epsilon_i &\sim \mathrm{N}(0, \psi^{-1}), \\ (\varpi_i|\nu) &\sim \mathcal{G}a(\nu/2, \nu/2), \end{aligned}$$

where the β_i's are the unknown regression coefficients and the ϖ_i's are independent of the ϵ_i's. Since the ϖ_i's are independent of the ϵ_i's, we have $\epsilon_i/\sqrt{\varpi_i} \sim t_{(\nu,0,\psi^{-1})}$, that is, the errors have a t distribution with ν degrees of freedom and scale parameter ψ^{-1}. The advantage of writing the model in this way is that, conditioning on the ϖ_i, the sampling errors are again Gaussian, but with different precisions.

In order to allow each variable to be in and out of the model, we use a mixture of a Gaussian distribution and a point mass at zero for the prior of each regression coefficient, as follows:

$$\beta_k \sim (1 - \omega)\delta_0 + \omega \mathrm{N}\left(0, \frac{S_Y^2}{S_{X_k}^2}\sigma_\beta^2\right),$$

where ω is the prior probability for each variable of being in the model, $S_{X_k}^2$ is the empirical variance of the kth predictor, S_Y^2 is the empirical variance of the response variables and σ_β^2 is a common variance parameter. The prior for the variance parameter σ_β^2 is taken to be uniform on the interval $[0, 1]$; arguments for priors of this kind are given by Raftery et al. (1997). The prior for the scaling parameter of the t-distribution, ψ is taken to be improper, $\pi(\psi) \propto \psi^{-1}$. We also tried a spread-out proper prior for ψ and the results were almost identical. The prior for the intercept β_0 is taken to be Gaussian with a large variance centered at the least squares estimate for the full model with all variables, $\hat{\beta}_0$, namely $\mathrm{N}(\hat{\beta}_0, 20\,\mathrm{se}(\hat{\beta}_0)^2)$. Finally, the prior for the degrees of freedom ν is uniform on the set $\{1, 2, 4, 8, 16, 32\}$ as suggested by Besag and Higdon (1999). For each β_k we use $w = 0.5$ which makes every model equally likely a priori.

As in Example 2, the full conditional of β_k can be derived explicitly. All we need is to derive both componentwise full conditionals $\pi_1(\beta_k|\cdots)$ and $\pi_2(\beta_k|\cdots)$ for the first and second components, respectively, where $\beta_k|\cdots$ means β_k conditioning on everything else in the model. In the case of the first component $\beta_k = 0$, and it follows that

$$\pi_1(\beta_k|\cdots) \propto L(\mathbf{y}|\beta_0, \beta_k = 0, \boldsymbol{\beta}_{-k}, \psi) \cdot 1$$

and the normalizing constant is $c_1 = L(\mathbf{y}|\beta_k = 0, \boldsymbol{\beta}_{-k}, \psi)$, where $\boldsymbol{\beta}_{-k}$ is the vector of all regression coefficients except the kth one. Similarly, the componentwise full conditional, π_2, is given by

$$\pi_2(\beta_k|\cdots) \propto L(\mathbf{y}|\beta_0, \boldsymbol{\beta}, \psi)\sqrt{\frac{\psi_{\beta_k}}{2\pi}} \exp\left(-\frac{\psi_{\beta_k}}{2}\beta_k^2\right),$$

where $\psi_{\beta_k}^{-1} = S_Y^2/S_{X_k}^2\sigma_\beta^2$, which is again a Gaussian distribution. As in Example 2, the normalizing constant, $c_2 = \int L(\mathbf{y}|\beta_0, \boldsymbol{\beta}, \psi)\sqrt{\psi_{\beta_k}/(2\pi)} \exp\left(-0.5\psi_{\beta_k}\beta_k^2\right) d\beta_k$, is again available analytically. After straightforward calculation of c_2, the full conditional can be obtained using (3.3) and is given by

$$(\beta_k|\ldots) \sim (1 - w_k^*)\delta_0 + w_k^*\mathrm{N}\left(\psi\sum_i r_{ik}/(\psi\sum_i \varpi_i X_{ik}^2 + \psi_{\beta_k}), (\psi\sum_i \varpi_i X_{ik}^2 + \psi_{\beta_k})^{-1}\right),$$

where

$$w_k^* = 1 - \frac{1 - w}{1 - w + w\sqrt{\psi_\beta/(\psi\sum_i w_i X_{ik}^2 + \psi_{\beta_k})} \exp(0.5(\psi\sum r_{ik})^2/(\psi\sum_i \varpi_i X_{ik}^2 + \psi_{\beta_k}))},$$

with

$$r_{ik} = \varpi_i(y_i - \beta_0 - \sum_{j \neq k} \beta_j X_{ij})X_{ik}.$$

As in Example 2, in some nonconjugate settings (e.g., non-Gaussian likelihood and/or prior), the componentwise full conditional π_2 might not be explicitly available and the Metropolis–Hastings algorithm provides a good alternative. We now compare the Metropolis–Hastings algorithm with the kernels K_1 and K_2 as defined in the previous example. For a given β_k, the mean of the continuous component of the proposal q_{21} was set to the least squares estimate of the corresponding coefficient based on the full model. The width of each proposal was chosen to maximize the proportion of moves between components.

To illustrate the variable selection method, we use the Stack Loss data of Brownlee (1965), previously considered by many authors including Daniel and Wood (1980), Atkinson (1985) and, in a Bayesian framework, by Hoeting et al. (1996). It consists of 21 days of operation from a plant for the oxidation of ammonia as a stage in the production of nitric acid. The response is called "stack loss," defined as the percentage of unconverted ammonia that escapes from the plant. There are three independent variables, $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$.

Table 2. Comparison of the estimated posterior inclusion probabilities of each regression coefficient for Example 3. Estimates were computed from 10,000 iterations with 1,000 burn-in iterations. The standard deviations were computed by dividing a chain of 10 million iterations into 1,000 batches of 10,000 iterations each. The “truth” was obtained from ten billion iterations on the basis of which the estimates from the three algorithms agreed to within three digits. PMBC (percent moves between components) is the percentage of time the chain moved from one component to an other. The Gibbs sampler performed better than the Metropolis–Hastings samplers. RB refers to estimates computed via Rao–Blackwellization.

	Truth	MH (K_1)		MH (K_2)		Gibbs			
		Est.	sd	Est.	sd	Est.	sd	RB-Est.	RB-sd
$\Pi(\beta_1 \neq 0 \mid \mathbf{y})$	1.000	1.000	0.015	1.000	0.005	1.000	0.005	1.000	0.004
$\Pi(\beta_2 \neq 0 \mid \mathbf{y})$	0.839	1.000	0.315	0.912	0.131	0.859	0.118	0.858	0.118
$\Pi(\beta_3 \neq 0 \mid \mathbf{y})$	0.141	0.244	0.230	0.182	0.093	0.097	0.085	0.097	0.085
PMBC % (β_1)		0.0		0.0		0.0			
PMBC % (β_2)		0.4		0.5		1.0			
PMBC % (β_3)		1.1		1.9		2.5			
CPU (μs/iter)		67		68		45			

The airflow $\mathbf{X}_1$ measures the rate of operation of the plant. The nitric oxides produced are absorbed in a counter-current absorption tower: $\mathbf{X}_2$ is the inlet temperature of cooling water circulating through coils in this tower and $\mathbf{X}_3$ is proportional to the concentration of acid in the tower. Small values of the response correspond to efficient absorption of the nitric oxides. The general consensus with the Stack Loss data is that the predictor $\mathbf{X}_3$ (acid concentration) should be dropped from the model and that observations 1, 3, 4, and 21 are outliers.

As in the previous example, the posterior inclusion probability for each regression coefficient, defined as the posterior probability that the coefficient is not equal to zero, can be easily estimated from the MCMC output as the proportion of zeros. Again, in the case of the Gibbs sampler one can obtain a more efficient estimate by Rao–Blackwellization when averaging the $1 - w^*$ values computed at each iteration.

The algorithms are compared in Table 2. This time, the Gibbs sampler outperformed the other two algorithms in terms of standard error for the estimated posterior inclusion probabilities. This is not surprising. Variables move in and out of the model over the course of the MCMC run, changing the relative estimates of the β’s and making it hard to construct an efficient proposal. The Gibbs sampler is automatic and depends on the current value of the other coefficients currently in the model. The average number of moves between components is higher for the Gibbs sampler than for the other two algorithms, and the estimated variances for the estimate of the posterior probabilities are smaller. Finally, the computing time is significantly reduced. Note that this time the improvement from the RB estimates is not great. This is due to the fact that, in this model, there are more parameters than in the previous one and the marginalization of β represents only a small portion of the overall number of parameters.

In this example, we are usually interested in the posterior probabilities of all $2^3 = 8$ possible models. These can easily be computed from the posterior samples of the β_k’s, and

Table 3. Estimated posterior model probabilities for the *Stack Loss* Data. The total posterior probability for the other models visited was less than 0.01.

Models	1	1,2	1,3	1,2,3
Post. prob.	0.13	0.72	0.02	0.13

are given in Table 3. The posterior probabilities are consistent with the general consensus about the data.

The posterior mode of the number of degrees of freedom, ν, is 4, suggesting that the observations are much heavier tailed than Gaussian. The posterior weights, ϖ, from the model are summarized in Table 4. It shows that observations 1, 3, 4, and 21 are downweighted by our model.

Finally, even though we have chosen to update each regression coefficient in turn, it would be possible to update the β_k's in blocks of 2 or more. In particular, the full conditionals can be derived explicitly and the Gibbs sampler could be used. For example, the full conditional of any two regression coefficients, $\boldsymbol{\beta}_{j,k} = (\beta_j, \beta_k)$, is now a mixture of four mutually singular distributions of the form,

$$(\boldsymbol{\beta}_{j,k}|\ldots) \sim w_1^*\delta_0\delta_0 + w_2^*\mathrm{N}\left(\mu_j^*, \psi_j^{*-1}\right)\delta_0 + w_3^*\delta_0\mathrm{N}\left(\mu_k^*, \psi_k^{*-1}\right) + w_4^*\mathrm{N}_2\left(\boldsymbol{\mu}_{j,k}^*, \boldsymbol{\Psi}_{j,k}^{*\;-1}\right),$$

where $w_1^*, w_2^*, w_3^*, w_4^*$, μ_j^*, μ_k^*, ψ_j^*, ψ_k^*, and $\boldsymbol{\mu}_{j,k}^*$, $\boldsymbol{\Psi}_{j,k}^*$ are constants depending on the data and the parameters that are being conditioned upon. The first component sets both regression coefficients to zero, the second (resp. third) sets β_k (resp. β_j) to zero while generating a new β_j (resp. β_k) from the componentwise full conditional of $[\beta_j|\beta_j > 0, \beta_k = 0]$ (resp. $[\beta_k|\beta_j = 0, \beta_k > 0]$), and the last component generates new regression coefficients from the componentwise full conditional of $[(\beta_j, \beta_k)|\beta_j > 0, \beta_k > 0]$. See Appendix A.2 for details. Similarly, the full conditional of more than two β's can be obtained but the number of components grows exponentially with the numbers of β's jointly updated.

Example 4: Three-way comparison in gene expression data. We now consider an application that arises in the analysis of gene expression microarray data. DNA microarrays allow the monitoring of thousands of genes simultaneously under different biological or experimental conditions. One of the main tasks with microarrays is the identification of genes that are expressed differentially under the different conditions. Hedenfalk et al. (2001) con-

Table 4. Estimated posterior weights, that is, posterior means of the ϖ's, associated with each observation of the Stack Loss Data. Observations with small weights are downweighted. Observations 1, 3, 4, and 21 have smaller weights, suggesting that they might be outliers.

Obs.	1	2	3	4	5	6	7	8	9	10	11
Weight	0.78	1.08	0.74	0.59	1.13	1.05	1.09	1.16	1.08	1.16	1.15
Obs.	12	13	14	15	16	17	18	19	20	21	
Weight	1.12	0.96	1.01	1.14	1.21	1.19	1.20	1.15	1.04	0.49	

Table 5. Log transformed measurements of one gene of the BRCA dataset. Each row corresponds to a different condition, and each row entry to a different tumor sample.

BRCA1	BRCA2	SPORADIC
−2.74	−1.51	1.47
−2.18	0.14	−0.81
−1.74	0.10	−1.69
−1.94	0.55	−1.06
0.29	−0.45	−1.32
−1.18	−0.67	−2.00
−1.40	−0.38	−1.18
	−0.60	

ducted a study to examine breast cancer tissues from patients carrying mutations in the predisposing genes, BRCA1 or BRCA2, or from patients not expected to carry a hereditary mutation. They examined 22 breast cancer tumor samples: 7 tumors with BRCA1, 8 tumors with BRCA2 and 7 SPORADIC tumors, that is, tumors with neither mutation. The goal of the experiment was to study the expression patterns of 3,226 genes under the three conditions and to detect genes whose expression changed in at least one of the conditions. For illustrative purposes, we show the results for one gene. A complete analysis would require fitting a similar model to each gene; this was done by Gottardo et al. (2006) in more general settings of gene expression experiments. The measurements for this gene in the three samples are given in Table 5.

In order to detect differential expression we consider the following model:

$$\begin{aligned} y_{ci} &= \mu_c + \epsilon_{ci}, \\ (\epsilon_{ci} | \psi_{\epsilon_c}) &\sim \mathrm{N}(0, \psi_{\epsilon_c}^{-1}), \end{aligned}$$

where μ_c represents the mean expression level of the gene under condition c, $i = 1, \ldots, n_c$ and $c = 1, 2, 3$ (BRCA1, BRA2, SPORADIC). We wish to test the null hypothesis $\mu_1 = \mu_2 = \mu_3$. In this example the alternative hypothesis is more complex due to the number of possible patterns. The prior distribution needs to include all such possible patterns. We therefore consider the following prior, whose density is

$$\begin{aligned} (\boldsymbol{\mu} | \boldsymbol{\psi}_{\boldsymbol{\mu}}, \mathbf{w}) \sim\ & w_1 \mathrm{N}(\mu_1 | 0, \psi_{\mu_{123}}^{-1}) \mathbf{1}_{[\mu_1 = \mu_2 = \mu_3]} \\ & + w_2 \mathrm{N}(\mu_1 | 0, \psi_{\mu_1}^{-1}) \mathrm{N}(\mu_2 | 0, \psi_{\mu_{23}}^{-1}) \mathbf{1}_{[\mu_1 \neq \mu_2 = \mu_3]} \\ & + w_3 \mathrm{N}(\mu_2 | 0, \psi_{\mu_2}^{-1}) \mathrm{N}(\mu_1 | 0, \psi_{\mu_{13}}^{-1}) \mathbf{1}_{[\mu_1 = \mu_3 \neq \mu_2]} \\ & + w_4 \mathrm{N}(\mu_3 | 0, \psi_{\mu_3}^{-1}) \mathrm{N}(\mu_1 | 0, \psi_{\mu_{12}}^{-1}) \mathbf{1}_{[\mu_1 = \mu_2 \neq \mu_3]} \\ & + w_5 \mathrm{N}(\mu_1 | 0, \psi_{\mu_1}^{-1}) \mathrm{N}(\mu_2 | 0, \psi_{\mu_2}^{-1}) \mathrm{N}(\mu_3 | 0, \psi_{\mu_3}^{-1}) \mathbf{1}_{[\mu_1 \neq \mu_2 \neq \mu_3]}, \end{aligned} \tag{4.5}$$

where $\boldsymbol{\psi}_{\boldsymbol{\mu}} = (\psi_{\mu_1}, \psi_{\mu_2}, \psi_{\mu_3}, \psi_{\mu_{12}}, \psi_{\mu_{13}}, \psi_{\mu_{23}}, \psi_{\mu_{123}})$ is the vector of precisions and $\mathbf{w}$ is the vector of probabilities for the five patterns constrained to sum to one. This defines a proper distribution with respect to the following σ-finite measure on $\mathbb{R}^3$, namely

$$\nu(\cdot) = \lambda_1(\Delta \cap \cdot) + \lambda_2(P_{\mu_1} \cap \cdot) + \lambda_2(P_{\mu_2} \cap \cdot) + \lambda_2(P_{\mu_3} \cap \cdot) + \lambda_3(\cdot),$$

where Δ is the line $\mu_1 = \mu_2 = \mu_3$, P_{μ_1} is the plane $\mu_2 = \mu_3$, P_{μ_2} is the plane $\mu_1 = \mu_3$, P_{μ_3} is the plane $\mu_1 = \mu_2$ and λ_k denotes the k-dimensional Lebesgue measure. In this case, we defined the distribution directly in terms of the density given by (4.5). The target distribution is $\pi(\boldsymbol{\mu}, \psi|\mathbf{y})$, but as before the update for ψ does not present anything complicated and we are thus only concerned with $\boldsymbol{\mu}$.

This time the prior density for $\boldsymbol{\mu}$ is a mixture of five singular Gaussian distributions, and given the Gaussian likelihood, the full conditional can once again be derived explicitly. Using (3.3), we need to derive each componentwise full conditional π_i, $i = 1, \dots, 5$. Here we show only how to derive π_2; other π_i's can be obtained in a similar fashion. For the second component, given that $\mu_1 \neq \mu_2 = \mu_3$, we have

$$\begin{aligned}\pi_2(\boldsymbol{\mu}|\mathbf{y}, \psi) &\propto L(\mathbf{y}|\mu_1, \mu_2, \psi)\frac{\sqrt{\psi_{\mu_1}\psi_{\mu_{23}}}}{2\pi}\exp\left(-\frac{\psi_{\mu_1}}{2}\mu_1^2 - \frac{\psi_{\mu_{23}}}{2}\mu_2^2\right)\\ &\propto \mathrm{N}(\mu_1|\mu_1^*, \psi_1^{*-1})\mathrm{N}(\mu_2|\mu_{23}{}^*, {\psi_{23}^*}^{-1}),\end{aligned}$$

where the normalizing constant c_2 is given by

$$c_2 = \int_{\mu_1}\int_{\mu_2} L(\mathbf{y}|\mu_1, \mu_2, \psi)\sqrt{\psi_{\mu_1}\psi_{\mu_{23}}}/(2\pi)\exp(-0.5\psi_{\mu_1}\mu_1^2 - 0.5\psi_{\mu_{23}}\mu_2^2)d\mu_1 d\mu_2,$$

and the parameters μ_1^*, ψ_1^* and $\mu_{23}{}^*$, ψ_{23}^* are defined in Appendix A.3. Note that μ_3 does not appear in the likelihood as it is equal to μ_2. Similarly, each componentwise full conditional can be obtained and the full conditional is again a mixture of five mutually singular Gaussian distributions with new means, precisions and proportions; see Appendix A.3.

Thus, one can simply generate a new $\boldsymbol{\mu}$ by first selecting a new component at random according to the new proportions, and then simulating $\boldsymbol{\mu}$ from the corresponding Gaussian distribution while imposing the respective constraint. For example, if the second component is selected, then new values for μ_1, μ_2 are generated from the Gaussian distribution above while μ_3 is set equal to μ_2.

In nonconjugate settings the full conditional might not be available and the Metropolis–Hastings algorithm is an alternative. Here, it is harder to construct an efficient proposal for the Metropolis–Hastings algorithm than in the last example because of the greater number of mutually singular components. To try to maximize the between-component acceptance rate, we used a proposal based on local moves. The local move structure is described by the graph given in Figure 1. We use a proposal of the form given by Equations (3.4), with $p_{ij} > 0$ if there is an edge between i and j (Figure 1). We consider only one kernel, which is similar to the kernel K_2 of Examples 2 and 3, adapted to the local move structure.

From the current value of $\boldsymbol{\mu}$, a new component is randomly chosen from among all the accessible components, based on Figure 1. Given the new component, new values are generated for $\boldsymbol{\mu}$ from a Gaussian proposal centered at the least squares estimates for the corresponding component. The width of the proposal was taken to be the same for all the Gaussian proposals, and its value was chosen by maximizing the proportion of moves between components. For example, given that the current $\boldsymbol{\mu}$ is from the first component, that is, $\mu_1 = \mu_2 = \mu_3$, we first select one of the three accessible components with probability 1/3 each. Then we generate new values for $\boldsymbol{\mu}$, independently of the current values, from

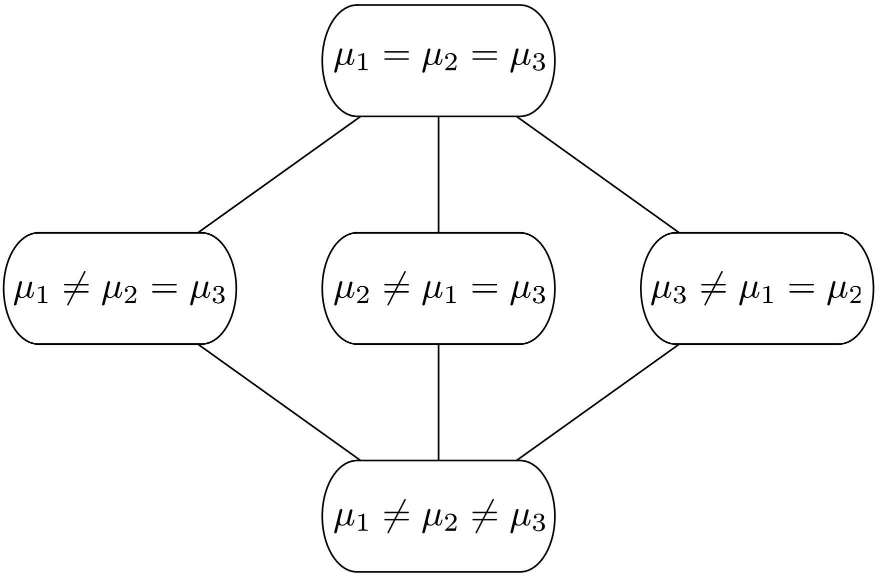
$\mu_1 = \mu_2 = \mu_3$
$\mu_1 \neq \mu_2 = \mu_3$
$\mu_2 \neq \mu_1 = \mu_3$
$\mu_3 \neq \mu_1 = \mu_2$
$\mu_1 \neq \mu_2 \neq \mu_3$

Figure 1. Local move graph for the three-way comparison proposal.

a Gaussian proposal centered at the constrained least squares estimates. For example, if the second component is chosen, then the constraint is $\mu_2 = \mu_3$. We have chosen such an independence proposal for simplicity, but it would be possible to derive more efficient proposals. As with K_2, within a given component of the mixture each parameter is updated using a componentwise Gibbs step based on the full conditional for that component.

Similarly to the two previous examples, the model posterior probabilities can be easily estimated from the MCMC output as the proportion of time spent in the corresponding model/component. As before, in the case of the Gibbs sampler one can obtain a more efficient estimate by Rao–Blackwellization when averaging the w^* values at each iteration, as given in Appendix A.3. Table 6 summarizes the estimates of the posterior probabilities for each component using the Metropolis–Hastings algorithm and the Gibbs sampler. The estimates of the probabilities agree well but the standard deviations are much smaller for the Gibbs sampler. The average number of moves between components is much greater for the Gibbs sampler than for Metropolis–Hastings, indicating better mixing. Additionally, the Gibbs sampler required less computing time.

Note that we have chosen to update the components of $\boldsymbol{\mu}$ jointly, both for efficiency and to give an example of a block Gibbs update. However, it would be possible to update each parameter component in turn as in Example 3.

5. RELATIONSHIP WITH THE REVERSIBLE JUMP SAMPLER

Green (1995) introduced a Markov chain Monte Carlo method for Bayesian model determination for the situation where the dimensionality of the parameter vector is not fixed.

Table 6. Comparison of the model posterior probabilities for the five models computed with each algorithm in Example 4. The estimates were computed from 10,000 iterations with 1,000 burn-in iterations. The standard deviations were computed by dividing a chain of 10 million iterations into 1,000 batches of 10,000 iterations each. The truth was estimated from 10 million iterations on the basis of which the estimates from the three algorithms agreed to within three digits. PMBC (percent of moves between components) is the percentage of time the chain moved from one component to an other. The estimates of the posterior probabilities agree well. The standard deviations are smaller for the Gibbs sampler. RB refers to estimates computed via Rao-Blackwellization. For ease of comparison and clarity, all standard deviations are multiplied by 10^4.

	Truth	MH		Gibbs				RJ	
		Est.	sd	Est.	sd	RB-Est.	RB-sd	Est.	sd
$\Pi(\mu_1 = \mu_2 = \mu_3 \mid \mathbf{y})$	0.085	0.089	90	0.087	30	0.087	20	0.089	90
$\Pi(\mu_1 \neq \mu_2 = \mu_3 \mid \mathbf{y})$	0.023	0.027	40	0.023	20	0.021	4	0.025	30
$\Pi(\mu_2 \neq \mu_2 = \mu_3 \mid \mathbf{y})$	0.850	0.826	130	0.847	40	0.851	20	0.843	120
$\Pi(\mu_1 = \mu_2 \neq \mu_3 \mid \mathbf{y})$	0.007	0.009	20	0.008	10	0.007	5	0.008	10
$\Pi(\mu_1 \neq \mu_2 \neq \mu_3 \mid \mathbf{y})$	0.035	0.050	100	0.038	20	0.035	0.3	0.034	40
PMBC (%)		6		24				8	
CPU (μs/iter)		12		7				13	

Following the notation of Green (1995), we assume that we have a countable collection of models, $\{\mathcal{M}_k : k \in \mathcal{K}\}$. Model $\mathcal{M}_k$ has a vector $\boldsymbol{\theta}_k$ of unknown parameters assumed to lie in $\mathbb{R}^{n_k}$, where the dimension n_k may vary from model to model. Bayesian inference about k and $\boldsymbol{\theta}_k$ is based on the joint posterior $\pi(k, \boldsymbol{\theta}_k|\mathbf{y})$, which can be decomposed as

$$\pi(k, \boldsymbol{\theta}_k|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta}_k, k)\pi(\boldsymbol{\theta}_k|k)p(k).$$

Using this formulation, the sample space can be represented by $S = \cup_{k \in \mathcal{K}} \{k\} \times \mathbb{R}^{n_k}$. Let $\mathbf{x} = (k, \boldsymbol{\theta}_k)$, and let $\pi(\mathbf{x}) \equiv \pi(\mathbf{x}|\mathbf{y})$ denote the the target distribution. Even though the dimension of $\mathbf{x}$ is allowed to change, Green (1995) showed that it is still possible to use the Metropolis–Hastings algorithm to form an irreducible and aperiodic Markov chain with stationary distribution π.

We now describe the reversible jump method in terms of random numbers as described in a more recent paper (Green 2003). At some current state $\mathbf{x}$, we generate r random numbers $\mathbf{u}$ from a known joint density g, and then form the proposed new state as some suitable deterministic function of the current state and the random numbers: $\mathbf{x}' = h(\mathbf{x}, \mathbf{u})$. The reverse transformation will be made with the aid of random numbers $\mathbf{u}'$ generated from some probability distribution g', giving $\mathbf{x} = h'(\mathbf{x}', \mathbf{u}')$. Assuming that the transformation from $(\mathbf{x}, \mathbf{u})$ to $(\mathbf{x}', \mathbf{u}')$ is a diffeomorphism, Green (1995) showed that a valid choice for the acceptance probability in the usual Metropolis–Hastings algorithm is given by

$$\alpha(\mathbf{x}, \mathbf{x}') = \min\left\{1, \frac{\pi(\mathbf{x}')g'(\mathbf{u}')}{\pi(\mathbf{x})g(\mathbf{u})}\left|\frac{\partial(\mathbf{x}', \mathbf{u}')}{\partial(\mathbf{x}, \mathbf{u})}\right|\right\}, \tag{5.1}$$

where $\pi(\mathbf{x}) \equiv \pi(k, \boldsymbol{\theta}_k|\mathbf{y})$.

The reversible jump formulation was introduced to handle cases where the dimension of the parameter vector $\boldsymbol{\theta}$ can change from model to model. In Section 3, we showed how reversible jump can also be viewed as a general Metropolis–Hastings algorithm where one fixes the number of parameters to be the same and allows some of the parameters to vanish or to lie in a hyperplane, reducing the dimension of the support. One example of this is provided by nested models with linear constraints on the parameters, and MCMC with mutually singular distributions could be applied to that case. This is also the case in Example 2, as we will show below. The Jacobian term present in (5.1) results from the change of variable induced by the diffeomorphism when a new value is proposed. If one designs a move that involves a change of variable, one should include the Jacobian term in (5.1) for the detailed balance condition to be satisfied (Green 1995).

Thinking about the problem in terms of mutually singular distributions allows us to use standard MCMC algorithms without worrying about the dimension matching. For example, we can use the usual Gibbs sampler when the full conditional is available. This is not possible with the reversible jump formulation, as pointed out by Green (1995) and Robert and Casella (1999, p. 287). We have shown that if one considers the right dominating measure, it is easy to establish that the Gibbs sampler is irreducible.

Example 2: Testing a normal mean (continued). In this example we have two competing models, $\mathcal{M}_0 : \mu = 0$ and $\mathcal{M}_1 : \mu \neq 0$. For the first model $\theta_0 = (\psi)$, and for the second model we have one more parameter, $\boldsymbol{\theta}_1 = (\mu, \psi)$. In the notation of Green (1995), the sample space is $S = \{0\} \times \mathbb{R} \cup \{1\} \times \mathbb{R}^2$. As ψ is common to the two models, we shall be concerned only with the update of $\boldsymbol{\theta}$ conditional on ψ. At the current state $\mathbf{x} = (k, \boldsymbol{\theta}_k)$, we can generate a new value $\mathbf{x}'$ according to

$$\mathbf{x}' = \begin{cases} 0 & \text{if} \quad k = 1 \\ (1, u) & \text{if} \quad k = 0, \end{cases}$$

where u is a random deviate with distribution g. In this case the acceptance probability reduces to

$$\alpha(\mathbf{x}, \mathbf{x}') = \begin{cases} 1 \wedge r(\mathbf{x}, \mathbf{x}') & \text{if} \quad k = 1 \\ 1 \wedge r(\mathbf{x}', \mathbf{x})^{-1} & \text{if} \quad k = 0, \end{cases} \tag{5.2}$$

where

$$r(\mathbf{x}, \mathbf{x}') = \frac{f(\mathbf{y}|\psi)(1-w)g(\mu)}{f(\mathbf{y}|\mu, \psi)\mathrm{N}(\mu|0, \psi_\mu^{-1})w}. \tag{5.3}$$

On the other hand, following our framework, we could use the prior mixture distribution given by (4.2), whose density is $(1 - w)\mathbf{1}_{\{0\}} + w\mathrm{N}(\mu|0, \psi_\mu^{-1})(1 - \mathbf{1}_{\{0\}})$, and the proposal

$$q(\mu, \mu') = \begin{cases} 1_{\{0\}}(\mu') & \text{if} \quad \mu \neq 0 \\ g(\mu') & \text{if} \quad \mu = 0, \end{cases}$$

which is absolutely continuous with respect to $(\delta_0 + \lambda)$. It is easy to see that the acceptance probability is the same as the one given by (5.2); the two formulations are equivalent. Thinking about the problem with a common dominating measure allows us to use the usual Metropolis–Hastings algorithm and the Gibbs sampler.

Usually, in reversible jump algorithms, there are two main types of moves: between-model moves and within-model moves. Even though the within-model moves are not always necessary for the algorithm to be ergodic, they can greatly improve the performance of the sampler. The Metropolis–Hastings algorithms used in Examples 2 and 3 could be seen as reversible jump algorithms, where q_{21} is the between model proposal, q_{22} is the within model proposal, and $g(u)$ is $\mathrm{N}(u; \hat{\mu}, \sigma^2)$ as in the definition of q_{21}. In Example 2, where it was easy to construct an efficient proposal, the reversible jump performed slightly better than the Gibbs sampler (with mutually singular distributions) in terms of mixing, but took substantially more computer time for the same number of iterations and was not as accurate as Gibbs with Rao–Blackwellization. Overall, for the same amount of computer time, the Gibbs sampler was more efficient. In Example 3, reversible jump performed relatively poorly compared to the Gibbs sampler, with much bigger standard deviations and much more computer time. We now turn back to Example 4 where it might be possible to design better moves using the reversible jump formulation.

Example 4: Three-way comparison (continued). In this example we have five competing models,

$$\begin{aligned}
\mathcal{M}_1 &: \mu_1 = \mu_2 = \mu_3\\
\mathcal{M}_2 &: \mu_1 \neq \mu_2 = \mu_3\\
\mathcal{M}_3 &: \mu_1 = \mu_3 \neq \mu_2\\
\mathcal{M}_4 &: \mu_1 = \mu_2 \neq \mu_3\\
\mathcal{M}_5 &: \mu_1 \neq \mu_2 \neq \mu_3,
\end{aligned}$$

which correspond to the five components given by the nodes of the local move graph (Figure 1). In a reversible jump framework, each model would be viewed as having a different number of parameters: the first model has one parameter, the second has two parameters, and so on. The parameter vectors for the five models can be written as

$$\begin{aligned}
\boldsymbol{\theta}_1 &= (\mu_{123})\\
\boldsymbol{\theta}_2 &= (\mu_1, \mu_{23})\\
\boldsymbol{\theta}_3 &= (\mu_2, \mu_{13})\\
\boldsymbol{\theta}_4 &= (\mu_3, \mu_{12})\\
\boldsymbol{\theta}_5 &= (\mu_1, \mu_2, \mu_3),
\end{aligned}$$

and the sample space is given by $S = \cup_i \{i\} \times \boldsymbol{\Theta}_i$, where $\boldsymbol{\Theta}_1 = \mathbb{R}$, $\boldsymbol{\Theta}_2 = \boldsymbol{\Theta}_3 = \boldsymbol{\Theta}_4 = \mathbb{R}^2$ and $\boldsymbol{\Theta}_5 = \mathbb{R}^3$.

Similarly to Example 3, if g as used in (5.1) is the proposal used in the Metropolis–Hastings algorithm with mutually singular distribution, that is, Gaussian centered at the least squares estimates, one can show that the acceptance probability of the reversible jump is the same as in the Metropolis-Hastings algorithm, and so the two algorithms are equivalent in this case.

One of the strengths of reversible jump is the ability to design elaborate moves, which might lead to the inclusion of a Jacobian term in the acceptance ratio. In this example, we use the common split-merge move (Richardson and Green 1997), where some parameters are merged to form a new one. Again, we use the local move graph to jump from one model to the other.

A move between model 1 and model 2 would be made as follows,

$$\mathbf{x}' = \begin{cases} (1, \frac{\mu_1+\mu_{23}}{2}) & \text{if} \quad k=2 \\ (2, \mu_{123}-u, \mu_{123}+u) & \text{if} \quad k=1, \end{cases}$$

where u is a random deviate with distribution g. In this case the acceptance probability reduces to

$$\alpha(\mathbf{x},\mathbf{x}') = \begin{cases} 1 \wedge r(\mathbf{x},\mathbf{x}') & \text{if} \quad k=2 \\ 1 \wedge r(\mathbf{x},\mathbf{x}')^{-1} & \text{if} \quad k=1, \end{cases}$$

where

$$r(\mathbf{x},\mathbf{x}') = \frac{f(\mathbf{y}|\psi,\mu_{123})w_1\mathrm{N}(\mu_{123}|0,\psi^{-1}_{\mu_{123}})}{f(\mathbf{y}|\psi,\mu_1,\mu_{23})w_2\mathrm{N}(\mu_1;0,\psi^{-1}_{\mu_1})\mathrm{N}(\mu_{23}|0,\psi^{-1}_{\mu_{23}})}\,\frac{p_{12}g((\mu_1-\mu_{23})/2)}{p_{21}}\,\frac{1}{2},$$

and p_{ij} is the probability of proposing a move from model i to model j. Note the Jacobian term in the acceptance ratio due to the change of variable induced by the merge move. The acceptance ratio for a move between 1 and 3 (or 4) would be the same with obvious changes in notation.

Similarly, a move between 2 and 5 would be as follows:

$$\mathbf{x}' = \begin{cases} (2, \mu_1, \frac{\mu_2+\mu_3}{2}) & \text{if} \quad k=5 \\ (5, \mu_1, \mu_{23}-u, \mu_{23}+u) & \text{if} \quad k=2, \end{cases}$$

where u is a random deviate with distribution g. In this case the acceptance probability reduces to

$$\alpha(\mathbf{x},\mathbf{x}') = \begin{cases} 1 \wedge r(\mathbf{x},\mathbf{x}') & \text{if} \quad k=5 \\ 1 \wedge r(\mathbf{x},\mathbf{x}')^{-1} & \text{if} \quad k=2, \end{cases}$$

where

$$r(\mathbf{x},\mathbf{x}') = \frac{f(\mathbf{y}|\psi,\mu_1,\mu_{23})w_2\mathrm{N}(\mu_1|0,\psi^{-1}_{\mu_1})\mathrm{N}(\mu_{23}|0,\psi^{-1}_{\mu_{23}})}{f(\mathbf{y}|\psi,\mu_1,\mu_2,\mu_3)w_5\mathrm{N}(\mu_1|0,\psi^{-1}_{\mu_1})\mathrm{N}(\mu_2|0,\psi^{-1}_{\mu_2})\mathrm{N}(\mu_3|0,\psi^{-1}_{\mu_3})} \times \frac{p_{25}g((\mu_2-\mu_3)/2)}{p_{52}}\,\frac{1}{2}.$$

The acceptance ratio for a move between 5 and 3 (or 4) would be the same with obvious changes in notation. Here, we choose g to be a Gaussian distribution with mean 0 and variance σ^2, with σ^2 chosen to maximize the proportion of moves between components. As with traditional reversible jump algorithms, within a given model each parameter is updated using a Gibbs step (as full conditionals are available). Table 6 shows the results obtained with the merge-split algorithm (last two columns). The performance of the algorithm is better than that of the Metropolis–Hastings algorithm with a proposal of the form (3.4). However, the improvement is not great and the Gibbs sampler with mutually singular distributions still does far better with much less computing time.

6. DISCUSSION

We have introduced a framework for the use of MCMC algorithms with mixtures of mutually singular distributions. We showed how one can use the usual Metropolis–Hastings algorithm to form an ergodic chain with stationary distribution $\Pi = \sum_{i \in I} w_i \Pi_i$ where the Π_i's are mutually singular distributions. We have analyzed three examples in which the method was easy to apply.

However, because of the singularity between the different components, the choice of a good proposal is harder than in the usual setting. The same problem arises with the reversible jump formulation and there has been a great deal of work on efficient construction of reversible jump proposal distributions (Green and Mira 2001; Brooks et al. 2003). Using a simple example we have shown the relationship between our formulation and the reversible jump formulation. Indeed it was possible to derive an algorithm with the same acceptance probability. Which formulation is to be preferred will depend on the application. Our formulation is convenient in the sense that the number of parameters remains the same and we do not have to worry about dimension matching.

When full conditionals are available, the Gibbs sampler can be used and can bring great improvement. Using the Gibbs sampler, no tuning is necessary, which can be a considerable advantage in problems where the number of parameters is large. This was illustrated in the gene expression problem, Example 4. In that example we used only one gene, but in practice one would want to use the same algorithm for thousands of (conditionally independent) genes. Both computation and tuning would be a serious problem. In some cases where the full conditional is not available, it might be available for a distribution that approximates the target distribution. The full conditional of the approximate distribution could be used as a proposal in a Metropolis–Hastings algorithm for the target. This could potentially lead to a high acceptance rate; see Gottardo et al. (2006) for an application. In the past few years there has been some progress towards automatic transdimensional algorithms (Green 2003; Hastie 2004), and we believe that the Gibbs sampler with mutually singular distributions is a step in that direction.

It might be hard to write the prior as a mixture of mutually singular distributions in some complex model selection problems where there is a large difference of dimension between the models. Moreover, full conditionals might not be available and good proposals might be hard to derive. The reversible jump formulation allows for clever moves between models by introducing a Jacobian term in the acceptance ratio. For example, in the mixture problem with an unknown number of components, Richardson and Green (1997) used moment matching conditions to move from one model to the other. It would be hard to formulate this problem in terms of mutually singular distributions. Bayesian analysis of mixture models with an unknown number of components is a problem where it is hard to devise moves with high acceptance rates. There has been some effort to try to create algorithms with better properties (Stephens 2000; Cappé et al. 2003; Brooks et al. 2003).

In this article, we have estimated each model posterior probability as the proportion of time spent in that model. This is what is commonly done in transdimensional MCMC. In the case of the Gibbs sampler we have seen that a simple Rao–Blackwellization step can improve the estimates. Other methods have also been proposed for getting better estimates of posterior model probabilities from reversible jump output; see Bartolucci et al. (2006).

Our formulation is different from the product space approach (Carlin and Chib 1995; Besag 1997; Godsill 2001; Dellaportas et al. 2002). In our approach we also keep the number of parameters fixed but we allow the dimension of the support of the distribution to vary. We only need to store as many parameters as in the reversible jump formulation. With our formulation, even though the number of parameters across models is the same, there is some redundancy in the parameters. Instead of varying the number of parameters, we vary the dimension of the support of the distribution. This is different from the product space approach where one needs to introduce pseudo-priors in order to keep the parameters that are not in the model currently visited. Our formulation is also more general than that of Petris and Tardella (2003) as (1) the Gibbs sampler can be used, and (2) it can deal with nonnested models. In the gene expression example, if the prior was restricted to the second ($\mu_1 \neq \mu_2 = \mu_3$) and third ($\mu_2 \neq \mu_3 = \mu_1$) components, their formulation could not be used because the models are not nested, whereas ours could.

A. APPENDIX

A.1 PROOF OF THEOREM 1

Proof: Since $\nu_i \perp \nu_j$ for $i \neq j$, we know that there exist sets S_i such that $\nu_i(S_i^c) = 0$ and $\nu_j(S_i) = 0$. Using the assumption that Π_i is dominated by ν_i, we have by the Radon–Nikodym Theorem,

$$\begin{aligned}\Pi_i(A) &= \int_A \frac{d\Pi_i}{d\nu_i}(x)\nu_i(dx)\\ &= \int_A \frac{d\Pi_i}{d\nu_i}(x)\mathbf{1}_{S_i}(x)\nu_i(dx)\\ &= \int_A \frac{d\Pi_i}{d\nu_i}(x)\mathbf{1}_{S_i}(x)(\sum_{k\in I}\nu_k)(dx).\end{aligned}$$

The result follows from the fact that $\Pi = \sum_{i\in I} w_i \Pi_i$ and the linearity of the integral operator. □

A.2 FULL CONDITIONAL FOR TWO REGRESSION COEFFICIENTS

When two regression coefficients β_j, β_k, are jointly updated, the prior used is of the form,

$$\begin{aligned}\Pi(\beta_j, \beta_k|\sigma_\beta^2) &= [(1-w_j)\delta_0 + w_j\mathrm{N}(0, \psi_{\beta_j}^{-1})][(1-w_k)\delta_0 + w_k\mathrm{N}(0, \psi_{\beta_k}^{-1})]\\ &= (1-w_j)(1-w_k)\delta_0\delta_0\\ &\quad +(1-w_j)w_k\delta_0\mathrm{N}(0, \psi_{\beta_k}^{-1})\\ &\quad +w_j(1-w_k)\mathrm{N}(0, \psi_{\beta_j}^{-1})\delta_0\\ &\quad +w_j w_k\mathrm{N}(0, \psi_{\beta_j}^{-1})\mathrm{N}(0, \psi_{\beta_k}^{-1}),\end{aligned}$$

where $\psi_{\beta_j}^{-1} = S_Y^2/S_{X_j}^2\sigma_\beta^2$. Thus we can write the prior as a mixture of four mutually singular distributions with respective weights $u_1 = (1-w_j)(1-w_k)$, $u_2 = (1-w_j)w_k$, $u_3 = w_j(1-w_k)$ and $u_4 = w_j w_k$. Again, given the full conjugacy of the prior to the likelihood, the full conditional can be derived explicitly and is of the form,

$$\begin{aligned}(\boldsymbol{\beta}_{j,k}|\ldots) \propto\ & u_1 c_1 \delta_0 \delta_0 \\ & + u_2 c_2 \mathrm{N}\left(\mu_j^*, {\psi_j^*}^{-1}\right)\delta_0 \\ & + u_3 c_3 \delta_0 \mathrm{N}\left(\mu_k^*, {\psi_k^*}^{-1}\right) \\ & + u_4 c_4 \mathrm{N}_2\left(\boldsymbol{\mu}_{j,k}^*, {\boldsymbol{\Psi}_{j,k}^*}^{-1}\right),\end{aligned} \tag{A.1}$$

where $\psi_j^* = (\psi \sum_i \varpi_i X_{ij}^2 + \psi_{\beta_j})^{-1}$ and $\mu_j^* = {\psi_j^*}^{-1}(\psi \sum_i r_{ij})$, similarly for ψ_k^* and μ_k^* replacing the index j with a k. Also

$$r_{ij} = \varpi_i(y_i - \beta_0 - \sum_{l \neq j,k} \beta_l X_{il}) X_{ij},$$

$$c_1 = \left(\frac{\psi}{2\pi}\right)^{n/2} \exp\left(-0.5\psi \sum_i \varpi_i (y_i - \beta_0 - \sum_{l\neq j,k} \beta_l X_{il})^2\right),$$

$$c_2 = c_1 \sqrt{\frac{\psi_{\beta_j}}{\psi_j^*}} \exp\left(0.5{\psi_j^*}^{-1}(\psi \sum_i r_{ij})^2\right),$$

$$c_3 = c_1 \sqrt{\frac{\psi_{\beta_k}}{\psi_k^*}} \exp\left(0.5{\psi_k^*}^{-1}(\psi \sum_i r_{ik})^2\right),$$

$$\boldsymbol{\Psi}_{j,k}^* = \mathbf{D}_{j,k} + \psi \tilde{\mathbf{X}}_{j,k}^t \tilde{\mathbf{X}}_{j,k},$$

$$\boldsymbol{\mu}_{j,k}^* = {\boldsymbol{\Psi}_{j,k}^*}^{-1} \tilde{\mathbf{X}}_{j,k}^t \mathbf{R}_{j,k},$$

$$c_4 = c_1 \sqrt{\frac{\psi_{\beta_j}\psi_{\beta_k}}{|\boldsymbol{\Psi}_{j,k}^*|}} \exp\left(-0.5\psi \tilde{\mathbf{R}}_{j,k}^t \tilde{\mathbf{X}}_{j,k} {\boldsymbol{\Psi}_{j,k}^*}^{-1} \tilde{\mathbf{X}}_{j,k}^t \tilde{\mathbf{R}}_{j,k}\right),$$

where $R_{j,k} = \sqrt{\psi}(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}_{-\{j,k\}}^t \boldsymbol{\beta}_{-\{j,k\}})$, $\mathbf{D}_{j,k} = \mathrm{diag}(\psi_{\beta_j}, \psi_{\beta_k})$, $\tilde{\mathbf{X}}_{j,k}$ is the matrix formed with the jth and kth columns of $\tilde{\mathbf{X}}$, $\tilde{\mathbf{X}}_{-\{j,k\}}$ is the matrix formed by removing the jth and kth columns of $\tilde{\mathbf{X}}$, and finally $\tilde{\mathbf{Y}}$ (resp. $\tilde{\mathbf{X}}$) is obtained by multiplying each element (resp. each row) by the corresponding $\sqrt{\varpi}_i$. Then, the full conditional is obtained by normalizing (A.1) so that the normalized weights, the w^*'s, sum up to one. In this case, the global dominating measure can be written as $\nu = \delta_0 \times \delta_0 + \delta_0 \times \lambda + \lambda \times \delta_0 + \lambda \times \lambda$, where $\nu_1 \times \nu_2$ represents the product measure of ν_1 with ν_2.

A.3 FULL CONDITIONAL IN THE THREE-WAY COMPARISON

In Example 4, the full conditional is again available, and is given by

$$
\begin{aligned}
(\boldsymbol{\mu}|\boldsymbol{\psi}_{\boldsymbol{\mu}}, \mathbf{w}) \propto\ & w_1 k_{123} \mathrm{N}(\mu_1|\mu_{123}^*, {\psi_{123}^*}^{-1}) \mathbf{1}_{[\mu_1=\mu_2=\mu_3]} \\
& + w_2 k_1 k_{23} \mathrm{N}(\mu_1|\mu_1^*, \psi_1^{*-1}) \mathrm{N}(\mu_2|{\mu_{23}}^*, {\psi_{23}^*}^{-1}) \mathbf{1}_{[\mu_1\neq\mu_2=\mu_3]} \\
& + w_3 k_2 k_{13} \mathrm{N}(\mu_2|\mu_2^*, \psi_2^{*-1}) \mathrm{N}(\mu_1|{\mu_{13}}^*, {\psi_{13}^*}^{-1}) \mathbf{1}_{[\mu_1=\mu_3\neq\mu_2]} \\
& + w_4 k_3 k_{12} \mathrm{N}(\mu_3|\mu_3^*, \psi_3^{-1}) \mathrm{N}(\mu_1|\mu_{12}^*, {\psi_{12}^*}^{-1}) \mathbf{1}_{[\mu_1=\mu_2\neq\mu_3]} \qquad \text{(A.2)} \\
& + w_5 k_1 k_2 k_3 \mathrm{N}(\mu_1|\mu_1^*, \psi_1^{*-1}) \mathrm{N}(\mu_2|\mu_2^*, \psi_2^{*-1}) \mathrm{N}(\mu_3|\mu_3^*, \psi_3^{*-1}) \mathbf{1}_{[\mu_1\neq\mu_2\neq\mu_3]}
\end{aligned}
$$

where

$$\psi_c^* = n_c\psi + \psi_{\mu_c}, \quad \mu_c^* = \psi\psi_c^{*-1}\sum_{i=1}^{n_c} y_{ci},$$

$$\psi_{sr}^* = (n_s + n_r)\psi + \psi_{\mu_{sr}}, \quad \mu_{sr}^* = \psi{\psi_{rs}^*}^{-1}(\sum_{i=1}^{n_s} y_{si} + \sum_{i=1}^{n_r} y_{ri}),$$

and

$$\psi_{123}^* = (n_1 + n_2 + n_3)\psi + \psi_{\mu_{123}}, \quad \mu_{123}^* = \psi{\psi_{123}^*}^{-1}\sum_{i,j} y_{ij}.$$

The constants k_i, k_{sr}, k_{123} are given by

$$k_i = \sqrt{\frac{\psi_{\mu_i}}{\psi_i^*}} \exp\left\{-0.5\psi\sum_{j=1}^{n_i} y_{ij}^2 + 0.5\psi_i^{*-1}\left(\psi\sum_{j=1}^{n_i} y_{ij}\right)^2\right\},$$

$$k_{sr} = \sqrt{\frac{\psi_{\mu_{sr}}}{\psi_{sr}^*}} \exp\left\{-0.5\psi\left(\sum_{j=1}^{n_s} y_{sj}^2 + \sum_{j=1}^{n_r} y_{rj}^2\right) + 0.5{\psi_{sr}^*}^{-1}\psi^2\left(\sum_{j=1}^{n_s} y_{sj} + \sum_{j=1}^{n_r} y_{rj}\right)^2\right\}$$

and

$$k_{123} = \sqrt{\frac{\psi_{\mu_{123}}}{\psi_{123}^*}} \exp\left\{-0.5\psi\sum_{i,j} y_{ij}^2 + 0.5{\psi_{123}^*}^{-1}\psi^2\left(\sum_{i,j} y_{ij}\right)^2\right\}.$$

It follows that the full conditional is again a mixture of five mutually singular distributions, whose weights, w^*'s, can be obtained by normalizing (A.2).

ACKNOWLEDGMENTS

The authors thank Julian Besag, Charlie Geyer, Peter Hoff, Matthew Stephens and Jeffrey Rosenthal for helpful discussions, Luke Bornn for proofreading the manuscript, and three anonymous referees and the editor for suggestions that improved an earlier version of the article. This research was supported by NIH Grants 8 R01 EB002137 02 and 1 R01 HD054511 01 A1.

[Received September 2005. Revised December 2007.]

REFERENCES

Atkinson, A. (1985), *Plots, Transformations, and Regression,* Oxford: Clarendon Press.

Bartolucci, F., Scaccia, L., and Mira, A. (2006), "Efficient Bayes Factor Estimation from the Reversible Jump Output," *Biometrika*, 93, 41–52.

Besag, J. (1997), Discussion of "Bayesian Analysis of Mixtures with an Unknown Number of Components" by S. Richardson and P. Green, *Journal of the Royal Statistical Society,* Series B, 59, 774.

Besag, J. E., and Higdon, D. M. (1999), "Bayesian Analysis of Agricultural Field Experiments" (with discussion), *Journal of the Royal Statistical Society*, Series B, 61, 691–746.

Billingsley, P. (1995), *Probability and Measure* (3rd ed.), New York: Wiley.

Brockwell, A. E., and Kadane, J. B. (2005), "Identification of Regeneration Times in MCMC Simulation, with Application to Adaptive Schemes," *Journal of Computational and Graphical Statistics*, 14, 436–458.

Brooks, S., Giudici, P., and Roberts, G. (2003), "Efficient Construction of Reversible Jump MCMC Proposal Distributions," *Journal of the Royal Statistical Society*, Series B, 65, 3–55.

Brownlee, K. (1965), *Statistical Theory and Methodology in Science and Engineering* (2nd ed.), New York: Wiley.

Cappé, O., Robert, C., and Ryden, T. (2003), "Reversible Jump, Birth-and-Death and More General Continuous Time Markov Chain Monte Carlo Samplers," *Journal of the Royal Statistical Society*, Series B, 65, 679–679.

Carlin, B. P., and Chib, S. (1995), "Bayesian Model Choice via Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Series B, 57, 473–484.

Chan, K. S., and Geyer, C. J. (1994), Comment on "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, 22, 1747–1758.

Daniel, C., and Wood, F. (1980), *Fitting Equations to Data,* New York: Wiley.

Dellaportas, P., Forster, J., and Ntzoufras, I. (2002), "On Bayesian Model and Variable Selection Using MCMC," *Statistics and Computing*, 12, 27–36.

Escobar, M. D., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

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

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, 721–742.

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

George, E. I., and McCulloch, R. E. (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–374.

Geweke, J. (1996), "Variable Selection and Model Comparison in Regression," in *Bayesian Statistics 5—Proceedings of the Fifth Valencia International Meeting*, pp. 609–620.

Godsill, S. J. (2001), "On the Relationship Between Markov Chain Monte Carlo Methods for Model Uncertainty," *Journal of Computational and Graphical Statistics*, 10, 230–248.

Gottardo, R., and Raftery, A. E. (2004), "Markov Chain Monte Carlo Computations with Mixture of Singular Distributions," Technical report, Statistics Department, University of Washington.

Gottardo, R., Raftery, A. E., Yeung, K. Y., and Bumgarner, R. (2006), "Bayesian Robust Inference for Differential Gene Expression in Microarrays with Multiple Samples," *Biometrics*, 62, 10–18.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

——— (2003). "Trans-dimensional Markov Chain Monte Carlo," in *Highly Structured Stochastic Systems*, eds. P.J. Green, N. L. Hjort, and S. Richardson, Oxford.

Green, P., and Mira, A. (2001), "Delayed Rejection in Reversible Jump Metropolis–Hastings," *Biometrika*, 4, 1035–1053.

Grenander, U., and Miller, M. I. (1994), "Representations of Knowledge in Complex Systems (Disc: p. 581–603). *Journal of the Royal Statistical Society*, Series B, 56, 549–581.

Hastie, D. (2004), "Towards Automatic Reversible Jump Markov Chain Monte Carlo," PhD thesis, Bristol University.

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

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O., Wilfond, B., Borg, A., and Trent, J. (2001), "Gene-Expression Profiles in Hereditary Breast Cancer," *The New England Journal of Medicine*, 344, 539–548.

Hoeting, J. A., Raftery, A. E., and Madigan, D. (1996), "A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression," *Computational Statistics & Data Analysis*, 22, 251–270.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Madigan, D., and York, J. (1995), "Bayesian Graphical Models for Discrete Data," *International Statistical Review*, 63, 215–232.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M., Teller, A., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, 21, 1087–1091.

Neal, R. M. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

Petris, G., and Tardella, L. (2003), "A Geometric Approach to Transdimensional Markov Chain Monte Carlo," *The Canadian Journal of Statistics*, 31, 469–482.

Phillips, D. B., and Smith, A. F. M. (1995), "Bayesian Model Comparison via Jump Diffusions," in *Markov Chain Monte Carlo in Practice* (chap. 13), New York: Chapman and Hall.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.

Richardson, S., and Green, P. J. (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Components" (disc: P758-792) (corr: 1998v60 p661), *Journal of the Royal Statistical Society*, Series B, 59, 731–758.

Robert, C., and Casella, G. (1999), *Monte Carlo Statistical Methods*, Berlin: Springer Verlag.

Roberts, G. O., and Rosenthal, J. S. (2004), "Harris Recurrence of Metropolis-Within-Gibbs and Trans-dimensional Markov Chains," Technical report, University of Toronto.

Roberts, G. O., and Tweedie, R. L. (1996), "Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis Algorithms," *Biometrika*, 83, 95–110.

Smith, M., and Kohn, R. (1996), "Nonparametric Regression using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–343.

Stephens, M. (2000), "Bayesian Analysis of Mixture Models with an Unknown Number of Components: An Alternative to Reversible Jump Methods," *The Annals of Statistics*, 28, 40–74.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (disc: P1728-1762), *The Annals of Statistics*, 22, 1701–1728.

——— (1998), "A Note on Metropolis–Hastings Kernels for General State Spaces," *Annals of Applied Probability*, 8, 1–9.